Data Analytics

Date: / / Page no:

Unit I

Statistics :-

The Branch of mathematics that transforms data into useful information for decision makers.

or

statistics is the science of conducting studies to.

- ↳ Collect
- ↳ organize
- ↳ Summarize
- ↳ analyze and
- ↳ draw Conclusion of data.

- Why we Should Study Statistics :

→ Statistics is the Science and also the art of learning from data. As a discipline it is Concerned with the collection, analysis, and interpretation of data, as well as the effective Communication and presentation of results relying on data.

→ Statistics lies at the heart of the kind of quantitative reasoning necessary for making important advance in the Sciences, Such as medicine and genetics, and for making important decisions in business & public policy.

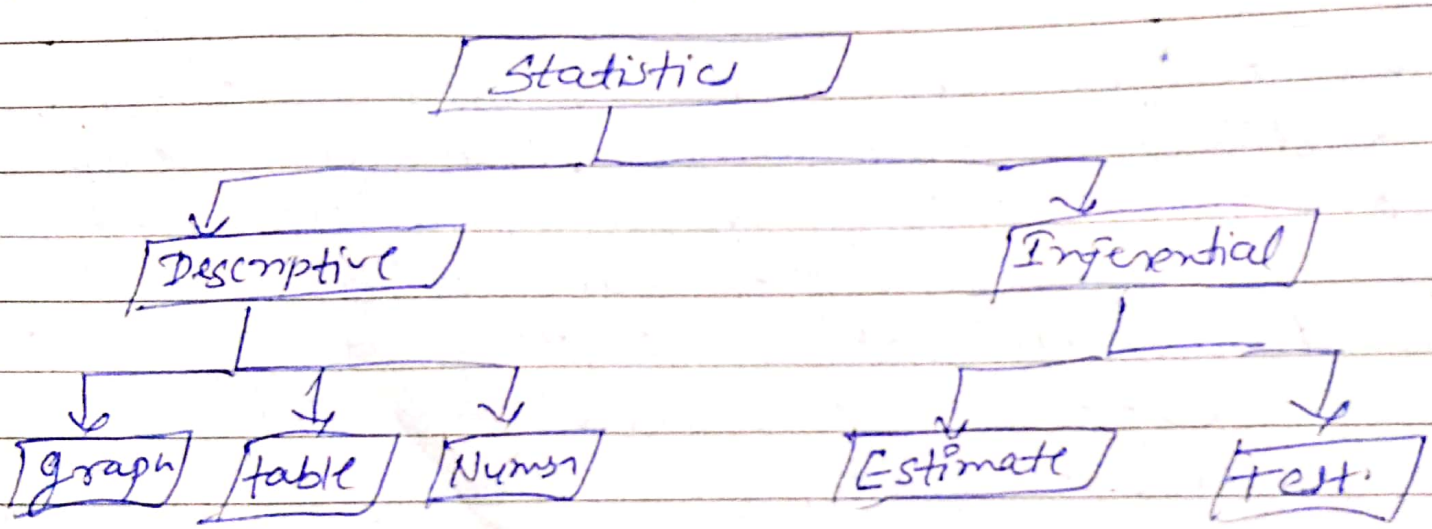
* Statistics is divided into two main areas depending on how data are used.

1) Descriptive statistics:- It describes the data Consists of the collection, organization, Table form summarization, and presentation of data.

Ex:- ~~30% People have A type blood~~ 1000 people died with Cholera in 2015

2) Inferential Statistics:- Consists of conclusion generalizing from sample to population, performing estimation and determining relationship among variables, and making predictions.

Ex:- 30% People have A type blood group.



- In inferential Statistics, the answers are never 100% accurate because the calculation we a Sample taken from the population. This Sample doesn't include every measurement from the population.

hypothesis tig :- Statistic analysis ke liye use karte hai.

Difference :-

Descriptive Statistics	Inferential Statistics
1) It give information that describe the data in some manner.	1) It make inferences about populations using data drawn from the population.
2) Organize, analyze and present data in a meaningful way.	2) Compare data, test hypothesis and make prediction (Estimation of parameters)
3) To describe a situation	3) To explain the chance of occurrence of an event.
4) It explain already known data and limited too sample or population having small size	4) It attempts to reach the conclusion to learn about the population.
5) charts, graphs and tables.	5) Probability.

* Probability :- = $\frac{\text{favorable outcome}}{\text{Total outcomes.}} = \frac{12}{20} = 0.6$

$\frac{8}{20} = 0.4$

Chance

Ex:- class 12 - boys
8 - girls
20 Total Student.

- from 20 student, a student will be selected, what is the prob that he will be boys

Sum of the probility or Total Probability $= 1$

Date: / / Page no:

Probability Distribution:-

- **Variable:-** A variable is a characteristic or attribute that can assume different values. (To vary karta hai)

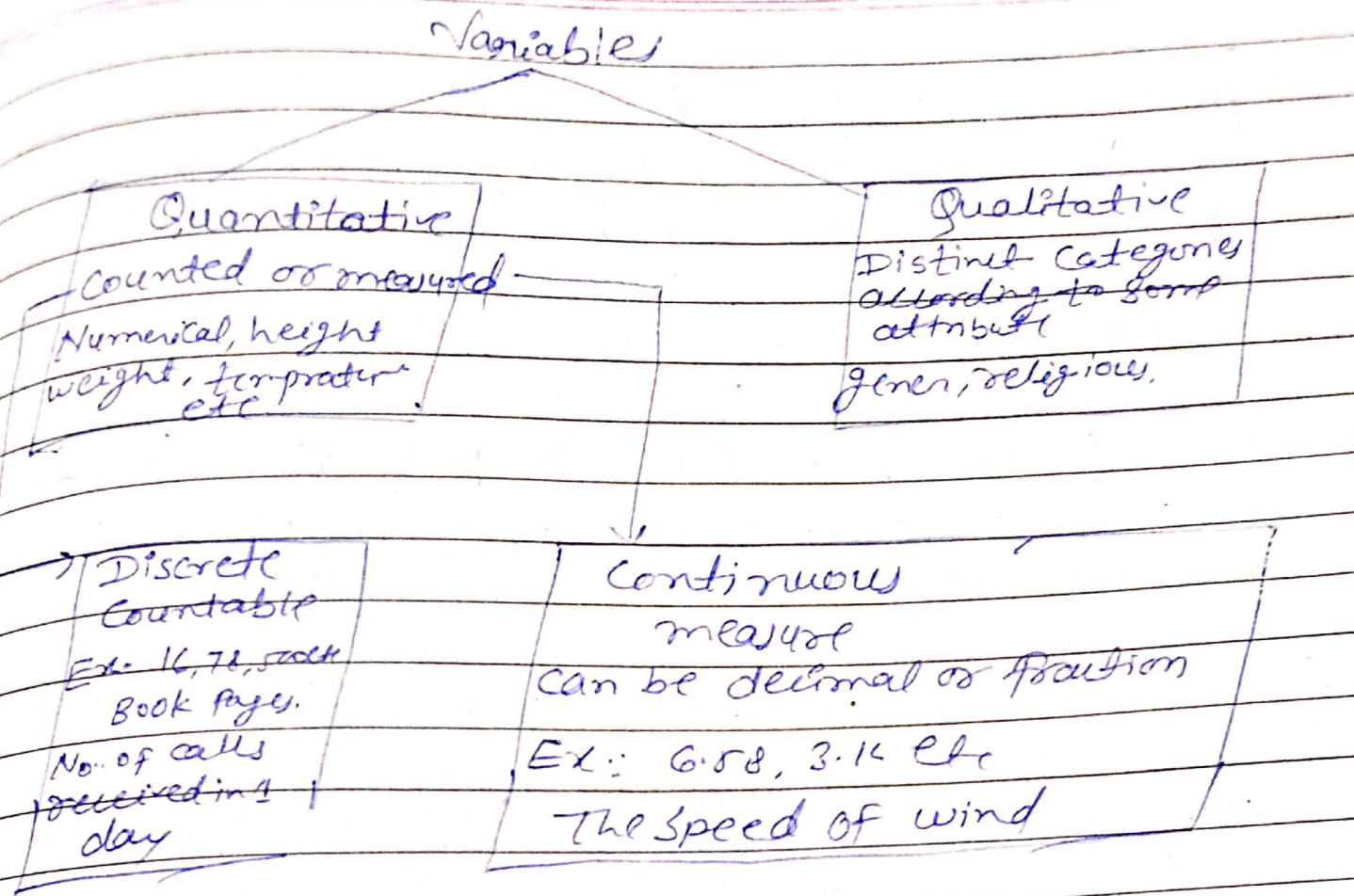
- **Data:-** The values that a variable can assume are called data.

- **Population:-** A population consists of all subjects (human or otherwise) that are studied. (all members of a defined group that we are studying or collecting information on for data driven decision)
  Ex:- Census example

- **Sample:-** A sample is a subset of the population. (A part of the population is called a sample.
  Ex.- may be biased (not confirm not 100%)

- **Random Variable:-** A variables whose values are determind by chance are called random variables.
  Ex:- Automobile insurance, claim suppose 5 every year

Date: / / Page no:



✱ Discrete probability Distribution :-

- A discrete probability distribution consists of the values a random variables can assume and the corresponding probability of the values.

Range of probability 0-1

5th insurance Random Variable
Random Variable ki kya probability hai usko discrete probability

children

Date: / / Page no:

Example:- Construct a probability distribution for a discrete random variable of the following sample space.

3 children

BBB BBG BGB GBB GGG GGB GBG BGG

Solution:- If X is the random variable for the number of girls, then it assume the value 0, 1, 2, 3.

No girls	1 Girls	2 girls	3 girls.
BBB	BBG, BGB, GBB	GGB, GBG, BGG	GGG
1/8	3/8	3/8	1/8

No. of girls X	0	1	2	3
probability (P(x))	1/8	3/8	3/8	1/8

probability $\frac{1}{8} + \frac{3}{8} + \frac{3}{8} + \frac{1}{8}$

$\frac{8}{8} = 1$

probability range = 0 – 1

* What is Hypothesis :-

A hypothesis is an assumption about the relations b/w variable. It is a tentative statement of the research problem or guess about the research outcome. proposing a statement pertaining to relationship b/w the two variables is called a hypothesis. Its validity is usually unknown.

- In my words. A prediction about the outcome of the research in the mind of researcher (Before starting the research work) this is called hypothesis.

* Criteria for hypothesis Construction :-

- It should never formulated in form of question
- It can be testable.
- It should be specific & precise.
- It should describe one issue only.

Type 1 Error :- If Ho is true, even then we are rejected it, then it is said to be type 1 Error.

Date: Page no.

Amtash = Abhishek
height Less.

Type of Hypothesis

Null hypothesis

Ho

$A = B$

Alternative hypothesis.

H_1 or Ha

opposite to Null

> Right tailed test

< Left tailed test

≠ Two tailed test

Type 1 Error

Type 2 Error.

Nothing

Ho

There is no significant difference b/w the mean score of intelligent test of group A & group B

Something

Ha

There is Signifi-

- If null hypothesis is accepted than alternative hypothesis is rejected.
- If null hypothesis is rejected than alternative hypothesis accepted.

Type 2 :- If Ho is false, even then we are accepted it then it is said to be type 2 Error

Date: / / Page no:

✱ Regression Analysis :-

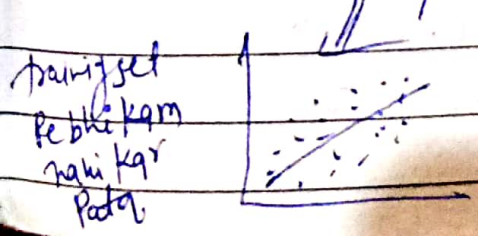
Regression analysis is one of the most used and most powerful multivariate Statistical technique for it infers the existence and form of a functional relationship in a population.

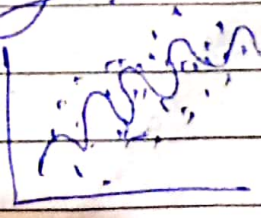
- Regression is a parametric technique used to predict Continuous (dependent) Variable given a set of independent Variables.
- It is parametric in nature because it makes certain assumption based on the data sets. If the data set follows those assumption, regression gives incredible results.

Point

D. Regression analysis is a statistical technique used to describe relationships among Variables.

→ Dependent & Independent Variable
→ outliers.
→ Multicollinearity
→ Underfiting & overfiting :-





Training set pe kam karta but testing time me kam nahi karta

Date: / / Page no:

Ex:- Industry task annual Sale nikalna annual Sale ko affect karne wale hazard factor honge annual Sale depende karta hai factors pe.

Annual Sale → dependent Variable
hazard factor → Independent Variable

- Regression analysis is trying to model the relationship b/w the dependent & Independent ~~and~~ variable.

