Program – 7

AIM: 1) To install and run Pig. 2) Write Pig Latin scripts to sort, group, join, project and filter data.

Introduction & Theory

About Pig

Pig is a high level scripting language that is used with Apache Hadoop. Pig enables data workers to write complex data transformations without knowing Java. Pig's simple SQL-like scripting language is called Pig Latin, and appeals to developers already familiar with scripting languages and SQL.

Pig is complete, so you can do all required data manipulations in Apache Hadoop with Pig. Through the User Defined Functions(UDF) facility in Pig, Pig can invoke code in many languages like JRuby, Jython and Java. You can also embed Pig scripts in other languages. The result is that you can use Pig as a component to build larger and more complex applications that tackle real business problems.

Pig works with data from many sources, including structured and unstructured data, and store the results into the Hadoop Data File System.

Pig scripts are translated into a series of MapReduce jobs that are run on the Apache Hadoop cluster.

Pig Features

- **Rich set of operators** − It provides many operators to perform operations like join, sort, filer, etc.
- **Ease of programming** − Pig Latin is similar to SQL and it is easy to write a Pig script if you are good at SQL.
- **Optimization opportunities** − The tasks in Apache Pig optimize their execution automatically, so the programmers need to focus only on semantics of the language.
- **Extensibility** − Using the existing operators, users can develop their own functions to read, process, and write data.
- **UDF's** − Pig provides the facility to create **User-defined Functions**in other programming languages such as Java and invoke or embed them in Pig Scripts.
- **Handles all kinds of data** − Apache Pig analyzes all kinds of data, both structured as well as unstructured. It stores the results in HDFS.

Pig versus MapReduce

Apache Pig	MapReduce
Apache Pig is a data flow language.	MapReduce is a data processing paradigm.
It is a high level language.	MapReduce is low level and rigid.

Program – 7

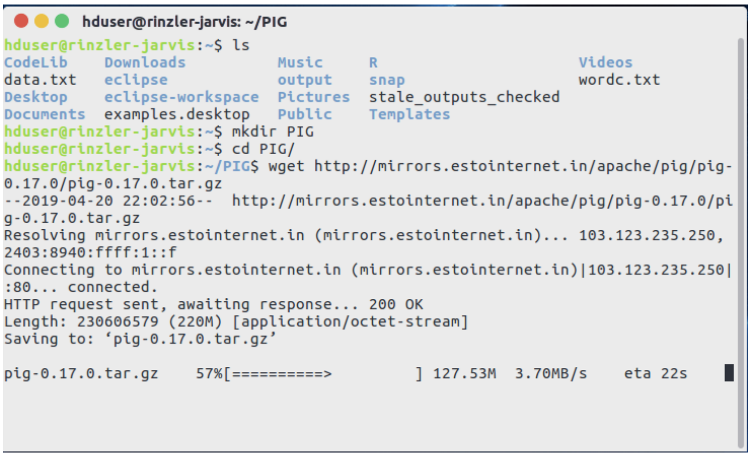
Performing a Join operation in Apache Pig is pretty simple.	It is quite difficult in MapReduce to perform a Join operation between datasets.
Any novice programmer with a basic knowledge of SQL can work conveniently with Apache Pig.	Exposure to Java is must to work with MapReduce.
Apache Pig uses multi-query approach, thereby reducing the length of the codes to a great extent.	MapReduce will require almost 20 times more the number of lines to perform the same task.
There is no need for compilation. On execution, every Apache Pig operator is converted internally into a MapReduce job.	MapReduce jobs have a long compilation process.

Installing Pig

Prerequisites:

