**Apache Spark 2.x – Data processing - Getting Started**

<https://www.edureka.co/blog/spark-architecture/#:~:text=Scala%20and%20Python.-,Spark%20Architecture%20Overview,Resilient%20Distributed%20Dataset%20(RDD)>

**Setup Spark Environment**

Let us review the details to setup of Spark Environment. We have already covered how to setup environment earlier.

* Pre-requisites
  + 64 bit Computer
  + At least 4 GB RAM and enough storage
  + 64-bit Operating System -Windows 10,Linux,Mac etc
  + We would recommend Ubuntu on top of Windows 10.You can either setup using Windows Subsystem for Linux or have virtual machine
* Setup Process
  + Go to [http://spark.apache.org/ 1](http://spark.apache.org/)
  + Download the tarball of your choice
  + Uncompress and untar in your favorite location
  + Make sure to setup environment variable so that you can use commands such as spark-shell,pyspark,spark-submit from anywhere
* Understand the spark layout

## **Spark & its Features**

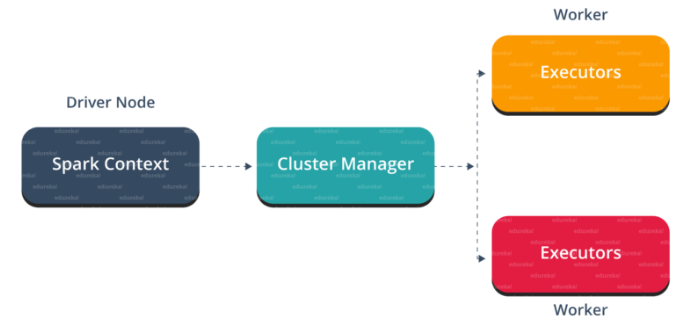
Apache Spark is an open source cluster computing framework for real-time data processing. The main feature of Apache Spark is its ***in-memory cluster computing*** that increases the processing speed of an application. Spark provides an interface for programming entire clusters with implicit **data parallelism and fault tolerance**. It is designed to cover a wide range of workloads such as batch applications, iterative algorithms, interactive queries, and streaming.

1. **Speed**Spark runs up to 100 times faster than Hadoop MapReduce for large-scale data processing. It is also able to achieve this speed through controlled partitioning.
2. **Powerful Caching**Simple programming layer provides powerful caching and disk persistence capabilities.
3. **Deployment**It can be deployed through ***Mesos, Hadoop via YARN, or Spark’s own cluster manager.***
4. **Real-Time**  
   It offers Real-time computation & low latency because of ***in-memory computation.***
5. **Polyglot**  
   Spark provides high-level APIs in Java, Scala, Python, and R. Spark code can be written in any of these four languages. It also provides a shell in Scala and Python.

