Problem Statement 1

Find out the top 10 trains which have longest routes.

```
REGISTER 'piggybank.jar';

trains = load 'train_details.csv' USING
org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');


b = foreach trains generate (chararray)$0 as tno,(chararray)$1 as tname,(int)$7 as dist;


grp_tno = group b by tno;


max_dist = foreach grp_tno generate group ,MAX(b.dist);


tno_tname = foreach b generate tno , tname;


dist_tno_tname = DISTINCT tno_tname;


join_rel = join max_dist by $0, dist_tno_tname by $0;


final = order join_rel by $1 desc;


final2 = LIMIT final 10;


result = foreach final2 generate $0 ,$3 ,$1;

dump result;
```

In Line 1: We are registering the *piggybank* jar in order to use the CSVExcelStorage class.

In relation **train**, we are loading the dataset using CSVExcelStorage because of its effective technique to handle double quotes and headers.

In relation **b**, we are generating the columns that are required for processing and explicitly typecasting each of them.

In relation **grp_tno**, we are grouping relation b by “tno.”

In relation **max_dist**, we are generating the grouped column and the maximum distance.

In relation **tno_tname**, we are generating train no and train name from the relation **b**.

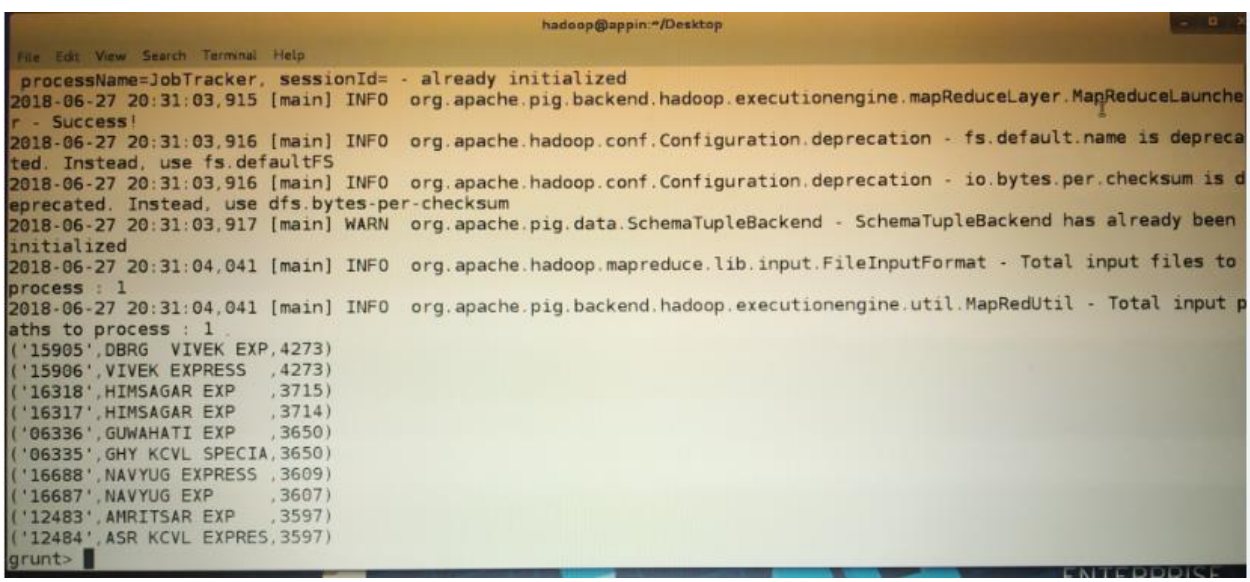
In relation **dist_tno_tname**, we are generating distinct train no . and train name from the previous relation.

In relation **join_rel**, we are joining **max_dist** and **dist_tno_tname** based on a common column, i.e., “tno”

In relation **final**, **final2** is used to order and limit the result to top 10.

In relation **result**, we are arranging the sequence of column.

Finally, using dump, we are printing the result.



