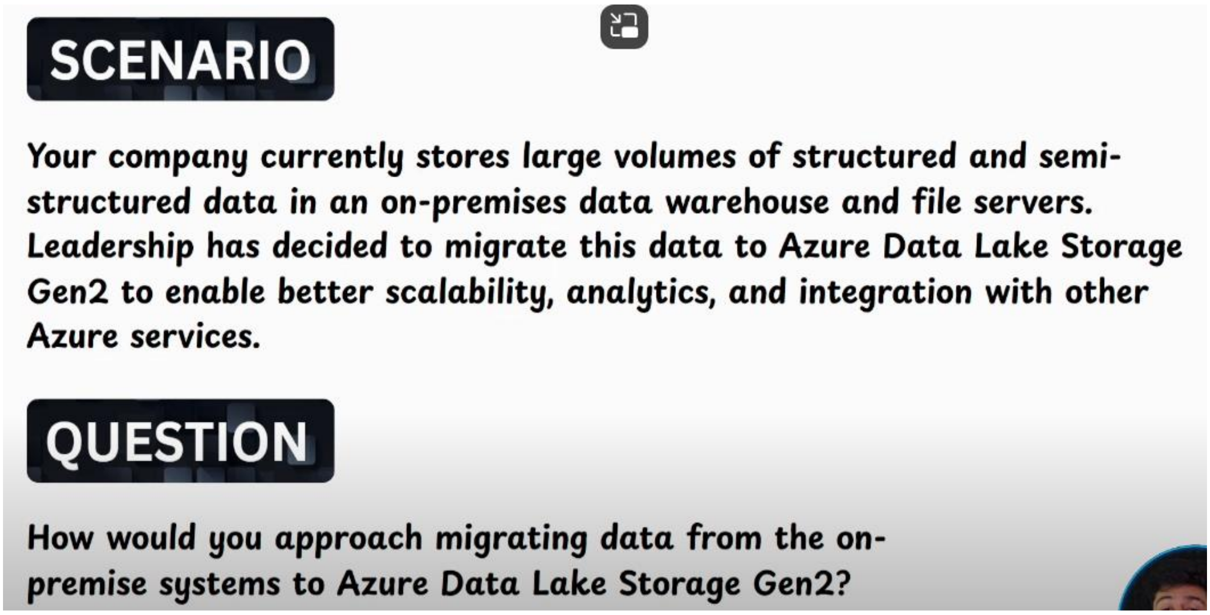


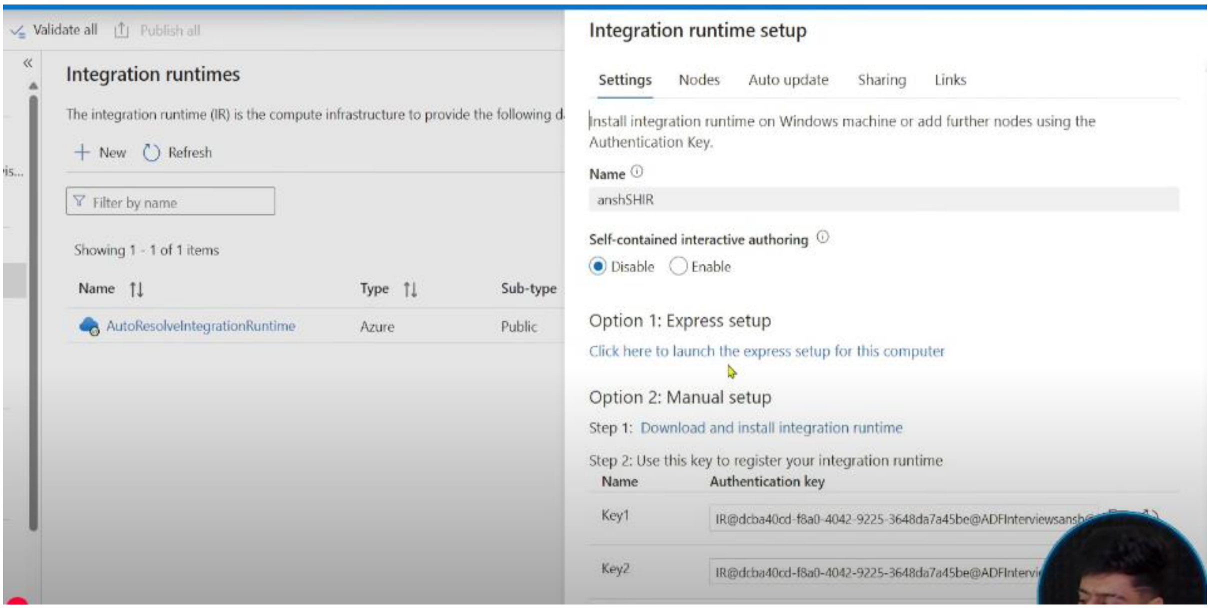
SCENARIO

Your company currently stores large volumes of structured and semi-structured data in an on-premises data warehouse and file servers. Leadership has decided to migrate this data to Azure Data Lake Storage Gen2 to enable better scalability, analytics, and integration with other Azure services.

QUESTION

How would you approach migrating data from the on-premise systems to Azure Data Lake Storage Gen2?

Use Self Hosted IR



Integration runtimes

The integration runtime (IR) is the compute infrastructure to provide the following d

Showing 1 - 1 of 1 items

Name	Type	Sub-type
AutoResolveIntegrationRuntime	Azure	Public

Integration runtime setup

Install integration runtime on Windows machine or add further nodes using the Authentication Key.

Name: anshSHIR

Self-contained interactive authoring: ● Disable ○ Enable

Option 1: Express setup

Click here to launch the express setup for this computer

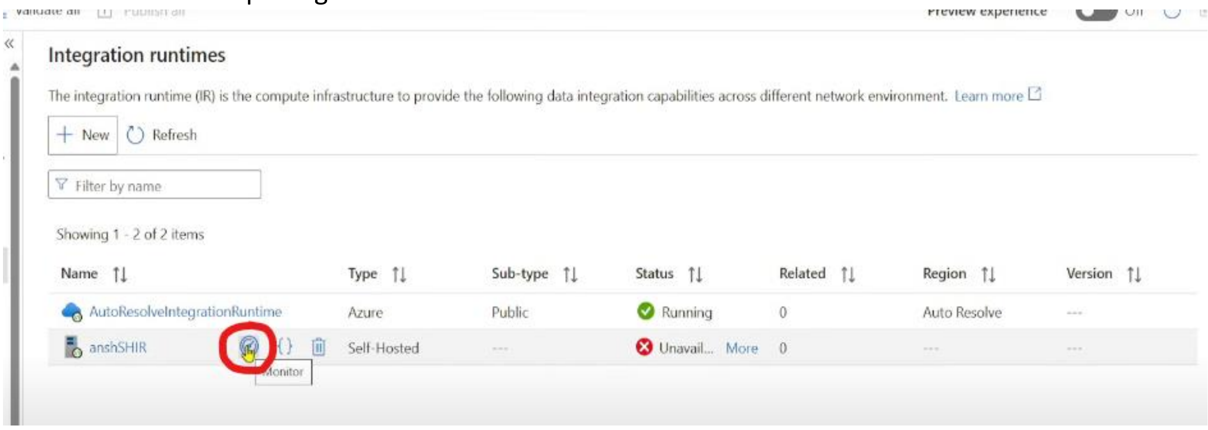
Option 2: Manual setup

Step 1: Download and install integration runtime

Step 2: Use this key to register your integration runtime

Name	Authentication key
Key1	IR@dcba40cd-f8a0-4042-9225-3648da7a45be@ADFInterviewsansh
Key2	IR@dcba40cd-f8a0-4042-9225-3648da7a45be@ADFIntervi

We need to install a package in our Local Machine

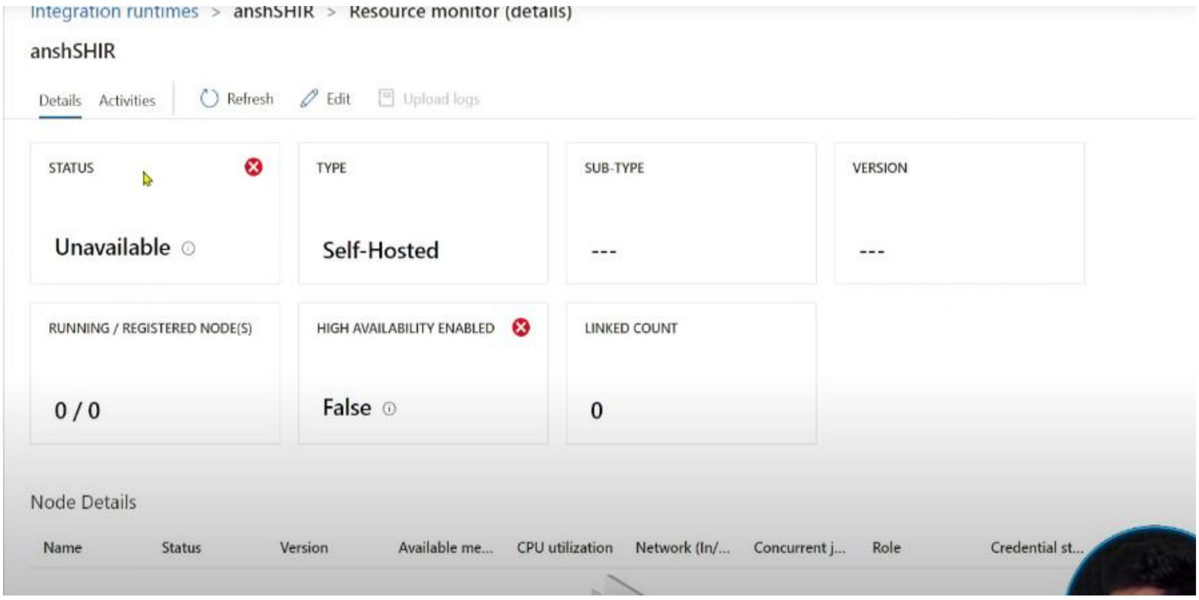


Integration runtimes

The integration runtime (IR) is the compute infrastructure to provide the following data integration capabilities across different network environment. Learn more

Showing 1 - 2 of 2 items

Name	Type	Sub-type	Status	Related	Region	Version
AutoResolveIntegrationRuntime	Azure	Public	✓ Running	0	Auto Resolve	---
anshSHIR	Self-Hosted	---	✗ Unavail... More	0	---	---



Integration runtimes > anshSHIR > Resource monitor (details)

anshSHIR

STATUS: Unavailable

TYPE: Self-Hosted

SUB-TYPE: ---

VERSION: ---

RUNNING / REGISTERED NODE(S): 0 / 0

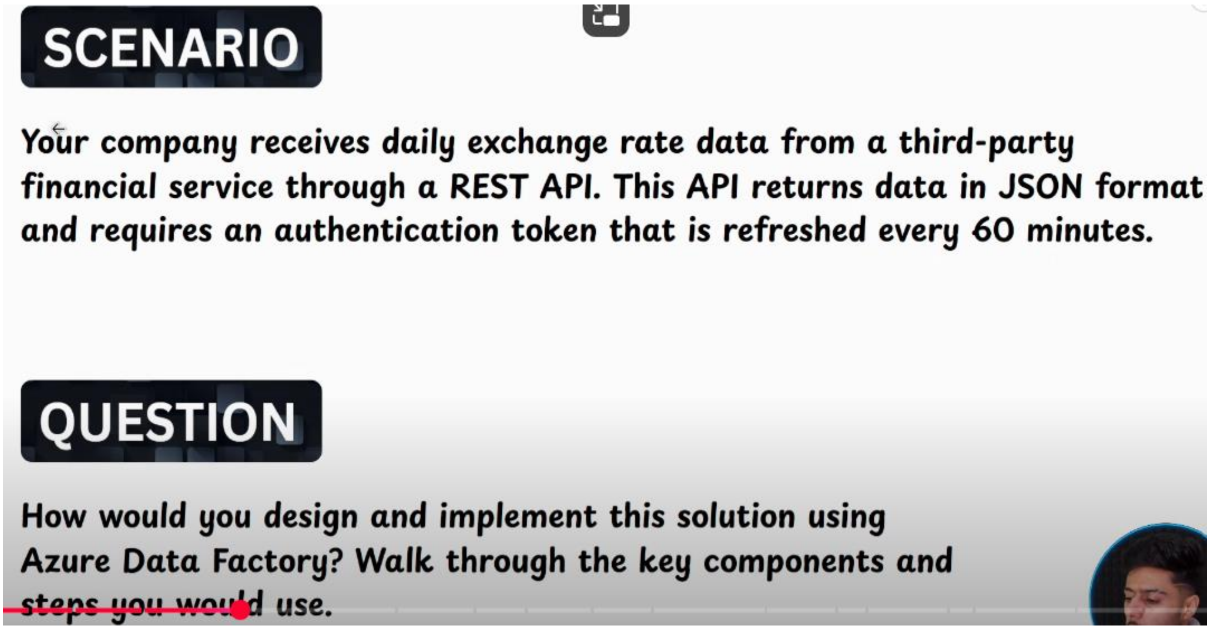
HIGH AVAILABILITY ENABLED: False

LINKED COUNT: 0

Node Details

Name	Status	Version	Available me...	CPU utilization	Network (In/...	Concurrent j...	Role	Credential st...

