Used Dataset

```
Year 2008
```

Solving Techniques

```
Simply using GROUP BY, AVG, MAX, COUNT, ORDER BY, FILTER BY to achieve all the
problem goal.
I also used nested FOREACH to count several filtered value.
```

Pig Latin

```
A = LOAD '2008_hdfs.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:chararray,
    TaxiIn:chararray,
    TaxiOut:chararray,
    Cancelled:chararray,
    CancellationCode:chararray,
    Diverted:chararray,
    CarrierDelay:int,
    WeatherDelay:int,
    NASDelay:int,
    SecurityDelay:int,
    LateAircraftDelay:int);

B = FOREACH A GENERATE Year, Month, ArrDelay, DepDelay, Origin, Dest,
CarrierDelay,WeatherDelay,NASDelay,SecurityDelay,LateAircraftDelay;

C_G = GROUP B BY Month;
-- Q1
ANA = FOREACH C_G GENERATE 'Month:', group, AVG(B.ArrDelay) , MAX(B.ArrDelay);

-- some markdown failure here
ANAO = ORDER ANA BY col2 DESC;
DUMP ANAO;
```

```
-- Q2
WEA = FILTER B BY NOT WeatherDelay == 0;
WEA1 = GROUP WEA BY Year;

ANB = FOREACH WEA1 GENERATE group,'WeatherDelayCount' ,
COUNT(WEA.WeatherDelay),'WeatherDelayAverage', AVG(WEA.WeatherDelay);

-- some markdown failure here
ANBO = ORDER ANB BY col4 DESC;

DUMP ANBO;
-- Q3
LOC = GROUP B BY Dest;
CAR = FOREACH LOC {

cd = FILTER B BY NOT CarrierDelay == 0;
wd = FILTER B BY NOT WeatherDelay == 0;
nd = FILTER B BY NOT NASDelay == 0;
sd = FILTER B BY NOT SecurityDelay == 0;
lad = FILTER B BY NOT LateAircraftDelay == 0;
GENERATE group ,'Delay ' , AVG(B.ArrDelay) ,'CarrierDelay ',
COUNT(cd),'WeatherDelay ' ,COUNT(wd),'NASDelay ',COUNT(nd),'SecurityDelay
',COUNT(sd) , 'LateAircraftDelay ' ,COUNT(lad);

}
-- some markdown failure here
ANCO = ORDER CAR BY col2 DESC;

DUMP ANCO;
```

Q1 Ans

```
December is the most delayed month.

Maximal delays were showed in the end of tuples.
(Month:,12,16.680505081496417,1655)
(Month:,6,13.266756009659792,1707)
(Month:,2,13.077836997760205,2461)
(Month:,3,11.19236458018227,1490)
(Month:,1,10.188855960349496,1525)
(Month:,7,9.975049681276131,1510)
(Month:,8,6.91091468997087,1359)
(Month:,4,6.807297481094145,2453)
(Month:,5,5.978448290248828,1951)
(Month:,11,2.015857969430839,1308)
(Month:,9,0.6977328787273043,1583)
(Month:,10,0.4154954706912698,1392)
```

Q2 Ans

```
99985 planes were influenced by weather delay, the average delayed time is 46.34
mins
(2008,WeatherDelayCount,99985,WeatherDelayAverage,46.34412161824274)
```

Q3 Ans

```
Refer to Q1, minimum delay occured in October.
```

Q4 Ans

```
TOP5 airports and most delays:
1.Marquette County Airport, LateAircraftDelay
2.North Bend Muni, NASDelay
3.Nantucket Memorial, NASDelay
4.Newark Intl, NASDelay
5.Capital, LateAircraftDelay

(MQT,Delay ,30.563365282215123,CarrierDelay ,91,WeatherDelay ,33,NASDelay
,200,SecurityDelay ,0,LateAircraftDelay ,301)
(OTH,Delay ,26.79233870967742,CarrierDelay ,143,WeatherDelay ,2,NASDelay
,171,SecurityDelay ,0,LateAircraftDelay ,116)
(ACK,Delay ,21.515081206496518,CarrierDelay ,85,WeatherDelay ,10,NASDelay
,115,SecurityDelay ,0,LateAircraftDelay ,54)
(EWR,Delay ,20.867524636637718,CarrierDelay ,8110,WeatherDelay ,2275,NASDelay
,38472,SecurityDelay ,140,LateAircraftDelay ,13409)
(SPI,Delay ,20.72508896797153,CarrierDelay ,142,WeatherDelay ,22,NASDelay
,98,SecurityDelay ,1,LateAircraftDelay ,175)
```

Other Information

```
There are several minus values in delay time.
How does that even possible?
```