- Architecture

=======================================================================

There are following components (Refer the diagram for their responsibilities)

- Driver

- Cluster Manager

- Executors on the Nodes

- Application Master (This is the part of the driver node in cluster mode, it handles all the resources.

To explain in simple manner, diagram describe this as part of driver only.)

------------------------------------------------------------------------------------------------------------------------

Types of deployment mode

- Cluster - Driver and Application Master are on the same node of the cluster.

- Client - Driver is on client machine and AM is on one of the node.

- Local - Everything is on one single machine.

------------------------------------------------------------------------------------------------------------------------

Data Locality in Spark

- PROCESS\_LOCAL - data and task is on same JVM

- NODE\_LOCAL - On same node but not on same JVM. Data is transferred between the process.

- RACK\_LOCAL - On different server. Network will happen no matter what. (S3)

- ANY – Somewhere on the network.

- NO\_PREF – data is accessed equally quickly from anywhere and has no locality preference

=======================================================================

=======================================================================

- RDD

=======================================================================

Collection of data elements on which parallel execution can be done.