MAPREDUCE GROOMING SESSION(22FEB17)

What is MapReduce?

What are Mappers ?

Mappers job is to filter and transform the data to the lowest logical level

Mapper input = 1000 records

Mapper output = 600 records

What is sort and Shuffle

Keys are sorted

Values are put together for each key from all the mappers

Values are sorted?

Yes/NO?

how does Reduce function work?

key1, (V1, V2...)

Key2, (V3, V4...)

What parameters are we passing in the reduce function?

default input Format class-textinputformat

input key - LongWritable - offset number of the record

input value - Text - full line or full record

For mapper output:-

job.setMapOutputKeyClass(Text.class);

job.setMapOutputValueClass(IntWritable.class);

job.setInputFormatClass(TextInputFormat.class);

job.setOutputFormatClass(TextOutputFormat.class);

For reducer output:-

job.setOutputKeyClass();

job.setOutputValueClass();

Map tasks = no of input splits

Reduce tasks = default=1

job.setNumReduceTasks(int);

output = 10 records in case 1 reducer

job.setNumReduceTasks(2);

part-r-00000 - 5

part-r-00001 - 5

What are combiners ?

Adv: traffic will be decreased, reduces the reducer tasks,reduces the load on reducers

one key - 1000 values

after combiner

one key - one value

What are partitioners ?

splits keys based on partition condition, and send to diff reducers

Partitioner will divide the data according to reducers

It runs before reducer

What are Distributed Cache?

Caching the memory in the memory of our cluster so that the data is available to all nodes present in that cluster. It is made available by creating hashmap

what are the different types of joins

map-side join and reduce-side join

In mapside join

can join 2 more files.

One Big Table + multiple small tables

Small table size is <=100mb

In a mapside join how many reducers are required?

0 reducers

Mapper output is the final output

job.setNumReduceTasks(0);-only mapper to be executed in our system\

join the data and extract the data from the map function before writing the output to the context class

What is the Reduce Side join

Joining 2 or more large data sets having a common key at reducer level

MultipleInputs.

important config files to configure hadoop

core-site.xml

it contains common properties for hdfs and mapreduce- to specify port numbers etc

Core-default.xml in hadoop 2.6.0-default values are specified/listed(to make changes we do it in core-site.xml)

fs.default.name

absolute path to the hdfs

hdfs://128.25.24.1:9000/niit-niit folder is created on the namenode port id 9000 on which our filesystem is running with ip address 128…

mapred-site.xml

it contains only mapreduce properties

Mapred-default.xml in hadoop 2.6.0

50070 - port no to run the Namenode

8088 - resource manager

19888 - job history server

hdfs-site.xml

for namenode, datanode and secondary namenode properties

hdfs-default.xml

dfs.replication=3,min=1,max=512

dfs.blocksize = 134217728

1024x1024x128 = 134217728

What is YARN?

in hadoop 1.x - mapreduce framework

it is a processing framework in hadoop 2.x and mapreduce is a part of yarn framework.

components of YARN?

Resource Manager - master

Node Manager - slaves

In hadoop 1.x we had job tracker and task tracker.

in 1.x number of jobs were limited

YARN flow

------------

1. user fires a job to RM

2. RM will a create a container on a single DN

3. Launch App Master in that container

4. Create containers on nodes where the blocks are located

5. Node Manager will run the tasks (Map or Reduce) in those containers

-Containers are the space being allocated to perform a job

-For every single query AM is created is not depend on DN

-New AppMaster is created for every single query. AM is destroyed once the work is completed.

-Job handling is done by AM. If any job gets failed, AM goes to the replicas and gets the job done

-AM will have the access to all the nodes