UNIT IV

BIG DATA ANALYTICS

- Data analytics life cycle, Data cleaning , Data transformation, Comparing reporting and analysis, Types of analysis, Analytical approaches, Data analytics using R, Exploring basic features of R, Exploring R GUI, Reading data sets, Manipulating and processing data in R, Functions and packages in R, Performing graphical analysis in R, Integrating R and Hadoop, Hive, Data analytics.

Objective

- To understand and apply the Analytical concept of Big data using R and Python.

Big data analytics

- Big data analytics refers to process of collecting ,organizing and analyzing large sets of Data to identify patterns and other important information.
- Helps organization to take strategic decisions.

Reporting and Analysis

- Are they same ????
- Reporting is the process in which data is organized and summarized in an easy to understand format.
- Analysis is process in which data and reports are examined to get insights from them.

Reporting

- A process in which raw data is transformed into useful information
- A tool used to monitor day-to day business operations
- Informational summaries are created in order to monitor how different areas of business performing

Reporting

- Key factors that define a report
  - Provide user the data that was asked for
  - Provide the data in standard and predefined format

Analysis

- An analysis is an interactive process of a person tackling a problem finding the data required to get an answer ,analyzing the data and interpreting the results to provide a recommendation for action.

- Key points that define analysis
  - Provides answers to the question being asked
  - Takes steps needed to get the answers to those questions
  - Customized to specific questions being addressed
  - Involves a person who guides the process

Data Analysis Life Cycle

- Phases
  - Understanding business
  - Understanding Data
  - Data Preparation
  - Modeling
  - Evaluation
  - Deployment

Understanding business

- Determine the business objectives
- Assess Situation
- Determine data mining goals
- Produce Project Plan

Data Understanding

- Collect Initial Data
- Describe data
- Explore data
- Verifying data

Data Preparation

- Select Data
- Clean data
- Construct Data
- Integrate Data
- Format Data

Modeling

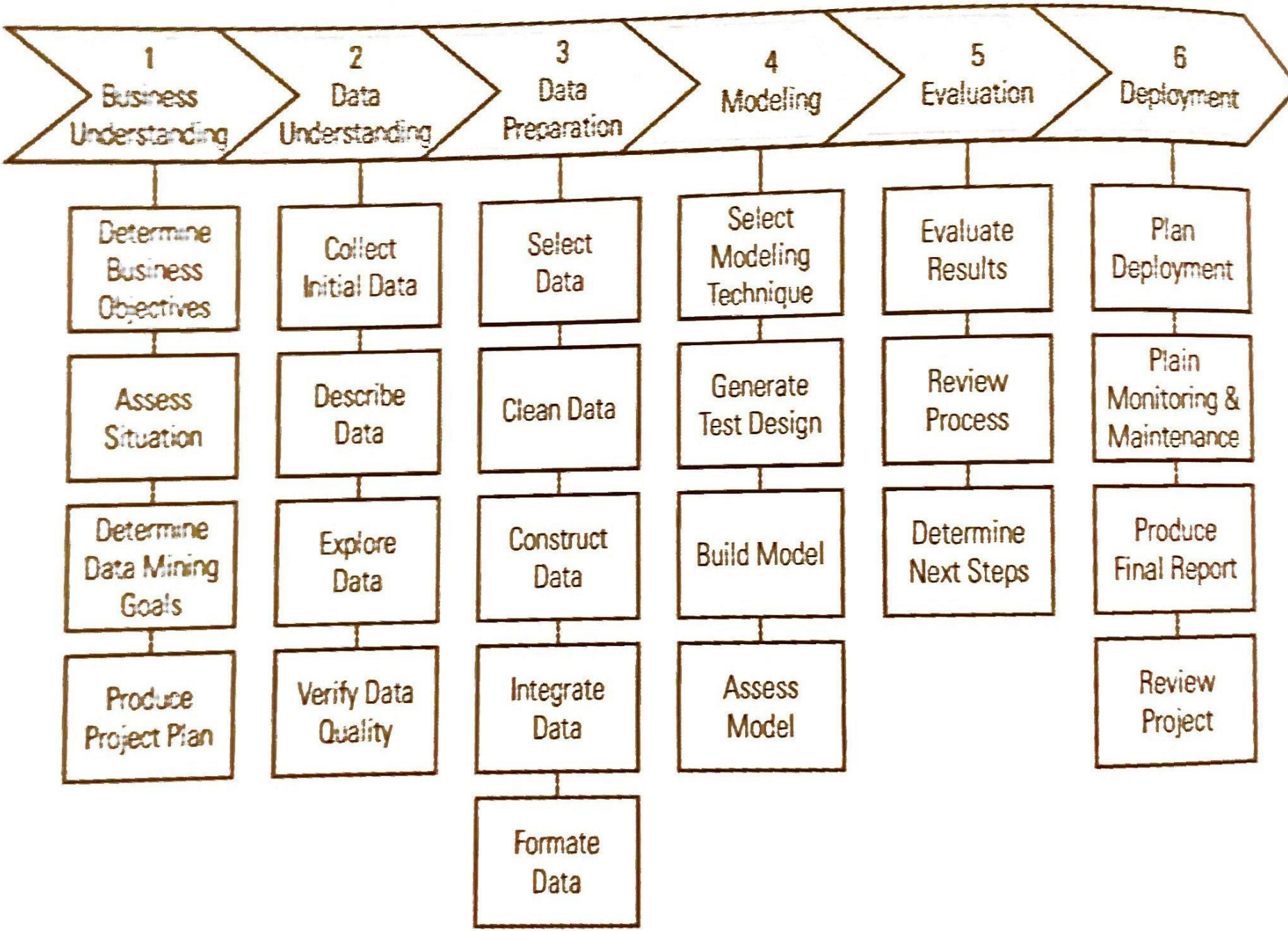
- Select Modeling Technique
- Generate Test Design
- Build Model
- Assess model

