Big Data Advance Project

Team Members:

1. **Chahat Kothari (ck3999)**
2. **Sajitha Mathi (sm12344)**
3. **Sarang P. Kadakia (sk11634)**

A Hybrid OLTP-DW System with an AI-Powered Analytics Interface

Abstract

This project focuses on the design and implementation of an end-to-end data management and analysis solution. The system will be composed of three primary components. First, a normalized Online Transaction Processing (OLTP) database will be developed to handle day-to-day business transactions with high efficiency and data integrity. Second, a denormalized Data Warehouse (DW) will be created, optimized specifically for analytical queries and reporting. An Extract, Transform, Load (ETL) pipeline will be built to incrementally transfer data from the OLTP to the DW. The final and most innovative component is an interactive LLM chatbot hosted on Streamlit. This chatbot will replace traditional business intelligence tools like Tableau by allowing users to ask questions and generate data visualizations and some insights in natural language, making data insights more accessible, dynamic and user-friendly.

Introduction

The modern business landscape is driven by data. However, the systems that capture transactional data are often not optimized for the complex analysis required for strategic decision-making. Our project addresses this challenge by creating a holistic solution that separates these functions. We will build a highly structured, normalized OLTP system for daily operations and a flexible, denormalized DW for in-depth analysis. This separation ensures that the performance of one system does not negatively impact the other. To bridge the gap between complex data and end-users, we will develop a cutting-edge LLM-powered chatbot. This tool will allow users to bypass the technical complexities of SQL and data visualization software, instead interacting with the data in a conversational, intuitive manner to retrieve insights.

Description & Executive Summary

This project outlines the development of a complete database system to serve both a business's daily transactional needs and its long-term analytical requirements. The core of the system is a two-part database architecture: a normalized Online Transaction Processing (OLTP) database for efficient data entry and retrieval, and a denormalized Data Warehouse (DW) for fast, complex querying. These two systems will be synchronized via an automated Extract, Transform, Load (ETL) pipeline that will capture and move updated data.

For example, while the OLTP will quickly process a new sale and a customer return, the DW

will be optimized for a monthly report that analyzes sales trends across all product lines and regions.

To provide a modern and user-friendly interface for data analysis, we will replace traditional reporting tools with an innovative LLM chatbot. This chatbot, hosted on Streamlit, will allow users to ask questions in plain English and receive instant data insights and visualizations. Its analytical capabilities will be powered by a Retrieval-Augmented Generation (RAG) pipeline, utilizing HNSW for efficient semantic search and intelligent keyword matching to ensure accurate and contextually relevant responses. The entire system is designed to be scalable, secure, and to provide a tangible demonstration of advanced database and AI principles.

Purpose

The purpose of this project is to apply advanced database principles to a real-world business case while integrating cutting-edge AI technology. We will demonstrate a comprehensive understanding of data management by building both an OLTP and a DW system. This dual-system approach showcases our ability to handle different data requirements such as transactional efficiency versus analytical insights.

Anticipated Benefits

The anticipated benefits of this project are:

- **Improved Data Accessibility:** The LLM chatbot will democratize data analysis by providing intuitive, conversational access to data insights.
- **Advanced Analytics:** The RAG pipeline with HNSW and keyword matching ensures accurate and contextually relevant responses for complex queries.
- **Increased Efficiency and Automation:** The automated ETL process streamlines data flow, eliminating manual effort and potential for human error.
- **Enhanced Data-Driven Decision Making:** The system empowers management to make faster, more informed decisions based on real-time data insights and predictive analytics.
  For instance:
  Imagine a manager at Awesome Inc. wants to understand why a specific product has a high return rate in certain regions. Instead of asking a data analyst to run a complex report, the manager can simply ask the chatbot: "Show me the return rate for Product X by region over the last quarter." The chatbot will not only provide the requested visualization but can also be further prompted to analyze the data, for example: "Are returns for Product X correlated with any specific promotions or weather events?" This direct, interactive access to data enables the manager to quickly identify and act on trends, such as adjusting marketing strategies or inventory levels, without any technical dependencies.

