

Project: Gold_Finance ETL

Introduction

Gold_Finance ETL: a robust data engineering project that automates financial time series data processing, transforms raw financial data into structured insights for business stakeholders and analysts.

Table of Contents

- 1. Introduction
- 2. Business Problem and Solution
- 3. Alpha Vantage Api
- 4. ERD Schema
- 5. Azure Solution Architecture
- 6. Azure Data Pipeline Flow
- 7. Azure Project Tech Stack
 - Azure Storage Setup
 - Azure Databricks Setup
 - Azure SQL Server and Database Setup
 - Setup Azure Data Factory
 - Azure Data Factory Pipeline
 - Error Handling in Azure
 - Error Notification
 - Azure DevOPs (CI/CD)
- 8. Onprem Solution Architecture
- 9. Onprem Data Pipeline Flow
- 10. Onprem Project Tech Stack
 - Pyspark on Vscode
 - Postgres
 - Airflow
- 11. Linux Commands

Business Problem Solution

S/N	Business Problem
1	High volume of rapidly changing financial data
2	Unstructured formats requiring significant cleaning
3	Time-intensive manual workflows prone to errors

S/N	Business Solution
1	Fully automated workflows for daily data integration and analysis
2	Aggregated metrics for informed decision-making

Alpha Vantage Api

++ TIME_SERIES_DAILY:

This API returns raw (as-traded) daily time series (date, daily open, daily high, daily low, daily close, daily volume) of the global equity specified, covering 20+ years of historical data. The OHLCV data is sometimes called "candles" in finance literature. If you are also interested in split/dividend-adjusted data, please use the Daily Adjusted API, which covers adjusted close values and historical split and dividend events.

Reference: <https://www.alphavantage.co/documentation/>

API Parameters

Required: **function**

The time series of your choice. In this case, **function=TIME_SERIES_DAILY**

Required: **symbol**

The name of the equity of your choice. For example: **symbol=IBM**

Optional: **outputsize**

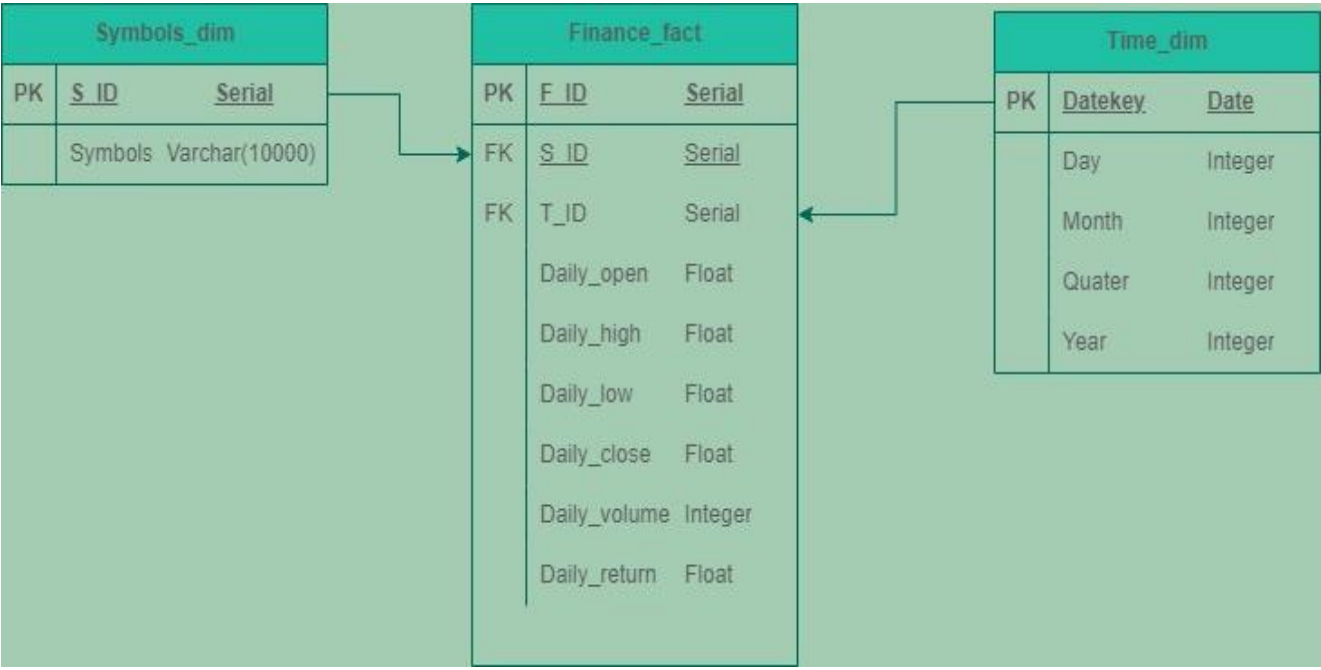
By default, **outputsize=compact**. Strings **compact** and **full** are accepted with the following specifications: **compact** returns only the latest 100 data points; **full** returns the full-length time series of 20+ years of historical data. The "compact" option is recommended if you would like to reduce the data size of each API call.

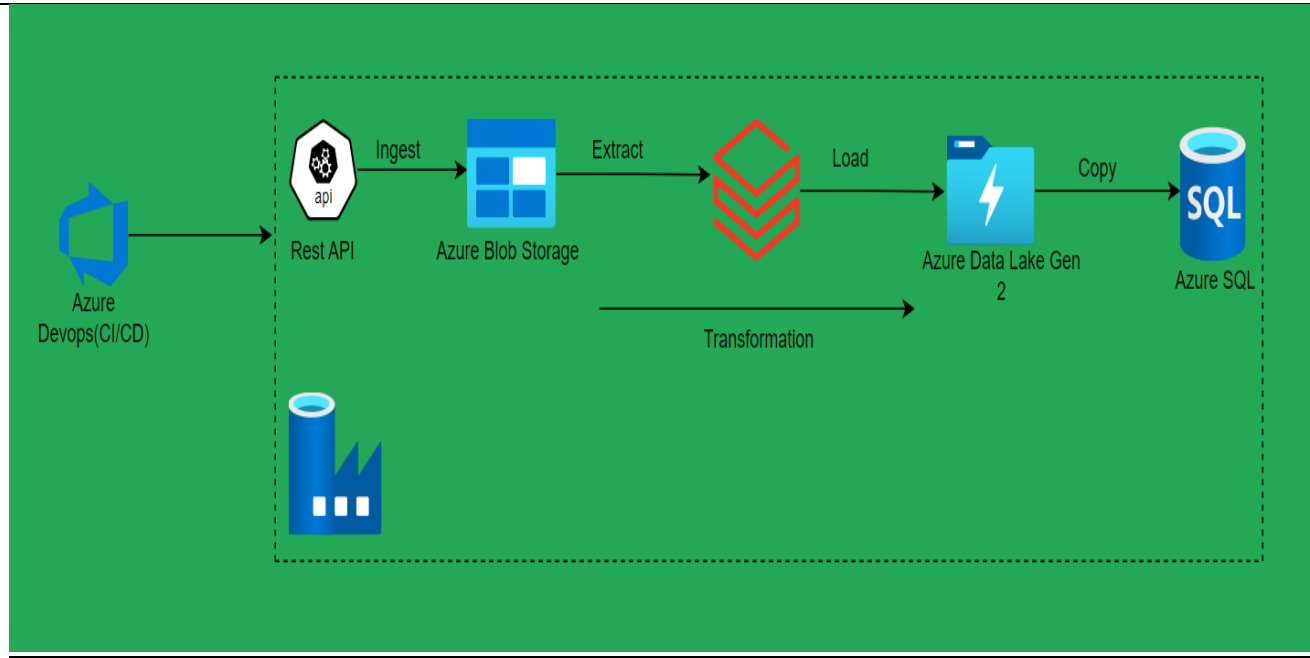
Optional: **datatype**

By default, **datatype=json**. Strings **json** and **csv** are accepted with the following specifications: **json** returns the daily time series in JSON format; **csv** returns the time series as a CSV (comma separated value) file.