⇒ Outliers: Suppose mean nikalna hai $\frac{1+2+3+90}{4}$

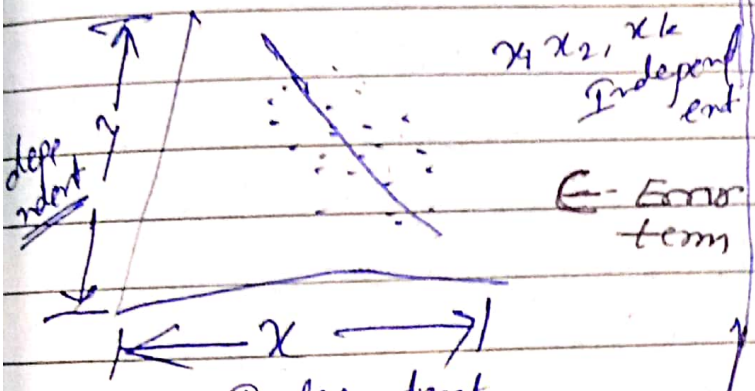
90 odd one out — 90 - outlier. / 90 ke hone se mean ~~the~~ expected aega

Outlier expected result ko wrong direction dete hai.

⇒ Multicollinearity:- Independent Variables when are corelated each other. mean independent Variable when they are sharing non linear relationship with each other (ek ko kuch)

Dependent → outcomes. Independent → Input.
Continuous Numeric Value.

Date / / Page no.

★ Linear Regression	Logistic Regression.
1) Dependent Variable are Continuous in nature. (numerical value)	1) Dependent Variable is binary
2) Linear Relationship. (dependent & Independent relationship)	2) 1 (True, Sucess) 0 (False, Fail)
3) 1 → I & 1 → D. Simple Linear Regression. $y = mx + c$ (slope, intercept) $y = \alpha_0 + \alpha_1 x_i$ (dependent, coefficient of regression, Independent)	3) Goal is to find the best fitting model for I & D Variable relationship.
4) 1 → D & > 1 → I ↓ multiple Linear R.	4) Independent Variables can be continuous or binary.
5) $y = \beta_1 + \beta_2 x_2 t \cdots \beta_k x_k t + \epsilon$ (dependent)	5) Also called logit Regression
	6) Used in machine learning
	7) deal with probability to measure the relation b/w dependent & independent Variable.
x badayenge to y bhi badhega means linear in nature. If x increase then y is increase.	Probability of Support ya ek a backbone use karta hai.

Date: / / Page no:

* Regression is a well known Statistical technique to model the predictive relationship b/w several independent variables (IVs) and one dependent variable.

- The objective is to find the best-fitting curve for a dependent variable in a multidimensional space, with each independent variable being a dimension.

* The key steps for regression are simple
  - List all the variables available for making the model
  - Establish a dependent variable of interest.
  - Examine visual relationship b/w variable of interest.
  - find a way to predict DV (dependent variable) using other variable

* Regression models are simple, versatile, visual/graphical tools with high predictive ability. They include non linear as well as binary predictions. Regression model should be used in conjunction with other data mining techniques to conform the findings.

Date: / / Page no:

Advantages of regression model:-

1) Regression model are easy to understand as they are built upon basic statistical principles such as correlation and least square error.

2) Regression models provide simple algebraic equation that are easy to understand & use.

3) The strength of the regression model is measured in terms of the correlation coefficient and other related statistical parameters.

4) Regression models can include all the variables that one wants to include in the model.

Date: / / Page no:

☆ ANOVA (Analysis of Variance) :-

In t-test we compare only two group.
ANOVA is a statistical method used to compare the mean of two or more groups.
Analysis of Variance is a collection of statistical models used to analyze the difference among group means and their associated procedures (such as 'variation' among and b/w groups) developed by statistician and Evolutionary biologist "Ronald fisher"

- factor (Variable)
- Level.

0mg	50mg	100mg
9	7	4
8	6	3
7	7	2
8	8	3
8	7	4
9	6	3

factor - Dosage.
Levels - 0mg, 50mg, 100mg.

Type of ANOVA

1) one-way ANOVA :- one factor with atleast two levels, (General Linear model (GLM-1)) levels are independent.

0mg	50mg	100mg
9	7	9
8	6	7
7	4	6
6	5	3

2) Repeated-measures Anova:- one factor with atleast two levels, levels are dependent.

Day1	Day 2	Day3
9	7	4
8	6	3
7	6	2
8	7	3

3) factorial Anova:- Two or more factor (each of which with atleast two levels), levels can be either independent, dependent or both (mixed)

gender	Day 1	Day 2	Day 3
		(Three 3 levels)	
men	9	3	5
	8	7	4
	7	6	3
women	7	6	8
	8	3	3
	9	2	2

(2 levels)

factor- gender / Day
↓ Two level ⇓ Three level
Level - male/female / Day1, Day2, Day3.

Date: / / Page no:

✱ Assumptions in ANOVA.

1) Normality of Sampling Distribution of mean:-
The distribution of sample mean is normally distributed.

2) Independence of Errors:-
Errors b/w case are independent of one another.

3) Absence of Outliers:-
outlying scores have been removed from the data set.

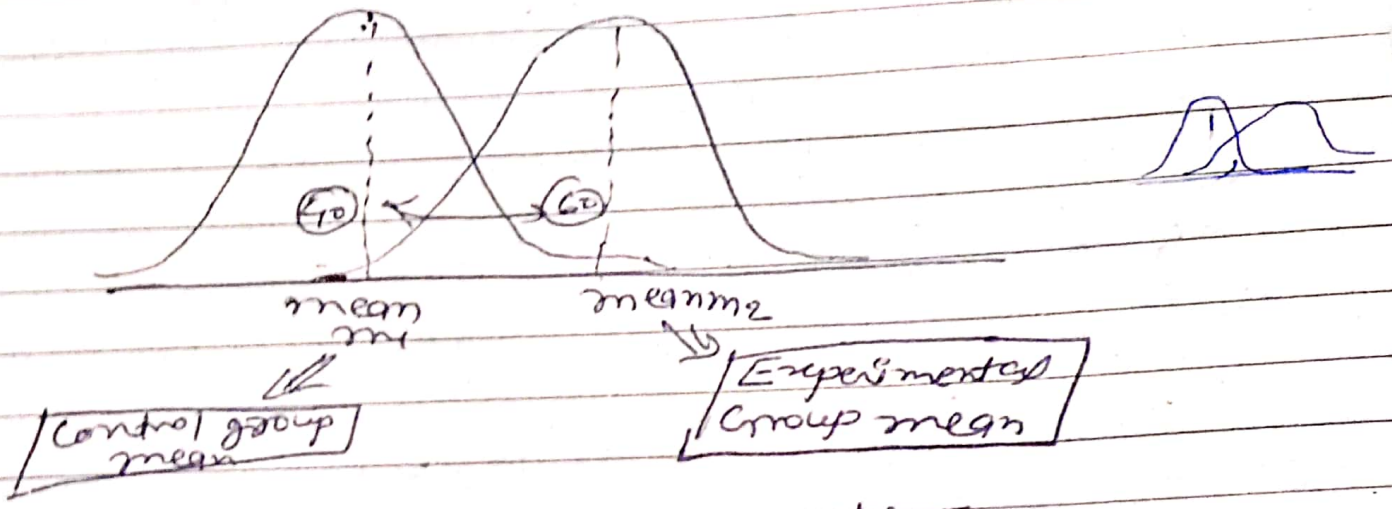
4) Homogenity of Variance:-
population Variance in different levels of each Independent Variable are equal.

✱ Example of ANOVA Testing:-

- A group of different therapies: Counseling, medication and biofeedback. You want to see if one therapy is better than the oth
- A manufacturer has two different process to make light bulbs. They want to know if one process is better than the other.

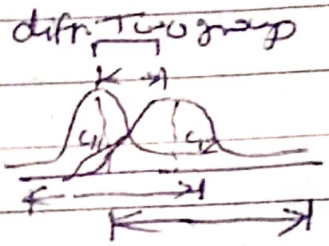
T-Test :

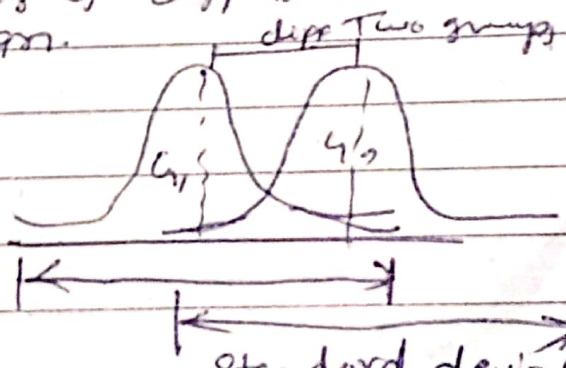
The t-test assesses whether the means of two groups are statistically different from each other.



formula for t test.

$$t = \frac{\text{Difference b/w mean of Two group}}{\text{Standard Error of difference b/w mean.}}$$





$$= \frac{M_1 - M_2}{SED}$$

Unit - II

Date: / / Page no: 1

What is Big data?

- Large amount of data. (Tera, Petta byte)
- Its a popular term used to express exponential growth of data.
- Big data is difficult to store, collect maintain, analyze and visualize.

Store (10GB aasani se, but 10 terabyte difficult) Copy hone me)
Collect (alag alag sources se data ata hai
maintain (Store to kiya hai but maintain thoda bhi idhar udhar] gayatoo
analyze (meaningful information exabtat karna mushkil)
Visualize (bar chart, graph)

- Big data is a collection of data sets so large and complex that it become difficult to process using on-hand database management tools.

The volume of data with the speed it is generated make it difficult for the current computing infrastructure to handle big data. To overcome this drawback, big data processing can be perform through a programming paradigm known as mapreduce.

- With the rapid development of Science & technology there are large amounts of data generated in real life. These big data come from all areas of human life and most of them are stored in computer.

☆ Big data characteristics. (3 V's Concepts).

- Volume :- Large amount of data. → The volume is related to size of data. At present data is in petta bytes and in near future it will be zetta byte.
- Velocity :- The rate at which data is getting generated. The velocity is related to the speed of data coming from different resources.
- Variety :- Different types of data [The variety of incoming data is not limited and is not known.]
  - Structured data Ex:- mysql
  - Semi- Structured Ex- xml
  - Unstructured data Ex. text, audio, video

☆ Big data Source :- There are several source of data including some new ones. Data from outside the organization may be incomplete and of a different quality & accuracy.

1) Social media (
2) Banks.
3) Instruments (Camera)
4) websites. (amazon)
5) Stock market (

☆ Use Cases of Big data.

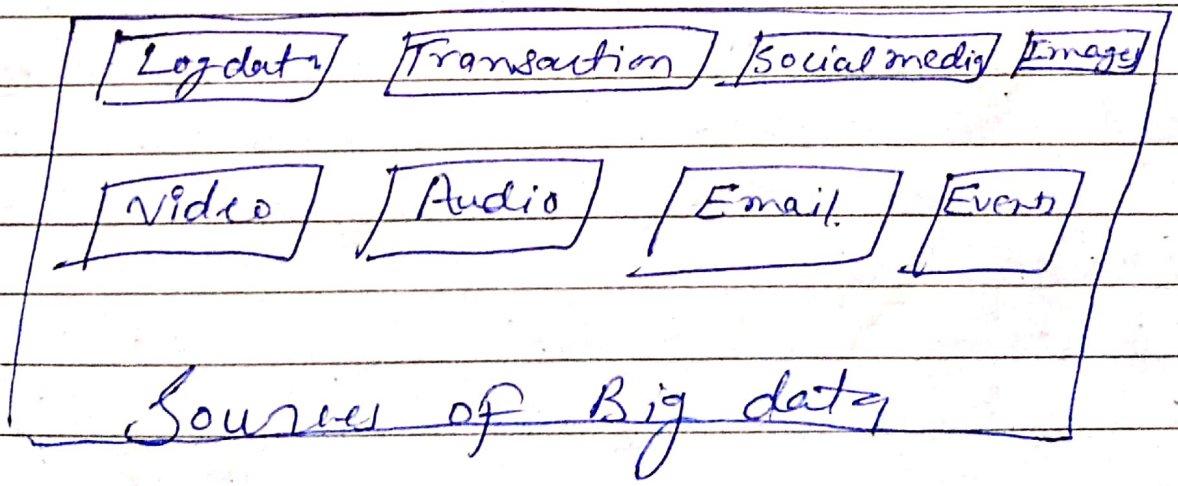
- Recommendation engine (amazon product search next time Re appearing)
- Analysing Call Detail Record (CDR) [Call center) telecom co.
- Fraud Detection [credit card fraud, online Bank)
- Market Basket Analysis [Kisi saman ke sath chijo ke sath rakhe
- Sentimental Analysis [Text data ko analyse Ek mudda Utha kar Social media par kar]

Date: / / Page no:

☆ There are Two types of Big data.

