

**Report title : Apache Pig Weekly Project - Airlines Delay**

**Pig \_ scripts**

**Group name: Ninjaz**

**Q1 : Pig script**

airlines = LOAD '/user/maria\_dev/Airlines/Airlines\_delay\_processed.csv' USING PigStorage(',') AS ( Year:int, Month:int,DepTime:float,CRSDepTime:int,ArrTime:float,

CRSArrTime:int,UniqueCarrier:chararray, FlightNum:int,TailNum:chararray,ActualElapsedTime:float,

CRSElapsedTime:float,AirTime:float,ArrDelay:float,DepDelay:float,Origin:chararray,Dest:chararray,

Distance:int,TaxiIn:int,TaxiOut:int,CarrierDelay:float);

airlines\_info = FOREACH airlines GENERATE FlightNum,ArrDelay,Distance;

distance= FILTER airlines\_info BY Distance > 500;

DUMP distance;

**Q2 : Pig script**

airlines = LOAD '/user/maria\_dev/Airlines/Airlines\_delay\_processed.csv' USING PigStorage(',') AS ( Year:int, Month:int,DepTime:float,CRSDepTime:int,ArrTime:float,

CRSArrTime:int,UniqueCarrier:chararray, FlightNum:int,TailNum:chararray,ActualElapsedTime:float,

CRSElapsedTime:float,AirTime:float,ArrDelay:float,DepDelay:float,Origin:chararray,Dest:chararray,

Distance:int,TaxiIn:int,TaxiOut:int,CarrierDelay:float);

airlines\_info = FOREACH airlines GENERATE UniqueCarrier,CarrierDelay;

carrier\_delay = GROUP airlines BY (UniqueCarrier);

Avg\_arr = FOREACH carrier\_delay GENERATE AVG(airlines.CarrierDelay);

DUMP Avg\_arr;

**Q3: Pig script**

airlines = LOAD '/user/maria\_dev/practice/Airlines.csv' USING PigStorage(',')

AS ( Year:int, Month:int,DepTime:float,CRSDepTime:int,ArrTime:float,

CRSArrTime:int,UniqueCarrier:chararray, FlightNum:int,TailNum:chararray,ActualElapsedTime:float,

CRSElapsedTime:float,AirTime:float,ArrDelay:float,DepDelay:float,Origin:chararray,Dest:chararray,

Distance:int,TaxiIn:int,TaxiOut:int,CarrierDelay:float);

airlines\_info = FOREACH airlines GENERATE FlightNum,ArrTime-DepTime,ActualElapsedTime;

DUMP airlines\_info;

**Q4 : Pig script**

airlines1 = LOAD '/user/maria\_dev/Airlines/Airlines\_delay\_processed.csv' USING PigStorage(',')

AS ( Year:int, Month:int,DepTime:float,CRSDepTime:int,ArrTime:float,

CRSArrTime:int,UniqueCarrier:chararray, FlightNum:int,TailNum:chararray,ActualElapsedTime:float,

CRSElapsedTime:float,AirTime:float,ArrDelay:float,DepDelay:float,Origin:chararray,Dest:chararray,

Distance:int,TaxiIn:int,TaxiOut:int,CarrierDelay:float);

airlines2 = LOAD '/user/maria\_dev/Airlines/Airlines\_delay\_processed.csv' USING PigStorage(',')

AS ( Year:int, Month:int,DepTime:float,CRSDepTime:int,ArrTime:float,

CRSArrTime:int,UniqueCarrier:chararray, FlightNum:int,TailNum:chararray,ActualElapsedTime:float,

CRSElapsedTime:float,AirTime:float,ArrDelay:float,DepDelay:float,Origin:chararray,Dest:chararray,

Distance:int,TaxiIn:int,TaxiOut:int,CarrierDelay:float);

join\_airlines= JOIN airlines1 by (FlightNum,Origin), airlines2 by (FlightNum,Dest);

origin\_filter= FILTER join\_airlines by airlines1::Origin matches 'IND';

dest\_filter= FILTER join\_airlines by airlines2::Dest matches 'BWI';

cogroup\_data = COGROUP origin\_filter by airlines1::Origin, dest\_filter by airlines2::Dest;

DUMP cogroup\_data;

**Q5 : Pig script**

airlines = LOAD '/user/maria\_dev/practice/Airlines.csv' USING PigStorage(',')

AS ( Year:int, Month:int,DepTime:float,CRSDepTime:int,ArrTime:float,

CRSArrTime:int,UniqueCarrier:chararray, FlightNum:int,TailNum:chararray,ActualElapsedTime:float,

CRSElapsedTime:float,AirTime:float,ArrDelay:float,DepDelay:float,Origin:chararray,Dest:chararray,

