**Data Engineering Capstone Project-1**

Introduction:

Your first major task is to work on data engineering project for one of the big corporation’s employee’s data from the 1980s and 1995s. All the database of employees from that period are provided six CSV files. In this project, you will design data model with all the tables to hold data, import the CSVs into a SQL database, transfer SQL database to HDFS/Hive, and perform analysis using Hive/Impala/Spark/SparkML using the data and create data and ML pipelines.

Project Description:

In this project, it is required to create end to end data pipeline and analyzing the data.

Technology Stack:

you are required to work on below technology Stack.

- MySQL (to create database)

- Linux Commands

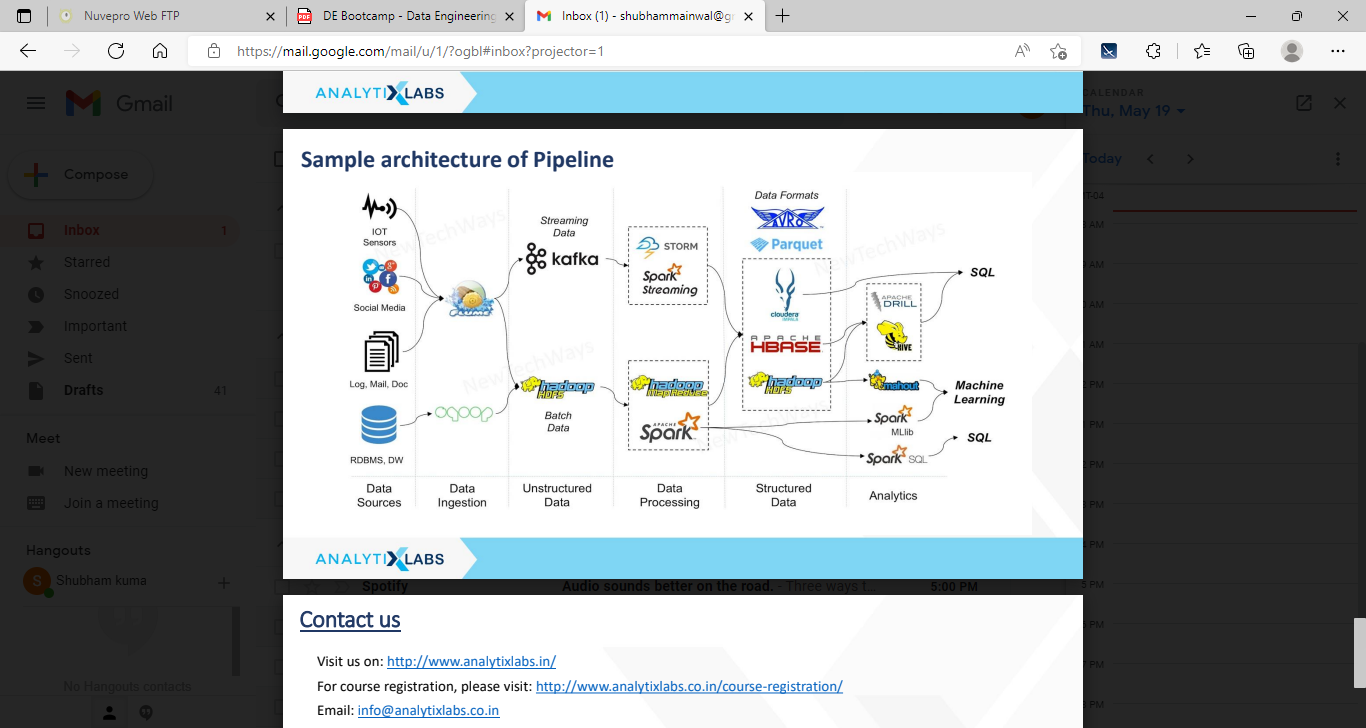
- Sqoop (Transfer data from MySQL Server to HDFS/Hive)

- HDFS (to store the data)

- Hive (to create database)

- SparkSQL (to perform the EDA)

- SparkML (to perform model building)



Data Description

Please find the details of all the tables

**a. Titles (titles.csv):**

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

title – Designation – Character – Not Null

**b. Employees (employees.csv):**

emp\_no – Employee Id – Integer – Not Null - PK

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

**c. Salaries (salaries.csv):**

emp\_no – Employee id – Integer – Not Null -PK

Salary – Employee’s Salary – Integer – Not Null

**d. Departments (departments.csv)**

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

dept\_name – Department Name – Character – Not Null

**e. Department Managers (dept\_manager.csv)**

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

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

**f. Department Employees (dept\_emp.csv)**

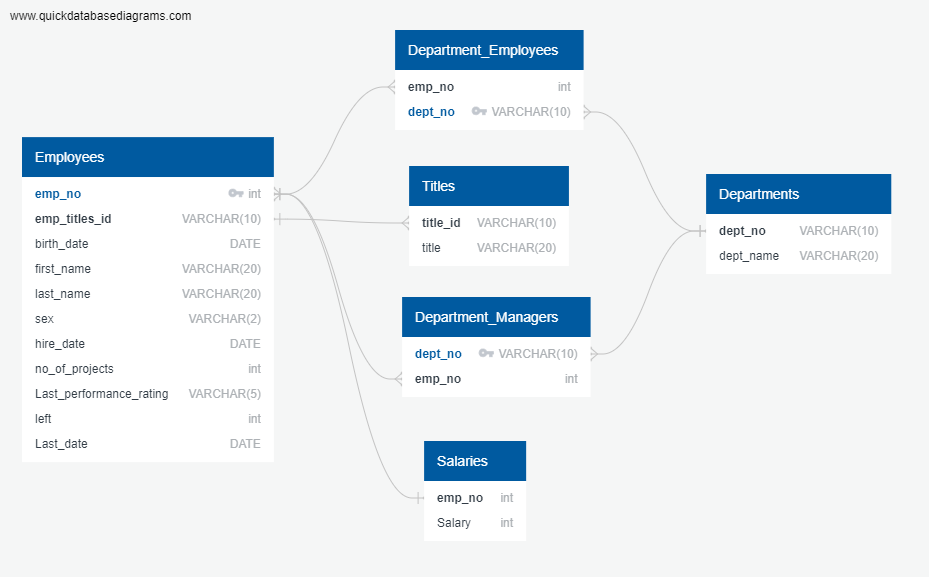
emp\_no – Employee id – Integer – Not Null

