DATE

Name : Shivam Sharma
Roll no : 14/1018649
Section : D
Semester : 5th
Date : 14/Sept/2020
Subject : Hadoop Fundamentals (TCS-561)

Q1.

A. Big data is term that describes large volume of data - both structured and unstructured. Big data can be analyzed for insights that lead to better decision and strategic business moves.

→ Characteristics :

(i) Volume : It refers to unimaginable amount of information generated every from social media, credit cards, etc.

(ii) Variety : It refers to hetrogenous sources and the nature of data, both structured and unstructured.

(iii) Velocity : Velocity refers to speed of generation of data How fast the data is generated and processed to meet the demands, determines real

Scanned by TapScanner

potential in data.

(iv) Variablity: This refers to inconsistency which can be shown by data at times, thus hampering the process of being able to handle and manage data effectively.

B. 1. Banking and Security:
→ These industry rely heavily on Big Data for risk Analytics, including Antimoney laundering, demand enterprise risk management and fraud mitigation.
→ Retail traders, Big banks, hedge funds, and other organizations use Big Data for trade analytics used in high frequency trading.

2. Transportation:
→ Government use big data for traffic control, route planning, intelligent transport system.
→ Many individuals use Big Data for route, planning to save fuel and time.

Scanned by TapScanner

DATE

Q2

A. Some data mining techniques are:

1. Tracking Patterns: This technique is usually a recognition of some aberration in your data happening at regular intervals or an ebb and flow of a certain variable over time.

2. Classification: It is more complex data mining technique that forces you to collect various attributes together into discernable categories, which you can then use to draw further conclusions.

3. Association: Association is related to tracking patterns, but it is more specific todependently linked variables.

4. Outlier detection: It simply recognize the overarching pattern can't give you clear understanding of your data set. You also need to be able to identify anomalies, or outliers in your data.

Scanned by TapScanner

DATE

5. Clustering: Clustering is very similar to classification, but involves grouping chunks of data together based on similarities.

6. Regression: It is used as of planning and modeling, is used to identify the likelihood of certain variable, given the presense of other variable.