So start installing using Manual setup (advisable), post installing provide the key.
In real world, we create VMs as our IR(instead of downloading in local machine, download the IR in VM), in case of throttling(CPU utilisation is 80-90%) scale up VM. Easy to scaleup VM instead of local.
Post that create a LS for the datastore using SHIR.



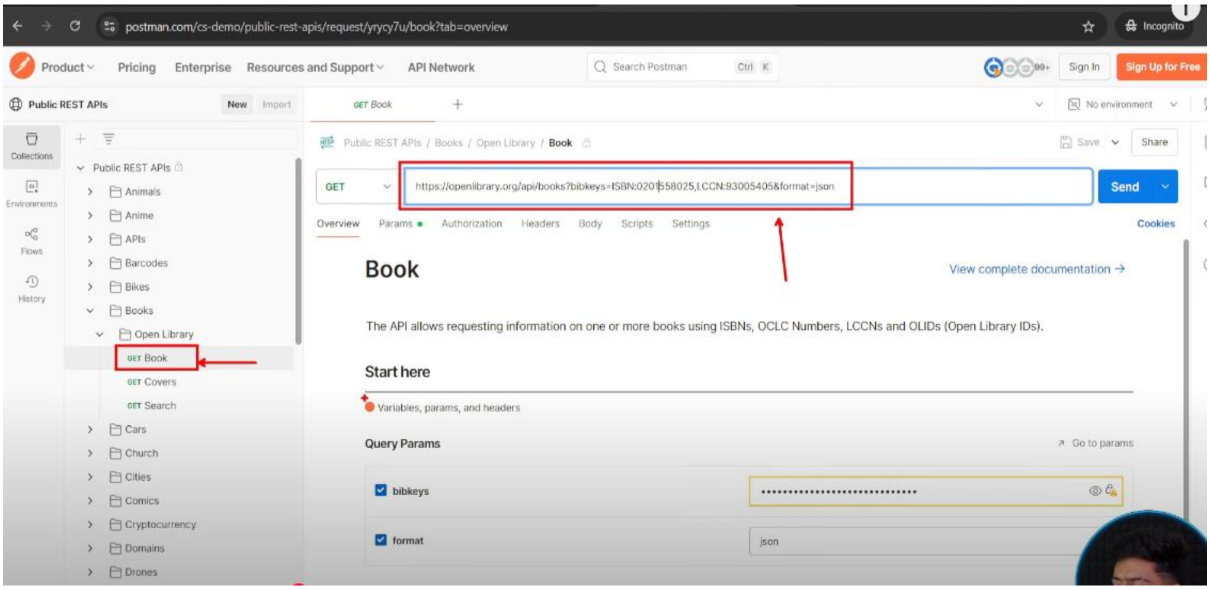
SCENARIO

Your company receives daily exchange rate data from a third-party financial service through a REST API. This API returns data in JSON format and requires an authentication token that is refreshed every 60 minutes.

QUESTION

How would you design and implement this solution using Azure Data Factory? Walk through the key components and steps you would use.

Sample API from Postman



GET https://openlibrary.org/api/books?bibkeys=ISBN:0201558025,LCCN:93005405&format=json

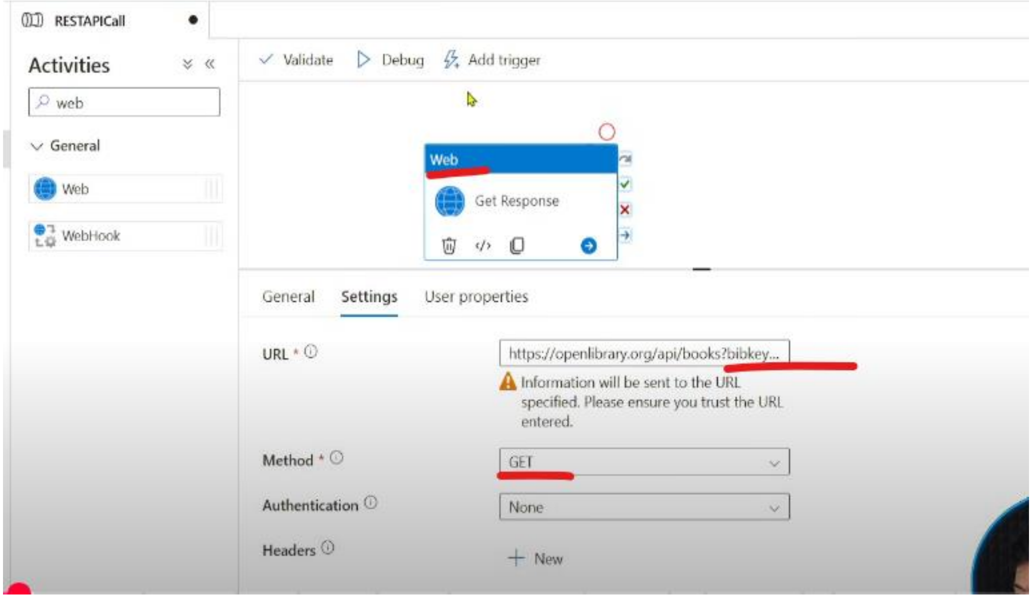
Book

The API allows requesting information on one or more books using ISBNs, OCLC Numbers, LCCNs and OLIDs (Open Library IDs).

Query Params

☑ bibkeys	••••••••••••••••••••••••••••••
☑ format	json

Query parameters: used for authentication of API



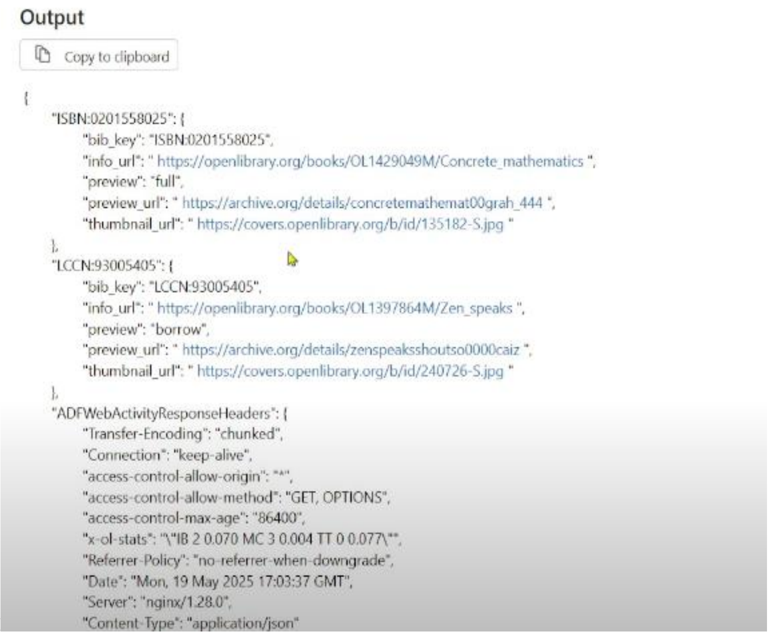
URL * https://openlibrary.org/api/books?bibkey...

Information will be sent to the URL specified. Please ensure you trust the URL entered.

Method * GET

Authentication None

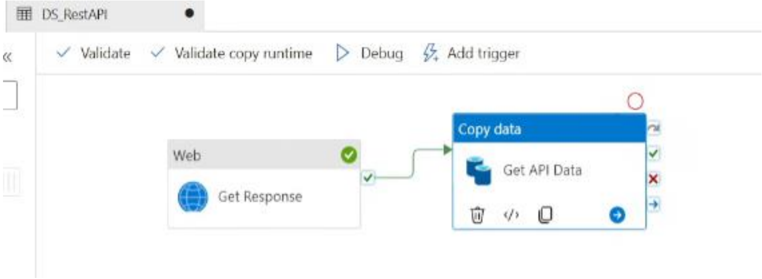
Headers



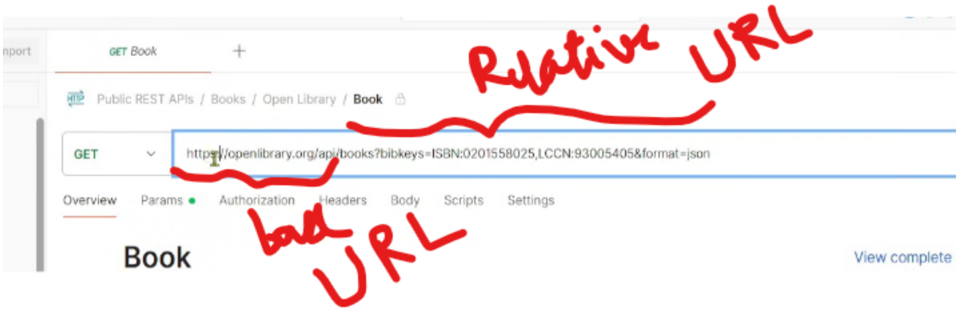
```
{
    "ISBN:0201558025": {
        "bib_key": "ISBN:0201558025",
        "info_url": " https://openlibrary.org/books/OL1429049M/Concrete_mathematics ",
        "preview": "full",
        "preview_url": " https://archive.org/details/concretemathemat00grah_444 ",
        "thumbnail_url": " https://covers.openlibrary.org/b/id/135182-S.jpg "
    },
    "LCCN:93005405": {
        "bib_key": "LCCN:93005405",
        "info_url": " https://openlibrary.org/books/OL1397864M/Zen_speaks ",
        "preview": "borrow",
        "preview_url": " https://archive.org/details/zenspeaksshoutso0000caiz ",
        "thumbnail_url": " https://covers.openlibrary.org/b/id/240726-S.jpg "
    },
    "ADFWebActivityResponseHeaders": {
        "Transfer-Encoding": "chunked",
        "Connection": "keep-alive",
        "access-control-allow-origin": "*",
        "access-control-allow-method": "GET, OPTIONS",
        "access-control-max-age": "86400",
        "x-ol-stats": "\"IB 2 0.070 MC 3 0.004 TT 0 0.077\"",
        "Referrer-Policy": "no-referrer-when-downgrade",
        "Date": "Mon, 19 May 2025 17:03:37 GMT",
        "Server": "nginx/1.28.0",
        "Content-Type": "application/json"
```

Debug to see output

Use a Copy activity to store the response



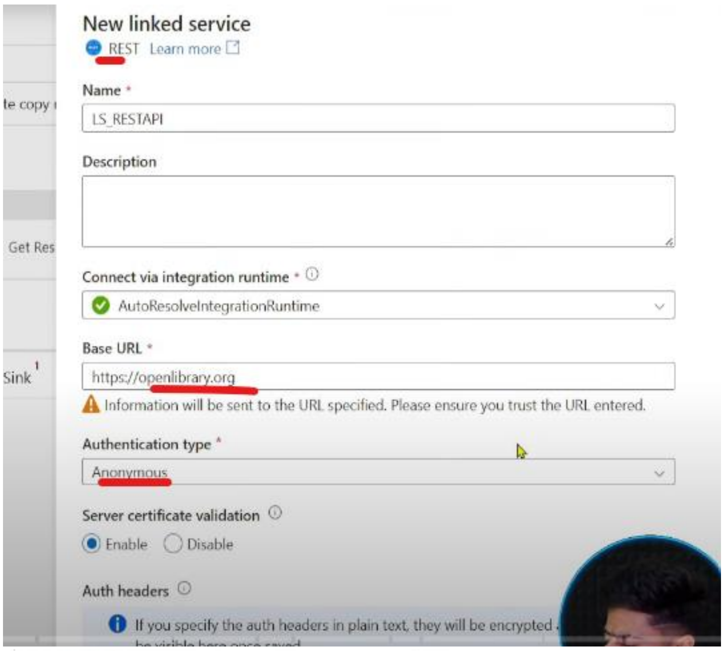
Source → REST dataset with REST LS



https://openlibrary.org/api/books?bibkeys=ISBN:0201558025,LCCN:93005405&format=json

Relative URL

base URL



New linked service

REST Learn more

Name *

LS_RESTAPI

Description

Connect via integration runtime *

AutoResolveIntegrationRuntime

Base URL *

https://openlibrary.org

Information will be sent to the URL specified. Please ensure you trust the URL entered.

Authentication type *

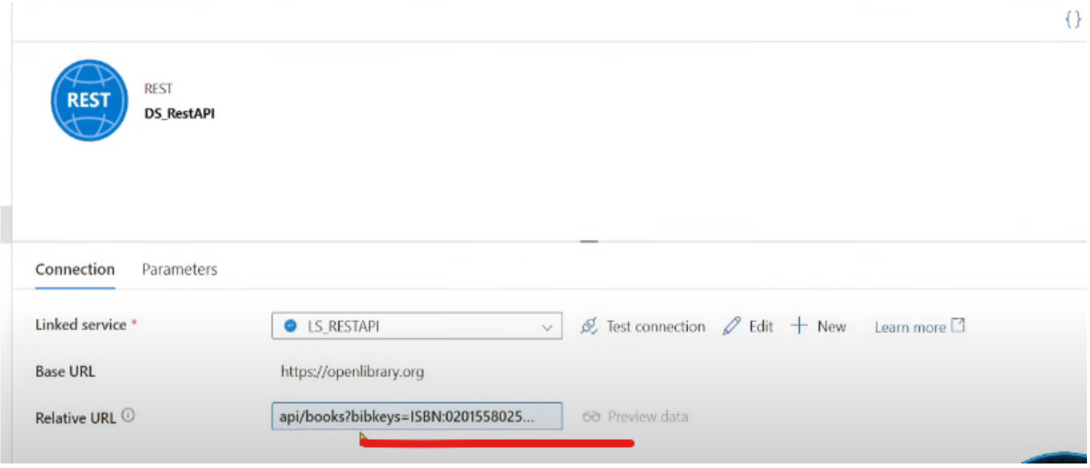
Anonymous

Server certificate validation

● Enable ○ Disable

Auth headers

If you specify the auth headers in plain text, they will be encrypted



REST

DS_RestAPI

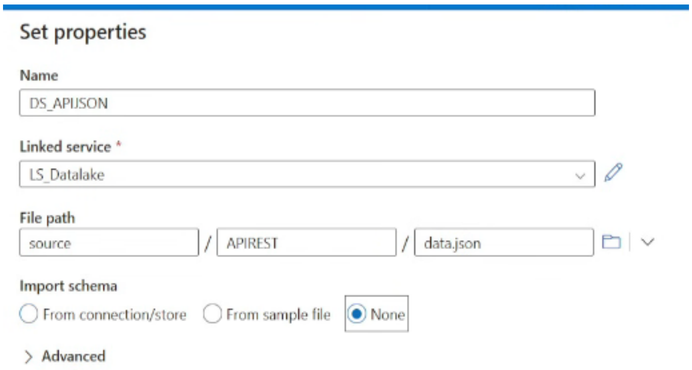
Connection Parameters

Linked service * LS_RESTAPI

Base URL https://openlibrary.org

Relative URL api/books?bibkeys=ISBN:0201558025...

