Apache Spark - written in Scala & runs in JVM

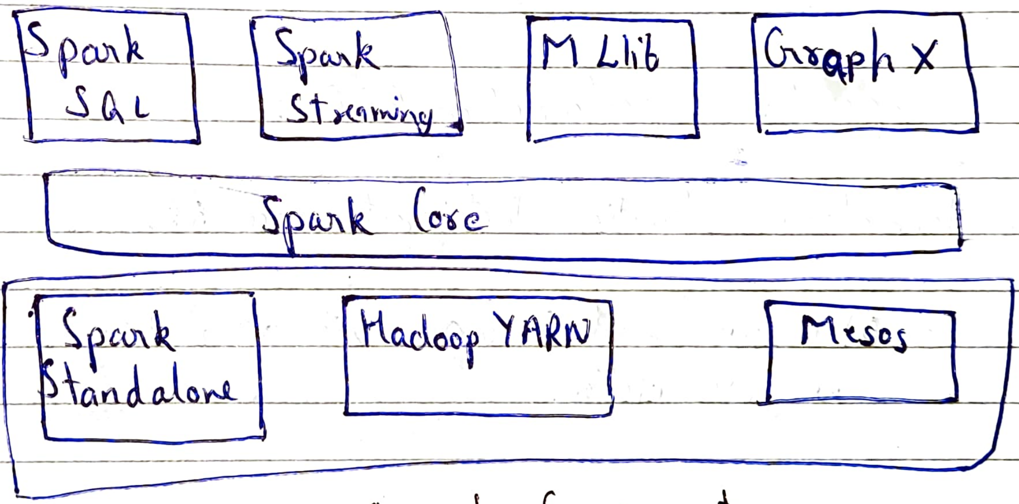
general purpose cluster computing sytem.

high-level API in Java, Scala, Python, R.

It also has abundant high-level tools for structured data processing, machine learning, graph processing & streaming.

6 components of Apache Spark ecosystem

- Spark core
- Spark SQL
- Spark streaming
- MLib
- Graphx
- SparkR

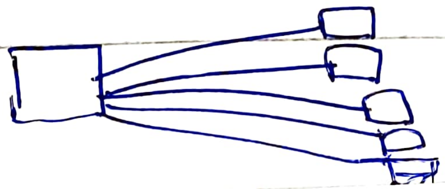


Spark Components

Apache Spark Core

It delivers speed by providing in-memory computation.

parallel & distributed computing



Key features —

- In charge of essential. I/o functionalities.
- Fault recovery
- Task dispatching
- It overcomes the snag ofmap-reduce by using in-memory computation.

Spark core is embedded with special collection called RDD (resilient distributed dataset).

Spark RDD handles partioning data across all the nodes in a cluster.

Two operations performed on RDDs:

Transformation → function that creates new RDD from the existing one's.

Action → When we want to work with actual dataset, then we use Action.

Spark SQL

It is a distributed framework for structured data processing

Features of Spark SQL

Cost band Optimizer

Mid-query fault-tolerance

Full compatibility with existing hive data.

Dataframes & SQL provide ways to access a variety of data sources.

Spark Streaming

Add-on to core spark API which allows scalable - high-throughput, fault-tolerant stream processing of live data streams

3 phases of Spark

(a) Gathering → Basic sources - Sources which are
→ Advanced sources sources like Kafka, Flume

