VINEET JOSHI
SEC: D ROLL NO: 29
1018633

Date
Page

HADOOP ASSIGNMENT

Ans. Characteristics of Big Data

(i) Volume:
Volume refers to the unimaginable amount of information generated every from social media cell phones, cars, credit cards, M2M sensors, Images, video ~~and whatnot~~ etc.

(ii) Variety:
It refers to heterogeneous sources and the nature of data, both structured and unstructured. This variety of unstructured data poses certain issues for storage, mining and analyzing data.

(iii) Velocity:
The term 'velocity' refers to the speed of generation of data. How fast the data is generated and processed to meet the demands, determines real potential in the data.

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

Some of the techniques of data mining are used to analyze the big data such as clustering, prediction and classification and decision tree etc.
Apache Hadoop, Apache Spark, Apache Storm,

FREEMIND

Date

Page

MongoDB, NOSQL, HPCC are the tools used to handle big data.

Ans2. for many industries, big data isn't a choice but a naturally shaped reality, as data the amount of structured and unstructured data is growing exponentially, along with a wide network of IOT devices capturing it.
The major business opportunities peresented by big data for any industry include:

- Automation: data driven IT infrastructures allow businesses to automate time-consuming proces-sesssuchas data collection and analysis.

- Trends and insights: big data reveals hidden opportunities and patterns that can be used to tailor products and services to end users' need-s to increase operational efficiency.

- Data-based decision-making: machines learn on big data to enable predictive analysis and drive informed decisions.

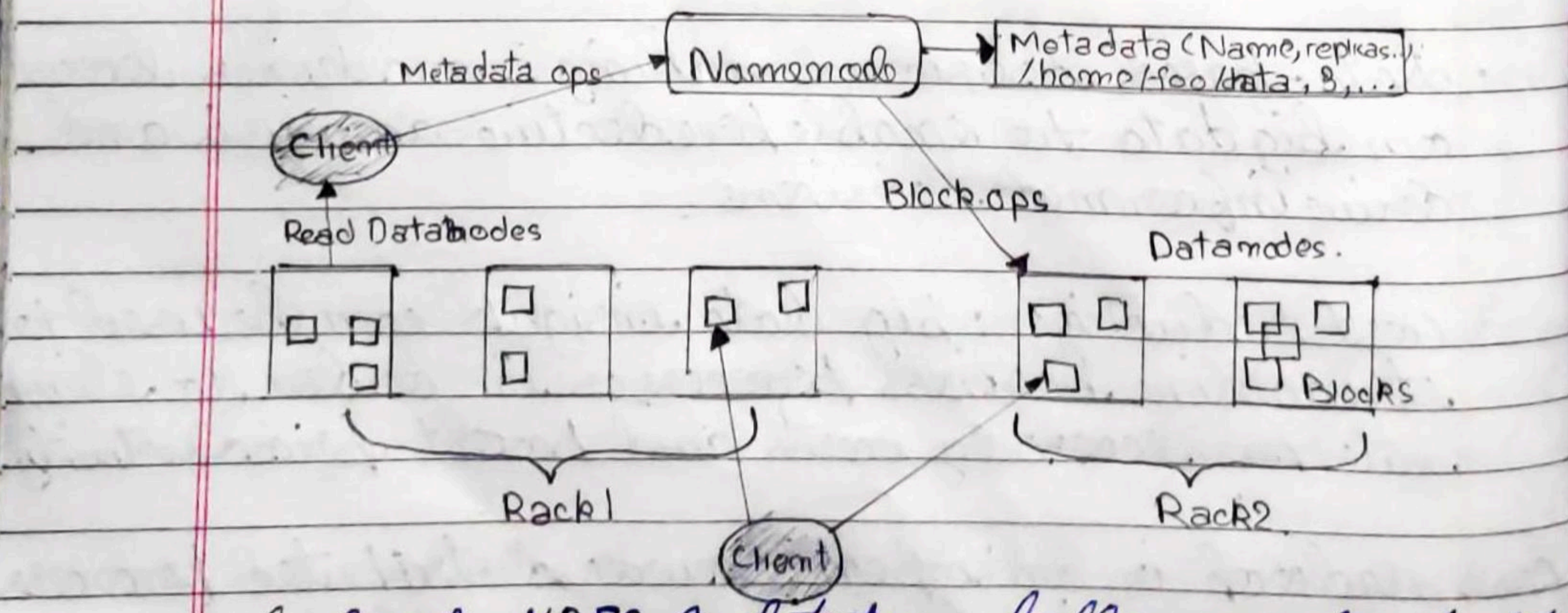
- Cost reduction: big data insights can be used to streamline business processes in order to elimi-nate unnecessary costs and boost productivity

Ans3. Hadoop is an open source distributed process-ing framework that manages data process-ing and storage for big data applications in scalable clusters of of computer servers.

Some myths about hadoop are:

- Hadoop is a database
- Hadoop is cheap.
- Hadoop requires MapReduce: Some users opt to deploy HDFS with Hive or HBase but not Map Reduce
- Mapreduce only controls analytics: it also handles parallel programming, fault tolerance, wide variety of coded logics.
- Hadoop is too risky for enterprise use.
- Hadoop is a complete, single product: Hadoop is an ecosystem, a family of opensource products.

Ans4 HDFS is the storage system of Hadoop framework. It is a distributed file system that can conveniently run on commodity hardware for processing unstructured data.



Hadoop HDFS Architecture follows a Master/Slave Architecture where a cluster comprises of a single Namenode (Master node) and all the other nodes are

FREEMIND

Date

Page

Data Nodes (Slave nodes). HDFS can be deployed on a broad spectrum of machines that support Java.

Ans.5 Hadoop ecosystem includes multiple components that support each stage of Big Data processing.

- Flume and Sqoop ingest data,
- Spark and MapReduce process data.
- Pig Hive and Impala analyze data.
- Hue and Cloudera Search help to explore data.
- Oozie manages the work flow of hadoop jobs.