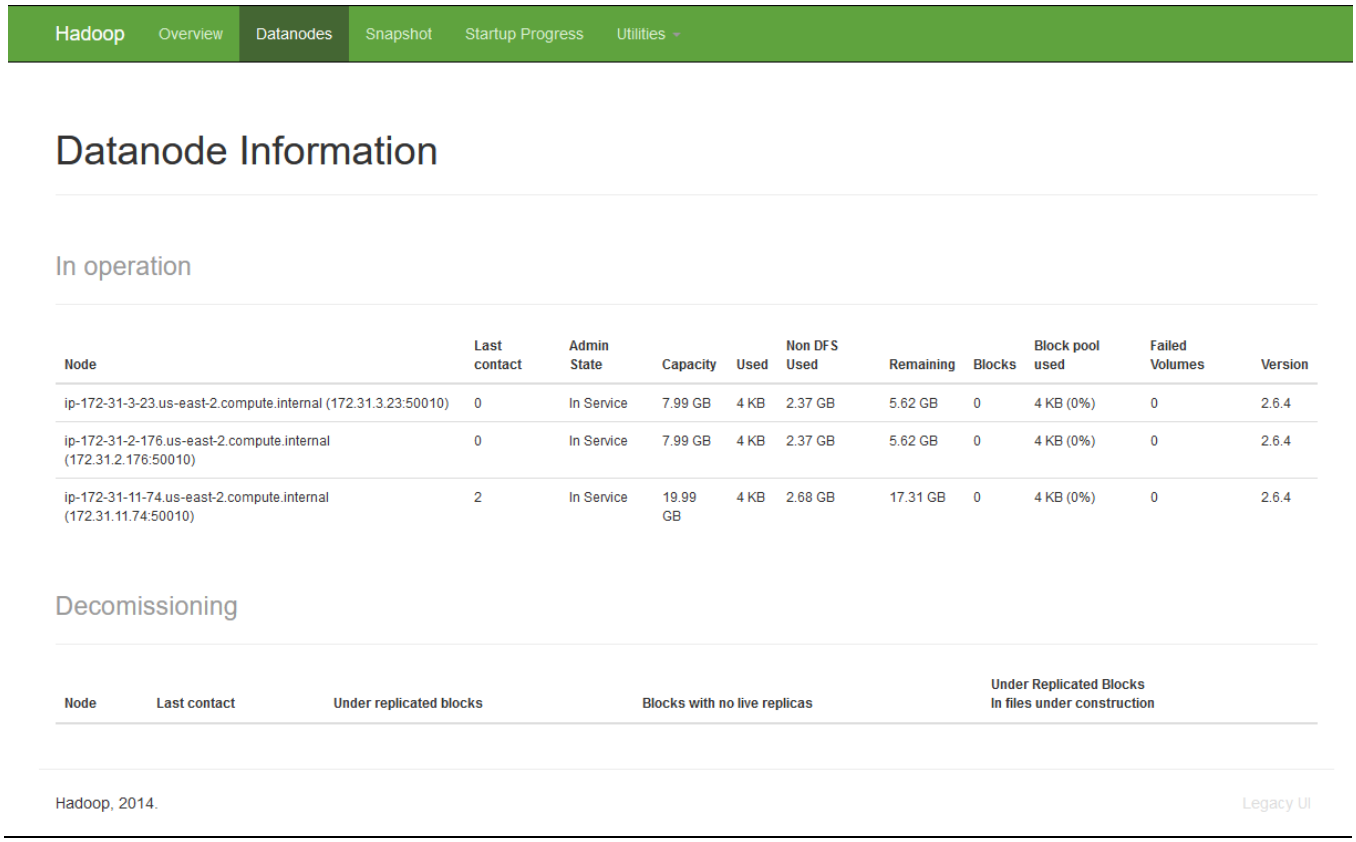
Keiland Pullen
DSC 555

DSC 333 and CSC 555 Take-home midterm

(due Saturday, May 21st)