## **Spark Architecture Overview**

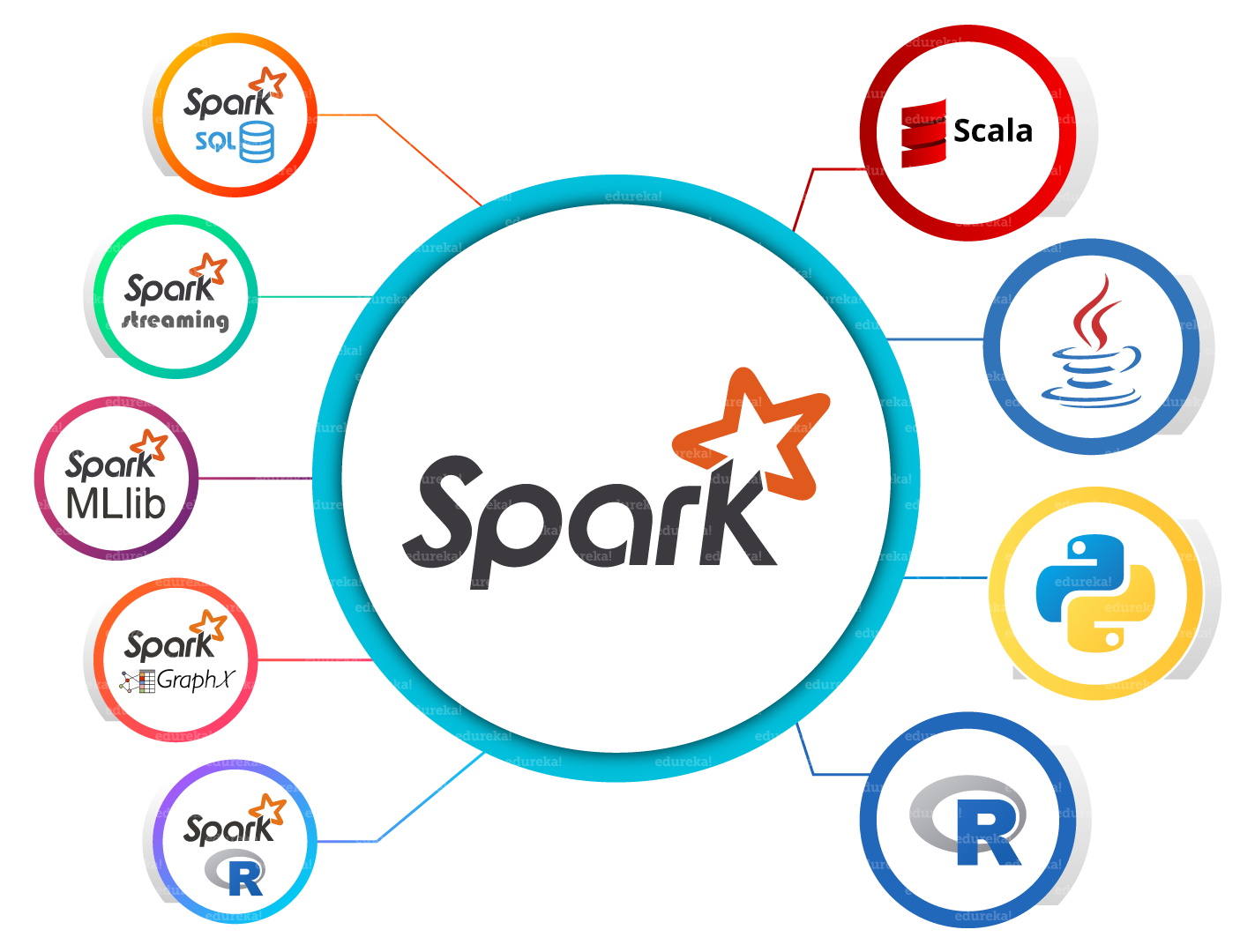
Apache Spark has a well-defined layered architecture where all the spark components and layers are loosely coupled. This architecture is further integrated with various extensions and libraries. Apache Spark Architecture is based on two main abstractions:

* Resilient Distributed Dataset (RDD)
* Directed Acyclic Graph (DAG)



## **Spark Eco-System**

As you can see from the below image, the spark ecosystem is composed of various components like Spark SQL, Spark Streaming, MLlib, GraphX, and the Core API component.



1. **park Core**Spark Core is the base engine for large-scale parallel and distributed data processing. Further, additional libraries which are built on the top of the core allows diverse workloads for streaming, SQL, and machine learning. It is responsible for memory management and fault recovery, scheduling, distributing and monitoring jobs on a cluster & interacting with storage systems.
2. **Spark Streaming**Spark Streaming is the component of Spark which is used to process real-time streaming data. Thus, it is a useful addition to the core Spark API. It enables high-throughput and fault-tolerant stream processing of live data streams.
3. **Spark SQL**Spark SQL is a new module in Spark which integrates relational processing with Spark’s functional programming API. It supports querying data either via SQL or via the Hive Query Language. For those of you familiar with RDBMS, Spark SQL will be an easy transition from your earlier tools where you can extend the boundaries of traditional relational data processing.
4. **GraphX**GraphX is the Spark API for graphs and graph-parallel computation. Thus, it extends the Spark RDD with a Resilient Distributed Property Graph. At a high-level, GraphX extends the Spark RDD abstraction by introducing the Resilient Distributed Property Graph (a directed multigraph with properties attached to each vertex and edge).
5. **MLlib (Machine Learning)**  
   MLlib stands for Machine Learning Library. Spark MLlib is used to perform machine learning in Apache Spark.
6. ***SparkR***It is an R package that provides a distributed data frame implementation. It also supports operations like selection, filtering, aggregation but on large data-sets.

