

EXPERIMENT 7

Name	Shreya Shetty
UID	20191410059
Batch	A
Class	TE IT
Subject	BDA

AIM: Installation and commands on Pig

THEORY:

Apache Pig is a tool/platform for creating and executing Map Reduce programs used with Hadoop. It is a tool/platform for analyzing large sets of data. You can say, Apache Pig is an abstraction over MapReduce. Programmers who are not so good at Java used to struggle working on Hadoop, majorly while writing MapReduce jobs. Apache Pig has its own language Pig Latin which is a boon for poor programmers.

The high-level procedural language used in Apache Pig platform is called Pig Latin. Apache Pig features 'Pig Latin' which is a relatively simpler language which can run over distributed datasets on Hadoop File System (HDFS). In Apache Pig, you need to write Pig scripts using Pig Latin language, which gets converted to MapReduce job when you run your Pig script. Apache Pig has various operators which are used to perform the tasks like reading, writing, processing the data.

Apache Pig Installation on Linux:

Below are the steps for Apache Pig Installation on Linux

Step 1 : Download Pig tar file.

Command: `wget http://www-us.apache.org/dist/pig/pig-0.16.0/pig-0.16.0.tar.gz`

```
edureka@localhost:~$ wget http://www-us.apache.org/dist/pig/latest/pig-0.16.0.tar.gz
--2016-11-18 17:46:31-- http://www-us.apache.org/dist/pig/latest/pig-0.16.0.tar.gz
Resolving www-us.apache.org (www-us.apache.org)... 140.211.11.105
Connecting to www-us.apache.org (www-us.apache.org)|140.211.11.105|:80..
. connected.
HTTP request sent, awaiting response... 200 OK
Length: 177279333 (169M) [application/x-gzip]
Saving to: 'pig-0.16.0.tar.gz'

pig-0.16.0.tar.gz  2%[          ]  4.80M  149KB/s  eta 9m 0s
```

Step 2 : Extract the tar file using tar command. In below tar command, x means extract an archive file, z means filter an archive through gzip, f means filename of an archive file.

Command: tar -xzf pig-0.16.0.tar.gz Command: ls

Step 3: Edit the “.bashrc” file to update the environment variables of Apache Pig. We are setting it so that we can access pig from any directory , we need not go to pig directory to execute pig commands. Also, if any other application is looking for Pig, it will get to know the path of Apache Pig from this file.

Command: sudo gedit .bashrc

Add the following at the end of the file:

```
# Set PIG_HOME
export PIG_HOME=/home/edureka/pig-0.16.0 export PATH=$PATH:/home/edureka/pig-0.16.0/bin export PIG_CLASSPATH=$HADOOP_CONF_DIR
```

Also, make sure that hadoop path is also set.

Run below command to make the changes get updated in same terminal. Command: source .bashrc

Step 4 : Check pig version. This is to test that Apache Pig got installed correctly . In case, you don't get the Apache Pig version, you need to verify if you have followed the above steps correctly.

Command: pig -version

```
[cloudera@quickstart ~]$ pig -version
log4j:WARN No appenders could be found for logger (org.apache.hadoop.util.Shell)
.
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more in
fo.
Apache Pig version 0.12.0-cdh5.4.2 (rexported)
compiled May 19 2015, 17:03:41
```

Step 5 : Check pig help to see all the pig command options.

Command: pig -help

```
[cloudera@quickstart ~]$ pig -help
log4j:WARN No appenders could be found for logger (org.apache.hadoop.util.Shell)
.
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more in
fo.

Apache Pig version 0.12.0-cdh5.4.2 (rexported)
compiled May 19 2015, 17:03:41

USAGE: Pig [options] [-] : Run interactively in grunt shell.
      Pig [options] -e[execute] cmd [cmd ...] : Run cmd(s).
      Pig [options] [-f[file]] file : Run cmds found in file.
options include:
  -4, -log4jconf - Log4j configuration file, overrides log conf
  -b, -brief - Brief logging (no timestamps)
  -c, -check - Syntax check
  -d, -debug - Debug level, INFO is default
  -e, -execute - Commands to execute (within quotes)
  -f, -file - Path to the script to execute
  -g, -embedded - ScriptEngine classname or keyword for the ScriptEngine
  -h, -help - Display this message. You can specify topic to get help for that
topic.
      properties is the only topic currently supported: -h properties.
```

Step 6: Run Pig to start the grunt shell. Grunt shell is used to run Pig Latin scripts.

Execution modes in Apache Pig:

MapReduce Mode – This is the default mode, which requires access to a Hadoop cluster and HDFS installation. Since, this is a default mode, it is not necessary to specify -x flag (you can execute pig OR pig -x mapreduce). The input and output in this mode are present on HDFS.

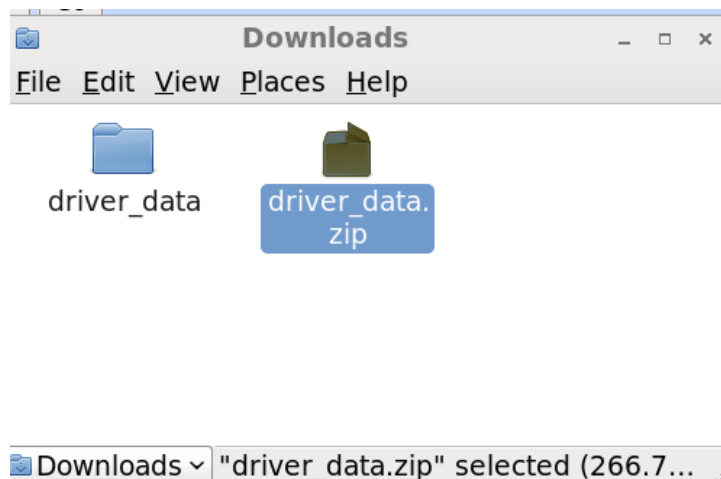
Local Mode – With access to a single machine, all files are installed and run using a local host and file system. Here the local mode is specified using ‘-x flag’ (pig -x local). The input and output in this mode are present on local file system.

