**A blue and white background

AI-generated content may be incorrect.**

A blue logo with a black background

AI-generated content may be incorrect.

**Azure Data Engineering**

**Project Report**

**Author:** Milan Tony

**Date:** 09-07-2025

**Contents**

[**1.** **Overview** 3](#_Toc202993128)

[**2.** **Setting Up the Azure Environment** 4](#_Toc202993129)

[**2.1.** **Resource Provisioning** 4](#_Toc202993130)

[**2.2.** **Security & IAM** 4](#_Toc202993131)

[**3.** **SQL Database Configuration** 5](#_Toc202993132)

[**4.** **Configuring Azure Data Factory** 6](#_Toc202993133)

[**4.1.** **Self-Hosted Integration Runtime** 6](#_Toc202993134)

[**4.2.** **Bronze Ingestion Pipeline** 6](#_Toc202993135)

[**5.** **Data Transformation with Databricks** 6](#_Toc202993136)

[**5.1.** **Cluster & Mount** 6](#_Toc202993137)

[**5.2.** **Bronze ➜ Silver (L1)** 6](#_Toc202993138)

[**5.3.** **Silver ➜ Gold (L2)** 6](#_Toc202993139)

[**5.4.** **Operationalizing via ADF** 7](#_Toc202993140)

[**6.** **Synapse Analytics Serving Layer** 7](#_Toc202993141)

[**7.** **Power BI Reporting** 8](#_Toc202993142)

[**8.** **Automation & Monitoring** 8](#_Toc202993143)

[**9.** **Security & Governance** 9](#_Toc202993144)

[**10.** **End-to-End Test** 9](#_Toc202993145)

[**11.** **Conclusion** 9](#_Toc202993146)

1. **Overview**

This project aims to bridge the knowledge gap in understanding customer demographics - particularly gender distribution - and its impact on product purchases. By building an automated Azure-based data pipeline we deliver a Power BI dashboard that surfaces Key Performance Indicators (KPIs) such as total products sold, total sales revenue and gender split, with flexible category and date filtering.

**Business Request**

In this project, a company has recognized a gap in understanding its customer demographics - specifically, the gender distribution within the customer base and how it might influence product purchases. With a significant amount of customer data stored in an on-premises SQL database, key stakeholders have requested a comprehensive KPI dashboard.

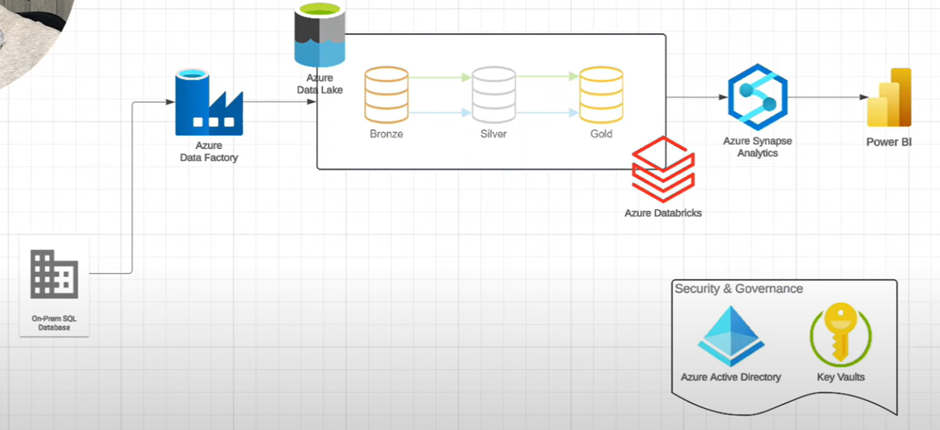
This dashboard should provide insights into sales by gender and product category, showing total products sold, total sales revenue, and a clear gender split among customers. Additionally, they need the ability to filter this data by product category and gender, with a user-friendly interface for date-based queries.

**Solution Overview**

To address this request, I built a robust data pipeline that extracts the on-premises data, loads it into Azure, and performs the necessary transformations to make the data more query-friendly. The transformed data will then feed into a custom-built report that meets all the specified requirements. This pipeline will be scheduled to run automatically every day, ensuring that stakeholders always have access to up-to-date and accurate data.

Database used: [AdventureWorksLT2022](https://learn.microsoft.com/en-us/sql/samples/adventureworks-install-configure?view=sql-server-ver17&tabs=ssms)

Project contents: <https://github.com/milantony05/data-engineering-project.git>



1. **Setting Up the Azure Environment**
   1. **Resource Provisioning**

|  |  |  |
| --- | --- | --- |
| **Resource** | **Purpose** | **Notes** |
| Resource Group (RG) | Logical container for all artefacts | ‑ |
| Azure Data Factory (ADF) | Orchestrate data movement & transformation | Created inside RG |
| Storage Account (Data Lake Gen2) | Persist raw (bronze), cleaned (silver) & curated (gold) data | Bronze container created |
| Azure Databricks | Spark-based transformations | Workspace auto-creates its own RG |
| Azure Synapse Analytics | Serverless SQL pool for BI serving layer | Connected to Data Lake |
| Azure Key Vault | Secrets management (SQL pwd, Databricks token) | Access via RBAC |

* 1. **Security & IAM**

• Key Vault Administrator role assigned to RG to allow secret creation.  
• Managed Identities & Entra ID security group used instead of individual principals for least-privilege governance.

