GoodLuck | Page No.
Date

Assignment: I.

1] What is Data?

Data is information that has been translated into a form that is efficient for movement or processing. Data can be in the form of text, figures, ~~imges~~ images, numbers, graphs or symbols. Eg Data might include individual prices, weights, addresses, ages, names, dates, etc. Data is now categorized as belonging to two camps: internal data (enterprise application data) & external data (eg web data).

2] What is Big Data?

1) Big Data is a application of data that is huge in volume yet growing exponentially with time.

2) It is a data with so large & complexity that none of traditional data management complexity that none of traditional data management tools can store it or process it efficiently. Big-data is also a data, with huge size.

Ex:- Student details. Company details.

3) Big Data is been used by Healthcare, Banking, Manufacture, Retail, Transportations.

What is an example of Big Data?

→ Healthcare

1) Big Data is slowly but surely making a major impact on huge healthcare industry.

2] Wearable devices & sensors collect patient data which is then fed in real time to individuals electronic health records.

3) Providers & practice organizations are now using Big Data for a number of purposes include
1) Predicting of epidemic outbreaks.
2) Early symptoms detection to avoid present diseases.
3] Electronic health records
4] Real time alerting
5] Enhancing patient engangement.
6] Predicting & prevention of serious medical conditions
7] Strategic planning
8] Research accelerating

2) Characteristic of Big Data.

3 V's.

1] Volume :
- The name Big Data itself is related to enormous size
- Big Data is a vast volumes of data generated machines social media platform network. human interaction & many more
- Facebook can generate approximately a billion message 4.5 billion times that like button messages 4.5 Billion times that records more than 350 million post new and uploded each day.

2] Velocity :
- It play on important roles compared to other.

- It means speed by which data is created in real-time.

- It contains linking of in coming data sets speed rates of changes & activity bursts.

- The primary aspect of aspect of Big Data is to provided demanding data rapidly.

- Big data velocity deals with speed at data flow from sources like application logs bussiness process, network & social media sites sensors mobile devices etc

3) Variety :

- Big Data can be structured unstructured and semi structured that are being collected from difference sources.

a] structured data :

- In structured schema along with all required colums
- It is tabular form.
- This data is stored RDBMS.

b) Semi-Structured :

- The schema is not appropriately defined.

c] JSON, XML, CSV & email

- OLTP (online transaction processing) system are built to work with semi structured data.
- It is stored in relation i.e tables.

c] Unstructured Data : All structured files log files.

audio files & images files are included in unstructured data

- Some organization have much data available they did not know how derive the value of data since the data is raw.

Q7] What is hadoop?

→ Hadoop is open source framework it is provided by apache to process & analyze huge volume of data it is written is java & used by google fb, linkedin, yahoo,

It is used offline processing, it can be scaled by adding nodes in cluster.

Q8] Advantages of hadoop

→

1] Fast :- In HDFS data distributed over cluster & are mapped which helps in faster retrival.

2] Scalable :- Hadoop cluster can be extended by just adding nodes in cluster.

3] Cost effective :- Hadoop is open source & uses commodity hardware to store data so it really cost effective as compared to traditional relational dbms.

4] Resiliant to failure :- It can replicated data over network normally data are replicated thrice but factors in & configurable

GoodLuck | Page No.
Date

Qu 9] Explain hadoop streaming?

Hadoop streaming is utility that map reduce script in any lang either Java or non-Java.

The article broughly explains headers streaming

Input directory name = input location for mapper.

Output directory name = input location for reducer.

mapper executable = command to run as mapper.

verbose = the verbose output

num Reduced Tasks = it specifies no of reducers