Sales Data Analytics & Pipeline POC

Summary

This report presents an analysis of sales performance across countries, and sales representatives. It highlights key insights, identifies data quality issues, outlines the data architecture, and recommends enhancements to scale and improve the solution.

Key Insights

- The dataset includes 27,000 sales records and 2,700 rebate records.
- **Revenue by Country:** Greece leads in net revenue, while the United States is lower performing.
- **Top Performing Sales Representative:** Janey Hanbury (Mexico) recorded the highest sales.
- **Sales Concentration:** Over 50% of total sales are driven by representatives from the United States and the United Kingdom.
- **Customer Distribution:** Customers are fairly evenly distributed across countries.
- **Cross-Border Purchases:** 87% of customers made purchases from a country different than their own, and 40% are from the same region but a different country.
- **Segment Performance:** The consumer product segment drives the majority of purchases across all countries. Across the Consumer and Corporate segments, Office Supplies is the dominant category, while Furniture leads in the Home Office segment.

Data Quality Assessment

- **Currency Variations:** Multiple currencies within the same dataset limited cross country comparisons.
- **Sales Representative Location Errors:** Misalignment between Sales Rep Country and Sales Rep Region. For example, U.S. reps are divided into Central, East, West, and South regions instead of a consistent North America region.
- **Customer Name Mismatches:** 69% of rebate records do not match customer names in the sales dataset.
- **Missing Representatives:** Germany, Canada, and Mexico lack dedicated sales reps; sales may be attributed to reps from other countries in the same region.
- **Unrealistic Transaction Amounts:** Several transactions had unrealistically high sale amounts (e.g., 1.124E12 for a standard office chair). These entries were flagged and excluded only in analyses where the amount is relevant, such as total sales calculations. 135 records in sales data and 27 in rebates data.

Data Architecture & Processing

Landing / Raw Layer:

- Raw files loaded without modification to preserve source data integrity.

Bronze Layer:

- Added metadata (ingestion time, source file path).
- Standardized column names.
- Stored as Delta Parquet files for durability and query efficiency.

Silver Layer:

- Cleaned column values and updated data types where necessary.
- Split multi-valued columns and removed unnecessary columns.
- Stored as Delta Parquet files for consistency and easy downstream processing.

Gold Layer:

- Fact table: Sales
- Dimension tables: Products, Customers, Sales Representatives
- Star schema enables efficient analytics, simplified reporting, and clear fact to dimension relationships.

Why the Medallion Architecture Delta Parquet?

The Medallion structure creates a clear progression from raw data to fully refined analytics ready tables. It strengthens governance, simplifies debugging, and builds trust in the final outputs because each layer adds structure without overwriting the previous one, making lineage and auditing easy to trace. Pairing this with Delta Parquet brings ACID transactions, schema enforcement, time travel, and efficient storage. It supports incremental processing, safe concurrent reads and writes, and version history, which makes the entire pipeline more reliable and scalable as the datasets grow.

Improvements & Future Development

While the current POC provides a solid foundation, several enhancements could make it more robust. Automated data validation and cleansing would address issues such as mismatched customer names, missing sales representatives, and inconsistent currencies. Implementing slowly changing dimensions (SCD) for dimension tables and incremental loads for the fact table would ensure accurate historical tracking and efficient updates. Analytics could be

further enhanced with trend analysis, predictive forecasting, and cohort segmentation to better understand cross-border behaviour and segment performance.

To handle growing data volumes over time, partitioning would be an effective strategy. The fact table could be partitioned by date, while dimension tables could be partitioned by frequently queried columns to improve query performance and manageability. Automating the pipeline using tools like Prefect or Airflow, combined with cloud-native storage, would improve scalability and overall performance. Additional improvements could include multi-currency support and creating views or stored procedures to enable ad-hoc analysis for business users. Collectively, these enhancements would transform the POC into a scalable, reliable, and insight-driven solution capable of supporting both operational and strategic decision-making.