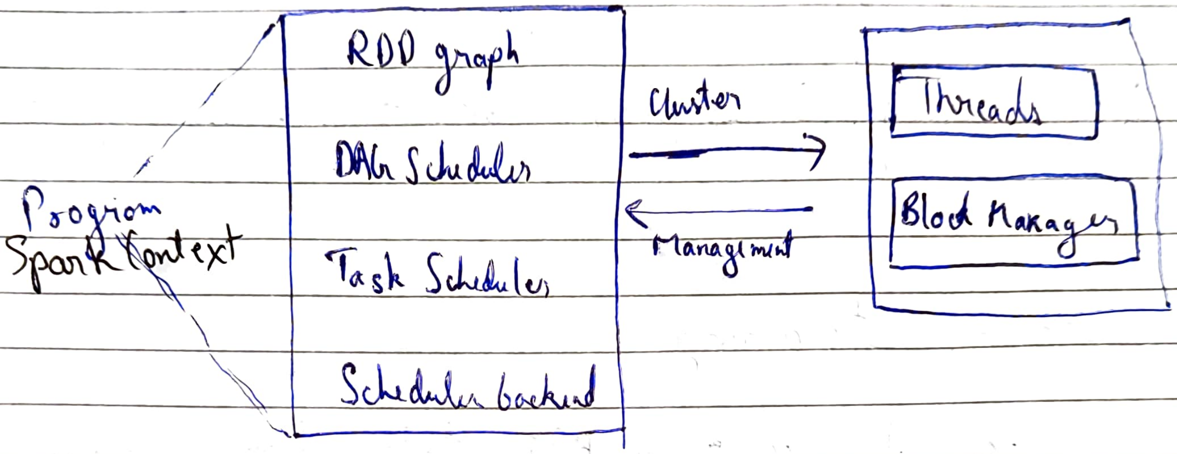
(b) Processing

(c) Data Storage → processed data is pushed out to file systems, databases & live dashboards.

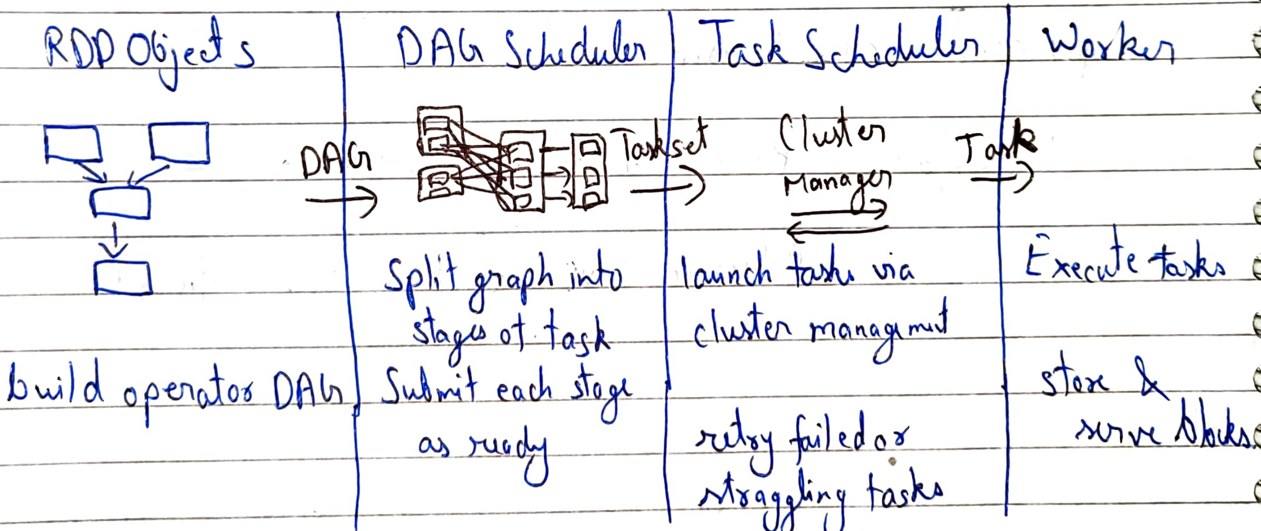
DStream in Spark signifies continuous stream of data.
DStream is internally a sequence of RDD's.

Spark MLlib(Machine learning library)

Machine learning library



~~Spark MLlib (Machine learning~~ How Spark Works —



RDD Objects	DAG Scheduler	Task Scheduler	Worker
build operator DAG	Split graph into stages of task Submit each stage as ready	launch tasks via cluster managment retry failed or straggling tasks	Execute tasks store & serve blocks