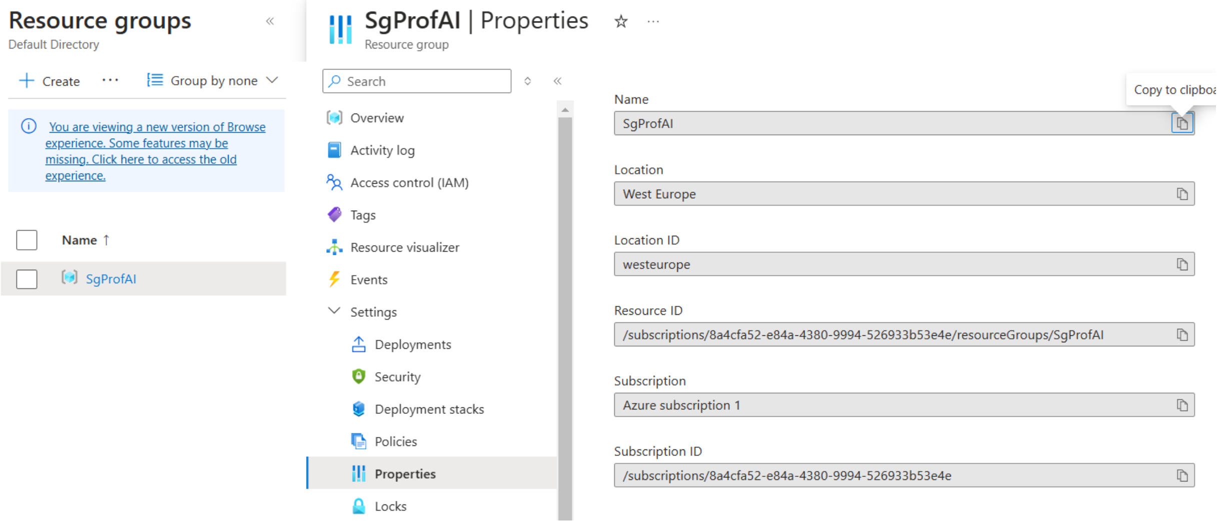
Exercise aim

The aim of this exercise is to construct a pipeline using Azure Data Factory that enables a CSV file to be uploaded to Azure Blob Storage, cleaned up and the columns remapped in Italian. It is essential that the process maintains only the columns 'Films', 'Genres' and 'Ratings', filters out films with ratings higher than 7 and ensures efficient and parallel data transfer. Furthermore, it is crucial to retain the metadata for any additional useful information.

Exercise performance

1. Setting up the Resource Group, Managed Identity and Key Vault

Once the subscription had been activated, a Resource Group was created by entering the desired subscription and location (Fig. 1). This was followed by the creation of a Managed Identity and Key Vault.
A Managed Identity was created with the objective of facilitating the management of access to Azure services (Fig. 2). A Key Vault was configured, granting access to the Managed Identity via access policies and setting a key for secure access to resources (Fig. 3).
The Managed Identity was assigned the role of 'Contributor' in the subscription, thereby ensuring access and management of all resources (Fig. 4).



