Apache spark

It’s an open-source data processing engine to store and process data in real time across various clusters using of computers using simple programming constructs

Spark stores data in RAM (But how much ?? ) and contains Resilient Distributed datasets (RDD) which saves time taken in reading and writing operations and hence it is 100 times faster than hadoop. And these RDD’s are designed to handle the failure of any node same like apache Kafka.

Ensure no data loss.

Below are the components of apache spark



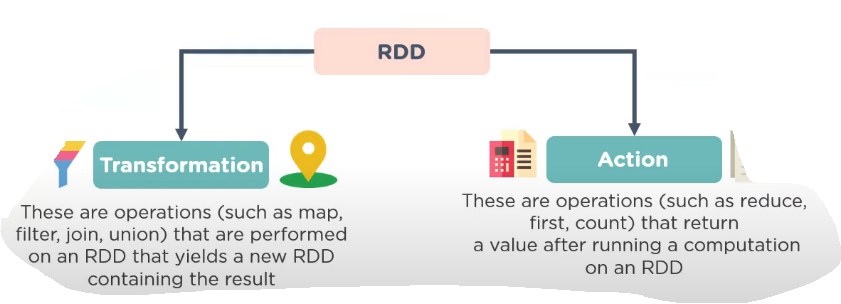
HDFS- hadoop distributed file system

Spark core is embedded with RDD (Resilient Distributed Datasets)

## RDD

These are same like collections , both RDD and Collections exists in memory

Resilient Distributed Data sets are an immutable, fault- tolerant, distributed collection of objects that can be operated in parallel.



RDD – same like collections –data sets= collection