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

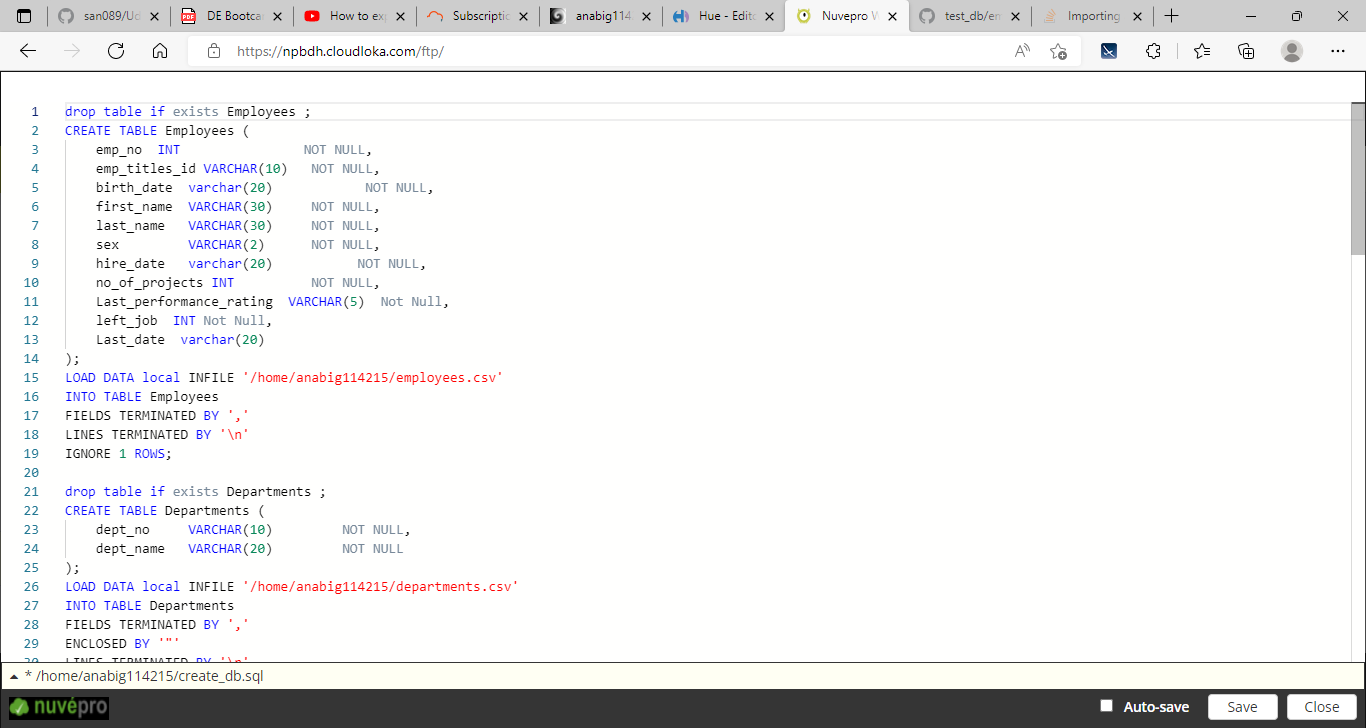
Project Objective

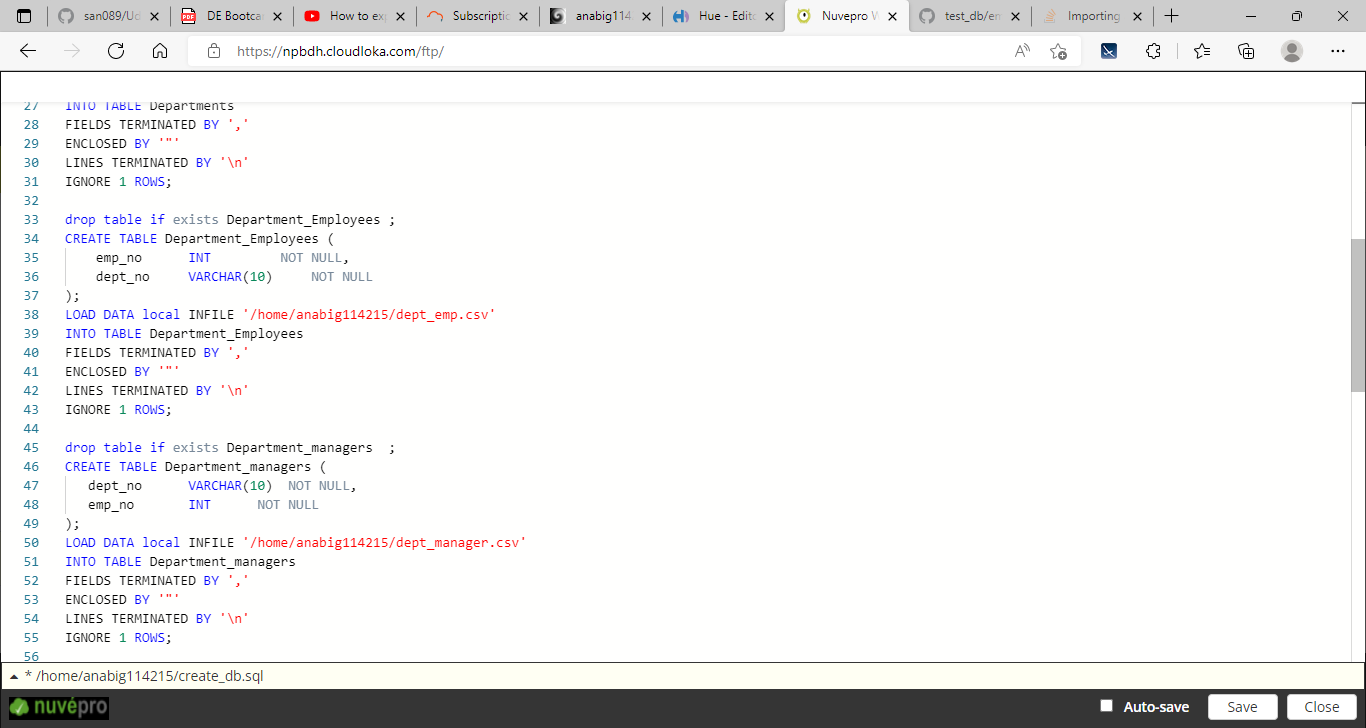
As part of this project, you are required to work on

