SPARK

Outline

- Introduction to Spark
- MR vs Spark
- Spark Components
- RDD Overview
- Spark Architecture

What is Spark?

Apache :

Spark™ is a fast and general engine for large-scale data processing.

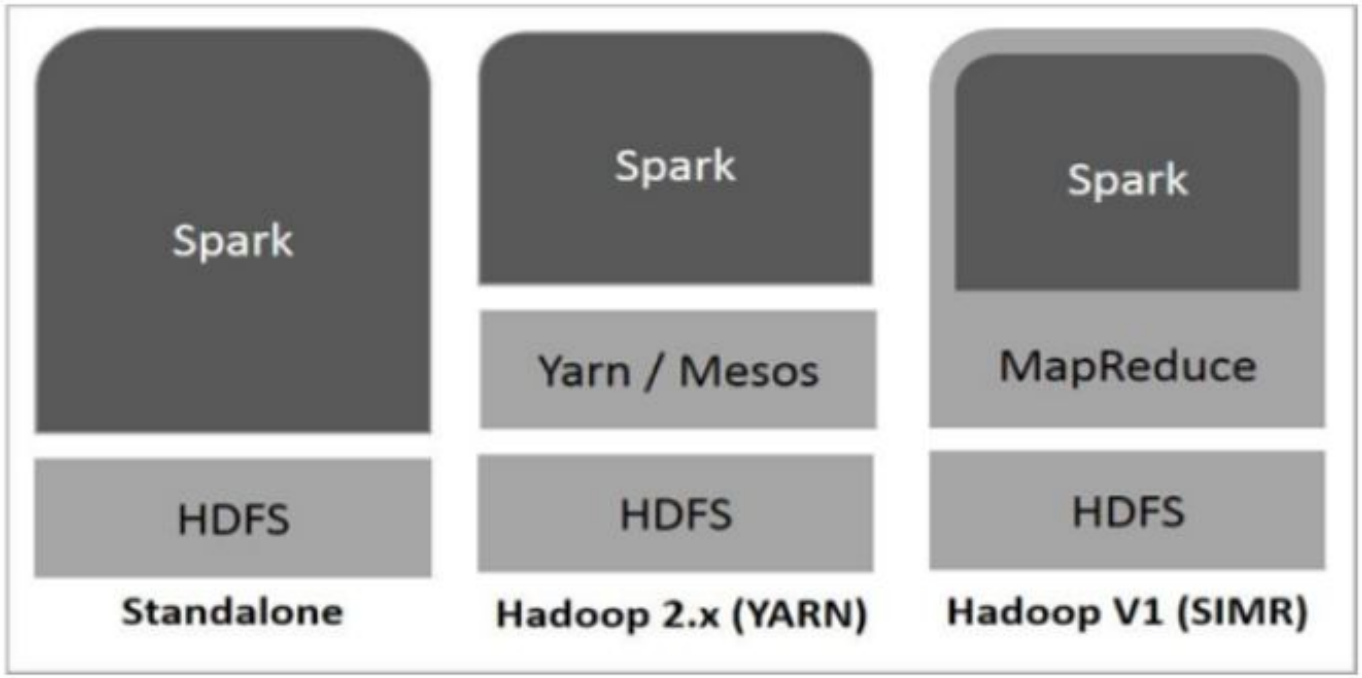
Datairicks:

Spark™ is a powerful open source processing engine built around speed, ease of use, and sophisticated analytics.