Part 1: Multi-node cluster



```
[ec2-user@ip-172-31-11-74 ~]$ time hadoop jar hadoop-2.6.4/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.6.4.jar  wordcount /data/bioproject.xml /data/wordcount1
22/05/19 23:20:05 INFO client.RMProxy: Connecting to ResourceManager at /172.31.11.74:8032
22/05/19 23:20:05 INFO input.FileInputFormat: Total input paths to process : 1
22/05/19 23:20:06 INFO mapreduce.JobSubmitter: number of splits:2
22/05/19 23:20:06 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1653000251899_0001
22/05/19 23:20:06 INFO impl.YarnClientImpl: Submitted application application_1653000251899_0001
22/05/19 23:20:06 INFO mapreduce.Job: The url to track the job: http://ip-172-31-11-74.us-east-2.compute.internal:8088/proxy/application_1653000251899_0001/
22/05/19 23:20:06 INFO mapreduce.Job: Running job: job_1653000251899_0001
22/05/19 23:20:18 INFO mapreduce.Job: Job job_1653000251899_0001 running in uber mode : false
22/05/19 23:20:18 INFO mapreduce.Job:  map 0% reduce 0%
22/05/19 23:20:31 INFO mapreduce.Job:  map 26% reduce 0%
22/05/19 23:20:34 INFO mapreduce.Job:  map 27% reduce 0%
22/05/19 23:20:37 INFO mapreduce.Job:  map 47% reduce 0%
22/05/19 23:20:40 INFO mapreduce.Job:  map 48% reduce 0%
22/05/19 23:20:43 INFO mapreduce.Job:  map 60% reduce 0%
22/05/19 23:20:44 INFO mapreduce.Job:  map 77% reduce 0%
22/05/19 23:20:46 INFO mapreduce.Job:  map 83% reduce 0%
22/05/19 23:20:51 INFO mapreduce.Job:  map 100% reduce 0%
22/05/19 23:20:54 INFO mapreduce.Job:  map 100% reduce 100%
22/05/19 23:20:55 INFO mapreduce.Job: Job job_1653000251899_0001 completed successfully
22/05/19 23:20:55 INFO mapreduce.Job: Counters: 49
        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=2
                Total time spent by all maps in occupied slots (ms)=54952
                Total time spent by all reduces in occupied slots (ms)=7463
                Total time spent by all map tasks (ms)=54952
                Total time spent by all reduce tasks (ms)=7463
                Total vcore-milliseconds taken by all map tasks=54952
                Total vcore-milliseconds taken by all reduce tasks=7463
                Total megabyte-milliseconds taken by all map tasks=56270848
                Total megabyte-milliseconds taken by all reduce tasks=7642112
        Map-Reduce Framework
```

```
                Total megabyte-milliseconds taken by all reduce tasks=7642112
        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)=715
                CPU time spent (ms)=42500
                Physical memory (bytes) snapshot=773029888
                Virtual memory (bytes) snapshot=6415982592
                Total committed heap usage (bytes)=533200896
        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    0m52.155s
user    0m4.226s
sys     0m0.290s
[ec2-user@ip-172-31-11-74 ~]$
```

```
sys     0m0.290s
[ec2-user@ip-172-31-11-74 ~]$ hadoop fs -du /data/wordcount1/
0         /data/wordcount1/_SUCCESS
20056175  /data/wordcount1/part-r-00000
[ec2-user@ip-172-31-11-74 ~]$
[ec2-user@ip-172-31-11-74 ~]$ hadoop fs -cat /data/wordcount1/part-r-00000 | grep arctic
&lt;I&gt;holarctica&lt;/I&gt;   28
&lt;I&gt;holarctica&lt;/I&gt;&lt;/B&gt;.       8
&lt;I&gt;holarctica&lt;/I&gt;,  1
&lt;I&gt;palearctica&lt;/I&gt;  4
&lt;i&gt;holarctica&lt;/i&gt;   1
(Antarctic      3
(Antarctica)    1
(Antarctica),   11
<Label>Antarctic        1
<Name>Antarctic 3
<Name>Antarctica        1
<Strain>Antarctic       1
<Title>Antarctic        5
Antarctic       137
Antarctic,      1
Antarctic.      2
Antarctic.</Description>        1
Antarctic.</Title>      1
Antarctic</Title>       4
Antarctica      16
Antarctica)</Title>     1
Antarctica,     9
Antarctica.     24
Antarctica.&#x0D;       3
Antarctica.</Description>       19
Antarctica</Description>        2
Antarctica</Name>       1
Antarctica</Title>      6
Palearctic      1
Project">Antarctic      1
Subarctic       11
abbr="Antarctic 1
antarctic       5
antarctica      17
antarctica&lt;/i&gt;&lt;/b&gt;.&#x0D;   2
antarctica,     4
antarctica</Name>       10
antarctica</OrganismName>       11
antarctica</Title>      1
antarcticum     32
antarcticum</Name>      3
antarcticum</OrganismName>      3
antarcticus     31
antarcticus&lt;/i&gt;   4
antarcticus&lt;/i&gt;&lt;/b&gt;.        1
antarcticus).   1
antarcticus,    1
antarcticus</Name>      5
antarcticus</OrganismName>      5
arctic  21
arctica 27
arctica&lt;/I&gt;)      2
arctica&lt;/i&gt;       3
arctica&lt;/i&gt;,      1
arctica.</Description>  2
```

