**ETL Pipeline for OLTP to OLAP Data Migration**

**Project Overview**

This project implements a ETL (Extract, Transform, Load) pipeline designed to migrate and transform data from an OLTP system to an OLAP data warehouse. The system handles both dimension and fact table updates, ensuring data is processed efficiently and accurately. It integrates data from MongoDB and PostgreSQL sources into a unified OLAP schema to support reporting and analytics.

**Key Features**

1. **Incremental Data Loading:**
   * Only new or changed data is loaded from the source, saving time and resources.
2. **Managing Dimension Tables:**
   * Keeps tables like users, products, and dates up to date while avoiding duplicates and fixing conflicts.
3. **Fact Table Processing:**
   * Adds transaction data to the reporting database efficiently by reusing already loaded information.
4. **Error Handling:**
   * Handles errors carefully and undoes changes if something goes wrong.
5. **Tracking Progress:**
   * Keeps a record of what data has already been processed to avoid repeating tasks.
6. **Logging and Monitoring:**
   * Keeps track of what the system is doing and its performance to help with problem-solving and improvements.

**System Architecture**

1. **Data Sources**:
   * **MongoDB**: Stores raw sales data.
   * **PostgreSQL (OLTP)**: Hosts transactional data.
2. **Destination**:
   * **PostgreSQL (OLAP)**: Centralized data warehouse optimized for reporting.
3. **Key Tables**:
   * **Staging Tables**: storage for raw data.
   * **Dimension Tables**: Users (dim\_users), Products (dim\_products), Dates (dim\_date).
   * **Fact Tables**: Sales transactions (fact\_sales).
4. **Reporting Tool:**
   * Tableau is used for reporting purpose

**ETL Workflow**

1. **Extract**:
   * **From MongoDB**: Gets sales data in small parts (batches) using a filter to fetch only new records based on the last processed ID.
   * **From OLTP PostgreSQL**: Retrieves user and product data in small batches
2. **Transform**:
   * Splits and organizes product categories into meaningful groups.
   * Fixes duplicates or conflicts in user and product data.
   * Creates unique keys for date information to ensure consistency.
3. **Load**:
   * Puts data into staging tables for cleaning and validation.
   * Updates or adds data into final OLAP tables, ensuring no duplicates with ON CONFLICT.

**Technical Details**

**Dimension Table Loading**

1. **Users**:
   * Extract user data from the OLTP system.
   * Use bulk inserts with ON CONFLICT to handle updates.
2. **Products**:
   * Joins products and categories during extraction to enrich data.
   * Handles hierarchical category splits into main, sub, and sub-sub categories.
3. **Dates**:
   * Dynamically generates and caches date-time entries to avoid redundant inserts.

**Fact Table Loading**

* Loads transactional data by joining user and product dimensions with cached lookups.
* Inserts records into the fact table while ensuring no duplicates using ON CONFLICT.

**Staging Data from MongoDB**

* Extracts raw sales data in batches.

**Integration with Tableau**

To make reports and track performance, two types of Tableau reports were created for this project:

1. **Live connection Report**: Tracks the real-time performance of the OLAP database.
2. **Extract-Based Report**: Makes Tableau work faster by using a saved copy of data

**Project Configuration**

* **Batch Size**: The pipeline processes data in configurable chunks (default: 10,000 rows).
* **Database Credentials**:
  + Managed using environment variables or configuration files for security.
* **Logging**:
  + Detailed logs capture each step of the ETL process for monitoring and debugging.

**Future Enhancements**

1. **Partitioning:**
   * Split large tables by date to make queries faster and more efficient.
2. **Parallel Processing:**
   * Process data in multiple parts at the same time to handle large datasets quicker.
3. **More Data Validations:**
   * Add checks before and after loading data to ensure everything is correct and complete.

**Conclusion**

This ETL pipeline is a strong and efficient system for moving and transforming data from OLTP systems to an OLAP data warehouse. It follows good practices to keep data accurate, supports detailed reports, and is built to handle future improvements easily