UNIT-1 Statistics Essential for Analytics

Q.1 Describe skewness & kurtosis

→

Skewness -

- It is measure of symmetry
- A data set is symmetric if it looks same to the left and right of the center point
- it mean lack of symmetry

first coefficient of skewness = $\frac{\text{mean} - \text{mode}}{s}$

second coefficient of skewness = $\frac{3(\text{mean} - \text{median})}{s}$

* Two Types of Skewness

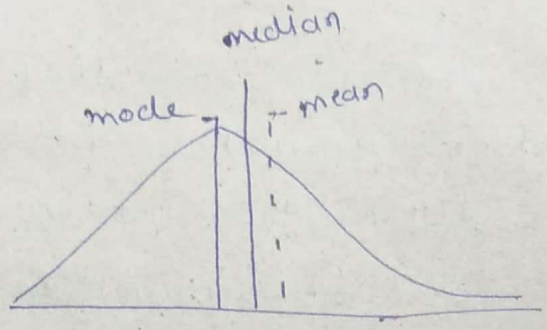
1) Positive skewness
2) Negative skewness

Scanned by CamScanner

1) Positive skewness -

- When the tail on the right side of the distribution is longer.

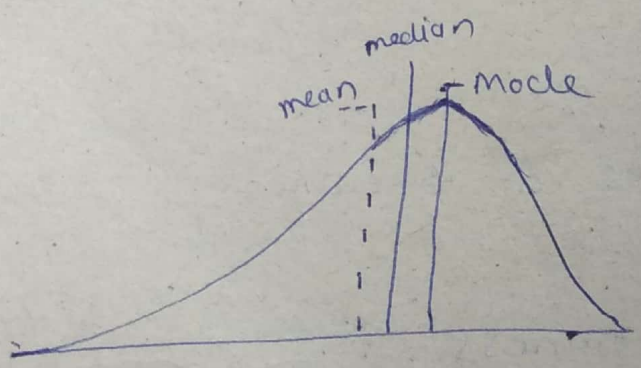
- The mean & median will be greater than mode



2) Negative skewness -

- When the tail on the left side of the distribution is longer than tail on the right side.

- The mean and median will be less than the mode.



Scanned by CamScanner

Kurtosis -

- The degree of peakedness of a distribution
- It determines the heaviness of the distribution tails
- Kurtosis of normal distribution is 3
- A positive value tells that there is heavy-tails (i.e a lot of values in tails)
- A negative value tells that there is light-tails (i.e a little data in tails)

* Types of Kurtosis

1) Leptokurtic - Distribution having relatively high peak. Kurtosis > 3

2) Platykurtic - Distribution curve is ~~not very~~ flat-topped kurtosis < 3

3) Mesokurtic - distribution curve is not very peaked or very flat-topped. it is normal distribution. value of kurtosis is 3

Scanned by CamScanner

Q.2 Define Probability, Probability distribution and Conditional Probability

→

Probability -

- Suppose that an event E can happen in h ways out of total of n possible equally likely ways

- Then the Probability of occurrence of the event is

$$P = Pr(E) = h/n$$

- The Probability of non-occurrence of the event is

$$q = Pr(\text{not } E) = n-h/n$$

- Probability is a number between 0 and 1; where

  0 - indicates impossibility

  1 - indicates certainty

Scanned by CamScanner

Probability Distribution -

- if a variable 'x' can assume discrete set of values $x_1, x_2, x_3 \ldots x_n$ with respective probabilities
- $P_1, P_2, \ldots P_n$ where $P_1 + P_2 + \ldots P_n = 1$ then we say that a discrete probability distribution for x has been defined
- The function $P(x)$ which has the respective value $P_1, P_2 \ldots P_n$ for $x = x_1, x_2 \ldots x_n$ is called the probability function

* Conditional Probability -

- if E_1 and E_2 are two events, the probability that E_2 occurs given that E_1 has occurred is denoted by $Pr\{E_2 | E_1\}$ is called the conditional probability of E_2 given that E_1 has occurred

$$Pr\{E_1 E_2\} = Pr\{E_1\} \cdot Pr\{E_2 | E_1\}$$

OR

$$Pr\{E_2 | E_1\} = Pr\{E_1 E_2\} / Pr\{E_1\}$$

Scanned by CamScanner

6 - if occurrence & non occurrence of E_1 does not affect the probability of occurrence of E_2

- then we say that E_1 & E_2 both are independent events; other wise they are dependent

$$Pr\{E_1 E_2\} = Pr\{E_1\} \cdot Pr\{E_2\}$$

Q. 3 - Difference between standard deviation & variance

→

Variance	Standard Deviation
1) Average squared differences from mean	The square root of the variance.
2) Measures Dispersion within the Data set.	it measures spread around the mean
3) Variance is not sub-additive	A measure spread for symmetrical distributions with no outliers
4) Variance also measure the volatility of Data of a population	Standard deviation is finance, is often called volatility
5) Variance measures how far the outcome varies from the mean	Standard deviation measures how far the normal standard deviation is from the expected value.
6) denoted by (σ^2)	6) denoted by (σ)

Scanned by CamScanner

Q.4 Define mean, mode, median. Calculate mean mode & meadian for given data.

Roll No	Mark of BPC
1	42
2	58
3	47
4	65
5	67
6	47
7	65
8	78
9	73
10	78
11	65
12	70
13	70
14	70
15	65
	$\Sigma 960$

→ **Mean** – The average of a set of Numerical values; as calculated by adding them together & dividing ~~the~~ by the number of terms is the set.

$$\bar{x} = \frac{\Sigma fx}{N}$$

Scanned by CamScanner

$$\bar{x} = \frac{960}{15}$$

$$\bar{x} = 64$$

2) <u>Median</u> -

The median of a set of number arranged in order of magitude (i-e in an array) is either the middle value or the arithmatic mean of the two middle value.

42 ; 47, 47, 58, 65, 65, 65, 65, 67, 70, 70, 70, 73
78, 78

<u>65</u> = is the middle value
i-e median

3) Mode -

The mode of a set of number is that value which occurs with the greatest frequency that is most common value

<u>65</u> = most common & with greater frequency occurence

Scanned by CamScanner

Q.5 ~~me~~ Explain measure of spread with example

→

Measure of Spread -

- The degree to which numerical data tend to spread about an average value is called the dispersion or variation & measures of spread are used to measure it.
- it describes how similar or different the set of observed values are for a Particular Variable
- Spread Can be measured if numerical data is arranged in an order.
- Summarizing data helps us to understand the data easily
- Measure of spread Summarise the data in a way that Shows how Scattered the Values are & how much they different from the mean Value

Scanned by CamScanner

Q.6 Explain measures of central tendency with example

→

Measures of Central tendency -

- Values which lie centrally within a set of data arranged according to magnitude or average are called measures of central Tendency

1) Arithmatic Mean -

- The average of set of numerical Values, as calculated by adding them together and dividing by the number of terms in the set.

2) The median -

- The median of a set of numbers arranged in order of magnitude (i.e in an array) is either the Middle Value or the arithmatic mean of the two middle values

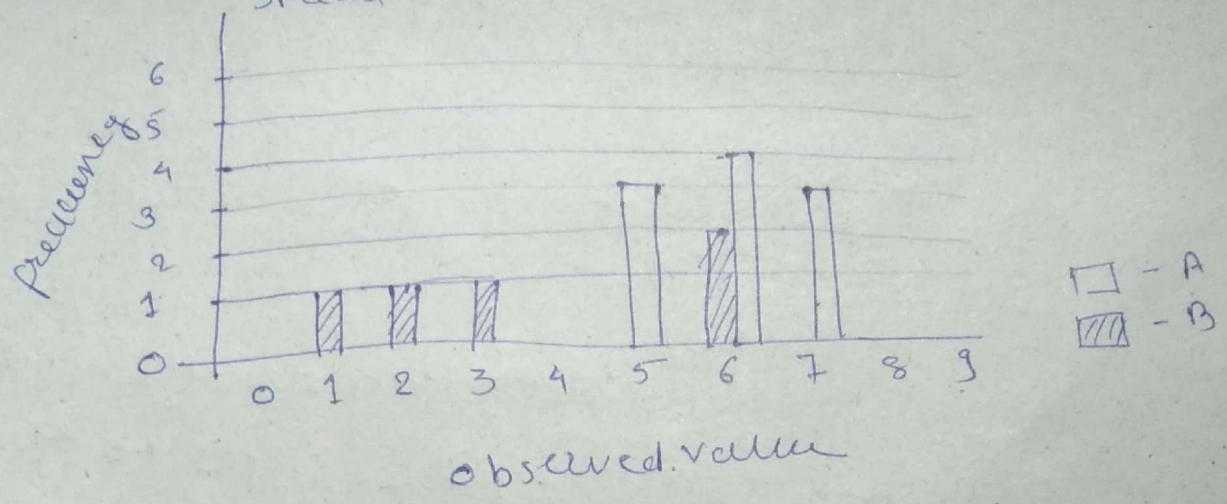
3) The Mode -

- The mode of a set of numbers is that value which occurs with the greatest frequency that is most common value

Scanned by CamScanner

Ex – set A: 5, 5, 5, 6, 6, 6, 6, 7, 7, 7 | set B: 1, 2, 3, 6, 6

spread of value in Data set A & Data set B



Q.7 Define Standard deviation and Varience. calculate standard deviation and varience for give data set.

