Unit I

- Sources and Nature of Data :-
Data is the Raw facts & figures. By nature, data are either quantitative or qualitative. Quantitative data are numerical and qualitative data are descriptive. Additionally data can also be graphic in nature.

- The source can be of two types :-
→ Statistical :- It refers to data, gathered for official purpose, administrative survey.
→ Non-statistical :- Collection of data for administrative purpose for private sector.
→ Internal source :- records from organisation itself.
→ External source :- data is collected from outside organization

- Classification of data
→ structured – organized, have relational keys, ex :- Relational data (SQL data)
→ semi structured :- have some organized properties, not reside in Relational DB (xml data, Email)
→ Unstructured :- not organized in predefined manner, not have relational database, ex :- word, PDF, media logs
→ meta-data :- data about data

- Characteristics of data
→ Accuracy & Precision
→ Legitimacy & Validity
→ Reliability & Consistency
→ Timeliness & Relevance
- Completeness & Comprehensiveness
- Availability & Accessibility
- Uniqueness

- Introduction to Big Data Platforms.
The aim around four letters
Scalability, Availability, Performance, Security. There are many tools for that such as Data Catalog platform, Data Ingestion Platform IoT Analytics platform.
→ Components of Big data platform :-
- Data Ingestion
- Computing
- Analytics
- Integration
- Data Governance
- Provide Accurate data
- scalability
- Price optimization
- Reduced Latency

Big Data definition :- huge amount of data which is growing exponentially with time.
Ex :- • New York Stock Exchange :- generate 1TB of new data per day.
• Social media :- fb, 500+TB of new data every day.

Types :- Structured, Unstructured, Semi structured → xml file
Data table in SQL
The output return by google

Need of data Analytics

→ Easy access to business information
→ Ad-hoc report Creation
→ Quick identification of Errors
→ Streamlined operational process
→ Details insights Anytime, Anywhere
→ Informed decision making.

Evolution of Analytic Scalability

→ The origins (40s, 50s, 60s)
- operational Research
- optimization problems
- Scheduling & resource allocation
- 24 hrs Compute time

→ The Internet goes global (90s)
- 1995: Amazon
- 1995: EBay
- 1998: Google
- Knowledge Data in DB (1996)
- Analytics :- log queries, data mining
- Operation :- ACID,

→ micro Batch & Event Streaming Analytics (10s, 20s)
- NewSQL (VoldDB)
- Micro-Batch (Spark Streaming)

→ Analytics goes Mainstream (70s, 80s)
- Relational database Born 1972: E.F. Codd relat. DB,
- 1986: first standardized SQL.
- Exploratory Data Analysis

→ The world goes Social (00s)
- Web apps go in hyper-growth
- 2003: LinkedIn
- 2004: Facebook
- Map-Reduce & Hadoop
- The rise of data scientists

→ Fast Data, APIs, mobile & IOT (10s)
- Whatsapp: in a day 31B msgs sent 700m photos
- New problems:- Hadoop is too slow SQL is not enough
- The RAM is the New Disk
- Popular Analytical Stack

Analytic process and Tools

Process :- It is used to examine the varied & large amount of data sets to uncover unknown correlations, hidden pattern, market trends, customer preferences and most useful information which makes & help organization to take business decision based on information.

Tools :- Excel, Tableau, R, python, SAS, Microstrategy, FineReport, IBM Cognos,

Lecture 1

Data Analytics-Introduction

In computing data, information has been translated into a form that is efficient for movement or processing.

Data can exist in a variety of forms as numbers or texts on pieces of paper. Bits and bytes stored in electronic memory facts stored in a person's mind.

Introduction to data

- **Example:**

10, 25, ..., Mathura, 10CSE3002, rk@glbajajgroup.org

Anything else?

- **Data vs. Information**

100.0, 0.0, 250.0, 150.0, 220.0, 300.0, 110.0

Is there any information?

Sources of data

- "Every day, we create 2.5 quintillion bytes of data
  - So much that 90% of the data in the world today has been created in the last two years alone.
  - The data come from several sources
    - sensors used to gather climate information
    - posts to social media sites,
    - digital pictures and videos
    - purchase transaction records
    - cell phone GPS signals

etc.