Problem Statement 2

Find out the top 10 trains which have max stoppage.

```
REGISTER  'piggybank.jar';

train = load  'train_details.csv'  USING
org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');


b = foreach train generate (chararray)$0 as tno,(chararray)$1 as tname,(int)$2 as islno;


grp_tno = group b by tno;


max_islno = foreach grp_tno generate group ,MAX(b.islno);


tno_tname = foreach b generate tno , tname;


dist_tno_tname = DISTINCT tno_tname;


join_rel = join max_islno by $0, dist_tno_tname by $0;


final = order join_rel by $1 desc;


final2 = LIMIT final 10;


result = foreach final2 generate $0 ,$3 ,$1;

dump result;
```

In Line 1: We are registering the *piggybank* jar in order to use the CSVExcelStorage class.

In relation **train**, we are loading the dataset using CSVExcelStorage because of its effective technique to handle double quotes and headers.

In relation **b**, we are generating the columns that are required for processing and explicitly typecasting each of them.

In relation **grp_tno**, we are grouping relation b by “tno.”

In relation **max_islno**, we are generating the grouped column and the maximum islno.

In relation **tno_tname**, we are generating train no and train name from the relation **b**.

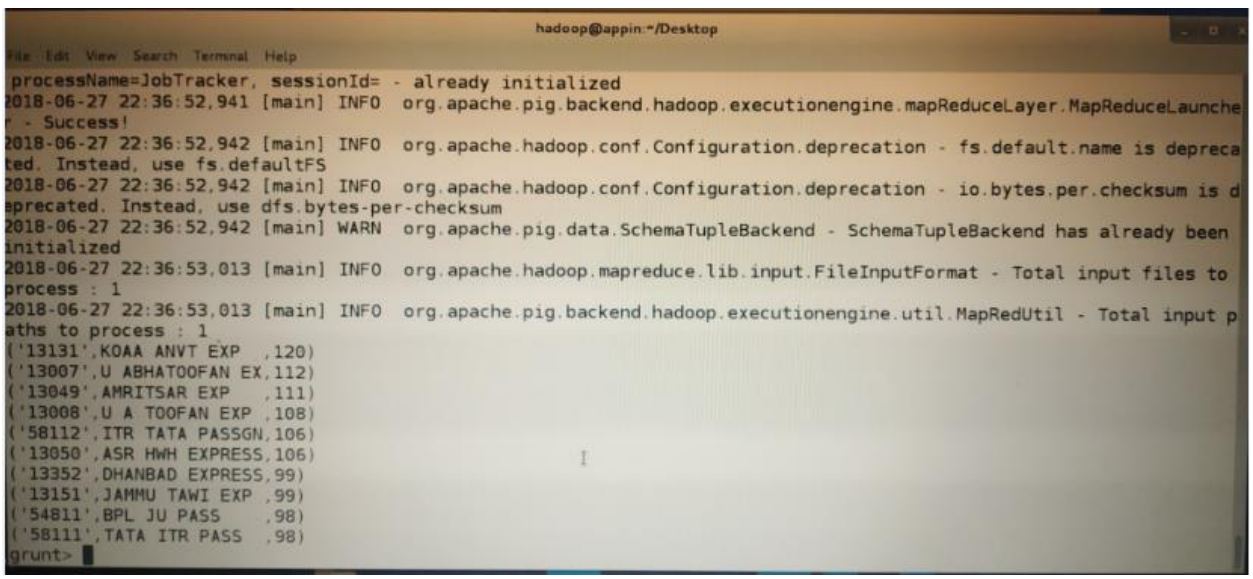
In relation **dist_tno_tname**, we are generating distinct train no . and train name from the previous relation.

In relation **join_rel**, we are joining **max_dist** and **dist_tno_tname** based on a common column, i.e., “tno”

In relation **final** and **final2 ,** is used to order and limit the result to top 10.

In relation **result**, we are arranging the sequence of column.

Finally, using dump, we are printing the result.



Problem Statement 3

Find out the top 10 mostly visited station.

```
REGISTER 'piggybank.jar';

train = load 'train_details.csv' USING
org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');
```

```
b = foreach train generate (chararray)$0 as tno,(chararray)$1 as tname,(chararray)$4 as sname;

c = DISTINCT b;

grpbystation = group c by sname;

d = foreach grpbystation generate group,COUNT(b1.tno);

e = order d by $1 desc;

result = LIMIT e 10;

dump result;
```

In Line 1: We are registering the *piggybank* jar in order to use the CSVExcelStorage class.