Command: pig -x local

```
[cloudera@quickstart ~]$ pig -x local
log4j:WARN No appenders could be found for logger (org.apache.hadoop.util.Shell).
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more info
.
2022-03-30 21:18:14,986 [main] INFO  org.apache.pig.Main - Apache Pig version 0.12
.0-cdh5.4.2 (rexported) compiled May 19 2015, 17:03:41
2022-03-30 21:18:14,987 [main] INFO  org.apache.pig.Main - Logging error messages
to: /home/cloudera/pig_1648700294959.log
2022-03-30 21:18:15,016 [main] INFO  org.apache.pig.impl.util.Utils - Default boot
up file /home/cloudera/.pigbootup not found
2022-03-30 21:18:15,382 [main] INFO  org.apache.hadoop.conf.Configuration.deprecate
ion - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-03-30 21:18:15,382 [main] INFO  org.apache.hadoop.conf.Configuration.deprecate
ion - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-03-30 21:18:15,384 [main] INFO  org.apache.pig.backend.hadoop.executionengine
.HExecutionEngine - Connecting to hadoop file system at: file:///
2022-03-30 21:18:15,991 [main] INFO  org.apache.hadoop.conf.Configuration.deprecate
ion - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2022-03-30 21:18:15,996 [main] INFO  org.apache.hadoop.conf.Configuration.deprecate
ion - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-03-30 21:18:15,996 [main] INFO  org.apache.hadoop.conf.Configuration.deprecate
ion - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-03-30 21:18:16,118 [main] INFO  org.apache.hadoop.conf.Configuration.deprecate
```

COMMANDS:

Step 1 : Download data from site, extract and add the data files to the newly created directory



Creating a new directory:

```
[cloudera@quickstart ~]$ hdfs dfs -mkdir /user/shreya_pig
```

Copying data files from Downloads:

```
[cloudera@quickstart ~]$ hdfs dfs -copyFromLocal /home/cloudera/Downloads/driver_data /user/shreya_pig
```

Listing contents in directory:

```
[cloudera@quickstart ~]$ hdfs dfs -ls /user/shreya_pig
Found 1 items
drwxr-xr-x - cloudera supergroup 0 2022-04-06 15:25 /user/shreya_pig/
driver_data
```

Step 2 : Create Your Script

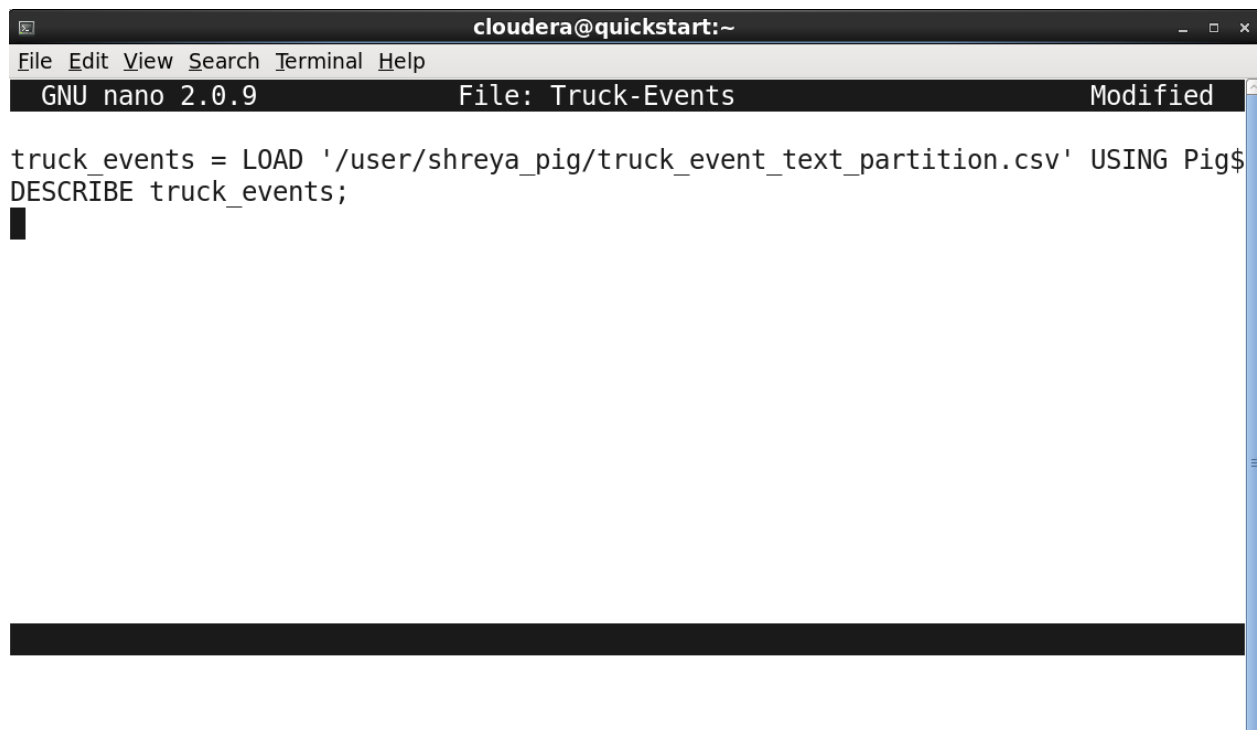
Open the Pig interface by clicking the Pig View in the views menu. On the left we can choose between our saved Pig Scripts, UDFs and the Pig Jobs executed in the past. To the right of this menu bar we see our saved Pig Scripts.

