COMPUTER ENGINEERING DEPARTMENT

BDA Assignment 1

COURSE: **B.E.** YEAR: **2020-2021** SEMESTER: **VII**

DEPT: **Computer Engineering**

SUBJECT CODE: **CSDLO7032** DATE OF ASSIGNMENT: **08-10-2021**

===

NAME: **AMEY MAHENDRA THAKUR** ROLL NO.: **50**

CLASS: **COMPS BE B** DATE OF SUBMISSION: **08-10-2021**

Sr. No.	Questions
1	Explain Hadoop Ecosystem with core components. Explain its Physical architecture. State Limitations of Hadoop.
2	What is MapReduce? Explain How Map and Reduce Work? What is Shuffling in MapReduce?
3	What is NoSQL? What are the business drivers for NoSQL? Discuss any two architectural patterns of NoSQL.

Amey

Signature of Student

AMEY THAKUR B - 50 Amey

Q1 Explain Hadoop Ecosystem with core components.
Explain its Physical architecture.
State limitations of Hadoop.

Ans:

- Hadoop is an open-source JAVA-based framework that allows storage and process big data in a distributed environment across clusters of computers using simple programming models.

- Core components of Hadoop are:

① HDFS: Maintaining the distributed file system. HDFS is the pillar of Hadoop that maintains the distributed file system. It makes it possible to store and replicate data across multiple servers.

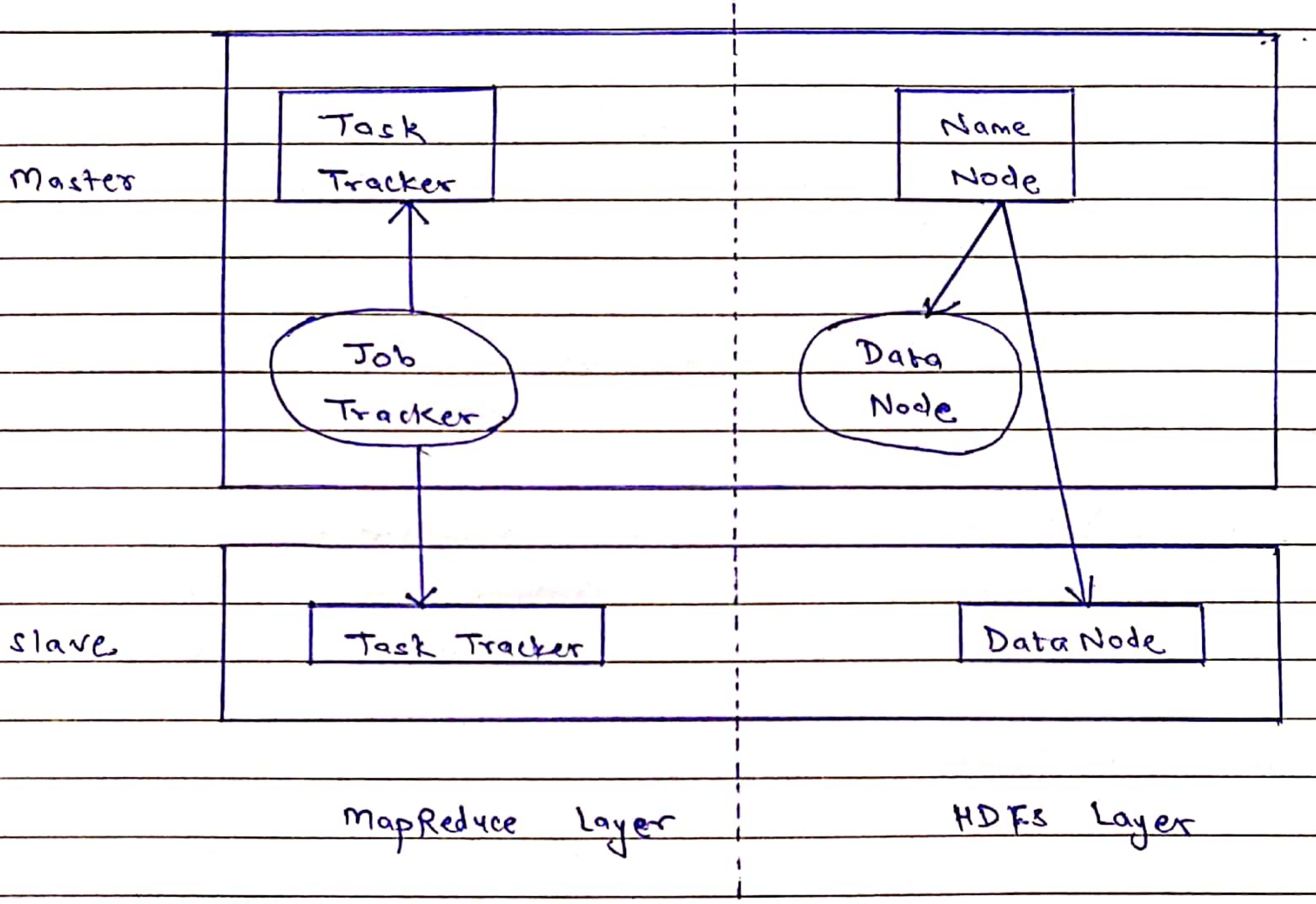
② YARN: Yet Another Resource Negotiator. It manages and schedules the resources and decides what should happen in each data node.

