MRCET CAMPUS

MALLA REDDY COLLEGE OF ENGINEERING & TECHNOLOGY

(AUTONOMOUSINSTITUTION – UGC, GOVT. OF INDIA)

Department of CSE
(Emerging Technologies)
(Data Science, Cyber Security & IOT)

B.TECH(R-22 Regulation)
(III YEAR–II SEM)
(2024-25)

DATA ANALYTICS
(R22A6703)

LECTURE NOTES

MALLA REDDY COLLEGE OF ENGINEERING & TECHNOLOGY
(Autonomous Institution–UGC,Govt. of India)
Recognizedunder2(f)and12(B)ofUGCACT1956
(Affiliated to JNTUH, Hyderabad, Approved by AICTE-Accredited by NBA & NAAC – 'A' Grade - ISO 9001:2015 Certified)
Maisammaguda,Dhulapally(PostVia.Hakimpet),Secunderabad–500100,TelanganaState,India

Department of Computer Science and Engineering

EMERGING TECHNOLOGIES

DATA ANALYTICS
(R22A6703)

LECTURE NOTES

Prepared by
Dr. M.V.Kamal, HOD, Professor
&
Dr. I. Nagaraju, Professor
&
P. Sreenivas, Associate Professor

On
25-12-2024

Department of Computer Science and Engineering

EMERGING TECHNOLOGIES

Vision

- "To be at the forefront of Emerging Technologies and to evolve as a Centre of Excellence in Research, Learning and Consultancy to foster the students into globally competent professionals useful to the Society."

Mission

The department of CSE (Emerging Technologies) is committed to:

- To offer highest Professional and Academic Standards in terms of Personal growth and satisfaction.
- Make the society as the hub of emerging technologies and thereby capture opportunities in new age technologies.
- To create a benchmark in the areas of Research, Education and Public Outreach.
- To provide students a platform where independent learning and scientific study are encouraged with emphasis on latest engineering techniques.

QUALITY POLICY

- To pursue continual improvement of teaching learning process of Undergraduate and Post Graduate programs in Engineering & Management vigorously.
- Toprovidestateofartinfrastructureandexpertisetoimpartthequalityeducationand research environment to students for a complete learning experiences.
- Developing students with a disciplined and integrated personality.
- To offer quality relevant and cost effective programmes to produce engineers as per requirements of the industry need.

For more information : www.mrcet.ac.in

MALLA REDDY COLLEGE OF ENGINEERING & TECHNOLOGY

DEPARTMENTOF COMPUTER SCIENCE - ET

M R C E T CAMPUS | AUTONOMOUS INSTITUTION - UGC, GOVT. OF INDIA
III Year B.Tech CSE (DS) - II Sem L/T/P/C 3/0/0/3
(R22A6703) DATA ANALYTICS

Course Objectives:

To explore the fundamental concepts of data analytics.

To learn the principles and methods of statistical analysis

To gain the knowledge on Big Data Techniques like Hadoop

To explore the Map Reduce and YARN techniques for Big Data Analytics

To understand the various search methods and visualization techniques

UNIT - I Data Management: Design Data Architecture and manage the data for analysis, understand various sources of Data like Sensors/Signals/GPS etc. Data Management, Data Quality (noise, outliers, missing values, duplicate data) and Data Processing & Processing.

UNIT - II Data Analytics: Introduction to Analytics, Introduction to Tools and Environment, Application of Modelling in Business, Databases & Types of Data and Variables, Data Modelling Techniques, Missing Imputations etc. Need for Business Modelling.

UNIT – III

Big data technologies and Databases: Hadoop – Requirement of Hadoop Framework - Design principle of Hadoop –Comparison with other system SQL and RDBMS- Hadoop Components – Architecture -Hadoop 1 vs Hadoop 2.

UNIT – IV

MapReduce and YARN framework: Introduction to MapReduce , Processing data with Hadoop using MapReduce, Introduction to YARN, Architecture, Managing Resources and Applications with Hadoop YARN.

Big data technologies and Databases: NoSQL: Introduction to NoSQL - Features and Types-Advantages & Disadvantages -Application of NoSQL.

UNIT - V Data Visualization: Pixel-Oriented Visualization Techniques, Geometric Projection Visualization Techniques, Icon-Based Visualization Techniques, Hierarchical Visualization Techniques, Visualizing Complex Data and Relations.

TEXT BOOKS:

1. Student's Handbook for Associate Analytics – II, III.
2. Data Mining Concepts and Techniques, Han, Kamber, 3rd Edition, Morgan Kaufmann Publishers.

REFERENCE BOOKS:

1.Data Mining Analysis and Concepts, M. Zaki and W. Meira 3. Mining of Massive Datasets, Jure Leskovec Stanford Univ. Anand Rajaraman Milliway Labs Jeffrey D Ullman Stanford Univ.

2. Seema Acharya and Subhashini Chellappan, "Big Data and Analytics", Wiley India Pvt. Ltd., 2016.

3. Mike Frampton, "Mastering Apache Spark", Packt Publishing, 2015.

Course Outcomes: After completion of this course students will be able to:

1. Understand the impact of data analytics for business decisions and strategy
2. Carry out data analysis/statistical analysis
3. To carry out standard data visualization and formal inference procedures
4. Design Data Architecture; Understand various Data Sources

MALLA REDDY COLLEGE OF ENGINEERING & TECHNOLOGY

DEPARTMENT OF COMPUTER SCIENCE - ET

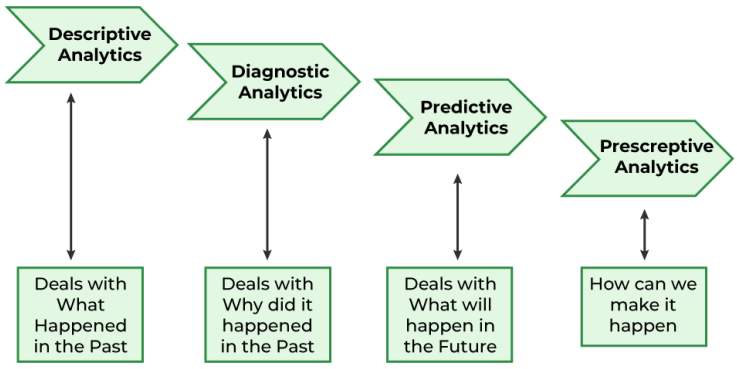
INDEX

UNIT - I

Data Management

Data analytics is the process of analyzing raw data to find trends and answer questions. It has a broad scope across the field. This process includes many different techniques and goals that can shift from industry to industry.