Sink → ADLS in JSON format



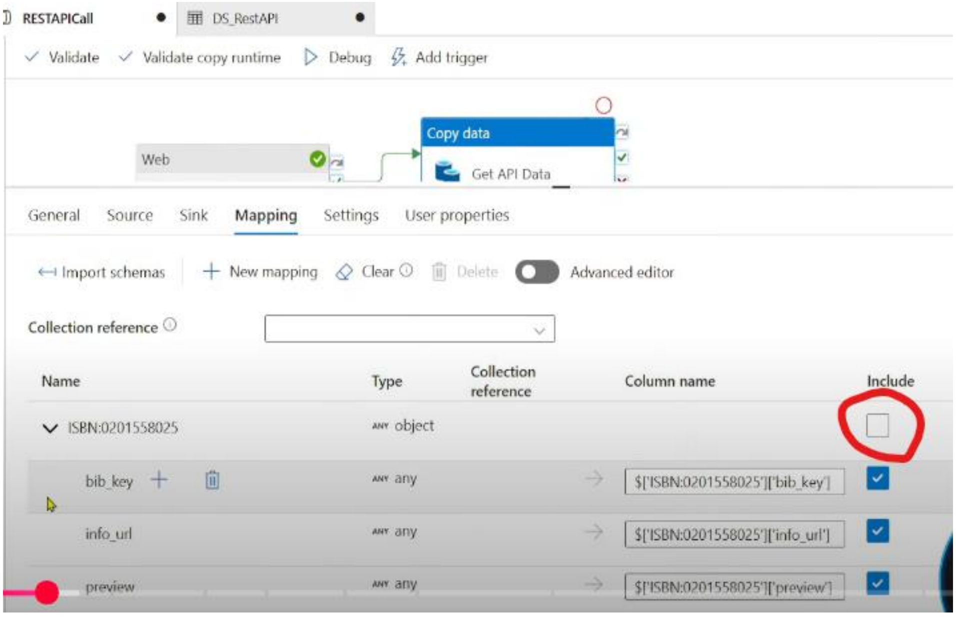
Set properties

Name: DS_APIJSON

Linked service *: LS_Datalake

File path: source / APIREST / data.json

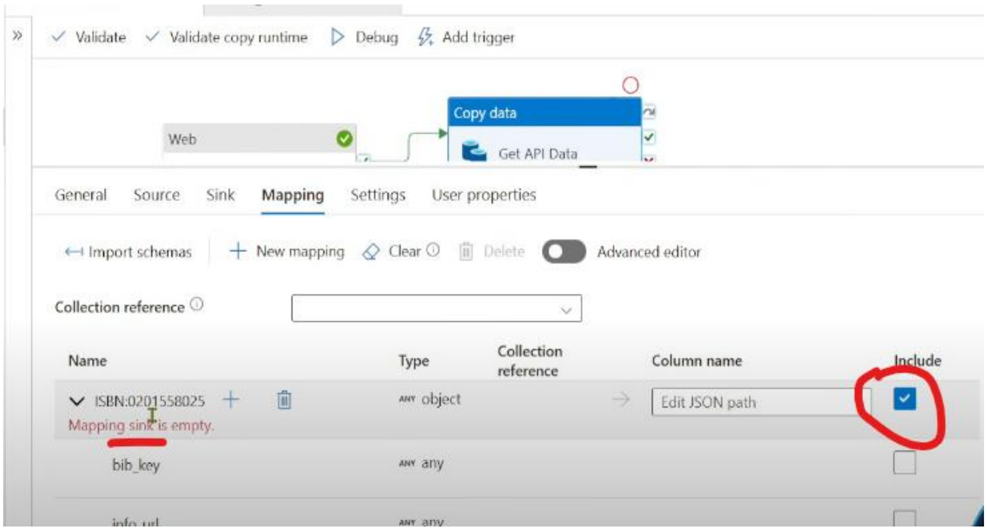
Import schema: ○ From connection/store ○ From sample file ● None



Collection reference:

Name	Type	Collection reference	Column name	Include
ISBN:0201558025	ANY object			□
bib_key	ANY any		$['ISBN:0201558025']['bib_key']	☑
info_url	ANY any		$['ISBN:0201558025']['info_url']	☑
preview	ANY any		$['ISBN:0201558025']['preview']	☑

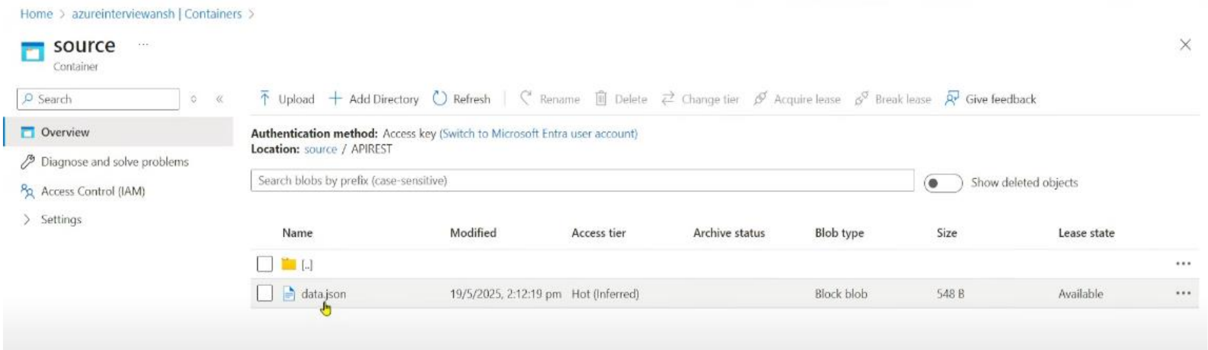
Include the parent object



Collection reference:

Name	Type	Collection reference	Column name	Include
ISBN:0201558025 Mapping sink is empty.	ANY object		Edit JSON path	☑
bib_key	ANY any			□
info_url	ANY any			□

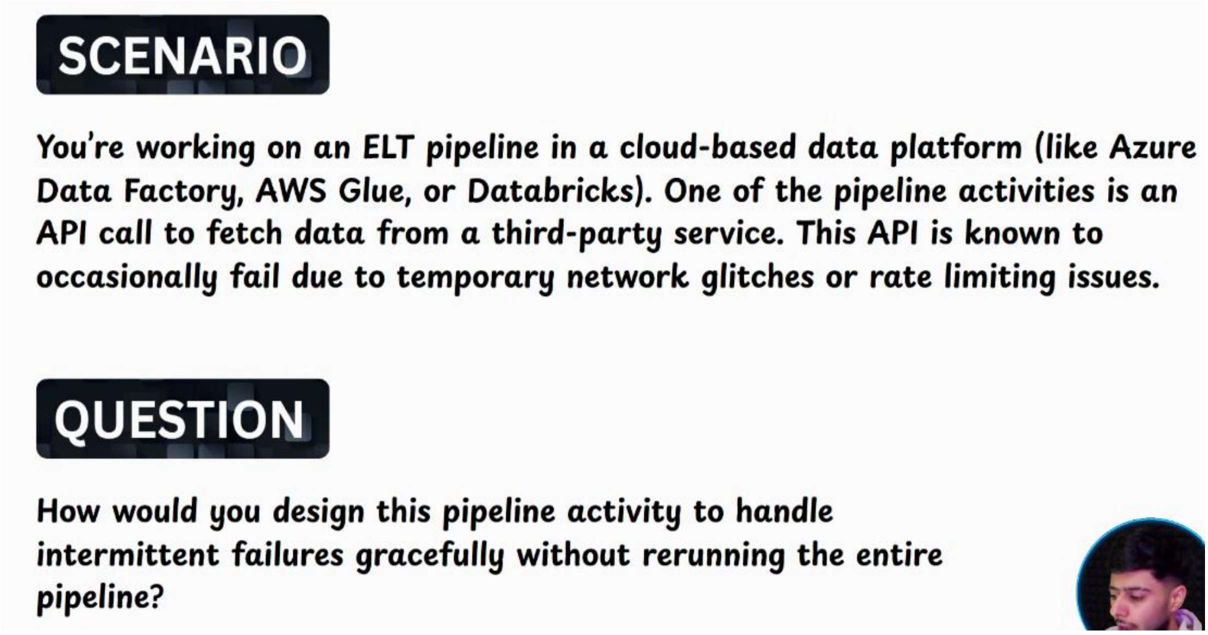
Now click on Import Schemas again → Now debug the pipeline



Authentication method: Access key (Switch to Microsoft Entra user account)
Location: source / APIREST

	Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
□	[..]						
□	data.json	19/5/2025, 2:12:19 pm	Hot (Inferred)		Block blob	548 B	Available

Additional: Now let's consider if the output of Web activity is in csv format(not in json format) →
Always use HTTP connection instead of REST in case of Delimited text
Because HTTP has an option of choosing Format (REST doesnot have)



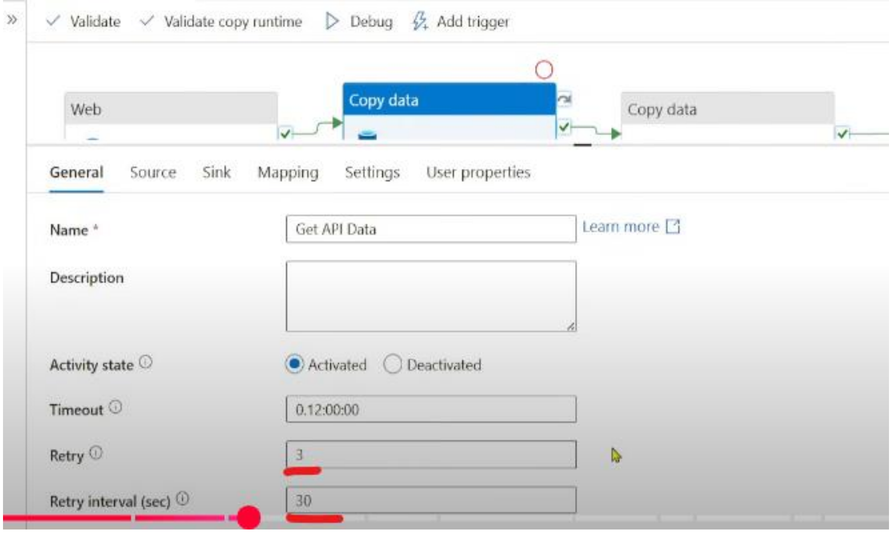
SCENARIO

You're working on an ELT pipeline in a cloud-based data platform (like Azure Data Factory, AWS Glue, or Databricks). One of the pipeline activities is an API call to fetch data from a third-party service. This API is known to occasionally fail due to temporary network glitches or rate limiting issues.

QUESTION

How would you design this pipeline activity to handle intermittent failures gracefully without rerunning the entire pipeline?