In relation **train**, we are loading the dataset using CSVExcelStorage because of its effective technique to handle double quotes and headers.

In relation **b**, we are generating the columns that are required for processing and explicitly typecasting each of them.

In relation **c**, we are generating DISTINCT relation **b**.

In relation **grpbystation**, we are grouping relation b by "tno."

In relation **d**, we are generating the grouped column and counting the train no.

In relation **e** & **result** is for ordering and finding the top 10 trains.

Finally, using dump, we are printing the result.

```
hadoop@appin:~/Desktop
File  Edit  View  Search  Terminal  Help
 processName=JobTracker, sessionId= - already initialized
2018-06-29 00:59:57,034 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLaunche
r - Success!
2018-06-29 00:59:57,034 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is depreca
ted. Instead, use fs.defaultFS
2018-06-29 00:59:57,035 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is d
eprecated. Instead, use dfs.bytes-per-checksum
2018-06-29 00:59:57,035 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been
initialized
2018-06-29 00:59:57,137 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to
process : 1
2018-06-29 00:59:57,137 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input p
aths to process : 1
(VIJAYAWADA JN   ,313)
(VADODARA JN     ,297)
(KANPUR CENTRAL  ,283)
(SURAT           ,265)
(ITARSI JN       ,261)
(AHMEDABAD JN    ,255)
(KALYAN JN       ,254)
(BHUSAVAL JN     ,246)
(NAGPUR          ,243)
(NEW DELHI       ,239)
grunt>
```

Problem Statement 4

Find out the top 10 visited destination station.

```
REGISTER  'piggybank.jar';

train = load  'train_details.csv'  USING
org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');


b = foreach train generate (chararray)$0 as tno,(chararray)$1 as tname,(chararray)$11 as dest;


c = DISTINCT b;


grpbydest = group c by dest;


d = foreach grpbydest generate group,COUNT(b1.tno);


e = order d by $1 desc;


result = LIMIT e 10;

dump result;
```

In Line 1: We are registering the *piggybank* jar in order to use the CSVExcelStorage class.

In relation **train**, we are loading the dataset using CSVExcelStorage because of its effective technique to handle double quotes and headers.

In relation **b**, we are generating the columns that are required for processing and explicitly typecasting each of them.

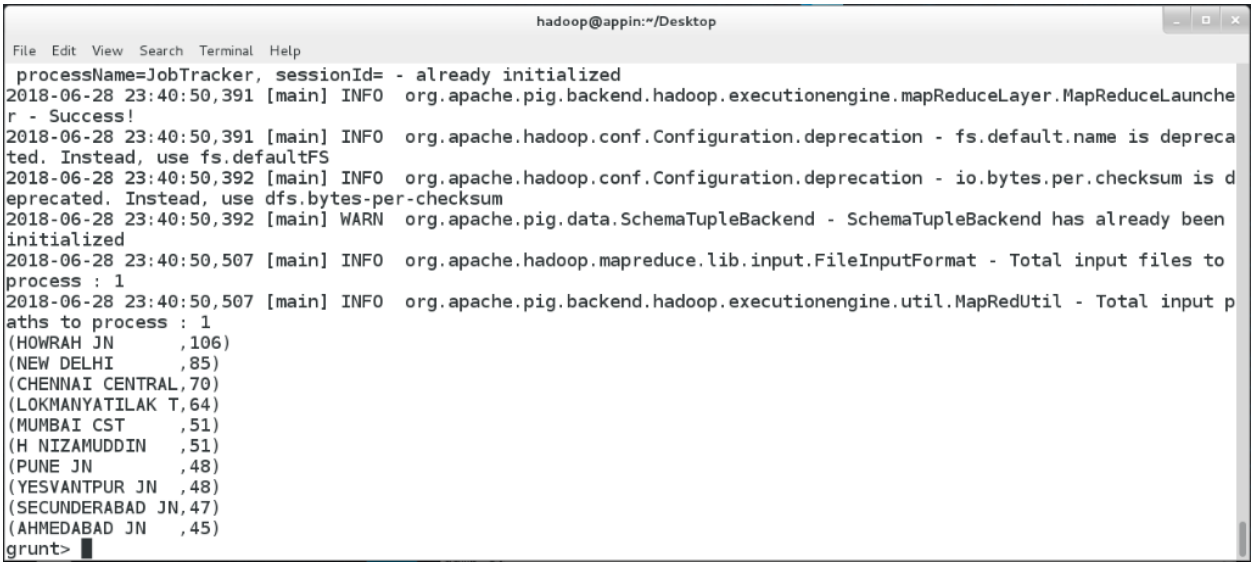
In relation **c**, we are generating the grouped column and counting the train no.