Click on the button "New Script", enter "Truck-Events" for the title of your script and leave the location path empty .

```
[cloudera@quickstart ~]$ nano Truck-Events
```

Step 3 : Define a relation

```
truck_events = LOAD '/user/maria_dev/truck_event_text_partition.csv' USING PigStorage(',');
DESCRIBE truck_events;
```



```
cloudera@quickstart:~
File Edit View Search Terminal Help
GNU nano 2.0.9 File: Truck-Events Modified

truck_events = LOAD '/user/shreya_pig/truck_event_text_partition.csv' USING PigStorage(',');
DESCRIBE truck_events;
█
```

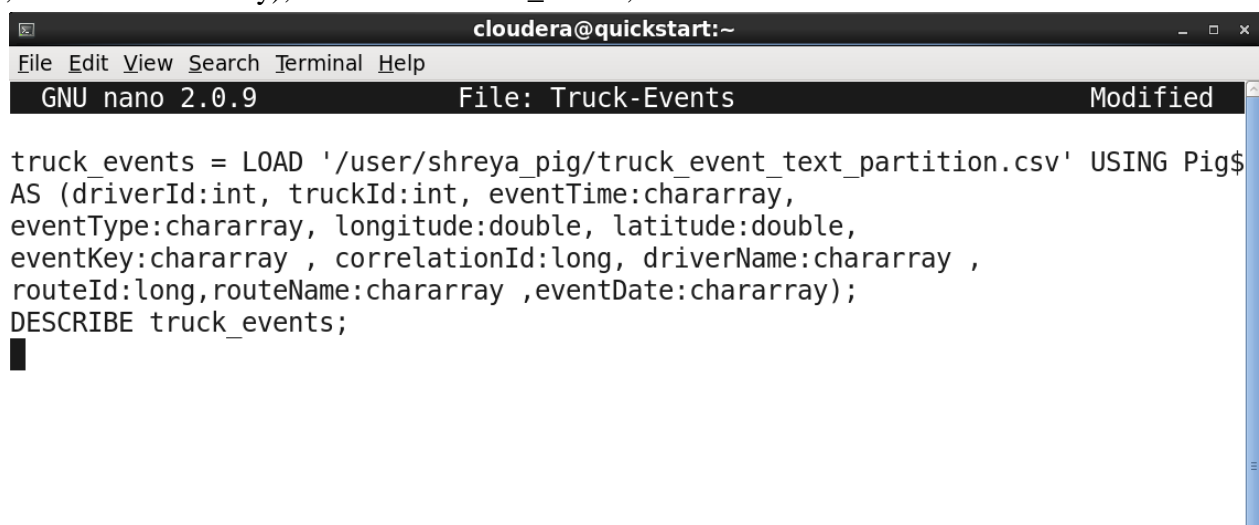
Step 4 : Save and Execute the Script

Click the Save button to save your changes to the script. Click Execute to run the script.

```
[cloudera@quickstart ~]$ pig -f Truck-Events
log4j:WARN No appenders could be found for logger (org.apache.hadoop.util.Shell)
.
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more info.
2022-04-06 15:43:41,064 [main] INFO org.apache.pig.Main - Apache Pig version 0.12.0-cdh5.4.2 (rexported) compiled May 19 2015, 17:03:41
2022-04-06 15:43:41,066 [main] INFO org.apache.pig.Main - Logging error messages to: /home/cloudera/pig_1649285020975.log
2022-04-06 15:43:47,335 [main] INFO org.apache.pig.impl.util.Utils - Default bootstrap file /home/cloudera/.pigbootstrap not found
2022-04-06 15:43:49,755 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-04-06 15:43:49,759 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-04-06 15:43:49,773 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://quickstart.cloudera:8020
2022-04-06 15:43:54,906 [main] INFO org.apache.pig.backend.hadoop.executionengine
```

Step 5: Define a Relation with a Schema

```
truck_events = LOAD '/user/maria_dev/truck_event_text_partition.csv' USING PigStorage(',')
AS (driverId:int, truckId:int, eventTime:chararray,
eventTypes:chararray, longitude:double, latitude:double, eventKey:chararray ,
correlationId:long, driverName:chararray , routeId:long,routeName:chararray
,eventDate:chararray); DESCRIBE truck_events;
```

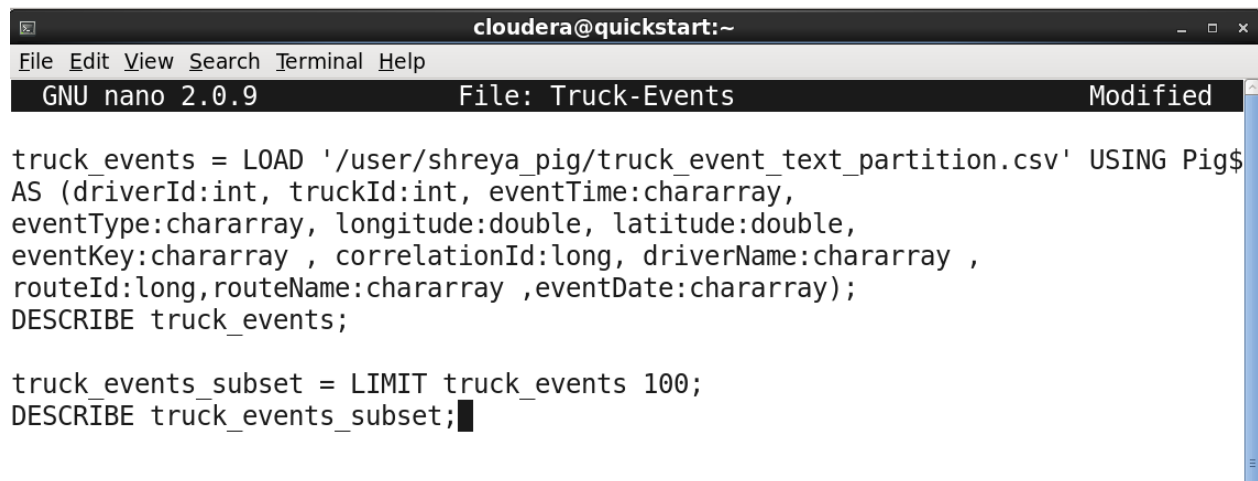
A screenshot of a terminal window titled 'cloudera@quickstart:~'. The terminal shows the GNU nano 2.0.9 editor with the file 'Truck-Events' open. The editor displays the Pig script from the previous block. The terminal window has a menu bar with 'File', 'Edit', 'View', 'Search', 'Terminal', and 'Help'. The status bar at the bottom shows 'GNU nano 2.0.9', 'File: Truck-Events', and 'Modified'. The script content is as follows:

```
truck_events = LOAD '/user/shreya_pig/truck_event_text_partition.csv' USING PigStorage(',')
AS (driverId:int, truckId:int, eventTime:chararray,
eventTypes:chararray, longitude:double, latitude:double,
eventKey:chararray , correlationId:long, driverName:chararray ,
routeId:long,routeName:chararray ,eventDate:chararray);
DESCRIBE truck_events;
```

```
[cloudera@quickstart ~]$ pig -f Truck-Events
log4j:WARN No appenders could be found for logger (org.apache.hadoop.util.Shell)
.
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more info.
2022-04-06 15:54:11,935 [main] INFO org.apache.pig.Main - Apache Pig version 0.12.0-cdh5.4.2 (rexported) compiled May 19 2015, 17:03:41
2022-04-06 15:54:11,936 [main] INFO org.apache.pig.Main - Logging error messages to: /home/cloudera/pig_1649285651914.log
2022-04-06 15:54:13,623 [main] INFO org.apache.pig.impl.util.Utils - Default bootstrap file /home/cloudera/.pigbootstrap not found
2022-04-06 15:54:14,264 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-04-06 15:54:14,264 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-04-06 15:54:14,272 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://quickstart.cloudera:8020
2022-04-06 15:54:15,657 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to map-reduce job tracker at: localhost:8021
2022-04-06 15:54:15,740 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
truck_events: {driverId: int,truckId: int,eventTime: chararray,eventType: chararray,longitude: double,latitude: double,eventKey: chararray,correlationId: long,driverName: chararray,routeId: long,routeName: chararray,eventDate: chararray}
```

