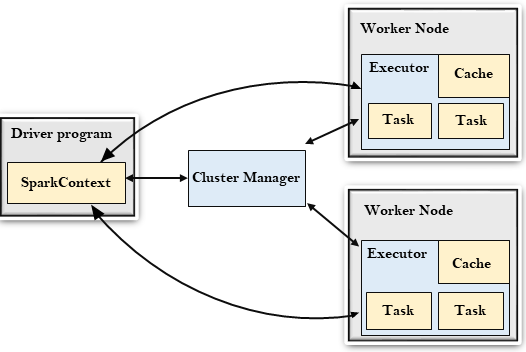
**Spark Architecture**



Spark architecture depend on

🡪RDD(Resilient Distributed Dataset)

R: it means fault tolerances ie due to some reason it fails to create rdd or local machine off, network outrage,maintain pointers to dataset it will reset its configuration automatically .

D: Each partition of the dataset can be done in each machine means 1 partition in machine another partition in another machine

D Dataset : set of data in a form of json csv parquet etc format

🡪DAG (Directed Acyclic Graph

The above image is based on dag

Driver Program { it has its own jvm and instruct to execute task }

The Driver Program is a process that runs the main() function of the application and creates the SparkContext object. The purpose of SparkContext is to coordinate the spark applications, running as independent sets of processes on a cluster.

To run on a cluster, the SparkContext connects to a different type of cluster managers and then perform the following tasks: -

🡪It acquires executors on nodes in the cluster.

🡪Then, it sends your application code to the executors. Here, the application code can be defined by JAR or Python files passed to the SparkContext.

🡪At last, the SparkContext sends tasks to the executors to run.

Cluster Manager{ group of many/single node and driver }

🡪The role of the cluster manager is to allocate resources across applications. The Spark is capable enough of running on a large number of clusters.

🡪It consists of various types of cluster managers such as Hadoop YARN, Apache Mesos and Standalone Scheduler.

🡪Here, the Standalone Scheduler is a standalone spark cluster manager that facilitates to install Spark on an empty set of machines.

Worker Node{ it may have multiple executor to execute task }

🡪The worker node is a slave node

🡪Its role is to run the application code in the cluster.

Executor{ it execute the task in its won jvm }

🡪An executor is a process launched for an application on a worker node.

🡪It runs tasks and keeps data in memory or disk storage across them.

🡪It read and write data to the external sources.

🡪Every application contains its executor.

Task{ a unit / workload given }

🡪A unit of work that will be sent to one executor.

Cluster

Driver

Core/Thread

Executor

Executor

Executor

Executor

Executor

Executor

Executor

Executor

Executor

Node

Node

Node