Phase 1
Zehong Zhuang
CSC 555

PART 1

```
[ec2-user@ip-172-31-3-189 ~]$ hadoop dfsadmin -report
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

Configured Capacity: 128798605312 (119.95 GB)
Present Capacity: 119618019328 (111.40 GB)
DFS Remaining: 119618002944 (111.40 GB)
DFS Used: 16384 (16 KB)
DFS Used%: 0.00%
Under replicated blocks: 0
Blocks with corrupt replicas: 0
Missing blocks: 0

-------------------------------------------------
Live datanodes (4):

Name: 172.31.3.189:50010 (ip-172-31-3-189.us-east-2.compute.internal)
Hostname: ip-172-31-3-189.us-east-2.compute.internal
Decommission Status : Normal
Configured Capacity: 32199651328 (29.99 GB)
DFS Used: 4096 (4 KB)
Non DFS Used: 2444898304 (2.28 GB)
DFS Remaining: 29754748928 (27.71 GB)
DFS Used%: 0.00%
DFS Remaining%: 92.41%
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: Mon Feb 18 16:25:38 UTC 2019


Name: 172.31.11.197:50010 (ip-172-31-11-197.us-east-2.compute.internal)
Hostname: ip-172-31-11-197.us-east-2.compute.internal
Decommission Status : Normal
Configured Capacity: 32199651328 (29.99 GB)
DFS Used: 4096 (4 KB)
Non DFS Used: 2244780032 (2.09 GB)
DFS Remaining: 29954867200 (27.90 GB)
DFS Used%: 0.00%
DFS Remaining%: 93.03%
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: Mon Feb 18 16:25:38 UTC 2019
```

```
Name: 172.31.3.4:50010 (ip-172-31-3-4.us-east-2.compute.internal)
Hostname: ip-172-31-3-4.us-east-2.compute.internal
Decommission Status : Normal
Configured Capacity: 32199651328 (29.99 GB)
DFS Used: 4096 (4 KB)
Non DFS Used: 2244075520 (2.09 GB)
DFS Remaining: 29955571712 (27.90 GB)
DFS Used%: 0.00%
DFS Remaining%: 93.03%
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: Mon Feb 18 16:25:38 UTC 2019


Name: 172.31.5.39:50010 (ip-172-31-5-39.us-east-2.compute.internal)
Hostname: ip-172-31-5-39.us-east-2.compute.internal
Decommission Status : Normal
Configured Capacity: 32199651328 (29.99 GB)
DFS Used: 4096 (4 KB)
Non DFS Used: 2246832128 (2.09 GB)
DFS Remaining: 29952815104 (27.90 GB)
DFS Used%: 0.00%
DFS Remaining%: 93.02%
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: Mon Feb 18 16:25:38 UTC 2019
```

```
[[ec2-user@ip-172-31-3-189 ~]$ time hadoop jar hadoop-2.6.4/share/hadoop/mapreduce/hadoop-mapreduce-e
xamples-2.6.4.jar  wordcount /data/bioproject.xml /data/wordcount1
19/02/18 16:30:41 INFO client.RMProxy: Connecting to ResourceManager at /172.31.3.189:8032
19/02/18 16:30:42 INFO input.FileInputFormat: Total input paths to process : 1
19/02/18 16:30:42 INFO mapreduce.JobSubmitter: number of splits:2
19/02/18 16:30:42 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1550507000577_0001
19/02/18 16:30:43 INFO impl.YarnClientImpl: Submitted application application_1550507000577_0001
19/02/18 16:30:43 INFO mapreduce.Job: The url to track the job: http://ip-172-31-3-189.us-east-2.com
pute.internal:8088/proxy/application_1550507000577_0001/
19/02/18 16:30:43 INFO mapreduce.Job: Running job: job_1550507000577_0001
19/02/18 16:30:49 INFO mapreduce.Job: Job job_1550507000577_0001 running in uber mode : false
19/02/18 16:30:49 INFO mapreduce.Job:  map 0% reduce 0%
19/02/18 16:31:01 INFO mapreduce.Job:  map 26% reduce 0%
19/02/18 16:31:04 INFO mapreduce.Job:  map 45% reduce 0%
19/02/18 16:31:07 INFO mapreduce.Job:  map 49% reduce 0%
19/02/18 16:31:10 INFO mapreduce.Job:  map 60% reduce 0%
19/02/18 16:31:11 INFO mapreduce.Job:  map 77% reduce 0%
19/02/18 16:31:13 INFO mapreduce.Job:  map 83% reduce 0%
19/02/18 16:31:16 INFO mapreduce.Job:  map 89% reduce 0%
19/02/18 16:31:17 INFO mapreduce.Job:  map 100% reduce 0%
19/02/18 16:31:20 INFO mapreduce.Job:  map 100% reduce 100%
19/02/18 16:31:20 INFO mapreduce.Job: Job job_1550507000577_0001 completed successfully
19/02/18 16:31:20 INFO mapreduce.Job: Counters: 50
        File System Counters
                FILE: Number of bytes read=59605201
                FILE: Number of bytes written=86827979
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=231153307
                HDFS: Number of bytes written=20056175
                HDFS: Number of read operations=9
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=1