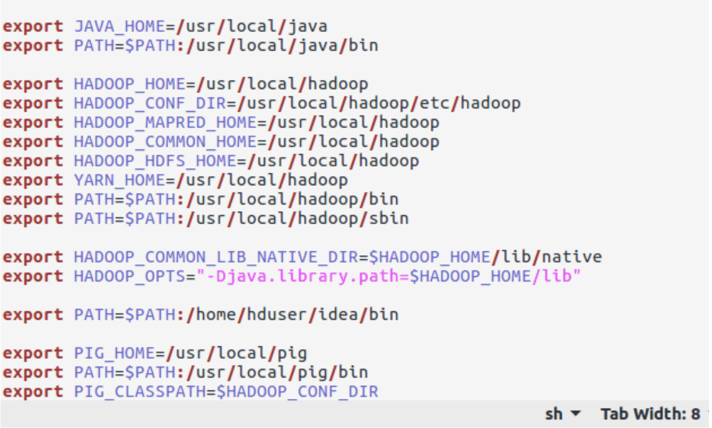
1. Java
2. Hadoop

1. Download the Pig Files from Apache.



```
hduser@rinzler-jarvis:~$ ls
CodeLib    Downloads          Music    R                    Videos
data.txt   eclipse            output   snap                 wordc.txt
Desktop    eclipse-workspace  Pictures stale_outputs_checked
Documents  examples.desktop   Public   Templates
hduser@rinzler-jarvis:~$ mkdir PIG
hduser@rinzler-jarvis:~$ cd PIG/
hduser@rinzler-jarvis:~/PIG$ wget http://mirrors.estointernet.in/apache/pig/pig-
0.17.0/pig-0.17.0.tar.gz
--2019-04-20 22:02:56--  http://mirrors.estointernet.in/apache/pig/pig-0.17.0/pi
g-0.17.0.tar.gz
Resolving mirrors.estointernet.in (mirrors.estointernet.in)... 103.123.235.250,
2403:8940:ffff:1::f
Connecting to mirrors.estointernet.in (mirrors.estointernet.in)|103.123.235.250|
:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: 230606579 (220M) [application/octet-stream]
Saving to: ‘pig-0.17.0.tar.gz’

pig-0.17.0.tar.gz    57%[==========>         ] 127.53M  3.70MB/s    eta 22s
```

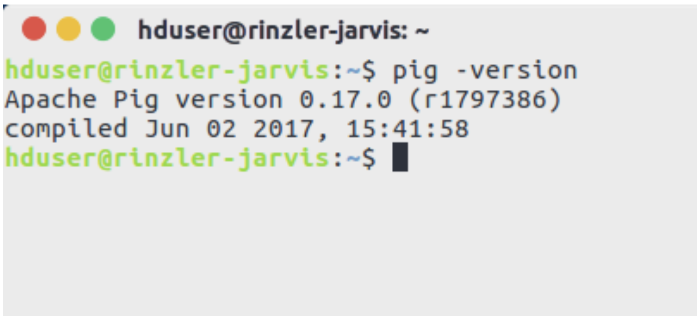
2. Extract the files to a convenient location. (/usr/local).
3. Edit the system variable to include the Pig files.



```
export JAVA_HOME=/usr/local/java
export PATH=$PATH:/usr/local/java/bin

export HADOOP_HOME=/usr/local/hadoop
export HADOOP_CONF_DIR=/usr/local/hadoop/etc/hadoop
export HADOOP_MAPRED_HOME=/usr/local/hadoop
export HADOOP_COMMON_HOME=/usr/local/hadoop
export HADOOP_HDFS_HOME=/usr/local/hadoop
export YARN_HOME=/usr/local/hadoop
export PATH=$PATH:/usr/local/hadoop/bin
export PATH=$PATH:/usr/local/hadoop/sbin

export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_HOME/lib/native
export HADOOP_OPTS="-Djava.library.path=$HADOOP_HOME/lib"

export PATH=$PATH:/home/hduser/idea/bin

export PIG_HOME=/usr/local/pig
export PATH=$PATH:/usr/local/pig/bin
export PIG_CLASSPATH=$HADOOP_CONF_DIR
```

