

1) Download and install Pig:

```
cd
wget http://rasinsrv07.cstcis.cti.depaul.edu/CSC555/pig-0.15.0.tar.gz
gunzip pig-0.15.0.tar.gz
tar xvf pig-0.15.0.tar
```

set the environment variables (this can also be placed in ~/.bashrc to make it permanent)

```
export PIG_HOME=/home/ec2-user/pig-0.15.0
export PATH=$PATH:$PIG_HOME/bin
```

Use the same vehicles file. Copy the vehicles.csv file to the HDFS if it is not already there.

Now run pig (and use the pig home variable we set earlier):

```
cd $PIG_HOME
bin/pig
```

Create the same table as what we used in Hive, assuming that vehicles.csv is in the home directory on HDFS:

```
VehicleData = LOAD '/user/ec2-user/vehicles.csv' USING PigStorage(',')
AS (barrels08:FLOAT, barrelsA08:FLOAT, charge120:FLOAT, charge240:FLOAT, city08:FLOAT);
```

You can see the table description by

```
DESCRIBE VehicleData;
```

Verify that your data has loaded by running:

```
VehicleG = GROUP VehicleData ALL;
Count = FOREACH VehicleG GENERATE COUNT(VehicleData);
DUMP Count;
```

How many rows did you get? (if you get an error here, it is likely because vehicles.csv is not in HDFS)

Create the same ThreeColExtract file that you have in the previous assignment, by placing barrels08, city08 and charge120 into a new file using PigStorage. You want the STORE command to record output in HDFS. (discussed in p457, Pig Chapter, "Data Processing Operator section")

NOTE: You can use this to get one column:

```
OneCol = FOREACH VehicleData GENERATE barrels08;
```

Verify that the new file has been created and report the size of the newly created file. (you can use **quit** to exit the grunt shell)

Submit a single document containing your written answers. Be sure that this document contains your name and “CSC 555 Assignment 3” at the top.

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder