hadoop

APACHE
Spark™

elasticsearch

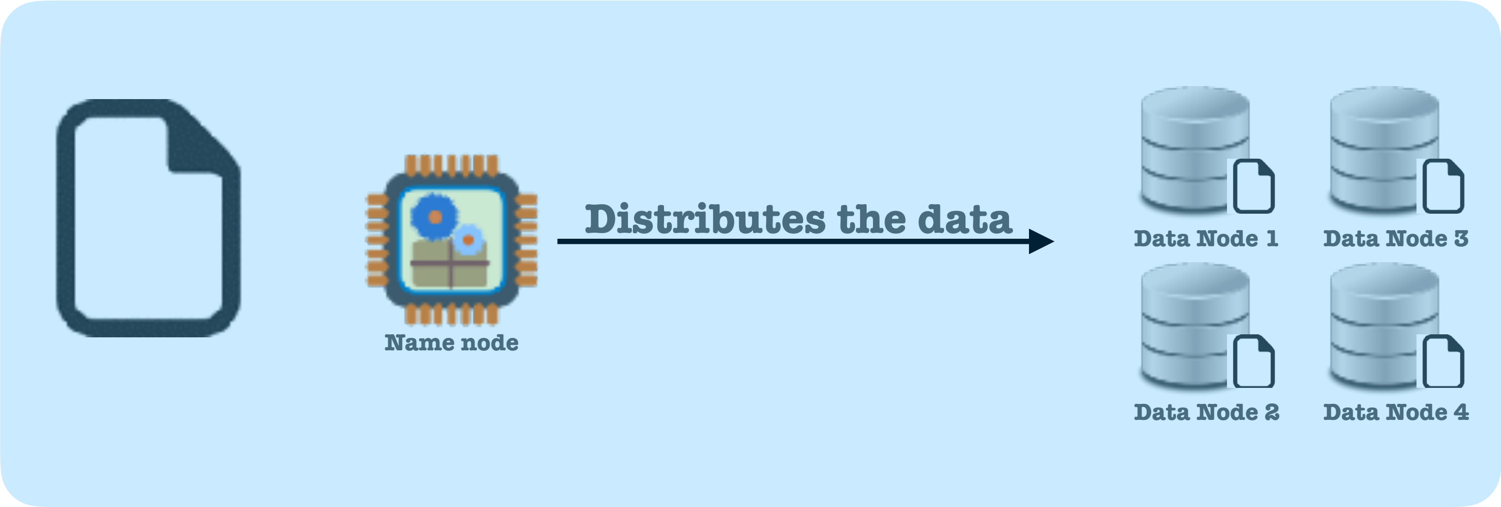
Topics

1. Hadoop
2. Map Reduce life cycle
3. Yarn
4. Why did spark came into picture?
5. What are the components of spark?
6. What are the programming languages supported by spark?
7. What are the operations present in the spark?
8. What is the architecture of spark?
9. What happens when you run your code?
10. What is RDD in spark?

1 Hadoop

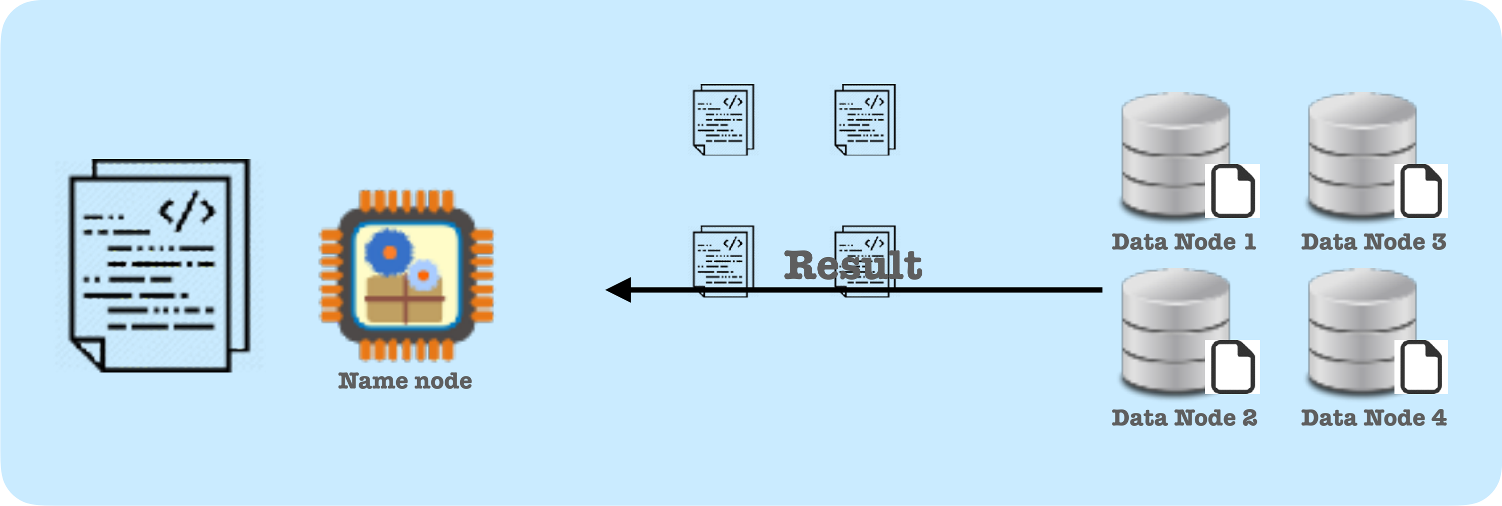
Let’s understand Hadoop by it’s components.
There are two main components present in the Hadoop those are:

