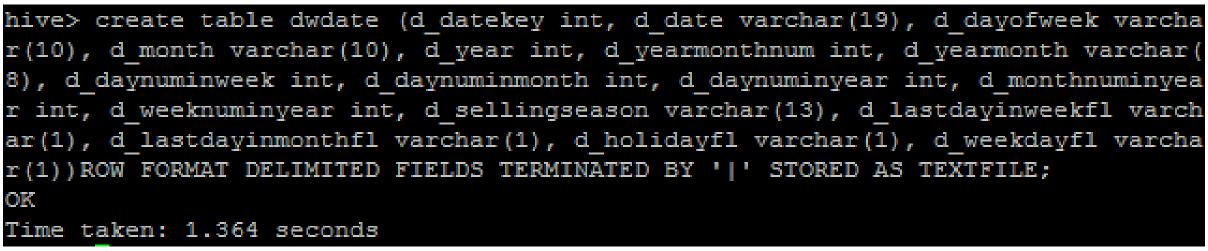
Hive

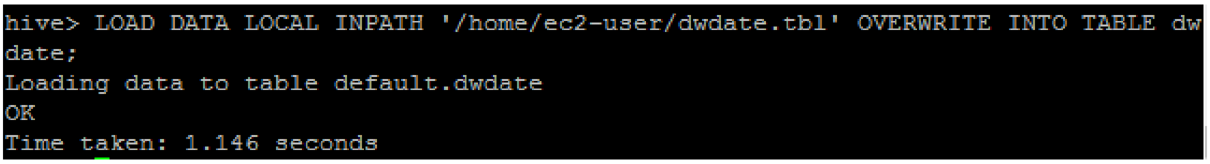
Sum and Transform

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 varchar(8), d_daynuminweek int, 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 '|' STORED AS TEXTFILE;



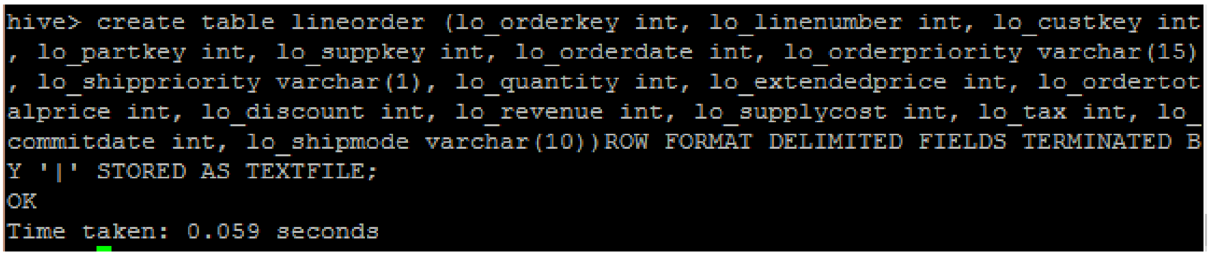
```
hive> create table dwdate (d_datekey int, d_date varchar(19), d_dayofweek varcha
r(10), d_month varchar(10), d_year int, d_yearmonthnum int, d_yearmonth varchar(
8), d_daynuminweek int, d_daynuminmonth int, d_daynuminyear int, d_monthnuminyea
r int, d_weeknuminyear int, d_sellingseason varchar(13), d_lastdayinweekfl varch
ar(1), d_lastdayinmonthfl varchar(1), d_holidayfl varchar(1), d_weekdayfl varcha
r(1))ROW FORMAT DELIMITED FIELDS TERMINATED BY '|' STORED AS TEXTFILE;
OK
Time taken: 1.364 seconds
```

LOAD DATA LOCAL INPATH '/home/ec2-user/dwdate.tbl' OVERWRITE INTO TABLE dwdate;



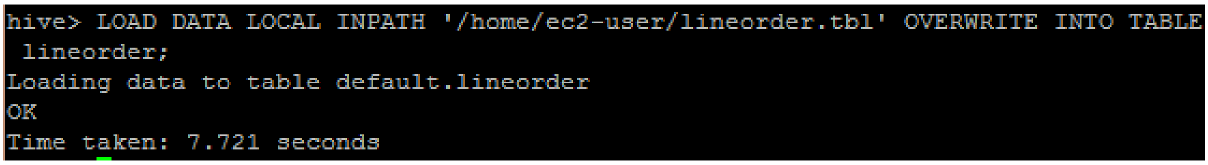
```
hive> LOAD DATA LOCAL INPATH '/home/ec2-user/dwdate.tbl' OVERWRITE INTO TABLE dw
date;
Loading data to table default.dwdate
OK
Time taken: 1.146 seconds
```