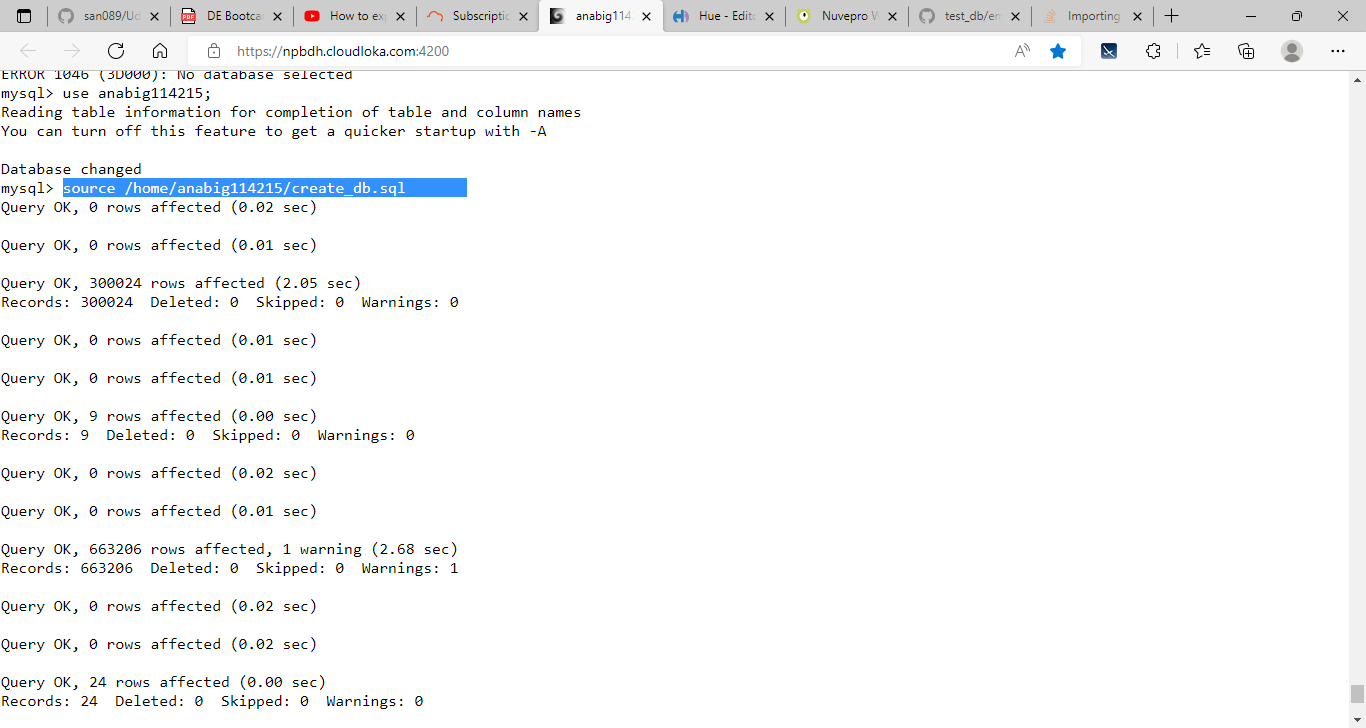
1. Entity Relation Diagram: Create data model as per your understanding from the data (you are required include tables names, relation between tables, column names, data types, primary & foreign keys etc.