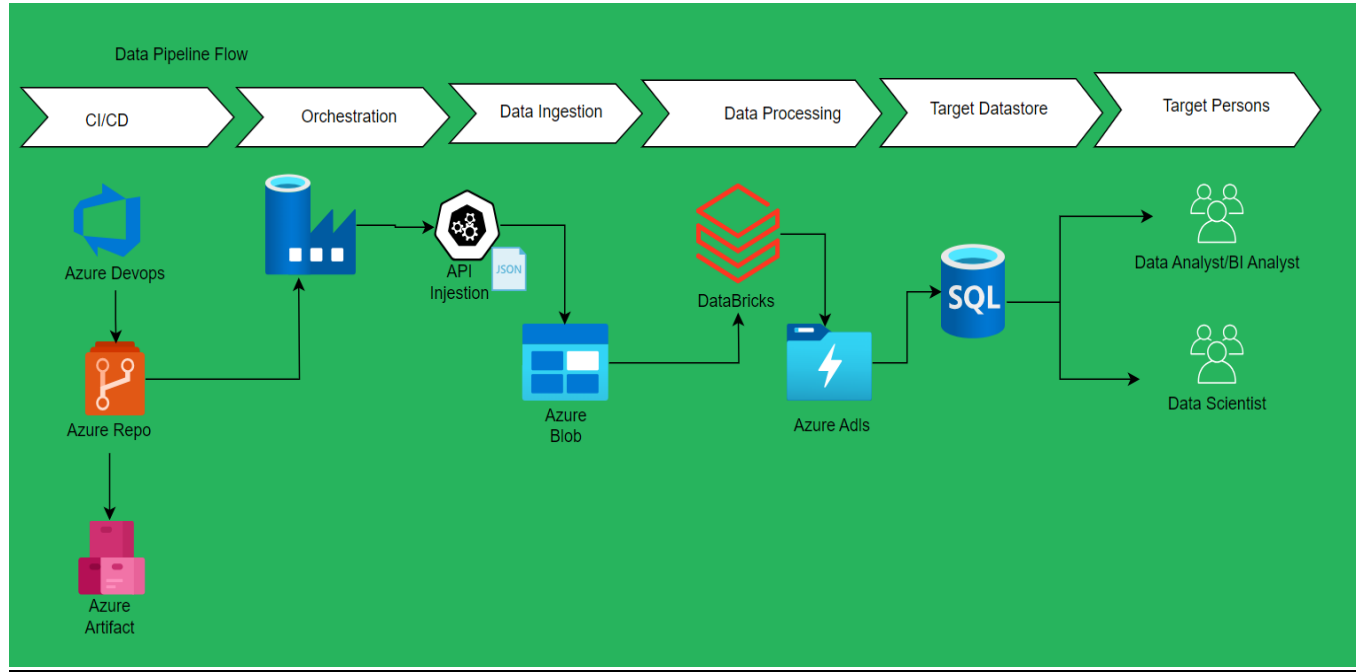
Required: **apikey**

ERD Schema



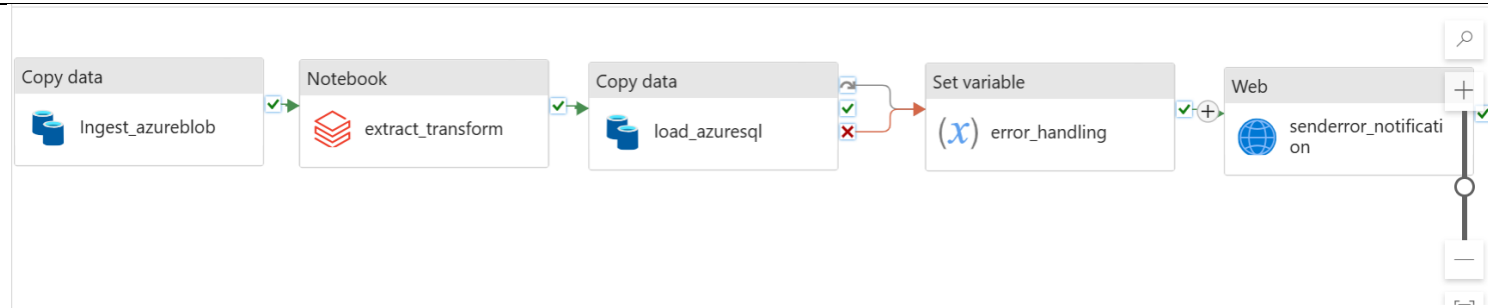


Data Pipeline Flow



Azure Project Tech Stack

S/N	Tech Stack	Details
1	Azure Storage Account	Data lake, Datawarehouse
2	Databricks	Transformation
3	Azure Data Factory	Orchestration
4	Azure Logic Apps	Error Handling and Notification
5	Azure DevOps	CI/CD



Azure DevOps AgroFarms / datamigration / Pipelines / Releases / GoldFinanceRelease / Release-1

Search

datamigration +

- Overview
- Boards
- Repos
- Pipelines
- Pipelines
- Environments
- Releases
- Library
- Task groups
- Deployment groups
- Project settings

GoldFinanceRelease > Release-1

Pipeline Variables History + Deploy Cancel Refresh Edit ...

Release

Continuous deployment

for Chinwe Uzoka

12/8/2024, 3:04 PM

Artifacts

_GoldFinance_etl

20241208.1

master

Stages

Dev

Succeeded

on 12/8/2024, 3:05 PM

AgroFarms / datamigration / Pipelines / Releases / GoldFinanceRelease

Search

All pipelines > GoldFinanceRelease

Save Create release View releases ...

Pipeline Tasks Variables Retention Options History

Artifacts + Add

_GoldFinance_etl

Schedule not set

Stages + Add

Dev 1 job, 1 task

Test 1 job, 1 task

Prod 1 job, 1 task

+++ Azure Storage Setup


Create Storage Account and create Containers for both blob storage and adls gen 2


Create a Storage Account:


- Click on create a resource


Microsoft Azure Search resources, services, and docs (G+/) Copilot


Azure services


 Create a resource


 Billing subscriptions


 Azure DevOps organizations

 Cost Management

 Storage accounts

 Logic apps

 Azure Database for MySQL...

 SQL ser

Resources

Recent Favorite

Name	Type
------	------

- Type storage account in the search bar and select storage account, fill the below as desired with respect to business use case.

Microsoft Azure Search resources, services, and docs (G+/) Copilot

Home > Create a resource > Marketplace > Storage account >

Create a storage account

Basics Advanced Networking Data protection Encryption Tags Review + create

Azure Storage is a Microsoft-managed service providing cloud storage that is highly available, secure, durable, scalable, and redundant. Azure Storage includes Azure Blobs (objects), Azure Data Lake Storage Gen2, Azure Files, Azure Queues, and Azure Tables. The cost of your storage account depends on the usage and the options you choose below. [Learn more about Azure storage accounts](#)

Project details

Select the subscription in which to create the new storage account. Choose a new or existing resource group to organize and manage your storage account together with other resources.

Subscription * Azure subscription 1

Resource group * dev

[Create new](#)

Instance details

Previous Next Review + create

Create Storage account for both blob storage and adls gen 2

- While creating a blob storage account make sure to uncheck the “enable hierarchical namespace”

Microsoft Azure

Search resources, services, and docs (G+)

Copilot

[Home](#) > [Create a resource](#) > [Marketplace](#) > [Storage account](#) >

Create a storage account

Basics

Advanced

Networking

Data protection

Encryption

Tags

Review + create

Security

Configure security settings that impact your storage account.

Require secure transfer for REST API operations ⓘ

☒

Allow enabling anonymous access on individual containers ⓘ

☐

Enable storage account key access ⓘ

☒

Default to Microsoft Entra authorization in the Azure portal ⓘ

☐

Minimum TLS version ⓘ

Version 1.2

▼

Permitted scope for copy operations

From any storage account

▼

Microsoft Azure

Search resources, services, and docs (G+)

Copilot

[Home](#) > [Create a resource](#) > [Marketplace](#) > [Storage account](#) >

Create a storage account

Enable storage account key access ⓘ

☒

Default to Microsoft Entra authorization in the Azure portal ⓘ

☐

Minimum TLS version ⓘ

Version 1.2

▼

Permitted scope for copy operations (preview) ⓘ

From any storage account

▼

Hierarchical Namespace

Hierarchical namespace, complemented by Data Lake Storage Gen2 endpoint, enables file and directory semantics, accelerates big data analytics workloads, and enables access control lists (ACLs) [Learn more](#)

Enable hierarchical namespace ⓘ

☐

Access protocols

Blob and Data Lake Gen2 endpoints are provisioned by default [Learn more](#)

☐

Previous

Next

Review + create

- While creating an adls Gen 2 storage account make sure to check the “enable hierarchical namespace”

Microsoft Azure

Search resources, services, and docs (G+/)

Copilot

uzokar
RITZY (UZOKARGMAIL.ONMICR...

Home > Create a resource > Marketplace > Storage account >

Create a storage account

Enable storage account key access

☒

Default to Microsoft Entra authorization in the Azure portal

☐

Minimum TLS version

Version 1.2

Permitted scope for copy operations (preview)

From any storage account

Hierarchical Namespace

Hierarchical namespace, complemented by Data Lake Storage Gen2 endpoint, enables file and directory semantics, accelerates big data analytics workloads, and enables access control lists (ACLs) [Learn more](#)

Enable hierarchical namespace

☒

Access protocols

Blob and Data Lake Gen2 endpoints are provisioned by default [Learn more](#)

☐

Previous

Next

Review + create

- Create a container for both blob storage and adls gen 2

Microsoft Azure

Search resources, services, and docs (G+/)

Copilot

uzokar@gmail.com
RITZY (UZOKARGMAIL.ONMICR...

