By OnlineInterviewQuestions.com

Cognizant Interview Questions on Hadoop

Q1. Explain the architecture of Hadoop Eco system?

Apache Hadoop is used to process a huge amount of data. The architecture of Apache Hadoop consists of Hadoop components and various technologies which is helpful to solve complex data problems easily.

The description of each component of the architecture of the Hadoop ecosystem are as follows:

Namenode controls the operation of data.

Datanode writes the data to local storage. To store all data at a single place is not always recommended, as it may cause loss of data in case of an outage situation.

Task tracker accepts tasks assigned to the slave node.

The map takes data from a stream and each line is processed after splitting it into various fields.

Reduce: The fields, obtained through Map are grouped together or concatenated with each other.

Q2. What is Incremental load in hive?

In the hive, Incremental load is generally used to implement slowly changing dimensions. When you migrate your data to the Hadoop Hive, you might usually keep the slowly changing tables to sync up tables with the latest data.

Q3. What is difference between MR1 and MR2?

MR stands for MapReduce. The Difference between MR1 and MR2 are as follows:

The earlier version of the map-reduce framework in Hadoop 1.0 is called MR1. The newer version of MapReduce is known as MR2. MR2 is more isolated and scalable as compared to the earlier MR1 system. MR2 is one kind of distributed application that runs the MapReduce framework on top of YARN. MapReduce performs data processing via YARN. Hence Yarn execution model is more generic than the earlier MapReduce model.

Q4. What is bucketing in Apache hive?

Bucketing is defined as a technique offered by Apache Hive to decompose data into more manageable parts, also known as buckets that enhance query performance. Bucketing allowed by partitioning, where partitions can be further divided into buckets.

Q5. What is executor in Apache spark?

Executors in Apache Spark are the worker nodes that assist the process of operating the individual tasks in the given job of Apache Spark. Executors are launched at the beginning of a spark application and run for the lifetime. Executors also offer in-memory storage for the RDDs of the Spark that are in return cached by the user programs through the Block manager.

Executors also send metrics of the heartbeats through the Heartbeat Sender Threads. Executors can also be identified through the hostname, id, classpath or environment. The executors in the backends extensively manage the executors present in the Apache spark.

Q6. What are cores in spark?

A **core** is the computation unit of the CPU. In spark, cores control the total number of tasks an executor can run. It is the base foundation of the entire spark project. It assists in different types of functionalities like scheduling, task dispatching, operations of input and output and many more. Ore in the spark is the engine for distributed execution with all the functionalities that are attached at the top.

The **core in the Apache spark** offers the entire functionalities like fault tolerance, monitoring, in-memory computation, management of the memory, and task scheduling.

Q7. What are HDFS and YARN?

HDFS: Hadoop File System or HDS is the main storage segment of the Hadoop. It stores various types of data in the distributed environment in the form of blocks. It follows the topology of slave and master. It is used to spread out in multiple machines to increase the trusts and decrease the costs.

Yarn: Yet Another Resource Negotiator or YARN is the execution system of the program that enhances the MapReduce (MR). YARN is used for scheduling, queuing and the management systems of the execution. It schedules the executions inside the containers. YARN is the framework for processing in Hadoop.

Q8. What is use 'jps' command in Hadoop?

The **Full form of JPS** is **Java Virtual Machine Process Status**. JPS offers all the instrumental hotspots that the JVM is running in the system. JPS is a type of command that is implemented to check out all the Hadoop

daemons like DataNode, NodeManager, NameNode, and ResourceManager that are currently running on the machine.

JPS command is used to check if a specific daemon is up or not. The command of JPS displays all the processes that are based on Java for a particular user. The command of JPS should run from the root to check all the operating nodes in the host.

Q9. What is NameNode in Hadoop?

NameNode is the foundation of the HDFS system. It stores all the directory tree of the files in a single file system and keeps track of where the data file is kept. It does not store the data within itself. The NameNode responds to the successful requests by returning the lists of the relevant DataNode servers.

The NameNode is also considered as the single point of failure for the HDFS cluster. The file system goes down when there is a failure of NameNode. The NameNode can be configured to store a single transaction log on a separate disk image.