No	mark	$x_i - \bar{x}$
1	42	-22
2	58	-6
3	47	-17
4	65	1
5	67	3
6	47	-17
7	65	1
8	78	14
9	73	9
10	78	14
11	65	1
12	70	6
13	70	6
14	70	6
15	65	1

Σ mark = 960

mean $\bar{x} = \frac{\Sigma f x}{N}$

$= 64$

median = 65

mode = 65

Scanned by CamScanner

1) Standard Deviation -

- it is measure of the amount of variation or dispersion of a set of values
- it tells how other elements from a group are spread out from the average (mean) or expected value
- Low standard deviation mean most of the number are close to the average
- High standard deviation mean the numbers are more spread out.

$$S.D = \sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}$$

$x_i - \bar{x}$	$(x_i - \bar{x})^2$
-22	484
-6	36
-17	289
1	1
3	9
-17	289
1	1
14	196
9	81
14	196
1	1
6	36
6	36
6	36
1	1
	$\sum 1692$

$$\sigma = \sqrt{\frac{\sum(1692)}{15}}$$

$$= \sqrt{112.8}$$

$$\sigma = 10.620$$

Scanned by CamScanner

Variance —

- it is square of standard deviation

$$\sigma^2 = S.D$$

$$\sigma = 10.620$$

$$\sigma^2 = 112.8$$

Q.8. Explain inferential statistics and descriptive statistics

→

1) Inferential statistics -

- It uses random sample of data from a Population to describe and to make inferences / conclusions about the Population

- it is used to study the relationships between variables within a sample, & you can make conclusions, generalizations or predictions about a bigger population

- Here sample accurately represent the Population

Ex- Sample data of small number of people and try to determine whether drug will work for entire Population

Scanned by CamScanner

2) Descriptive Statistics –

- it quantitatively describes or summarizes collection of information (which is either population or sample)
- Can not reach to any conclusions
- it includes numbers, charts, tables and graphs
- one can simply describe about current data using mean, standard deviation etc.
- it allows us to show data in a meaningfull way and provides ability to make a simple interpretation data.

Ex – From population 10,000 employees average salary of 30 employees is calculated.

Q.9 Write a note on poission distribution –

→

Poission Distribution –

- Discrete Probability distribution that expresses the Probability of a given number of events occurring in a fixed interval of time or space if these events occurs with a known constant mean rate and independently of the time since the last event

$$P(K \text{ events in interval}) = e^{-\lambda} \frac{\lambda^K}{K!}$$

Scanned by CamScanner

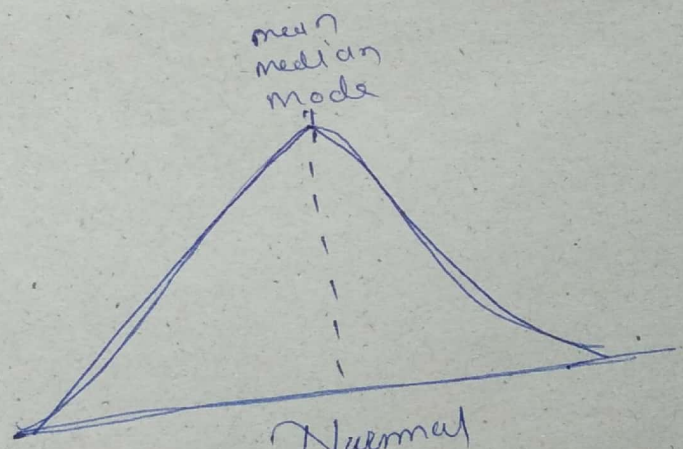
- Poisson Distribution to do interesting things like finding the Probability of a number of events in a time Period or finding the Probability of waiting some time until the next event

Q.10 Write a note on normal distribution

→

Normal Distribution -

- Also called as Gaussian Distribution/ Standard distribution
- The "Bell Curve" for data points indicate Normal distribution.



Normal Distribution

- it shows normal or symetric data distribution Here mean, mode, & median value are same Located at the center of the curve

Scanned by CamScanner

- it is a complete bell shape curve
- A data set is symmetric if it looks same to the left and right of the center point.

Q.10 Write on Bayesian inference

→

Bayesian inference -

- Bayesian inference is a method of statistical inference in which Bayes' theorm is used to update the probability for a hypothesis as more evidence or information becomes available.

Scanned by CamScanner

Q.11 Explain Interquartile range

→

Interquartile range -

- is a measure of Variability, based on dividing as set into quartiles.
- Quartiles divide an ordered dataset into four equal Parts.
- The values that divide each Part are called the first, second and third quartiles denoted by Q_1, Q_2, Q_3
- Is the measure of where bulk (majority) of the values lies
- $IQR = Q_3 - Q_1$

Q_1 - The lower quartile (Q_1) is the point betn the lowest 25% of value and the highest 75% of value. it is also called the 25th percentile

Q_2 - The second quartile (Q_2) is the middle of the data set, it is also called the 50th percentile, or the median

Scanned by CamScanner

Q_3 – The upper quartile (Q_3) is the point betn the lowest 75% and highest 25% of values it is also called the 75th percentile

Why to calculate IQR

- To find out outliers
- The higher the IQR, the more spread out the data points; the smaller the IQR the data points are bunched up around the mean

Scanned by CamScanner

UNIT - 2

Q.1 Differentiate betn Probability Sampling and Non Probability Sampling

→

Probability Sampling	Non-Probability Sampling
1) Probability sampling is a sampling technique, in which the subjects of the population get an equal opportunity to be selected as a representative sample	1) Non-Probability sampling is a method of sampling wherein, it is not known that which individual from the population will be selected as a sample
2) it also known as Random Sampling.	2) it is also known as Non Random sampling
3) in this samples select Randomly	3) in this samples select Arbitrarily
4) opportunity of selection is fixed and known	4) opportunity of selection is Not specified and unknown
5) Research is conclusive	5) Research is exploratary
6) Result is unbiased	6) Result is Biased.
7) Method is objective	7) Method is subjective
8) Inferences is statistical	8) Inferences is Analytical
9) Hypothesis is Tested	9) Hypothesis is Generated.

Scanned by CamScanner

Q.2 Write a note on Hypothesis testing

→

- It is a test used to determine whether the Hypothesis assumed for the sample of data stands true for the entire population or not.

Two Type of Hypothesis

1) Null hypothesis
2) Alternate hypothesis

1) Null Hypothesis -

- if we are to compare method A and method B about it's superiority and if proceed on the assumption that both methods are equally good, then this assumption is termed as Null Hypothesis

- it is symbolized as H_0

2) Alternate Hypothesis -

- if we think that method A is superior or the method B is inferior then it is termed as Alternate Hypothesis
- it is symbolized as H_1.

Scanned by CamScanner

- Alternate ~~ho~~ and Null Hypothesis are the statements about unknown Population para-meters.

- Rejecting the Null Hypothesis is accepting alternate hypothesis and vice versa.

- ~~In hypothesis testing the main~~

* During hypothesis Testing, we test 3 type of hypothesis given by.

1) $H_0 = \mu = \mu_0$ Against $H_1 : \mu \neq \mu_0$
(Two tailed test)

