Bootcamp2 Project Documentation

# Project Overview

The objective of the Bootcamp2 project is to design and implement a robust data pipeline for processing customer account data. This includes copying data from a backend team's storage account, performing necessary transformations using Databricks, and upserting data into a dedicated SQL pool table in Azure Synapse Analytics.

# Tools and Technologies Used:

- Azure Data Factory (ADF)

- Azure Databricks

- Azure Data Lake Storage (ADLS)

- Azure Synapse Analytics

- Parquet and Delta formats for data storage

A diagram of a process

AI-generated content may be incorrect.

# Detailed Implementation

## Step 1: Data Ingestion

Data is copied from the backend team's storage account to a designated raw (Bronze) container in the data lake using Azure Data Factory. Five datasets are involved: accounts.csv, customers.csv, loan\_payments.csv, loans.csv, transactions.csv.

## Step 2: Delta Processing

Data from the raw container is processed and cleaned using a Databricks notebook. Transformations include filling missing values, casting data types, and writing to a silver container for each dataset.

## Step 3: ETL Processing

Data is further processed using another Databricks notebook. This involves reading from the silver containers, applying business logic and transformations, and then writing to a gold container. Complex transformations such as aggregations to calculate the total balance across all accounts for each customer are implemented using SQL queries and PySpark dataframes.

## Step 4: Azure Synapse Analytics

The transformed data in the gold container is then loaded into Azure Synapse Analytics. A dedicated SQL pool is used to create target tables and configure data pipeline for upsert functionality. Data is upserted into Synapse, allowing for updates or insertions as necessary.

# Deliverables:

- Well-structured documentation with detailed explanations and screenshots.

- Relevant code snippets from the Databricks notebooks and ADF pipeline configurations.