1) Structured Data :- are no. and words that can be easily categorized and analyzed. These data are generated by things like network sensors embedded in electronic devices, Smart phones and global positioning System (GPS) devices
→ Structured data also include thing like Sale figures, account balance, and transaction data.

2) Unstructured Data :- Include most complex information, Such as Customer reviews from commercial websites, photos and other multimedia, and Comments on Social networking sites. These data cannot easily be Seprated into Categories or analyzed numerically.



Sources of Big data

★ ~~Drivers for Big data:-~~
★ Why is Big data important?

→ The importance of big data doesn't revolve around how much data you have, but what you do with it. You can take data from any source and analyze it to find answers that enable
1) Cost reductions,
2) time reductions.
3) New product development and optimized offerings. ~~and~~
4) Smart decision making.

Sources of Big data:-

a) Social media:- All activities on the web and social media are considered Store and are accessible. Email was the first major source of new data. google searches, facebook posts, tweets, youtube vidos ~~etc~~

b) organizations:- Buiness organizations & government are a major sources of data. E-Commerce systems, user-generated content, web access logs and many other sources of data generate valuable data for organization.

~~c) machine:-~~

Analyst:- collect data + analyze data = Report.
Various Source valiar- expert

Date: / / Page no: 1

✱ What is analytics:-

Analyzing or deep understand and Something using data.

→ Why to do analytics

[Business] [marketing]

[Critical Problems] [New Ideas]

- Gather Hidden ~~information~~ insights.
- Generate Reports.
- Perform market Analysis.
- Improve Business Requirement.

✱ Data Analytics:-

Data analytics refer to the techniques to analyse data to enhanced productivity and business gain

[Business administration] + [Exploratory data Analysis] =

[Growth in Business]

Past → future

[Descriptive analytics or observe] ⟹ [Predictive Analysis or Predict] ⟹ [Prescriptive analysis or Inferential]

Important of Bi

Date: / / Page no:

* Need for big data analytics:-

→ cost reduction
→ fast & Better Decision making
→ New products and Services.

* Big data Analytics Application. :-

1) Health Care :- E-medical, purchase data, Research data, Transactional

2) Eduction sector :→ Improving evaluation of Student Results.
- Analysing and Creating the Custom program.
- Computing the marks of Students.

3) E-Commerce :- Retailers]
- Predict Trends.
- optimize pricing
- forcast demand.
- personalized Stores
- Customer Servie.
- Sales generation.

4) Government :- E-Government
- Traffic optimization
- cyber Security & Intelligence
- Crime prediction & prevention.

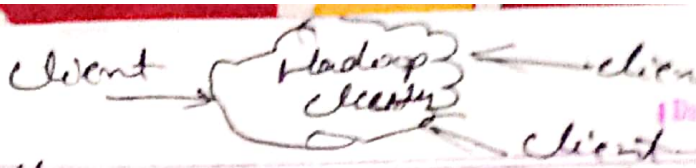
- weather forecasting
- Drug Evaluation
- Tax Compliance.

5) IOT :- - Retail
- Vehicles
- Industry
- Communication
- medicare.

Smart city concept.

6) Media & Entertainment :- - media Scheduling
- Audience intrests.
- Ad targetting

* Dat Big data Analytics :- Big data analytics is the often complex proces of examining large and varied data sets or big data to uncover information including hidden patterns, unknown correlations, market trends and customer preferences that can help organizations make informed business decisions.



"Hadoop is an open source software framework which huge data storage facility."

✡ Hadoop :-

Hadoop is an open source, Java based programming framework that supports the processing and storage of extremely large data sets in a distributed computing environment.

- Hadoop is capable to work on clusters (thousand of interconnected machines) and handle to thousand of terabytes data
- Hadoop provides rapid data transfer rate
- It is developed by doug cutting and mike cafarella in 2006.
- It was inspired by google's mapReduce programming framework → GFS (google file System)

Component present in Hadoop

- Name node (NN)
- Job tracker (JT)
- Secondary Name node (SNN)
- Data node (DN)
- Task tracker (TT)

- All this are daemony, which are present in Program (necessary)

3) YARN (Yet another resource Negotiator) :- Jobtracker Task Tracker machine fail ho jaye to uska load kisi dusre machine pe dena resource manage karna

✱ HDFS (Hadoop Distributed file System) :-

HDFS architecture :- It have 2 types of Nodes.

1) NameNode.
2) DataNode.

It works on master slave concept :- one master multiple slaves

one name node multiple Data node

1] Name node :-

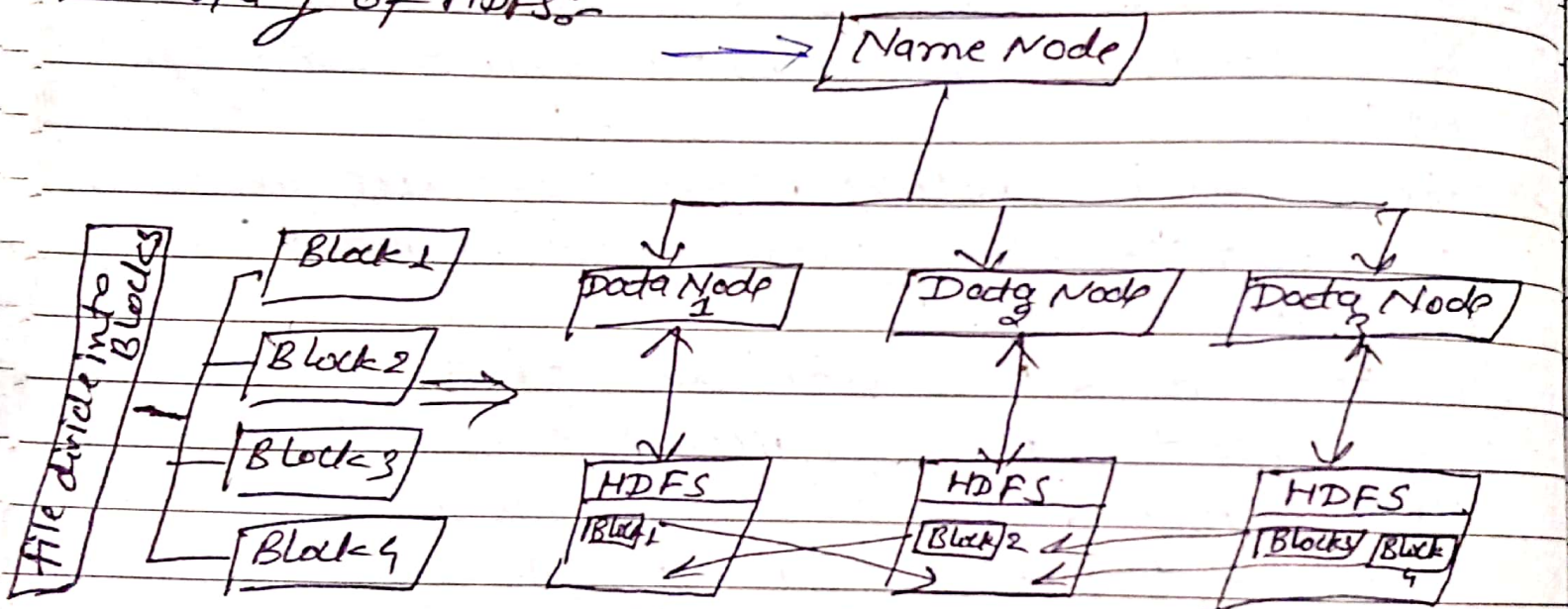
- Name node is also known as the master.
- Name node only stores the metadata of HDFS.
- Name node does not store the actual data or data set. The data stored in data nodes physically
- Name node knows the list of the blocks and its location for any given file in HDFS. With this information Name node knows how to Construct the file from blocks.
- Name Node is a Single point of failure in Hadoop clusters.
- NameNode is usually configured with a lot of memory (RAM). Because the block locations are help in main memory

(Slave)

2) Data Nodes:-

- Data Node is responsible for storing the actual data in HDFS.
- DataNode is also known as the slave.
- NameNode and DataNode are in constant communication.
- When a DataNode is down, it does not affect the availability of data or the cluster.
- Namenode will arrange for replication for the blocks managed by the dataNode that is not available.

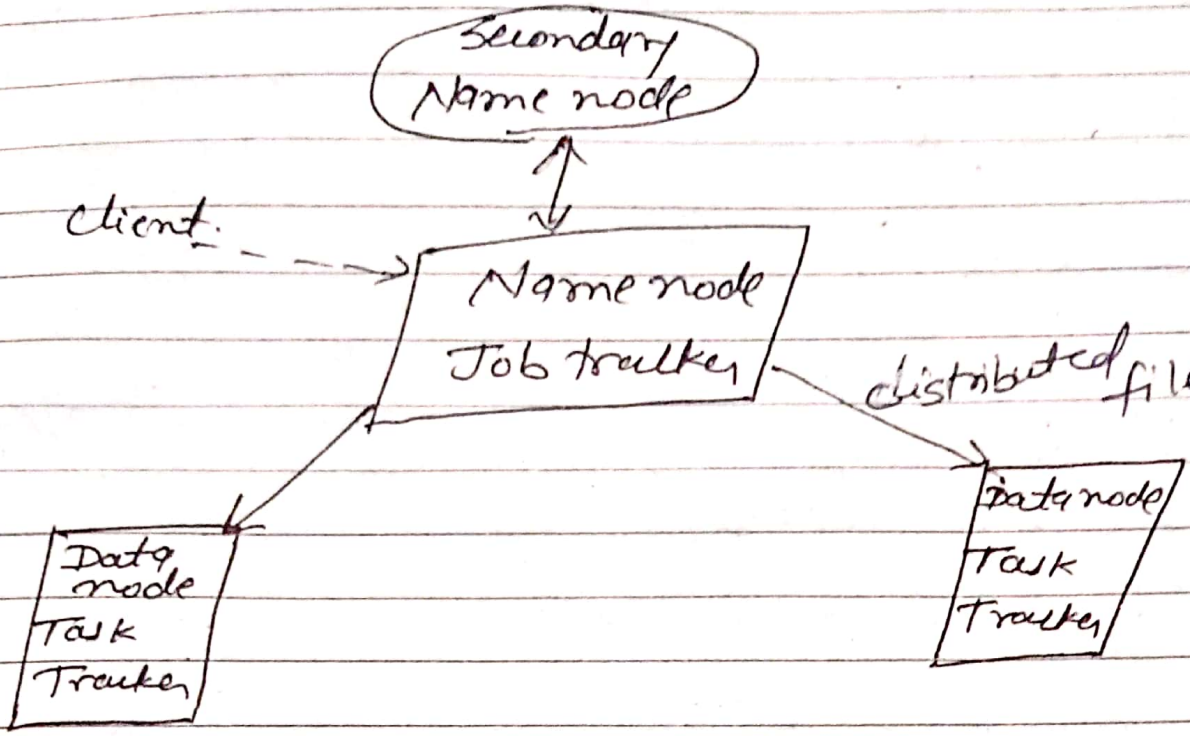
* working of HDFS:-



Storage & Replication of Block in HDFS.

Date: / / Page no: 1

* physical Architecture of hadoop:-



Process:- - client provides its Job request to hadoop
- This request is accepted by the name node.
- Name node is master in Hadoop.
- It also contains a Job tracker which is again a part of master.
- This Job is divided into tasks and tracker provide it to the data node.
- Now the data node is a slave and it passes task tracker which actually performs the task
- And Job tracker continuously communicates with task tracker and if anytime it fail to reply then it assumes that the task tracker may have crashed.