Home > goldfinanceetl

goldfinanceetl | Containers

Storage account

Search

+ Container

Change access level

Restore containers

Refresh

Delete

Give feedback

Diagnose and solve problems

Access Control (IAM)

Data migration

Events

Storage browser

Partner solutions

Data storage

Containers

File shares

Queues

Tables

Security + networking

Data management

Search containers by prefix

Show deleted containers

Name	Last modified	Anonymous access level	Lease state	
<input type="checkbox"/> \$logs	11/24/2024, 9:40:26 AM	Private	Available	...
<input type="checkbox"/> finance-transformed-data	11/24/2024, 9:45:10 AM	Private	Available	...

- An adls gen 2 Container created

Microsoft Azure Search resources, services, and docs (G+ /) Copilot

Home > goldfinanceetl | Containers >

finance-transformed-data ...
Container

Search Upload Add Directory Refresh Rename Delete Change tier Acquire lease Break lease

Overview
Diagnose and solve problems
Access Control (IAM)
Settings

Authentication method: Access key (Switch to Microsoft Entra user account)
Location: finance-transformed-data

Search blobs by prefix (case-sensitive) Show

Name	Modified	Access tier	Archive status
<input type="checkbox"/> _azuretmpfolder\$	12/1/2024, 11:16:44 ...		
<input type="checkbox"/> finance-alpha-vantage-transformed-delta	12/4/2024, 7:15:00 AM		

- A blob storage container created

Microsoft Azure Search resources, services, and docs (G+ /) Copilot

Home > financealphavantageblob | Containers >

finance-alpha-vantage-raw ...
Container

Search Upload Change access level Refresh Delete Change tier Acquire lease Break lease

Overview
Diagnose and solve problems
Access Control (IAM)
Settings

Authentication method: Access key (Switch to Microsoft Entra user account)
Location: finance-alpha-vantage-raw

Search blobs by prefix (case-sensitive) Show c

+ Add filter

Name	Modified	Access tier	Archive status
<input type="checkbox"/> data_09ac2a74-6f26-40d4-a1ae-cc05221e4437_dab...	11/25/2024, 5:42:06 ...	Hot (Inferred)	
<input type="checkbox"/> data_21fd3fa3-4665-4c77-b9b6-fef35bc3a542_7620...	12/2/2024, 12:49:14 ...	Hot (Inferred)	
<input type="checkbox"/> data_24f5878e-251a-46f5-86fe-91dbeb366381_0de...	12/2/2024, 11:43:17 ...	Hot (Inferred)	
<input type="checkbox"/> data_37be23b2-6b6f-4a5b-92b2-1444b66dd03b_b6...	12/1/2024, 8:56:07 AM	Hot (Inferred)	

+++ Azure Databricks Setup

Create Azure Databrick's Workspace

- Click on create a resource search for Azure Databricks and select Azure Databricks.

Microsoft Azure

Search resources, services, and docs (G+/)

Copilot

Home > Create a resource > Marketplace >

Azure Databricks

Microsoft

Azure Databricks

Microsoft | Azure Service

★ 4.5 (396 ratings)

Plan

Azure Databricks

Create

Overview

Plans

Usage Information + Support

Ratings + Reviews

Fast, easy, and collaborative Apache Spark-based analytics platform

Accelerate innovation by enabling data science with a high-performance analytics platform that's optimized for Azure.

Drive innovation and increase productivity

Bring teams together in an interactive workspace. From data gathering to model creation, use Databricks Notebooks to unify the process and instantly deploy to production. Launch your new Spark environment with a single click. Integrate effortlessly with a wide variety of data stores and services such as [Azure SQL Data Warehouse](#), [Azure Cosmos DB](#), [Azure Data Lake Store](#), [Azure Blob storage](#), and [Azure Event Hub](#). Add advanced artificial intelligence (AI) capabilities instantly and share your insights through rich

Microsoft Azure

Search resources, services, and docs (G+/)

Copilot

Home > Create a resource > Marketplace > Azure Databricks >

Create an Azure Databricks workspace

Basics

Networking

Encryption

Security & compliance

Tags

Review + create

Project Details

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.

Subscription * ⓘ

Azure subscription 1

Resource group * ⓘ

Create new

Instance Details

Workspace name *

Enter name for Databricks workspace

Region *

East US

Review + create

< Previous

Next : Networking >

Microsoft Azure

Search resources, services, and docs (G+)

Copilot

[Home](#) >

Azure Databricks

Ritzzy (uzokargmail.onmicrosoft.com)

+ Create

Manage view

Refresh

Export to CSV

Open query

Assign tags

Filter for any field...

Subscription equals all

Resource group equals all

Location equals all

Add filter

Showing 1 to 1 of 1 records.

No grouping


<input type="checkbox"/> Name ↑↓	Type ↑↓	Resource group ↑↓	Location ↑↓
<input type="checkbox"/> Gold_finance_etl	Azure Databricks Service	devprjt	UK West

Microsoft Azure

Search resources, services, and docs (G+)

Copilot

[Home](#) > [Azure Databricks](#) >

 **Gold_finance_etl**

Azure Databricks Service

Search

Delete

Overview

Activity log

Access control (IAM)

Tags

Diagnose and solve problems

Settings

Automation

Help


[Azure subscription 1](#)

Subscription ID

0411e30c-19bd-44ec-9143-5e16fad48005

Tags [\(edit\)](#)

[Add tags](#)



Launch Workspace

Upgrade to Premium

+++ Azure SQL Server Database Setup

Create Azure SQL Server:

- Click on create a resource, search for azure sql server and select azure sql.

Microsoft Azure


Search resources, services, and docs (G+/)

Copilot

[Home](#) > [Create a resource](#) > [Marketplace](#) >

Azure SQL

Microsoft



Azure SQL

Microsoft | Azure Service

★ 4.7 (6 ratings)

Plan

Azure SQL

Create

Overview

Plans

Usage Information + Support

Ratings + Reviews

Azure SQL allows you to create and manage your SQL Server resources from a single view, ranging from fully managed PaaS databases to IaaS virtual machines with direct database engine access. All deployment options enable you to bring your on-premises licenses to Azure using Azure Hybrid Benefit.

Databases
Single databases are optimized for modern application development of new cloud-born applications. Databases provide a fully managed SQL experience with extensive a easy to use manageability features.
Includes: single databases, elastic pools, and database servers

Managed instances

Microsoft Azure

Search resources, services, and docs (G+/)

Copilot

[Home](#) > [Create a resource](#) > [Marketplace](#) > [Azure SQL](#) > [Select SQL deployment option](#) > [Create SQL Database](#) >

Create SQL Database Server

Microsoft

Server details

Enter required settings for this server, including providing a name and location. This server will be created in the same subscription and resource group as your database.

Server name *


Enter server name

.database.windows.net

Location *

(US) East US

Authentication

 Azure Active Directory (Azure AD) is now Microsoft Entra ID. [Learn more](#)

OK

Create Azure SQL Database:

- Follow the prompt to create an sql database.

[Home](#) > [Create a resource](#) > [Marketplace](#) > [Azure SQL](#) > [Select SQL deployment option](#) >

Create SQL Database ...

Microsoft

Enter required settings for this database, including picking a logical server and configuring the compute and storage resources

Database name *

Enter database name

✖ Your database name can't end with '.' or '' , can't contain '<,>,*%,&,:\\,/?' or control characters

✖ Database name should not be empty

Server * ⓘ

gold-finance-ss (UK West)

[Create new](#)

Want to use SQL elastic pool? ⓘ

☐ Yes ☒ No

Workload environment

☒ Development
☐ Production

i Default settings provided for Development workloads. Configurations can be modified as needed.

[Review + create](#)

[Next : Networking >](#)

+++ Setup Azure Data Factory

Create Azure Data Factory

- Click the select a resource, search with azure data factory and select azure data factory, create an azure data factory by following the prompts, open the azure data studio to launch the adf.

Microsoft Azure


Search resources, services, and docs (G+)

Copilot

[Home](#) > [Create a resource](#) > [Marketplace](#) >

Data Factory

Microsoft



Data Factory

♥ Add to Favorites

Microsoft | Azure Service

★ 3.6 (605 ratings)

Plan

Data Factory

Create

Overview

Plans

Usage Information + Support

Ratings + Reviews

Integrate data silos with Azure Data Factory, a service built for all data integration needs and skill levels. Easily construct ETL and ELT processes code-free within the intuitive visual environment, or write your own code. Visually integrate data sources using more than 90+ natively built and maintenance-free connectors at no added cost. Focus on your data - the serverless integration service does the rest.

- No code or maintenance required to build hybrid ETL and ELT pipelines within the Data Factory visual environment
- Cost-efficient and fully managed serverless cloud data integration tool that scales on demand
- Azure security measures to connect to on-premises, cloud-based, and software-as-a-service apps with peace of mind
- SSIS integration runtime to easily rehost on-premises SSIS packages in the cloud using familiar SSIS tools

Microsoft Azure Search resources, services, and docs (G+/) Copilot uzokar@gmail.com RITZY (UZOKARGMAILONMICR...)

