

Final Project Phase-1

Nisarg Patel

1) Creating 4 Node cluster

```
ec2-user@ip-172-31-20-175:~$ jps
3197 DataNode
3426 Jps
3304 NodeManager
ec2-user@ip-172-31-20-175:~$

ec2-user@ip-172-31-24-165:~$ jps
3957 NodeManager
3058 Jps
3859 DataNode
ec2-user@ip-172-31-24-165:~$

ec2-user@ip-172-31-26-151:~$ jps
3050 Jps
2851 DataNode
2949 NodeManager
ec2-user@ip-172-31-26-151:~$

ec2-user@ip-172-31-26-136:~$ cat /logs/yarn-ec2-user-nodemanager-ip-172-31-26-136.out
[ec2-user@ip-172-31-26-136 ~]$ ha
hadoop                  hadoop-daemon.sh      halt
hadoop.cmd              hadoop-daemons.sh    hash
[ec2-user@ip-172-31-26-136 ~]$ hadoop dfsadmin -report
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

Configured Capacity: 33275133952 (30.99 GB)
Present Capacity: 25539850240 (23.79 GB)
DFS Remaining: 25539743744 (23.79 GB)
DFS Used: 106496 (104 KB)
DFS Used%: 0.00%
Under replicated blocks: 0
Blocks with corrupt replicas: 0
Missing blocks: 0

-----
Live datanodes (4):

Name: 172.31.26.151:50010 (ip-172-31-26-151.us-west-1.compute.internal)
Hostname: ip-172-31-26-151.us-west-1.compute.internal
Decommission Status : Normal
Configured Capacity: 8318783488 (7.75 GB)
DFS Used: 24576 (24 KB)
Non DFS Used: 1883418624 (1.75 GB)
DFS Remaining: 6435340288 (5.99 GB)
DFS Used%: 0.00%
DFS Remaining%: 77.36%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 1
Last contact: Thu May 10 10:42:36 UTC 2018

Name: 172.31.24.165:50010 (ip-172-31-24-165.us-west-1.compute.internal)
Hostname: ip-172-31-24-165.us-west-1.compute.internal
Decommission Status : Normal
Configured Capacity: 8318783488 (7.75 GB)
DFS Used: 24576 (24 KB)
Non DFS Used: 1883406336 (1.75 GB)
DFS Remaining: 6435352576 (5.99 GB)
DFS Used%: 0.00%
DFS Remaining%: 77.36%
```

```
ec2-user@ip-172-31-20-175:~$ jps
3197 DataNode
3426 Jps
3304 NodeManager
ec2-user@ip-172-31-20-175:~$

ec2-user@ip-172-31-24-165:~$ jps
3957 NodeManager
3058 Jps
3859 DataNode
ec2-user@ip-172-31-24-165:~$

ec2-user@ip-172-31-26-151:~$ jps
3050 Jps
2851 DataNode
2949 NodeManager
ec2-user@ip-172-31-26-151:~$

ec2-user@ip-172-31-26-136:~$ cat /logs/yarn-ec2-user-nodemanager-ip-172-31-26-136.out
DFS Remaining%: 77.36%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 1
Last contact: Thu May 10 10:42:36 UTC 2018

Name: 172.31.26.136:50010 (ip-172-31-26-136.us-west-1.compute.internal)
Hostname: ip-172-31-26-136.us-west-1.compute.internal
Decommission Status : Normal
Configured Capacity: 8318783488 (7.75 GB)
DFS Used: 28672 (28 KB)
Non DFS Used: 2084888576 (1.94 GB)
DFS Remaining: 6233866240 (5.81 GB)
DFS Used%: 0.00%
DFS Remaining%: 74.94%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 1
Last contact: Thu May 10 10:42:39 UTC 2018

Name: 172.31.20.175:50010 (ip-172-31-20-175.us-west-1.compute.internal)
Hostname: ip-172-31-20-175.us-west-1.compute.internal
Decommission Status : Normal
Configured Capacity: 8318783488 (7.75 GB)
DFS Used: 21672 (20 KB)
Non DFS Used: 1883570176 (1.75 GB)
DFS Remaining: 6435184640 (5.99 GB)
DFS Used%: 0.00%
DFS Remaining%: 77.36%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 1
Last contact: Thu May 10 10:42:39 UTC 2018

ec2-user@ip-172-31-26-136:~$
```

