Apache Spark

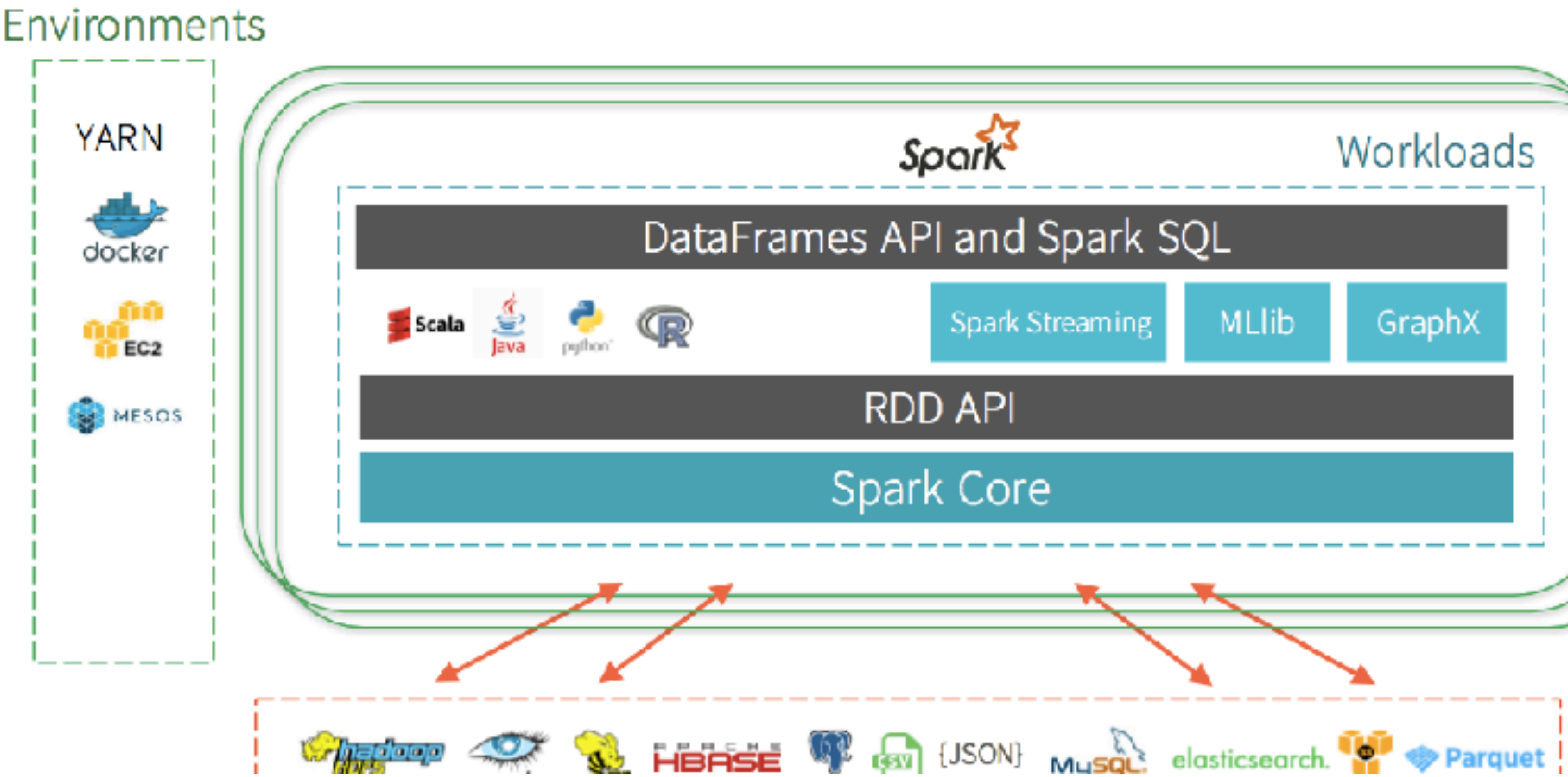
Ladle Patel

Apache Spark

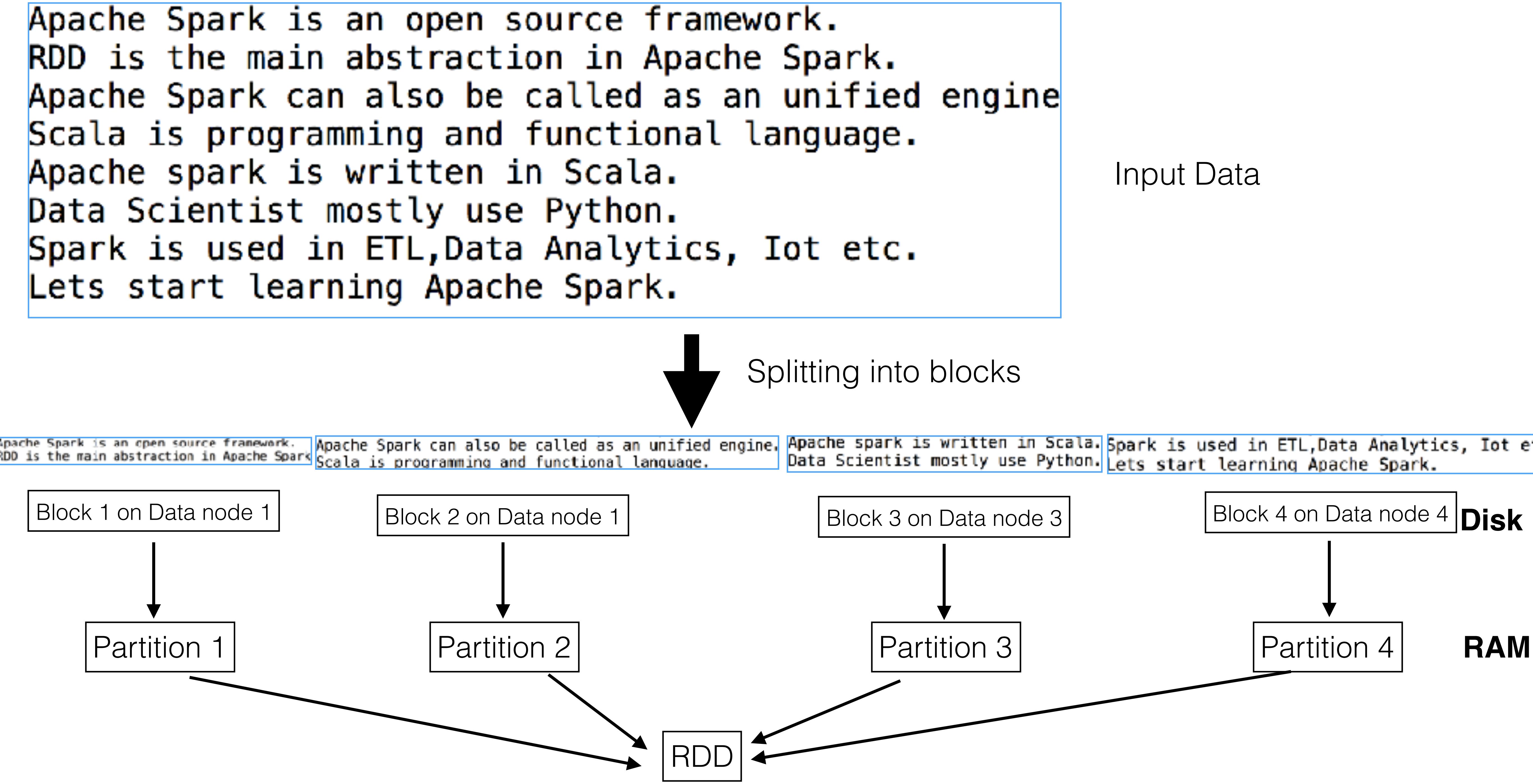
- Started as a research project at UC Berkeley in 2009 .
- Open Source License (Apache 2.0) .
- Latest Stable Release: Spark 2.1.0 released on Dec 28, 2016.
- 1130634 lines of code (68% Scala,16% Java,8% Java,8 % Other).
- Built by 1300+ developers from 200+ companies .