Home >

pcuretaildf Data factory (V2)

Search Delete

Overview

- Activity log
- Access control (IAM)
- Tags
- Diagnose and solve problems
- Settings
- Getting started
- Monitoring
- Automation
- Help


Succeeded

Location
East US

Subscription (move)
[Azure subscription 1](#)

Subscription ID
0411e30c-19bd-44ec-9143-5e16fad48005

[Quick start](#)



Azure Data Factory Studio


[Launch studio](#)

Integrated Runtime (IR):


- AutoResolve IR: To be used for all Azure services; the same IR can be used for all Azure services in a project and or multiple projects.
- Self-Hosted IR: To be used for onprem MySQL Server migration to Azure Cloud

Integration runtime setup


Integration Runtime is the native compute used to execute or dispatch activities. Choose what integration runtime to create based on required capabilities. [Learn more](#)



Azure, Self-Hosted
Perform data flows, data movement and dispatch activities to external compute.



Azure-SSIS
Lift-and-shift existing SSIS packages to execute in Azure.



Airflow (Preview)
Use this for running your existing DAGs

[Continue](#) [Cancel](#)

Linked Service: defines the connection information to a data store or compute.

New linked service

Data store

Compute



Azure Blob Storage



Azure Cosmos DB for
MongoDB



Azure Cosmos DB for
NoSQL



Azure Data Explorer
(Kusto)



Azure Data Lake Storage
Gen2



Azure Database for
MariaDB (Legacy)

Continue

Cancel

Dataset: to specify the location and structure of your data within a data store.

New dataset

In pipeline activities and data flows, reference a dataset to specify the location and structure of your data within a data store. [Learn more](#)

Select a data store

Search

<AllAzureDatabaseFileGeneric protocol>


Google BigQuery

Google Cloud Storage


Greenplum

APACHE
HBASE


HBase (Legacy)




HDFS



HTTP











Continue

Cancel

Select format

Choose the format type of your data

 Avro	 Binary	 DelimitedText
 Excel	 JSON	 ORC

Continue

Back

Cancel

Create Linked Service for Azure Blob Storage:
Create Dataset for Azure blob Storage:

Create Linked Service for ADLS Gen 2:
Create Dataset for ADLS Gen 2:

Create Linked Service for Azure Databricks:
Create Dataset for Azure Databricks:

Create Linked Service for Azure SQL Database:
Create Dataset for Azure SQL Database:

Linked services









Linked service defines the connection information to a data store or compute. [Learn more](#)

+ New








Filter by name

Annotations : Any

Showing 1 - 8 of 8 items

Name ↑↓	Type ↑↓	Related ↑↓	Annotations ↑↓
 AzureBlobStorage2	Azure Blob Storage	1	
 AzureDatabricks_Is	Azure Databricks	0	
 AzureDatabricks_main_Is	Azure Databricks	1	
 AzureDataLakeStorage1	Azure Data Lake Storage Gen2	2	
 AzureDataLakeStorage_Is	Azure Data Lake Storage Gen2	1	
 fin_alphavantage_raw_Is	Azure Blob Storage	1	
 fin_alphavantage_RestServiceap...	REST	1	
 SqlServer_Is	SQL server	1	

▲ Datasets 7

-  azureadls_ds
-  azureadls_ds2
-  azureadlsParquet_ds
-  azureblob_ds
-  fin_alpha_vantage_Jsonds
-  fin_alphavantage_RestResourceapi_ds
-  gold_finance_SqlServerTable_ds

+++++

Create Azure Data Factory Pipeline:

Each Time

- 1) Create new Linked Service for each Azure service
- 2) Create new Dataset for each Azure service data format.

Copy Activity

The Copy Activity is used to move data from one source to a destination in the cloud. Every Copy Activity has a source dataset and a sink dataset

Ingestion Layer into Bronze-Layer (Azure Blob Storage)

Step 1: Drag the copy activity to the pipeline workspace and configure the parameters

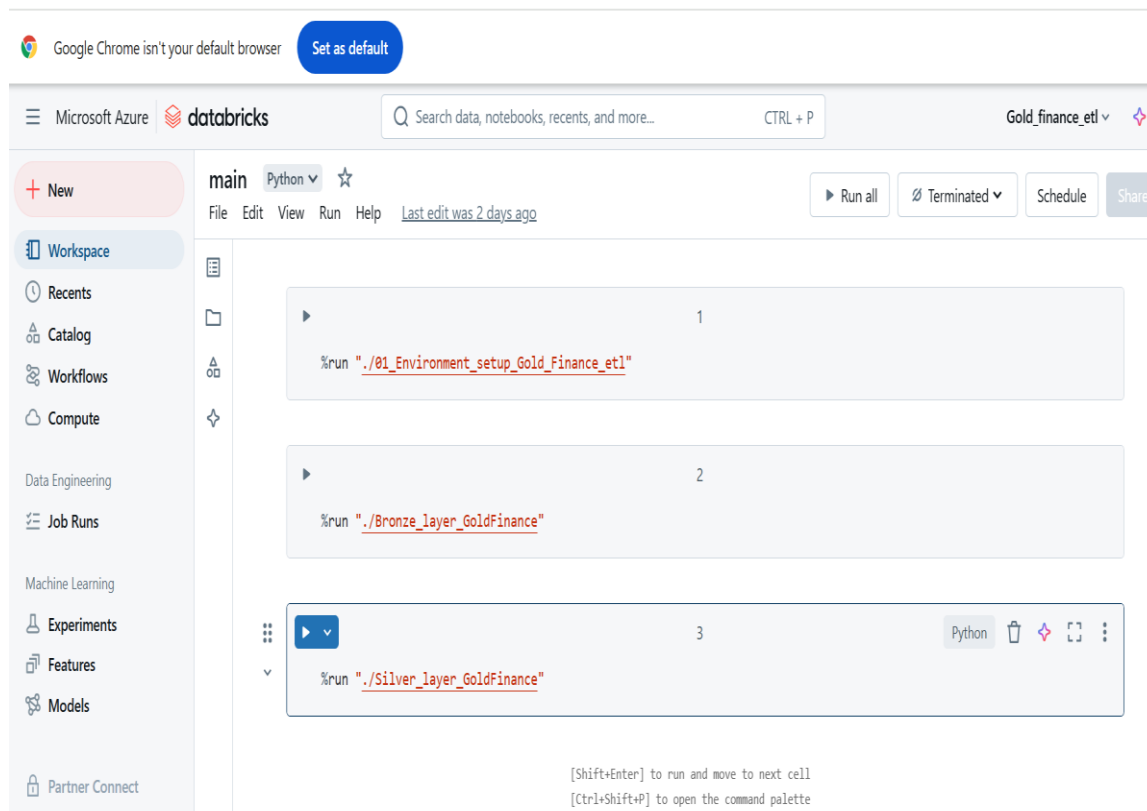
++Copy Activity Parameters: Select the respective datasets for the source and sink dataset. The copy Activity here ingests data from the Alpha vantage rest Api (Source dataset) to the azure blob storage (sink dataset)

Extraction/Transformation Layer into Silver-Layer (Azure Databricks)

Each Time

- 1) Mount new data storage in databricks (Refer to Databricks Notebook 01: Environment Setup)
- 2) Update Azure Storage Details Configuration in Environment Setup & Bronze layer code

+++ Use modularized coding as seen below:



Copy data from the Adls Gen 2 to Azure SQL

The copy activity is used to copy data from the adls (source dataset) to azure sql database (sink dataset).

+++ Error Handling Azure

The Set Variable activity is used for error handling in adf.

First: define the variable error:

Add new variable

Name

error

Type

String


Confirm

Cancel

+++ **Error Notification**

Logic Apps: Set up logic Apps, configure necessary email notification.

