# Retail Store Analytics

What I did in this project?

1. Created Azure SQL Server and Azure SQL DB
2. Executed SQL scripts to create data sources
3. Pushed SQL scripts and JSON files to GitHub repo
4. Created a storage account for ADLS Gen2 with namespace hierarchy
5. Created a container called **retail-container**
   1. Following the medallion architecture, I created 3 directories – bronze, silver and gold
   2. In the bronze folder, I have created sub folders for the data sources like **retail transactions, stores, products** and **customers**
6. Ingested the aforementioned raw data sources from source systems – Azure SQL DB, JSON file using Azure Data Factory
   1. Created a pipeline to ingest data and insert into the respective sub-folders within the bronze folder and kept as our raw data store
   2. Sunk the raw data in form of parquets because…
7. Created workspace for Azure Databricks
8. Created and started a compute cluster in Databricks
   1. Issues found: **SKUNotAvailable** because of the region associated with workspace
      1. **Fixed**: queried list of resources available by location (eastus), ResourceType, Name, Zones, and Restrictions. Selected the resource that has norestriction and within the zone in which the databricks workspace was created on
9. Wrote PySpark code to
   1. Mount data – read files available from ADLS
      1. How to get secret access key: Security + Networking -> Access Keys -> click copy on **Key** and paste in the parameter of the mounting script in databricks notebook