# UNIT I

# 1

# INTRODUCTION TO BIG DATA

**Unit Structure**





## 1.0 OBJECTIVES

Irrespective of the size of the enterprise whether it is big or small, data continues to be a precious and irreplaceable asset. Data is present in homogeneous sources as well as in heterogeneous sources. The need of the hour is to understand, manage, process, and take the data for analysis to draw valuable insights. Digital data can be structured, semi-structured or unstructured data.

Data generates information and from information we can draw valuable insight. As depicted in Figure 1.1, digital data can be broadly classified into structured, semi-structured, and unstructured data.

**1. Unstructured data:** This is the data which does not conform to a data model or is not in a form which can be used easily by a computer program. About 80% data of an organization is in this format; for example, memos, chat rooms, PowerPoint presentations, images, videos, letters. researches, white papers, body of an email, etc.

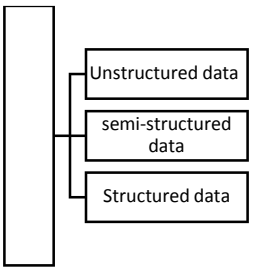



**Figure 1.1 classification of digital data**(Big Data and Analytics)

**2. Semi-structured data:** Semi-structured data is also referred to as self-describing structure. This is the data which does not conform to a data model but has some structure. However, it is not in a form which can be used easily by a computer program. About 10% data of an organization is in this format; for example, HTML, XML, JSON, email data etc.
Figure 1.1 classification of digital data

**3. Structured data:** When data follows a pre-defined schema/structure we say it is structured data. This is the data which is in an organized form (e.g., in rows and columns) and be easily used by a computer program. Relationships exist between entities of data, such as classes and their objects. About 10% data of an organization is in this format. Data stored in databases is an example of structured data.

## 1.1 INTRODUCTION TO BIG DATA

The "Internet of Things" and its widely ultra-connected nature are leading to a burgeoning rise in big data. There is no dearth of data for today's enterprise. On the contrary, they are mired in data and quite deep at that. That brings us to the following questions:

1. Why is it that we cannot forego big data?
2. How has it come to assume such magnanimous importance in running business?
3. How does it compare with the traditional Business Intelligence (BI) environment?
4. Is it here to replace the traditional, relational database management system and data warehouse environment or is it likely to complement their existence?"

Data is widely available. What is scarce is the ability to draw valuable insight.

**Some examples of Big Data:**

- There are some examples of Big Data Analytics in different areas such as retail, ITinfrastructure, and social media.
- Retail: As mentioned earlier, Big Data presents many opportunities to

improve sales andmarketing analytics.

- An example of this is the U.S. retailer Target. After analyzing consumer purchasing behavior, Target's statisticians determined that the retailer made a great deal of money from three main life-event situations.
- Marriage, when people tend to buy many new products
- Divorce, when people buy new products and change their spending habits
- Pregnancy, when people have many new things to buy and have an urgency to buy them. The analysis target to manage its inventory, knowing that there would be demand for specific products and it would likely vary by month over the coming nine- to ten-month cycles
- IT infrastructure: MapReduce paradigm is an ideal technical framework for many Big Data projects, which rely on large data sets with unconventional data structures.
- One of the main benefits of Hadoop is that it employs a distributed file system, meaning it can use a distributed cluster of servers and commodity hardware to process large amounts ofdata.

Some of the most common examples of Hadoop implementations are in the social media space, where Hadoop can manage transactions, give textual updates, and develop social graphs among millions of users.

Twitter and Facebook generate massive amounts of unstructured data and use Hadoop and its ecosystem of tools to manage this high volume.

**Social media:** It represents a tremendous opportunity to leverage social and professionalinteractions to derive new insights.

LinkedIn represents a company in which data itself is the product. Early on, Linkedln founderReid Hoffman saw the opportunity to create a social network for working professionals.

As of 2014, Linkedln has more than 250 million user accounts and has added many additionalfeatures and data-related products, such as recruiting, job seeker tools, advertising, and lnMaps, which show a social graph of a user's professional network.

## 1.2 CHARACTERISTICS OF DATA

As depicted in Figure 1.2, data has three key characteristics:

1. **Composition:** The composition of data deals with the structure of data, that is, the sources of data, the granularity, the types, and the nature of data as to whether it isstatic or real-time streaming.

2. **Condition:** The condition of data deals with the state of data, that is, "Can one use this data as is foranalysis?" or "Does it require cleansing for further enhancement and enrichment?"
3. **Context:** The context of data deals with "Where has this data been generated?" "Why was this datagenerated?" How sensitive is this data?"

"What are the events associated with this data?" and so on.

Small data (data as it existed prior to the big data revolution) is about certainty. It is about known datasources; it is about no major changes to the composition or context of data.

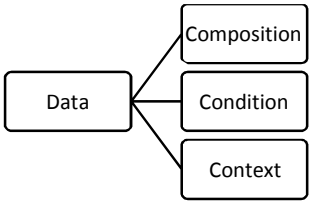



**Figure 1.2 Characteristics of data** (Big Data and Analytics)

Most often we have answers to queries like why this data was generated, where and when it was generated, exactly how we would like to use it, what questions will this data be able to answer, and so on. Big data is about complexity. Complexity in terms of multiple and unknown datasets, in terms of exploding volume, in terms of speed at which the data is being generated and the speed at which it needs to be processed and in terms of the variety of data (internal or external, behavioural or social) that is being generated.

## 1.3 EVOLUTION OF BIG DATA

1970s and before was the era of mainframes. The data was essentially primitive and structured. Relational databases evolved in 1980s and 1990s. The era was of data intensive applications. The World Wide Web (WWW) and the Internet of Things (IOT) have led to an onslaught of structured, unstructured, and multimedia data. Refer Table 1.1.

| | DATA GENERATIONAND STORAGE | DATA UTILZATION | DATA DRIVEN |
|---|---|---|---|
| COMPLEX AND UNSTRUCTURED | | | Structured data, Unstructured data, Multimedia data |
| COMPLEX AND RELATIONAL | | databases: Data-intensive applications | |
| PRIMITIVE AND STRUCTURED | Mainframes: Basic data storage | | |
| | 1970 and before | Relational (1980 and 1990s) | 2000 and beyond |

**Table 1.1 The evolution of big data (Big Data and Analytics)**

## 1.4 DEFINITION OF BIG DATA

- Big data is high-velocity and high-variety information assets that demand cost effective,innovative forms of information processing for enhanced insight and decision making.
- Big data refers to datasets whose size is typically beyond the storage capacity of and alsocomplex for traditional database software tools
- Big data is anything beyond the human & technical infrastructure needed to supportstorage, processing and analysis.
- It is data that is big in volume, velocity and variety. Refer to figure 1.3

**Variety**: Data can be structured data, semi-structured data and unstructured data. Data stored in a database is an example of structured data.HTML data, XML data, email data,

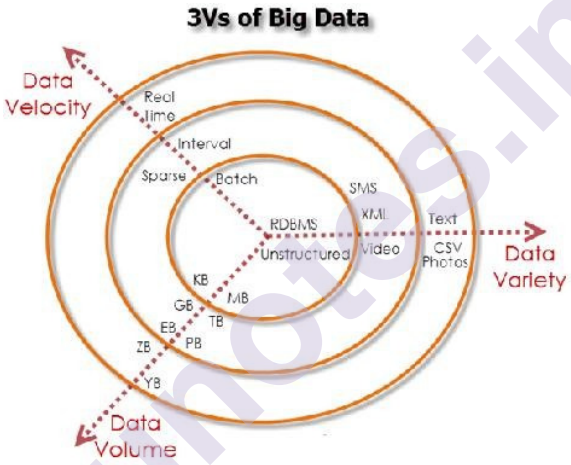



**Figure 1.3 Data: Big in volume, variety, and Velocity** (Big Data and Analytics)

CSV files are the examples of semi-structured data. Power point presentation, images, videos, researches, white papers, body of email etc are the examples of unstructured data.

**Velocity:** Velocity essentially refers to the speed at which data is being created in real- time. We have moved from simple desktop applications like payroll application to real-time processing applications.

**Volume:** Volume can be in Terabytes or Petabytes or Zettabytes.

Gartner Glossary **Big data** is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight and decision making.

For the sake of easy comprehension, we will look at the definition in three parts. Refer Figure 1.4.

**Part I of the definition:** "Big data is high-volume, high-velocity, and high-variety information assets" talks about voluminous data (humongous data) that may have great variety (a good mix of

structured, semi-structured. and unstructured data) and will require a good speed/pacefor storage, preparation, processing and analysis.

**Part II of the definition:** "cost effective, innovative forms of information processing" talks about embracing new techniques and technologies to capture (ingest), store, process, persist, integrate and visualize the high-volume, high-velocity, and high-variety data.

**Part III of the definition:** "enhanced insight and decision making" talks about deriving deeper, richer and meaningful insights and then using these insights to make faster and better decisions to gain business value and thus a competitive edge.
Data —> Information —> Actionable intelligence —> Better decisions —>Enhanced business value

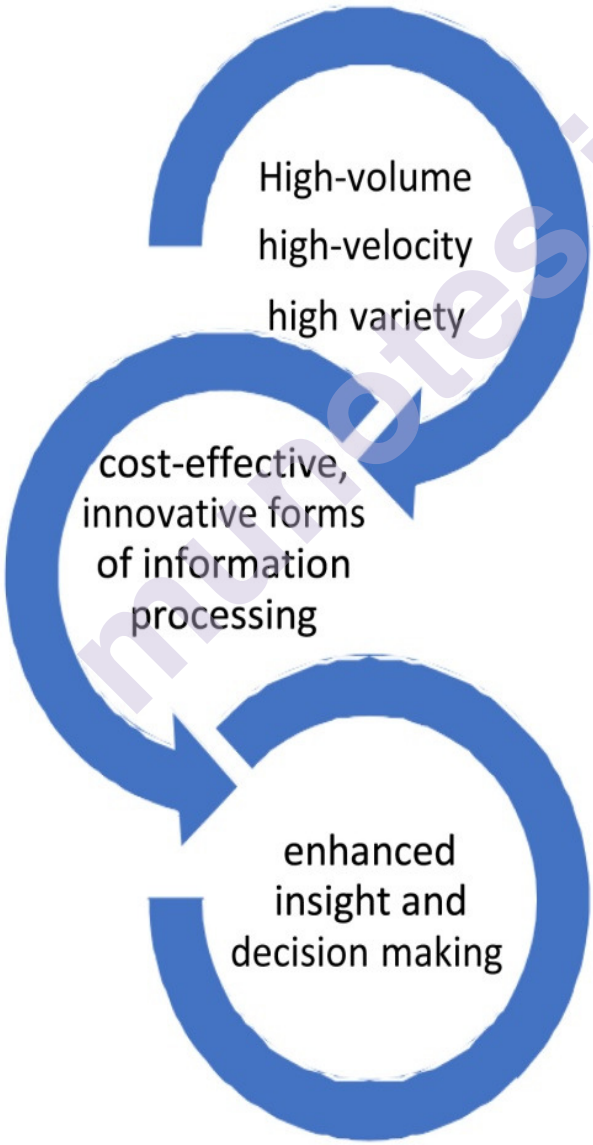



**Figure 1.4 Definition of big data – Gartner** (Big Data and Analytics)

## 1.5 CHALLENGES WITH BIG DATA

Refer figure 1.5. Following are a few challenges with big data:

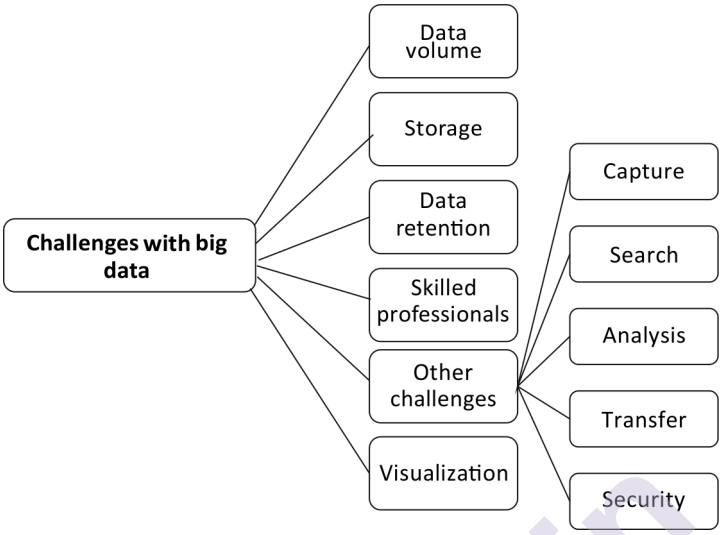



**Figure 1.5 Challenges with big data** (Big Data and Analytics)

**Data volume:** Data today is growing at an exponential rate. This high tide of data willcontinue to rise continuously. The key questions are –

"will all this data be useful for analysis?",

"Do we work with all this data or subset of it?",

"How will we separate the knowledge from the noise?" etc.

**Storage:** Cloud computing is the answer to managing infrastructure for big data as far as cost-efficiency, elasticity and easy upgrading / downgrading is concerned. This further complicates the decision to host big data solutions outside the enterprise.

**Data retention:** How long should one retain this data? Some data may require for log-term decision, but some data may quickly become irrelevant and obsolete.

**Skilled professionals:** In order to develop, manage and run those applications that generate insights, organizations need professionals who possess a high-level proficiencyin data sciences.

**Other challenges:** Other challenges of big data are with respect to capture, storage, search,analysis, transfer and security of big data.

**Visualization:** Big data refers to datasets whose size is typically beyond the storage capacity of traditional database software tools. There is no explicit definition of how bigthe data set should be for it to be considered bigdata. Data visualization(computer graphics) is becoming popular as a separate discipline. There are very few data visualization experts.

## 1.6 WHY BIG DATA?

The more data we have for analysis, the greater will be the analytical accuracy and the greater would be the confidence in our decisions based on these analytical findings. The analytical accuracy will lead a greater positive impact in terms of enhancing operational efficiencies, reducing cost and time, and originating new products, new services, and optimizing existing services. Refer Figure 1.6.

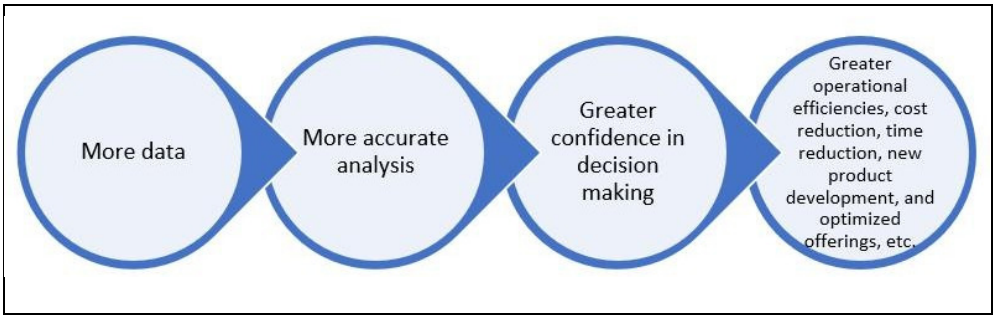



**Figure 1.6: Why big data?** (Big Data and Analytics)

## 1.7 DATA WAREHOUSE ENVIRONMENT

The data from these sources may differ in format.
Operational or transactional or day-to-day business data is gathered from Enterprise Resource Planning (ERP) systems, Customer Relationship Management (CRM), Legacy systems, and several third-party applications.

The data from these sources may differ in format.

This data is then integrated, cleaned up, transformed, and standardized through the process of Extraction, Transformation, and Loading (ETL).

The transformed data is then loaded into the enterprise data warehouse (available at the enterprise level) or data marts (available at the business unit/ functional unit or business process level).

Business intelligence and analytics tools are then used to enable decision making from the use of ad-hoc queries, SQL, enterprise dashboards, data mining, Online Analytical Processing etc. Refer Figure

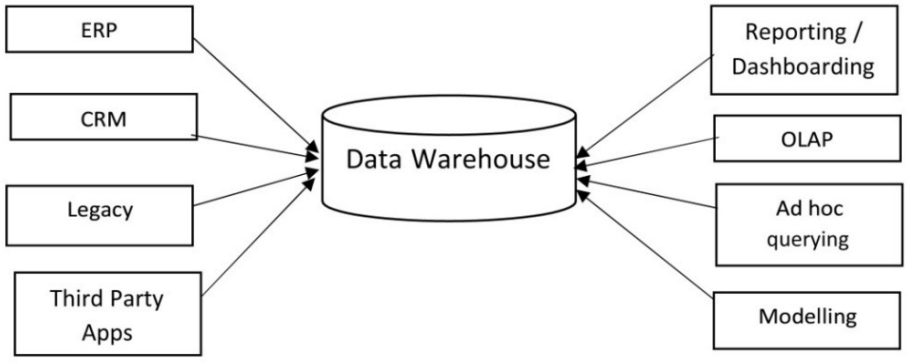



**Figure 1.7: Data Warehouse Environment** (Big Data and Analytics)

## 1.8 TRADITIONAL BUSINESS INTELLIGENCE (BI) VERSUS BIG DATA

Following are the differences that one encounters dealing with traditional Bl and big data.

In traditional BI environment, all the enterprise's data is housed in a central server whereas in a big data environment data resides in a distributed file system. The distributed file system scales by scaling in(decrease) or out(increase) horizontally ascompared to typical database server that scales vertically.

In traditional BI, data is generally analysed in an offline mode whereas in big data, it is analysed in both real-time streaming as well as in offline mode.

Traditional Bl is about structured data and it is here that data is taken to processing functions (move data to code) whereas big data is about variety: Structured, semi- structured, and unstructured data and here the processing functions are taken to the data(move code to data).

## 1.9 STATE OF THE PRACTICE IN ANALYTICS

Current business problems provide many opportunities for organizations to become more analytical and data driven, as shown in Table 1 ·2.

| Business Driver | Examples |
|---|---|
| Optimize business operations | Sales, pricing, profitability, efficiency |
| Identify business risk | Customer churn, fraud, default |
| Predict new business opportunities | Upsell, cross-sell, best new customer prospects |
| Comply with laws or regulatory Requirements | Anti-Money Laundering, Fair Lending, |

**TABLE 1.2 Business Drivers for Advanced Analytics**

(Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data)

The first three examples do not represent new problems.

Organizations have been trying to reduce customer churn, increase sales, and cross-sellcustomers for many years.

What is new is the opportunity to fuse advanced analytical techniques with Big Data to produce more impactful analyses for these traditional problems.

The last example portrays emerging regulatory requirements.
Many compliance and regulatory laws have been in existence for

decades, but additional requirements are added every year, which represent additional complexity and data requirements for organizations.

Laws related to anti-money laundering (AML) and fraud prevention require advanced analytical techniques to comply with and manage properly.

**Different types of analytics:**

1.9.1 BI Versus Data Science

1.9.2 Current Analytical Architecture (data flow)

1.9.3 Drivers of Big Data

1.9.4 Emerging Big Data Ecosystem and a New Approach to Analytics

1.9.5 BI Versus Data Science: Refer figure 1.8 for comparing BI with Data Science

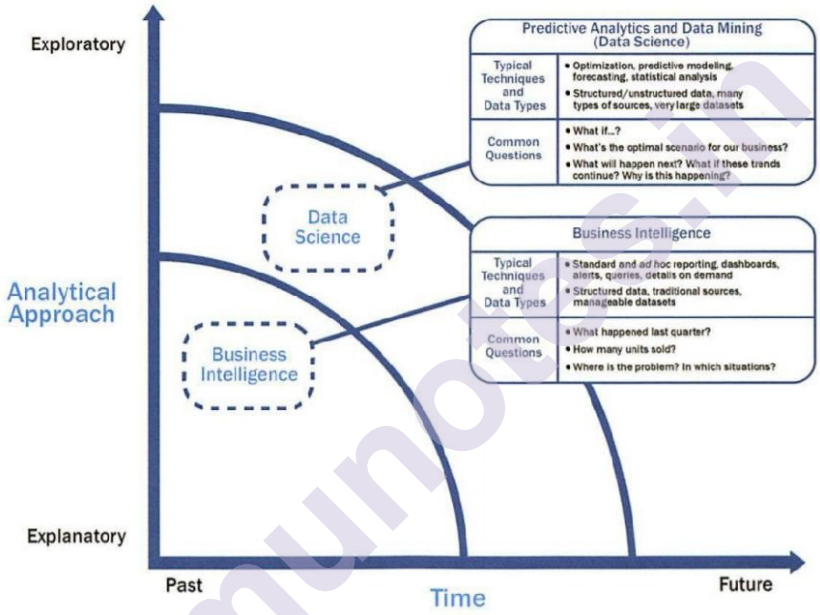



**Figure 1.8 Comparing BI with Data Science**

(Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data)

Tables – 1.3 and 1.4 explain the comparison between BI and Data Science.

| **Predictive Analytics and Data Mining (Data Science)** | |
|---|---|
| Typical Techniques and Data Types | • Optimization. predictive modelling, forecasting. statisticalanalysis<br>• Structured/unstructured data, many types of sources, verylarge datasets |
| Common Questions | • What if ... ?<br>• What's the optimal scenario for our business?<br>• What will happen next? What if these trends continue? Whyis this happening? |

**Table 1.3: Data Science** (Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data)

| Business Intelligence | |
|---|---|
| Typical Techniques and Data Types | • Standard and ad hoc reporting, dashboards, alerts, queries,details on demand<br>• Structured data. traditional sources. manageable datasets |
| Common Questions | • What happened last quarter?<br>• How many units sold?<br>• Where is the problem? Hey in which situation? |

**Table 1.4: BI**

(Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data)

1.9.1 Current Analytical Architecture: Figure 1.9 explains a typical analytical architecture.

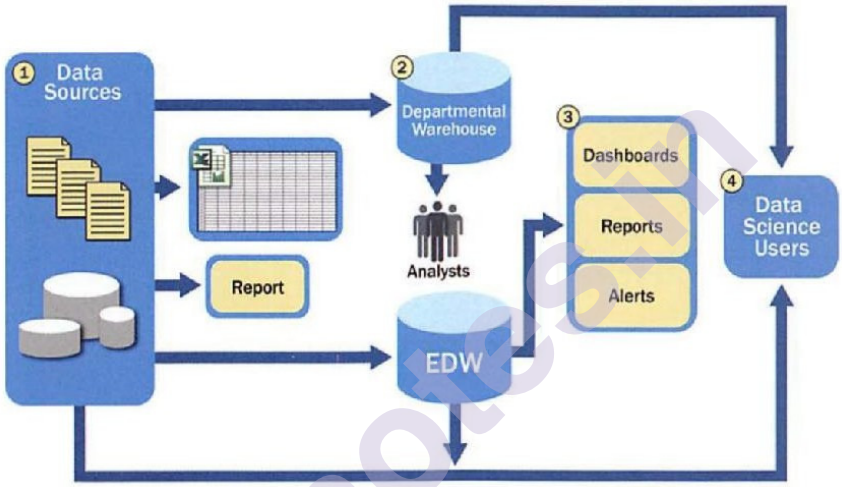



**Figure 1.9:Typical Analytical Architecture**

(Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data)

1. For data sources to be loaded into the data warehouse, data needs to be well understood, structured and normalized with the appropriate data type definitions.
2. As a result of this level of control on the EDW(enterprise data warehouse-on server or on cloud), additional local systems may emerge in the form of departmental warehouses and local data marts that business users create to accommodate their need for flexible analysis. However, these local systems reside in isolation, often are not synchronized or integrated with other data stores and may not be backed up.
3. In the data warehouse, data is read by additional applications across the enterprisefor Bl and reporting purposes.
4. At the end of this workflow, analysts get data from server. Because users generally are not allowed to run custom or intensive analytics on production databases, analysts create data extracts from the EDW to analyze data offline in R or other local analytical tools to store and process critical data, supporting enterprise applications and enabling corporate reporting activities.

Although reports and dashboards are still important for organizations, most traditional data architectures prevent data exploration and more sophisticated analysis.

### 1.9.3 Drivers of Big Data:

As shown in Figure 1.10, in the 1990s the volume of information was often measured in terabytes. Most organizations analyzed structured data in rows and columns and used relational databases and data warehouses to manage large amount of enterprise information.

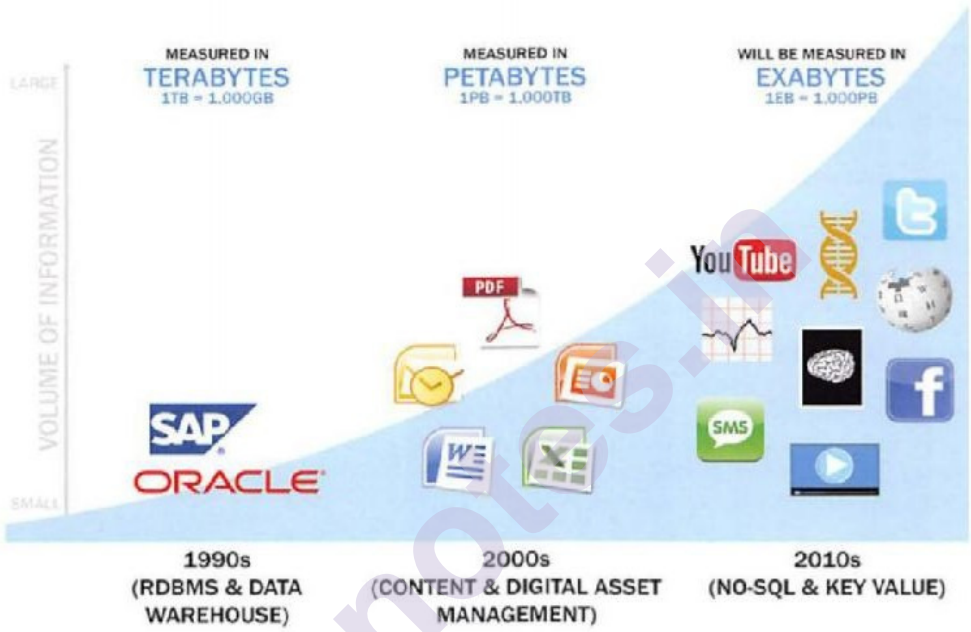



**Figure 1.10: Data Evolution and the Rise of Big Data Sources**
(Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data)

The following decade (2000) saw different kinds of data sources-mainly productivity and publishing tools such as content management repositories and networked attached storage systems-to manage this kind of information, and the data began to increase in size and started to be measured at petabyte scales.

In the 2010s, the information that organizations try to manage has broadened to include many other kinds of data. In this era, everyone and everything is leaving a digital footprint. These applications, which generate data volumes that can be measured in exabyte scale, provide opportunities for new analytics and driving new value for organizations. The data now comes from multiple sources, like Medical information, Photos and video footage, Video surveillance, Mobile devices, Smart devices, Nontraditional IT devices etc.

### 1.9.4 Emerging Big Data Ecosystem and a New Approach to Analytics


**Figure 1.11 – Emerging Big Data Ecosystem**
(Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data)

As the new ecosystem takes shape, there are four main groups of players within thisinterconnected web. These are shown in Figure 1.11.

1. Data devices and the "Sensornet" gather data from multiple locations and continuously generate new data about this data. For each gigabyte of new datacreated, an additional petabyte of data is created about that data.

For example, consider someone playing an online video game through a PC, game console, or smartphone. In this case, the video game provider captures data about the skill and levels attained by the player. Intelligent systems monitor and log how and when the user plays the game. As a consequence, the game provider can fine-tune thedifficulty of the game, suggest other related games that would most likely interest the user, and offer additional equipment and enhancements for the character based on the user's age, gender, and interests. This information may get stored locally or uploaded to the game provider's cloud to analyze the gaming habits and opportunities for upsell and cross-sell and identify typical profiles of specific kinds of users.

Smartphones provide another rich source of data. In addition to messaging and basic phone usage, they store and transmit data about Internet usage, SMS usage, and real- time location. This metadata can be used for analyzing traffic patterns by scanning thedensity of smartphones in locations to track the speed of cars or the relative traffic congestion on busy roads. In this way, GPS devices in cars can give drivers real-time updates and offer alternative routes to avoid traffic delays.

Retail shopping loyalty cards record not just the amount an individual spends, but the locations of stores that person visits, the kinds of products purchased, the stores where goods are purchased most often,

and the combinations of products purchased together.Collecting this data provides insights into shopping and travel habits and the likelihood of successful advertisement targeting for certain types of retail promotions.

2. Data collectors include sample entities that collect data from the device and users.

   Data results from a cable TV provider tracking the shows a person watches, which TV channels someone will and will not pay for to watch on demand, and the prices someone is willing to pay for premium TV content

   Retail stores tracking the path a customer takes through their store while pushing a shopping cart with an RFID chip so they can gauge which products get the most foottraffic using geospatial data collected from the RFID chips

3. Data aggregators make sense of the data collected from the various entities from the "SensorNet" or the "Internet of Things." These organizations compile data from the devices and usage patterns collected by government agencies, retail stores and websites. ln turn, they can choose to transform and package the data as products to sell to list brokers, who may want to generate marketing lists of people who may be good targets for specific ad campaigns.

4. Data users / buyers: These groups directly benefit from the data collected and aggregated by others within the data value chain. Retail banks, acting as a data buyer, may want to know which customers have the highest likelihood to apply for a second mortgage or a home equity line of credit.

To provide input for this analysis, retail banks may purchase data from a data aggregator. This kind of data may include demographic information about people living in specific locations; people who appear to have a specific level of debt, yet still have solid credit scores (or other characteristics such as paying bills on time and having savings accounts) that can be used to infer credit worthiness; and those who are searching the web for information about paying off debts or doing home remodeling projects. Obtaining data from these various sources and aggregators will enable a more targeted marketing campaign, which would have been more challenging before Big Data due to the lack of information or high-performing technologies.

Using technologies such as Hadoop to perform natural language processing on unstructured, textual data from social media websites, users can gauge the reaction to events such as presidential campaigns. People may, for example, want to determine public sentiments toward a candidate by analyzing related blogs and online comments. Similarly, data users may want to track and prepare for natural disasters by identifying which areas a hurricane affects first and how it moves, based

on which geographic areas are tweeting about it or discussing it via social media.

## 1.10 KEY ROLES FOR THE NEW BIG DATA ECOSYSTEM

Refer figure 1.12 for Key roles of the new big data ecosystems.

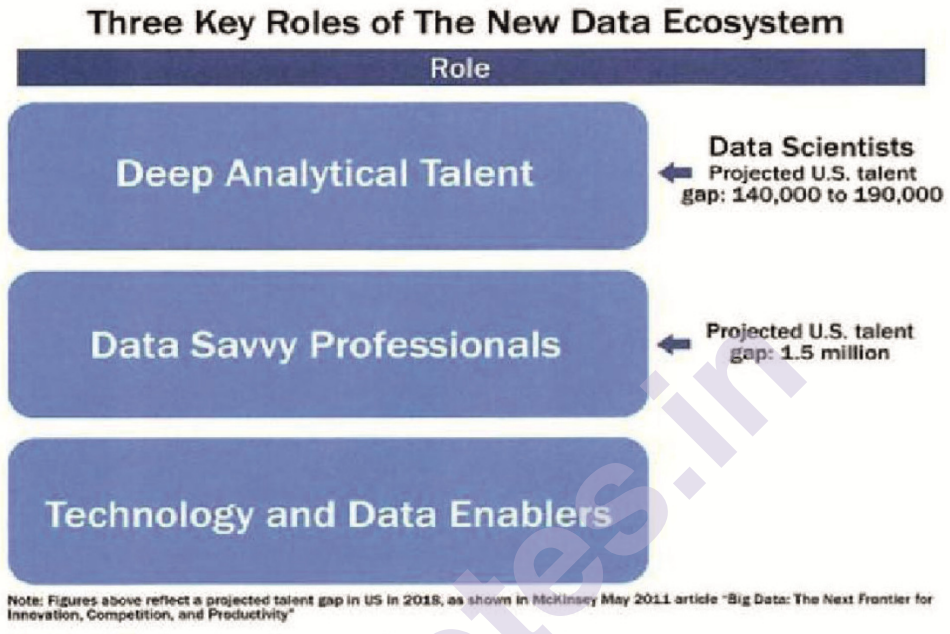



**Figure 1.12 – Key roles of the new big data ecosystems**
**(Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data)**

1. Deep Analytical Talent is technically savvy, with strong analytical skills. Members possess a combination of skills to handle raw, unstructured data and to apply complex analytical techniques at massive scales.

2. This group has advanced training in quantitative disciplines, such as mathematics, statistics, and machine learning. To do their jobs, members need access to a robust analytic sandbox or workspace where they can perform large-scale analytical dataexperiments.

   Examples of current professions fitting into this group include statisticians, economists,mathematicians, and the new role of the Data Scientist.

3. Data Savvy Professionals-has less technical depth but has a basic knowledge of statistics or machine learning and can define key questions that can be answered using advanced analytics.

   These people tend to have a base knowledge of working with data, or an appreciation for some of the work being performed by data scientists and others with deep analytical talent.

Examples of data savvy professionals include financial analysts, market research analysts, life scientists, operations managers, and business and functional managers.

4. Technology and Data Enablers- This group represents people providing technical expertise to support analytical projects, such as provisioning and administrating analytical sandboxes, and managing large-scale data architectures that enable widespread analytics within companies and other organizations.

   This role requires skills related to computer engineering, programming, and databaseadministration.

   These three groups must work together closely to solve complex Big Data challenges.
   Most organizations are familiar with people in the latter two groups mentioned, but thefirst group, Deep Analytical Talent, tends to be the newest role for most and the least understood.

   For simplicity, this discussion focuses on the emerging role of the Data Scientist. It describes the kinds of activities that role performs and provides a more detailed view of the skills needed to fulfill that role.

**Activities of data scientist:**

There are three recurring sets of activities that data scientists perform:

Reframe business challenges as analytics challenges. Specifically, this is a skill to diagnose business problems, consider the core of a given problem, and determine whichkinds of analytical methods can be applied to solve it.

Design, implement, and deploy statistical models and data mining techniques on Big Data. This set of activities is mainly what people think about when they consider the roleof the Data Scientist: namely, applying complex or advanced analytical methods to a variety of business problems using data.

Develop insights that lead to actionable recommendations. It is critical to note that applying advanced methods to data problems does not necessarily drive new businessvalue. Instead, it is important to learn how to draw insights out of the data and communicate them effectively.

**Profile of a data scientist:**

Data scientists are generally thought of as having five main sets of skills and behavioralcharacteristics, as shown in Figure 1-13:

Quantitative skill: such as mathematics or statistics

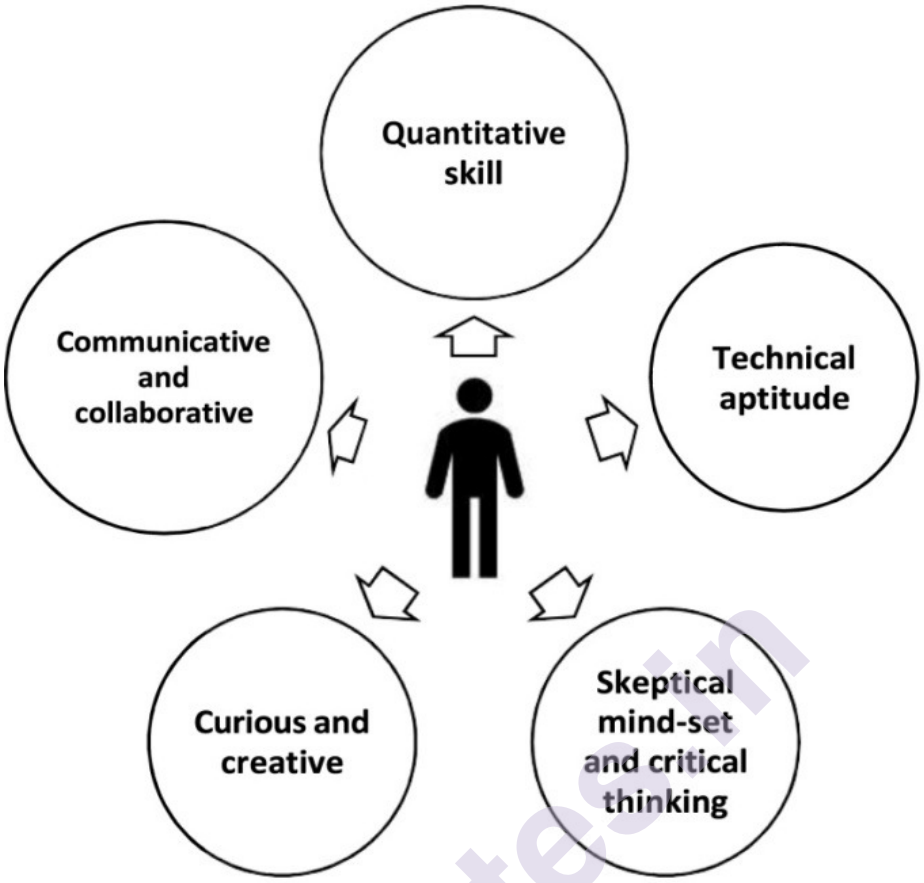



**Figure 1.13 - Data scientist**

**(Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data)**

**Technical aptitude:** namely, software engineering, machine learning, and programmingskills

**Skeptical mind-set and critical thinking:** It is important that data scientists can examinetheir work critically rather than in a one-sided way. Curious and creative: Data scientists are passionate about data and finding creative waysto solve problems and portray information.

**Communicative and collaborative:** Data scientists must be able to understand the business value in a clear way and collaboratively work with other groups, including project sponsors and key stakeholders.

Data scientists are generally comfortable using this blend of skills to acquire, manage, analyze, and visualize data and tell compelling stories about it.

## 1.11 EXAMPLES OF BIG DATA ANALYTICS

There are three examples of Big Data Analytics in different areas: retail, IT infrastructure, and social media.

1. **Retail**: As mentioned earlier, Big Data presents many opportunities to improve sales and marketing analytics.

An example of this is the U.S. retailer Target. After analyzing consumer purchasing behavior, Target's statisticians determined that the retailer made a great deal of money from three main life-event situations.

- Marriage, when people tend to buy many new products.
- Divorce, when people buy new products and change their spending habits.
- Pregnancy, when people have many new things to buy and have an urgency to buy them. The analysis target to manage its inventory, knowing that there would be demand for specific products and it would likely vary by month over the coming nine- to ten-month cycles.

2. **IT infrastructure:** MapReduce paradigm is an ideal technical framework for many Big Data projects, which rely on large data sets with unusual data structures.

   One of the main benefits of Hadoop is that it employs a distributed file system, meaning it can use a distributed cluster of servers and commodity hardware to process large amounts of data.

   Some of the most common examples of Hadoop implementations are in the social media space, where Hadoop can manage transactions, give textual updates, and develop social graphs among millions of users.

   Twitter and Facebook generate massive amounts of unstructured data and use Hadoop and its ecosystem of tools to manage this high volume.

3. **Social media:** It represents a tremendous opportunity to leverage social and professional interactions to derive new insights.

   LinkedIn represents a company in which data itself is the product. Early on, LinkedIn founder Reid Hoffman saw the opportunity to create a social network for working professionals.

   As of 2014, LinkedIn has more than 250 million user accounts and has added many additional features and data-related products, such as recruiting, job seeker tools, advertising, and ln Maps, which show a social graph of a user's professional network.

## SUMMARY

In this chapter you have learnt about What is big, its Characteristics, when it got evolved its definition, challenges. Also, Data Warehouse environment, what are the differences between Traditional Business Intelligence and Big Data, State of Practice in Analytics, Key

roles for New Big Data Ecosystems and some examples of big Data Analytics.

## REVIEW QUESTIONS

1. Define big data. Why is big data required? How does traditional BI environment differ from big data environment?
2. What are the challenges with big data?
3. Define big data. Why is big data required? Write a note on data warehouse environment.
4. What are the three characteristics of big data? Explain the differences between Bl and Data Science.
5. Describe the current analytical architecture for data scientists.
6. What are the key roles for the New Big Data Ecosystem?
7. What are key skill sets and behavioral characteristics of a data scientist?

## RFERENCES


- (n.d.). In S. A. Subhashini Chellappan, *Big Data and Analytics* (First ed.). Wiley.
- (n.d.). In *Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data.* Manufactured in the United States of America: John Wiley & Sons, Inc.


.

*****

2

# BIG DATA ANALYTICS

**Unit Structure**



## 2.0 OBJECTIVES

Big Data is creating significant new opportunities for organizations to derive new value and create competitive advantage from their most valuable asset: information. For businesses, Big Data helps drive efficiency, quality, and personalized products and services, producing improved levels of customer satisfaction and profit. For scientific efforts, Big Data analytics enable new avenues of investigation with potentially richer results and deeper insights than previously available. In many cases, Big Data analytics integrate structured and unstructured data with Realtime feeds and queries, opening new paths to innovation and insight.

## 2.1 INTRODUCTION TO BIG DATA ANALYTICS

Big Data Analytics is...

1 Technology-enabled analytics: Quite a few data analytics and visualization tools are available in the market today from leading vendors such as IBM, Tableau, SAS, R Analytics, Statistica, World Programming Systems (WPS), etc. to help process and analyze your big data.

2. About gaining a meaningful, deeper, and richer insight into your business to steer it in the right direction. understanding the customer's demographics to cross-sell and up-sell to them, better leveraging the services of your vendors and suppliers, etc.

3. About a competitive edge over your competitors by enabling you with findings thatallow quicker and better decision-making.

4. A tight handshake between three communities: IT, business users, and data scientists.Refer Figure 3.3.

5. Working with datasets whose volume and variety exceed the current storage and processing capabilities and infrastructure of your enterprise.

About moving code to data. This makes perfect sense as the program for distributed processing is tiny (just a few KBs) compared to the data (Terabytes or Petabytes today and likely to be Exabytes or Zettabytes in the near future).

## 2.2 CLASSIFICATION OF ANALYTICS

There are basically two schools of thought:

1 Those that classify analytics into basic, operationalized, advanced and Monetized.

2 Those that classify analytics into analytics 1.0, analytics 2.0, and analytics 3.0.

### 2.2.1. First School of Thought

It includes Basic analytics, Operationalized analytics, Advanced analytics and Monetizedanalytics.

**Basic analytics:** This primarily is slicing and dicing of data to help with basic business insights. This is about reporting on historical data, basic visualization, etc.