Distance:int,TaxiIn:int,TaxiOut:int,CarrierDelay:float);

airlines\_info = FOREACH airlines GENERATE FlightNum,Origin,Dest;

origin\_filter = FILTER airlines\_info by Origin matches 'IND';

DUMP origin\_filter;

**Q6 : Pig script**

airlines = LOAD '/user/maria\_dev/Airlines/Airlines\_delay\_processed.csv' USING PigStorage(',') AS ( Year:int, Month:int,DepTime:float,CRSDepTime:int,ArrTime:float,

CRSArrTime:int,UniqueCarrier:chararray, FlightNum:int,TailNum:chararray,ActualElapsedTime:float,

CRSElapsedTime:float,AirTime:float,ArrDelay:float,DepDelay:float,Origin:chararray,Dest:chararray,

Distance:int,TaxiIn:int,TaxiOut:int,CarrierDelay:float);

carrier\_delay = GROUP airlines BY (UniqueCarrier);

max\_delay = FOREACH carrier\_delay GENERATE MAX(airlines.CarrierDelay);

limited\_result = limit max\_delay 1;

DUMP limited\_result;

**Q7 : Pig script**

airlines = LOAD '/user/maria\_dev/Airlines/Airlines\_delay\_processed.csv' USING PigStorage(',') AS ( Year:int, Month:int,DepTime:float,CRSDepTime:int,ArrTime:float,

CRSArrTime:int,UniqueCarrier:chararray, FlightNum:int,TailNum:chararray,ActualElapsedTime:float,

CRSElapsedTime:float,AirTime:float,ArrDelay:float,DepDelay:float,Origin:chararray,Dest:chararray,

Distance:int,TaxiIn:int,TaxiOut:int,CarrierDelay:float);

carrier\_delay = GROUP airlines BY (UniqueCarrier);

min\_delay = FOREACH carrier\_delay GENERATE MIN(airlines.CarrierDelay);

limited\_result = limit min\_delay 1;

DUMP limited\_result;

**Q8: Pig script**

airlines = LOAD '/user/maria\_dev/Airlines/Airlines\_delay\_processed.csv' USING PigStorage(',') AS ( Year:int, Month:int,DepTime:float,CRSDepTime:int,ArrTime:float,

CRSArrTime:int,UniqueCarrier:chararray, FlightNum:int,TailNum:chararray,ActualElapsedTime:float,

CRSElapsedTime:float,AirTime:float,ArrDelay:float,DepDelay:float,Origin:chararray,Dest:chararray,

Distance:int,TaxiIn:int,TaxiOut:int,CarrierDelay:float);

carrier = GROUP airlines BY (UniqueCarrier);

count\_carrier = FOREACH carrier GENERATE FLATTEN(group) as (UniqueCarrier), COUNT(airlines.UniqueCarrier);

DUMP count\_carrier;

**Q9 : Pig script**

airlines = LOAD '/user/maria\_dev/Airlines/Airlines\_delay\_processed.csv' USING PigStorage(',') AS ( Year:int, Month:int,DepTime:float,CRSDepTime:int,ArrTime:float,

CRSArrTime:int,UniqueCarrier:chararray, FlightNum:int,TailNum:chararray,ActualElapsedTime:float,

CRSElapsedTime:float,AirTime:float,ArrDelay:float,DepDelay:float,Origin:chararray,Dest:chararray,

Distance:int,TaxiIn:int,TaxiOut:int,CarrierDelay:float);

airlines\_info = FOREACH airlines Generate FlightNum,ActualElapsedTime;

ActualElapsedTime\_ordered = ORDER airlines\_info BY ActualElapsedTime ASC;

STORE ActualElapsedTime\_ordered INTO 'ActualElapsedTime\_ASC\_ordered';

**Q10 : Pig script**

airlines = LOAD '/user/maria\_dev/Airlines/Airlines\_delay\_processed.csv' USING PigStorage(',') AS

( Year:int, Month:int,DepTime:float,CRSDepTime:int,ArrTime:float,

CRSArrTime:int,UniqueCarrier:chararray, FlightNum:int,TailNum:chararray,ActualElapsedTime:float,

CRSElapsedTime:float,AirTime:float,ArrDelay:float,DepDelay:float,Origin:chararray,Dest:chararray,

Distance:int,TaxiIn:int,TaxiOut:int,CarrierDelay:float);

airlines\_info = FOREACH airlines Generate FlightNum,CRSElapsedTime;

CRSElapsedTime\_ordered = ORDER airlines\_info BY CRSElapsedTime ASC;

STORE CRSElapsedTime\_ordered INTO 'CRSElapsedTime\_ASC\_ordered';