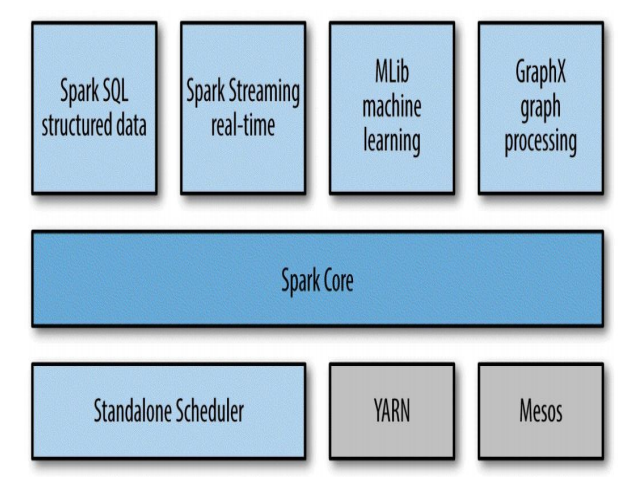
**Basic Definitions –**

Spark – Fast in memory cluster computation engine used for computation of big data

RDD – Resilient Distributed Datasets. Spark’s main programming abstraction. Collection of items distributed among many nodes that can be manipulated in parallel.

**Spark Ecosystem –**



Spark Core – Responsible for basic functionalities in spark like interacting with storage units, task scheduling, in-memory management and fault recovery.

Spark SQL – Spark’s package to work with structured data. It allows querying in SQL as well as HQL (Hive Query Language). It allows intermix of SQL with programmatic skills of java, Python and scala, thus, making it easier for complex computations. ‘Shark’ is the older version of Spark SQL.

Spark Streaming – Spark’s component to process live stream of data and batch processing.

MLib – Contains a library for common machine learning functionality.

GraphX – Library for manipulating graphs and performing graph parallel computations. It also has a library of common graph algorithms.

Cluster Managers for spark – Standalone cluster, Mesos cluster and YARN cluster.