2021F-T1 BDM 1003 - Big Data Tools 01 (DSMM Group 1)

Each Question Carries 5 Marks

1. What is Kafka?

Answer: Kafka is open-source software platform developed to manage the event Streaming or real time data feeds. It is written in Scala and java. Event streaming is the process of collecting or capturing the data from various sources like database, IOT, cloud Services in the form of the events. These recorded events are used for analytics, real-time processing, replying to these events or just route these feeds or event to different location. Kafka implements the data pipeline which ensures the continuous flow of data wherever and whenever needed.

2. List Kafka Components

Answer: The major Kafka Components are

1. Broker
2. Zookeeper
3. Producer
4. Consumer

3. Briefly explain about file formats

Answer: Some of the file formats used in Big data are

1. **Text files**
   Text files are simple human readable files. It consumes more space when a number are stored as strings. It is difficult to represent binary data in text files such as images
2. **Sequence Files**
   They are more efficient than the text file and can be used to store the binary data. They store the data as key values pair in a binary container. Sequence files are not human readable
3. **Avro**
   Due to optimized binary encoding Avro files have efficient storage. Schema meta data can be embedded within the file which makes the file more readable. It is widely accepted inside and outside the Hadoop ecosystem. Schema can adapt to the changes. Avro can read and write from many languages such as java and Scala.
4. **Parquet**
   Parquet is a columnar format. It is supported in pig, hive, spark MapReduce. Schema can be embedded within the file like Avro. It uses advance optimization and reduce storage space and increase performance. They are most efficient while adding multiple records at the same time.
5. **ORC (Optimized Row columnar)**
   It provided efficient way to store Hive data. They are column oriented and ready-heavy workloads. It is compatible in HiveQL and support serialization. Meta data are stored using Protocol Buffers, which allows addition and removal of fields

4. What is Spark RDD?

Answer: Spark RDD (Resilient Distributed Dataset) are the datasets which can be stored in memory on worker nodes thus ensures the improved performance. Data are distributed among the

various nodes and recovers upon the failure. RDD supports two types of operations Transformations and actions. Since RDD are immutable any transformation operation performed on it returns the new RDD instead of modifying the existing ones. After the action operations are performed on RDD they return the new value. Whenever the action function is applied on the RDD, they data processing request are evaluated, and value is returned at the same time.

5. Explain Spark DAG Engine

Answer: Dag are sequence of the operation which are carried out on node which is a RDD and edge are represented as transformation which are performed on the data. Transformation is the process which change the state of the data partition from one stage to another and acyclic operation are those which cannot be reverted. The abstraction provide by the DAG removes the complex multiple stage process implemented by Map reduce and improves the performance

6. List 10 Spark transformations

Answer: Some of the Spark Transformation are

1. map
2. flatmap
3. filter
4. mapPartitions
5. mappartitionWithindex
6. union
7. intersection
8. distinct
9. groupBykey
10. reduceBykey

7. What is HMaster

Answer: HMaster is a component which acts as a master in master slave environment in HBase architecture. It associates regions with the region server and manages the Data Definition language operations (DDL) as well. HMaster monitors and controls all the regions server present in the cluster. Multiple background threads are running in the HMaster to perform load balancing and controlling failover

8. What are regions in HBase

Answer: Regions are components in Hbase architecture which are responsible for handling, managing and read write HBase operation on set of regions. It had all the rows starting from the start key to the end key which are assigned to that region.

9. What is Zookeeper?

Answer: Zookeeper is the open source Apache projects which provides the service for managing, naming configuration information, distributed synchronization and group services to the large cluster in distributed systems.

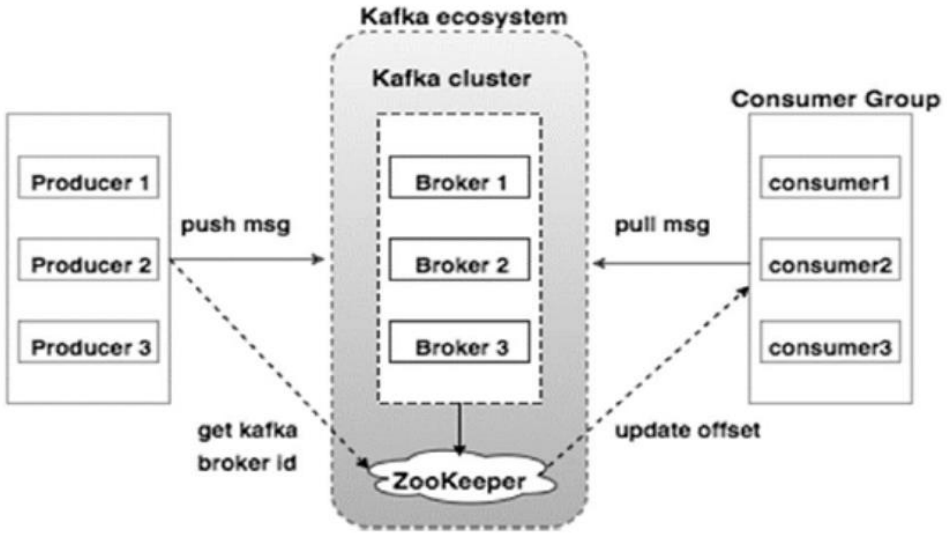
10. What is MongoDB?

