GCP – ENABLE DATAPROC AND STORAGE WITH HIVE

Beeline command shell

1)Load data in the hdfs using wget – o command

2)create hive table

create external table if not exists flight\_data(year int, month int, DayofMonth int, DayOfWeek int, DepTime int, CRSDepTime int, ArrTime int, CRSArrTime int, UniqueCarrier varchar(20), FlightNum varchar(20), TailNum varchar(20), ActualElapsedTime int, CRSElapsedTime int, AirTime int, ArrDelay int, DepDelay int, Origin varchar(20), Dest varchar(20), Distance int, TaxiIn int, TaxiOut int, Cancelled string, CancellationCode varchar(20), Diverted string, CarrierDelay int, WeatherDelay int, NASDelay int, SecurityDelay int, LateAircraftDelay int)

ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n' LOCATION 'gs://final\_project\_bigdata/airlines\_data/'TBLPROPERTIES ("skip.header.line.count"="1","hive.lazysimple.extended\_boolean\_literal"="true");

3) Create another result table with year, carrier,Arrdelay,depdelay , Totaldelay and No delay

create table flight\_table2 as select Year, UniqueCarrier AS Carrier, ArrDelay+DepDelay AS TotalDelay, ArrDelay, DepDelay, IF((ArrDelay<=0) AND (DepDelay<=0),'Y','N') AS NoDelay from flight\_data

4)copy the table data from hdfs to server /tmp in csv format

beeline -u jdbc:hive2://localhost:10000 -n scott -p tiger --outputformat=csv2 -e "SELECT \* FROM flight\_table2" > export.csv

gsutil cp export2.csv gs://final\_project\_bigdata/output/

5)Download in local system from cloud storage

6) Create local repository in Rapid Miner tool

7) create two folders in the repo : data and processes

8) In the repo import flight\_2008.csv data from local storage

9) Add operators – Retrieve, filter ,model and connect node to result .