# Debre Berhan University Vacancy 2024

# BIG DATA ANALYTICS ASSIGNMENT

# GROUP MEMBERS ID NO\_ 1. DEMIREWU MANIDEFRO ………………………………..DBU1501123 2. LAMROT SOLOMON ……………………………………….DBU1501315 3. MIKIYAS TOLKO …………………………………………..…DBU1501375 4. NAHOM TESHOME ………………………………………….DBU1500984 5. YONAS ABEBE …………………………………………………..DBU1501655 6. ABAYNEWUABERU …………………………………..……….DBU1501615 7. YONAS HABTAMU ……………………………………………DBU1501578

# 

# SUMITTED TO: inst.Derbew

# SUBMISSION DATE: 29,OCT 2025

# Design and Implementation of a Data Pipeline and ETL Process for NYC Yellow Taxi Trip Data

## 1. Introduction

This project presents the end-to-end design and implementation of an ETL (Extract, Transform, Load) pipeline for the NYC Yellow Taxi Trip dataset. The primary objective was to extract raw trip data, clean and transform it into a usable format, and load it into a PostgreSQL database for analytics. Through this project, we demonstrate practical expertise in data engineering, Python-based automation, and business intelligence pipeline design.

## 2. Objectives

* • Efficiently extract and process large-scale datasets using Python and Pandas.
* • Perform data cleaning, transformation, and enrichment for analytical readiness.
* • Load transformed data into PostgreSQL with a well-structured schema.
* • Maintain data integrity through reference tables and foreign key relationships.
* • Enable querying and CRUD operations for flexible data management.

## 3. Data Source

Dataset: NYC Yellow Taxi Trip Data (January 2025)  
Size: Approximately 3.47 million rows  
Key Columns: Vendor ID, Pickup & Dropoff timestamps, Passenger count, Trip distance, Payment type, Fare, Tip, Taxes, and Surcharges.  
  
Business Relevance: This dataset supports analyses such as ride demand forecasting, driver performance evaluation, and revenue optimization.

## 4. ETL Process

### 4.1 Extract

• Data was read from a Parquet file for speed and memory efficiency.  
• Conducted initial inspection to detect missing values, duplicates, and format inconsistencies.

### 4.2 Transform

• Data Cleaning: Replaced or removed missing and inconsistent values in key fields.  
• Derived Features: Created new metrics such as trip duration, pickup hour, and pickup date.  
• Standardization: Renamed columns and enforced consistent data types.  
• Validation: Filtered out invalid or zero-distance trips to maintain accuracy.

### 4.3 Load

• Implemented chunked data loading to handle millions of rows without memory overload.  
• Successfully loaded ~3.47 million records into PostgreSQL table 'trips'.  
• Added indexes on key columns to boost query performance.

## 5. Database Management

Designed a relational schema consisting of:  
- Main table: trips  
- Reference tables: vendors, rate\_codes, payment\_types, and locations.  
  
Foreign key constraints were enforced to ensure referential integrity. CRUD operations were demonstrated, including:  
• Inserting new trip records  
• Querying trips by pickup or payment type  
• Updating fare or tip details  
• Deleting invalid or test data

## 6. Results

* • Built a robust, scalable ETL pipeline for large datasets.
* • Achieved a fully cleaned, enriched, and standardized dataset.
* • Database schema design ensured data consistency and efficient querying.
* • Verified CRUD functionality for business analytics workflows.

## 7. Conclusion

This project successfully delivers a functional and maintainable ETL pipeline using Python, Pandas, and PostgreSQL. It demonstrates the complete data lifecycle—from ingestion to transformation and storage—while ensuring data quality, integrity, and usability. The final system forms a foundation for data-driven decision-making and can be extended for real-time analytics.

## 8. Future Enhancements

* • Automate the ETL pipeline for scheduled or continuous ingestion.
* • Implement data validation and anomaly detection modules.
* • Optimize the database with partitioned tables and advanced indexing.
* • Integrate with BI tools (e.g., Power BI, Tableau) for interactive dashboards and reporting.