>>

 When a HTTP request is received

Parameters

Settings

Code view

About

HTTP URL

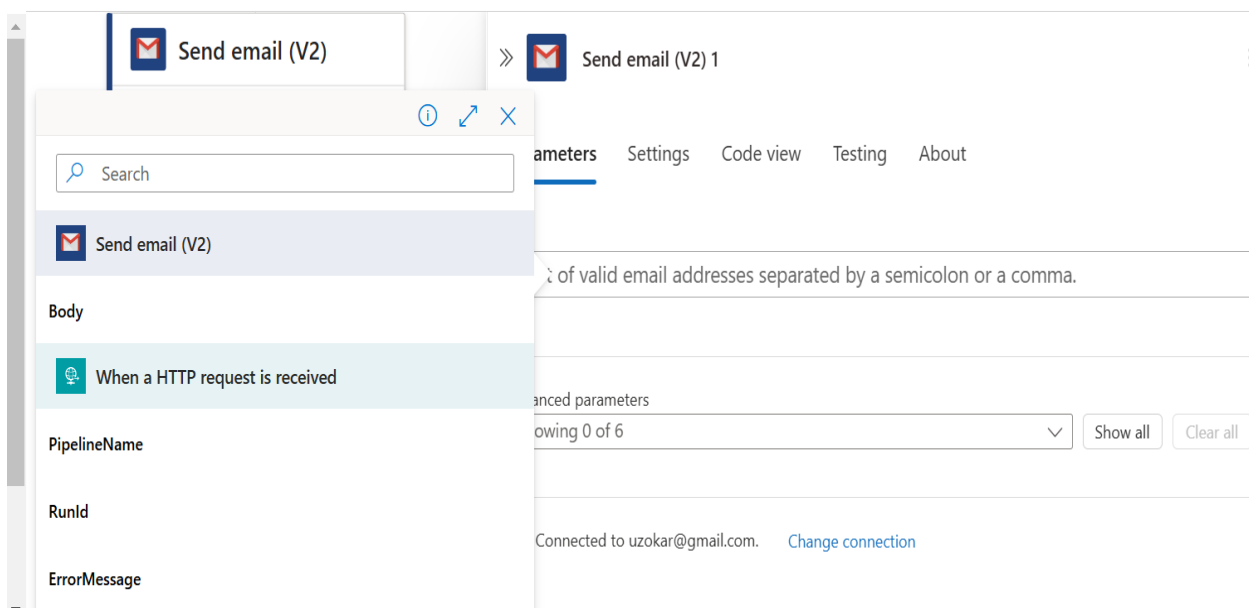
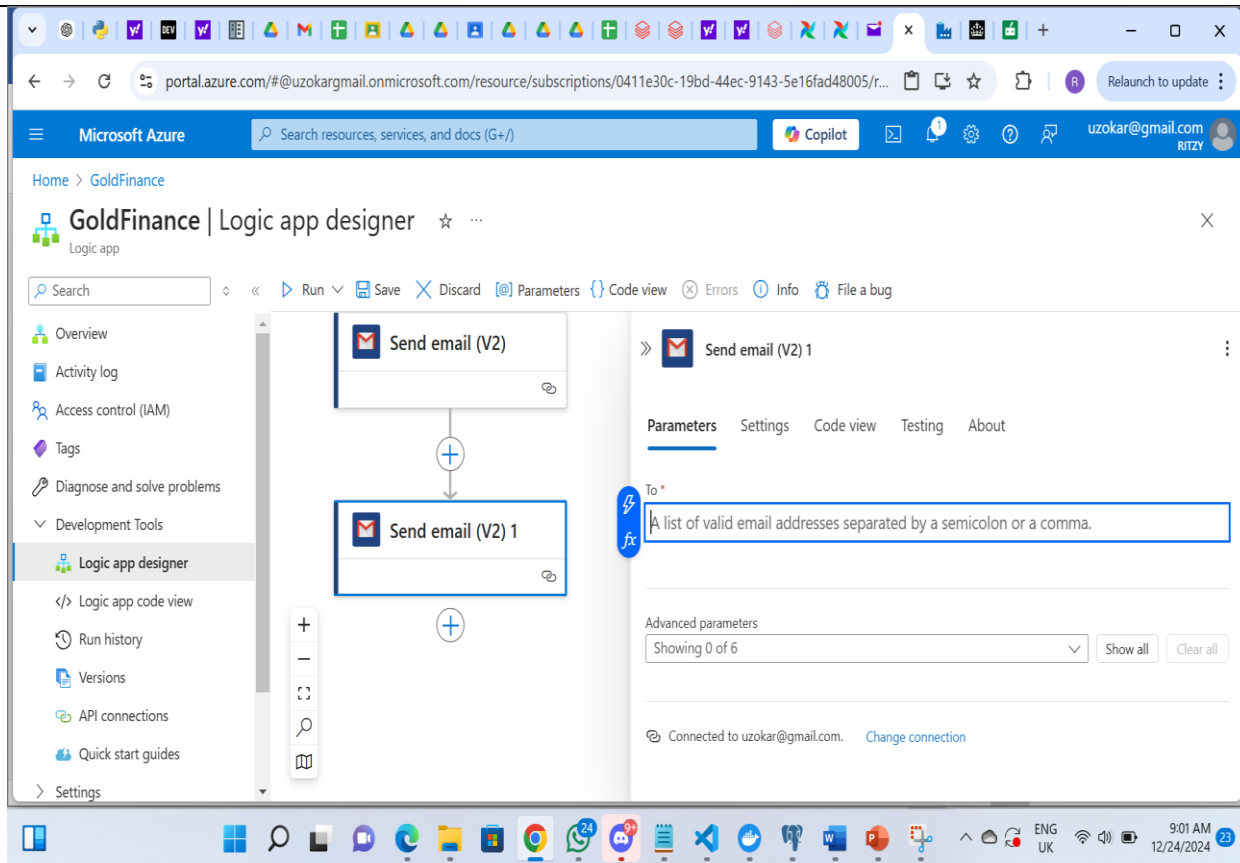
https://prod-24.ukwest.logic.azure.com:443/workflows/22220cd5fbaf47bbba43578883808e11/trigg...

Method

POST

Request Body JSON Schema

```
{
  "type": "object",
  "properties": {
    "PipelineName": {
      "type": "string"
    },
    "RunId": {
      "type": "string"
    },
    "ErrorMessage": {
      "type": "string"
    }
  }
}
```



++ Web Activity: Fill with Logic Apps parameters.

+++ Azure DevOps (CI/CD)

Step 1: Create an Azure DevOps Organization

Step 2: Fill this form to create a free parallelism request(<https://aka.ms/azpipelines-parallelism-request>) and submit for approval, it is only when it is approved that one can be able to work on the azure Organization.

forms.office.com/pages/responsepage.aspx?id=v4j5cvGGr0GRqy180BHbR5zsR558741CrNi6q8iTpANURUhKMVA3WE4wMFhHRExTVlpET1BEMI...

Azure DevOps Parallelism Request

This form is for users to request increased parallelism in Azure DevOps.

Please consider that it could take 4-5 business days to proceed the request. We are working on improving this process at the moment. Sorry for the inconvenience.

* Required

1. What is your name? *

2. What is your email address? *

Step 3: Create a Project in the organization

Step 4: Create Repos for pull request and Merge to the main branch

Azure DevOps

AgroFarms / datamigration / Repos / Pull requests / GoldFinance_etl

Datamigration

Overview

Boards

Repos

Files

Commits

Pushes

Branches

Tags

Pull requests

Advanced Security

Project settings

New pull request

feature into master

Overview Files Commits

Title

Updating pipeline: Gold_Finance_etl_pl

Description

Updating pipeline: Gold_Finance_etl_pl

38/4000

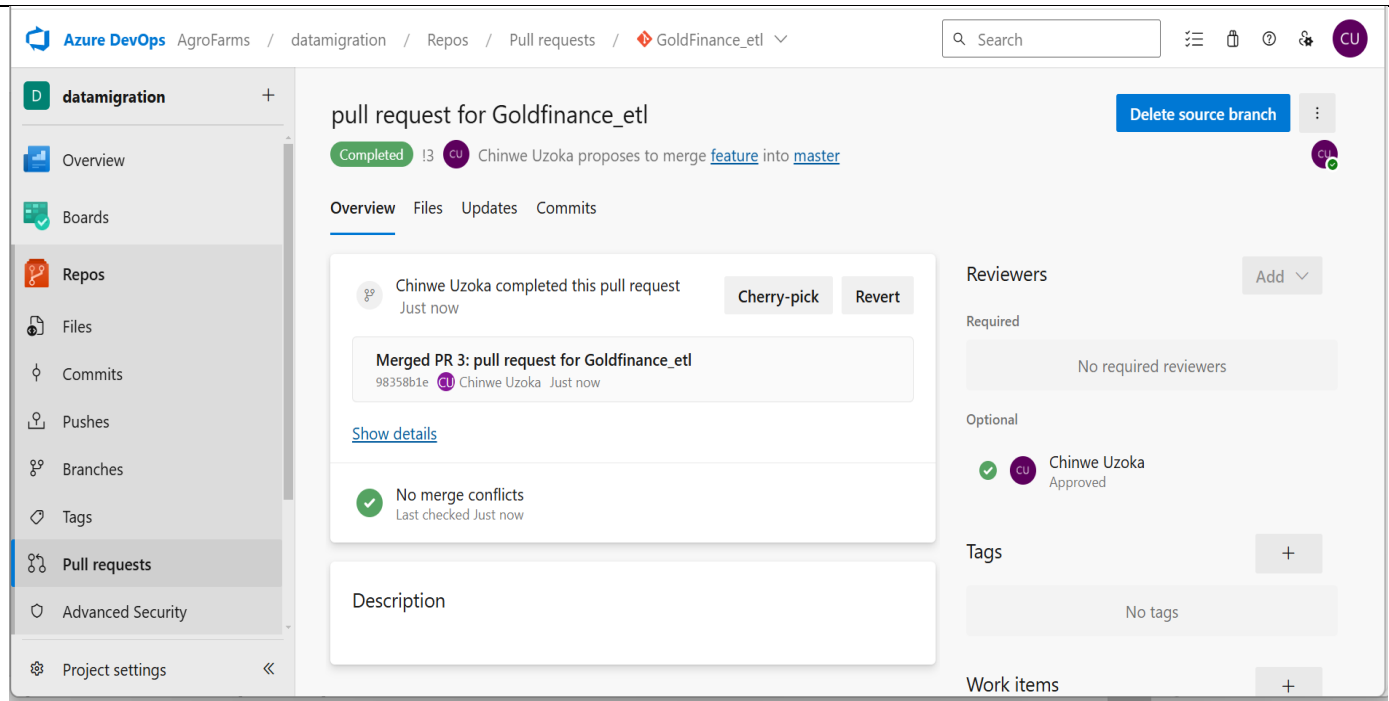
Markdown supported. Drag & drop, paste, or select files to insert.