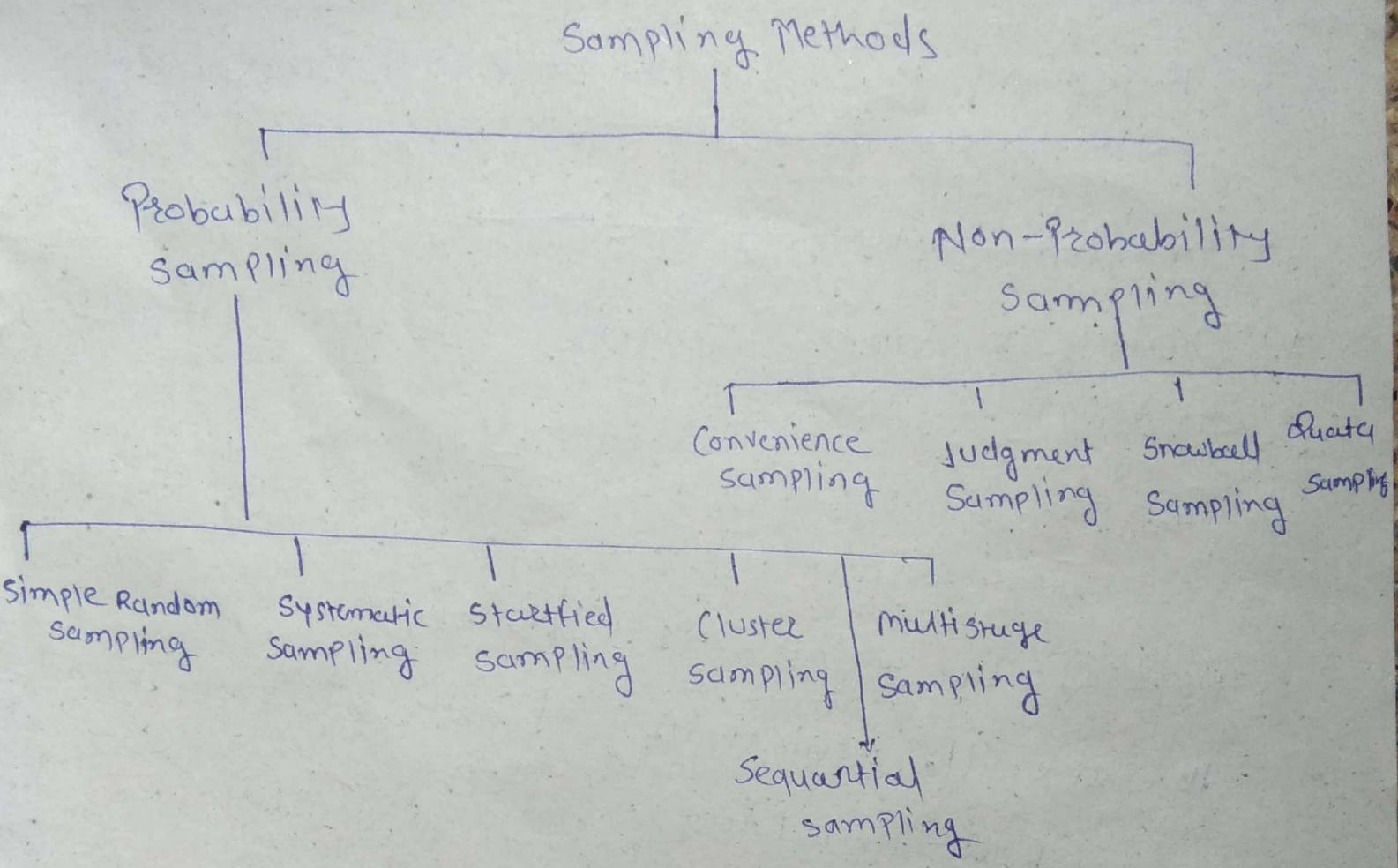
2) $H_0 = \mu = \mu_0$ Against $H_1 : \mu > \mu_0$ OR
$H_0 : \mu <= \mu_0$ Against $H_1 : \mu > \mu_0$
(Right tailed test)

3) $H_0 : \mu = \mu_0$ Against $H_1 : \mu < \mu_0$ OR
$H_0 : \mu >= \mu_0$ Against $H_1 : \mu < \mu_0$
(left taild test)

Scanned by CamScanner

Q.3 Explain different Sampling methods

→



1) Probability Sampling -

- Also called as random sampling or Chance sampling
- Each item from population has an equal chance of inclusion in the sample.
- Each item / element in the population has a known and equal probability of selection
- Every element is selected independently of every other element.

Scanned by CamScanner

f) Sequential Sampling

- Number of samples is more than two. one can go on taking samples one after another as long as one desires to do so

2) Non-Probability Sampling -

- Also called as deliberate sampling, Purposive Sampling and Judgment Sampling
- No Predefined system for selection of a sample
- Each item from a population has no equal chance of being selected in a sample
- The sample is not a proportion of the population & there is no system in selecting the sample, The selection depends upon situation

a) Convenience Sampling -

- The process of including whoever happens to be available at the time, also called "accidental" or "haphazard" Sampling

b) Judgment Sampling

- A researcher exerts same effort in selecting a sample that seems to be most appropriate for the study

Scanned by CamScanner

a) Simple Random sampling -

- It give each possible sample combination an equal probability of being picked up and each item in the entire population to have an equal chance of being include in the sample

b) Systematic Sampling -

- select every ith item on a list
- First unit is selected randomly and the remaining units of the sample are selected at fixed intervals
- it is spread more evenly over the entire population

c) Stratified sampling -

- if population is not homogeneous then the population is divided into several sub-populations that are individually more homogeneous than the total population

d) Cluster Sampling -

- The total population is divided into a number of relatively small subdivisions (i.e clusters) of smaller units and then some of these clusters are randomely selected to be part of sample.

e) Multistage Sampling

- is the taking of samples in stages using smaller and smaller sampling units at each stage

Scanned by CamScanner

c) Snowball sampling -

- Selection of additional respondents is based on referrals from the initial respondents and used to sample from low incidence or rare populations

d) Quota sampling

- The population is divided into classes on the basis of relevant control characteristics
- A quota of sample units is established for each cell
- A convenience sample is drawn for each cell until the quota is met.

Q.4 Describe Point estimation and hypothesis testing

→

i) Point estimation -

- it is a single valued estimation of an unknown parameter of a population
- Using sample mean (i.e $\bar{x}$) we can estimate population mean (i.e μ)
- using sample variance (s^2) and modified sample variance (s^2_1) we can estimate unknown population variance (σ^2)

Scanned by CamScanner

- if there are more than 1 estimators for an unknown parameter then good estimator is chosen by considering following properties

* Properties of good estimator

1) ~~unbai~~ unbiasedness -

2) consistency

3) sufficiency

4) Efficiency

2) Hypothesis Testing

- it is a test used to determine whether the hypothesis assumed for the sample of data stands true for the entire population or not.

- Two Type of Hypothesis

1) Null hypothesis

2) Alternate hypothesis

- Alternate & NULL hypothesis are the statements about unknown population parameters

- Rejecting the NULL hypothesis is accepting Alternate hypothesis & vice versa

Scanned by CamScanner

Q.5 Define Data. Explain Type of data.

→

Data –

- facts and statistics collected together for reference or analysis
- E.g personal details, marks of students
- Geographical points collected by your mobile (GPS) wherever you have travelled etc.

There are 2 - Types of Data:

1) Quantitative Data
2) Qualitative Data

1) Quantitative Data –

- Data are measures of values or counts and are expressed as numbers.
- Quantitative data are data about numeric variable
- e.g – how many; how much; or how often;

Scanned by CamScanner

2) Qualitative Data -

- Data are measures of types and may be represented by a name, symbol, or a number code.

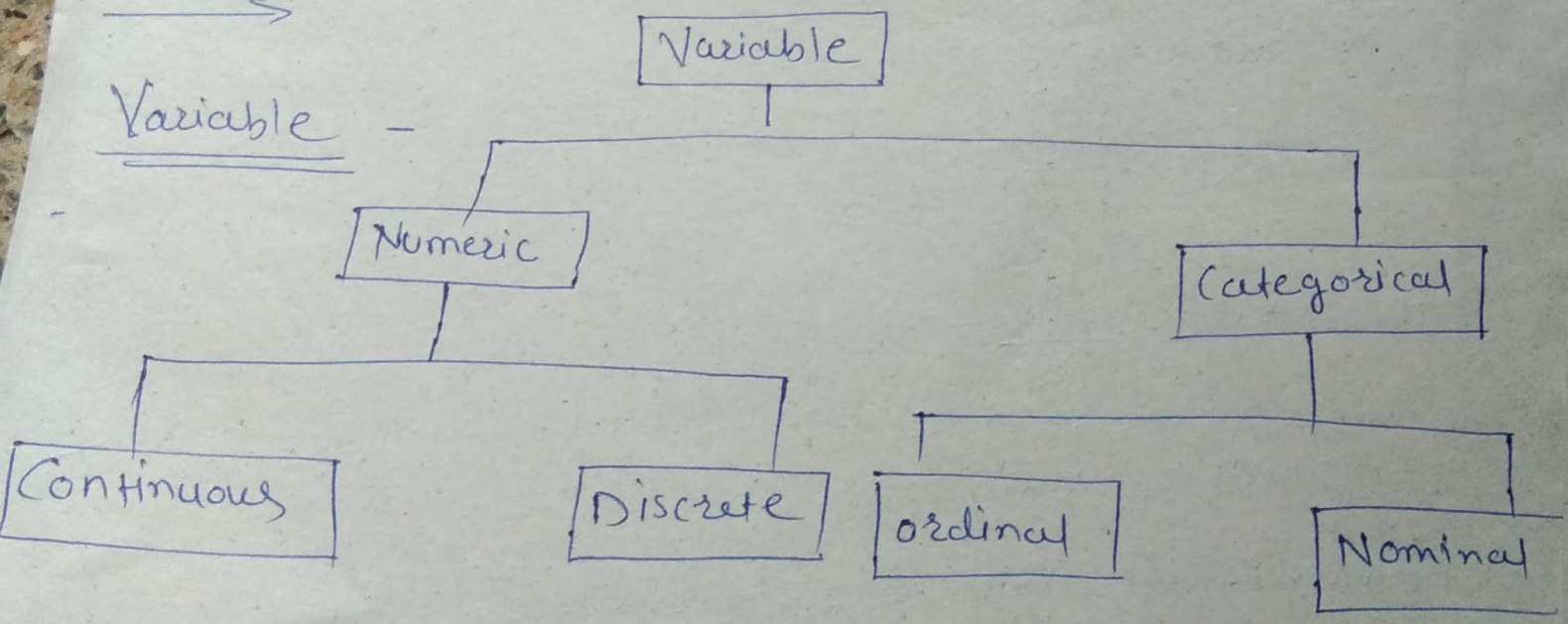
- Qualitative data are data about categorical variable (e.g what type)