Step 6 : Define a new relation from an existing relation `truck_events_subset = LIMIT`

`truck_events 100; DESCRIBE truck_events_subset`



```
cloudera@quickstart:~
File Edit View Search Terminal Help
GNU nano 2.0.9 File: Truck-Events Modified

truck_events = LOAD '/user/shreya_pig/truck_event_text_partition.csv' USING Pig$
AS (driverId:int, truckId:int, eventTime:chararray,
eventTime:chararray, longitude:double, latitude:double,
eventKey:chararray , correlationId:long, driverName:chararray ,
routeId:long,routeName:chararray ,eventDate:chararray);
DESCRIBE truck_events;

truck_events_subset = LIMIT truck_events 100;
DESCRIBE truck_events_subset;
```



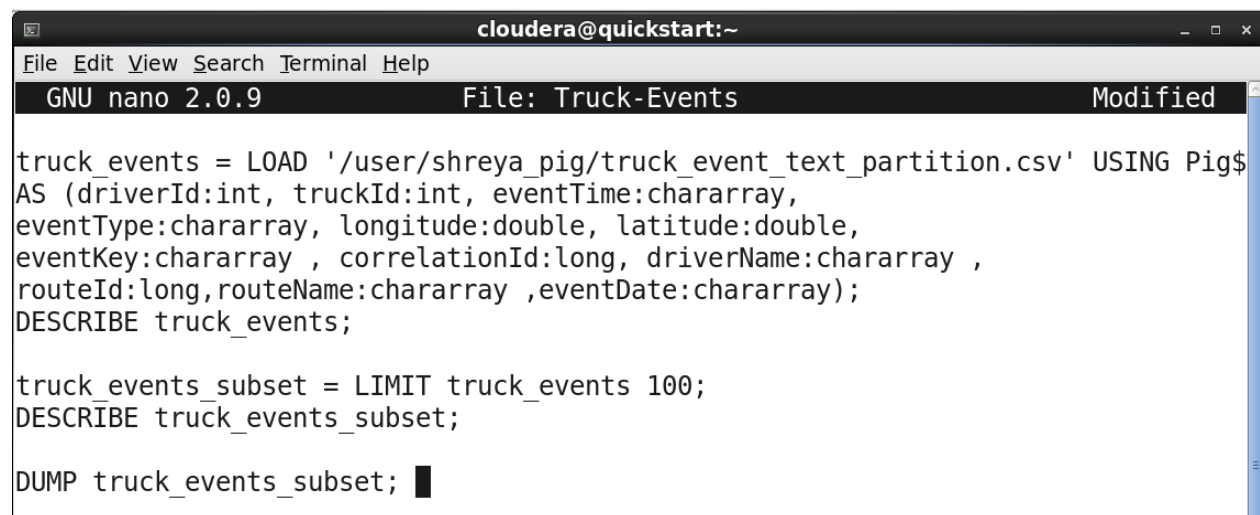
```

[cloudera@quickstart ~]$ pig -f Truck-Events
log4j:WARN No appenders could be found for logger (org.apache.hadoop.util.Shell)
.
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more in
fo.
2022-04-06 15:57:32,280 [main] INFO  org.apache.pig.Main - Apache Pig version 0.
12.0-cdh5.4.2 (rexported) compiled May 19 2015, 17:03:41
2022-04-06 15:57:32,281 [main] INFO  org.apache.pig.Main - Logging error message
s to: /home/cloudera/pig_1649285852251.log
2022-04-06 15:57:33,900 [main] INFO  org.apache.pig.impl.util.Utils - Default bo
otup file /home/cloudera/.pigbootup not found
2022-04-06 15:57:34,602 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
2022-04-06 15:57:34,602 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-04-06 15:57:34,602 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.HExecutionEngine - Connecting to hadoop file system at: hdfs://quickstart.clo
udera:8020
2022-04-06 15:57:36,031 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.HExecutionEngine - Connecting to map-reduce job tracker at: localhost:8021
truck_events: {driverId: int,truckId: int,eventTime: chararray,eventType: charar
ray,longitude: double,latitude: double,eventKey: chararray,correlationId: long,d
riverName: chararray,routeId: long,routeName: chararray,eventDate: chararray}
truck_events_subset: {driverId: int,truckId: int,eventTime: chararray,eventType:
chararray,longitude: double,latitude: double,eventKey: chararray,correlationId:
long,driverName: chararray,routeId: long,routeName: chararray,eventDate: charar
ray}
-

