**Apache Spark:**

* Successor of Hadoop
* Both real time data and batch processing
* General Purpose clustering system
* Supports in-memory calculations- 100 times faster than Hadoop- reducing the number of read-write operations into the disk
* Offers high level APIs in Java, Python, Scala, R and SQL

**Hadoop**

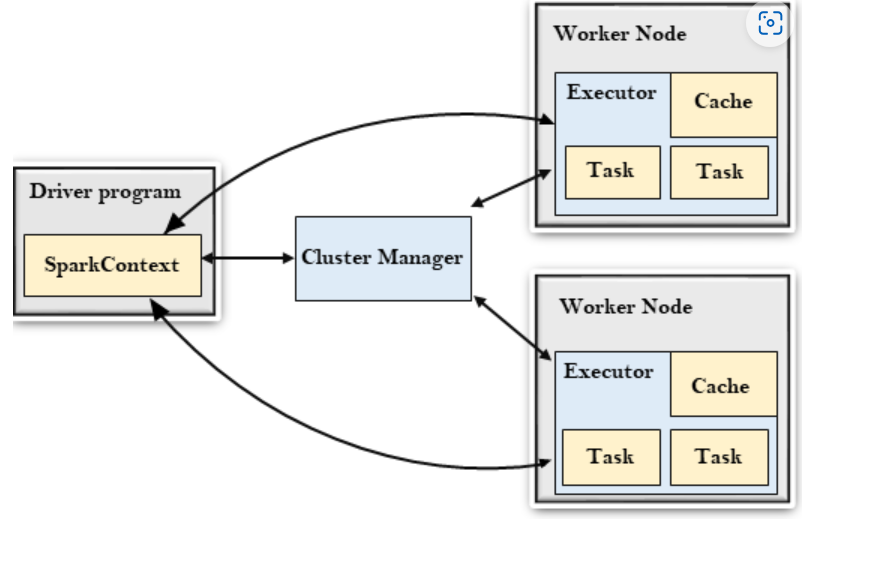
**HDFS:** to store data -distributed manner

**MR:** Map Reduce

**Limitation:**

* On disk computation:
* Batch data

**Spark Architecture:**



Cluster Manager:

-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.

Spark Context

* Entry point of spark functionality
* To communicate with cluster and to create RDD