- e.g - What type

Q.6 Define Variable. Explain types of variable

→

Variable -



Scanned by CamScanner

1) Numeric Variables

- have values that describe a measurable quantity as a number like 'how many' or 'how much'

a) Continuous variable -

- it takes value from a certain set of real numbers e.g height, time, age and temperature

b) Discrete variable -

- value is based on a count which is taken from a set of distinct whole values.

- No fraction values. E.g number of registered car, number of ~~children~~ children etc

2) Categorical Variable.

- have values that describe a 'quality or characteristic', like 'what type' or which category

a) ordinal variable -

- it takes value that can be logically ordered or ranked. E.g academic grades (A, B, C etc) clothing size (M, L, XL), Attitude (strongly agree, agree)

Scanned by CamScanner

b) Nominal Variable -

- Value that is not able to be organized in a logical sequence.
- Eg - business type, eye color, region etc.

Scanned by CamScanner

B.D.C - Unit - 3

Fundamentals of Big Data.

Q.1 What is big data ? Explain different types of data.

→ Big Data -

- Big data is the term that describes the large volume of data - both structured and unstructured, which is analyzed computationally to get Patterns, trends and associations

- This term used for which data set that are difficult to store, Process, and Manage the available data base using traditional data Processing Appl^n

- As data is big it requires different approaches in terms of - techniques, tools & architectures

* Major Sources of Big Data
  1) E-commerce
  2) Social media
  3) Web logs
  4) Medical Records
  5) cell phone records
  6) Transaction Records

Scanned by CamScanner

* Type of Big Data.

1) Structured –

- Any data that can be stored accessed and Processed in a fixed format is called structured data. Ex – data stored in RDBMS. 'SQL' is used to manage that data.

2) Semi – Structured –

- Data which does not have a formal structure of a data model, i.e table definition in a relational DBMS but it has some organizational Properties like tags to seperate the elements.
Ex – JSON and XML file format.

3) UnStructured –

- Data which have unknown form and cannot be stored in RDBMS and can not be analyzed unless it is transformed into a structured format.

Ex – Multimedia contents as Images, audio, video.

- Almost 80% data generated is unstructured data.

Scanned by CamScanner

Q.2 Explain different categories of NoSQL database with Example

→ - There are 4- Main categories of NoSQL database are as follows

1) key-value store
2) Column oriented
3) Graph Database
4) Document oriented.

1) key-Value store -

- Are the simplest
- ~~Every name (i.e "key") along with its value~~
- Every item in the database is stored as an attribute name (i.e "key") along with its value.
- It stores data in Hash Table where each key is unique, and the value can be BLOB, string etc.
- it is most basic type of data base
- Useful to stored basic Information as Customer information, shopping cart contents etc

Ex — Riak, Tokyocabinet, Redis server, Memcached, Scalaris

Scanned by CamScanner

2) Column oriented -

- it works on column
- Stores data from each column seperatly instead of rows
- Good example of indexing
- it allow for quicker scan for only small no. of column
- it is easy to add new column but adding entire record is difficult.
- it delivers high Performance on aggregation queries as somm, COUNT, AVG, MIN etc

Ex — Hbase, Cassandra, Hypertable etc

- The concept was ~~given~~ taken from google Big table

3) Graph Databases -

- Comparatively complex type
- it stores entities along with relations among those entities
- The entities are stored as a node with relationship as edges
- Every node and edge has unique identifier
- Traversing relationship is faster.
- mostly used for social network, logistics spatial data.

Ex — Neo4J, infinite Graph, Orient DB, FlockD

Scanned by CamScanner

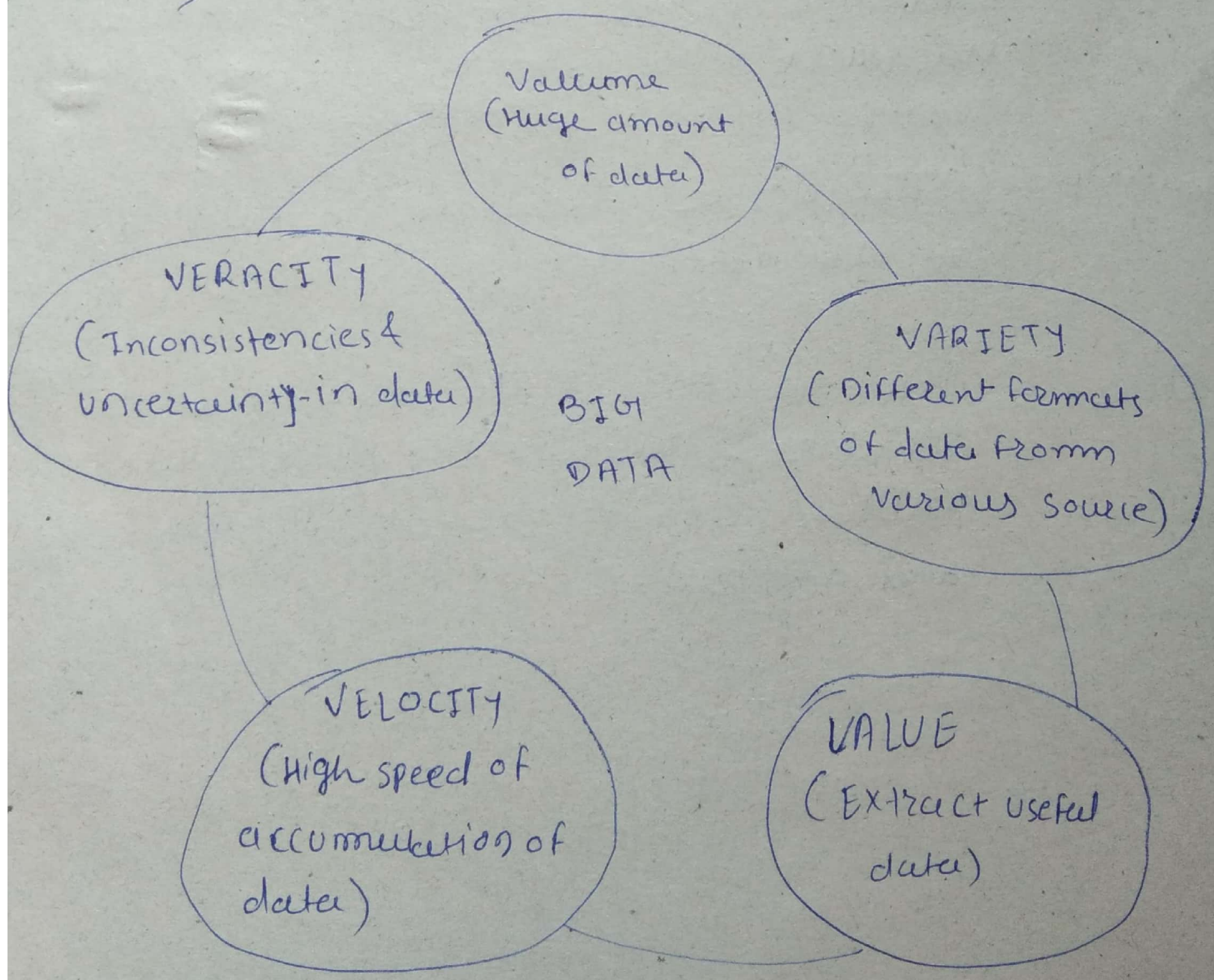
4) Document Oriented -

- one step complex from key-value stores
- it care little about normalization as long as the data is in a structure that makes sense
- Document have number of key value pairs
- JSON or XML or BSON format are majorly used to stored such -key-value pairs

Ex — MangoDB (JSON), Couch DB, Lotus Notes etc.