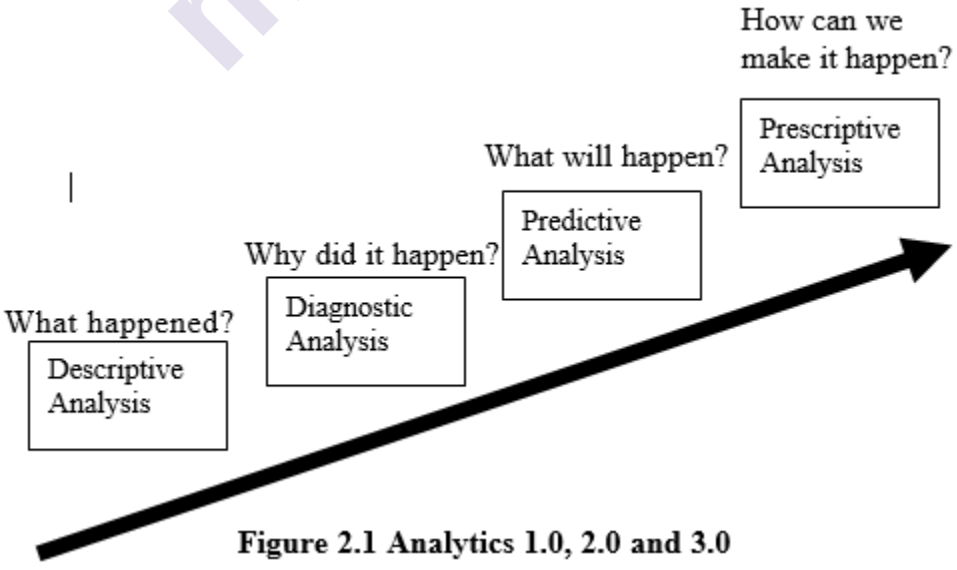



**Figure 2.1 Analytics 1.0, 2.0 and 3.0**

**Operationalized analytics:** It is operationalized analytics if it gets woven into the enterprisesbusiness processes.

**Advanced analytics:** This largely is about forecasting for the future by way of predictive andprescriptive modelling.

**Monetized analytics:** This is analytics in use to derive direct business revenue.

### 2.2.2 Second School of Thought:

Let us take a closer look at analytics 1.0, analytics 2.0, and analytics 3.0. Refer Table 2.1. Figure 2.1 shows the subtle growth of analytics from Descriptive → Diagnostic → Predictive → Perspective analytics.

| Analytics 1.0 | Analytics 2.0 | Analytics 3.0 |
|---|---|---|
| **Era: mid 1990s to 2009** Descriptive statistics (report on events, occurrences, etc. of the past) | **2005 to 2012** Descriptive statistics predictive statistics (use data from the past to make predictions for the future) | **2012 to present** Descriptive + predictive + prescriptive statistics (use data from the past to make prophecies for the future and at the same time make recommendations to leverage the situation to one's advantage) |
| key questions asked:<br>What happened?<br>Why did it happen? | key questions asked:<br>What happened?<br>Why will it happen? | Key questions asked:<br>What will happen?<br>When will it happen?<br>Why will it happen?<br>What should be the action taken to take advantage of what will happen? |
| Data from legacy systems. ERP, CRM, and 3rd party applications. | Big data | A blend of big data and data from legacy systems, ERP, CRM, and 3rd party applications. |
| Small and structured data sources. Data stored in enterprise data warehouses or data marts. | Big data is being taken up seriously. Data is mainly unstructured, arriving at a much higher pace. This fast flow of data entailed that the influx of big volume data had to be stored and processed rapidly, often on massive parallel servers running Hadoop. | A blend of big data and traditional analytics to yield insights and offerings with speed and impact. |
| Data was internally sourced. | Data was often externally sourced. | Data is both being internally and externally sourced. |

| | | |
|---|---|---|
| Relational databases | Database appliances, Hadoop clusters, SQL to Hadoop environments, etc. | In memory analytics, in database processing, agile analytical methods, machine learning techniques etc. |

**Table 2.1Analytics 1.0, 2.0 and 3.0** (Big Data and Analytics)

## 2.3 CHALLENGES OF BIG DATA

There are mainly seven challenges of big data: scale, security, schema, Continuous availability, Consistency, Partition tolerant and data quality.

**Scale:** Storage (RDBMS (Relational Database Management System) or NoSQL (Not only SQL)) is one major concern that needs to be addressed to handle the need for scaling rapidlyand elastically. The need of the hour is a storage that can best withstand the attack of large volume, velocity and variety of big data. Should you scale vertically or should you scale horizontally?

**Security:** Most of the NoSQL big data platforms have poor security mechanisms (lack of proper authentication and authorization mechanisms) when it comes to safeguarding big data. A spot that cannot be ignored given that big data carries credit card information, personal information and other sensitive data.

**schema:** Rigid schemas have no place. We want the technology to be able to fit our big data and not the other way around. The need of the hour is dynamic schema. Static (pre-defined schemas) are obsolete.

**Continuous availability:** The big question here is how to provide 24/7 support because almostall RDBMS and NoSQL big data platforms have a certain amount of downtime built in.

**Consistency:** Should one opt for consistency or eventual consistency?
Partition tolerant: How to build partition tolerant systems that can take care of both hardwareand software failures?

**Data quality:** How to maintain data quality- data accuracy, completeness, timeliness, etc.? Dowe have appropriate metadata in place?

## 2.4 IMPORTANCE OF BIG DATA

Let us study the various approaches to analysis of data and what it leads to.

**Reactive-Business Intelligence:** What does Business Intelligence (BI) help us with? It allows the businesses to make faster and better decisions by providing the right information to the right person at the right time in

the right format. It is about analysis of the past or historical data and then displaying the findings of the analysis or reports in the form of enterprise dashboards, alerts, notifications, etc. It has support for both pre-specified reports as well as adhoc querying.

**Reactive - Big Data Analytics:** Here the analysis is done on huge datasets but the approach isstill reactive as it is still based on static data.

**Proactive - Analytics:** This is to support futuristic decision making by use of data mining predictive modelling, text mining, and statistical analysis on. This analysis is not on big data as it still the traditional database management practices on big data and therefore has severe limitations on the storage capacity and the processing capability.

**Proactive - Big Data Analytics:** This is filtering through terabytes, petabytes, exabytes of information to filter out the relevant data to analyze. This also includes high performance analytics to gain rapid insights from big data and the ability to solve complex problems using more data.

## 2.5 BIG DATA TECHNOLOGIES

Following are the requirements of technologies to meet challenges of big data:

- The first requirement is of cheap and ample storage.
- We need faster processors to help with quicker processing of big data. Affordable open source distributed big data platforms, such as Hadoop.
- Parallel processing, clustering, virtualization, large grid environments (to distribute processing to a number of machines), high connectivity, and high throughputs(rate at whichsomething is processed).
- Cloud computing and other flexible resource allocation arrangements.

## 2.6 DATA SCIENCE

Data science is the science of extracting knowledge from data. In other words, it is a scienceof drawing out hidden patterns amongst data using statistical and mathematical techniques.

It employs techniques and theories drawn from many fields from the broad areas of mathematics, statistics, information technology including machine learning, data engineering, probability models, statistical learning, pattern recognition and learning, etc.

Data Scientist works on massive datasets for weather predictions, oil drillings, earthquake prediction, financial frauds, terrorist network and

activities, global economic impacts, sensor logs, social media analytics, customer churn, collaborative filtering(prediction about interest on users), regression analysis, etc. Data science is multi-disciplinary. Refer to Figure 2.2.

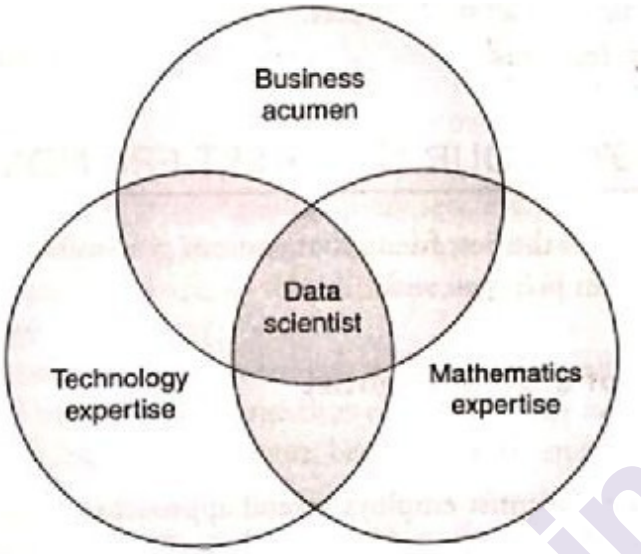



**Figure 2.2 Data Scientist** (Big Data and Analytics)

### 2.6.1 Business Acumen(expertise) Skills:

A data scientist should have following ability to play the role of data scientist.

- Understanding of domain
- Business strategy
- Problem solving
- Communication
- Presentation
- Keenness

### 2.6.2 Technology Expertise:

Following skills required as far as technical expertise is concerned.

- Good database knowledge such as RDBMS.
- Good NoSQL database knowledge such as MongoDB, Cassandra, HBase, etc.
- Programming languages such as Java. Python, C++, etc.
- Open-source tools such as Hadoop.
- Data warehousing.
- Data mining
- Visualization such as Tableau, Flare, Google visualization APIs, etc.

### 2.6.3 Mathematics Expertise:

The following are the key skills that a data scientist will have to have to comprehend data,interpret it and analyze.

- Mathematics.
- Statistics.
- Artificial Intelligence (AI).
- Algorithms.
- Machine learning.
- Pattern recognition.
- Natural Language Processing.
- To sum it up, the data science process is
- Collecting raw data from multiple different data sources.
- Processing the data.
- Integrating the data and preparing clean datasets.
- Engaging in explorative data analysis using model and algorithms.
- Preparing presentations using data visualizations.
- Communicating the findings to all stakeholders.
- Making faster and better decisions.

## 2.7 RESPONSIBILITIES

Refer figure 2.3 to understand the responsibilities of a data scientist.

**Data Management:** A data scientist employs several approaches to develop the relevant datasets for analysis. Raw data is just "RAW", unsuitable for analysis. The data scientist works on it to prepare to reflect the relationships and contexts. This data then becomes useful for processing and further analysis.

**Analytical Techniques:** Depending on the business questions which we are trying to find answers to and the type of data available at hand, the data scientist employs a blend of analytical techniques to develop models and algorithms to understand the data, interpret relationships, spot trends, and reveal patterns.

**Business Analysis:** A data scientist is a business analyst who distinguishes cool facts from insights and is able to apply his business expertise and domain knowledge to see the results inthe business context.

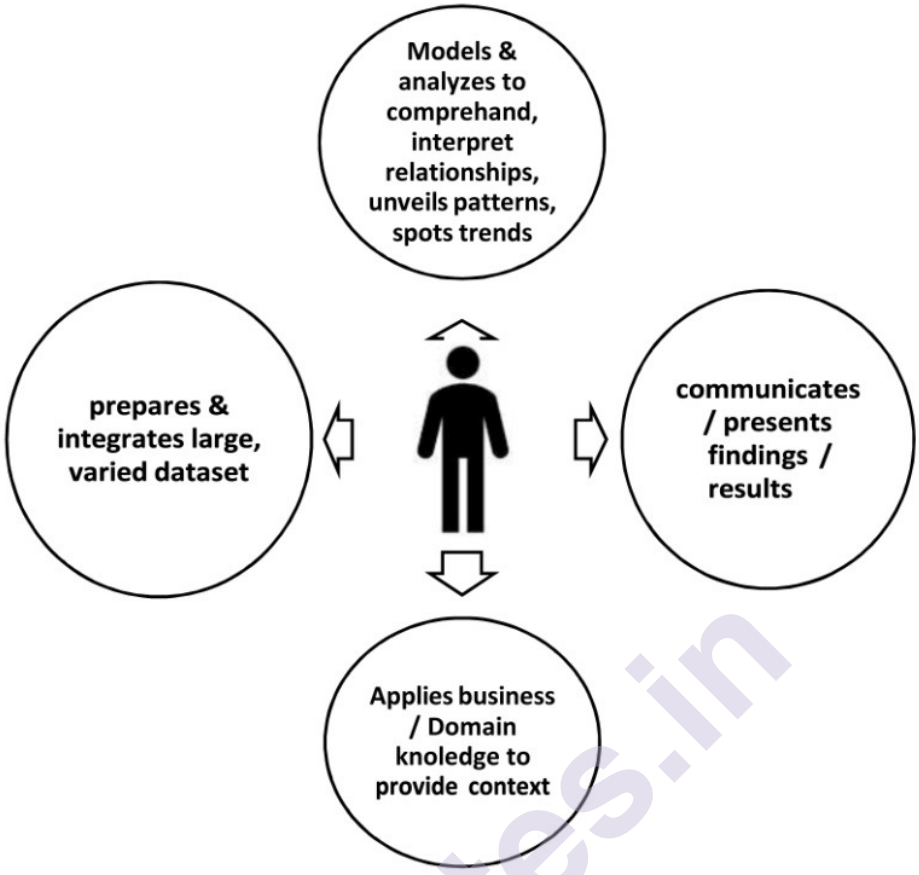



**Figure 2.3 Data scientist: your new best friend!!!**
(Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data)

**Communicator:** He is a good presenter and communicator who is able to communicate the results of his findings in a language that is understood by the different business stakeholders.

## 2.8 SOFT STATE EVENTUAL CONSISTENCY

**ACID property in RDBMS:**

**Atomicity:** Either the task (or all tasks) within a transaction are performed or none of them are. This is the all-or-none principle. If one element of a transaction fails the entire transaction fails.

**Consistency:** The transaction must meet all protocols or rules defined by the system at all times. The transaction does not isolate those protocols and the database must remain in a consistent state at the beginning and end of a transaction; there are never any half-completedtransactions.

**Isolation:** No transaction has access to any other transaction that is in an intermediate or unfinished state. Thus, each transaction is independent unto itself. This is required for both performance and consistency of transactions within a database.

**Durability:** Once the transaction is complete, it will persist as complete

and cannot be undone; it will survive system failure, power loss and other types of system breakdowns.

BASE (Basically Available, Soft state, Eventual consistency). In a system where BASE is the prime requirement for reliability, the activity/potential (p) of the data (H) changes;
it *essentially* slows down.

**Basically Available:** This constraint states that the system does guarantee the availability of the data as regards CAP Theorem; there will be a response to any request. But, that responsecould still be 'failure' to obtain the requested data or the data may be in an inconsistent or changing state, much like waiting for a check to clear in your bank account.

**Eventual consistency:** The system will *eventually* become consistent once it stops receiving input. The data will propagate to everywhere it should sooner or later, but the system will continue to receive input and is not checking the consistency of every transaction before it moves onto the next one. Werner Vogel's article "Eventually Consistent – Revisited" coversthis topic is much greater detail.

**Soft state:** The state of the system could change over time, so even during times without input there may be changes going on due to 'eventual consistency,' thus the state of the system is always 'soft.'

## 2.9 DATA ANALYTICS LIFE CYCLE

**Here is a brief overview of the main phases of the Data Analytics:**

**Phase 1- Discovery:** In Phase 1, the team learns the business domain, including relevant history such as whether the organization or business unit has attempted similar projects in thepast from which they can learn. The team assesses the resources available to support the project in terms of people, technology, time and data. Important activities in this phase include framing the business problem as an analytics challenge that can be addressed in subsequent phases and formulating initial hypotheses (IHs) to test and begin learning the data.

**Phase 2- Data preparation:** Phase 2 requires the presence of an analytic sandbox, in which theteam can work with data and perform analytics for the duration of the project. The team needs to execute extract, load, and transform (ELT) or extract, transform and load (ETL) to get data into the sandbox. The ELT and ETL are sometimes abbreviated as ETLT. Data should be transformed in the ETLT process so the team can work with it and analyze it. In this phase, the team also needs to familiarize itself with the data thoroughly and take steps tocondition the data.

**Phase 3-Model planning:** Phase 3 is model planning, where the team

determines the methods, techniques and workflow it intends to follow for the subsequent model building phase. The team explores the data to learn about the relationships between variables and subsequently selects key variables and the most suitable models.

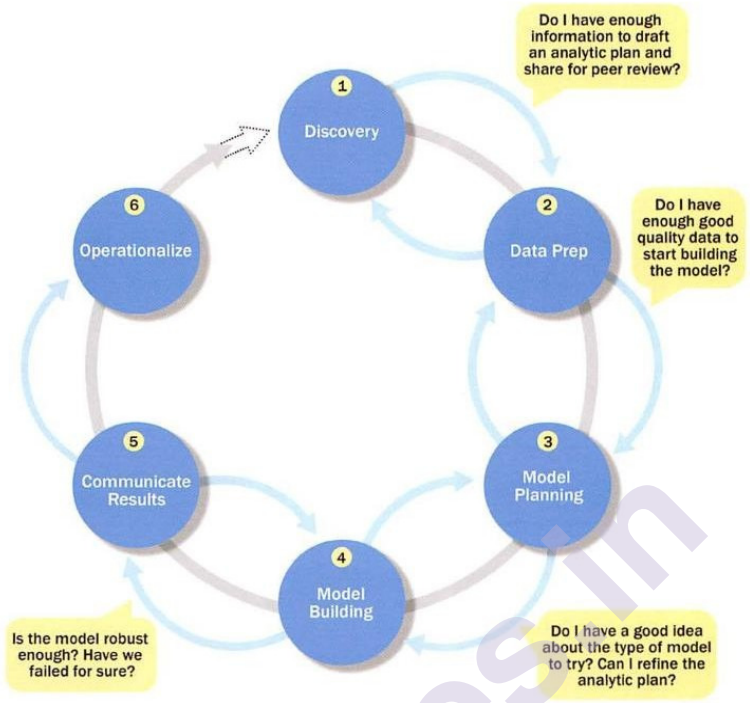



**Figure 2.4 - Overview of Data Analytical Lifecycle**

(Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data)

**Phase 4-Model building:** In Phase 4, the team develops data sets for testing, training, and production purposes. In addition, in this phase the team builds and executes models based on the work done in the model planning phase. The team also considers whether its existing tools will suffice for running the models, or if it will need a more robust environment for executing models and workflows (for example, fast hardware and parallel processing, if applicable).

**Phase 5-Communicate results:** In Phase 5, the team, in collaboration with major stakeholders, determines if the results of the project are a success or a failure based on the criteria developed in Phase 1. The team should identify key findings, quantify the business value, and develop a narrative to summarize and convey findings to stakeholders.

**Phase 6-0perationalize:** In Phase 6, the team delivers final reports, briefings, code and technical documents. In addition, the team may run a pilot project to implement the models ina production environment.

## SUMMARY

In this chapter you have learnt about What is big data analytics, its classification, challenges, and importance of Big Data. Also, Big Data Technologies, What is Data Science, the responsibilities of a data

scientist, soft state eventual consistency and Data Analytics Life Cycle.

## REVIEW QUESTIONS

1. What is big data analytics? Explain in detail with its example.
2. Write a short note on Classification of Analytics.
3. Describe the Challenges of Big Data.
4. Write a short note on data science and data science process.
5. Write a short note on soft state eventual consistency.
6. What are different phases of the Data Analytics Lifecycle? Explain each in detail.

## REFERENCES


- (n.d.). In S. A. Subhashini Chellappan, *Big Data and Analytics* (First ed.). Wiley.
- (n.d.). In *Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data.* Manufactured in the United States of America: John Wiley & Sons, Inc.


.

*****

# UNIT II

# 3

# OVERVIEW OF CLUSTERING

**Unit Structure**



## 3.1 OVERVIEW OF CLUSTERING

In general, clustering is the use of *unsupervised* techniques for grouping similar objects. In machine learning, unsupervised refers to the problem of finding hidden structure within unlabeled data. Clustering techniques are unsupervised in the sense that the data scientist does not determine, in advance, the labels to apply to the clusters. The structure of the data describes the objects of interest and determines how best to group the objects.

For example, based on customers' personal income, it is straightforward to divide the customers into three groups depending on arbitrarily selected values. The customers could be divided into three groups as follows:

Earn less than $10,000, Earn between $10,000 and $99,999 Earn $100,000 or more

In this case, the income levels were chosen somewhat subjectively based on easy-to communicate points of delineation. However, such groupings do not indicate a natural affinity of the customers within each group. In other words, there is no inherent reason to believe that the customer making $90,000 will behave any differently than the customer making $110,000. As additional dimensions are introduced by adding more variables about the customers, the task of finding meaningful groupings becomes more complex. For instance, suppose variables such as age, years of education, household size, and annual purchase expenditures were considered along with the personal income variable. What are the natural occurring groupings of customers? This is the type of question that clustering analysis can help answer.

Clustering is a method often used for exploratory analysis of the data. In clustering, there are no predictions made. Rather, clustering methods find the similarities between objects according to the object attributes and group the similar objects into clusters. Clustering techniques are utilized in marketing, economics, and various branches of science. A popular clustering method is k-means.

## 3.2-K-MEANS

Given a collection of objects each with n measurable attributes, *k-means* is an analytical technique that, for a chosen value of k, identifies k clusters of objects based on the objects' proximity to the center of the k groups. The center is determined as the arithmetic average (mean) of each cluster's n-dimensional vector of attributes. Below figure illustrates three dusters of objects with two attributes. Each object in the dataset is represented by a small dot color-coded to the closest large dot, the mean of the cluster.

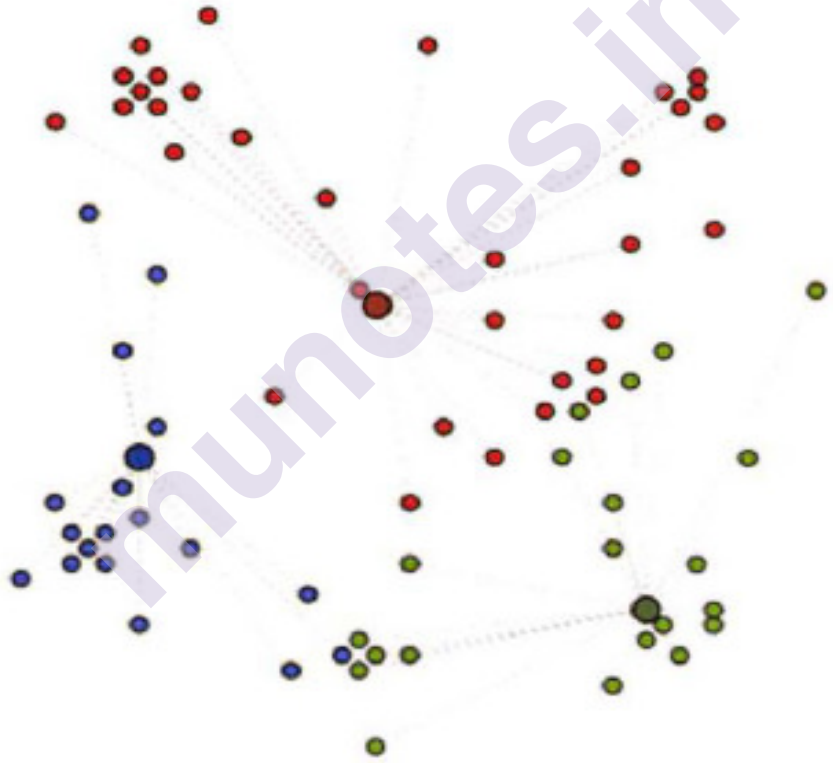


**Figure 3.1 Possible K-means clusters for k=3**

### 3.2.1-Use Cases:

Clustering is often used as a lead-in to classification. Once the clusters are identified, labels can be applied to each cluster to classify each group based on its characteristics. Some specific applications of k-means are image processing, medical and customer segmentation.

**Image Processing:**

Video is one example of the growing volumes of unstructured data being collected. Within each frame of a video, k-means analysis can be used to identify objects in the video. For each frame, the task is to determine which pixels are most similar to each other. The attributes of each pixel can include brightness, color, and location, the x and y coordinates in the frame. With security video images, for example, successive frames are examined to identify any changes to the clusters. These newly identified dusters may indicate unauthorized access to a facility.

**Medical:**

Patient attributes such as age, height, weight, systolic and diastolic blood pressures, cholesterol level, and other attributes can identify naturally occurring clusters. These dusters could be used to target individuals for specific preventive measures or clinical trial participation. Clustering, in general, is useful in biology for the classification of plants and animals as well as in the field of human genetics.

**Customer Segmentation:**

Marketing and sales groups use k-means to better identify customers who have similar behaviors and spending patterns. For example, a wireless provider may look at the following customer attributes: monthly bill, number of text messages, data volume consumed, minutes used during various daily periods, and years as a customer. The wireless company could then look at the naturally occurring clusters and consider tactics to increase sales or reduce the customer *churn rate,* the proportion of customers who end their relationship with a particular company.

**3.2.2 Overview of the Method:**

To illustrate the method to find k clusters from a collection of M objects with n attributes, the two-dimensional case (n = 2) is examined. It is much easier to visualize the k-means method in two dimensions.

Because each object in this example has two attributes, it is useful to consider each object corresponding to the point $(x_i, y_i)$, where x and y denote the two attributes and i = 1, 2 ... M. For a given cluster of m points ($m \leq M$), the point that corresponds to the cluster's mean is called a centroid.

The k-means algorithm to find k dusters can be described in the following four steps.

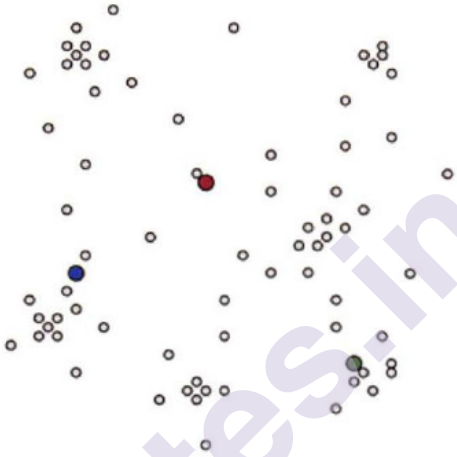
1. Choose the value of k and the k initial guesses for the centroids.
   In this example, k = 3, and the initial centroids are indicated by the points shaded in red, green, and blue in figure 2.2.

2. Compute the distance from each data point ($x_i$ ,$y_i$ ) to each centroid. Assign each point to the closest centroid. This association defines the first k dusters.

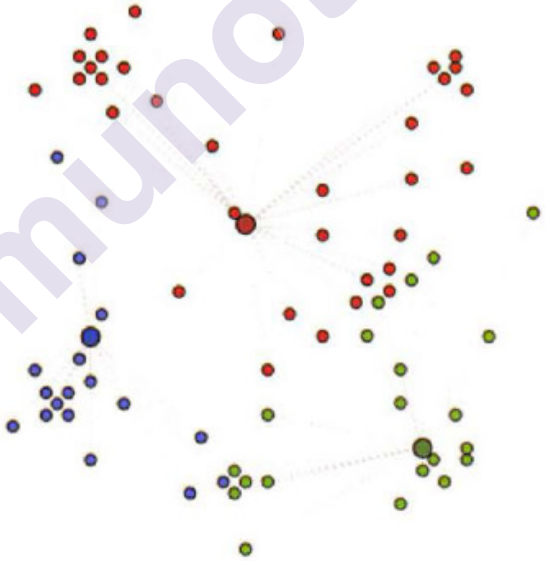
In two dimensions, the distance, *d,* between any two points, ($x_1$,$y_1$) and ($x_2$,$y_2$), in the Cartesian plane is typically expressed by using the Euclidean distance measure provided in Equation.

$$d= \sqrt{(x1-y1)^2+(x2-y2)^2}$$

In figure 3.2, the points closest to a centroid are shaded the corresponding color.



**3.2 Initial starting points for the centroids**



**3.3 Points are assigned to the closest centroid**

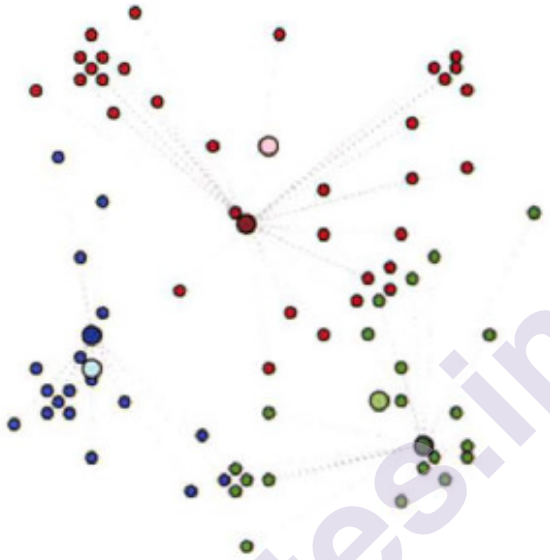
3. Compute the centroid, the center of mass, of each newly defined cluster from Step *2.*

In Figure 2-4, the computed centroids in Step 3 are the lightly shaded points of the corresponding color. In two dimensions, the centroid ($x_c$,$y_c$) of the m points in a k-means duster is calculated as follows in

Equation.

$$P(Xc, Yc) = \left( \frac{\sum_{i=1}^{m} X_i}{m}, \frac{\sum_{i=1}^{m} Y_i}{m} \right)$$

Thus, $(x_c, y_c)$ is the ordered pair of the arithmetic means of the coordinates of the m points in the cluster. In this step, a centroid is computed for each of the k clusters.



**3.4a Compute the mean of each cluster**

4. Repeat Steps 2 and 3 until the algorithm converges to an answer
   a. Assign each point to the closest centroid computed in Step 3.
   b. Compute the centroid of newly defined clusters.
   c. Repeat until the algorithm reaches the final answer.

### 3.2.3 Determining the Number of Clusters:

In k-means, k clusters can be identified in a given dataset, but what value of k should be selected? The value of k can be chosen based on a reasonable guess or some predefined requirement. However, even then, it would be good to know how much better or worse having k clusters versus k-1 or k+1 cluster would be in explaining the structure of the data. Next, a heuristic using the Within Sum of Squares (WSS) metric is examined to determine a reasonably optimal value of k. Using the distance function, WSS is defined as shown below.

$$WSS = \sum_{i=1}^{M} d\left(p_1, q^i\right)^2 = \sum_{i=1}^{M} \sum_{i=1}^{n} \left(p_i - q_i^l\right)^2$$

In other words, WSS is the sum of the squares of the distances between each data point and the closest centroid. The term $\boldsymbol{q^{(i)}}$ indicates the closest centroid that is associated with the $i^{th}$ point. If the points are

relatively close to their respective centroids, the WSS is relatively small. Thus, if k +1 clusters do not greatly reduce the value of WSS from the case with only k clusters, there may be little benefit to adding another cluster.

### 3.2.4 Diagnostics:

The heuristic using WSS can provide at least several possible k values to consider. When the number of attributes is relatively small, a common approach to further refine the choice of k is to plot the data to determine how distinct the identified clusters are from each other. In general, the following questions should be considered.

- Are the dusters well separated from each other?
- Do any of the dusters have only a few points?
- Do any of the centroids appear to be too close to each other?

In the first case, ideally the plot would look like the one shown in below figure, when n = 2.

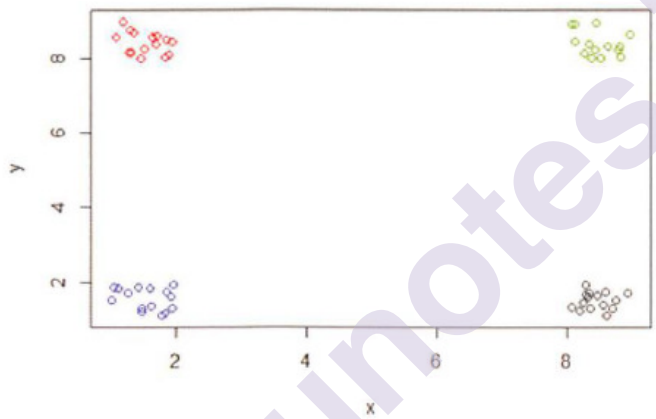
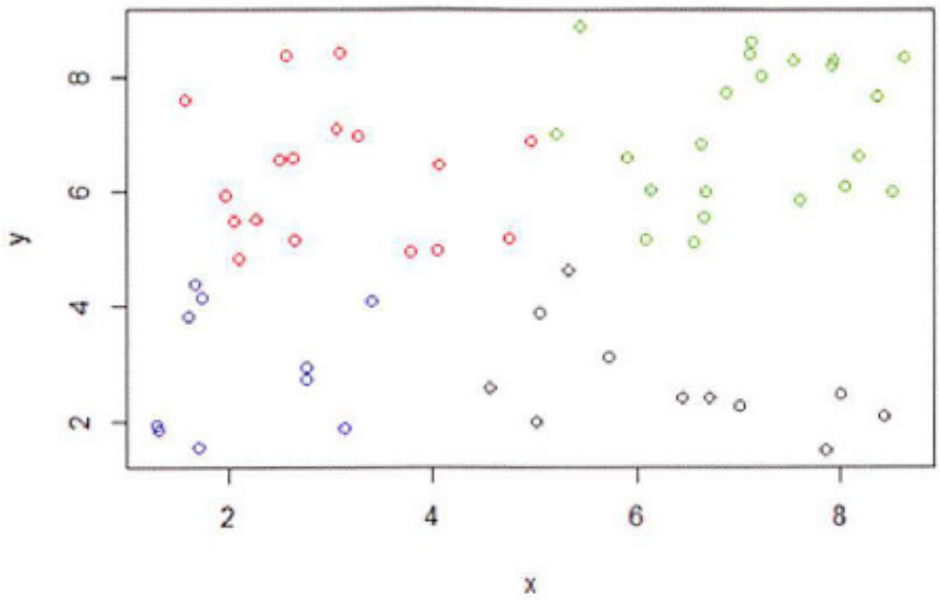



**Figure: 3.4b Example of distinct clusters**

The clusters are well defined, with considerable space between the four identified clusters. However, in other cases, such as in below figure, the clusters may be close to each other, and the distinction may not be so Obvious




.**Figure 3.5 Example of less obvious cluster**

### 3.2.5 Reasons to Choose and Cautions:

K-means is a simple and straightforward method for defining clusters. Once clusters and their associated centroids are identified, it is easy to assign new objects (for example, new customers) to a cluster based on the object's distance from the closest centroid. Because the method is unsupervised, using k-means helps to eliminate subjectivity from the analysis.

Although k-means is considered an unsupervised method, there are still several decisions that the practitioner must make:

a. What object attributes should be included in the analysis?

b. What unit of measure (for example, miles or kilometers) should be used for each attribute?

c. Do the attributes need to be rescaled so that one attribute does not have a disproportionate effect on the results?

d. What other considerations might apply?

**a. Object Attributes:**

Regarding which object attributes (for example, age and income) to use in the analysis, it is important to understand what attributes will be known at the time a new object will be assigned to a cluster. For example, information on existing customers' satisfaction or purchase frequency may be available, but such information may not be available for potential customers.

The Data Scientist may have a choice of a dozen or more attributes to use in the clustering analysis. Whenever possible and based on the data, it is best to reduce the number of attributes to the extent possible. Too

many attributes can minimize the impact of the most important variables. Also, the use of several similar attributes can place too much importance on one type of attribute. For example, if five attributes related to personal wealth are included in a clustering analysis, the wealth attributes dominate the analysis and possibly mask the importance of other attributes, such as age.

When dealing with the problem of too many attributes, one useful approach is to identify any highly correlated attributes and use only one or two of the correlated attributes in the clustering analysis.

Another option to reduce the number of attributes is to combine several attributes into one measure. For example, instead of using two attribute variables, one for Debt and one for Assets, a Debt to Asset ratio could be used, This option also addresses the problem when the magnitude of an attribute is not of real interest, but the relative magnitude is a more important measure.

**b. Units of Measure:**

From a computational perspective, the k-means algorithm is somewhat indifferent to the units of measure for a given attribute (for example, meters or centimeters for a patient's height). However, the algorithm will identify different clusters depending on the choice of the units of measure.

For example, suppose that k-means is used to cluster patients based on age in years and height in centimeters. For k=2, below figure illustrates the two clusters that would be determined for a given dataset.

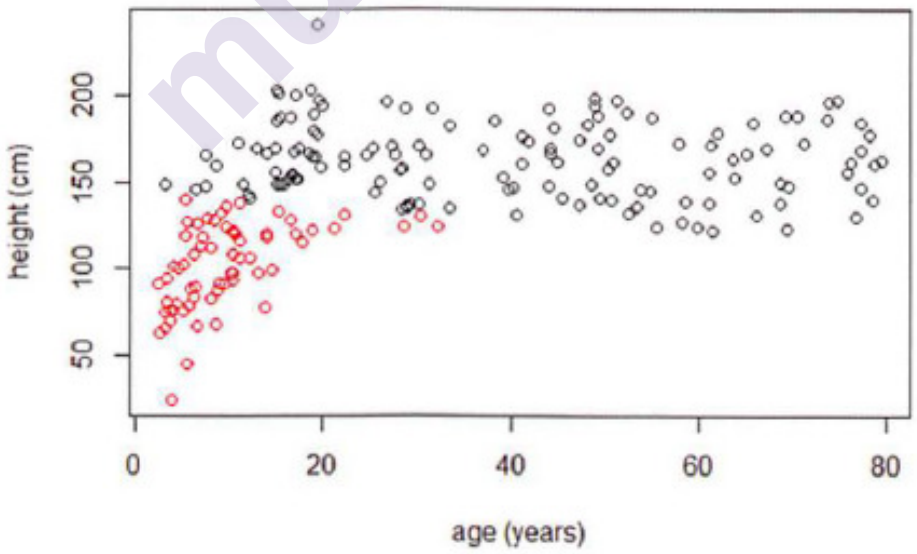



**Figure 3.6. Cluster with height expressed in centimeters**

But if the height was rescaled from centimeters to meters by

dividing by 100, the resulting dusters would be slightly different, as illustrated in below Figure.

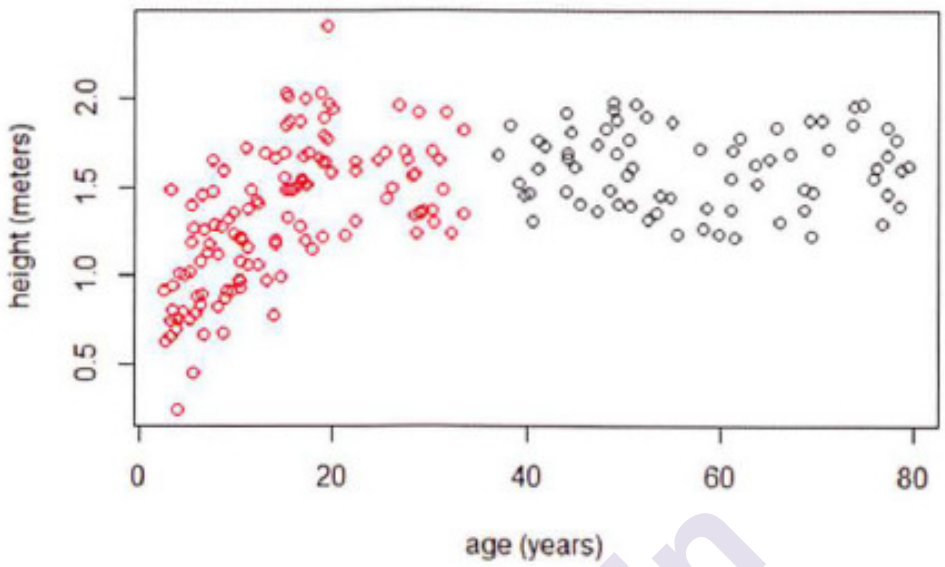



**Figure 3.7. Cluster with height expressed in meters**

**c. Rescaling:**

Attributes that are expressed in dollars are common in clustering analyses and can differ in magnitude from the other attributes. For example, if personal income is expressed in dollars and age is expressed in years, the income attribute, often exceeding 510,000, can easily dominate the distance calculation with ages typically less than 100 years.

Although some adjustments could be made by expressing the income in thousands of dollars (for example, 10 for 510,000), a more straightforward method is to divide each attribute by the attribute's standard deviation. The resulting attributes will each have a standard deviation equal to 1 and will be without units.

Returning to the age and height example, the standard deviations are 23.1 years and 36.4 cm, respectively. Dividing each attribute value by the appropriate standard deviation and performing the k-means analysis yields the result shown in Figure 3.8.

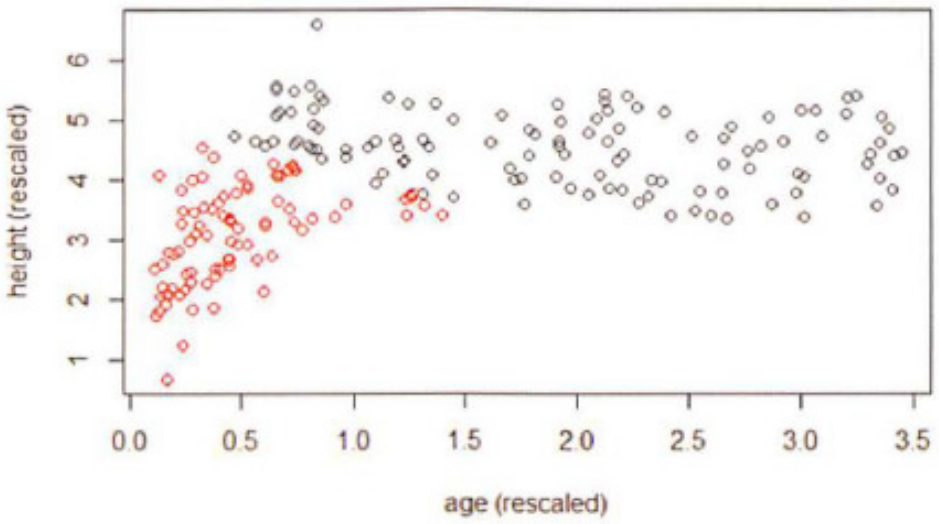



**Figure 3.8.Cluster with rescaled attributes**

In many statistical analyses, it is common to transform typically skewed data, such as income, with long tails by taking the logarithm of the data. Such transformation can also be applied in k-means, but the Data Scientist needs to be aware of what effect this transformation will have.

**d. Additional Considerations:**

The k-means algorithm is sensitive to the starting positions of the initial centroid. Thus, it is important to rerun the k-means analysis several times for a particular value of k to ensure the cluster results provide the overall minimum WSS. As we know, this task is accomplished in R by using the nstart option in the kmeans () function call.

K-means clustering is applicable to objects that can be described by attributes that are numerical with a meaningful distance measure. Interval and ratio attribute types can certainly be used. However, k-means does not handle categorical variables well. For example, suppose a clustering analysis is to be conducted on new car sales. Among other attributes, such as the sale price, the color of the car is considered important. Although one could assign numerical values to the color, such as red = 1, yellow = 2, and green = 3, it is not useful to consider that yellow is as close to red as yellow is to green from a clustering perspective. In such cases, it may be necessary to use an alternative clustering methodology.

## 3.3 ASSOCIATION RULES

An unsupervised learning method called association rules. This is a descriptive, not predictive, method often used to discover interesting relationships hidden in a large dataset. The disclosed relationships can be represented as rules or frequent item sets. Association rules are commonly used for mining transactions in databases.

Here are some possible questions that association rules can answer:

- Which products tend to be purchased together?
- Of those customers who are similar to this person, what products do they tend to buy?
- Of those customers who have purchased this product, what other similar products do they tend to view or purchase?

### 3.3.1 Overview:

Below figure shows the general logic behind association rules. Given a large collection of transactions (depicted as three stacks of receipts in the figure), in which each transaction consists of one or more items, association rules go through the items being purchased to see what items are frequently bought together and to discover a list of rules that describe the purchasing behavior. The goal with association rules is to discover interesting relationships among the items. The relationships that are interesting depend both on the business context and the nature of the algorithm being used for the discovery.


**Figure 3.9 The general logic behind association rules**

Each of the uncovered rules is in the form X —> Y, meaning that when item X is observed, item Y is also observed. In this case, the left-hand side (LHS) of the rule is X, and the right-hand side (RH5) of the rule is Y.

Using association rules, patterns can be discovered from the data that allow the association rule algorithms to disclose rules of related product purchases. The uncovered rules are listed on the right side of Figure. The first three rules suggest that when cereal is purchased, 90% of the time milk is purchased also. When bread is purchased, 40% of the time milk is purchased also. When milk is purchased, 23% of the time cereal is also purchased.

In the example of a retail store, association rules are used over transactions that consist of one or more items. In fact, because of their popularity in mining customer transactions, association rules are

sometimes referred to as ***market basket analysis.*** Each transaction can be viewed as the shopping basket of a customer that contains one or more items. This is also known as an itemset. The term ***itemset*** refers to a collection of items or individual entities that contain some kind of relationship. This could be a set of retail items purchased together in one transaction, a set of hyperlinks clicked on by one user in a single session, or a set of tasks done in one day. An itemset containing ***k*** items is called a ***k-itemset*** denoted by {item1,item 2***,*** . . . item k}**.**

### 3.3.2 Apriori Algorithm:

The Apriori algorithm takes a bottom-up iterative approach to uncovering the frequent itemsets by first determining all the possible items (or 1-itemsets, for example ***{bread}, {eggs}, {milk},*** …) and then identifying which among them are frequent.

Assuming the minimum support threshold (or the minimum support criterion) is set at 0.5, the algorithm identifies and retains those itemsets that appear in at least 50% of all transactions and discards (or "prunes away") the itemsets that have a support less than 0.5 or appear in fewer than 50% of the transactions.

In the next iteration of the Apriori algorithm, the identified frequent 1-itemsets are paired into 2-itemsets (for example, {bread, ***eggs***}, {bread, ***milk}, {eggs, milk},***…) and again evaluated to identify the frequent 2-itemsets among them.