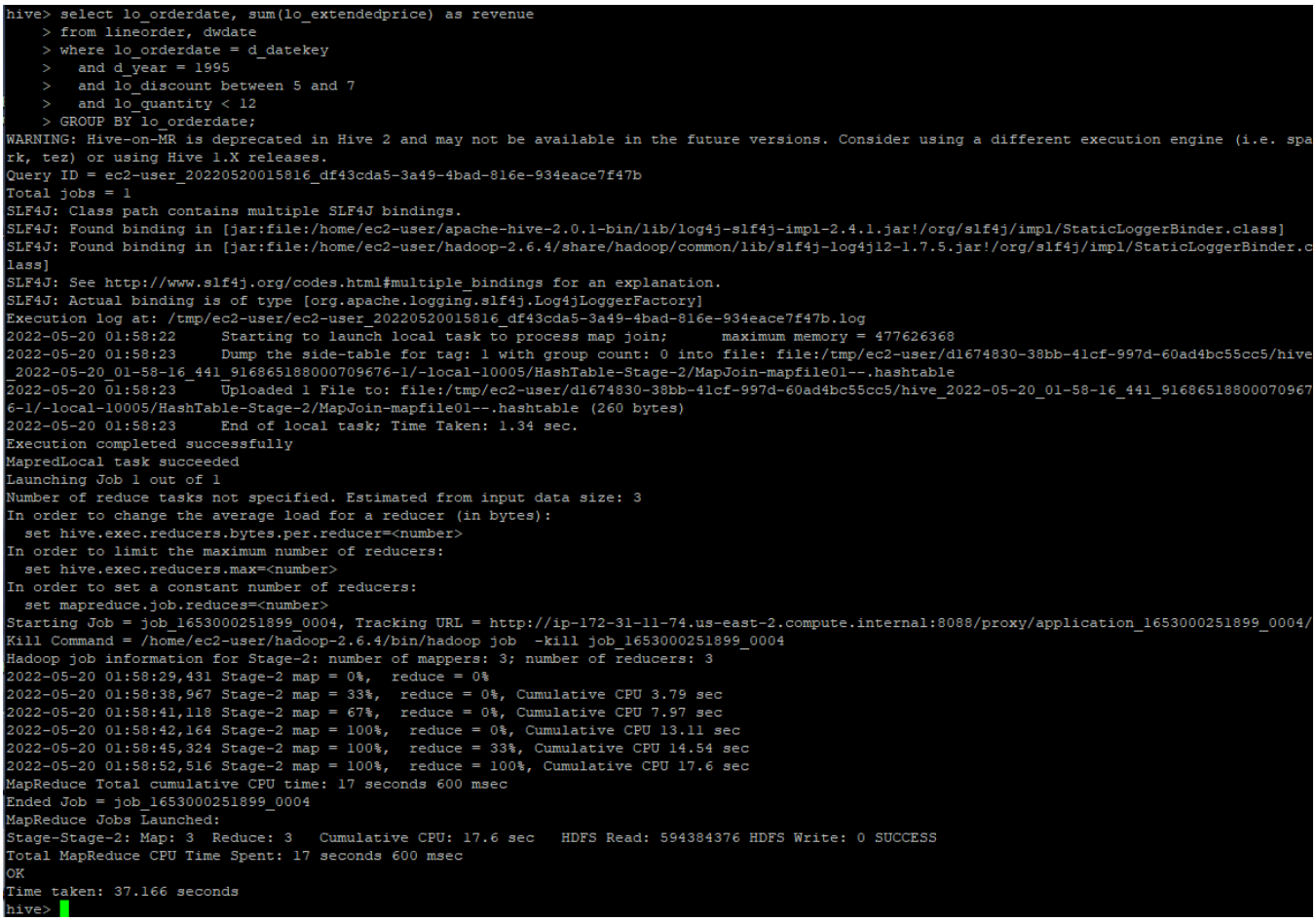
```
arctica.</Description>  2
arctica</Name>  5
arctica</OrganismName>  5
arcticus        31
arcticus&lt;/i&gt;      2
arcticus</Name> 4
arcticus</OrganismName> 4
holarctica      77
humans.Antarctic        1
palearctica     66
palearctica</Name>      1
sub-Antarctic   4
sub-arctic      4
subantarctic    1
subantarcticus  7
subantarcticus</Name>   1
subantarcticus</OrganismName>   1
subarctic       21
[ec2-user@ip-172-31-11-74 ~]$
```

