**What is Apache Spark?**

Apache Spark is an open-source distributed computing framework designed for fast, large-scale data processing.

Key points:

* Developed originally at UC Berkeley.
* Can process batch, streaming, machine learning, and graph data.
* Works in-memory (but can spill to disk) → much faster than Hadoop MapReduce.
* Runs on Hadoop (YARN), Kubernetes, Mesos, or standalone cluster manager.

**Features of Spark**

| **Feature** | **Description** |
| --- | --- |
| **In-memory computation** | Data stays in memory between steps → faster than disk-heavy systems like Hadoop MapReduce |
| **Speed** | 10–100x faster than Hadoop MapReduce for many workloads |
| **Rich APIs** | Supports Scala, Java, Python (PySpark), R, SQL |
| **Unified engine** | Handles batch, streaming (Structured Streaming), ML (MLlib), graphs (GraphX) |
| **Lazy evaluation** | Builds an execution plan; runs only when action is called |
| **Fault tolerance** | Uses lineage + DAG → can recompute lost data |
| **Flexible deployment** | Runs on Hadoop YARN, Kubernetes, Mesos, or standalone |
| **Multiple data sources** | Reads from HDFS, S3, Cassandra, JDBC, Kafka, etc. |

**What is a transformation?**

A transformation creates a new RDD or DataFrame from an existing one.

**Key points**:

* Lazy → doesn't compute immediately.
* Builds a DAG of execution.

**Examples**

| **Narrow** | **Wide** |
| --- | --- |
| map | reduceByKey |
| filter | groupByKey |
| flatMap | join |
| union (if no shuffle) | distinct |

**What is an action?**

An action triggers computation and returns result to driver or writes output.

**Examples**

| **Returns value** | **Outputs data** |
| --- | --- |
| collect() → returns all data to driver | saveAsTextFile() |
| count() → returns count | saveAsSequenceFile() |
| first() → first element |  |
| take(n) → first n elements |  |

Actions cause Spark to execute the DAG of transformations.