Justification

This project is justified as a valuable exercise in applying both core database principles and modern AI concepts. It demonstrates a holistic understanding of data architecture by implementing a two-part system, an OLTP for efficiency and a DW for analysis. The project's innovation lies in its integration of an LLM chatbot, which showcases a forward-thinking approach to business intelligence and makes data analysis more accessible and intuitive.

Objectives (SMART goal)

The objectives of this project are:

- **Specific**: Develop an end-to-end data management and analytics system for "Awesome Inc." consisting of an OLTP database (normalized up to 3NF), a Data Warehouse (DW) in star schema, an ETL pipeline with Change Data Capture (CDC), and a reporting interface. Instead of Tableau, our system will feature a Streamlit-based LLM chatbot for natural-language querying and visualization.

- **Measurable**:

    a. OLTP database will capture and process at least 100+ sales and return transactions with triggers maintaining `TBL_LAST_DT`.

    b. ETL pipeline will incrementally load updated records into the DW and support at least 3 full load cycles during testing.

    c. DW will successfully answer 5+ predefined analytical queries (e.g., sales by region, product return rate).

    d. Streamlit chatbot will correctly interpret and execute at least 10 natural-language queries, producing SQL outputs and visualizations.

- **Achievable**: The project leverages widely used database systems (Oracle/MySQL for OLTP, PostgreSQL/Oracle for DW), Python-based ETL with CDC, and existing LLMs (open-source like LLaMA 2 or closed-source like GPT-4). This ensures feasibility within the semester timeline.

- **Relevant**: This project applies core advanced database concepts required by the course (OLTP design, DW star schema, ETL, CDC, triggers, PL/SQL) while introducing innovation by integrating an AI chatbot for analytics, replacing traditional BI tools.

- **Time-Bound**: The full system (OLTP, DW, ETL, and Streamlit chatbot) will be designed, implemented, and demonstrated by Dec 8, 2025, with the final report and

presentation submitted by Dec 11, 2025.

Scope

A. Product Scope

The deliverable system will include:

- **OLTP Database**: Normalized schema (up to 3NF) with tables for Customers, Products, Stores, Sales, and Returns; each table will have `TBL_LAST_DT` with triggers for CDC.
- **Data Warehouse (DW)**: Star schema with central FACT_SALES and supporting dimensions (DIM_DATE, DIM_CUSTOMER, DIM_PRODUCT, DIM_STORE).
- **ETL Pipeline (CDC)**: Automated process to extract incremental changes from OLTP, transform them, and load them into the DW.
- **Frontend Application**:
    a. **OLTP Layer**: Data entry forms (sales, returns) with role-based access.
    b. **DW Layer**: Analytics dashboards and a Streamlit chatbot that accepts natural-language queries, generates SQL, and visualizes results.
- **Security & Validation**: Role-based user management, password handling, and data integrity enforcement through constraints and triggers.

B. Project Scope

- **In-Scope**:
    a. Database schema design and implementation (OLTP and DW).
    b. ETL development with CDC logic using `TBL_LAST_DT`.
    c. Development of a frontend (Streamlit) for data entry and analytics.
    d. Integration of an LLM (open- or closed-source) for natural-language queries.
    e. Sample reports/dashboards for key business questions (e.g., top-selling products, regional sales trends, product return rates).

- **Out-of-Scope**:
    a. Enterprise-scale deployment (prototype only).
    b. Real business datasets (synthetic datasets will be used).
    c. Custom LLM training (only pre-trained model integration).
    d. Advanced predictive analytics beyond basic trends and descriptive insights.

