**Project Report: End-to-End Azure Cloud Data Engineering & Analytics Pipeline**

**Project Overview**

|  |  |
| --- | --- |
| **Project Title** | **Hybrid Cloud Data Migration and ETL/ELT Analytics Platform** |
| Analyst/Owner | Md Akram Khan |
| Date | December 2025 |
| Architecture Pattern | Medallion Architecture (Raw, Transformed) |
| Key Achievement | Established an enterprise-grade, monitored data pipeline from On-Premises to Azure Cloud. |

## **1. Executive Summary From CSV to Cloud BI**

This project demonstrates the design and deployment of a complete cloud-based data ingestion, transformation, storage, and analytics ecosystem. Starting with on-premise CSV datasets, the entire workflow was migrated to Azure Cloud using modern data engineering best practices. The architecture leverages **Azure Data Factory** for ingestion, **Azure Data Lake Gen2** for storage, **Azure Databricks** for scalable processing, **Azure Synapse** Analytics for querying, and **Power BI** for business intelligence. **Automated monitoring**, notifications, security, and **version control** were implemented to deliver an enterprise-grade solution.

### **1.1. Business Problem**

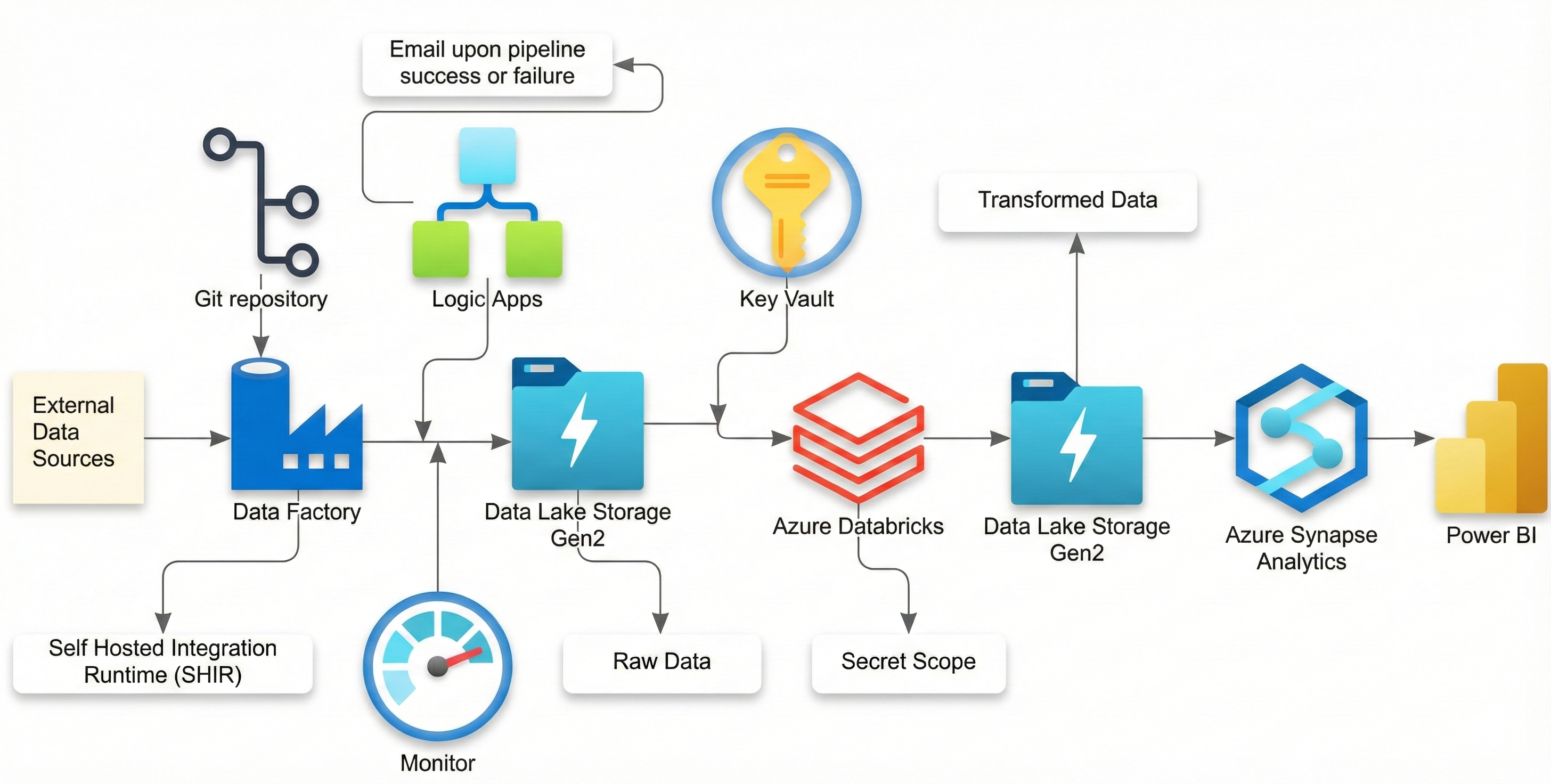
The organization maintained multiple on-premises CSV files (accounts.csv, products.csv, sales\_teams.csv, data\_dictionary.csv, sales\_pipeline.csv).  
 The goals were:

