Azure Meal Project | ETL Pipeline

Overview

The Meal Project is a cloud-based ETL solution hosted on Azure. It efficiently manages raw data stored in a blob container, processes the data using Function Apps, and ingests the results into a SQL Server database. This processed data is then connected to Power BI, where a report is generated to provide valuable insights. Figure 1 illustrates the architecture diagram of the pipeline.

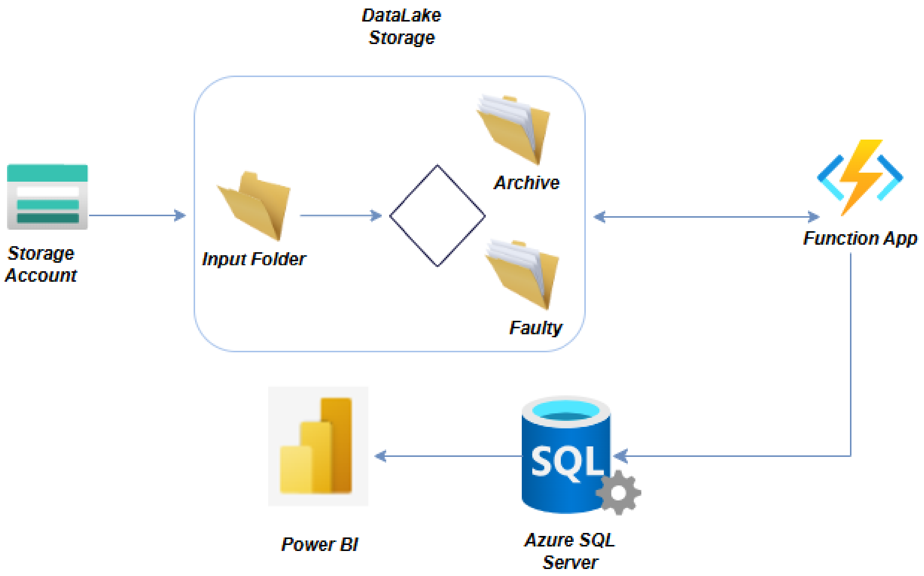


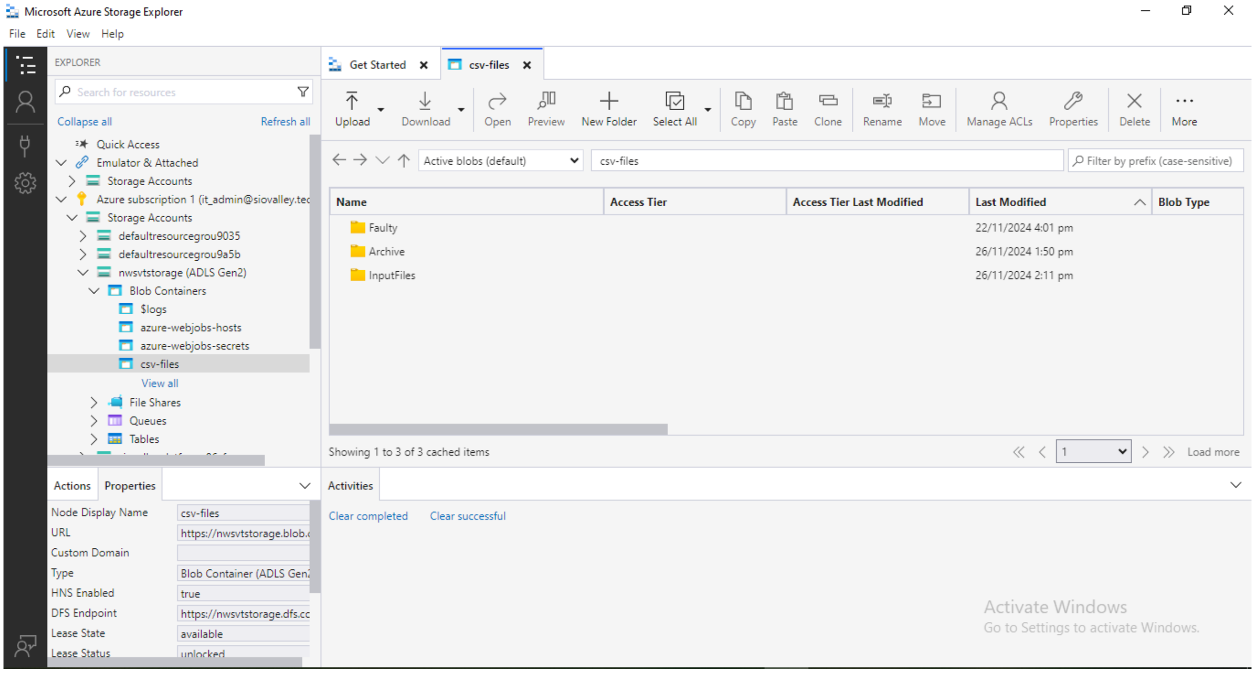
Figure 1: Architecture Design of Meal Project ETL