At each iteration, the algorithm checks whether the support criterion can be met; if it can, the algorithm grows the itemset, repeating the process until it runs out of support or until the itemsets reach a predefined length. Let variable $C_k$ be the set of candidate k-itemsets and variable $L_k$ be the set of k-itemsets that satisfy the minimum support. Given a transaction database D, a minimum support threshold □, and an optional parameter N indicating the maximum length an itemset could reach, Apriori iteratively computes frequent itemsets $L_{k+1}$ , based on $L_k$.

```
1  Apriori (D, δ, N)
2    k ← 1
3    L_k ← {1-itemsets that satisfy minimum support δ}
4    while L_k ≠ ∅
5      if ∄N ∨ (∃N ∧ k < N)
6        C_{k+1} ← candidate itemsets generated from L_k
7        for each transaction t in database D do
8          increment the counts of C_{k+1} contained in t
9        L_{k+1} ← candidates in C_{k+1} that satisfy minimum support δ
10       k ← k+1
11   return ⋃_k L_k
```

### 3.3.3 Evaluation of Candidate Rules:

Frequent itemsets from the previous section can form candidate rules such as X implies Y (X —> Y). *Confidence* is defined as the measure of certainty or trustworthiness associated with each discovered rule. Mathematically, confidence is the percent of transactions that contain both X and Y out of all the transactions that contain X

$$Confidence(X \rightarrow Y) = \frac{Support(X \Lambda Y)}{Support(X)}$$

For example, if ***{bread, eggs, milk}*** has a support of 0.15 and ***{bread, eggs}*** also has a support of 0.15, the confidence of rule ***{bread, eggs }->{ milk}*** is 1, which means 100% of the time a customer buys bread and eggs, milk is bought as well. The rule is therefore correct for 100% of the transactions containing bread and eggs.

A relationship may be thought of as interesting when the algorithm identifies the relationship with a measure of confidence greater than or equal to a predefined threshold. This predefined threshold is called the *minimum confidence.* A higher confidence indicates that the rule (X —> Y) is more interesting or more trustworthy, based on the sample dataset.

Even though confidence can identify the interesting rules from all the candidate rules, it comes with a problem. Given rules in the form of X -> Y, confidence considers only the antecedent (X) and the cooccurrence of X and Y; it does not take the consequent of the rule (Y) into concern. Therefore, confidence cannot tell if a rule contains true implication of the relationship or if the rule is purely coincidental. X and Y can be statistically independent yet still receive a high confidence score. Other measures such as lift and leverage are designed to address this issue.

*Lift* measures how many times more often X and Y occur together than expected if they are statistically independent of each other. Lift is a measure of how X and Y are really related rather than coincidentally happening together

$$Lift(X \rightarrow Y) = \frac{Support(X \Lambda Y)}{Support(X) * Support(Y)}$$

Lift is 1 if X and Y are statistically independent of each other. In contrast, a lift of X —> Y greater than 1 indicates that there is some usefulness to the rule. A larger value of lift suggests a greater strength of the association between X and Y.

Assuming 1,000 transactions, with (milk, eggs} appearing in 300 of them, {milk} appearing in 500, and {eggs} appearing in 400, then *Lift(milk—>eggs)* = 0.3/(0.5*0.4) = 1.5. If {bread} appears in 400 transactions and {milk, bread} appears in 400, then *Lift(milk —>bread)* = 0.4/(0.5*0.4) = 2. Therefore it can be concluded that milk and bread have a stronger association than milk and eggs.

*Leverage* is a similar notion, but instead of using a ratio, leverage uses the difference. Leverage measures the difference in the probability of X and Y appearing together in the dataset compared to what would be expected if X and Y were statistically independent of each other.

Leverage(X —> Y) = Support(X ∧Y ) - Support(X)* Support(Y)

In theory, leverage is 0 when X and Y are statistically independent of each other. If X and Y have some kind of relationship, the leverage would be greater than zero. A larger leverage value indicates a stronger relationship between X and Y. For the previous example, *Leverage{milk —> eggs)* = 0.3-(0.5*0.4) = 0.1 and *Leverage(milk -> bread*)=0.4 - (0.5 * 0.4) = 0.2. It again confirms that milk and bread have a stronger association than milk and eggs.

Confidence is able to identify trustworthy rules, but it cannot tell whether a rule is coincidental.

### 3.3.4 Applications of Association Rules:

The term *market basket analysis* refers to a specific implementation of association rules mining that many companies use for a variety of purposes, including these:

- Broad-scale approaches to better merchandising—what products should be included in or excluded from the inventory each month
- Cross-merchandising between products and high-margin or high-ticket items
- Physical or logical placement of product within related categories of products
- Promotional programs—multiple product purchase incentives managed through a loyalty card program

Besides market basket analysis, association rules are commonly used for recommender systems and clickstream analysis.

Many online service providers such as Amazon and Netflix use recommender systems. Recommender systems can use association rules to discover related products or identify customers who have similar interests. For example, association rules may suggest that those customers who have bought product A have also bought product B, or those customers who have bought products A, B, and C are more similar to this customer. These findings provide opportunities for retailers to cross-sell their products.

Clickstream analysis refers to the analytics on data related to web browsing and user clicks, which is stored on the client or the server side. Web usage log files generated on web servers contain huge amounts of information, and association rules can potentially give useful knowledge to web usage data analysts. For example, association rules may suggest that website visitors who land on page X click on links A, B, and C much more often than links D, E, and F. This observation provides valuable

insight on how to better personalize and recommend the content to site visitors.

### 3.3.5 Validation and Testing:

After gathering the output rules, it may become necessary to use one or more methods to validate the results in the business context for the sample dataset. The first approach can be established through statistical measures such as confidence, lift, and leverage. Rules that involve mutually independent items or cover few transactions are considered uninteresting because they may capture spurious relationships.

Confidence measures the chance that X and Y appear together in relation to the chance X appears. Confidence can be used to identify the interestingness of the rules.

Lift and leverage both compare the support of X and Y against their individual support. While mining data with association rules, some rules generated could be purely coincidental. For example, if 95% of customers buy X and 90% of customers buy Y, then X and Y would occur together at least 85% of the time, even if there is no relationship between the two.

Another set of criteria can be established through subjective arguments. Even with a high confidence, a rule may be considered subjectively uninteresting unless it reveals any unexpected profitable actions. For example, rules like {***paper}->{pencil***} may not be subjectively interesting or meaningful despite high support and confidence values. In contrast, a rule like ***{diaper}->{beer}*** that satisfies both minimum support and minimum confidence can be considered subjectively interesting because this rule is unexpected and may suggest a cross-sell opportunity for the retailer.

### 3.3.6 Diagnostics:

Although the Apriori algorithm is easy to understand and implement, some of the rules generated are uninteresting or practically useless. Additionally, some of the rules may be generated due to coincidental relationships between the variables. Measures like confidence, lift, and leverage should be used along with human insights to address this problem.

Another problem with association rules is that, in Phase 3 and *4* of the Data Analytics Lifecycle, the team must specify the minimum support prior to the model execution, which may lead to too many or too few rules. In related research, a variant of the algorithm can use a predefined target range for the number of rules so that the algorithm can adjust the minimum support accordingly.

Apriori algorithm is one of the earliest and the most fundamental algorithms for generating association rules. The Apriori algorithm reduces the computational workload by only examining itemsets that meet the specified minimum threshold. However, depending on the size of the dataset, the Apriori algorithm can be computationally expensive. For each level of support, the algorithm requires a scan of the entire database to obtain the result. Accordingly, as the database grows, it takes more time to compute in each run. Here are some approaches to improve Apriori's efficiency:

- **Partitioning:** Any itemset that is potentially frequent in a transaction database must be frequent in at least one of the partitions of the transaction database.
- **Sampling:** This extracts a subset of the data with a lower support threshold and uses the subset to perform association rule mining.
- **Transaction reduction:** A transaction that does not contain frequent fc-itemsets is useless in subsequent scans and therefore can be ignored.
- **Hash-based itemset counting:** If the corresponding hashing bucket count of a fc-itemset is below a certain threshold, the/c-itemset cannot be frequent.
- **Dynamic itemset counting:** Only add new candidate itemsets when all of their subsets are estimated to be frequent.

## Summary

In above chapter we study the different concept About different Analytical Theory and Methods taking the overview of methods k-means clustering and Given a collection of objects each with n measurable attributes Using R to Perform a K-means Analysis and algorithm we study all these thing in detail with diagram.

## Review Question

1. Explain Analytical Theory and Methods
2. Write in detail concept of K-means.
3. Write a short note on Diagnostics
4. Explain Units of Measure
5. What is mean by Additional Considerations
6. What are the Additional Algorithms

## REFERENCES


- Big Data and Analytics, Subhashini Chellappan Seema Acharya, Wiley First addition

- Data Analytics with Hadoop, *An Introduction for Data Scientists, Benjamin Bengfort and Jenny Kim*O'Reilly 2016

- Big Data and Hadoop V.K Jain KhannaPublishing First 2018

- https://towardsdatascience.com

- Data Science & Big Data Analytics Discovering, Analyzing, Visualizing and Presenting Data EMC Education Services Published by John Wiley & Sons, Inc


*****

# 4

# REGRESSION ANALYSIS

**Unit Structure**



## 4.0 OBJECTIVES

**To Study and Understand the following concept**

- Regression
- Linear Regression
- Logistic Regression
- Additional Regression Models.

**Regression Analysis:**

In general, regression analysis attempts to explain the influence that a set of variables has on the outcome of another variable of interest. Often, the outcome variable is called a *dependent variable* because the outcome depends on the other variables. These additional variables are sometimes called the *input variables* or the *independent variables.* Regression analysis is useful for answering the following kinds of questions:

- What is a person's expected income?
- What is the probability that an applicant will default on a loan?

Linear regression is a useful tool for answering the first question, and logistic regression is a popular method for addressing the second.

Regression analysis is a useful explanatory tool that can identify the input variables that have the greatest statistical influence on the

outcome. With such knowledge and insight, environmental changes can be attempted to produce more favorable values of the input variables. For example, if it is found that the reading level of 10-year-old students is an excellent predictor of the students' success in high school and a factor in their attending college, then additional emphasis on reading can be considered, implemented, and evaluated to improve students' reading levels at a younger age.

## 4.1 LINEAR REGRESSION

Linear regression is an analytical technique used to model the relationship between several input variables and a continuous outcome variable. A key assumption is that the relationship between an input variable and the outcome variable is linear. Although this assumption may appear restrictive, it is often possible to properly transform the input or outcome variables to achieve a linear relationship between the modified input and outcome variables.

A linear regression model is a probabilistic one that accounts for the randomness that can affect any particular outcome. Based on known input values, a linear regression model provides the expected value of the outcome variable based on the values of the input variables, but some uncertainty may remain in predicting any particular outcome.

### 4.1.1 Use Cases:

Linear regression is often used in business, government, and other scenarios. Some common practical applications of linear regression in the real world include the following:

- **Real estate:** A simple linear regression analysis can be used to model residential home prices as a function of the home's living area. Such a model helps set or evaluate the list price of a home on the market. The model could be further improved by including other input variables such as number of bathrooms, number of bedrooms, lot size, school district rankings, crime statistics, and property taxes

- **Demand forecasting:** Businesses and governments can use linear regression models to predict demand for goods and services. For example, restaurant chains can appropriately prepare for the predicted type and quantity of food that customers will consume based upon the weather, the day of the week, whether an item is offered as a special, the time of day, and the reservation volume. Similar models can be built to predict retail sales, emergency room visits, and ambulance dispatches.

- **Medical:** A linear regression model can be used to analyze the effect of a proposed radiation treatment on reducing tumor sizes. Input variables might include duration of a single radiation treatment, frequency of radiation treatment, and patient attributes such as age or weight.

### 4.1.2 Model Description:

As the name of this technique suggests, the linear regression model assumes that there is a linear relationship between the input variables and the outcome variable. This relationship can be expressed as shown in Equation

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots\ldots\ldots + \beta p_{-1} x\ p_{-1} + \cdot$$

where:
y is the outcome variable
$x_j$ are the input variables, for j=1,2,...,p-1
$\beta_0$ is the value of y when each $x_j$ equals zero
$\beta_j$ is the change in y based on a unit change in $x_j$ for j=1,2,...,p-1
$\cdot$ is a random error term that represents the difference in the linear model and a particular observed value for y.

Suppose it is desired to build a linear regression model that estimates a person's annual income as a function of two variables—age and education—both expressed in years. In this case, income is the outcome variable, and the input variables are age and education.

However, it is also obvious that there is considerable variation in income levels for a group of people with identical ages and years of education. This variation is represented by $\cdot$ in the model. So, in this example, the model would be expressed as shown in Equation.

$$\text{Income} = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Education} + \cdot$$

***Linear Regression Model (Ordinary Least Squares):***

In the linear model, the $\beta_2$s represent the unknown p parameters. The estimates for these unknown parameters are chosen so that, on average, the model provides a reasonable estimate of a person's income based on age and education. In other words, the fitted model should minimize the overall error between the linear model and the actual observations. Ordinary Least Squares (OLS) is a common technique to estimate the parameters.

To illustrate how OLS works, suppose there is only one input variable, x, for an outcome variable y. Furthermore, *n* observations of (x, *y)* are obtained and plotted in below Figure.

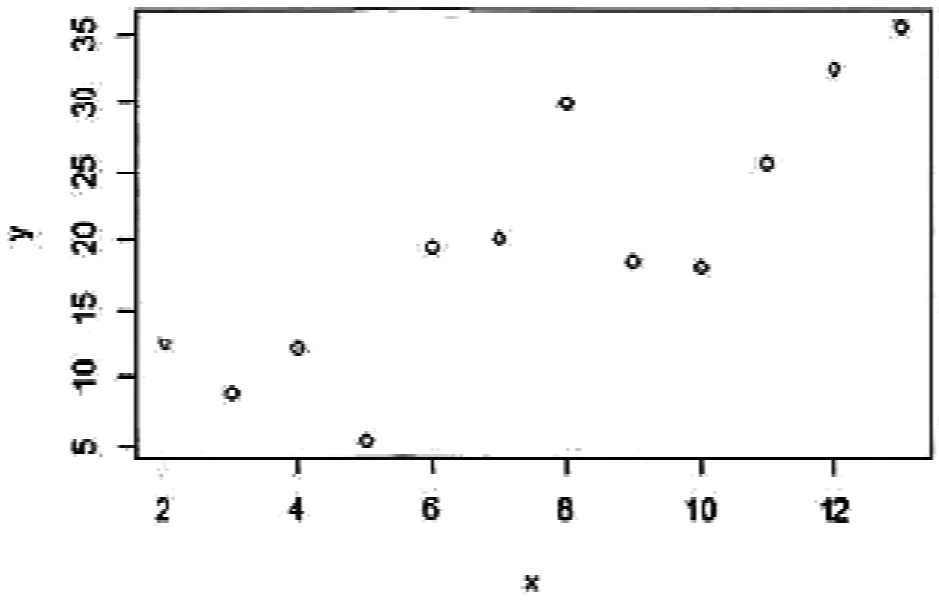



**Figure 4.1 Scatterplot of y versus x**

The goal is to find the line that best approximates the relationship between the outcome variable and the input variables. With OLS, the objective is to find the line through these points that minimizes the sum of the squares of the difference between each point and the line in the vertical direction. In other words, find the values of $\beta_0$ and $\beta_1$, such that the summation shown in Equation is minimized.

$$\sum_{i=1}^{n}\left[y_i-(\beta_0+\beta_1 x_i)\right]^2$$

The n individual distances to be squared and then summed are illustrated in below figure. The vertical lines represent the distance between each observed y value and the line $y = \beta_0 + \beta_1 x_1$

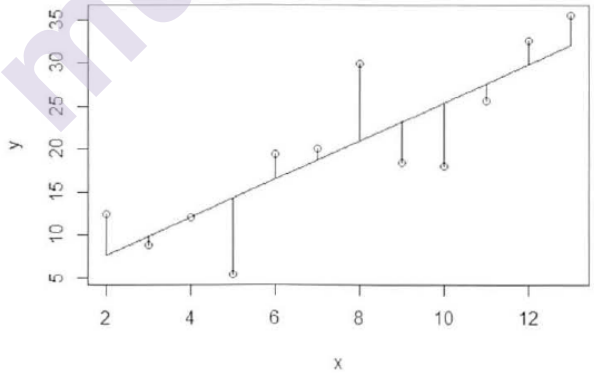



**Figure 4.2: Scatterplot of y versus x with vertical distance from the observed points to a fitted line**

**Linear Regression Model (with Normally Distributed Errors):**

In the normal model description, there were no assumptions made about the error term; no additional assumptions were necessary for OLS to provide estimates of the model parameters. However, in most linear regression analyses, it is common to assume that the error term is a

normally distributed random variable with mean equal to zero and constant variance. Thus, the linear regression model is expressed as shown in Equation.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots\ldots\ldots + \beta_{p-1} x_{p-1} + \cdot$$

where:
y is the outcome variable
$x_j$ are the input variables, for j=1,2,...,p-1
$\beta_0$ is the value of y when each $x_j$ equals zero
$\beta_j$ is the change in y based on a unit change in $x_j$ for j=1,2,...,p-1
$\cdot$ ~N(0,$\sigma^2$) and the $\cdot$ s are independent of each other

This additional assumptions yields the following result about the expected value of y, E(y) for given ($x_1$,$x_2$,…….$x_{p-1}$):

$$\begin{aligned} E(y) &= E\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots\ldots + \beta_{p-1} x_{p-1} + \varepsilon\right) \\ &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots\ldots + \beta_{p-1} x_{p-1} + E(\varepsilon) \\ &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots\ldots + \beta_{p-1} x_{p-1} \end{aligned}$$

Because $\beta_i$ and $x_i$ are constants, the E(y) is the value of the linear regression model for the given ($x_1$, $x_2$,... $x_{p-1}$). Furthermore, the variance of y, V(y), for given ($x_1$, $x_2$,... $x_p$) is this.

$$\begin{aligned} V(y) &= V\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{p-1} x_{p-1} + \varepsilon\right) \\ &= 0 + V(\varepsilon) = \sigma^2 \end{aligned}$$

Thus, for a given ($x_1$, $x_2$,... $x_{p-1}$), y is normally distributed with mean $\beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots\ldots\ldots + \beta_{p-1} x_{p-1}$ and variance $\sigma^2$. For a regression model with just one input variable, below figure illustrates the normality assumption on the error terms and the effect on the outcome variable, y, for a given value of x.

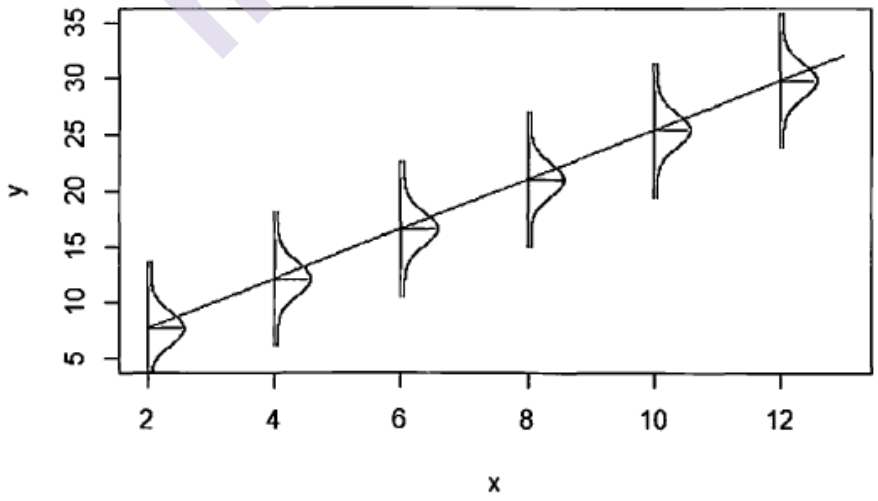



**Figure 4.3: Normal distribution about y for a given value of x**

### 4.1.3 Diagnostics:

The use of hypothesis tests, confidence intervals, and prediction intervals is dependent on the model assumptions being true. Following are Some tools and techniques that can be used to validate a fitted linear regression model.

### a. Evaluating the Linearity Assumption:

A major assumption in linear regression modeling is that the relationship between the input variables and the outcome variable is linear. The most fundamental way to evaluate such a relationship is to plot the outcome variable against each input variable. If the relationship between *Age* and *Income* is represented as illustrated in Figure 4.3, a linear model would not apply.

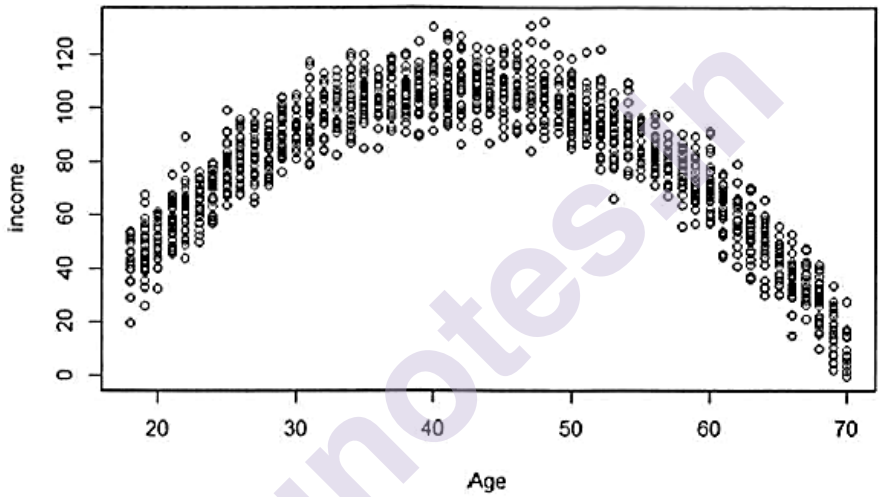



**Figure 4.3: Income as a quadratic function of Age**

In such a case, it is often useful to do any of the following:

- Transform the outcome variable.
- Transform the input variables.
- Add extra input variables or terms to the regression model.

Common transformations include taking square roots or the logarithm of the variables. Another option is to create a new input variable such as the age squared and add it to the linear regression model to fit a quadratic relationship between an input variable and the outcome.

### b. Evaluating the Residuals:

As stated previously, it is assumed that the error terms in the linear regression model are normally distributed with a mean of zero and a constant variance. If this assumption does not hold, the various inferences that were made with the hypothesis tests, confidence intervals, and prediction intervals are suspect.

**c. Evaluating the Normality Assumption:**

The residual plots are useful for confirming that the residuals were centered on zero and have a constant variance. However, the normality assumption still has to be validated.

**d. N-Fold Cross-Validation:**

To prevent overfitting a given dataset, a common practice is to randomly split the entire dataset into a training set and a testing set. Once the model is developed on the training set, the model is evaluated against the testing set. When there is not enough data to create training and testing sets, an N-fold cross-validation technique may be helpful to compare one fitted model against another. In N-fold cross-validation, the following occurs:

- The entire dataset is randomly split into N datasets of approximately equal size.
- A model is trained against N - 1 of these datasets and tested against the remaining dataset. A measure of the model error is obtained.
- This process is repeated a total of N times across the various combinations of N datasets taken N - 1 at a time. Recall:

$$\binom{N}{N-1} = N$$

- The observed N model errors are averaged over the N folds.

The averaged error from one mode! is compared against the averaged error from another model. This technique can also help determine whether adding more variables to an existing model is beneficial or possibly overfitting the data.

## 4.2 LOGISTIC REGRESSION

In linear regression modeling, the outcome variable is a continuous variable. When the outcome variable is categorical in nature, logistic regression can be used to predict the likelihood of an outcome based on the input variables. Although logistic regression can be applied to an outcome variable that represents multiple values, but we will examine the case in which the outcome variable represents two values such as true/false, pass/fail, or yes/no.

For example, a logistic regression model can be built to determine if a person will or will not purchase a new automobile in the next 12 months. The training set could include input variables for a person's age, income, and gender as well as the age of an existing automobile. The training set would also include the outcome variable on whether the person purchased a new automobile over a 12-month period. The logistic regression model provides the likelihood or probability of a person making a purchase in the next 12 months.

### 4.2.1 Use Cases:

The logistic regression model is applied to a variety of situations in both the public and the private sector.

Some common ways that the logistic regression model is used include the following:

- **Medical:** Develop a model to determine the likelihood of a patient's successful response to a specific medical treatment or procedure. Input variables could include age, weight, blood pressure, and cholesterol levels.

- ***Finance:*** Using a loan applicant's credit history and the details on the loan, determine the probability that an applicant will default on the loan. Based on the prediction, the loan can be approved or denied, or the terms can be modified.

- ***Marketing:*** Determine a wireless customer's probability of switching carriers (known as churning) based on age, number of family members on the plan, months remaining on the existing contract, and social network contacts. With such insight, target the high-probability customers with appropriate offers to prevent churn.

- ***Engineering:*** Based on operating conditions and various diagnostic measurements, determine the probability of a mechanical part experiencing a malfunction or failure. With this, probability estimate, schedule the appropriate preventive maintenance activity.

### 4.2.2 Model Description:

Logistic regression is based on the logistic function (y), as given by following equation

$$f(y) = \frac{e^r}{1+e^r} \; for -\infty < y < \infty \; \ldots\ldots\ldots\ldots\ldots..(4.7)$$

Note that as $y \to \infty, f(y) \to 1,$ and as $y \to -\infty, f(y) \to 0.$ So as Figure 4.4 illustrates, the value of the logistic function(y) varies from 0 to 1 y increases,

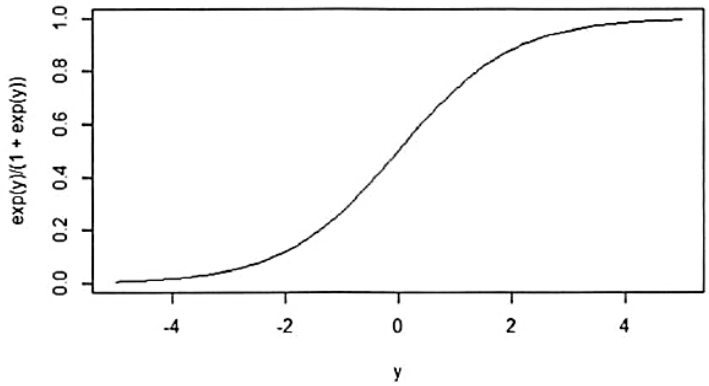



**Figure 4.4 The logistic function**

Because the range of f(y) is (0, 1), the logistic function appears to be an appropriate function to model the probability of a particular outcome occurring. As the value of y increases, the probability of the outcome occurring increases. In any proposed model, to predict the likelihood of an outcome, y needs to be a function of the input variables. In logistic regression, y is expressed as a linear function of the input variables. In other words, the formula shown in equation 4-8 applies

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 ..... + \beta_{p-1} x_{p-1} \text{ ……..4.8}$$

Then based on the input variables ($x_1$, $x_2$,……. $x_{p-1}$): the probability of an event is shown in equation 4.9

$$p(x_1, x_2, ....., x_{p-1}) = f(y) = \frac{e^y}{1+e^y} \text{ for } -\infty < y < \infty \text{ ……..4.9}$$

Equation 4.8 is comparable to Equation 4.1 used in linear regression modeling. However, one difference is that the values of y are not directly observed. Only the value of f(y) in terms of success or failure (typically expressed as 1 or 0, respectively) is observed.

Using p to denote f(y), Equation 6.9 can be written in the form provided equation 6-10

$$\text{In}\left(\frac{p}{1-p}\right) = y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ..... + \beta_{p-1} x_{p-1} \text{ ……..4.10}$$

The quantity in $\left(\frac{p}{1-p}\right)$, in equation 4.10 is known as log odds ratio or logit of p. Techniques such as Maximum Likelihood Estimation (MLE) are used to estimate the model parameters. MLE determines the values of the model parameters that maximize the chances of observing the given dataset.

**Customer Churn Example:**

A wireless telecommunications company wants to estimate the probability that a customer will churn (switch to a different company) in the next six months. With a reasonably accurate prediction of a person's likelihood of churning, the sales and marketing groups can attempt to retain the customer by offering various incentives. Data on 8,000 current and prior customers was obtained. The variables collected for each customer follow:

- *Age* (years)
- *Married*(true/false)
- *Duration* as a customer (years)
- *Churned_contacts*(count)—Number of the customer's contacts that have churned (count)
- *churned* (true/false)—Whether the customer churned

After analyzing the data and fitting a logistic regression model, *Age* and *Churned_contacts* were selected as the best predictor variables. Equation 6-11 provides the estimated model parameters.

$$y = 3.50 - 0.16 * Age + 0.38 * Churned\_contacts \qquad (6.11)$$

Using the fitted model from Equation 4-1, below table provides the probability of a customer churning based on the customer's age and the number of churned contacts.

| Customer | Age (Years) | Churned_Contacts | y | Prob. of Churning |
|---|---|---|---|---|
| 1 | 50 | 1 | –4.12 | 0.016 |
| 2 | 50 | 3 | –3.36 | 0.034 |
| 3 | 50 | 6 | –2.22 | 0.098 |
| 4 | 30 | 1 | –0.92 | 0.285 |
| 5 | 30 | 3 | –0.16 | 0.460 |
| 6 | 30 | 6 | 0.98 | 0.727 |
| 7 | 20 | 1 | 0.68 | 0.664 |
| 8 | 20 | 3 | 1.44 | 0.808 |
| 9 | 20 | 6 | 2.58 | 0.930 |

**Table 4.1 Estimated churn probabilities**

Based on the fitted model, there is a 93% chance that a 20-year-old customer who has had six contacts churn will also churn.

### 4.2.3 Diagnostics:

### Deviance and the Pseudo-$R^2$:

In logistic regression, deviance is defined to be -2*logL, where L is the maximized value of the likelihood function that was used to obtain the parameter estimates. In the R output, two deviance values are provided. The null deviance is the value where the likelihood function is based only on the intercept term ( $y = \beta_0$ ). The residual deviance is the value where the likelihood function is based on the parameters in the specified logistic model, shown in Equation 4-12

$$y = \beta_0 + \beta_1 * Age + \beta_2 * Churned\_contacts \ldots\ldots\ldots\ldots\ldots\ldots (4\text{-}12)$$

A metric analogous to $R^2$ in linear regression can be computed in equation 4-13

$$pseudo - R^2 = 1 - \frac{residual\ dev}{null\ \text{dev.}} = \frac{null\ \text{dev.} - res.dev.}{null\ \text{dev.}}$$

The pseudo-$R^2$is a measure of how well the fitted model explains the data as compared to the default model of no predictor variables and

only an intercept term. A pseudo-$R^2$ value near 1 indicates a good fit over the simple null model.

**Deviance and the Log-likelihood Ratio Test:**

In the pseudo-$R^2$ calculation, the -2 multiplier simply divide out. So, it may appear that including such multiplier does not provide a benefit.

However, the multiplier in the deviance definition is based on the log-likelihood test statistic shown in Equation 4-14.;

$$T = -2 * \log\left(\frac{L_{null}}{L_{alt}}\right)$$

$$= -2 * \log\left(L_{null}\right) - (-2) * \log\left(L_{alt}\right)$$

Where T is approximately Chi-squared distributed $\left(x^2_{p-1}\right)$ with

K degree off freedom (df) = $df_{null}$ - $df_{alternate}$

The previous description of the log-likelihood test statistic applies to any estimation using MLE. As can be seen in equation 4-15, in the logistic regression case,

$$T = null\ deviance - residdual\ deviance \sim x^2_{p-1} \quad \text{.....(4-14)}$$

Where *p* is a number of parameters in the fitted model
So, in the hypothesis test, a large value of T would indicate that the fitted model is significantly better than the null model that uses only the intercept term.

In the churn example, the log likelihood ratio statistic would be like this:
T= 8387.3-5359.2=3028.1 with 2 degrees of freedom and a corresponding p-value that is essentially zero.

So far, the log-likelihood ratio test discussion has focused on comparing a fitted model to the default model of using only the intercept. However, the log-likelihood ratio test can also compare one fitted model to another.

**Receiver Operating Characteristic (ROC) Curve:**

Logistic regression is often used as a classifier to assign class labels to a person, item, or transaction based on the predicted probability provided by the model. In the Churn example, a customer can be classified with the label called Churn if the logistic model predicts a high probability that the customer will churn. Otherwise, a Remain label is assigned to the customer. Commonly, 0.5 is used as the default probability threshold to distinguish between any two class labels. However, any threshold value

can be used depending on the preference to avoid false positives (for example, to predict Churn when actually the customer will Remain) or false negatives (for example, to predict Remain when the customer will actually Churn).

**Histogram of the Probabilities:**

It can be useful to visualize the observed responses against the estimated probabilities provided by the logistic regression. Figure 4-2 provides overlaying histograms for the customers who churned and for the customers who remained as customers. With a proper fitting logistic model, the customers who remained tend to have a low probability of churning. Conversely, the customers who churned have a high probability of churning again. This histogram plot helps visualize the number of items to be properly classified or mis- dassified. In the Churn example, an ideal histogram plot would have the remaining customers grouped at the left side of the plot, the customers who churned at the right side of the plot, and no overlap of these two groups.

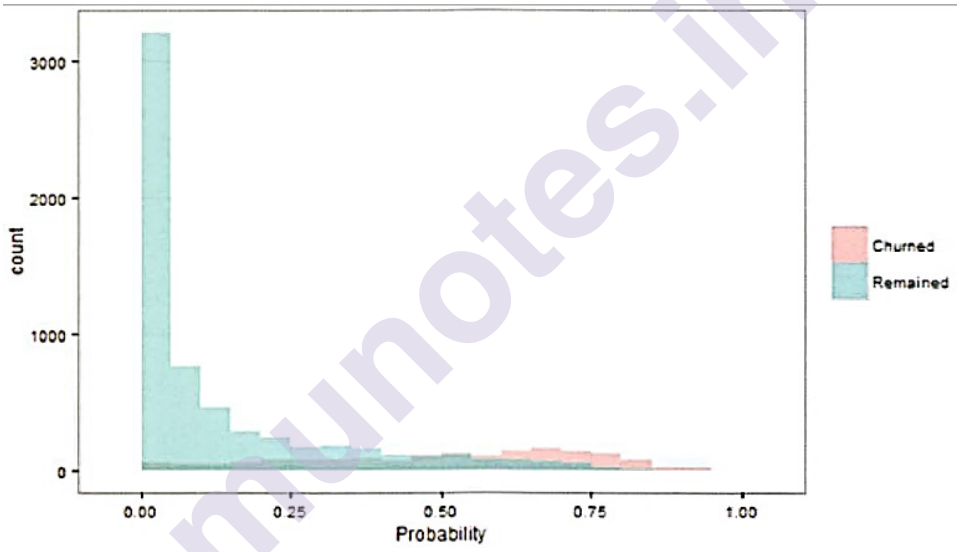



**Figure 4.2 : Customer counts versus estimated churn probability**

## 4.3 REASONS TO CHOOSE AND CAUTIONS

Linear regression is suitable when the input variables are continuous or discrete, including categorical data types, but the outcome variable is continuous. If the outcome variable is categorical, logistic regression is a better choice.

Both models assume a linear additive function of the input variables. If such an assumption does not hold true, both regression techniques perform poorly, Furthermore, in linear regression, the assumption of normally distributed error terms with a constant variance is important for many of the statistical inferences that can be considered. If

the various assumptions do not appear to hold, the appropriate transformations need to be applied to the data.

Although a collection of input variables may be a good predictor for the outcome variable, the analyst should not infer that the input variables directly cause an outcome. For example, it may be identified that those individuals who have regular dentist visits may have a reduced risk of heart attacks. However, simply sending someone to the dentist almost certainly has no effect on that person's chance of having a heart attack. It is possible that regular dentist visits may indicate a person's overall health and dietary choices, which may have a more direct impact on a person's health.

Use caution when applying an already fitted model to data that falls outside the dataset used to train the model. The linear relationship in a regression model may no longer hold at values outside the training dataset. For example, if income was an input variable and the values of income ranged from $35,000 to $90,000, applying the model to incomes well outside those incomes could result in inaccurate estimates and predictions.

If several of the input variables are highly correlated to each other, the condition is known as *multicollinearity.* Multicollinearity can often lead to coefficient estimates that are relatively large in absolute magnitude and may be of inappropriate direction (negative or positive sign). When possible, the majority of these correlated variables should be removed from the model or replaced by a new variable that is a function of the correlated variables.

## UNIT END QUESTIONS

1. What is clustering? Explain in detail. Also explain any two of its applications.
2. Describe the steps to find k clusters using k-means algorithm.
3. How to generalize the k-means algorithm? Also write a short note on determining the number of clusters.
4. Write a short note on association rules.
5. What is the role of support in apriori algorithm? Also explain how the Apriori property works with a neat diagram.
6. Find the associative rule using Apriori algorithm; if there are four transactions – T1, T2, T3 and T4 for itemsets {A,B,C},{A,C},{A,D} and {B,E,F} respectively and minimum support and confidence are 50 %.
7. What is Linear regression? Explain in detail. Also explain any two of its applications.

8. Write a short note on linear regression model. Also apply Ordinary least Squares (OLS) technique to estimate the parameters.

9. Explain Linear Regression Model with Normally Distributed Errors.

10. What is Logistic regression? Explain in detail. Also explain any two of its applications.

11. Describe logistic regression model with respect to logistic function.

## REFERENCES


- Big Data and Analytics, Subhashini Chellappan Seema Acharya, Wiley First addition
- Data Analytics with Hadoop, *An Introduction for Data Scientists, Benjamin Bengfort and Jenny Kim*O'Reilly 2016
- Big Data and Hadoop V.K Jain Khanna Publishing First 2018


*****

# UNIT III

# 5

# ANALYTICAL THEORY AND METHODS

**Unit Structure**





## 5.0 OBJECTIVES

- To study classification techniques used in data analytics
- To develop the understanding of decision trees
- To develop the understanding of Naïve Bayes
- To analyse the diagnostics of classifiers

## 5.1 DECISION TREES

A ***decision tree*** (also called ***prediction tree)*** uses a tree structure to specify sequences of decisions and consequences. Given input $X = \{x_1,x_2,...x_n\}$, the goal is to predict a response or output variable *Y*. Each member of the set $\{x_1,x_2,...x_n\}$ is called an ***input variable.*** The prediction can be achieved by constructing a decision tree with test points and branches. At each test point, a decision is made to pick a specific branch and traverse down the tree. Eventually, a final point is reached, and a prediction can be made. Due to its flexibility and easy visualization, decision trees are commonly deployed in data mining applications for classification purposes.

The input values of a decision tree can be categorical or

continuous. A decision tree employs a structure of test points (called ***nodes)*** and branches, which represent the decision being made. A node without further branches is called a ***leaf node.*** The leaf nodes return class labels and, in some implementations, they return the probability scores. A decision tree can be converted into a set of decision rules. In the following example rule, *income* and *mortgage_amount* are input variables, and the response is the output variable default with a probability score.

IF income <50,000 AND
mortgage_amount > 100K
THEN default = True WITH
PROBABILITY 75%

Decision trees have two varieties: *classification trees* and *regression trees.* Classification trees usually apply to output variables that are categorical—often binary—in nature, such as yes or no, purchase or not purchase, and so on. Regression trees, on the other hand, can apply to output variables that are numeric or continuous, such as the predicted price of a consumer good or the likelihood a subscription will be purchased.

### 5.1.1-Overview of a Decision Tree:

Figure 5-1 shows an example of using a decision tree to predict whether customers will buy a product. The term *branch* refers to the outcome of a decision and is visualized as a line connecting two nodes. If a decision is numerical, the "greater than" branch is usually placed on the right, and the "less than" branch is placed on the left. Depending on the nature of the variable, one of the branches may need to include an "equal to" component.

*Internal nodes* are the decision or test points. Each internal node refers to an input variable or an attribute. The top internal node is called the root. The decision tree in Figure 5-1 is a binary tree in that each internal node has no more than two branches. The branching of a node is referred to as a *split.*

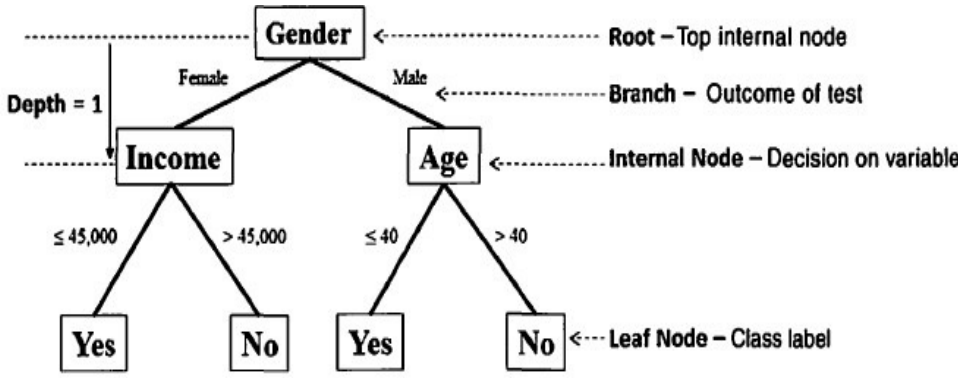



**Figure 5-1 Example of a decision tree**

The depth of a node is the minimum number of steps required to reach the node from the root. In Figure 5-1 for example, nodes Income and Age have a depth of one, and the four nodes on the bottom of the tree have a depth of two.

Leaf nodes are at the end of the last branches on the tree. They represent class labels—the outcome of all the prior decisions. The path from the root to a leaf node contains a series of decisions made at various internal nodes.

The decision tree in Figure 5-1 shows that females with income less than or equal to $45,000 and males 40 years old or younger are classified as people who would purchase the product. In traversing this tree, age does not matter for females, and income does not matter for males.

**Where decision tree is used?**

- Decision trees are widely used in practice.
- To classify animals, questions (like cold-blooded or warm-blooded, mammal or not mammal) are answered to arrive at a certain classification.
- A checklist of symptoms during a doctor's evaluation of a patient.
- The artificial intelligence engine of a video game commonly uses decision trees to control the autonomous actions of a character in response to various scenarios.
- Retailers can use decision trees to segment customers or predict response rates to marketing and promotions.
- Financial institutions can use decision trees to help decide if a loan application should be approved or denied. In the case of loan approval, computers can use the logical if - then statements to predict whether the customer will default on the loan.



### 5.1.2 The General Algorithm of Decision Tree :

In general, the objective of a decision tree algorithm is to construct a tree T from a training set S. If all the records in *S* belong to some class *C* ***(subscribed=***yes, for example), or if *S* is sufficiently pure (greater than a preset threshold), then that node is considered a leaf node and assigned the label *C*. The ***purity*** of a node is defined as its probability of the corresponding class.

In contrast, if not all the records in *S* belong to class *C* or if *S* is not sufficiently pure, the algorithm selects the next most informative attribute *A* (duration, marital, and so on) and partitions *S* according to A's values. The algorithm constructs subtrees *T1 T2*... for the subsets of *S* recursively until one of the following criteria is met:

- All the leaf nodes in the tree satisfy the minimum purity threshold.
- The tree cannot be further split with the preset minimum purity threshold.
- Any other stopping criterion is satisfied (such as the maximum depth of the tree).

The first step in constructing a decision tree is to choose the most informative attribute. A common way to identify the most informative attribute is to use entropy-based methods. The entropy methods select the most informative attribute based on two basic measures:

- *Entropy,* which measures the *impurity* of an attribute
- *Information gain,* which measures the *purity* of an attribute

Given a class X and its label $x \in x$, let P(x) be the probability of x. $H_x$ the entropy of X, is defined as shown in equation 5-1.

$$H_x = -\sum_{\forall x \in X} P(x) \log_2 P(x)$$

Equation 5-1 shows that entropy $H_x$ becomes 0 when all p(x) is 0 or 1. For a binary classification (True or false), $H_x$ is zero if p(x) is the probability of each label x is neither zero or one. On the other hand, $H_x$ achieves the maximum entropy when all the class lables are equally probable. For a binary classification, $H_x$ =1 if the probability of all class lables is 50/50. The maximum entropy increases as the number of possible outcome increases.

