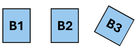
resource manger (master)

Yarn -container (AM)

apllication ID :

Logic File : sample.py



Node manager


```
df1=spark.read("sample.txt", 5)
df2=df1.select

no. of blocks =no. of tasks
```

```
rdd = sc.textFile(" sample.txt")
rdd1=rdd.flatMap( lambda s:s.split(" "))
rdd2=rdd1.map ( lambda z: (z,1))

rdd3=rdd2.reduceByKey( lambda x,y : (x+y))
```

Spark Sql	SQL
Data --->dataframe/table	ACID properties
OLAP (historical)	Create a table --> insert into OLTP (transactional)
schema on write	schema on read
no update or delete, merge	
Join + Union	
databricks (Delta Lakes) DW + DL	

Nalini, 24
Arpita,23

AWS , Azure , GCP
Cloud

Amazon : EMR (Elastic map reduce)
big data platform , Hadoop , Spark , Hive , Flink

Compute : pay as per use, EMR, s3, EC2

add remove the nodes , click

Storage :
HDFS (disk , DN)
EMR : HDFS , S3 Buckets

Hadoop : master worker

HDFS : Master : Namenode

Worker: datanode

YARN: MAster - Resource Manager

Worker : Node manager

Map Reduce:
Master : Job tracker
worker : task tracker

EMR

Master Node (yarn RM ,
Spark Driver , Hadoop
Namenode)

Core Nodes (DN)
(Node manager , DN ,
executors)

Task Node (compute)

Storage

1. HDFS
core nodes (data , disk)

2 . Amazon s3 , bucket
spark , hive, flink
EMR -compute

Cluster

1. EMR Cluster on EC2
specific nodes

2. EMR serverless

uniform :
instance type

xsmall , xlarge, 2xlarge

fixing , price

instance fleets

master , core , task

70 on spot , 30 on demand

BLOB

Binary Large Objects

raw format

Small File Problems

BS =128 MB

Executors --task

spark.read.file("/user/sample")

spark.read.("s3://aws-emr-studio-698614186346-us-west-2/customers.csv

sample.py >> demo.py>>AWS