```

Step 7 : View the Data

To view the data of a relation, use the DUMP command. DUMP truck_events_subset;



```

cloudera@quickstart:~
File Edit View Search Terminal Help
GNU nano 2.0.9 File: Truck-Events Modified
truck_events = LOAD '/user/shreya_pig/truck_event_text_partition.csv' USING Pig$
AS (driverId:int, truckId:int, eventTime:chararray,
eventType:chararray, longitude:double, latitude:double,
eventKey:chararray , correlationId:long, driverName:chararray ,
routeId:long,routeName:chararray ,eventDate:chararray);
DESCRIBE truck_events;

truck_events_subset = LIMIT truck_events 100;
DESCRIBE truck_events_subset;

DUMP truck_events_subset; █

```



```
Output(s):
Successfully stored 100 records (14966 bytes) in: "hdfs://quickstart.cloudera:8020/tmp/temp175069
2286/tmp1397866081"
```

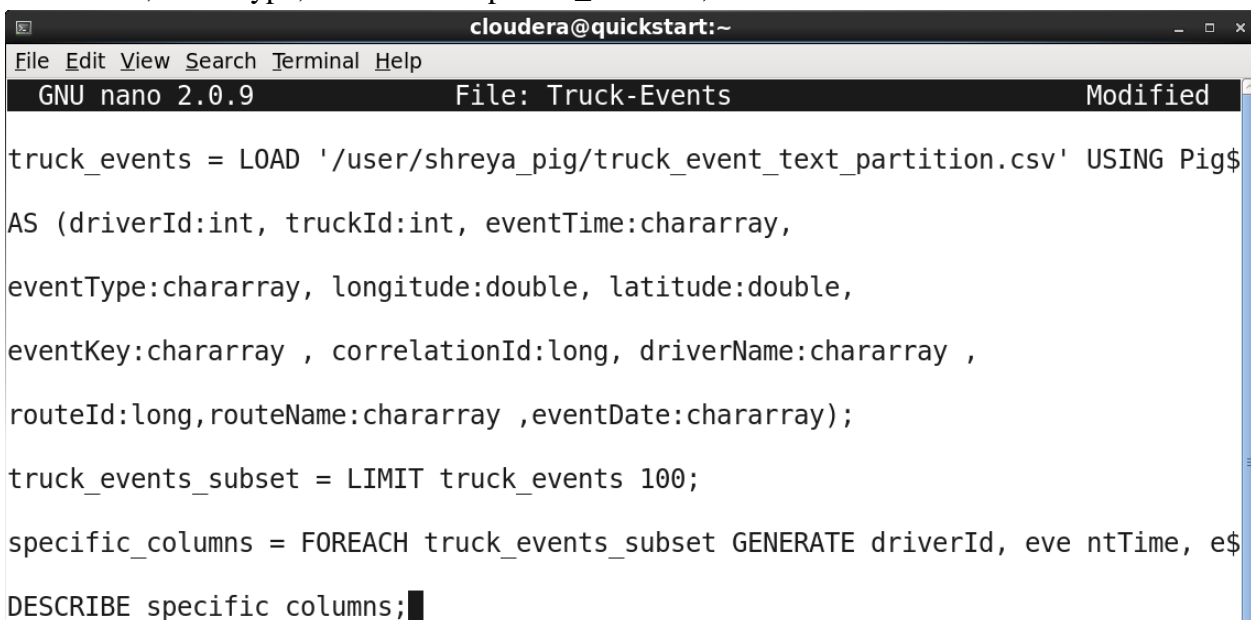
```
Counters:
Total records written : 100
Total bytes written : 14966
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0
```

```
Job DAG:
job_1649242432845_0007 ->      job_1649242432845_0008,
job_1649242432845_0008
```

```
(32,42,59:22.5,Normal,-90.37,35.21,32|42|9223370572464813296,3660000000000000000,Ryan Templeton,1
090292248,Peoria to Cedar Rapids Route 2,2016-05-27-22)
(32,42,59:24.2,Normal,-90.94,35.03,32|42|9223370572464811596,3660000000000000000,Ryan Templeton,1
090292248,Peoria to Cedar Rapids Route 2,2016-05-27-22)
(32,42,59:31.2,Normal,-92.09,34.8,32|42|9223370572464804645,3660000000000000000,Ryan Templeton,10
90292248,Peoria to Cedar Rapids Route 2,2016-05-27-22)
(32,42,59:32.1,Normal,-91.93,34.81,32|42|9223370572464803676,3660000000000000000,Ryan Templeton,1
090292248,Peoria to Cedar Rapids Route 2,2016-05-27-22)
(32,42,59:33.7,Normal,-91.38,34.83,32|42|9223370572464802114,3660000000000000000,Ryan Templeton,1
090292248,Peoria to Cedar Rapids Route 2,2016-05-27-22)
(,eventTime,eventType,,,eventKey,,driverName,,routeName,eventDate)
```

Step 8 : Select specific columns from a relation One of the key uses of Pig is data transformation.

```
truck_events = LOAD '/user/maria_dev/truck_event_text_partition.csv' USING PigStorage(',')
AS (driverId:int, truckId:int, eventTime:chararray, eventType:chararray, longitude:double,
latitude:double, eventKey:chararray , correlationId:long, driverName:chararray ,
routeId:long,routeName:chararray ,eventDate:chararray); truck_events_subset = LIMIT
truck_events 100; specific_columns = FOREACH truck_events_subset GENERATE driverId, eventTime, eventKey,
eventType; DESCRIBE specific_columns;
```