Examples

Social media and networks

(All of us are generating data)

Scientific instruments

(Collecting all sorts of data)

Sensor technology and networks

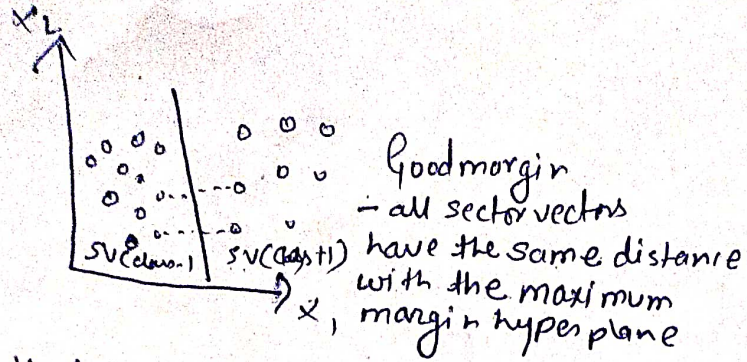
(Measuring all kinds of data)

Analytics

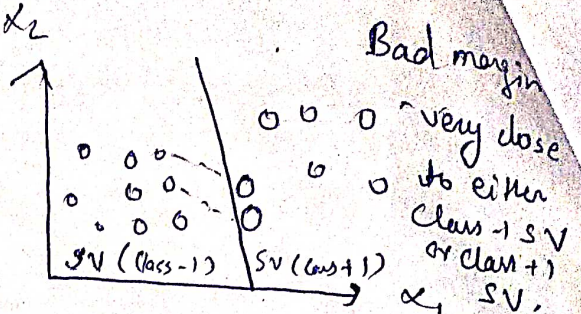
- Analytics is the discovery, interpretation and communication of meaningful patterns in data and applying those patterns towards effective decision making.
- Analytics encompasses multidimensional fields that use mathematics, statistics, predictive modelling, and machine learning techniques to find meaningful patterns and knowledge in recorded data.

What is Data Analytics?

- Data Analytics the science of examining raw data to conclude that information.
- Data Analytics involves applying an algorithmic or mechanical process to derive insights and, for example, running through several data sets to look for meaningful correlations between each other.



Good margin
- all sector vectors have the same distance with the maximum margin hyperplane



Bad margin
- very close to either Class -1 SV or Class +1 SV.

* Analysis of time series

Software used :- tableau, R, pyton

→ Time series :- is a specific way of Analyzing a sequence of data points collected over an interval of time.

→ An Analysts record the data points at consistent intervals over a set period of time rather than just recording the data points randomly.

→ It shows how variable change over time.

→ It is used for forecasting - predicting future data based on historical data.

→ ex :- Schools analyzed five years of student achievement data to identify student progress over time.

→ This analysis is used for non-Stationary data.
- that fluctuate over time
- like finance, retail, economics, stock market
- ex :- trading algorithms, weather data, Rainfall measurement, Temp. reading, Heart Rate monitoring, Interest Rates.

↳ Linear System Analysis :- each data point Y_t, can be viewed as linear combination of past or future values or differences
ex :- things that change slowly, (height of a river measured every hour if there isn't a flash flood). model like this :-

$$x_t = .9 x_{t-1} + .81 X_{t-2} + \ldots + \epsilon_t$$

↳ Non-linear time series :- much more complex, whole host of tools for getting it right with ARIMA proces. because it extract dynamical information about the succession of values in a data set
- ex :- wind speed, water wave, heart rate, stock market, blood pressure

↳ ARIMA models :-
- Univariate model used to better understand a single time-dependent variable. such as temp. over time.
- includes terms to account for moving averages, seasonal differences, operators and autoregressive terms within the model.

- It is used in several industries to allow organizations and companies to make better decisions and verify and disprove existing theories or models. The focus of Data Analytics lies in inference, which is the process of deriving conclusions that are solely based on what the researcher already knows.

The process of Data Analytics

- Analysis refers to breaking a whole into separate components for individual examination.
- Data analysis is a process for obtaining raw data and converting it into valuable information for users' decision-making.
- There are several phases that can be applied
  - Data requirements,
  - Data collection,
  - Data Processing, Data Cleaning,
  - Exploratory data analysis,
  - Modeling and algorithms, data product and communications

Applications of Data Analytics

- **Healthcare**

The main challenge for hospitals with cost pressures tightens is to treat as many patients as they can efficiently, keeping in mind the improvement of the quality of care. Instrument and machine data are being used increasingly to track and optimize patient flow, treatment, and equipment used in hospitals. It is estimated that there will be a 1% efficiency gain that could yield more than $63 billion in global healthcare savings.