                Rack-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=45794
                Total time spent by all reduces in occupied slots (ms)=6086
                Total time spent by all map tasks (ms)=45794
                Total time spent by all reduce tasks (ms)=6086
                Total vcore-milliseconds taken by all map tasks=45794
                Total vcore-milliseconds taken by all reduce tasks=6086
                Total megabyte-milliseconds taken by all map tasks=46893056
                Total megabyte-milliseconds taken by all reduce tasks=6232064
        Map-Reduce Framework
                Map input records=5284546
                Map output records=18562366
                Map output bytes=279356680
                Map output materialized bytes=26902454
                Input split bytes=208
                Combine input records=20053191
                Combine output records=2673165
```

```
                FILE: Number of write operations=0
                HDFS: Number of bytes read=231153307
                HDFS: Number of bytes written=20056175
                HDFS: Number of read operations=9
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=1
                Rack-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=45794
                Total time spent by all reduces in occupied slots (ms)=6086
                Total time spent by all map tasks (ms)=45794
                Total time spent by all reduce tasks (ms)=6086
                Total vcore-milliseconds taken by all map tasks=45794
                Total vcore-milliseconds taken by all reduce tasks=6086
                Total megabyte-milliseconds taken by all map tasks=46893056
                Total megabyte-milliseconds taken by all reduce tasks=6232064
        Map-Reduce Framework
                Map input records=5284546
                Map output records=18562366
                Map output bytes=279356680
                Map output materialized bytes=26902454
                Input split bytes=208
                Combine input records=20053191
                Combine output records=2673165
                Reduce input groups=1040390
                Reduce shuffle bytes=26902454
                Reduce input records=1182340
                Reduce output records=1040390
                Spilled Records=3855505
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=448
                CPU time spent (ms)=40590
                Physical memory (bytes) snapshot=725405696
                Virtual memory (bytes) snapshot=6459305984
                Total committed heap usage (bytes)=516947968
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=231153099
        File Output Format Counters
                Bytes Written=20056175

real    0m40.151s
user    0m3.609s
sys     0m0.271s
```

The performance is faster than before, and it should be 4 times faster since there are 4 nodes (1 master + 3 workers) are working at the problem.

PART 2

```
different execution engine (i.e. spark, tez) or using Hive 1.X releases.
hive> CREATE TABLE dwdate (d_datekey int,d_date varchar(19), d_dayofweek varchar(10), d_month varchar
(10), d_year int, d_yearmonthnum int, d_yearmonth int, d_daynuminweek int, d_daynuminmonth int, d_day
numinyear int, d_monthnuminyear int, d_weeknuminyear int, d_sellingseason varchar(13), d_lastdayinwee
kf1 varchar(1), d_lastdayinmonthf1 varchar(1), d_holidayf1 varchar(1), d_weekdayf1 varchar(1)) ROW FO
RMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE;
hive> LOAD DATA LOCAL INPATH '/home/ec2-user/dwdate.tbl' OVERWRITE INTO TABLE dwdate;
Loading data to table default.dwdate
OK
Time taken: 0.683 seconds
hive> SELECT COUNT(*) FROM dwdate;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider
 using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = ec2-user_20190219080939_7eb0596d-d79c-4839-85e8-2700dfc158c9
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1550558351319_0007, Tracking URL = http://ip-172-31-3-189.us-east-2.compute.intern
al:8088/proxy/application_1550558351319_0007/
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job  -kill job_1550558351319_0007