Deployment

- Plan Deployment
- Plan Monitoring and maintenance
- Produce Final Report
- Review Project

Evaluation

- Evaluate Results
- Review Process
- Determine Next Steps


1
Business Understanding
2
Data Understanding
3
Data Preparation
4
Modeling
5
Evaluation
6
Deployment
Determine Business Objectives
Assess Situation
Determine Data Mining Goals
Produce Project Plan
Collect Initial Data
Describe Data
Explore Data
Verify Data Quality
Select Data
Clean Data
Construct Data
Integrate Data
Formate Data
Select Modeling Technique
Generate Test Design
Build Model
Assess Model
Evaluate Results
Review Process
Determine Next Steps
Plan Deployment
Plain Monitoring & Maintenance
Produce Final Report
Review Project

Data Cleaning

- The process of altering data in a given storage resource to make sure that it is accurate and correct
- Also known as data cleansing or data scrubbing

- Invalid Values
- Formats
- Missing Values
- Misspelling

Fixing the issue

- Visualization
- Outlier Detection
- Validation code

Transformation

- Bucketing
  - Reduce the effects of minor observation errors
  - Data set is divided into intervals and replaced by categorical values.
- Scaling

Types of Analysis

- Descriptive
  - Describes the data set
- Exploratory
  - Discover previously unknown relationships
- Inferential
  - Test theories related to real world problems.
- Predictive
  - Based on historical data estimating future events

R

- Free open source analytical package
- Features
  - Object Oriented
  - Can be linked with common programming languages like C++,Java
  - Deals with mathematical and complex statistical tasks.

Installing R

- On Ubuntu
  - 1) R on Ubuntu
  - sudo apt update
  - sudo apt install r-base

  - 2) RStudio (IDE used for R) installation
  - From https://www.rstudio.com/products/rstudio/download/#download download tar.gz file for your Ubuntu
  - `sudo tar -zxvf rstudio file name .tar.gz`

- On Windows
  - 1)R on windows
  - Down load for windows from
  - https://cran.r-project.org/bin/windows/base/
  - Click on the exe downloaded
  - Follow the instruction and choose default settings for the installation.
  - 2) RStudio (IDE used for R) installation of Window
  - Download R studio from
  - https://www.rstudio.com/products/rstudio/download/#download
  - Run the .exe file and follow the installation instructions.

Reading data sets

- Reading a file data
  - data=scan(file='test.txt')
  - Data=scan(file.choose())
- Reading large data sets
  - read.csv(file, sep=' ',header=TRUE,row.names)
  - read.csv2
    - sep is by default ;
  -

Pacakges

- Collections of R functions, data in well defined format
- To install a package
  - Install.packages(“name of the package”)
- To load the package
  - Library(package name)