Capstone Project - Employee Data Analysis

Objective

The organization wanted to know about their employee retention rate and other useful insights through analysis of their employee database.
And to create an end to end pipeline which includes data modeling, data engineering and data analysis to effientienly answer the business needs of the organization.

Data Description

The dataset comprised of six csv files. It contains information about the employees from 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 Stack Used

- MySQL (to create the database)
- Sqoop (Transfer data from MySQL Server to HDFS/Hive)
- HDFS (to store the data)
- Hive (to create database)
- Impala (to perform the EDA)
- SparkSQL (to perform the EDA)
- SparkML (to create the ML model)

Steps followed

1) Login to mysql
   mysql -u anabig11424 -pBigdata123

2) show databases; #to show the available databases
   use anabig11424; #to use a specific database

3) Create tables for retail data using codes
   a. upload files in the ‘Data’ folder to ftp (https://npbdh.cloudloka.com/ftp).
   b. run the sql commands to create tables and load data into them. The 'mysql.sql' file from the ‘Program Files’ should contain all the queries needed for creation and loading of data.

d. exit from mysql by using ‘quit’ command.

4) create a directory in hdfs and a directory in local to store the dataset as well as tables schema which will be import from mysql by using below commands:
    - creating directory in hdfs. (hdfs dfs -mkdir)
    - creating directory in local (mkdir)

5) Use sqoop to import data from sql to hdfs. Incase we want to import file to a directory, make sure to remove files with the same name as source from the target directory. The 'sqoop.sh' file from the ‘Program Files’ should contain all the queries needed for importing data from sql to hdfs.

6) Transfer schema from local system to hdfs.

7) Uploaded the data to PySpark and carried out data analysis using Spark SQL. (‘spark_ml’ file from the Progam Files)

8) Open the ‘spark_ml.ipynb’ file for running all the Spark and ML commands.

Outputs

Outputs of my analysis are present in the Output folder

Challenges Faced

- Data fromat related challenges.
- Collecting the data and transferring to HDFS.
- Debugging alot of errors.
- Technology related issues.
- Creating the ML pipeline