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2019-02-19 08:09:46,064 Stage-1 map = 0%,  reduce = 0%
2019-02-19 08:09:50,255 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.07 sec
2019-02-19 08:09:55,517 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 2.43 sec
MapReduce Total cumulative CPU time: 2 seconds 430 msec
Ended Job = job_1550558351319_0007
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 2.43 sec   HDFS Read: 241432 HDFS Write: 5 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 430 msec
OK
2556
Time taken: 17.221 seconds, Fetched: 1 row(s)
hive> CREATE TABLE Lineorder (lo_orderkey int, lo_linenumber int, lo_custkey int, lo_partkey int, lo_
suppkey int, lo_orderdate int, lo_orderpriority varchar(15), lo_shippriority varchar(1), lo_quantity
int, lo_extendedprice int, lo_ordertotalprice int, lo_discount int, lo_revenue int, lo_supplycost int
, lo_tax int, lo_commitdate int, lo_shipmode varchar(10)) ROW FORMAT DELIMITED FIELDS TERMINATED BY '
,' STORED AS TEXTFILE;
OK
Time taken: 0.443 seconds
hive> LOAD DATA LOCAL INPATH '/home/ec2-user/lineorder.tbl' OVERWRITE INTO TABLE Lineorder;
Loading data to table default.lineorder
OK
Time taken: 9.387 seconds
hive> SELECT COUNT(*) FROM Lineorder;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider
 using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = ec2-user_20190219082644_11ea7fc0-d422-43fc-a952-c4dbeffa3cfc
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1550558351319_0008, Tracking URL = http://ip-172-31-3-189.us-east-2.compute.intern
al:8088/proxy/application_1550558351319_0008/
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job  -kill job_1550558351319_0008
Hadoop job information for Stage-1: number of mappers: 3; number of reducers: 1
2019-02-19 08:26:50,653 Stage-1 map = 0%,  reduce = 0%
2019-02-19 08:26:55,986 Stage-1 map = 67%,  reduce = 0%, Cumulative CPU 2.1 sec
2019-02-19 08:26:57,024 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 7.89 sec
```

```
hive> CREATE TABLE Part (p_partkey int, p_name varchar(22), p_mfgr varchar(6), p_category varchar(7),
  p_brand1 varchar(9), p_color varchar(11), p_type varchar(25), p_size int, p_container varchar(10)) R
OW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE;
OK
Time taken: 0.621 seconds
hive> LOAD DATA LOCAL INPATH '/home/ec2-user/part.tbl' OVERWRITE INTO TABLE Part;
Loading data to table default.part
OK
Time taken: 0.964 seconds
hive> CREATE TABLE Supplier (s_suppkey int, s_name varchar(25), s_address varchar(25), s_city varchar
(10), s_nation varchar(15), s_region varchar(12), s_phone varchar(15)) ROW FORMAT DELIMITED FIELDS TE
RMINATED BY ',' STORED AS TEXTFILE;
OK
Time taken: 0.052 seconds
hive> LOAD DATA LOCAL INPATH '/home/ec2-user/supplier.tbl' OVERWRITE INTO TABLE Supplier;
Loading data to table default.supplier
OK
Time taken: 0.162 seconds
```

1.2
22.798s

```
hive> 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;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider
 using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = ec2-user_20190219082922_5d517a12-0ec1-49b1-ae87-8013c2708e5a
Total jobs = 1
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/ec2-user/apache-hive-2.0.1-bin/lib/log4j-slf4j-impl-2.4.1.jar
!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/ec2-user/hadoop-2.6.4/share/hadoop/common/lib/slf4j-log4j12-1
.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
Execution log at: /tmp/ec2-user/ec2-user_20190219082922_5d517a12-0ec1-49b1-ae87-8013c2708e5a.log
2019-02-19 08:29:26     Starting to launch local task to process map join;      maximum memory = 4776
26368
2019-02-19 08:29:28     Dump the side-table for tag: 1 with group count: 0 into file: file:/tmp/ec2-u
ser/e47734ff-6b56-4a65-8084-e15900c32b17/hive_2019-02-19_08-29-22_552_3803481498719796837-1/-local-10
005/HashTable-Stage-2/MapJoin-mapfile01--.hashtable
2019-02-19 08:29:28     Uploaded 1 File to: file:/tmp/ec2-user/e47734ff-6b56-4a65-8084-e15900c32b17/h
ive_2019-02-19_08-29-22_552_3803481498719796837-1/-local-10005/HashTable-Stage-2/MapJoin-mapfile01--.