As an example of a binary random variable, consider tossing a coin with known, not necessarily fair, probabilities of coming up heads or tails. The corresponding entropy graph is shown in Figure 5-5. Let $x$ = 1 represent heads and $x$ = 0 represent tails. The entropy of the unknown result of the next toss is maximized when the coin is fair. That is, when heads and tai ls have equal probability $P(x = 1) = P(x = 0) = 0.5$, entropy $H_X$ = - (0.5 x log2 0.5 + x log2 0.5) = 1. On the other hand, if the coin is not fair, the probabilities of heads and tails would not be equal and there would be less uncertainty. As an extreme case, when the probability of tossing a head is equal to 0 or 1, the entropy is minimized to 0. Therefore, the entropy for a completely pure variable is 0 and is 1 for a set with equal occurrences for both the classes (head and tail, or yes and no)

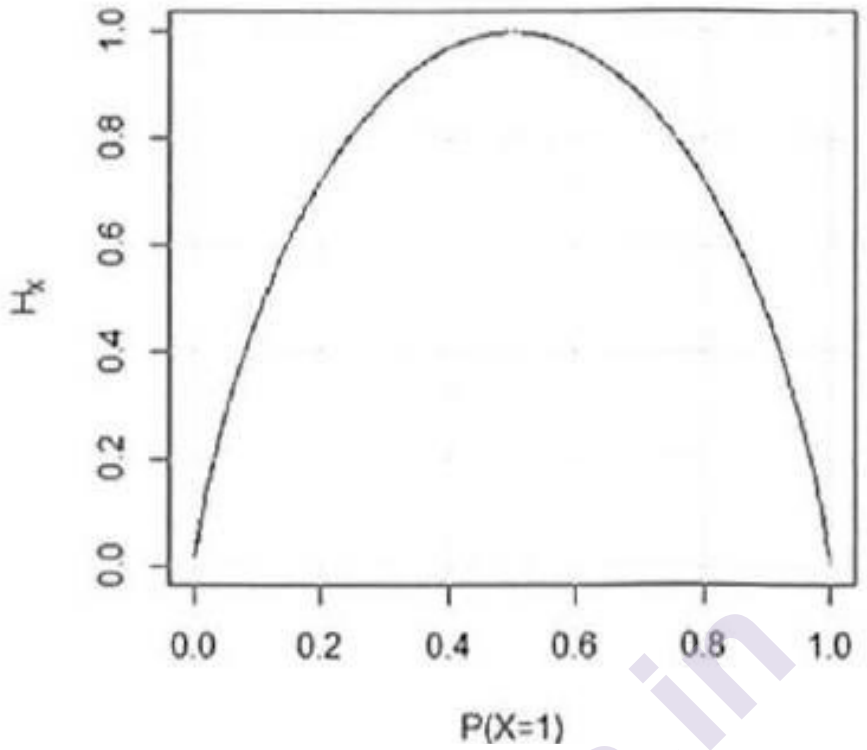



**Figure 5.5 Entropy of coin flips, where X=1 represents heads**

The next step is to identify the conditional entropy for each attribute. Given an attribute X, its value x, its outcome Y, and its value y, conditional entropy $H_{y/x}$ is the remaining entropy of Y given X, formally defined as shown in Equation 5.2.

$$H_{y/x} = \sum_{X} P(x) H(Y/X = x)$$
$$= -\sum_{\forall x \in X} P(x) \sum_{\forall y \in Y} P(Y/X) \log_2 P(Y/X) \ldots\ldots\ldots\ldots\ldots \text{(5-2)}$$

The information gain of an attribute A is defined as the difference between the base entropy and the conditional entropy of the attribute, as shown in figure 5-3.

$$InfoGain_A = H_s - H_{S/A} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(5-3)}$$

Information gain compares the degree of purity of the parent node before a split with the degree of purity of the child node after a split. At each split, an attribute with the greatest information gain is considered the most informative attribute. Information gain indicates the purity of an attribute.

### 5.1.3 Decision Tree Algorithms:

Multiple algorithms exist to implement decision trees, and the methods of tree construction vary with different algorithms. Some popular algorithms include ID3.

***ID3 Algorithm :***

ID3 (or Iterative Dichotomiser 3) is one of the first decision tree algorithms, and it was developed by John Ross Quinlan. Let A be a set of categorical input variables, P be the output variable (or the predicted class), and *T* be the training set. The ID3 algorithm is shown here.

```
1   ID3 (A, P, T)
2    if T∈φ
3      return φ
4    if all records in T have the same value for P
5      return a single node with that value
6    if A∈φ
7      return a single node with the most frequent value of P in T
8    Compute information gain for each attribute in A relative to T
9    Pick attribute D with the largest gain
10   Let {d1,d2...dm} be the values of attribute D
11   Partition T into {T1,T2...Tm} according to the values of D
12   return a tree with root D and branches labeled d1,d2...dm
           going respectively to trees ID3(A-{D}, P, T1),
           ID3(A-{D}, P, T2), . . . ID3(A-{D}, P, Tm)
```

### 5.1.4 Evaluating a Decision Tree:

Decision trees use *greedy algorithms,* in that they always choose the option that seems the best available at that moment. At each step, the algorithm selects which attribute to use for splitting the remaining records. This selection may not be the best overall, but it is guaranteed to be the best at that step. This characteristic reinforces the efficiency of decision trees. However, once a bad split is taken, it is propagated through the rest of the tree. To address this problem, an ensemble technique (such as random forest) may randomize the splitting or even randomize data and come up with a multiple tree structure, these trees then vote for each class, and the class with the most votes is chosen as the predicted class.

There are a few ways to evaluate a decision tree, First, evaluate whether the splits of the tree make sense. Conduct sanity checks by validating the decision rules with domain experts, and determine if the decision rules are sound.

Having too many layers and obtaining nodes with few members might be signs of overfitting. In overfitting, the model fits the training set well, but it performs poorly on the new samples in the testing set. For decision tree learning, overfitting can be caused by either the lack of training data or the biased data in the training set. Two approaches can help avoid overfitting in decision tree learning.

- Stop growing the tree early before it reaches the point where all the training data is perfectly classified.
- Grow the full tree, and then post-prune the tree with methods such as reduced-error pruning and rule- based post pruning.

Decision trees are computationally inexpensive, and it is easy to classify the data. The outputs are easy to interpret as a fixed sequence of simple tests. Decision trees are able to handle both numerical and categorical attributes and are robust with redundant or correlated variables. Decision trees can handle categorical attributes with many distinct values, such as country codes for telephone numbers. Decision trees can also handle variables that have a nonlinear effect on the outcome, so they work better than linear models (for example, linear regression and logistic regression) for highly nonlinear problems.

The structure of a decision tree is sensitive to small variations in the training data. Although the dataset is the same, constructing two decision trees based on two different subsets may result in very different trees. If a tree is too deep, overfitting may occur, because each split reduces the training data for subsequent splits.

Decision trees are not a good choice if the dataset contains many irrelevant variables. This is different from the notion that they are robust with redundant variables and correlated variables. If the dataset contains redundant variables, the resulting decision tree ignores all but one of these variables because the algorithm cannot detect information gain by including more redundant variables. On the other hand, if the dataset contains irrelevant variables and if these variables are accidentally chosen as splits in the tree, the tree may grow too large and may end up with less data at every split, where overfitting is likely to occur. To address this problem, feature selection can be introduced in the data preprocessing phase to eliminate the irrelevant variables.

Although decision trees are able to handle correlated variables, decision trees are not well suited when most of the variables in the training set are correlated, since overfitting is likely to occur. To overcome the issue of instability and potential overfitting of deep trees, one can combine the decisions of several randomized shallow decision trees—the basic idea of another classifier called random forest or use ensemble methods to combine several weak learners for better classification.

For binary decisions, a decision tree works better if the training dataset consists of records with an even probability of each result. In other words, the root of the tree has a 50% chance of either classification. This occurs by randomly selecting training records from each possible classification in equal numbers.

When using methods such as logistic regression on a dataset with many variables, decision trees can help determine which variables are the most useful to select based on information gain. Then these variables can be selected for the logistic regression. Decision trees can also be used to prune redundant variables.

## 5.2 NAIVE BAYES

Naive Bayes is a probabilistic classification method based on Bayes' theorem. Bayes' theorem gives the relationship between the probabilities of two events and their conditional probabilities.

A naive Bayes classifier assumes that the presence or absence of a particular feature of a class is unrelated to the presence or absence of other features. For example, an object can be classified based on its attributes such as shape, color, and weight.

The input variables are generally categorical, but variations of the algorithm can accept continuous variables, There are also ways to convert continuous variables into categorical ones. This process is often referred to as the *discretization of continuous variables.* For an attribute such as *income,* the attribute can be converted into categorical values as shown below.

- **Low Income:** income **<** $10,000
- ***Working Class:*** $10,000 < income < $50,000
- ***Middle Class:*** $50,000 < income < $1,000,000
- ***Upper Class:*** income >$1,000,000

The output typically includes a class label and its corresponding probability score. The probability score is not the true probability of the class label, but it's proportional to the true probability.

Because naive Bayes classifiers are easy to implement and can execute efficient. Spam filtering is a classic use case of naive Bayes text classification. Bayesian spam filtering has become a popular mechanism to distinguish spam e-mail from legitimate e-mail.

Naive Bayes classifiers can also be used for fraud detection. In the domain of auto insurance, for example, based on a training set with attributes such as driver's rating, vehicle age, vehicle price, historical claims by the policy holder, police report status, and claim genuineness, naive Bayes can provide probability- based classification of whether a new claim is genuine.

## 5.3 BAYES' THEOREM

The *conditional probability* of event C occurring, given that event A has already occurred, is denoted as P(C|A*)*, which can be found using the formula in Equation 5-6.

$$\mathrm{P}(C/A) = \frac{P(A \cap C)}{\mathrm{P}(\mathrm{A})} \ldots\ldots\ldots\ldots\ldots\ldots (5\text{-}6)$$

Equation 5-7 can be obtained with some minor algebra and

substitution of the conditional probability

$$\mathrm{P}(C/A) = \frac{P(A/C).P(C)}{\mathrm{P}(\mathrm{A})} \quad \ldots\ldots\ldots\ldots\ldots(5\text{-}7)$$

Where c is the class label $C \in \{c_1, c_2, .....c_n\}$ and A is observed attributes $A = \{a_1, a_2, .....a_m\}$ Equation 5-7 is the most common form of the Baye's theorem.

Mathematically, Bayes' theorem gives the relationship between the probabilities of C and A, P(C) and P(A), and the conditional probabilities of C given A and A, given C, namely P(C/A) and P(A/C)

## 5.4 DIAGNOSTICS

Unlike logistic regression, naive Bayes classifiers can handle missing values. Naive Bayes is also robust to irrelevant variables—variables that are distributed among all the classes whose effects are not pronounced.

The model is simple to implement even without using libraries. The prediction is based on counting the occurrences of events, making the classifier efficient to run. Naive Bayes is computationally efficient and is able to handle high-dimensional data efficiently. .In some cases naive Bayes even outperforms other methods. Unlike logistic regression, the naive Bayes classifier can handle categorical variables with many levels. Recall that decision trees can handle categorical variables as well, but too many levels may result in a deep tree. The naive Bayes classifier overall performs better than decision trees on categorical values with many levels. Compared to decision trees, naive Bayes is more resistant to overfitting, especially with the presence of a smoothing technique.

One problem of the Laplace smoothing is that it may assign too much probability to unseen events. To address this problem, Laplace smoothing can be generalized to use $\varepsilon$ instead of 1, where typically $\varepsilon \in [0,1]$ see equation 5-8.

$$\mathrm{P}``(x) = \frac{count(x) + \varepsilon}{\sum_x [count(x) + \varepsilon]} \quad \ldots\ldots\ldots\ldots(5.8)$$

Smoothing techniques are available in most standard software packages for native Bayes classifiers. However, if for some reason (like performance concerns) the native Bayes classifiers needs to be coded directly into an application, the smoothing and logarithm calculations should be incorporated into the implementation.

Despite the benefits of naive Bayes, it also comes with a few disadvantages. Naive Bayes assumes the variables in the data are conditionally independent. Therefore, it is sensitive to correlated variables

because the algorithm may double count the effects. As an example, assume that people with low income and low credit tend to default. If the task is to score "default" based on both income and credit as two separate attributes, naive Bayes would experience the double-counting effect on the default outcome, thus reducing the accuracy of the prediction.

Although probabilities are provided as part of the output for the prediction, naive Bayes classifiers in general are not very reliable for probability estimation and should be used only for assigning class labels. Naive Bayes in its simple form is used only with categorical variables. Any continuous variables should be converted into a categorical variable with the process known as discretization.

## 5.5 DIAGNOSTICS OF CLASSIFIERS

Classifiers methods can be used to classify instances into distinct groups according to the similar characteristics they share. Each of these classifiers faces the same issue: how to evaluate if they perform well. A few tools have been designed to evaluate the performance of a classifier. Such tools are not limited to the three classifiers but rather serve the purpose of assessing classifiers in general.

A *confusion matrix* is a specific table layout that allows visualization of the performance of a classifier. Table 5-6 shows the confusion matrix for a two-class classifier. *True positives* (TP) are the number of positive instances the classifier correctly identified as positive. *False positives* (FP) are the number of instances in which the classifier identified as positive but in reality are negative. *True negatives* (TN) are the number of negative instances the classifier correctly identified as negative, *False negatives* (FN) are the number of instances classified as negative but in reality are positive. In a two-class classification, a preset threshold may be used to separate positives from negatives. TP and TN are the correct guesses. A good classifier should have large TP and TN and small (ideally zero) numbers for FP and FN.

| | | **Predicted class** | |
|---|---|---|---|
| | | *P* | *N* |
| **Actual Class** | *P* | True Positives (TP) | False Negatives (FN) |
| | *N* | False Positives (FP) | True Negatives (TN) |

The accuracy (or the overall success rate) is a metric defining the rate at which a model has classified the records correctly. It is defined as the sum of TP and TN divided by the total number of instances, as shown in Equation 4.9.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \ldots\ldots\ldots\ldots (4.9)$$

A good model should have a high accuracy score, but having a high accuracy score alone does not guarantee the model is well established. The ***true positive rate*** (TPR) shows what percent of positive instances the classifier correctly identified. It's also illustrated in Equation 4.10

$$TPR = \frac{TP}{TP + FN} \ldots\ldots\ldots\ldots$$

The false positive rate (FPR) shows what percent of negatives. The classifier marked as positive. The FPR is also called the false alarm rate or the type one 1 error rate and is shown in equation 7-20

$$FPR = \frac{FP}{FP + TN} \ldots\ldots\ldots\ldots$$

The false negative rate (FNR) shows what percent of positives the classifier marked as negatives. It is also known as the miss rate type II error rate and is shown in equation 7-21. Note that the sum of TPR and FNR is 1

$$FNR = \frac{FN}{TP + FN}$$

A well-performed model should have a high TPR that is ideally 1 and a low FPR and FNR that are ideally 0. In some cases, a model with a TPR of 0.95 and an FPR of 0.3 is more acceptable than a model with a TPR of 0.9 and an FPR of 0.1 even if the second model is more accurate overall. *Precision* is the percentage of instances marked positive that really are positive, as shown in Equation 7-22.

$$\Pr ecision = \frac{TP}{TP + FP}$$

ROC curve is a common tool to evaluate classifiers. The abbreviation stands for receiver operating characteristic, a term used in signal detection to characterize the trade-off between hit rate and false-alarm rate over a noisy channel. A ROC curve evaluates the performance of a classifier based on the TP and FP, regardless of other factors such as class distribution and error costs.

Related to the ROC curve is the area under the curve (AUC). The AUC is calculated by measuring the area under the ROC curve. Higher AUC scores mean the classifier performs better. The score can range from 0.5 (for the diagonal line TPR=FPR) to 1.0 (with ROC passing through the top-left corner).

## 5.6 ADDITIONAL CLASSIFICATION METHODS

Besides the two classifiers introduced in this chapter, several other methods are commonly used for classification, including bagging, boosting, random forest, and support vector machines (SVM).

Bagging (or bootstrap aggregating) uses the bootstrap technique that repeatedly samples with replacement from a dataset according to a uniform probability distribution. "With replacement" means that when a sample is selected for a training or testing set, the sample is still kept in the dataset and may be selected again. Because the sampling is with replacement, some samples may appear several times in a training or testing set, whereas others may be absent. A model or base classifier is trained separately on each bootstrap sample, and a test sample is assigned to the class that received the highest number of votes.

Similar to bagging, boosting (or AdaBoost) uses votes for classification to combine the output of individual models. In addition, it combines models of the same type. However, boosting is an iterative procedure where a new model is influenced by the performances of those models built previously. Furthermore, boosting assigns a weight to each training sample that reflects its importance, and the weight may adaptively change at the end of each boosting round. Bagging and boosting have been shown to have better performances [S] than a decision tree.

Random forest is a class of ensemble methods using decision tree classifiers. It is a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. A special case of random forest uses bagging on decision trees, where samples are randomly chosen with replacement from the original training set.

SVM is another common classification method that combines linear models with instance-based learning techniques. Support vector machines select a small number of critical boundary instances called support vectors from each class and build a linear decision function that separates them as widely as possible, SVM by default can efficiently perform linear classifications and can be configured to perform nonlinear classifications as well.

## SUMMARY

- A decision tree (also called prediction tree) uses a tree structure to specify sequences of decisions and consequences. Given input X = {x1,x2,...xn}, the goal is to predict a response or output variable Y. Each member of the set {x1,x2,...xn} is called an input variable.
- Internal nodes are the decision or test points. Each internal node refers to an input variable or an attribute
- The objective of a decision tree algorithm is to construct a tree T from a training set S.
- ID3 (or Iterative Dichotomiser 3) is one of the first decision tree algorithms, and it was developed by John Ross Quinlan
- Decision trees use greedy algorithms, in that they always choose the option that seems the best available at that moment. At each step, the algorithm selects which attribute to use for splitting the remaining records.
- Decision trees are computationally inexpensive, and it is easy to classify the data. The outputs are easy to interpret as a fixed sequence of simple tests.
- Although decision trees are able to handle correlated variables, decision trees are not well suited when most of the variables in the training set are correlated, since overfitting is likely to occur.
- Naive Bayes is a probabilistic classification method based on Bayes' theorem. Bayes' theorem gives the relationship between the probabilities of two events and their conditional probabilities.
- Naive Bayes is also robust to irrelevant variables—variables that are distributed among all the classes whose effects are not pronounced.
- Naive Bayes assumes the variables in the data are conditionally independent. Therefore, it is sensitive to correlated variables because the algorithm may double count the effects
- A confusion matrix is a specific table layout that allows visualization of the performance of a classifier
- Bagging (or bootstrap aggregating) uses the bootstrap technique that repeatedly samples with replacement from a dataset according to a uniform probability distribution.

## UNIT END QUESTIONS

A. Where decision tree is used?

B. Explain the General Algorithm of Decision Tree.

C. Write down the ID3 Algorithm.

D. Explain the Bayes' Theorem.

E. What are Diagnostics of Classifiers?

F. Give a brief account Classification Methods used in data analytics.

## REFERENCES


- Data Analytics with Hadoop -An Introduction for Data Scientists by Benjamin Bengfort and Jenny Kim
- The Data Revolution: Big Data, Open Data, Data Infrastructures, And Their Consequences By Rob Kitchin
- Big Data: Principles and Best Practices of Scalable Real-Time Data Systems By Nathan Marz And James Warren
- Big Data Demystified By David Stephenson


*****

6

# TIME SERIES AND TEXT ANALYSIS

**Unit Structure**





## 6.0 OBJECTIVES

- To analyse and interpret Time series data
- To understand Box-Jenkins Methodology and its applications to time series data
- To study the applications of Autoregressive Integrated Moving Average methodology
- To analyse the various techniques used in text analysis
- To learn how to categorizing documents by topics

## 6.1 OVERVIEW OF TIME SERIES ANALYSIS

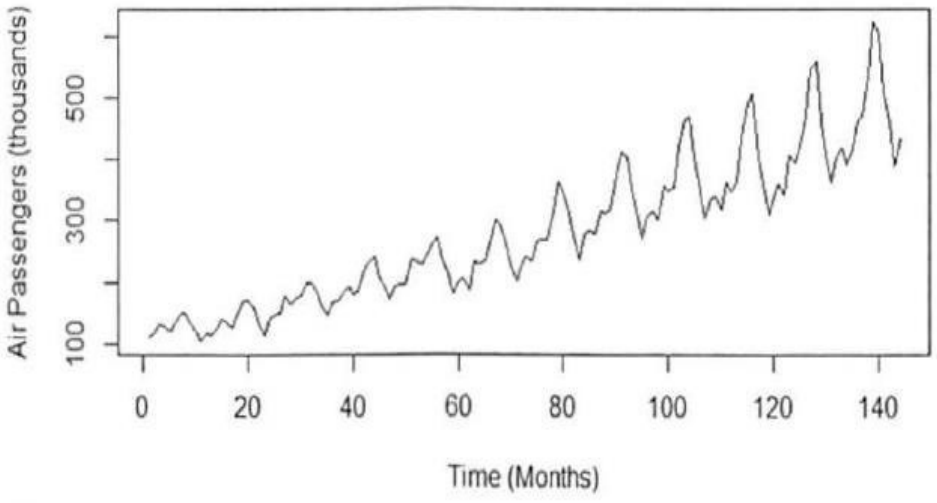



**Figure: 6.1 Monthly international airline passengers**

Time series analysis attempts to model the underlying structure of observations taken over time, A time series, denoted Y = a + bX , is an ordered sequence of equally spaced values over time. For example, Figure 6-1 provides a plot of the monthly number of international airline passengers over a 12-year period. In this example, the time series consists of an ordered sequence of 144 values.

Following are the goals of time series analysis:

- Identify and model the structure of the time series.
- Forecast future values in the time series.

  Time series analysis has many applications in finance, economics, biology, engineering, retail, and manufacturing. Here are a few specific use cases:

- **Retail sales:** For various product lines, a clothing retailer is looking to forecast future monthly sales. These forecasts need to account for the seasonal aspects of the customer's purchasing decisions.
- **Spare parts planning:** Companies' service organizations have to forecast future spare part demands to ensure an adequate supply of parts to repair customer products. To forecast future demand, complex models for each part number can be built using input variables such as expected part failure rates, service diagnostic effectiveness, forecasted new product shipments, and forecasted trade-ins/decommissions.
- **Stock trading:** Some high-frequency stock traders utilize a technique called pairs trading. In pairs trading, an identified strong positive correlation between the prices of two stocks is used to detect a market opportunity. Suppose the stock prices of Company A and Company B consistently move together. Time series analysis can be applied to the difference of these companies' stock prices over time. A statistically

larger than expected price difference indicates that it is a good time to buy the stock of Company A and sell the stock of Company B, or vice versa.

### 6.1.1 Box-Jenkins Methodology:

A time series consists of an ordered sequence of equally spaced values over time. Examples of a time series are monthly unemployment rates, daily website visits, or stock prices every second. A time series can consist of the following components:

- Trend
- Seasonality
- Cyclic
- Random

The **trend** refers to the long-term movement in a time series. It indicates whether the observation values are increasing or decreasing over time. Examples of trends are a steady increase in sales month over month oran annual decline of fatalities due to car accidents.

The **seasonality** component describes the fixed, periodic fluctuation in the observations over time. As the name suggests, the seasonality component is often related to the calendar. For example, monthly retail sales can fluctuate over the year due to the weather and holidays.

A **cyclic** component also refers to a periodic fluctuation, but one that is not as fixed as in the case of a seasonality component. For example, retails sales are influenced by the general state of the economy. Thus, a retail sales time series can often follow the lengthy boom-bust cycles of the economy.

Although noise is certainly part of this **random** component, there is often some underlying structure to this random component that needs to be modeled to forecast future values of a given time series.

The Box-Jenkins methodology for time series analysis involves the following three main steps:

**1)** Condition data and select a model.

   a. Identify and account for any trends or seasonality in the time series,

   b. Examine the remaining time series and determine a suitable model.

**2)** Estimate the model parameters.

**3)** Assess the model and return to Step 1, if necessary.

## 6.2 ARIMA MODEL (AUTOREGRESSIVE INTEGRATED MOVING AVERAGE)

As stated in the first step of the Box-Jenkins methodology, it is necessary to remove any trends or seasonality in the time series. This step is necessary to achieve a time series with certain properties to which autoregressive and moving average models can be applied. Such a time series is known as a stationary time series. A stationary time series is one whose properties do not depend on the time at which the series is observed. A time series, yt, for t = 1,2,3,...,, is a **stationary time series** if the following three conditions are met:

(a) The expected value (mean) of yt, is a constant for all values of t.

(b) The variance of yt, is finite.

(c) The covariance of yt and yt+h depends only on the value of h= 0,1,2, ...for all t.

The covariance of yt and yt+h is a measure of how the two variables, yt and yt+h vary together. It is expressed in Equation 8-1.

$$\text{cov}(y_t , y_{t+h}) = E[(y_t - \mu_t)(y_{t+h} - \mu_{t+h})] \quad (8.1)$$

If two variables are independent of each other, their covariance is zero. If the variables change together in the same direction, the variables have a positive covariance. Conversely, if the variables change together in the opposite direction, the variables have a negative covariance.

For a stationary time series, by condition (a), the mean is a constant, say $\mu$. So, for a given stationary sequence, yt, the covariance notation can be simplified to what's shown in Equation 6-2.

$$\text{cov}(h) = E[(y_t - \mu)(y_{t+h} - \mu)] \quad (8.2)$$

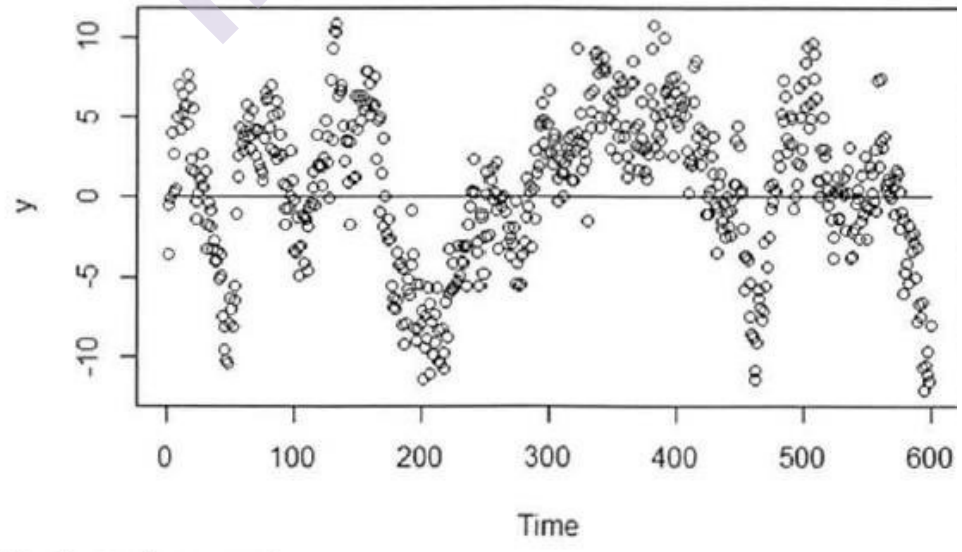


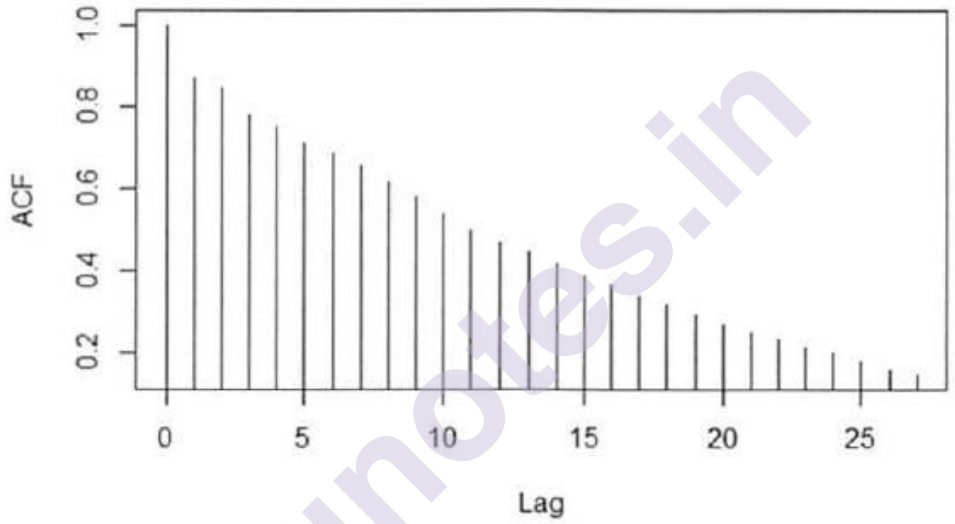

**Figure 6.2 A plot of a stationary series**

So the constant variance coupled with part (a), $E[y_t]=\mu$, for all t and some constant $\mu$, suggests that a stationary time series can look like Figure 6-2. In this plot, the points appear to be centered about a fixed constant, zero, and the variance appears to be somewhat constant over time.

### 6.2.1 Autocorrelation Function (Acf):

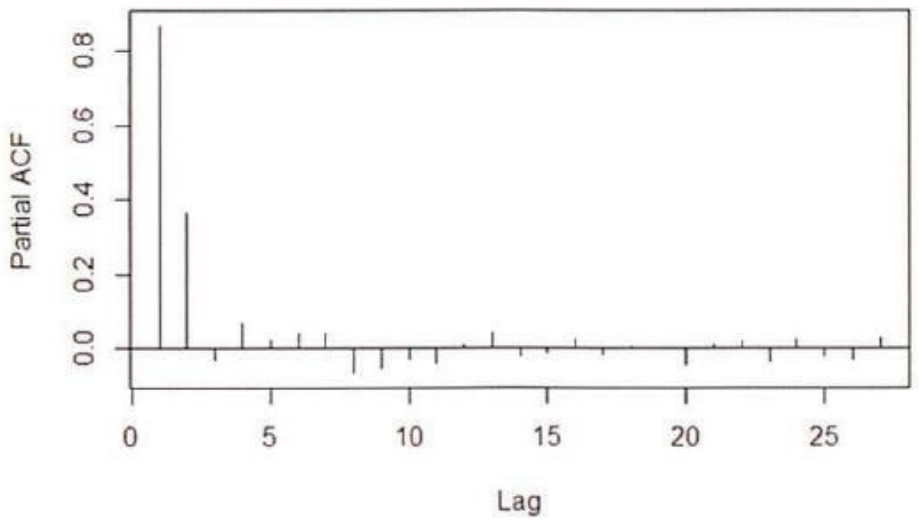
The plot of *autocorrelation function (ACF)* provides insight into the covariance of the variables in the time series and its underlying structure. For a stationary time series, the ACF is defined as shown in Equation 8-4

$$ACF(h)=\frac{cov(y_t, y_{t+h})}{\sqrt{cov(y_t, y_t)\,cov(y_{t+h}, y_{t+h})}}=\frac{cov(h)}{cov(0)} \qquad (8\text{-}4)$$




Because the cov(0) is the variance, the ACF is analogous to the correlation function of two variables, corr*(yt , yt+h)*, and the value of the ACF falls between -1 and 1.Thus, the closer the absolute value of ACF(h) is to 1, the more useful yt can be as a predictor of yt+h. The plot of the ACF is provided in Figure 8-3 for stationary time series.

By convention, the quantity *h* in the ACF is referred to as the *lag,* the difference between the time points t and *t +h.* At lag 0, the ACF provides the correlation of every point with itself. So ACF(0) always equals 1. According to the ACF plot, at lag 1 the correlation between *y, and* yt-1 is approximately 0.9, which is very close to 1. So yt-1 appears to be a good predictor of the value of y.




### 6.2.1 Moving Average Models:

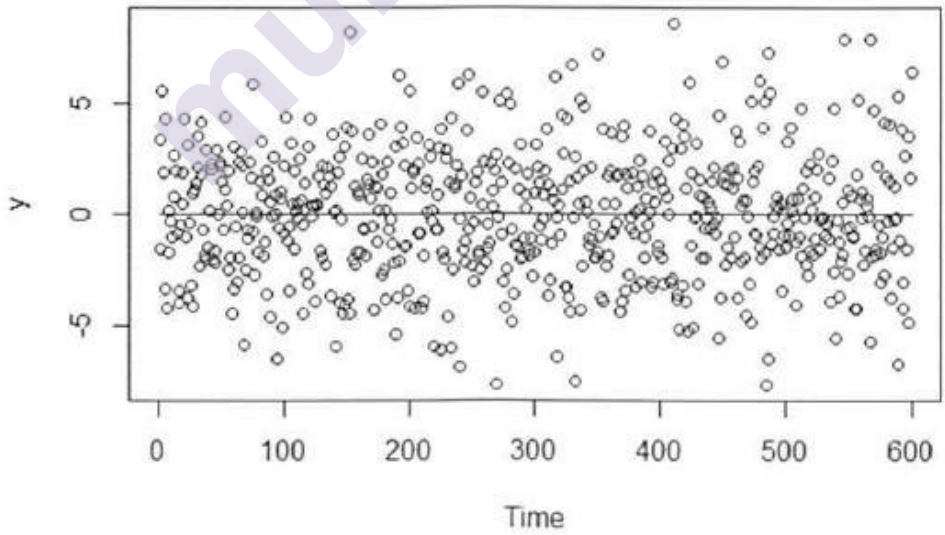
Fora time series, yt centered at zero, a moving average model of order q, denoted MA(q), is expressed as shown in Equation 8-9.

$y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q}$ (8.9)

Where $\theta_k$ is a constant for k = 1, 2…..,q $\theta_q \neq 0$

$\varepsilon_t \sim N(0,\sigma^2)$ for all t

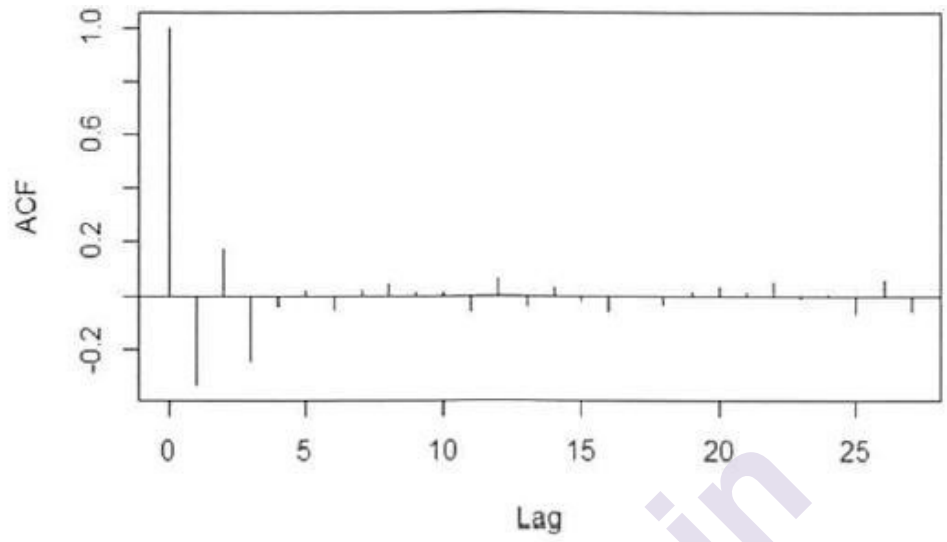
In an MA(q) model, the value of a time series is a linear combination of the current white noise term and the prior q white noise terms. So earlier random shocks directly affect the current value of the time series. For MA(q) models, the behavior of the ACF and PACF plots are somewhat swapped from the behavior of these plots for AR(p) models. Fora simulated MA(3) time series of the form $y_t = \varepsilon_t + 0.4\varepsilon_{t-1} + 1.1\varepsilon_{t-2} - 2.5\varepsilon_{t-3}$




$\varepsilon_t \sim N(0,1)$, Figure 6-5 provides the scatterplot of the simulated data overtime.

Figure 6-6 provide; the ACF plot for the simulated data. Again, the ACF(O) equals 1, because any variable is perfectly correlated with itself.

At lags 1, 2, and 3, the value of the ACF is relatively large in absolute value compared to the subsequent terms. In an autoregressive model, the ACF slowly decays, but for an MA(3) model, the ACF somewhat abruptly cuts off after lag 3. in general, this pattern can be extended to any MA(q) model.




### 6.2.2 Arma And Arima Models :

In general, the data scientist does not have to choose between an AR(p) and an MA(q) model to describe a time series. In fact, it is often useful to combine these two representations into one model. The combination of these two models for a stationary time series results in an Autoregressive Moving Average model, ARMA(p,q), which is expressed as shown in Equation 8-15.

$$y_t = \delta + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q} \quad \text{(8-15)}$$

where $\delta$ is a constant for a nonzero-centered time series

$\phi_j$ is a constant for j = 1, 2, ..., p

$\phi_p \neq 0$

$\theta_k$ is a constant for k = 1, 2, ..., q

$\theta_q \neq 0$

$\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$ for all t

If $p$ = 0 and q≠0, then the ARMA(p,q) model is simply an AR(p) model. Similarly, if $p \neq 0$ and q=0, then the ARMA(p,q) model is an MA(q) model.

To apply an ARMA model properly, the time series must be a stationary one. If detrending using a linear or higher order regression model does not provide a stationary series, a second option is to compute the difference between successive y-values. This is known as *differencing*. In

other words, for the n values in a given time series compute the differences as shown in Equation 8-16

$$d_t = y_t - y_{t-1} \text{ for } t=2,3,\ldots..,n \quad (8.16)$$

Because the need to make a time series stationary is common, the differencing can be included (integrated) into the ARMA model definition by defining the *Autoregressive Integrated Moving Average* model, denoted ARIMA(p,d,q). The structure of the ARIMA model is identical to the expression in Equation 8-15, but the ARMA(p.q) model is applied to the time series, *yt*, after applying differencing d times.

### 6.2.3 Reasons to Choose and Cautions:

One advantage of ARIMA modeling is that the analysis can be based simply on historical time series data for the variable of interest. Various input variables need to be considered and evaluated for inclusion in the regression model for the outcome variable. Because ARIMA modeling, in general, ignores any additional input variables, the forecasting process is simplified.

The minimal data requirement also leads to a disadvantage of ARIMA modeling; the model does not provide an indication of what underlying variables affect the outcome. For example, if ARIMA modeling was used to forecast future retail sales, the fitted model would not provide an indication of what could be done to increase sales.

One caution in using time series analysis is the impact of severe shocks to the system. In the gas production example, shocks might include refinery fires, international incidents, or weather-related impacts such as hurricanes. Such events can lead to short-term drops in production, followed by persistently high increases in production to compensate for the lost production or to simply capitalize on any price increases.

Along similar lines of reasoning, time series analysis should only be used for short-term forecasts.

## 6.3 ADDITIONAL METHODS

Additional time series methods include the following:

- **Autoregressive Moving Average with Exogenous inputs (ARM AX)** is used to analyze a time series that is dependent on another time series. For example, retail demand for products can be modeled based on the previous demand combined with a weather-related time series such as temperature or rainfall.
- **Spectral analysis** is commonly used for signal processing and other engineering applications. Speech recognition software uses such techniques to separate the signal for the spoken words from the overall signal that may include some noise.

- **Generalized Autoregressive Conditionally Heteroscedastic (GARCH)** is a useful model for addressing time series with nonconstant variance or volatility. GARCH is used for modeling stock market activity and price fluctuations.
- **Kalman filtering** is useful for analyzing real-time inputs about a system that can exist in certain states. Typically, there is an underlying model of how the various components of the system interact and affect each other. A Kalman filter processes the various inputs, attempts to identify the errors in the input, and predicts the current state.
- **Multivariate time series analysis** examines multiple time series and their effect on each other. Vector ARIMA (VARIMA) extends ARIMA by considering a vector of several time series at a particular time, t. VARIMA can be used in marketing analyses that examine the time series related to a company's price and sales volume as well as related time series for the competitors

## 6.4 TEXT ANALYSIS

Text analysis, sometimes called text analytics, refers to the representation, processing, and modeling of textual data to derive useful insights. An important component of text analysis is text mining, the process of discovering relationships and interesting patterns in large text collections.

Text analysis suffers from the curse of high dimensionality. Text analysis often deals with textual data that is far more complex. A *corpus* (plural: corpora) is a large collection of texts used for various purposes in Natural Language Processing (N LP). Another major challenge with text analysis is that most of the time the text is not structured.

## 6.5 TEXT ANALYSIS STEPS

A text analysis problem usually consists of three important steps:

- Parsing
- Search and Retrieval
- Text Mining.

**Parsing** is the process that takes unstructured text and imposes a structure for further analysis. The unstructured text could be a plain text file, a weblog, an Extensible Markup Language (XML) file, a Hyper Text Markup Language (HTML) file, or a Word document. Parsing deconstructs the provided text and renders it in a more structured way for the subsequent steps.

**Search and retrieval** is the identification of the documents in a corpus that contain search items such as specific words, phrases, topics, or

entities like people or organizations. These search items are generally called key terms. Search and retrieval originated from the field of library science and is now used extensively by web search engines.

**Text mining** uses the terms and indexes produced by the prior two steps to discover meaningful insights pertaining to domains or problems of interest. With the proper representation of the text, many of the techniques such as clustering and classification, can be adapted to text mining. For example, the k-means can be modified to cluster text documents into groups, where each group represents a collection of documents with a similar topic. The distance of a document to a centroid represents how closely the document talks about that topic. Classification tasks such as sentiment analysis and spam filtering are prominent use cases for the naive Bayes. Text mining may utilize methods and techniques from various fields of study, such as statistical analysis, information retrieval, data mining, and natural language processing.

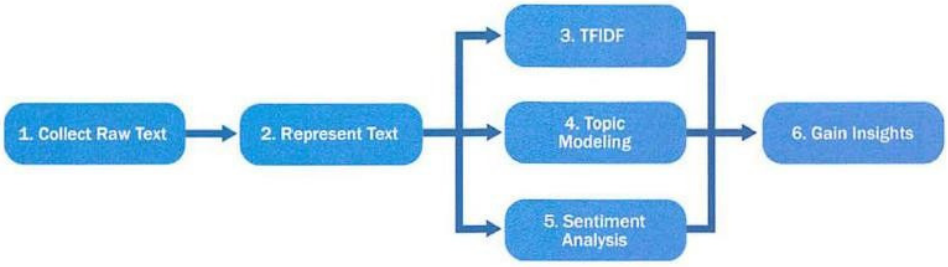
Note that, in reality, all three steps do not have to be present in a text analysis project. If the goal is to construct a corpus or provide a catalog service, for example, the focus would be the parsing task using one or more text preprocessing techniques, such as part-of-speech (POS) tagging, named entity recognition, lemmatization, or stemming. Furthermore, the three tasks do not have to be sequential. Sometimes their orders might even look like a tree.

## 6.6 A TEXT ANALYSIS EXAMPLE

Consider the fictitious company ACME, maker of two products: bPhone and bEbook. ACME is in strong competition with other companies that manufacture and sell similar products. To succeed, ACME needs to produce excellent phones and eBook readers and increase sales.
One of the ways the company does this is to monitor what is being said about ACME products in social media. In other words, what is the buzz on its products? ACME wants to search all that is said about ACME products in social media sites, such as Twitter and Facebook, and popular review sites, such as Amazon and Consumer Reports. It wants to answer questions such as these.

- Are people mentioning its products?
- What is being said? Are the products seen as good or bad? If people think an ACME product is bad, why? For example, are they complaining about the battery life of the bPhone, or the response time in their bEbook?

  ACME can monitor the social media buzz using a simple process based on the three steps of text analysis. This process is illustrated in Figure 9-1, and it includes following the modules.




1. **Collect raw text:** This corresponds to Phase 1 and Phase 2 of the Data Analytic Lifecycle. In this step, the Data Science team at ACME monitors websites for references to specific products. The websites may include social media and review sites. The team could interact with social network application programming interfaces (APIs) process data feeds, or scrape pages and use product names as keywords to get the raw data. Regular expressions are commonly used in this case to identify text that matches certain patterns. Additional filters can be applied to the raw data for a more focused study. For example, only retrieving the reviews originating in New York instead of the entire United States would allow ACME to conduct regional studies on its products. Generally, it is a good practice to apply filters during the data collection phase. They can reduce I/O workloads and minimize the storage requirements.