@ # Link work items

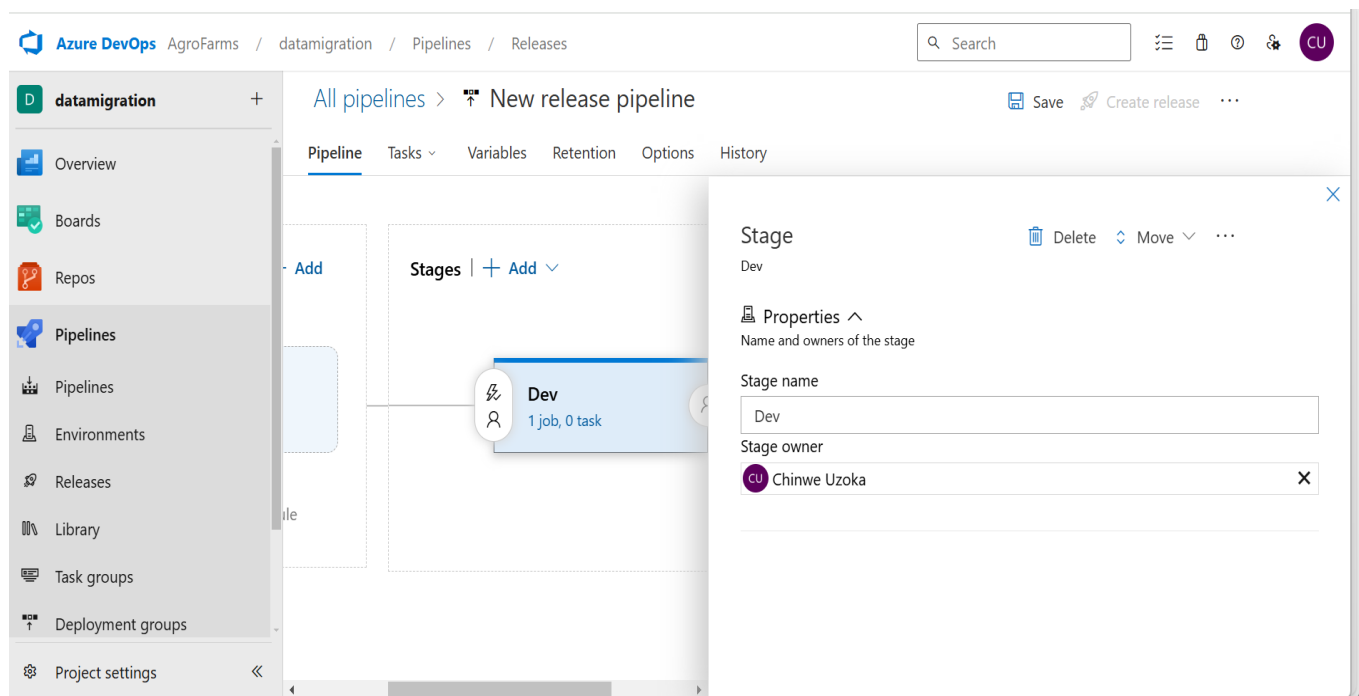
Updating pipeline: Gold_Finance_etl_pl

Reviewers

Add required reviewers



Step 4: adf_publish:1) Validate the pipeline 2) Generate Arm template and keep it in adf_publish branch 3) Deploy code to Adf Dev Live mode.
+++ Use the adf build.yml file to build the pipeline and generate an Arm template.



Azure DevOps AgroFarms / datamigration / Pipelines / Releases

Search

CU

datamigration +

Overview
Boards
Repos
Pipelines
Environments
Releases
Library
Task groups
Deployment groups
Project settings

All pipelines > GoldFinanceRelease

Save Create release ...

Pipeline Tasks Variables Retention Options History

Artifacts | + Add

Stages | + Add

_GoldFinance_etl

Dev
1 job, 0 task

Schedule not set

Azure DevOps AgroFarms / datamigration / Pipelines / Releases / GoldFinanceRelease

Search

CU

All pipelines > GoldFinanceRelease

Save Create release View releases ...

Pipeline Tasks Variables Retention Options History

Pipeline variables

Variable groups
Predefined variables

Filter by keywords Scope

Name	Value
factoryName	pcuretaildf
AzureBlobStorage2_connectionString	
AzureDataLakeStorage1_accountKey	
AzureDataLakeStorage1s_accountKey	
AzureDatabricks_Is_accessToken	
AzureDatabricks_main_Is_accessToken	
SqlServer_Is_password	
fin_alphavantage_raw_Is_connectionString	
AzureDataLakeStorage1_properties_typePrope...	https://pcupos.dfs.core.windows.net/

List Grid

Azure DevOps AgroFarms / datamigration / Pipelines / Releases

Search

CU

datamigration +

Overview
Boards
Repos
Pipelines
Environments
Releases
Library
Task groups
Deployment groups
Project settings

All pipelines > GoldFinanceRelease

Save Create release ...

Pipeline Tasks Variables Retention Options History

Continuous deployment trigger

Build: _GoldFinance_etl

Enabled

Creates a release every time a new build is available.

Build branch filters

Type Build branch Build tags

Include master

+ Add

Pull request trigger

Build: _GoldFinance_etl

Disabled

Azure DevOps AgroFarms / datamigration / Pipelines / Releases / GoldFinanceRelease / Release-1

Search

CU

datamigration +

- Overview
- Boards
- Repos
- Pipelines
- Pipelines
- Environments
- Releases
- Library
- Task groups
- Deployment groups
- Project settings

GoldFinanceRelease > Release-1

Help

Pipeline Variables History + Deploy Cancel Refresh Edit ...

Release

Continuous deployment

for Chinwe Uzoka
12/8/2024, 3:04 PM

Artifacts

_GoldFinance_etl
20241208.1
master

Stages

Dev

Succeeded

on 12/8/2024, 3:05 PM

AgroFarms / datamigration / Pipelines / Releases / GoldFinanceRelease

Search

CU

All pipelines > GoldFinanceRelease

Save Create release View releases ...