hashtable (260 bytes)
2019-02-19 08:29:28     End of local task; Time Taken: 1.183 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1550558351319_0009, Tracking URL = http://ip-172-31-3-189.us-east-2.compute.intern
al:8088/proxy/application_1550558351319_0009/
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job  -kill job_1550558351319_0009
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2019-02-19 08:29:33,919 Stage-2 map = 0%,  reduce = 0%
2019-02-19 08:29:39,114 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 1.16 sec
2019-02-19 08:29:44,285 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 2.51 sec
MapReduce Total cumulative CPU time: 2 seconds 510 msec
Ended Job = job_1550558351319_0009
MapReduce Jobs Launched:
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 2.51 sec   HDFS Read: 16617 HDFS Write: 3 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 510 msec
OK
NULL
Time taken: 22.798 seconds, Fetched: 1 row(s)
```

1.3
22.814s

```
hive> 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;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider
 using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = ec2-user_20190219083105_cb194b0c-db3c-473c-aa0a-3ebd7fe24a30
Total jobs = 1
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/ec2-user/apache-hive-2.0.1-bin/lib/log4j-slf4j-impl-2.4.1.jar
!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/ec2-user/hadoop-2.6.4/share/hadoop/common/lib/slf4j-log4j12-1
.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
Execution log at: /tmp/ec2-user/ec2-user_20190219083105_cb194b0c-db3c-473c-aa0a-3ebd7fe24a30.log
2019-02-19 08:31:09     Starting to launch local task to process map join;      maximum memory = 4776
26368
2019-02-19 08:31:10     Dump the side-table for tag: 1 with group count: 0 into file: file:/tmp/ec2-u
ser/e47734ff-6b56-4a65-8084-e15900c32b17/hive_2019-02-19_08-31-05_475_7774735655416569277-1/-local-10
005/HashTable-Stage-2/MapJoin-mapfile11--.hashtable
2019-02-19 08:31:10     Uploaded 1 File to: file:/tmp/ec2-user/e47734ff-6b56-4a65-8084-e15900c32b17/h
ive_2019-02-19_08-31-05_475_7774735655416569277-1/-local-10005/HashTable-Stage-2/MapJoin-mapfile11--.
hashtable (260 bytes)
2019-02-19 08:31:10     End of local task; Time Taken: 1.218 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1550558351319_0010, Tracking URL = http://ip-172-31-3-189.us-east-2.compute.intern
al:8088/proxy/application_1550558351319_0010/
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job  -kill job_1550558351319_0010
Hadoop job information for Stage-2: number of mappers: 3; number of reducers: 1
2019-02-19 08:31:16,569 Stage-2 map = 0%,  reduce = 0%
2019-02-19 08:31:21,863 Stage-2 map = 33%,  reduce = 0%, Cumulative CPU 2.6 sec
2019-02-19 08:31:24,154 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 11.79 sec
2019-02-19 08:31:27,234 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 13.33 sec
MapReduce Total cumulative CPU time: 13 seconds 330 msec
Ended Job = job_1550558351319_0010
MapReduce Jobs Launched:
Stage-Stage-2: Map: 3  Reduce: 1   Cumulative CPU: 13.33 sec   HDFS Read: 594368473 HDFS Write: 3 SUC
CESS
Total MapReduce CPU Time Spent: 13 seconds 330 msec
OK
NULL
Time taken: 22.814 seconds, Fetched: 1 row(s)
```

2.1
80.669s

```
hive> 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;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider
 using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = ec2-user_20190219084310_6d04a9ed-4e42-4696-b0be-4eb1f7978cd9
Total jobs = 6
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/ec2-user/apache-hive-2.0.1-bin/lib/log4j-slf4j-impl-2.4.1.jar
!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/ec2-user/hadoop-2.6.4/share/hadoop/common/lib/slf4j-log4j12-1
.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
Execution log at: /tmp/ec2-user/ec2-user_20190219084310_6d04a9ed-4e42-4696-b0be-4eb1f7978cd9.log
2019-02-19 08:43:16     Starting to launch local task to process map join;      maximum memory = 4776
26368
2019-02-19 08:43:17     Dump the side-table for tag: 1 with group count: 0 into file: file:/tmp/ec2-u
ser/e22048d6-e387-4984-bd5c-ff23bfb5b1b1/hive_2019-02-19_08-43-10_433_8712731078040772434-1/-local-10
014/HashTable-Stage-13/MapJoin-mapfile31--.hashtable
2019-02-19 08:43:17     Uploaded 1 File to: file:/tmp/ec2-user/e22048d6-e387-4984-bd5c-ff23bfb5b1b1/h
ive_2019-02-19_08-43-10_433_8712731078040772434-1/-local-10014/HashTable-Stage-13/MapJoin-mapfile31--
.hashtable (260 bytes)
2019-02-19 08:43:17     End of local task; Time Taken: 1.238 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 1 out of 6
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1550558351319_0011, Tracking URL = http://ip-172-31-3-189.us-east-2.compute.intern
al:8088/proxy/application_1550558351319_0011/
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job  -kill job_1550558351319_0011
Hadoop job information for Stage-13: number of mappers: 3; number of reducers: 0
2019-02-19 08:43:24,083 Stage-13 map = 0%,  reduce = 0%
2019-02-19 08:43:29,287 Stage-13 map = 33%,  reduce = 0%, Cumulative CPU 2.59 sec
2019-02-19 08:43:34,428 Stage-13 map = 100%,  reduce = 0%, Cumulative CPU 11.8 sec
MapReduce Total cumulative CPU time: 11 seconds 800 msec
Ended Job = job_1550558351319_0011
Stage-15 is filtered out by condition resolver.