* Migrate on-premise data to Azure Cloud
* Build a scalable, automated data ingestion pipeline
* Ensure data quality through transformation in Databricks
* Store processed data in Data Lake for analytics
* Build business dashboards for revenue insights
* Implement monitoring, logging, automation, notifications

## **2. Solution Architecture and Implementation**

### **2.1. Architecture Overview**

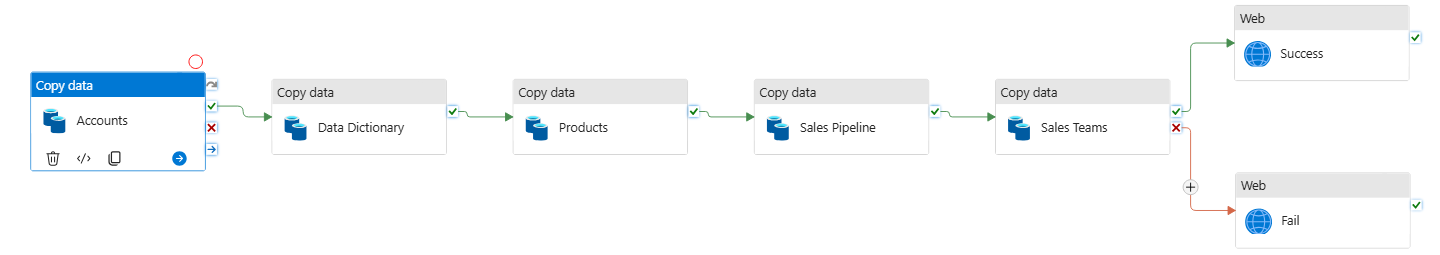
The platform is designed to handle hybrid data movement, processing, and advanced analytics in a secure and scalable manner.



### **2.2. Key Services Utilized**

|  |  |  |
| --- | --- | --- |
| **Category** | **Service** | **Function** |
| **Ingestion** | Azure Data Factory (ADF) | Orchestration of the ETL pipeline; uses Self-Hosted Integration Runtime (SHIR) for secure on-prem connectivity. |
| **Storage** | Azure Data Lake Storage Gen2 (ADLS Gen2) | Primary storage with Hierarchical Namespace enabled; used for Raw and Transformed data containers. |
| **Security/**  **Monitoring** | Azure Logic Apps, Azure Monitor | Provides email notifications for pipeline success/failure and offers deep operational monitoring. |
| **Processing** | Azure Databricks | Cloud-scale processing for data cleaning, validation, and transformation using PySpark. |
| **Analytics** | Azure Synapse Analytics | Serverless SQL pools for complex querying, aggregation, and analytical insights on cleaned data. |
| **Visualization** | Power BI | Interactive business intelligence dashboard for data consumption and insight generation. |

