% Running the Following Using Command Line

% Format HDFS and Make New Directory
./bin/hdfs dfs –mkdir /user
./bin/hdfs dfs –mkdir /user/uber

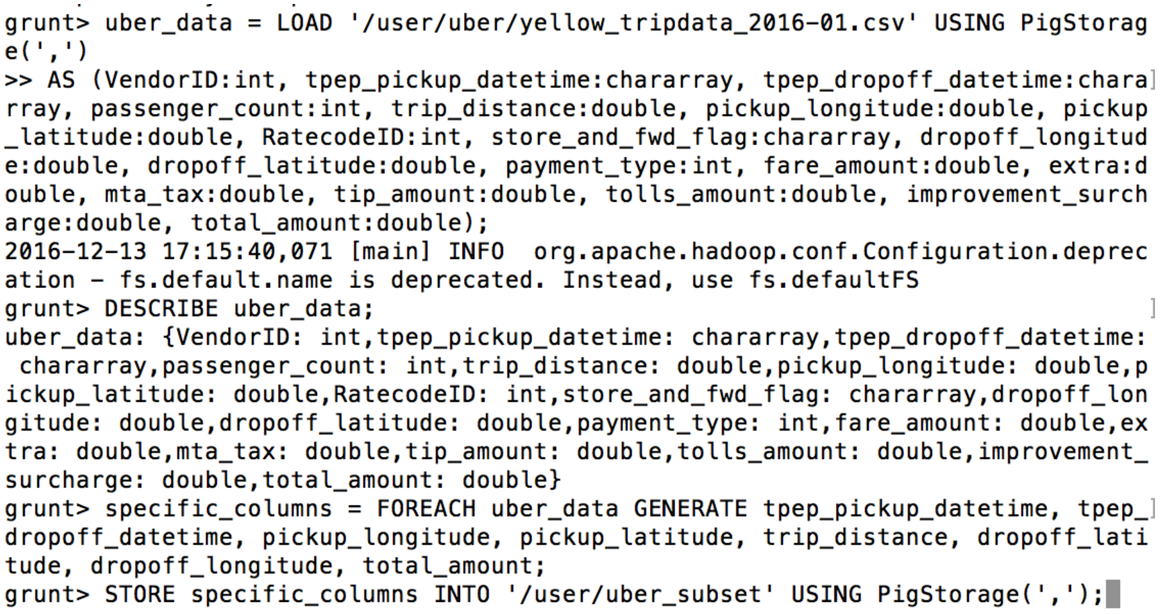
% Check Hadoop FileSystem at:
%Resource Manager: http://localhost:50070

% Upload Dataset
./bin/hadoop fs -put /Users/jingyiyuan/Downloads/yellow_tripdata_2016-06.csv /user/uber

% Starting PIG
PIG

% Running the Following Using PIG

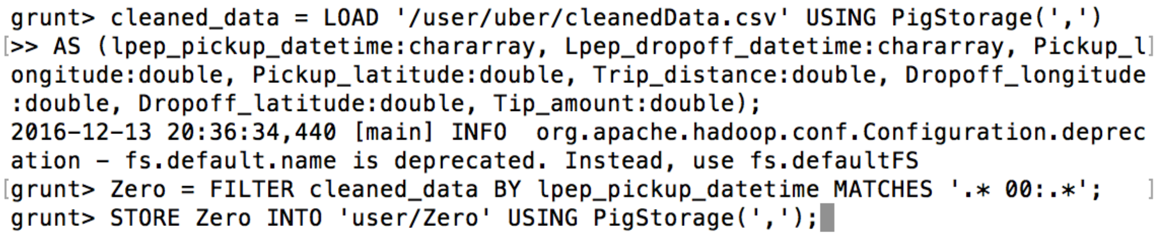
% Clean the Dataset and Get the Needed Features
uber_data=LOAD ‘usr/uber/yellow_tripdata_2016-01.csv’ USING PigStorage(‘,’)
DESCRIBE uber_data;
specific_columns=FOREACH uber_data GENERATE tpep_pickup_datetime, tpep_dropoff_datetime, pickup_longitude, pickup_latitude, trip_distance, dropoff_latitude, dropoff_longitude, total_amount;
STORE specific_columns INTO '/user/uber_subset' USING PigStorage(',');



```
grunt> uber_data = LOAD '/user/uber/yellow_tripdata_2016-01.csv' USING PigStorag
e(',')
>> AS (VendorID:int, tpep_pickup_datetime:chararray, tpep_dropoff_datetime:chara
rray, passenger_count:int, trip_distance:double, pickup_longitude:double, pickup
_latitude:double, RatecodeID:int, store_and_fwd_flag:chararray, dropoff_longitud
e:double, dropoff_latitude:double, payment_type:int, fare_amount:double, extra:d
ouble, mta_tax:double, tip_amount:double, tolls_amount:double, improvement_surch
arge:double, total_amount:double);
2016-12-13 17:15:40,071 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> DESCRIBE uber_data;
uber_data: {VendorID: int,tpep_pickup_datetime: chararray,tpep_dropoff_datetime:
 chararray,passenger_count: int,trip_distance: double,pickup_longitude: double,p
ickup_latitude: double,RatecodeID: int,store_and_fwd_flag: chararray,dropoff_lon
gitude: double,dropoff_latitude: double,payment_type: int,fare_amount: double,ex
tra: double,mta_tax: double,tip_amount: double,tolls_amount: double,improvement_
surcharge: double,total_amount: double}
grunt> specific_columns = FOREACH uber_data GENERATE tpep_pickup_datetime, tpep_
dropoff_datetime, pickup_longitude, pickup_latitude, trip_distance, dropoff_lati
tude, dropoff_longitude, total_amount;
grunt> STORE specific_columns INTO '/user/uber_subset' USING PigStorage(',');
```

