(/)

Tutorials (/tutorials.html)

Beginners Guide to Apache Pig

Ready to Get Started?

Download Sandbox **(/downloads/hortonworks-sandbox.html)**

Introduction

In this tutorial you will gain a working knowledge of Pig through the hands-on experience of creating Pig scripts to carry out essential data operations and tasks.

We will first read in two data files that contain driver data statistics, and then use these files to perform a number of Pig operations including:

- Define a relation with and without `schema`
- Define a new relation from an `existing relation`
- `Select` specific columns from within a relation
- `Join` two relations
- Sort the data using `'ORDER BY'`
- FILTER and Group the data using `'GROUP BY'`

Prerequisites

- Downloaded and deployed the **Hortonworks Data Platform (HDP) (https://www.cloudera.com/downloads/hortonworks-sandbox/hdp.html?utm_source=mktg-tutorial)** Sandbox
- **Learning the Ropes of the HDP Sandbox (/tutor sandbox.html)**
- Allow yourself around one hour to complete this tu

Cloudera Data Platform (CDP) is our easy-to-use, integrated analytics and data management platform. …

2

Outline

What is 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.

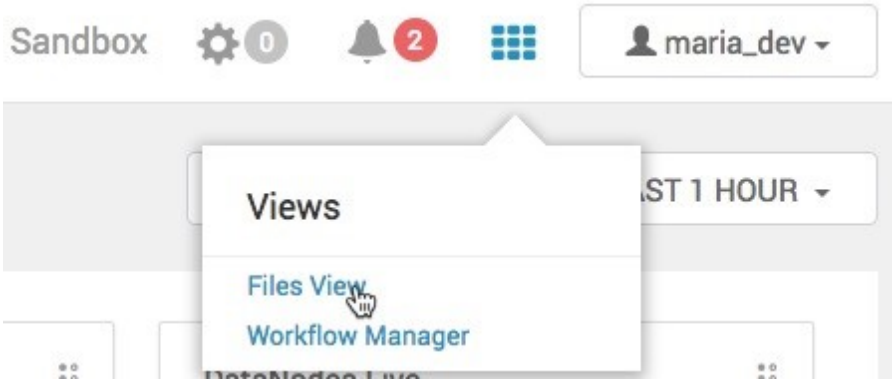
Pig scripts are translated into a series of MapReduce jobs that are run on the Apache Hadoop cluster.

Download the Data

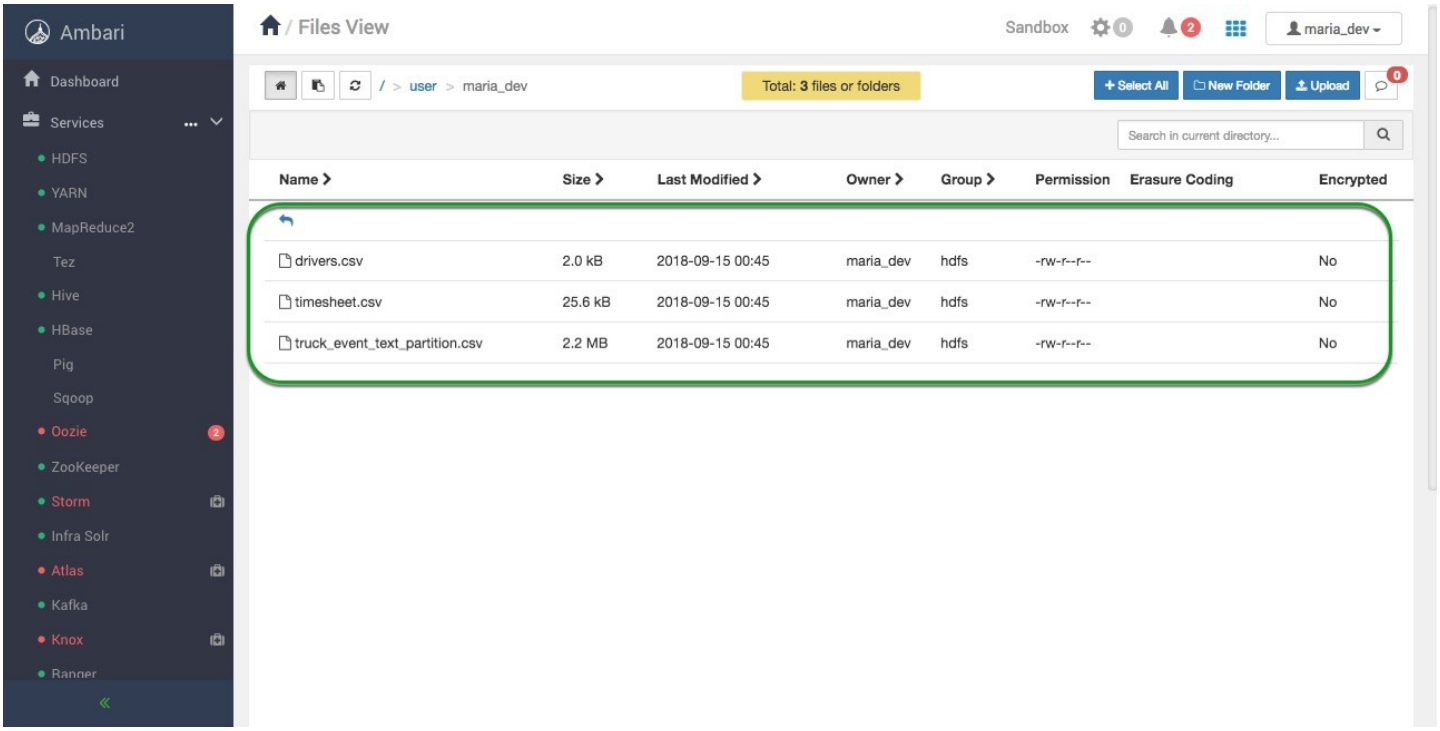
Download the driver data file **from here (/content/dam/www/marketing/tutorials/beginners-guide-to-apache-pig/assets/driver_data.zip)**. Once you have the file you will need to unzip the file into a directory. We will be uploading two csv files - `truck_event_text_partition.csv` and `drivers.csv` .

Upload the data files

Select the `HDFS Files view` from the Off-canvas menu at the top. That is the `views menu`. The HDFS Files view allows you to view the Hortonworks Data Platform(HDP) file store. The HDP file system is separate from the local file system.



Navigate to `/user/maria_dev` or a path of your choice, click `Upload` and `Browse`, which brings up a dialog box where you can select the `drivers.csv` file from your computer. Upload the `truck_event_text_partition.csv` file in the same way. When finished, notice that both files are now in HDFS.