The data analytics process has components that can help a variety of initiatives. By combining these components, a successful data analytics initiative can help answer business questions related to historical trends, future predictions and decision making.



The data used was not as much of as it is today, the data then could be so easily stored and managed by all the users and business enterprises on a single computer, because the data never exceeded to the extent of 19 exabytes but now in this era, the data has increased about 2.5 quintillion per day.

Most of the data is generated from social media sites like Facebook, Instagram, Twitter, etc, and the other sources can be e-business, e-commerce transactions, hospital, school, bank data, etc. This data is impossible to manage by traditional data storing techniques. Either the data being generated from large-scale enterprises or the data generated from an individual, each and every aspect of data needs to be analysed to benefit yourself from it. But how do we do it? Well, that's where the term 'Data Analytics' comes in.

Data Analytics important :

Data Analytics has a key role in improving your business as it is used to gather hidden insights, Interesting Patterns in Data, generate reports, perform market analysis, and improve business requirements

Role of Data Analytics

Gather Hidden Insights – Hidden insights from data are gathered and then analyzed with respect to business requirements.

Generate Reports – Reports are generated from the data and are passed on to the respective teams and individuals to deal with further actions for a high rise in business.

Perform Market Analysis – Market Analysis can be performed to understand the strengths and weaknesses of competitors.

Improve Business Requirement – Analysis of Data allows improving Business to customer requirements and experience.

Data and architecture design:

Data architecture in Information Technology is composed of models, policies, rules or standards that govern which data is collected, and how it is stored, arranged, integrated, and put to use in data systems and in organizations.

A data architecture should set data standards for all its data systems as a vision or a model of the eventual interactions between those data systems .

Data architectures address data in storage and data in motion; descriptions of data stores, data groups and data items; and mappings of those data artifacts to data qualities, applications, locations etc. Essential to realizing the target state, Data Architecture describes how data is processed, stored, and utilized in a given system. It provides criteria for data processing operations that make it possible to design data flows and also control the flow of data in the system.

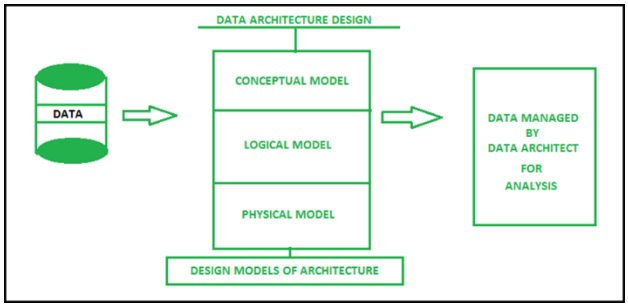
The Data Architect is typically responsible for defining the target state, aligning during development and then following up to ensure enhancements are done in the spirit of the original blueprint.

The Data Architect breaks the subject down by going through 3 traditional architectural processes:

Conceptual model: It is a business model which uses Entity Relationship (ER) model for relation between entities and their attributes.

Logical model: It is a model where problems are represented in the form of logic such as rows and column of data, classes, xml tags and other DBMS techniques

. Physical model: Physical models holds the database design like which type of database technology will be suitable for architecture.



Various constraints and influences will have an effect on data architecture design. These include enterprise requirements, technology drivers, economics, business policies and data processing need.

Enterprise requirements: These will generally include such elements as economical and effective system expansion, acceptable performance levels (especially system access speed), transaction reliability, and transparent data management. In addition, the conversion of raw data such as transaction records and image files into more useful information forms through such features as data warehouses is also a common organizational requirement, since this enables managerial decision making and other organizational processes. One of the architecture techniques is the split between managing transaction data and (master) reference data. Another one is splitting data capture systems from data retrieval systems (as done in a data warehouse).

Technology drivers: These are usually suggested by the completed data architecture and database architecture designs. In addition, some technology drivers will derive from existing organizational integration frameworks and standards, organizational economics, and existing site resources (e.g. previously purchased software licensing).

Economics: These are also important factors that must be considered during the data architecture phase. It is possible that some solutions, while optimal in principle, may not be potential candidates due to their cost. External factors such as the business cycle, interest rates, market conditions, and legal considerations could all have an effect on decisions relevant to data architecture.

Business policies: Business policies that also drive data architecture design include internal organizational policies, rules of regulatory bodies, professional standards, and applicable governmental laws that can vary by applicable agency. These policies and rules will help describe the manner in which enterprise wishes to process their data.

Data processing needs These include accurate and reproducible transactions performed in high volumes, data warehousing for the support of management information systems (and potential data mining), repetitive periodic reporting, ad hoc reporting, and support of various organizational initiatives as required (i.e. annual budgets, new product development) The General Approach is based on designing the Architecture at three Levels of Specification.

Understand various sources of the Data:

Data can be generated from two types of sources namely Primary and Secondary Sources of Primary Data. Data collection is the process of acquiring, collecting, extracting, and storing the voluminous

amount of data which may be in the structured or unstructured form like text, video, audio, XML files, records, or other image files used in later stages of data analysis. In the process of big data analysis, “Data collection” is the initial step before starting to analyze the patterns or useful information in data.

The data which is to be analyzed must be collected from different valid sources. The data which is collected is known as raw data which is not useful now but on cleaning the impure and utilizing that data for further analysis forms information, the information obtained is known as “knowledge”. Knowledge has many meanings like business knowledge or sales of enterprise products, disease treatment, etc. The main goal of data collection is to collect information-rich data.

Data collection starts with asking some questions such as what type of data is to be collected and what is the source of collection. Most of the data collected are of two types known as qualitative data which is a group of non-numerical data such as words, sentences mostly focus on behaviour and actions of the group and another one is quantitative data which is in numerical forms and can be calculated using different scientific tools and sampling data.

Tools used in data analytics :

Data Analytics in the market, many tools have emerged with various functionalities for this purpose. Either open-source or user-friendly, the top tools in the data analytics market are as follows.