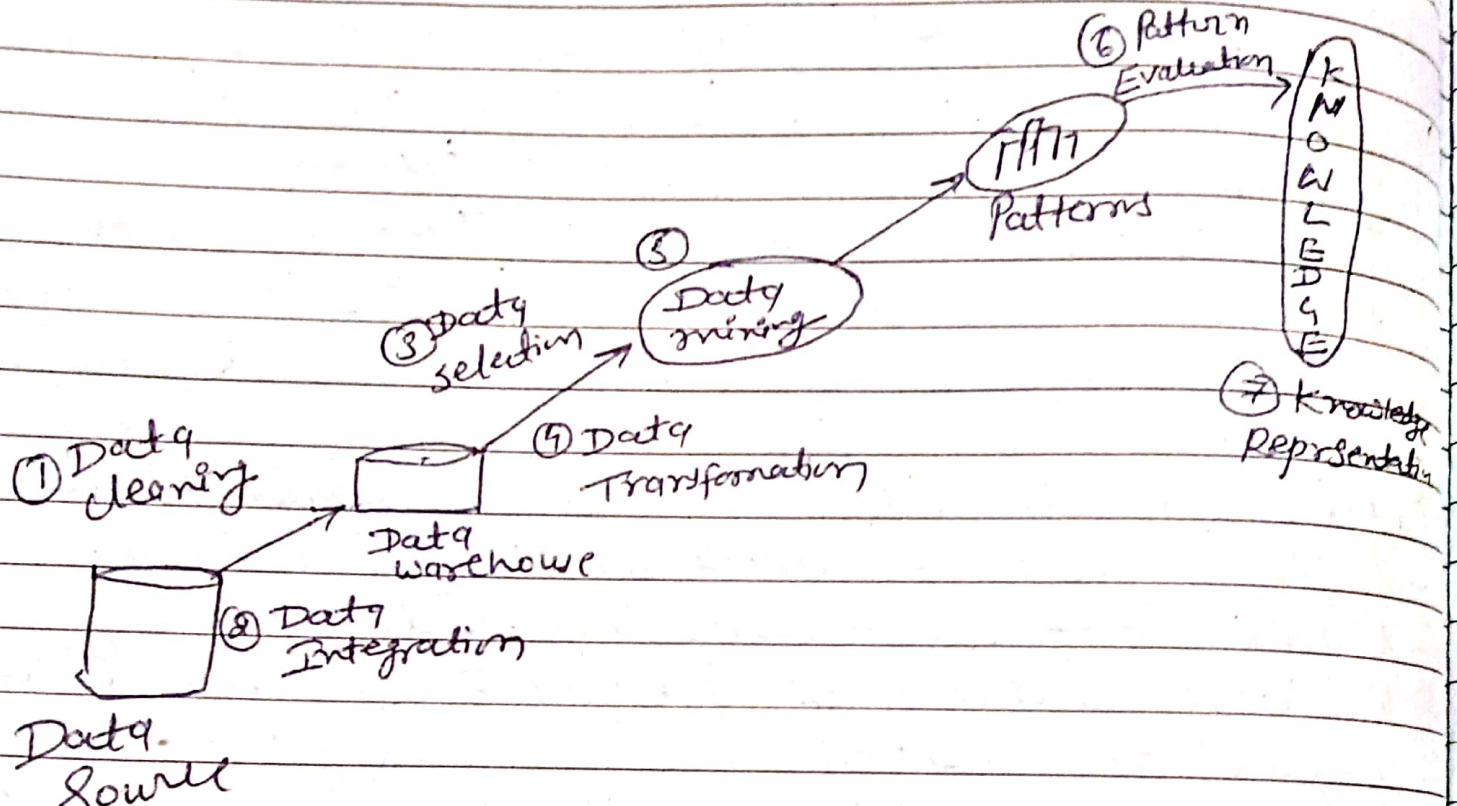
Useful data konikalna.

Date: / / Page no:

✱ Data Discovery :-

Data discovery A combination of processes and technology that enables the detection of patterns in data.

KDD :- Knowledge Discovery from Data :-



- Data cleaning :-
  → To remove noise and inconsistent data example parsing the data.
  → cleaning is performed for detection of syntax error.

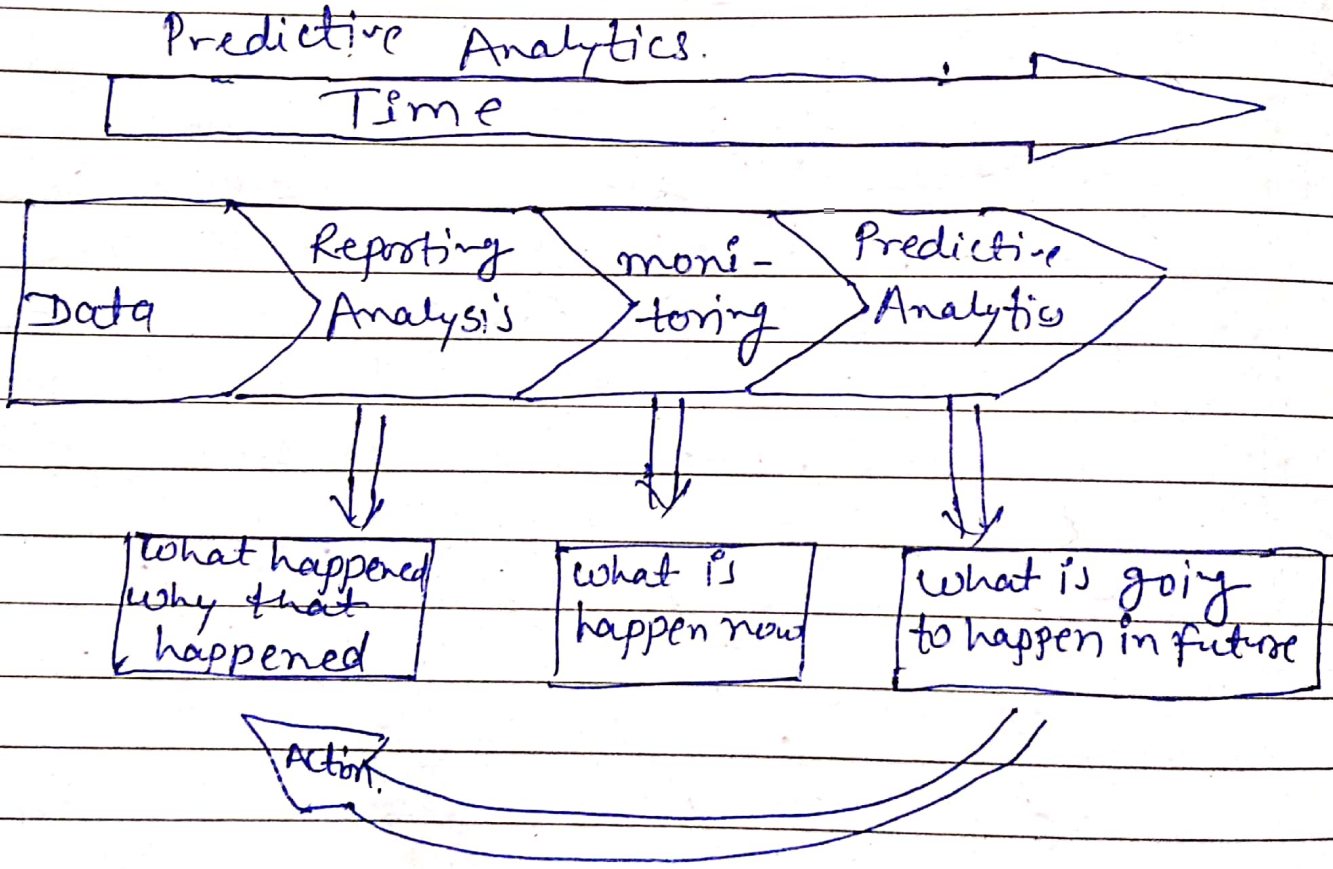
Date: / / Page no:

- Data Integration :- where multiple data Source are combined.
- Data selection:- where data relevant to the analysis task are retrived from the database.
- Data Transformation:- where data are transformed or ~~Consolidated~~ consolidated (~~ESSE~~) int form appropriate for mining by performing summery or aggregation operation for instance.
- Data mining:- An essential process where intelligent method are applied inorder to extract data patterns.
- Pattern Evaluation:- To identify the truely intresting patterns representing knowledge base on some interresting measures
- Knowledge Representation:- where visualization & knowledge representation technique are used to present the mined knowledge to the user.

✡ Predictive analytics :-

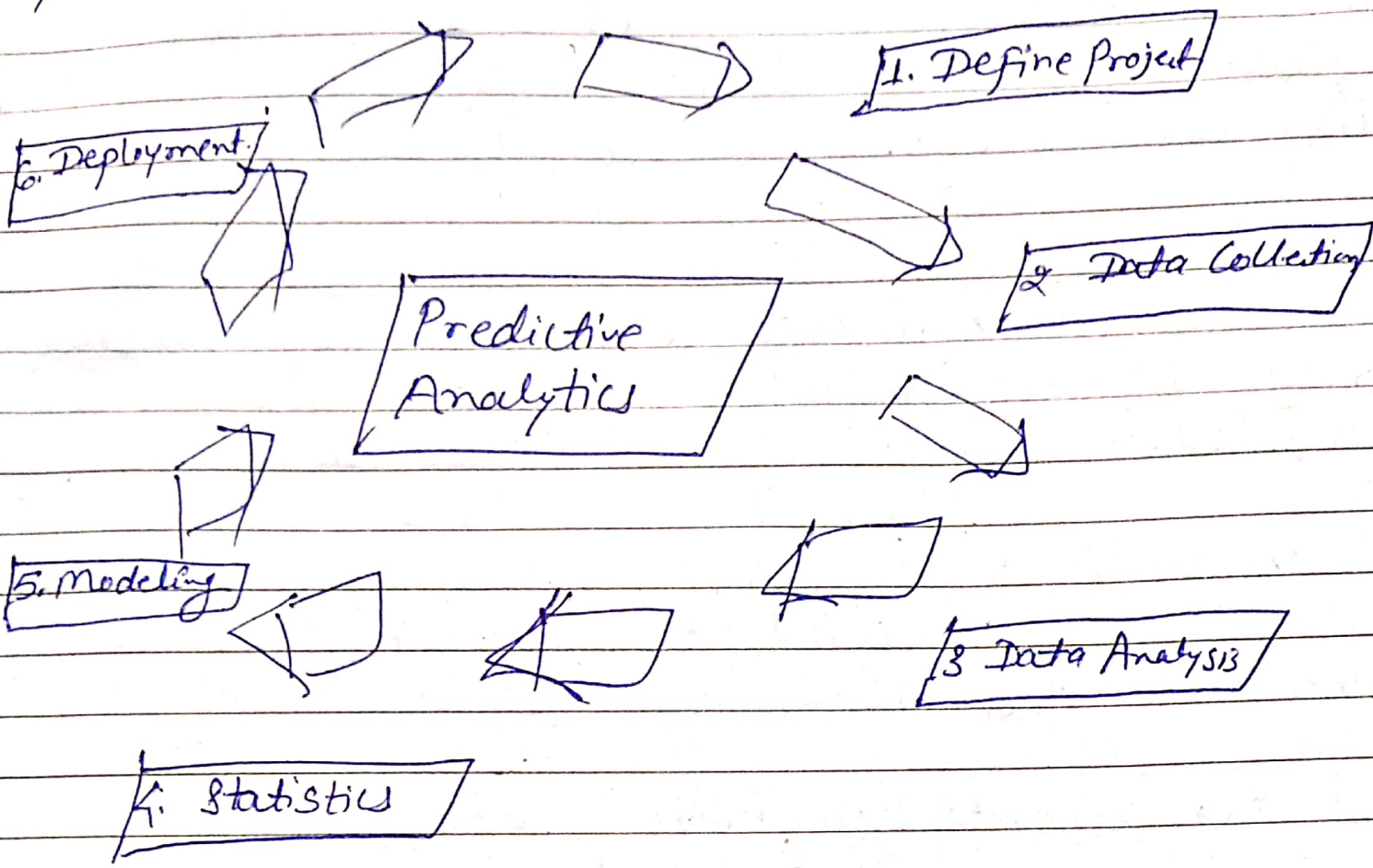
Predictive analytics is the branch of the advanced analytics which is used to make prediction about unknown future events.

- predictive analytics use many techniques from data mining, statistics, modeling, machine learning and artificial intelligence to analyze current data to make prediction about future



[predictive Analytics Value chain]

⇒ Predictive analytics process :-



1) Define project :- Define the project outcomes, deliverables, Scoping of the effort, bwines objectives, identify the data sets which are going to be wed.

2) Data Collection :- Data mining for predictive analytic prepares data from multiple Sources for analysis. This provides a complete view of the customes intractions.

Date: / / Page no:

3) Data Analysis:-
Data analysis is the process of inspecting, cleaning, transforming, and modeling data with the objective of discovering useful information arriving at conclusions.

4) Statistics:- Statistical Analysis enables to Validate the assumptions, hypotheses and test them with using Standard Statistical models.

5) Modeling:- predictive modeling provides the ability to automatically create accurate predictive model about future.