Create Your Script

> Note: In this tutorial Vi is used; however, any text editor will work as long as the files we create are stored on the Sandbox.

Navigate to **http://sandbox-hdp.hortonworks.com:4200/ (http://sandbox-**

hdp.hortonworks.com:4200/) (Shell-In-A-Box) and sign in as root.

Next, change users to **maria_dev** and change directories to maria_dev's home directory:

```
su maria_dev
cd
```

Now create a new file where we will create the Pig Script:

```
vi Truck-Events
```

> Note: Press **i** to enter insert mode in Vi.

Define a relation

In this step, you will create a script to load the data and define a relation.

On line 1 define a relation named `truck_events` that represents all the truck events

On line 2 use the `DESCRIBE` command to view the `truck_events` relation

The completed code will look like:

```
truck_events = LOAD '/user/maria_dev/truck_event_text_partition.csv' USING
DESCRIBE truck_events;
```

> **Note:** In the LOAD script, you can choose any directory path. Verify the folders have been created in HDFS Files View.

Save and Execute the Script

To save your changes while on Vi press `esc` and type `:x` then enter.

To execute the script return to `~/` and submit the file we just created to pig:

```
pig -f Truck-Events
```

```
[root@sandbox-hdp ~]# pig -f Truck-Events
```

This action creates one or more MapReduce jobs. After a moment, the script starts and the page changes.

When the job completes, result output. Notice truck_events does not have a schema because we did not define one when loading the data into relation truck_events.

```
[root@sandbox-hdp ~]# pig -f Truck-Events
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
18/09/15 09:11:17 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
18/09/15 09:11:17 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
18/09/15 09:11:17 INFO pig.ExecTypeProvider: Trying ExecType : TEZ_LOCAL
18/09/15 09:11:17 INFO pig.ExecTypeProvider: Trying ExecType : TEZ
18/09/15 09:11:17 INFO pig.ExecTypeProvider: Picked TEZ as the ExecType
2018-09-15 09:11:17,833 [main] INFO  org.apache.pig.Main - Apache Pig version 0.16.0.3.0.0.0-1634 (rUnversion
ed directory) compiled Jul 12 2018, 20:39:28
2018-09-15 09:11:17,833 [main] INFO  org.apache.pig.Main - Logging error messages to: /root/pig_1537002677831
.log
2018-09-15 09:11:18,618 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /root/.pigbootup no
t found
2018-09-15 09:11:18,683 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connect
ing to hadoop file system at: hdfs://sandbox-hdp.hortonworks.com:8020
2018-09-15 09:11:19,300 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-Truck-Even
ts-203b8bc6-2b62-43fd-a301-6bb00ad68544
2018-09-15 09:11:19,300 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.e
nabled set to false
Schema for truck_events unknown.
2018-09-15 09:11:19,977 [main] INFO  org.apache.pig.Main - Pig script completed in 2 seconds and 388 millisec
onds (2388 ms)
[root@sandbox-hdp ~]#
```

Define a Relation with a Schema

Let's use the above code but this time with a schema. Modify line 1 of your script and add the following `AS` clause to define a schema for the truck events data. Open Vi and enter the following script:

```
truck_events = LOAD '/user/maria_dev/truck_event_text_partition.csv' USING
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 = 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);
DESCRIBE truck_events;
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
:x
```

Save and execute the script again.

> Note: Recall that we used `:x` to save the script and `pig -f Truck-Events` to run the job.

This time you should see the schema for the truck_events relation:

```
[                ~]# pig -f Truck-Events
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
18/09/15 09:20:21 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
18/09/15 09:20:21 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
18/09/15 09:20:21 INFO pig.ExecTypeProvider: Trying ExecType : TEZ_LOCAL
18/09/15 09:20:21 INFO pig.ExecTypeProvider: Trying ExecType : TEZ
18/09/15 09:20:21 INFO pig.ExecTypeProvider: Picked TEZ as the ExecType
2018-09-15 09:20:21,994 [main] INFO  org.apache.pig.Main - Apache Pig version 0.16.0.3.0.0.0-1634 (rUnversioned directory) com
piled Jul 12 2018, 20:39:28
2018-09-15 09:20:21,994 [main] INFO  org.apache.pig.Main - Logging error messages to: /root/pig_1537003221992.log
2018-09-15 09:20:22,685 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /root/.pigbootup not found
2018-09-15 09:20:22,756 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop fil
e system at: hdfs://sandbox-hdp.hortonworks.com:8020
2018-09-15 09:20:23,317 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-Truck-Events-e9658df4-2f30-
472c-ba0c-36c1a4211b53
2018-09-15 09:20:23,317 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to fal
se
truck_events: {driverId: int,truckId: int,eventTime: chararray,eventType: chararray,longitude: double,latitude: double,eventKe
y: chararray,correlationId: long,driverName: chararray,routeId: long,routeName: chararray,eventDate: chararray}
2018-09-15 09:20:24,055 [main] INFO  org.apache.pig.Main - Pig script completed in 2 seconds and 297 milliseconds (2297 ms)
[                ~]#
```

Define a new relation from an existing relation

You can define a new relation based on an existing one. For example, define the following

truck_events_subset relation, which is a collection of 100 entries (arbitrarily selected) from the truck_events relation. Add the following line to the end of your code:

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

```
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);
DESCRIBE truck_events;
truck_events_subset = LIMIT truck_events 100;
DESCRIBE truck_events_subset;
~
~
~
~
~
~
~
:x
```

Save and execute the code. Notice `truck_events_subset` has the same schema as `truck_events`, because `truck_events_subset` is a subset of `truck_events` relation.

```
[root@sandbox-hdp ~]# pig -f Truck-Events
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
18/09/15 09:24:45 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
18/09/15 09:24:45 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
18/09/15 09:24:45 INFO pig.ExecTypeProvider: Trying ExecType : TEZ_LOCAL
18/09/15 09:24:45 INFO pig.ExecTypeProvider: Trying ExecType : TEZ
18/09/15 09:24:45 INFO pig.ExecTypeProvider: Picked TEZ as the ExecType
2018-09-15 09:24:45,732 [main] INFO  org.apache.pig.Main - Apache Pig version 0.16.0.3.0.0.0-1634 (rUnversioned directory) compiled Jul 1
2 2018, 20:39:28
2018-09-15 09:24:45,732 [main] INFO  org.apache.pig.Main - Logging error messages to: /root/pig_1537003485730.log
2018-09-15 09:24:46,578 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /root/.pigbootup not found
2018-09-15 09:24:46,659 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at
: hdfs://sandbox-hdp.hortonworks.com:8020
2018-09-15 09:24:47,435 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-Truck-Events-2b113721-8bd8-44b7-8edf-0
20d8bc49f47
2018-09-15 09:24:47,435 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
1 truck_events: {driverId: int,truckId: int,eventTime: chararray,eventType: chararray,longitude: double,latitude: double,eventKey: chararra
y,correlationId: long,driverName: chararray,routeId: long,routeName: chararray,eventDate: chararray}
2 truck_events_subset: {driverId: int,truckId: int,eventTime: chararray,eventType: chararray,longitude: double,latitude: double,eventKey: c
hararray,correlationId: long,driverName: chararray,routeId: long,routeName: chararray,eventDate: chararray}
2018-09-15 09:24:48,469 [main] INFO  org.apache.pig.Main - Pig script completed in 2 seconds and 966 milliseconds (2966 ms)
[root@sandbox-hdp ~]#
```

