**📁 AZ Adventure Works Project Documentation**

**Project Type:** End-to-End Azure Data Engineering Project  
**Cloud Platform:** Microsoft Azure

**📌 Project Overview**

This project demonstrates the implementation of a full-scale data architecture pipeline using Microsoft Azure. It involves ingesting, transforming, and analyzing data across **Bronze**, **Silver**, and **Gold** layers using services such as **Azure Data Factory**, **Azure Databricks**, and **Azure Synapse Analytics**.

**1. ☁️ Azure Data Architecture Overview**

**🔷 Introduction**

Azure Data Architecture focuses on managing data storage, processing, and analytics in the cloud.

**📌 Example:**

Creating a cloud-based **data warehouse** that handles both **real-time** transactions and **batch** processing.

**2. 🔐 Azure Free Account Setup**

**📋 Steps:**

* Visit [Azure Free Trial](https://azure.microsoft.com/en-us/free/)
* Sign up using your Microsoft account
* Get $200 free credits to explore services like SQL Database, Blob Storage, etc.

**3. 🔎 Data Understanding via APIs**

Understand and consume external APIs into your pipeline.

**📌 Example:**

* Base URL: https://github.com/
* Relative Path: girald02/az\_project\_01/tree/main
* Fetching external CSV data via API using Azure Data Factory.

**4. 🟫 Bronze Layer: Raw Data Ingestion**

**🔧 Tools: Azure Data Factory (ADF)**

* **Linked Services**
* **Datasets**
* **Pipelines**
  + Static pipelines
  + Dynamic pipelines with parameters

**📥 Example:**

Ingesting raw CSV files into **Azure Data Lake Storage (ADLS)** – Bronze container.

**5. 🔨 Azure Resources Setup**

* Create Azure Blob Storage
* Set up SQL Database
* Provision a Virtual Machine (optional for testing or hosting services)

**6. ⚙️ Azure Data Factory Pipeline (ADF)**

**Static Pipeline**

* Fixed source and destination

**Dynamic Pipeline**

* Parameterized pipeline for iterating over datasets or dates

**7. 🪙 Silver Layer: Data Transformation using Databricks**

**🔐 Accessing Data Lake from Databricks**

1. **Register App in Microsoft Entra** (formerly Azure AD)
2. Get the following:
   * app\_id, object\_id, secret
3. Assign **Storage Blob Data Contributor** role to the App
4. Use OAuth configs in Databricks to authenticate

**🔁 Spark Configuration**

spark.conf.set("fs.azure.account.auth.type.<storage-account>.dfs.core.windows.net", "OAuth")

spark.conf.set("fs.azure.account.oauth.provider.type.<storage-account>.dfs.core.windows.net", "org.apache.hadoop.fs.azurebfs.oauth2.ClientCredsTokenProvider")

spark.conf.set("fs.azure.account.oauth2.client.id.<storage-account>.dfs.core.windows.net", "<application-id>")

spark.conf.set("fs.azure.account.oauth2.client.secret.<storage-account>.dfs.core.windows.net", service\_credential)

spark.conf.set("fs.azure.account.oauth2.client.endpoint.<storage-account>.dfs.core.windows.net", "https://login.microsoftonline.com/<directory-id>/oauth2/token")

**8. 🧪 Data Processing with PySpark**

**🔹 Read Data from Bronze**

df = spark.read.format("csv").option("header", "true").option("inferSchema", "true").load("<bronze\_path>")

**📘 Examples:**

df\_cal = spark.read.format("csv")\

.option("header", "true").option("inferSchema", "true")\

.load("abfss://bronze@storagedatalakeproject1.dfs.core.windows.net/AdventureWorks\_Calendar")

**✏️ PySpark Column Operations**

from pyspark.sql.functions import \*

df\_cal = df\_cal.withColumn('Month', month(col('Date')))\

.withColumn('Year', year(col('Date')))\

.withColumn('Day', day(col('Date')))

**9. 📦 Write to Silver Layer (Parquet Format)**

df\_cal.write.format("parquet")\

.mode("append")\

.option("path", "<silver\_path>")\

.save()

**💡 Add FullName Column:**

df\_cus = df\_cus.withColumn('FullName', concat\_ws(' ', col('Prefix'), col('FirstName'), col('LastName')))

**10. 🔁 Data Transformation Requirements**

**✅ Requirement 1: Parse Date Format**

df\_sales\_year = df\_sales\_year.withColumn('StockDate', to\_timestamp(col('StockDate')))

**✅ Requirement 2: Replace Characters**

df\_sales\_year = df\_sales\_year.withColumn('OrderNumber', regexp\_replace('OrderNumber', 'S', 'T'))

**✅ Requirement 3: Multiply Columns**

df\_sales\_year = df\_sales\_year.withColumn('Total', col('OrderLineItem') \* col('OrderQuantity'))

**11. 📊 Sales Aggregation Example**

df\_aggregated = df\_sales\_year.groupBy("OrderDate").agg(count('OrderNumber').alias('total\_order'))

display(df\_aggregated)

**12. 🧠 Azure Synapse Analytics**

Unified platform combining:

* Azure Data Factory
* Apache Spark
* SQL Data Warehousing

**Identity Access:**

Create user access for querying via Synapse Studio using managed identity.

**13. 📂 SQL Pool (Serverless + Dedicated)**

Example to read from Parquet in Silver layer:

SELECT \*

FROM OPENROWSET(

BULK 'https://storagedatalakeproject1.blob.core.windows.net/silver/AdventureWorks\_Calendar',

FORMAT = 'PARQUET'

) AS [result];

**14. 👑 Gold Layer: Metadata & Reporting**

**✅ Create Schema:**

CREATE SCHEMA gold;

**✅ Create View:**

CREATE VIEW gold.calendar AS

SELECT \* FROM OPENROWSET(

BULK 'https://storagedatalakeproject1.blob.core.windows.net/silver/AdventureWorks\_Calendar',

FORMAT = 'PARQUET'

) AS Quer1;

**15. 🔐 External Tables (CETAS)**

**Required:**

* Master Key
* Scoped Credential
* External Data Source
* External File Format

**Sample:**

CREATE MASTER KEY ENCRYPTION BY PASSWORD='<StrongPassword>';

CREATE DATABASE SCOPED CREDENTIAL AppCred WITH IDENTITY = 'Mary5', SECRET = '<StrongPassword>';

CREATE DATABASE SCOPED CREDENTIAL credGirald

WITH IDENTITY = 'Managed Identity';

CREATE EXTERNAL DATA SOURCE sourceGold

WITH (LOCATION = 'https://storagedatalakeproject1.blob.core.windows.net/gold', CREDENTIAL = credGirald);

**16. 📉 Power BI Integration**

* Use **Serverless SQL Endpoint** from Synapse:
* azgiraldproject-synapse-ondemand.sql.azuresynapse.net
* Load data directly into Power BI
* Build visualizations: Pie charts, Bar charts, Trend lines, etc.

**17. 💼 Azure Data Engineer Interview Preparation**

* ADF concepts: Pipelines, Triggers, Integration Runtimes
* Databricks usage, Spark transformations
* ADLS layers: Bronze, Silver, Gold
* Synapse Serverless vs Dedicated
* Power BI Integration
* Managing secrets with Key Vault