- **Travel**

Data analytics can optimize the buying experience through mobile/ weblog and social media data analysis. Travel sights can gain insights into the customer's desires and preferences. Products can be up-sold by correlating the current sales to the subsequent browsing increase browse-to-buy conversions via customized packages and offers. Personalized travel recommendations can also be delivered by data analytics based on social media data.

- **Gaming**

Data Analytics helps in collecting data to optimize and spend within as well as across games. Game companies gain insight into the dislikes, relationships, and the likes of the users.

- **Energy Management**

Most firms use data analytics for energy management, including smart-grid management, energy optimization, energy distribution, and building automation in utility companies. The application is centred on controlling and monitoring network devices, dispatch crews, and

→ Analyzing Vs Reporting

Analysis :- The process of exploring data and reports in order to extract meaningful insights, which can be used to better understand and improve business performance.

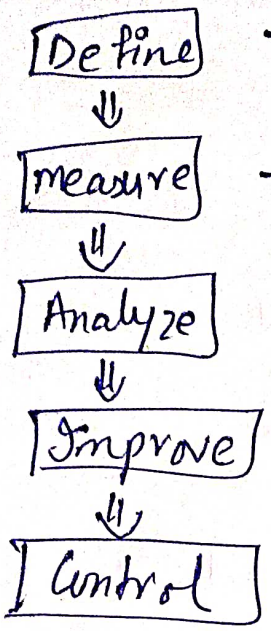
Reporting :- The process of organizing data into informational summaries in order to monitor how different areas of a business are performing.

→ Modern data Analytic tools

- R programming :- It is a leading analytical tool in the industry widely used for statistic and data modeling. It can easily manipulates the data. and presents in different ways.
- Tableau Public
  It is a free s/w that connects any data source be it Corporate data warehouse, Microsoft Excel or web based data and create data visualizations, maps, dashboards etc. They can also be shared through social media or with Client.
- Python: object oriented Scripting language, easy to read, write & maintain and is a free open source tool.
- SAS :- Stastical Analysis System, a s/w for data analysis. & report writing.

Applications :-

Data Analytics Process Showing 5 Steps



→ 100% editable,

Work of Data Analyst

Collect Data
Analyse Data
Create Reports

Role of Data Analytics

→ Gather Hidden Insights
→ Generate Reports
→ Perform Market Analysis
→ Improve Business Requirement

● It is a technique used to Analyze data to enhance productivity & Business gain.

→ Data extracted from sources
→ Get Cleaned
→ Categorized

Types of Data Analytics

1. Descriptive Analysis
2. Predictive Analysis
3. Diagnostic Analysis
4. Prescriptive Analysis

Benifits of Data Analyst

→ Proactivity & Participating needs.
→ Mitigating Risk & fraud
→ Delivering Relative Products
→ Personalisation & Service
→ Optimizing & Improving the Customer Experience

Why Data

→ Define why you need.
→ Collecting data
→ Clean unnecessary data
→ Begin Analyzing
→ Interpret result & Apply them

Tools used in Data Analytics

→ R
→ Python
→ OpenRefine :- known as google Refine, clean messy data
→ Tableau Pubic
→ Microsoft Excel

→ Data Life Cycle ⟶ SEMMA methodology
→ Business Understanding → used for data mining
→ Data Understanding → sample, Explore, Modify, Model, Access
→ Data Preparation →
→ Modeling
→ Evaluation
→ Deployment

R-programming language used
- which is based on S-prog. lang.
- easy to use language
- provide graphical capabilities
- interactive interface for doing statistics

Module I :- Data Gathering & Preparation

Module II :- • Data Cleaning
→ Clean noisy data
→ Correct inconsistencies in data
→ transformation to correct the wrong data
→

• -Consistency
→ usability of data
• Transaction Consistency
• Application Consistency
• Missing data

R-programming lang. based on the name of authors
Robert Gentleman & Ross Ihaka
→ for statical analysis
→ Graphic representation
→ Reporting