Word Count time taken for 4 node cluster compare to 1 node cluster

```
real    0m36.320s
user    0m3.824s
sys     0m0.195s
[ec2-user@ip-172-31-26-136 ~]$
Bytes Written=20056175

real    1m20.660s
user    0m4.240s
sys     0m0.164s
[ec2-user@ip-172-31-26-246 ~]$
```

2)

Time Taken to run query 1.2, 1.3, 2.1

```
2018-05-12 07:30:04,807 Stage-2 map = 0%, reduce = 0%
2018-05-12 07:30:13,250 Stage-2 map = 25%, reduce = 0%, Cumulative CPU 2.83 sec
2018-05-12 07:30:14,284 Stage-2 map = 50%, reduce = 0%, Cumulative CPU 6.43 sec
2018-05-12 07:30:17,383 Stage-2 map = 75%, reduce = 0%, Cumulative CPU 9.99 sec
2018-05-12 07:30:18,419 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 14.86 sec
2018-05-12 07:30:19,455 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 16.25 sec
MapReduce Total cumulative CPU time: 16 seconds 360 msec
Ended Job = job_1526109783349_0001
MapReduce Jobs Launched:
Stage-Stage-2: Map: 4 Reduce: 1 Cumulative CPU: 16.36 sec
9110 HDFS Write: 12 SUCCESS
Total MapReduce CPU Time Spent: 16 seconds 360 msec
OK
14215822897
Time taken: 34.389 seconds, Fetched: 1 row(s)
hive>

--Q1.2 Simplified to remove expression in sum
select sum(lo_extendedprice) as revenue
from lineorder, dwdate
where lo_orderdate = d_datekey
and d_yearmonth = 'Jan1993'
and lo_discount between 5 and 6
and lo_quantity between 25 and 35;

2018-05-12 07:39:12,041 Stage-2 map = 0%, reduce = 0%
2018-05-12 07:39:18,298 Stage-2 map = 25%, reduce = 0%, Cumulative CPU 2.83 sec
2018-05-12 07:39:19,332 Stage-2 map = 50%, reduce = 0%, Cumulative CPU 6.43 sec
2018-05-12 07:39:22,451 Stage-2 map = 75%, reduce = 0%, Cumulative CPU 9.99 sec
2018-05-12 07:39:23,517 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 14.86 sec
2018-05-12 07:39:25,594 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 16.25 sec
MapReduce Total cumulative CPU time: 16 seconds 250 msec
Ended Job = job_1526109783349_0002
MapReduce Jobs Launched:
Stage-Stage-2: Map: 4 Reduce: 1 Cumulative CPU: 16.25 sec
594378990 HDFS Write: 11 SUCCESS
Total MapReduce CPU Time Spent: 16 seconds 250 msec
OK
4435791464
Time taken: 27.44 seconds, Fetched: 1 row(s)
hive>

--Q1.3 Simplified to remove expression in sum
select sum(lo_extendedprice) as revenue
from lineorder, dwdate
where lo_orderdate = d_datekey
and d_weeknuminyear = 6 and d_year = 1994
and lo_discount between 5 and 8
and lo_quantity between 36 and 41;

361416497      1998      MFGR#128
318769573      1998      MFGR#129
Time taken: 108.893 seconds, Fetched: 280 row(s)
hive>

--Q2.1 No simplifications
select sum(lo_revenue), d_year, p_brand1
from lineorder, dwdate, part, supplier
where lo_orderdate = d_datekey
and lo_partkey = p_partkey
and lo_suppkey = s_suppkey
and p_category = 'MFGR#12'
and s_region = 'AMERICA'
group by d_year, p_brand1
order by d_year, p_brand1;
```