Make pipeline robust enough to retry from the failed activity



Name * Get API Data

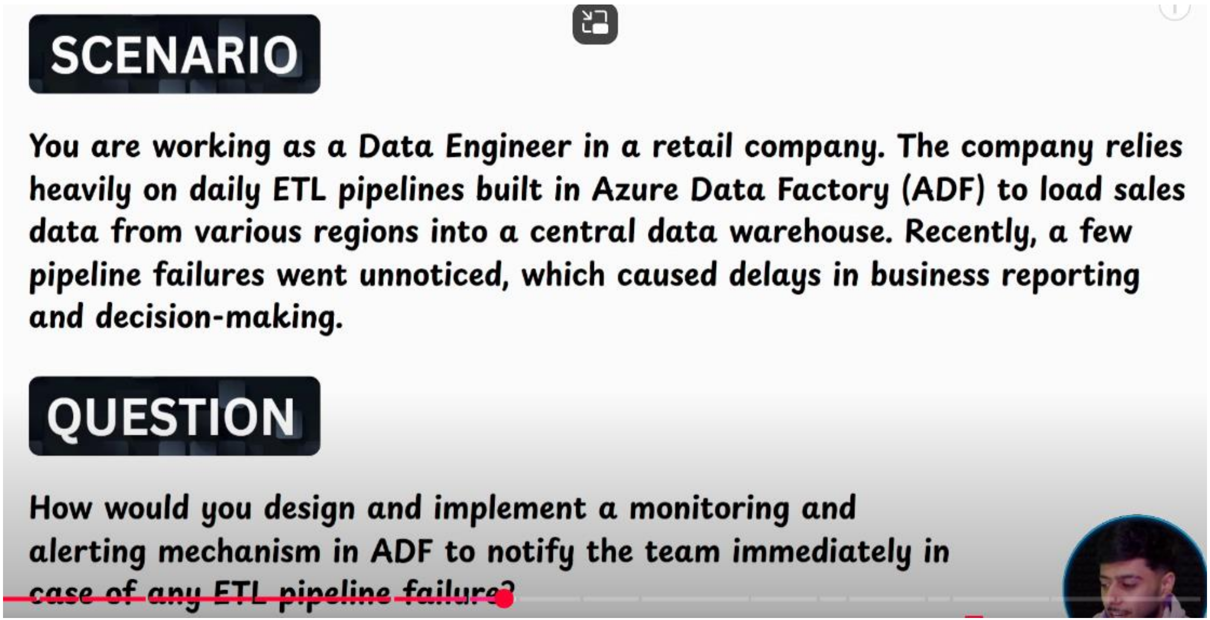
Description

Activity state ● Activated ○ Deactivated

Timeout 0.12:00:00

Retry 3

Retry interval (sec) 30

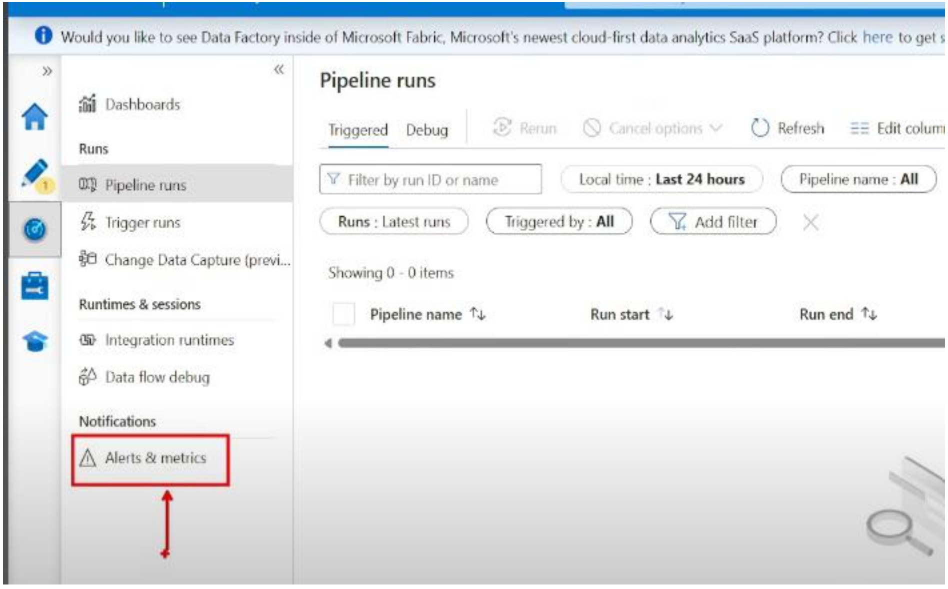


SCENARIO

You are working as a Data Engineer in a retail company. The company relies heavily on daily ETL pipelines built in Azure Data Factory (ADF) to load sales data from various regions into a central data warehouse. Recently, a few pipeline failures went unnoticed, which caused delays in business reporting and decision-making.

QUESTION

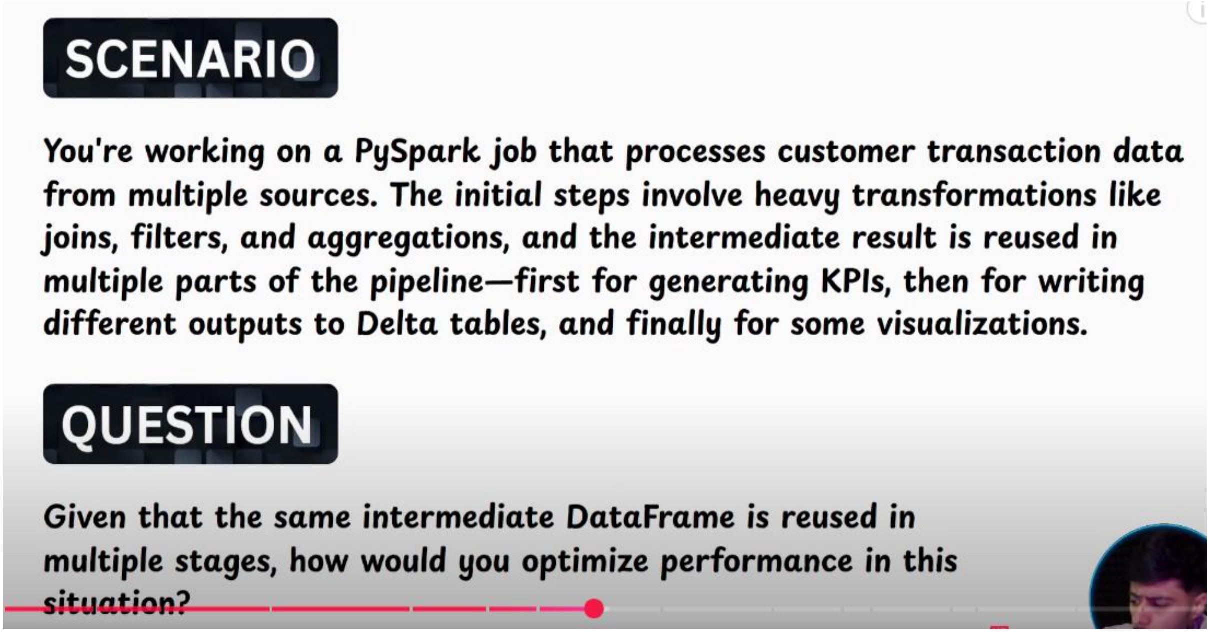
How would you design and implement a monitoring and alerting mechanism in ADF to notify the team immediately in case of any ETL pipeline failure?



Would you like to see Data Factory inside of Microsoft Fabric, Microsoft's newest cloud-first data analytics SaaS platform? Click here to get s

Showing 0 - 0 items

Setup Alerts & metrics



SCENARIO

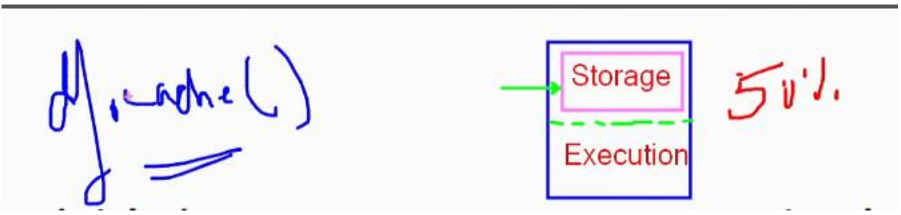
You're working on a PySpark job that processes customer transaction data from multiple sources. The initial steps involve heavy transformations like joins, filters, and aggregations, and the intermediate result is reused in multiple parts of the pipeline—first for generating KPIs, then for writing different outputs to Delta tables, and finally for some visualizations.

QUESTION

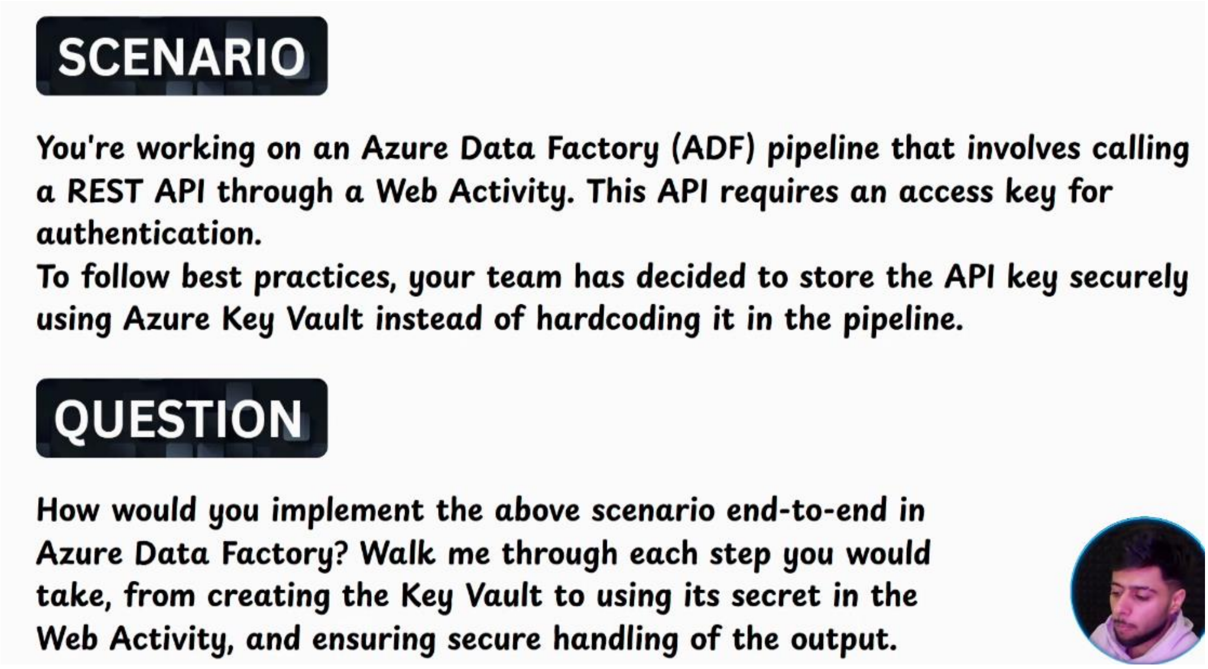
Given that the same intermediate DataFrame is reused in multiple stages, how would you optimize performance in this situation?

Whenever we create a dataframe, executor doesn't hold the dataframe in the execution memory for long time, it will eliminate dataframe immediately if it is assigned a different task or need to create a new df from an existing df.

Use cache() or persist()



Execution memory will be cleaned everytime there is a new job, so it will follow the DAG and start from the first step, so its better to store in storage memory.
Make sure to use unpersist() post completion.



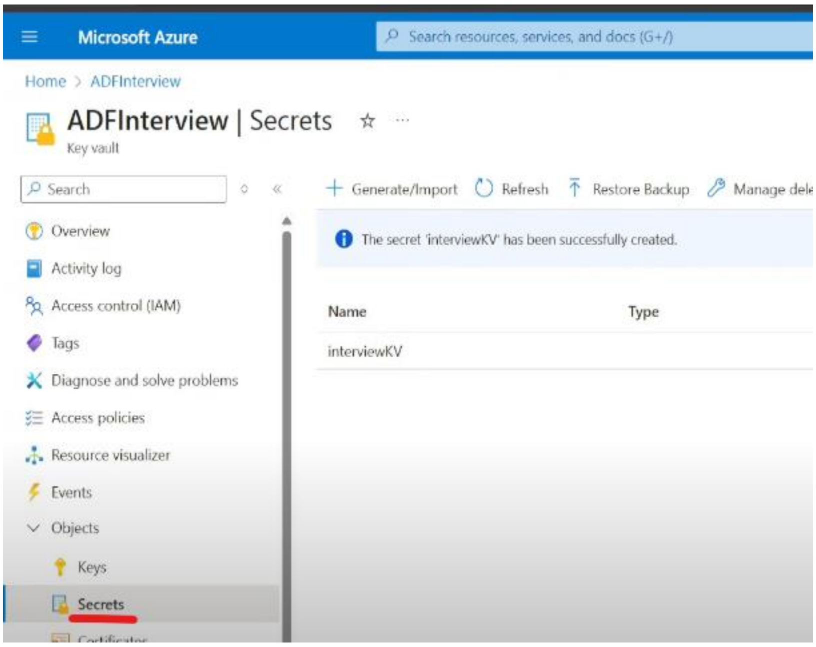
SCENARIO

You're working on an Azure Data Factory (ADF) pipeline that involves calling a REST API through a Web Activity. This API requires an access key for authentication.
To follow best practices, your team has decided to store the API key securely using Azure Key Vault instead of hardcoding it in the pipeline.

QUESTION

How would you implement the above scenario end-to-end in Azure Data Factory? Walk me through each step you would take, from creating the Key Vault to using its secret in the Web Activity, and ensuring secure handling of the output.

