Case Study on Flight Data
Some insights on the U.S. Airline data using Apache Pig

Understanding Data
The U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics (BTS) tracks the on-time performance of domestic flights operated by large air carriers. Summary information on the number of on-time, delayed, cancelled, and diverted flights appears in DOT's monthly Air Travel Consumer Report, published about 30 days after the month's end.

The data format is comma separated values. These are 2 different datasets, i.e., delayedflights.csv and airports.csv. Let us understand one at a time.
1. *delayedflights.csv*
2. *airports.csv*

There are 29 columns in delayedflights.csv dataset. Please check flight_description file for details about schema/fields.
For *airports.csv*
• iata: the international airport abbreviation code
• name of the airport
• city and country in which airport is located.
• lat and long: the latitude and longitude of the airport

Some exploration ideas with Pig:
1. Find out the top 5 most visited destinations.
2. Which month has seen the most number of cancellations due to bad weather?
3. Top ten origins with the highest AVG departure delay
4. Which route (origin & destination) has seen the maximum diversion?

Running Pig
Pig contains multiple modes that can be specified to configure how Pig scripts and Pig statements will be executed.

Execution Modes
Pig has two execution modes: local and MapReduce.
Running Pig in local mode only requires a single machine. Pig will run on the local host and access the local filesystem. To run Pig in local mode, use the -x local flag:
$ pig -x local ...

Running Pig in MapReduce mode requires access to a Hadoop cluster. MapReduce mode executes Pig statements and jobs on the cluster and accesses HDFS. To run Pig in MapReduce mode, simply call Pig from the command line or use the -x

mapreduce flag:
$ pig ...
or
$ pig -x mapreduce

1) Find out the top 5 most visited destinations.
Answer :

```
flight_data = load '/user/support1161/delayedflights.csv' USING PigStorage(',');
gen_flight_data = foreach flight_data generate (int)$1 as year, (int)$10
as flight_num, (chararray)$17 as origin,(chararray) $18 as dest;
fiter_flight_not_null = filter gen_flight_data by dest is not null;
grp_dest = group fiter_flight_not_null by dest;
gen_count_dest = foreach grp_dest generate group,
COUNT(fiter_flight_not_null.dest);
ord_count_desc = order gen_count_dest by $1 DESC;
lmt_dest_cnt = LIMIT ord_count_desc 5;
airport_data = load '/user/support1161/airports.csv' USING PigStorage(',');
gen_airport_data = foreach airport_data generate (chararray)$0 as dest,
(chararray)$2 as city, (chararray)$4 as country;
joined_table = join lmt_dest_cnt by $0, gen_airport_data by dest;
dump joined_table;
```

2) Which month has seen the most number of cancellations due to bad weather?
Answer :

```
flight_data = load '/user/support1161/delayedflights.csv' USING PigStorage(',');
gen_flight_data = foreach flight_data generate (int)$2 as month,(int)$10
as flight_num,(int)$22 as cancelled,(chararray)$23 as cancel_code;
fltr_data = filter gen_flight_data by cancelled == 1 AND cancel_code =='B';
grp_mnth = group fltr_data by month;
gen_grp = foreach grp_mnth generate group, COUNT(fltr_data.cancelled);
ord_yr= order gen_grp by $1 DESC;
Result = limit ord_yr 1;
dump Result;
```

3) Top ten origins with the highest AVG departure delay
Answer :

```
A = load '/user/support1161/delayedflights.csv' USING PigStorage(',');
B1 = foreach A generate (int)$16 as dep_delay, (chararray)$17 as origin;
```

```
C1 = filter B1 by (dep_delay is not null) AND (origin is not null);
D1 = group C1 by origin;
E1 = foreach D1 generate group, AVG(C1.dep_delay);
Result = order E1 by $1 DESC;
Top_ten = limit Result 10;
Lookup = load '/user/support1161/airports.csv' USING PigStorage(',');
Lookup1 = foreach Lookup generate (chararray)$0 as origin, (chararray)$2 as
city, (chararray)$4 as country;
Joined = join Lookup1 by origin, Top_ten by $0;
Final = foreach Joined generate $0,$1,$2,$4;
Final_Result = ORDER Final by $3 DESC;
dump Final_Result;
```

4) Which route (origin & destination) has seen the maximum diversion?

Answer :

```
A = load '/user/support1161/delayedflights.csv' USING PigStorage(',');
B = FOREACH A GENERATE (chararray)$17 as origin, (chararray)$18 as
dest, (int)$24 as diversion;
C = FILTER B BY (origin is not null) AND (dest is not null) AND (diversion == 1);
D = GROUP C by (origin,dest);
E = FOREACH D generate group, COUNT(C.diversion);
F = ORDER E BY $1 DESC;
result = limit F 10;
dump result;
```