Stage-16 is selected by condition resolver.
Stage-2 is filtered out by condition resolver.
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/ec2-user/apache-hive-2.0.1-bin/lib/log4j-slf4j-impl-2.4.1.jar
!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/ec2-user/hadoop-2.6.4/share/hadoop/common/lib/slf4j-log4j12-1
.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
Execution log at: /tmp/ec2-user/ec2-user_20190219084310_6d04a9ed-4e42-4696-b0be-4eb1f7978cd9.log
2019-02-19 08:43:39     Starting to launch local task to process map join;      maximum memory = 4776
26368
```

```
2019-02-19 08:43:40     Dump the side-table for tag: 0 with group count: 0 into file: file:/tmp/ec2-u
ser/e22048d6-e387-4984-bd5c-ff23bfb5b1b1/hive_2019-02-19_08-43-10_433_8712731078040772434-1/-local-10
012/HashTable-Stage-11/MapJoin-mapfile20--.hashtable
2019-02-19 08:43:40     Uploaded 1 File to: file:/tmp/ec2-user/e22048d6-e387-4984-bd5c-ff23bfb5b1b1/h
ive_2019-02-19_08-43-10_433_8712731078040772434-1/-local-10012/HashTable-Stage-11/MapJoin-mapfile20--
.hashtable (260 bytes)
2019-02-19 08:43:40     End of local task; Time Taken: 0.9 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 3 out of 6
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1550558351319_0012, Tracking URL = http://ip-172-31-3-189.us-east-2.compute.intern
al:8088/proxy/application_1550558351319_0012/
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job  -kill job_1550558351319_0012
Hadoop job information for Stage-11: number of mappers: 1; number of reducers: 0
2019-02-19 08:43:46,357 Stage-11 map = 0%,  reduce = 0%
2019-02-19 08:43:52,547 Stage-11 map = 100%,  reduce = 0%, Cumulative CPU 3.49 sec
MapReduce Total cumulative CPU time: 3 seconds 490 msec
Ended Job = job_1550558351319_0012
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/ec2-user/apache-hive-2.0.1-bin/lib/log4j-slf4j-impl-2.4.1.jar
!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/ec2-user/hadoop-2.6.4/share/hadoop/common/lib/slf4j-log4j12-1
.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
Execution log at: /tmp/ec2-user/ec2-user_20190219084310_6d04a9ed-4e42-4696-b0be-4eb1f7978cd9.log
2019-02-19 08:43:57     Starting to launch local task to process map join;      maximum memory = 4776
26368
2019-02-19 08:43:58     Dump the side-table for tag: 1 with group count: 0 into file: file:/tmp/ec2-u
ser/e22048d6-e387-4984-bd5c-ff23bfb5b1b1/hive_2019-02-19_08-43-10_433_8712731078040772434-1/-local-10
008/HashTable-Stage-4/MapJoin-mapfile01--.hashtable
2019-02-19 08:43:59     Uploaded 1 File to: file:/tmp/ec2-user/e22048d6-e387-4984-bd5c-ff23bfb5b1b1/h
ive_2019-02-19_08-43-10_433_8712731078040772434-1/-local-10008/HashTable-Stage-4/MapJoin-mapfile01--.