2. **Represent text:** Convert each review into a suitable document representation with proper indices, and build a corpus based on these indexed reviews. This step corresponds to Phases 2 and 3 of the Data Analytic Lifecycle.

3. **Comput:** the usefulness of each word in the reviews using methods such as TFIDF. This and the following two steps correspond to Phases 3 through 5 of the Data Analytic Lifecycle.

4. **Categorize documents by topics :** This can be achieved through topic models (such as latent Dirichlet allocation).

5. **Determine sentiments of the reviews**-. Identify whether the reviews are positive or negative. Many product review sites provide ratings of a product with each review. If such information is not available, techniques like sentiment analysis can be used on the textual data to infer the underlying sentiments.

6. **Review the results and gain greater insights**- This step corresponds to Phase 5 and 6 of the Data Analytic Lifecycle. Marketing gathers the results from the previous steps. Find out what exactly makes people love or hate a product. Use one or more visualization techniques to report the findings. Test the soundness of the conclusions and operationalize the findings if applicable.

## 6.7 COLLECTING RAW TEXT

In Data Analytic Lifecycle discovery is the first phase. In it, the Data Science team investigates the problem, understands the necessary data sources, and formulates initial hypotheses. Correspondingly, for text analysis, data must be collected before anything can happen, The Data Science team starts by actively monitoring various websites for user-generated contents. The user-generated contents being collected could be related articles from news portals and blogs, comments on ACME'S products from online shops or reviews sites, or social media posts that contain keywords ibPhone or **bEbook.** Regardless of where the data comes from, it's likely that the team would deal with semi-structured data such as HTML web pages, Really Simple Syndication (RSS) feeds, XML, or JavaScript Object Notation (JSON) files. Enough structure needs to be imposed to find the part of the raw text that the team really cares about. In the brand management example, ACME is interested in what the reviews say about **bPhone** or **bEbook** and when the reviews are posted. Therefore, the team will actively collect such information.

Many websites and services offer public APIs for third-party developers to access their data. For example, the Twitter API allows developers to choose from the Streaming API or the REST API to retrieve public Twitter posts that contain the keywords **bPhone** or **bEbook.** Developers can also read tweets in real time from a specific user or tweets posted near a specific venue. The fetched tweets are in the JSON format.
Many news portals and blogs provide data feeds that are in an open standard format, such as RSS or XML.

If the plan is to collect user comments on ACME'S products from online shops and review sites where APIs or data feeds are not provided, the team may have to write web scrapers to parse web pages and automatically extract the interesting data from those HTML files. A web scraper is a software program (bot) that systematically browses the World Wide Web, downloads web pages, extracts useful information, and stores it somewhere for further study.

The team can then construct the web scraper based on the identified patterns. The scraper can use the curl tool to fetch HTML source code given specific URLs, use XPath and regular expressions to select and extract the data that match the patterns, and write them into a data store.

Regular expressions can find words and strings that match particular patterns in the text effectively and efficiently. The general idea is that once text from the fields of interest is obtained, regular expressions can help identify if the text is really interesting for the project. In this case, do those fields mention bPhone, bEbook, or ACME? When matching the text, regular expressions can also take into account capitalizations, common misspellings, common abbreviations, and special formats for e-mail addresses, dates, and telephone numbers.

## 6.8 REPRESENTING TEXT

In this data representation step, raw text is first transformed with text normalization techniques such as tokenization and case folding. Then it is represented in a more structured way for analysis.

*Tokenization* is the task of separating words from the body of text. Raw text is converted into collections of tokens after the tokenization, where each token is generally a word.

A common approach is tokenizing on spaces. For example, witthe tweet shown previously:

I once had a gf back in the day. Then the bPhone came out lol tokenization based on spaces would output a list of tokens.

(I, once, had, a, gf, back, in, the, day., Then, the, bPhone, came, out, lol) Another way is to tokenize the text based on punctuation marks and spaces. In this case, the previous tweet would become:

{I, once, had, a, gf, back, in, the, day, ., Then, the, bPhone, came, out, lol} However, tokenizing based on punctuation marks might not be well suited to certain scenarios. For example, if the text contains contractions such as ***we*** 'll, tokenizing based on punctuation will split them into separated words ***we*** and ll.

Tokenization is a much more difficult task than one may expect. For example, should words like ***state-of - the - art,*** Wi -Fi,and ***San Francisco*** be considered one token or more?

Another text normalization technique is called *case folding,* which reduces all letters to lowercase (or the opposite if applicable). For the previous tweet, after case folding the text would become this:

i once had a gf back in the day. then the bphone came out lol
One needs to be cautious applying case folding to tasks such as information extraction, sentiment analysis, and machine translation. For example, when ***General Motors*** becomes ***general*** and ***motors,*** the downstream analysis may very likely consider them as separated words rather than the name of a company.

After normalizing the text by tokenization and case folding, it needs to be represented in a more structured way. A simple yet widely used approach to represent text is called *bag-of-words.* Given a document, bag-of-words represents the document as a set of terms, ignoring information such as order, context, inferences, and discourse.

Bag-of-words takes quite a naive approach, as order plays an important role in the semantics of text. With bag- of-words, many texts

with different meanings are combined into one form. For example, the texts "a dog bites a man" and "a man bites a dog" have very different meanings, but they would share the same representation with bag-of-words.

Besides extracting the terms, their morphological *features* may need to be included. The morphological features specify additional information about the terms, which may include root words, affixes, part-of-speech tags, named entities, or intonation (variations of spoken pitch). The features from this step contribute to the downstream analysis in classification or sentiment analysis.

Sometimes creating features is a text analysis task all to itself. One such example is *topic modeling*. Topic modeling provides a way to quickly analyze large volumes of raw text and identify the latent topics. Topic modeling may not require the documents to be labeled or annotated. It can discover topics directly from an analysis of the raw text.

It is important not only to create a representation of a document but also to create a representation of a corpus. Most corpora come with metadata, such as the size of the corpus and the domains from which the text is extracted. Some corpora (such as the Brown Corpus) include the information content of every word appearing in the text. *Information content* (IC) is a metric to denote the importance of a term in a corpus.

## 6.9 TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY (TFIDF)

TFIDF is widely used in information retrieval and text analysis. Instead of using a traditional corpus as a knowledge base, TFIDF directly works on top of the fetched documents and treats these documents as the "corpus." TFIDF is robust and efficient on dynamic content, because document changes require only the update of frequency counts.

Given a term *t* and a document d = (t1,t2,t3…tn) containing *n* terms, the simplest form of term frequency of r in dean be defined as the number of times f appears in *d,* as shown in Equation 9-1.

$$TF_1(t,d)=\sum_{i=1}^{n} f(t,t_i) \qquad t_i \in d; |d| = n$$

where

$$f(t,t')=\begin{cases} 1, & \text{if } t=t' \\ 0, & \text{otherwise} \end{cases} \tag{9-1}$$

$$TF_2(t,d)=\log[TF_1(t,d)+1] \tag{9-2}$$

Similarly, the logarithm can be applied to word frequencies whose distribution also contains a long tail, as shown in Equation 9-2.

Because longer documents contain more terms, they tend to have higher term frequency values, They also tend to contain more distinct terms. These factors can conspire to raise the term frequency values of longer documents and lead to undesirable bias favoring longer documents. To address this problem, the term frequency can be normalized. For example, the term frequency of term t in document *d* can be normalized based on the number of terms in das shown in Equation 9-3.

$$TF_3(t,d) = \frac{TF_1(t,d)}{n} \qquad |d| = n \qquad (9\text{-}3)$$

Indeed, that is the intention of them verted *document frequency* (IDF). The IDF inversely corresponds to the *document frequency* (DF), which is defined to be the number of documents in the corpus that contain a term. Let a corpus *D* contain *N* documents. The document frequency of a term t in corpus *D = [dyd2,...d*N3 is defined as shown in Equation 9-4.

The precise base of the logarithm is not material to the ranking of a term. Mathematically, the base constitutes a constant multiplicative factor towards the overall result.

The TFIDF (or TF-IDF) is a measure that considers both the prevalence of a term within a document (TF) and the scarcity of the term over the entire corpus (IDF). The TFIDF of a term *t* in a document dis defined as the term frequency of *t* in *d* multiplying the document frequency of *t* in the corpus as shown in Equation 9-7:

$$TFIDF(t,d) = TF(t,d) \times IDF(t) \qquad (9\text{-}7)$$

TFIDF is efficient in that the calculations are simple and straightforward, and it does not require knowledge of the underlying meanings of the text. But this approach also reveals little of the inter-document or intra-document statistical structure.

## 6.10 CATEGORIZING DOCUMENTS BY TOPICS

A topic consists of a cluster of words that frequently occur together and share the same theme. Document grouping can be achieved with clustering methods such as k-means clustering or classification methods such as support vector machines, or naive Bayes, However, a more feasible and prevalent approach is to use topic modeling. Topic modeling provides tools to automatically organize, search, understand, and summarize from vast amounts of information. Topic models are statistical models that examine words from a set of documents, determine the themes over the text, and discover how the themes are associated or change over time. The process of topic modeling can be simplified to the following.

1. Uncover the hidden topical patterns within a corpus.

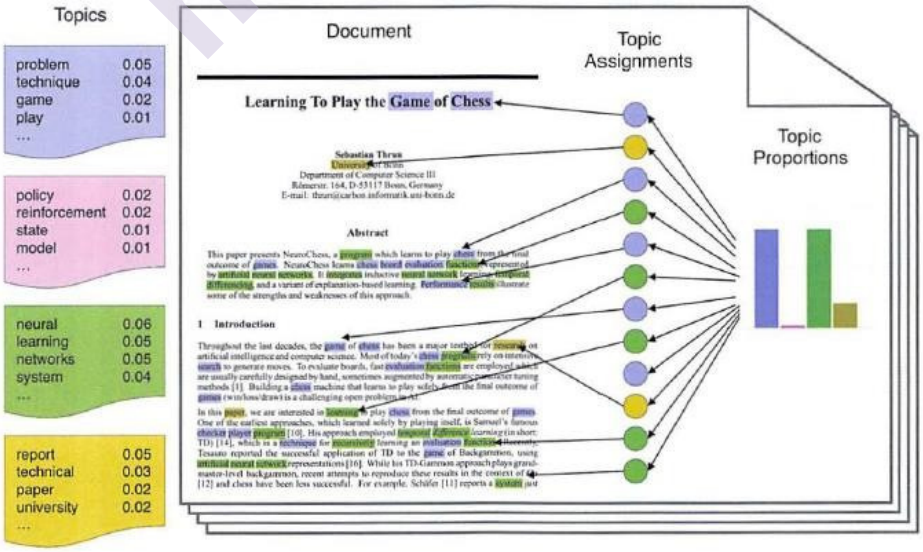
2. Annotate documents according to these topics.
3. Use annotations to organize, search, and summarize texts.

A **topic** is formally defined as a distribution over a fixed vocabulary of words, Different topics would have different distributions over the same vocabulary. A topic can be viewed as a cluster of words with related meanings, and each word has a corresponding weight inside this topic.

The simplest topic model is latent Dirichlet allocation (LDA), a generative probabilistic model of a corpus proposed by David M. Blei and two other researchers. In generative probabilistic modeling, data is treated as the result of a generative process that includes hidden variables. LDA assumes that there is a fixed vocabulary of words, and the number of the latent topics is predefined and remains constant. LDA assumes that each latent topic follows a Dirichlet distribution over the vocabulary, and each document is represented as a random mixture of latent topics.

Figure 6-4 illustrates the intuitions behind LDA. The left side of the figure shows four topics built from a corpus, where each topic contains a list of the most important words from the vocabulary. The four example topics are related to problem, policy, neural, and report. For each document, a distribution over the topics is chosen, as shown in the histogram on the right. Next, a topic assignment is picked for each word in the document, and the word from the corresponding topic (colored discs) is chosen. In reality, only the documents (as shown in the middle of the figure) are available. The goal of LDA is to infer the underlying topics, topic proportions, and topic assignments for every document.

Many programming tools provide software packages that can perform LDA over datasets. R comes with an *Ida* package that has built-in functions and sample datasets.



## 6.11 DETERMINING SENTIMENTS

Sentiment analysis refers to a group of tasks that use statistics and natural language processing to mine opinions to identify and extract subjective information from texts.

Intuitively, to conduct sentiment analysis, one can manually construct lists of words with positive sentiments (such as brilliant, awesome, and spectacular) and negative sentiments (such as awful, stupid, and hideous). Related work has pointed out that such an approach can be expected to achieve accuracy around 60%, and it is likely to be outperformed by examination of corpus statistics.

Classification methods such as naive Bayes, maximum entropy (MaxEnt), and support vector machines (SVM) are often used to extract corpus statistics for sentiment analysis. Related research has found out that these classifiers can score around 80% accuracy on sentiment analysis over unstructured data. One or more of such classifiers can be applied to unstructured data, such as movie reviews or even tweets.

Depending on the classifier, the data may need to be split into training and testing sets. One way for splitting data is to produce a training set much bigger than the testing set. For example, an 80/20 split would produce 80% of the data as the training set and 20% as the testing set.

Next, one or more classifiers are trained over the training set to learn the characteristics or patterns residing in the data. The sentiment tags in the testing data are hidden away from the classifiers. After the training, classifiers are tested over the testing set to infer the sentiment tags. Finally, the result is compared against the original sentiment tags to evaluate the overall performance of the classifier.

Classifiers determine sentiments solely based on the datasets on which they are trained. The domain of the datasets and the characteristics of the features determine what the knowledge classifiers can learn. For example, lightweight is a positive feature for reviews on laptops but not necessarily for reviews on wheelbarrows or textbooks. In addition, the training and the testing sets should share similar traits for classifiers to perform well. For example, classifiers trained on movie reviews generally should not be tested on tweets or blog comments.

Note that an absolute sentiment level is not necessarily very informative. Instead, a baseline should be established and then compared against the latest observed values. For example, a ratio of 40% positive tweets on a topic versus 60% negative might not be considered a sign that a product is unsuccessful if other similar successful products have a similar ratio based on the psychology of when people tweet.

## SUMMARY

- Time series analysis attempts to model the underlying structure of observations taken over time.
- The goals of time series analysis:Identify and model the structure of the time series and forecast future values in the time series
- The trend refers to the long-term movement in a time series. It indicates whether the observation values are increasing or decreasing over time.
- The seasonality component describes the fixed, periodic fluctuation in the observations over time.
- A cyclic component also refers to a periodic fluctuation, but one that is not as fixed as in the case of a seasonality component.
- Although noise is certainly part of this random component, there is often some underlying structure to this random component that needs to be modeled to forecast future values of a given time series.
- The plot of autocorrelation function (ACF) provides insight into the covariance of the variables in the time series and its underlying structure.
- Text analysis is text mining, the process of discovering relationships and interesting patterns in large text collections.
- A corpus (plural: corpora) is a large collection of texts used for various purposes in Natural Language Processing (NLP).
- Parsing is the process that takes unstructured text and imposes a structure for further analysis.
- Search and retrieval is the identification of the documents in a corpus that contain search items such as specific words, phrases, topics, or entities like people or organizations.
- Text mining uses the terms and indexes produced by the prior two steps to discover meaningful insights pertaining to domains or problems of interest.

## UNIT END QUESTIONS

1. Write a note on Box-Jenkins Methodology
2. Explain the ARIMA Model technique
3. Explain the steps involved in Text analysis with example.
4. What is tokenization? Explain how it is used in text analysis.
5. Describe the Term Frequency-Inverse Document Frequency method.

6. Explain the steps involved in categorizing documents by topic.
7. Write a note on determining sentiments of documents using text analysis.
8. Write a short note on decision tree.
9. How to predict whether customers will buy a product or not? Explain with respect to decision tree.
10. Explain a probabilistic classification method based on Naive Bayes' theorem.
11. John flies frequently and likes to upgrade his seat to first class. He has determined that if he checks in for his flight at least two hours early, the probability that he will get an upgrade is 0.75; otherwise, the probability that he will get an upgrade is 0.35. With his busy schedule, he checks in at least two hours before his flight only 40% of the time. Suppose John did not receive an upgrade on his most recent attempt. What is the probability that he did not arrive two hours early? Find it with respect to on Bayes' theorem.
12. Describe additional classification methods other than decision tree and Bayes' theorem.
13. How to model a structure of observations taken over time? Explain with respect to Time series analysis. Also explain any two of its applications.
14. What are the components of time series? Explain each of them. Also write the main steps of Box-Jenkins methodology for time series analysis.
15. Explain Autoregressive Integrated Moving Average Model in detail.
16. Explain additional time series methods other than Box-Jenkins methodology and Autoregressive Integrated Moving Average Model.
17. What are major challenges with text analysis? Explain with examples.
18. What are various text analysis steps? Explain in detail.
19. Describe ACME's Text Analysis Process.
20. What is the use of Regular Expressions? Explain any five regular expressions with its description and example.
21. How to normalize the text using tokenization and case folding? Explain in detail. Also explain about Bag-of-words approach.
22. How to retrieve information and applying text analysis? Explain with respect to Term Frequency.
23. What is the critical problem in using Term frequency? How can it be fixed?
24. How to categorize documents by topics? Explain in detail.

25. What is sentiment analysis? How it can be carried out? Explain it in detail.

## REFERENCES


- Data Analytics with Hadoop -An Introduction for Data Scientists by Benjamin Bengfort and Jenny Kim
- The Data Revolution: Big Data, Open Data, Data Infrastructures, And Their Consequences By Rob Kitchin
- Big Data: Principles and Best Practices of Scalable Real-Time Data Systems By Nathan Marz And James Warren
- Big Data Demystified By David Stephenson




*****

# UNIT IV

# 7

# DATA PRODUCT & BIG DATA OPERATING SYSTEM

**Unit Structure**



## 7.0 OBJECTIVES

The objectives of this chapter are to:

- Introduce the concept of a data product and evolve its definition as we go ahead.
- Introduce Hadoop as a means to solve the problem of processing data at a scale
- Define & understand the difference in the Data Science Pipeline as well as the Big Data Pipeline
- Understand the Architecture of the Big Data Operating System: Hadoop
- Understand the Hadoop file system and way to perform Distributed Computation
- Understand what is Map Reduce

## 7.1 INTRODUCTION

- This chapter explains the concept data products, its necessity and explains how Hadoop can be used for implementing it and for data science.

- The requirements for distributed storage and computation has resulted in the creation of Hadoop and its evolution into an operating system for big data.

## 7.2 INTRODUCTION TO DATA PRODUCT

- The volume of data being generated every second is very overwhelming.
- Data is increasingly changing how we work, play, and entertain ourselves, and technology has come to describe every facet of life, from the food we prepare to our online social interactions.
- Yet we expect highly personalized and finely tuned products and services well suited to our behavior and nature which has resulted in creating an opportunity in the market with a new technology—the data product.
- Data products are created using processing chains in data science, thoughtful application of complex algorithms possibly predictive or inferential being applied to a specific dataset.
- **Defining a Data Product**
  - Traditionally a data product is any application combining data with algorithms.
  - Writing a software is not just combining data with algorithms, speaking of data product, it is the combination of data and statistical algorithms useful for generating inferences or predictions. Ex. Facebook's "People You May Know"
  - But this definition limits data products to single software instances (ex, any web application),
  - A data product is not just a name for a app driven by data but also an data application which uses the data to acquire its value and in the process creates additional data as output. it's a data product, not just an application with data.
  - A data product is a cost-effective engine that extracts value from data while also generating more data.
- Data products have been described as systems that learn from data and can self-adapt in a number of ways.
- The Nest Thermostat, for example, is a data system that derives its value from sensor data, schedules heating and cooling, and collects and validates new sensor observations.
- Data products are economic engines that self-adapt and uses the data to acquire its value and in the process creates additional data while it makes inferences or predictions upon new data by influencing human behavior with this very data.

- Data products are no longer programs that run on the web interface, they are becoming an important part of every domain of activity in the current modern world.

## 7.3 USING HADOOP TO BUILD DATA PRODUCTS AT SCALE

- In this era of data product the job of the data scientist is to build it.

  The experimental methodology is this typical analytical workflow as pointed by data scientists in creating a data product is:

  Ingestion → Wrangling → Modeling → Reporting & Visualization.

- The data science pipeline and is human designed and augmented by the use of languages like R and Python
- When we create a data product it allows data to become big in size, fast in execution, and enables larger variety of data for computation which in turn help to derive insights and does not involve human interaction.
- **Using Large Datasets as an Advantage**
  - Humans have extraordinary vision for large-scale patterns, such as woods and clearings visible through the foliage.
  - Statistical methodologies allow us to deal with both noisy and meaningful data by defining them with aggregations and indices or inferentially by performing the analysis directly.
  - As our ability to gather data has increased, so has the requirement for more generalization.
  - Smart grids, quantified selves, mobile technology, sensors, and wired homes all require personalised statistical inference.
  - Scale is measured by the number of facets that must be explored in addition to the amount of data—a forest view for individual trees.
  - Hadoop is distinct due to the economics of data processing as well as the fact that it is a platform.
  - Hadoop's release was interesting in that it came at a time when the world needed a solution for large-scale data analytics.

- **Creating Data Products using Hadoop**
- Hadoop has been developed by tech giants such as Google, Facebook, and Yahoo to deal with big data challenges
- Data issues are no longer limited to tech behemoths; they also impact commercial and public organisations of all sizes, from large companies to startups, federal agencies to cities, and perhaps even individuals.
- Computing services are also becoming more available and affordable.

- Data scientists can get on-demand, instant access to clusters of large sizes by using different cloud computing platforms like Google Compute Engine or Amazon EC2 at a fraction of the cost of traditional data centres and with no requirement of doing data center management..
- Big data computing is being made democratic and more open to everyone by Hadoop
- Data analytics at large scale have historically been available only to social networks such as Facebook and Twitter, but now they are also available to individual brands or artists.
- Connected homes and mobile devices, as well as other personal sensors, are producing vast quantities of personal data, raising questions about privacy, among other items.
- In 2015, British researchers founded the Hub of All Things (HAT). It is a customised data collection that tackles the problem of data ownership and offers a solution for personal data aggregation.
- New data problems are emerging, and a data product is needed to address these questions.
- Applications like ShotSpotter & Location and HAT offer an application interface and decision-making tools to help people derive value from data and create new data.
- Conventional software development workflows are insufficient for working with large datasets, but Big Data workflows and Hadoop have allowed and personalised these applications.

### 7.4.1 The Data Science Pipeline:

Characteristics of data science pipeline are as follows:

- Human - driven, is concerned with the development of practical data visualisations.
- Having a workflow with the aim to produce results that enable humans to make decisions.

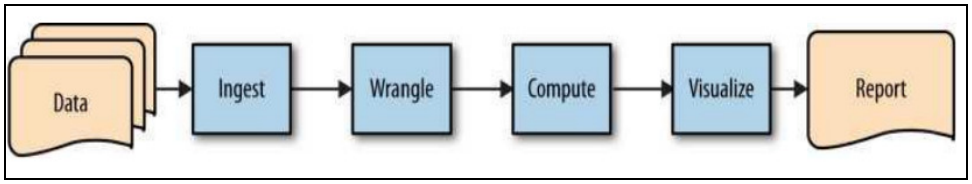



**Fig 7.1: The Data Science Pipeline**

(Ref - Chapter 1, Fig 7.1 - Data Analytics with Hadoop - An Introduction for Data Scientists)

- An analyst takes in large volume of data performs some operations on it to convert it into a normal form so that can we can perform different calculations to finally present the results in a visual manner.

- With the overwhelming growth rate in the volume and velocity at which many businesses are now generating data, this human-powered model is not scalable.

### 7.4.2 The Big Data Pipeline:

- We implement a machine learning feedback loop into the data science pipeline to create a framework that enables the development of scalable, automated data analysis and insight generation solutions.
- The new framework is the big data pipeline which is
  - Not human-driven,
  - Is an iterative model
  - has four primary phases
  - ensures scalability and automation

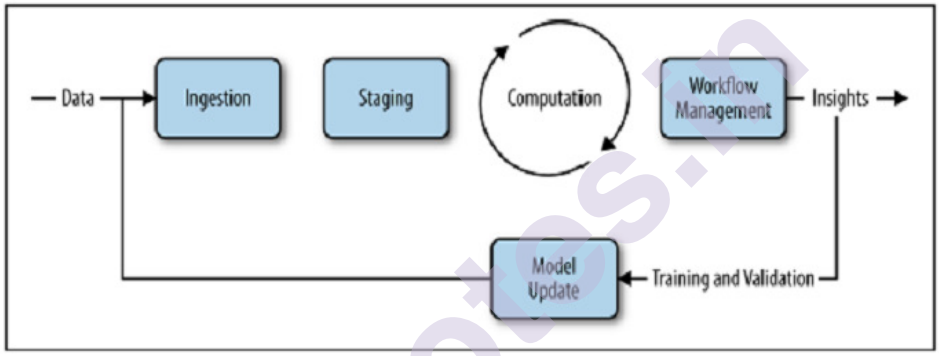



**Fig 7.2 The Big Data Pipeline**

(Ref - Chapter 1, Fig 1.1 - Data Analytics with Hadoop - An Introduction for Data Scientists)

- The 4 stages of the Big Data Pipeline are:
  - staging,
  - ingestion,
  - computation,
  - workflow management
- In its most basic form, this model, like the data science pipeline, takes raw data and transforms it into insights.
- This stage generates a reusable data product as the output by transforming the ingestion, staging, and computation phases into an automated workflow.
- A feedback system is often needed during the workflow management stage, that gives the output of one job can be automatically fed in as the data input for the next, allowing for self-adaptation.
- The **ingestion phase** involves both the model's initialization and the model's device interaction with users.

- Users may define data source locations or annotate data during the initialization process
- While interacting, users will receive the predictions given by the model and in turn give important feedback to strengthen the model
- The **staging step** requires executing transformations on data to make it usable and storeable, allowing it to be processed.
- The tasks of staging include data normalization, standardization & data management.
- The **computation phase** takes maximum time while executing the key responsibilities of extracting insights from data, conducting aggregations or reports, and developing machine learning models for recommendations, regressions, clustering, or classification.
- The **workflow management phase** involves tasks such as abstraction, orchestration, and automation, which enables to operationalize the performance of workflow steps. The final output is supposed to be an program that is automated that can be run as desired.

### 7.4.3 The Hadoop Ecosystem:

- Hadoop Platform has specifically transformed into an ecosystem consisting of a variety different tools that operationalize the different parts of the Big Data pipeline.
- Kafka and Sqoop, for example, are designed for data extraction and ingestion, enabling relational databases to be imported into Hadoop or distributed message queues for processing on-demand.
- Data warehouses in Hadoop, such as Hive and HBase, allow for large-scale data management.
- Hapdoop makes use of libraries like Spark's GraphX and MLlib, as well as Mahout, which provide analytical packages for large-scale computation and validation.

## 7.5 HADOOP : THE BIG DATA OPERATING SYSTEM

- Hadoop systems ensure that the criteria for a distributed Big Data Operating System are met, as well as that Hadoop is a data management system that works as expected while processing analytical data.
- Hadoop has mainly been used to store and compute massive, heterogeneous datasets stored in data lakes rather than warehouses, as well as for rapid data processing and prototyping.
- Basic knowledge of distributed computing and storage is needed to fully understand the working of Hadoop and how to build data processing algorithms and workflows.

- Hadoop distributes the computational processing of a large dataset to several machines that each run on their own chunk of data in parallel to perform computation at scale.
- The following conditions must be fulfilled by a distributed system:
  - **Fault tolerance** - A system part failure does not result in the whole system failing. The system should be able to degrade into a less productive state in a graceful manner. The failed system part should be able to rejoin the system if it recovers.
  - **Recoverability** - No data should be lost when a malfunction occurs no matter how big or small.
  - **Scalability -** As the load increases (data & computation), the output decreases, not fails; increasing resources should result in a proportional increase in power.
  - **Continuity -** The failure of one job or task should not affect the final result.
- Hadoop tackles the above specifications using a variety of abstract principles such as:
  - **Clusters** - working out how to manage data storage and distributed computing in a cluster.
  - **Data distribution -** As data is applied to the cluster and stored on several nodes, it is distributed instantly. To reduce network traffic, each node processes locally stored data
  - **Data Storage -** Data is held in typically 128 MB fixed-size blocks, and copies of each block are made several times for achieving redundancy and data protection.
  - **Jobs -** In Hadoop, a job is any computation performed; jobs may be divided into several tasks, with each node performing the work on a single block of data.
  - **Programming Language -** Jobs written in high level allow us to ignore low level details, allowing developers to concentrate their attention only on data and computation.
  - **Fault tolerance -** When task replication is used, jobs are fault tolerant, ensuring that the final computation is not incorrect or incomplete if a single node or task fails.
  - **Communication -** The amount of communication occurring between nodes should be kept at minimum and should be done in a transparent manner by the system. To avoid inter-process dependencies leading to deadlock situation every task should be executed independently and nodes should not communicate during processing to ensure it.
  - **Work Allocation -** Master programmes divide work among worker nodes so that they can all run in parallel on their own slice of the larger dataset.

## 7.6 HADOOP ARCHITECTURE

- Hadoop Architecture consists of two primary components:

  1. **HDFS (or DFS) is the Hadoop Distributed File System**

     It implements the fundamentals of distributed storage and is in charge of handling data around the cluster's discs.

  2. **YARN (Yet Another Resource Negotiator)**

  - It implements computation in a distributed environment.
  - YARN acts as a cluster resource manager, allocating computing resources to applications that require distributed computing.
  - Hadoop Architecture in Figure below:

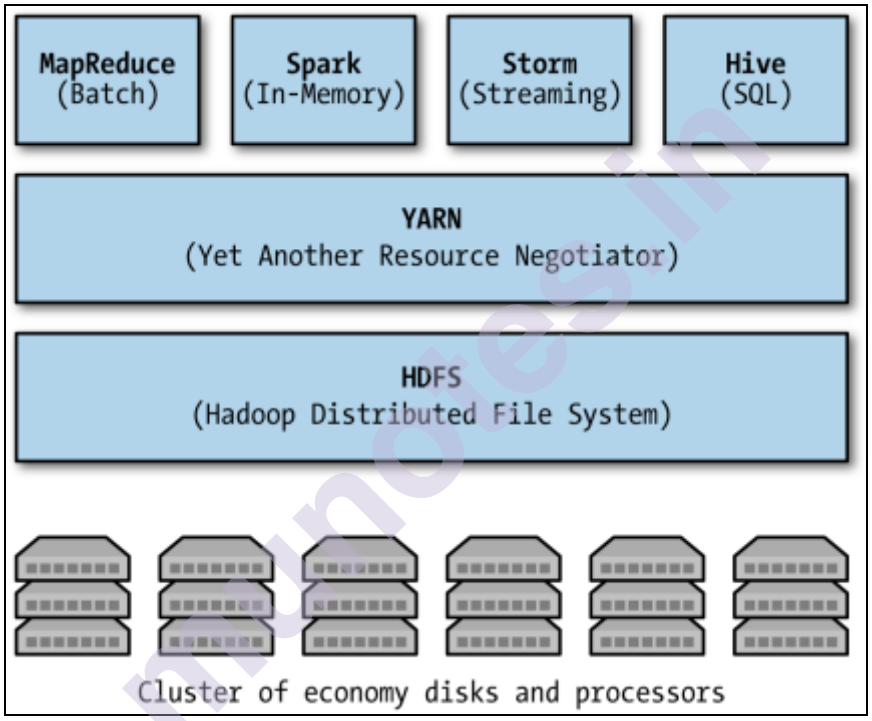



**Fig 7.3 : Hadoop Architecture**

(Ref - Chapter 2, Fig 7.3 - Data Analytics with Hadoop - An Introduction for Data Scientists)

- Together, HDFS and YARN together form a platform. This can be used for creating big data applications as it provides an operating system for big data. The two collaborate to reduce network traffic in the cluster, mainly by guaranteeing that data is kept local to the necessary computation. Both data and tasks are duplicated to ensure error tolerance, recoverability, and accuracy. To provide scalability and low-level clustering programming information, the cluster is managed centrally.

### 7.6.1 Hadoop Cluster:

- Hadoop is not a piece of hardware but is a cluster of machines that function together in a synchronised manner.
- Hadoop is the software that runs on a cluster—it include the DFS - HDFS, and the cluster resource manager - YARN, which run in background on a group of machines are collectively as six different types of background services.
- A cluster is a collection of machines running HDFS and YARN, with nodes representing individual machines. Several daemon (background) processes implement YARN and HDFS in the background and do not require user input
- A cluster may have a single node or thousands, but they all scale horizontally, meaning that the cluster's capacity and performance increase linearly as more nodes are added.
- Hadoop processes are background services that execute throughout the time on a cluster node. It accepts input and output across the network. Each of Hadoop processes has its own allocation of system resources and is managed independently.
- Nodes are two types, each of which is differentiated by the process or processes that it executes:
- **Master nodes**
    - These nodes provide Hadoop workers with organising resources and are usually the cluster's entry points.
    - Communication would fall apart without masters, and would not be possible to have distributed storage tasks or computations.
- **Worker nodes**
    - These are the majority of the cluster's machines.
    - Services of Worker nodes include accepting requests from master nodes, such as storing or retrieving data or running a specific programme.
    - In a distributed computation, the analysis is parallelized across worker nodes.
- Both HDFS and YARN have multiple master services. These services are responsible for coordinating worker services which run on each worker node.

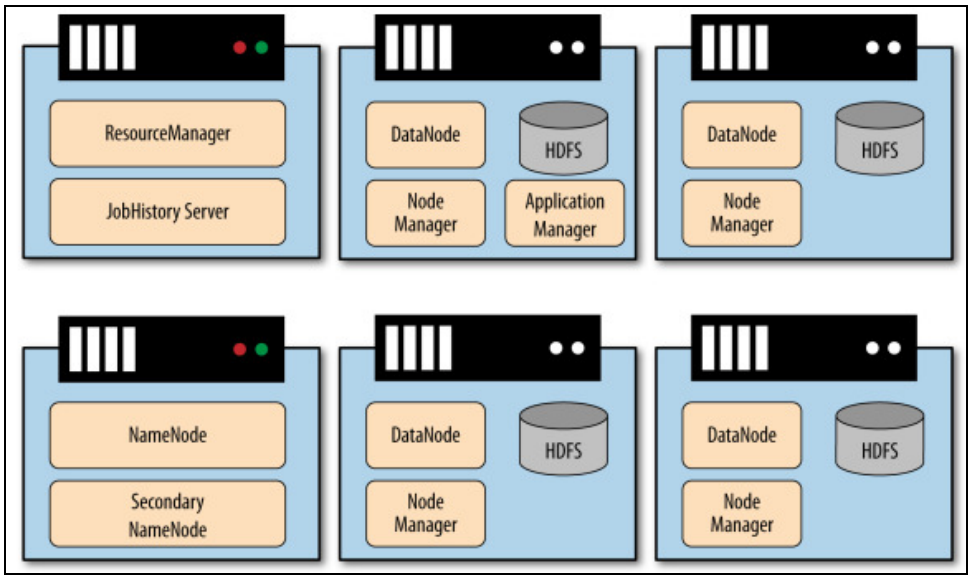



**Fig 7.4: A cluster in Hadoop containing two master & four workers nodes together implementing the six primary Hadoop services**

(Ref - Chapter 2, Fig 2.2 - Data Analytics with Hadoop - An Introduction for Data Scientists)

HDFS has the following master, worker services:

- **NameNode (Master service)**
  - Keeps the file system's directory tree, file metadata, and the locations of all files in the cluster.
  - Clients who want to use HDFS must first request information from the NameNode in order to find the required storage nodes.
- **Secondary NameNode (Master service)**
  - On behalf of the NameNode, conducts housekeeping and checkpointing.
  - It is not a backup NameNode, despite its name.
- **DataNode (Worker service)**
  - Stores and manages HDFS blocks on the local disk.
  - Reports health and status of individual data stores back to the NameNode.
- When a client application requests data from HDFS, it must first make a request to the NameNode for the data to be located on disc.
- Instead of storing data or transferring data from DataNode to client, the NameNode simply functions as a traffic cop, guiding clients to the necessary DataNodes.

- Following are the master and worker services provided by YARN:
- **ResourceManager (Master service)**
  - Controls job scheduling on the cluster by allocating and monitoring available cluster resources, such as physical assets like memory and processor cores, to applications.
- **ApplicationMaster (Master service)**
  - The ResourceManager schedules the execution of a particular program on the cluster, and this portion coordinates its execution.
- **NodeManager (Worker service)**
  - On a each individual node, it runs and manages processing activities, as well as reporting on their health and status..
- Similar to how HDFS works, clients that wish to execute a job must first request resources from the ResourceManager, which assigns an application-specific ApplicationMaster for the duration of the job. The ApplicationMaster is responsible for tracking the execution of the job, while the ResourceManager is responsible for tracking the status of the nodes, and each individual NodeManager creates containers and executes tasks within them.
- Pseudo-distributed mode is a single node cluster. All Hadoop daemons are run on a single machine as if it were a cluster, but network traffic is routed via the local loopback network interface. The advantages of a distributed architecture aren't realised in this mode, but it's a great way to build without having to worry about handling multiple machines.

### 7.6.2 Hadoop Distributed File System (HDFS):

- HDFS doubles the amount of storage space available from a single computer by storing it via a cluster of low-cost, unreliable devices.
- HDFS is a layer of software that sits on top of a native file system, allowing it to communicate with local file systems and generalising the storage layer.
- HDFS was built with the aim of storing large files while still allowing for real-time access to data.
- For storing raw input data for computation, intermediate results between computational phases, and overall job results, HDFS is the best choice.
- HDFS is not good as a data backend for applications requiring real-time updates, interactive analysis and record based transactional support.
- Following are few characteristics of HDFS:
  - HDFS is best suited to a small number of very large files—for example, millions of large files (greater than 100 MB in size)

rather than billions of smaller files that would otherwise occupy the same amount of space.

- HDFS follows the WORM (write once, read many) pattern and does not permit random file appends or writes.
- HDFS is designed for large-scale, continuous file reading rather than random reading or collection.

- **HDFS Blocks**
  - HDFS files are divided into blocks, which are usually 64 MB or 128 MB in size, but this is configurable at runtime, and high-performance systems typically use 256 MB block sizes.
  - Equivalent to the block size on a single disc file system, the block size in HDFS is the smallest amount of data that can be read or written to. Files that are smaller than the block size, unlike blocks on a single disc, do not fill the entire block.
  - Blocks allow very large files to be split across multiple machines and distributed at runtime. To allow for more efficient distributed processing, separate blocks from the same file will be stored on different machines.
  - The DataNodes will duplicate the blocks. The replication is three-fold by design, but this can be modified at runtime. As a result, each block of data resides on three different computers and three different discs, and the data will not be lost even though two nodes fail.
  - The cluster's potential data storage capacity is just a third of the available disc space due to replication.
- **HDFS Data Management**
  - The master NameNode keeps track of the file's blocks and their locations.
  - The NameNode communicates with the DataNodes, which are processes that house the blocks in the cluster.
  - Each file's metadata is stored in the NameNode master's memory for fast lookups, and if the NameNode stops or fails, the entire cluster becomes unavailable.
  - The Secondary NameNode is not a substitute for the NameNode; rather, it handles the NameNode's housekeeping such as periodically combining a snapshot of the current data space with the edit log to prevent the edit log from becoming too large.
  - The function of the edit log is used to maintain data integrity and avoid data loss; in case the NameNode fails, this combined record can be used to restore the state of the DataNodes.

### 7.6.3 Workload & Resource Manager (YARN):

- The original version of Hadoop offered MapReduce on HDFS where the MapReduce job/workload management functions were highly coupled to the cluster/resource management functions. As a result, other computing models or applications were unable to use the cluster infrastructure for execution of distributed workloads.
- YARN separates workload and resource management so that many applications can share a single, unified resource management service. Hadoop is no longer a uniquely oriented MapReduce platform, but a full-fledged multi-application, big data operating system, thanks to YARN's generalised job and resource management capabilities.
- The basic concept behind YARN is to separate the resource management and workload management roles into separate daemons.

## 7.7 WORKING WITH DISTRIBUTED COMPUTATION – MAPREDUCE

- Although YARN has allowed Hadoop to become a general-purpose distributed computing platform, MapReduce (also known as MR) was the first Hadoop computational system.
- MapReduce is a straightforward but effective computational system for fault-tolerant distributed computing across a cluster of centrally controlled machines. It accomplishes this by using a "functional" programming style that is essentially parallelizable, allowing several independent tasks to perform a function on local groups of data and then combining the results.
- Functional programming is a programming methodology that guarantees stateless evaluation of unit computations. This implies that functions are closed, in the sense that they do not exchange state and depend solely on their inputs. Data is transferred between functions by using the output of one function as the input of a completely different function.
- MapReduce provides the two functions that distribute work and aggregate results called map and reduce
- **Map Function**
    - MapReduce offers the map and reduce functions, which distribute work and aggregate results.
    - A map function takes a list of key/value pairs as input and works on each pair separately.
    - The map operation is where the core analysis or processing takes place, as this is the function that sees each individual element in the dataset

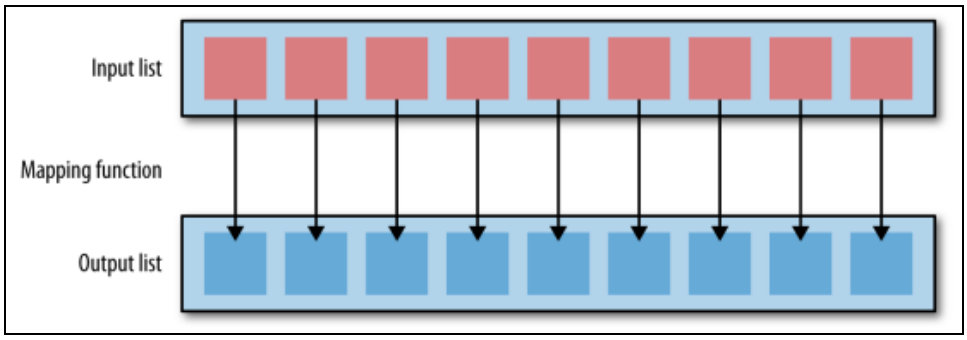



**Fig 7.5 - A map function**

(Ref - Chapter 2, Fig 2.3 - Data Analytics with Hadoop - An Introduction for Data Scientists)

- **Reduce Function**
  - Any emitted key/value pairs will be grouped by key after the map phase, and those key/value groups will be used as input for per-key minimization functions.
  - When a reduce function is applied to an input set, the output is a single, aggregated value.

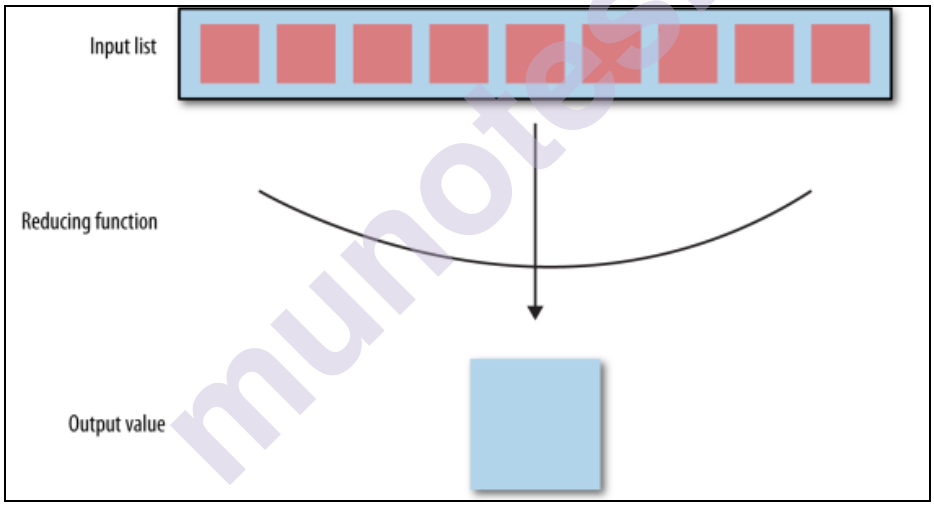





**Fig 7.6 - A reduce function**

(Ref - Chapter 2, Fig 2.4 - Data Analytics with Hadoop - An Introduction for Data Scientists)

- **MapReduce Framework**
  - Hadoop MapReduce is a software framework for composing jobs that run in parallel on a cluster and process large quantities of data, and is the native distributed processing framework that ships with Hadoop.