Key Roles for Data Analytics Project

Contrary project → 20 people to fulfil roles

1. Business Use:
   - → A user who understand the main area of the project and is also basically benefited from the result.
   - → The user give advice & Consult the team
2. Project Sponsor:
   - → who is responsible to initiate the projects.
   - → provide the funds
3. Project Manager
   - → Person ensure the purpose of project. is met on time & expected quality
4. Business Intelligent Analyst.
   - → Provide business domain perfection based on a detailed & deep understanding of the data.
   - → The person create reports and knows the data feed & sources
5. DB Administrator:
   - → Arrange & Maintain the dB environment to Support Analytic needs
   - → providing permission to key dB or tables. and make sure about the security stages.
6. Data Engineer
   - → Gasps the technical Skills to assist with tuning SQL queries for data management & data extraction.
   - → work with data Scientist to help build data in Correct ways for analysis
7. Data Scientist
   - → It facilitates with the subject matter expertise for analytical techniques data modelling & applying correct analytical techniques for a business
   - → ensures overall analytical objectives are met.
   - → outline and apply analytical methods & process towards the data available for the Concerned project.

Need of Data Analytics Lifecycle:-

- Designed for Big Data Problems and data science Projects.
- Step-by-step methodology
- to organize the activities and task

Phase I :- Discovery

- data science team learn & Investigate the problem
- Develop context & Understanding
-

Phase2 : Data Preparation

- Steps to explove, preprolems
- It requires the presence of an analytic sandbox,
- Server tools used are :- Hadoop, OpenRefine, etc.

Phase 3 : Model Planning

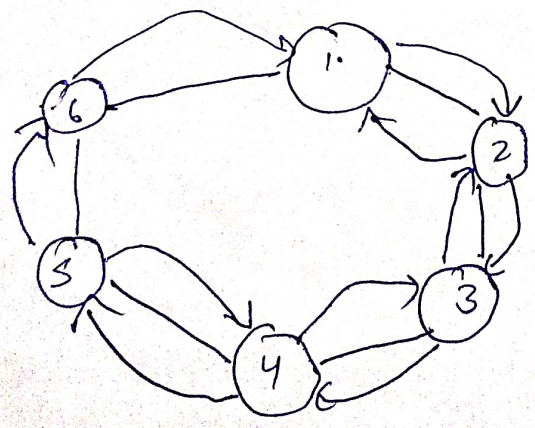
- Team explores data to learn about relationships between variables with most suitable model.
- develop data sets for training, testing & Production purposes.
- Tools used :- Matlab, Stastica

Phase 4 : Model Building -

- Develop datasets for testing, training & Production purpose.
- Free or open source tools Octave, WEKA
- Commercial tools - Matlab, STASTICA

Phase 5 : Communication Results

- After executing model team need to Compare outcomes of modeling to criteria established for Success or failure
-



Phase 6 : operationalize

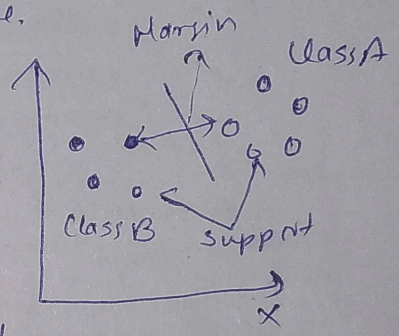
- The team Communicates benefits of project more broadly and sets up project to deploy.
- Free or open Soure tools :- Octave, WEKA, SQL

Unit-II

Support Vector & kernel Methods.

SVM :- flexible supervised Machine learning algorithm which are used both for classification & regression.
→ generally used in classification problem.

Support vector Machine model is basically a representation of different classes in a hyperplane in multidimensional space.
→ The hyperplane will be generated in an iterative manner by SVM so that the error can be minimized.
→ The goal of SVM is to divide the datasets into classes to find a maximum marginal hyperplane.



SVM Kernals:-
SVM algorithm is implemented with kernels that transforms an input data space into the required form.
→ SVM uses a technique called the kernal trick in which kernel takes a low dimension input space and transform it into a higher dimensional space.
→ The following are some of the types of kernel methods used by SVM.

- Linear kernel
  It can be used as a dot product between any two observation
  formula :- $k(x, x_i) = sum(x * x_i)$
- Polynomial kernel
  It is more generalized form of linear kernal & distinguish curved or nonlinear input space.
  formula :- $k(X, X_i) = 1 + Sum(X * X_i) \wedge d$
  Here d is the degree of polynomial, which we need to specify manually in learning algorithm.
- Radial Basis function (RBF) kernel
  RBF kernel, mostly used in SVM classification, maps input space in indefinite dimension space.
  formula :- $K(X, X_i) = exp(-gamma * Sum(X - X_i \wedge 2))$

Gamma ranges from 0 to 1, default value is 0.1