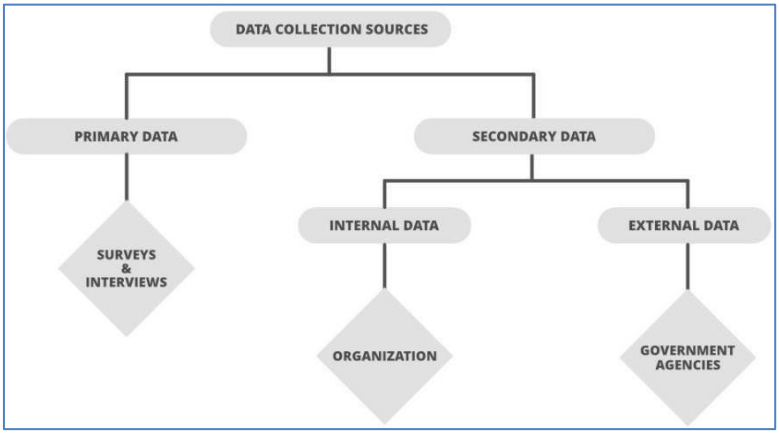
- R programming
- Python
- Tableau Public
- Qlik View
- SAS
- Microsoft Excel
- Rapid Miner
- KNIME
- Open Refine
- Apache Spark

Missing data : (or missing values) is defined as the data value that is not stored for a variable in the observation of interest. The problem of missing data is relatively common in almost all research and can have a significant effect on the conclusions that can be drawn from the data

Data duplication : is the process of creating one or more identical versions of data, either intentionally, such as for planned backups, or unintentionally

Data collection :

1. Primary data
2. Secondary data



1. Primary data:

The data collected must be according to the demand and requirements of the target audience on which analysis performed otherwise it would be a burden in the data processing.

Few methods of collecting primary data:

1. Interview method:

- The data collected during this process is through interviewing the target audience by a person called interviewer and the person who answers the interview is known as the interviewee.
- Some basic business or product related questions are asked and noted down in the form of notes, audio, or video and this data is stored for processing.
- These can be both structured and unstructured like personal interviews or formal interviews through telephone, face to face, email, etc.

2. Survey method:

- The survey method is the process of research where a list of relevant questions are asked and answers are noted down in the form of text, audio, or video.
- The survey method can be obtained in both online and offline model like through website forms and email. Then that survey answers are stored for analyzing data. Examples are online surveys or surveys through social media polls.

3. **Observation method:**

- The observation method is same method of data collection in which the researcher keenly observes the behaviour and practices of the target audience using some data collecting tool and stores the observed data in the form of text, audio, video, or any raw formats.
- In this method, the data is collected directly by posting a few questions on the participants. For example, observing a group of customers and their behaviour towards the products. The data obtained will be sent for processing.

Internal source:

These types of data can easily be found within the organization such as market record, as sales record, transactions, customer data, accounting resources, etc. The cost and time consumption is less in obtaining internal sources.

- **Accounting resources**- This gives so much information which can be usedby the marketing researcher. They give information about internal factors.
- **Sales Force Report**- It gives information about the sales of a product. The information provided is from outside the organization.
- **Internal Experts**-These are people who are heading the various departments. They can give an idea of how a particular thing is working.
- **Miscellaneous Reports**-These are what information you are getting from operational reports. If the data available within the organization are unsuitable or in adequate,the marketer should extend the search to external secondary data sources.

External source:

The data which can't be found at internal organizations and can begained through external third-party resources is external source data. The cost and time consumption are more because this contains a huge amount of data. Examples of external sources are Government publications, news publications, Registrar General of India, planning commission, international labour bureau, syndicate services, and other non-governmental publications.

1. **Government Publications-**

- Government sources provide an extremely rich pool of data for the researchers. In addition, many of these data are available free of cost on internet websites. There are number of government agencies generating data.

  It is an office which generates demographic data. It includes details of gender, age,

occupation etc.

2. **Central Statistical Organization-**
   - This organization publishes the national accounts statistics. It contains estimates of national income for several years, growth rate, and rate of major economic activities. Annual survey of Industries is also published by the CSO.
   - It gives information about the total number of workers employed, production units, material used and value added by the manufacturer.
3. **Director General of Commercial Intelligence-**
   - This office operates from Kolkata. It gives information about foreign trade i.e. import and export. These figures are provided region-wise and country-wise.

Data Management process :

Data management is the practice of collecting, keeping, and using data securely, efficiently, and cost- effectively. The goal of data management is to help people, organizations, and connected things optimize the use of data within the bounds of policy and regulation so that they can make decisions and take actions that maximize the benefit to the organization.

Managing digital data in an organization involves a broad range of tasks, policies, procedures, and practices.

Cloud Computing means storing and accessing the data and programs on remote servers that are hosted on the internet instead of the computer's hard drive or local server. Cloud computing is also referred to as Internet-based computing, it is a technology where the resource is provided as a service through the Internet to the user. The data that is stored can be files, images, documents, or any other storable document.

The following are some of the Operations that can be performed with Cloud Computing

- Storage, backup, and recovery of data
- Delivery of software on demand
- Development of new applications and services
- Streaming videos and audio

Software as a Service is the most common cloud service used by organizations. A 2020 Virayo study found that 80 percent of organizations use one or more SaaS applications in their business. While using SaaS services, you don't have to install any software on your computer. Instead, you can easily access them on the cloud where they are stored. So, if you want to do some urgent work and do not have your laptop with you, all you need is an internet connection and a browser to access the required tools.

Features of PaaS

PaaS has several features. Some of them include:

- **Auto-scaling:** Since it is based on virtualization technology, resources can be scaled up and down at your convenience.
- **Resource sharing:** PaaS allows resource sharing amongst different development teams.
- **Accessibility:** It also allows several users to access the platform with the same development application.
- **Time-saving:** PaaS offers developers pre-coded components and several development tools, which saves their time and resources.
- **Integrations:** PaaS allows the integration of databases and web services

Features of IaaS include:

- **Dynamic scaling:** IaaS allows dynamic and flexible scaling of resources as they are available in an as-a-service model.
- **Platform virtualization:** IaaS uses platform virtualization technology to provide cloud computing infrastructure.
- **Costs:** The services are available on a pay-as-you-go basis. Therefore you pay only for the resources you use.
- **Control:** IaaS users have complete control over their infrastructure and IT platform.

Benefits of IaaS:

- **Scaling:** The IaaS services are available 24*7*365. You can easily scale globally and enhance application performance.
- **Enhanced security:** The data is secure and can only be accessed by authorized people. IaaS also allows you to keep backups in case of data loss.
- **Automation:** IaaS easily automates the deployment of various resources, such as networks, and servers.
- **Saves time and cost:** Users save a lot of time and cost since all the hardware maintenance is done by the vendor.
- **Flexibility:** IaaS vendors allow users to purchase the features they need and scale up and down at their convenience.

Data quality refers to the reliability, accuracy, completeness, and consistency of data. High-quality data is free from errors, inconsistencies, and inaccuracies, making it suitable for reliable decision-making and analysis.