Create AKV → Grant KV Administrator role→ create secret



Microsoft Azure

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

Home > ADFInterview

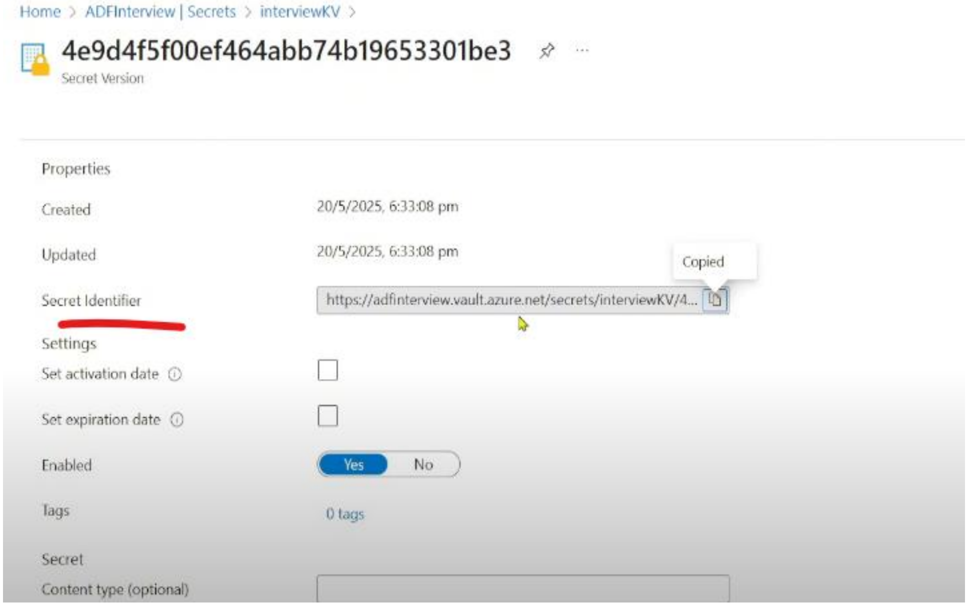
ADFInterview | Secrets

Key vault

The secret 'interviewKV' has been successfully created.

Name	Type
interviewKV	

I want to fetch this info using Web activity



Home > ADFInterview | Secrets > interviewKV >

4e9d4f5f00ef464abb74b19653301be3

Secret Version

Properties

Created	20/5/2025, 6:33:08 pm
Updated	20/5/2025, 6:33:08 pm
Secret Identifier	https://adfinterview.vault.azure.net/secrets/interviewKV/4...

Copied

Settings

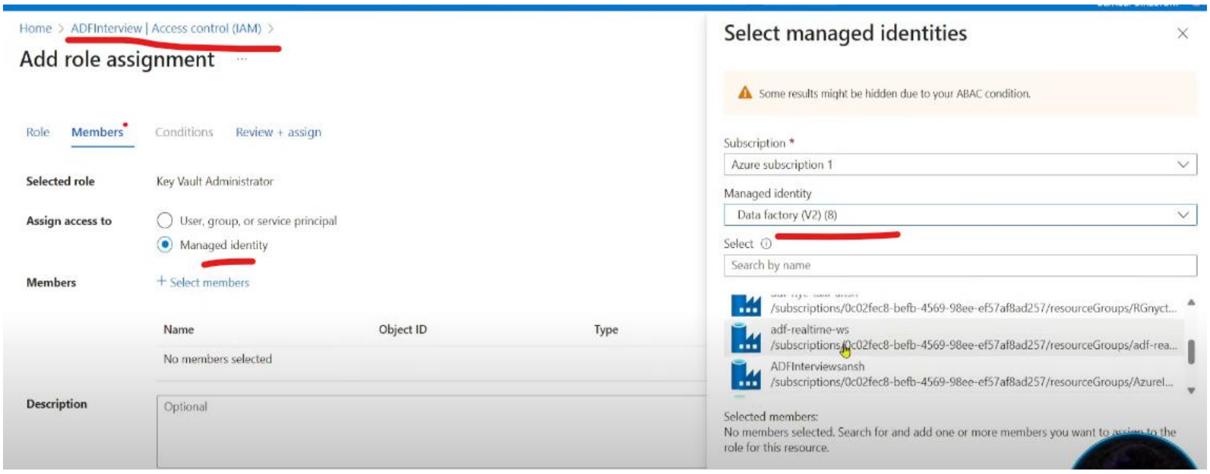
Set activation date	□
Set expiration date	□
Enabled	[✓ Yes] No
Tags	0 tags

Secret

Content type (optional)

Copy the secret identifier

Grant ADF access to AKV (Go to AKV→ IAM → KV Administrator role→ Managed Identity)



Home > ADFInterview | Access control (IAM) >

Add role assignment

Selected role	Key Vault Administrator
Assign access to	○ User, group, or service principal ● Managed identity
Members	+ Select members

Name	Object ID	Type
No members selected		

Description

Optional

Select managed identities

Some results might be hidden due to your ABAC condition.

Subscription *

Azure subscription 1

Managed identity

Data factory (V2) (8)

Select

Search by name

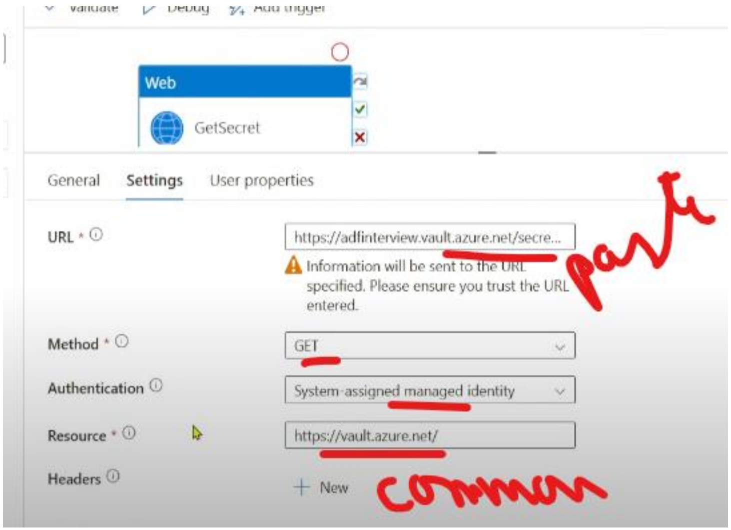
/subscriptions/0c02fec8-befb-4569-98ee-ef57af8ad257/resourceGroups/RGnyct...

adf-realtime-ws
/subscriptions/0c02fec8-befb-4569-98ee-ef57af8ad257/resourceGroups/adf-rea...

ADFInterviewsansh
/subscriptions/0c02fec8-befb-4569-98ee-ef57af8ad257/resourceGroups/Azurel...

Selected members:

No members selected. Search for and add one or more members you want to assign to the role for this resource.



URL * https://adfinterview.vault.azure.net/secre...

⚠ Information will be sent to the URL specified. Please ensure you trust the URL entered.

Method * GET

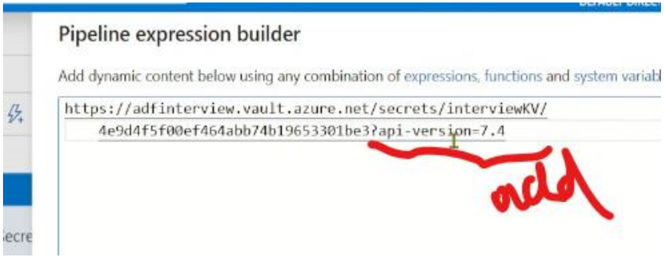
Authentication System-assigned managed identity

Resource * https://vault.azure.net/

Headers + New

part

common

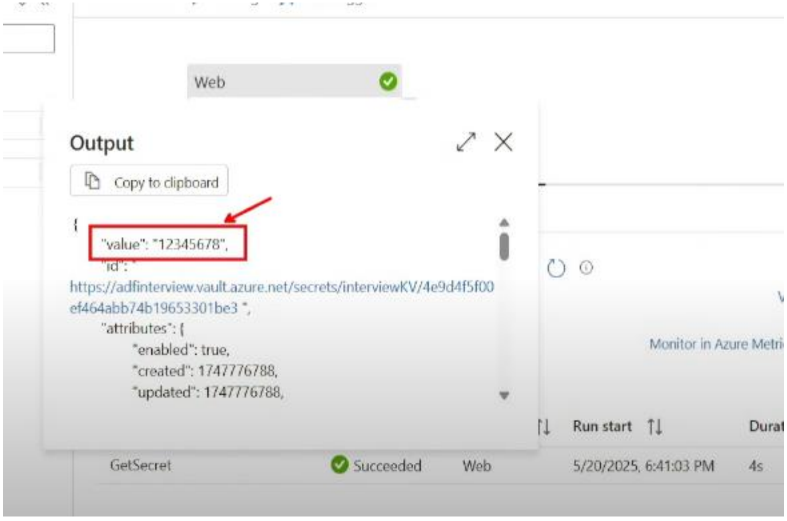


Pipeline expression builder

Add dynamic content below using any combination of expressions, functions and system variabl

```
https://adfinterview.vault.azure.net/secrets/interviewKV/
    4e9d4f5f00ef464abb74b19653301be3?api-version=7.4
```

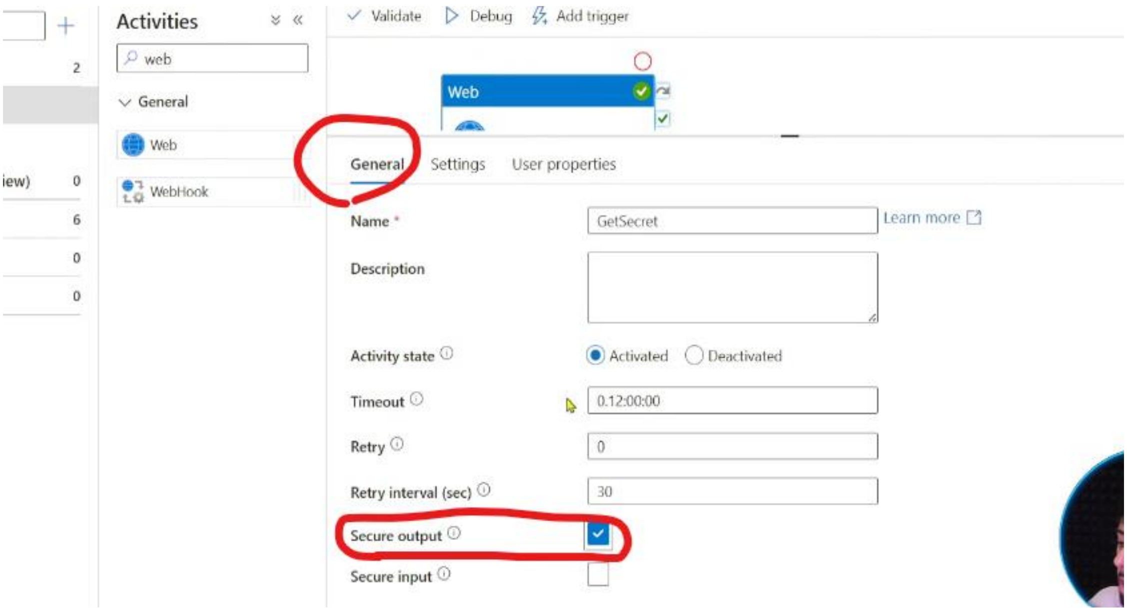
add



```
{
    "value": "12345678",
    "id": "
https://adfinterview.vault.azure.net/secrets/interviewKV/4e9d4f5f00
ef464abb74b19653301be3 ",
    "attributes": {
        "enabled": true,
        "created": 1747776788,
        "updated": 1747776788,
```