View the Data

To view the data of a relation, use the `DUMP` command. Add the following `DUMP` command to your Pig script, then save and execute it again:

```
DUMP truck_events_subset;
```

```
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);
DESCRIBE truck_events;
truck_events_subset = LIMIT truck_events 100;
DESCRIBE truck_events_subset;
DUMP truck_events_subset;
~
~
~
~
~
~
:x
```

The command requires a MapReduce job to execute, so you will need to wait a minute or two for the job to complete. The output should be 100 entries from the contents of `truck_events_text_partition.csv` (and not necessarily the ones shown below, because again, entries are arbitrarily chosen):

```
(23,68,59:27.4,Normal,-91.47,41.74,23|68|9223370572464808375,3660000000000000000,Adam Diaz,160405074,Joplin to Kansas City Route 2,2016-05-27-22)
(13,89,59:27.6,Normal,-91.59,41.7,13|89|9223370572464808244,3660000000000000000,Joe Niemiec,927636994,Des Moines to Chicago.kml,2016-05-27-22)
(27,105,59:27.7,Normal,-91.19,38.83,27|105|9223370572464808146,3660000000000000000,Mark Lochbihler,1325562373,Springfield to KC Via Columbia Route 2,2016-05-27-2
2)
(28,39,59:27.7,Normal,-90.52,39.71,28|39|9223370572464808114,3660000000000000000,Olivier Renault,137128276,Springfield to KC Via Hanibal Route 2,2016-05-27-22)
(25,96,59:27.8,Normal,-89.65,36.37,25|96|9223370572464807986,3660000000000000000,Jean-Philippe Player,371182829,Memphis to Little Rock,2016-05-27-22)
(24,97,59:27.9,Normal,-87.67,41.87,24|97|9223370572464807926,3660000000000000000,Don Hilborn,1090292248,Peoria to Ceder Rapids Route 2,2016-05-27-22)
(30,58,59:28.0,Normal,-91.59,41.7,30|58|9223370572464807766,3660000000000000000,Dan Rice,160779139,Des Moines to Chicago Route 2,2016-05-27-22)
(11,74,59:28.3,Normal,-87.67,41.87,11|74|9223370572464807516,3660000000000000000,Jamie Engesser,1567254452,Saint Louis to Memphis Route2,2016-05-27-22)
(23,68,59:28.4,Normal,-91.63,41.72,23|68|9223370572464807444,3660000000000000000,Adam Diaz,160405074,Joplin to Kansas City Route 2,2016-05-27-22)
(13,89,59:28.5,Normal,-91.24,41.67,13|89|9223370572464807315,3660000000000000000,Joe Niemiec,927636994,Des Moines to Chicago.kml,2016-05-27-22)
(16,12,59:28.8,Normal,-91.63,41.72,16|12|9223370572464807036,3660000000000000000,Tom McCuch,1961634315,Saint Louis to Memphis,2016-05-27-22)
(15,51,59:28.8,Normal,-92.09,34.8,15|51|9223370572464807016,3660000000000000000,Rohit Bakshi,1384345811,Joplin to Kansas City,2016-05-27-22)
(12,104,59:29.1,Normal,-90.07,35.68,12|104|9223370572464806666,3660000000000000000,Paul Codding,24929475,Peoria to Ceder Rapids,2016-05-27-22)
(17,15,59:29.2,Normal,-91.56,38.93,17|15|9223370572464806616,3660000000000000000,Eric Mizell,1927624662,Springfield to KC Via Columbia,2016-05-27-22)
(27,105,59:29.3,Normal,-91.56,38.93,27|105|9223370572464806546,3660000000000000000,Mark Lochbihler,1325562373,Springfield to KC Via Columbia Route 2,2016-05-27-2
2)
(13,89,59:29.5,Normal,-90.82,41.66,13|89|9223370572464806335,3660000000000000000,Joe Niemiec,927636994,Des Moines to Chicago.kml,2016-05-27-22)
(16,12,59:29.6,Normal,-91.78,42.23,16|12|9223370572464806205,3660000000000000000,Tom McCuch,1961634315,Saint Louis to Memphis,2016-05-27-22)
(23,68,59:29.9,Normal,-91.78,42.23,23|68|9223370572464805905,3660000000000000000,Adam Diaz,160405074,Joplin to Kansas City Route 2,2016-05-27-22)
(11,74,59:30.0,Normal,-88.42,41.11,11|74|9223370572464805845,3660000000000000000,Jamie Engesser,1567254452,Saint Louis to Memphis Route2,2016-05-27-22)
(22,87,59:30.1,Normal,-92.31,34.78,22|87|9223370572464805680,3660000000000000000,Nadeem Asghar,1198242881, Saint Louis to Chicago Route2,2016-05-27-22)
(21,109,59:30.4,Normal,-94.38,38.99,21|109|9223370572464805406,3660000000000000000,Jeff Markham,1594289134,Memphis to Little Rock Route 2,2016-05-27-22)
(13,89,59:30.5,Normal,-90.64,41.56,13|89|9223370572464805356,3660000000000000000,Joe Niemiec,927636994,Des Moines to Chicago.kml,2016-05-27-22)
(16,12,59:30.5,Normal,-91.78,42.23,16|12|9223370572464805265,3660000000000000000,Tom McCuch,1961634315,Saint Louis to Memphis,2016-05-27-22)
(17,15,59:30.8,Normal,-92.03,38.97,17|15|9223370572464805005,3660000000000000000,Eric Mizell,1927624662,Springfield to KC Via Columbia,2016-05-27-22)
(23,68,59:30.8,Normal,-91.63,41.72,23|68|9223370572464804995,3660000000000000000,Adam Diaz,160405074,Joplin to Kansas City Route 2,2016-05-27-22)
(32,42,59:31.2,Normal,-92.09,34.8,32|42|9223370572464804645,3660000000000000000,Ryan Templeton,1090292248,Peoria to Ceder Rapids Route 2,2016-05-27-22)
(13,89,59:31.4,Normal,-90.2,41.59,13|89|9223370572464804436,3660000000000000000,Joe Niemiec,927636994,Des Moines to Chicago.kml,2016-05-27-22)
(24,97,59:31.4,Normal,-88.77,40.76,24|97|9223370572464804420,3660000000000000000,Don Hilborn,1090292248,Peoria to Ceder Rapids Route 2,2016-05-27-22)
