**Data Engineering – Final Assessment**

**Business Scenario:**

Finance is a field that is concerned with the allocation (investment) of assets and liabilities over space and time, often under conditions of risk or uncertainty. Finance can also be defined as the art of money management. Participants in the market aim to price assets based on their risk level, fundamental value, and their expected rate of return.

The dataset consists of customer information of L&T financial services. It is a finance dataset, which consists of customers demographics, loan disbursed, asset cost being purchased and the customers previous account and loan history. The dataset also consists of the state and branch id of L&T from where the loan was disbursed, the customer’s account history. It also contains the CNS score and score description provided by the Credit Bureaus of India.

**Challenges:**

It is a challenge for any financial services to target the right people for disbursing the loan. The credit team must analyse various details like CIBIL score, payment history (if available), credit history, geographical location, profession, income, age, education etc. of the customers. This will help in understanding whether the person is capable of paying back the loan amount. Which in turn reduces its NPAs and increases the profitability.

You need to assess what data is available and perform some exploratory and descriptive analytics to identify interesting and useful patterns, trends, and insights.

**What is Expected?**

After completing this final project, you will apply various data engineering skills and techniques. You've recently joined the organization, facing a business challenge that requires data engineering skiils on real data sets. Based on your findings, the next level of analysis will be charted out.

Here are some indicative types of data engineering techniques you can perform. Please note that this is not an exhaustive list, you may add more

You will perform the various tasks that professional data analysts do as part of their jobs, including:

* Data collection from multiple sources
* Data wrangling and data preparation
* Exploratory data analysis
* Data Storage and retrieval

**Data Discovery ,Data Preparation, Data Manipulation, Data Storage**

* Identify data source types
* Identify duplicate values in the dataset
* Remove duplicate values from the dataset.
* Identify missing values in the dataset.
* Determine the missing values in the dataset.
* Normalize data in the dataset.
* Identify silver , bronze , gold layer attributes

The project will come to a close with a presentation of your data engineering report to relevant organisation stakeholders. An executive summary, your analysis, and a conclusion will all be included in the report. Both the final output and your work for the various Data engineering process phases will be taken into consideration as part of evaluation.

Your ability to use Jupyter Notebooks, SQL, Relational Database Management Systems (RDBMS), Big Data Tools , Kafka , Airflow , NoSQL ,cloud computing environments like Azure and GCP, and Python libraries like Pandas, Numpy, and others will be put to the test as part of this project.

**NOTE:** Results must be backed with appropriate inferences and insights.

**Data Dictionary:**

|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| disbursed\_amount | Amount of Loan disbursed |
| asset\_cost | Cost of the Asset |
| ltv | Loan to Value of the asset |
| branch\_id | Branch where the loan was disbursed |
| Date.of.Birth | Date of birth of the customer |
| Employment.Type | Employment Type of the customer (Salaried/Self Employed) |
| DisbursalDate | Date of disbursement |
| State\_ID | State of disbursement |
| Employee\_code\_ID | Employee of the organization who logged the disbursement |
| MobileNo\_Avl\_Flag | if Mobile no. was shared by the customer, then flagged as 1 |
| Aadhar\_flag | if aadhar was shared by the customer then flagged as 1 |
| PAN\_flag | if pan was shared by the customer then flagged as 1 |
| VoterID\_flag | if voter was shared by the customer then flagged as 1 |
| Driving\_flag | if DL was shared by the customer then flagged as 1 |
| Passport\_flag | if passport was shared by the customer then flagged as 1 |
| PERFORM\_CNS.SCORE | Bureau Score |
| DELINQUENT.ACCTS.IN.LAST.SIX.MONTHS | Number of delinquent accounts in the last six months |
| CREDIT.HISTORY.LENGTH | Credit history in terms of years |
| NO.OF\_INQUIRIES | Number of inquiries made by the customer |
| loan\_default | Payment default in the first EMI on due date |

**Tasks :**

**Data Discovery , Processing , Loading Task :**

**Task 1**

* 1. Identify mongo DB data model from csv file in terms of database, collections, documents

**database(investmentDB)** --databases hold one or more collections of documents. To select a database to use

## Collections(LoadData)-

MongoDB stores documents in collections. Collections are analogous to tables in relational databases.

**Documents** ( specifically BSON documents) which are gathered together in collections. A database stores one or more collections of documents.

* 1. Populate the above data to mongo DB datastore

--connect mongodb Atlas--

mongosh "mongodb+srv://cluster0.qwmnbnd.mongodb.net/" --apiVersion 1 --username kripal

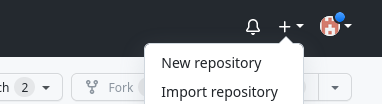
**Mongoimport -d database2 -c LoanData --type csv --file LoanData.csv --headerline**

* 1. Create five reports . Example : report about percentage of default etc

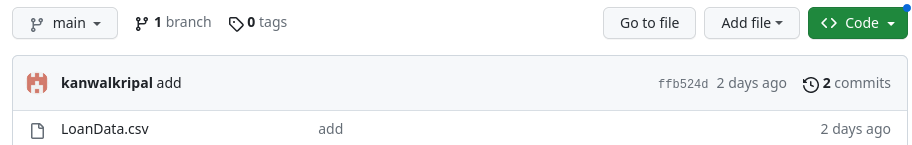
**Data Versioning Task :**

**Task 1 Create a repository in github and upload the dataset**

* 1. Create a repository in github.com

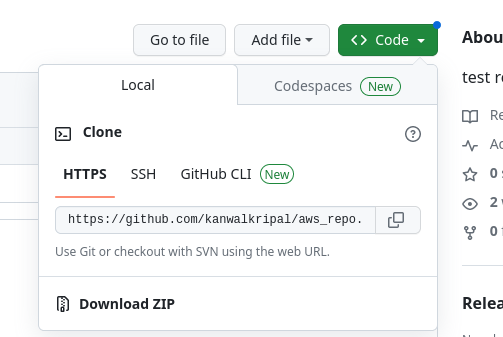


* 1. Upload the dataset to github.com through browser



**Task 2 Generate the Reports and upload them to github**

1. Clone the repository that you created in task 1



Open Terminal.

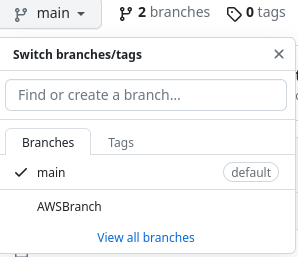
Change the current working directory to the location where you want the cloned directory.

Type git clone, and then paste the URL you copied earlier.

$ git clone https://github.com/kanwalkripal/aws\_repo.git

Press Enter to create your local clone.

1. Create a branch for this activity ,Create a file with all the queries from the above tasks.



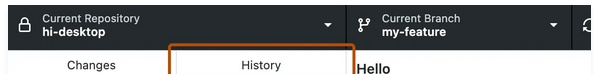
$ git branch -a

git add LoanData.csv

git stash

1. List the version history ( commit history) , status of the local repository

On the History tab, click the commit you'd like to review.



1. Commit the file to the branch

git commit -m "added something new"

1. Push the changes to the github

Git Push

**Cloud Data Engineering Tasks :**

1. Upload the dataset to aws s3
2. Create aws managed mongo db instance and create five reports
3. Create aws code commit repos and perform the same steps as mentioned in Data Versioning Task
4. Create aws emr instance and perform three data engineering tasks such cleaning etc . Example : Removing column values with Unknown , Imputing the column values