In assignment 2, the execution time was 1 minute and 28 seconds using a single node instance. For this case, using a three node instance, the execution time was 52 seconds. In theory, it seems that the execution time could have been faster. To investigate, I used the "hadoop dfsadmin –report" shell command. After reviewing the report, I noticed that the "Configured Capacity" for each worker node is different. This could be a factor as to why the execution speed on the three nodes was not as fast as I expected. Another factor could be the actual network. The following is the report that was returned:

```
[ec2-user@ip-172-31-11-74 ~]$ hdfs dfsadmin -report
Configured Capacity: 38616895488 (35.96 GB)
Present Capacity: 30412926976 (28.32 GB)
DFS Remaining: 29906210816 (27.85 GB)
DFS Used: 506716160 (483.24 MB)
DFS Used%: 1.67%
Under replicated blocks: 0
Blocks with corrupt replicas: 0
Missing blocks: 0

-------------------------------------------------
Live datanodes (3):

Name: 172.31.3.23:50010 (ip-172-31-3-23.us-east-2.compute.internal)
Hostname: ip-172-31-3-23.us-east-2.compute.internal
Decommission Status : Normal
Configured Capacity: 8577331200 (7.99 GB)
DFS Used: 155664384 (148.45 MB)
Non DFS Used: 2546814976 (2.37 GB)
DFS Remaining: 5874851840 (5.47 GB)
DFS Used%: 1.81%
DFS Remaining%: 68.49%
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 19 23:38:26 UTC 2022


Name: 172.31.11.74:50010 (ip-172-31-11-74.us-east-2.compute.internal)
Hostname: ip-172-31-11-74.us-east-2.compute.internal
Decommission Status : Normal
Configured Capacity: 21462233088 (19.99 GB)
DFS Used: 253186048 (241.46 MB)
Non DFS Used: 3109011456 (2.90 GB)
DFS Remaining: 18100035584 (16.86 GB)
DFS Used%: 1.18%
DFS Remaining%: 84.33%
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 19 23:38:27 UTC 2022


Name: 172.31.2.176:50010 (ip-172-31-2-176.us-east-2.compute.internal)
Hostname: ip-172-31-2-176.us-east-2.compute.internal
Decommission Status : Normal
Configured Capacity: 8577331200 (7.99 GB)
DFS Used: 97865728 (93.33 MB)
Non DFS Used: 2548142080 (2.37 GB)
DFS Remaining: 5931323392 (5.52 GB)
DFS Used%: 1.14%
DFS Remaining%: 69.15%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
```

Part 2: Hive

1) Total time to execute query is 37.166 seconds, see the following screen shot:



2) Perform the following transform operation using SELECT TRANSFORM on the dwdate table by creating a new table. The new dwdate table will combine d_daynuminweek, d_daynuminmonth, and d_daynuminyear into a single column in the new table using a delimiter of your choice. You should also eliminate the following 1 column: d_lastdayinmonthfl. The final table will have fewer columns than the original table because you merge 3 columns into 1 and remove 1 column.