Adding the file :-

ADD FILE /home/ec2-user/10char.py

Now we will do select transform :-

```

INSERT OVERWRITE TABLE customer3 SELECT
TRANSFORM(c_custkey,c_name,c_address,c_city,c_nation,c_region,c_phone,c_mktsegment)
USING 'python 10char.py' AS
(c_custkey,c_name,c_address,c_city,c_nation,c_region,c_phone,c_mktsegment) FROM customer;

```

10Char.py python file :-

```

GNU nano 2.5.3                               File: 10char.py
import sys
import re

for line in sys.stdin:
    line = line.strip().split('\t')
    c_custkey = line[0]
    c_name = line[1]
    c_address = line[2]
    if len(c_address) > 10:
        c_address = c_address[:10]
        line[2] = c_address
    c_city = line[3]
    if c_city[-2] != ' ':
        c_city = c_city[:-1] + ' ' + c_city[-1]
        line[3] = c_city
    c_nation = line[4]
    c_region = line[5]
    c_phone = line[6]
    c_mktsegment = line[7]

    print '\t'.join([c_custkey,c_name,c_address,c_city,c_nation,c_region,c_phone,c_mktsegment])

```

Describe Table:-

```

hive> describe customer;
OK
c_custkey          int
c_name             varchar(25)
c_address          varchar(25)
c_city             varchar(10)
c_nation           varchar(15)
c_region           varchar(12)
c_phone            varchar(15)
c_mktsegment       varchar(10)
Time taken: 0.041 seconds, Fetched: 8 row(s)
hive> describe customer3;
OK
c_custkey          int
c_name             varchar(25)
c_address          varchar(25)
c_city             varchar(12)
c_nation           varchar(15)
c_region           varchar(12)
c_phone            varchar(15)
c_mktsegment       varchar(10)
Time taken: 0.038 seconds, Fetched: 8 row(s)
hive>

```

Output:

```

ec2-user@ip-172-31-26-136:~/apache-hive-2.0.1-bin
hive> select c_address,c_city from customer3 where c_custkey < 20;
OK
j5JsirBM9P      MOROCCO  0
487LWldovn     JORDAN   1
fkRGN8n ARGENTINA 7
4u58h f EGYPT   4
hwBtxkoBF      CANADA   5
glS,pzDen      SAUDI ARA 2
8OkMVLQldK     CHINA    0
j,pZ,Qp,qt     PERU     6
vgIql8H6zo     INDIA    6
Vf mQ6Ug9U     ETHIOPIA 9
cG48rYjF3A     UNITED KI 3
Sb4gxKs7       JORDAN   5
Ez3ax0D5Hn     CANADA   8
h3GFMzeFf      ARGENTINA 0
3y4KK4Ccfn     UNITED KI 0
P2IQMff18e     IRAN     5
Js JrVHNAY     BRAZIL   6
YyukcsqIxl     FRANCE   0
y00XPkiuSW     CHINA    3
Time taken: 0.083 seconds, Fetched: 19 row(s)

```

3)

```

lorder = LOAD '/data/lineorder.tbl' USING PigStorage('|') AS
(lo_orderkey:int,lo_linenummer:int,lo_custkey:int,lo_partkey:int,lo_suppkey:int,lo_orderdate:int,lo_
orderpriority:chararray,lo_shippriority:chararray,lo_quantity:int,lo_extendedprice:int,lo_ordertotalp
rice:int,lo_discount:int,lo_revenue:int,lo_supplycost:int,lo_tax:int,lo_commitdate:int,lo_shipdate:ch
ararray);

```

QUERY 0.1

```

grouplorder = group lorder all;
avgrevenue = FOREACH grouplorder GENERATE AVG(lorder.lo_revenue);

```

```

2018-05-13 00:10:53,206 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2018-05-13 00:10:53,207 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not ge
2018-05-13 00:10:53,215 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-05-13 00:10:53,215 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to proc
(3634300.709514323)
grunt> █

```

Average lo_revenue = 3634300.7095

Time to run query =

```

ecutionEngine - Connecting to hadoop file system at: hdfs://172.31.26.136/
2018-05-14 03:48:31,798 [main] INFO org.apache.hadoop.conf.Configuration.deprecation
- fs.default.name is deprecated. Instead, use fs.defaultFS
2018-05-14 03:48:31,915 [main] INFO org.apache.pig.Main - Pig script completed in 1 s
econd and 998 milliseconds (1998 ms)
[ec2-user@ip-172-31-26-136 pig-0.15.0]$ █

