Data Engineering PySpark

Spark Architecture

Spark Components:

- Spark SQL: For structured data processing and SQL queries.
- Spark ML: For machine learning pipelines.
- Spark Graph: For graph processing and analysis.
- Spark Streaming: For real-time data processing.

Data APIs:

- DataFrame/Dataset APIs: Unified API for structured and semi-structured data.

Spark SQL Engine:

- Catalyst Optimizer: Optimizes queries for efficient execution.
- Tungsten: In-memory execution engine for faster performance.

Spark Core:

- Resilient Distributed Dataset (RDD): Fundamental data abstraction for fault-tolerance and parallelism.
- Supported Languages: Scala, Python, Java, and R.

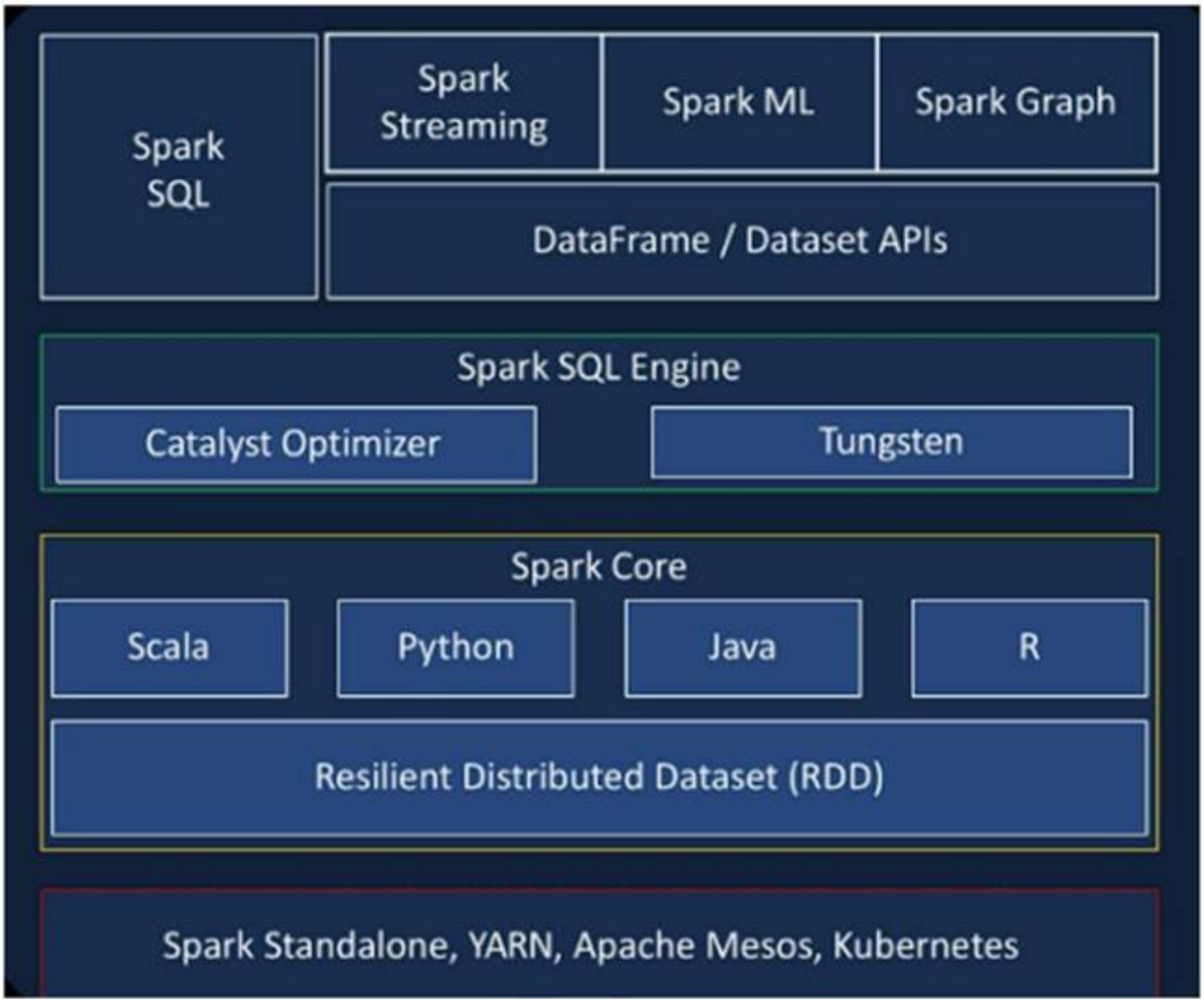
Cluster Managers:

- Spark Standalone: Self-contained cluster manager.
- YARN: Resource management framework for Hadoop ecosystem.
- Apache Mesos: Cluster manager for resource sharing across different frameworks.
- Kubernetes: Container orchestration platform for deploying and managing Spark clusters.

Key Points:

- Spark is a unified analytics engine for big data processing.
- It offers a variety of components for different data processing needs.
- It supports multiple programming languages for flexibility.
- It leverages distributed processing for scalability and fault-tolerance.

- It provides a variety of cluster management options for different deployment scenarios.


Spark SQL
Spark Streaming
Spark ML
Spark Graph
DataFrame / Dataset APIs
Spark SQL Engine
Catalyst Optimizer
Tungsten
Spark Core
Scala
Python
Java
R
Resilient Distributed Dataset (RDD)
Spark Standalone, YARN, Apache Mesos, Kubernetes