

P-4: Work with the database

Global Inbound and Outbound Travel

Juilee Patil – NUID 002724809

Raksha Israni – NUID: 002925990

Dristi Dani – NUID: 002756885

Ashwin Kumar Kuchibhotla – NUID: 002655594













Introduction

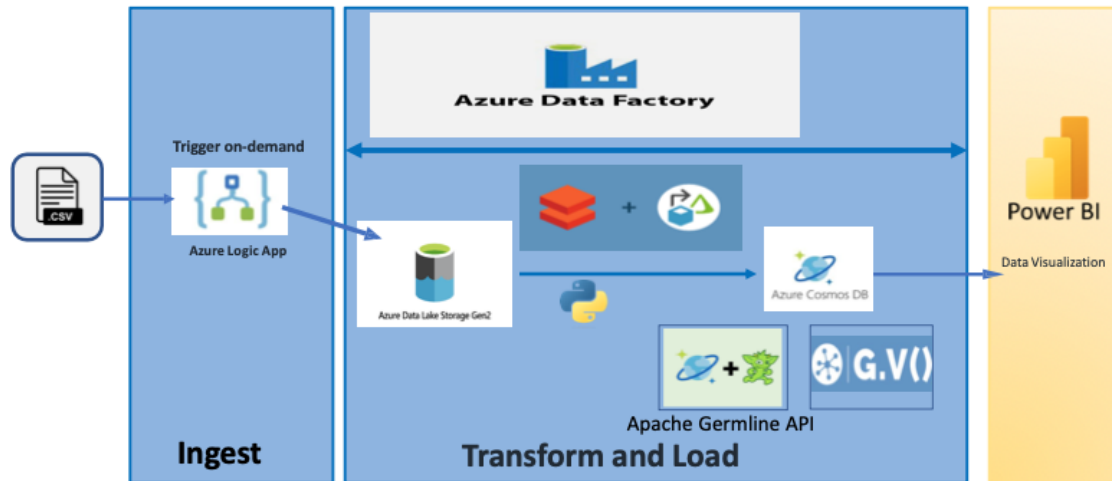
To implement the project, we have followed our architecture diagram. We have used many services available on Azure cloud platform which are listed below.

1. Azure Logic Apps
2. Azure Databricks
3. Azure Blob Storage
4. Azure Data Factory
5. Azure Cosmos DB

Recent

Favorite

Name	Type	Last Viewed
 Team4LogicApp	Logic App (Standard)	a few seconds ago
 Team4adf	Data factory (V2)	a few seconds ago
 Team4Databricks	Azure Databricks Service	a few seconds ago
 team4docdb	Azure Cosmos DB account	16 minutes ago
 team4graphdb	Azure Cosmos DB account	18 minutes ago
 team4bloba	Storage account	7 hours ago
 Team4RG	Resource group	10 hours ago
 team4sa	Storage account	3 days ago
 team4rg8c26	Storage account	4 days ago
 Azure subscription 1	Subscription	5 days ago
 LogicAppRG	Resource group	5 days ago
 DefaultResourceGroup-EUS	Resource group	6 days ago
See all		



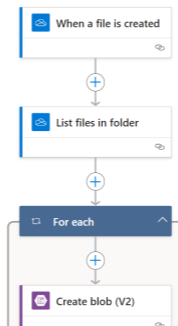
Data Refresh

For the P4 submission, we have implemented the data refresh through Azure Data Factory and logic apps.

Implementation

For loading new file through Logic App workflow -onetoblob

- 1.The Logic app will run when a new file is uploaded. It is scheduled to check every day.



Folder *

/ADBMS_PROJECT/Team4InputData

Advanced parameters

Showing 2 of 2

Show all

Clear all

Include Subfolders

No

Infer Content Type

Yes

How often do you want to check for items?