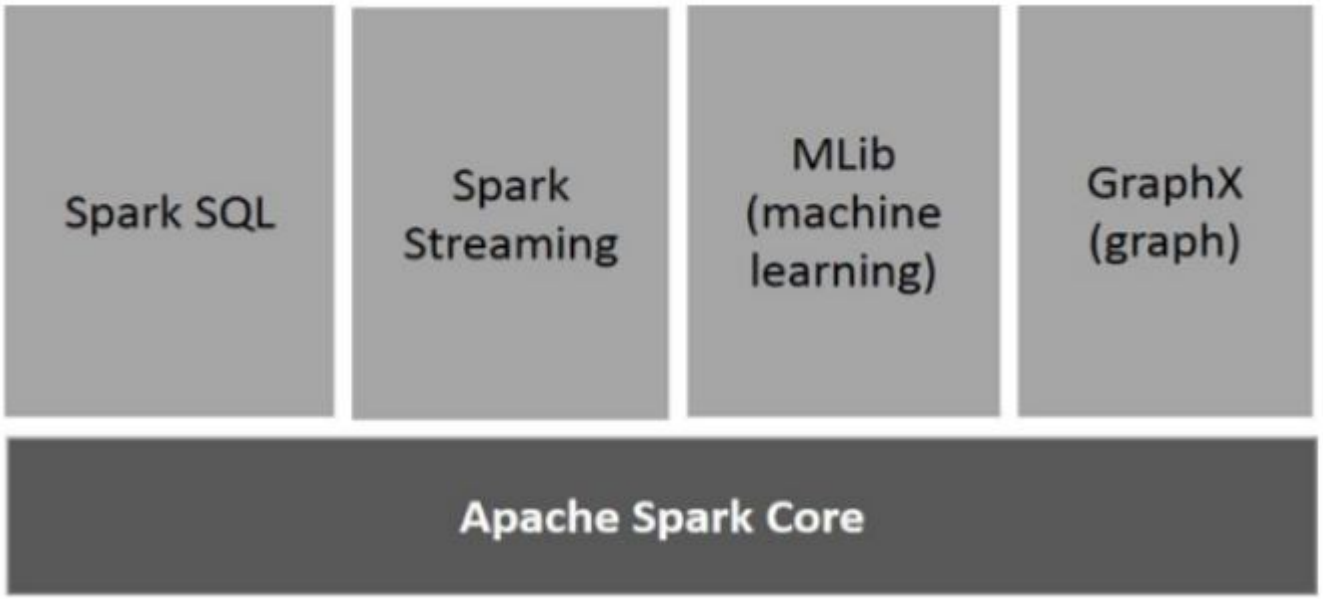
Spark is open source distributed computing engine for data processing and data analytics.