1. HDFS - Hadoop Distributed File System



HDFS helps to store the big data in distributed environment so that you can process it in parallel.

2. Map Reduce - Processing layer



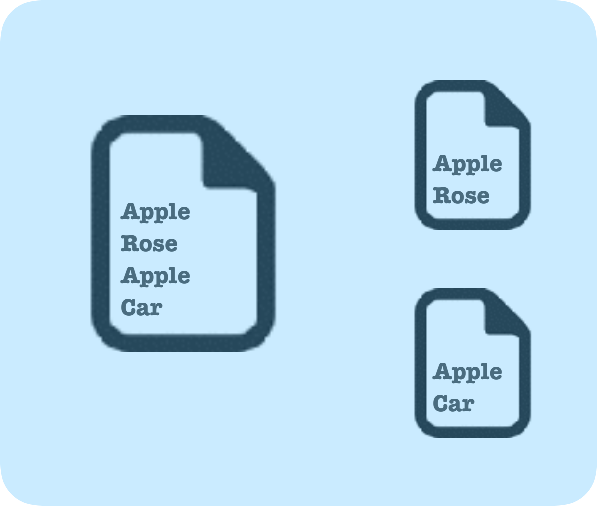
Map reduce helps to process the big data in distributed environment in parallel.

2 Map Reduce life cycle

https://www.youtube.com/watch?v=5cbkRX8Jes8

Whole map reduce process happens through following steps:
Let's take word count example and see life cycle of map reduce