Answer: Mongo DB is NOSQL database. It is a document-based database where there are documents and collections. Collection is the like table in RDBMS. A collection exists within a single database.

There are number of fields in the collections. MongoDB has the concept of shards to maintain replication in the distributed environment. The maximum size of the BSON document is 16 MB which makes sure that RAM and bandwidth is not consumed that much during transmission.

Each Question Carries 10 Marks:

1. Explain Kafka Architecture



The above picture reveals the major components in kafka architecture. They are

5. Broker
6. Zookeeper
7. Producer
8. Consumer

Broker

A single kafka broker can handle hundred and thousands of reads and writes per second and ever broker can handle TB of message without having any impact on the performance. They are stateless thus the state of the Broker are managed by zookeeper. To maintain the load balance kafka consists of multiple brokers and broker lead election is done by zookeeper.

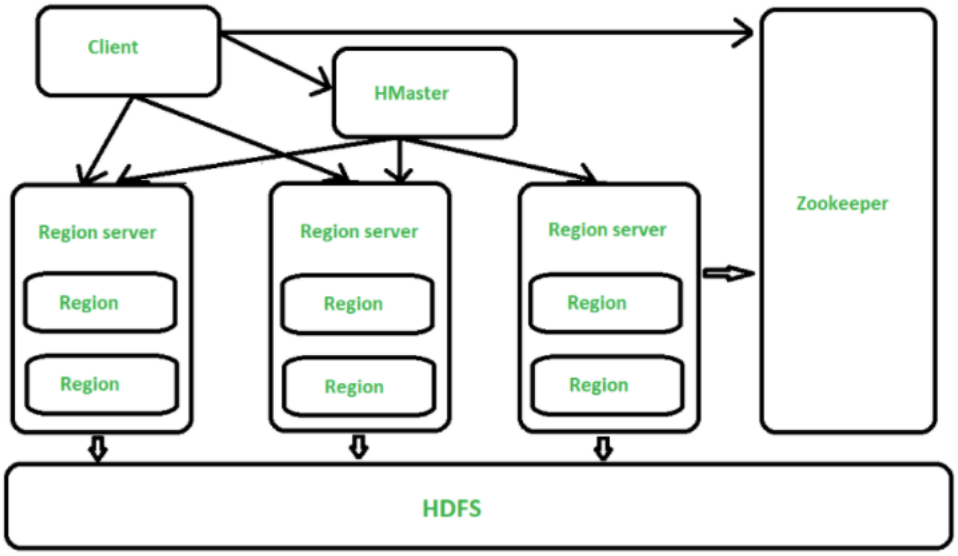
Zookeeper

Zookeeper maintains the states of the broker. The status information of the broker is regulated by zookeeper. The main task of the Zookeeper is to notify producer and consumer in the time of the failure of the consumer and producer so that they can rely on another new broker and start coordinating their task with other broker.

Producer
Kafka producer do not wait for the acknowledgements from the broker. They push the data to the brokers. Whenever the new broker is started, it sends the message to the broker as soon as possible in the amount the broker can handle.

Consumer
They have to keep track of the amount of the message they consumed since brokers are stateless. They use partition offset to do that. If a consumer acknowledges any message, it means that it have consumed all messages prior to that. Simply by looking at its offset value consumer can know how much message has been consumed and gives a asynchronous pull request to the broker.

2. Explain HBase Architecture



There are three main components in HBase architecture

HMaster
HMaster acts as a master in master slave environment. It associates regions with the region server and manages the Data Definition language operations(DDl) as well. Hmaster monitors and controls all the regions server present in the cluster. Multiple background threads are running in the Hmaster to perform load balancing and controlling failover

Region Server
Regions on the regions server are responsible for many things such as managing reads and writes and other Hbase operations on their regions. A regions server can contain multiple regions and each regions server run on the datanode of Hadoop. Regions consists of the distributed tables which in turn records the information of the column families. Regions are the building blocks of the HBase.

Zookeeper
Clients communicates to region server via zookeeper. Basically, it performs the state management and co-ordinates between components. It mainly provides configuration information manages distributed synchronization and is responsible for failure notification.