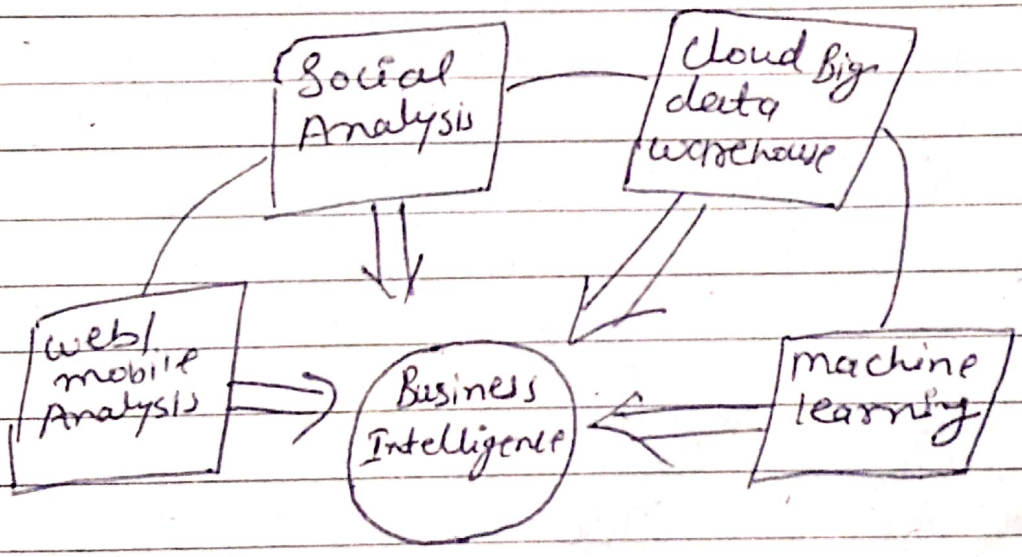
6) Deployment:- predictive model deployment provides the option to deploy the analytical results into every day decision making process to get results.

~~7) model~~

★ Mobile Business Intelligence :-

Mobile Business Intelligence (BI) is the ability to access BI-related data such as KPIs, business metrics, and dashboards on mobile devices.

- Mobile BI is a system comprising both technical and organizational elements that present historical and/or real-time information to its user for analysis on mobile devices such as smartphones and tablets (not laptop), to enable effective decision making and management support, for the overall purpose of increasing firm performance.



(Date: / / Page no:

- The definition of mobile BI Refers to the access and use of information via mobile device. with the increasing use of mobile devices for business not only in management positions, mobile BI is able to bring business intelligence and analytics closer to the user when done properly.
  - whether during a train journey, ~~in the airport departure~~ or during a meeting break information can be consumed almost anywhere and anytime with mobile BI.

★ Benefits of mobile BI:-

1) The first major benefits is the ability for end users to access information in their mobile BI System at anytime and from any location.

2) Improves their daily operations

3) React more quickly.

4) Speeds up the decision making.

5) Better time utilization.

6)

Date: / / Page no:

BI	Data Science
1) Backward approach.	1) Forward approach.
2) Provides reports but that can't predict future.	2) Provides patterns and reports ~~that state~~ how the data will look in future.
3) Preplanned and restricted data → (No explination)	3) No-preplanny data is retrived from source as per requirement.
4) Helps to answer the quations you know	4) Help you to discover new quation.
5) Busines user.	5) Data Scientists.
6) What happend?	6) What will happen?
7) Input: past data output: present Solution	7) Input: Dynamic requisite data output: future predict.

~~Information~~ technology:

Main frame computing → Mini computing → personal computing.

Desktop internet → mobile internet.

- Mobile BI is defined as "The capability that enable the mobile workforce to gain busines insights through information analysis using application optimized for mobile devices"

Date: / / Page no:

☆ What is Crowdsourcing:-

Crowdsourcing is the process of getting work or funding, usually online, from a crowd of people.

- ~~The word~~ It is a combination of the two words 'crowd' and 'out-sourcing'. The idea is to take work and outsource it to a crowd of workers.

~~Crowdsourcing & quality~~: The priple of Crowdsourcing is that more heads are better than one.

- By canvassing (वोट मांगना) a large crowd of people for ideas, skills, or participation, the quality of content and idea generation will be superior.
- The idea is generally to introduce new or more developed skill sets or a larger work force to achive some specific goal.

☆ Benefits of Crowdsourcing:-

Crowdsourcing is a powerful buiness marketing tool as it allows an organization to leverage the creativity and resources of its own audience in promoting and growing the company for free.

→ Crowd Sourcing increase the productivity of a company while minimising labor expenses
→ The internet is a time proven strategy for Soliciting feedback from an active and passionate Consumer base.

☆ Inter & Trans-firewall analytics:-

Over the last 100 years, Supply chain has evolved to connect multiple companies and enables them to collaborate to create ~~enormous~~ Value to the end-consumer through concept like

→ CPFR (Collaborative planning, forecasting and Replenishment):- A Collection of busines pratice that leverage the internet and electronic data interchange to reduce inventories and expenses while improving Customer Services.

→ VMI (Vendor managed Inventory):- A technique used by customers in which manufacturers receive sales data to forecast Consumer demand more accurately.

- There are instances where a retailer and a social company can come together to share insights on consumer behaviour that will benefit both concerns.
- Some of the more progressive companies will take this a step further and work on leveraging the large volume of data outside the firewall such as social media, location data etc.

→ In other words, it will not be long before internal data and insight from within the enterprise firewall is no longer a differentiator. We call this trend the move from intra to inter and trans firewall analytics.

* Information management:- Is the process of collecting, storing, managing and maintaining information in all its forms.

- Information management may also be called information asset management.
- Information can be in the form of physical data (such as papers, documents and books) or digital data assets.

(Big data technologies)

Date: / / Page no:

- Unstructured Information management Architecture (UIMA): This is one of the elements in the 'Secret Sauce' behind IBM Watson's System that reads massive amounts of data and organizes for just in time processing.

- Watson beat the quiz program champion in 2011 and is now used for many business application, like diagonosis in healthcare situations. Natural language processing is another capability that helps extend the power of big data technologies.

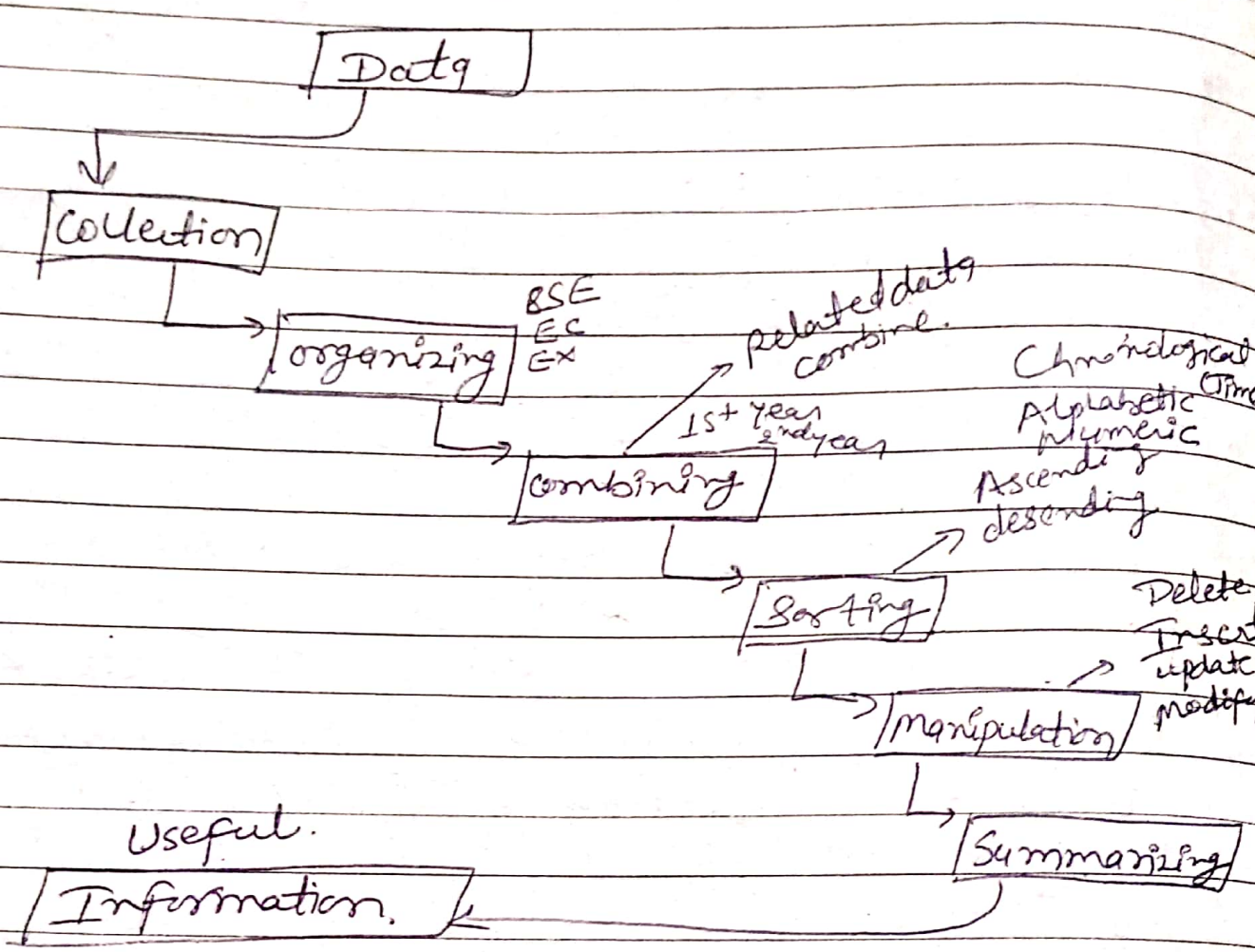
(work on Concept human being dekhaty hai, samajata hai than deside best result. Virtual assistance kai bhi sawal ka sahi Jawab dega.

Unit - III

Date: / / Page no:

☆ Processing Big data:-

- Data processing:-



dissimilar / different

Date: / / Page no:

* Integrating disparate data Stores :-

- Data integration :- Data integration involves combining data from several disparate sources, which are stored using various technologies and provide a unified view of the data.
- Data integration become increasingly important in case of merging systems of two companies or consolidating applications within one company to provide a unified view of the company's data assets. The later initiatives is often called a data warehouse

* Data Integration areas :- data integration is a term covering several distinct sub-area such as

- Data warehouse
- Data migration (Data migration is the process of transfering data from one system to another)
- ~~Enterprise application / Information integration~~
- master data management (MDM represents a set of tools and process used by an enterprise to consistently manage their non-transactional data)

Date: / / Page no

* Data lake Storage on aws: The most Secure, durable, and Scalable Storage to build your data lake through Amazon S3 (Simple Storage Service).

- Amazon S3 :- is the largest ~~and most perfor~~ Storage Service for Structured and unstructured data and the Storage Service of choice to build a data lake with amazon S3, you can cost-effectively build and Scale a data lake of any size in a Secure environment where data is protected by 99.9% of durability.

- ~~with a data lake~~
- with a data lake built on Amazon S3, you can use native AWS Service to run big data analytics, AI, ML, high performance computing (HPC) and media data processing applications to gain insight from your unstructured data sets.

- In the modern data marketplace, disparate data Sources are largely what we refer to as unstructured in nature, making up the bulk of "big data" Volumes. Databases, data warehouses, and data lakes are all

governed in unique ways. Hadoop brings different data types together in one place but does not guarantee any substantive forms of organization. (original data)

* Mapping data to the programming framework :-

- The MapReduce programming framework use two tasks common in functional programming: Map and Reduce. MapReduce is a new parallel processing framework and hadoop is its open-source implementation on a single computing node or on clusters.

- Big data processing can be performed through a programming paradigm known as Map-Reduce. MapReduce is a programming model designed for processing large volumes of data in parallel by dividing the work into a set of independent tasks.

- It is based on "break problem up into smaller sub-problem" strategy and can be exposed via SQL & SQL-Based BI tools.

- MapReduce is a programming model and an associated implementation for processing and generating large data sets.
- User specify a map function that processes a value pair to generate a set of intermediate value pairs, and a Reduce function that merges all intermediate values associated with the same intermediate key.
- MapReduce processing consists of two phase MapPhase and Reduce phase. Each phase of MapReduce consists of key-value pairs as input and output. Hadoop divides the input to a MapReduce job into the fixed-size pieces called input splits.
- Map() function $(K_1, V_1) \rightarrow$ List (K_2, V_2)
  It performs filtering and sorting of task into queue.
- Reduce() function. It performs a summary operation of best candidate resource for task execution.

Date: / / Page no:

✡ Transforming data for processing :-

- The main purpose of data transformation Data tr and feature extraction is to enhance the data in such a way that it increase the likelihood that the classification algorithm will be able to make meanigful predicting.

Data pre-processing techniques:

- Preprocessing of data involves a set of key tasks that demand extensive computational infrastructure.

- The bigger the data sets, more complex mechanism are needed to proces it before analysis and visualization.

- Pre-processing prepares the data and make the analysis feasible while improving the effectiveness of the result.

Date: / / Page no:

- following are some crucial steps involved in data pre-processing.
  - Data cleansing
  - Data Normalization
  - Data Transformation
  - Missing value imputation
  - Noise identification.
  - Minimizing the preprocessing tasks.