Name
SgProfAI

Location
West Europe

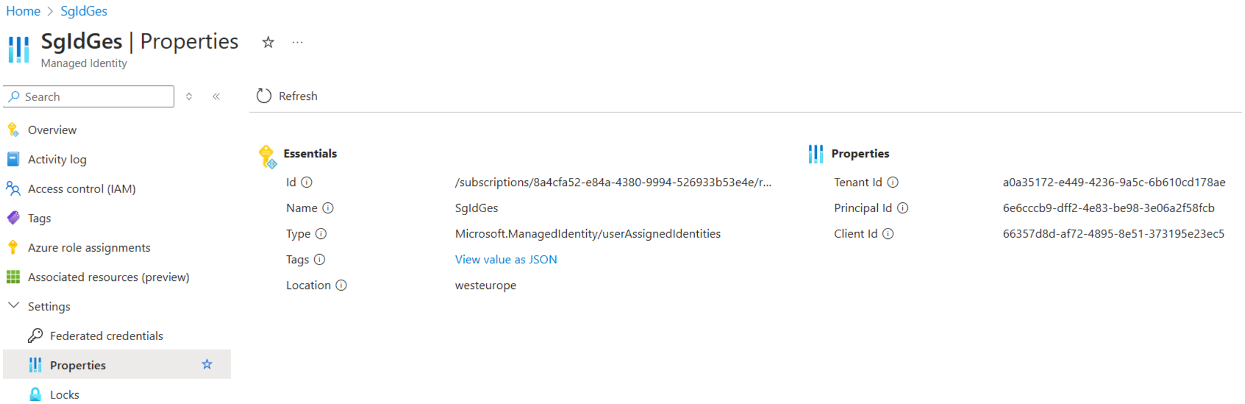
Location ID
westeurope

Resource ID
/subscriptions/8a4cfa52-e84a-4380-9994-526933b53e4e/resourceGroups/SgProfAI

Subscription
Azure subscription 1

Subscription ID
/subscriptions/8a4cfa52-e84a-4380-9994-526933b53e4e

Fig. 1 – Resource Group properties.



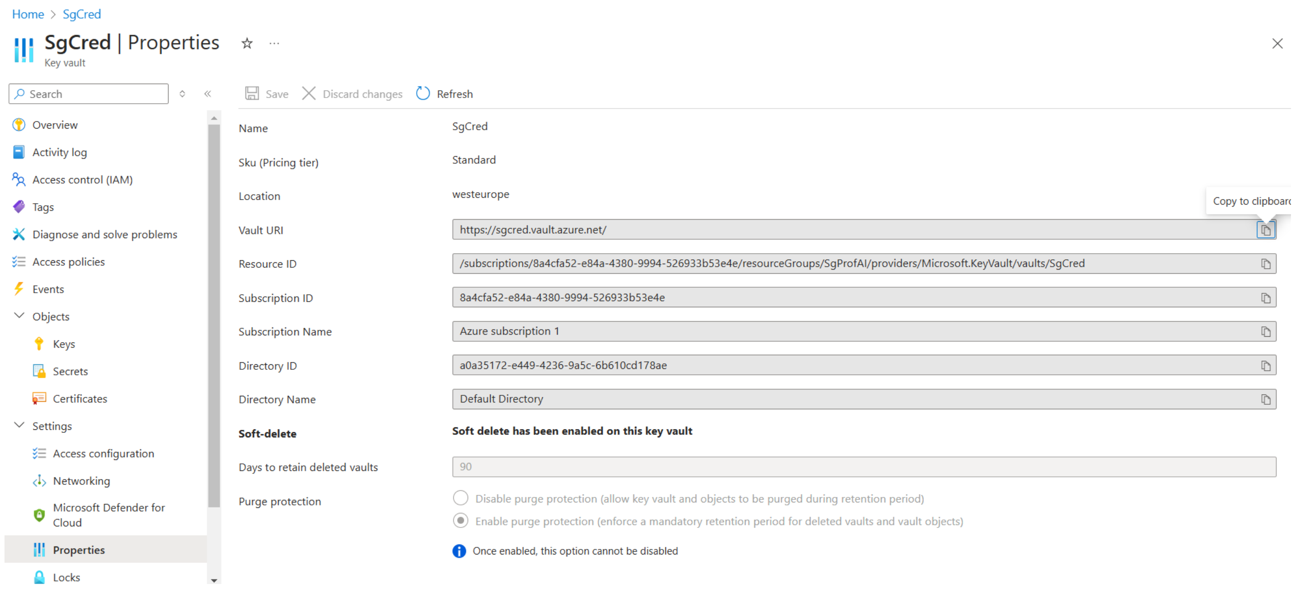
Essentials

Id	/subscriptions/8a4cfa52-e84a-4380-9994-526933b53e4e/r...
Name	SgIdGes
Type	Microsoft.ManagedIdentity/userAssignedIdentities
Tags	View value as JSON
Location	westeurope