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





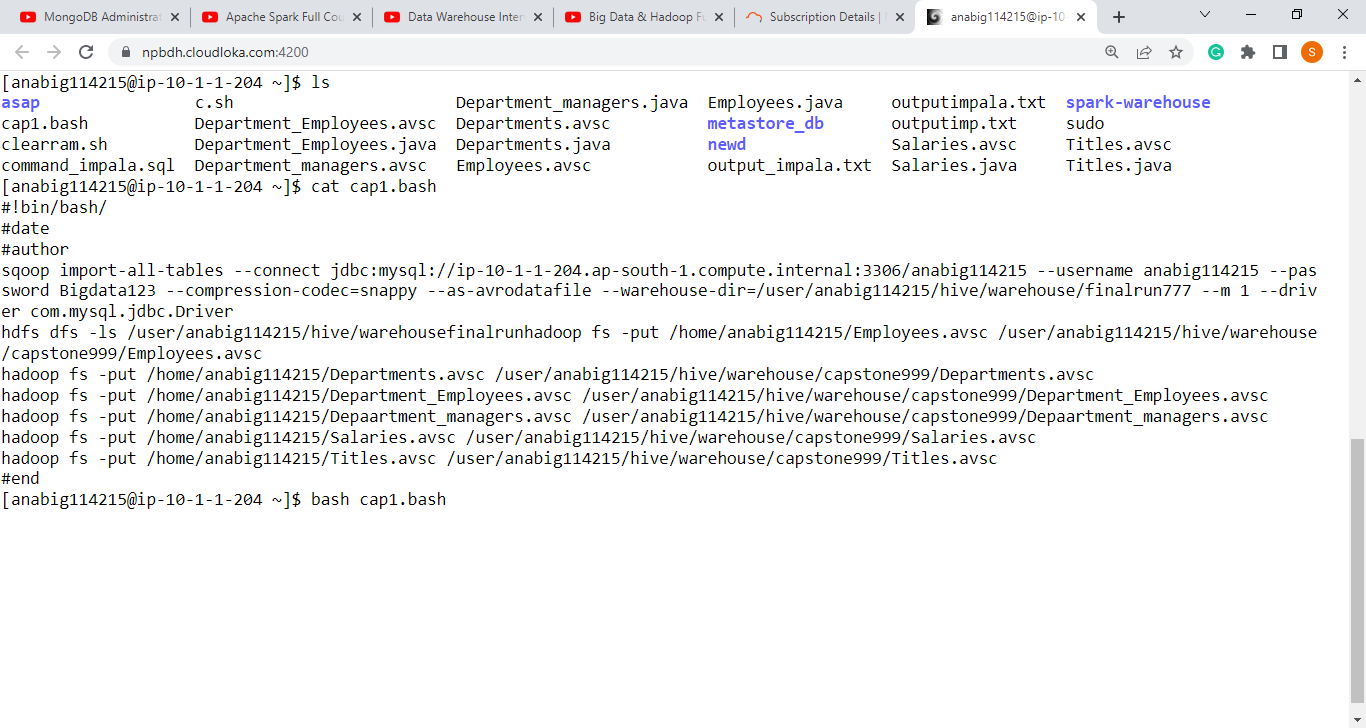




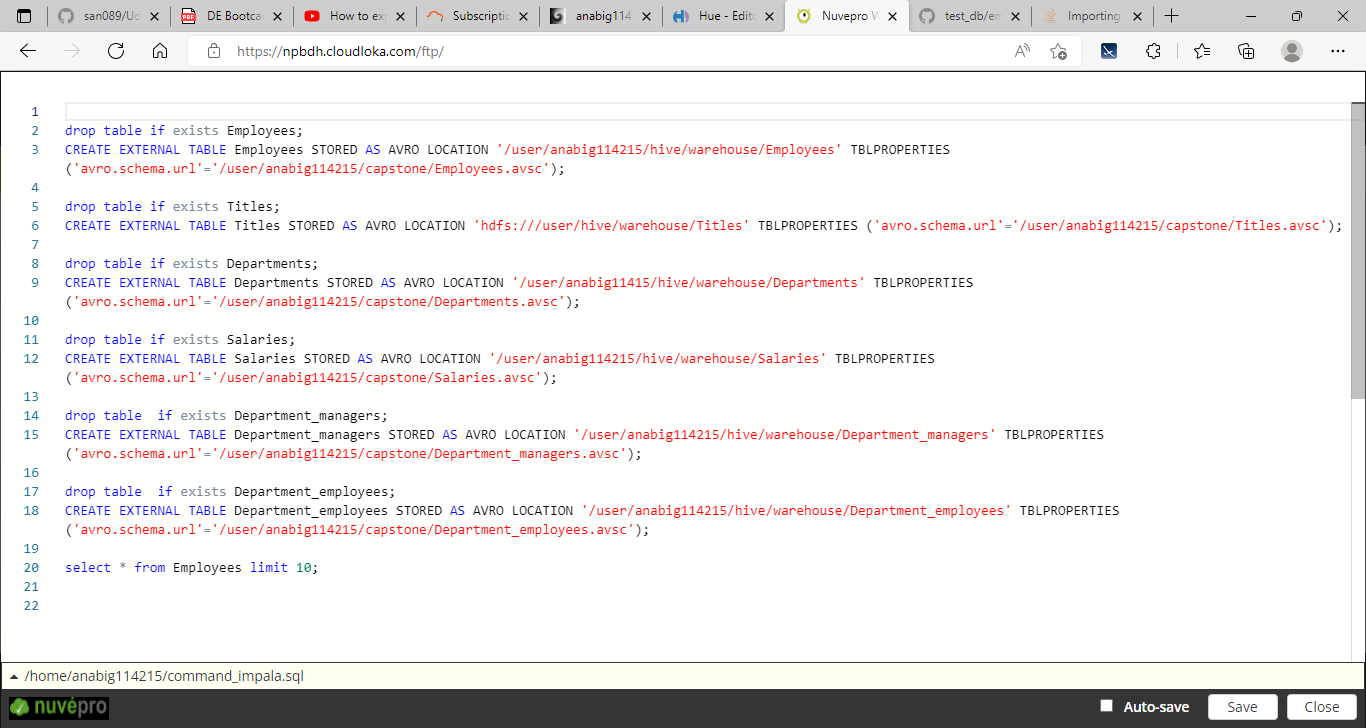
1. Create Sqoop job to transfer the data from MySQL to HDFS (Data required to store in Parquet/Avro/Json format