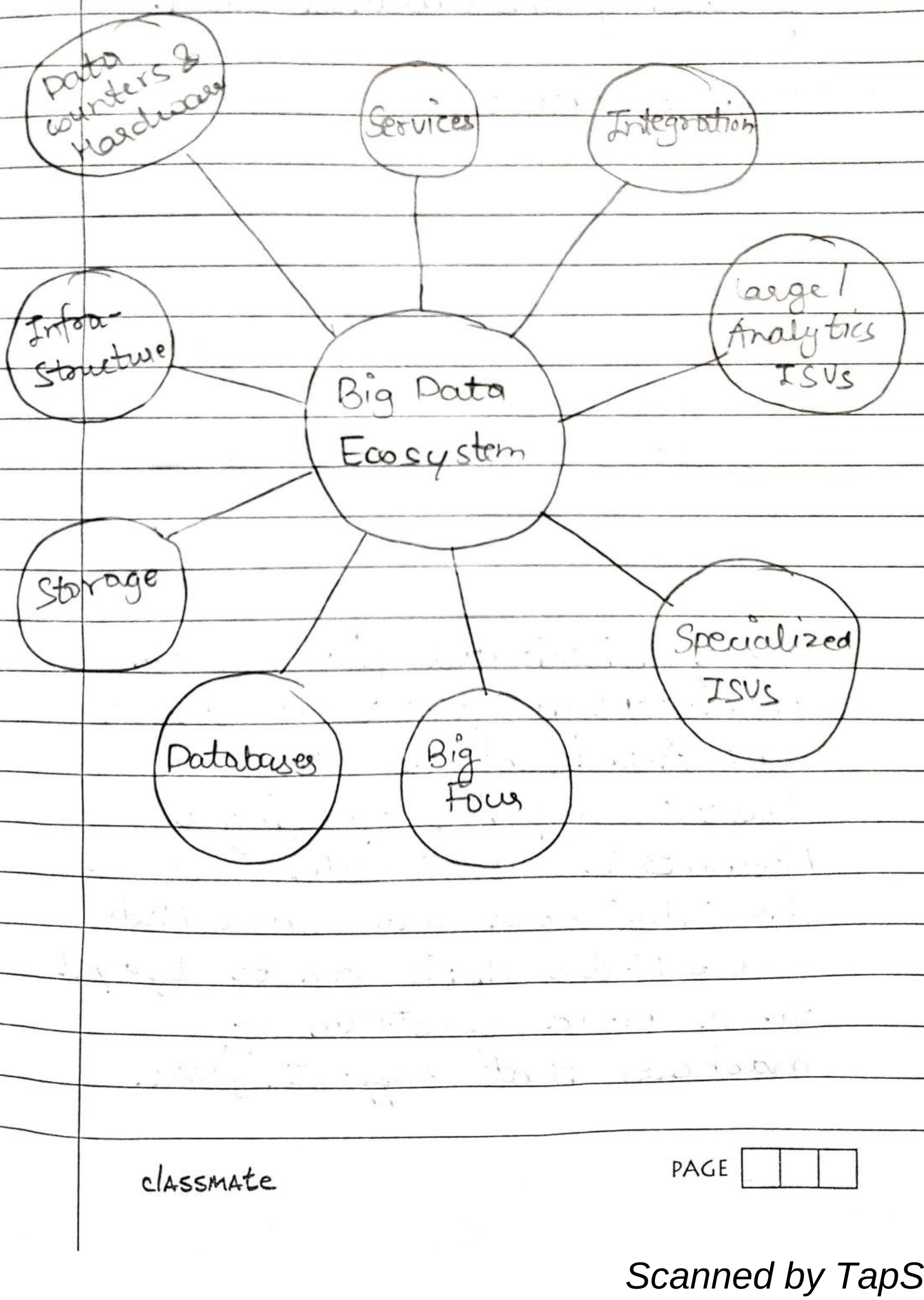
7. Prediction: Prediction is one of the most valuable data mining technique since it is used to project the type of data you will see in future. Just recognizing and understanding historical trends is enough to chart a accurate prediction of what will happen in future.

B. Big Data Ecosystem: A Big Data Ecosystem is collection of infra-structure, analytics and application used to capture and analyze data. Data Ecosystem provide companies with data that they rely on to understand their customer and to make better pricing, operations,

Scanned by TapScanner

DATE

and marketing decisions. The term Ecosystem is used rather that 'Environment' because like real ecosystem data ecosystems are intended to evolve over time.



