**HADOOP Interview Questions**

### What are the three modes in which Hadoop can run?

### Why is HDFS fault-tolerant? -> Practise HDFS command line statement

### Explain the architecture of HDFS.

### What is the difference between a federation and high availability?

### If you have an input file of 350 MB, how many input splits would HDFS create and what would be the size of each input split?

1. What is the difference between Block & Input Split in HDFS?

### How does rack awareness work in HDFS?

### Which command will help you find the status of blocks and FileSystem health?

### What would happen if you store too many small files in a cluster on HDFS?

### What role do RecordReader, Combiner, and Partitioner play in a MapReduce operation?

### Why is MapReduce slower in processing data in comparison to other processing frameworks Spark?

### What do you mean by map-side join and reduce-side join in MapReduce? Very Important Qus

### Explain the process of spilling in MapReduce.

### Explain how YARN allocates resources to an application with the help of its architecture.

### In a cluster of 10 DataNodes, each having 16 GB RAM and 10 cores, what would be the total processing capacity of the cluster?

### How does NameNode tackle DataNode failures?

### What will you do when NameNode is down?

### What is a checkpoint? - Very Imp Qus

### What is “speculative execution” in Hadoop?

**Spark Interview Questions**

### Explain the concept of Resilient Distributed Dataset (RDD).

1. What is the difference between RDD, Dataframe and Dataset

### How do we create RDDs in Spark?

### What is Executor Memory in a Spark application?

### Define Partitions in Apache Spark.

### What operations does RDD support?

### What do you understand by Transformations in Spark?

### What do you understand by Pair RDD?

### How can Apache Spark be used alongside Hadoop?

### What is RDD Lineage or Lineage Graph?

### What is Spark Driver?

### What is Spark Executor?

1. What is spark context?

### What do you understand by worker node?

### Can you use Spark to access and analyze data stored in Cassandra databases?

### How can you minimize data transfers when working with Spark?

### What are broadcast variables? Very imp

### Explain accumulators in Apache Spark.

### What is a DStream in Apache Spark?

1. Whar is caching & persistence in Spark?

### What are the various levels of persistence in Apache Spark?

### Does Apache Spark provide checkpoints?

### What do you understand by Lazy Evaluation?

### Explain a scenario where you will be using Spark Streaming.

1. What are the important optimization techniques we can use in Spark?
2. How to take care of Driver out of memory & executor out of memory failures? (Very Imp)
3. How to deal with Data Skewness in Spark?
4. What is the difference between repartition and coalesce?

**Hive Interview Questions**

### Where does the data of a Hive table gets stored?

### What is a metastore in Hive?

### What is the default database provided by Apache Hive for metastore?

1. *Suppose I have installed Apache Hive on top of my Hadoop cluster using default metastore configuration. Then, what will happen if we have multiple clients trying to access Hive at the same time?*

### What is the difference between external table and managed table? Very imp

### When should we use SORT BY instead of ORDER BY?

### What is a partition in Hive? - Very imp

### Why do we perform partitioning in Hive?

### What is the difference between Static & dynamic partitioning and when is it used?

1. *Suppose, I create a table that contains details of all the transactions done by the customers of year 2016:* CREATE TABLE transaction\_details (cust\_id INT, amount FLOAT, month STRING, country STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY ‘,’ ;

*Now, after inserting 50,000 tuples in this table, I want to know the total revenue generated for each month. But, Hive is taking too much time in processing this query. How will you solve this problem and list the steps that I will be taking in order to do so?*

### How Hive distributes the rows into buckets?

### **What are Buckets in Hive?**

1. How to implement map-side join in hive?
2. What broadcast join in Hive?

### **What are the uses of Hive Explode?**

1. Difference between ORC,Avro,Parquet file formats?
2. How to read json file in HIVE?
3. What SerDe in Hive?
4. How can we write custom UDF’s in Hive?
5. Difference between schema on Read & schema on Write?

**SQOOP Interview Questions**

1. **What is Sqoop Import? Explain its purpose.**
2. **How can I import large objects (BLOB and CLOB objects) in Apache Sqoop?**
3. **How can you execute a free-form SQL query in Sqoop to import the rows in a sequential manner?**
4. **If the source data gets updated every now and then, how will you synchronize the data in HDFS that is imported by Sqoop?**
5. **How can we import data from particular row or column? What is the destination types allowed in Sqoop import command?**
6. **What is the process to perform an incremental data load in Sqoop?**
7. **Can free-form SQL queries be used with Sqoop import command? If yes, then how can they be used?**
8. **What is the importance of eval tool?**
9. **How can you import only a subset of rows from a table?**
10. **What is the role of JDBC driver in Sqoop?**
11. **How can you schedule a sqoop job using Oozie?**
12. **What is the importance of — the split-by clause in running parallel import tasks in sqoop?**
13. **What is the purpose of sqoop-merge?**
14. **How can you avoid importing tables one-by-one when importing a large number of tables from a database?**
15. **How can you control the number of mappers used by the sqoop command?**
16. **How will you update the rows that are already exported?**
17. **What is Sqoop Validation?**
18. Difference Between Apache Sqoop vs Flume.