(16,12,59:31.4,Normal,-91.63,41.72,16|12|9223370572464804385,3660000000000000000,Tom McCuch,1961634315,Saint Louis to Memphis,2016-05-27-22)
(29,66,59:31.6,Normal,-96.21,36.19,29|66|9223370572464804196,3660000000000000000,Teddy Choi,803014426,Wichita to Little Rock Route 2,2016-05-27-22)
(11,74,59:31.8,Normal,-89.17,40.38,11|74|9223370572464804026,3660000000000000000,Jamie Engesser,1567254452,Saint Louis to Memphis Route2,2016-05-27-22)
(19,26,59:32.1,Normal,-96.21,36.19,19|26|9223370572464803715,3660000000000000000,Ajay Singh,1962261785,Wichita to Little Rock.kml,2016-05-27-22)
(32,42,59:32.1,Normal,-91.93,34.81,32|42|9223370572464803676,3660000000000000000,Ryan Templeton,1090292248,Peoria to Ceder Rapids Route 2,2016-05-27-22)
(30,58,59:32.4,Normal,-89.88,41.77,30|58|9223370572464803425,3660000000000000000,Dan Rice,160779139,Des Moines to Chicago Route 2,2016-05-27-22)
(29,66,59:32.5,Normal,-95.99,36.17,29|66|9223370572464803336,3660000000000000000,Teddy Choi,803014426,Wichita to Little Rock Route 2,2016-05-27-22)
(26,57,59:32.5,Normal,-92.08,37.81,26|57|9223370572464803302,3660000000000000000,Michael Aube,1325712174,Saint Louis to Tulsa Route2,2016-05-27-22)
(28,39,59:33.0,Normal,-91.59,39.68,28|39|9223370572464802830,3660000000000000000,Olivier Renault,137128276,Springfield to KC Via Hanibal Route 2,2016-05-27-22)
(19,26,59:33.0,Normal,-95.99,36.17,19|26|9223370572464802825,3660000000000000000,Ajay Singh,1962261785,Wichita to Little Rock.kml,2016-05-27-22)
(21,109,59:33.1,Normal,-94.37,38.0,21|109|9223370572464802705,3660000000000000000,Jeff Markham,1594289134,Memphis to Little Rock Route 2,2016-05-27-22)
(29,66,59:33.2,Normal,-95.76,36.08,29|66|9223370572464802564,3660000000000000000,Teddy Choi,803014426,Wichita to Little Rock Route 2,2016-05-27-22)
(30,58,59:33.4,Normal,-89.6,41.76,30|58|9223370572464802454,3660000000000000000,Dan Rice,160779139,Des Moines to Chicago Route 2,2016-05-27-22)
(15,51,59:33.5,Normal,-91.74,34.89,15|51|9223370572464802316,3660000000000000000,Rohit Bakshi,1384345811,Joplin to Kansas City,2016-05-27-22)
(32,42,59:33.7,Normal,-91.38,34.83,32|42|9223370572464802114,3660000000000000000,Ryan Templeton,1090292248,Peoria to Ceder Rapids Route 2,2016-05-27-22)
(19,26,59:33.7,Normal,-95.76,36.08,19|26|9223370572464802066,3660000000000000000,Ajay Singh,1962261785,Wichita to Little Rock.kml,2016-05-27-22)
(14,25,59:34.0,Normal,-94.3,37.66,14|25|9223370572464801796,3660000000000000000,Adis Cesir,160405074,Joplin to Kansas City Route 2,2016-05-27-22)
(26,57,59:34.2,Normal,-92.48,37.8,26|57|9223370572464801635,3660000000000000000,Michael Aube,1325712174,Saint Louis to Tulsa Route2,2016-05-27-22)
(16,12,59:34.3,Normal,-91.05,41.72,16|12|9223370572464801496,3660000000000000000,Tom McCuch,1961634315,Saint Louis to Memphis,2016-05-27-22)
(15,51,59:34.4,Normal,-91.38,34.83,15|51|9223370572464801425,3660000000000000000,Rohit Bakshi,1384345811,Joplin to Kansas City,2016-05-27-22)
(10,85,59:34.6,Normal,-92.61,37.7,10|85|9223370572464801186,3660000000000000000,George Vetticaden,1390372503,Saint Louis to Tulsa,2016-05-27-22)
(31,18,59:34.7,Normal,-94.31,37.31,31|18|9223370572464801146,3660000000000000000,Rommel Garcia,1594289134,Memphis to Little Rock Route 2,2016-05-27-22)
(29,66,59:35.1,Normal,-95.42,35.87,29|66|9223370572464800713,3660000000000000000,Teddy Choi,803014426,Wichita to Little Rock Route 2,2016-05-27-22)
(15,51,59:35.1,Normal,-91.14,34.96,15|51|9223370572464800666,3660000000000000000,Rohit Bakshi,1384345811,Joplin to Kansas City,2016-05-27-22)
(12,104,59:35.3,Normal,-89.65,36.37,12|104|9223370572464800526,3660000000000000000,Paul Codding,24929475,Peoria to Ceder Rapids,2016-05-27-22)
(20,41,59:35.5,Normal,-88.96,42.25,20|41|9223370572464800335,3660000000000000000,Chris Harris,160779139,Des Moines to Chicago Route 2,2016-05-27-22)
(31,18,59:35.6,Normal,-94.46,37.16,31|18|9223370572464800225,3660000000000000000,Rommel Garcia,1594289134,Memphis to Little Rock Route 2,2016-05-27-22)
(27,105,59:35.6,Normal,-92.85,38.93,27|105|9223370572464800175,3660000000000000000,Mark Lochbihler,1325562373,Springfield to KC Via Columbia Route 2,2016-05-27-2
2)
(14,25,59:35.8,Normal,-94.46,37.16,14|25|9223370572464800006,3660000000000000000,Adis Cesir,160405074,Joplin to Kansas City Route 2,2016-05-27-22)
(26,57,59:35.9,Normal,-92.74,37.6,26|57|9223370572464799895,3660000000000000000,Michael Aube,1325712174,Saint Louis to Tulsa Route2,2016-05-27-22)
(18,16,59:36.3,Normal,-92.42,39.76,18|16|9223370572464799486,3660000000000000000,Grant Liu,1565885487,Springfield to KC Via Hanibal,2016-05-27-22)
2018-09-15 09:31:45,671 [main] INFO  org.apache.pig.Main - Pig script completed in 4 seconds and 177 milliseconds (4177 ms)
[root@sandbox-hdp ~]#
```