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;



```
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_ordertot
alprice 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 B
Y '|' STORED AS TEXTFILE;
OK
Time taken: 0.059 seconds
```

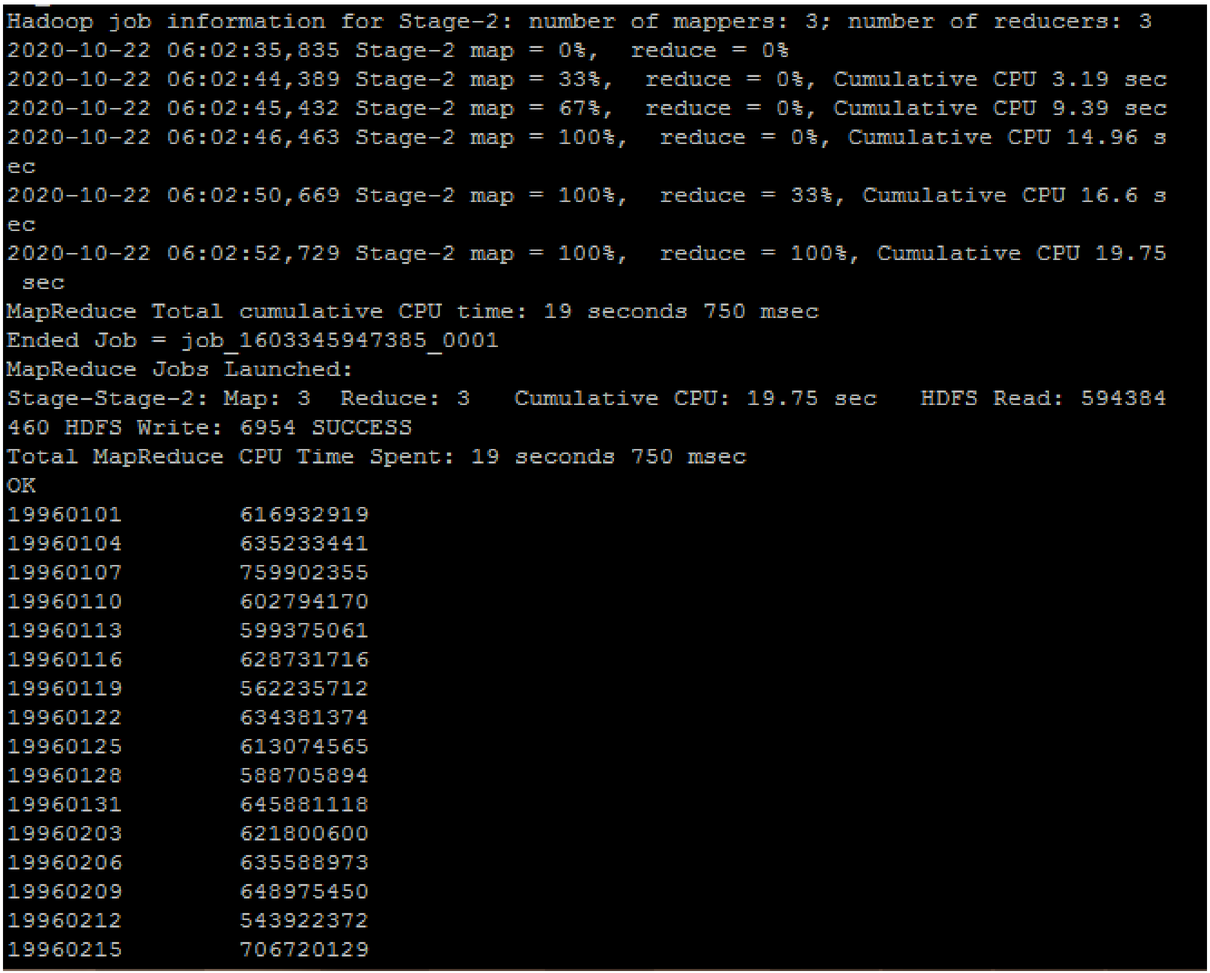
LOAD DATA LOCAL INPATH '/home/ec2-user/lineorder.tbl' OVERWRITE INTO TABLE lineorder;