for this you can use shell file to automate things

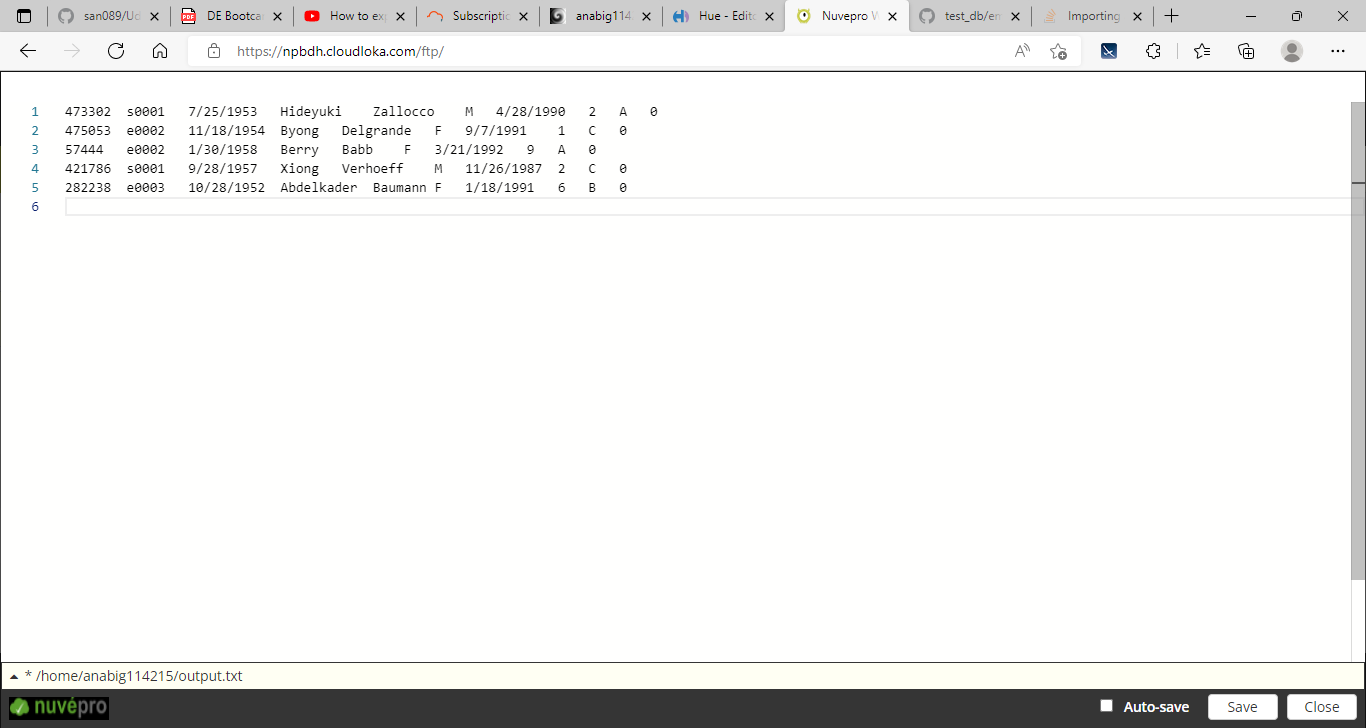
bash cap1.bash



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



Output file

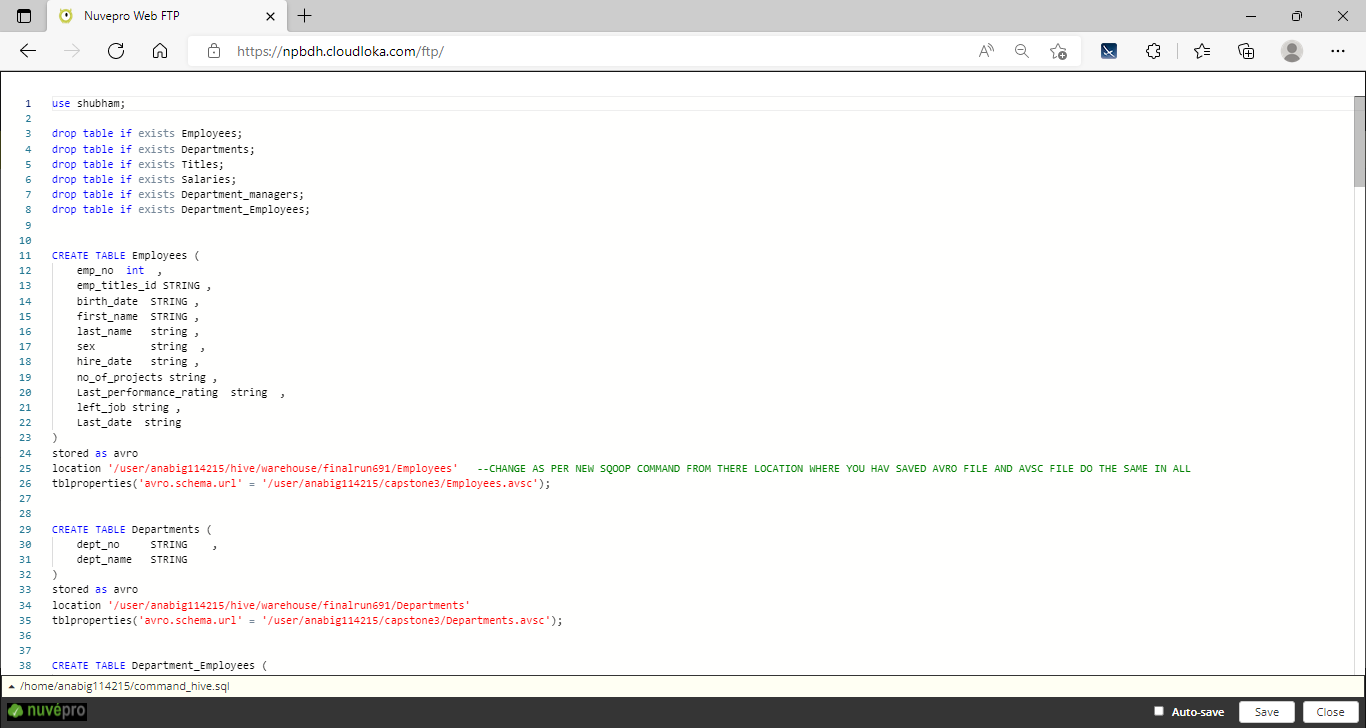


As it’s avro file

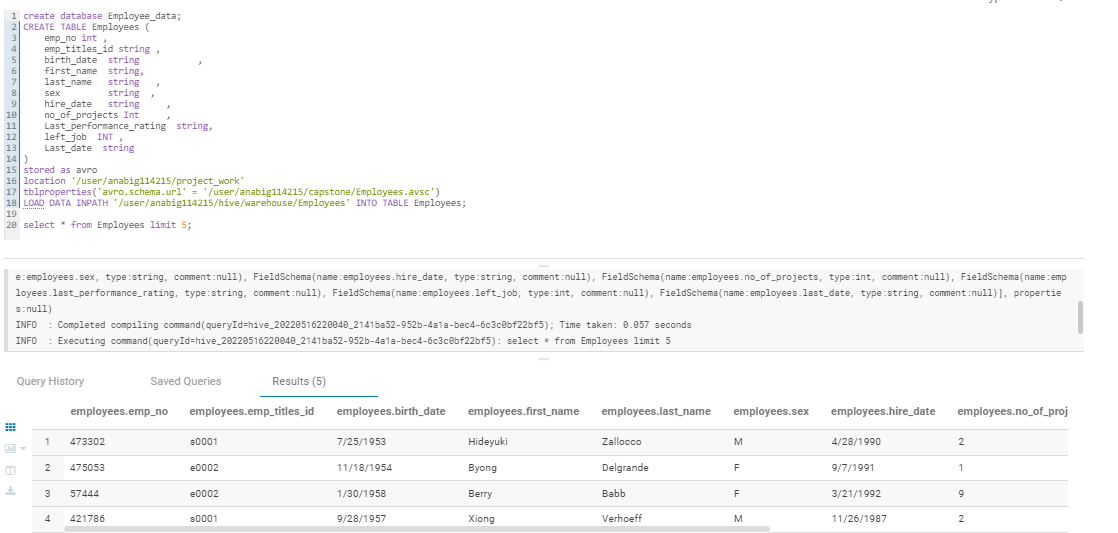
And its schema and java file which tells you to how sqoop imported the data

Then you need to transfer these only. avsc files to hdfs and through scoop you have already sent the .avro data right