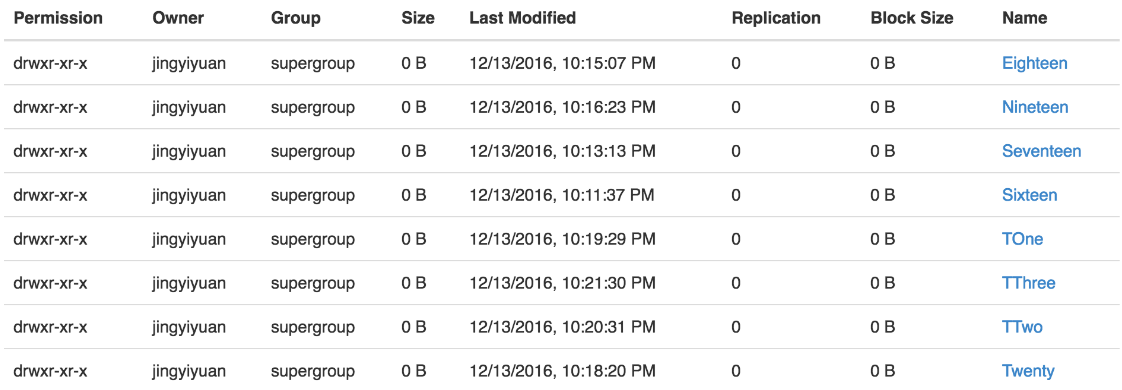
```
% splitting the data into 24 subsets
cleaned_data = LOAD '/user/uber/cleanedData.csv' USING PigStorage(',')
AS (lpep_pickup_datetime:chararray, Lpep_dropoff_datetime:chararray,
Pickup_longitude:double, Pickup_latitude:double, Trip_distance:double,
Dropoff_longitude:double, Dropoff_latitude:double, Tip_amount:double);

Zero = FILTER cleaned_data BY lpep_pickup_datetime MATCHES '.* 00:.*';
STORE Zero INTO 'user/Zero' USING PigStorage(',');

One = FILTER cleaned_data BY lpep_pickup_datetime MATCHES '.* 01:.*';
STORE One INTO 'user/One' USING PigStorage(',');

Two = FILTER cleaned_data BY lpep_pickup_datetime MATCHES '.* 02:.*';
STORE Two INTO 'user/Two' USING PigStorage(',');

Three = FILTER cleaned_data BY lpep_pickup_datetime MATCHES '.* 03:.*';
STORE Three INTO 'user/Three' USING PigStorage(',');

Four = FILTER cleaned_data BY lpep_pickup_datetime MATCHES '.* 04:.*';
STORE Four INTO 'user/Four' USING PigStorage(',');

Five = FILTER cleaned_data BY lpep_pickup_datetime MATCHES '.* 05:.*';
STORE Five INTO 'user/Five' USING PigStorage(',');

Six = FILTER cleaned_data BY lpep_pickup_datetime MATCHES '.* 06:.*';
STORE Six INTO 'user/Six' USING PigStorage(',');

Seven = FILTER cleaned_data BY lpep_pickup_datetime MATCHES '.* 07:.*';
STORE Seven INTO 'user/Seven' USING PigStorage(',');

Eight = FILTER cleaned_data BY lpep_pickup_datetime MATCHES '.* 08:.*';
STORE Eight INTO 'user/Eight' USING PigStorage(',');

Nine = FILTER cleaned_data BY lpep_pickup_datetime MATCHES '.* 09:.*';
STORE Nine INTO 'user/Nine' USING PigStorage(',');

Ten = FILTER cleaned_data BY lpep_pickup_datetime MATCHES '.* 10:.*';
STORE Ten INTO 'user/Ten' USING PigStorage(',');
```

```
Eleven = FILTER cleaned_data BY lpep_pickup_datetime MATCHES '.* 11:.*';
STORE Eleven INTO 'user/Eleven' USING PigStorage(',');

Twelve = FILTER cleaned_data BY lpep_pickup_datetime MATCHES '.* 12:.*';
STORE Twelve INTO 'user/Twelve' USING PigStorage(',');

Thirteen = FILTER cleaned_data BY lpep_pickup_datetime MATCHES '.* 13:.*';
STORE Thirteen INTO 'user/Thirteen' USING PigStorage(',');

Fourteen = FILTER cleaned_data BY lpep_pickup_datetime MATCHES '.* 14:.*';
STORE Fourteen INTO 'user/Fourteen' USING PigStorage(',');

Fifteen = FILTER cleaned_data BY lpep_pickup_datetime MATCHES '.* 15:.*';