Pipeline Tasks Variables Retention Options History

Artifacts + Add

_GoldFinance_etl

Schedule not set

Stages + Add

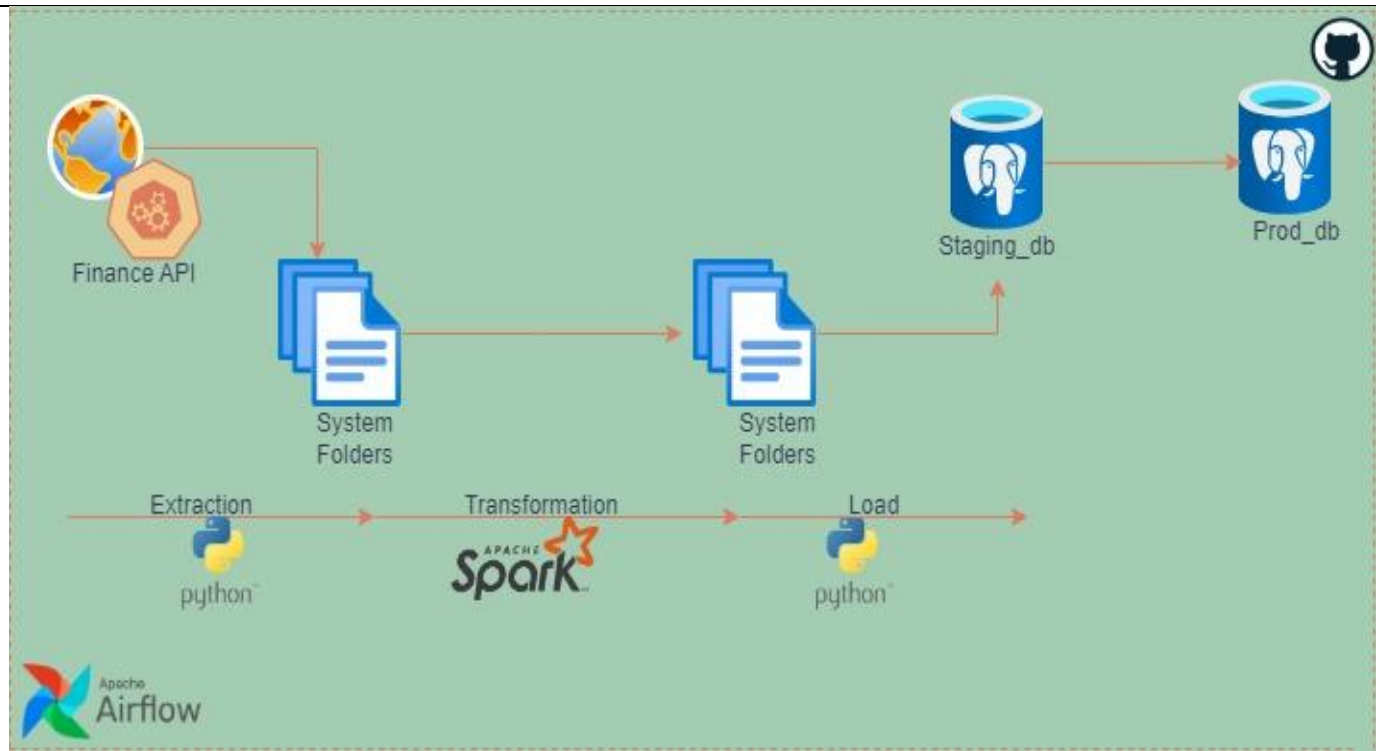
Dev
1 job, 1 task

Test
1 job, 1 task

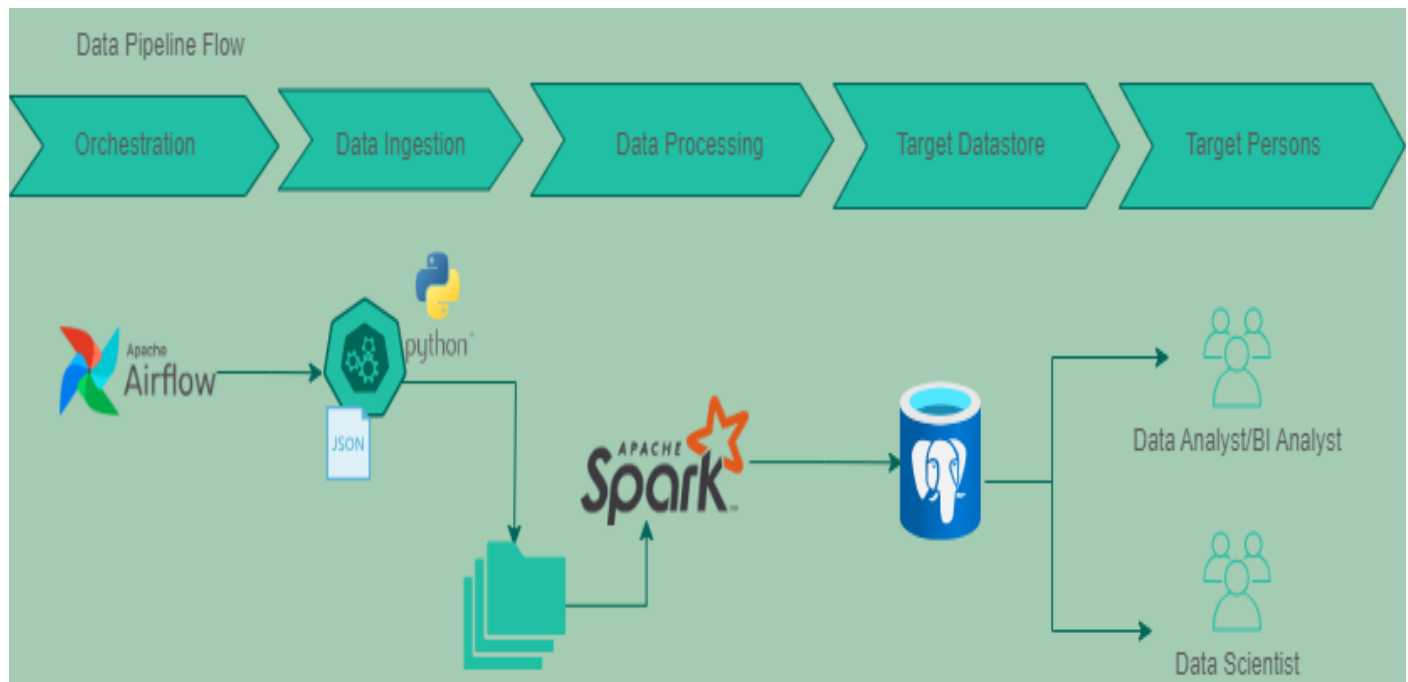
Prod
1 job, 1 task

2.0 Onprem Project: Gold_Finance ETL

Onprem Solution Architecture



Onprem Data Pipeline Flow



Onprem Project Tech Stack

S/N	Tech Stack	Details
1	Vscode	Python programming
2	Pyspark on Vscode	Transformation
3	Postgres	Database
4	Airflow	Orchestration

+++ Pyspark Vscode

- **Purpose:** Run Spark and custom Python scripts for data ingestion and transformation.
- **Aim:**
 - Writing Spark jobs to process large-scale datasets.

- Using Python scripts to ingest, transform and load data to PostgreSQL.

+++PostgreSQL

- **Purpose:** A relational database for storing metadata, structured data, and managing transactions.
- **Aim:**
 - Querying relational datasets using SQL.

Storing and retrieving structured data for analysis or visualization

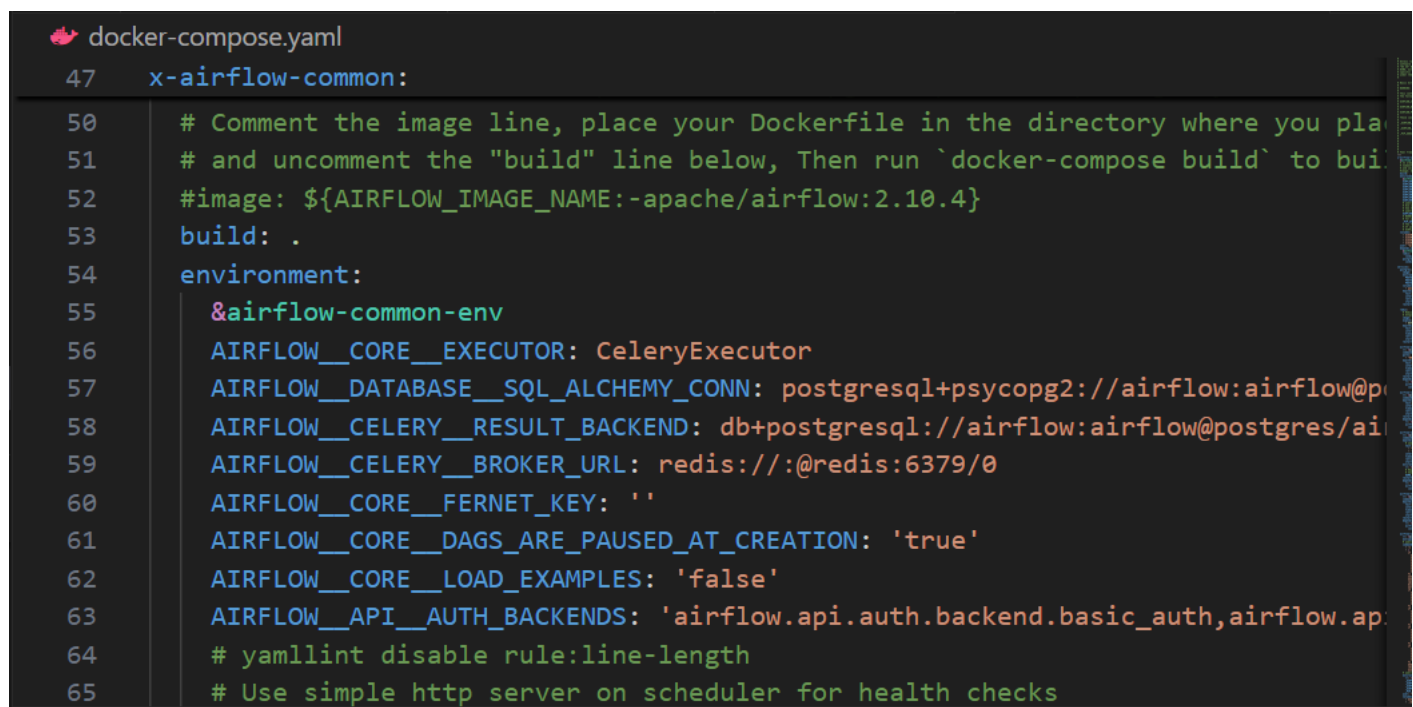
+++ Airflow

- **Purpose:** Workflow orchestration and ETL pipeline management.
- **Aim:**
 - Building and scheduling DAGs (Directed Acyclic Graphs) to automate workflows.
 - Managing and monitoring pipelines through the Airflow web interface.
- **Web Interface URL:** <http://localhost:8080>

