GLOBAL SALES ANALYTICS PROJECT – END-TO-END DATA WAREHOUSING AND BUSINESS INTELLIGENCE SOLUTION

This document presents a comprehensive description of the Global Sales Analytics Project, an advanced initiative developed to simulate, prepare, store, analyze, and visualize large-scale transactional data in a controlled environment. The project was designed as a showcase of professional capabilities spanning data engineering, relational database development, SQL analytics, and modern business intelligence reporting. The outcome is a complete, portfolio-grade case study that demonstrates how raw data can be transformed into clear, actionable insights suitable for executive decision-making.

Project Purpose and Context

The primary purpose of the Global Sales Analytics Project was to demonstrate the full lifecycle of a modern data analytics solution. The objective was to replicate an enterprise scenario where a business needs to generate and process high-volume sales data, store it reliably, perform complex aggregations, and present findings through interactive reporting tools. This project was developed to validate proficiency across the entire pipeline of data handling and analysis.

In many real-world organizations, sales operations generate vast amounts of data every day. This data is often spread across multiple regions, product lines, and time periods, making it essential to design robust systems capable of handling volume, ensuring data quality, and supporting business stakeholders in their decision-making processes. The project captures this reality by simulating a multi-year dataset containing over one million records with attributes representing geographies, products, revenues, costs, and profitability measures.

Data Generation Process

The project began with the creation of a large synthetic dataset intended to closely resemble real-world sales transactions. A custom Python script was written to generate records covering multiple years, regions such as North America, Europe, and Asia, and a diverse set of product categories. Each record included essential transaction information, such as a unique identifier, the date of the sale, the quantity of units sold, and the unit price.

Calculated fields such as total revenue and profit were derived to support later analysis of financial performance.

To ensure realism, randomization logic was included to vary quantities, prices, and profit margins within sensible business ranges. For example, profit margins were generated as fractional values between 5% and 35%, reflecting typical variations in retail and e-commerce scenarios. The simulation process also included categorical fields like revenue classification to facilitate segmentation and comparison across different tiers of transaction volume.

The final dataset consisted of precisely 1,048,575 rows. Each entry maintained integrity across all columns and was structured in a tabular format compatible with further cleaning, transformation, and loading into a relational database.

Data Preparation and Cleaning

Once the synthetic dataset was generated, it was exported to a CSV file. Before moving the data into a warehouse environment, a thorough cleaning and preparation process was applied. This step was essential to eliminate any inconsistencies and ensure compliance with relational database standards.

A critical aspect of preparation was the standardization of date formats. Because MySQL requires date fields to be in the YYYY-MM-DD format, the Python cleaning script parsed all dates and restructured them accordingly. Additional validation was performed to confirm that all numeric fields, such as quantity, unit price, revenue, and profit, contained valid numeric values and no unexpected characters or blanks.

Further feature engineering was implemented to create derived columns such as Year and Month, enabling time-based aggregation in SQL queries and business intelligence dashboards. The final cleaned file was saved as a dedicated CSV, ready for ingestion into the MySQL environment.

Relational Database Design

The next phase of the project involved establishing a structured relational database schema. MySQL was chosen as the platform due to its widespread use in production environments and strong support for high-volume transactional data storage.

A dedicated schema named global_sales_db was created. Within this schema, a primary table called sales_transactions was defined. This table included well-structured fields such as transaction ID, date, region, country, product, quantity, unit price, cost, revenue, profit, profit margin, year, month, and revenue category. Each field was assigned an appropriate data type, including integers for identifiers and quantities, decimals for monetary amounts,

and strings for categorical fields. Constraints were defined to enforce data integrity, including a primary key on the transaction ID.

With the schema in place, the cleaned CSV file was placed in the secure upload directory as required by MySQL's secure file privilege settings. The data was ingested using the LOAD DATA INFILE command. Special care was taken to ensure there were no duplicate primary keys and that all records adhered to the schema's constraints. Verification queries confirmed the successful import of all 1,048,575 records.

SQL Query Development

A major component of the project involved developing a suite of advanced SQL queries designed to produce clear, actionable insights from the dataset. These queries were designed to answer common business questions such as:

- How does revenue vary by region and month?
- Which countries generate the highest total profit?
- What is the performance of each product in terms of revenue and profitability?
- How do different revenue categories contribute to the overall sales portfolio?
- What is the average profit margin for each product line?
- Which individual transactions generated the most revenue?

Each query was written carefully to leverage grouping, aggregation, ordering, and filtering operations. For example, the monthly revenue by region query grouped data by year, month, and region, summing total revenue and ordering results chronologically. Similarly, the product performance summary counted the number of transactions per product, calculated total revenue and profit, and sorted products in descending order of revenue contribution.

All queries were consolidated into a single SQL script file, providing a reusable asset that can be executed on any instance of the dataset. This approach also demonstrates the ability to produce well-organized, maintainable query libraries in a professional analytics environment.

Power BI Business Intelligence Reporting

With the data stored and accessible via SQL, the project advanced to the reporting phase. Power BI Desktop was used to build a robust business intelligence report connected directly to the MySQL warehouse.

The report was divided into two main sections:

Executive Overview Page:
This section was designed for high-level stakeholders to quickly assess the overall health and performance of sales operations. It featured:

- Three KPI cards displaying total revenue, total profit, and average profit margin.
- A bar chart visualizing revenue distribution across regions.
- A line chart illustrating monthly revenue trends segmented by year.
- Slicers enabling dynamic filtering by year and month.

Product and Country Insights Page:
This page was intended for more detailed exploration. It included:

- A geographic map displaying revenue by country.
- A bar chart ranking top products by revenue.
- A second bar chart showing the average profit margin per product.
- Slicers for region, product, and year to allow in-depth analysis.

Visual formatting adhered to professional design standards, with consistent colors, clear titles, and data labels to ensure clarity and usability.

Project Deliverables and Organization

The project was carefully structured to produce a complete set of deliverables suitable for inclusion in a professional portfolio. The deliverables include:

- The cleaned dataset file containing all generated transactions.
- The MySQL schema definition and fully populated table.
- A consolidated SQL script with all analytical queries.
- The Power BI report file (.pbix) containing all visuals and interactivity.
- Optional exported screenshots or PDF reports for demonstration purposes.

This package can be shared via Google Drive or presented to prospective employers, clients, or collaborators as proof of advanced data analytics capability.

Competencies Demonstrated

The Global Sales Analytics Project highlights a combination of competencies highly relevant to modern analytics and business intelligence roles, including:

- Data simulation and preparation for large-scale analysis.
- Relational database design and high-volume data ingestion.
- SQL query development for operational and strategic insights.
- Interactive dashboard creation using industry-standard BI tools.
- Effective communication of data-driven findings through professional documentation.

The project reflects a disciplined approach to structuring, validating, and presenting data in a format that is clear, impactful, and immediately relevant to business decision-making processes.

Conclusion

The Global Sales Analytics Project is a comprehensive demonstration of end-to-end analytics capability, beginning with data generation and culminating in actionable visual reporting. Each phase of the project was executed with attention to accuracy, clarity, and professional presentation. The resulting solution can be adapted to real-world sales operations or extended to additional datasets and reporting requirements. It serves as a strong example of what can be achieved by combining data engineering, SQL development, and modern business intelligence practices into a cohesive, results-oriented project.