**Sakina-GDE\_Homework Report: PySpark ETL Practice**

[**https://colab.research.google.com/drive/1KxSBhvrbYg4EhC8hGFnYjyGjYCrmuNM9?usp=sharing**](https://colab.research.google.com/drive/1KxSBhvrbYg4EhC8hGFnYjyGjYCrmuNM9?usp=sharing)

**Objective:**  
To perform a basic ETL (Extract, Transform, Load) process using PySpark on a sample dataset.

**1. Setup**

* Installed **PySpark** to enable Spark functionalities
* Checked the current working directory
* Listed the files and folders to verify setup

A screenshot of a computer

AI-generated content may be incorrect.

**2. Ingestion (Data Download)**

* Downloaded a sample dataset cars.csv using the wget command:

A close-up of a computer code

AI-generated content may be incorrect.

**3. Extraction (ETL Step 1)**

* Created a **SparkSession** to start working with PySpark:
* Loaded the cars.csv file into a DataFrame:
* Displayed the first 11 rows to verify data ingestion:

A screenshot of a computer

AI-generated content may be incorrect.

**4. Transformation (ETL Step 2)**

* Added a new column **"Warranty"** based on the value in the **"Origin"** column:
  + If Origin = "US" → Warranty = "10 Years"
  + Else → Warranty = "5 Years"

A screenshot of a computer

AI-generated content may be incorrect.

**5. Load (ETL Step 3)**

* Saved the transformed DataFrame into a folder named cars\_output in CSV format with headers:

A white rectangular object with black text

AI-generated content may be incorrect.

**Conclusion**

Successfully completed a full **ETL cycle** using PySpark:

* **Extracted** data from a CSV file,
* **Transformed** data by adding a calculated column,
* **Loaded** the result back as a new CSV output.