Select specific columns from a relation

Delete the `DESCRIBE truck_events` , `DESCRIBE truck_events_subset` and `DUMP truck_events_subset` commands from your Pig script; you will no longer need those. One of the

key uses of Pig is data transformation. You can define a new relation based on the fields of an existing relation using the `FOREACH` command. Define a new relation `specific_columns`, which will contain only the `driverId`, `eventTime` and `eventType` from relation `truck_events_subset`. Now the completed code is:

```
truck_events = LOAD '/user/maria_dev/truck_event_text_partition.csv' USING
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, eventTim
DESCRIBE specific_columns;
```

```
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, eventType;
DESCRIBE specific_columns;
~
~
~
~
~
:x
```

Save and execute the script and your output will look like the following:

```
[                ~]# pig -f Truck-Events
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
18/09/15 09:35:53 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
18/09/15 09:35:53 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
18/09/15 09:35:53 INFO pig.ExecTypeProvider: Trying ExecType : TEZ_LOCAL
18/09/15 09:35:53 INFO pig.ExecTypeProvider: Trying ExecType : TEZ
18/09/15 09:35:53 INFO pig.ExecTypeProvider: Picked TEZ as the ExecType
2018-09-15 09:35:53,827 [main] INFO  org.apache.pig.Main - Apache Pig version 0.16.0.3.0.0.0-1634 (rUnversioned directory) c
ompiled Jul 12 2018, 20:39:28
2018-09-15 09:35:53,827 [main] INFO  org.apache.pig.Main - Logging error messages to: /root/pig_1537004153826.log
2018-09-15 09:35:55,085 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /root/.pigbootup not found
2018-09-15 09:35:55,263 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop f
ile system at: hdfs://sandbox-hdp.hortonworks.com:8020
2018-09-15 09:35:57,230 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-Truck-Events-48ce07d6-6c2
b-45e1-b043-84ac347cc90a
2018-09-15 09:35:57,230 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to f
alse
specific_columns: {driverId: int,eventTime: chararray,eventType: chararray}
2018-09-15 09:36:03,291 [main] INFO  org.apache.pig.Main - Pig script completed in 9 seconds and 857 milliseconds (9857 ms)
[                ~]#
```

Store relationship data into a HDFS File

In this step, you will use the `STORE` command to output a relation into a new file in HDFS. Enter the

following command to output the `specific_columns` relation to a folder named `output/specific_columns` :

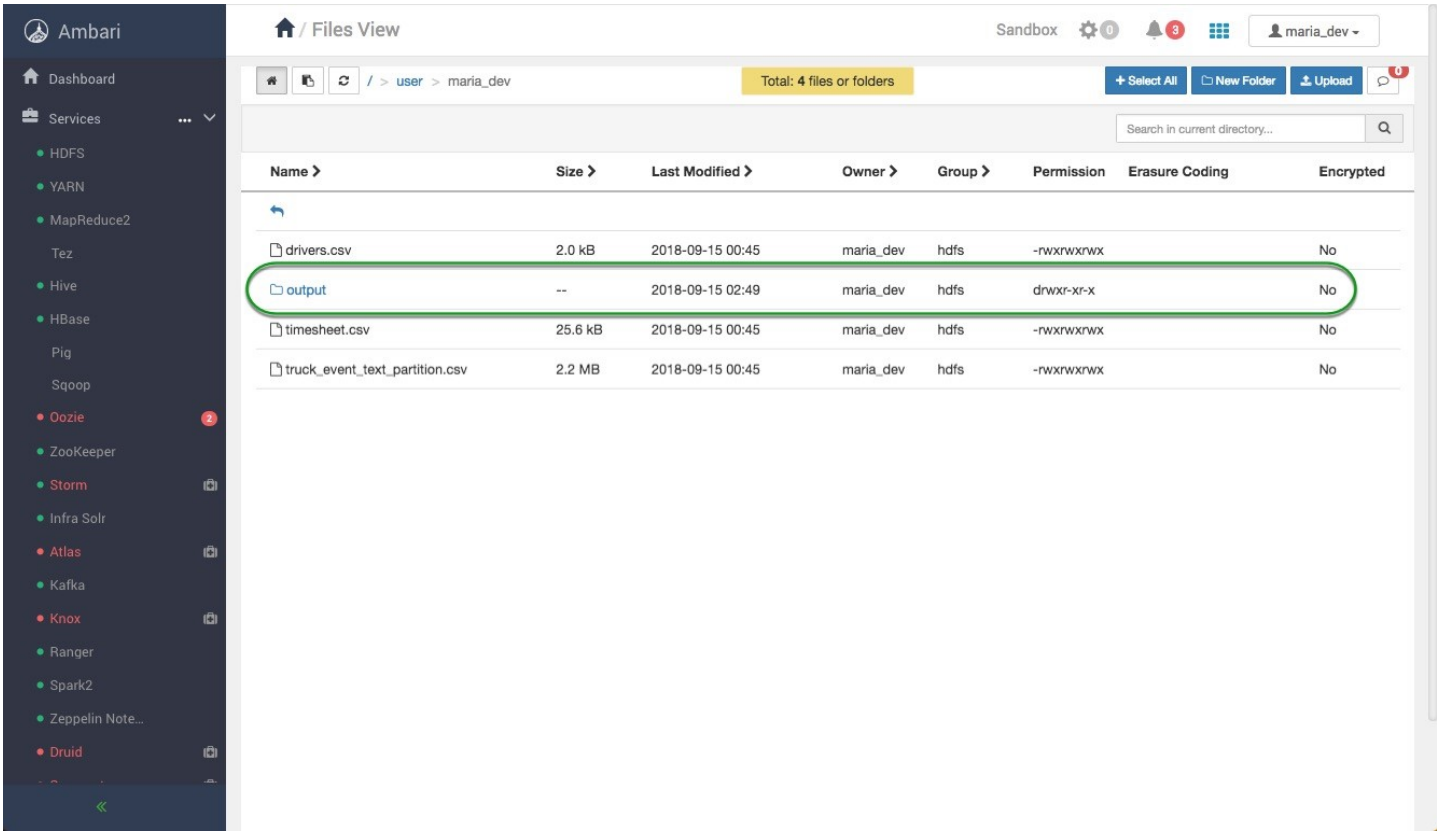
```
STORE specific_columns INTO 'output/specific_columns' USING PigStorage(','
```

```
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, eventType;
DESCRIBE specific_columns;
STORE specific_columns INTO 'output/specific_columns' USING PigStorage(',');
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
:x
```

