Case study

ETL pipeline stages

Car data is sourced from different data providers.
Frequency – Monthly. We get full data including historical data every month.
Data quality – Dirty with many null values.
Data types of same data element is different between different data soruces (Ex. ID is char in one of the sources and int in another source).
We need to design an ETL pipeline that reads the data from a server, integrate data from different sources, cleanse, process and load the data in the target database.

1. **Please draw the ETL pipeline stages from extraction and until loading. (No coding required to be done).**
2. **Mention how you would handle the data quality issues.**

Application

Consider the following data model (simplified for the purpose of case study)

Car attirbutes

- Car id
- Brand – BMW, Audi, etc.,
- Model – C-Class, A3, etc,
- Model year – 2019, 2020
- Fuel type – Petrol, Diesel, Electric
- Transmission type – Manual, Automatic
- Engine HP – 140, 150, 160, etc.,
- Showroom price

Agreement attributes

- Agreement id
- Car id
- Residual value at the end of lease period
- Lease duration in months
- Lease rental/month
- Lease start date
- Lease end date
- Car resale price

In this Problem, you have to write an application in Python to answer the following business queries.

1. Which is the most popular car brand leased in a particular year
2. Average lease duration for each car brand
3. Which brand car lease agreements are generally terminated early (hint – Lease end date < (Lease start date + Lease duration in months)
4. Keeping aside other factors, how good are the residual value predictions (hint – Compare residual value with the car resale price)
5. Calculate the proft/loss of an agreement