Unified Engine

Apache Spark is an open-source distributed general-purpose cluster computing framework with in-memory data processing engine that can do ETL, analytics, machine learning and graph processing on large volumes of data at rest (batch processing) or in motion (streaming processing) with rich high-level APIs for the programming languages: Scala, Python, Java and R



Main Abstraction in Spark



```
Apache Spark is an open source framework.
RDD is the main abstraction in Apache Spark.
Apache Spark can also be called as an unified engine
Scala is programming and functional language.
Apache spark is written in Scala.
Data Scientist mostly use Python.
Spark is used in ETL,Data Analytics, Iot etc.
Lets start learning Apache Spark.
```

Input Data

Splitting into blocks

```
pache Spark is an open source framework.
DD is the main abstraction in Apache Spark
```

```
Apache Spark can also be called as an unified engine.
Scala is programming and functional language.
```

```
Apache spark is written in Scala.
Data Scientist mostly use Python.
```

```
Spark is used in ETL,Data Analytics, Iot e
Lets start learning Apache Spark.
```

Block 1 on Data node 1

Block 2 on Data node 1

Block 3 on Data node 3

Block 4 on Data node 4

Disk

Partition 1

Partition 2

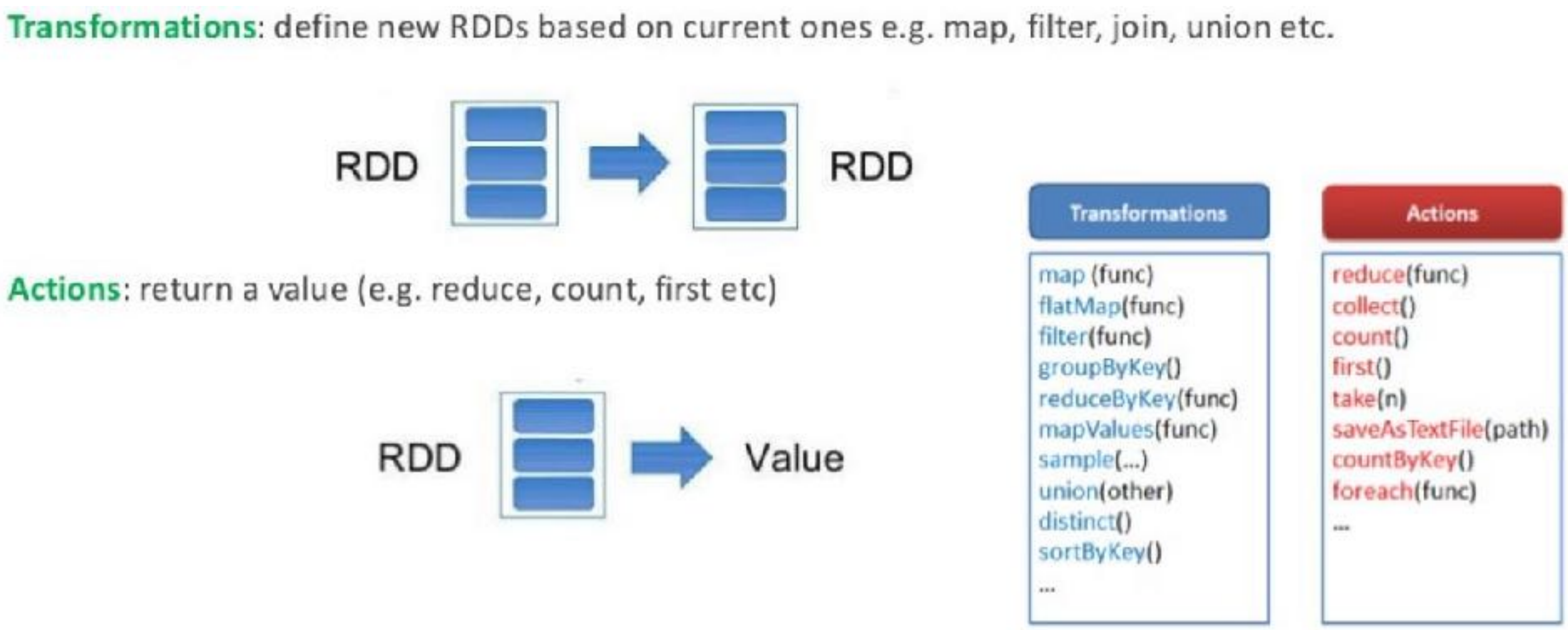
Partition 3

Partition 4

RAM

RDD

Operation on RDD



Transformations: define new RDDs based on current ones e.g. map, filter, join, union etc.