- **It was originally developed at UC Berkeley in 2009**

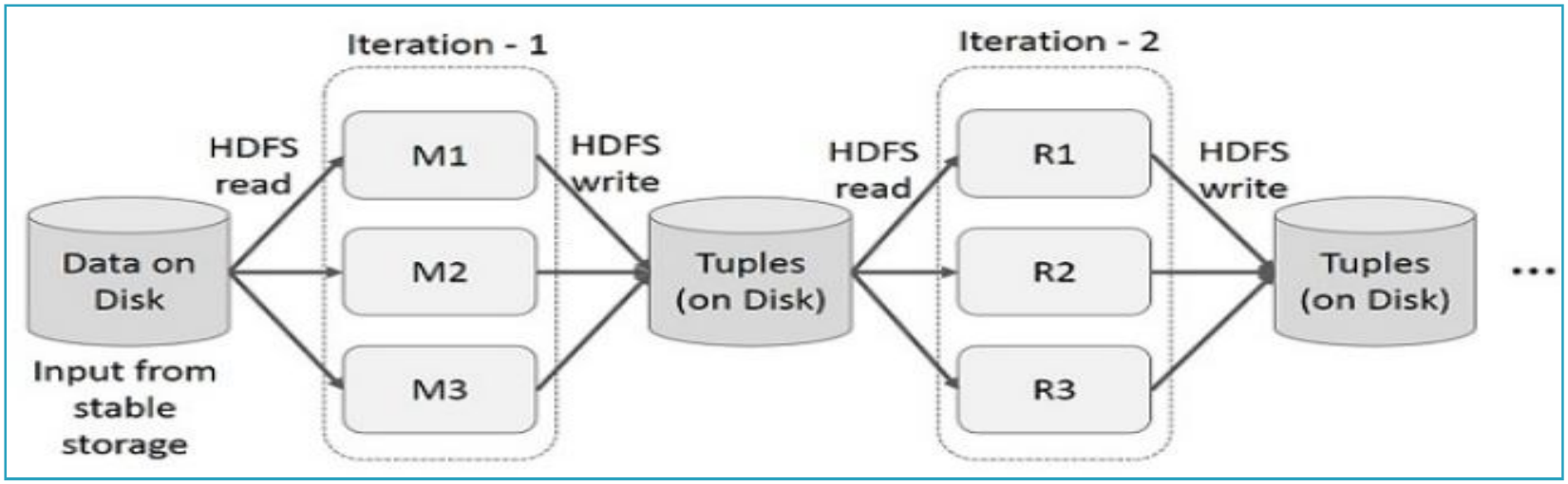
Spark Built on Hadoop

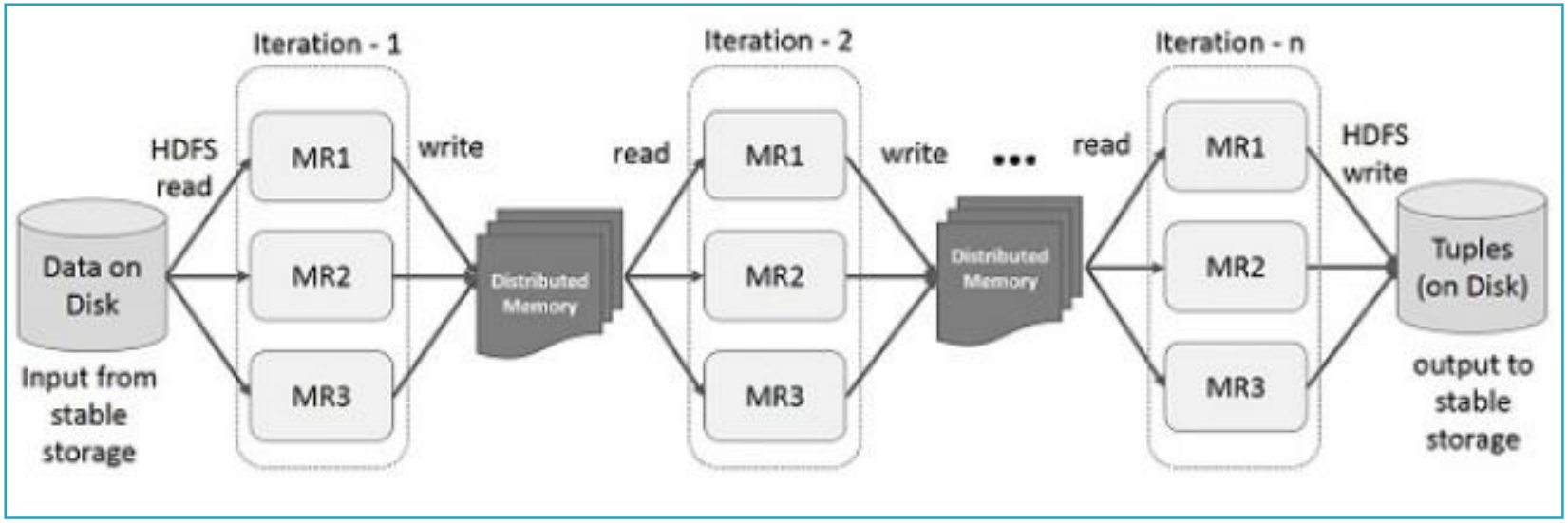


