**CAPSTONE PROJECT DOCUMENTATION**

**BUSINESS OBJECTIVE**

As a process of business objective to this project it is said to remain of the database of employees from that period are six CSV files. In this project, you will design the tables to hold data in the CSVs, import the CSVs into a SQL database, import to HDFS/Hive, and perform analysis using Hive/Impala/Spark/SparkML using the data and create pipelines.

**Technology Stack**

As a part of technology stack you are required to work on below technology Stack.

- MySQL (to create database)

- Linux Commands

- 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 perform model building

**DATA USED**

-Departments

- dept\_emp

-dept\_manager

-employees

-salaries

-titles

**DATA DESCRIPTION**

1. Titles (titles.csv):

title\_id – Unique id of type of employee (designation id) – Character – Not Null

Title – Designation – Character – Not Null

1. Employees (employees.csv):

emp\_no – Employee Id – Integer – Not Null

emp\_titles\_id – designation id – Not Null

birth\_date – Date of Birth – Date Time – Not Null

first\_name – First Name – Character – Not Null

last\_name – Last Name – Character – Not Null

sex – Gender – Character – Not Null

hire\_date – Employee Hire date –Date Time -Not Null

no\_of\_projects – Number of projects worked on – Integer – Not Null Last\_performance\_rating – Last year performance rating – Character – Not Null

left – Employee left the organization – Boolean – Not Null

Last\_date - Last date of employment (Exit Date) – Date Time

1. Salaries (salaries.csv):

emp\_no – Employee id – Integer – Not Null

Salary – Employee’s Salary – Integer – Not Null

1. Departments (departments.csv)

dept\_no - Unique id for each department – character – Not Null

dept\_name – Department Name – Character – Not Null

1. Department Managers (dept\_manager.csv)

dept\_no - Unique id for each department – character – Not Null

emp\_no – Employee number (head of the department ) – Integer – Not Null

1. Department Employees (dept\_emp.csv)

emp\_no – Employee id – Integer – Not Null

dept\_no - Unique id for each department – character – Not Null

**PROJECT OBJECTIVE:**

As part of this project, you are required to work on

1. Create data model as per your understanding from the data (you are required include tables names, relation between tables, column names, data types, primary & foreign keys etc.) Tip: You can create ER diagram either in EXCEL or using the link https://www.quickdatabasediagrams.com/ (Preferably in this app)

2. Create database & tables in MySQL server as per the above ER Diagram

3. Create Sqoop job to transfer the data from MySQL to HDFS (Data required to store in Parque/Avro/Json format)

4. Create database in Hive as per the above ER Diagram and load the data into Hive tables

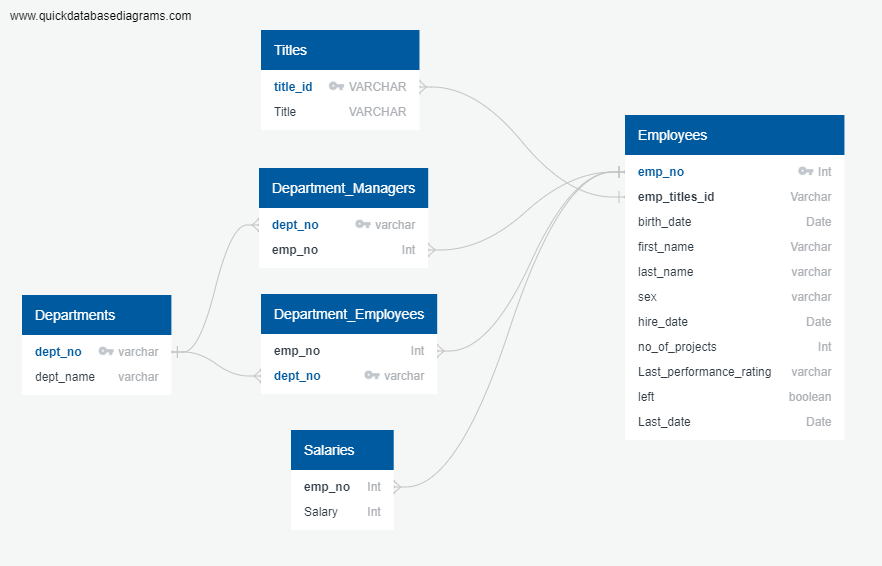
5. Work on Exploratory data analysis as per the analysis requirement using Impala and Spark SQL.

6. Build ML Model as per the requirement.

7. Create entire data pipeline and ML pipe line

8. Upload the entire project repository including source code to Github (you are required to create github account if you don’t have it

**E R DIAGRAM**

****

**ARCHITECTURE OF PIPELINE**

1. **DATABASE CREATION:**

a. Connecting to Linux terminal to access MySQL server for database creation.

b. In the database create 6 table corresponding to the files, with proper schema.

c. Upload the data into the HDFS FTP

d. Load the data into the table using appropriate delimiter.

1. **DATA AND SCHEMA TRANSFER TO HDFS AND HIVE**

a. Select a compressed file format (AVRO) for the data to be transferred.

b. Use Sqoop command to transfer all table data to a specified location into HDFS directory.

c. In Linux shell, transfer the .avsc schema file to another specified HDFS directory.

1. **HIVE DATABASE AND TABLE CREATION**

a. Create a database and external table for the table data imported.

b. Load the data into external table using appropriate path and SerDes.

c. Create views from the tables for specified purposes.

1. **HIVE AND SPARKSQL EXPLORATORY DATA ANALYSIS**

a. Use a database.

b. Do EDA on the business problem asked.

c. Open Jupyter notebook and create Spark Session instance.

d. Use SparkSQL to perform EDA of the same business problems.

e. Perform proper visualisation.

1. **SPARKML MODEL BUILDING**

a. Import all the required libraries in Jupyter Labs

b. Data preparation to clean out duplicate and null values.

c. Proper formatting of table fields.

d. Create categorical and continous features.

e. Do Encoding on the categorical data and transform the data.

f. Slit the dataset into train and test set.

g. Perform Logistic Regression and Random Forest Classifier.

h. Compare the result.

**CHALLENGES FACED DURING THE PROJECT EXECUTION:**

1. Creation of employee table with the specified data types was a hurdle
2. Data transformation in spark ML
3. ML Pipeline creation