- As a job configuration, the system exposes a Java API that allows developers to define HDFS input and output positions, map and reduce functions, and other job parameters.
- Jobs are compiled and packaged into a JAR, which is submitted to the Resource Manager by the job client—usually via the command line. The Resource Manager will then schedule the tasks, monitor them, and provide the status back to the client.
- Typically, a Map Reduce application is composed of three Java classes: a Job, a Mapper, and a Reducer.
- Mappers and reducers handle the details of computation on key/value pairs and are connected through a shuffle and sort phase. The Job is responsible of configuring the input and output data format by specifying the InputFormat and OutputFormat classes of data being serialized to and from HDFS.

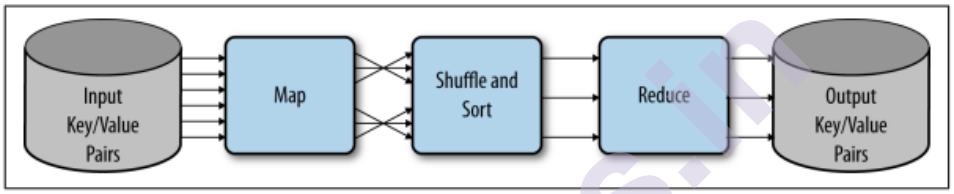



**Fig. 7.7- Stages of a MapReduce Framework**

(Ref - Chapter 2, Fig 2.5 - Data Analytics with Hadoop - An Introduction for Data Scientists)

- The stages a MapReduce Framework are as follows:
  - **Phase I -** Local data is loaded as key/value pairs from HDFS into a mapping process..
  - **Phase II** - Mappers can generate nil or multiple key/value pairs, to map a given key to computed values.
  - **Phase III -** The pairs are then subjected to sort or shuffle operation, selection of operation depends on the key and passed to the reducer so that it can access all of the values for a given key.
  - **Phase IV -** Reducers then reduce the map's output by supplying output with nil or multiple final key/value pairs

## SUMMARY

In this chapter we introduced the concept of data product, its several definitions. We were introduced to Hadoop as a solution to the problem of processing data at a scale. We defined the Data Science & Big Data Pipelines. We introduced and explained in detail the concept of Big Data Operating System: Hadoop and understood the Hadoop file system along with MapReduce

## UNIT END QUESTION

1. Explain the concept of Data Product.
2. How can Hadoop be used to build Data Products at scale?
3. Write a short note on : The Data Science Pipeline
4. Write a short note on : The Big Data Pipeline
5. Write a short note on : Hadoop : The Big Data Operating System
6. Explain Hadoop Architecture
7. Explain the different master and worker services in Hadoop
8. Explain the concept of Hadoop Cluster
9. Explain Hadoop Distributed File System
10. Explain the concept of MapReduce
11. Explain the MapReduce Framework

## REFERENCES


- The book - Data Analytics with Hadoop - An Introduction for Data Scientists authored by Benjamin Bengfort and Jenny Kim


*****

# 8

# HADOOP STREAMING & IN-MEMORY COMPUTATION WITH SPARK

**Unit Structure**





## 8.0 OBJECTIVES

The objectives of this chapter are to:

- Explain the concept of Hadoop Streaming
- Explain the Advanced Map Reduce concepts
- Explain the basics of Apache Spark
- Explain the components of Spark Stack
- Explain the concept of Resilient Distributed Datasets (RDD) in Spark
- Explain the working of a typical Spark application

## 8.1 INTRODUCTION

- Hadoop Streaming is an important tool that allows data scientists to program in R or Python instead of Java to immediately start using Hadoop and MapReduce
- Advanced Map Reduce concepts such as Combiners, Partitioners and Job Chaining play a large role in MapReduce algorithms and optimizations and are essential to understanding Hadoop
- Apache Spark will remain the primary method for interaction with cluster for any new Hadoop user. Apache Spark is the first distributed

computing platform that is not only fast & general-purpose but also popular particularly because of features such as speed and adaptability. It uses a data model called Resilient Distributed Datasets (RDD) that stores the required data in the memory for the duration of computation thereby eliminating intermediate writes to disk.

## 8.2 HADOOP STREAMING

- Hadoop MapReduce framework allows programmers to specify HDFS input /output locations, map-reduce functions, and other parameters through a Java API that provides them as a job configuration.
- However, to use the MapReduce framework Java is not the only option. Hadoop Streaming is a utility written in Java that allows programmers to make the executable of the mapper and reducer independent of programming languages. With Hadoop Streaming shell utilities, R, or Python, scripts can all be used to compose MapReduce jobs.
- Hadoop Streaming is a utility, that is packaged as a JAR file which comes with the Hadoop MapReduce distribution. Concept of Streaming is a normal Hadoop job that is passed to the cluster through the job client allowing you to also specify arguments such as the HDFS input and output paths, along with the mapper and reducer executable.
- Streams in Hadoop Streaming refer to the concept of standard Unix streams: stdin, stdout, and stderr. Streaming utilizes the standard Unix streams for input and output to perform a MapReduce job, hence the name Streaming.
- Input to both mappers and reducers is read from stdin, which a Python process can access
- Hadoop expects the Python mappers and reducers to write their output key/value pairs to stdout.
- Following figure demonstrates the streaming process in a MapReduce context.

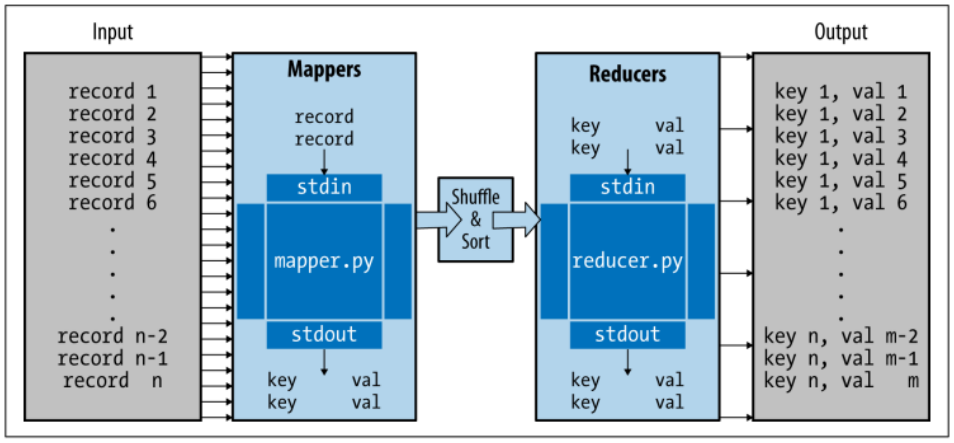



**Fig. 8.1-Streaming in Hadoop using Python**

(Ref - Chapter 3, Fig 3.1 - Data Analytics with Hadoop - An Introduction for Data Scientists)

- When Streaming executes a job, each mapper task will launch the supplied executable inside of its own process.
- The mapper then converts the input data into lines of text and pipes it to the stdin of the external process while simultaneously collecting output from stdout.
- The input conversion is usually a simple and straight forward serialization of the value as data is being read from HDFS having each line as a new value.
- The mapper expects output to be in a string key/value format, where the key is separated from the value by some separator character, tab (\t) by default. If there is no separator, then the mapper considers the output to only be a key with a null value.
- The reducer is launched as its separate executable once the output from the mappers is shuffled and sorted ensuring that each key is sent to the same reducer.
- The output of key/value strings from the mapper are then streamed to the reducer as input through stdin, matching the output from the mapper, and ensures to be grouped by key.
- The output given by the reducer to stdout is supposed to have the same key, separator, and value format as that of the mapper.
- To write Hadoop jobs using Python, we are required to create two separate Python files, mapper.py and a reducer.py. Inside each of these files we have to include the statement import sys to enable access to stdin and stdout.
- The code will accept input as a string, then parse it and after converting for each number or complex data type, it needs to serialize the output as a string.

## 8.3 ADVANCED MAP REDUCE CONCEPTS

- The concepts such as combiners, partitioners, and job chaining play a large role in MapReduce algorithms and optimizations.
- **Combiners** are the primary MapReduce optimization technique
- **Partitioners** are a technique for ensuring there is no bottleneck in the reduce step
- **Job Chaining** is a technique for putting together larger algorithms and data flows.

### 8.3.1 Combiners:

- Mappers produce a lot of intermediate data that must be sent over the network to be shuffled, sorted, and reduced. Since networking is a physical resource, transmission of large amounts of data can lead to job delays resulting in memory bottlenecks
- Combiners are the primary mechanism to solve this problem, and are essentially intermediate reducers that are associated with the mapper output. Combiners reduce network traffic by performing a mapper-local reduction of the data before forwarding it on to the appropriate reducer.

### 8.3.2 Partitioners:

- Partitioners control how keys and their values get sent to individual reducers by dividing up the keyspace.
- The default behaviour is the HashPartitioner, which is often all that is needed. By computing the hash of the key the partitioner allocates keys evenly to each reducer and then assigns the key to a keyspace that is determined by the number of reducers.
- Given a uniformly distributed keyspace, each reducer will get a relatively equal workload. The problem occurs when there is a key imbalance caused when a large number of values are associated with one key. In such a situation, a major portion of the reducers are unutilized, and the benefit of reduction using parallelism is lost.
- A custom partitioner can ease this problem by dividing the keyspace according to some other semantic structure besides hashing.

### 8.3.3 Job Chaining:

- Most complex algorithms cannot be described as a simple map and reduce, so in order to implement more complex analytics, a technique called job chaining is required.
- If a complex algorithm can be decomposed into several smaller MapReduce tasks, then these tasks can be chained together to produce a complete output.
- Job chaining is therefore the combination of many smaller jobs into a complete computation by sending the output of one or more previous jobs into the input of another.
- Linear job chaining produces complete computations by sending the output of one or more MapReduce jobs as the input to another

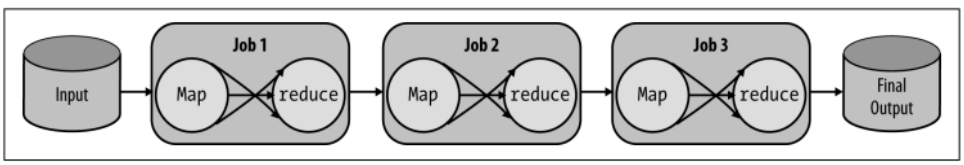



**Fig.8.2- Job Chaining**

(Ref - Chapter 3, Fig 3.2 - Data Analytics with Hadoop - An Introduction for Data Scientists)

- Linear job chaining is a simplification of the more general form of job chaining, which is expressed as a data flow where jobs are dependent on one or more previous jobs.
- Complex jobs are represented as directed acyclic graphs (DAG) that describe how data flows from an input source through each job to the next job and finally as final output

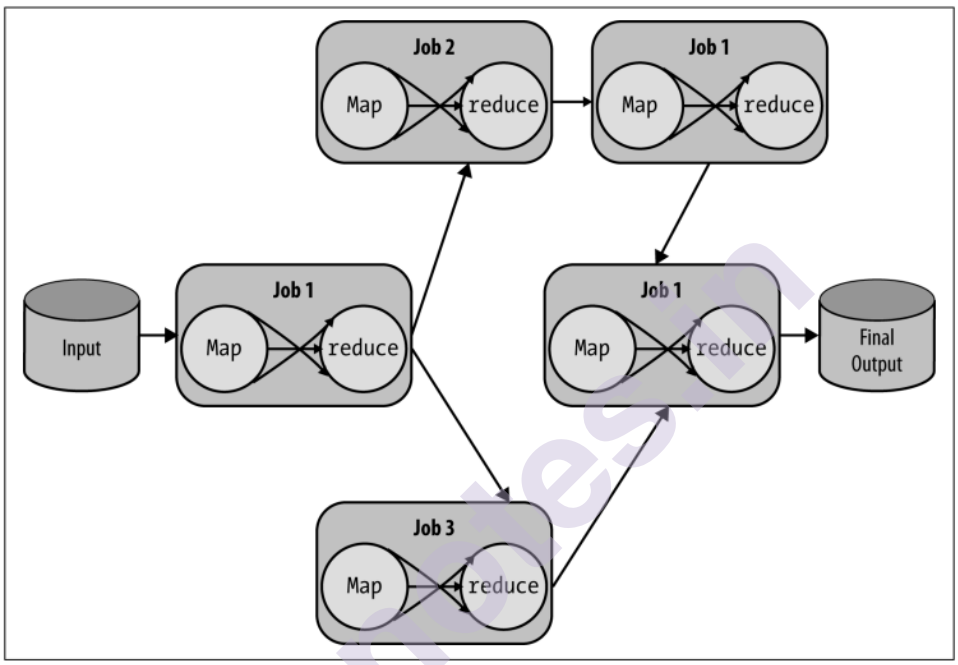



**Fig. 8.3 - An extension of linear chaining Data flow job chaining**

(Ref - Chapter 3, Fig 3.3 - Data Analytics with Hadoop - An Introduction for Data Scientists)

## 8.4 SPARK BASICS

- Apache Spark is a lightning-fast cluster computing technology designed for fast computation. This framework was built above the Hadoop MapReduce extending the MapReduce model to use a variety computations efficiently which includes Interactive Queries and Stream Processing.
- Apache Spark enables distributed programming that is similar to the MapReduce model through an API for that is designed for faster execution on iterative algorithms and interactive queries.
- Spark achieves the goal of executing computations with super speed by caching data in the memory of the cluster nodes initially that may be needed for computation. This allows Spark to execute iterative

algorithms and get back to it without needing to reload it again from disk.

- Spark stores the dataset in memory for the duration of execution of the application thus does not need to reload data during iterations.
- Spark utilizes Hadoop platform in two ways – First by using it for **storage** and second is to use it for **processing**. As Spark has its own computation for cluster management, it uses Hadoop for storage purposes only.
- While Spark is implemented in Scala it provides programming APIs in Scala, Java, R, and Python.

## 8.5 THE SPARK STACK

- Apache Spark – It is a computing platform that is distributed and can execute in a without a cluster called standalone mode. Spark is mainly focused on computation rather than storage of data and as executed in a cluster implementing cluster management and data warehousing tools.
- When Spark is built with Hadoop the task of allocation and management of cluster resources is done by YARN using its ResourceManager. This way Spark can then access any kind of Hadoop data source like HBase, Hive, etc
- Core Spark module is the one that provides Sparks primary programming concepts to programmers and includes the API that contains the functionality to for creating definitions of Resilient Distributed Datasets (RDD).

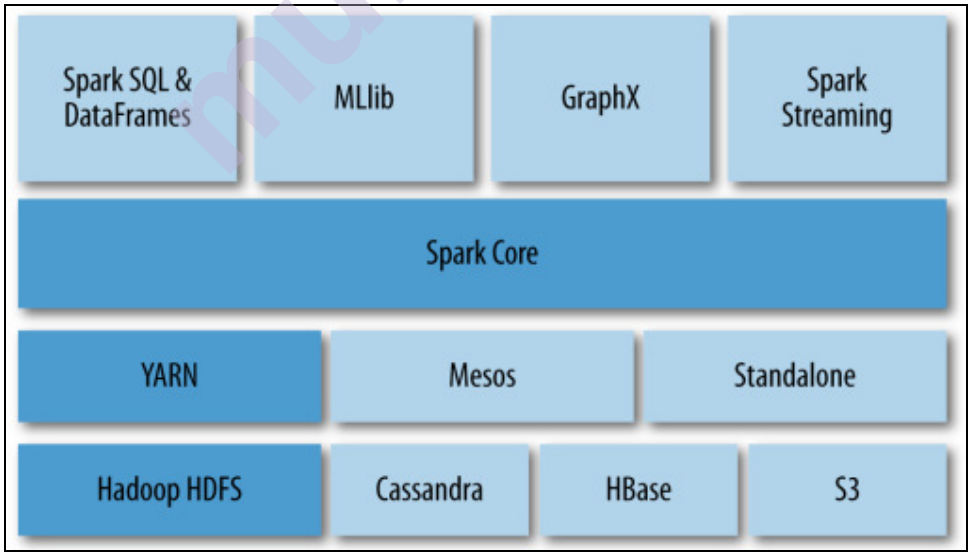



**Fig. 8.4 - Spark framework**

(Ref - Chapter 4, Fig 4.1 - Data Analytics with Hadoop - An Introduction for Data Scientists)

- The primary components of Spark Core are described as follows:
- **Spark SQL**
  - It is a component which provides necessary support for data of types structured and semi-structured.
  - Provides API to interact with Spark.
  - Libraries included provide structured data-processing using DataFrames.
- **Spark Streaming**
  - Enables real time processing and manipulation of unbounded streams of data
  - There are many streaming data libraries available for handling real-time data.
- Spark Streaming makes sure that programmers get advantage by allowing data interaction in a manner similar to interacting with a normal RDD as data starts to come in.
- **MLlib**
  - A library containing machine learning algorithms implemented as Spark operations on RDD.
  - MLib provides developers with scalable algorithms like the ones in machine learning like neural networks, decision tress, classification, regression, etc.
- **GraphX**
  - It provides set of algorithms and tools that allow manipulation of graphs and perform parallel operations on graphs
  - It provides extension to the RDD API to include functions for manipulation of graphs, creation subgraphs, or accessing all vertices in a path.

## 8.6 RESILIENT DISTRIBUTED DATASETS (RDD)

- Spark does not deal with distributed data storage but it depends upon Hadoop to provide storage functionality and itself uses Resilient Distributed Datasets to make distributed computation more reliable .
- RDD is a concept suggesting a collection of objects that is read-only & partitioned across a set of machines.
- RDD can be recreated using knowledge of the sequence of applications of transformations to earlier RDD and are hence fault tolerant and can be accessed using parallel operations. RDD can be

read and written to distributed storages and also provide ability to be cached in the memory of worker nodes for future iterations.

- The feature of in-memory caching helps achieve massive speedups in execution and facilitates iterative computing required in the case of machine learning analyses.
- RDD are executed using functional programming concepts which use the concepts of map and reduce. New RDDs can be created by simply of loading data or by making any transformation to an existing collection of data resulting in generation of a new RDD.
- The knowledge of the sequence of applications of transformations to RDD defines its lineage, and the transformations can be reapplied to the complete collection in to recover from failure as they are immutable.
- Spark API is a collection of operations that is used for creation, transformation and export of RDD
- RDD can be operated upon using transformations and actions.
- **Transformations** – These consist of operations that are applied to an existing RDD for the creation of a new one—for example, application of a filter operation to an RDD for generation of a smaller RDD.
- **Actions** are operations that will return the computed result to the Spark driver program —This results in a coordinating or aggregating of all partitions of an RDD.
- In context of the MapReduce model, **map** is a transformation while **reduce** is an action. In case of map transformation after passing a function to each object present in the RDD it gives an output which is a mapping from old RDD to a new one. In case of reduce operation the RDD has to be partitioned again and an aggregate value like sum or average has to be computed and returned.

## 8.7 A TYPICAL SPARK APPLICATION

- A typical Spark application will execute a dataset in parallel mode throughout the cluster into the RDD

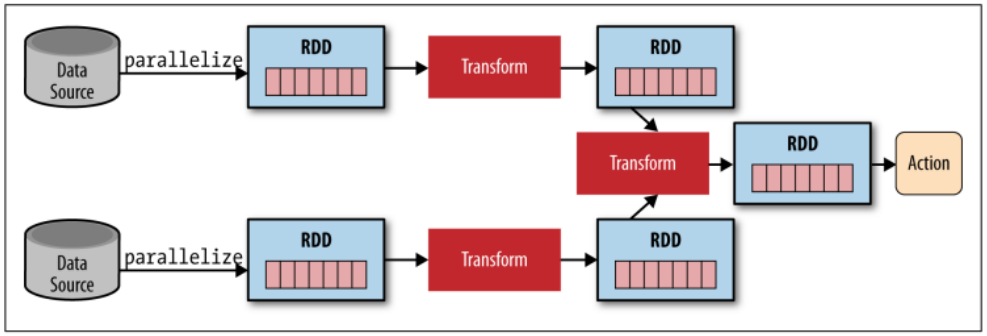



**Fig. 8.5 - Typical Spark Application**

(Ref - Chapter 4, Fig 4.2 - Data Analytics with Hadoop - An Introduction for Data Scientists)

- The code Spark in applications is written in a driver program which is executed on the drivers local machine upon its submission, and when acted upon, this driver code is first distributed across the entire cluster and then its execution is done by workers on their respective partitions. For combining of the results they are sent back to the driver program.
- The driver program creates RDDs by executing a dataset in parallel from a Hadoop data source then it applies transformations to obtain new RDD then finally the action on the new RDD to get the output
- Programming Spark involves the following data flow sequence:
    1. Definition of RDD by accessing data present on disk which involves parallel execution of certain tasks like transformation of an existing RDD. The task of Caching is important in Spark as storing of RDDs inside the memory of a node allows us to achieve super fast access for the performing calculations.
    2. Invoke operations on the RDD by passing closured functions to each element of the RDD. Spark library provides large number of high-level operators other than map and reduce.
    3. Use the output RDDs with aggregation actions like collect, save, count, etc. Progress can be made only when the aggregation has been computed on the cluster and hence Actions signal the end of the computation.

- **The Spark Execution Model**
    - The Spark Execution model brings about its execution through the interaction of the following components: driver, YARN, and workers.
    - The SparkContext in a driver program coordinates the independent execution of processes in Spark applications
    - The context in SparkContext will connect to a cluster manager for allocation of system resources.
    - Management of every worker in the cluster is done by an executor, and management of the executor is done by the SparkContext.
    - The executor coordinates computation, storage and caching on every machine.

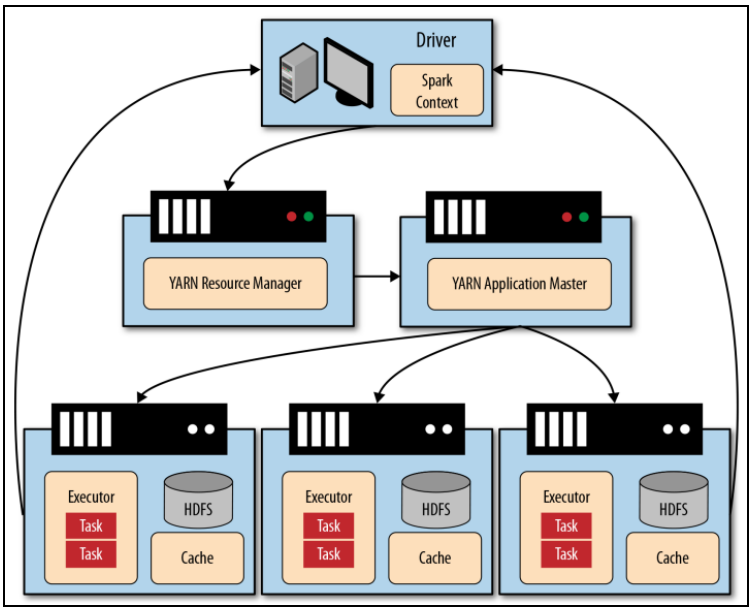



**Fig. 8.6 - The Spark Execution Model**

(Ref - Chapter 4, Fig 4.3 - Data Analytics with Hadoop - An Introduction for Data Scientists)

## SUMMARY

In this chapter we got a thorough understanding of the Hadoop concepts such as Hadoop Streaming, Advanced Map Reduce. We were introduced with the basics of Apache Spark, the components of the Spark Stack and the concept of Resilient Distributed Datasets (RDD) in Spark. Finally the chapter was concluded with the description of working of a typical Spark application.

## UNIT END QUESTION

1. Explain the concept of Hadoop Streaming
2. Explain the following Advanced Map Reduce concepts:

   a) Combiners b) Partitioners c) Job Chaining
3. Explain the Basics of Apache Spark
4. Explain the components of Spark Stack
5. Explain the concept of Resilient Distributed Datasets
6. Write short note on : A typical Spark Application
7. Write short note on : The Spark Execution model
8. What is data science pipeline? Explain in detail with a neat diagram.
9. How to refactor the data science pipeline into an iterative model? Explain all its phases with a neat diagram.
10. List the requirements of distributed system in order to perform computation at scale.

11. How Hadoop addresses these requirements?
12. Write a short note on Hadoop architecture.
13. Explain with a neat diagram a small Hadoop cluster with two master nodes and four workers nodes that implements all six primary Hadoop services.
14. Write a short note on Hadoop Distributed File System.
15. How basic interaction can be done in Hadoop distributed file system? Explain any five basic file system operations with its appropriate command.
16. What are various types of permissions in Hadoop distributed file system? What are different access levels? Write and explain commands to set various types and access levels. What is a caveat with file permissions on HDFS?
17. Explain functionality of map() function and reduce() function with a neat diagram in a MapReduce context.
18. How MapReduce can be implemented on a Cluster? Explain its all phases with a neat diagram.
19. Explain the details of data flow in a MapReduce pipeline executed on a cluster of a few nodes with a neat diagram.
20. Write a short note on job chaining.
21. Demonstrate the process of Hadoop streaming in a MapReduce context.
22. Demonstrate the process of Computing on CSV Data with Hadoop Streaming.
22. Demonstrate the process of executing a Streaming job on a Hadoop cluster.
23. Write a short note on Combiners in advanced MapReduce context.
24. Write a short note on Partitioners in advanced MapReduce context.
25. Write a short note on Job Chaining in advanced MapReduce context.
26. Write in brief about Spark. Also write and explain its primary components.



## REFERENCES


- Data Analytics with Hadoop - An Introduction for Data Scientists by Benjamin Bengfort and Jenny Kim
- Apache Spark https://www.tutorialspoint.com/apache_spark/index.html
- RDD - https://spark.apache.org/docs/latest/rdd-programming-guide.html
- Apache Spark - https://spark.apache.org/docs/latest/quick-start.html


*****

# UNIT V

# 9

# DISTRIBUTED ANALYSIS AND PATTERNS

**Unit Structure**



## 9.0 OBJECTIVES

To Study and Understand the following concept

- Distributed Analysis and Patterns,
- Computing with Keys
- Design Patterns
- Last-Mile Analytics

## 9.1 A DISTRIBUTED ANALYSIS AND PATTERNS

MapReduce and Spark allow developers and data scientists the ability to easily con- duct data parallel operations, where data is distributed to multiple processing nodes and computed upon simultaneously, then reduced to a final output. YARN provides simple task parallelism by allowing a cluster to perform multiple different operations simultaneously by allocating free computational resources to perform individual tasks. Parallelism reduces the amount of time required to perform a single computation, thereby unlocking datasets that are measured in petabytes, analyzed at thou- sands of records per second, or composed of multiple heterogeneous data sources. However, most parallel operations like the ones described to this point are simple, leading to the question, how can data scientists conduct advanced data analysis at scale?

The primary principle of conducting large-scale analytics can be summarized by the quip from Creighton Abrams: “When eating an elephant, take one bite at a time.” Whereas single operations take many small bites of the data, these operations must be composed into a step-by-

step sequence called a data flow to be organized into more meaningful results. Data flows may fork and merge, allowing for both task and data parallelism if two operations can be computed simultaneously, but the sequence must maintain the property that data is fed sequentially from an input data source to a final output. For that reason, data flows are described as directed acyclic graphs (DAGs). It is important, therefore, to realize that if an algorithm, analysis, or other non-trivial computation can be expressed as a DAG, then it can be parallelized on Hadoop.

Unfortunately, it also quickly becomes apparent that many algorithms aren't easily converted into DAGs, and are therefore unsuitable for this type of parallelism. Algo- rithms that cannot be described as a directed data flow include those that maintain or update a single data structure throughout the course of computation (requiring some shared memory) or computations that are dependent on the results of another at intermediate steps (requiring intermediate interprocess communication). Algorithms that introduce cycles, particularly iterative algorithms that are not bounded by a finite number of cycles, are also not easily described as DAGs.

There are tools and techniques that address requirements for cyclicity, shared mem- ory, or interprocess communication in both MapReduce and Spark, but to make use of these tools, algorithms must be rewritten to a distributed form. Rather than rewrite algorithms, a less technical but equally effective approach is usually employed: design a data flow that decomposes the input domain into a smaller output that fits into the memory of a single machine, run the sequential algorithm on that output, then vali- date that analysis across the cluster with another data flow (e.g., to compute error).

It is because of the widespread use of this approach that Hadoop is often said to be a preprocessor that unlocks the potential of large datasets by reducing them into increasingly manageable chunks through every operation. A common rule of thumb is use either MapReduce or Spark to articulate data down to a computational space that can fit into 128 GB of memory (a cost-effective hardware requirement for a sin- gle machine). This rule is often called "last-mile" computing because it moves data from an extremely large space to a place close enough, the last mile, that allows for accurate analyses or application-specific computations.

In this chapter, we explore patterns for parallel computations in the context of data flows that reduce or decompose the computational space into a more manageable one. We begin by discussing key-based computations, a requirement for MapReduce and also essential to Spark. This leads us to a discussion of patterns for summariza- tion, indexing, and filtering, which are key components to most decomposition algorithms. In this context, we will discuss applications for statistical summarization, sampling, search, and binning. We conclude by surveying three preprocessing techniques for computing regression, classification, and clustering style analyses

This chapter also presents standard algorithms that are used routinely for data analytics, including statistical summarization (the parallel "describe" command), parallel grep, TF-IDF, and canopy clustering. Through these examples, we will clarify the basic mechanics of both MapReduce and Spark.

## 9.1 COMPUTING WITH KEYS

The first step toward understanding how data flows work in practice is to understand the relationship between key/value pairs and parallel computation. In MapReduce, all data is structured as key/value pairs in both the map and reduce stages. The key requirement relates primarily to reduction, as aggregation is grouped by the key, and parallel reduction requires partitioning of the key space—in other words, the domain of key values such that a reducer task sees all values for that key. If you don't necessarily have a key to group by (which is actually very common), you could reduce to a single key that would force a single reduction on all mapped values. However, in this case, the reduce phase would not benefit from parallelism.

Although often ignored (especially in the mapper, where the key is simply a document identifier), keys allow the computation to work on sets of data simultaneously. Therefore, a data flow expresses the relation of one set of values to another, which should sound familiar, especially presented in the context of more traditional data management—structured queries on a relational database. Similar to how you would not run multiple individual queries for an analysis of different dimensions on a data- base like PostgreSQL, MapReduce and Spark computations look to perform grouping operations in parallel, as shown by the mean computation grouped by key in Figure 9-1.

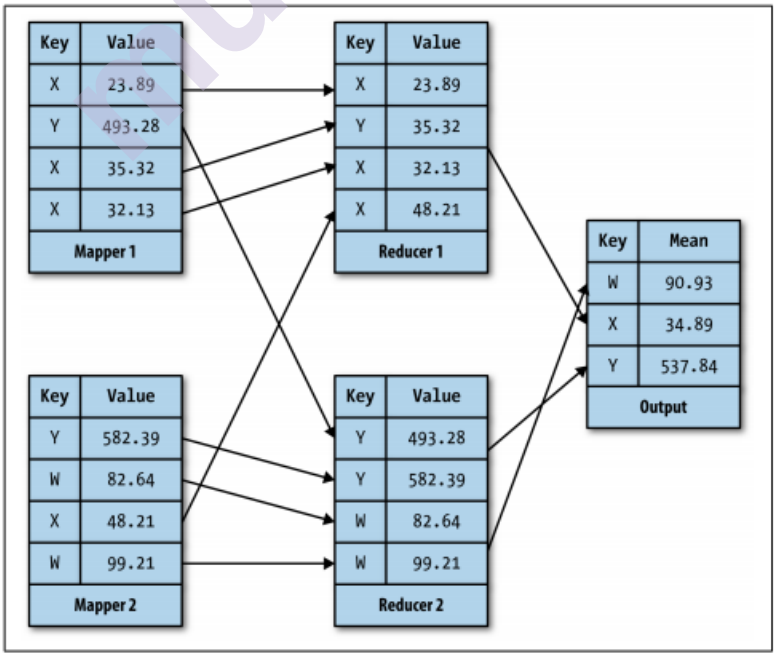



Figure 9-1. Keys allow parallel reduction by partitioning the key space to

multiple reducers

Moreover, keys can maintain information that has already been reduced at one stage in the data flow, automatically parallelizing a result that is required for the next step in computation. This is done using compound keys—a technique discussed in the next section that shows that keys do not need to be simple, primitive values. Keys are so useful for these types of computations, in fact, that although they are not strictly required in computations with Spark (an RDD can be a collection of simple values), most Spark applications require them for their analyses, primarily using groupByKey, aggregateByKey, sortByKey and reduceByKey actions to collect and reduce.

### 9.1.1 Compound Keys:

Keys need not be simple primitives such as integers or strings; instead, they can be compound or complex types so long as they are both *hashable* and *comparable*. Comparable types must at the very least expose some mechanism to determine equality and some method of ordering. Comparison is usually accomplished by mapping some type to a numeric value (e.g., months of the year to the integers 1-12) or through a lexical ordering. Hashable types in Python are any immutable type, the most notable of which is the tuple. Tuples can contain mutable types (e.g., a tuple of lists), however, so a hashable tuple is one that is composed of immutable types. Mutable types such as lists and dictionaries can be transformed into immutable tuples:

```
# Transform a list into a tuple
key = tuple(['a', 'b', 'c'])

# Transform a dictionary into a tuple of tuples
key = {'a': 1, 'b': 2}
key = tuple(key.items())
```

Compound keys are used in two primary ways: to facet the keyspace across multiple dimensions and to carry key-specific information forward through computational stages that involve the values alone. Consider web log records of the following form:

```
local - - [30/Apr/1995:21:18:07 -0600] "GET 7448.html HTTP/1.0" 404 -
local - - [30/Apr/1995:21:18:42 -0600] "GET 7448.html HTTP/1.0" 200 980
remote - - [30/Apr/1995:21:22:56 -0600] "GET 4115.html HTTP/1.0" 200 1363
remote - - [30/Apr/1995:21:26:29 -0600] "GET index.html HTTP/1.0" 200 2881
```

Web log records are a typical data source of big data computations on Hadoop, as they represent per-user clickstream data that can be easily mined for insight in a variety of domains; they also tend to be very large, dynamic semistructured datasets, well suited to operations in Spark and MapReduce. Initial computation on this dataset requires a frequency analysis; for example, we can decompose the text into two daily time series, one for local traffic and the other for remote traffic using a compound key:

```
import re
from datetime import datetime

# Parse datetimes in the log record
dtfmt = "%d/%b/%Y:%H:%M:%S %z"

# Parse log records using a regular expression
linre = re.compile(r'^(\w+) \- \- \[(.+)\] "(.+)" (\d+) ([\d\-]+)$')

def parse(line):
    # Match the log record against our regular expression
    match = linre.match(line)
    if match is not None:
        # The regular expression has groups to extract the source, timestamp,
        # the request, the status code, and the byte size of the response.
        parts = match.groups()

        # Parse the datetime and return the source, along with the year and day.
        date  = datetime.strptime(parts[1], dtfmt).timetuple()
        return (parts[0], date.tm_year, date.tm_yday)
```

Mapping yields the following data from the preceding dataset:

```
('local', 1995, 120)    1
('local', 1995, 120)    1
('remote', 1995, 120)    1
('remote', 1995, 120)    1
```

**Compound data serialization:**

The final consideration when using compound keys (and complex values) is to understand *serialization* and *deserialization* of the compound data. *Serialization* is the process of turning an object in memory into a stream of bytes such that it can be written to disk or transmitted across the network *(deserialization* is the reverse process). This process is essential, particularly in MapReduce, as keys and values are written (usually as strings) to disk between map and reduce phases

By default in Spark, the Python API uses the pickle module for serialization, which means that any data structures you use must be pickle-able. With MapReduce Streaming, you must serialize both the key and the value as a string, separated by a specified character, by default a tab (\t).

One common first attempt is to simply serialize an immutable type (e.g., a tuple) using the built-in str function, converting the tuple into a string that can be easily pickled or streamed. The problem then shifts to deserialization; using the ast (abstract syntax tree) in the Python standard library, we can use the literal_eval function to evaluate stringified tuples back into Python tuple types as follows:

```
import ast

def map(key, val):
    # Parse the compound key, which is a tuple.
    key = ast.literal_eval(key)

    # Write out the new key as a string
    return (str(key), val)
```

As both keys and values get more complex. Other data structures for serialization is Base64-encoded JSON because it is compact, uses only ASCII characters, and is easily serialized and deserialized with the standard library as follows:

```
import json
import base64

def serialize(data):
    """
    Returns the Base64-encoded JSON representation of the data (keys or values)
    """
    return base64.b64encode(json.dumps(data))

def deserialize(data):
    """
    Decodes Base64 JSON-encoded data
    """
    return json.loads(base64.b64decode(data))
```

However, take care when using more complex serial representations; often there is a trade-off in the computational complexity of serialization versus the amount of space used.

### 9.1.2-Keyspace Patterns:

The notion of computing with keys allows you to manage sets of data and their relations. However, keys are also a primary piece of the computation, and as such, they must be managed in addition to the data. There several patterns that impact the *keyspace,* specifically the explode, filter, transform, and identity patterns.

For the following examples, we will consider a dataset of orders whose key is the order ID, customer ID, and timestamp, and whose value is a list of universal product codes (UPCs) for the products purchased in the order as follows:

```
1001, 1063457, 2014-09-16 12:23:33, 098668259830, 098668318865
1002, 0171488, 2014-12-11 03:05:03, 098668318865
1003, 1022739, 2015-01-03 13:01:54, 098668275427, 098668331789, 098668274321
```

### Transforming the keyspace:

The most common key-based operation is a transformation of the input key domain, which can be conducted either in a map or a reduce. The most common transformation functions are direct assignment, compounding, splitting, and inversion.

Direct assignment drops the input key, which is usually entirely ignored, and constructs a new key from the input value or another source (e.g., a random key). Compounding constructs or adds to a compound key, increasing the faceting of the key relation. Splitting breaks apart a compound key and uses only a smaller piece of it. It is, however, appropriate to also drop unneeded data and eliminate extraneous information via compounding or splitting.

For example, in order to sort a dataset by value rather than by key, it is necessary to first map the inversion of the key and value, perform a sortByKey or utilize the shuffle and sort in MapReduce, then reinvert in the reduce or with another map. Consider a job to sort our orders by the number of products in each order, along with the date, which will use all of the keyspace transformations identified earlier

```
# Load orders into an RDD and parse the CSV
orders = sc.textFile("orders.csv").map(split)

# Key assignment: (orderid, customerid, date), products
orders = orders.map(lambda r: ((r[0], r[1], r[2]), r[3:]))

# Compute the order size and split the key to orderid, date
orders = orders.map(lambda (k, v): ((k[0], parse_date(k[2])), len(v)))

# Invert the key and value to sort
orders = orders.map(lambda (k, v): ((v, k[1]), k[0]))

# Sort the orders by key
orders = orders.sortByKey(ascending=False)

# Reinvert the key/value space so that we key on order ID again
orders = orders.map(lambda (k,v ): (v, k))

# Get the top ten order IDs by size and date
print orders.take(10)
```



This example is perhaps a bit verbose for the required task, but it does demonstrate each type of transformation as follows:

1. First, the dataset is loaded from a CSV using the split method.
2. At this point, orders is only a collection of lists, so we assign keys by breaking the value into the IDs and date as the key, and associate it with the list of products as the value.
3. The next step is to get the length of the products list (number of products ordered) and to parse the date, using a closure that wraps a date format for date time.strptime; note that this method splits the compound key and eliminates the customer ID, which is unnecessary.
4. In order to sort by order size, we need to invert the size value with the key, also splitting the date from the key so we can also sort by date.
5. After performing the sort, this function reinverts so that each order can be identified by size and date.

The following snippet demonstrates what happens to the first record throughout each map in the Spark job:

```
0. "1001, 1063457, 2014-09-16 12:23:33, 098668259830, 098668318865"
1. [1001, 1063457, 2014-09-16 12:23:33, 098668259830, 098668318865]
2. ((1001, 1063457, 2014-09-16 12:23:33), [098668259830, 098668318865])
3. ((1001, datetime(2014, 9, 16, 12, 23, 33), 2)
4. ((2, datetime(2014, 9, 16, 12, 23, 33)), 1001)
5. (1001, (2, datetime(2014, 9, 16, 12, 23, 33)))
```

Through this series of transformations, the client program can then take the top 10 orders by size and date, and print them out after the distributed computation.

**The explode mapper:**

The explode mapper generates multiple intermediate key/value pairs for a single input key. Generally, this is done by a combination of a key shift and splitting of the value into multiple parts. An explode mapper can also generate many intermediate pairs by dividing a value into its constituent parts and reassigning them with the key. We can explode the list of products per order value to order/product pairs, as in the following code:

```
def order_pairs(key, products):
    # Returns a list of order id, product pairs
    pairs = []

    for product in products:
        pairs.append((key[0], product))

    return pairs

orders = orders.flatMap(order_pairs)
```

Applying this mapper to our input dataset produces the following output:

```
1001, 098668259830
1001, 098668318865
1002, 098668318865
1003, 098668275427
1003, 098668331789
1003, 098668274321
```

Note the use of the flatMap operation on the RDD, which is specifically designed for explode mapping. It operates similarly to the regular map; however, the function can yield a sequence instead of a single item, which is then chained into a single collection (rather than an RDD of lists).

**The filter mapper:**

Filtering is often essential to limit the amount of computation performed in a reduce stage, particularly in a big data context. It is also

used to partition a computation into two paths of the same data flow, a sort of data-oriented branching in larger algorithms that is specifically designed for extremely large datasets.

Spark provides a filter operation that takes a function and transforms the RDD such that only elements on which the function returns True are retained. This example shows a more advanced use of a closure and a general filter function that can take any year. The partial function creates a closure whose year argument to year_filter is always 2014, allowing for a bit more versatility. MapReduce code is similar but requires a bit more logic:

```
def YearFilterMapper(Mapper):

    def __init__(self, year, **kwargs):
        super(YearFilterMapper, self).__init__(**kwargs)
        self.year = year

    def map(self):
        for key, value in self:
            if parse_date(key[2]).year == self.year:
                self.emit(key, value)

if __name__ == "__main__":
    mapper = YearFilterMapper(2014)
    mapper.map()
```

It is completely acceptable for a mapper to not emit anything, therefore the logic for a filter mapper is to only emit when the condition is met.

**The identity pattern:**

The final keyspace pattern that is commonly used in MapReduce is the Identity function. This is simply a pass-through, such that identity mappers or reducers return the same value as their input (e.g., as in the *identity function,* f(x) = x). Identity mappers are typically used to perform multiple reductions in a data flow. When an identity reducer is employed in MapReduce, it makes the job the equivalent of a sort on the keyspace. Identity mappers and reducers are implemented simply as follows:

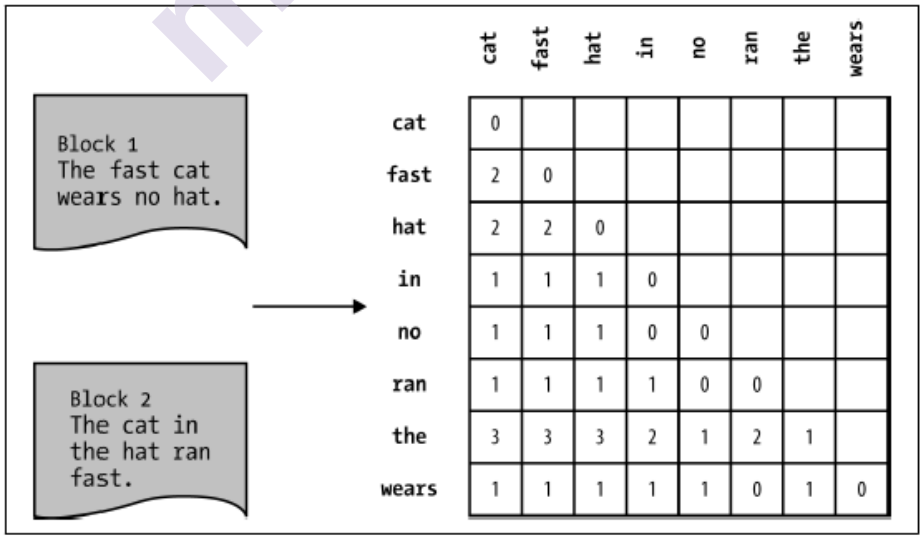
```
class IdentityMapper(Mapper):

    def map(self):
        for key, value in self:
            self.emit(key, value)


class IdentityReducer(Reducer):

    def reduce(self):
        for key, values in self:
            for value in values:
                self.emit(key, value)
```

Identity reducers are generally more common because of the optimized shuffle and sort in MapReduce. However, identity mappers are also very important, particularly in chained MapReduce jobs where the output of one reducer must immediately be reduced again by a secondary reducer.