→ Data cleaning:- cleansing the data is usually the first step in data processing and is done to remove the unwanted elements as well as to reduce the size of the data sets.
  - This will make it easier for the algorithm to analyze it.

→ Data Normalization:- Normalization is the process of reorganizing data in a database so that it meets basic requirements:
1) There is no redundancy of data (all data is stored in only one place),
2) data dependencies are logical (all related data items are store together).

Date: / / Page no:

- Normalization Increase result performance
Normalization is also known as data normalization.
NF (Normal forms)
1 NF
2 NF
3 NF
BCNF
5 NF

- Data Transformation:- Data transformation is the process of converting data or information from one format to another, usually from the format of a source system into the required format of a new destination system.

- The usual reason for this data migration is the adoption of a new system that totally different from the previous one.

Data transformation ka main goal yahi hota he ki system ko is kabil banana ki vo new system ko adopt kr sake

- Data transformation involves two key phases:

→ Data Mapping :- The assignment of elements from the source base or system toward the destination to capture all transformations that occur

- This is more complicated when there are complex transformations like many-to-one or one-to-many rules for transformation.

→ Code generation :- The creation of the actual transformation program. The resulting data map specification is used to create an executable program to run on computer systems.

→ Missing value imputation

imputation : In Statistics, imputation is the process of replacing missing data with substituted values

Date: / / Page no:

- Missing value are data that haven't been extracted or stored due to budget restrictions or other limitations in the data extraction process.
- Missing value is not something to be ignored as it could skew your results.
- fixing the missing value issue is challenging handling it without utmost care could easily lead to complication in data handling and wrong conclusion.
- ML (machine learning technique) is used to missing value imputation.
  (Intruption in data is known as noise)

⇒ Noise identification :- Data gathering is not always perfect, but the data mining algorithm would always assume it to be.

- Data with noise can seriously affect the quality of the results, ~~tackling~~ tackling this issue is crucial.
- Noise can affect the input features, output or both in most cases.
- There are two popular approaches to remove from Data
  → Data polishing methods are used to eliminate ~~from~~ the noise.

Date: / / Page no:

→ The other method involves using noise filters that can identify and remove instances with noise from the data.

⇒ Minimizing and pre-processing data.

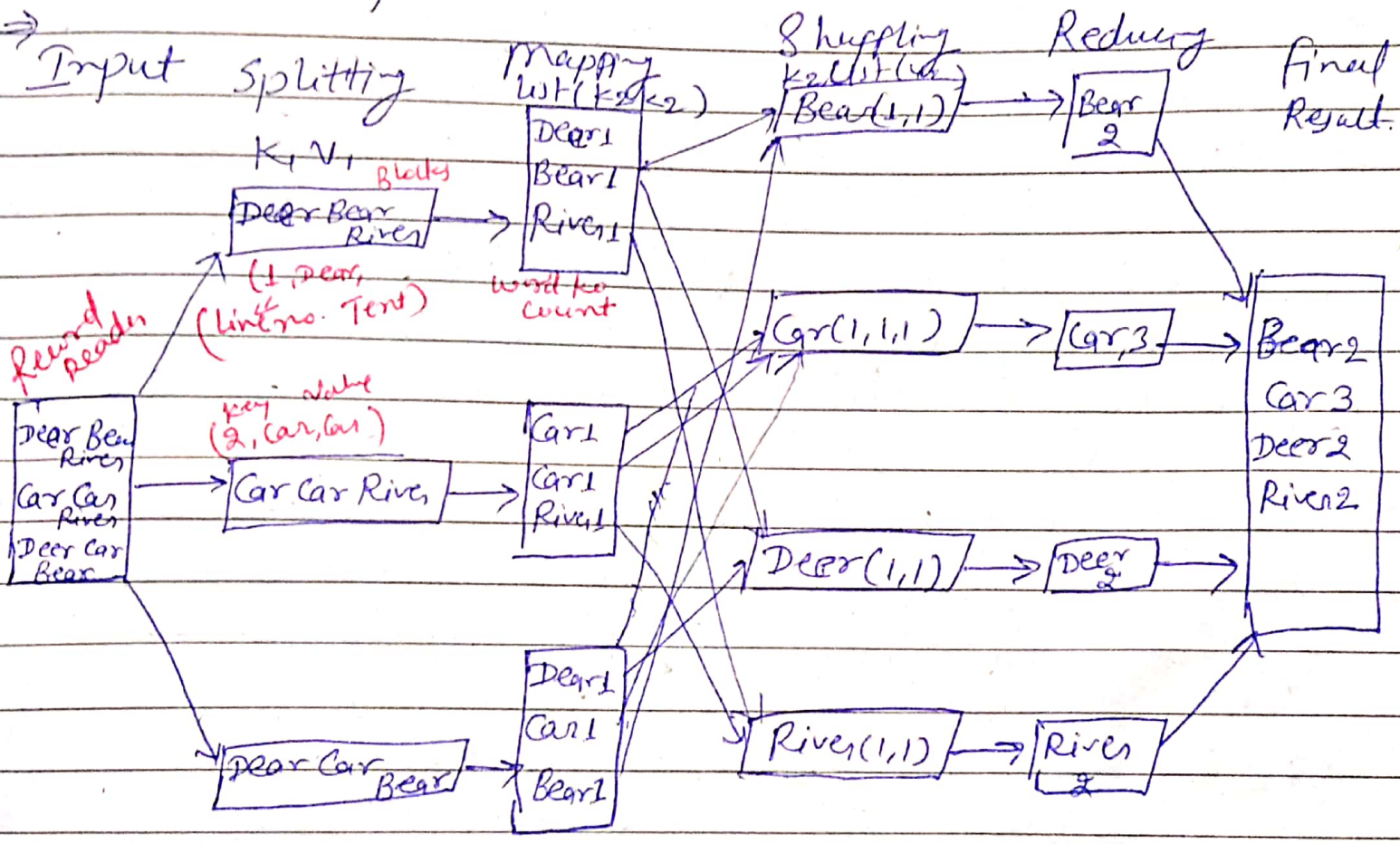
- preparing the data for your data analysis algorithm can involve many more processes depending on the application unique demands.
- However basic process like cleansing, redundant and normalization can be avoided in most cases if you choose right source for data extraction.
  → But it is highly unlikely that a raw source can give you clean data.

MapReduce divides a task into small parts and assigns them to many computers. Date: / / Page no:

✱ How MapReduce works:-

- The Map task take a set of data and converts it into another set of data, where individual elements are broken down into tuples (key-value pairs).

- The Reduce task take the output from the Map as an input and combines those data tuples (key-value pairs) into a smaller set of tuples.

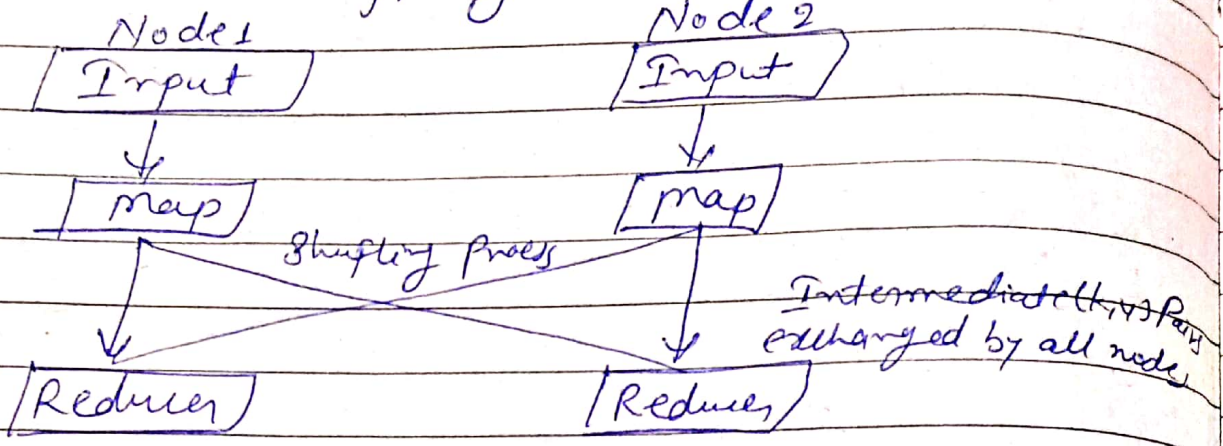


[The overall MapReduce word count process].

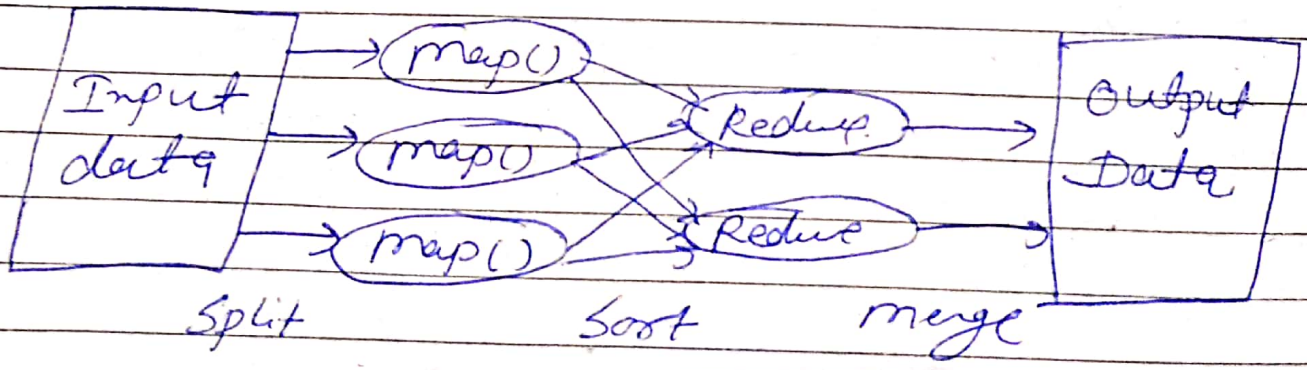
Date: / / Page no:

- Shuffling:

The process of exchanging the intermediate outputs from the map tasks to where they are required by the reducers is known as "Shuffling".



- MapReduce programming model uses two components: a Job tracker (master node) and many task trackers (Slave nodes). The Job Tracker is responsible for accepting job requests, for splitting the data input. The task tracker executes tasks as ordered by the master node. ~~The task can be either a map or a reduce.~~



Date: / / Page no:

☆ Subdividing data in preparation for Hadoop MapReduce:-

The whole process goes through four phases of execution namely, Splitting, Mapping, Shuffling, and Reducing.

- Input Splits:- An input to a MapReduce job is divided into fixed-size pieces called input splits. Input split is a chunk of the input that is consumed by a single map.

- Mapping:- This is the very first phase in the execution of map-reduce program. In this phase data in each split is passed to a mapping function to produce output value.
- In our example, a job of mapping of mapping phase is to count number of occurence of each world from input splits and prepare a list in the form of <key value> (word, frequency).

- Shuffling :- This phase consumes the output of mapping phase. Its task is to consolidate the relevant records from mapping phase output.
  - our example, the same words are clubbed together along with their respective frequency.