Q.3 Explain Characteristics of Big data (5 V's)

→



Scanned by CamScanner

Q 1) Volume - Unimaginable amount of data generated every second by multiple sources i.e extremely larg amount of data is generated & need to be handle by different big data technologies

2) Velocity - It measures How fast data is Produced and modified and the speed with which it needs to Processed

3) Variety - Big-Data is generated in multiple Varieties. Compared with traditional data like Phone numbers, addresses the latest data is in the form of Image, videos, text documents etc

4) Veracity - The degree of reliability that the data has to offer

5) Value - Need to concentrate on valuable, reliable & trustworthy data that further needs to be stored, Processed and analyzed.

Scanned by CamScanner

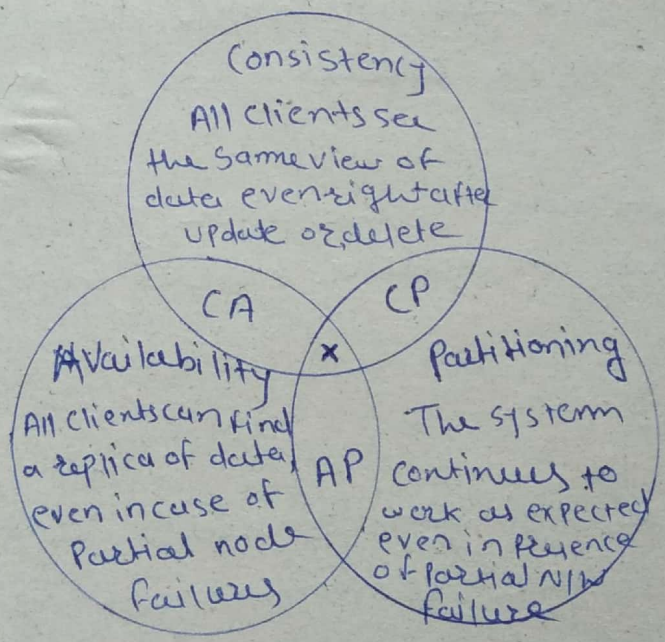
Q.4 Explain CAP theorem

→

CAP Theorem -

- Also called as Brewer's theorm
- It states that it is impossible for a distributed data store to offer more that two out of three qua

1) Consistency
2) Availability
3) Partition Tolerance



Scanned by CamScanner

1) Consistency -

- The data should remain Consistent after any write operation is performed.

- Database changes are Propagated to all nodes

- Performing a read operation will return the value of the most recent write operation causing all nodes to return the same data

2) Availability -

- The database should always be available and responsive. it should not have downtime

- System remains operational 100% of the time

- Every client gets a response, regardless of the states of any individual node in the system.

3) Partition Tolerance -

- it means the system should continue to function even if the communication among servers is not stable

- That the network will not stop functioning even with partitions.

- A lost or temporarily delayed connection between two nodes

- Partitions tolerance means that the cluster must continue to work despit any

Scanned by CamScanner

number of communication break downs between nodes in the system

Q.5. Differentiate between RDBMS and NoSQL. What are different types in which data is stored in NoSQL database?

→

RDBMS	NoSQL
1) RDBMS stands for Relational Database Management System	No-SQL stands for No-Structural Query Language
2) In RDBMS use only Structured data i.e Relational Database	2) In this use un-Structured data i.e Distributed Database
3) it is vertically scalable	3) it is Horizontally Scalable
4) fixed or Predefined schema	4) Dynamic schema.
5) Not Suitable for hierarchical data storage	5) Best suitable for hierarchical data storage
6) Canbe used complex Queries	6) Not good for complex Queries
7) it is used for small-data sets	7) it is used for Big-data set.

Scanned by CamScanner

* different type in which data is stored in NO-SQL

1) key-value pair Based
2) Column-oriented Graph
3) Graph based.
4) Document-oriented

Scanned by CamScanner

Q.7 State and explain different fallbacks of traditional RDBMS in handling and Processing Big Data Provide solution to overcome fallbacks of RDBMS

→

Fall Backs of traditional RDBMS

- Traditional RDBMS is used only to manage structured and semi-structured data; maximum big data is in unstructured ~~data~~ form

- Big data is generated at a very high speed RDBMS lacks in high velocity because it's designed to handle steady data

Big Data Provide Solution

1) No SQL
2) Hadoop
3) Map-Reduce

Scanned by CamScanner

Q.6 Explain how big data is handled using any real world Example

Scanned by CamScanner

Q.8 - Explain different Challenges in handling big data

→ *) Big Data Challenges

1) Dealing with data growth
- Data is increasing exponentially
- Data generation is sky-rocketing
- Challenge is not to store but to manage that data

2) Real-time can be compleze
- Data is not "stagnant"
- Manage every second generated data

3) Data security
- Data collected from multiple sources may not be trustworthy, so it can impact on organi-zational data

4) Talent gap in Big Data
- Lack of skills available in market for big data technologies

5) Syncing across data sources
- Data gathered from multiple resources must be synched (i.e data must have same data definition, metadata etc)

6) Spending a Huge Amount of Money
- on-Premise deployment
- cloud deployment

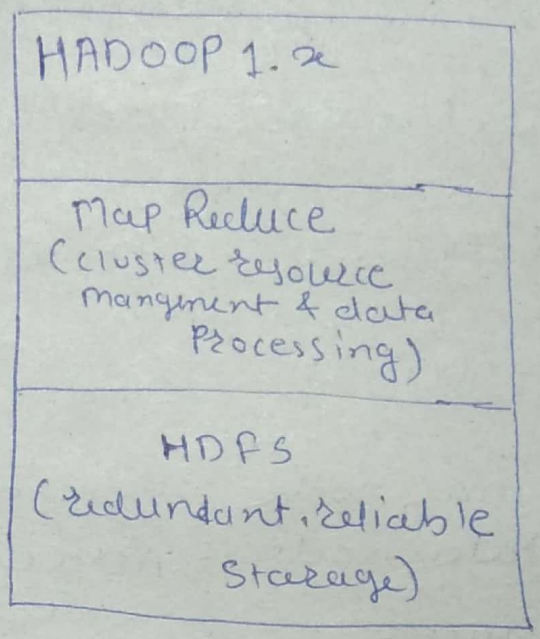
7) Scaling up and down big data according to current demand.

Scanned by CamScanner

unit - 4

Q.1 Explain Hadoop 1.x ecosystem

→



- Apache Hadoop V.1.x has the following two major components
  1) HDFS (HDFS V1)
  2) MapReduces (MR V1)
- in Hadoop V.1.x these two are also known as Two pillars of Hadoop
- Hadoop 1.x framework Supports only Map Reduce Processing (MR) tool and does not Support any other non-MapReduce tools.
- There is no separate setup to do the resource Management. MR does both data Processing & cluster resource management.

Scanned by CamScanner

- Scalability is ~~1 ~~ 4000 nodes per cluster
- entire name space manage By the single Name node
- Name Node manage metadata of cluster
- Name Node failure affects the stack

* Hadoop Ecosystem 1.0 (1.x)

Core Component.

1) HDFS
   a) Name Node - Meta data of Data node
   b) Data Node - Actual data.
2) Map Reduce -
   a) Job Tracker
   b) Task Tracker

1) HDFS -

- HDFS is a Hadoop Distributed file System where our BigData is stored using Commodity Hardware.

- it is designed to work with large Datasets with default block size is 64-MB (we can changes it as per our project requirements)

- HDFS Component is again divided into two Sub-Components:

Scanned by CamScanner

a) Name Node -

it is placed in master node. it is used to store meta data about data nodes e.g. How many data blocks are stored in data nodes. Which data nodes have data, slave node details, data nodes locations. timestamps etc

b) Data Node -

it is located in slaves. It is used to store applications actual data. it stores data in data slots of size 64-MB default size

2) Map-Reduce -

- Batch Processing Prog model.
- it is used for file Processing in Hadoop
- it has two sub-components

a) Job Tracker -

- it is used to assign map-Reduce tasks To Task Trackers in the cluster of nodes.

- Sometimes, it assign same task to other Task tracker if Previous Task Tracker has faild.