The following code is used to create the second table, dwdate2:

```
Create table dwdate2(
 d_datekey          int,
  d_date            varchar(19),
  d_dayofweek       varchar(10),
  d_month           varchar(10),
  d_year            int,
  d_yearmonthnum    int,
  d_yearmonth       varchar(8),
  d_daynuminweek    varchar(15),
  d_daynuminmonth   int,
  d_daynuminyear    int,
  d_monthnuminyear  int,
  d_weeknuminyear   int,
  d_sellingseason   varchar(13),
  d_lastdayinweekfl varchar(1),
  d_lastdayinmonthfl varchar(1),
  d_holidayfl       varchar(1),
  d_weekdayfl       varchar(1)
)
ROW FORMAT DELIMITED FIELDS
TERMINATED BY '\t' STORED AS TEXTFILE;
```

The following is python code used for the transformation, dwdate_transform.py code and a screen shot:

```
#!/usr/bin/python
import sys

for line in sys.stdin:
    line = line.strip().split('\t')
    seven = line[7]
    eight  = line[8]
    nine   = line[9]
    d = '-'

    line[7] = seven + d + eight + d + nine

    del line[14]
    del line[9]
    del line[8]

    print '\t'.join(line)
```

```
#!/usr/bin/python
import sys

for line in sys.stdin:
        line = line.strip().split('\t')
        seven = line[7]
        eight  = line[8]
        nine   = line[9]
        d = '-'

        line[7] = seven + d + eight + d + nine

        del line[14]
        del line[9]
        del line[8]

        print '\t'.join(line)
```

```
hive> add file /home/ec2-user/dwdate_transform.py;
Added resources: [/home/ec2-user/dwdate_transform.py]
hive> insert overwrite table dwdate2 SELECT TRANSFORM (d_datekey, d_date, d_dayofweek, d_month, d_year, d_yearmonthnum, d_yearmonth, d_daynuminweek, d_daynuminmonth, d_daynu
minyear, d_monthnuminyear, d_weeknuminyear, d_sellingseason, d_lastdayinweekfl, d_lastdayinmonthfl, d_holidayfl, d_weekdayfl) using 'dwdate_transform.py' as (d_datekey, d_da
te, d_dayofweek, d_month, d_year, d_yearmonthnum, d_yearmonth, d_daynuminweek, d_daynuminmonth, d_daynuminyear, d_monthnuminyear, d_weeknuminyear, d_sellingseason, d_lastday
inweekfl, d_lastdayinmonthfl, d_holidayfl, d_weekdayfl) 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_20220520160830_913940c6-eb3d-4a00-a4ee-ee6b107792c9
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1653055965520_0012, Tracking URL = http://ip-172-31-11-74.us-east-2.compute.internal:8088/proxy/application_1653055965520_0012/
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job  -kill job_1653055965520_0012
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2022-05-20 16:08:35,714 Stage-1 map = 0%,  reduce = 0%
2022-05-20 16:08:41,953 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.29 sec
MapReduce Total cumulative CPU time: 2 seconds 290 msec
Ended Job = job_1653055965520_0012
Stage-4 is selected by condition resolver.
Stage-3 is filtered out by condition resolver.
Stage-5 is filtered out by condition resolver.
Moving data to: hdfs://172.31.11.74/user/hive/warehouse/dwdate2/.hive-staging_hive_2022-05-20_16-08-30_838_8036125505059529635-1/-ext-10000
Loading data to table default.dwdate2
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1   Cumulative CPU: 2.29 sec   HDFS Read: 241130 HDFS Write: 243866 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 290 msec
OK
Time taken: 12.272 seconds
hive> select * from dwdate2 limit 8;
OK
19920101        January 1, 1992 Thursday        January 1992    199201  Jan1992 5-1-1   1       1       1       1       Winter  0       1       1       1
19920102        January 2, 1992 Friday  January 1992    199201  Jan1992 6-2-2   2       2       1       1       Winter  0       1       0       1
19920103        January 3, 1992 Saturday        January 1992    199201  Jan1992 7-3-3   3       3       1       1       Winter  1       1       0       0
19920104        January 4, 1992 Sunday  January 1992    199201  Jan1992 1-4-4   4       4       1       1       Winter  0       1       0       0
19920105        January 5, 1992 Monday  January 1992    199201  Jan1992 2-5-5   5       5       1       1       Winter  0       1       0       1
19920106        January 6, 1992 Tuesday January 1992    199201  Jan1992 3-6-6   6       6       1       1       Winter  0       1       0       1
19920107        January 7, 1992 Wednesday       January 1992    199201  Jan1992 4-7-7   7       7       1       2       Winter  0       1       0       1
19920108        January 8, 1992 Thursday        January 1992    199201  Jan1992 5-8-8   8       8       1       2       Winter  0       1       0       1
Time taken: 0.049 seconds, Fetched: 8 row(s)
hive>
```

