**Apache Spark**

Spark is an open-source, distributed computing system that provides a fast and general-purpose cluster-computing framework for big data processing.

Spark architecture

**Driver:** The central coordinator of a Spark application, responsible for defining the application logic and creating the SparkContext.

**Spark Context:** The entry point for a Spark application, coordinating tasks across the cluster, communicating with the cluster manager, and managing resources.

**Cluster Manager:** Manages resources across the cluster, allocating tasks to worker nodes.

**Worker nodes**: Individual machines in the cluster that execute tasks. Each worker node runs an executor process, performing computations and storing intermediate data.

**Spark Session:** The unified entry point for reading data, executing SQL queries, and managing the Spark environment. It encompasses the functionality of SparkContext, SQLContext, and HiveContext.

**Three APIs of Spark:**

**Dataset** - is a strongly-typed extension of DataFrame API Introduced in Spark 2.0.

Datasets provide type safety and compile-time checks while maintaining the benefits of a DataFrame's high-level abstraction.

**DataFrame** - DataFrame is a higher-level abstraction built on top of RDD,representing a distributed collection of data organized into named columns

**Resilient Distributed DataFrame** - RDD is the fundamental data structure in Spark, representing an immutable distributed collection of objects. RDDs can be processed in parallel and are fault-tolerant, allowing for distributed computing across a Spark cluster.