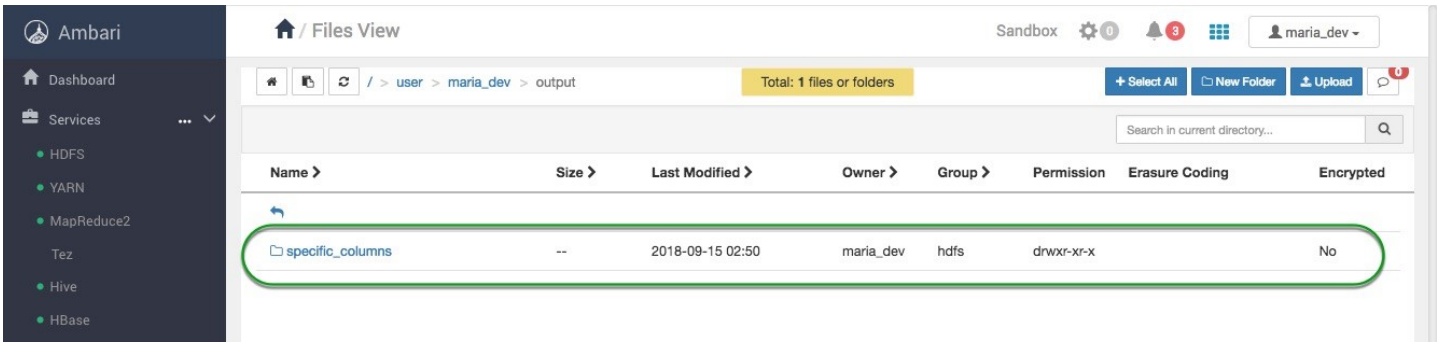
Save and Execute the script. Again, this requires a MapReduce job (just like the DUMP command), so you will need to wait a minute for the job to complete.

Once the job is finished, go to `HDFS Files view` and look for a newly created folder called `"output"` under `/user/maria_dev` :

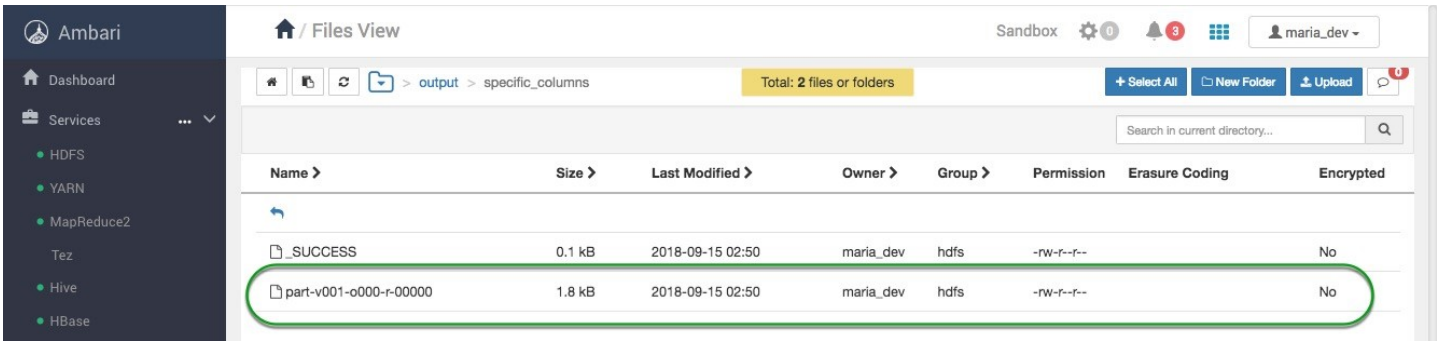
> **Note:** If you didn't use the default path above, then the new folder will exist in the path you created.



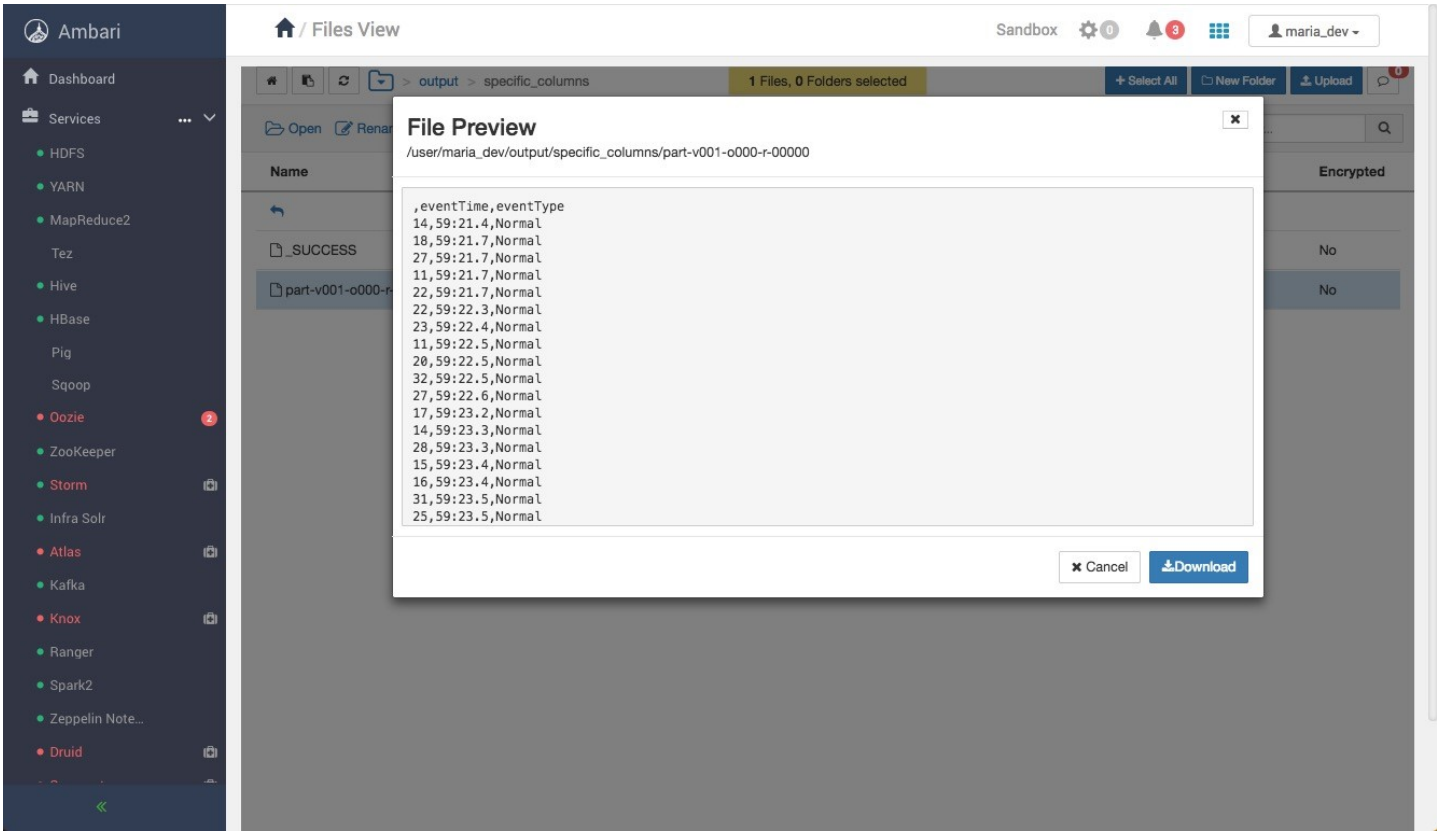
Click on "output" folder. You will find a sub-folder named "specific_columns" .