The above is a screen-shot that displays adding the transform code, inserting the data, and a query to view the output:

- hive> add file /home/ec2-user/dwdate_transform.py;

- hive> insert overwrite table dwdate2 SELECT TRANSFORM (d_datekey, d_date, d_dayofweek, d_month, d_year, d_yearmonthnum, d_yearmonth, d_daynuminweek, d_daynuminmonth, d_daynuminyear, d_monthnuminyear, d_weeknuminyear, d_sellingseason, d_lastdayinweekfl, d_lastdayinmonthfl, d_holidayfl, d_weekdayfl) using 'dwdate_transform.py' as (d_datekey, d_date, d_dayofweek, d_month, d_year, d_yearmonthnum, d_yearmonth, d_daynuminweek, d_daynuminmonth, d_daynuminyear, d_monthnuminyear, d_weeknuminyear, d_sellingseason, d_lastdayinweekfl, d_lastdayinmonthfl, d_holidayfl, d_weekdayfl) from dwdate;

- hive> select * from dwdate2 limit 8;

In this output above, the transform code was altered to comment out the “del” lines. If the “del” lines are included, then the final output contains Null values, see the following image:

```
hive> select * from dwdate2 limit 8;
OK
19920101        January 1, 1992 Thursday        January 1992    199201  Jan1992 5-1-1   1       1       NULL    0       1       1       NULL    NULL    NULL
19920102        January 2, 1992 Friday  January 1992    199201  Jan1992 6-2-2   1       1       NULL    0       0       1       NULL    NULL    NULL
19920103        January 3, 1992 Saturday        January 1992    199201  Jan1992 7-3-3   1       1       NULL    1       0       0       NULL    NULL    NULL
19920104        January 4, 1992 Sunday  January 1992    199201  Jan1992 1-4-4   1       1       NULL    0       0       0       NULL    NULL    NULL
19920105        January 5, 1992 Monday  January 1992    199201  Jan1992 2-5-5   1       1       NULL    0       0       1       NULL    NULL    NULL
19920106        January 6, 1992 Tuesday January 1992    199201  Jan1992 3-6-6   1       1       NULL    0       0       1       NULL    NULL    NULL
19920107        January 7, 1992 Wednesday       January 1992    199201  Jan1992 4-7-7   1       2       NULL    0       0       1       NULL    NULL    NULL
19920108        January 8, 1992 Thursday        January 1992    199201  Jan1992 5-8-8   1       2       NULL    0       0       1       NULL    NULL    NULL
Time taken: 0.064 seconds, Fetched: 8 row(s)
```

The issue that I encountered was how to alter the final table so that it would not include the “d_lastdayinmonthfl” column.