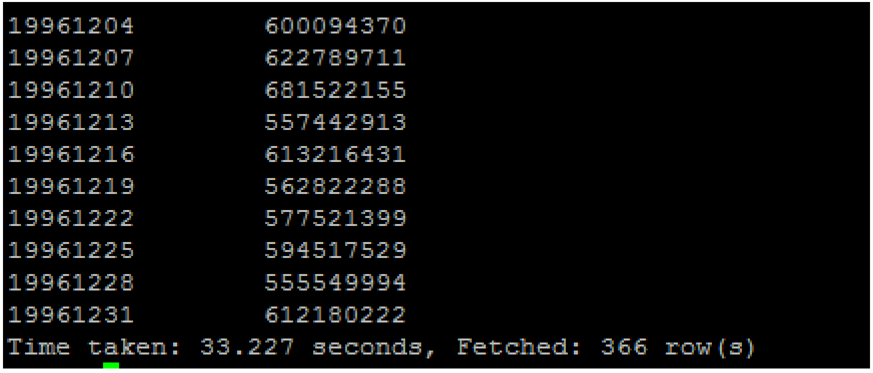
```
hive> LOAD DATA LOCAL INPATH '/home/ec2-user/lineorder.tbl' OVERWRITE INTO TABLE
 lineorder;
Loading data to table default.lineorder
OK
Time taken: 7.721 seconds
```

```
select lo_orderdate, sum(lo_extendedprice) as revenue
from lineorder, dwdate
where lo_orderdate = d_datekey
  and d_year = 1996
```

  and lo_discount between 2 and 4
  and lo_quantity < 25
GROUP BY lo_orderdate;



```
Hadoop job information for Stage-2: number of mappers: 3; number of reducers: 3
2020-10-22 06:02:35,835 Stage-2 map = 0%,  reduce = 0%
2020-10-22 06:02:44,389 Stage-2 map = 33%,  reduce = 0%, Cumulative CPU 3.19 sec
2020-10-22 06:02:45,432 Stage-2 map = 67%,  reduce = 0%, Cumulative CPU 9.39 sec
2020-10-22 06:02:46,463 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 14.96 s
ec
2020-10-22 06:02:50,669 Stage-2 map = 100%,  reduce = 33%, Cumulative CPU 16.6 s
ec
2020-10-22 06:02:52,729 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 19.75
 sec
MapReduce Total cumulative CPU time: 19 seconds 750 msec
Ended Job = job_1603345947385_0001
MapReduce Jobs Launched:
Stage-Stage-2: Map: 3  Reduce: 3   Cumulative CPU: 19.75 sec   HDFS Read: 594384
460 HDFS Write: 6954 SUCCESS
Total MapReduce CPU Time Spent: 19 seconds 750 msec
OK
19960101        616932919
19960104        635233441
19960107        759902355
19960110        602794170
19960113        599375061
19960116        628731716
19960119        562235712
19960122        634381374
19960125        613074565
19960128        588705894
19960131        645881118
19960203        621800600
19960206        635588973
19960209        648975450
19960212        543922372
19960215        706720129
```

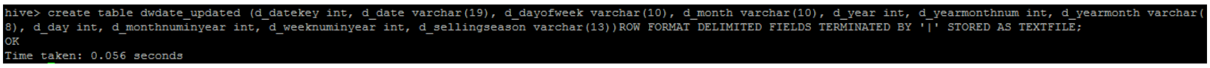


```
19961204        600094370
19961207        622789711
19961210        681522155
19961213        557442913
19961216        613216431
19961219        562822288
19961222        577521399
19961225        594517529
19961228        555549994
19961231        612180222
Time taken: 33.227 seconds, Fetched: 366 row(s)
```

