IBM Data Analysis Using Pig, Hcatalog & Hive

==========

Below is the link to download the INPUT dataset.

https://drive.google.com/file/d/0B_Qjau8wv1KoZUd2NHFjU0xBUVE/view

Input Dataset Fields are mentioned below:

Age,Attrition,BusinessTravel,DailyRate,Department,DistanceFromHome,Education,EducationField,EmployeeCount,EmployeeNumber,

EnvironmentSatisfaction,Gender,HourlyRate,JobInvolvement,JobLevel,JobRole,JobSatisfaction,MaritalStatus,MonthlyIncome,

MonthlyRate,NumCompaniesWorked,Over18,OverTime,PercentSalaryHike,PerformanceRating,RelationshipSatisfaction,StandardHours,

StockOptionLevel,TotalWorkingYears,TrainingTimesLastYear,WorkLifeBalance,YearsAtCompany,YearsInCurrentRole,

YearsSinceLastPromotion,YearsWithCurrManager

We will use below fields for POC:

EmployeeNumber

Department

Overtime

Gender

MonthlyIncome

YearsSinceLastPromotion

[cloudera@quickstart ~]$ hadoop fs -copyFromLocal /home/cloudera/Downloads/ibm.csv /user/cloudera/hadoop/pig/

Data cleansing using Apache Pig

==========

//Running Pig in mapreduce mode

```
[cloudera@quickstart ~]$ pig
```

//Registering piggybank jar

```
grunt> REGISTER '/home/cloudera/Downloads/piggybank-0.17.0.jar';
```

// Loading the input dataset to pig

```
grunt> A = LOAD '/user/cloudera/hadoop/pig/ibm.csv' USING
org.apache.pig.piggybank.storage.CSVExcelStorage(',','YES_MULTILINE','NOCHANG
E','SKIP_INPUT_HEADER');
```

//Selecting fewer columns that we may need for analytics

```
grunt> B = FOREACH A GENERATE (CHARARRAY)$4 as dept, (INT)$9 as empnum,
(CHARARRAY)$11 as gender, (INT)$18 as income, (CHARARRAY)$22 as overtime,
(INT)$33 as lastpromotion;
```

//Removal of null and empty values

```
grunt> C = FILTER B by (dept!= 'NULL')
   AND NOT(dept MATCHES '')
   AND (empnum is not null)
   AND (gender!= 'NULL') AND NOT(gender MATCHES '')
   AND (income IS NOT NULL)
   AND (overtime!='NULL')
   AND NOT(overtime MATCHES '')
   AND(lastpromotion IS NOT NULL);
```

/* This is it from the cleaning part, we need to send cleansed data to Hive for further analysis.

Loading validated data into Hive table using HCatalog

Once the data is validated, we must send it to Hive. For this, we will use HCatalog.

First, you must start the Hive metastore services using below command.*/

```
[cloudera@quickstart ~]$ hive -service metastore

hive> CREATE DATABASE IF NOT EXISTS ibmdatabase;

hive> use ibmdatabase;

hive> CREATE TABLE IF NOT EXISTS ibmanalysis(
dept STRING,
empnum INT,
gender STRING,
income INT,
overtime STRING,
lastpromotion INT)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE;
```

/* Now is the time when we will be loading the data from Pig to Hive using HCatalog. Let's go back to Pig grunt shell and fire below command. The column names must be same for both pig and hive before load operation.*/

```
grunt> STORE C INTO 'ibmdatabase.ibmanalysis' USING
org.apache.hive.hcatalog.pig.HCatStorer();
```

Using HCatStorer() , we are loading the data into Hive at ibmdatabase.ibmanalysis location(where ibmdatabase is the database name and ibmanalysis is the Hive table).

We can go back to Hive and check whether the data got loaded in table or not.

Run the below command.

```
hive> use ibmdatabase;

hive> Select * from ibmanalysis;
```

Problem Statement 1:

Q1) Find out the employee number and dept of employee who does overtime?

```
hive> SELECT empnum, dept from ibmanalysis where overtime = 'Yes';
```

Problem Statement 2:

Q2) Find out last 5 employees based on last promotion received?

```
hive> SELECT empnum,lastpromotion from ibmanalysis order by lastpromotion DESC
limit 5;
```

Problem Statement 3

Q3) Find out the list of employee whose income is more than the average income of all the employee's present in the same department?

```
hive> SELECT a.empnum, a.dept, a.income from ibmanalysis a
INNER join
(SELECT dept, avg(income) as averageincome from ibmanalysis  group by dept) temp
ON (a.dept = temp.dept)
WHERE (a.income >= temp.averageincome);
```
