**Project Report: SQL ETL Pipeline Simulation**

**1. Introduction**

In modern data-driven environments, clean and structured data is essential for generating insights. This project simulates a real-world **ETL (Extract, Transform, Load)** pipeline using SQL to clean and manage an e-learning dataset. It demonstrates essential data engineering and analytics skills using MySQL Workbench, preparing data from raw CSV format through multiple stages into a production-ready format for analysis.

**2. Abstract**

This simulation project involves importing an unrefined e-learning courses dataset into a **staging table**, followed by a series of **data cleaning, transformation**, and **auditing** operations. Cleaned data was then inserted into a structured **production table**, with triggers logging inserts and updates into an **audit log**. The final dataset was analyzed to extract actionable insights such as course trends, instructor ratings, and platform effectiveness.

**3. Tools Used**

* **MySQL Workbench** - for SQL scripting, ETL simulation, data analysis and insight generation.
* **CSV Files (online\_courses.csv)** - raw data source. (generated)
* **Excel** - for preliminary review and export handling.
* **Languages Used -**  Standard SQL (DDL, DML, Aggregates, Triggers and Stored Logic)

**4. Steps Involved in Building the Project**

1. **Dataset Understanding & Cleaning**

* Analyzed the raw CSV containing 1000 rows of online course data.
* Identified fields with inconsistent formatting (e.g., ratings like "3.5 stars", prices as "$29.99", durations as "10 hours").
* Cleaned missing/null records and removed duplicate course entries.

1. **ETL Pipeline Construction**

* **Extract:** Imported raw data into staging\_courses.
* **Transform:** Used SQL functions to trim text, parse ratings, extract numeric values, and normalize boolean fields (certifications).
* **Load:** Inserted validated data into courses table (production environment).

1. **Trigger-Based Audit Logging**

* Created two triggers: AFTER INSERT to log every new course into audit\_log.

AFTER UPDATE to log modifications like rating updates.

* Ensures traceability and data integrity.

1. **Insightful SQL Analysis**

* Top 5 most enrolled courses.
* Average rating per category.
* Paid vs. free course distributions.
* Top instructors by average rating.
* Platforms offering highly rated and popular courses.

1. **Documentation & Export**

* Final datasets (cleaned\_courses, audit\_log) exported as CSVs.
* Analytical outputs documented as screenshots and result sets.

1. **Conclusion**

This project successfully simulates a **professional-grade SQL ETL pipeline**, emphasizing **data cleaning**, **production-grade table design**, **automated auditing**, and **business-oriented analytics**. It demonstrates strong command of SQL for both data engineering and insight extraction. The structure, automation, and clarity of this project make it ideal for real-world applications and portfolio showcasing.