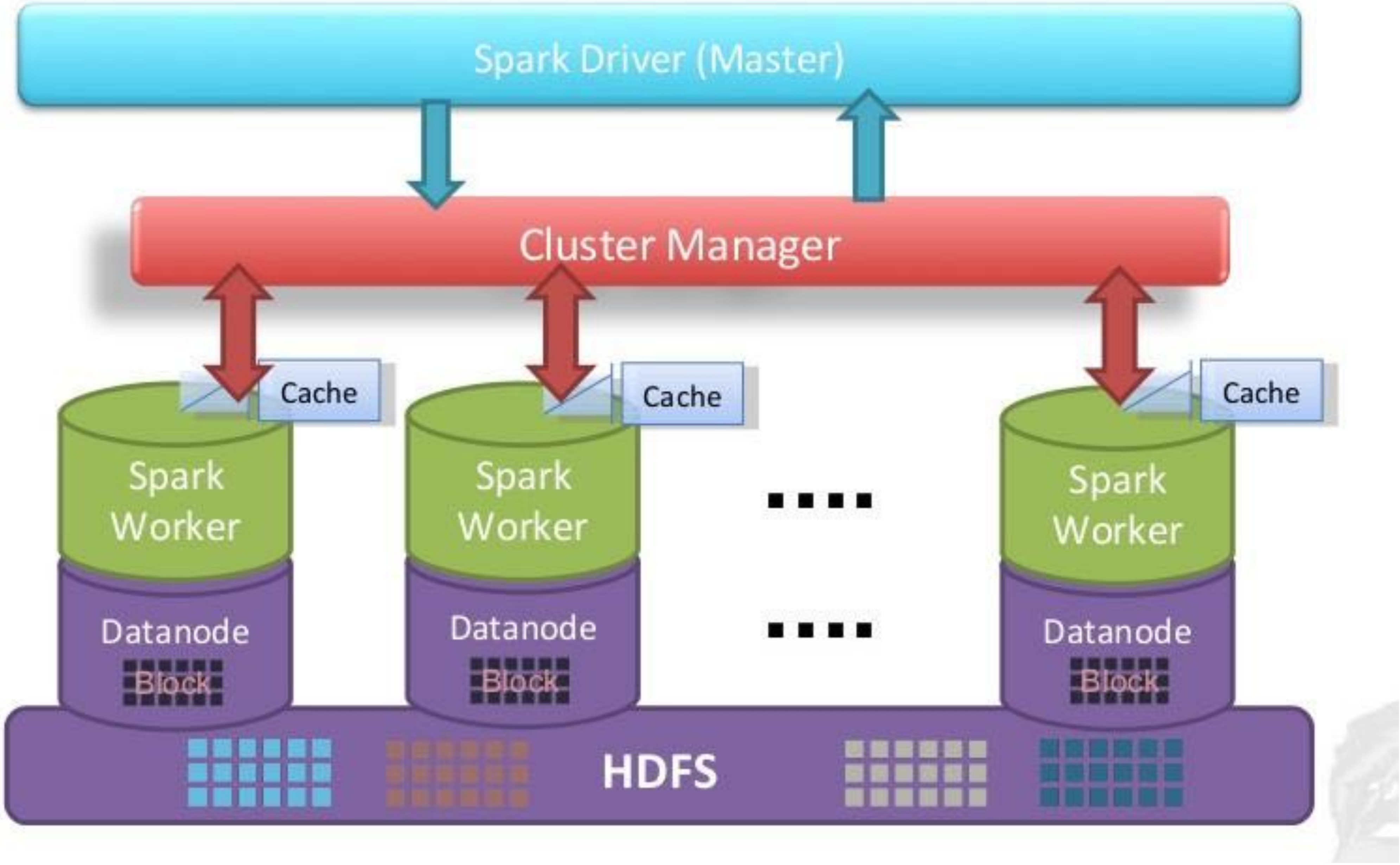
RDD
RDD

Actions: return a value (e.g. reduce, count, first etc)

