Page No.

Date: / /20

Sheet 1 -:BDA:-

Q.1 What is Big Data & its key characteristics?

Ans:- Big data is high-volume, high-velocity & high-variety info assets that demand cost-effective.

Or :-

The defination of big data is data that contains greater variety, arriving in increasing volumes & with more velocity. This is also known as the three Vs.

characteristics :-

1. volume
2. velocity
3. variety
4. veracity
5. value

1. volume :-

volume, the first of the 5 V's of big data, refers to the amount of data that exists. volume is like the base of big data, as it is the initial size & amount of data that is collected.

2. velocity :-

The next of the 5 V's of big data is velocity. It refers to how quickly data is generated & how quickly that data moves. for eg in health-carce, there are many medical devices made to monitor ~~paiten~~ patients & collect data.

3. variety :-

The next v in the five 5 V's of big data is variety. variety refers to the diversity of data types.

4. veracity :-

veracity is the fourth v in the 5 v's of big data. It refers to the quality & accuracy of data. for example, concerning the medical field, if data about what drugs & patient is taking is incomplete, then the patient's life may be endangered.

5. value:-

The last v in the 5 v's of big data is value. This refers to the value that big data can provide, & it relates directly to what organizations can do with that collected data.

SQ/LQ.2. Explain the following big data tools.
↳ List of the various tools available for big data.

Ans :- 1. NOSQL
2. MapReduce
3. Storage
4. Servers
5. processing

NoSQL :-

NoSQL database is used to refer a non-SQL or non relational database. NoSQL database doesn't use tables for storing data. It is generally used to store big data & real-time web applications.

→ It supports query language.
→ It provides fast performance.
→ It provides horizontal scalability.
→ DatabasesMangoDB, CouchDB, Big Table, Redis, Zookeper.

MAX

MapReduce :-

MapReduece is a programming model for writing application that can process big data in parallel on multiple nodes.

→ Hadoop, Hive, pig, caffeine, S4, MapR.

Storage :-

Big data storage is a storage infrastructure that is designed specifically to store, manage & retrieve massive amounts of data or big data.

→ S3, HDFS, GIDFS

Servers :-

Big data servers are dedicated servers configured for working with big data.

A big data server must have high processing, power for storage, retrival & analytics.

→ EC2, Google app Engine, elastic

processing :-

Big data processing is a set of techniques or programming models to access large-scale data to extract useful info. for supporting & providing decisions.

→ R, Yahoo! pipes, mechanical, Turk, Datameer, Bigsheets.

S.Q. 3. Define data & its importance.

Ans Data is defined as facts or figures, or information that's stored in or used by a computer. An eg of data is info. collected for a research paper. An e.g of data is email.

importance :-

Data helps in make better decisions.
Data helps one improve processes.
Data helps one understand consumers & the market

S.Q 4. Define data analytics & its type.

Ans Data analytics is the process of examining data sets in order to find trends & draw conclusion about the info. they contain.

Types :-

1. Descriptive
2. Diagnostic
3. prescriptive
4. predictive

1. Descriptive :-

Descriptive analytics is one of the most common forms of analytics that companies use to stay updated on current trends & the company's operational performance.

e.g :- Data Queries, Reports, Data dashboard etc.

2. Diagnostic :-

Diagnostic analytic is one of the more advanced types of big data analytics that you can use to investigate data & content.

→ it uses techniques such as:

Data discovery, Data mining etc..

3. Prescriptive Analytics :-
Set of techniques to indicate the best course of action.
It tells what decision to make to optimize the outcome.
The goal of prescriptive analytics is to enable:
Quality improvements
Service enhancements
cost reduction
Increasing productivity

4. predictive :-
As the name suggests, this type of data analytics is all about making predictions about future outcomes based on insights from data.
Techniques :- Linear regression
Time series analysis & forecasting
Data mining.

9Q 5. Justify the importance of Data Analytics.?
ns :- → Determining credit risk
→ Developing new medicines
→ Finding more efficient ways to deliver products & services.
→ preventing fraud
→ uncovering cyber threats
→ Retaining the most valuable customers.

L.Q.5. Explain the 4 level of Data.

Ans 1. Nominal - lowest level of measurement.
2. Ordinal
3. Interval
4. Ratio - Highest level of measurement.

1. Nominal :-

Nominal values represent discreate units & are used to lebal variables, that have no quantitative value.

e.g:- ~~nated~~ most common examples include male/female, hair color, nationalities, name of people.