## **Resilient Distributed Dataset(RDD)**

RDDs are the building blocks of any Spark application. RDDs Stands for:

* **Resilient:** Fault tolerant and is capable of rebuilding data on failure
* ***Distributed:*** Distributed data among the multiple nodes in a cluster
* **Dataset:** Collection of partitioned data with values
* It is a layer of abstracted data over the distributed collection. It is immutable in nature and follows [*lazy transformations*](https://www.edureka.co/blog/spark-tutorial/#Spark_Features).
* Now you might be wondering about its working. Well, the data in an RDD is split into chunks based on a key. RDDs are highly resilient, i.e, they are able to recover quickly from any issues as the same data chunks are replicated across multiple executor nodes. Thus, even if one executor node fails, another will still process the data. This allows you to perform your functional calculations against your dataset very quickly by harnessing the power of multiple nodes.
* Moreover, once you create an RDD it becomes ***immutable***. By immutable I mean, an object whose state cannot be modified after it is created, but they can surely be transformed.

Talking about the distributed environment, each dataset in RDD is divided into logical partitions, which may be computed on different nodes of the cluster. Due to this, you can perform transformations or actions on the complete data parallelly. Also, you don’t have to worry about the distribution, because Spark takes care of that.

There are two ways to create RDDs − parallelizing an existing collection in your driver program, or by referencing a dataset in an external storage system, such as a shared file system, HDFS, HBase, etc.

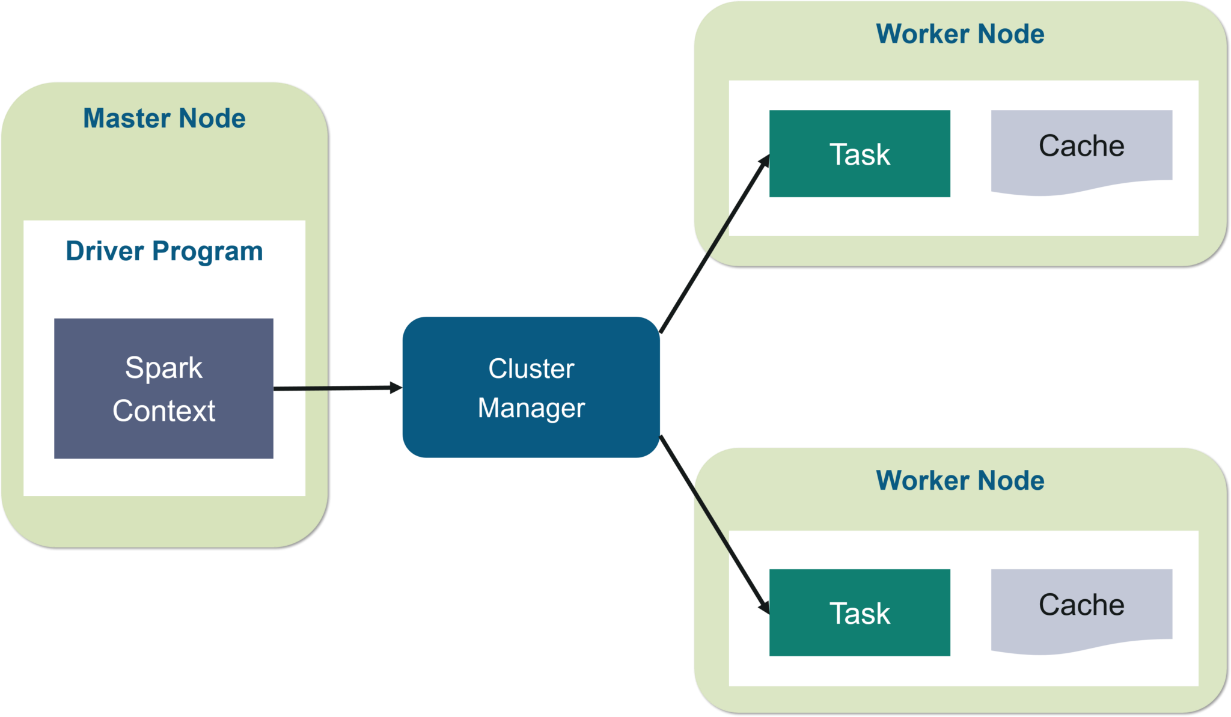
With RDDs, you can perform two types of operations:

1. **Transformations:**They are the operations that are applied to create a new RDD.
2. **Actions:** They are applied on an RDD to instruct Apache Spark to apply computation and pass the result back to the driver.