Then go into the shell then command\_hive.sql file and for sample the code is like this



Sample output:

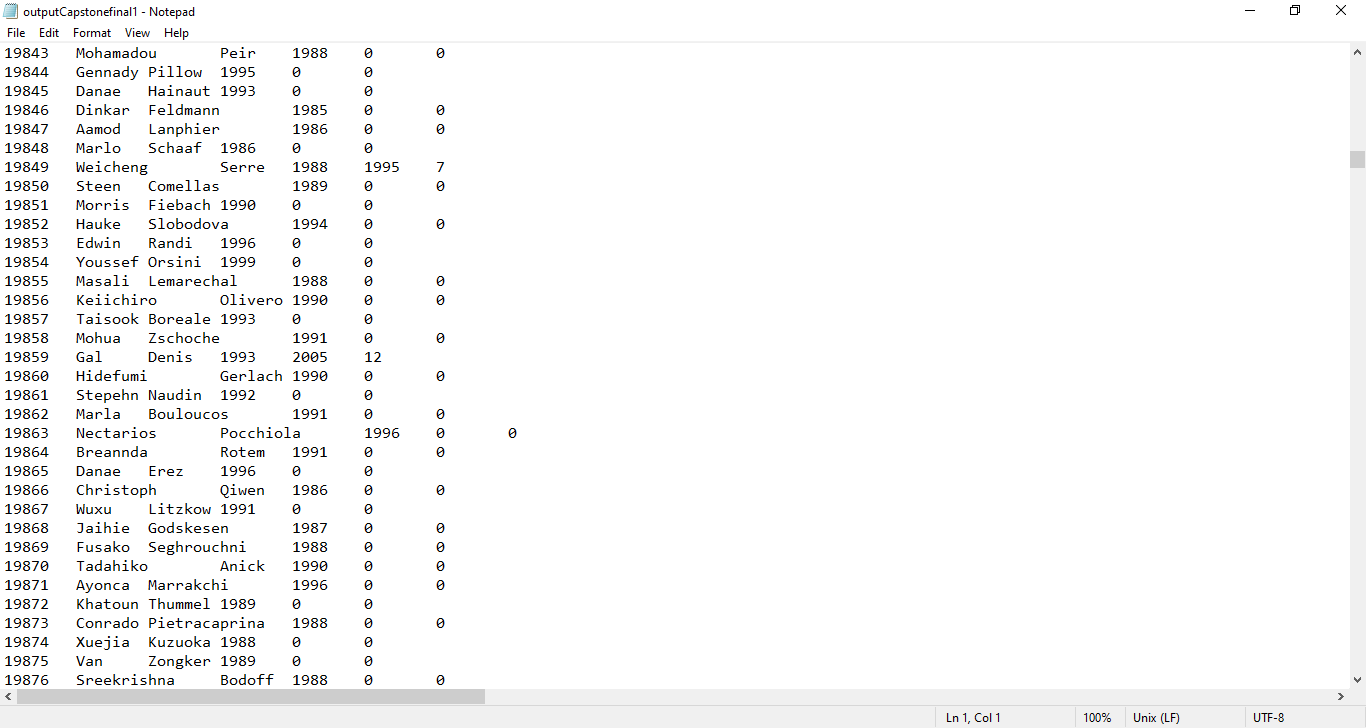


1. Analysis is done with command\_hive.sql file as we are making it as pipeline so analysis is done automatically. Using hive –f command\_hive.sql > output.txt you will get the ouput data and analysis stored in the file and which is in the local system.