```

```
groupdiscount = group lorder BY lo_discount;
outputdiscount = FOREACH groupdiscount GENERATE lorder.lo_discount, COUNT(
lorder.lo_extendedprice );
```

```

10), (10), (10), (10), (10), (10), (10), (10), (10), (
), (10), (10), (10), (10), (10), (10), (10), (10), (10
(10), (10), (10), (10), (10), (10), (10), (10), (10),
0), (10), (10), (10), (10), (10), (10), (10), (10), (1
, (10), (10), (10), (10), (10), (10), (10), (10), (10)
}, 545815)

```

```
ExecutionEngine - Connecting to hadoop file system at: hdfs://172.31.26.136/
2018-05-14 03:56:23,282 [main] INFO org.apache.hadoop.conf.Configuration.deprecation
fs.default.name is deprecated. Instead, use fs.defaultFS
2018-05-14 03:56:23,397 [main] INFO org.apache.pig.Main - Pig script completed in 1 s
second and 889 milliseconds (1889 ms)
```

```
grunt> filterdiscount = FILTER lorder BY lo_discount < 3;
grunt> groupquantity = group filterdiscount BY lo_quantity;
grunt> outputquantity = FOREACH groupquantity GENERATE(filterdiscount.lo_quantity), SUM(filterdiscount.lo_revenue);
grunt>
```

[illegible]

```
- fs.default.name is deprecated. Instead, use fs.defaultFS
2018-05-14 03:53:43,674 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://172.31.26.136/
2018-05-14 03:53:44,583 [main] INFO org.apache.hadoop.conf.Configuration.deprecation
- fs.default.name is deprecated. Instead, use fs.defaultFS
2018-05-14 03:53:44,708 [main] INFO org.apache.pig.Main - Pig script completed in 1 second and 887 milliseconds (1887 ms)
```

4)

Query 0.2

Mapper

```
ec2-user@ip-172-31-26-136:~  
GNU nano 2.5.3 File: countmapper.py  
  
import sys  
import fileinput  
  
for line in sys.stdin:  
    data = line.strip().split('|')  
    lo_discount = data[11]  
    lo_extendedprice = data[9]  
    lo_discount = int(lo_discount)  
    print "{0}\t{1}".format(lo_discount,1)
```

Reducer

```
ec2-user@ip-172-31-26-136:~  
GNU nano 2.5.3 File: countreducer.py  
  
import sys  
import fileinput  
  
curr_count = 0  
currentkey = None  
  
for line in sys.stdin:  
    line = line.strip().split('\t')  
    key, count = line  
    count=int(count)  
  
    if currentkey == line:  
        curr_count += count  
    else:  
        if currentkey:  
            print "{0}\t{1}".format(currentkey,curr_count)  
            currentkey = key  
            curr_count = count  
  
if currentkey == key:  
    print "{0}\t{1}".format(currentkey,curr_count)
```

Query 0.3

Mapper

```
ec2-user@ip-172-31-26-136:~  
GNU nano 2.5.3 File: summapper.py  
  
import sys  
import fileinput  
  
for line in sys.stdin:  
    data = line.strip().split('|')  
    lo_quantity = data[8]  
    lo_extendedprice = data[9]  
    lo_discount = data[11]  
    lo_revenue = data[12]  
    lo_discount = int(lo_discount)  
  
    if lo_discount < 3:  
        print "{0}\t{1}".format(lo_quantity,lo_revenue)
```

Reducer

```
import sys

total = 0
lastkey = None

for line in sys.stdin:
    line = line.strip().split("\t")
    key = line[0]
    value = line[1]
    lastkey = key

    if lastkey == None or lastkey == key:
        total = total+value
    else:
        print "{0}\t{1}".format(lastkey,total)
        total = value

if key == lastkey:
    print "{0}\t{1}".format(lastkey,total)
```