4. Check Pig version to check if its working properly.

Program – 7

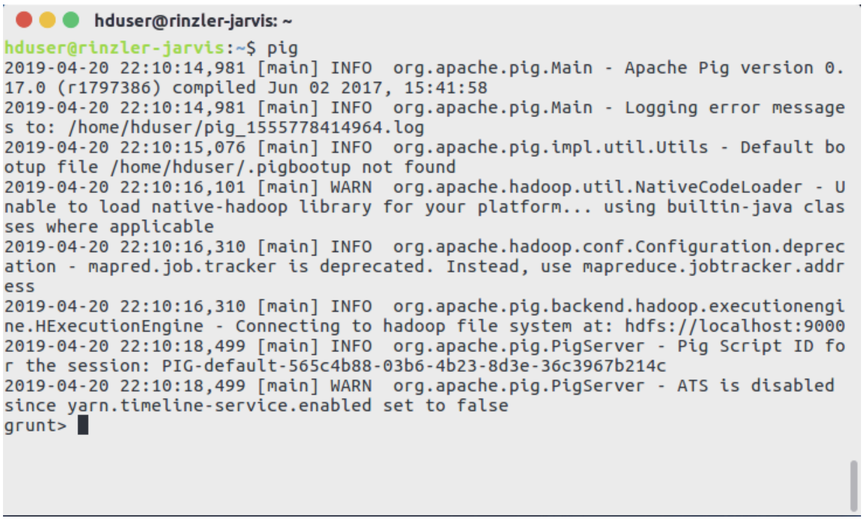


```
hduser@rinzler-jarvis: ~
hduser@rinzler-jarvis:~$ pig -version
Apache Pig version 0.17.0 (r1797386)
compiled Jun 02 2017, 15:41:58
hduser@rinzler-jarvis:~$ 
```

5. Run Pig.

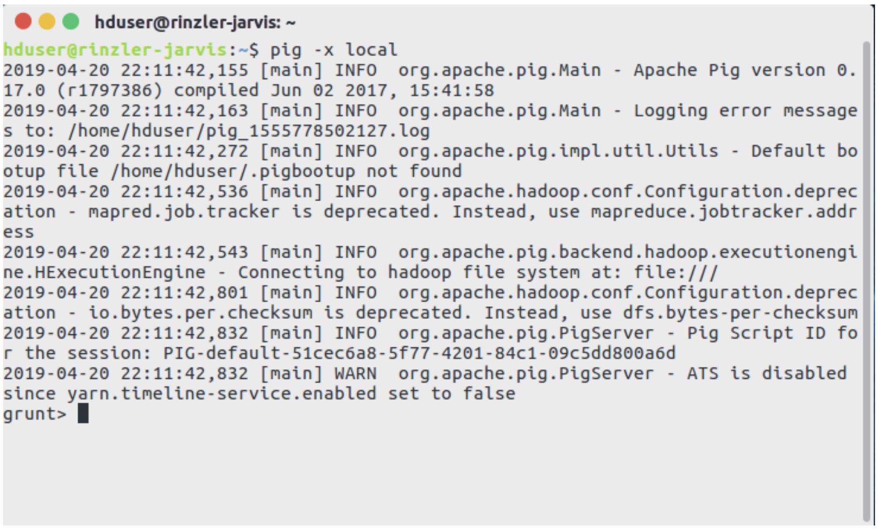
Execution modes in Apache Pig:

- *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.



```
hduser@rinzler-jarvis: ~
hduser@rinzler-jarvis:~$ pig
2019-04-20 22:10:14,981 [main] INFO  org.apache.pig.Main - Apache Pig version 0.
17.0 (r1797386) compiled Jun 02 2017, 15:41:58
2019-04-20 22:10:14,981 [main] INFO  org.apache.pig.Main - Logging error message
s to: /home/hduser/pig_1555778414964.log
2019-04-20 22:10:15,076 [main] INFO  org.apache.pig.impl.util.Utils - Default bo
otup file /home/hduser/.pigbootup not found
2019-04-20 22:10:16,101 [main] WARN  org.apache.hadoop.util.NativeCodeLoader - U
nable to load native-hadoop library for your platform... using builtin-java clas
ses where applicable
2019-04-20 22:10:16,310 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
2019-04-20 22:10:16,310 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.HExecutionEngine - Connecting to hadoop file system at: hdfs://localhost:9000
2019-04-20 22:10:18,499 [main] INFO  org.apache.pig.PigServer - Pig Script ID fo
r the session: PIG-default-565c4b88-03b6-4b23-8d3e-36c3967b214c
2019-04-20 22:10:18,499 [main] WARN  org.apache.pig.PigServer - ATS is disabled
since yarn.timeline-service.enabled set to false
grunt> 
```