Properties

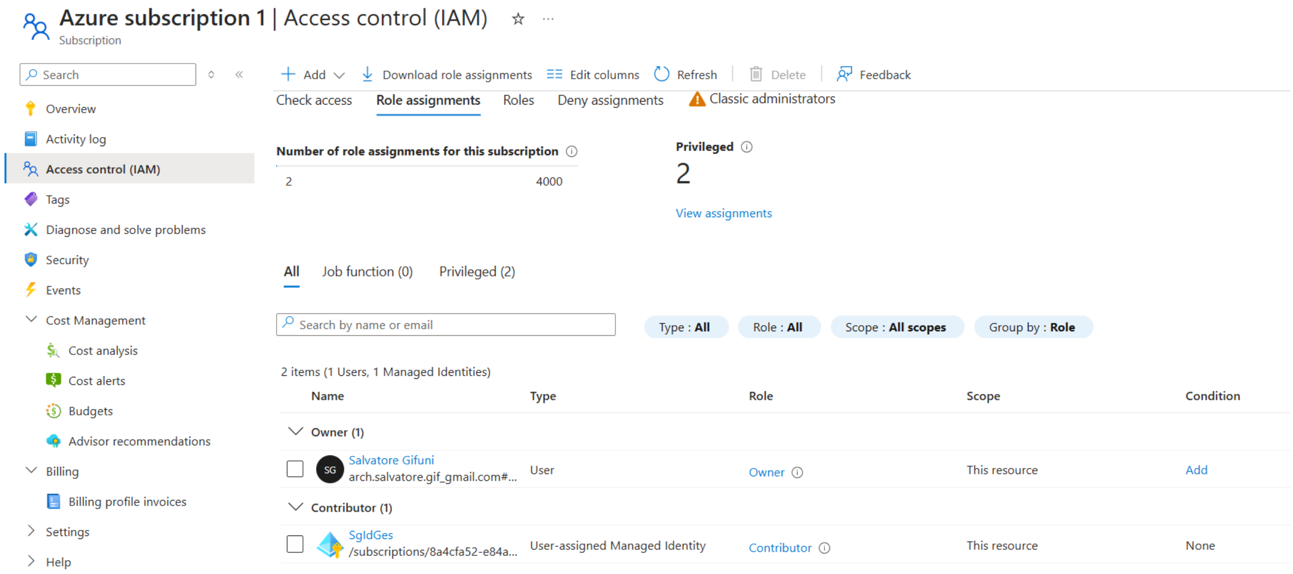
Tenant Id	a0a35172-e449-4236-9a5c-6b610cd178ae
Principal Id	6e6cccb9-dff2-4e83-be98-3e06a2f58fcb
Client Id	66357d8d-af72-4895-8e51-373195e23ec5

Fig. 2 - Managed Identity properties.



Name: SgCred
Sku (Pricing tier): Standard
Location: westeurope
Vault URI: https://sgcred.vault.azure.net/
Resource ID: /subscriptions/8a4cfa52-e84a-4380-9994-526933b53e4e/resourceGroups/SgProfAI/providers/Microsoft.KeyVault/vaults/SgCred
Subscription ID: 8a4cfa52-e84a-4380-9994-526933b53e4e
Subscription Name: Azure subscription 1
Directory ID: a0a35172-e449-4236-9a5c-6b610cd178ae
Directory Name: Default Directory
Soft-delete: **Soft delete has been enabled on this key vault**
Days to retain deleted vaults: 90
Purge protection:
○ Disable purge protection (allow key vault and objects to be purged during retention period)
● Enable purge protection (enforce a mandatory retention period for deleted vaults and vault objects)
Once enabled, this option cannot be disabled

Fig. 3 - Key Vault properties.