Spark Components

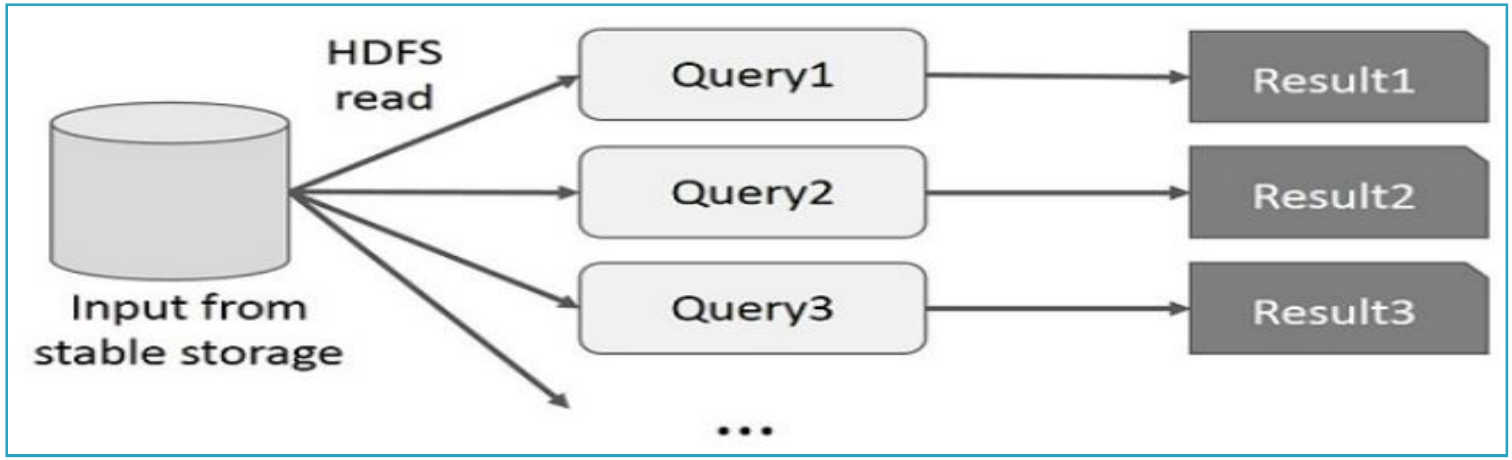


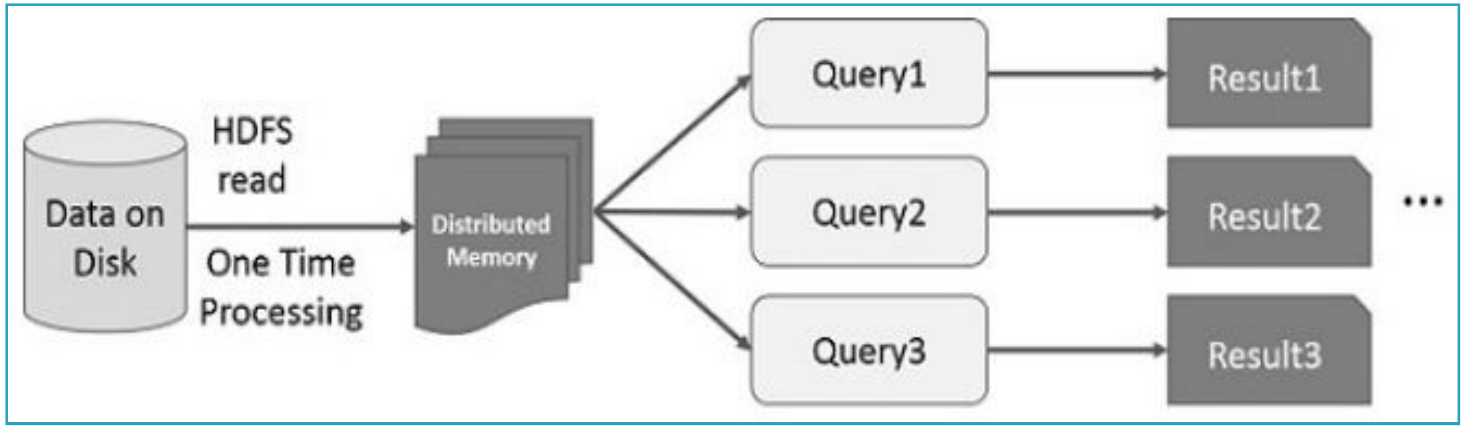
Spark vs MR – Iterative operation use case