1. Splitting

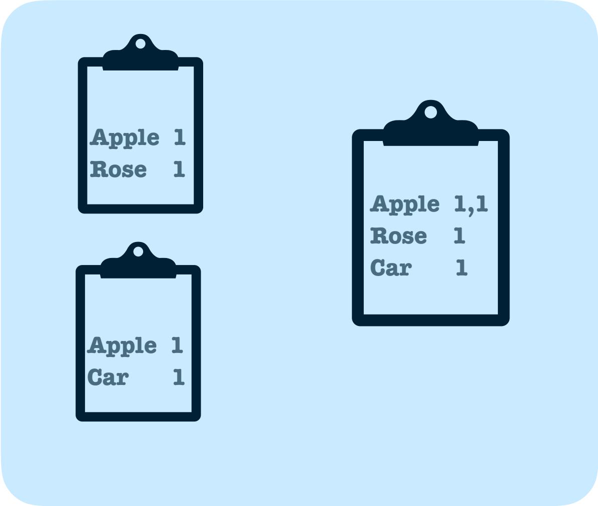


Apple
Rose
Apple
Car

Apple
Rose

Apple
Car

3. Shuffling

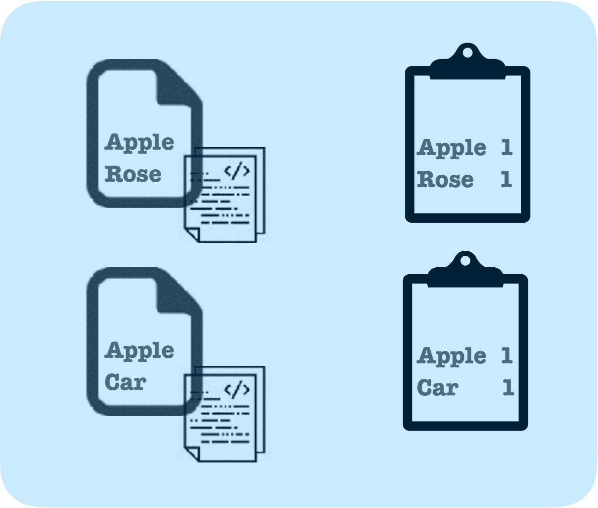


Apple 1
Rose 1

Apple 1
Car 1

Apple 1,1
Rose 1
Car 1

2. Mapping



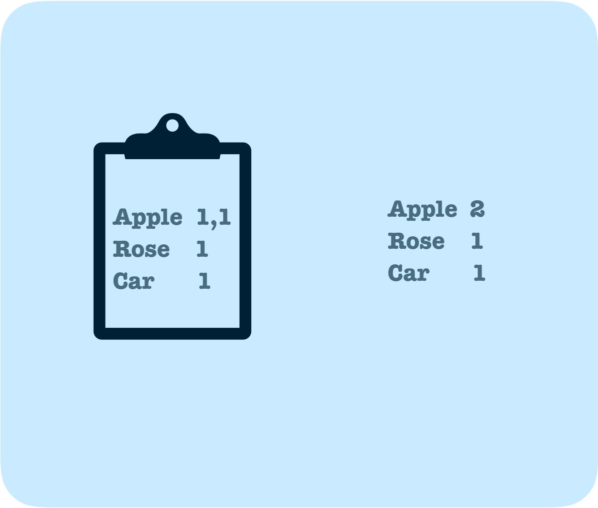
Apple
Rose

Apple
Car

Apple 1
Rose 1

Apple 1
Car 1

4. Reducing



Apple 1,1
Rose 1
Car 1

Apple 2
Rose 1
Car 1

Name node

Data node

HDFS - Hadoop Distributed File System

Map reduce

3 Yarn

YARN performs all your processing activities by allocating resources and scheduling tasks.

It has two major daemons, i.e. ResourceManager and NodeManager.

ResourceManager is a present in each cluster and runs on the master machine.

NodeManager is present on each node and runs on each slave machine.

Name node

Data node

HDFS - Hadoop Distributed File System

Map reduce

ResourceManager

NodeManager

4 Why did spark came into picture?

Spark is 100 times faster than map reduce due to in memory storage and rich APIs

Hadoop supports only batch processing and does not support for real time processing

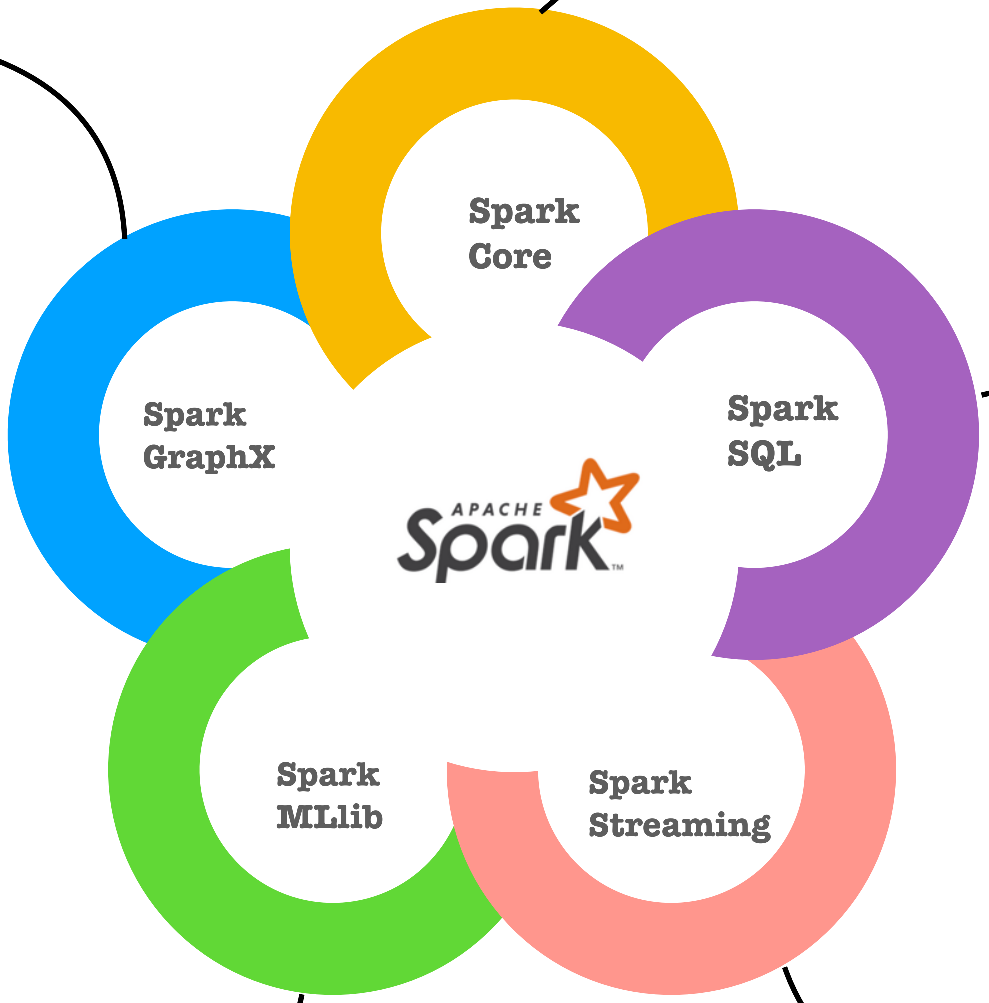
Apache Storm / S4 can only perform stream processing and does not support for batch processing