Scanned by CamScanner

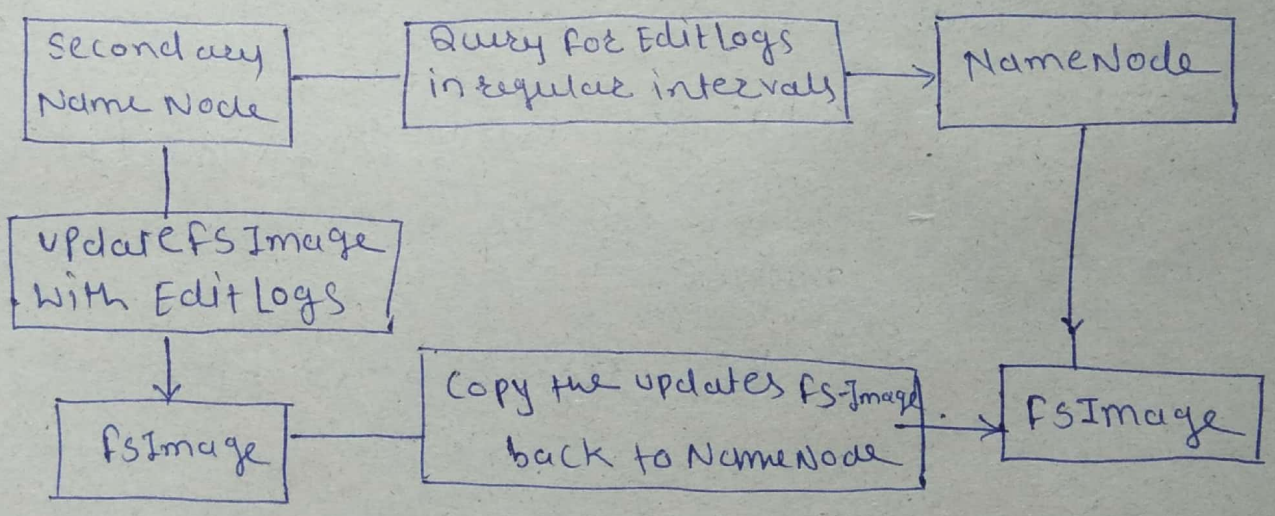
- Job Tracker maintain status of all Task tracker.

b) Task tracker -

- it is located in slave node ~~it is~~
- it executes the tasks which are assigned by job Tracker and sends the status of those tasks to Job Tracker.

Q.2 Explain HDFS Architecture.

→



- Data needs to be stored in HDFS is first divided into data blocks. Size of data block can be 64-MB (by default, which can be configured)

Scanned by CamScanner

1) Name Node -

- it maintains & manages data nodes
- it maintain metadata about data stored among different data nodes. like -
Location of blocks, the size of files, permission, hierarchy etc.

- Two files are associated with metadata -

a) fsimage - it contains complete status of file system from the start of name node

b) EditLogs - it contains recent modification made to file system, which wil eventually be updated in FSImage
Ex - if file / data block is deleted the namenode will immediately record this in EditLogs

- it regularly receives Heart Beat & a block report to ensure data node are alive

- it ~~regularly~~ is responsible to take care of replication factor.

Scanned by CamScanner

2) Data Node -

- The actual data is stored on DataNodes.
- The DataNodes perform the low-level read & write requests from the file system clients
- They send heart beats to the NameNodes periodically to report the overall health of HDFS, by default, this frequency is set to 3 seconds.

3) Secondary NameNode -

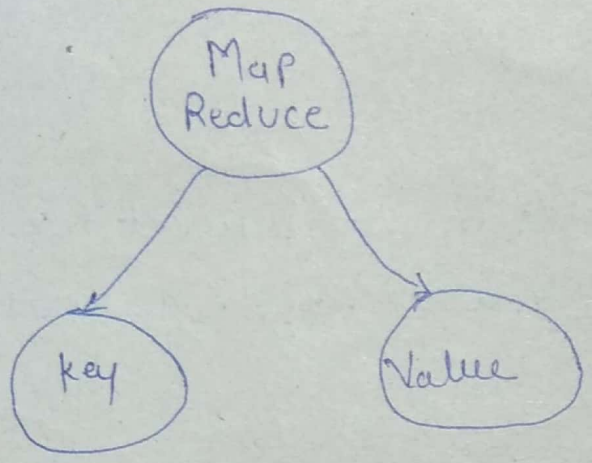
- The secondary NameNode is one which constantly reads all the file systems and metadata from the RAM of the NameNode and writes it into the hard disk or the file system.
- it is responsible for combining the Edit logs with FSImage from the NameNode

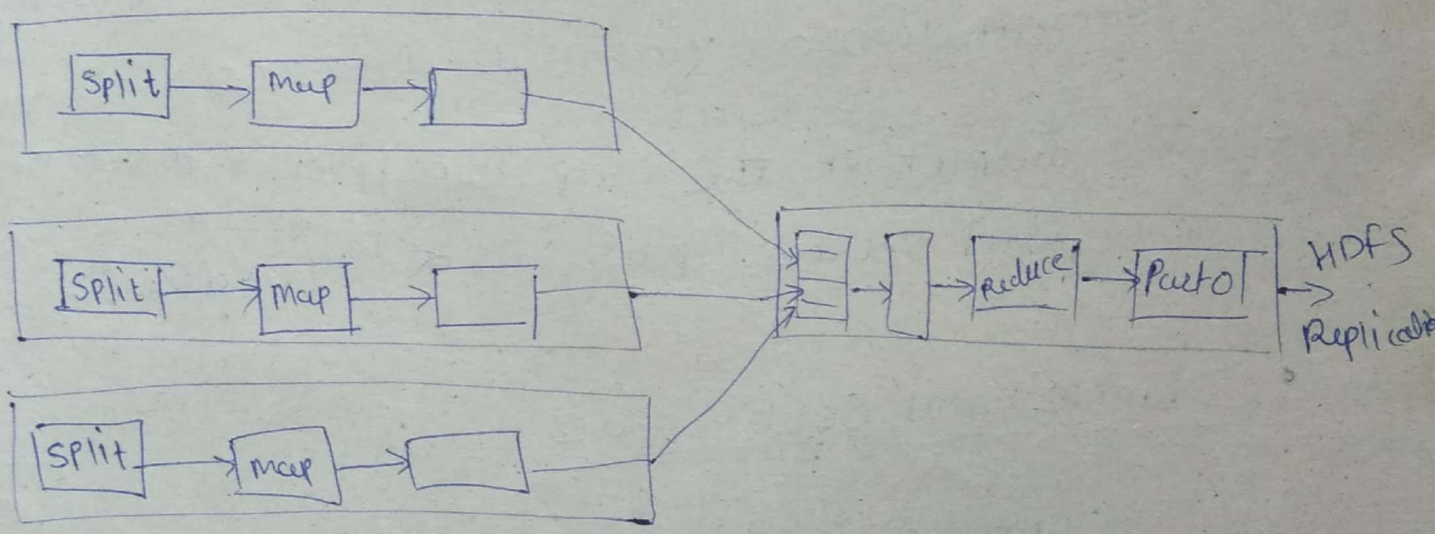
* Namenode and DataNode are java processes that run on the master slave machines respectively.

Scanned by CamScanner

Q 3 Explain Map Reduce architecture

→





* Phase in Map Reduce Algorithm

- Map Reduce is a distributed Data Processing algorithm introduced by Google

- In MapReduce, each phase has key-Value Pairs as input & output.

Scanned by CamScanner

* Map Phase

- first step in Map Reduce algorithm
- Two Sub steps :

• Splitting –

Takes input dataset from source and divide into smaller Sub-datasets

• Mapping –

Takes smaller sub-datasets as an input and Perform required action or computation on each Sub-datasets

- The output of the Map function is a set of key and value pairs as <key, Value>

• Shuffle and Sort Phase –

Shuffling is the grouping of the data from various nodes based on the key

- Two Sub Steps –

• Merging –

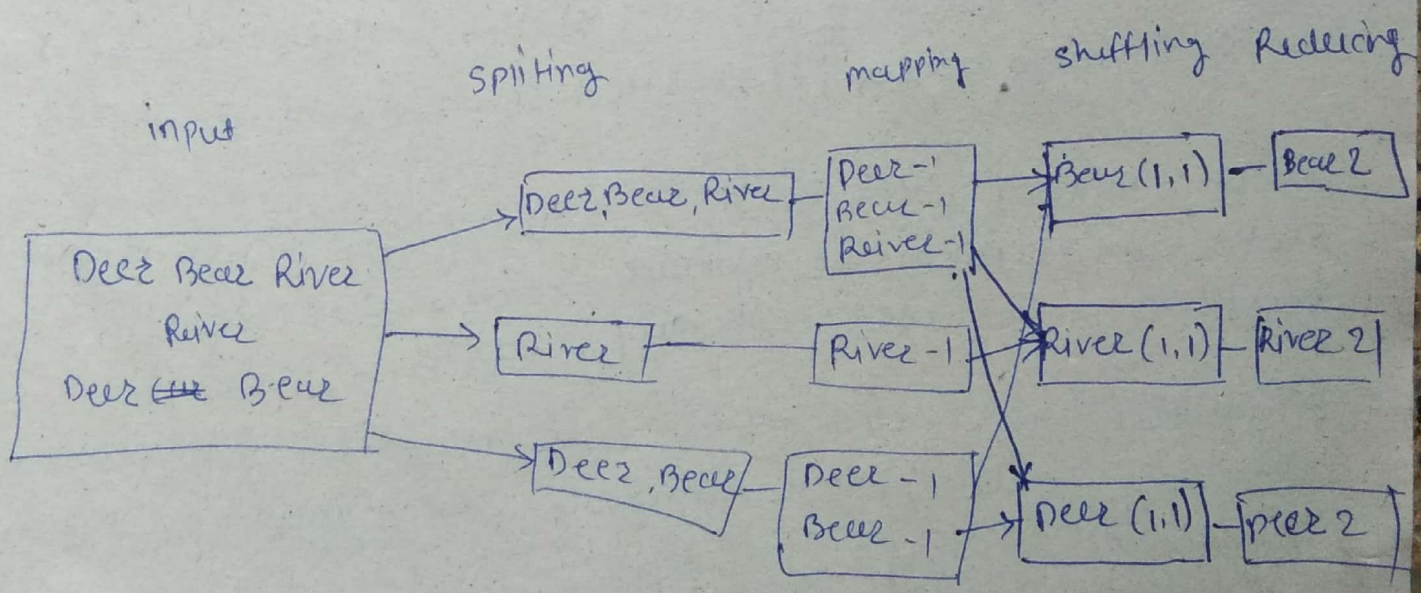
Combines all key-value pairs which have same keys and return <key, List <Value>>

Scanned by CamScanner

- **Sorting** –
  Takes output from Merging step and sort all key-value pairs by using keys.
  This step also returns <key, List<value>> output but with sorted key-value pairs

* Reducer Phase

- Reduce Phase is the final step in MapReduce Algorithm
- Reducer combines all these values together & provide single output value for the specific key
- This phase performs only one step - Reduce step.