Click on "specific_columns" folder. You will see an output file called "part-r-00000" :



Click on the file "part-r-00000" and then click on Open :



Perform a join between 2 relations

In this step, you will perform a `join` on two driver statistics data sets: `truck_event_text_partition.csv` and the `driver.csv` files. Drivers.csv has all the details for the driver like driverId, name, ssn, location, etc.

You have already defined a relation for the events named truck_events. Create a new Pig script named `"Pig-Join"` . Then define a new relation named `drivers` then join truck_events and drivers by `driverId` and describe the schema of the new relation `join_data` . The completed code will be:

```
truck_events = LOAD '/user/maria_dev/truck_event_text_partition.csv' USING
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;
```

```
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;
~
~
~
~
~
~
~
~
~
:x
```

Save the script and execute it:

```
pig -f Truck-Events | tee -a joinAttributes.txt
cat joinAttributes.txt
```

```
[maria_dev@sandbox-hdp ~]$ cat joinAttributes.txt
join_data: {truck_events::driverId: int,truck_events::truckId: int,truck_events::eventTime: chararray,t
ruck_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::routeId: long,truck_events::routeName: chararray,truck_events::eventDate: chararray,drivers::d
riverId: int,drivers::name: chararray,drivers::ssn: chararray,drivers::location: chararray,drivers::cer
tified: chararray,drivers::wage_plan: chararray}
[maria_dev@sandbox-hdp ~]$
```

Notice `join_data` contains all the fields of both `truck_events` and `drivers` .

Sort the data using “ORDER BY”

Use the `ORDER BY` command to sort a relation by one or more of its fields. Create a new Pig script named `"Pig-Sort"` from maria_dev home directory enter:

```
vi Pig-Sort
```

Next, enter the following commands to sort the drivers data by name then date in ascending order:

```
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;
```

Save and execute the script. Your output should be sorted as shown here:

```
(23,Adam Diaz,928312208,P.O. Box 260- 6127 Vitae Road,Y,hours)
(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,Nadeem Asghar,783204269,154-9147 Aliquam Ave,Y,hours)
(40,Nicolas Maillard,208510217,1027 Quis Rd.,Y,hours)
(28,Olivier Renault,959908181,P.O. Box 243- 6509 Erat. Avenue,Y,hours)
(12,Paul Coddin,198041975,Ap #622-957 Risus. Street,Y,hours)
(42,Randy Gelhausen,853302254,145-4200 In- Avenue,Y,hours)
(15,Rohit Bakshi,239005227,648-5681 Dui- Rd.,Y,hours)
(31,Rommel Garcia,858912101,P.O. Box 945- 6015 Sociis St.,Y,hours)
(32,Ryan Templeton,290304287,765-6599 Egestas. Av.,Y,hours)
(38,Scott Shaw,386411175,276 Lobortis Road,Y,hours)
(33,Sridhara Sabbella,967409015,Ap #477-2507 Sagittis Avenue,Y,hours)
(29,Teddy Choi,185502192,P.O. Box 106- 7003 Amet Rd.,Y,hours)
(16,Tom McCuch,363303105,P.O. Box 313- 962 Parturient Rd.,Y,hours)
(37,Wes Floyd,190504074,P.O. Box 269- 9611 Nulla Street,Y,hours)
```

Filter and Group the data using “GROUP BY”

The `GROUP` command allows you to group a relation by one of its fields. Create a new Pig script named `“Pig-Group”` . Then, enter the following commands, which group the `truck_events` relation by the `driverId` for the `eventType` which are not ‘Normal’.

```
truck_events = LOAD '/user/maria_dev/truck_event_text_partition.csv' USING
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);
filtered_events = FILTER truck_events BY NOT (eventType MATCHES 'Normal');
grouped_events = GROUP filtered_events BY driverId;
DESCRIBE grouped_events;
DUMP grouped_events;
```

```
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);
filtered_events = FILTER truck_events BY NOT (eventType MATCHES 'Normal');
grouped_events = GROUP filtered_events BY driverId;
DESCRIBE grouped_events;
DUMP grouped_events;
~
~
~
~
~
:x
```