The screenshot shows a terminal window titled "cloudera@quickstart:~" with a nano 2.0.9 editor open. The file being edited is "Truck-Events". The content of the file is the Pig Latin script from the previous block, with the cursor at the end of the "DESCRIBE specific_columns;" line.

```
cloudera@quickstart:~
File Edit View Search Terminal Help
GNU nano 2.0.9 File: Truck-Events Modified
truck_events = LOAD '/user/shreya_pig/truck_event_text_partition.csv' USING Pig$
AS (driverId:int, truckId:int, eventTime:chararray,
eventType:chararray, longitude:double, latitude:double,
eventKey:chararray , correlationId:long, driverName:chararray ,
routeId:long,routeName:chararray ,eventDate:chararray);
truck_events_subset = LIMIT truck_events 100;
specific_columns = FOREACH truck_events_subset GENERATE driverId, eventTime, e$
DESCRIBE specific_columns;
```

```
[cloudera@quickstart ~]$ pig -f Truck-Events
log4j:WARN No appenders could be found for logger (org.apache.hadoop.util.Shell)
.
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more info.
2022-04-06 16:04:37,442 [main] INFO org.apache.pig.Main - Apache Pig version 0.12.0-cdh5.4.2 (rexported) compiled May 19 2015, 17:03:41
2022-04-06 16:04:37,443 [main] INFO org.apache.pig.Main - Logging error messages to: /home/cloudera/pig_1649286277421.log
2022-04-06 16:04:38,953 [main] INFO org.apache.pig.impl.util.Utils - Default bootstrap file /home/cloudera/.pigbootstrap not found
2022-04-06 16:04:39,553 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-04-06 16:04:39,553 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-04-06 16:04:39,557 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://quickstart.cloudera:8020
2022-04-06 16:04:40,401 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to map-reduce job tracker at: localhost:8021
2022-04-06 16:04:40,401 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-04-06 16:04:40,869 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-04-06 16:04:40,871 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-04-06 16:04:41,030 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-04-06 16:04:41,031 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-04-06 16:04:41,136 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-04-06 16:04:41,141 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-04-06 16:04:41,232 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-04-06 16:04:41,232 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
specific_columns: {driverId: int,eventTime: chararray,eventType: chararray}
```

Step 9 : Store relationship data into a HDFS File

```
STORE specific_columns INTO 'output/specific_columns' USING PigStorage(',');
```

```
Store_truck_events_subset_specific
```

```
cloudera@quickstart:~  
File Edit View Search Terminal Help  
GNU nano 2.0.9 File: Truck-Events Modified  
truck_events = LOAD '/user/shreya_pig/truck_event_text_partition.csv' USING Pig$  
AS (driverId:int, truckId:int, eventTime:chararray,  
eventType:chararray, longitude:double, latitude:double,  
eventKey:chararray , correlationId:long, driverName:chararray ,  
routeId:long,routeName:chararray ,eventDate:chararray);  
truck_events_subset = LIMIT truck_events 100;  
specific_columns = FOREACH truck_events_subset GENERATE driverId, eventTime, ev$  
DESCRIBE specific_columns;  
STORE specific_columns INTO 'output/specific_columns' USING PigStorage( ',');  
Store_truck_events_subset_specific  
^G Get Help ^O WriteOut ^R Read File ^Y Prev Page ^K Cut Text ^C Cur Pos  
^X Exit ^J Justify ^W Where Is ^V Next Page ^U UnCut Text ^T To Spell
```

Output(s):
Successfully stored 100 records (1803 bytes) in: "hdfs://quickstart.cloudera:8020/user/cloudera/output/specific_columns"

Counters:
Total records written : 100
Total bytes written : 1803
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1649242432845_0005 -> job_1649242432845_0006,
job_1649242432845_0006

Step 10 : Perform a join between 2 relations

```
truck_events = LOAD '/user/maria_dev/truck_event_text_partition.csv' USING PigStorage(',')  
AS (driverId:int, truckId:int, eventTime:chararray,  
eventType:chararray, longitude:double, latitude:double, eventKey:chararray ,  
correlationId:long, driverName:chararray , routeId:long,routeName:chararray  
,eventDate:chararray);  
drivers = LOAD '/user/maria_dev/drivers.csv' USING PigStorage(',') AS (driverId:int,  
name:chararray , ssn:chararray,  
location:chararray, certified:chararray, wage_plan:chararray); join_data = JOIN truck_events  
BY (driverId), drivers BY (driverId); DESCRIBE join_data;
```

```

cloudera@quickstart:~
File Edit View Search Terminal Help
GNU nano 2.0.9 File: Truck-Events

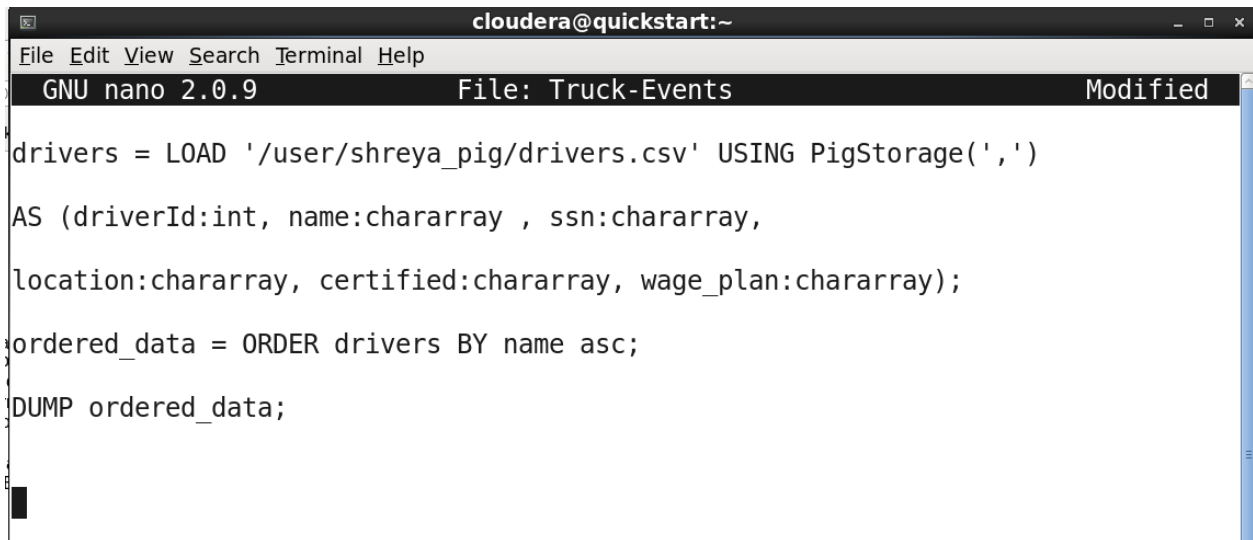
truck_events = LOAD '/user/shreya_pig/truck_event_text_partition.csv' USING Pig$
AS (driverId:int, truckId:int, eventTime:chararray,
eventType:chararray, longitude:double, latitude:double, eventKey:chararray , co$
drivers = LOAD '/user/shreya_pig/drivers.csv' USING PigStorage(',')
AS (driverId:int, name:chararray , ssn:chararray, location:chararray,
certified:chararray, wage_plan:chararray);
join_data = JOIN truck_events BY (driverId), drivers BY (driverId);
DESCRIBE join_data;

[cloudera@quickstart ~]$ pig -f Truck-Events
log4j:WARN No appenders could be found for logger (org.apache.hadoop.util.Shell)
.
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more in
fo.
2022-04-06 16:12:35,025 [main] INFO org.apache.pig.Main - Apache Pig version 0.
12.0-cdh5.4.2 (rexported) compiled May 19 2015, 17:03:41
2022-04-06 16:12:35,025 [main] INFO org.apache.pig.Main - Logging error message
s to: /home/cloudera/pig_1649286754994.log
2022-04-06 16:12:37,020 [main] INFO org.apache.pig.impl.util.Utils - Default bo
otup file /home/cloudera/.pigbootup not found
2022-04-06 16:12:37,725 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
2022-04-06 16:12:37,725 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-04-06 16:12:37,726 [main] INFO org.apache.pig.backend.hadoop.executionengi
2022-04-06 16:12:39,270 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-04-06 16:12:39,277 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
2022-04-06 16:12:39,417 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-04-06 16:12:39,423 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
2022-04-06 16:12:39,514 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-04-06 16:12:39,515 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
join_data: {truck_events::driverId: int,truck_events::truckId: int,truck_events:
:eventTime: chararray,truck_events::eventType: chararray,truck_events::longitude
: double,truck_events::latitude: double,truck_events::eventKey: chararray,truck_
events::correlationId: long,truck_events::driverName: chararray,truck_events::ro
uteId: long,truck_events::routeName: chararray,truck_events::eventDate: chararra
y,drivers::driverId: int,drivers::name: chararray,drivers::ssn: chararray,driver
s::location: chararray,drivers::certified: chararray,drivers::wage_plan: chararr
ay}