Backend

The backend will consist of two core database environments and supporting middleware for data movement and processing:

- **OLTP Database (Transactional Layer):**
    a. **Technology:** Oracle or MySQL (for relational integrity, concurrent transactions, and ACID compliance).
    b. **Design:** Highly normalized schema (up to 3rd Normal Form) with tables for Customers, Products, Stores, Sales, and Returns.
    c. **Key Features:**
        i. Enforces data validation through constraints and triggers.
        ii. Includes a TBL_LAST_DT column in each table for tracking incremental changes.
        iii. Supports real-time transaction management for multi-user environments.
    d. **Enhancements:** Triggers automatically update timestamps to support downstream ETL.
- **Data Warehouse (Analytical Layer):**
    a. **Technology:** PostgreSQL, Snowflake, or another DW-optimized platform.
    b. **Design:** Denormalized star schema with a central FACT_SALES table linked to dimension tables (DIM_DATE, DIM_CUSTOMER, DIM_PRODUCT, DIM_STORE).
    c. **Optimization:** Partitioning, indexing, and materialized views for faster query response.
    d. **Role:** Acts as the single source of truth for reporting and analytics.
- **ETL Middleware:**
    a. **Technology:** Python-based ETL pipelines (using pandas or frameworks like Apache Airflow).

b. **Function:**
    i. Extract new/updated records from OLTP using TBL_LAST_DT.
    ii. Transform records into DW-compatible schema.
    iii. Load data into DW tables incrementally.

Frontend

The frontend is designed as a **Streamlit web application**, enabling users to interact with both structured reporting tools and natural-language query features.

- **User Interface:**
    a. **Login & Role-based Access:** Store managers, sales associates, and executives each have distinct privileges.
    b. **Data Entry Forms:** Simple and intuitive UI for inserting or viewing sales and return transactions (directly linked to OLTP).
    c. **Analytics & Visualization:** Dashboards displaying KPIs such as sales trends, return rates, and regional performance using charts (bar, line, maps).
    d. **Natural Language Query:** An integrated LLM backend interprets user queries (e.g., “Show me the top 5 products by sales in Q2”) and dynamically generates SQL queries or visualizations.
- **LLM Integration:**
    a. **LLM Options:**
        i. **Open-source models** (e.g., LLaMA 2, Falcon, or Mistral) hosted locally or on a secure VM for cost efficiency and data privacy.
        ii. **Closed-source APIs** (e.g., OpenAI GPT-4 or Anthropic Claude) for higher accuracy and enterprise-level support.
    b. **Functionality:**
        i. Translates natural language queries into SQL for the DW.
        ii. Generates dynamic visualizations (via Matplotlib, Plotly, or Altair in Streamlit).
        iii. Provides text-based insights alongside visual analytics.
- **Hosting & Deployment:**
    a. Streamlit app hosted on a cloud platform (AWS, Azure, or GCP).
    b. Secure connections to databases (via SSL/TLS).
    c. Scalable deployment using Docker containers or Kubernetes for multi-user access.

Technical Solution Overview

The solution architecture integrates **transactional processing, analytical processing, and AI-driven analytics** into a seamless ecosystem:

- **Data Flow:**
    a. **Step 1:** Users input sales/returns data through the Streamlit forms → stored in OLTP DB.
    b. **Step 2:** ETL pipeline extracts incremental changes from OLTP → transforms → loads into DW.
    c. **Step 3:** Users query the DW via the Streamlit dashboard or natural-language interface.
    d. **Step 4:** LLM interprets natural language → generates SQL/graph → executes on DW → displays results interactively.
- **Security & Auditability:**
    a. Database-level security (role-based permissions, encrypted connections).
    b. Application-level security (hashed passwords, session tokens, role-based access control).
    c. Comprehensive audit trails for all transactions.
- **Scalability:**
    a. **OLTP Layer:** Scales vertically for transaction integrity.
    b. **DW Layer:** Scales horizontally for large-scale analytical queries.
    c. **Streamlit + LLM:** Containerized deployment supports concurrent multi-user workloads.
- **Advantages Over Tableau (Streamlit + LLM):**
    a. Fully open-source, cost-efficient, and customizable.
    b. Natural-language interface removes the barrier of learning BI tools.
    c. Extensible architecture: new visualizations or AI models can be plugged in easily.