2. Ordinal :-

Ordinal values represent discreate & ordered units. It is therefore nearly the same as nominal data, except that it's ordering matteres.

e.g:- Example of Ordinal values included :-
Socio economic status ("low income", "middle income", "high income") etc.

3. Interval :-

An interval is an ordered scale in which the diff" bet" measurements is a meaingful quantity but the measurements do not have a true zero point.

e.g:- Temperature in Fahernheit & celcius
Year

4. Ratio :-

A ratio scale is an ordereed scale in which the diff" bet" the measurements is a meaingful quantity

and the measurements have a true zero point.
e.g:- weight, Age, salary.

Q.6 Define Data Analytics. Also explain the similarities and/or difference between Analysis & Analytics.

Ans Data Analytics :-

Data analytics is the process of examining data sets in order to find trends & draw conclusion about the info. they contain.

~~Similarities~~

Difference :-

	Data Analytics	Data Analysis
1.	It is described as a traditional form or generic form of analytics	It is described as a particularized form a analytics.
2.	It includes several stages like the collection of data & then the inspection of business data is done.	To process data, firstly, raw data is defined in a meaniful manner, then data cleaning & conversion are done to get meaningful info. from raw data.
3.	It supports decision making by analyzing enterprise data.	It analyzes the data by focusing on insights into business data.
4.	~~Such~~ It uses various tools to process data such as python, Excel etc.	It uses diff" tools to analyze data such as Rapid Miner, Node XL etc.

Data Analytics	Data Analysis
5. Descriptive analysis cannot be performed on this.	5. Descriptive analysis can be performed on this.
6. Relations with the help of this.	6. One cannot find anonymous relations with the help of this.
7. It does not deal with inferential analysis.	7. It supports inferential analysis.

Similarities :-

Slide-3

S.Q 1. with example, differentiate sample & population of data involved in a use-case.

	sample	population
1.	part of the group.	1. whole group
2.	Groupe we do know about.	2. Group we want to know about.
3.	characteristics are called statistics.	3. characteristics are called parameters.
4.	statistic are always known.	4. parameters are generally unknown.
5.	statistics change with the sample.	5. parameters are called fixed.
6.	e.g:- All students studying in class XII is a sample, whereas those students belong to a given school is population.	6. All Tv viewers in the country / world.
7.	Normally, a sample is obtained in such a way as to be representative of the population.	7. All people in the country/ world is not a population!

S.Q 2. Defining Statistical Inference.

Ans :- Statistical Inference :- (e.g problem Q)

Statistical inference is the process of using sample statistisad to make decisions about population.

e.g :- In the content of TRP overall frequency of the various of happiness.

L.Q 1. Consider the following data (in increasing order)

13, 16, 16, 16, 17, 18, 23, 27, 34, 34, 35, 36, 40, 40, 41, 41, 41, 44, 44, 48, 58, 61, 67, 70, 70, 75, 77, 81, 82, 84, 90.

(a) what is the mean of the data? what is the median?

Slide – 1

L.Q. 1:- Discuss the database ~~the advantages of adopting~~ transparency & its type in distributed database systems.

Ans :- Data Transparency is the characteristics of data being used with integrity, lawfully, fairly & traceably, for valid purpose.

Individuals & businesses should know that data is being collected, who can access it, how it's being used & how they can interact with it.

OR

Transparency :-

Transparency is the separation of the higher level semantics of a system from the lower level implementation issues.

Fundamental issue is to provide data independence in the distributed environment.

There are 3 types of transparency

1. Network Transparency (Distribution)
2. Replication "
3. Fragmentation "

1. Network Transparency :-

Network transparency is the process of sending or accessing data over a network in such a way that the information is not visible to user.

communicating with a local or remote host, System network or software.

It can provide remote data & computing resources to a local user without providing intermediate network info.

2. Replication Transparency :-

Replication Transparency is the ability to create multiple copies of objects without any effect of the replication seen by applications that use the objects.

It should not be possible for an application to determine the no. of replicas, or to be see the identities of specific replica instances.

3. Fragmentation Transparency :-

Fragmentation is a database server feature that allows you to control where data is stored at the table level.

Fragmentation enables you to define groups of rows or index keys within a table according to some algorithms or scheme.

L.Q 2. With examples discuss the database fragmentation types, Also, list the rules to checks the fragments correctness.

Ans :- Database fragmentation :-

Fragmentation is a database server feature that allows you to control where data is stored at the tabel level.

e.g :- Account (Acc-No, Balance, Branch-Name, Type)

Types :-

There are 3 types of fragmentation.