Apache Impala / Apache Tez can only perform interactive processing

Neo4j / Apache Giraph can only perform graph processing

Hence in the industry, there is a big demand for a powerful engine that can process the data in real-time (streaming) as well as in batch mode.

5 What are the components of spark?



This is where you do transformations and actions on your data

GraphX is a library for manipulating graphs and performing graph-parallel computations.

It allows querying data via SQL as well as the apache Hive variant of SQL—called the Hive Query Language

MLlib provides multiple types of machine learning algorithms.

Spark Streaming is a Spark component that enables processing of live streams of data

6

What are the programming languages supported by spark?

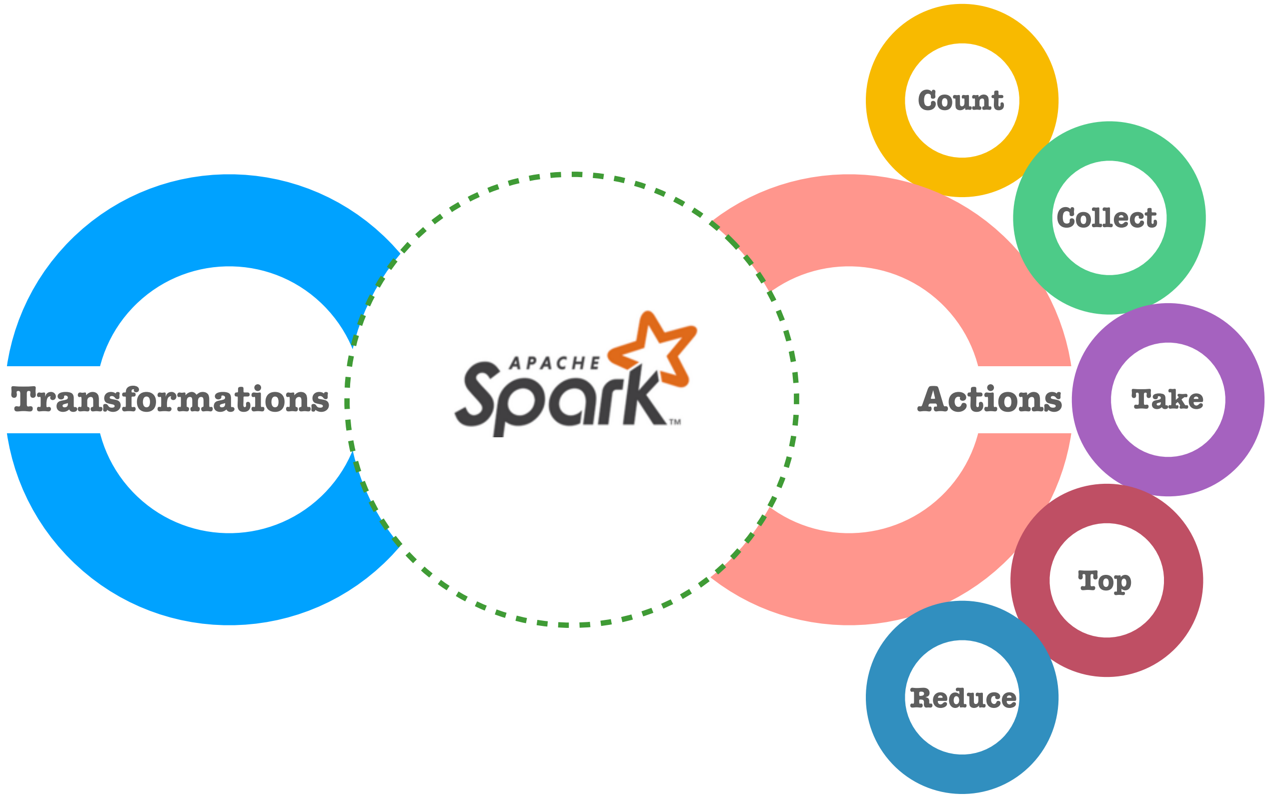
Scala

Java

Python

R

7 What are the operations present in the spark?



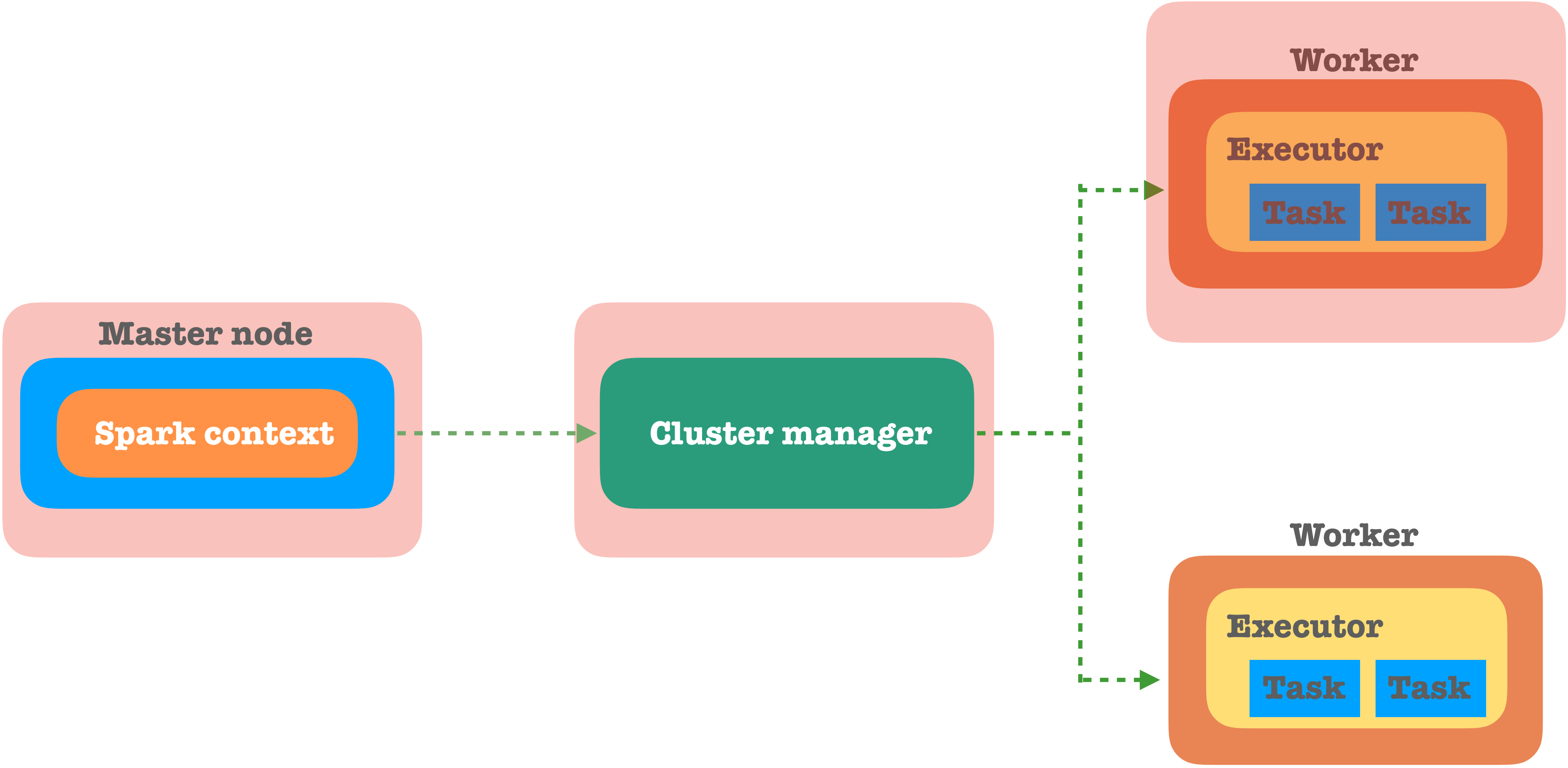
Spark Transformation is a function that produces new RDD from the existing RDDs.