hashtable (260 bytes)
2019-02-19 08:43:59     End of local task; Time Taken: 1.199 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 4 out of 6
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1550558351319_0013, Tracking URL = http://ip-172-31-3-189.us-east-2.compute.intern
al:8088/proxy/application_1550558351319_0013/
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job  -kill job_1550558351319_0013
Hadoop job information for Stage-4: number of mappers: 1; number of reducers: 1
2019-02-19 08:44:04,646 Stage-4 map = 0%,  reduce = 0%
2019-02-19 08:44:09,861 Stage-4 map = 100%,  reduce = 0%, Cumulative CPU 1.1 sec
2019-02-19 08:44:13,981 Stage-4 map = 100%,  reduce = 100%, Cumulative CPU 2.24 sec
MapReduce Total cumulative CPU time: 2 seconds 240 msec
Ended Job = job_1550558351319_0013
Launching Job 5 out of 6
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1550558351319_0014, Tracking URL = http://ip-172-31-3-189.us-east-2.compute.intern
al:8088/proxy/application_1550558351319_0014/
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job  -kill job_1550558351319_0014
Hadoop job information for Stage-5: number of mappers: 1; number of reducers: 1
2019-02-19 08:44:20,780 Stage-5 map = 0%,  reduce = 0%
2019-02-19 08:44:24,914 Stage-5 map = 100%,  reduce = 0%, Cumulative CPU 0.86 sec
2019-02-19 08:44:30,053 Stage-5 map = 100%,  reduce = 100%, Cumulative CPU 1.95 sec
MapReduce Total cumulative CPU time: 1 seconds 950 msec
Ended Job = job_1550558351319_0014
MapReduce Jobs Launched:
Stage-Stage-13: Map: 3   Cumulative CPU: 11.8 sec   HDFS Read: 594358966 HDFS Write: 288 SUCCESS
Stage-Stage-11: Map: 1   Cumulative CPU: 3.49 sec   HDFS Read: 17147181 HDFS Write: 96 SUCCESS
Stage-Stage-4: Map: 1  Reduce: 1   Cumulative CPU: 2.24 sec   HDFS Read: 12341 HDFS Write: 96 SUCCESS
Stage-Stage-5: Map: 1  Reduce: 1   Cumulative CPU: 1.95 sec   HDFS Read: 5890 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 19 seconds 480 msec
OK
Time taken: 80.669 seconds
```

Hive Transformation

```
hive> CREATE TABLE Customer (c_custkey int, c_name varchar(25), c_address varchar(50), c_city varchar
(30), c_nation varchar(15), c_region varchar(12), c_phone varchar(15), c_mktsegmet varchar(10)) ROW F
ORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE;
OK
Time taken: 0.045 seconds
hive> LOAD DATA LOCAL INPATH '/home/ec2-user/customer.tbl' OVERWRITE INTO TABLE Customer;
Loading data to table default.customer
OK
Time taken: 0.619 seconds
hive>
```

```
hive> INSERT OVERWRITE TABLE CustomerT SELECT TRANSFORM (c_custkey, c_name, c_ad
dress, c_city, c_nation, c_region, c_phone,c_mktsegmet) USING 'python location.p
y' AS (c_custkey, c_name, c_address, c_city, c_nation, c_region, c_phone, c_mkts
egmet) FROM Customer;
```

```
[ec2-user@ip-172-31-3-189 ~]$ cat location.py
#!/usr/bin/python
import sys

for line in sys.stdin:
        cols=line.strip()
        adress=cols[2]
        city=cols[3].split('/t')
        ncity=[''.join(x) for x in zip(city[0::2],city[1::2])]
        for i in address:
                if len(i)>8:
                        i=i[0:8]
                else:
                        pass
        for a in ncity:
                if a.startswith('UNITED'):
                        a=a[0:6]+ ' ' + a[6:8] + ' ' +'#' +a[-1]
                else:
                        a=a[0:len(a)-1]+' '+ '#' +a[-1]
```

```
hive> add file /home/ec2-user/location.py;
Added resources: [/home/ec2-user/location.py]
```

```
hive> INSERT OVERWRITE TABLE CustomerT SELECT TRANSFORM (c_custkey, c_name, c_ad
dress, c_city, c_nation, c_region, c_phone,c_mktsegmet) USING 'python location.p
y' AS (c_custkey, c_name, c_address, c_city, c_nation, c_region, c_phone, c_mkts
egmet) FROM Customer;
```

PART 3

```
grunt> lineorder = LOAD 'user/ec2-user/lineorder.tbl' USING PigStorage('|') AS(lo_orderkey:int, lo_li
nenumber:int, lo_custkey:int, lo_partkey:int, lo_suppkey:int, lo_orderdate:int, lo_orderpriority:char
array, lo_shippriority:chararray, lo_quantity:int, lo_extendedprice:int, lo_ordertotalprice:int, lo_d
iscount:int, lo_revenue:int, lo_supplycost:int,lo_tax:int,lo_commitdate:int, lo_shipmode:chararray);
2019-02-18 21:08:36,338 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.na
me is deprecated. Instead, use fs.defaultFS
grunt> DESCRIBE lineorder;
lineorder: {lo_orderkey: int,lo_linenumber: int,lo_custkey: int,lo_partkey: int,lo_suppkey: int,lo_or
derdate: int,lo_orderpriority: chararray,lo_shippriority: chararray,lo_quantity: int,lo_extendedprice
: int,lo_ordertotalprice: int,lo_discount: int,lo_revenue: int,lo_supplycost: int,lo_tax: int,lo_comm
itdate: int,lo_shipmode: chararray}
```

0.1

```
Details at logfile: /home/ec2-user/pig-0.15.0/pig_1550559223978.log