Blob Storage

The raw data from the user or client is stored in Azure Blob Storage, an object storage solution that accommodates various file types, such as .csv and .txt files, organized into separate folders. Given that our input files are primarily .csv, this service provides an optimal solution for our needs. Additionally, Azure Blob Storage is easily accessible through the Microsoft Azure Storage Explorer, which offers a user-friendly interface for managing data. Users can utilize this tool to upload and review files within their designated folders, as illustrated in Figure 2.

In our project, we have created two directories into the blob container ‘csv-files’ named

1. Archived: contains files that are processed
2. InputFiles: contains file that will be processed and be sent to achieve
3. FaultyFiles: contains unprocessed files because they have ambiguity in data.



Name	Access Tier	Access Tier Last Modified	Last Modified	Blob Type
Faulty			22/11/2024 4:01 pm	
Archive			26/11/2024 1:50 pm	
InputFiles			26/11/2024 2:11 pm	

Showing 1 to 3 of 3 cached items

Node Display Name	csv-files
URL	https://nwsvtstorage.blob.
Custom Domain	
Type	Blob Container (ADLS Gen2
HNS Enabled	true
DFS Endpoint	https://nwsvtstorage.dfs.cc
Lease State	available
Lease Status	unlocked

Figure 2: Microsoft Storage Explorer view of blob-container ‘csv-files’

FunctionApp

The Azure Function App is a serverless service that executes in response to specific events. Our Function App is configured with a timer trigger that activates at midnight. The entire setup has been developed using Visual Studio Code, where we manage the repository and handle deployments. You can find the repository at the link below.

Repo Link: https://drive.google.com/file/d/1W3EpT-qBYnMJsFcxvrv7viwnlNKLylyO/view?usp=sharing

The code runs in the following sequence

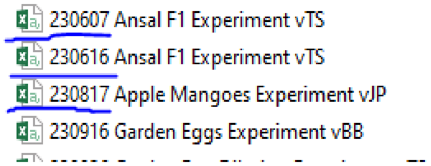
1. Look for the input file in the ‘InputFiles’ directory in blob container
2. Process the data based on ‘WLD’ and the defined calculations
3. Ingest the data into Azure SQL Server Database by making connection
4. Next step is Data Ingestion, where if any files have missing columns or other issues, they will not be processed and will be moved to the 'Faulty' folder. Files without any ambiguities will be successfully processed and moved from the 'InputFiles' directory to the 'Archived' directory.

Azure SQL Server Database

Our final processed data is stored in Azure SQL Server Database. Here we have declared two tables

TABLE 1: File_Metadata

- dbo.File_Metadata: This table contains Files data such filename etc.
- Here group_id used to group data entries by their corresponding sample and File_Code is basically number which is mention in title of each file as see in below image

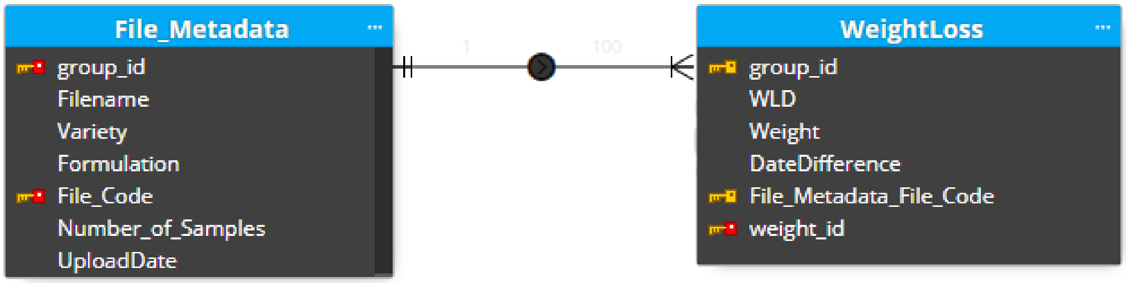


230607 Ansal F1 Experiment vTS
230616 Ansal F1 Experiment vTS
230817 Apple Mangoes Experiment vJP
230916 Garden Eggs Experiment vBB