- *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.



```
hduser@rinzler-jarvis: ~
hduser@rinzler-jarvis:~$ pig -x local
2019-04-20 22:11:42,155 [main] INFO  org.apache.pig.Main - Apache Pig version 0.
17.0 (r1797386) compiled Jun 02 2017, 15:41:58
2019-04-20 22:11:42,163 [main] INFO  org.apache.pig.Main - Logging error message
s to: /home/hduser/pig_1555778502127.log
2019-04-20 22:11:42,272 [main] INFO  org.apache.pig.impl.util.Utils - Default bo
otup file /home/hduser/.pigbootup not found
2019-04-20 22:11:42,536 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
2019-04-20 22:11:42,543 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.HExecutionEngine - Connecting to hadoop file system at: file:///
2019-04-20 22:11:42,801 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2019-04-20 22:11:42,832 [main] INFO  org.apache.pig.PigServer - Pig Script ID fo
r the session: PIG-default-51cec6a8-5f77-4201-84c1-09c5dd800a6d
2019-04-20 22:11:42,832 [main] WARN  org.apache.pig.PigServer - ATS is disabled
since yarn.timeline-service.enabled set to false
grunt> 
```

Program – 7

Pig Operations

```
truck_events = LOAD '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 '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);

ordered_data = ORDER join_data BY truck_events::driverName asc;

filtered = FILTER ordered_data BY NOT (eventType MATCHES 'Normal');

grouped_events = GROUP filtered BY truck_events::driverId;

grouped_events_subset = LIMIT grouped_events 5;

DUMP grouped_events;

STORE grouped_events INTO 'outut.txt' USING PigStorage(';');
```

Output

```
10;{(10,85,59:46.9,Overspeed,-
95.5,36.37,10|85|9223370572464788896,3660000000000000000,George
Vetticaden,1390372503,Saint Louis to Tulsa,2016-05-27-22,10,George
Vetticaden,621011971,244-4532 Nulla
Rd.,N,miles),(10,85,00:13.1,Unsafe tail distance,-
91.18,38.22,10|85|9223370572464762694,3660000000000000000,George
Vetticaden,1390372503,Saint Louis to Tulsa,2016-05-27-22,10,George
Vetticaden,621011971,244-4532 Nulla
Rd.,N,miles),(10,85,00:39.7,Overspeed,-
94.23,37.09,10|85|9223370572464736126,3660000000000000000,George
Vetticaden,1390372503,Saint Louis to Tulsa,2016-05-27-22,10,George
Vetticaden,621011971,244-4532 Nulla Rd.,N,miles)}
```

```
11;{(11,74,59:47.3,Unsafe tail distance,-
89.63,39.84,11|74|9223370572464788546,3660000000000000000,Jamie
Engesser,1567254452,Saint Louis to Memphis Route2,2016-05-27-
22,11,Jamie Engesser,262112338,366-4125 Ac
Street,N,miles),(11,74,00:49.6,Lane Departure,-
89.71,37.47,11|74|9223370572464726246,3660000000000000000,Jamie
Engesser,1567254452,Saint Louis to Memphis Route2,2016-05-27-
22,11,Jamie Engesser,262112338,366-4125 Ac
Street,N,miles),(11,74,00:14.1,Lane Departure,-
88.77,40.76,11|74|9223370572464761716,3660000000000000000,Jamie
Engesser,1567254452,Saint Louis to Memphis Route2,2016-05-27-
22,11,Jamie Engesser,262112338,366-4125 Ac
Street,N,miles),(11,74,00:23.1,Unsafe tail distance,-
88.42,41.11,11|74|9223370572464752715,3660000000000000000,Jamie
Engesser,1567254452,Saint Louis to Memphis Route2,2016-05-27-
22,11,Jamie Engesser,262112338,366-4125 Ac
Street,N,miles),(11,74,00:32.0,Unsafe tail distance,-
90.2,38.65,11|74|9223370572464743846,3660000000000000000,Jamie
```

