1. Explain Architecture of Spark?

The apache spark follows master-slave architecture that consists of a driver, which runs as a master node, and many executors that run across as worker nodes in the cluster.

2. Difference between Hadoop and spark

HADOOP	SPARK
Hadoop's mapreduce operates by reading data from disk which makes the processing speed slower than spark	spark performs in-memory processing hence the processing speed is way more than hadoop
Efficient for batch processing	Efficient for real time processing
Cost is low when compared to spark as it uses any disk space	Cost is high when compared to hadoop as it uses RAM
Hadoop is written in JAVA, more time to execute	Spark is written in SCALA, less time to execute

3. Difference between RDD, Dataframe, Dataset

RDD	Dataframe	Dataset
Program how to do (no optimization)	Program how to do (with optimization)	Program how to do (with optimization)
OOPS style API	SQL style API	SQL style API
Strongly type	Less type safety	Strong type safety
Give compile time error	Give runtime error	Give compile time error
No schema	Structured schema	Structured schema

4. Explain the similarities in all API of Spark

- Distributed in nature,
- Fault tolerant,
- In-memory parallel processing,
- Immutable,
- Use lazy evaluation techniques

5. What is Transformation? Explain in detail

In spark, transformation refers to the operation that takes an input dataset(RDD or dataframe or dataset) and produces a new dataset as output. Transformations are lazy , meaning they do not execute immediately instead they form a series of transformations to be executed later when an action is called.

Some of the transformation are filter, join, map, groupby, union, flatmap etc.,

6. What are Actions in spark? Explain in detail

Action in spark is the operation that triggers the execution of transformation which produces the output in the form of a new dataset(RDD or dataframe or dataset). These are not lazy as result they will initiate the computation as soon as the action is called.

Some of the action are collect, count, reduce, show, save etc.,

7. What is the Wide Transformation ?, explain with example

Wide transformation in spark involves shuffling and redistributing data across partitions.

Example:

Partition1 Dept = Sales Dept = HR Drpt = IT	Partition2 Dept = Sales Dept = HR Drpt = IT	Partition3 Dept = Sales Dept = HR Drpt = IT

After applying filter it will create a new dataset as below

Partition1 Dept = Sales Dept = Sales Drpt = Sales	Partition2 Dept = HR Dept = HR Drpt = HR	Partition3 Dept = IT Dept = IT Drpt = IT

After applying groupby transformation and count action will get the output

This is an example of wide transformation as data is shuffled across partition before getting output from the dataset.

8. What is Narrow Transformation? Explain with example

Wide transformation in spark does not involve shuffling and redistributing data across partitions.

Example:

Partition1 emp_id = 100 emp_id = 200 emp_id = 300	Partition2 emp_id = 200 emp_id = 400 emp_id = 600	Partition3 emp_id = 200 emp_id = 500

To fetch emp_id, we will use a filter which does not involve shuffling data across partitions and will give direct output when action is called, hence the above is an example of narrow transformation.

9. Write down the query of wide n narrow transformation with example?

Wide transformation:

Partition1 Dept = Sales Dept = HR Drpt = IT	Partition2 Dept = Sales Dept = HR Drpt = IT	Partition3 Dept = Sales Dept = HR Drpt = IT

query:

DF2 = DF1.groupby('Dept').count()

Narrow transformation:

Partition1 emp_id = 100 emp_id = 200 emp_id = 300	Partition2 emp_id = 200 emp_id = 400 emp_id = 600	Partition3 emp_id = 200 emp_id = 500

query:

DF2 = DF1.filter(DF1.emp_id=200)

10. Explain Kerberos Architecture

Kerberos is a network authentication protocol designed to provide strong authentication for client/server applications.

It consist of three main components

1. Client
2. Key Distribution Center(KDC)
3. Server

Steps:

- Client requests authentication to access the service.
- The request will be communicated with KDC at the initial stage which contains Authentication Server(AS) and Ticket Granting Server(TGS) has its main components.
- The request is first sent from KDC to AS for checking authenticity of the user as the process of checking AS will sent ticket(encrypted) to the client where it will be decrypted using hash code and sent back to AS for validation , if the client is valid client the request is then further connected with TGS for providing service ticket(secret key).
- This service ticket can be used by the client for accessing the services.