Apache spark
company software tool.

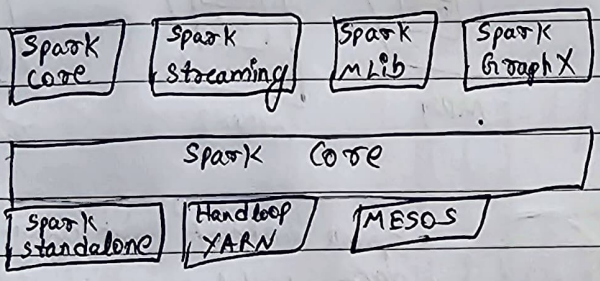
Apache Spark

Apache Spark → It is a general purpose cluster computing system. It provides us high-level API in Java, Scala, Pyton and R.

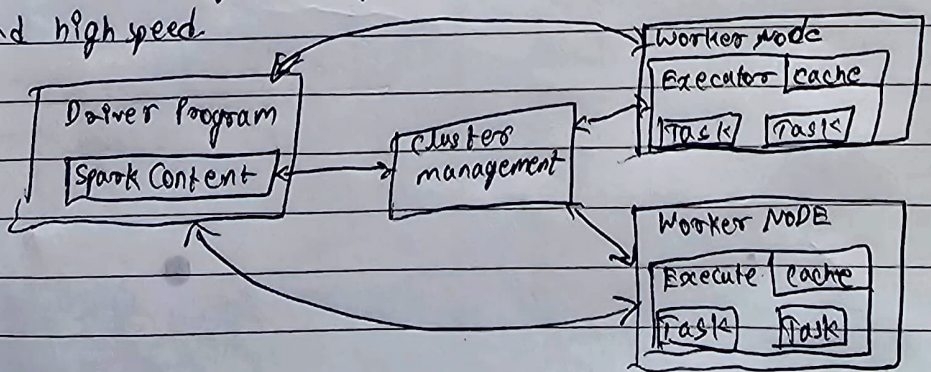
Spark SQL | Spark Streaming | Spark MLib | Spark GraphX | Spark R

→ Spark is a backend tool, where it we will help us to get the large amount of data (size of 1024TB) fastly.

Components of Apache:



- Apache spark core:- It is the foundational and essential component of the Apache spark framework. It serves as the underlying engine for distributed data processing in spark.
- Apache Spark SQL:- This component is a distributed framework for structed data processing. It works to access structed and semi-structed information.
- Apache spark streaming:- It is an add-on to core spark API which allows scalable, high-troughput, fault-tordent stream processing of live data streams.
- Apache Spark Mlib:- MLib (Machine learning library) in spark is a scable machine learning library that discusses both high-quality algorithm and high speed.



- Spark Context :- represents the connection to a spark cluster, and can be used to create RDD's, accumulators and broadcast variables on that cluster.
- DAGScheduler :- computes a DAG of stages for each job and submits them to taskScheduler determines preferred locations for tasks and finds minimum schedule to run the jobs.
- TaskScheduler :- responsible for sending tasks to the cluster, running them retrying if there are failures and mitigating straggles.
- SchedulerBackend :- backend interface for scheduling systems that allows plugging in different implementations