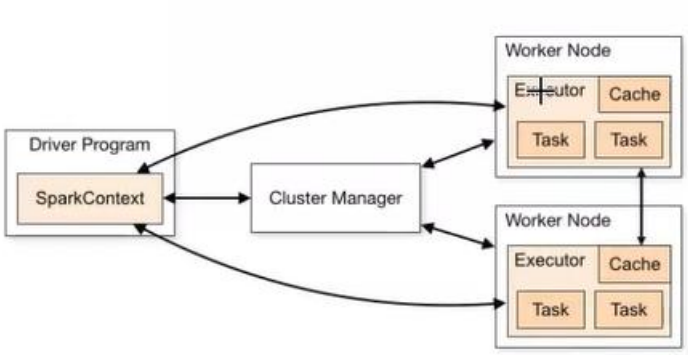
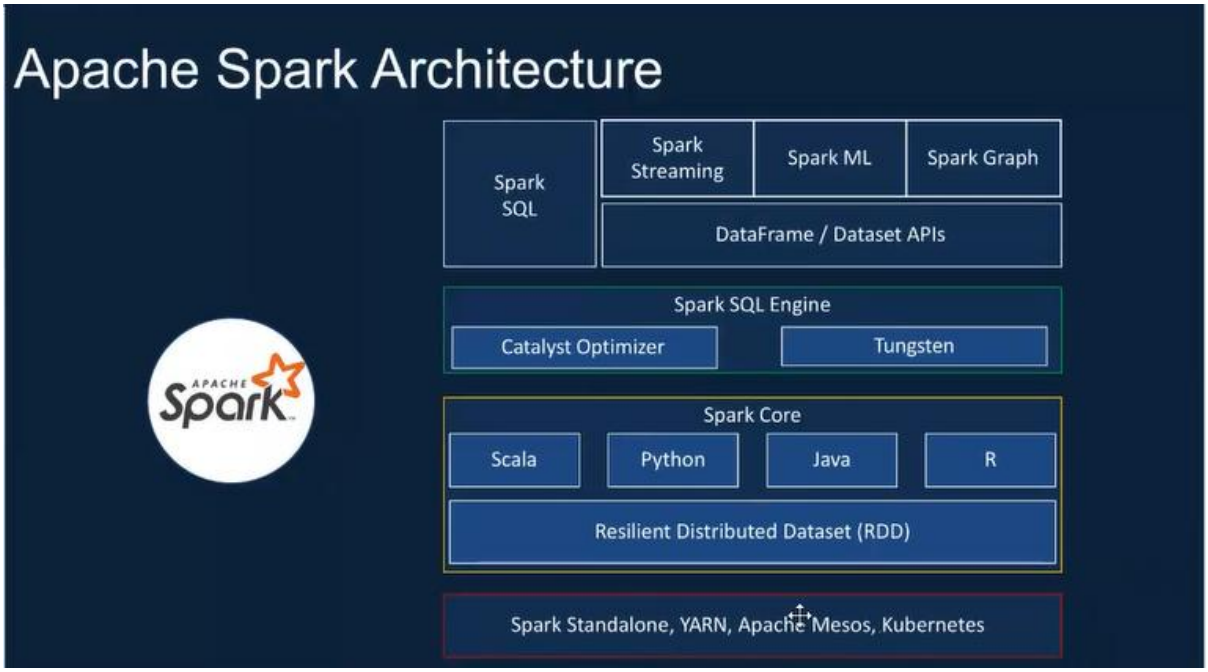




Spark can be bult on HDFS--->StandAlone

Spark can be built on yarn/mesos and HDFS

Spark can be built on MapReduce and HDFS

Spark core is module where entire spark is going to run,It is underlying execution engine for spark platform.

Spark is an analytical engine

sub systems: 1.Spark SQL 2.Spark streaming 3.Spark matlib 4.

Databricks comes with inbuilt spark.It has all inbuilt spark modules.

But in jupyter notebook,it doesn't provide inbuilt spark

Spark cluster architecture

spark context <----------->Cluster manager<--->Worker nodes

(Driver program) [Executor,cache,tasks]

Pyspark Intro

Data frames can be created in pyspark

1.From Existing RDD

2.From external json,excel

1.Create pyspark using existing RDD

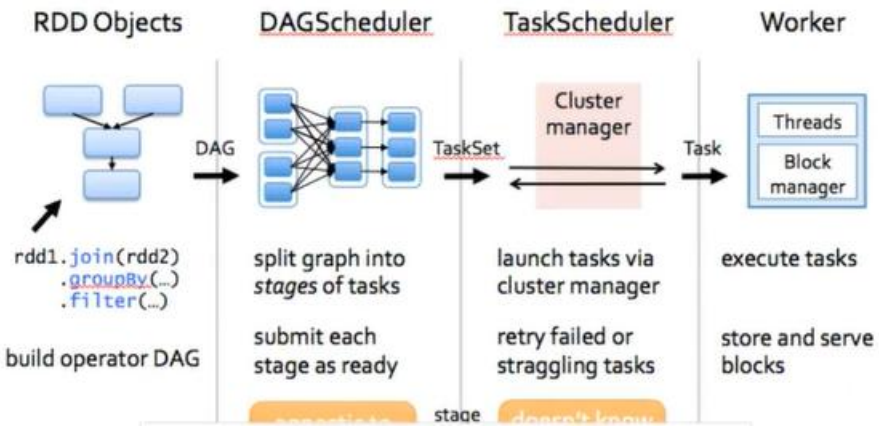
.parallelize() method is used

Session is place where program is gping ot run in spark

Input data→spark rdd→spark dataframe

1. 2.

Jobs run on cluster



###Code

```
#initializing the program
from pyspark import SparkContext
from pyspark.sql import SparkSession
sc=SparkContext.getOrCreate()
spark=SparkSession.builder.appName('Pyspark session').getOrCreate()

#Creating rdd
rdd=sc.parallelize([('C',85,76,87,91),('B',85,76,87,91),("A",85,78,96,92),("A",92,76,89,96)])
print(type(rdd))

#Creating dataframe
sub=['Division','English','Mathematics','Physics','Chemistry']
marks_df=spark.createDataFrame(rdd,schema=sub)
print(type(marks_df))
print(rdd)
marks_df.show()
marks_df.printSchema()



#initializing the program
from pyspark import SparkContext
from pyspark.sql import SparkSession
sc=SparkContext.getOrCreate()
spark=SparkSession.builder.appName('Pyspark session').getOrCreate()
data=spark.read.csv("/FileStore/tables/loandata_1_.csv",header=True,inferSchema=True)
data.show()
display(data)
```

------Lunch break

https://spark.apache.org/examples.html