Step 1: On vscode, open a bash terminal and paste the code below

```
curl -Lfo 'https://airflow.apache.org/docs/apache-airflow/2.10.4/docker-compose.yaml'
```

Adjust the yaml file as per below screenshot



```
docker-compose.yaml
47  x-airflow-common:
48
49      # Comment the image line, place your Dockerfile in the directory where you place this file
50      # and uncomment the "build" line below, Then run `docker-compose build` to build the image
51      #image: ${AIRFLOW_IMAGE_NAME:-apache/airflow:2.10.4}
52      build: .
53
54      environment:
55          &airflow-common-env
56          AIRFLOW__CORE__EXECUTOR: CeleryExecutor
57          AIRFLOW__DATABASE__SQL_ALCHEMY_CONN: postgresql+psycopg2://airflow:airflow@postgres/airflow
58          AIRFLOW__CELERY__RESULT_BACKEND: db+postgresql://airflow:airflow@postgres/airflow
59          AIRFLOW__CELERY__BROKER_URL: redis://:@redis:6379/0
60          AIRFLOW__CORE__FERNET_KEY: ''
61          AIRFLOW__CORE__DAGS_ARE_PAUSED_AT_CREATION: 'true'
62          AIRFLOW__CORE__LOAD_EXAMPLES: 'false'
63          AIRFLOW__API__AUTH_BACKENDS: 'airflow.api.auth.backend.basic_auth,airflow.api.auth.backend.session'
64          # yamllint disable rule:line-length
65          # Use simple http server on scheduler for health checks
```

Step 2: Create a Dockerfile and a requirements.txt file.

Step 3: Paste the below in the bash terminal

```
mkdir -p ./dags ./logs ./plugins ./config
```

```
echo -e "AIRFLOW_UID=$(id -u)" > .env
```

Step 4: Create scripts as per below structure

```
myenv/
├── dags/
│   ├── etl_pipeline_dag.py
│   ├── modules/
│   │   ├── __init__.py
│   │   ├── extract.py
│   │   ├── transform.py
│   │   ├── load.py
│   │   ├── helpers.py
│   ├── configs/
│   │   ├── dev_config.json
│   │   ├── prod_config.json
│   └── tests/
│       ├── test_extract.py
│       ├── test_transform.py
│       └── test_load.py
```


<https://airflow.apache.org/docs/apache-airflow/stable/howto/docker-compose/index.html#fetching-docker-compose-yaml>

Step 5: Run the below command on bash terminal



docker compose up airflow-init

docker build .

docker compose up

 Airflow

[DAGs](#) [Cluster Activity](#) [Datasets](#) [Security](#) [Browse](#) [Admin](#) [Docs](#)

 21:35 UTC 

List Connection



Search

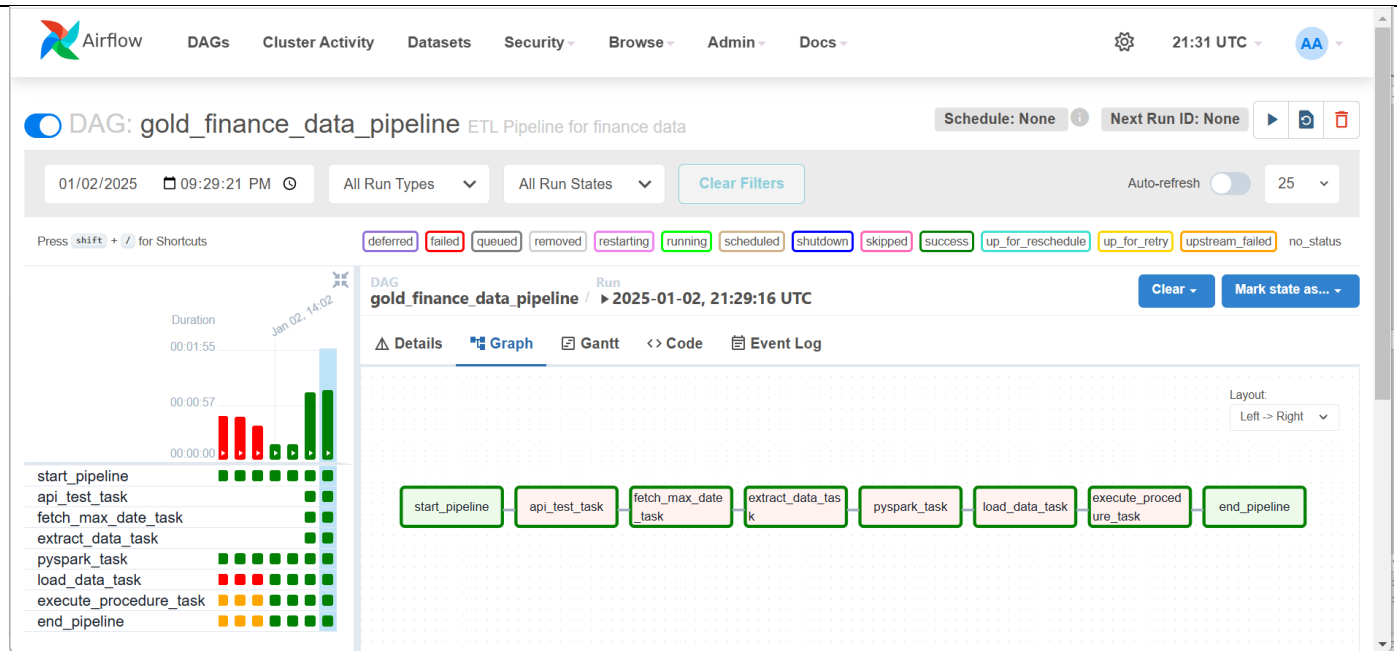
+

 Actions

←

Record Count: 2

<input type="checkbox"/>	Conn Id	Conn Type	Description	Host	Port	Is Encrypted	Is Extra Encrypted
<input type="checkbox"/>	  my_postgres_connection	postgres		host.docker.internal	5432	False	False



Linux Commands

```
rita@DESKTOP-1F00B3O:~/Finance_ETL/myenv$ ls -ld ./dags/tmp/raw
```

```
drwxrwxrwx 2 rita root 4096 Dec 28 17:54 ./dags/tmp/raw
```

```
rita@DESKTOP-1F00B3O:~/Finance_ETL/myenv$ ls -ld ./dags/tmp/transformed_data
```

```
drwxr-xr-x 2 rita root 4096 Dec 28 17:55 ./dags/tmp/transformed_data
```

```
rita@DESKTOP-1F00B3O:~/Finance_ETL/myenv$ chmod 777 ./dags/tmp/transformed_data
```

```
rita@DESKTOP-1F00B3O:~/Finance_ETL/myenv$ ls -ld ./dags/tmp/transformed_data
```

```
drwxrwxrwx 2 rita root 4096 Dec 28 17:55 ./dags/tmp/transformed_data
```

```
rita@DESKTOP-1F00B3O:~/Finance_ETL/myenv$ docker exec -u root -it myenv-airflow-worker-1 /bin/bash
```

```
root@77bcd09e75b5:/app# ls -ld /app/dags/tmp/raw
```

```
drwxrwxrwx 1 airflow root 4096 Dec 28 17:54 /app/dags/tmp/raw
```

```
rita@DESKTOP-1F00B3O:~/Finance_ETL/myenv$ docker ps
```

CONTAINER ID	IMAGE	COMMAND	CREATED	STATUS	PORTS
3673140a474c	myenv-airflow-scheduler	"/usr/bin/dumb-init ..."	14 hours ago	Up 9 minutes (healthy)	8080/tcp
77bcd09e75b5	myenv-airflow-worker	"/usr/bin/dumb-init ..."	14 hours ago	Up 9 minutes (healthy)	8080/tcp
e89d6ad6b639	myenv-airflow-webserver	"/usr/bin/dumb-init ..."	14 hours ago	Up 9 minutes (healthy)	0.0.0.0:8080->8080/tcp
b979ae8ff5c3	myenv-airflow-triggerer	"/usr/bin/dumb-init ..."	14 hours ago	Up 9 minutes (healthy)	8080/tcp

a03182e4283a 6379/tcp	redis:7.2-bookworm myenv-redis-1	"docker-entrypoin...	4 days ago	Up 10 minutes (healthy)
5b6d7ef142a6 myenv-postgres-1	postgres:13	"docker-entrypoin...	4 days ago	Up 10 minutes (healthy) 5432/tcp
(myenv) rita@DESKTOP-1F00B3O:/tmp\$ mkdir -p /tmp/spark_temp, sudo chmod 777 /tmp/spark_temp				

Project: Gold_Finance ETL: End				

Team Alpha				