EMPLOYEE DATA ANALYSIS

Business Objective

Employee retention data analysis assists businesses in improving the quality of new employees, establishing high-performing sales teams, forecasting future staffing needs, and implementing more effective training. Employee data is crucial because it informs your firm about employee viewpoints, which can lead to better decisions, increased employee productivity, and decreased attrition rates.

Why is it a Big Data Engineering problem?

We're creating, building, testing, and maintaining complicated data processing systems for a big data set in this project. As a result, it's a big data engineering problem.

Data

Six csv files comprise the dataset. It contains information about employees of a large firm from the 1980s through the 1990s.

a. Titles (titles.csv)

b. Employees (employees.csv)

c. Salaries (salaries.csv)

d. Departments (departments.csv)

e. Department Managers (dept_manager.csv)

f. Department Employees (dept_emp.csv)

Technology

SQL, Sqoop, Hive, Impala, PySpark (Spark SQL and Spark ML), Jupyter notebook.

Steps followed

a. Prerequisite: upload the data and all other files required to the "https://npbdh.cloudloka.com/ftp/" page.

1. Log in to mysql using : mysql -u anabig11425 -pBigdata123

2. Change the database: mysql> Use anabig11425;

3. Create 'mysql.sql' file. It should contain all the queries needed for creation and loading of data. In sql, run the command: mysql> source cap1.sql

4. Use sqoop to import data from sql to hdfs. Suppose we want to import file to a directory, make sure to remove files with the same name as source from the target directory.

if /anabig11425/hive/warehouse is the target dir and table1 is the source table, run the command:

[anabig11425@ip-10-1-1-204 ~]$ hdfs dfs -rm -r /anabig11425/hive/warehouse/table1

5. for importing, run command:

[anabig11425@ip-10-1-1-204 ~]$ sh cap1g.sh

6. transfer schema from local system to hdfs

hdfs dfs -put /home/anabig114237/table1.avsc /user/anabig114237/new/table1.avsc

7. Run the following command.

hive -f command_impala.sql > output.txt

The Output of the Above Command will be Save in the output.txt

b. Uploaded data to PySpark and carried out data analysis using Spark SQL.

c. ML pipeline is built using Spark ML.

Final Outcome

A pipeline of analysis done using impala and ML pipeline built using Jupyter notebook(random forest and logistic regression models were built).

Challenges faced

a. HUE platform being extremely slow at times.

b. Problems created due to incorrect format of dates.

