**Bootcamp Project 2 — Transactions and Loan Data for a Customer**

**Project Objective**

Design and implement a robust, scalable data pipeline for processing customer account and loan data using modern Azure services:

* Ingest data from **ADLS Gen2 (Bronze Layer)**
* Mount storage using **Service Principal Authentication**
* Clean and transform using **Databricks Notebooks (Silver Layer)**
* Create **Delta Tables with SCD Type 1 (Gold Layer)**
* Enable downstream analytics and **visualization in Power BI**

**Tools and Technologies Used**

* **Azure Data Lake Storage Gen2 (ADLS Gen2)**
* **Service Principal** for secure authentication.
* **Azure Key Vault**
* **Azure Databricks (PySpark and SQL)**
* **Delta Lake**
* **Power BI**

**ARCHITECTURE DIAGRAM** **A diagram of a computer process

Description automatically generated**

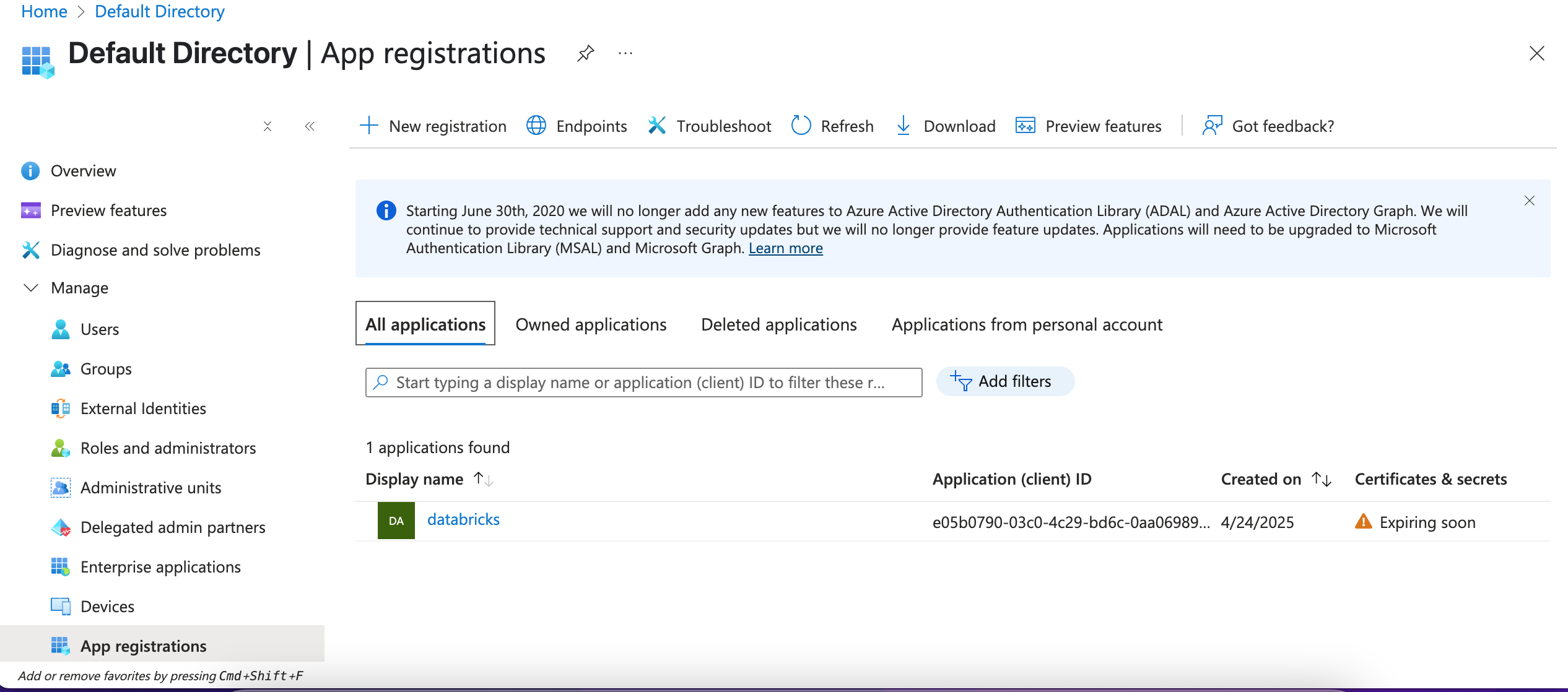
**Project Steps**

**Step 1: Data Ingestion (Bronze Layer)**

* **Source Files:**
  + accounts.csv
  + customers.csv
  + loan\_payments.csv
  + loans.csv
  + transactions.csv
* **Sink:**
  + ADLS Gen2 Raw (Bronze) Container
* **Reference:**
  + [Kaggle AI Bank Dataset](https://www.kaggle.com/datasets/varunkumari/ai-bank-dataset)

**Step 2: Mounting Storage & Data Cleaning (Silver Layer)**

* **Mount ADLS Gen2 Storage** to Databricks using Service Principal credentials stored in Azure Key Vault

Microsoft Entra ID 🡪Manage 🡪 App Registration 🡪Create a service Principle 🡪 copy application ID and Tenant ID 🡪 go to clien services and create a new secrete  A screenshot of a computer

Description automatically generatedcopy secrete value and create secrete (app ID and app Value ) in keyvalue A screenshot of a computer

Description automatically generatedCreate a scope in databricks A screenshot of a computer

Description automatically generatedNow create a mount point in adls gen2 A screen shot of a computer code

Description automatically generated Read the raw CSV files into Spark DataFrames.

Remove poor quality data and prepare consistent Parquet files.

**Actions Taken:**

* Dropped all **null values** (.na.drop()) to avoid incomplete records.
* Removed **duplicate records** (.dropDuplicates()).
* Saved each cleaned dataset as **Parquet** into the **Silver layer** (/mnt/edw/silver/...).

A screenshot of a computer code