			Run start	Durat
GetSecret	Succeeded	Web	5/20/2025, 6:41:03 PM	4s

We can see the value, ideally we should not



Name * GetSecret

Description

Activity state ● Activated ○ Deactivated

Timeout 0.12:00:00

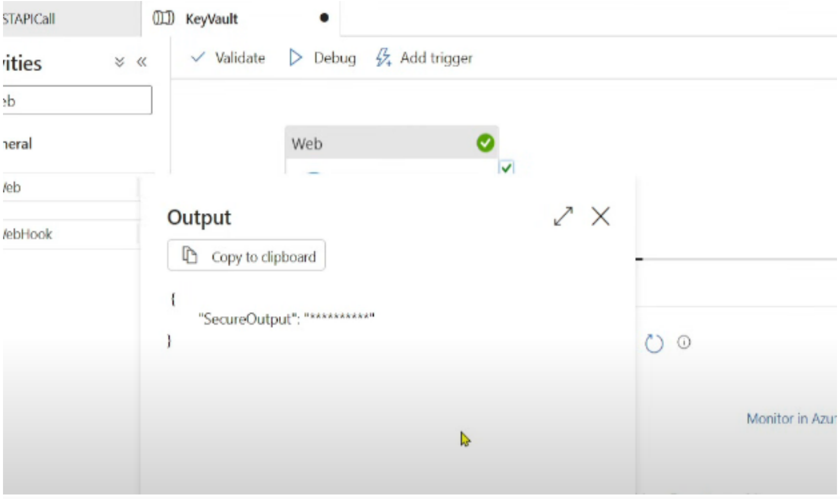
Retry 0

Retry interval (sec) 30

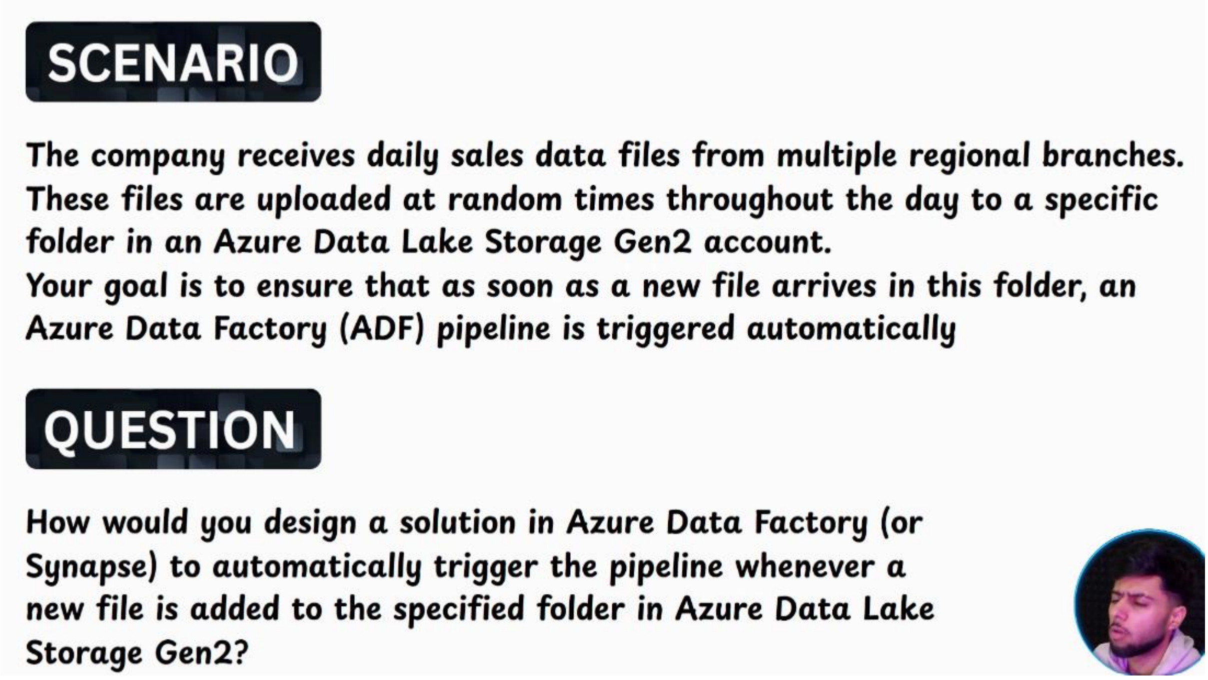
Secure output ☑

Secure input □

Enable secure output



```
{
    "SecureOutput": "**********"
}
```



SCENARIO

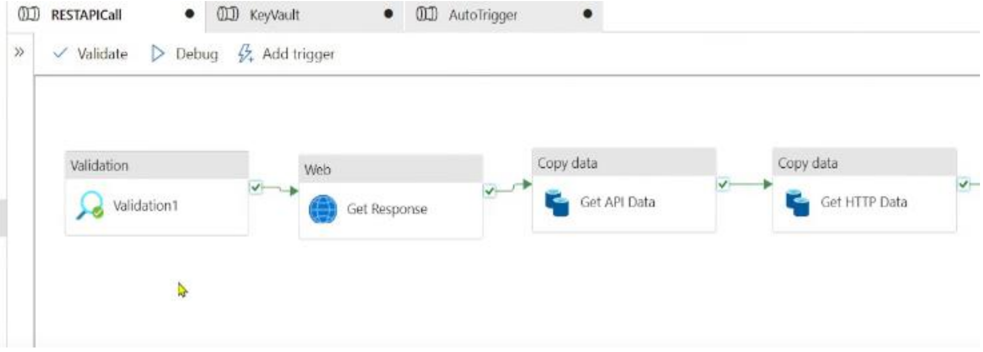
The company receives daily sales data files from multiple regional branches. These files are uploaded at random times throughout the day to a specific folder in an Azure Data Lake Storage Gen2 account.

Your goal is to ensure that as soon as a new file arrives in this folder, an Azure Data Factory (ADF) pipeline is triggered automatically

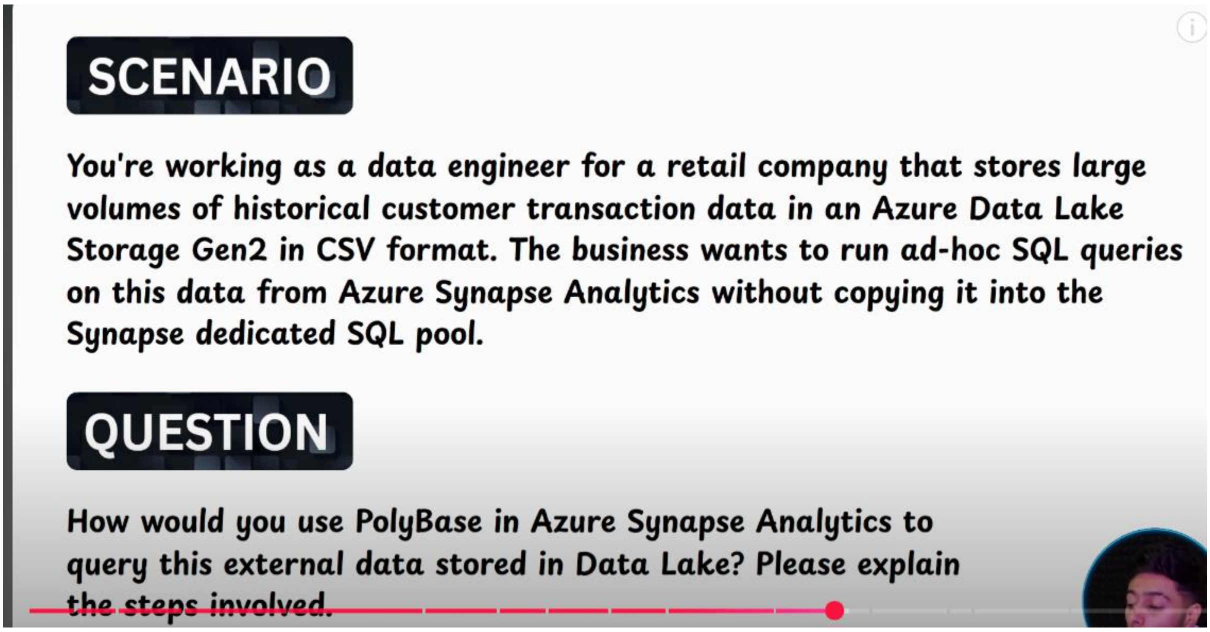
QUESTION

How would you design a solution in Azure Data Factory (or Synapse) to automatically trigger the pipeline whenever a new file is added to the specified folder in Azure Data Lake Storage Gen2?

M1 : Use Storage Events Trigger
M2 : Use Validation activity



Validation activity will be continuously searching for the file and then run the subsequent steps



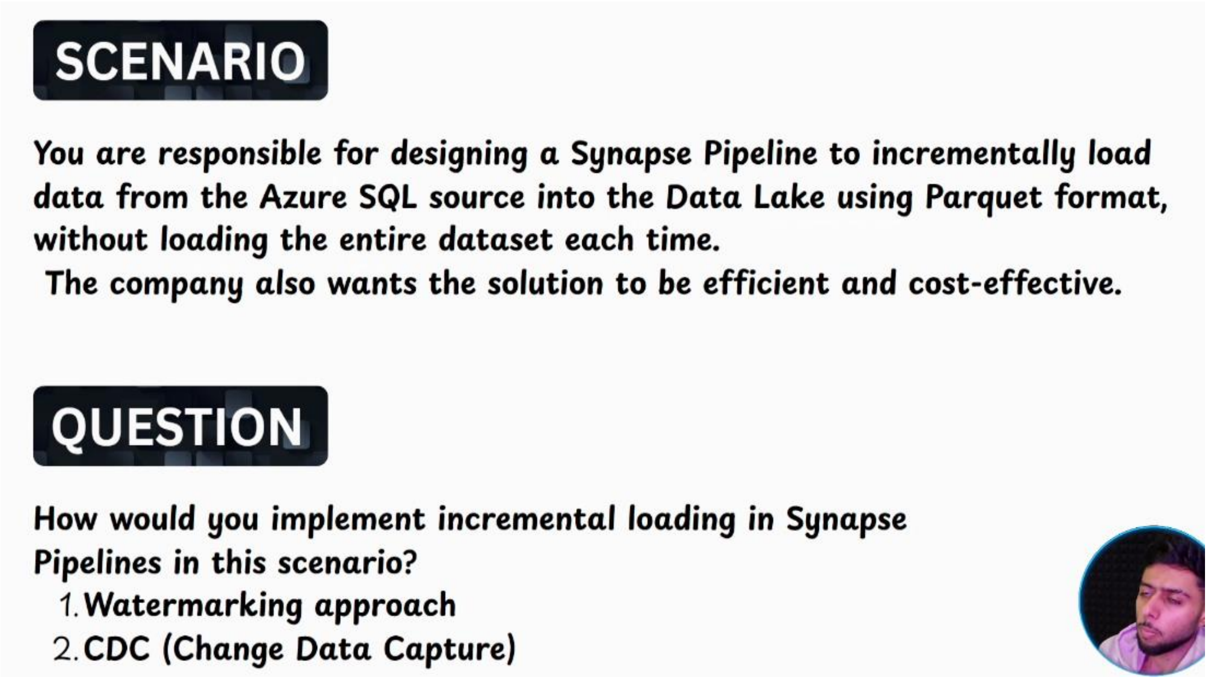
SCENARIO

You're working as a data engineer for a retail company that stores large volumes of historical customer transaction data in an Azure Data Lake Storage Gen2 in CSV format. The business wants to run ad-hoc SQL queries on this data from Azure Synapse Analytics without copying it into the Synapse dedicated SQL pool.

QUESTION

How would you use PolyBase in Azure Synapse Analytics to query this external data stored in Data Lake? Please explain the steps involved.

In serverless SQL pool, we use Polybase
Step 1: Create Master key for database
Step 2: Create credential
Step 3: Create External Data Source
Step 4: Create External File Format
Step 5: Create External Table



SCENARIO

You are responsible for designing a Synapse Pipeline to incrementally load data from the Azure SQL source into the Data Lake using Parquet format, without loading the entire dataset each time.

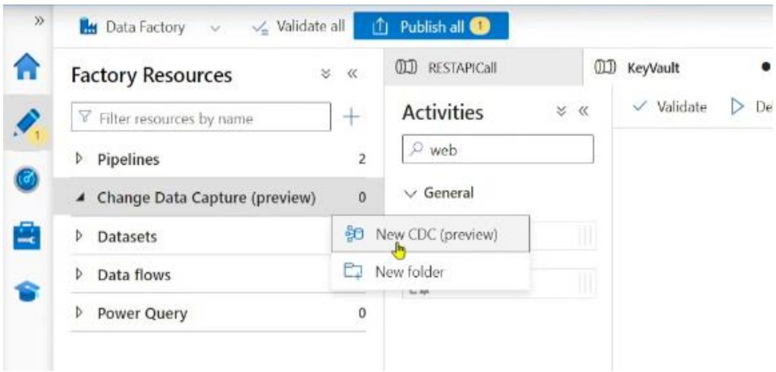
The company also wants the solution to be efficient and cost-effective.

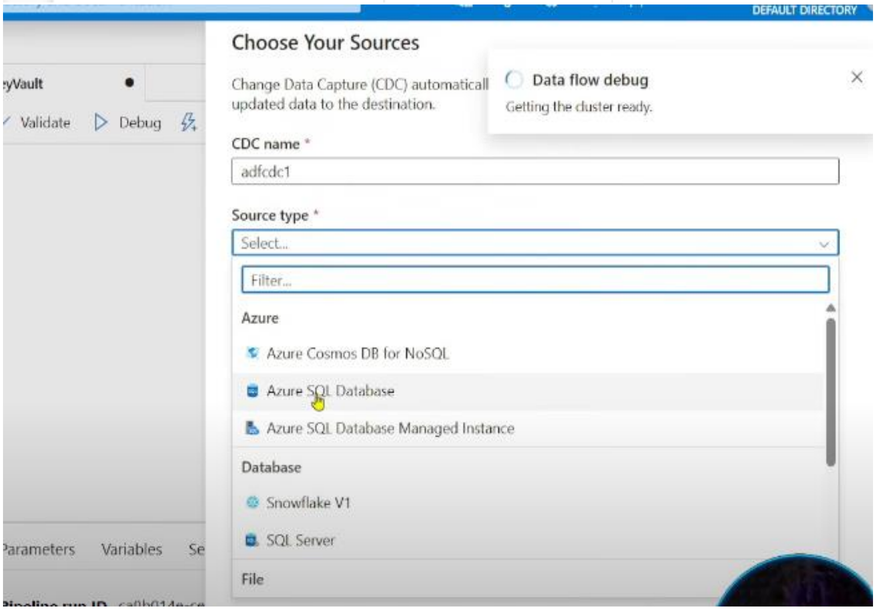
QUESTION

How would you implement incremental loading in Synapse Pipelines in this scenario?

1. Watermarking approach
2. CDC (Change Data Capture)

CDC as of now works when we have SQL db as a source





Choose Your Sources

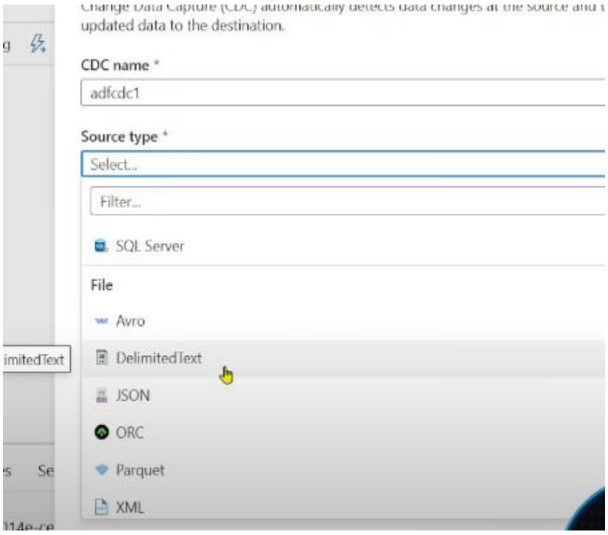
Change Data Capture (CDC) automaticall updated data to the destination.