Save and execute the script. Notice that the data for eventType which are not Normal is grouped together for each driverId.

```
(10,{(10,85,59:46.9,Overspeed,-95.5,36.37,10|85|922337057246478896,366000000000000000,George Vetticaden,1390372503,Saint Louis to T
ulsa,2016-05-27-22),(10,85,00:39.7,Overspeed,-94.23,37.09,10|85|9223370572464736126,366000000000000000,George Vetticaden,1390372503,
Saint Louis to Tulsa,2016-05-27-22),(10,85,00:13.1,Unsafe tail distance,-91.18,38.22,10|85|9223370572464762694,366000000000000000,Ge
orge Vetticaden,1390372503,Saint Louis to Tulsa,2016-05-27-22)})
(11,{(11,74,59:56.4,Lane Departure,-87.67,41.87,11|74|9223370572464779456,366000000000000000,Jamie Engesser,1567254452,Saint Louis t
o Memphis Route2,2016-05-27-22),(11,74,59:47.3,Unsafe tail distance,-89.63,39.84,11|74|9223370572464788546,366000000000000000,Jamie
Engesser,1567254452,Saint Louis to Memphis Route2,2016-05-27-22),(11,74,00:05.4,Unsafe following distance,-89.74,39.1,11|74|922337057
2464770396,366000000000000000,Jamie Engesser,1567254452,Saint Louis to Memphis Route2,2016-05-27-22),(11,74,00:14.1,Lane Departure,-
88.77,40.76,11|74|9223370572464761716,366000000000000000,Jamie Engesser,1567254452,Saint Louis to Memphis Route2,2016-05-27-22),(11,
74,00:23.1,Unsafe tail distance,-88.42,41.11,11|74|9223370572464752715,366000000000000000,Jamie Engesser,1567254452,Saint Louis to M
emphis Route2,2016-05-27-22),(11,74,00:32.0,Unsafe tail distance,-90.2,38.65,11|74|9223370572464743846,366000000000000000,Jamie Enge
sser,1567254452,Saint Louis to Memphis Route2,2016-05-27-22),(11,74,59:38.0,Unsafe tail distance,-89.17,40.38,11|74|92233705724647977
96,366000000000000000,Jamie Engesser,1567254452,Saint Louis to Memphis Route2,2016-05-27-22),(11,74,59:29.1,Overspeed,-88.07,41.48,1
1|74|9223370572464806746,366000000000000000,Jamie Engesser,1567254452,Saint Louis to Memphis Route2,2016-05-27-22),(11,74,00:49.6,La
ne Departure,-89.71,37.47,11|74|9223370572464726246,366000000000000000,Jamie Engesser,1567254452,Saint Louis to Memphis Route2,2016-
05-27-22),(11,74,00:41.0,Lane Departure,-90.07,35.68,11|74|9223370572464734786,366000000000000000,Jamie Engesser,1567254452,Saint Lo
uis to Memphis Route2,2016-05-27-22)})
(12,{(12,104,00:47.6,Unsafe following distance,-90.0,37.72,12|104|9223370572464728186,366000000000000000,Paul Codding,24929475,Peori
a to Ceder Rapids,2016-05-27-22)})
(13,{(13,89,00:47.7,Lane Departure,-89.03,41.92,13|89|9223370572464728156,366000000000000000,Joe Niemiec,927636994,Des Moines to Chi
cago.kml,2016-05-27-22)})
(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)})
(15,{(15,51,00:48.8,Lane Departure,-90.04,35.19,15|51|9223370572464727025,366000000000000000,Rohit Bakshi,1384345811,Joplin to Kansa
s 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 Mem
phis,2016-05-27-22)})
(17,{(17,15,00:48.4,Lane Departure,-90.79,38.83,17|15|9223370572464727374,366000000000000000,Eric Mizell,1927624662,Springfield to K
C Via Columbia,2016-05-27-22)})
(18,{(18,16,00:47.2,Overspeed,-94.28,39.53,18|16|9223370572464728575,366000000000000000,Grant Liu,1565885487,Springfield to KC Via H
anibal,2016-05-27-22)})
(19,{(19,26,00:48.6,Unsafe following distance,-94.57,35.37,19|26|9223370572464727224,366000000000000000,Ajay Singh,1962261785,Wichit
a to Little Rock.kml,2016-05-27-22)})
(20,{(20,41,00:46.9,Overspeed,-89.03,41.92,20|41|9223370572464728915,366000000000000000,Chris Harris,160779139,Des Moines to Chicago
 Route 2,2016-05-27-22)})
(21,{(21,109,00:46.8,Unsafe tail distance,-88.07,41.48,21|109|9223370572464729016,366000000000000000,Jeff Markham,1594289134,Memphis
 to Little Rock Route 2,2016-05-27-22)})
(22,{(22,87,00:46.5,Unsafe tail distance,-90.04,35.19,22|87|92233705724647292[illegible]6,366000000000000000,Nadeem Asghar,1198242881, Saint L
ouis to Chicago Route2,2016-05-27-22)})
(23,{(23,68,00:47.8,Lane Departure,-89.52,40.7,23|68|9223370572464727994,366000000000000000,Adam Diaz,160405074,Joplin to Kansas Cit
y Route 2,2016-05-27-22)})
(24,{(24,97,00:48.6,Lane Departure,-89.17,40.38,24|97|9223370572464727226,366000000000000000,Don Hilborn,1090292248,Peoria to Ceder
Rapids Route 2,2016-05-27-22)})
(25,{(25,96,00:40.1,Overspeed,-89.54,36.84,25|96|9223370572464735726,366000000000000000,Jean-Philippe Player,371182829,Memphis to Li
```

The results of the Pig Job will show all non-Normal events grouped under each driverId.

Congratulations! You have successfully completed the tutorial and well on your way to pigging on Big Data.

Further Reading

Apache Pig (/tutorials/how-to-process-data-with-apache-pig.html)

Welcome to Apache Pig! (https://pig.apache.org/)

Pig Latin Basics (https://pig.apache.org/docs/r0.12.0/basic.html#store)

(https://www.facebook.com/cloudera/) **(https://twitter.com/cloudera)** **(https://www.linkedin.com/company/cloudera)**

Partners (/partners.html)

Resources (/resources.html)

Community (https://cloudera-production.okta.com/app/template_saml_2_0/exk1az52keQqZWEnE0x7/sso/saml?RelayState=http%3A%2F%2Fcommunity.cloudera.com%2F)

Documentation (http://docs.cloudera.com)

Careers (/about/careers.html)

Contact Us (/contact-sales.html)

US: +1 888 789 1488 (tel:18887891488)
Outside the US: +1 650 362 0488 (tel:16503620488)

English

© 2020 Cloudera, Inc. All rights reserved. **Terms & Conditions (/legal/terms-and-conditions.html) | Privacy Policy and Data Policy (/legal/policies.html) | Unsubscribe / Do Not Sell My Personal Information (/unsubscribe.html)**
Apache Hadoop (http://hadoop.apache.org/) and associated open source project names are trademarks of the **Apache Software Foundation (http://apache.org/)**. For a complete list of trademarks, **click here (/legal/terms-and-conditions.html#trademarks)**.