3. Explain SQL and NOSQL Differences
Answer :There are multiple differences between SQL and NoSQL some of them are

1. SQL databases are generally called relational databases whereas NoSQL databases are called non-relational databases.
2. Traditional DBMS use SQL syntax and queries to perform OLAP and deriving insights, NOSQL was designed to meet high performance and modern application need.
3. SQL are table-based databases whereas NoSQL can be document based, key-value based or Graph Databases.
4. SQL have predefined Schema and deal with structured data, NoSQL has dynamic Schema and deals with all structured, semi structured and unstructured data.
5. SQl are vertically Scalable, NoSQL are horizontally scalable.
6. Oracle, Postgres, MySQL are some examples of SQl and Mongo DB, Cassandra , Hbase are examples of No SQL.
7. SQL follows ACID (Atomicity, Consistency, Isolation and Durability) properties and NOSQL follows CAP(consistency, availability, partition tolerance) properties

4. Explain Spark Architecture

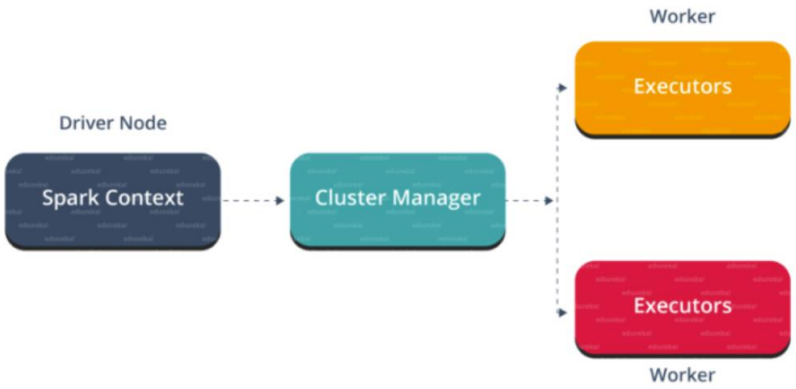


Fig: Spark Architecture

Spark architecture follows the master slave approach. There is a single master multiple slaves in the system. The architecture can be explained with two major abstractions

1. Resilient Distributed Datasets (RDD)
2. Directed Acyclic Graph (DAG)

Resilient Distributed Datasets
These are the datasets which can be stored in memory on worker nodes thus ensures the improved performance. Data are distributed among the various nodes and recovers upon the failure. RDD supports two types of operations Transformations and actions. Since RDD are immutable any transformation operation performed on it returns the new RDD instead of modifying the existing ones. After the action operations are performed on RDD they return the new value. Whenever the

action function is applied on the RDD, they data processing request are evaluated, and value is returned at the same time.

Directed Acyclic Graph (DAG)
Transformation is the process which change the state of the data partition from one stage to another and acyclic operation are those which cannot be reverted. Dag are sequence of the operation which are carried out on node which is a RDD and edge are represented as transformation which are performed on the data. The abstraction provide by the DAG removes the complex multiple stage process implemented by Map reduce and improves the performance.

The major components is Spark are

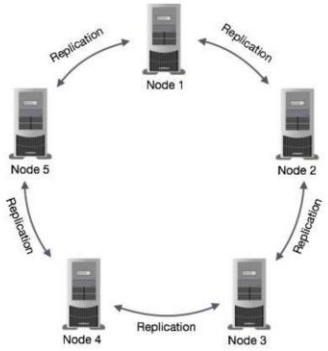
1. Spark Driver
2. Executors
3. Cluster Manager

Spark Driver
The major responsibility of the spark driver is to co-ordinate the tasks and work. It is an Application JVM process thus considered as Master Node. A driver splits the tasks into schedule which will execute on the executor.

Executor
Executors are responsible for execution of the job scheduled by the driver and they store data in the cache. They execute the task and report the status of the execution to the driver. A spark application has its own sets of executors.

Cluster Manager
Cluster Manager are tied to the physical devices rather than the processes. Cluster manager comes in between the spark driver and executors. Cluster Manager is responsible for managing the cluster of machines for spark application. When the spark application is about to run resources are requested from the cluster manager. The resources might be for the driver or executors no matter what cluster manager is responsible for handling all the underlying machine on which the application is running

5. Explain Cassandra Architecture



Cassandra was designed to handle the huge workloads on the distributed environment with single node of failure. It is peer to peer based distributed architecture. Data is distributed between the node in the cluster.

The major components in the Cassandra architecture are described below

Node
Node is the place where actual data are stored. It is the basic components in Cassandra. Each node has the ability to run independently but in interconnected fashion.

Data Center
Data Center are actually the collection of the nodes that are related

Cluster
Cluster is the components which has sets of Datacenter.

Commit log
Every commit or write operation are written in the commit log and acts as a crash-recovery mechanism in Cassandra.

Mem-table
After the data is written in the commit log then they are written in the Mem-table. Data are reside there temporarily. Mem-table is memory resident table. In some scenario, there will be multiple mem table for single column family.

SS Table
Whenever some threshold value is reached in storing data in Mem table is reached then data is flushed to the SS table. It is a disk file.

Bloom Filter
Bloom Filter are the cache. After every query bloom filter are accessed. In simple terms they are non-deterministic algorithm to check if a element Is present in set or not.