- Reducing :- In this phase, output values from the Shuffling phase are aggregated. This phase combines values from Shuffling phase and returns a single output value. In short, this phase summarizes the complete dataset.
  - In our example, this phase aggregates the values from Shuffling phase i.e., calculate total occurance of each word

☆ How MapReduce organizes work?
- Hadoop divide the job into tasks. There are two types of tasks:

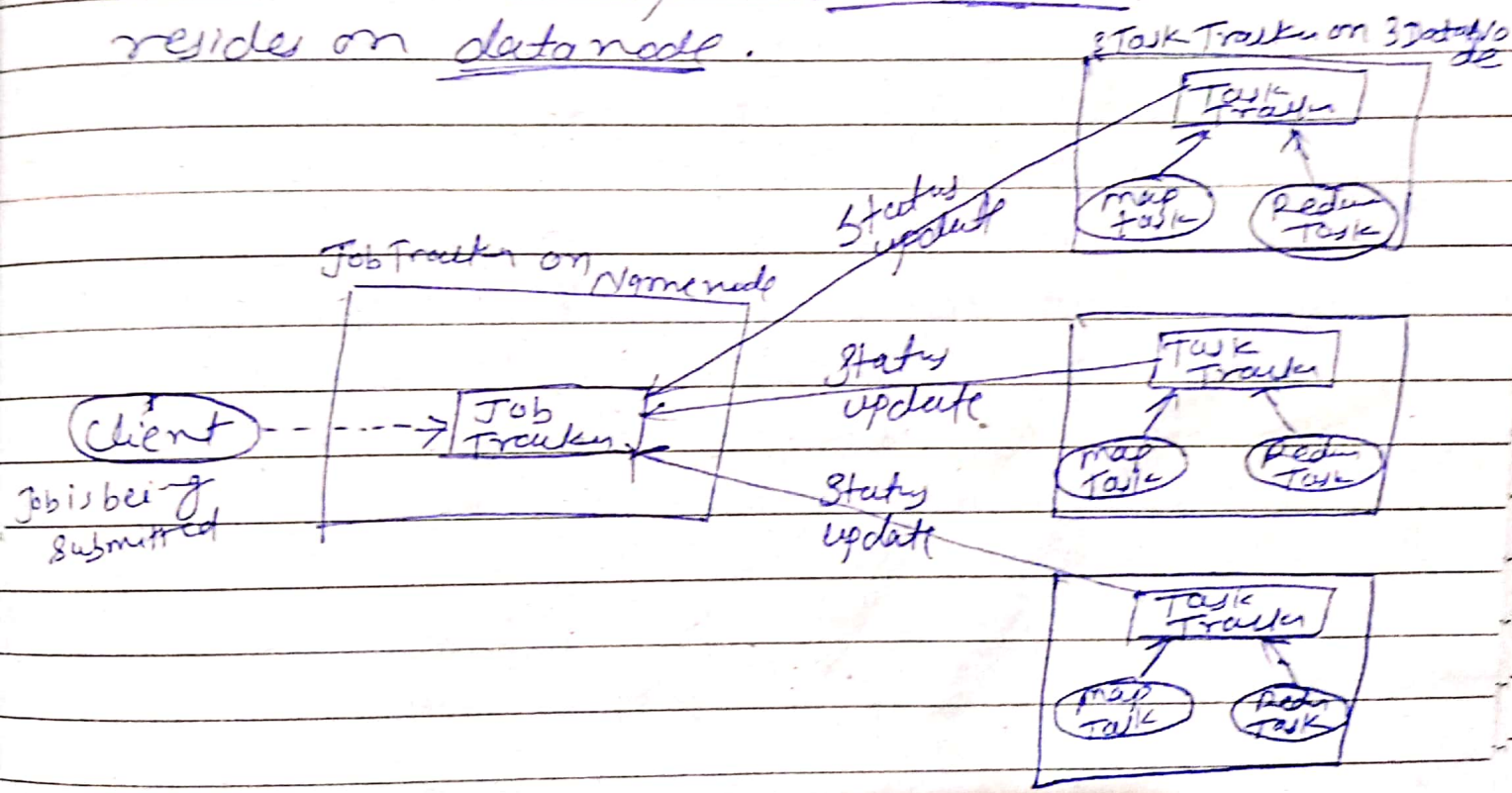
1) Map tasks (Splits & Mapping)
2) Reduce tasks (Shuffling, Reducing)

- The complete execution process (execution of map and reduce tasks, both) is controlled by two types of entities called a.

1) Job tracker: Acts like a master (responsible for complete execution of submitted job)

2) Multiple Task Trackers: Acts like slaves, each of them performing the Job

- For every job submitted for execution in the system, there is one Job tracker that resides on Namenode and there are multiple task trackers which resides on datanode.



Date: / / Page no:

- A job is divided into multiple tasks which are then run onto multiple data node in a cluster.
- It is the responsibility of job tracker to coordinate the activity by scheduling tasks to run on different data nodes.
- Execution of individual task is then to look after by task tracker, which resides on every data node executing part of the job.
- Task tracker responsibility is to send the progress report to the job tracker.
- In addition, task tracker periodically send "heartbeat" signal to the jobtracker so as to notify him of the current state of the system.
- Thus job tracker keeps track of the overall progress of each job. In the event of task failure, the job tracker can reschedule it on a different task tracker.

Unit IV

Date: / / Page no:

Hadoop MapReduce

* Employing Hadoop MapReduce:-

Hadoop Configuration mode: Three Hadoop Configuration files :- Settings to run hadoop cluster mapreduce and Spark.

- Hadoop cluster can be run in any of three supported modes:

- Local (Standalone) mode:-

By default, Hadoop is configured to run in a non-distributed mode, as a single java process where instead of HDFS, local file system is used for the purpose of input & output. Here we do not perform configurations, all services run on single JVM. This is useful for debugging.

- Pseudo-Distributed operations:-

Hadoop can also be run on a single-node in a pseudo-distributed mode, where each hadoop daemon runs in a seprate java process.

Date: / / Page no:

- It requires Configuration of three Hadoop installation file, hdfs-site.xml, core-site.xml and mapred-site.xml. Here data node, namenode, job tracker, task tracker all suppose to be run on single node. This mode is useful for development and testing HDFS.

- Fully Distributed :-

All Hadoop Services run in different JVM's belonging to one cluster. All daemons get executed in different node that form a multinode cluster.

Installing a hadoop cluster typically involves unpacking the software on all the machine in the cluster or installing it via a packaging system as appropriate for the operating system.

* Creating the components of Hadoop map-reduce jobs:-

We can create a hadoop MapReduce job with Spring data apache hadoop by following these steps:

1) Get the required dependencies by using maven.

2) Create the mapper component.

3) Create the reducer component.

4) Configure the application context

5) Load the application context when the application starts.

- Create the mapper component: A mapper is a component that divides the original problem into smaller problems that are easier to solve. We can create a customer mapper component by extending the class and overriding its map() method.

Date: / / Page no:

- Creating the reducer component: A reducer is a component that removes the unwanted intermediate values and passes forward only the relevant key-value pairs. We can implement our reducer by extending the class and overriding it reduce() method.

- Configuring the application context:- Because Spring data apache hadoop 1.0.0.M2 does not support java configuration, we have to configure the application context of our application by using XML.

- Loading the application context when the application starts:- We can execute the created hadoop job by loading the application context when our application is started. We can do this by creating new classpathXmlApplicationContext object and providing the name of our application context configuration file as a constructor parameter.

Distributed data processing across seven forms:-

Distributed data processing is a computer-networking method in which multiple computer across different locations share computer-processing capability. This is in contrast to a single, centralized server managing and providing processing capability to all connected systems.

- Computers that comprise the distributed data-processing network are located at different locations but interconnected by mean of wireless or satellite links.

Advantages of data processig:

- Lower cost :- Larger organization invest in expensive mainframe and Superlomputers to function as centralized servers.

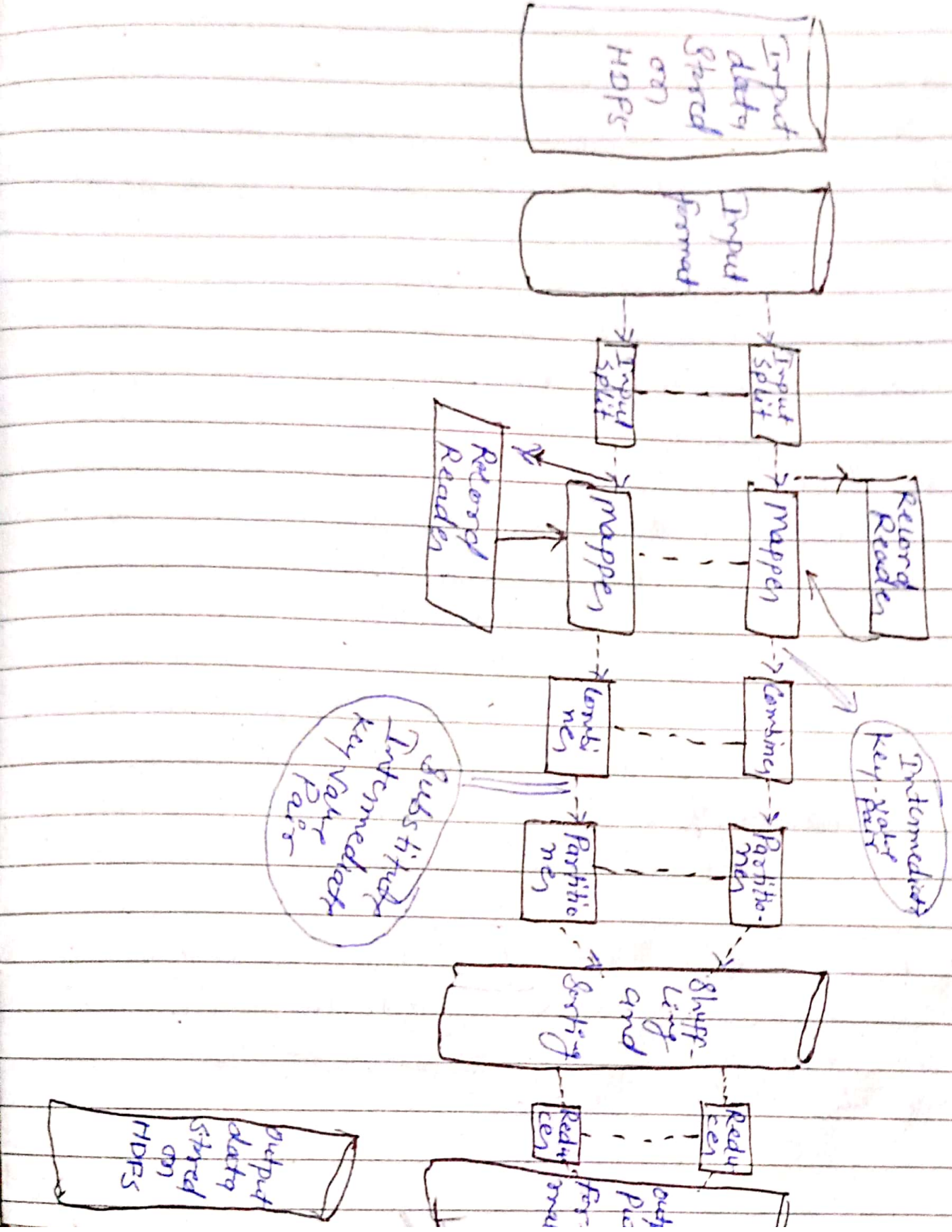
- Reliable :- Distributed data processig is more reliable, since multiple control centers are spread across different machines.

- Improve performance and Reduced processing time:- Single computers are limited in their performance and efficiency. An easy way to increase performance is by adding another computer to a network.

- Flexible:- Individual computers that comprise a distributed network are present at different geographical locations. For example, an organizational distributed network comprising of three computer can have each machine in different branch.

Date: / / Page no:

✱ Executing Hadoop MapReduce Jobs:-



- Hadoop MapReduce is the data processing layer. It process the huge amount of structured and un-structured data stored in HDFS. MapReduce process data in parallel by dividing the job into the set of independent tasks. So parallel processing improves speed and reliability.

- Steps of MapReduce Job execution flow:
  MapReduce process the data in various phase with the help of different components.

1) Input files :- In input file data for MapReduce Job is stored. In HDFS, input file reside. Line based log files and binary format can also be used.

2) Input format :- After that input format defines how to split and read these input files. It selects the file or other objects for input.

3) Input splits :- It represents the data which will be processed by an individual mapper. For each split, one map task is created.

4) Record Reader:- It Communicate with the InputSplit. And then Converts the data into key-value pairs suitable for reading by the mapper.

5] Mapper:- It processes input record produced by the Record Reader and generates intermediate key-value pairs.

6) Combiner:- Combiner is mini-reducer which performs local aggregation on the mapper output. It minimizes the data transfer b/w mapper and reducer.

7) Partitioner:- Partitioner comes into the existence if we are working with more than one reducer.

8) Shuffling & Sorting:- After partitioning, the output is shuffled to the reducer node. The shuffling is the physical movement of the data which is done over the n/w.

9) Reducer:- Reducer then takes set of intermediate key-value pairs produced by the mapper as a input.

Date: / / Page no:

10) Record writer :- It writes these output key-value pair from the Reducer phase to the output files.

11) Output format :- Output format defines the way how Record reader writes these output key-value pairs in output files.

* The Building Blocks of Hadoop MapReduce distinguishing Hadoop daemons :-

Hadoop employs a master/slave architecture for both distributed Storage and distributed computation. The distributed Storage System is called the Hadoop Distributed File System (HDFS).

- on a fully Configured cluster, "running Hadoop" means running a set of daemons, or resident programs, on the different servers in your network.

- These daemons have specific roles, some exists only on one server, some exist across multiple servers. These daemons include:

  - NameNode
  - DataNode
  - Secondary Namenode
  - Job Tracker
  - Task Tracker

1) Namenode :- The Namenode is the master of HDFS that directs the slave DataNode daemons to perform the low-level I/O tasks.
It is the bookkeeper of HDFS, it keeps track of how your files are broken down into file blocks, which nodes store those blocks and the overall health of the distributed file system.

2) DataNode :- Each slave machine in your cluster will host a datanode daemon to perform the grunt work of the distributed file system - reading and writing HDFS block to actual files on the local file system.

3) Secondary NameNode (SNN):- The SNN is an assistant daemon for monitoring the state of the cluster HDFS. Like the NameNode, each cluster has one SNN and it typically resides on its own machine as well. No other DataNode or TaskTracker daemons run on the same server.

