**Data Engineering – Final Assessment**

**Business Scenario:**

There is an automobile company XYZ from India which aspires to enter the Indian used car market by setting up their company locally to give competition to their counterparts.

**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:**

|  |  |
| --- | --- |
| **Column Name** | **Description** |
| Sales\_ID | Sales ID |
| name | Name of the used car |
| year | Year of the car purchase |
| selling\_price | Current selling price for used car |
| km\_driven | Total km driven |
| Region | Region where it is used |
| State or Province | State or Province where it is used |
| City | City where it is used |
| fuel | Fuel type |
| seller\_type | Who is selling the car |
| transmission | Transmission type of the car |
| owner | Owner type |
| mileage | Mileage of the car |
| engine | engine power |
| max\_power | max power |
| seats | Number of seats |
| sold | used car sold or not |

**Tasks :**

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

**Task 1**

* + 1. Move the dataset onto HDFS.
    2. Create a hive table for the dataset and load the data into the table and create three reports. Example report about most profitable regions based on the number of cars sold etc

**Task 2**

1. Create a data frame for the dataset using PySpark SQL
2. Perform data engineering operations like remove missing values and save the result to hdfs
3. Create five reports . Example report about most profitable regions based on the number of cars sold etc
4. Save reports to HDFS in parquet format

**Task 3**

* 1. Identify mongo DB data model from csv file in terms of database, collections, documents
  2. Populate the above data to mongo DB datastore
  3. Create five reports . Example : report about most profitable regions based on the number of cars sold etc

**Data Versioning Task :**

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

* 1. Create a repository in github.com
  2. 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
2. Create a branch for this activity ,Create a file with all the queries from the above tasks.
3. List the version history ( commit history) , status of the local repository
4. Commit the file to the branch
5. Push the changes to the github

**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