grunt> Userset = GROUP Lineorder BY lo_orderkey;
grunt> UserRating = FOREACH Userset GENERATE AVG(Lineorder.lo_revenue);
grunt> DUMP UserRating;
Statistics:

HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
2.6.4   0.15.0  ec2-user        2019-02-19 07:10:14     2019-02-19 07:11:25     GROUP_BY

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTimM
inReduceTime    AvgReduceTime   MedianReducetime        Alias   Feature Outputs
job_1550558351319_0003  5       1       57      30      51      56      27      27      27      27  L
ineorder,UserRating,Userset     GROUP_BY,COMBINER       hdfs://172.31.3.189/tmp/temp457811316/tmp-185
9662845,

Input(s):
Successfully read 6001215 records (594331255 bytes) from: "/user/ec2-user/lineorder.tbl"

Output(s):
Successfully stored 1500000 records (19500000 bytes) in: "hdfs://172.31.3.189/tmp/temp457811316/tmp-1
859662845"

Counters:
Total records written : 1500000
Total bytes written : 19500000
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1550558351319_0003


2019-02-19 07:11:25,554 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceMa
nager at /172.31.3.189:8032
2019-02-19 07:11:25,558 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application sta
te is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2019-02-19 07:11:25,605 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceMa
nager at /172.31.3.189:8032
2019-02-19 07:11:25,612 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application sta
te is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2019-02-19 07:11:25,640 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceMa
nager at /172.31.3.189:8032
2019-02-19 07:11:25,645 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application sta
te is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2019-02-19 07:11:25,672 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.Map
ReduceLauncher - Success!
2019-02-19 07:11:25,672 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.na
me is deprecated. Instead, use fs.defaultFS
2019-02-19 07:11:25,673 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] w
as not set... will not generate code.
2019-02-19 07:11:25,685 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total in
put paths to process : 1
2019-02-19 07:11:25,685 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil -
Total input paths to process : 1
(2786386.8333333335)
(4469446.0)
```

0.2

```
DiscountSet = GROUP Lineorder BY lo_discount;                                                           ;
grunt> DiscountPrice = FOREACH DiscountSet GENERATE Lineorder.lo_discount, COUNT(Lineorder.lo_extende
dprice);
grunt> DUMP DiscountPrice;
Statistics:

HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
2.6.4   0.15.0  ec2-user        2019-02-19 07:21:10     2019-02-19 07:22:36     GROUP_BY

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTimM
inReduceTime    AvgReduceTime   MedianReducetime        Alias   Feature Outputs
job_1550558351319_0004  5       1       29      10      19      17      64      64      64      64  D
iscountPrice,DiscountSet,Lineorder      GROUP_BY        hdfs://172.31.3.189/tmp/temp457811316/tmp-114
4581449,

Input(s):
Successfully read 6001215 records (594331255 bytes) from: "/user/ec2-user/lineorder.tbl"

Output(s):
Successfully stored 11 records (16913123 bytes) in: "hdfs://172.31.3.189/tmp/temp457811316/tmp-114458
1449"

Counters:
Total records written : 11
Total bytes written : 16913123
Spillable Memory Manager spill count : 12
Total bags proactively spilled: 11
Total records proactively spilled: 5161341

Job DAG:
job_1550558351319_0004


2019-02-19 07:22:36,673 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceMa
nager at /172.31.3.189:8032
2019-02-19 07:22:36,685 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application sta
te is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2019-02-19 07:22:36,723 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceMa
nager at /172.31.3.189:8032
2019-02-19 07:22:36,727 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application sta
te is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2019-02-19 07:22:36,767 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceMa
nager at /172.31.3.189:8032
2019-02-19 07:22:36,777 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application sta
te is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2019-02-19 07:22:36,810 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.Map
ReduceLauncher - Success!
2019-02-19 07:22:36,810 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.na
me is deprecated. Instead, use fs.defaultFS
2019-02-19 07:22:36,811 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] w
as not set... will not generate code.