In relation **grpbydest**, we are grouping relation b by "dest" i.e. destination.

In relation **d,** we are generating DISTINCT relation **b**.

In relation **e** and **result** is for ordering and finding the top 10 trains.

Finally, using dump, we are printing the result.



```
hadoop@appin:~/Desktop
File  Edit  View  Search  Terminal  Help
 processName=JobTracker, sessionId= - already initialized
2018-06-28 23:40:50,391 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLaunche
r - Success!
2018-06-28 23:40:50,391 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is depreca
ted. Instead, use fs.defaultFS
2018-06-28 23:40:50,392 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is d
eprecated. Instead, use dfs.bytes-per-checksum
2018-06-28 23:40:50,392 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been
initialized
2018-06-28 23:40:50,507 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to
process : 1
2018-06-28 23:40:50,507 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input p
aths to process : 1
(HOWRAH JN      ,106)
(NEW DELHI      ,85)
(CHENNAI CENTRAL,70)
(LOKMANYATILAK T,64)
(MUMBAI CST     ,51)
(H NIZAMUDDIN   ,51)
(PUNE JN        ,48)
(YESVANTPUR JN  ,48)
(SECUNDERABAD JN,47)
(AHMEDABAD JN   ,45)
grunt>
```

Problem Statement 5

Find out the top 5 states of India which have max growth in national highway.

```
REGISTER  'piggybank.jar';

hways = load  'national_highway.csv'  USING
org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');
```

```
b = foreach hways generate (chararray)$0 as sname,(int)$1 as y1,(int)$4 as y4;

c = foreach b generate sname , (y4-y1) as growth;

remove_row = filter c by sname != 'All INDIA';

d = order remove_row by growth desc;

final = foreach d generate $0 , growth;

result = LIMIT final 5;

dump result;
```

In Line 1: We are registering the *piggybank* jar in order to use the CSVExcelStorage class.

In relation **hways**, we are loading the dataset using CSVExcelStorage because of its effective technique to handle double quotes and headers.

In relation **b**, we are generating the columns that are required for processing and explicitly typecasting each of them.

In relation **c**, we are generating the state name & difference of growth between the year 2009 & 2012.

In relation **remove_row**, we are filtering the data based on State name i.e., sname != 'ALL INDIA'.

In relation **d**, we are ordering growth in descending order.

In relation **final**, we are generating State name & growth.

In relation **result**, finding the top 5 States.

Finally, using dump, we are printing the result.

```
hadoop@appin:~/Desktop
File  Edit  View  Search  Terminal  Help
2018-06-28 22:54:04,016 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with process
Name=JobTracker, sessionId= - already initialized
2018-06-28 22:54:04,018 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with process
Name=JobTracker, sessionId= - already initialized
2018-06-28 22:54:04,019 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with process
Name=JobTracker, sessionId= - already initialized
2018-06-28 22:54:04,022 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Succ
ess!
2018-06-28 22:54:04,022 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Ins
tead, use fs.defaultFS
2018-06-28 22:54:04,023 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecate
d. Instead, use dfs.bytes-per-checksum
2018-06-28 22:54:04,023 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initiali
zed
2018-06-28 22:54:04,118 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process
: 1
2018-06-28 22:54:04,118 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to
process : 1
(Rajasthan,1545)
(Uttar Pradesh,1044)
(Gujarat,787)
(Bihar,463)
(Madhya Pradesh,394)
grunt>
```