Spark action is a function that is used to get some information of the data without creating new RDD.

1	10
2	20
3	30
4	40
5	50
6	60

1	6
2	
3	
4	
5	
6	

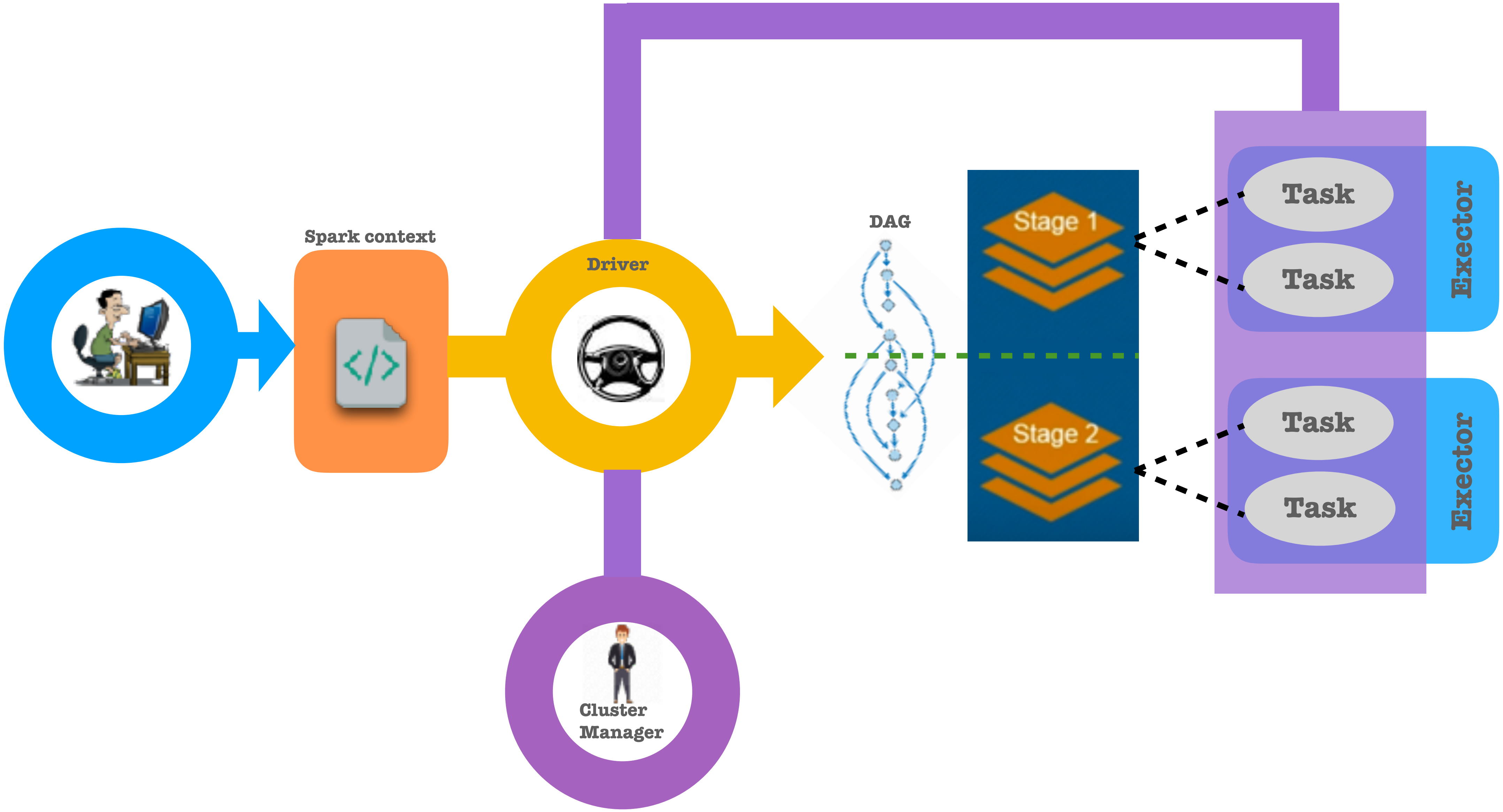
8 What is the architecture of spark?