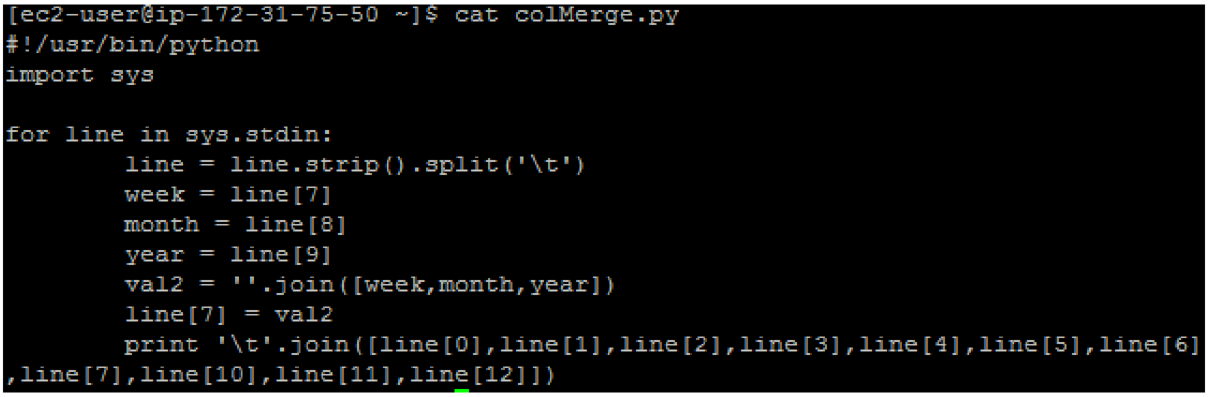
The select query takes ~33 seconds to execute the HiveQL statement and retrieve the result.

Section b)

create table dwdate_updated (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_day int, d_monthnuminyear int, d_weeknuminyear int, d_sellingseason varchar(13))ROW FORMAT DELIMITED FIELDS TERMINATED BY '|' STORED AS TEXTFILE;

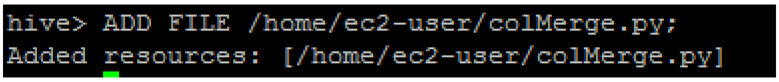


```
hive> create table dwdate_updated (d_datekey int, d_date varchar(19), d_dayofweek varchar(10), d_month varchar(10), d_year int, d_yearmonthnum int, d_yearmonth varchar(
8), d_day int, d_monthnuminyear int, d_weeknuminyear int, d_sellingseason varchar(13))ROW FORMAT DELIMITED FIELDS TERMINATED BY '|' STORED AS TEXTFILE;
OK
Time taken: 0.056 seconds
```



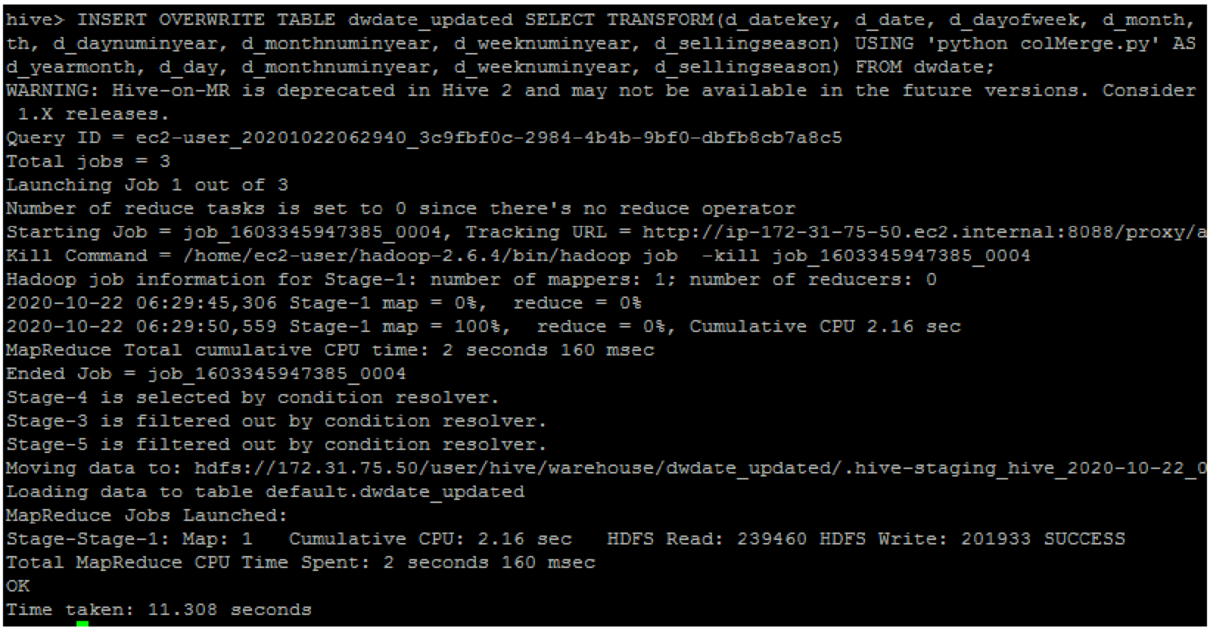
```
[ec2-user@ip-172-31-75-50 ~]$ cat colMerge.py
#!/usr/bin/python
import sys

for line in sys.stdin:
        line = line.strip().split('\t')
        week = line[7]
        month = line[8]
        year = line[9]
        val2 = ''.join([week,month,year])
        line[7] = val2
        print '\t'.join([line[0],line[1],line[2],line[3],line[4],line[5],line[6]
,line[7],line[10],line[11],line[12]])
```