Data quality encompasses various aspects, including correctness, timeliness, relevance, and adherence to predefined standards. Organizations prioritize data quality to ensure that their information assets meet the required standards and contribute effectively to business processes and decision-making. Effective data quality management involves processes such as data profiling, cleansing, validation, and monitoring to maintain and improve data integrity.

Data Quality vs Data Integrity

Oversight of data quality is only one component of data integrity, which includes many other elements as well. Keeping data valuable and helpful to the company is the main objective of data integrity. To achieve data integrity, the following four essential elements are necessary:

- Data Integration: The smooth integration of data from various sources is very much essential.
- Data Quality: A vital aspect of maintaining data integrity is verifying that the information is complete, legitimate, unique, current, and accurate.
- Location Intelligence: when location insights are included in the data, it gains dimension and therefore becomes more useful and actionable.
- Data Enrichment: By adding more information from outside sources, such customer, business, and geographical data, data enrichment may improve the context and completeness of data.

Outlier :

Outliers are data points that lie outside the majority of the data in a particular data set. These values might be much higher or lower in value than other points and may impact the results of the data analysis in ways that misrepresent the data sample. By learning how to identify and handle outliers, data analysts can increase the likelihood that their analysis will accurately reflect the validity and reliability of their results.

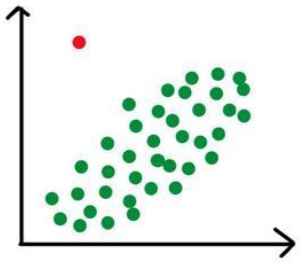
The 3 Different Types of Outliers

In statistics and data science, there are three generally accepted categories which all outliers fall into:

- Type 1: Global Outliers (aka Point Anomalies)
- Type 2: Contextual Outliers (aka Conditional Anomalies)
- Type 3: Collective Outliers

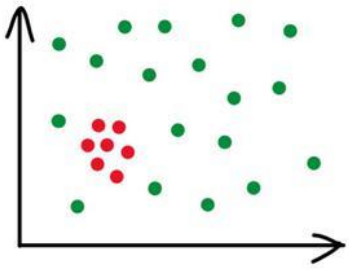
1. Global Outliers

1. Definition: Global outliers are data points that deviate significantly from the overall distribution of a dataset.
2. Causes: Errors in data collection, measurement errors, or truly unusual events can result in global outliers.
3. Impact: Global outliers can distort data analysis results and affect machine learning model performance.
4. Detection: Techniques include statistical methods (e.g., z-score, Mahalanobis distance), machine learning algorithms (e.g., isolation forest, one-class SVM), and data visualization techniques.
5. Handling: Options may include removing or correcting outliers, transforming data, or using robust methods.
6. Considerations: Carefully considering the impact of global outliers is crucial for accurate data analysis and machine learning model outcomes



2. Collective Outliers

1. Definition: Collective outliers are groups of data points that collectively deviate significantly from the overall distribution of a dataset.
2. Characteristics: Collective outliers may not be outliers when considered individually, but as a group, they exhibit unusual behavior.
3. Detection: Techniques for detecting collective outliers include clustering algorithms, density-based methods, and subspace-based approaches.
4 Impact: Collective outliers can represent interesting patterns or anomalies in data that may require special attention or further investigation.
5. Handling: Handling collective outliers depends on the specific use case and may involve further analysis of the group behavior, identification of contributing factors, or considering contextual information.
6. Considerations: Detecting and interpreting collective outliers can be more complex than individual outliers, as the focus is on group behavior rather than individual data points. Proper understanding of the data context and domain knowledge is crucial for effective handling of collective outliers



3. Contextual Outliers

1. Definition: Contextual outliers are data points that deviate significantly from the expected behavior within a specific context or subgroup.

2. Characteristics: Contextual outliers may not be outliers when considered in the entire dataset, but they exhibit unusual behavior within a specific context or subgroup.

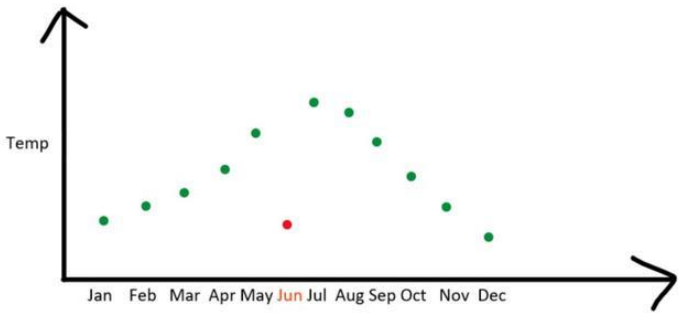
3. Detection: Techniques for detecting contextual outliers include contextual clustering, contextual anomaly detection, and context-aware machine learning approaches.

4. Contextual Information: Contextual information such as time, location, or other relevant factors are crucial in identifying contextual outliers.

5. Impact: Contextual outliers can represent unusual or anomalous behavior within a specific context, which may require further investigation or attention.

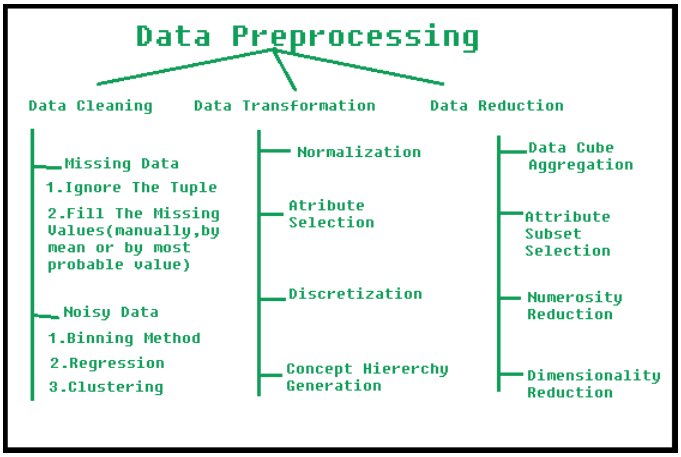
6. Handling: Handling contextual outliers may involve considering the contextual information, contextual normalization or transformation of data, or using context-specific models or algorithms.

7. Considerations: Proper understanding of the context and domain-specific knowledge is crucial for accurate detection and interpretation of contextual outliers, as they may vary based on the specific context or subgroup being considered.



Data pre processing:

Preprocessing in Data Mining: Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format.



Data Preprocessing

1. Data Cleaning: The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc.

- Missing Data: This situation arises when some data is missing in the data. It can be handled in various ways.
  Some of them are:
  - Ignore the tuples: This approach is suitable only when the dataset we have is quite large and multiple values are missing within a tuple.
  - Fill the Missing values: There are various ways to do this task. You can choose to fill the missing values manually, by attribute mean or the most probable value.