1 2 3

- Worker nodes are slave nodes whose job is to execute the task.
- Task is then executed in executor of worker node
- Once the tasks has been completed worker will return the result back to spark context

9 What happens when you run your code?



1. Client submits user application code to spark context. This user code consist of Transformations and actions.
2. Driver will convert this user code to logically directed graph called DAG.
3. Driver will also convert this DAG into many stages.
4. In order to execute these stages it also creates tasks.
5. Now that tasks are created in order to execute them driver will ask resources from cluster manager.
6. Cluster manager will give resources/ executors to driver to execute the tasks.

10 What is RDD in spark?

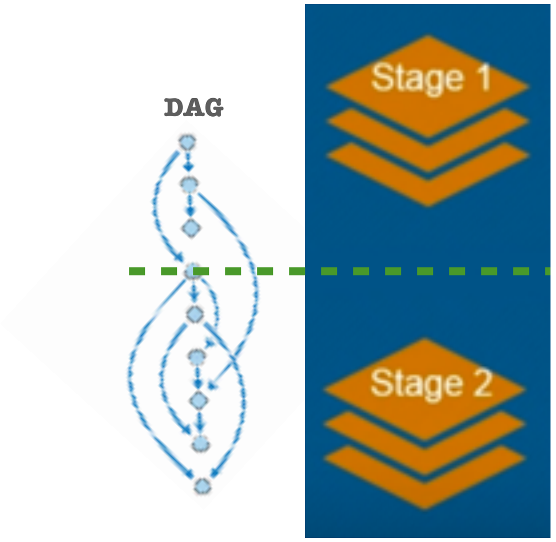
Resilient Distributed Datasets

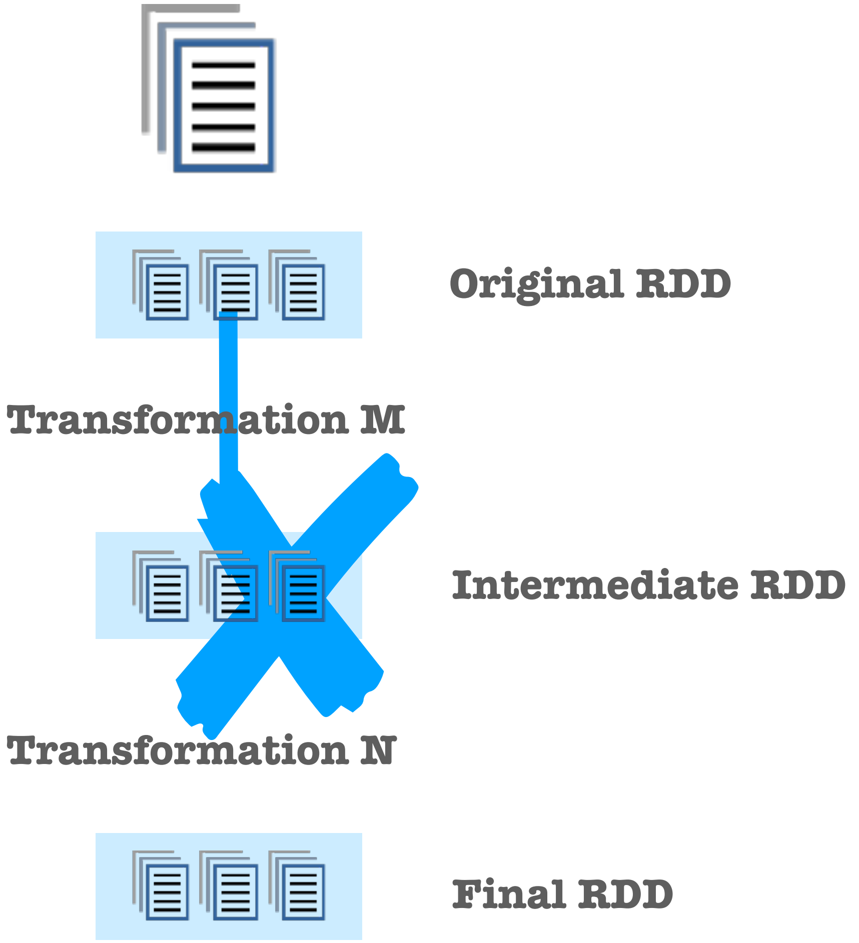
It is kind of a data structure.