### **2.3. Data Ingestion Pipeline (ADF)**

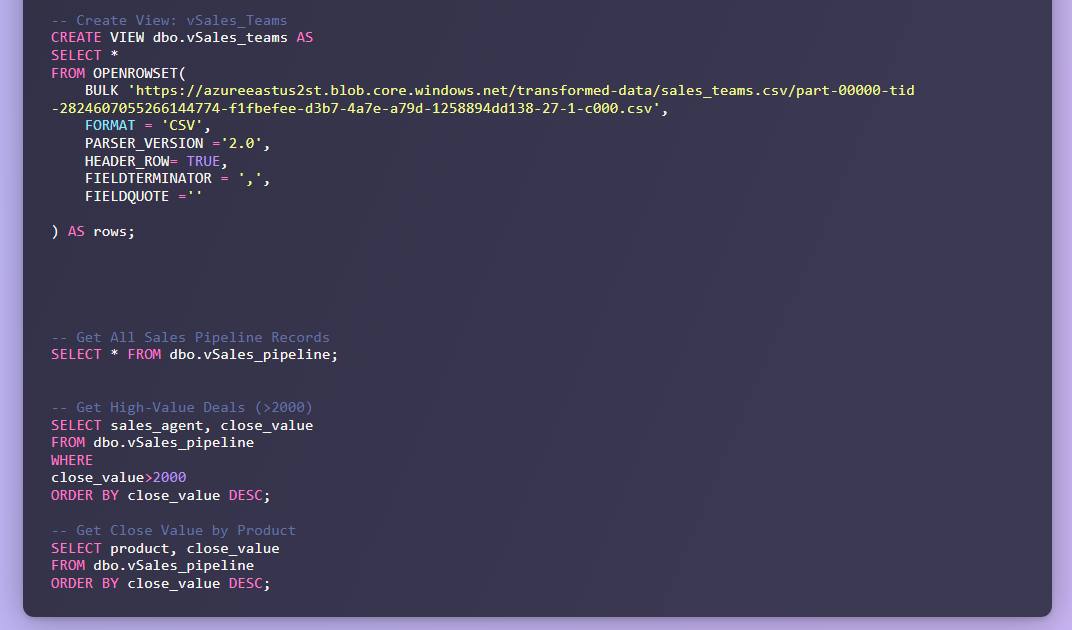