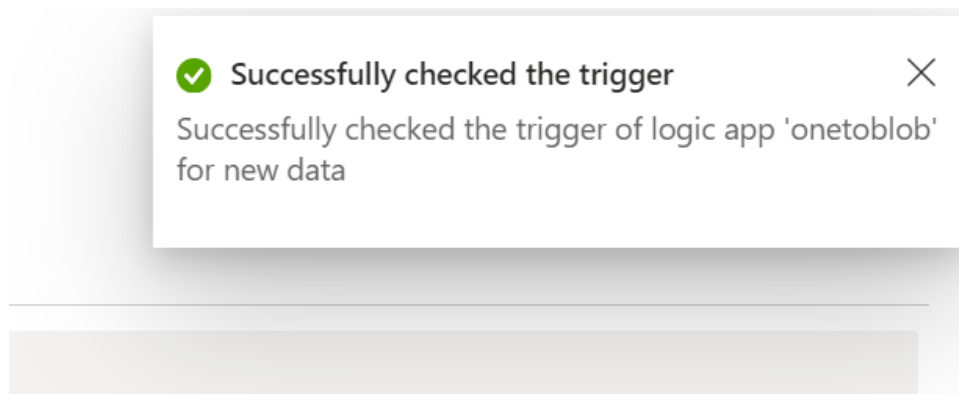
Recurrence *

Interval *

1

Frequency *

Day



jhts

2.New file added on one drive

Shared with you > ADBMS_PROJECT > Team4InputData

Name	Modified	File size	Sharing
Business_Merged.csv	12 days ago	115 KB	Shared
Inbound1.csv	About an hour ago	95.2 KB	Shared
InboundData.csv	12 days ago	3.74 MB	Shared
Pleasure_Merged.csv	12 days ago	1.05 MB	Shared
Student_Merged.csv	12 days ago	13.0 KB	Shared

3.Trigger is fired when a new file is added. The screenshot shows that the new file upload is succeeded

28-day run history			Edit columns	All runs
Start	Duration	Status		
Apr 9, 04:48 PM (34 sec ago)	00:00:07	Succeeded		
Apr 9, 04:44 PM (4 min ago)	00:00:10	Test succeeded		

4. After the successful trigger we can verify that the file gets uploaded to blob.

<input type="checkbox"/>	Output				
<input type="checkbox"/>	Business_Merged.csv	30/3/2024, 1:07:05 am	Hot (Inferred)	Block blob	141.
<input type="checkbox"/>	Inbound1.csv	9/4/2024, 4:49:29 pm	Hot (Inferred)	Block blob	95.1
<input type="checkbox"/>	InboundData.csv	30/3/2024, 1:07:42 am	Hot (Inferred)	Block blob	5.02
<input type="checkbox"/>	Output	9/4/2024, 11:34:14 am	Hot (Inferred)	Block blob	0.8
<input type="checkbox"/>	Pleasure_Merged.csv	29/3/2024, 1:15:54 am	Hot (Inferred)	Block blob	1.05
<input type="checkbox"/>	Student_Merged.csv	29/3/2024, 1:15:29 am	Hot (Inferred)	Block blob	12.9

For the refresh through the ADF, we have followed the below mentioned steps.

1. Created ADF trigger to run the databricks notebook to load latest data in document db and cosmosdb

The screenshot displays the 'Edit trigger' configuration window in the Microsoft Azure Data Factory portal. The trigger is named 'Team4Trigger' and is described as a 'Scheduled trigger to run the databricks notebook'. It is configured as a 'ScheduleTrigger' with a start date of '3/30/2024, 10:58:00 AM' and a time zone of 'Eastern Time (US & Canada) (UTC-5)'. The recurrence is set to 'Every 1 Week(s)'. The background shows the Data Factory interface with 'Factory Resources' and 'Activities' panels.

2. Scheduled the trigger to run only on weekdays at 11:30 am.

Advanced recurrence options

Run on these days

Sun	Mon	Tue	Wed	Thu	Fri	Sat
-----	-----	-----	-----	-----	-----	-----

Execute at these times (i)

Hours ×

Minutes ×

Schedule execution times

11:30

☐ Specify an end date

- Once the trigger is created and scheduled as per the requirements, we can check the status of the trigger and monitor. Below is the screenshot of trigger running successfully at 11:30 am on a week day.

The second entry is a manual trigger that can be ignored. The first entry with Triggered by value as “Team4Trigger” is the scheduled trigger.

The screenshot displays the Microsoft Azure Data Factory interface for the 'Team4adf' workspace. The 'Pipeline runs' section is active, showing a list of runs for the 'Team4ADF' pipeline. The table below summarizes the visible runs:

Pipeline name	Run start	Run end	Duration	Triggered by	Status
Team4ADF	4/9/2024, 11:30:00 AM	4/9/2024, 11:40:28 AM	10m 28s	Team4Trigger	Succeed
Team4ADF	4/9/2024, 11:22:31 AM	4/9/2024, 11:32:58 AM	10m 27s	Manual trigger	Failed

The interface also shows navigation options like 'Triggered', 'Debug', 'Rerun', and 'Cancel options', along with filters for 'Status' and 'Triggered by'.

Showing 1 - 2 items

Last refreshed 0 minutes ago

<input type="checkbox"/>	Pipeline name ↑↓	Run start ↑↓	Run end ↑↓	Duration	Triggered by	Status ↑↓
<input type="checkbox"/>	Team4ADF	4/9/2024, 11:30:00 AM	4/9/2024, 11:40:28 AM	10m 28s	Team4Trigger	✔ Succeeded
<input type="checkbox"/>	Team4ADF	4/9/2024, 11:22:31 AM	4/9/2024, 11:32:58 AM	10m 27s	Manual trigger	✖ Failed 🔔