- Noisy Data: Noisy data is a meaningless data that can't be interpreted by machines.It can be generated due to faulty data collection, data entry errors etc. It can be handled in following ways :
  - Binning Method: This method works on sorted data in order to smooth it. The whole

data is divided into segments of equal size and then various methods are performed to complete the task. Each segmented is handled separately. One can replace all data in a segment by its mean or boundary values can be used to complete the task.

- Regression:Here data can be made smooth by fitting it to a regression function.The regression used may be linear (having one independent variable) or multiple (having multiple independent variables).
- Clustering: This approach groups the similar data in a cluster. The outliers may be undetected or it will fall outside the clusters.

2. Data Transformation: This step is taken in order to transform the data in appropriate forms suitable for mining process. This involves following ways:

- Normalization: It is done in order to scale the data values in a specified range (-1.0 to 1.0 or 0.0 to 1.0)
- Attribute Selection: In this strategy, new attributes are constructed from the given set of attributes to help the mining process.
- Discretization: This is done to replace the raw values of numeric attribute by interval levels or conceptual levels.
- Concept Hierarchy Generation: Here attributes are converted from lower level to higher level in hierarchy. For Example-The attribute “city” can be converted to “country”.

3. Data Reduction: Data reduction is a crucial step in the data mining process that involves reducing the size of the dataset while preserving the important information. This is done to improve the efficiency of data analysis and to avoid overfitting of the model. Some common steps involved in data reduction are:

- Feature Selection: This involves selecting a subset of relevant features from the dataset. Feature selection is often performed to remove irrelevant or redundant features from the dataset. It can be done using various techniques such as correlation analysis, mutual information, and principal component analysis (PCA).
- Feature Extraction: This involves transforming the data into a lower-dimensional space while preserving the important information. Feature extraction is often used when the original features are high-dimensional and complex. It can be done using techniques such as PCA, linear discriminant analysis (LDA), and non-negative matrix factorization (NMF).
- Sampling: This involves selecting a subset of data points from the dataset. Sampling is often

used to reduce the size of the dataset while preserving the important information. It can be done using techniques such as random sampling, stratified sampling, and systematic sampling.

- Clustering: This involves grouping similar data points together into clusters. Clustering is often used to reduce the size of the dataset by replacing similar data points with a representative centroid. It can be done using techniques such as k-means, hierarchical clustering, and density-based clustering.
- Compression: This involves compressing the dataset while preserving the important information. Compression is often used to reduce the size of the dataset for storage and transmission purposes. It can be done using techniques such as wavelet compression, JPEG compression, and gif compression.

How is Data Preprocessing Used?

This we have earlier noted is one of the reasons data preprocessing is important in the earlier stages of the development of machine learning and AI applications. While in AI context data preprocessing is applied in order to optimize the methods used to cleanse, transform and structure the data in a way that will enhance the accuracy of a new model with less computing power used.

An excellent data preprocessing step will help develop a set of components or tools that can be utilized to quickly prototype on a set of ideas or even run experiments on improving business processes or customer satisfaction. For instance, preprocessing can enhance the manner in which data is arranged for a recommendation engine by enhancing the age ranges of customers that are used for categorisation.

It can also make the process of developing and enhancing data easier for more enhanced BI which is beneficial to the business. For instance, small size, category or regions of the customers may have different behaviors across regions. Backend processing the data into the correct formats might enable BI teams to integrate such findings into BI dashboard.

In a broad concept, data preprocessing is a sub-process of web mining which is used in customer relationship management (CRM). There's usually the possibility of pre-processing of the Web usage logs in order to arrive at meaningful data sets referred to as user transactions which are actually a set of groups of URL references. Sessions may be stored to make user identification possible as well as the websites requested and their sequence and time of use.

UNIT – II

Data Analytics

Introduction to Analytics :

Data has been the buzzword for ages now. Either the data being generated from large-scale enterprises or the data generated from an individual, each and every aspect of data needs to be analyzed to benefit yourself from it.

Why is Data Analytics important?

1. Data Mining

Most simply stated, data mining is a process used to extract usable data from a large dataset. Data mining involves data collection, warehousing and computer processing. In order to segment and evaluate the data, data mining uses advanced algorithms.

Real-Life Scenario: Data mining is often used in the health care industry during patient clinical trials. The algorithms can evaluate behavioral patterns of large amounts of data for interpretation, knowledge building and decision making.

2. Text Analytics

Text analytics is the process of drawing meaning out of written communication. Usually, text analytics software relies on text mining and natural language processing (NLP) algorithms to find patterns and meaning.

Real-Life Scenario: Text analytics is used to build the auto-correct function on your mobile device. It will not only correct your spelling, but also predict what you're going to type next based on linguistic analysis and data pattern recognition.

3. Data Visualization

Data visualization presents a clear picture of what the data actually means. Using bar graphs, pie charts, tables and other visuals, data visualization makes the data easier for those making business decisions to comprehend.

Real-Life Scenario: Data visualizations are part of our everyday lives on IoT devices – and you probably don't even realize it. Think about the exercise rings on your smartwatch, the energy-use trends from your smart thermostat and the weekly screen time charts on your phone.

4. Business Intelligence

Business intelligence (BI) is the end game. It leverages analytics tools to convert data to actionable insights. Often paired with data visualization techniques, BI provides decision makers with detailed intel about the state of the business

Data Analytics Tool	**How It's Used**
Artificial Intelligence	Makes decisions that can provide a plausible likelihood in achieving a goal
NoSQL Database	Delivers a method for accumulation and retrieval of data
R Programming	Assists data scientists in designing statistical software
Data Lakes	Accumulates data without transforming it into structured data
Predictive Analytics	Predicts future behavior via prior data
Apache Spark	Generates big data transformation via Python, R, Scala and Java
Prescriptive Analytics	Provides guidance about what to do to achieve a desired outcome
In-Memory Database	Saves time by omitting the requirements to access hard drives
Hadoop Ecosystem	Ingests, stores, analyzes and maintains large data sets
Blockchain	Distributed ledger technologies have proven valuable in managing data challenges
Microsoft Excel	Aggregates data to create reports and easy-to-use dashboards

Different Components of Data Analytics

Generally, there are three stages of data analytics: collection and storage, process and organization, and finally, analysis and visualization. In other words, it starts with identifying the data, then progresses to organizing it in a way that makes sense, and ends with identifying patterns and trends that mean something.