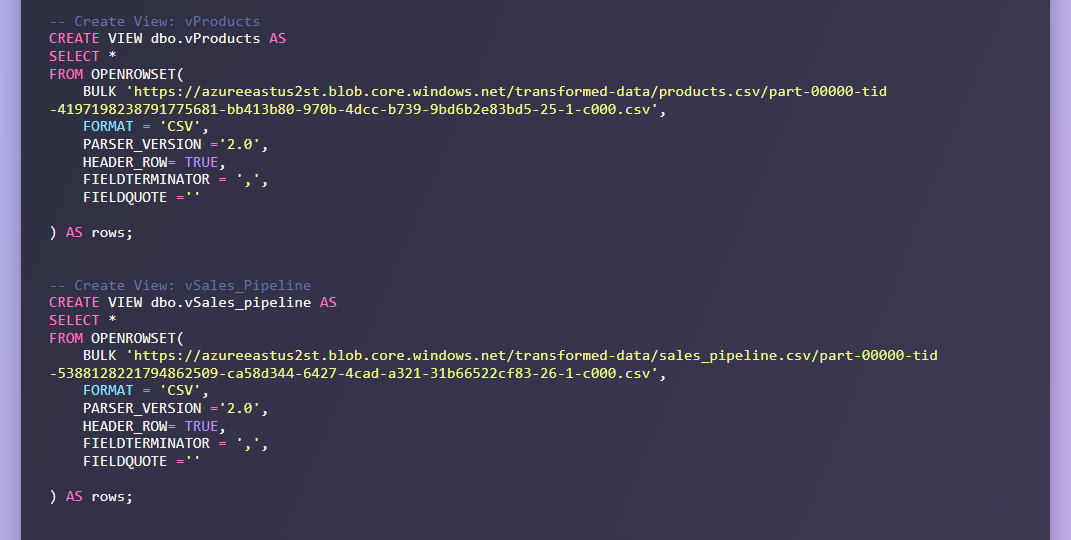
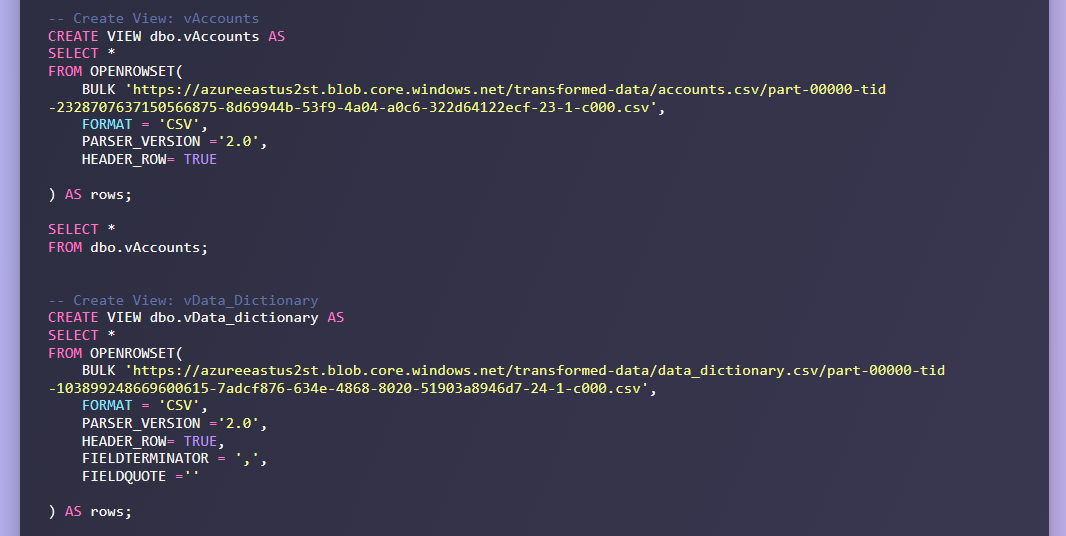
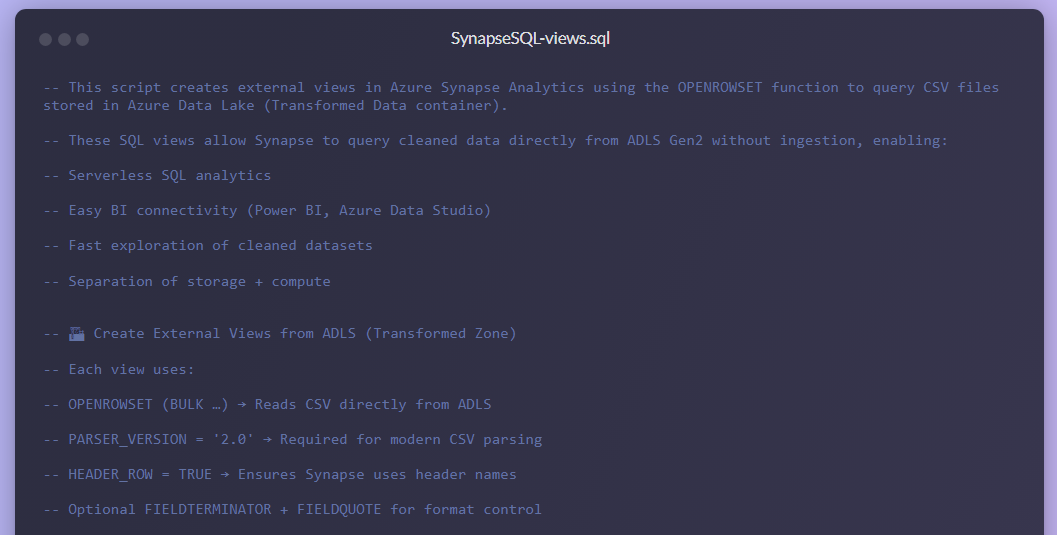