Scanned by CamScanner

Q.4 Differentiate between Hadoop 1.x and Hadoop 2.x ecosystem

→

Hadoop 1.x	Hadoop 2.x
1) Support MapReduce (MR) Processing Model only. Does not support Non-MR tools	1) Allows to work in MR as well as other distributed Computing models like Spark, Hama, Giraph, Message Passing Interface) MPI & HBase Co Processor
2) MR does both Processing and Cluster resource management	2) YARN (Yet Another Resource Negotiator) does cluster resource management & Processing is done using different processing models
3) Has limited Scaling of nodes Limited 4000 nodes per cluster	3) Has better Scalability Scalable up to 40000 nodes per cluster
4) Works on concepts of slots - slots can run either a map task or a Reduce task only	4) Work on concepts containers containers can run generic tasks
5) A single Namenode to manage the entire name space	5) multiple namenode servers manage multiple Namespaces
6) Has single-point-of-failure (SPOF) because of single Namenode and in the case of Namenode failure, needs manual intervention to overcome	6) Has to feature to overcome SPOF with a standby name node & in the case of Name node failure, it is configured for automatic recovery.

Scanned by CamScanner

1) component used in H.1.x has a HDFS & map Reduce

2) Daemons H.1.x
1) Name Node
2) Data Node
3) secondary name Node
4) Job tracker
Task Tracker

3) H 1.x in this map Reduce performe two task
Data mangment +
Data Processing

— HDFS [Data stored]

4) Low performance

5) it has Node ~~clust~~ limitation in cluster
work on [GBS)

2) Component used in H2.x
HDFS & YARN

2) H.2.x
1) name Node
2) Data Node
3) secondary namenode
4) Resource manager
5) Node manager

4) H2.x in this map Reduce
- HDF (Data stored)
- YARN (Resource Mangment
Resources Availability Check.
for Run Task.

5) High Performance

6) it does not has a limitation
(Tb's or Pb's)

Scanned by CamScanner

Q.5 Explain YARN architechture

→

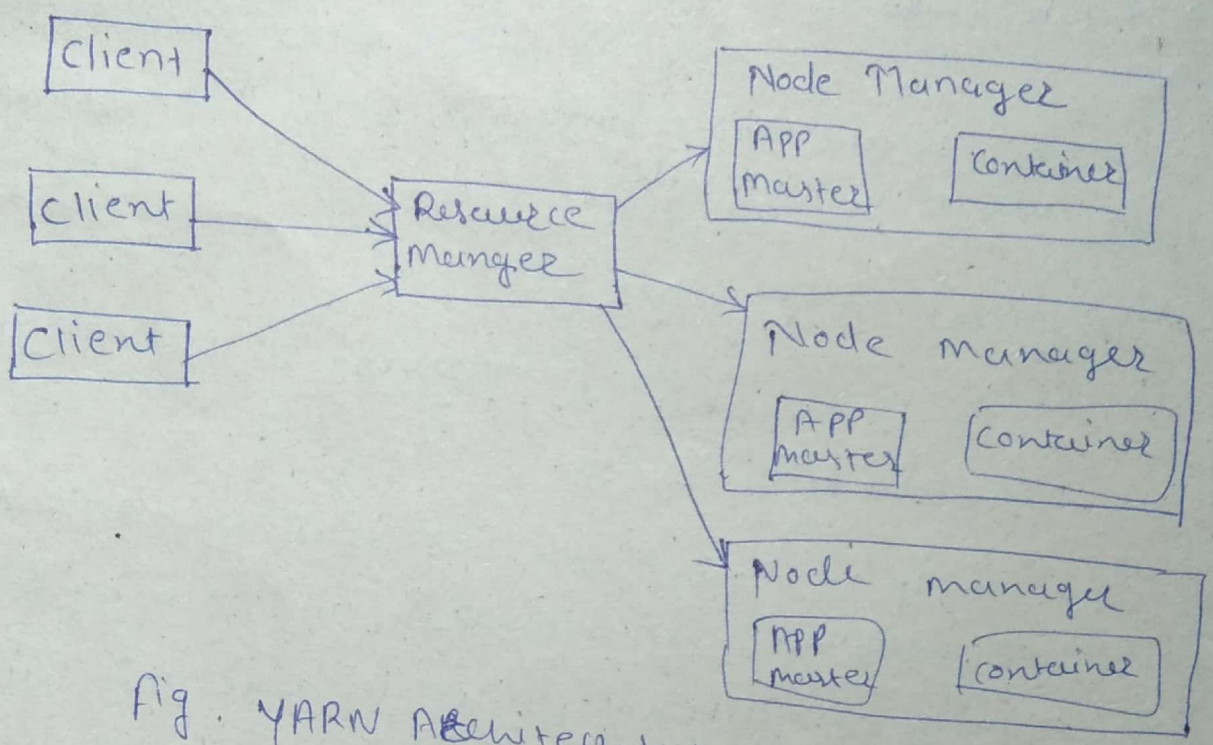


fig. YARN Architecture

1) Resource Manager -

- is a per-cluster level component
- it has all the information about available resources (i.e CPU, memory) with all slave nodes
- it manage assignments of resources among all competing Applications
- Stand by Resources Manager can be available in case of failure

Scanned by CamScanner

3) Application Manager

* Sub components -

a) Job Scheduler -

Allocate resources to each slave node (on which appl^n is running) with the given scheduling method.

b) Application Manager -

- To monitor the progress of submitted application like Map Reduce job

- it is responsible for starting application master (available on slave node), monitoring & restarting them on different nodes in case of failures

2) Node Manager -

- Is the per machine / node framework who is responsible for containers monitoring their resource usage (CPU, memory, disk, network & reporting the same to the Resource Manager

- Primary goal is to launch and manage application containers assigned to it by the resource manager.

Scanned by CamScanner

3) Application Manager –

- Dynamically lunched one per application. it is short life. once application is executed it will be dynamically destroyed
- Request for new container to resource manager is initiated by Application master.
- Functionality of Application master depends differently as compare to spark application master.