ADD FILE /home/ec2-user/colMerge.py;



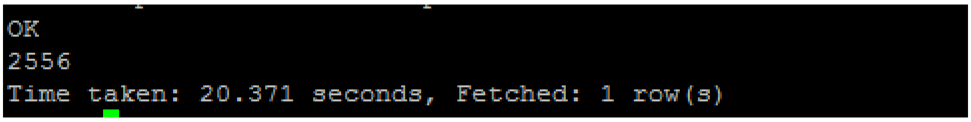
```
hive> ADD FILE /home/ec2-user/colMerge.py;
Added resources: [/home/ec2-user/colMerge.py]
```

INSERT OVERWRITE TABLE dwdate_updated 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) USING 'python colMerge.py' AS (d_datekey, d_date, d_dayofweek, d_month, d_year, d_yearmonthnum, d_yearmonth, d_day, d_monthnuminyear, d_weeknuminyear, d_sellingseason) FROM dwdate;



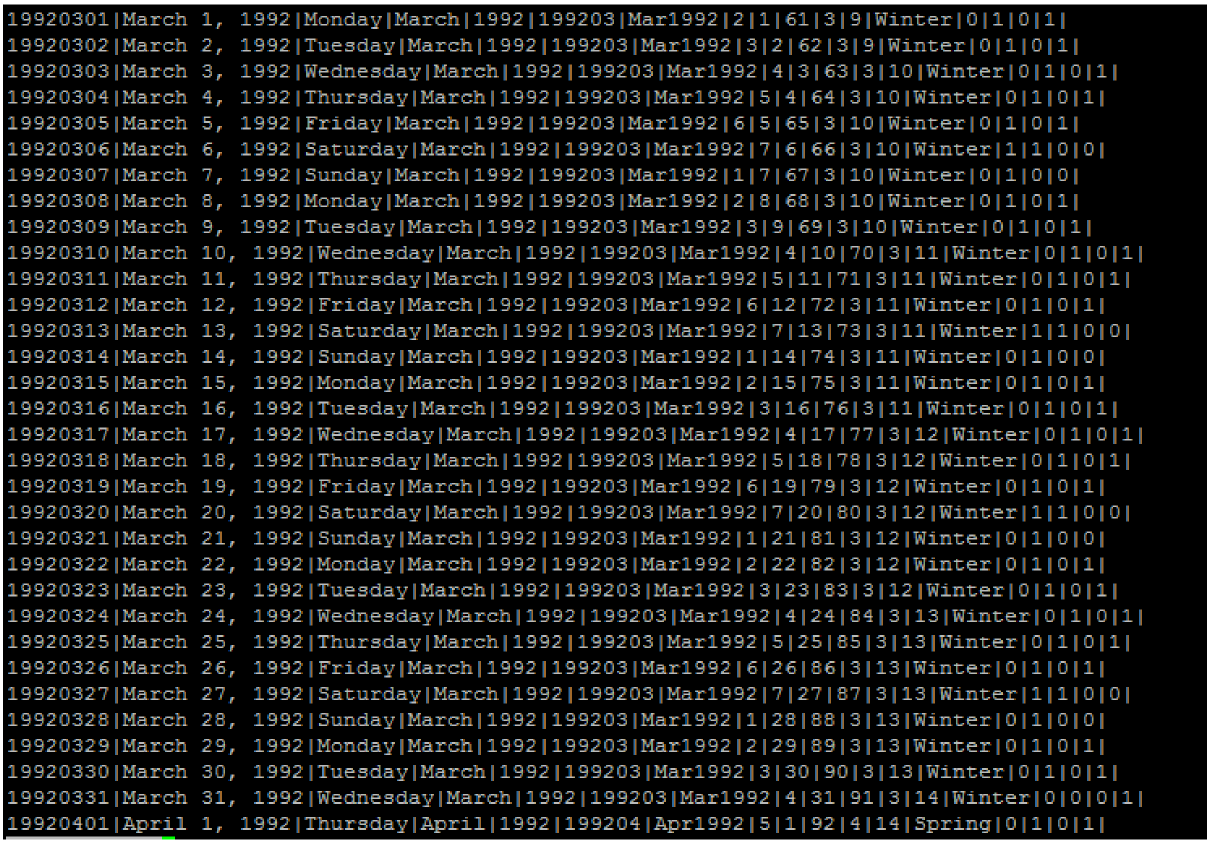
```
hive> INSERT OVERWRITE TABLE dwdate_updated SELECT TRANSFORM(d_datekey, d_date, d_dayofweek, d_month,
th, d_daynuminyear, d_monthnuminyear, d_weeknuminyear, d_sellingseason) USING 'python colMerge.py' AS
d_yearmonth, d_day, d_monthnuminyear, d_weeknuminyear, d_sellingseason) FROM dwdate;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider
 1.X releases.
Query ID = ec2-user_20201022062940_3c9fbf0c-2984-4b4b-9bf0-dbfb8cb7a8c5
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_1603345947385_0004, Tracking URL = http://ip-172-31-75-50.ec2.internal:8088/proxy/a
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job  -kill job_1603345947385_0004
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2020-10-22 06:29:45,306 Stage-1 map = 0%,  reduce = 0%
2020-10-22 06:29:50,559 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.16 sec
MapReduce Total cumulative CPU time: 2 seconds 160 msec
Ended Job = job_1603345947385_0004
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://172.31.75.50/user/hive/warehouse/dwdate_updated/.hive-staging_hive_2020-10-22_0
Loading data to table default.dwdate_updated
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1   Cumulative CPU: 2.16 sec   HDFS Read: 239460 HDFS Write: 201933 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 160 msec
OK
Time taken: 11.308 seconds
```