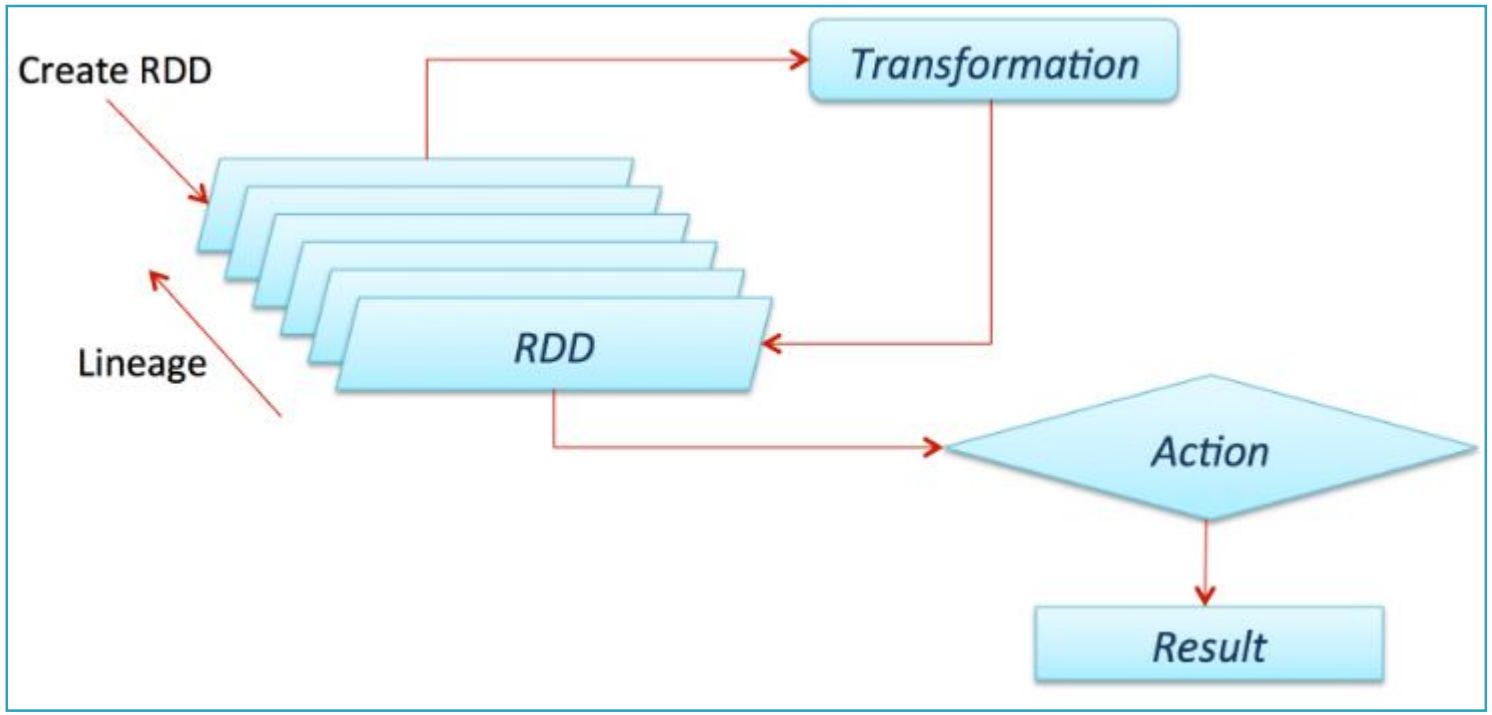


Spark vs MR – Interactive operation use case

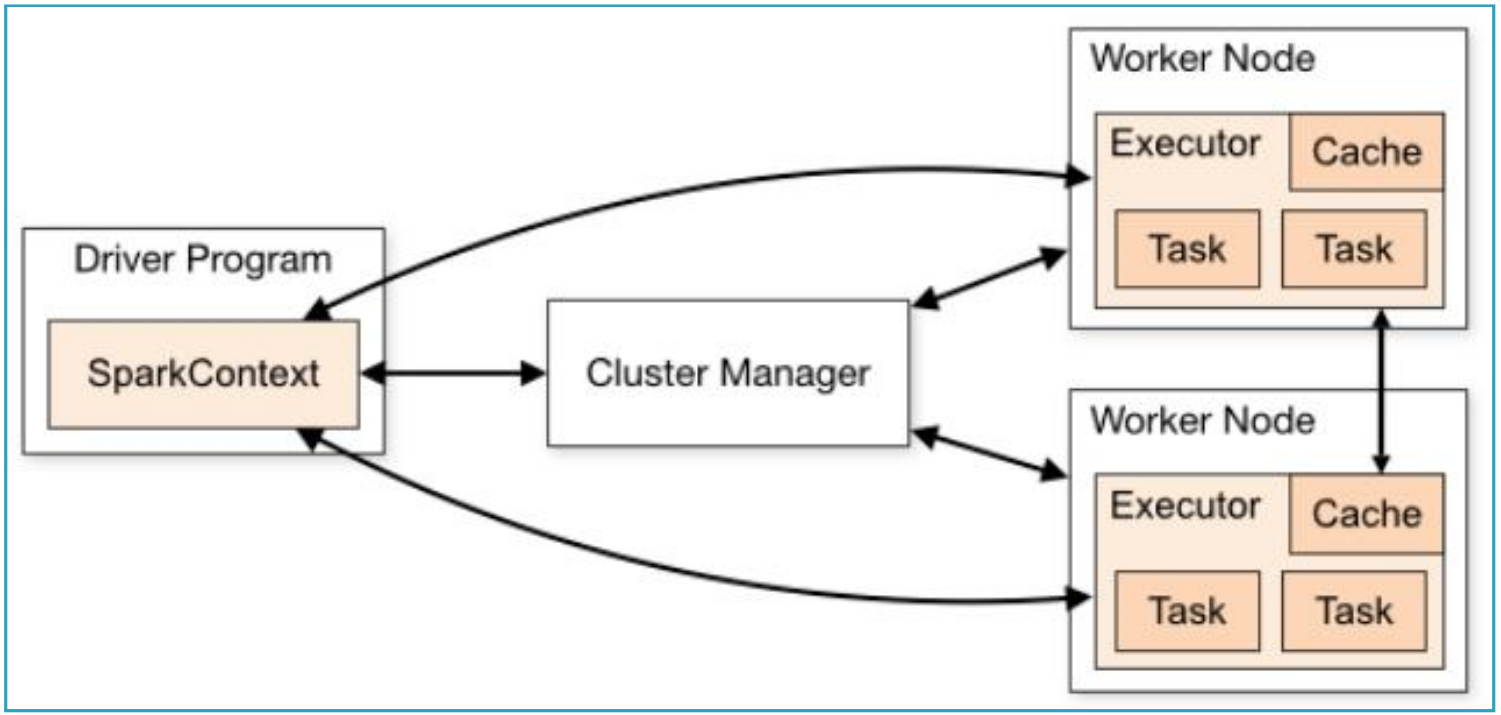