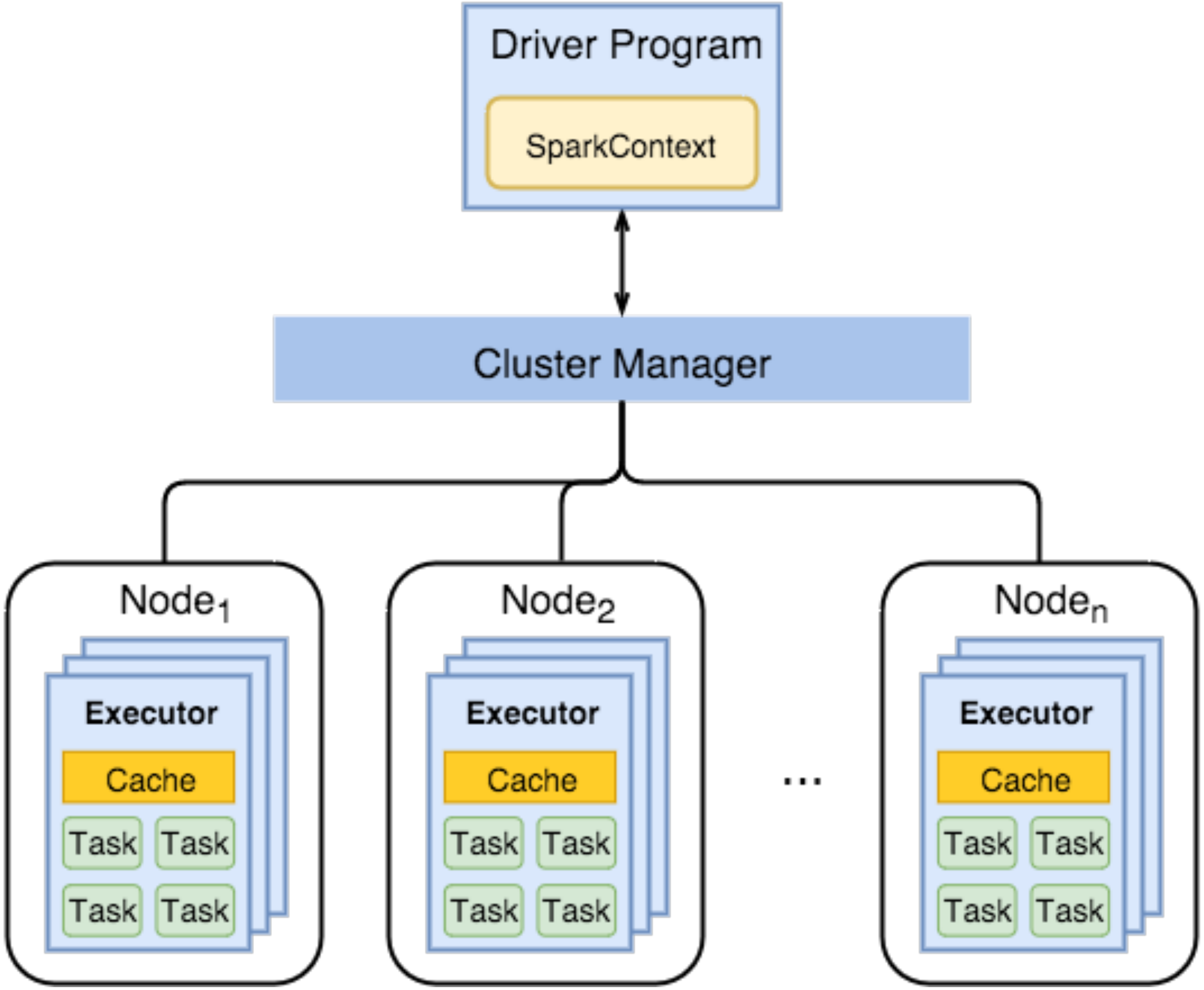
RDD
Value

Transformations	Actions
map (func)	reduce(func)
flatMap(func)	collect()
filter(func)	count()
groupByKey()	first()
reduceByKey(func)	take(n)
mapValues(func)	saveAsTextFile(path)
sample(...)	countByKey()
union(other)	foreach(func)
distinct()	...
sortByKey()	
...	