select count(1) from dwdate_updated;



```
OK
2556
Time taken: 20.371 seconds, Fetched: 1 row(s)
```

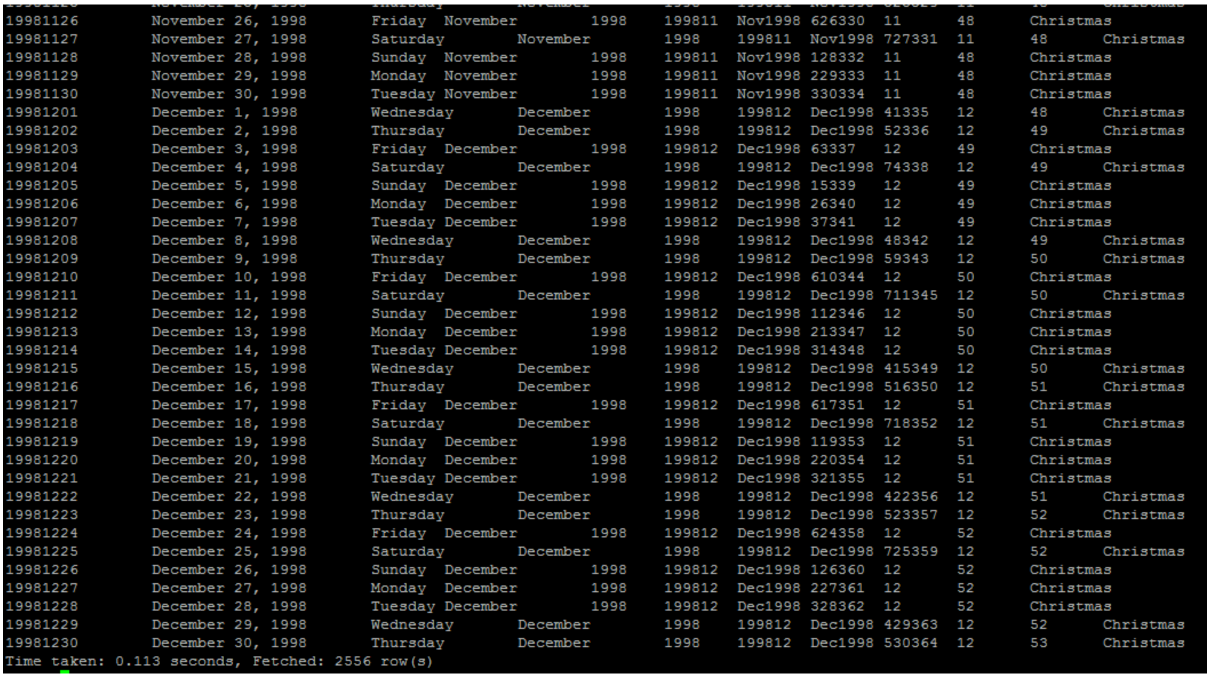
Original Data



```
19920301|March 1, 1992|Monday|March|1992|199203|Mar1992|2|1|61|3|9|Winter|0|1|0|1|
19920302|March 2, 1992|Tuesday|March|1992|199203|Mar1992|3|2|62|3|9|Winter|0|1|0|1|
19920303|March 3, 1992|Wednesday|March|1992|199203|Mar1992|4|3|63|3|10|Winter|0|1|0|1|
19920304|March 4, 1992|Thursday|March|1992|199203|Mar1992|5|4|64|3|10|Winter|0|1|0|1|
19920305|March 5, 1992|Friday|March|1992|199203|Mar1992|6|5|65|3|10|Winter|0|1|0|1|
19920306|March 6, 1992|Saturday|March|1992|199203|Mar1992|7|6|66|3|10|Winter|1|1|0|0|
19920307|March 7, 1992|Sunday|March|1992|199203|Mar1992|1|7|67|3|10|Winter|0|1|0|0|
19920308|March 8, 1992|Monday|March|1992|199203|Mar1992|2|8|68|3|10|Winter|0|1|0|1|
19920309|March 9, 1992|Tuesday|March|1992|199203|Mar1992|3|9|69|3|10|Winter|0|1|0|1|
19920310|March 10, 1992|Wednesday|March|1992|199203|Mar1992|4|10|70|3|11|Winter|0|1|0|1|
19920311|March 11, 1992|Thursday|March|1992|199203|Mar1992|5|11|71|3|11|Winter|0|1|0|1|
19920312|March 12, 1992|Friday|March|1992|199203|Mar1992|6|12|72|3|11|Winter|0|1|0|1|
19920313|March 13, 1992|Saturday|March|1992|199203|Mar1992|7|13|73|3|11|Winter|1|1|0|0|
19920314|March 14, 1992|Sunday|March|1992|199203|Mar1992|1|14|74|3|11|Winter|0|1|0|0|
19920315|March 15, 1992|Monday|March|1992|199203|Mar1992|2|15|75|3|11|Winter|0|1|0|1|