### 9.1.3 Pairs versus Stripes:

There are two ways that matrices are commonly represented: by *pairs* and by *stripes*. Both pairs and stripes are examples of key-based computation. To explain the motivation behind this example, consider the problem of building a word co-occurrence matrix for a text-based corpus.

The word co-occurrence matrix as shown in Figure 9-2 is a square matrix of size NxN, where N is the vocabulary (the number of unique words) in the corpus. Each cell Wij contains the number of times both word $w_i$ and word $w_j$ appear together in a sentence, paragraph, document, or other fixed-length window. This matrix is sparse, particularly with aggressive stopword filtering because most words only co-occur with very few other words on a regular basis.




| | cat | fast | hat | in | no | ran | the | wears |
|---|---|---|---|---|---|---|---|---|
| cat | 0 | | | | | | | |
| fast | 2 | 0 | | | | | | |
| hat | 2 | 2 | 0 | | | | | |
| in | 1 | 1 | 1 | 0 | | | | |
| no | 1 | 1 | 1 | 0 | 0 | | | |
| ran | 1 | 1 | 1 | 1 | 0 | 0 | | |
| the | 3 | 3 | 3 | 2 | 1 | 2 | 1 | |
| wears | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 |

**Figure 9.2 A word co-occurrence matrix demonstrates the frequency of terms apperaring together in the same block of text such as a sentence**

The *pairs* approach maps every cell in the matrix to a particular

value, where the pair is the compound key i, j. Reducers therefore work on per-cell values to produce a final, cell-by-cell matrix. This is a reasonable approach, which yields output where each *Wij* is computed upon and stored separately. Using a sum reducer, the mapper is as follows:

```
from itertools import combinations

class WordPairsMapper(Mapper):

    def map(self):
        for docid, document in self:
            tokens = list(self.tokenize(document))
            for pair in combinations(sorted(tokens), 2):
                self.emit(pair, 1)
```

Here is the input:

```
"See Spot run, run Spot, run!"
```

The pairs of the word co-occurrence matrix would be aggregated as follows:

```
(run, run), 6
(run, see), 3
(run, spot), 6
(see, run), 3
(see, spot), 2
(spot, run), 6
(spot, see), 2
(spot, spot), 1
```

While the pairs approach is easy to implement and understand, it causes a lot of intermediate pairs that must be transmitted across the network both during the MapReduce shuffle and sort phase, and during groupByKey operations to shuffle values between partitions in an RDD. Moreover, the pairs approach is not well suited to computations that require an entire row (or column) of data.

The stripes approach was initially conceived as an optimization to reduce the number of intermediate pairs and reduce network communication in order to make jobs faster. However, it also quickly became an essential tool in many algorithms that need to perform fast per-element computations—for example, relative frequencies or other statistical operations. Instead of pairs, a per-term associative array (a Python dictionary) is constructed in the mapper and emitted as a value:

```
from collections import Counter

class WordStripeMapper(Mapper):

    def map(self):
        for docid, document in self:
            tokens = list(self.tokenize(document))

            for i, term in enumerate(tokens):
```

```
                # Create a new stripe for each term
                stripe = Counter()

                for j, token in enumerate(tokens):
                    # Don't count the term's co-occurrence with itself
                    if i != j:
                        stripe[token] += 1

                # Emit the term and the stripe
                self.emit(term, stripe)


class StripeSumReducer(Reducer):

    def reduce(self):
        for key, values in self:
            stripe = Counter()

            # Add all the mapped counters together
            for value in values:
                for token, count in value.iteritems():
                    # For each token, count add the collector stripe
                    stripe[token] += count

            self.emit(key, stripe)
```

The stripes approach is not only more compact in its representation, but also generates fewer and simpler intermediary keys, thus optimizing sorting and shuffling of data or other optimizations. However, the stripes object is heavier, both in terms of processing time as well as the serialization requirements, particularly if the stripes get very large. There is a limit to the size of a stripe, particularly in very dense matrices, which may take a lot of memory to track single occurrences.

## 9.2 DESIGN PATTERNS

Design patterns are a special term in software design: generic, reusable solutions for a particular programming challenge. We can explore functional design patterns for solving parallel computations in both MapReduce and Spark. These patterns show a generic strategy and principle that can be used in more complex or domain-specific roles.

Donald Miner and Adam Shook explore 23 design patterns for common MapReduce jobs. They loosely categorize them as follows:

**Summarization:**
Provide a summary view of a large dataset in terms of aggregations, grouping, statistical measures, indexing, or other high-level views of the data.

**Filtering:**
Create subsets or samples of the data based on a fixed set of criteria, without modifying the original data in any way.

**Data Organization:**
Reorganize records into a meaningful pattern in a way that doesn't necessarily imply grouping. This task is useful as a first step to further computations.

**Joins:**
Collect related data from disparate sources into a unified whole.

**Metapatterns:**
Implement job chaining and job merging for complex or optimized computations. These are patterns associated with other patterns.

**Input and output:**
Transform data from one input source to a different output source using data manipulation patterns, either internal to HDFS or from external sources.

### 9.2.1 Summarization:

Summarization attempts to describe the largest amount of information about a dataset as simply as possible. We are accustomed to executive summaries that highlight the primary take-aways of a longer document without getting into the details. Similarly, descriptive statistics attempt to summarize the relationships between observations by measuring their central tendency (mean, median), their dispersion (standard deviation), the shape of their distribution (skewness), or the dependence of variables on each other (correlation).

MapReduce and Spark in principle apply a sequence of summarizations distilling the most specific form of the data (each individual record) to a more general form. Broadly speaking, we are most familiar with summarization as characterized by the following operations:

- Aggregation (collection to a single value such as the mean, sum, or maximum)
- Indexing (the functional mapping of a value to a set of values)
- Grouping (selection or division of a set into multiple sets)

**Aggregation:**

An aggregation function in the context of MapReduce and Spark is one that takes two input values and produces a single output value and is also *commutative* and *associative* so that it can be computed in parallel. Addition and multiplication are commutative and associative, whereas subtraction and division are not.

Aggregation is the general application of an operation on a collection to create a smaller collection (gathering together), and reduction is generally considered an operation that reduces a collection into a single

value. Aggregation can also be thought of as the application of a series of smaller reductions. With this context, it's easy to see why associativity and commutativity are necessary for parallelism.

Consider the standard dataset descriptors: mean, median, mode, minimum, maximum, and sum. Of these, *summation, minimum,* and *maximum* are easily implemented because they are both associative and commutative. *Mean, median,* and *mode,* however, are not. Although there are parallel approximations for these computations, it is important to be aware that some care should be taken when performing these types of analyses.

**Statistical summarization:**

We can simplify and summarize large datasets by grouping instances into keys and describing the per-key properties. Rather than implementing a MapReduce job for each descriptive metric individually (costly), we're going to run all six jobs together in a single batch, computing the count, sum, mean, standard deviation, and range (minimum and maximum).

The basic strategy will be to map a collection of counter values for each computation we want to make on a per-key basis. The reducer will then apply each operation independently to each item in the value collection, using each as necessary to compute the final output (e.g., mean depends on both a count and a sum). Here is the basic outline for such a mapper:

```
class StatsMapper(Mapper):

    def map(self):
        for key, value in self:
            try:
                value = float(value)
                self.emit(key, (1, value, value ** 2))
            except ValueError:
                # Could not parse the value, ignore.
                pass
```

In this case, the three operations that will be directly reduced are count, sum, and sum of squares. Therefore, this mapper emits on a per-key basis, a 1 for count, the value for summation, and the square of the value for the sum of the squares. The reducer uses the count and sum to compute the mean, the value to compute the range, and the count, sum, and sum of squares to compute the standard deviation as follows:

```
from ast import literal_eval as make_tuple

class StatsReducer(Reducer):

    def reduce(self):
        for key, values in self:
            # Parse the values from the mapper
            values = make_tuple(values)
             count   = 0
             delay   = 0.0
             square  = 0.0
             minimum = None
             maximum = None

             for value in values:
                 count  += value[0]
                 delay  += value[1]
                 square += value[2]
    if minimum is None or value[1] < minimum:
        minimum = value[1]

    if maximum is None or value[1] > maximum:
        maximum = value[1]

mean   = delay / float(count)
stddev = math.sqrt((square-(delay**2)/count)/count-1)

self.emit(key, (count, mean, stddev, minimum, maximum))
```

The reducer utilizes the ast.literal_eval mechanism of deserialization to parse the value tuple, then performs a *single* loop over the data values to compute the various sums, minimums, and maximums.
Our mapper must extend the value with minimum and maximum counters, such that the minimum and maximum values are tracked with each value through the reduction as follows:

```
    def counters(item):
        """
        Parses a key/value pair into the key and summary counters.
        A counter is as follows: (count, total, square, minimum, maximum).
        """
        key, value = item # Break apart the item tuple

        try:
            value = float(value)
            self.emit(key, (1, value, value ** 2, value, value))
        except ValueError:
            # Could not parse the value, ignore.
            pass

    def aggregation(first, second):
        """
        For two (key, counter) items, perform summary aggregations.
        """
        count1, total1, squares1, min1, max1 = first
        count2, total2, squares2, min2, max2 = second
      minimum = min((min1, min2))
      maximum = max((max1, max2))
      count   = count1 + count2
      total   = total1 + total2
      squares = squares1 + squares2

      return (count, total, squares, minimum, maximum)

def summary(aggregate):
    """
    Compute summary statistics from aggregation.
    """
    (key, (count, total, square, minimum, maximum)) = aggregate

    mean   = total / float(count)
    stddev = math.sqrt((square-(total**2)/count)/count-1)

    return (key, (count, mean, stddev, minimum, maximum))


def main(sc):
    """
    Primary analysis mechanism for Spark application
    """

    # Given a dataset of key/value pairs, map to counters
    dataset = dataset.map(counters)

    # Perform summary aggregation by key
    dataset = dataset.reduceByKey(aggregation)
    dataset = dataset.map(summary)

    # Write the results out to disk
    dataset.saveAsTextFile("dataset-summary")
```



We can't simply perform our final computation during the aggregation, and another map is needed to finalize the summarization across the (much smaller) aggregated RDD.

The describe example provides a useful pattern for computing multiple features simultaneously and returning them as a vector. This pattern is reused often, particularly in the machine learning context, where multiple procedures might be required in order to produce an instance to train on (e.g., quadratic computations, normalization, imputation, joins, or more specific machine learning tasks). Understanding the difference between aggregation implementations in MapReduce versus Spark can make a lot of difference in tracking down bugs and porting code from MapReduce to Spark and vice versa.

### 9.2.2 Indexing:

In contrast to aggregation-based summarization techniques, *indexing* takes a many- to-many approach. While aggregation collects several records into a single record, indexing associates several records to one or more indices. In databases, an index is a specialized data structure that is used for fast lookups, usually a binary-tree (B-Tree). In Hadoop/Spark, indices perform a similar function, though rather than being maintained and updated, they are typically generated as a first step to downstream computation that will require fast lookups.

Text indexing has a special place in the Hadoop algorithm pantheon due to Hadoop's original intended use for creating search applications. When dealing with only a small corpus of documents, it may be possible to scan the documents looking for the search term like grep does. However, as the number of documents and queries increases, this quickly becomes unreasonable. In this section, we take a look at two types of text- based indices, the more common inverted index, as well as term frequency-inverse document frequency (TF-IDF), a numerical statistic that is associated with an index and is commonly used for machine learning.

**Inverted index:**

An *inverted index* is a mapping from an index term to locations in a set of documents (in contrast to forward indexing, which maps from documents to index terms). In full text search, the index terms are search terms: usually words or numbers with stopwords removed (e.g., very common words that are meaningless in search). Most search engines also employ some sort of stemming or lemmatization: multiple words with the same meaning are categorized into a single word class (e.g., "running", "ran", "runs" is indexed by the single term "run").

The search example shows the most common use case for an inverted index: it quickly allows the search algorithm to retrieve the subset of documents that it must rank and return without scanning every single document. For example, for the query "running bear", the index can be used to look up the intersection of documents that contain the term "running" and the term "bear". A simple ranking system might then be employed to return documents where the search terms are close together rather than far apart in the document (though obviously modern search ranking systems are far more complex than this).

The search example can be generalized, however, to a machine learning context. The index term does not necessarily have to be text; it can be any piece of a larger record. Moreover, the task of using an index to simplify or speed up downstream computation (like the ranking) is common. Depending on how the index is created, there can be a trade-off between performance and accuracy, or, given a stochastic index, between precision and recall.

we would use an identity reducer and the following mapper (note that the same algorithm is easily implemented with Spark):

```
class CharacterIndexMapper(Mapper):

    def map(self):
        for row in self:
            row = row.split("\t")            # split the tab parts
            if not len(row) >= 3: continue   # ensure we have data

            if row[1] != "":
                # If we have a character, emit the name and the docid/lineno.
                self.emit(row[1], row[0])
```

The output of the character index job is a list of character names, each of which corresponds to a list of lines where that character starts speaking. This can be used as a lookup table or as input to other types of analysis.

**TF-IDF:**

*Term frequency-inverse document frequency* (TF-IDF) is now probably the most commonly used form of text-based summarization and is currently the most commonly used feature of documents in text-based machine learning. TF-IDF is a metric that defines the relationship between a term (a word) and a document that is part of a larger corpus. In particular, it attempts to define how important that word is to that particular document given the word's relative frequency in other documents.

We include this algorithm with indexing for a similar reason that we included the simpler inverted indexing example: it creates a data structure that is typically used for downstream computations and machine learning. Moreover, this more complex example highlights something we've only touched upon in other sections: the use of job chaining to compute a single algorithm. With that in mind, let's take a look at the MapReduce implementation of TF-IDF.

Typically, the tf-idf weight is composed by two terms: the first computes the normalized Term Frequency (TF), aka. the number of times a word appears in a document, divided by the total number of words in that document; the second term is the Inverse Document Frequency (IDF),

computed as the logarithm of the number of the documents in the corpus divided by the number of documents where the specific term appears.

- **TF: Term Frequency**, which measures how frequently a term occurs in a document. Since every document is different in length, it is possible that a term would appear much more times in long documents than shorter ones. Thus, the term frequency is often divided by the document length (aka. the total number of terms in the document) as a way of normalization:

*TF(t) = (Number of times term t appears in a document) / (Total number of terms in the document).*

- **IDF: Inverse Document Frequency**, which measures how important a term is. While computing TF, all terms are considered equally important. However it is known that certain terms, such as "is", "of", and "that", may appear a lot of times but have little importance. Thus we need to weigh down the frequent terms while scale up the rare ones, by computing the following:

*IDF(t) = log_e(Total number of documents / Number of documents with term t in it).*

Consider a document containing 100 words wherein the word *cat* appears 3 times. The term frequency (i.e., tf) for *cat* is then (3 / 100) = 0.03. Now, assume we have 10 million documents and the word *cat* appears in one thousand of these. Then, the inverse document frequency (i.e., idf) is calculated as log(10,000,000 / 1,000) = 4. Thus, the Tf-idf weight is the product of these quantities: 0.03 * 4 = 0.12.

### 9.2.3 Filtering:

*Filtering* is one of the primary methods of coarse-grained data reduction for downstream computation. Unlike aggregation methods, which reduce the input space through a high-level overview over a set of groups, filtering is intended to reduce the computational space through omission. In fact, many filtering tasks are a perfect fit for map-only jobs, which do not require reducers because mappers are so well suited to this task. This can be considered filtering by predicate or by selection, similar to a where clause in a SQL statement.

Other filtering tasks may leverage reducers in order to accumulate a representative dataset or to perform some per-values filtering constraint. Examples of this style of filtering include finding the n-largest or n-smallest values, performing deduplication, or subselection. A very common filtering task in analytics is sampling: creating a smaller, representative dataset that is well distributed relative to the larger dataset (depending on the type of distribution you are expecting to achieve). Data-oriented subsamples are used in development, to validate machine-

learning algorithms (e.g., cross-validation) or to produce other statistical computations (e.g., power).

Generically we might implement filtering as a function that takes a single record as input. If the evaluation returns true, the record is emitted; otherwise, it is dropped. In this section, we explore sortless n-largest/smallest, sampling techniques, as well as more advanced filtering using Bloom filters to improve performance.

**Top n records:**

The top n records (and conversely the bottom n records) methodology is a cardinality comparison filter that requires both a mapper and reducer to work. The basic principle is to have each mapper yield its top n items, and then the reducer will similarly choose the top n items from the mappers. If n is relatively small (at least in comparison to the rest of the dataset) a single reducer should be able to handle this computation with ease because at most n records will come from each mapper:

```
import bisect

class TopNMapper(Mapper):

    def __init__(self, n, *args, **kwargs):
        self.n = n
        super(TopNMapper, self).__init__(*args, **kwargs)

    def map(self):
        items = []
        for value in self:
            # maintain sorted list of items
            bisect.insort(items, value)

        for item in items[-self.n:]:
            # emit the top n values from the mapper
            self.emit(None, item)


class TopNReducer(object):

    def __init__(self, n, *args, **kwargs):
        self.n = n
        super(TopNReducer, self).__init__(*args, **kwargs)

    def reduce(self):
        items = []
        for _, values in self:
            for value in values:
                bisect.insort(items, value)

        for item in items[-self.n:]:
            # emit the top n values from the mapper
            self.emit(None, item)
```

The primary benefit of this methodology is that a complete sort does not have to occur over the entire dataset. Instead, the mappers each

sort their own subset of the data, and the reducer sees only n times the number of mappers worth of data.

**Simple random sample:**

Simple random samples are subsets of a dataset where each record is equally likely to belong to the subset. In this case, the evaluation function does not care about the content or structure of the record, but instead utilizes some random number generator to evaluate whether to emit the record. The question is how to ensure that every element has an equal likelihood of being selected.

A first approach if we don't exactly need a specific sample of size n but rather some percentage of records is to simply use a random number generator to produce a number and compare it to the desired threshold size. Generally speaking, random number generators return a value between 0 and 1—so direct comparison to a percentage will yield the intended result! For example, if we want to sample 20% of our dataset, we might write a mapper as follows:

```
import random

class PercentSampleMapper(Mapper):

    def __init__(self, *args, **kwargs):
        self.percentage = kwargs.pop("percentage")
        super(PercentSampleMapper, self).__init__(*args, **kwargs)

    def map(self):
        for _, val in self:
            if random.random() < self.percentage:
                self.emit(None, val)

if __name__ == '__main__':
    mapper = PercentSampleMapper(sys.stdin, percentage=0.20)
    mapper.map()
```

**Bloom filtering:**

A *bloom filter* is an efficient probabilistic data structure used to perform set membership testing. A bloom filter is really no different from any other evaluation function, except that a preliminary computation must be made to gather "hot values", which we would like to filter against. The benefit is that a bloom filter is compact and fast to test membership.

Bloom filters suffer, however, from false positives—in other words, saying something belongs to the set when it does not; however, they guarantee that any exclusion does not belong in the membership set—there are no false negatives). If you're willing to have some fuzziness, most bloom filters can be constructed with a threshold for the probability of a false negative, by increasing or decreasing the size of the bloom filter.

In order to construct a bloom filter, you will first have to build it. Bloom filters work by applying several hashes to input data, then by setting bits in a bit array according to the hash. Once the bit array is constructed, it can be used to test membership by applying hashes to the test data and seeing if the relevant bits are 1 or not. The bit array construction can either be parallelized by using rules to map distinct values to a reducer that constructs the bloom filter, or it can be a living, versioned data structure that is maintained by other processes.

In this example, we will use a third-party library, pybloomfiltermmap, which can be installed using pip. Let's consider an example in which we are including tweets based on whether they contain a hashtag or @ reply that is in a whitelist of terms and usernames. In order to create the bloom filter, we load our data from disk, and save the bloom filter to mmap file as follows:

```
from pybloomfilter import BloomFilter

bloom = BloomFilter(1000000, 0.1, 'twitter.bloom')

for prefix, path in (('#', 'hashtags.txt'), ('@', 'handles.txt')):
    with open(path, 'r') as f:
        for word in f:
            bloom.add(prefix + word.strip())
```

After reading our hashtags and Twitter handles from files on, our bloom filter will be written to disk in a file called *twitter.bloom*.
To employ this in a Spark context:

```
ELEMS = re.compile(r'[#@][\w\d]+')

def tweet_filter(tweet, bloom=None):
    for elem in ELEMS.findall(tweet['text']):
        if elem in bloom.value:
            return True

# Load the bloom filter from disk and parallelize it
bloom = sc.broadcast(BloomFilter.open('twitter.bloom'))

# Load JSON tweets from disk and parse them
tweets = sc.textFile('tweets').map(json.loads)
tweets = tweets.filter(partial(tweet_filter, bloom=bloom))
```

Bloom filters are potentially the most complex data structure that you will use on a regular basis performing analytics in Hadoop.

## 9.3 TOWARD LAST-MILE ANALYTICS

Many machine learning techniques use generalized linear models (GLM) under the hood to estimate a response variable given some input

data and an error distribution. The most commonly used GLM is a linear regression (others include logistic and Poisson regressions), which models the continuous relationship between a dependent variable *Y* and one or more independent variables, X. That relationship is encoded by a set of *coefficients* and an *error term* as follows:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n + \epsilon$$

We can state that the computation of the  coefficients is the primary goal of fitting the model to existing data. This is generally done via an optimization algorithm that finds the set of coefficients that minimizes the amount of error given some dataset with observations for *X* and Y. Note that linear regression can be considered a *supervised* machine learning method, as the "correct" answers are known in advance.

Optimization algorithms like ordinary least squares or stochastic gradient descent are iterative; that is, they make multiple passes over the data. In a big data context, reading a complete dataset multiple times for each optimization iteration can be prohibitively time consuming, particularly for on-demand analytics or development. Spark makes things a bit better with distributed machine learning algorithms and inmemory computing exposed in its MLlib. However, for extremely large datasets, or smaller time windows, even Spark can take too long; and if Spark doesn't have the model or distributed algorithm you'd like to implement, then the many gotchas of distributed programming could limit your analytical choices.

The general solution is the one: decompose your problem by transforming the input dataset into a smaller one, until it fits in memory. Once the dataset is reduced to an in-memory computation, it can be analyzed using standard techniques, then validated across the entire dataset. For a linear regression, we could take a simple random sample of the dataset, perform feature extraction on the sample, build our linear model, then validate the model by computing the mean square error of the entire dataset.



### 9.3.1 Fitting a Model:

Consider a specific example where we have a dataset that originates from news articles or blog posts and a prediction task where we want to determine the number of comments in the next 24 hours. Given the raw HTML pages from a web crawl, the data flow may be as follows:

1. Parse HTML page for metadata and separate the main text and the comments.
2. Create an index of comments/commenters to blog post associated with a timestamp.
3. Use the index to create instances for our model, where an instance is a blog post and the comments in a 24-hour sliding window.

4. Join the instances with the primary text data (for both comments and blog test).
5. Extract the features of each instance (e.g., number of comments in the first 24 hours, the length of the blog post, bag of words features, day of week, etc.).
6. Sample the instance features.
7. Build a linear model in memory using Scikit-Learn or Statsmodels.
8. Compute the mean squared error or coefficient of determination across the entire dataset of instance features.

At this point, let's assume that through techniques we've already learned we've managed to arrive at a dataset that has all features extracted. Using the sampling technique, we can take a smaller dataset and save it to disk, and build a linear model with Scikit-Learn:

```
        import pickle
        import numpy as np
        from sklearn import linear_model

        # Load data from tab-delimited file on disk
        data = np.loadtxt('sample.txt')

        # The target is the first column (the key) and X is the value
        y = data[:,0]
        X = data[:,1:]

        # Instantiate and fit our linear model
        clf = linear_model.Ridge(alpha=1.0, fit_intercept=True)
        clf.fit(X, y)
# Write the model as a pickle to disk
with open('clf.pickle', 'wb') as f:
    pickle.dump(clf, f)
```

This snippet of code uses the np.loadtxt function to load our sample data from disk, which in this case must be a tab-delimited file of instances where the first column is the target value and the remaining columns are the features. This type of output matches what might happen when key/value pairs are written to disk from Spark or MapReduce, although you will have to collect the data from the cluster into a single file, and ensure it is correctly formatted.

### 9.3.2 Validating Models:

In order to use this model in the cluster to evaluate our performance, we have two choices. First, we could write the Scikit-Learn linear model properties, clf.coef_ (coefficients) and clf.intercept_ (error term) to disk and then load those parameters into our MapReduce or Spark job and compute the error ourselves. However, this requires us to implement a prediction function for every single model we may want to use. Instead, we will use the pickle module to dump the model to disk,

then load it to every node in the cluster to make our prediction.

In order to validate our model, we must compute the mean square error (MSE) across the entire dataset. Error is defined as the difference between the actual and predicted values.

```
import pickle

class MSEMapper(Mapper):

    def __init__(self, model, *args, **kwargs):
        super(MSEMapper, self).__init__(*args, **kwargs)

        # Load our model from disk
        with open(model, 'rb') as f:
            self.clf = pickle.load(f)


    def map(self):
        for row in self:
            # Parse the floating-point values in the row
            row = map(float, row)
            y = row[0]
            X = row[1:]

            yhat = self.clf.predict(x)

            self.emit(_, (y-yhat) ** 2)
```

In Spark, we can use an accumulator to sum the square error, and broadc model across the cluster as follows:

```
def cost(row, clf=None):
    """
    Computes the square error given the row.
    """
    return (row[0] - clf.predict(row[1:])) ** 2

def main(sc):
    """
    Primary analysis mechanism for Spark application
    """

    # Load the model from the pickle file
    with open('clf.pickle', 'rb') as f:
```

```
    clf = sc.broadcast(model.load(f))


# Create an accumulator to sum the squared error
sum_square_error = sc.accumulator(0)

# Load and parse the blog data
blogs = sc.textFile("blogData").map(float)

# Map the cost function and accumulate the sum
error = blogs.map(partial(cost, clf=clf))
error.foreach(lambda cost: sum_square_error.add(cost))

# Print and compute the mean
print sum_square_error.value / error.count()
```

## SUMMARY

This chapter we learn the concept of Distributed Analysis and Patterns using MapReduce, Spark and YARN concept for big data analytics. How to Computing with Keys, Compound Keys using final consideration when using compound keys (and complex values) is to understand serialization and deserialization of the compound data and different design pattern,. Concept of Summarization, Aggregation, Indexing, Inverted index, etc.. used in MapReduce.

## REVIEW QUESTIONS

1. What are the Distributed Analysis and Patterns.
2. What are the Computing with Keys.
3. Write a short note on Compound Keys
4. Explain Compound data serialization
5. What are the identity pattern..
6. Write a short note on Pairs versus Stripes
7. Explain different Design Patterns.
8. What is mean by Summarization.

## REFERENCES


- Data Science & Big Data Analytics Discovering, Analyzing, Visualizing and Presenting Data EMC Education Services Published by John Wiley & Sons, Inc


*****

# 10

# DATA MINING AND WAREHOUSING

**Unit Structure**



## 10.0 OBJECTIVES

To Study and Understand the following concept

- Data Mining and Warehousing
- Structured Data Queries with Hive
- HBase
- Data Ingestion
- Importing Relational data with Sqoop
- Injesting stream data with flume.
- Analytics with higher level APIs
- Pig
- Spark's higher level APIs.



## 10.1 DATA MINING AND WAREHOUSING

It's estimated that ETL consumes 70–80% of data warehousing costs, risks, and implementation time. This overhead makes it costly to perform even modest levels of data analysis prototyping or exploratory analysis. RDBMSs present another limitation in the face of the rapidly expanding diversity of data types that we need to store and analyze, which can be unstructured (emails, multimedia files) or semi-structured (clickstream data) in nature. The velocity and variety of this data often demands the ability to evolve the schema in a "just-in-time" manner, which is very tough to support in a traditional DW.

It's for these reasons that Hadoop has become the most disruptive technology in the data warehousing and data mining space. Hadoop's separation of storage from processing enables organizations to store their raw data in HDFS without necessitating ETLs to conform the data into a single unified data model. Moreover, with YARN's generalized processing layer, we're able to directly access and query the raw data from multiple perspectives and using different methods (SQL, non-SQL) as appropriate for the particular use case. Hadoop thus not only enables exploratory analysis and data mining prototyping, it opens the floodgates to new types of data and analysis.

## 10.2 STRUCTURED DATA QUERIES WITH HIVE

Apache Hive is a "data warehousing" framework built on top of Hadoop. Hive provides data analysts with a familiar SQL-based interface to Hadoop, which allows them to attach structured schemas to data in HDFS and access and analyze that data using SQL queries. Hive has made it possible for developers who are fluent in SQL to leverage the scalability and resilience of Hadoop without requiring them to learn Java or the native MapReduce API.

Hive provides its own dialect of SQL called the Hive Query Language, or HQL. HQL supports many commonly used SQL statements, including data definition statements (DDLs), data manipulation statements (DMSs), and data retrieval queries. Hive also supports integration of custom user-defined functions, which can be written in Java or any language supported by Hadoop Streaming, that extend the built-in functionality of HQL.

Hive commands and HQL queries are compiled into an *execution plan* or a series of HDFS operations and/or MapReduce jobs, which are then executed on a Hadoop cluster. Thus, Hive has inherited certain limitations from HDFS and MapReduce that constrain it from providing key online transaction processing (OLTP) features that one might expect from a traditional database management system. In particular, because HDFS is a write-once, read-many (WORM) file system and does not provide in-place file updates, Hive is not very efficient for performing row-level inserts, updates, or deletes.

Additionally, Hive queries entail higher-latency due to the overhead required to generate and launch the compiled MapReduce jobs on the cluster; even small queries that would complete within a few seconds on a traditional RDBMS may take several minutes to finish in Hive.

On the plus side, Hive provides the high-scalability and high-throughput that you would expect from any Hadoop-based application, and as a result, is very well suited to batch-level workloads for online analytical processing (OLAP) of very large datasets at the terabyte and petabyte scale.

### 10.2.1 The Hive Command-Line Interface (CLI):

Hive's installation comes packaged with a handy command-line interface (CLI), which we will use to interact with Hive and run our HQL statements. To start the Hive CLI from the *$HIVE_HOME*:

```
~$ cd $HIVE_HOME
 /srv/hive$ bin/hive
```

This will initiate the CLI and bootstrap the logger and Hive history file, and finally display a Hive CLI prompt:

```
hive>
```

At any time, you can exit the Hive CLI using the following command:

```
hive> exit;
```

Hive can also run in non-interactive mode directly from the command line by passing the filename option, -f, followed by the path to the script to execute:

```
~$ hive -f ~/hadoop-fundamentats/hive/init.hqt
~$ hive -f ~/hadoop-fundamentats/hive/top_50_ptayers_by_homeruns.hqt
>> ~/homeruns.tsv
```

Additionally, the quoted-query-string option, -e, allows you to run inline commands from the command line:

```
~$ hive -e 'SHOW DATABASES;'
```

You can view the full list of Hive options for the CLI by using the -H flag:

```
~$ hive -H

usage: hive
 -d,--define <key=value>          Variable substitution to apply to hive
                                      commands. e.g. -d A=B or --define A=B
```

…..

### 10.1.2 Hive Query Language (HQL):

#### Creating a database

Creating a database in Hive is very similar to creating a database in a SQL-based RDBMS, by using the `CREATE DATABASE` or `CREATE SCHEMA` statement:

```
hive> CREATE DATABASE log_data;
```

When Hive creates a new database, the schema definition data is stored in the *Hive metastore*. Hive will raise an error if the database already exists in the metastore; we can check for the existence of the database by using `IF NOT EXISTS`:

```
hive> CREATE DATABASE IF NOT EXISTS log_data;
```

We can then run `SHOW DATABASES` to verify that our database has been created. Hive will return all databases found in the metastore, along with the default Hive database:

```
hive> SHOW DATABASES;
OK
default
log_data
Time taken: 0.085 seconds, Fetched: 2 row(s)
```

Additionally, we can set our working database with the `USE` command:

```
hive> USE log_data;
```

Now that we've created a database in Hive, we can describe the layout of our data by creating table definitions within that database.

#### Creating tables

Hive provides a SQL-like `CREATE TABLE` statement, which in its simplest form takes a table name and column definitions:

```
CREATE TABLE apache_log (
    host STRING,
    identity STRING,
   user STRING,
   time STRING,
   request STRING,
   status STRING,
   size STRING,
   referer STRING,
   agent STRING
);
```



However, because Hive data is stored in the file system, usually in HDFS or the local file system, the **CREATE** TABLE command also takes optional clauses to specify the row format with the **ROW FORMAT** clause that tells Hive how to read each row in the file and map to our columns. For example, we could indicate that the data is in a delimited file with fields delimited by the tab character:

```
hive> CREATE TABLE shakespeare (
    lineno STRING,
    linetext STRING
)
ROW FORMAT DELIMITED
    FIELDS TERMINATED BY '\t';
```

Fortunately, Hive provides a way for us to apply a regex to known record formats to *deserialize* or parse each row into its constituent fields. We'll use the Hive serializer-deserializer row format option, SERDE, and the contributed RegexSerDe library to specify a regex with which to

deserialize and map the fields into columns for our table. We'll need to manually add the *hive-serde* JAR from the *lib* folder to the current hive session in order to use the RegexSerDe package:

```
hive> ADD JAR /srv/hive/tib/hive-serde-0.13.1.jar;
```

And now let's drop the apache_tog table that we created previously, and re-create it to use our custom serializer:

```
hive> DROP TABLE apache_log;

hive> CREATE TABLE apache_log (
    host STRING,
    identity STRING,
    user STRING,
    time STRING,
    request STRING,
    status STRING,
    size STRING,
    referer STRING,
    agent STRING
)
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.RegexSerDe'
WITH SERDEPROPERTIES ("input.regex" = "([^ ]*) ([^ ]*) ([^ ]*) (-|\\[[^\\]]
*\\])([^ \"]*|\"[^\"]*\") (-|[0-9]*) (-|[0-9]*)(?: ([^ \"]*|\".*\") ([^ \"]

   *|\".*\"))?", "output.format.string" = "%1$s %2$s %3$s %4$s %5$s %6$s %7$s
   %8$s %9$s")
   STORED AS TEXTFILE;
```

Once we've created the table, we can use DESCRIBE to verify our table definition:

```
hive> DESCRIBE apache_log;
OK
host                    string                  from deserializer
identity                string                  from deserializer
user                    string                  from deserializer
time                    string                  from deserializer
request                 string                  from deserializer
status                  string                  from deserializer
size                    string                  from deserializer
referrer                string                  from deserializer
agent                   string                  from deserializer
Time taken: 0.553 seconds, Fetched: 9 row(s)
```

| Type | Description | Example |
|---|---|---|
| TINYINT | 8-bit signed integer, from -128 to 127 | 127 |
| SMALLINT | 16-bit signed integer, from -32,768 to 32,767 | 32,767 |
| INT | 32-bit signed integer | 2,147,483,647 |
| BIGINT | 64-bit signed integer | 9,223,372,036,854,775,807 |
| FLOAT | 32-bit single-precision float | 1.99 |
| DOUBLE | 64-bit double-precision float | 3.14159265359 |
| BOOLEAN | True/false | true |
| STRING | 2 GB max character string | hello world |
| TIMESTAMP | Nanosecond precision | 1400561325 |

**Table 10.1 Hive primitive data type**

| Type | Description | Example |
|---|---|---|
| ARRAY | Ordered collection of elements. The elements in the array must be of the same type. | `recipients ARRAY<email:STRING>` |
| MAP | Unordered collection of key/value pairs. Keys must be of primitive types and values can be of any type. | `files MAP<filename:STRING, size:INT>` |
| STRUCT | Collection of elements of any type. | `address STRUCT<street:STRING, city:STRING, state:STRING, zip:INT>` |

**Table 10.2 Hive complex data type**

**Loading data:**

Hive does not perform any verification of the data for compliance with the table schema, nor does it perform any transformations when loading the data into a table. Data loading in Hive is done in batch-oriented fashion using a bulk LOAD DATA command or by inserting results from another query with the INSERT command. To start,

let's copy our Apache log data file to HDFS and then load it into the table we created earlier:

```
~$ hadoop fs -mkdir statistics
~$ hadoop fs -mkdir statistics/log_data
~$ hadoop fs -copyFromLocal ~/hadoop-fundamentals/data/log_data/apache.log \
    statistics/log_data/
```

You can verify that the *apache.log* file was successfully uploaded to HDFS with the tail command:

~$ hadoop fs –tail statistics/log_data/apache.log

Once the file has been uploaded to HDFS, return to the Hive CLI and use the log_data database:

```
~$ $HIVE_HOME/bin/hive

hive> use log_data;
OK
Time taken: 0.221 seconds
```

We'll use the LOAD DATA command and specify the HDFS path to the logfile, writing the contents into the apache_log table:

```
hive> LOAD DATA INPATH 'statistics/log-data/apache.log'
OVERWRITE INTO TABLE apache_log;

Loading data to table log_data.apache_log
rmr: DEPRECATED: Please use 'rm -r' instead.
Deleted hdfs://localhost:9000/user/hive/warehouse/log_data.db/apache_log
Table log_data.apache_log stats: [numFiles=1, numRows=0, totalSize=52276758,
rawDataSize=0]
OK
Time taken: 0.902 seconds
```

INPATH takes an argument to a path on the default file system (in this case, HDFS). We can also specify a path on the local file system by using LOCAL INPATH instead.

### 10.1.3 Data Analysis with Hive:

SQL GROUP BY query can be used to computes the number of hits per calendar month:

```
hive> SELECT
        month,
        count(1) AS count
      FROM (SELECT split(time, '/')[1] AS month FROM apache_log) l
      GROUP BY month
      ORDER BY count DESC;
OK
Mar 99717
Sep 89083
Feb 72088
Aug 66058
Apr 64984
May 63753
Jul 54920
Jun 53682
Oct 45892
Jan 43635
Nov 41235
Dec 29789
NULL    1903
Time taken: 84.77 seconds, Fetched: 13 row(s)
```

we can easily perform other ad hoc queries on any of the other fields for example:
hive> SELECT host, count(1) AS count FROM apache_log GROUP BY host ORDER BY count;

In addition to count, Hive also supports other aggregate functions to compute the sum, average, min, max as well as statistical aggregations for variance, standard deviation, and covariance of numeric columns.

When using these built-in aggregate functions, you can improve the performance of the aggregation query by setting the following property to true:
hive> SET hive.map.aggr = true;

We can create new tables to store the results returned by these queries for later record-keeping and analysis:

```
hive> CREATE TABLE
        remote_hits_by_month
        AS
        SELECT
            month,
            count(1) AS count
        FROM (
            SELECT split(time, '/')[1] AS month
            FROM apache_log
            WHERE host == 'remote'
            ) l
        GROUP BY month
        ORDER BY count DESC;
```

**Aggregations and joins:**

Consider the US flight data *ontime_flights.tsv*. Each row of the on-time flight data in *ontime_flights.tsv* includes an integer value that represents the code for AIRLINE_ID (such as 19805) and a string value that represents the code for CARRIER (such as "AA"). AIRLINE_ID codes can be joined with the corresponding code in the *airlines.tsv* file in which each row contains the code and corresponding description:
19805 American Airlines Inc.: AA

Accordingly, CARRIER codes can be joined with the corresponding code in *carriers.tsv*, which contains the code and corresponding airline name and effective dates:
AA American Airlines Inc. (1960 - )
Assuming that we've uploaded our data files to HDFS or local file system and created required tables with data.

To get a list of airlines and their respective average departure delays, we can simply perform a SQL JOIN on flights and airlines on the airline code and then use the aggregate function AVG() to compute the average depart_delay grouped by the airline description:

```
hive> SELECT
        a.description,
        AVG(f.depart_delay)
      FROM airlines a
      JOIN flights f ON a.code = f.airline_code
      GROUP BY a.description;
```

## 10.2 HBase

We know that while Hive provides a familiar data manipulation paradigm within Hadoop, it doesn't change the storage and processing paradigm, which still utilizes HDFS and MapReduce in a batch-oriented fashion. HBase is part of the Hadoop ecosystem which offers random real-time read/write access to data in the Hadoop File System.

### 10.2.1 NoSQL and Column-Oriented Databases:

NoSQL is a broad term that generally refers to non-relational databases and encompasses a wide collection of data storage models, including graph databases, document databases, key/value data stores and column family databases. HBase is classified as a column-family or column-oriented database, modeled on Google's BigTable architecture. This architecture allows HBase to provide:

- Random (row-level) read/write access
- Strong consistency
- "Schema-less" or flexible data modeling

The schema-less trait is a result of how HBase approaches data modeling, which is very different from how relational databases approach data modeling. HBase organizes data into *tables* that contain *rows*. Within a table, rows are identified by their unique *row key*, which do not have a data type and are instead stored and treated as a *byte array*. Row keys are similar to the concept of primary keys in relational databases, in that they are automatically indexed; in HBase, table rows are sorted by their row key and because row keys are byte arrays, almost anything can serve as a row key from strings to binary representations of longs or even serialized data structures.

HBase stores its data as key/value pairs, where all table lookups are performed via the table's row key, or unique identifier to the stored record data. Data within a row is grouped into *column families,* which consist of related columns. Visually, you can picture an HBase table that holds census data for a given population where each row represents a person and is accessed via a unique ID rowkey, with column families for personal data which contains columns for name and address, and demographic info which contains columns for birthdate and gender. This example is shown in Figure 10.3.