Program – 7

	Engesser,1567254452,Saint Louis to Memphis Route2,2016-05-27-22,11,Jamie Engesser,262112338,366-4125 Ac Street,N,miles),(11,74,59:29.1,Overspeed,-88.07,41.48,11|74|9223370572464806746,366000000000000000,Jamie Engesser,1567254452,Saint Louis to Memphis Route2,2016-05-27-22,11,Jamie Engesser,262112338,366-4125 Ac Street,N,miles),(11,74,00:05.4,Unsafe following distance,-89.74,39.1,11|74|9223370572464770396,366000000000000000,Jamie Engesser,1567254452,Saint Louis to Memphis Route2,2016-05-27-22,11,Jamie Engesser,262112338,366-4125 Ac Street,N,miles),(11,74,00:41.0,Lane Departure,-90.07,35.68,11|74|9223370572464734786,366000000000000000,Jamie Engesser,1567254452,Saint Louis to Memphis Route2,2016-05-27-22,11,Jamie Engesser,262112338,366-4125 Ac Street,N,miles),(11,74,59:56.4,Lane Departure,-87.67,41.87,11|74|9223370572464779456,366000000000000000,Jamie Engesser,1567254452,Saint Louis to Memphis Route2,2016-05-27-22,11,Jamie Engesser,262112338,366-4125 Ac Street,N,miles),(11,74,59:38.0,Unsafe tail distance,-89.17,40.38,11|74|9223370572464797796,366000000000000000,Jamie Engesser,1567254452,Saint Louis to Memphis Route2,2016-05-27-22,11,Jamie Engesser,262112338,366-4125 Ac Street,N,miles)}
3.	12;{(12,104,00:47.6,Unsafe following distance,-90.0,37.72,12|104|9223370572464728186,366000000000000000,Paul Codding,24929475,Peoria to Ceder Rapids,2016-05-27-22,12,Paul Coddin,198041975,Ap #622-957 Risus. Street,Y,hours)}
4.	13;{(13,89,00:47.7,Lane Departure,-89.03,41.92,13|89|9223370572464728156,366000000000000000,Joe Niemiec,927636994,Des Moines to Chicago.kml,2016-05-27-22,13,Joe Niemiec,139907145,2071 Hendrerit. Ave,Y,hours)}
5.	14;{(14,25,00:48.4,Unsafe following distance,-91.63,41.72,14|25|9223370572464727394,366000000000000000,Adis Cesir,160405074,Joplin to Kansas City Route 2,2016-05-27-22,14,Adis Cesir,820812209,Ap #810-1228 In St.,Y,hours)}
6.	15;{(15,51,00:48.8,Lane Departure,-90.04,35.19,15|51|9223370572464727025,366000000000000000,Rohit Bakshi,1384345811,Joplin to Kansas City,2016-05-27-22,15,Rohit Bakshi,239005227,648-5681 Dui- Rd.,Y,hours)}
7.	16;{(16,12,00:48.9,Lane Departure,-89.52,40.7,16|12|9223370572464726925,366000000000000000,Tom McCuch,1961634315,Saint Louis to Memphis,2016-05-27-22,16,Tom McCuch,363303105,P.O. Box 313- 962 Parturient Rd.,Y,hours)}
8.	17;{(17,15,00:48.4,Lane Departure,-90.79,38.83,17|15|9223370572464727374,366000000000000000,Eric Mizell,1927624662,Springfield to KC Via Columbia,2016-05-27-22,17,Eric Mizell,123808238,P.O. Box 579- 2191 Gravida. Street,Y,hours)}
9.	18;{(18,16,00:47.2,Overspeed,-94.28,39.53,18|16|9223370572464728575,366000000000000000,Grant Liu,1565885487,Springfield to KC Via Hanibal,2016-05-27-22,18,Grant Liu,171010151,Ap #928-3159 Vestibulum Av.,Y,hours)}
10.	19;{(19,26,00:48.6,Unsafe following distance,-94.57,35.37,19|26|9223370572464727224,366000000000000000,Ajay Singh,1962261785,Wichita to Little Rock.kml,2016-05-27-22,19,Ajay Singh,160005158,592-9430 Nonummy Avenue,Y,hours)}

