**Hadoop (HDFS, MapReduce, YARN)**

What are the core components of Hadoop?

What are the characteristics of Hadoop?

What is the purpose of HDFS?

YARN?

MapReduce?

Explain the architecture of HDFS?

What is the purpose of Namenode? Datanode?

What is the default block size in HDFS?

Default replications?

What is the architecture of YARN?

What is the purpose of ResourceManager? NodeManager?

Purpose of Scheduler? ApplicationsManager? Container? Applications Master?

What are the different types of input formats?

What is a columnar file format?

**Hive**

What is Hive?

-          Hive gives an SQL-like interface to query data that interfaces with Hadoop

Where is the default location of Hive's data in HDFS?

-          /user/hive/warehouse

What is an External table?

-          A table whose data is stored outside of Hive’s structure. The data is not lost when drop is used on an external table

What is a managed table?

-          A table whose data is stored in hive. If dropped, the data is removed.

What is a Hive partition?

-          A way of dividing a table into related parts based on the values of particular columns

Provide an example of a good column or set of columns to partition on.

-          A good column to partition on is one where there might be many duplicate elements so as not to create too many directories  for small partitions

What's the benefit of partitioning?

-          Queries on smaller pieces of data become much faster if you only have to search through one or few partitions

What does a partitioned table look like in HDFS?

-          Each partition is stored in its own directory in HDFS with each row as a file within

What is a Hive bucket?

-          Bucketing in Hive groups the data by ranges of values of a column or combination of columns.

What does it mean to have data skew and why does this matter when bucketing?

-          Skew data is data that is very far off from the rest of the data in a table. For example, a field having a very large number when all the other values in the column are quite small. If skew data is not handled, the ranges for the buckets might not be optimized.

What does a bucketed table look like in HDFS?

-          The buckets are stored as separate files with the rows that fall into each bucket contained within.

What is the Hive metastore?

-          Service that stores metadata related to Apache Spark and other services

What is beeline?

-          A client that allows the execution of Hive queries through multiple concurrent connections and supports authentication

Hive Syntax questions: How do we....

create a table?

-          CREATE TABLE <table\_name> (column\_names DATATYPE,…) row format deliminated fields terminated by ‘<deliminator>’ stored as textfile;

load data into a table?

-          LOAD DATA INPATH ‘<path>’ INTO TABLE <table\_name>

query data in a table?

-          SELECT <columns> FROM <table> …

filter the records from a query?

-          WHERE clause

group records and find the count in each group?

-          SELECT COUNT(\*) FROM <table> WHERE…

write the output of a query to HDFS?

-          INSERT OVERWRITE DIRECTORY ‘<path>’ SELECT …

specify we're reading from a csv file?

-          CREATE TABLE <table>(<columns> <datatypes>)row format delimited fields terminated by ‘,’

**Spark : Cluster Computing with Working Sets**

What does Cluster Computing refer to?

-          Cluster computing refers to the process of sharing computation tasks among multiple machines which works together to act as a single machine

What is a Working Set?

-          A working set is a part of the resources that the current process is using. It is the amount of memory that a process requires in a given time interval

What does RDD stand for?

-          Resilient Distributed dataset

What does it mean when we say an RDD is a collection of objects partitioned across a set of machines?

-          An RDD is distributed, meaning the data is stored across multiple nodes. This helps the data to be resilient. Now, the data itself that is contained in the RDD is not copied across to create resilience, but the lineage of the RDD is, so that the data that is accessed after an action can still be accessed.

Why do we say that MapReduce has an acyclic data flow?

-          MapReduce has a data flow that is not in a cycle or a closed path. The data transformations are similar to that of functional programming, where the process of transforming the data happens programmatically.

Explain the deficiency in using Hive for interactive analysis on datasets. How does Spark alleviate this problem?

-          Hive is more suited for batch processing with large amounts of data. It is not good with streaming data. As such, it runs slower on those interactive analysis processes. Spark, as it runs on RAM, can process streaming data much faster.

What is the lineage of an RDD?

-          The list of transformations that the RDD has gone through.

RDDs are lazy and ephemeral. What does this mean?

-          The processing of transformations do not occur until an action is called, and the dataset is discarded from memory after use

What are the 4 ways provided to construct an RDD?

-          Parallelize,

What does it mean to transform an RDD?

-          To transform an RDD is to create a new RDD from an existing one. Transformations are lazy, meaning that the data is not actually transformed until an action is called.

What does it mean to cache an RDD?

-          Caching is an optimization tactic. It saves the data from an RDD action instead of deleting that data right away so that the RDD can be used again on a later transformation or action. This can be helpful if many different transformations and actions are called from a common RDD.

What does it mean to perform a parallel operation on an RDD?

-          Since the RDD is partitioned, operations can be done in parallel, where multiple executors can perform the same action across those partitions

Why does Spark need special tools for shared variables, instead of just declaring, for instance, var counter=0?

-          Because operations happen in parallel across multiple partitions, shared variables need to be able to be accessed in each of those places in memory

What is a broadcast variable?

-          A broadcast variable is a read-only variable that is shared on all nodes. They are immutable and distributed.