But when it comes to business, we can take these stages a bit further. To start, before we begin sourcing data, we need to engage in some business analytics. We need to ask questions about our objectives and desired outcomes before we identify the type of data we need to gather.

We also need to consider the people and the processes making this analysis happen. Do we need more qualified people? Do we need more training? And how will we share our findings internally and externally?

As businesses are continuing to make digital transformations, the components of data analytics can be seen more as a comprehensive data strategy, with the following components:

1. Address the specific business needs.
2. Determine where the data exists and how it will be gathered.
3. Take inventory of the technical infrastructure needed to support the sourcing of data.
4. Identify how to turn data into actionable insights.
5. Look at the necessary processes and required skillsets of your people.
6. Ensure the right people have access to the right data.
7. Define the business value by creating a roadmap.

Data Analytics Applications

Data in business :

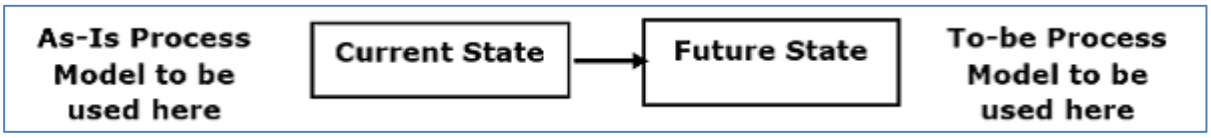
In Data Analytics there are many advantages of data, but without the proper data analytics tools and processes, you can't access these benefits. Raw data is also very important and you need data analytics to unlock the potential of raw data and converted into useful information for the business.
Example –
Record of the potential customer, records of customers like name, address.
Data in healthcare :
Data is extremely useful in this field of medical and healthcare. Most of the medical devices are big data-oriented. In Data Analytics uses of data has gone to such an extent that in the healthcare sector each record or you can say data is very essential where doctors can check person through the heart and temperature monitoring watch which is critical information of any patients and kept to be as data fitted on patient's hand and prescribe him with related medicines.
Example –
Patient records like name, address, contact no. etc., treatment records, Records of Doctor's profile are the examples in healthcare.
Data in media and entertainment :
The business model runs on collecting and creating the content, further analyzing it, then marketing and distribution of the content. We can run through customer's data along with observable data and gather even minute information to create a customer's detailed profile. The benefits of big data in the media and entertainment industry include forecasting what the target audience wants, planning, optimization, expanding acquisition, and retention suggest content on-demand and new.
Example –
Records of the team, the time duration of media project, location, etc.
Data in transportation :
Data in transportation is very crucial. For proper communication and for proper synchronization of transport medium you need data and to analyze the information you need data analytics. Data potential is to analyze how many passengers traveled from any source to destination and with the help of data analytics it can be processed in real-time for the smooth functioning of transportation.
Example –
feedback of customer, transport time, source and destination records, customer traveled history, etc.
Data in banking :
Banking is a very crucial sector. Data here is very beneficial and helps in fraud detection in the banking system. Using big data, we can search for all the illegal activities that have taken place and can identify the misuse of credit and debit cards, business precision, you can say for customer statistics modification, and in public analytics for business.
Example –
Employee records, Bank name address, and branch name, customer account records, transaction history, etc.

Application of Modelling in Business :

Purpose of Business Modelling

Business modelling is used to design current and future state of an enterprise. This model is used by the Business Analyst and the stakeholders to ensure that they have an accurate understanding of the current “As-Is” model of the enterprise.

It is used to verify if, stakeholders have a shared understanding of the proposed “To-be of the solution.



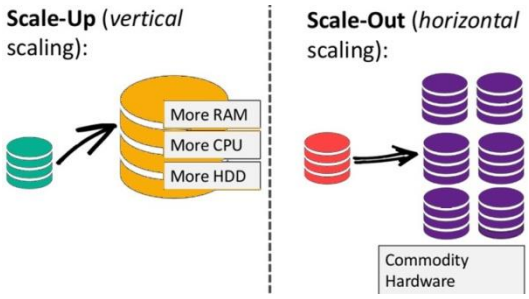
Analyzing requirements is a part of business modelling process and it forms the core focus area. Functional Requirements are gathered during the “Current state”. These requirements are provided by the stakeholders regarding the business processes, data, and business rules that describe the desired functionality which will be designed in the Future State.

Databases & Types of Data and variables

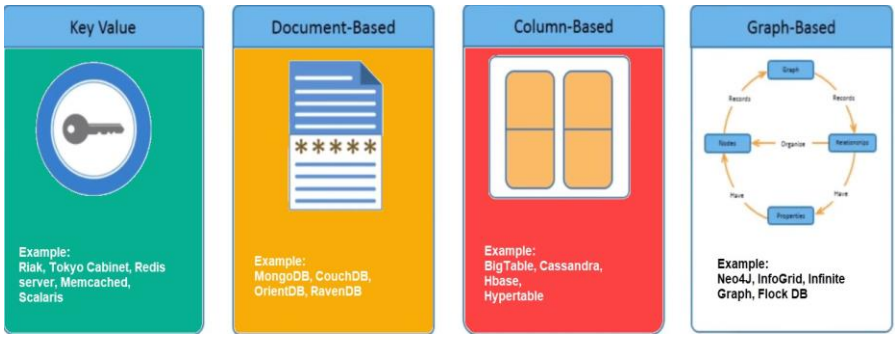
Relational Database Management System: RDBMS is a software system used to maintain relational databases. Many relational database systems have an option of using the SQL.

NoSQL:

- NoSQLDatabaseisanon-relationalDataManagementSystem,thatdoesnotrequire a fixed schema. It avoids joins, and is easy to scale. The major purpose of using a NoSQL database is for distributed data stores with humongous data storage needs. NoSQL is used for Big data and real-time web apps. For example, companies like Twitter, Facebook and Google collect terabytes of user data every single day.
- **NoSQL database** stands for "Not Only SQL" or "Not SQL." Though a better term would be "NoREL", NoSQL caught on. Carl Strozz introduced the NoSQL concept in 1998.
- Traditional RDBMS uses SQL syntax to store and retrieve data for further insights. Instead, a NoSQL database system encompasses a wide range of database technologies that can store structured, semi-structured, unstructured and polymorphic data.
- The concept of NoSQL databases became popular with Internet giants like Google, Facebook, Amazon, etc. whodealwith hugevolumesof data. The systemresponse time becomes slow when you use RDBMS for massive volumes of data.
- To resolve this problem, we could"scale up"our systems byupgradingour existing hardware. This process is expensive. The alternative for this issue is to distribute databaseloadonmultiplehostswhenevertheloadincreases.Thismethodisknown as "scaling out."