Data flow debug

Getting the cluster ready.

CDC name *: adfcdc1

Source type *: Select...



Change Data Capture (CDC) automatically detects data changes at the source and t
updated data to the destination.

CDC name *

adfcdc1

Source type *

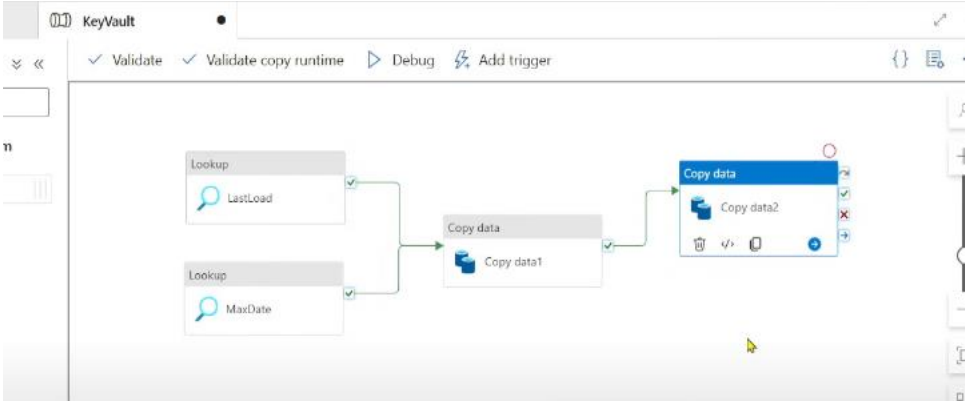
Select...

Enable CDC in the DB

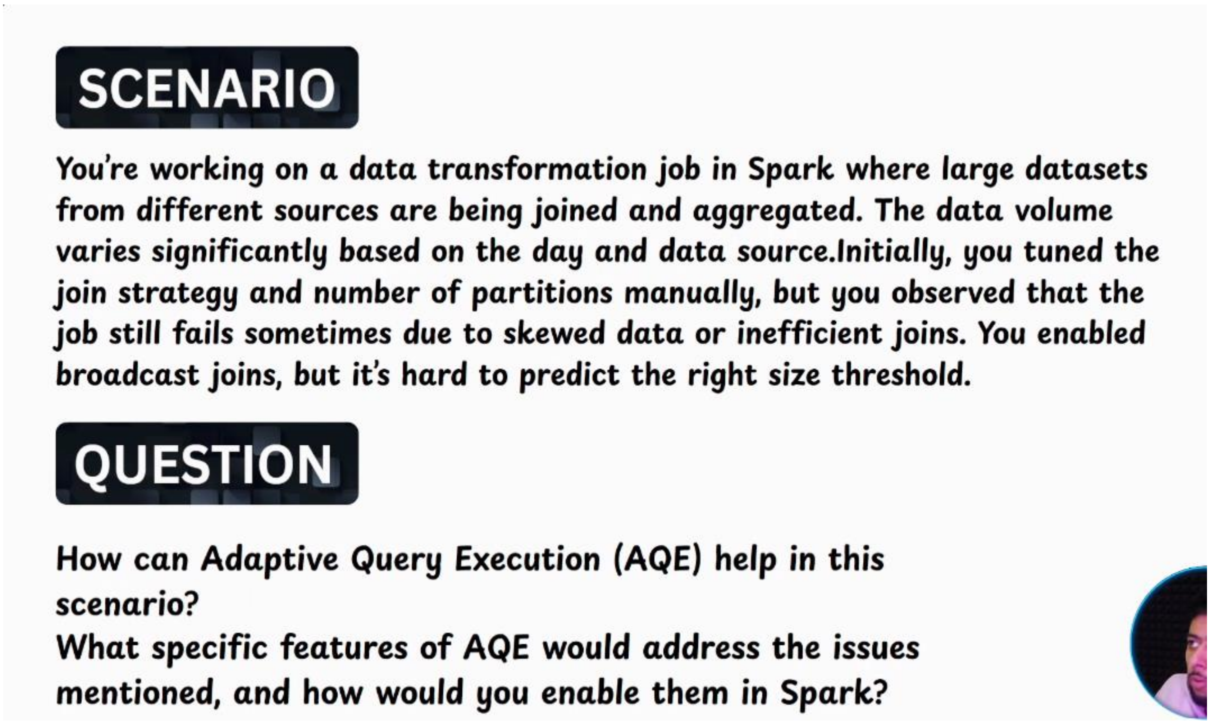
Enable CDC in the Table

Connect the Table

Watermarking



Instead of Stored Procedure, store the data in file



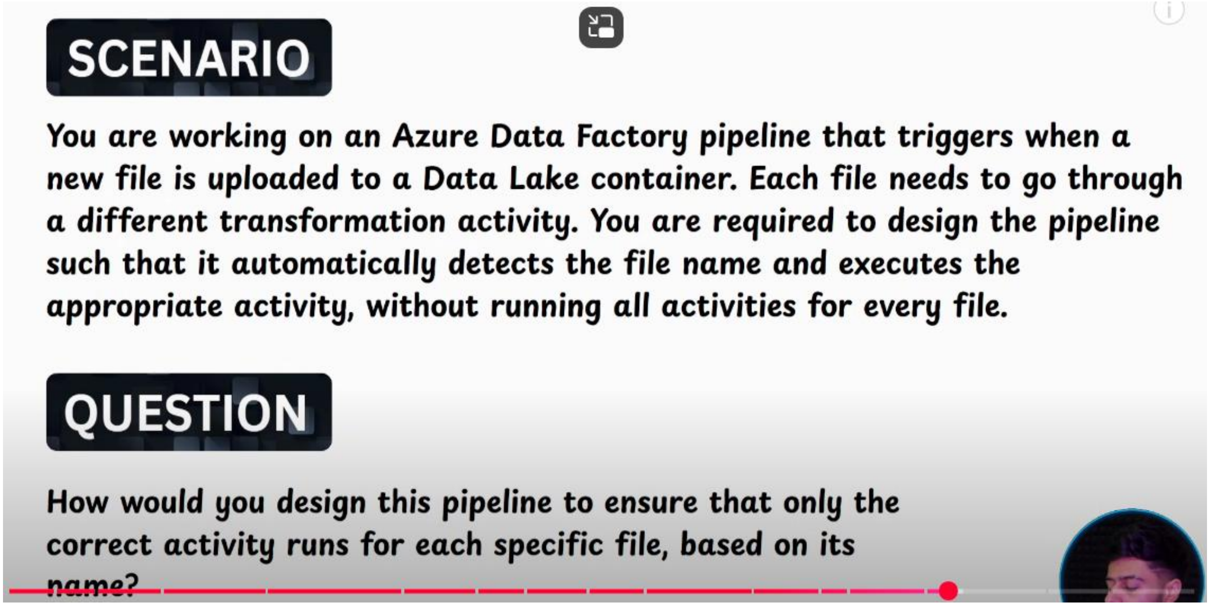
SCENARIO

You're working on a data transformation job in Spark where large datasets from different sources are being joined and aggregated. The data volume varies significantly based on the day and data source.Initially, you tuned the join strategy and number of partitions manually, but you observed that the job still fails sometimes due to skewed data or inefficient joins. You enabled broadcast joins, but it's hard to predict the right size threshold.

QUESTION

How can Adaptive Query Execution (AQE) help in this scenario?
What specific features of AQE would address the issues mentioned, and how would you enable them in Spark?

AQE : calculates query statistics during runtime based on that it addresses the manual threshold for broadcast join
(Before AQE, LP→PP) With AQE, LP→PP→Query statistics
Addresses skewness, Dynamically coalesces the partitions

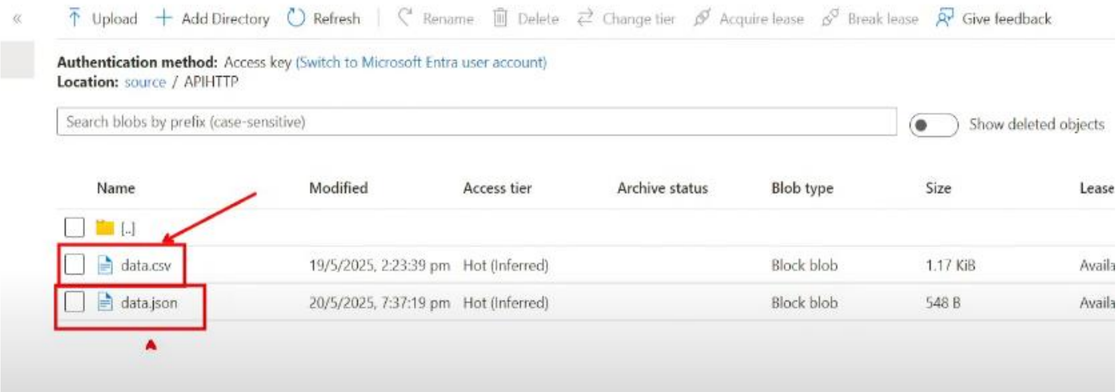


SCENARIO

You are working on an Azure Data Factory pipeline that triggers when a new file is uploaded to a Data Lake container. Each file needs to go through a different transformation activity. You are required to design the pipeline such that it automatically detects the file name and executes the appropriate activity, without running all activities for every file.

QUESTION

How would you design this pipeline to ensure that only the correct activity runs for each specific file, based on its name?



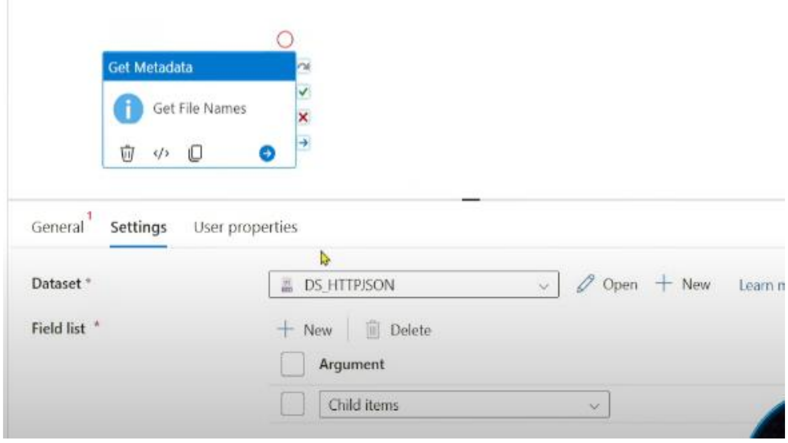
Authentication method: Access key (Switch to Microsoft Entra user account)
Location: source / APIHTTP

Search blobs by prefix (case-sensitive)

Show deleted objects

Name	Modified	Access tier	Archive status	Blob type	Size	Lease
□ [..]						
□ data.csv	19/5/2025, 2:23:39 pm	Hot (Inferred)		Block blob	1.17 KiB	Availa
□ data.json	20/5/2025, 7:37:19 pm	Hot (Inferred)		Block blob	548 B	Availa

Want to perform different sets of activities for each of these files

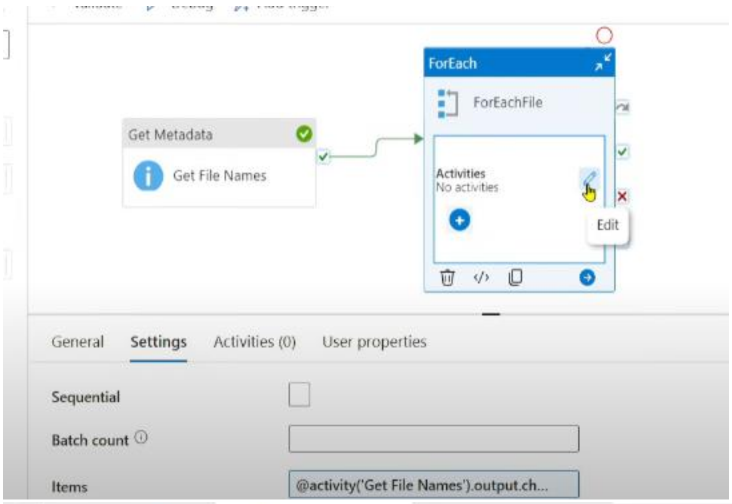


Dataset * DS_HTTPJSON

Field list *

□ Argument

□ Child items

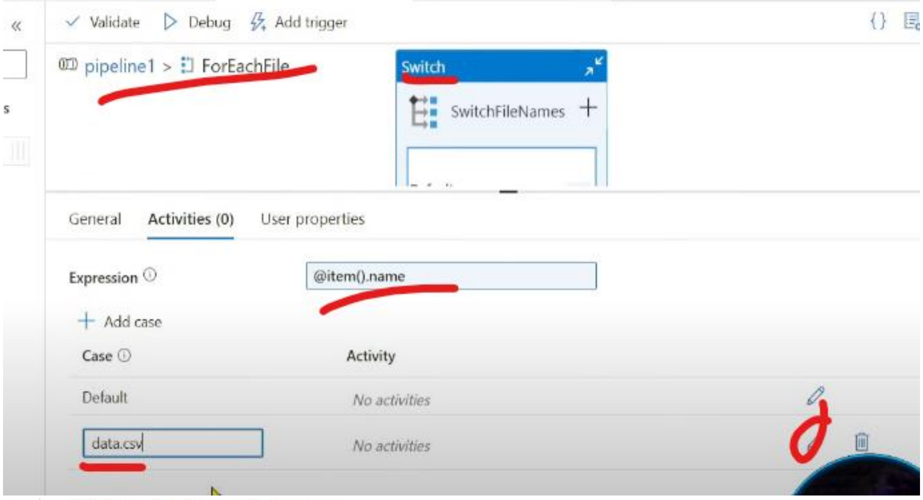


General | **Settings** | Activities (0) | User properties

Sequential □

Batch count

Items @activity('Get File Names').output.ch...

