Big Data Analytics Techniques and Applications Homework II

0556562 陳鴻君

Q1: Compute the average delays and find the maximal delays for each month by using data of all years.

Here I used departure delay as main delay. To compute average delay for each month, I group origin data by month first. Then calculate average values.

Code section:

```
airplane_2008 = LOAD '2008.csv' USING PigStorage(',') AS (
    Year:chararray,
    Month:chararray,
    DayofMonth:chararray,
    DayOfWeek:chararray,
    DepTime:chararray,
    CRSDepTime:chararray,
    ArrTime:chararray,
    CRSArrTime:chararray,
    UniqueCarrier:chararray,
    FlightNum:chararray,
    TailNum:chararray,
    ActualElapsedTime:chararray,
    CRSElapsedTime:chararray,
    AirTime:chararray,
    ArrDelay:int,
    DepDelay:int,
    Origin:chararray,
    Dest:chararray,
    Distance:int,
    TaxiIn:chararray,
    TaxiOut:chararray,
    Cancelled:chararray,
    CancellationCode:chararray,
    Diverted:chararray,
    CarrierDelay:int,
    WeatherDelay:int,
    NASDelay:int,
    SecurityDelay:int,
    LateAircraftDelay:int);


airplane_monthly = GROUP airplane_2008 BY Month;

avg_delay_monthly = FOREACH airplane_monthly GENERATE  AVG(airplane_2008.DepDelay);

DUMP avg_delay_monthly;

MAX_delay_monthly = FOREACH airplane_monthly GENERATE  MAX(airplane_2008.DepDelay);

DUMP MAX_delay_monthly;
```

Result:

Month	Average delay	Maximal delay
1 月	(11.47609595943289)	(1355)
2 月	(13.706226305045202)	(2457)
3 月	(12.49126948010275)	(1521)
4 月	(8.201132754082797)	(2467)
5 月	(7.642741440912969)	(1952)
6 月	(13.609818079614008)	(1710)
7 月	(11.807544712497146)	(1518)
8 月	(9.61475257451315)	(1367)
9 月	(3.961818849518357)	(1552)
1 0 月	(3.803487686795168)	(1369)
1 1 月	(5.420469498039744)	(1286)
1 2 月	(17.30437978049954)	(1597)

Q2: How many plane delays were caused by weather? Please also show the average delays.
I group original data by year. And count how many delays caused by weather.

Code section:

```
airplane_weather = FILTER airplane_2008 BY NOT WeatherDelay == 0;
Wdelay_yearly = GROUP airplane_weather BY Year;

Wdelay_count = FOREACH Wdelay_yearly GENERATE
COUNT(airplane_weather.WeatherDelay );

DUMP Wdelay_count;

Wdelay_avg = FOREACH Wdelay_yearly GENERATE AVG(airplane_weather.WeatherDelay );

DUMP Wdelay_avg;
```

Result:
99985 planes delay were caused by weather.
average weather delay is 46.34412161824274

Q3: Which is the best month of a year to fly with minimum delays?
By the result in Q1. October is the best month of a year to fly with minimum delay.

Q4: List top 5 airports (using IATA airport code) with largest average delay and show which type of delay occurs most for each of the top 5 airport.

Code section:

```
airport = GROUP airplane_2008 BY Origin;

result = FOREACH e {
  g = FILTER airplane_2008 BY NOT CarrierDelay == 0;
  h = FILTER airplane_2008 BY NOT WeatherDelay == 0;
  i = FILTER airplane_2008 BY NOT NASDelay == 0;
  j = FILTER airplane_2008 BY NOT SecurityDelay == 0;
  k = FILTER airplane_2008 BY NOT LateAircraftDelay == 0;

  GENERATE
group,AVG(airplane_2008.DepDelay),COUNT(g),COUNT(h),COUNT(i),COUNT(j),COUNT(k);

}
DUMP result;
```

Result:

UniqueCarrier	Average delay	Carrier delay	Weather delay	NAS delay	Security delay	Late Aircraft delay	Airport name	most type of delay
ACK	29.8544152744630 1	50	15	113	0	69	Nantucket Memorial	NAS delay
PUB	27.0	0	2	2	0	0	Pueblo Memorial	Weather & NAS delay
CEC	24.1862068965517 24	72	10	239	0	164	Jack McNamara	NAS delay
PIR	22.8	1	0	1	0	0	Pierre Regional	Carrier & NAS delay
SPI	22.3145161290322 6	75	4	191	2	105	Capital	NAS delay