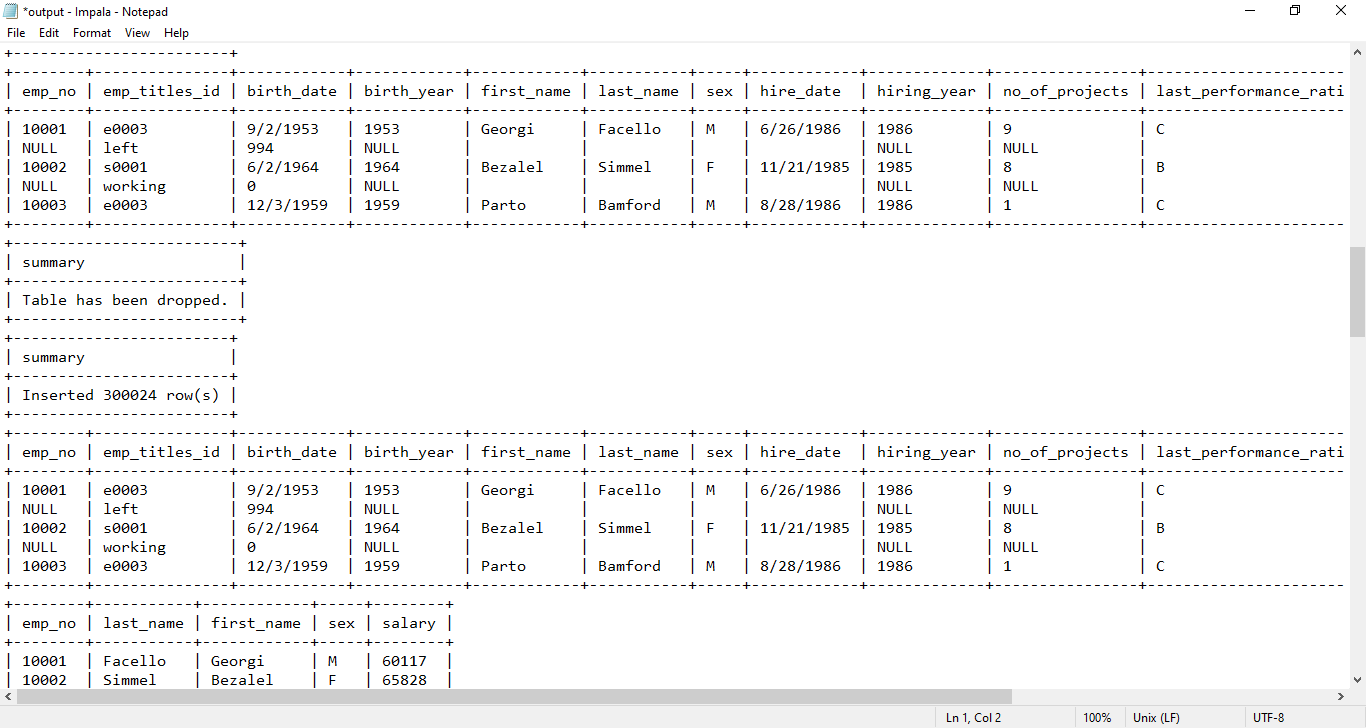
For impala use

impala-shell -i ip-10-1-2-103.ap-south-1.compute.internal -d shubham -f command\_impala.sql > outputimp.txt

hive ouput:

