Salary Data Processing & Analysis Project

Data Overview

This project demonstrates the implementation of a fully automated, end-to-end data pipeline using a suite of Azure services. The primary objective is to ingest raw salary data, process it through structured layers (Bronze, Silver, and Gold), and deliver meaningful insights through a Power BI dashboard.

The pipeline leverages Azure's scalable ecosystem to manage data storage, transformation, and visualization seamlessly. Automation is achieved using Azure Data Factory to orchestrate notebook execution on Azure Databricks, ensuring efficiency and consistency across each stage.

Technologies Used

- ✓ **Azure Blob Storage** – for raw data ingestion
- ✓ **Azure Data Lake Storage Gen2** – structured data storage (Bronze, Silver, Gold)
- ✓ **Azure Databricks (PySpark)** – data transformation and processing
- ✓ **Azure SQL Database** – analytical data storage
- ✓ **Azure Data Factory** – pipeline orchestration and automation
- ✓ **Power BI** – data visualization and reporting

Project Objective

The objective of this project is to:

- ✓ Build a structured data pipeline using PySpark.
- ✓ Implement the Medallion Architecture for efficient data storage and processing.
- ✓ Clean and transform raw data into meaningful insights.
- ✓ Aggregate data and store results in SQL Server.
- ✓ Visualize the final outputs using Power BI to aid in business decision-making through key performance indicators (KPIs).

Key Insights & Findings

Quarterly Salary Trends

- ✓ The **sum of total salary value increases steadily** from Q1 to Q4, indicating either consistent hiring, salary increments, or both across the year.

Monthly Salary Variation

- ✓ There is **noticeable fluctuation in monthly salary payouts**, with a dip in February and peaks around May and August.
- ✓ This could reflect bonuses, seasonal payouts, or variable pay structures.

Top Earners

- ✓ The **top 10 employees** listed (e.g., Employee 18517, 48604, 24898) account for **significant portions of total salary**, showing a skewed salary distribution possibly due to high-level positions or long-tenured staff.

Departmental Employee Distribution

- ✓ The pie chart shows a **broad spread of employees across departments**, with Departments 16, 37, 1, and 6 having the highest headcounts.
- ✓ No single department dominates the headcount, suggesting a fairly balanced organizational structure.

Monthly Salary Consistency

- ✓ The **sum of salary by month number** remains relatively stable across months, with some slight dips, possibly due to variable leave/bonus months or payroll adjustments.

Average Salary Indicator

- ✓ The **average salary amount is around 74.92K**, giving a quick snapshot of pay scale across the organization.

Limitations or Challenges Faced

Small Dataset: The data volume was limited, which might not fully reflect real-world complexities.

Local Environment Setup: Installing and configuring SQL Server and Power BI on a local machine required careful handling of compatibility and dependencies.

Data Quality Issues: Raw files contained missing values, especially in salary records, which required careful cleaning to ensure accurate transformation.

No Real-Time Processing: The pipeline was batch-based and not configured for streaming or real-time ingestion.

Future Scope Scalability

Extend the project to handle larger datasets or integrate it with cloud platforms like Azure Databricks, Azure Data Factory and Azure SQL.

Automation: Automate the entire ETL process using workflow schedulers like Azure Data Factory or triggers.

Real-Time Dashboard: Enable near real-time reporting by integrating stream processing tools such as Kafka or Spark Structured Streaming.

Advanced Analytics: Incorporate predictive analytics and machine learning to forecast salary trends and workforce requirements.

Security & Audit Logging: Implement robust logging, error handling, and access control mechanisms for production readiness.

Conclusion

This project demonstrates a full-fledged data engineering pipeline using industry-standard tools and architecture. By leveraging the Medallion Architecture, it successfully:Ingested and stored raw data in a structured manner.

Cleaned and enriched the data for analysis.

Aggregated key metrics for business insights.

Delivered an interactive Power BI dashboard to stakeholders.