19920316|March 16, 1992|Tuesday|March|1992|199203|Mar1992|3|16|76|3|11|Winter|0|1|0|1|
19920317|March 17, 1992|Wednesday|March|1992|199203|Mar1992|4|17|77|3|12|Winter|0|1|0|1|
19920318|March 18, 1992|Thursday|March|1992|199203|Mar1992|5|18|78|3|12|Winter|0|1|0|1|
19920319|March 19, 1992|Friday|March|1992|199203|Mar1992|6|19|79|3|12|Winter|0|1|0|1|
19920320|March 20, 1992|Saturday|March|1992|199203|Mar1992|7|20|80|3|12|Winter|1|1|0|0|
19920321|March 21, 1992|Sunday|March|1992|199203|Mar1992|1|21|81|3|12|Winter|0|1|0|0|
19920322|March 22, 1992|Monday|March|1992|199203|Mar1992|2|22|82|3|12|Winter|0|1|0|1|
19920323|March 23, 1992|Tuesday|March|1992|199203|Mar1992|3|23|83|3|12|Winter|0|1|0|1|
19920324|March 24, 1992|Wednesday|March|1992|199203|Mar1992|4|24|84|3|13|Winter|0|1|0|1|
19920325|March 25, 1992|Thursday|March|1992|199203|Mar1992|5|25|85|3|13|Winter|0|1|0|1|
19920326|March 26, 1992|Friday|March|1992|199203|Mar1992|6|26|86|3|13|Winter|0|1|0|1|
19920327|March 27, 1992|Saturday|March|1992|199203|Mar1992|7|27|87|3|13|Winter|1|1|0|0|
19920328|March 28, 1992|Sunday|March|1992|199203|Mar1992|1|28|88|3|13|Winter|0|1|0|0|
19920329|March 29, 1992|Monday|March|1992|199203|Mar1992|2|29|89|3|13|Winter|0|1|0|1|
19920330|March 30, 1992|Tuesday|March|1992|199203|Mar1992|3|30|90|3|13|Winter|0|1|0|1|
19920331|March 31, 1992|Wednesday|March|1992|199203|Mar1992|4|31|91|3|14|Winter|0|0|0|1|
19920401|April 1, 1992|Thursday|April|1992|199204|Apr1992|5|1|92|4|14|Spring|0|1|0|1|
```