TABLE 2: WeightLoss

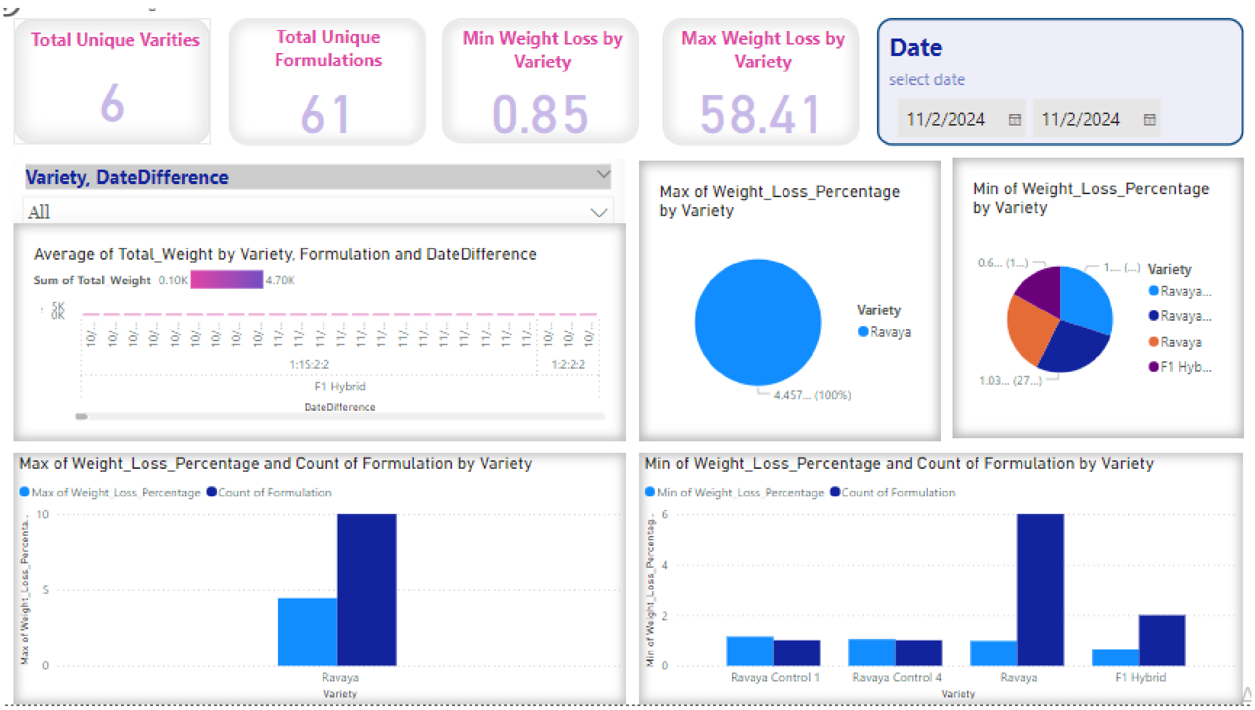
- dbo.WeightLoss: contains data weight for that day, weight_id which identify unique record.

Database Schema Diagram:



Power BI Report

Below this Power BI Dashboard provides a comprehensive view of the weight loss data dashboard including the following visualizations and data insights: Maximum and Minimum Weight Loss, Varieties and Total Formulations, daily variations.



Total Unique Varities

6

Total Unique Formulations

61

Min Weight Loss by Variety

0.85

Max Weight Loss by Variety

58.41

Date

select date

11/2/2024 11/2/2024

Variety, DateDifference

All

Average of Total_Weight by Variety, Formulation and DateDifference
Sum of Total_Weight 0.10K 4.70K
5K
0K
1:15:2:2
1:2:2:2
F1 Hybrid
DateDifference
Max of Weight_Loss_Percentage by Variety
Variety
Ravaya
4,457... (100%)
Min of Weight_Loss_Percentage by Variety
0.6... (1...)
1... (...)
1.03... (27...)
Variety
Ravaya...
Ravaya...
Ravaya
F1 Hyb...
Max of Weight_Loss_Percentage and Count of Formulation by Variety
Max of Weight_Loss_Percentage
Count of Formulation
Max of Weight_Loss_Percenta...
10
5
0
Ravaya
Variety
Min of Weight_Loss_Percentage and Count of Formulation by Variety
Min of Weight_Loss_Percentage
Count of Formulation
Min of Weight_Loss_Percentag...
6
4
2
0
Ravaya Control 1
Ravaya Control 4
Ravaya
F1 Hybrid
Variety