Description automatically generated Joining Datasets and Creating a Master Data Frame

Create a combined view containing necessary customer + account + transaction + loan + payment information.

**Actions Taken:**

* Performed multiple joins:
  + Joined Accounts and Customers on customer\_id
  + Joined the result with Transactions on account\_id
  + Joined further with Loans on customer\_id
  + Finally joined with Loan Payments on loan\_id
* Selected only required columns:
  + account\_id, transaction\_id, customer\_id, loan\_id, payment\_id, amount, date
* Removed duplicates on the combination of selected fields.

## A screenshot of a computer program Description automatically generated Store the enriched, trusted master dataset into the **Gold Layer** in **Delta** format. **SCD Type 1 Implementation (Slowly Changing Dimension)**

**Goal:**  
Maintain the Gold Delta Tables by handling changes (update old records, insert new ones).

**Actions Taken:**

* Created a hash\_key (using CRC32) on all columns to detect changes easily.
* Used **Delta Lake MERGE** operation:
  + **When Matched** → Update the record if changed.
  + **When Not Matched** → Insert as new record.
* Managed audit columns:
  + created\_date, created\_by
  + updated\_date, updated\_by

**Create a Delta Table (if not exists)**

* Creates a **Delta table** for customer data.

Stores data in **Delta format**, enabling ACID transactions & time travel.

* A screenshot of a computer

  Description automatically generatedDefine Source and Target path A screenshot of a computer

  Description automatically generatedread the data A screen shot of a computer

  Description automatically generatedGenerate hashkey A screenshot of a computer screen

  Description automatically generatedload data and perform anti join A screenshot of a computer code

  Description automatically generated **Merge Changes into the Delta Table**
* **Merges new and updated records into Delta table**.
* **Updates existing records** where ID matches.
* **Inserts new records** where no match is found. A screenshot of a computer code

  Description automatically generateddisplay the updated record A screenshot of a computer

  Description automatically generated A screenshot of a computer

  Description automatically generated