Part 3: Pig

Convert and load the data into Pig, implementing and timing the following queries:

a. SELECT lo_discount, AVG(lo_extendedprice)
FROM lineorder
GROUP BY lo_discount;

b. SELECT lo_quantity, SUM(lo_revenue)
FROM lineorder
WHERE lo_discount > 8 AND lo_quantity < 23
GROUP BY lo_quantity;

3a.)

```
lineorder = LOAD '/user/lineorder.tbl' USING PigStorage('|')
AS (lo_orderkey:int, lo_linenumber: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_ordertotalprice:int, lo_discount:int, lo_revenue:int, lo_supplycost:int, lo_tax:int, lo_commitdate:int,
lo_shipmode:chararray);

lineord2 = GROUP lineorder BY lo_discount;
lineordAVG = FOREACH lineord2 GENERATE lineorder.lo_discount, AVG(lineorder.lo_extendedprice);
DUMP lineordAVG;
```

```
Details at logfile: /home/ec2-user/pig-0.15.0/pig_1653168567932.log
2022-05-21 21:32:17,177 [main] INFO  org.apache.pig.Main - Pig script completed in 2 minutes, 49 seconds and 404 milliseconds (169404 ms
)
[ec2-user@ip-172-31-11-74 pig-0.15.0]$
```

Time to execute: 2 mins. 49 secs and 404 ms

3b.)

```
lineorder = LOAD '/user/lineorder.1M.tbl' USING PigStorage('|')
AS (lo_orderkey:int, lo_linenumber: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_ordertotalprice:int, lo_discount:int, lo_revenue:int, lo_supplycost:int, lo_tax:int, lo_commitdate:int,
lo_shipmode:chararray);

lineordFilter = FILTER lineorder BY lo_discount>8 AND lo_quantity<23;
lineord3 = GROUP lineordFilter BY lo_quantity;
lineordSUM = FOREACH lineord3 GENERATE group as lo_quantity, SUM(lineordFilter.lo_revenue) ;

STORE lineordSUM into 'lineordSUM' USING PigStorage('_');
```

```
2022-05-21 21:54:06,453 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2022-05-21 21:54:06,471 [main] INFO  org.apache.pig.Main - Pig script completed in 1 minute, 59 seconds and 513 milliseconds (119513 ms)
[ec2-user@ip-172-31-11-74 pig-0.15.0]$
```

```
[ec2-user@ip-172-31-11-74 pig-0.15.0]$ cd
[ec2-user@ip-172-31-11-74 ~]$ hadoop fs -ls /user/ec2-user/lineordSUM
Found 2 items
-rw-r--r--   2 ec2-user supergroup          0 2022-05-21 21:54 /user/ec2-user/lineordSUM/_SUCCESS
-rw-r--r--   2 ec2-user supergroup        318 2022-05-21 21:54 /user/ec2-user/lineordSUM/part-r-00000
[ec2-user@ip-172-31-11-74 ~]$
```

Time to execute: 1 minute, 59 secs and 513 ms

File size: 318

Part 4: Hadoop Streaming

```
[ec2-user@ip-172-31-11-74 ~]$ hadoop fs -mkdir /data
[ec2-user@ip-172-31-11-74 ~]$ hadoop fs -put lineorder.tbl /data/
[ec2-user@ip-172-31-11-74 ~]$ hadoop fs -put dwdate.tbl /data/
[ec2-user@ip-172-31-11-74 ~]$
[ec2-user@ip-172-31-11-74 ~]$ hadoop fs -ls /data
Found 2 items
-rw-r--r--   2 ec2-user supergroup     229965 2022-05-22 00:51 /data/dwdate.tbl
-rw-r--r--   2 ec2-user supergroup  594313001 2022-05-22 00:51 /data/lineorder.tbl
[ec2-user@ip-172-31-11-74 ~]$
```

