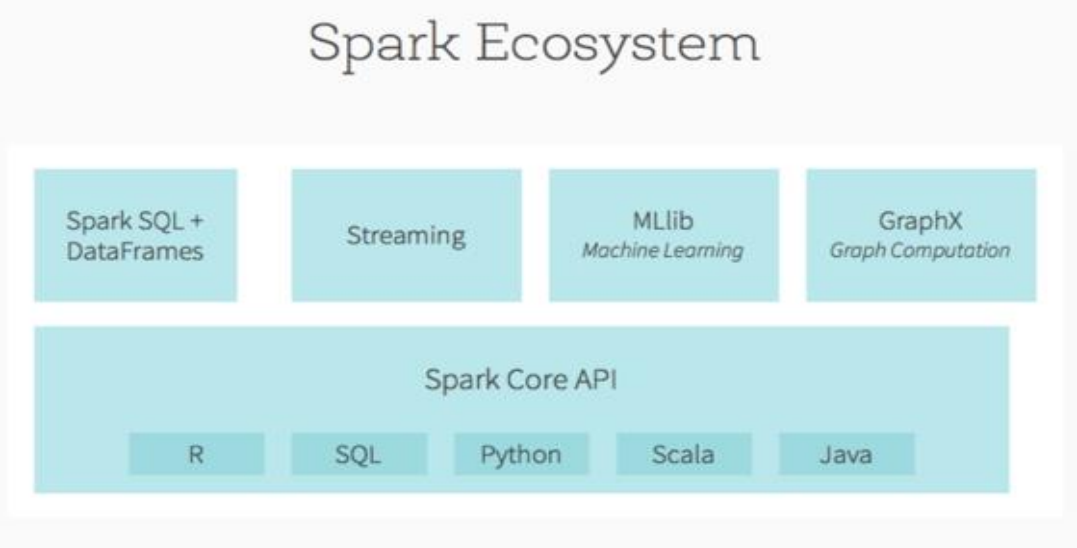
Implementation of Kmean on Spark

Apache spark is a cluster computing platform designed to be fast and general purpose usage. One of the main features Spark offers for speed is the ability to run computations in memory, but the system is also more efficient than MapReduce for complex applications running on disk.

The entire Spark Ecosystem are made up of various components in the heart lies the spark core which is mainly built using Scala but various APIs are provided for development in Python, Java and other languages too. On top of Spark core various modules are built for specific purposes, these are, Spark SQL, Streaming, MLlib, GraphX.



Problem of using Iterative algorithm on Map reduce implementation.

Since MapReduce uses coarse-grained tasks to do its work which are too heavy for iterative algorithms. Another problem with MapReduce is that MapReduce has no awareness of total pipeline of Map Plus Reduce steps, so it can't cache intermediate data to disk between each step. Due to this limitation for each iteration we have do a disk read, which result in slow algorithm convergence.

In this project, we ran Kmean on Spark with spark standalone used as a resource allocator. Our main aim was to take iterative algorithm here in this case Kmean and run it on spark instead on Hadoop framework which in return will give us improved performance.

Input file

I have also submitted in the zip folder a file named input.csv which will be fed as an input data in the program.

Running Command

Spark-submit - -master yarn - -deploy-mode client - -executor-memory 1g -- name Kmean_spark Kmean_spark.py > output.txt

Outfile

The program will generate an output file as outfile.csv which will contain the predicted and the actual class value along with the accuracy

Future Scope and Work

- For future analysis which can be taken forward from this is that we can run spark over some well-known and powerful resource negotiator like Yarn and Mesos.
- We can run this Kmean Spark Application on cluster with three or more node and gauze by what fraction we increased our efficiency.

Reference

https://www.safaribooksonline.com/library/view/learning-spark/9781449359034/ch01.html