4) Job Tracker:- once you submit your code to your cluster, the JobTracker determines the execution plan by determining which file to process, assigns nodes to different task, and monitors all the tasks as they running. There is only one JobTracker daemon per hadoop cluster.

5) TaskTracker:- Each task tracker is responsible for executing the individual tasks that the JobTracker assigns. Although there is a single TaskTracker per slave node, each TaskTracker can spawn multiple JVM to handle many map or reduce task in parallel.

* Investigating the hadoop distributed file system selecting appropriate execution models:

- Single Node (Local Mode or Standalone mode):-

Standalone mode is the default mode in which hadoop run. Standalon mode is mainly used for debugging where you don't really use HDFS.

- You can use input and output both as a local file system in standalon mode.
- Standalone mode is usually the fastest hadoop modes as it use the local file system for all the input & output.

- Pseudo-distributed Mode:- The pseudo-distributed mode is also known as single-node cluster where both NameNode and DataNode will reside on the same machine.

- In pseudo-distributed mode, all the hadoop daemons will be running on a single. Such configuration is mainly used while testing when we don't need to think about the resources and other users sharing the research resources.

- In this architecture, a seprate JVM is spawned for every hadoop components as they could communicate across network sockets, effectively producing a fully functioning and optimized mini-cluster on a sigle ~~host~~ host.

- **Fully distributed mode:-** As the Name suggests, this mode involve the code running on an actual hadoop cluster. It is mode in which you see the actual power of hadoop, when you run your code against a very large input on 1000s of servers.

- It is always difficult to debug a MapReduce program as you have mappers running on different machine with

different piece of input. You can never know where the mappers are going to run eventually. Also with large inputs it is likely that the data will be irregular in its format.

Unit V

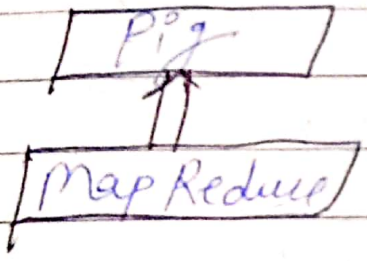
Date: / / Page no: 1

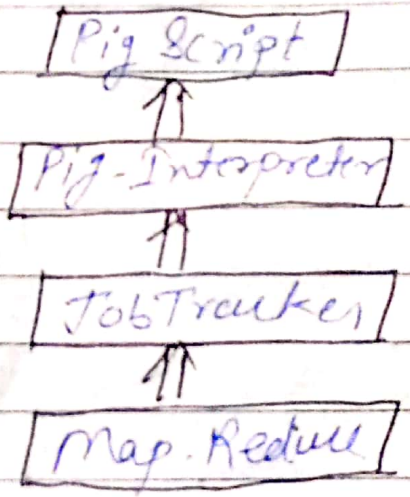
Big data tools and Technique.

★ Installing and Running Pig.

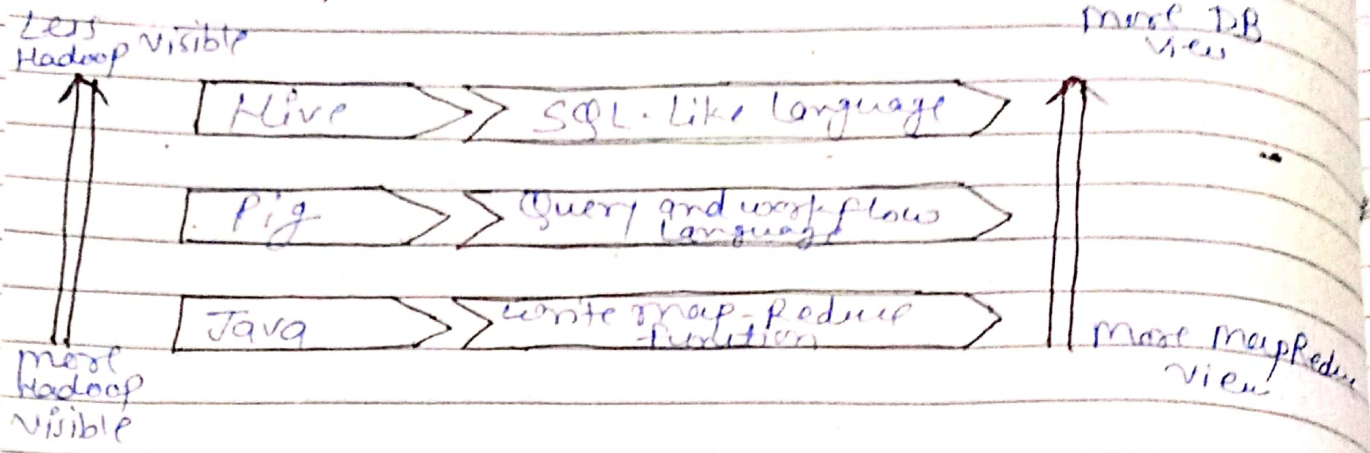
Apache pig :-

- Apache pig is an abstraction over map Reduce. It is a tool/platform which is used to analyze larger sets of data representing them as data flows.
- Pig provides a high-level language known as Pig Latin.
- Compiles down to MapReduce jobs.
- Developed by Yahoo. open source language.





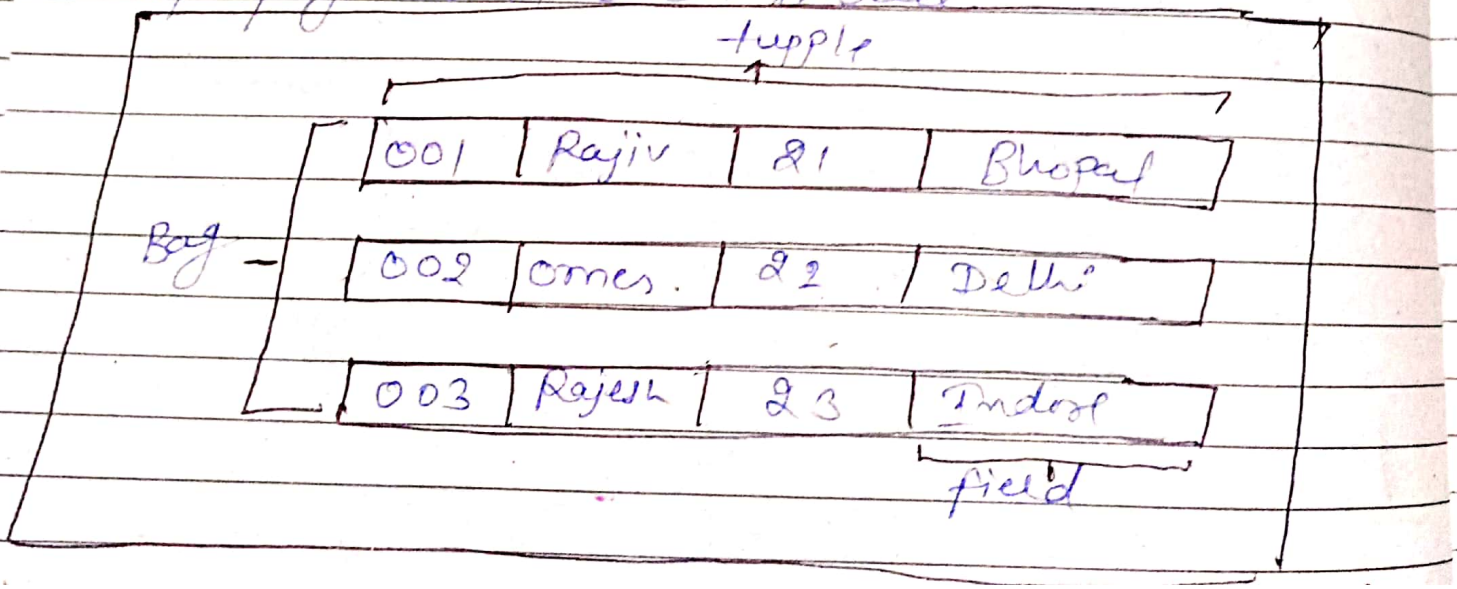
★ Levels of Abstraction :-



★ Pig Latin Data model :-

The data model of Pig Latin is fully nested and it allows complex non-atomic data types such as map and tuple.

- given below is the diagrammatical representation of Pig Latin data model



Date: / / Page no:

- **Atom:-** Any single value in pig Latin, irrespective of their data, type is known as an Atom. It is stored as string and can be used as string and number. int, long, float, double, chararray and bytearray are the atomic value of pig.

- **Field:-** A piece of data or a simple atomic value is known as field.

- **Tuple:-** A record that is formed by an ordered set of fields is known as tuple, the fields can be of any type. A tuple is similar to a row in a table of RDBMS.

- **Bag:-** A Bag is an unordered set of tuples. In other words, a collection of tuple (non-unique) is known a bag.

Date: / / Page no:

* Comparision with databases:-

- With the range of technologies in the big data world, there is often confusion to choose from them. It is required to handle huge databases efficiently with big data, and the options for managing and querying data are also needed.

- When it comes to managing databases, SQL (Structured Query Language) is the old friend well tried and tested by everyone for data analysis. But the complicated world of Hadoop needs high-level data analysis model.

- ~~Though~~

- Though old SQL still is the favorite of many and is popularly used in numerous organizations, Apache Hive and pig have become the buzz terms in the big data world today.

- These tools provide easy alternatives to carry out the complex programming of MapReduce helping data developers and analysts.

Date: / / Page no:

Comparision Pig, Hive, SQL.

Criteria	Hive	Pig	SQL
1) Language used	Uses hive QL a declarative language	Uses pig latin a procedural data flow language	SQL itself is a declarative language.
2) Definition	An open source built with an analytical focus used for analytical queries.	An open source and high-level data flow language with a multi-query approach	General purpose database language for analytical and transactional queries.
3) Suitable for.	Batch processing OLAP (online Analytical processing)	Complex & nested data structure	Business demand for fast data analysis
4) Developed by	Facebook	Yahoo	oracle.
5) Operational for	Structured data	Structured & Semi-Structured	Relational database manage.
6) Mainly used by.	Data Analysts.	Researcher & programmes.	Data analysts Data Scientists and programmes

* Apache Pig - User defined functions :-

- Apache pig provides extensive Support for User defined Functions (UDFs). Using these UDFs we can define our own functions and use them.
- The UDF Support is provided in six programming languages namely, Java, Jython, python JavaScript, Ruby and Groovy.
- For writing UDFs complete Support is provided in java and limited Support is provided in all the remaining languages.
- Using Java, you can write UDFs involving all parts of the processing like data Load/Store, Column transformation, and aggregation.
- In apache pig, we also have a Java repository for UDFs named piggybank. Using piggybank we can access Java UDFs written by other users, and contribute our own UDFs.

Date / / Page no

- Type of UDF's in Java:-

While writing UDFs using java, we can create and use the following three types of functions:

- Filter functions:- The filter functions are used as conditions in filter statements. These functions accept a pig value as input and return a boolean value.
- Eval functions:- The Eval functions are used in FOREACHGENERATE statements. These functions accept a pig value as input and return a pig result.
- Algebraic functions:- The algebraic functions act on inner bags in a FOREACHGENERATE statement. These functions are used to perform full MapReduce operations on an inner bag.

Date: / / Page no:

☆ Apache Pig operators :-

The apache Pig operator is a high-level procedural language for querying large data sets using hadoop and the Map Reduce platform.

- A pig latin statement is an operator that takes a relation as input and produces another relation as output.

- These operators are the main tools for pig latin provides to operate on the data. They allow you to transform it by sorting, grouping, joining, projecting and filtering.

Relational operators :-

Relational operators are the main tools pig latin provides to operate on the data. It allows you to transform the data by sorting, grouping, joining, projecting and filtering.

Date / / Page no:

* Relational operations :-

Loading and storing.

- **LOAD :-** To load the data from the file System (local / HDFS) into a relation.
- **STORE :-** To save a relation to the file System (local / HDFS).

Filtering.

- **FILTER :-** To remove unwanted rows from a relation.
- **DISTINCT :-** To remove duplicate rows from a relation.
- **FOREACH, GENERATE :-** To generate data transformation based on Column of data.
- **STREAM :-** To transform a relation using an external program.

Grouping & Joining.

- **Join :-** To join two or more relations.

Date: / / Page no:

- ~~SPLIT~~ :
- COGROUP :- To group the data in two or more relation.
- GROUP :- To group the data in a single relation.
- CROSS :- To create the cross product of two or more relations.

~~Sorting~~ :

- ORDER :- To arrange a relation in a sorted order based on one or more field (ascending or descending).
- LIMIT :- To get a limited number of tuples from a relation.

Combining & Spliting :-

- UNION :- To combine two or more relation into single relation.
- SPLIT :- To split a single relation into two or more relations.