Explain Architecture of Spark?
It uses Master-Slave Architecture consists of a driver, which runs as a master node, and many executors that run across as worker nodes in the cluster. It can be used for batch processing and real time processing.

RDD- Resilient Distributed Dataset
DAG - Directed Acyclic Graph

Difference between Hadoop and Spark?

Hadoop	**Spark**
Developed based on Distributed Computing	Developed based on Hadoop for big data because of slow processing
Store and Process big data	Faster Processing
Fault Tolerance by keeping copies of data objects in the disk	Follows the lineage process where RDD failed , can be computed again and does not lead to more data copies in memory.
Map-reduce program sare primarily written in java	Programs are written in scala supports functional and OOPs.

Difference between RDD, Dataframe , Dataset.
RDD: Resilient Distributed Dataset is a fundamental abstraction in SPark, represents immutable distributed collection of objects partitioned across the nodes of a cluster.
They provide a low-level API for distributed data processing, allowing users to perform transformations and actions on distributed datasets.
RDD offers strong typing and are available in JAVA, Scala, and Python API.

DataFrame: is a higher level abstraction introduced in Spark 1.3
Dataframes are built on top of RDD but provide a more structured and user-friendly API, similar to data frames in R or pandas in Python.
DataFrame are available in JAVA, Scala, Python, and R API.

Dataset:
Dataset is an extension of DataFrame API introduced in Spark 1.6, aiming to provide the benefits of both RDD and DataFrame.
Datasets are available in JAVA and Scala API.

Explain the similarities in all API of Spark.

Distributed Computing
Fault Tolerance
Lazy Evaluation

Immutability
Data Abstraction
Parallelism
Optimization

What is Transformation? Explain in detail.
Transformation is an operation that is applied to an existing distributed dataset to create new dataset. Transformations are lazy operations, meaning they do not compute their results immediately. Instead it builds a DAG of computational steps that define the transformation to be applied to the dataset. The actual computation is deferred until an action is called the dataset.

What is Action in spark? Explain in detail.
Action is an operation that triggers the execution of the previously defined lazy transformations on a distributed dataset and computes a result to be returned to the driver program or stored in external storage. Characteristics:
Eager Evaluation
Result Computation
Spark UI and Monitoring
Performance Consideration.

What is Wide Transformation? Explain with example.
Wide transformations required data to be shuffled across partitions, often involving data exchange between partitions.It typically involve a stage boundary and can be more expensive in terms of computational overhead.
Example: groupby, join, reducebykey, sortbykey, etc.

What is Narrow Transformation? Explain with example.
Narrow Transformation do not shuffle data across partitions. They operate on each partition independently and can be executed in parallel.
Example: map, filter, flatmap, mappartitions, etc.

Write down the query of wide and narrow transformation with example.

Explain Kerberos Architecture.
Kerberos provides a centralized authentication server whose function is to authenticate users to servers and servers to users. In Kerberos server and database is used for client authentication. Kerberos runs as a third-party trusted server known as the Key Distribution Center.
Main Components:
Authentication Server:
Initial authentication and ticket for Ticket Granting Service.

Database:
Verifies the access rights of users in the database.

Ticket Granting Server:
Issues the ticket for the Server.