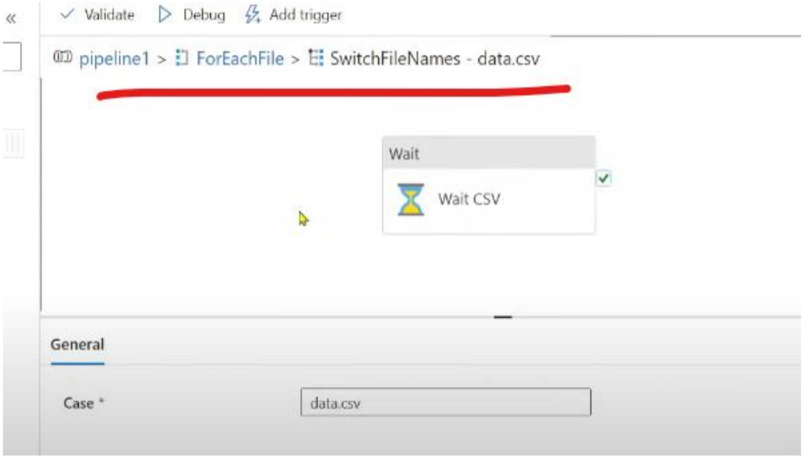


General | **Activities (0)** | User properties

Expression @item().name

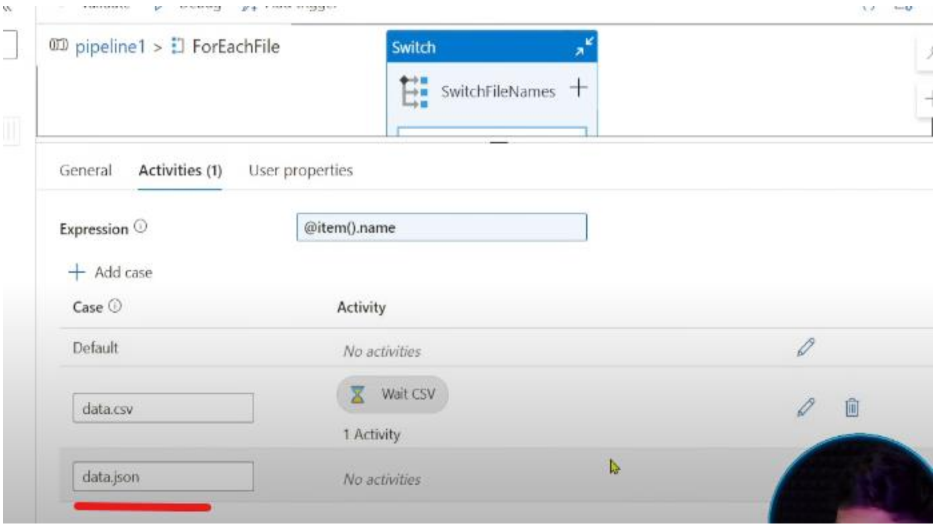
+ Add case

Case	Activity
Default	*No activities*
data.csv	*No activities*



General

Case * data.csv



pipeline1 > ForEachFile

Switch

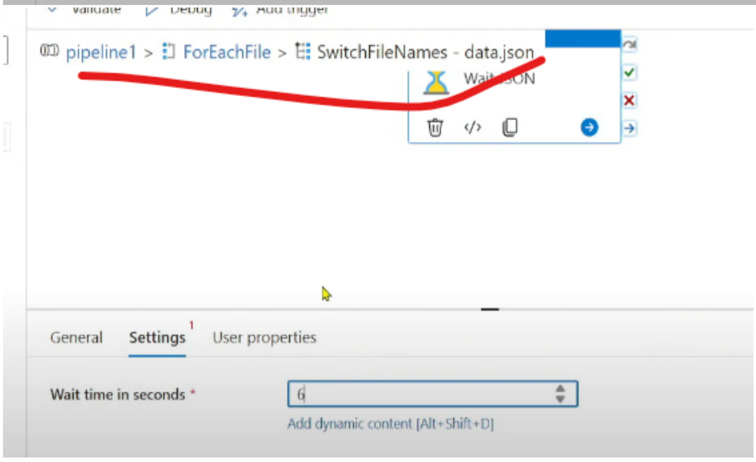
SwitchFileNames

General | Activities (1) | User properties

Expression: @item().name

+ Add case

Case	Activity
Default	No activities
data.csv	Wait CSV 1 Activity
data.json	No activities



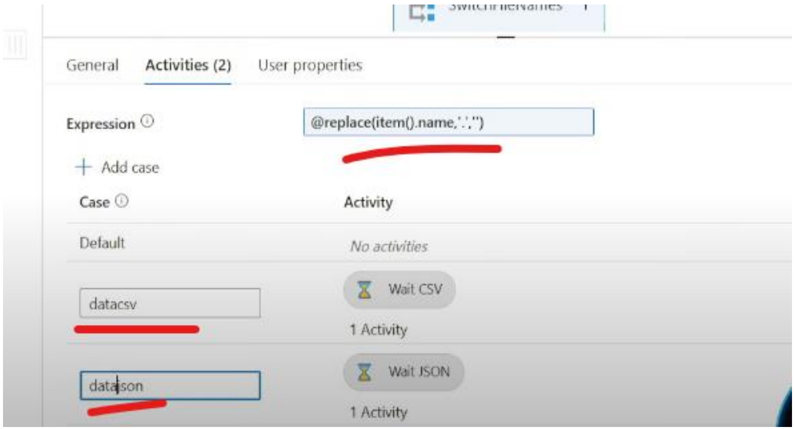
pipeline1 > ForEachFile > SwitchFileNames - data.json

General | Settings | User properties

Wait time in seconds *: 6

Add dynamic content [Alt+Shift+D]

We cant use . in case conditions

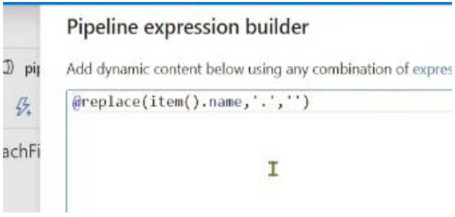


General | Activities (2) | User properties

Expression: @replace(item().name,'.','')

+ Add case

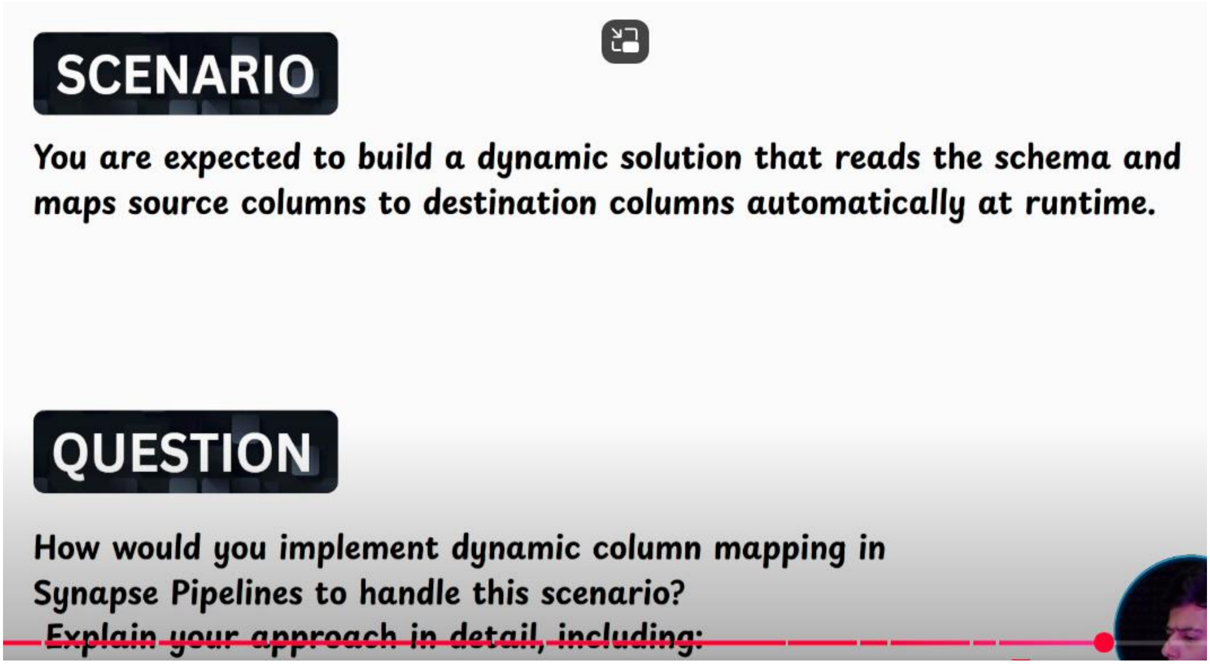
Case	Activity
Default	No activities
datacsv	Wait CSV 1 Activity
datajson	Wait JSON 1 Activity



Pipeline expression builder

Add dynamic content below using any combination of expres

```
@replace(item().name,'.','')
```



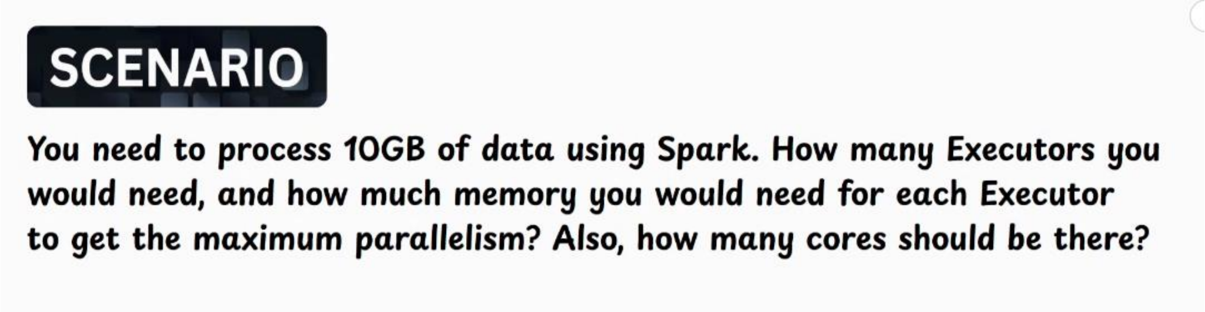
SCENARIO

You are expected to build a dynamic solution that reads the schema and maps source columns to destination columns automatically at runtime.

QUESTION

How would you implement dynamic column mapping in Synapse Pipelines to handle this scenario?
Explain your approach in detail, including:

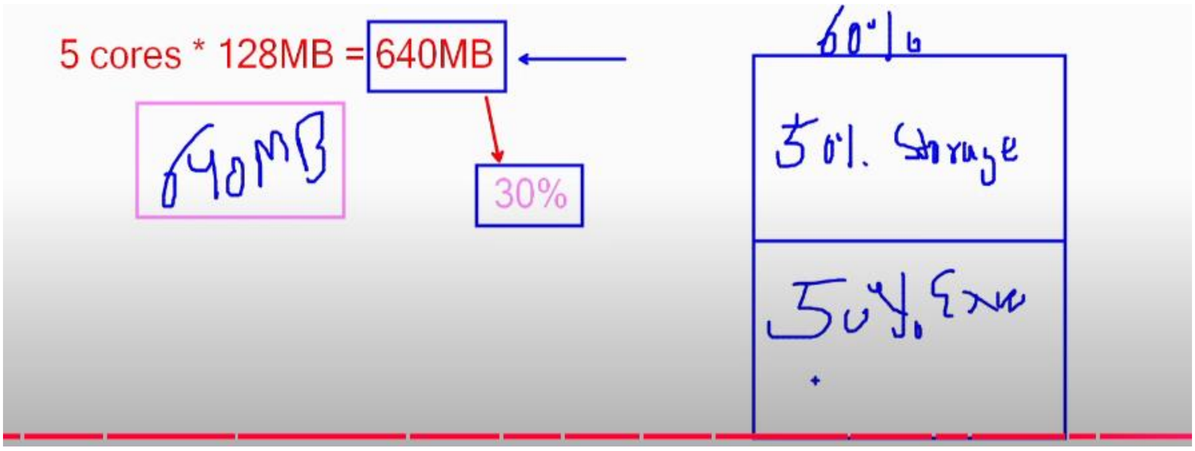
Copy the translator code



SCENARIO

You need to process 10GB of data using Spark. How many Executors you would need, and how much memory you would need for each Executor to get the maximum parallelism? Also, how many cores should be there?

S1: 10GB → 80 partitions
S2: Leverage max parallelism→ 80 partitions → 80 cores
S3: No of cores→ 5 per executor → 80/5 → 16 executor
S4: Memory for each Executor → In each executor 5 cores, each core will process 1 partition sized 128MB i.e. 128*5 = 640MB
640MB is not the total memory, 640MB is required just to execute the task, we get 30% memory to execute the task.



Spark pool memory is 60%, 40% off heap.
Out of 60%, 50% is storage memory/caching & 50% is execution memory.

So 640MB is equivalent to 50% execution memory.
So we multiply 640 with 3.5-4 to get the total executor memory

640MB
3.5-4
= 2.6GB

Total Executor memory = 2.6 GB