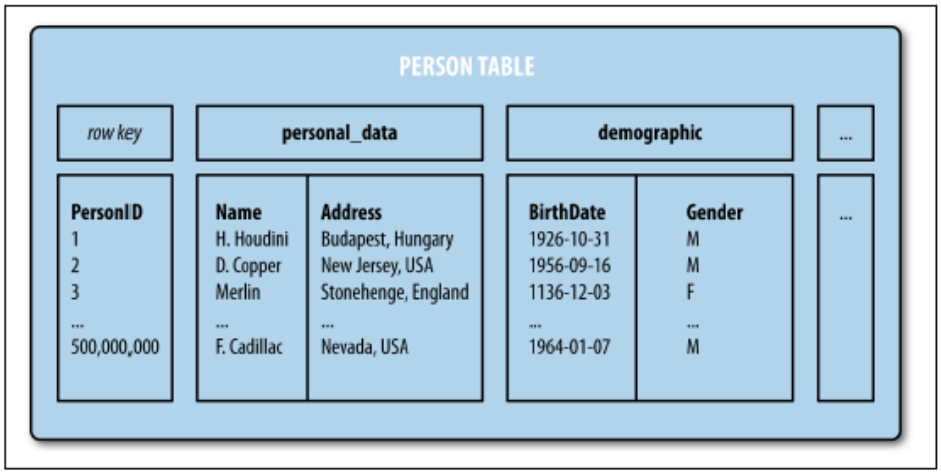
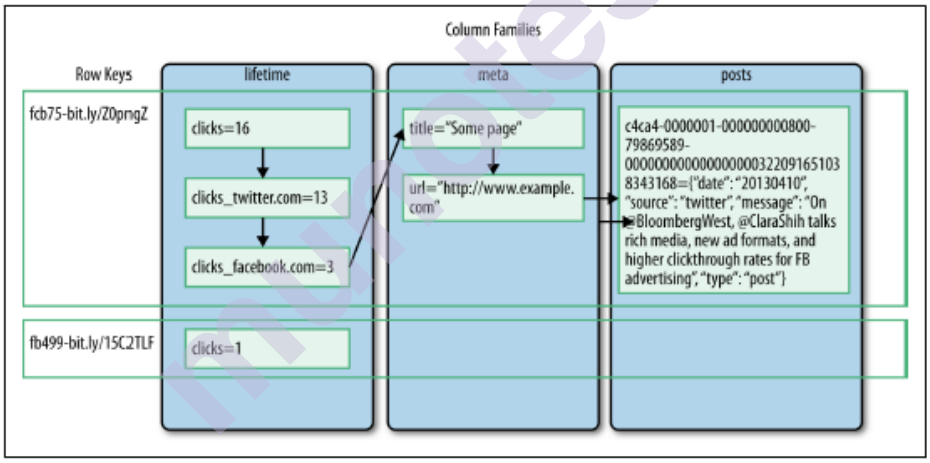


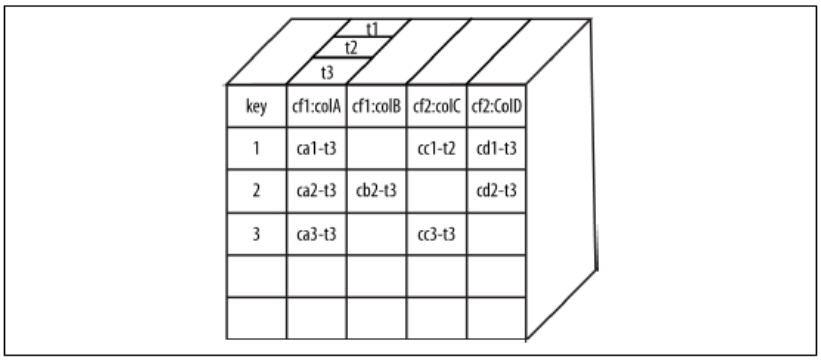

**Figure 10.3 Census data as an HBase schema**

Storing data in columns rather than rows has particular benefits for data warehouses and analytical databases where aggregates are computed over large sets of data with potentially sparse values, where not all columns values are present. However, the actual columns that make up a row can be determined and created on an as-needed basis. In fact, each row can have a different set of columns. Figure 6-2 shows an example HBase table with two rows where first row key utilizes three column families and the second row key utilizes just one column.




**10.4 Social media events with sparse columns**

Another interesting feature of HBase and BigTable-based column-oriented databases is that the table cells, or the intersection of row and column coordinates, are versioned by timestamp, stored as a long integer representing milliseconds since January 1, 1970 UTC. HBase is thus also described as being a multidimensional map where time provides the third dimension, as shown in Figure 6-3. The time dimension is indexed in decreasing order, so that when reading from an HBase store, the most recent values are found first. The contents of a cell can be referenced by a *{rowkey, column, timestamp}* tuple, or we can scan for a range of cell values by time range.




**10.5 HBase timestamp versioning**

### 10.2.2 Real-Time Analytics with HBase:

HBase schemas can be created or updated with the HBase Shell or with the Java API, using the HBaseAdmin interface class. Additionally, HBase supports a number of other clients that can be used to support non-Java programming languages, including a REST API interface, Thrift, and Avro. These clients act as proxies that wrap the native Java API.

### Generating a schema:

When designing schemas in HBase, it's important to think in terms of the column- family structure of the data model and how it affects data access patterns. Furthermore, because HBase doesn't support joins and provides only a single indexed rowkey, we must be careful to ensure that the schema can fully support all use cases. Often this involves de-normalization and data duplication with nested entities.

But HBase allows dynamic column definition at runtime, we have quite a bit of flexibility even after table creation to modify and scale our schema.

### Namespaces, tables, and column families:

First, we need to declare the table name, and at least one column-family name at the time of table definition. We can also declare our own optional namespace to serve as a logical grouping of tables, analogous to a database in relational database systems. If no namespace is declared, HBase will use the default namespace.

```
hbase> create 'linkshare', 'link'
0 row(s) in 1.5740 seconds
```

We just created a single table called `linkshare` in the default namespace with one column-family, named `link`. To alter the table after creation, such as changing or adding column families, we need to first disable the table so that clients will not be able to access the table during the `alter` operation:

```
hbase> disable 'linkshare'
0 row(s) in 1.1340 seconds

hbase> alter 'linkshare', 'statistics'
Updating all regions with the new schema...
1/1 regions updated.
Done.
0 row(s) in 1.1630 seconds
```

We can then re-enable the table using the `enable` command:

```
hbase> enable 'linkshare'
0 row(s) in 1.1930 seconds
```

And then use the `describe` command to verify that the table contains the two expected column families with the default configurations:

```
hbase> describe 'linkshare'

Table linkshare is ENABLED
COLUMN FAMILIES DESCRIPTION
{NAME => 'link', DATA_BLOCK_ENCODING => 'NONE', BLOOMFILTER => 'ROW',
REPLICATION_SCOPE => '0', COMPRESSION => 'NONE', VERSIONS => '1',
TTL => 'FOREVER', MIN_VERSIONS => '0', KEEP_DELETED_CELLS => 'FALSE',
BLOCKSIZE => '65536', IN_MEMORY => 'false', BLOCKCACHE => 'true'}
{NAME => 'statistics', DATA_BLOCK_ENCODING => 'NONE',
BLOOMFILTER => 'ROW', REPLICATION_SCOPE => '0', VERSIONS => '1',
COMPRESSION => 'NONE', MIN_VERSIONS => '0', TTL => 'FOREVER',
KEEP_DELETED_CELLS => 'FALSE', BLOCKSIZE => '65536',
IN_MEMORY => 'false', BLOCKCACHE => 'true'}
2 row(s) in 0.1290 seconds
```

**Row keys:**

Before we insert row data, we need to determine how to design our row key. By default, HBase stores rows in sorted order by row key, so that similar keys are stored to the same RegionServer. While this enables faster range scans, it could also lead to uneven load on particular servers during read/write operations. For the current example, let's assume that we will use the unique reversed link URL for the row key.

## Inserting data with put:

We can insert, or `put`, a value in a cell at the specified table/row/column and optionally timestamp coordinates. To put a cell value into table `linkshare` at row with row key `org.hbase.www` under column-family link and column title marked with the current timestamp, do:

```
hbase> put 'linkshare', 'org.hbase.www', 'link:title', 'Apache HBase'
hbase> put 'linkshare', 'org.hadoop.www', 'link:title', 'Apache Hadoop'
hbase> put 'linkshare', 'com.oreilly.www', 'link:title', 'O\'Reilly.com'
```

The `put` operation works great for inserting a value for a single cell, but for incrementing frequency counters, HBase provides a special mechanism to treat columns as counters. Otherwise, under heavy load, we could face significant contention for these rows as we would need to lock the row, read the value, increment it, write it back, and finally unlock the row for other writers to be able to access the cell.[9]

To increment a counter, we use the command `incr` instead of `put`:

```
hbase> incr 'linkshare', 'org.hbase.www', 'statistics:share', 1
(COUNTER VALUE is now 1)

hbase> incr 'linkshare', 'org.hbase.www', 'statistics:like', 1
(COUNTER VALUE is now 1)
```

The last option passed is the increment value, which in this case is 1. Incrementing a counter will return the updated counter value, but you can also access a counter's current value any time using the `get_counter` command, specifying the table name, row key, and column:

```
hbase> incr 'linkshare', 'org.hbase.www', 'statistics:share', 1
(COUNTER VALUE is now 2)

hbase> get_counter 'linkshare', 'org.hbase.www', 'statistics:share', 'dummy'
COUNTER VALUE = 2
```

The `get_counter` method is used to decode the byte-array value of the counter and return the integer value. Unfortunately, the latest HBase build includes a bug in the shell command for getting the counter value, which expects a fourth argument that is never used. As a result, we'll need to pass in a fourth dummy value:

```
hbase> get_counter 'linkshare', 'org.hbase.www', 'statistics:share', 'dummy'
COUNTER VALUE = 2
```

HBase provides two general methods to retrieve data from a table: the `get` command performs lookups by row key to retrieve attributes for a specific row, and the `scan` command, which takes a set of filter specifications and iterates over multiple rows based on the indicated specifications.

### Get row or cell values

In its simplest form, the `get` command accepts the table name followed by the row key, and returns the most recent version timestamp and cell value for all columns in the row:

```
hbase> get 'linkshare', 'org.hbase.www'

COLUMN                          CELL
 link:title                     timestamp=1422145743298, value=Apache HBase
 statistics:like                timestamp=1422153344211,
     value=\x00\x00\x00\x00\x00\x00\x00\x1F
 statistics:share               timestamp=1422153337498,
     value=\x00\x00\x00\x00\x00\x00\x00\x02
3 row(s) in 0.0310 seconds
```

Note that the `statistics:share` column returns the value in its byte array representation, printing the separate bytes as hexadecimal values. To display the integer representation of the counter value, use the `get_counter` command mentioned in the previous section.

The `get` command also accepts an optional dictionary of parameters to specify the column(s), timestamp, timerange, and version of the cell values we want to retrieve. For example, we can specify the column(s) of interest:

```
hbase> get 'linkshare', 'org.hbase.www', {COLUMN => 'link:title'}
hbase> get 'linkshare', 'org.hbase.www', {COLUMN => ['link:title',
    'statistics:share']}
```

There is also a shortcut to specify column parameters in a `get` by just appending the comma-delimited column names after the row key:

```
hbase> get 'linkshare', 'org.hbase.www', 'link:title'
hbase> get 'linkshare', 'org.hbase.www', 'link:title', 'statistics:share'
hbase> get 'linkshare', 'org.hbase.www', ['link:title', 'statistics:share']
```

To specify a time range of values we are interested in, we pass in a `TIMERANGE` parameter with start and end timestamps in milliseconds:

```
hbase> get 'linkshare', 'org.hbase.www', {TIMERANGE => [1399887705673,
    1400133976734]}
```



## Scan rows:

Using `scan` is similar to using the `get` command; it also accepts `COLUMN`, `TIMESTAMP`, `TIMERANGE`, and `FILTER` parameters. However, instead of specifying a single row key, you can specify an optional `STARTROW` and/or `STOPROW` parameter, which can be used to limit the scan to a specific range of rows. If neither `STARTROW` nor `STOPROW` are provided, the `scan` operation will scan through the entire table.

You can, in fact, call `scan` with the table name to display all the contents of a table:

```
hbase> scan 'linkshare'

ROW                 COLUMN+CELL
com.oreilly.www     column=link:title, timestamp=1422153270279,
value=O'Reilly.com
org.hadoop.www      column=link:title, timestamp=1422153262507,
value=Apache Hadoop
org.hbase.www       column=link:title, timestamp=1422145743298,
```

```
value=Apache HBase
org.hbase.www       column=statistics:like, timestamp=1422153344211,
    value=\x00\x00\x00\x00\x00\x00\x00\x1F
org.hbase.www       column=statistics:share, timestamp=1422153337498,
    value=\x00\x00\x00\x00\x00\x00\x00\x02
3 row(s) in 0.0290 seconds
```

Keep in mind that the rows in HBase are stored in lexicographical order.[10] For example, numbers going from 1 to 100 will be ordered like this:

```
1,10,100,11,12,13,14,15,16,17,18,19,2,20,21,...,9,91,92,93,94,95,96,97,98,99
```

Let's retrieve the `link:title` column but limit our scan to the rows starting with row key `org.hbase.www`:

```
hbase> scan 'linkshare', {COLUMNS => ['link:title'], STARTROW => 'org.hbase.www'}

ROW                 COLUMN+CELL
 org.hbase.www      column=link:title, timestamp=1453184861236,
 value=Apache HBase
1 row(s) in 0.0250 seconds
```

**Filters:**

HBase provides a number of filter classes that can be applied to further filter the row data returned from a get or scan operation. These filters can provide a much more efficient means of limiting the row data returned by HBase and offloading the row-filtering operations from the client to the server. Some of HBase's available filters include:

- ***RowFilter:*** Used for data filtering based on row key values
- ***ColumnRangeFilter:*** Allows efficient intra-row scanning, can be used to get a slice of the columns of a very wide row.
- **SingleColumnValueFilter:** Used to filter cells based on column value
- ***RegexStringComparator:*** Used to test if a given regular expression matches a cell value in the column.

The HBase Java API provides a Filter interface and abstract FilterBase class plus a number of specialized Filter subclasses. Custom filters can also be created by subclassing the FilterBase abstract class and implementing the key abstract methods.

To begin, we need to import the necessary classes, including the org.apache.hadoop.hbase.util.Bytes to convert our column family, column, and values into bytes, and the filter and comparator classes:

```
hbase> import org.apache.hadoop.hbase.util.Bytes
hbase> import org.apache.hadoop.hbase.filter.SingleColumnValueFilter
hbase> import org.apache.hadoop.hbase.filter.BinaryComparator
hbase> import org.apache.hadoop.hbase.filter.CompareFilter
```

Next, we'll create a filter that limits the results to rows where the `statistics:like` counter column value is greater than or equal to 10:

```
hbase> likeFilter = SingleColumnValueFilter.new(Bytes.toBytes('statistics'),
    Bytes.toBytes('like'),
    CompareFilter::CompareOp.valueOf('GREATER_OR_EQUAL'),
    BinaryComparator.new(Bytes.toBytes(10)))
```

And because we don't have a value for this column for every row, we need to set a flag that tells this filter to skip any rows without a value in this column:

```
hbase> likeFilter.setFilterIfMissing(true)
```

At this point, we can run our `scan` operation with the filter we configured:

```
hbase> scan 'linkshare', { FILTER => likeFilter }

ROW                 COLUMN+CELL
org.hbase.www       column=link:title, timestamp=1422145743298,
    value=Apache HBase
org.hbase.www       column=statistics:like, timestamp=1422153344211,
    value=\x00\x00\x00\x00\x00\x00\x00\x1F
org.hbase.www       column=statistics:share, timestamp=1422153337498,
  value=\x00\x00\x00\x00\x00\x00\x00\x02
1 row(s) in 0.0470 seconds
```

This should return all rows that contain a column value for `statistics:like` that is greater than or equal to 10; this should include row key `com.oreilly.www` in this example.

## 10.3 DATA INGESTION

One of Hadoop's greatest strengths is that it's inherently schemaless and can work with any type or format of data regardless of structure (or lack of structure) from any source, as long as you implement Hadoop's Writable or DBWritable interfaces and write your MapReduce code to parse the data correctly. However, in cases where the input data is already structured because it resides in a relational database, it would be convenient to leverage this known schema to import the data into Hadoop in a more efficient manner than uploading CSVs to HDFS and parsing them manually.

Sqoop (SQL-to-Hadoop) is designed to transfer data between relational database management systems (RDBMS) and Hadoop. It automates most of the data transformation process, relying on the RDBMS to provide the schema description for the data to be imported.

While Sqoop works very well for bulk-loading data that already resides in a relational database into Hadoop, many new applications and systems involve fast-moving data streams like application logs, GPS tracking, social media updates, and sensor-data that we'd like to load directly into HDFS to process in Hadoop. In order to handle and process the high-throughput of event-based data produced by these systems, we need the ability to support continuous ingestion of data from multiple sources into Hadoop.

Apache Flume was designed to efficiently collect, aggregate, and move large amounts of log data from many different sources into a centralized data store. While Flume is most often used to direct streaming log data into Hadoop, usually HDFS or HBase, Flume data sources are actually quite flexible and can be customized to transport many types of event data, including network traffic data, social media-generated data, and sensor data into any Flume-compatible consumer.

### 10.3.1 Importing Relational Data with Sqoop:

Sqoop (SQL-to-Hadoop) is a relational database import and export tool created by Cloudera, and is now an Apache top-level project. Sqoop is designed to transfer data between a relational database like MySQL or Oracle, into a Hadoop data store, including HDFS, Hive, and HBase. It automates most of the data transfer process by reading the schema information directly from the RDBMS. Sqoop then uses MapReduce to import and export the data to and from Hadoop.

Sqoop gives us the flexibility to maintain our data in its production state while copying it into Hadoop to make it available for further analysis without modifying the production database.

### 10.3.2 Importing from MySQL to HDFS:

When importing data from relational databases like MySQL, Sqoop reads the source database to gather the necessary metadata for the data being imported. Sqoop then submits a map-only Hadoop job to transfer the actual table data based on the metadata that was captured in the previous step. This job produces a set of serialized files, which may be delimited text files, binary format (e.g., Avro), or Sequence Files containing a copy of the imported table or datasets. By default, the files are saved as comma-separated files to a directory on HDFS with a name that corresponds to the source table name.

Assuming that you have MySQL with a database called energydata and a table called average_price_by_state:

Before we proceed to run the sqoop import command, verify that HDFS and YARN are started with the jps command:

```
~$ sudo su hadoop
hadoop@ubuntu:~$ jps

4051 NodeManager
31134 Jps
3523 DataNode
3709 SecondaryNameNode
3375 NameNode
3921 ResourceManager
```

At this point, we can import the data in table `average_price_by_state` into HDFS by using the `import` command. We can specify the source database's connection string, username, and tablename with the `--connect` option, `--username` option, and `--table` option, respectively. We'll set the optional `-m` flag to 1 to indicate that this job should use a single map task:

```
/srv/sqoop$ sqoop import --connect jdbc:mysql://localhost:3306/energydata
    --username root --table average_price_by_state -m 1

15/01/20 22:47:35 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6
15/01/20 22:47:35 INFO manager.MySQLManager: Preparing to use a MySQL
    streaming resultset.
15/01/20 22:47:35 INFO tool.CodeGenTool: Beginning code generation
15/01/20 22:47:36 INFO manager.SqlManager: Executing SQL statement:
    SELECT t.* FROM `average_price_by_state` AS t LIMIT 1

(output truncated)

15/01/25 22:47:53 INFO mapreduce.ImportJobBase: Transferred 200.4287 KB in
    15.3718 seconds (13.0387 KB/sec)
15/01/25 22:47:53 INFO mapreduce.ImportJobBase: Retrieved 3272 records.
```

### 10.2.3 Importing from MySQL to Hive:

Sqoop provides a couple ways to do this, either exporting to HDFS first and then loading the data into Hive using the LOAD DATA HQL command in the Hive shell, or by using Sqoop to directly create the tables and load the relational database data into the corresponding tables in Hive.

Sqoop can generate a Hive table and load data based on the defined schema and table contents from a source database, using the import command. However, because Sqoop still actually utilizes MapReduce to implement the data load operation, we must first delete any preexisting data directory with the same output name before running the import tool:

```
/srv/sqoop$ hadoop fs -rm -r /user/hadoop/average_price_by_state
```

We can then run Sqoop's `import` command, passing it the JDBC connection string to the database, the table name, field delimiter, line terminator, and null string value:

```
/srv/sqoop$ sqoop import --connect jdbc:mysql://localhost:3306/energydata
    --username root --table average_price_by_state
    --hive-import --fields-terminated-by ','
    --lines-terminated-by '\n' --null-string 'null' -m 1

(output truncated)

15/01/20 00:14:37 INFO hive.HiveImport: Table default.average_price_by_state stats:
    [numFiles=2, numRows=0, totalSize=205239, rawDataSize=0]
15/01/20 00:14:37 INFO hive.HiveImport: OK
15/01/20 00:14:37 INFO hive.HiveImport: Time taken: 0.435 seconds
15/01/20 00:14:37 INFO hive.HiveImport: Hive import complete.
15/01/20 00:14:37 INFO hive.HiveImport: Export directory is empty, removing it.
```

In local mode, Hive will create a *metastore_db* directory within the file system location from which it was run; After above query *metastore_db* will be created under the *SQOOP_HOME* (*/srv/sqoop*). Open the Hive shell and verify that the table average_price_by_state was created:

```
/srv/sqoop$ hive

hive> DESC average_price_by_state;

OK
year                    int
state                   string
sector                  string
residential             double
commercial              double
industrial              double
transportation          double
other                   double
total                   double
Time taken: 0.858 seconds, Fetched: 9 row(s)
```

### 10.3.4 Importing from MySQL to HBase:

HBase is designed to handle large volumes of data for a large number of concurrent clients that need real-time access to row-level data. Sqoop's import tool allows us to import data from a relational database to HBase. As with Hive, there are two approaches to importing this data. We can import to HDFS first and then use the HBase CLI or API to load the data into an HBase table, or we can use the **--hbase-table** option to instruct Sqoop to directly import to a table in HBase.

In this example, the data that we want to offload to HBase is a table of weblog stats where each record contains a primary key composed of the pipe-delimited IP address and year, and a column for each month that contains the number of hits for that IP and year. You can find the CSV named *weblogs.csv* in the GitHub repo's */data* directory. Download this CSV and load it into a MySQL table. Consider we have table *weblogs* in **logadata** database in MQSQL.

Again, we need to verify that Hadoop is running, as well as HBase daemons:

```
~$ cd $HBASE_HOME
/srv/hbase$ bin/start-hbase.sh
```

We can then run Sqoop's `import` command, passing it the JDBC connection string to the database, the table name, HBase table name, column family name, and row key name:

```
sqoop import --connect jdbc:mysql://localhost:3306/logdata
    --table weblogs --hbase-table weblogs --column-family traffic
    --hbase-row-key ipyear --hbase-create-table -m 1

(output truncated)

15/01/20 00:33:01 INFO mapreduce.ImportJobBase: Transferred 0 bytes in
19.0716 seconds (0 bytes/sec)
15/01/20 00:33:01 INFO mapreduce.ImportJobBase: Retrieved 27300 records.
```

## 10.4 INGESTING STREAMING DATA WITH FLUME

Flume is designed to collect and ingest high volumes of data from multiple data streams into Hadoop. A very common use case for Flume is the collection of log data, such as collecting web server log data emitted from multiple application servers, and aggregating it in HDFS for later search or analysis. However, Flume isn't restricted to simply consuming and ingesting log data sources, but can also be customized to transport massive quantities of event data from any custom event source. In both cases, Flume enables us to incrementally and continuously ingest streaming data *as it is written* into Hadoop, rather than writing custom client applications to batch-load the data into HDFS, HBase, or other Hadoop data sink. Flume provides a unified yet flexible method of pushing data from many fast-moving, disparate data streams into Hadoop.

Flume's flexibility is derived from its inherently extensible data flow architecture. In addition to flexibility, Flume is designed to maintain both fault-tolerance and scalability through its distributed architecture. Flume provides multiple failover and recovery mechanisms, although the default "end-to-end" reliability mode that guarantees that accepted events will eventually be delivered is generally the recommended setting.

### 10.4.1 Flume Data Flows:

Flume expresses the data ingestion pathway from origin to destination as a *data flow*. In a data flow, a unit of data or *event* (e.g., a single log statement) travels from a source to the next destination via a sequence of hops. This concept of data flow is expressed even in the simplest entity in a Flume flow, a Flume *agent.* A Flume agent is a single unit within a Flume data flow (actually, a JVM process), through which events propagate once initiated at an external source. Agents consist of three configurable components: the *source, channel,* and *sink,* as shown in Figure 10-6.

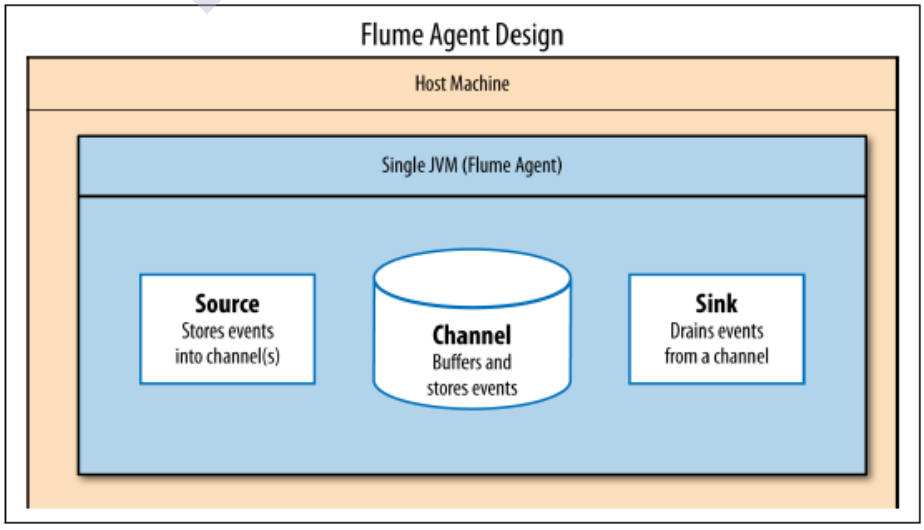



**Figure 10.6. flume agent design**

A Flume source is configured to listen for and consume events from one or more external data sources (not to be confused with a Flume source), which are configured by setting a name, type, and additional optional parameters for each data source. For example, we could configure up a Flume agent's source to accept events from an Apache access log by running a tail -f /etc/httpd/logs/access_log command. This type of source is called an exec source because it requires Flume to execute a Unix command to retrieve events.

When the agent consumes an event, the Flume source writes it to a channel, which acts as a storage queue that stores and buffers events until they are ready to be read. Events are written to channels transactionally, meaning that a channel keeps all events queued until they have been consumed and the corresponding transactions are explicitly closed. This enables Flume to maintain durability of data events even if an agent goes down.

Flume sinks eventually read and remove events from the channel and forward them to their next hop or final destination. Sinks can thus be configured to write its output as a streaming source for another Flume agent, or to a data store like HDFS or HBase.

Using this source-channel-sink paradigm, we can easily construct a simple singleagent Flume data flow to consume events from an Apache access log and write the log events to HDFS, as shown in Figure 10-7.

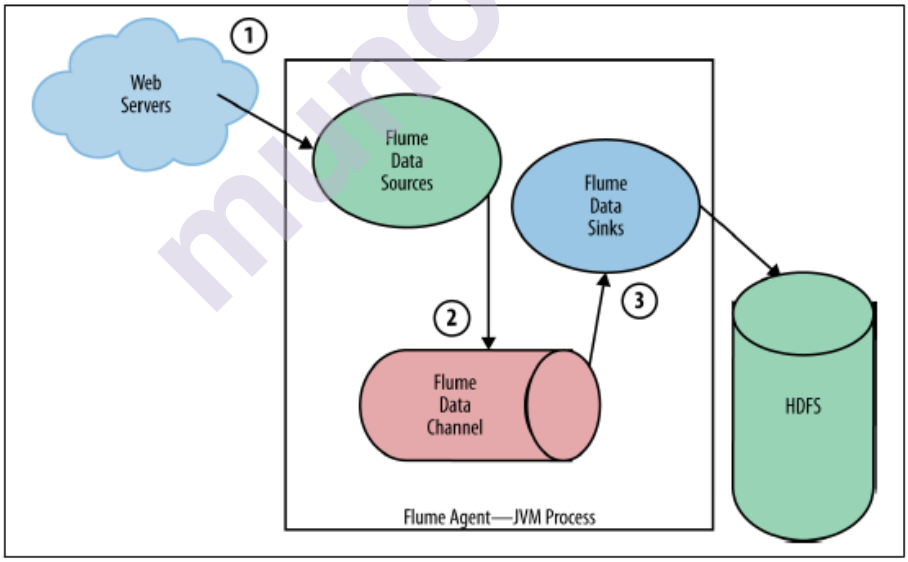



**Figure 10.7.Simple Flume data flow**

But because Flume agents are so adaptable and can even be configured to have multiple sources, channels, and sinks, we can actually construct multi-agent data flows by chaining several Flume agents together, as shown in Figure 10.8.

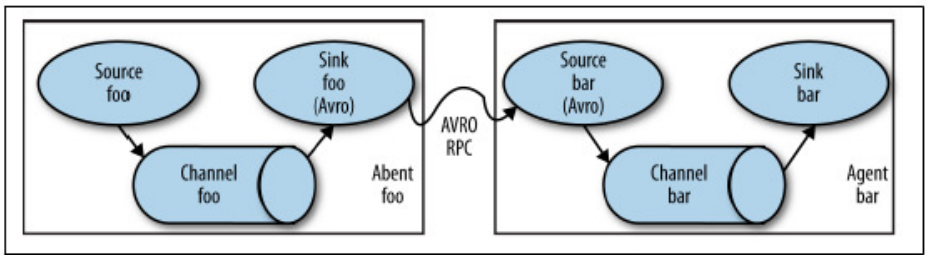



**Figure 10.8. Multi-agent Flume data flow**

There's almost no boundaries around how Flume agents can be organized into these complex data flows, although certain patterns and topologies of Flume data flows have emerged to handle common scenarios when dealing with a streaming data- processing architecture. For instance, a common scenario in log collection is when a large number of log producing clients are writing events to several Flume agents, which we call "first-tier" agents, as they are consuming data at the layer of the external data source(s). If we want to write these events to HDFS, we can set up each of the first-tier agents' sinks to write to HDFS, but this could present several problems as the first-tier scales out. Because several disparate agents are writing to HDFS independently, this data flow wouldn't be able to handle periodic bursts of data writes to the storage system and could thus introduce spikes in load and latency.

## 10.5 ANALYTICS WITH HIGHER-LEVEL APIS

### 10.5.1 Pig:

Pig, like Hive, is an abstraction of MapReduce, allowing users to express their data processing and analysis operations in a higher-level language that then compiles into a MapReduce job. Pig is now a top-level Apache Project that includes two main platform components:

- Pig Latin, a procedural scripting language used to express data flows.
- The Pig execution environment to run Pig Latin programs, which can be run in local or MapReduce mode and includes the Grunt command-line interface.

Pig Latin is procedural in nature and designed to enable programmers to easily implement a series of data operations and transformations that are applied to datasets to form a data pipeline. While Hive is great for use cases that translate well to SQL-based scripts, SQL can become unwieldy when multiple complex data transformations are required. Pig Latin is ideal for implementing these types of multistage data flows, particularly in cases where we need to aggregate data from multiple sources and perform subsequent transformations at each stage of the data processing flow.

Pig Latin scripts start with data, apply transformations to the data until the script describes the desired results, and execute the entire data processing flow as an optimized MapReduce job. Additionally, Pig supports the ability to integrate custom code with user-defined functions (UDFs) that can be written in Java, Python, or JavaScript, among other supported languages. Pig thus enables us to perform near arbitrary transformations and ad hoc analysis on our big data using comparatively simple constructs.

It is important to remember the earlier point that Pig, like Hive, ultimately compiles into MapReduce and cannot transcend the limitations of Hadoop's batch-processing approach. However, Pig does provide us with powerful tools to easily and succinctly write complex data processing flows, with the fine-grained controls that we need to build real business applications on Hadoop.

### 10.5.2 Pig Latin:

The following script loads Twitter tweets with the hashtag #unitedairlines over the course of a single week. The data file, *united_airlines_tweets.tsv*, provides the tweet ID, permalink, date posted, tweet text, and Twitter username. The script loads a dictionary, *dictionary.tsv*, of known "positive" and "negative" words along with sentiment scores (1 and -1, respectively) associated to each word. The script then performs a series of Pig transformations to generate a sentiment score and classification, either POSITIVE or NEGATIVE, for each computed tweet:

```
grunt> tweets = LOAD 'united_airlines_tweets.tsv' USING PigStorage('\t')
    AS (id_str:chararray, tweet_url:chararray, created_at:chararray,
    text:chararray, lang:chararray, retweet_count:int, favorite_count:int,
    screen_name:chararray);
grunt> dictionary = LOAD 'dictionary.tsv' USING PigStorage('\t')
    AS (word:chararray, score:int);
grunt> english_tweets = FILTER tweets BY lang == 'en';
grunt> tokenized = FOREACH english_tweets GENERATE id_str,
    FLATTEN( TOKENIZE(text) ) AS word;
grunt> clean_tokens = FOREACH tokenized GENERATE id_str,
    LOWER(REGEX_EXTRACT(word, '[#@]{0,1}(.*)', 1)) AS word;
grunt> token_sentiment = JOIN clean_tokens BY word, dictionary BY word;
grunt> sentiment_group = GROUP token_sentiment BY id_str;
grunt> sentiment_score = FOREACH sentiment_group
    GENERATE group AS id, SUM(token_sentiment.score) AS final;
grunt> classified = FOREACH sentiment_score
    GENERATE id, ( (final >= 0)? 'POSITIVE' : 'NEGATIVE' ) AS classification,
    final AS score;
grunt> final = ORDER classified BY score DESC;
grunt> STORE final INTO 'sentiment_analysis';
```

### 10.5.3 Data Types in pig:

**Table 10.1. Pig scalar types**

| Category | Type | Description | Example |
|---|---|---|---|
| Numeric | `int` | 32-bit signed integer | 12 |
| | `long` | 64-bit signed integer | 34L |
| | `float` | 32-bit floating-point number | 2.18F |
| | `double` | 64-bit floating-point number | 3e-17 |
| Text | `chararray` | String or array of characters | hello world |
| Binary | `bytearray` | Blob or array of bytes | N/A |

**Table 10.2 Pig relational operators**

| Category | Operator | Description |
|---|---|---|
| Loading and storing | `LOAD` | Loads data from the file system or other storage source |
| | `STORE` | Saves a relation to the file system or other storage |
| | `DUMP` | Prints a relation to the console |
| Filtering and projection | `FILTER` | Selects tuples from a relation based on some condition |
| | `DISTINCT` | Removes duplicate tuples in a relation |
| | `FOREACH...GENERATE` | Generates data transformations based on columns of data. |
| | `MAPREDUCE` | Executes native MapReduce jobs inside a Pig script |
| | `STREAM` | Sends data to an external script or program |
| | `SAMPLE` | Selects a random sample of data based on the specified sample size |
| Grouping and joining | `JOIN` | Joins two or more relations |
| | `COGROUP` | Groups the data from two or more relations |
| | `GROUP` | Groups the data in a single relation |
| | `CROSS` | Creates the cross-product of two or more relations |
| Sorting | `ORDER` | Sorts the relation by one or more fields |
| | `LIMIT` | Limits the number of tuples returned from a relation |
| Combining and splitting | `UNION` | Computes the union of two or more relations |
| | `SPLIT` | Partitions a relation into two or more relations |



### 10.5.4 User-Defined Functions:

Pig provides extensive support for such user-defined functions (UDFs), and currently provides integration libraries for six languages: Java, Jython, Python, JavaScript, Ruby, and Groovy. In this scenario, we would like to write a custom eval UDF in java that will allow us to convert the score classification evaluation into a function.

```
package com.statistics.pig;

import java.io.IOException;

import org.apache.pig.EvalFunc;
import org.apache.pig.backend.executionengine.ExecException;
import org.apache.pig.data.Tuple;

public class Classify extends EvalFunc {

    @Override
    public String exec(Tuple input) throws IOException {
        if (args == null || args.size() == 0) {
            return false;
        }
        try {
            Object object = args.get(0);
            if (object == null) {
                return false;
            }
            int i = (Integer) object;
            if (i >= 0) {
                return new String("POSITIVE");
            } else {
                return new String("NEGATIVE");
            }
        } catch (ExecException e) {
            throw new IOException(e);
        }
    }
}
```

To use this function, we need to compile it, package it into a JAR file, and then register the JAR with Pig by using the REGISTER operator:

```
grunt> REGISTER statistics-pig.jar;
```

We can then invoke the function in a command:

```
grunt> classified = FOREACH sentiment_score GENERATE id,
    com.statistics.pig.Classify(final) AS classification, final AS score;
```

### 10.5.5 Wrapping Up:

Pig can be a powerful tool for users who prefer a procedural programming model. It provides the ability to control data checkpoints in the pipeline, as well as fine-grained controls over how the data is processed at each step. This makes Pig a great choice when you require more flexibility in controlling the sequence of operations in a data flow (e.g., an extract, form, and load, or ETL, process), or when you are working with semi-structured data that may not lend itself well to Hive's SQL syntax.

## 10.6 SPARK'S HIGHER-LEVEL APIS

In practice, a typical analytic workflow will entail some combination of relational queries, procedural programming, and custom processing, which means that most end-to-end Hadoop workflows involve integrating several disparate components and switching between different programming APIs. Spark, in contrast, provides two major programming advantages over the MapReduce-centric Hadoop stack:

- Built-in expressive APIs in standard, general-purpose languages like Scala, Java, Python, and R
- A unified programming interface that includes several built-in higher-

level libraries to support a broad range of data processing tasks, including complex interactive analysis, structured querying, stream processing, and machine learning.

### 10.6.1 Spark SQL:

Spark SQL is a module in Apache Spark that provides a relational interface to work with structured data using familiar SQL-based operations in Spark. It can be accessed through JDBC/ODBC connectors, a built-in interactive Hive console, or via its built-in APIs. The last method of access is the most interesting and powerful aspect of Spark SQL; because Spark SQL actually runs as a library on top of Spark's Core engine and APIs, we can access the Spark SQL API using the same programming interface that we use for Spark's RDD APIs, as shown in Figure 10-9.

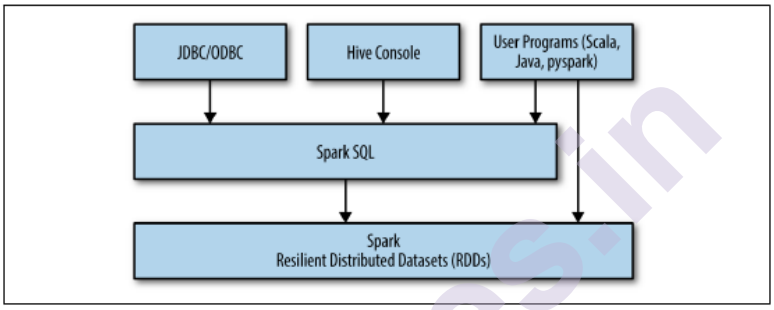



**Figure 10.9. Spark SQL interface**

This allows us to seamlessly combine and leverage the benefits of relational queries with the flexibility of Spark's procedural processing and the power of Python's analytic libraries, all in one programming environment.

Let's write a simple program that uses the Spark SQL API to load JSON data and query it. You can enter these commands directly in a running pyspark shell or in a Jupyter notebook that is using a pyspark kernel; in either case, ensure that you have a running SparkContext, which we'll assume is referenced by the variable sc.

To begin, we'll need to import the SQLContext class from the *pyspark.sql* package.

*from pyspark.sql import SQLContext*
*sqlContext = SQLContext(sc)*

With the file properly formatted, we can easily load its contents by calling sqlCon text.read.json and passing it the path to the file:

*parking = sqlContext.read.json('../data/sf_parking/sf_parking_clean.json')*

In order to run a SQL statement against our dataset, we must first register it as a temporary named table:

*parking.registerTempTable("parking")*

This allows us to run additional table and SQL methods, including

show, which will display the first 20 rows of data in a tabular format:
parking.show()

To execute a SQL statement on the parking table, we use the sql method, passing it the full query.

### 10.6.2 DataFrames:

DataFrames are the underlying data abstraction in Spark SQL. The data frame concept should be very familiar to users of Python's Pandas or R, and in fact, Spark's DataFrames are interoperable with native Pandas (using pyspark) and R data frames (using SparkR). In Spark, a DataFrame also represents a tabular collection of data with a defined schema. The key difference between a Spark DataFrame and a dataframe in Pandas or R is that a Spark DataFrame is a distributed collection that actually wraps an RDD; you can think of it as an RDD of row objects.

Additionally, DataFrame operations entail many optimizations under the hood that not only compile the query plan into executable code, but substantially improve the performance and memory-footprint over comparable handcoded RDD operations. In fact, in a benchmark test that compared the runtimes between DataFrames code that aggregated 10 million integer pairs against equivalent RDD code, DataFrames were not only found to be up to 4–5x faster for these workloads, but they also close the performance gap between Python and JVM implementations.

The concise and intuitive semantics of the DataFrames API coupled with the performance optimizations provided by its computational engine was the impetus to make DataFrames the main interface for all of Spark's modules, including Spark SQL, RDDs, MLlib, and GraphX. In this way, the DataFrames API provides a unified engine across all of Spark's data sources, workloads, and environments, as shown in Figure 10-10.

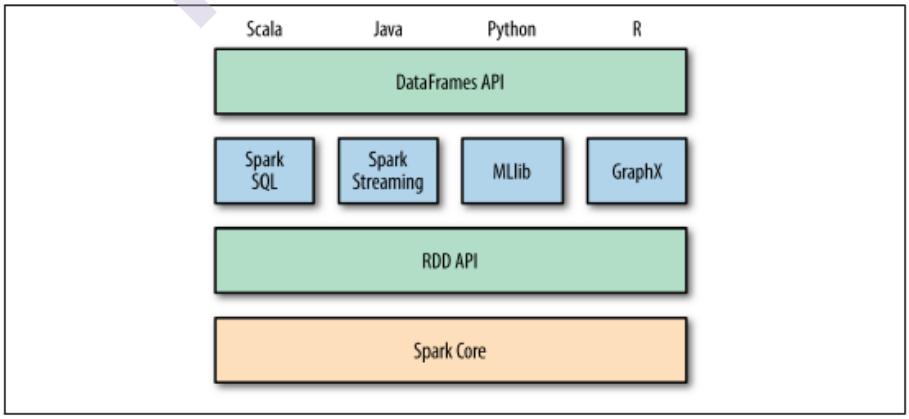



**Figure 10.10 Data Frames as sparks unified interface**

**Example of chaining several simple DataFrame operations:**

```
from pyspark.sql import functions as F

aggr_by_type = parking.select("primetype", "secondtype", "regcap") \
                      .where("trim(primetype) != ''") \
                      .groupBy("primetype", "secondtype") \
                      .agg(
                         F.count("*").alias("count"),
                         F.round(F.avg("regcap"), 0).alias("avg_spaces")
                       ) \
                       .sort("count", ascending=False)
```

The advantage of this approach over raw SQL is that we can easily iterate on a complex query by successively chaining and testing operations. Additionally, we have access to a rich collection of built-in functions from the DataFrames API, including the count, round, and avg aggregation functions that we used previously. The pyspark.sql.functions module also contains several mathematical and statistical utilities that include functions for:

- Random data generation
- Summary and descriptive statistics
- Sample covariance and correlation
- Cross tabulation (a.k.a. contingency table)
- Frequency computation
- Mathematical functions

## UNIT END QUESTIONS

1. Explain in detail how Keys allow parallel reduction by partitioning the keyspace to multiple reducers.
2. What is the functionality of the explode mapper? Explain in detail with example.
3. What is the functionality of the filter mapper? Explain in detail with example.
4. What is the functionality of the identity pattern? Explain in detail with example.
5. Write in brief about design pattern. Explain each of its category.
6. Consider a specific example where we have a dataset that originates from news arti-cles or blog posts and a prediction task where we want to determine the number of comments in the next 24 hours. Then how the data flow will be?
7. Write a command for the following in Hive Query Language:

   i) changing directory to HIVE_HOME

ii) creating a database iii) creating a table iv) loading data in a table

v) counting number of rows in a table and vi) exiting the Hive CLI

8. Write and explain with suitable example any three data analysis commands with Hive.

9. What is the major drawback of conventional relational approach for many data analytics applications? How can it be resolved? Explain in detail.

10. How HBase schema can be created? How data can be inserted? and Cell values can be fetched? Explain with suitable example.

11. Which different types of filters can be used in HBase? Explain its entire procedure with appropriate commands.

12. Write the entire procedure with appropriate commands for importing data from MySQL to HDFS.

13. Write the entire procedure with appropriate commands for importing data from MySQL to Hive.

14. Write the entire procedure with appropriate commands for importing data from MySQL to HBase.

15. Explain Flume Data Flows with a neat diagram.

16. How to construct a simple singleagent Flume data flow to consume events from an Apache access log and write the log events to HDFS? Explain with a neat diagram.

17. Explain with example Relations, tuples and Filtering in context of Pig.

18. Explain with example Projection in context of Pig.

19. Explain with example Grouping and joiningin context of Pig.

20. Explain with example Storing and outputting datain context of Pig.

21. Write and describe various Pig relational operators.

22. Explain Spark SQL interface architecture with a neat diagram.

## REFERENCES


- Data Science & Big Data Analytics Discovering, Analyzing, Visualizing and Presenting Data EMC Education Services Published by John Wiley & Sons, Inc


|

*****