```

Step 11 : Sort the data using “ORDER BY”

```
drivers = LOAD '/user/maria_dev/drivers.csv' USING PigStorage(',') AS (driverId:int,  
name:chararray , ssn:chararray,  
location:chararray, certified:chararray, wage_plan:chararray); ordered_data = ORDER drivers  
BY name asc;  
DUMP ordered_data;
```

A screenshot of a terminal window titled "cloudera@quickstart:~". The window shows the GNU nano 2.0.9 editor with a file named "Truck-Events". The script content is:

```
drivers = LOAD '/user/shreya_pig/drivers.csv' USING PigStorage(',')  
AS (driverId:int, name:chararray , ssn:chararray,  
location:chararray, certified:chararray, wage_plan:chararray);  
ordered_data = ORDER drivers BY name asc;  
DUMP ordered_data;
```

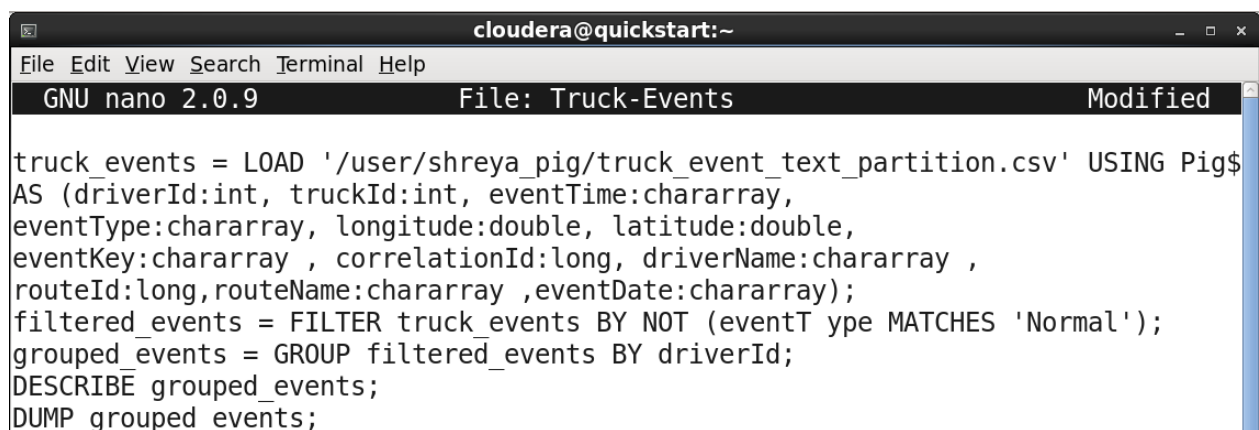
```
[cloudera@quickstart ~]$ pig -f Truck-Events  
log4j:WARN No appenders could be found for logger (org.apache.hadoop.util.Shell)  
.  
log4j:WARN Please initialize the log4j system properly.  
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more in  
fo.  
2022-04-06 16:16:33,014 [main] INFO org.apache.pig.Main - Apache Pig version 0.  
12.0-cdh5.4.2 (rexported) compiled May 19 2015, 17:03:41  
2022-04-06 16:16:33,015 [main] INFO org.apache.pig.Main - Logging error message  
s to: /home/cloudera/pig_1649286992987.log  
2022-04-06 16:16:34,583 [main] INFO org.apache.pig.impl.util.Utils - Default bo  
otup file /home/cloudera/.pigbootup not found  
2022-04-06 16:16:35,148 [main] INFO org.apache.hadoop.conf.Configuration.deprec  
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr  
ess  
2022-04-06 16:16:35,148 [main] INFO org.apache.hadoop.conf.Configuration.deprec  
ation - fs.default.name is deprecated. Instead, use fs.defaultFS  
2022-04-06 16:16:35,148 [main] INFO org.apache.pig.backend.hadoop.executionengi  
ne.HExecutionEngine - Connecting to hadoop file system at: hdfs://quickstart.clo  
udera:8020  
2022-04-06 16:16:36,636 [main] INFO org.apache.pig.backend.hadoop.executionengi  
ne.HExecutionEngine - Connecting to map-reduce job tracker at: localhost:8021
```



```
(14,Adis Cesir,820812209,Ap #810-1228 In St.,Y,hours)
(19,Ajay Singh,160005158,592-9430 Nonummy Avenue,Y,hours)
(36,Andrew Grande,245303216,Ap #685-9598 Egestas Rd.,Y,hours)
(20,Chris Harris,921812303,883-2691 Proin Avenue,Y,hours)
(30,Dan Rice,282307061,Ap #881-9267 Mollis Avenue,Y,hours)
(43,Dave Patton,977706052,3028 A- St.,Y,hours)
(39,David Kaiser,967706052,9185 At Street,Y,hours)
(24,Don Hilborn,254412152,4361 Ac Road,Y,hours)
(35,Emil Siemes,971401151,321-2976 Felis Rd.,Y,hours)
(17,Eric Mizell,123808238,P.O. Box 579- 2191 Gravida. Street,Y,hours)
(34,Frank Romano,391407216,Ap #753-6814 Quis Ave,Y,hours)
(10,George Vetticaden,621011971,244-4532 Nulla Rd.,N,miles)
(18,Grant Liu,171010151,Ap #928-3159 Vestibulum Av.,Y,hours)
(41,Greg Phillips,308103116,P.O. Box 847- 5961 Arcu. Road,Y,hours)
(11,Jamie Engesser,262112338,366-4125 Ac Street,N,miles)
(25,Jean-Philippe Playe,913310051,P.O. Box 812- 6238 Ac Rd.,Y,hours)
(21,Jeff Markham,209408086,Ap #852-7966 Facilisis St.,Y,hours)
(13,Joe Niemiec,139907145,2071 Hendrerit. Ave,Y,hours)
(27,Mark Lochbihler,392603159,8355 Ipsum St.,Y,hours)
(26,Michael Aube,124705141,P.O. Box 213- 8948 Nec Ave,Y,hours)
(22,Michael Aube,124705141,P.O. Box 213- 8948 Nec Ave,Y,hours)
```