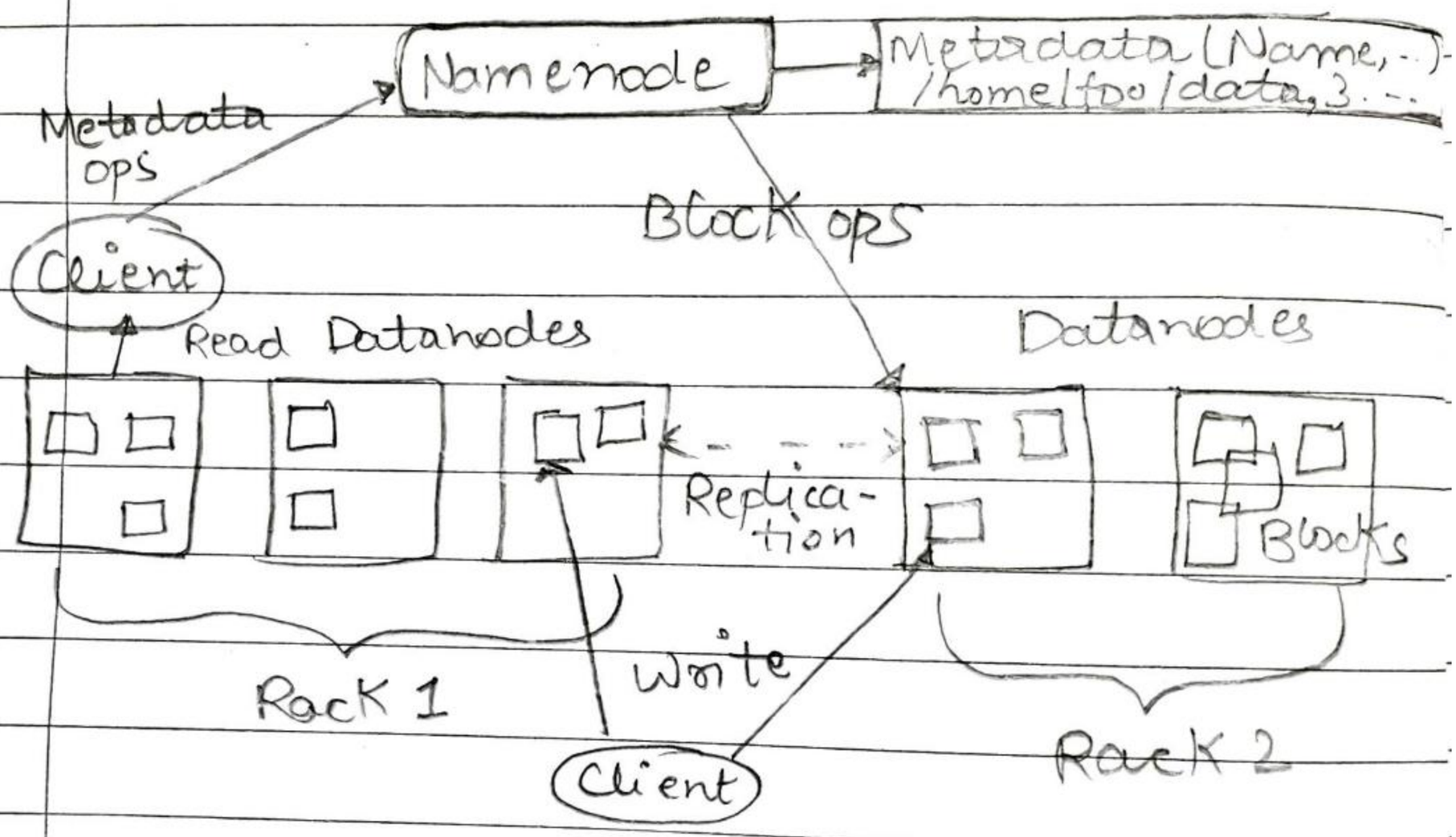
Scanned by TapScanner

DATE

Q5

A. HDFS is a storage system of Hadoop framework. It is distributed file system that can convinientlly run on commodity hardware for processing unstructured data.

→ HDFS Architecture:



Hadoop Distributed file system Architecture follows a raster/ slave Architecture where a cluster comprises of single Namenode (raster node) and all the other nodes are Data Nodes (Slave Nodes) HDFS can be deployed on a broad spectrum of machines that support java.

Scanned by TapScanner

DATE

B. Secondary NameNode in hadoop is specially dedicated node in HDFS cluster whose main function is to take checkpoints of file system metadata present on namenode. It is not a backup namenode.
As the NameNode is single point of failure in HDFS, if NameNode fails entire HDFS file system is lost. In order to overcome this Hadoop Implemented Secondary NameNode whose main function is to store a copy of FsImage file and edits log file.
Secondary NameNode is not a true backup Namenode and can not serve primary NameNode's operations.

→ Secondary NameNode Functions:

- Stores a copy of FsImage and edits log.
- Periodically applies edits log records to FsImage file & refreshes edit logs.
- If NameNode is failed, File System Metadata can be recovered from last saved FsImage on secondary NameNode.
- Check pointing of File System Metadata is performed.

Scanned by TapScanner