STORE Fifteen INTO 'user/Fifteen' USING PigStorage(',');

Sixteen = FILTER cleaned_data BY lpep_pickup_datetime MATCHES '.* 16:.*';
STORE Sixteen INTO 'user/Sixteen' USING PigStorage(',');

Seventeen = FILTER cleaned_data BY lpep_pickup_datetime MATCHES '.* 17:.*';
STORE Seventeen INTO 'user/Seventeen' USING PigStorage(',');

Eighteen = FILTER cleaned_data BY lpep_pickup_datetime MATCHES '.* 18:.*';
STORE Eighteen INTO 'user/Eighteen' USING PigStorage(',');

Nineteen = FILTER cleaned_data BY lpep_pickup_datetime MATCHES '.* 19:.*';
STORE Nineteen INTO 'user/Nineteen' USING PigStorage(',');

Twenty = FILTER cleaned_data BY lpep_pickup_datetime MATCHES '.* 20:.*';
STORE Twenty INTO 'user/Twenty' USING PigStorage(',');

TOne = FILTER cleaned_data BY lpep_pickup_datetime MATCHES '.* 21:.*';
STORE TOne INTO 'user/TOne' USING PigStorage(',');

TTwo = FILTER cleaned_data BY lpep_pickup_datetime MATCHES '.* 22:.*';
STORE TTwo INTO 'user/TTwo' USING PigStorage(',');

TThree = FILTER cleaned_data BY lpep_pickup_datetime MATCHES '.* 23:.*';
STORE TThree INTO 'user/TThree' USING PigStorage(',');
```



```
grunt> cleaned_data = LOAD '/user/uber/cleanedData.csv' USING PigStorage(',')
[>> AS (lpep_pickup_datetime:chararray, Lpep_dropoff_datetime:chararray, Pickup_l]
ongitude:double, Pickup_latitude:double, Trip_distance:double, Dropoff_longitude
:double, Dropoff_latitude:double, Tip_amount:double);
2016-12-13 20:36:34,440 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
[grunt> Zero = FILTER cleaned_data BY lpep_pickup_datetime MATCHES '.* 00:.*';    ]
grunt> STORE Zero INTO 'user/Zero' USING PigStorage(',');
```

% Download the 24 Subsets From Hadoop File System



Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
drwxr-xr-x	jingyiyuan	supergroup	0 B	12/13/2016, 10:15:07 PM	0	0 B	Eighteen
drwxr-xr-x	jingyiyuan	supergroup	0 B	12/13/2016, 10:16:23 PM	0	0 B	Nineteen
drwxr-xr-x	jingyiyuan	supergroup	0 B	12/13/2016, 10:13:13 PM	0	0 B	Seventeen
drwxr-xr-x	jingyiyuan	supergroup	0 B	12/13/2016, 10:11:37 PM	0	0 B	Sixteen
drwxr-xr-x	jingyiyuan	supergroup	0 B	12/13/2016, 10:19:29 PM	0	0 B	TOne
drwxr-xr-x	jingyiyuan	supergroup	0 B	12/13/2016, 10:21:30 PM	0	0 B	TThree
drwxr-xr-x	jingyiyuan	supergroup	0 B	12/13/2016, 10:20:31 PM	0	0 B	TTwo
drwxr-xr-x	jingyiyuan	supergroup	0 B	12/13/2016, 10:18:20 PM	0	0 B	Twenty