## **Working of Spark Architecture**

As you have already seen the basic architectural overview of Apache Spark, now let’s dive deeper into its working.

In your **master node**, you have the driver program, which drives your application. The code you are writing behaves as a driver program or if you are using the interactive shell, the shell acts as the driver program.

                                  Fig: Spark Architecture

Inside the driver program, the first thing you do is, you create a ***Spark Context.*** Assume that the Spark context is a gateway to all the Spark functionalities. It is similar to your database connection. Any command you execute in your database goes through the database connection. Likewise, anything you do on Spark goes through Spark context.

Now, this Spark context works with the **cluster manager** to manage various jobs. The driver program & Spark context takes care of the job execution within the cluster. A job is split into multiple tasks which are distributed over the worker node. Anytime an RDD is created in Spark context, it can be distributed across various nodes and can be cached there.

**W*orker nodes*** are the slave nodes whose job is to basically execute the tasks. These tasks are then executed on the partitioned RDDs in the worker node and hence returns back the result to the Spark Context.

Spark Context takes the job, breaks the job in tasks and distribute them to the worker nodes. These tasks work on the partitioned RDD, perform operations, collect the results and return to the main Spark Context.

If you increase the number of workers, then you can divide jobs into more partitions and execute them parallelly over multiple systems. It will be a lot faster.

With the increase in the number of workers, memory size will also increase & you can cache the jobs to execute it faster.

To know about the workflow of Spark Architecture, you can have a look at the **infographic** below

**STEP 1:**The client submits spark user application code. When an application code is submitted, the driver implicitly converts user code that contains transformations and actions into a logically directed acyclic graph called ***DAG.***At this stage, it also performs optimizations such as pipelining transformations.

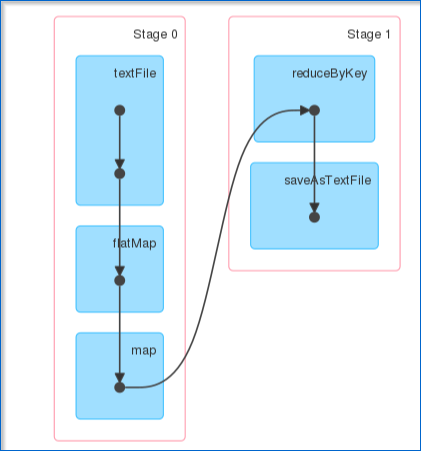
**STEP 2:** After that, it converts the logical graph called DAG into physical execution plan with many stages. After converting into a physical execution plan, it creates physical execution units called tasks under each stage. Then the tasks are bundled and sent to the cluster

**STEP 3:** Now the driver talks to the cluster manager and negotiates the resources. Cluster manager launches executors in worker nodes on behalf of the driver. At this point, the driver will send the tasks to the executors based on data placement. When executors start, they register themselves with drivers. So, the driver will have a complete view of executors that are executing the task.

**STEP 4:** During the course of execution of tasks, driver program will monitor the set of executors that runs. Driver node also schedules future tasks based on data placement.

This was all about Spark Architecture. Now, let’s get a hand’s on the working of a Spark shell.

* On clicking the task that you have submitted, you can view the Directed Acyclic Graph (DAG) of the completed job.

                            Fig: DAG Visualization

* Examples of Transformations
  + Row Level Transformations - map,filter,flatMap etc
  + Aggregations - reduceByKey,aggregrateByKey
  + Joins - join,leftOuterJoin,rightOuterJoin
  + Sorting - sortByKey
  + Ranking - groupByKey followed by flatMao with a lambda function
  + Except for Row Level Transformtions,most of the other transformations have to go through the shuffle phase and trigger new stage.
  + Row Level Transformations also known as Narrow Transformations.
  + Transformations that trigger shuffle and new stage are also called as wide transformations.
  + Example of Actions
    - Preview Data:take,takeSample,top,takeOrdered
    - Convert into Python List:collect
    - Total Aggregration:reduce
    - Writing into Files:saveAsTextFile,saveAsSequenceFile

n the earlier version of Spark,we have core API at the bottom and all the higher level modules work with core API.Examples of core API are a map,reduce,join,groupByKey etc.But with Spark 2,Data frames and Spark SQL has become the core module.