Types of NoSQL Databases:



Differences between SQL and NoSQL :

1. SQL	**2.** NoSQL
3. RELATIONAL DATABASE MANAGEMENT SYSTEM (RDBMS)	**4.** Non-relational or distributed database system.
5. These databases have fixed or static or predefined schema	**6.** They have a dynamic schema
7. These databases are not suited for hierarchical data storage.	**8.** These databases are best suited for hierarchical data storage.
9. These databases are best suited for complex queries	**10.** These databases are not so good for complex queries
11. Vertically Scalable	**12.** Horizontally scalable
13. Follows ACID property	**14.** Follows CAP(consistency, availability, partition tolerance)
15. Examples: MySQL, PostgreSQL, Oracle, MS-SQL Server, etc	**16.** Examples: MongoDB, HBase, Neo4j, Cassandra, etc

Data Modelling Techniques in Data Analytics :

Importance of Data Modeling in Data Warehouses

- Improved Data Quality: A well-structured data model helps ensure data consistency, accuracy, and reliability, which are critical for generating meaningful insights.
- Efficient Data Retrieval: By organizing data into logical structures, data modeling enables faster and more efficient data retrieval, which is essential for timely decision-making.
- Scalability: A robust data model allows for easy scaling of the data warehouse as the volume of data grows, ensuring that performance remains optimal.
- Reduced Redundancy: Proper data modeling helps eliminate data redundancy, reducing storage costs and simplifying data management.

Types of Data Models

Data modeling for data warehouses typically involves three main types of models:

1. Conceptual Data Model

- Purpose: Provides a high-level overview of the business entities and their relationships without going into technical details.
- Components: Entities, relationships, and attributes.
- Example: A conceptual model might define entities like "Customer," "Product," and "Sales" and illustrate the relationships between them.

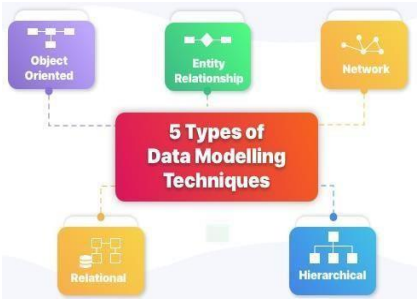
2. Logical Data Model

- Purpose: Represents the logical structure of the data, including the relationships between entities and the data types for each attribute, without considering physical storage.
- Components: Tables, columns, relationships, and constraints.
- Example: A logical model might define tables such as "Customer," "Product," and "Sales" with their respective columns like "CustomerID," "ProductID," and "SaleDate."

3. Physical Data Model

- Purpose: Specifies how the data will be physically stored in the database, including indexing, partitioning, and data storage mechanisms.
- Components: Tables, indexes, partitions, and storage settings.
- Example: A physical model might define storage settings for the "Sales" table, such as partitioning by date to improve query performance

Data Modelling Techniques :



Hierarchical Data Models

Hierarchical data models represent data in a tree-like structure, where information is organized in levels with parent-child relationships. A simplified example is a family tree: the highest level is the parent, followed by branches for children, then grandchildren, and so on. Each element in the hierarchy is called a node, and connections between them are called links.

Hierarchical data models excel at representing data with a natural parent-child hierarchy, providing an intuitive and efficient way to organize, store, and access information.

Relational Data Models

Relational data models, the foundation of relational databases (RDBMS), offer a structured and organized way to represent and manage data. They organize data into relations, also known as tables, where each relation holds records (rows) related to a specific entity or concept.

Relational data models provide a well-established and versatile approach for managing data, particularly for structured and well-defined information. Their advantages make them a popular choice for a wide range of applications across various sectors.

Object-oriented Data Models

Object-oriented data models (OODMs) offer a unique perspective on organizing and managing data, taking inspiration from the principles of object-oriented programming (OOP).

OODMs are similar to graph data models. The main difference is that OODMs focus on individual objects and their internal coherence, with relationships emerging through object interactions.

Need for Business Modelling :

The main purpose of Business Model is to assist the company in developing a plan which will establish and validate critical points of line in the business. This includes activities such as resources, customer relationships, revenue and expenses.

Business Models are important because it helps because

1. The target market is clear

Business Model guides you through the process for determining value preposition and will make you understand how your product can satisfy the customer. The clear and simple business model helps to determine target which will prioritize.

2. The product created is fixed

By following a precise model, there is a lot of clarity in the business model and creating products. The system becomes transparent.

3. Preparing a strategy becomes easier

The business model helps to determine the business strategy automatically. The system does not attract consumers but it teaches how to significantly develop close relationships with producers.

4. Anticipating Competition

Without a business model, the companies will find difficult to find a business position in the market. Due to this the company faces competition. With proper business plans , companies can make strategies for acquiring resources and selling best products to the consumers. So it is very important to determine the right business model.

UNIT – III

Big data technologies and Databases

1. Hadoop

Hadoop is an open source software programming framework for storing a large amount of data and performing the computation. Its framework is based on Java programming with some native code in C and shell scripts.

Hadoop is an open-source software framework that is used for storing and processing large amounts of data in a distributed computing environment. It is designed to handle big data and is based on the MapReduce programming model, which allows for the parallel processing of large datasets.

- HDFS (Hadoop Distributed File System): This is the storage component of Hadoop, which allows for the storage of large amounts of data across multiple machines. It is designed to work with commodity hardware, which makes it cost-effective.
- YARN (Yet Another Resource Negotiator): This is the resource management component of Hadoop, which manages the allocation of resources (such as CPU and memory) for processing the data stored in HDFS.
- Hadoop also includes several additional modules that provide additional functionality, such as Hive (a SQL-like query language), Pig (a high-level platform for creating MapReduce programs), and HBase (a non-relational, distributed database).
- Hadoop is commonly used in big data scenarios such as data warehousing, business intelligence, and machine learning. It's also used for data processing, data analysis, and data mining. It enables the distributed processing of large data sets across clusters of computers using a simple programming model.

2. Requirement of Hadoop Framework

HDFS (Hadoop Distributed File System):

It is a data storage system. Since the data sets are huge, it uses a distributed system to store this data. It is stored in blocks where each block is 128 MB. It consists of NameNode and DataNode. There can only be one NameNode but multiple DataNodes.