streamMapper.py :

```
#!/usr/bin/python

import sys

for line in sys.stdin:
    dayofweek     = ""
    datekey       = ""
    month         = ""
    orderdate     = ""
    extendedprice = ""

    line = line.strip()
    vals = line.split("|")

    if vals[2].endswith('day'):
        #print('ends with day and is from dwdate: ' + vals[2])
        #print('Day of week is: ' + vals[2] )
        #print('d_month is : ' + vals[3] )
        datekey   = vals[0]
        dayofweek = vals[2]
        month     = vals[3]
        year      = vals[4]

        if year == '1995':
            print '%s\t%s\t%s' % (datekey,dayofweek,month)

    else:
       #print('is from lineorder table: ' + vals[2])
       #print('orderdate is: ' + vals[5] )
       #print('extended price is: ' + vals[9] )
       orderdate = vals[5]
       discount = vals[11]
       quantity = vals[8]
       extendedprice = vals[9]

       if (int(discount) > 5 and int(discount) < 7 ) and (int(quantity) < 12 ) :
           print '%s\t%s' % (orderdate,extendedprice)
```

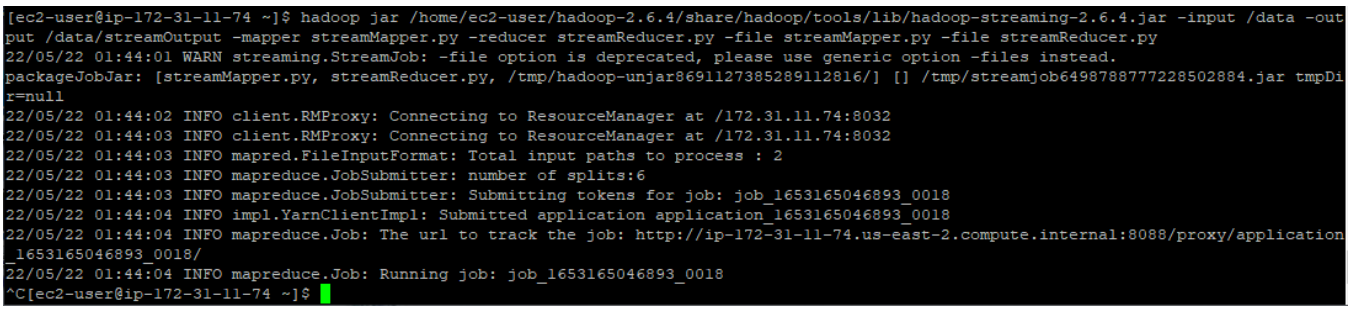
streamReducer.py :

```
#!/usr/bin/python

import sys

currentKey  = None
total       = None
val_lo      = None
val_date    = None
# lenLo       = 0
# lenDate     = 0

for line in sys.stdin:
    line  = line.strip().split('\t')
    key   = line[0]
    value = line[1:]
    val_1  = line[1]
    #val_2  = line[2]
    #val_3  = line[3]

    if currentKey == key:
        if val_1.endswith('day'):
            val_date.append(value)
        else:
            val_lo.append(value)

        print '%s\t%s' % (currentKey, value)
        #print (currentKey, '\t', value)

    else:
       if currentKey:
          len_Lo   = len(val_lo)
          len_Date = len(val_date)

          if (len_Lo * len_Date > 0):
              print '%s\t%s' % (currentKey, value)
              #print (currentKey, '\t', value)
       val_lo   = []
       val_date = []
       currentKey = key

       if val_1.endswith('day'):
           val_date = []
           val_date = [value]
       else:
           val_lo = []
           val_lo = [value]


len_lo_last   = len(val_lo)
len_date_last = len(val_date)
if (len_lo_last * len_date_last > 0):
      print '%s\t%s\t%s\t%s' % (currentKey, val_1, val_2, val_3)
```

Command-line :

Unfortunately, the following command-line results in a hanging Hadoop. After several attempts, I executed “Ctrl-C” on the command-line to stop execution

[ec2-user@ip-172-31-11-74 ~]$ hadoop jar /home/ec2-user/hadoop-2.6.4/share/hadoop/tools/lib/hadoop-streaming-2.6.4.jar -input /data -output /data/streamOutput -mapper streamMapper.py -reducer streamReducer.py -file streamMapper.py -file streamReducer.py



Execution time :

I was unable to capture the execution time as the code causes Hadoop to continuously hang.