* Core - Transformations and Actions -APIs such as map,reduce,join,filter etc.They typically work on RDD
* Spark SQL and Data Frames -APIs and Spark SQL interface for batch processing on top of Data Frames or Data Sets(not available for Python)
* Structured Streaming - APIs and Spark SQL interface for stream data processing on top of Data Frames
* Machine Learning Pipelines - Machine Learning data pipelines to apply Machine Learning algorithms on top of Data Frames
* GraphX Pipelines
* We can build applications using different programming languages such as Scala,Python,Java,R etc leveraging Spark APIs of the above-mentioned modules.

**Spark Data Structures**

We need to deal with 2 types of data structures in Spark - RDD and Data Frames. We will see both in detail as we proceed further.

* RDD is there for quite some time and it is the low-level data structure which spark uses to distribute the data between tasks while data is being processed.
* RDD can be created using SparkContext APIs such as textFile.
* RDD will be divided into partitions while data being processed. Each partition will be processed by one task.
* The number of RDD partitions is typically based on HDFS block size which is 128MB by default.We can control the number of minimum partitions by using additional arguments while invoking APIs such as textFile.
* Data Frame is nothing but RDD with the structure. We should be able to access the attributes of Data Frame using names.
* Typically we read data from file systems such as HDFS,S3,Azure Blob,Local file systems etc
* Based on the file formats we need to use different APIs available in spark to read data into RDD or Data Frame.
* Spark uses HDFS APIs to read from,and/or write data to the underlying file system.

**Simple Application**

Let us start with a simple application to understand details related to architecture using pyspark.

* As we have multiple versions of Spark on our lab and we are exploring Spark 2 we need to export SPARK\_MAJOR\_VERSION Spark 2
* For this demo, we will disable dynamic allocation by setting spark.dynamicAllocation.enable to false.
* Launch pyspark using YARN and disabling dynamic allocation( also,use spark.ui.port as well to specify unique port).
* Develop a simple word count program by reading data from /public/randomtextwriter/part-m-00000
* Save output to /user/training

# Spark Framework

Let us understand the execution modes as well as different components of the Spark Framework. Also, we will recap some important aspects of YARN.

Execution Modes

Following are the differnt execution modes supported by Spark.

* Local(for development)
* Standalone(for development)
* Mesos
* YARN

As our clusrer uses YARN,let us recap some important aspects of YARN.

* YARN uses Master(Resourse Manager) and Slave(Node Managers) Architecture.
* YARN primarily takes care of resource management and scheduling the tasks.
* FOr earn YARN Application,there will be an application master and set of containers created to process the data.
* We can plugin different distributed frameworks into YARN,such as Map Reduce, Spark etc.
* Spark creates executors to process the data and these executors will be managed by Resource Manager & per job Application Master.

## Execution Framework

Let us understand the Spark exection by running wordcount program using RDD.

* Driver Program
* Spark Context
* Executors
* Executor Cache
* Exector Tasks
* job
* Stage
* Task(Executor Tasks)

|  |
| --- |
| data = sc.textFile('/public/randomtextwriter/part-m-00000') |
|  | wc = data. \ |
|  | flatMap(lambda line: line.split(' ')). \ |
|  | map(lambda word: (word, 1)). \ |
|  | reduceByKey(lambda x, y: x + y) |
|  | wc. \ |
|  | map(lambda rec: rec[0] + ',' + str(rec[1])). \ |
|  | saveAsTextFile('/user/training/core/wordcount') |

|  |
| --- |
| from pyspark.sql.functions import split, explode |
|  | data = spark.read.text('/public/randomtextwriter/part-m-00000') |
|  | wc = data.select(explode(split(data.value, ' ')).alias('words')). \ |
|  | groupBy('words'). \ |
|  | agg(count('words').alias('wc')) |
|  | wc.write.csv('/user/training/df/wordcount') |