2019-02-19 07:22:36,835 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total in
put paths to process : 1
2019-02-19 07:22:36,835 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil -
Total input paths to process : 1
({(0),(0),(0),(0),(0),(0),(0),(0),(0),(0),(0),(0),(0),(0),(0),(0),(0),(0),(0),(0),(0),(0),(0),(0),(0)
,(0),(0),(0),(0),(0),(0),(0),(0),(0),(0),(0),(0),(0),(0),(0),(0),(0),(0),(0),(0),(0),(0),(0),(0),(0),
```

0.3

```
grunt> Discountset = FILTER Lineorder BY lo_discount < 3;
grunt> Quantityset = GROUP Discountset BY lo_quantity;
grunt> DUMP Quantityset;
grunt> Results = FOREACH Quantityset GENERATE Discountset.lo_quantity, SUM(Discountset.lo_revenue);
grunt> DUMP Results;
Statistics:

HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
2.6.4   0.15.0  ec2-user        2019-02-19 07:51:50     2019-02-19 07:52:41     GROUP_BY,FILTER

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTimM
inReduceTime    AvgReduceTime   MedianReducetime        Alias   Feature Outputs
job_1550558351319_0006  5       1       33      11      25      31      27      27      27      27  D
iscountset,Lineorder,Quantityset,Results        GROUP_BY        hdfs://172.31.3.189/tmp/temp457811316
/tmp-773613275,

Input(s):
Successfully read 6001215 records (594331255 bytes) from: "/user/ec2-user/lineorder.tbl"

Output(s):
Successfully stored 50 records (4878628 bytes) in: "hdfs://172.31.3.189/tmp/temp457811316/tmp-7736132
75"

Counters:
Total records written : 50
Total bytes written : 4878628
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1550558351319_0006


2019-02-19 07:52:41,764 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceMa
nager at /172.31.3.189:8032
2019-02-19 07:52:41,778 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application sta
te is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2019-02-19 07:52:41,807 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceMa
nager at /172.31.3.189:8032
2019-02-19 07:52:41,814 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application sta
te is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2019-02-19 07:52:41,834 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceMa
nager at /172.31.3.189:8032
2019-02-19 07:52:41,837 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application sta
te is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2019-02-19 07:52:41,855 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.Map
ReduceLauncher - Success!
2019-02-19 07:52:41,856 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.na
me is deprecated. Instead, use fs.defaultFS
2019-02-19 07:52:41,856 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] w
as not set... will not generate code.
2019-02-19 07:52:41,860 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total in
put paths to process : 1
2019-02-19 07:52:41,860 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil -
Total input paths to process : 1
({(1),(1),(1),(1),(1),(1),(1),(1),(1),(1),(1),(1),(1),(1),(1),(1),(1),(1),(1),(1),(1),(1),(1),(1),(1)
```

PART 4

0.2

```
#!/usr/bin/python

import sys, datetime

for line in  sys.stdin:
        line=line.strip()
        vals=line.split('|')
        print(vals)
[ec2-user@ip-172-31-3-189 ~]$
```

```
#!/usr/bin/env python


import sys, datetime
curr_id=None
price_cnt=0
id=None

for line in sys.stdin:
        line=line.strip()
        ln=line.split(',')
        id=ln[0]
        if curr_id==id:
                price_cnt+=1
        else:
                if curr_id:
                        print(curr_id, price_cnt)
                curr_id=id
                curr_cnt=0
if curr_id == id:
        print ('%s\t%d' % (curr_id, price_cnt))

[[ec2-user@ip-172-31-3-189 ~]$ hadoop jar hadoop-streaming-2.6.4.jar -input /data
 /0.2 -output /data/output_11 -mapper '02Mapper.py' -reducer '02Reducer.py' -file
  02Reducer.py -file 02Mapper.py
```

0.3

```
[ec2-user@ip-172-31-3-189 ~]$ cat 03Mapper.py
#!/usr/bin/python

import sys, datetime

for line in sys.stdin:
        line = line.strip()
        vals=line.split('|')
        discount=vals[11]
        revenue=int(vals[12])
        quantity=vals[8]
        print("%s\t%s\t%s" % (quantity, revenue, discount))
```

```
[ec2-user@ip-172-31-3-189 ~]$ cat 03Reducer.py
#!/usr/bin/env python

import sys, datetime

curr_id=None
curr_cnt=0
id=None
rev=[]
for line in sys.stdin:
        line=line.strip()
        ln=line.split('t')
        id=int(ln[0])

        if curr_id == id and ln[2]<3:
                curr_cnt+=1
                rev.append(ln[1])
        else:
                if curr_id:
                        print ('%s\t%d' % (curr_id, sum(rev)))
                curr_id=id
                curr_ct=0
if curr_id == id:
        print('%s\t%d' % (curr_id, sum(rev)))
```

```
[ec2-user@ip-172-31-3-189 ~]$ hadoop jar hadoop-streaming-2.6.4.jar -input /data
/0.2 -output /data/output_1 -mapper '03Mapper.py' -reducer '03Reducer.py' -file
03Reducer.py -file 03Mapper.py
```