Step 12 : Filter and Group the data using “GROUP BY”

```
truck_events = LOAD '/user/maria_dev/truck_event_text_partition.csv' USING PigStorage(',')
AS (driverId:int, truckId:int, eventTime:chararray,
eventKey:chararray, longitude:double, latitude:double, eventKey:chararray ,
correlationId:long, driverName:chararray , routeId:long,routeName:chararray
,eventDate:chararray);
filtered_events = FILTER truck_events BY NOT (eventTime MATCHES 'Normal');
grouped_events = GROUP filtered_events BY driverId;
DESCRIBE grouped_events; DUMP grouped_events;
```



The screenshot shows a terminal window titled "cloudera@quickstart:~". Inside, the GNU nano 2.0.9 editor is open with the file "Truck-Events". The editor contains the following Pig Latin code:

```
truck_events = LOAD '/user/shreya_pig/truck_event_text_partition.csv' USING Pig$
AS (driverId:int, truckId:int, eventTime:chararray,
eventKey:chararray, longitude:double, latitude:double,
eventKey:chararray , correlationId:long, driverName:chararray ,
routeId:long,routeName:chararray ,eventDate:chararray);
filtered_events = FILTER truck_events BY NOT (eventTime MATCHES 'Normal');
grouped_events = GROUP filtered_events BY driverId;
DESCRIBE grouped_events;
DUMP grouped_events;
```

```
[cloudera@quickstart ~]$ pig -f Truck-Events
log4j:WARN No appenders could be found for logger (org.apache.hadoop.util.Shell)
.
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more in
fo.
2022-04-06 16:21:49,588 [main] INFO  org.apache.pig.Main - Apache Pig version 0.
12.0-cdh5.4.2 (rexported) compiled May 19 2015, 17:03:41
2022-04-06 16:21:49,588 [main] INFO  org.apache.pig.Main - Logging error message
s to: /home/cloudera/pig_1649287309566.log
2022-04-06 16:21:50,997 [main] INFO  org.apache.pig.impl.util.Utils - Default bo
otup file /home/cloudera/.pigbootup not found
2022-04-06 16:21:51,633 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
2022-04-06 16:21:51,633 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-04-06 16:21:51,634 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.HExecutionEngine - Connecting to hadoop file system at: hdfs://quickstart.clo
udera:8020
2022-04-06 16:21:52,978 [main] INFO  org.apache.pig.backend.hadoop.executionengi
```

```

-----
(13,{(13,89,00:47.7,Lane Departure,-89.03,41.92,13|89|9223370572464728156,366000000000000000,Joe
Niemiec,927636994,Des Moines to Chicago.kml,2016-05-27-22)})
(14,{(14,25,00:48.4,Unsafe following distance,-91.63,41.72,14|25|9223370572464727394,366000000000
0000000,Adis Cesir,160405074,Joplin to Kansas City Route 2,2016-05-27-22)})
(15,{(15,51,00:48.8,Lane Departure,-90.04,35.19,15|51|9223370572464727025,366000000000000000,Roh
it Bakshi,1384345811,Joplin to Kansas City,2016-05-27-22)})
(16,{(16,12,00:48.9,Lane Departure,-89.52,40.7,16|12|9223370572464726925,366000000000000000,Tom
McCuch,1961634315,Saint Louis to Memphis,2016-05-27-22)})
(17,{(17,15,00:48.4,Lane Departure,-90.79,38.83,17|15|9223370572464727374,366000000000000000,Eri
c Mizell,1927624662,Springfield to KC Via Columbia,2016-05-27-22)})
(18,{(18,16,00:47.2,Overspeed,-94.28,39.53,18|16|9223370572464728575,366000000000000000,Grant Li
u,1565885487,Springfield to KC Via Hanibal,2016-05-27-22)})
(19,{(19,26,00:48.6,Unsafe following distance,-94.57,35.37,19|26|9223370572464727224,366000000000
0000000,Ajay Singh,1962261785,Wichita to Little Rock.kml,2016-05-27-22)})
(20,{(20,41,00:46.9,Overspeed,-89.03,41.92,20|41|9223370572464728915,366000000000000000,Chris Ha
rris,160779139,Des Moines to Chicago Route 2,2016-05-27-22)})
-----

```

CONCLUSION:

In this experiment, various Pig commands have been successfully implemented for data transformations on the given dataset.