1. **Azure Setup:** ADLS Gen2 initialized with two containers: raw-data and transformed-data.
2. **Hybrid Connectivity:** A **Self-Hosted Integration Runtime (SHIR)** was deployed on the on-premises machine to enable secure, encrypted data transfer with ADF.
3. **Pipeline:** An ADF pipeline was constructed to perform five sequential "Copy Data" activities, moving each CSV file from the on-premises location to the **raw-data** container in ADLS Gen2.

## **3. Data Transformation and Analytics** **3.1. Data Transformation in Databricks**

1. **Data Cleaning :** Azure Databricks was connected to the ADLS Gen2. Notebooks (PySpark/Python) were developed to read the raw CSV files, perform data quality checks, handle missing values, standardize formats, and enrich the data.
2. **Output:** The cleansed, production-ready data was then written back to the **transformed-data** container in ADLS Gen2 in a performant format (e.g., Parquet).

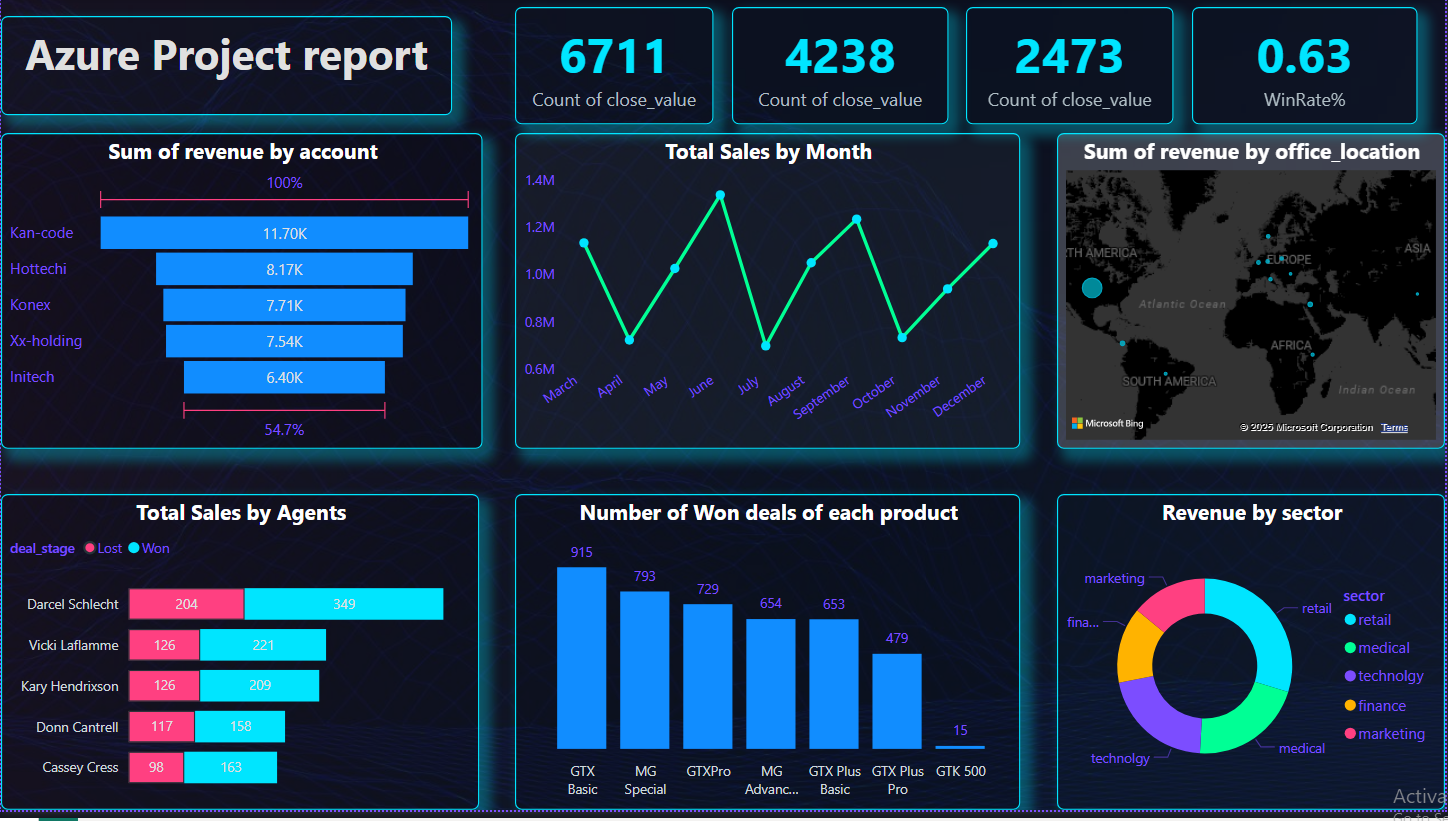
### **3.2. Synapse SQL Analysis and Key Metric Derivation**



1. **Data Warehousing:** Azure Synapse Analytics was configured to connect to the transformed-data container, leveraging its serverless SQL pool capabilities to query the data directly (Querying Lakehouse).
2. **SQL Insights:** Complex analytical queries were executed in Synapse to derive key business metrics (e.g., Win Rate, total close value, revenue by sector).

## **4. Business Value and Visualization**

The project culminated in a Power BI dashboard, providing executives with immediate, actionable intelligence based on the clean, trusted data from the Silver layer.

****

**Key Business Metrics Delivered:**

|  |  |  |
| --- | --- | --- |
| **Metric** | **Value** | **Impact** |
| **Total Deals Close** | 6,711 | Volume of activity processed. |
| **Total Close Value** | $4.238 Million | $4.238 Million, Top-line revenue figure. |
| **Overall Win Rate** | 63% | Key indicator of sales efficiency and process health. |
| **Top Segments** | Retail, Medical, Technology | Identified highest-performing market sectors. |

## **5. Conclusion**

This project validates proficiency in designing and deploying a secure, highly-available, and modern data platform on Azure. The full solution demonstrates expertise across hybrid connectivity, large-scale data processing, operational reliability, and delivering final business intelligence.