③ MapReduce: MapReduce is a programming model that was first used by Google for indexing its search operations.
It works based on two functions.
① Map ()
② Reduce ()
that pass data quickly and efficiently.

AMEY THAKUR B-50 Amey

Hadoop Architecture

- The hadoop architecture is a package of the file system, MapReduce engine and HDFS.
- A Hadoop cluster consists of a single master and multiple slave nodes. The master node includes Job Tracker, Task Tracker and DataNode whereas a slave node includes DataNode and Task Tracker.



AMEY THAKUR B - 50 Amey.

Limitations of Hadoop :

1. Issues with small files.
2. Slow processing speed
3. Support for batch processing only
4. No Real -Time processing
5. Iterative Processing
6. Security Issues
7. No caching
8. Vulnerable by Nature

AMEY THAKUR B - 50 Amey

Q2. What is MapReduce?
Explain how Map and Reduce work?
What is shuffling in MapReduce?

Ans:

- MapReduce is the processing layer in Hadoop. It is a software framework designed for processing huge volumes of data in parallel by dividing the task into the set of independent tasks.
- The inputs are in form of a list and output from the framework is also in form of a list

The efficiency and powerfulness of Hadoop are due to mapReduce framework parallel processing.

Map Phase:

- In map phase, the user defined map function processes input data. In map function, the user puts the business logic. The output from the map phase is intermediate outputs and is stored on the local disk.

Reduce Phase:

- This phase is combination of the shuffle phase and reduce phase. In reduce phase, output from the map stage is passed to Reducer where they are aggregated. The outpur of Reduce phase is the final output.

AMEY THAKUR B - 50 Amey

Shuffling

- The process of transferring data from mappers to reducers is shuffling. Then it transfers map output to reducer as input. In MapReduce shuffling and sorting occurs simultaneously to summarize the mapper intermediate output.

AMEY THAKUR B-50 Amey.

Q3. What is NoSQL?
What are the business drivers for NoSQL?
Discuss any two architectural patterns of NoSQL.

Ans:

- NoSQL databases refer to any non-relational database. NoSQL databases are used in real-time web applications and big data and their use are increasing over time.
- NoSQL databare has high scalability as it implements horizontal scaling.
- Auto replication feature in NoSQL databases makes it highly available in case of any failure data replicate itself to previous consistent state.

NoSQL business drivers

- Many organizations supporting single CPU relational systems have come to a crossroad: they need their organization to change. Businesses have found values in rapidly capturing and analyzing large amounts of variable data, and making intermediate changes in their business based on information they receive.
  1. Volume
  2. Velocity
  3. Agility
  4. Variability

All these are frequently associated with the NoSQL movement.

AMEY THAKUR B-50 Amey

Architectural patterns of NoSQL

① Key - value store

- A key-value store is a simple database that when presented with a simple string (the key) returns an arbitrary large BLOB of data. It is a simple way to associate a large data file with a simple text string.
- Typical uses: Dictionary, image store, document ~~file~~ file store, query cache, lookup tables.

② Graph stores:

- Graph stores are important in application that need to analyze relationships between objects or visit all nodes in a graph in a particular manner. It is basically a way to store nodes and arcs of a graph.
- Typical uses: Social networking query, friends-of-friends query, inference, rules systems and pattern matching.