Program – 7

```
20;{(20,41,00:46.9,Overspeed,-89.03,41.92,20|41|9223370572464728915,3660000000000000000,Chris Harris,160779139,Des Moines to Chicago Route 2,2016-05-27-22,20,Chris Harris,921812303,883-2691 Proin Avenue,Y,hours)}
21;{(21,109,00:46.8,Unsafe tail distance,-88.07,41.48,21|109|9223370572464729016,3660000000000000000,Jeff Markham,1594289134,Memphis to Little Rock Route 2,2016-05-27-22,21,Jeff Markham,209408086,Ap #852-7966 Facilisis St.,Y,hours)}
22;{(22,87,00:46.5,Unsafe tail distance,-90.04,35.19,22|87|9223370572464729286,3660000000000000000,Nadeem Asghar,1198242881, Saint Louis to Chicago Route2,2016-05-27-22,22,Nadeem Asghar,783204269,154-9147 Aliquam Ave,Y,hours)}
23;{(23,68,00:47.8,Lane Departure,-89.52,40.7,23|68|9223370572464727994,3660000000000000000,Adam Diaz,160405074,Joplin to Kansas City Route 2,2016-05-27-22,23,Adam Diaz,928312208,P.O. Box 260- 6127 Vitae Road,Y,hours)}
24;{(24,97,00:48.6,Lane Departure,-89.17,40.38,24|97|9223370572464727226,3660000000000000000,Don Hilborn,1090292248,Peoria to Ceder Rapids Route 2,2016-05-27-22,24,Don Hilborn,254412152,4361 Ac Road,Y,hours)}
25;{(25,96,00:40.1,Overspeed,-89.54,36.84,25|96|9223370572464735726,3660000000000000000,Jean-Philippe Player,371182829,Memphis to Little Rock,2016-05-27-22,25,Jean-Philippe Playe,913310051,P.O. Box 812- 6238 Ac Rd.,Y,hours)}
26;{(26,57,00:48.8,Overspeed,-95.99,36.17,26|57|9223370572464727046,3660000000000000000,Michael Aube,1325712174,Saint Louis to Tulsa Route2,2016-05-27-22,26,Michael Aube,124705141,P.O. Box 213- 8948 Nec Ave,Y,hours)}
27;{(27,105,00:48.0,Unsafe following distance,-90.79,38.83,27|105|9223370572464727846,3660000000000000000,Mark Lochbihler,1325562373,Springfield to KC Via Columbia Route 2,2016-05-27-22,27,Mark Lochbihler,392603159,8355 Ipsum St.,Y,hours)}
28;{(28,39,00:47.5,Overspeed,-94.28,39.53,28|39|9223370572464728273,3660000000000000000,Olivier Renault,137128276,Springfield to KC Via Hanibal Route 2,2016-05-27-22,28,Olivier Renault,959908181,P.O. Box 243- 6509 Erat. Avenue,Y,hours)}
29;{(29,66,00:47.8,Overspeed,-94.57,35.37,29|66|9223370572464728016,3660000000000000000,Teddy Choi,803014426,Wichita to Little Rock Route 2,2016-05-27-22,29,Teddy Choi,185502192,P.O. Box 106- 7003 Amet Rd.,Y,hours)}
30;{(30,58,00:49.3,Unsafe following distance,-89.03,41.92,30|58|9223370572464726546,3660000000000000000,Dan Rice,160779139,Des Moines to Chicago Route 2,2016-05-27-22,30,Dan Rice,282307061,Ap #881-9267 Mollis Avenue,Y,hours)}
31;{(31,18,00:47.6,Lane Departure,-88.07,41.48,31|18|9223370572464728166,3660000000000000000,Rommel Garcia,1594289134,Memphis to Little Rock Route 2,2016-05-27-22,31,Rommel Garcia,858912101,P.O. Box 945- 6015 Sociis St.,Y,hours)}
32;{(32,42,00:48.7,Unsafe following distance,-90.04,35.19,32|42|9223370572464727106,3660000000000000000,Ryan Templeton,1090292248,Peoria to Ceder Rapids Route 2,2016-05-27-22,32,Ryan Templeton,290304287,765-6599 Egestas. Av.,Y,hours)}
```

Program – 7

Findings and Learnings:

1. We learned about Apache Pig.

2. We learned about the advantages of Apache Pig.

3. We compared Map Reduce and Pig.

4. We learnt how to write scripts in Pig.