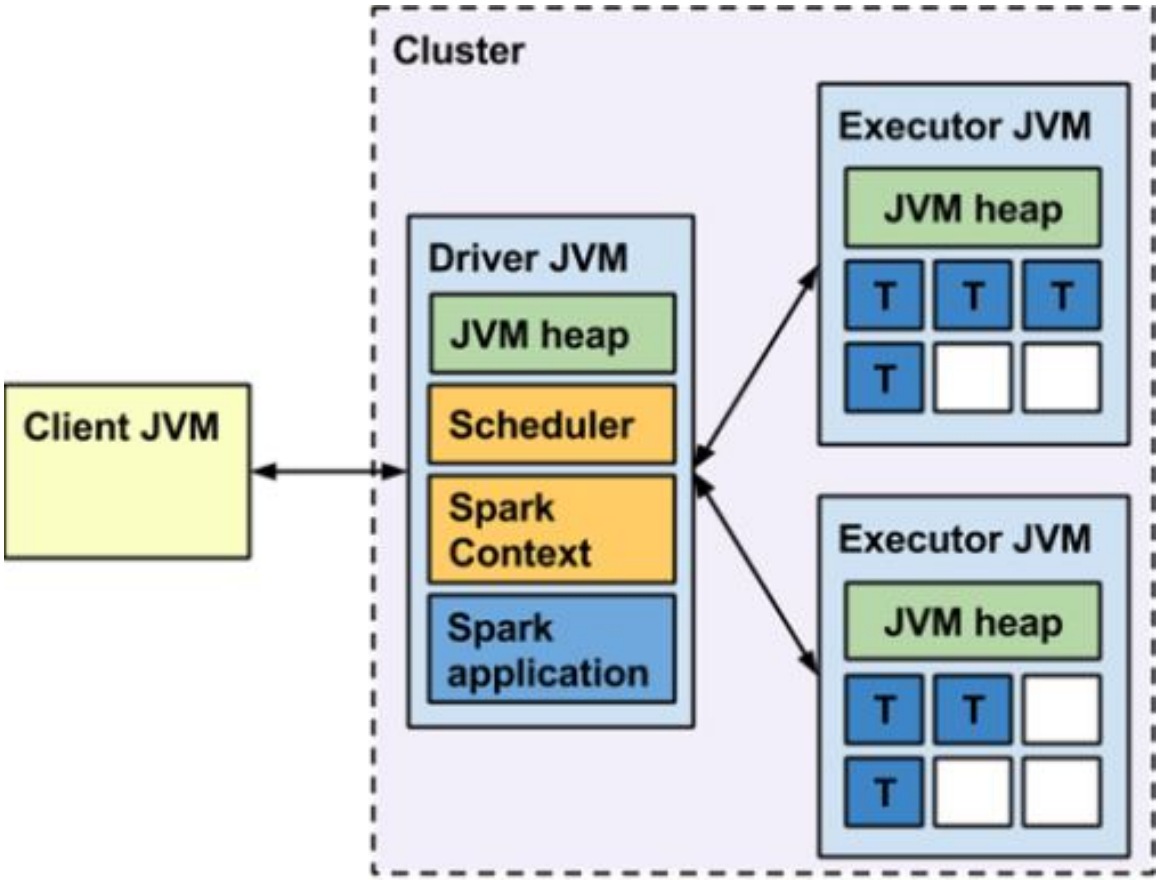
Spark RDD



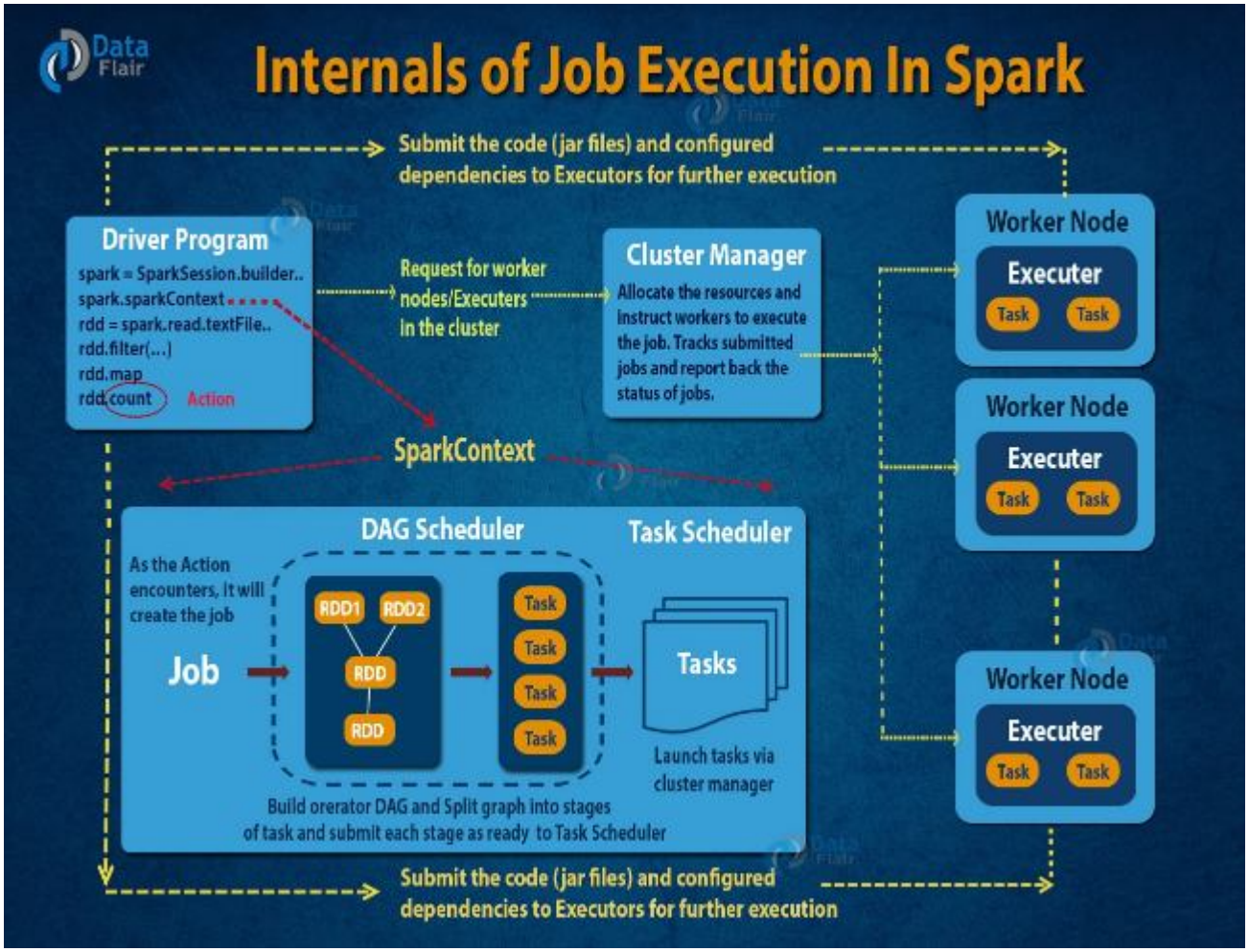
Spark Cluster Architecture



Spark Cluster Architecture (contd.)



Internals of Job Execution



Thank You!