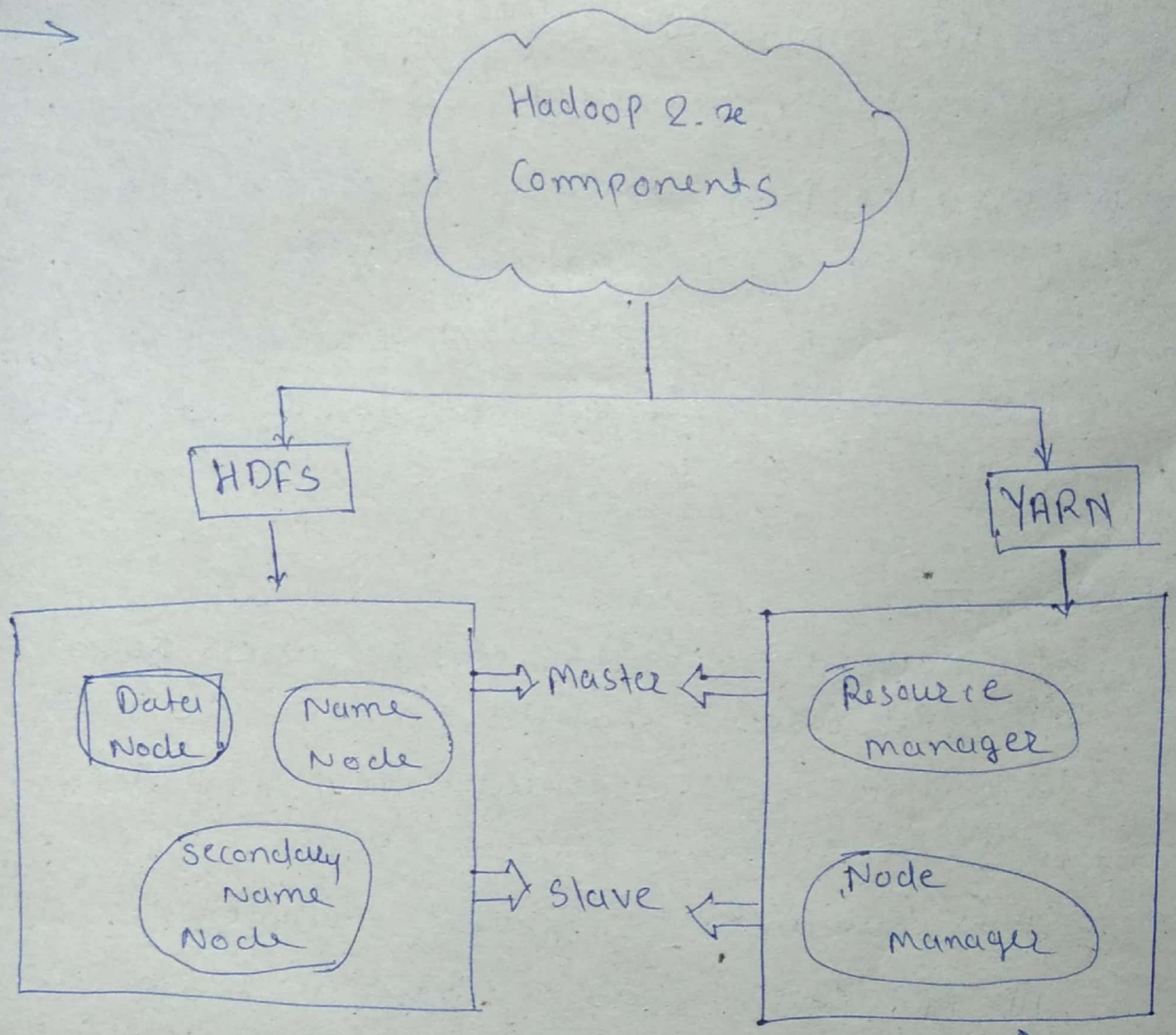
4) Container –

- it is collection of physical resources such as RAM, CPU cores and disks on a single nodes.
- it is a place where actual execution is done
- it is supervised by Node manager
- it is sheduled by Resource manager
- An Application / job will run on one or more containers
- In container single task (a unit of application / job) is executed.

Scanned by CamScanner

Q.6 Explain ~~Hadoop 2.x eco~~ core component of Hadoop 2.x.

→



These were called as Job Tracker & Task Tracker in Hadoop 1.x system

Scanned by CamScanner

1) HDFS -

- HDFS is used for Storage ~~and Manage~~ system
- HDFS holds very large amount of data and provides easier access.
- To store such huge data the files are stored across multiple machines. i-e mean in distributed (fusion) manner
- These files are stored in redundant fusion to rescue the systems from possible data losses in case of failure.
- HDFS also makes applications available to parallel processing

Sub Component

a) Data Node -

- The actual data is stored on Data Nodes
- The Data Nodes perform the Low-Level read and write requests from the file system's clients.
- They send heartbeats to the Name Node periodically to report the overall health of HDFS, by default, this frequency is set to 3 seconds.

Scanned by CamScanner

b) Name Node -

- it maintains and manages data nodes
- it maintains metadata about data stored among different data nodes. metadata means information about - the location of blocks stored, the size of files, permission hierarchy etc.
- it is responsible to take care of replication factor.

c) Secondary Name Node -

- The Secondary NameNode is one which constantly reads all the file systems & metadata from the RAM of the NameNode and writes it into the hard disk or the file system
- it is responsible for combining the Edit Logs with fs Image from the Name Node

2) YARN -

- Brain of Hadoop ecosystem. It perform all processing activities by allocating resources and scheduling tasks.
- YARN monitors different data processing engines like graph processing, interactive processing to run and process data stored in HDFS
- It is a processing system

Scanned by CamScanner

* Sub Components

a) Resource Manager -

- is a per-cluster level component
- it has all the information about available resources (i.e CPU, memory) with all slave nodes.
- it manages assignments of resources among all competing applications
- Stand by Resource manager can be available in case of failure

b) Node Manager -

- is the per machine/node framework who is responsible for container, monitoring their resource usage (CPU, memory, disk network) and reporting the same to the Resource manager
- Primary goal is to launch and manage application containers assigned to it by the resource manager

Scanned by CamScanner

c) Job Tracker -

- Any operation can be considered as Job. example - Read a text file is a job. This is handle by Job Tracker.

d) Task Tracker -

- A job can have many tasks. Like connection to file is on of the task, Reading the data is other task, displaying / Processing the data is another task. These are managed by Task Tracker.

Scanned by CamScanner

Q.7 Explain Hadoop 2.x ecosystem

Scanned by CamScanner

Q.8 What is Hadoop ? Explain different Hadoop key Characteristics

→

Hadoop -

- It is an open source framework for writing and running distributed applications that process large amount of data.

- It allow to store Big Data in a distributed environment, so that it can be processed in parallel

- Bassically 2 components in Hadoop -

1) HDFS
2) YARN

* Hadoop key Characteristics

1) Accessible -

Hadoop runs on Large clusters of commodity machines or on cloud computing services such as Amazon's Elastic compute cloud [EC2]

2) Robust -

Because it is intended to run on the assumption of frequent hardware malfunctions

Scanned by CamScanner

3) Scalable -

Hadoop Scales linearly to handle larger data by adding more nodes to the cluster

4) simple - Hadoop allows users to quickly write efficient parallel code.

5) faster - Hadoop is extremely good at high volume batch processing because of it's ability to do parallel processing.

6) fault Talerance - The data sent to one individual node and the same data also replicates on other nodes in the same cluster.

7) Reliability -

Due to data replication in the cluster, data is reliably stored on the cluster of machine despite machine failure

8) flexibility -

Hadoop manages data whether structure or unstructured, encoded or formatted, or any other type of data.

Scanned by CamScanner

Q.9 State difference between RDBMS and Hadoop

→

RDBMS	Hadoop
1) Mainly for structured data	1) used for structured, semi structured and unstructured data
2) Average size of data storage is in (GBS)	2) used for large data set (Tbs and Pbs)
3) SQL Language is used	3) HQL (Hive Query Language) is used.
4) Required on write (static schema)	4) Required on reading (dynamic schema)
5) ~~Read~~ Data Read are fast	5) Both reads and writes are fast
6) work on Relational Table	6) work on key/value pair.
7) it is vertically scalable	7) it is Horizontaly scalable
8) Throughput is low	8) Throughput is High.
9) Integrity is High	9) Integrity is low

Scanned by CamScanner

Q.10 Explain brief history of hadoop

→

- Project 'Nutch' was created to handle billions of searches and indexing millions of web pages

- in oct 2003 google releases a paper on google file system (GFS)

- 2004 - Google Releases a paper on MapReduce

- 2005 : Nutch used GFS and MapReduce

- 2006 : Yahoo created Hadoop based on GFS and MapReduce

  - Yahoo deploys 300 machines & within this year reaches 600 machines

- 2007 : - Yahoo Started using Hadoop on a 1000 node cluster

  - Hadoop includes HBase

  [- Yahoo runs 2 clusters of 1000 machine]

- 2009 : - Hadoop becomes capable enough to sort a peta byte

  - Map Reduce & HDFS become separate subproject

Scanned by CamScanner

2010 :- Apache Hive & Pig released

2011 - Apache zookeeper released.

2012 - Apache Hadoop 1.0 version Released

2013 - Apache Hadoop 2.2 version Realeased.

2014 - Apache Hadoop 2.6 version Realeased

2015 - Apache Hadoop 2.7 version Realeased.

2017 - Apache Hadoop 3.0 version Realeased

2018 - Apache Hadoop 3.1 version Realeased.

Scanned by CamScanner

Q.11 State advantages of YARN Component.

→

1) It Provides central resource management.

2) cluster utilization -

- YARN allows to use the Hadoop cluster in dynamic way, rather than in static manner

3) Scalability -

- it Provides Power of scalability to the Hadoop cluster. YARN'S RM (Resource Manager) is the central Controling Authority for resource management and allocation decisions

4) compatibility -

- YARN Provides compatibility to those Projects that are working with MapReduce in Hadoop 1.0 can easily move on to Hadoop 2.0

5) YARN allows to Process data using different Processing engines (e.x - Spark) apart from jus MapReduce.

Scanned by CamScanner

* pipeline.

1) data Extraction
2) Row Process → process Data
3) Data Rangling → meaniful format
4) Data Analysis → Analysis
5) Data visualization → graphy&all format Represent.

Scanned by CamScanner