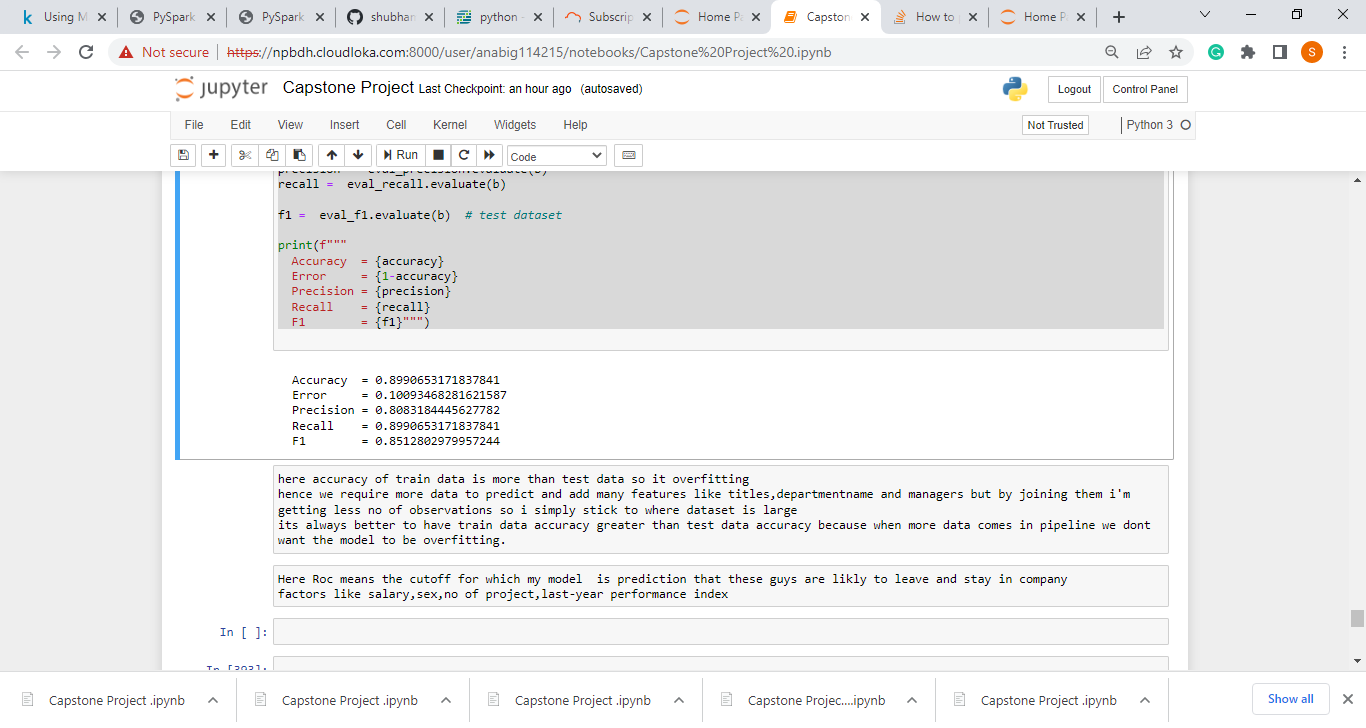


Impala output file

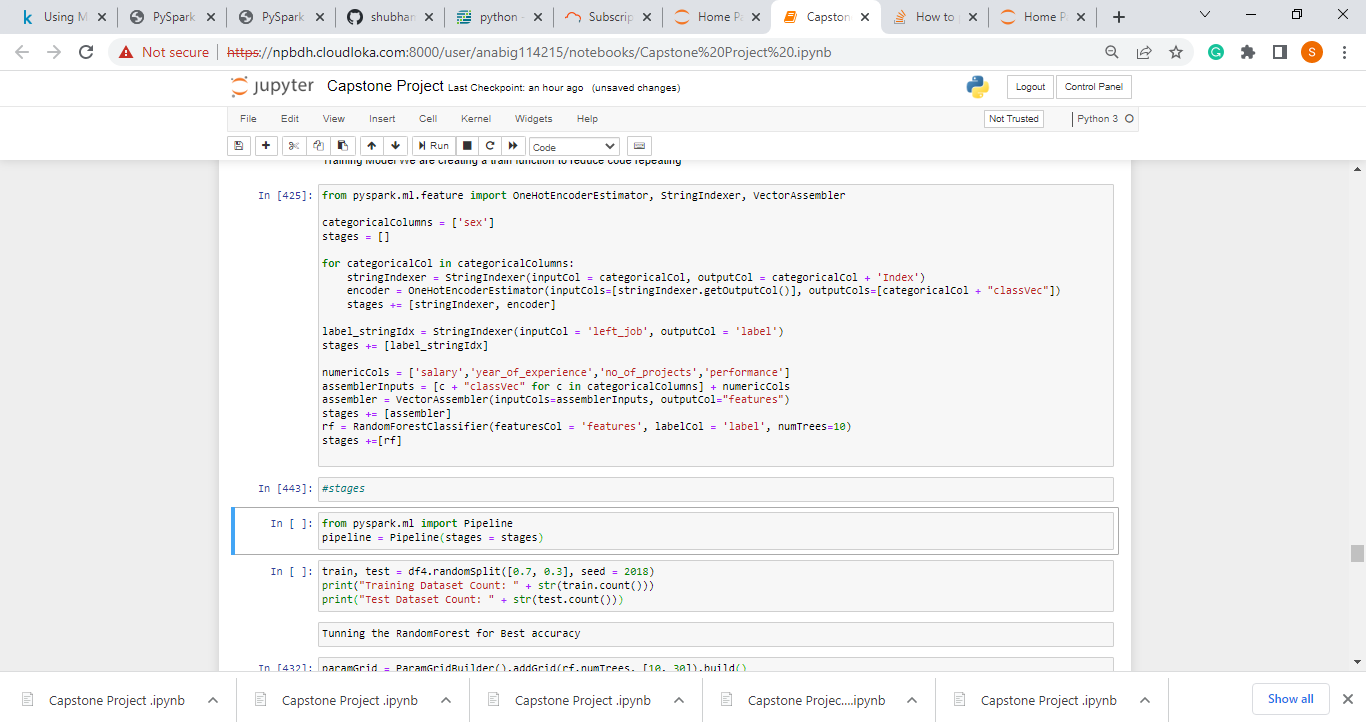


1. Build ML Model as per the requirement.

Problem is the company wants to the Attrition rate of employees with 89% accuracy my will ml will predict the values based on variables.



1. Create entire data pipeline and ML pipe line



**Challenges:**

1. Impala and hive mix usage because of the common syntax.
2. While creating pipeline in impala facing issues like memory limit so need to reduce the query concurrency.
3. Thinking how to reduce more as pipeline is as better with minor manual things
4. Data Collection from data base for picking up right variables as Schema should be defined properly because we are working with avro file.
5. Automation of all the things in sql then in hive just little mistakes can take lot of time.
6. At running the code for the output it is very challenging as pipeline takes a lot of time to give you fruit full results
7. Data has been provided in encoded format, converting them into more understandable format like year of experience, hiring year and last year rating which is ordinal variable.
8. Selecting important features from 24 columns out of which most the data has no correlation with (left\_job) as there can be only few possibilities that it depends on as per dataset
9. Identify the Positive Drivers to understand driving factors of the case
10. Finding right tuning parameters for accuracy