Information Technology Institute
Data Management Track
End to End ETL pipeline using Python
(Intake 44/2024)

Objective:

The goal of this project is to design and implement an end-to-end Extract, Transform, Load (ETL) process for a bicycle store using Python. The ETL pipeline will handle data from various sources including databases, data lakes, and APIs, and will involve data extraction, data quality checks, transformations, and loading into a structured data model for analytics.

Milestones:

Stage 1: Database and Data Lake Setup

- Database Creation: Initialize a PostgreSQL database with schemas and tables for storing order and item data.
- Data Lake Configuration: Set up folders within a cloud storage solution to store additional data such as CSV files from different departments.

Stage 2: Data Extraction

- Database Extraction: Extract data from PostgreSQL using custom SQL queries.
- Data Lake Extraction: Read files from the data lake folder structure.
- API Integration: Fetch real-time exchange rates and store them in the landing folder.
- Data Consolidation: Combine all extracted data into a single CSV file per dataset, enrich them with metadata like extraction timestamp and data source.

Stage 3: Data Quality Checks

- Null Checks: Identify and handle null values in essential fields.
- Duplicate Checks: Detect and remove duplicate rows.
- Data Validation: Ensure data types and values are within expected ranges (e.g., price ranges, date limits).
- Preparation for Staging: Store cleaned and validated data in 'staging_1' folder for transformation.

Stage 4: Data Transformation

- Currency Conversion: Merge latest currency exchange rates to calculate local prices.
- Delivery Metrics: Add columns to track delivery performance, such as late deliveries and latency days.
- Locality Flag: Determine if customers are local based on proximity to stores.
- Lookup Tables: Resolve ambiguous columns by creating and utilizing lookup tables for order statuses.
- Transformed Data Staging: Output transformed data to 'staging_2' for further processing.

Stage 5: Data Modeling and Visualization

- Data Merging: Integrate orders, items, and product details into a unified dataset for deeper analysis.
- Visualization Creation: Develop at least three types of visualizations to illustrate key metrics and trends.
- Documentation: Document the modeling techniques and visualization choices.
- Result Compilation: Organize final datasets and visualizations in 'Information Mart' and 'Visualization' folders respectively.

Final Deliverables:

- Documented Python scripts and Jupyter notebooks for each stage of the ETL process.
- Clean and transformed datasets ready for analysis in structured folders.
- Visualizations and analytical reports demonstrating the insights derived from the data.

Assignment Type:

This project is an individual assignment designed to demonstrate proficiency in handling complex data workflows using Python and various data processing libraries.