1. Horizontal fragmentation
2. vertical "
3. Hybrid "

1. Horizontal fragmentation :-

Horizontal fragmentation refers to the process of dividing a table horizontally by assigning each row or (a group of rows) of relation to one or more fragments.

These fragments are then be assigned to different sides in the distributed system.

[e.g :- slide no. 1 page no. 18]

2. vertical fragmentation :-

vertical fragmentation refers to the process of decomposing a table vertically by attributes are columns.

In this fragmentation, some of the attributes are stored in one system & the rest are stored in other systems.

[e.g :- slide no. 1 page no. 19]

3. Hybrid fragmentation :-

The combination of vertical fragmentation of a table followed by further horizontal fragmentation of some fragments is called mixed or hybrid fragmentation.

For defining this type of fragmentation we use the SELECT & the PROJECT operations of relational algebra.

[e.g :- slide no. 1 page no. 20]

checks the fragments correctness - [slide no. 1 page no. 21]

Rules

1. completeness
2. Reconstruction
3. Disjointness

L.Q 3. With proper syntax & example discuss the following Relational operators:
(a) selection
(b) projection
(c) Union
(d) Intersection
(e) set difference
(f) cartesian product
(g) Join

Ans: (a) selection :—

selection operator (σ) is a unary operator in relational algebra that performs a selection operation.

Syntax:- $\sigma_F(R) = \{t \in R \mid F(t) \text{ is true}\}$

where,

R is a relation,

t " " tuple variable

F " " formula consisting.

eg:- select tuples from a relation "Books" where subject is "database"

$\sigma_{subject = "database"}(Books)$

→ or

produce a horizontal subset of the operand relation.

(b) projection :-

produce a vertical slice of a relation

Syntax :-

$\Pi_{A1,\ldots,An} = \{t[A_1, \ldots A_n] \mid t \in R\}$

where,

R is a relation

t " " tuples variable

Date: / /20

{A1,...,An} is a subset of attributes of R over which the projection will be performed.]
[eg. slide no. 1 page no. 37

(c) Union :-

Similar to set union

Syntax :-

$$R \cup S = \{ t \mid t \in R \text{ or } t \in S \}$$

where

R & S are relations,
t is tuple variable.

eg :- Table 1 :-

Reg no.	Branch	Section
1	CSE	A
2	CIVIL	B
3	ECE	C
4	MECH	D

Table 2 :-

Reg no.	Branch	Section
1	CIVIL	A B
2	CSE	B A
3	EEE	C C

To display all the reg no. of Table 1 & 2.
command is reg no. (Table 1) ∪ (Table 2)

output

(T1) Reg no.	(T2) Branch	Section
1	CSE	A
2	CIVIL	B
3	ECE	C
4	MECH	D

OR

Select Student_Name from Sci_Students.
UNION
Select Student_Name from Dance._Students.

(d) Intersection :-

Intersection ~~on~~ typical set intersection operator is denoted by ∩ symbol & it is used to select common rows (tuples) from two tables relations.

Syntax :-

$$R \cap S = \{ t \mid t \in R \ \& \ t \in S \}$$
$$= R - (R - S)$$

R, S union - compatible

e.g :-

Select student_name from sci_students
INTERSECT
Select student_Name from Dance_students

(e) Set difference :-

The set difference operators takes the two sets & returns the value that are in the first set but not the second set.

Syntax :-

$$R - S = \{ t \mid t \in R \text{ and } t \notin S \}$$

e.g :- Select student_Name from sci_students
MINUS
Select student_Name from Dance_Student

(f) cartesian product :-

Given Relations
R of degree k_1, cardinality n_1
S " " k_2, " n_2

Syntax :-

$$R \times S = \{ t[A_1, \ldots, A_{k_1}, A_{k_1+1}, \ldots, A_{k_1+k_2}] \mid t[A_1, \ldots, A_{k_1}] \in R \text{ and } t[A_{k_1+1}, \ldots, A_{k_1+k_2}] \in S \}$$

e.g :-

Date : / /20

(g) Join :-

Join operation combines the rel^n R_1 & R_2 with respect to a condition.

Syntax :-

$$R \bowtie_{F(R.Ai, S.Bj)} = S$$

where, Ro

R, S are relations

F(R.Ai, S.Bj) is a formula defined as that of selection

A derivative of cartesian product

$$R \bowtie_F S = \sigma_F (R \times S)$$

e.g :- — — —

L.Q.4:- what is HDFS? Discuss the Hadoop Components & its Architecture.