Features:

The storage is distributed to handle a large data pool

Distribution increases data security

It is fault-tolerant, other blocks can pick up the failure of one block

MapReduce:

The **MapReduce framework** is the processing unit. All data is distributed and processed parallelly. There is a MasterNode that distributes data amongst SlaveNodes. The SlaveNodes do the processing and send it back to the MasterNode.

Features:

Consists of two phases, Map Phase and Reduce Phase.

Processes big data faster with multiples nodes working under one CPU

YARN (yet another Resources Negotiator):

It is the resource management unit of the **Hadoop framework**. The data which is stored can be processed with help of YARN using data processing engines like interactive processing. It can be used to fetch any sort of data analysis.

Features:

It is a filing system that acts as an Operating System for the data stored on HDFS

It helps to schedule the tasks to avoid overloading any system

Design principle of Hadoop

Open-source – Apache Hadoop is an open source project. It means its code can be modified according to business requirements.

Distributed Processing – As data is stored in a distributed manner in HDFS across the cluster, data is processed in parallel on cluster of nodes.

Fault Tolerance – By default 3 replicas of each block is stored across the cluster in Hadoop and it can be changed also as per the requirement. So if any node goes down, data on that node can be recovered from other nodes easily. Failures of nodes or tasks are recovered automatically by the framework. This is how Hadoop is fault tolerant.

Reliability – Due to replication of data in the cluster, data is reliably stored on the cluster of machine despite machine failures. If your machine goes down, then also your data will be stored reliably.

High Availability – Data is highly available and accessible despite hardware failure due to multiple copies of data. If a machine or few hardware crashes, then data will be accessed from other path.

Scalability – Hadoop is highly scalable in the way new hardware can be easily added to the nodes. It also provides horizontal scalability which means new nodes can be added on the fly without any downtime.

Comparison with other system SQL and RDBMS

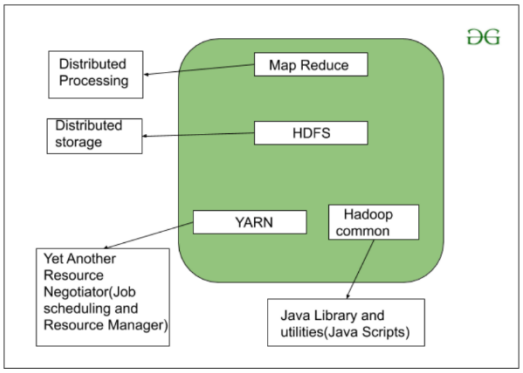
Database management system, as the name suggests, is a management system that is used to manage the entire flow of data, i.e, the insertion of data or the retrieval of data, how the data is inserted into the database, or how fast the data should be retrieved, so DBMS takes care of all these features, as it maintains the uniformity of the database as well does the faster insertions as well as retrievals.

RDBMS on the other hand is a type of DBMS, as the name suggests it deals with relations as well as various key constraints. So here we have tables which are called schema and we have rows which are called tuples. It also aids in the reduction of data redundancy and the preservation of database integrity.

Relational Database Management System is an **advanced** version of a DBMS.

Hadoop Components :

- MapReduce
- HDFS(Hadoop Distributed File System)
- YARN(Yet Another Resource Negotiator)
- Common Utilities or Hadoop Common



1. MapReduce

MapReduce nothing but just like an Algorithm or a data structure that is based on the YARN framework. The major feature of MapReduce is to perform the distributed processing in parallel in a Hadoop cluster which Makes Hadoop working so fast. When you are dealing with Big Data, serial processing is no more of any use. MapReduce has mainly 2 tasks which are divided phase-wise:

In first phase, **Map** is utilized and in next phase **Reduce** is utilized

Map Task:

RecordReader The purpose of *recordreader* is to break the records. It is responsible for providing key-value pairs in a Map() function. The key is actually is its locational information and value is the data associated with it.

- **Map:** A map is nothing but a user-defined function whose work is to process the Tuples obtained from record reader. The Map() function either does not generate any key-value pair or generate multiple pairs of these tuples.
- **Combiner:** Combiner is used for grouping the data in the Map workflow. It is similar to a Local reducer. The intermediate key-value that are generated in the Map is combined with the help of this combiner. Using a combiner is not necessary as it is optional.
- **Partitionar:** Partitional is responsible for fetching key-value pairs generated in the Mapper Phases. The partitioner generates the shards corresponding to each reducer. Hashcode of each key is also fetched by this partition.

Reduce Task

Shuffle and Sort: The Task of Reducer starts with this step, the process in which the Mapper generates the intermediate key-value and transfers them to the Reducer task is known as *Shuffling*. Using the Shuffling process the system can sort the data using its key value.

Once some of the Mapping tasks are done Shuffling begins that is why it is a faster process and does not wait for the completion of the task performed by Mapper.

- **Reduce:** The main function or task of the Reduce is to gather the Tuple generated from Map and then perform some sorting and aggregation sort of process on those key-value depending on its key element.
- **OutputFormat:** Once all the operations are performed, the key-value pairs are written into the file with the help of record writer, each record in a new line, and the key and value in a space-separated manner.

Sr. No.	Key	Hadoop 1	Hadoop 2
1	New Components and API	As Hadoop 1 introduced prior to Hadoop 2 so has some less components and APIs as compare to that of Hadoop 2.	On other hand Hadoop 2 introduced after Hadoop 1 so has more components and APIs as compare to Hadoop 1 such as YARN API,YARN FRAMEWORK, and enhanced Resource Manager.
2	Support	Hadoop 1 only supports MapReduce processing model in its architecture and it does not support non MapReduce tools.	On other hand Hadoop 2 allows to work in MapReducer model as well as other distributed computing models like Spark, Hama, Giraph, Message Passing Interface) MPI & HBase coprocessors.
3	Resource Management	Map reducer in Hadoop 1 is responsible for processing and cluster-resource management.	On other hand in case of Hadoop 2 for cluster resource management YARN is used while processing management is done using different processing models.
4	Scalability	As Hadoop 1 is prior to Hadoop 2 so comparatively less scalable than Hadoop 2 and in context of scaling of nodes it is limited to 4000 nodes per cluster	On other hand Hadoop 2 has better scalability than Hadoop 1 and is scalable up to 10000 nodes per cluster.

Sr. No.	Key	Hadoop 1	Hadoop 2
5	Implementation	Hadoop 1 is implemented as it follows the concepts of slots which can be used to run a Map task or a Reduce task only.	On other hand Hadoop 2 follows concepts of containers that can be used to run generic tasks.
6	Windows Support	Initially in Hadoop 1 there is no support for Microsoft Windows provided by Apache.	On other hand with an advancement in version of Hadoop Apache provided support for Microsoft