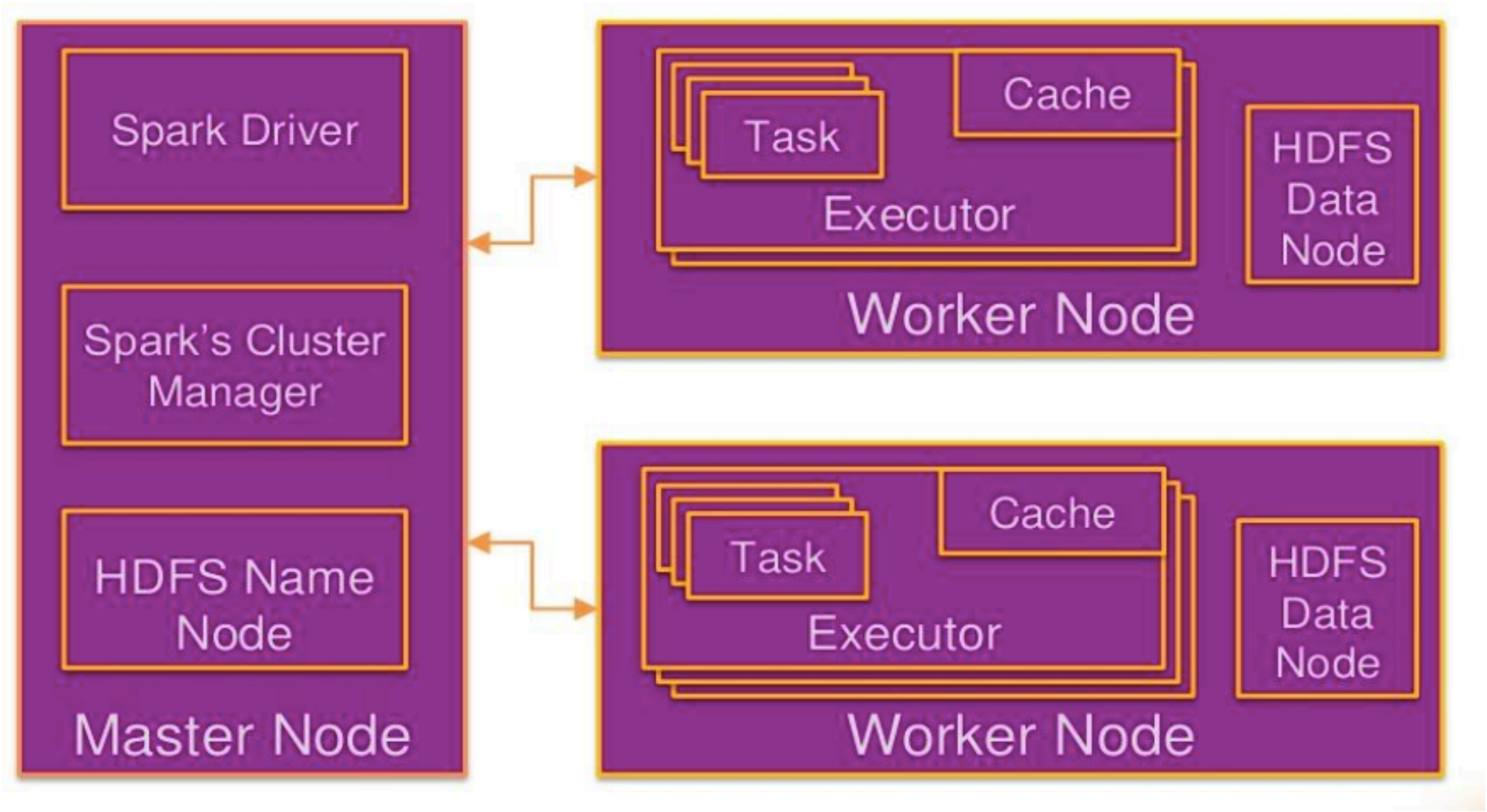
Architecture



Architecture Cont..



Architecture Cont..



Terminologies

- Job: A piece of code which reads some input from HDFS or local, performs some computation on the data and writes some output data.

- Driver: The program/process which has main() method and responsible for running the Job over the Spark Engine.It also Creates SparkContext to schedule jobs execution and negotiate with cluster manager

- Spark context: Object which sets up internal services and establishes a connection to a Spark execution environment.Once a SparkContext is created you can use it to create RDDs, accumulators and broadcast variables, access Spark services and run jobs (until SparkContext is stopped).Its an entry point in Spark application.

- RDD :Is a resilient and distributed collection of records spread over one or many partitions.

Terminologies Cont..

- Master: The machine on which the Driver program runs.
- Slave: The machine on which the Executor program runs.
- Executor: The process responsible for executing a task.It runs tasks scheduled by driver.Stores computation results in memory, on disk or off-heap interact with storage systems.
- Cluster Manager:Cluster Manager is responsible for monitoring the cluster and provides the resources for executors. Ex : Mesos, YARN and Spark Standalone

Thank You.