Transformed Data

select * from dwdate_updated;



```
19981126        November 26, 1998       Friday  November        1998    199811  Nov1998 626330  11      48      Christmas
19981127        November 27, 1998       Saturday        November        1998    199811  Nov1998 727331  11      48      Christmas
19981128        November 28, 1998       Sunday  November        1998    199811  Nov1998 128332  11      48      Christmas
19981129        November 29, 1998       Monday  November        1998    199811  Nov1998 229333  11      48      Christmas
19981130        November 30, 1998       Tuesday November        1998    199811  Nov1998 330334  11      48      Christmas
19981201        December 1, 1998        Wednesday       December        1998    199812  Dec1998 41335   12      48      Christmas
19981202        December 2, 1998        Thursday        December        1998    199812  Dec1998 52336   12      49      Christmas
19981203        December 3, 1998        Friday  December        1998    199812  Dec1998 63337   12      49      Christmas
19981204        December 4, 1998        Saturday        December        1998    199812  Dec1998 74338   12      49      Christmas
19981205        December 5, 1998        Sunday  December        1998    199812  Dec1998 15339   12      49      Christmas
19981206        December 6, 1998        Monday  December        1998    199812  Dec1998 26340   12      49      Christmas
19981207        December 7, 1998        Tuesday December        1998    199812  Dec1998 37341   12      49      Christmas
19981208        December 8, 1998        Wednesday       December        1998    199812  Dec1998 48342   12      49      Christmas
19981209        December 9, 1998        Thursday        December        1998    199812  Dec1998 59343   12      50      Christmas
19981210        December 10, 1998       Friday  December        1998    199812  Dec1998 610344  12      50      Christmas
19981211        December 11, 1998       Saturday        December        1998    199812  Dec1998 711345  12      50      Christmas
19981212        December 12, 1998       Sunday  December        1998    199812  Dec1998 112346  12      50      Christmas
19981213        December 13, 1998       Monday  December        1998    199812  Dec1998 213347  12      50      Christmas
19981214        December 14, 1998       Tuesday December        1998    199812  Dec1998 314348  12      50      Christmas
19981215        December 15, 1998       Wednesday       December        1998    199812  Dec1998 415349  12      50      Christmas
19981216        December 16, 1998       Thursday        December        1998    199812  Dec1998 516350  12      51      Christmas
19981217        December 17, 1998       Friday  December        1998    199812  Dec1998 617351  12      51      Christmas
19981218        December 18, 1998       Saturday        December        1998    199812  Dec1998 718352  12      51      Christmas
19981219        December 19, 1998       Sunday  December        1998    199812  Dec1998 119353  12      51      Christmas
19981220        December 20, 1998       Monday  December        1998    199812  Dec1998 220354  12      51      Christmas
19981221        December 21, 1998       Tuesday December        1998    199812  Dec1998 321355  12      51      Christmas
19981222        December 22, 1998       Wednesday       December        1998    199812  Dec1998 422356  12      51      Christmas
19981223        December 23, 1998       Thursday        December        1998    199812  Dec1998 523357  12      52      Christmas
19981224        December 24, 1998       Friday  December        1998    199812  Dec1998 624358  12      52      Christmas
19981225        December 25, 1998       Saturday        December        1998    199812  Dec1998 725359  12      52      Christmas
19981226        December 26, 1998       Sunday  December        1998    199812  Dec1998 126360  12      52      Christmas
19981227        December 27, 1998       Monday  December        1998    199812  Dec1998 227361  12      52      Christmas
19981228        December 28, 1998       Tuesday December        1998    199812  Dec1998 328362  12      52      Christmas
19981229        December 29, 1998       Wednesday       December        1998    199812  Dec1998 429363  12      52      Christmas
19981230        December 30, 1998       Thursday        December        1998    199812  Dec1998 530364  12      53      Christmas
Time taken: 0.113 seconds, Fetched: 2556 row(s)
```

As can be seen in the above screenshots, the data is transformed and inserted into DWUAT_UPDATED table. The new column DWUAT_UPDATED.D_DAY is created by merging d_daynuminweek, d_daynuminmonth, and d_daynuminyear into a single column.

The final table has 6 columns lesser than the original table.