All the data residing in this data structure will be distributed into multiple different servers and this will help us to parallel computing.

RDD is immutable collection of objects

RDDs are fault tolerance.





Original RDD

Transformation M

Intermediate RDD

Transformation N

Final RDD

Master node

Cluster manager

Worker

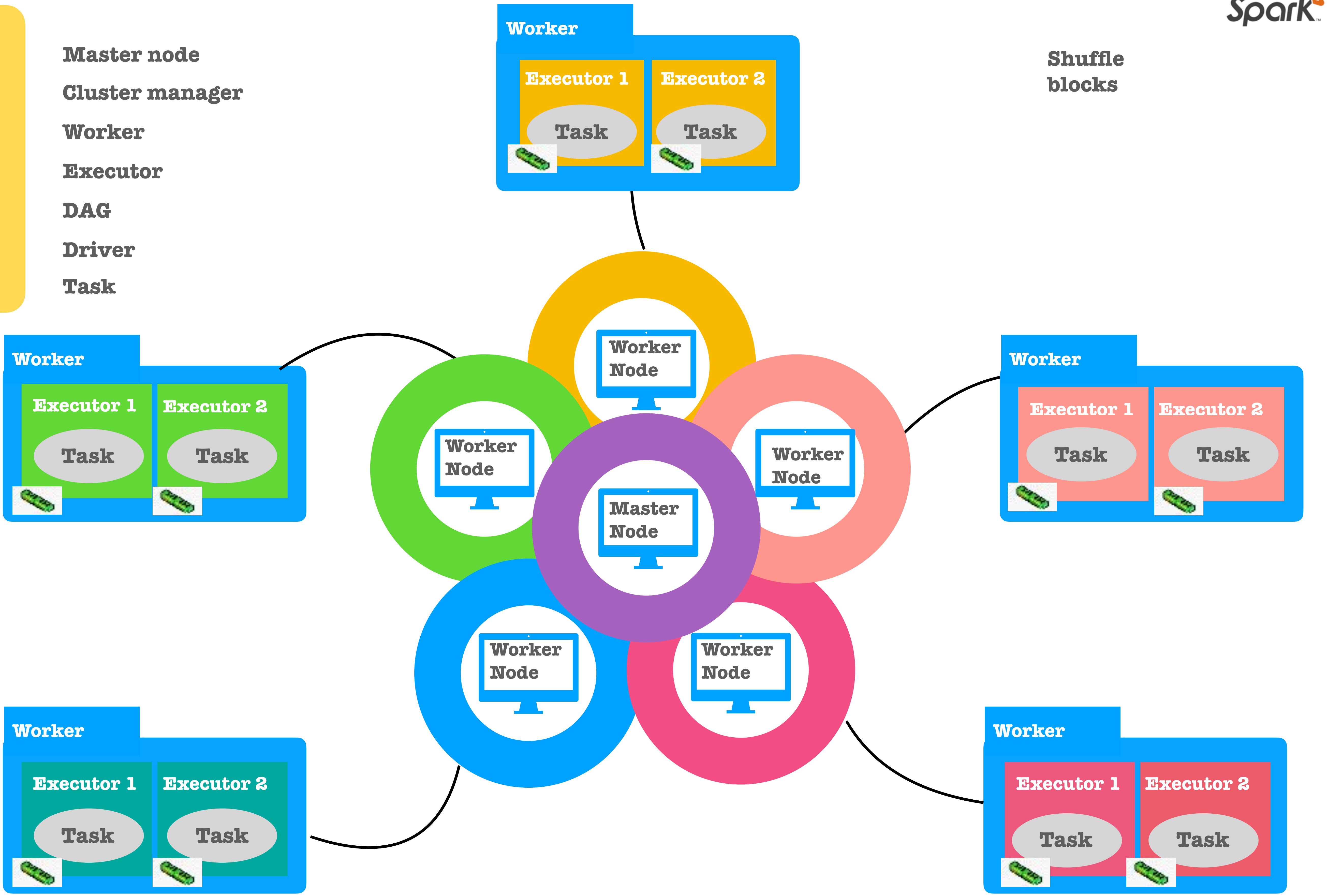
Executor

DAG

Driver

Task

Shuffle blocks



Summary

1. Hadoop
2. Map Reduce life cycle
3. Yarn
4. Why did spark came into picture?
5. What are the components of spark?
6. What are the programming languages supported by spark?
7. What are the operations present in the spark?
8. What is the architecture of spark?
9. What happens when you run your code?
10. What is RDD in spark?