Project - Pig

Summary: Using multiple datasets, the data was filtered in Pig to yield certain results. The first result shows the merging/joining of certain fields within the datasets. In the second case, the data was filtered to yield the highest number of orders. Lastly, a pig script was written to find the hour of the day when the highest number of tweets were generated by users on a certain day (in this case, a random day was chosen, March 6, 2010).

 # Retrieving the data from multiple files in Pig and joining data to extract certain information. Below is the script to complete this task.

```
grunt> a = load '/user/pig/Customer.txt'using PigStorage ('\t')  AS (CustomerID:chararray, Cu
stomerName:chararray, ContactName:chararray, Address:chararray, Country:chararray);
```

```
grunt> b = load '/user/pig/Orders.txt' using PigStorage (' ')  AS (OrderID:chararray, Item:ch
ararray, OrderDate:chararray, CustomerID:chararray, ShipperID:chararray);
```

```
grunt> c = load '/user/pig/Shipper.txt' using PigStorage (',')  AS (ShipperID:chararray, Name
:chararray);
```

```
grunt> ab = join a by CustomerID, b by CustomerID;
grunt> abc = join ab by ShipperID, c by ShipperID;
grunt> abcfinal = foreach abc generate b::OrderID..b::OrderDate, a::CustomerName,c::Name;
grunt> dump abcfinal;
```

Result:

The result below shows the Order ID, Item, Order Date merged with Customer Name and Name of the Shipper.

```
(001,Watch,01/01/2016,JackPiper,FastShipper)
(002,TV,02/01/2016,JackPiper,ShipperOnTime)
(003,DVDPlayer,03/01/2016,JackPiper,PerfectShipper)
(004,BlueRayPlayer,04/01/2016,ArchieLee,Purolator)
(005,Iphone,05/01/2016,ArchieLee,Fedex)
(006,IMac,06/01/2016,BenJones,UPS)
(007,IPod,07/01/2016,JohnSmith,CanadaPost)
```

 # Below is the script to find the top 2 customers are found with the highest number of orders (based on customer ID and number of orders).

```
grunt> abcg = group b by CustomerID;
grunt> abcgg = foreach abcg generate group as CustomerID, COUNT (b) as cnt;
grunt> abcggo = order abcgg by cnt desc;
grunt> abcggol = limit abcggo 2;
grunt> dump abcggol;
```

Result:

The top 2 customers are shown below with their Customer ID followed by the number of orders.

```
(12,3)
(13,2)
```

 # Below is the script where the hour of the day is determined when the highest number of tweets were generated on March 6 2010.

```
grunt> a = load '/user/pig/full_text.txt' AS (id:chararray, ts:chararray, location:chararray,
 lat:float, lon:float, tweet:chararray);
grunt> b = foreach a generate id, ts, SUBSTRING(ts,0,10) as date;
grunt> c = filter b by date == '2010-03-06';
grunt> d = foreach c generate id, SUBSTRING(ts,11,13) as hour;
grunt> e = group d by hour;
grunt> f = foreach e generate group as hour, COUNT(d) as cnt;
grunt> g = order f by cnt desc;
grunt> h = limit g 1;
grunt> dump h;
```

Result:

Hence at 0100 hours, there were 3913 tweets (which were the highest on March 6, 2010).

```
(01,3913)
```