Eric Johnson / Ching-Yung Lin
efj2106@columbia.edu

PIG Installation

1) Download PIG Binary

1a) Change to your Hadoop installation director and download the compiled PIG binary

```
cd /usr/local/Cellar/hadoop/2.7.2

wget http://apache.spinellicreations.com/pig/pig-0.16.0/pig-0.16.0.tar.gz
```

1a) Unpackge PIG

```
sudo tar -xzvf pig-0.16.0.tar.gz
```

1b) Set you path for PIG

```
export PATH=/<my-path-to-pig>/pig-n.n.n/bin:$PATH

for mac OSX

export HADOOP_HOME=/usr/local/Cellar/hadoop/2.7.2

export PATH=/usr/local/Cellar/hadoop/2.7.2/pig-0.16.0/bin:$PATH

export JAVA_HOME=/Library/Java/JavaVirtualMachines/jdk1.8.0_101.jdk/Contents/Home
```

1c) Download Sample Dataset

```
wget https://github.com/hortonworks/tutorials/raw/hdp-2.5/driver_data.zip

unzip driver_data.zip

./bin/hadoop fs -mkdir /user/pig_example

./bin/hadoop fs -put /driver_data/truck_event_text_partition.csv /user/pig_example

./bin/hadoop fs -put /driver_data/truck_event_text_partition.csv /user/pig_example
```

2) PIG Examples

2a) Load truck data and define a schema to use in PIG

```
truck_events = LOAD '/user/pig_example/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;
```

2b) Load truck data and define a schema to use in PIG

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

2c) Describe a subset of the data (specific columns)

```
specific_columns = FOREACH truck_events_subset GENERATE driverId, eventTime, eventType;

DESCRIBE specific_columns;
```

2d) Store a subset into an HDFS location

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

2e) Perform a join using multiple tables

```
truck_events = LOAD '/user/pig_example/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);
```

Eric Johnson / Ching-Yung Lin
efj2106@columbia.edu

```
drivers =  LOAD '/user/pig_example/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;
```

2e) Perform a calculation using GROUP BY

```
truck_events = LOAD '/user/pig_example/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;
```