1. **SQL Database Configuration**

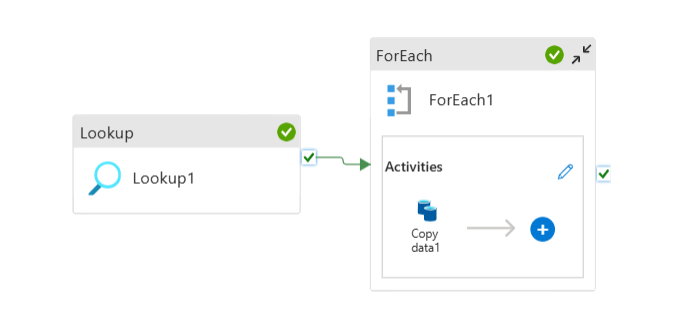
On-premises SQL Server Express hosts AdventureWorksLT2022 sample DB. Steps:

1. Install SQL Server, SSMS and restore AdventureWorksLT2022 backup.
2. Create SQL login user (milan) with SELECT rights on SalesLT schema.
3. Store credentials in Key Vault (sql-username, sql-password).
4. **Configuring Azure Data Factory**
   1. **Self-Hosted Integration Runtime**

Installed on the on-prem machine to enable outbound copy from SQL Server to Azure.

* 1. **Bronze Ingestion Pipeline**

1. Lookup activity enumerates all tables under SalesLT schema using dynamic SQL.
2. ForEach + CopyData iterates through list, extracting each table into Parquet files at bronze/<schema>/<table>/<table>.parquet.
3. Trigger: scheduled daily; also manual reruns for schema/permission fixes.

****

1. **Data Transformation with Databricks**
   1. **Cluster & Mount**

Single-node cluster (DBR 11.x), credential passthrough enabled. Data Lake containers mounted to /mnt/bronze, /mnt/silver, /mnt/gold.

* 1. **Bronze ➜ Silver (L1)**

• Convert ModifiedDate & any datetime columns to date.  
• Write each refined table to Delta format in silver layer.

* 1. **Silver ➜ Gold (L2)**

• Standardize column names to UPPER\_SNAKE\_CASE.  
• Perform joins/aggregations to build star-schema facts & dimensions.  
• Persist as Delta in gold layer.

* 1. **Operationalizing via ADF**

ADF notebook activities chain after ingestion pipeline: Bronze-to-Silver followed by Silver-to-Gold, using Databricks linked service authenticated via Key Vault-held PAT.

A screenshot of a computer

AI-generated content may be incorrect.

1. **Synapse Analytics Serving Layer**
2. Create gold\_db in serverless SQL pool.
3. External views (dbo.<table>) created automatically by stored procedure CreateSQLServerlessView\_gold, executed per table through Synapse pipeline with Get Metadata → ForEach → Stored Proc activities.
4. Views auto-reflect underlying Delta updates; pipeline rerun needed only for schema changes.

**A screenshot of a computer

AI-generated content may be incorrect.**

1. **Power BI Reporting**

• Power BI Desktop connects via DirectQuery to Synapse serverless endpoint.  
• Data model defines relationships between fact & dimension views (e.g., SalesOrderDetail ↔ Product).  
• Report visuals: KPI cards (Total Products, Total Sales), Donut chart (Gender distribution), slicers for Title (gender proxy) & Product Category.  
• Refresh triggered manually or via Power BI service after pipeline completes.

**A screenshot of a computer

AI-generated content may be incorrect.**

1. **Automation & Monitoring**

|  |  |  |
| --- | --- | --- |
| **Layer** | **Mechanism** | **Frequency** |
| Ingestion (SQL ➜ Bronze) | ADF Schedule Trigger | Daily @ 02:00 UTC |
| Transform (Bronze ➜ Gold) | Chained ADF activities | Follows ingestion |
| Synapse View Refresh | Not required (views are virtual) | N/A |
| Power BI Dataset Refresh | Pro licence scheduled | After pipeline |

ADF & Synapse provide run-level monitoring with retry & alert configuration. Delta time-travel in silver/gold offers data-quality rollback.

1. **Security & Governance**

• Entra ID security group controls contributor rights across RG.  
• Managed identities used for Synapse ↔ Data Lake & ADF ↔ Key Vault interactions.  
• RBAC roles: Storage Blob Data Contributor (Synapse), Key Vault Secrets User (ADF), etc.

1. **End-to-End Test**

Inserted new Product row in on-prem SQL. Scheduled trigger executed full pipeline; Power BI refresh displayed incremented product count, verifying data freshness.

1. **Conclusion**

The delivered Azure data platform ingests on-prem transactional data, applies scalable Spark transformations, serves analytics through serverless SQL and visualizes insights in Power BI - all orchestrated and secured via Azure-native services. Stakeholders now possess a gender-aware sales dashboard refreshed daily, enabling data-driven marketing and inventory decisions.