Number of role assignments for this subscription: 2 / 4000
Privileged: 2

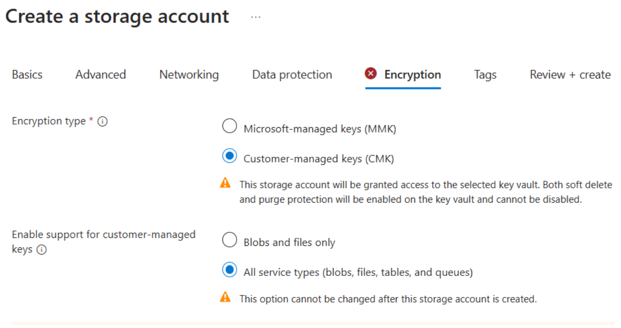
2 items (1 Users, 1 Managed Identities)

Name	Type	Role	Scope	Condition
Owner (1)				
Salvatore Gifuni arch.salvatore.gif_gmail.com#...	User	Owner	This resource	Add
Contributor (1)				
SgIdGes /subscriptions/8a4cfa52-e84a...	User-assigned Managed Identity	Contributor	This resource	None

Fig. 4 - Setting the 'Contributor' role to the Managed Identity in the subscription.

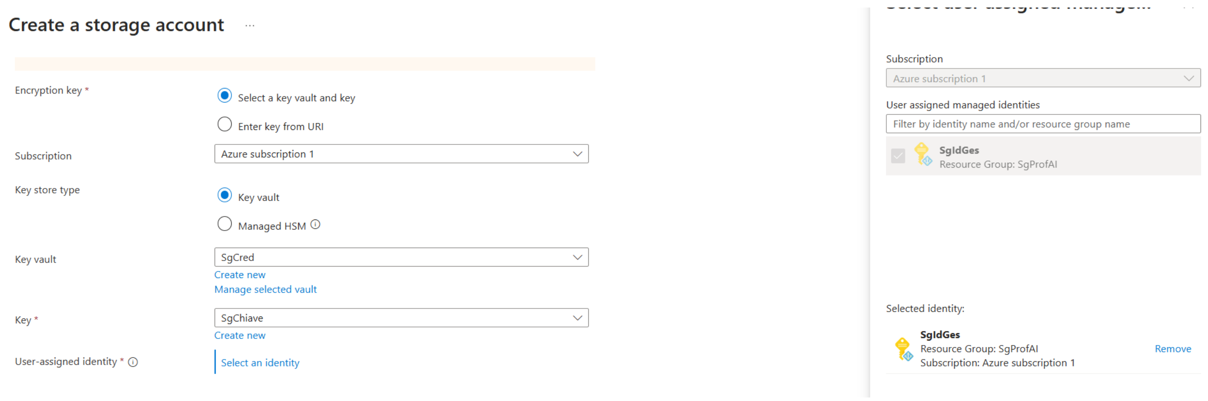
2. Azure Storage Account creation

A storage account was created, paying particular attention to the security settings in the 'Encryption' section (Fig. 5).
Three containers were created: one for the input file, one for files with intermediate changes and one for the final file.



Encryption type *:
○ Microsoft-managed keys (MMK)
● Customer-managed keys (CMK)
This storage account will be granted access to the selected key vault. Both soft delete and purge protection will be enabled on the key vault and cannot be disabled.

Enable support for customer-managed keys:
○ Blobs and files only
● All service types (blobs, files, tables, and queues)
This option cannot be changed after this storage account is created.



Create a storage account

Encryption key *
● Select a key vault and key
○ Enter key from URI

Subscription: Azure subscription 1

Key store type
● Key vault
○ Managed HSM

Key vault: SgCred
Create new
Manage selected vault

Key *: SgChiave
Create new

User-assigned identity *: Select an identity

Subscription: Azure subscription 1

User assigned managed identities: Filter by identity name and/or resource group name

☑ SgIdGes
Resource Group: SgProfAI

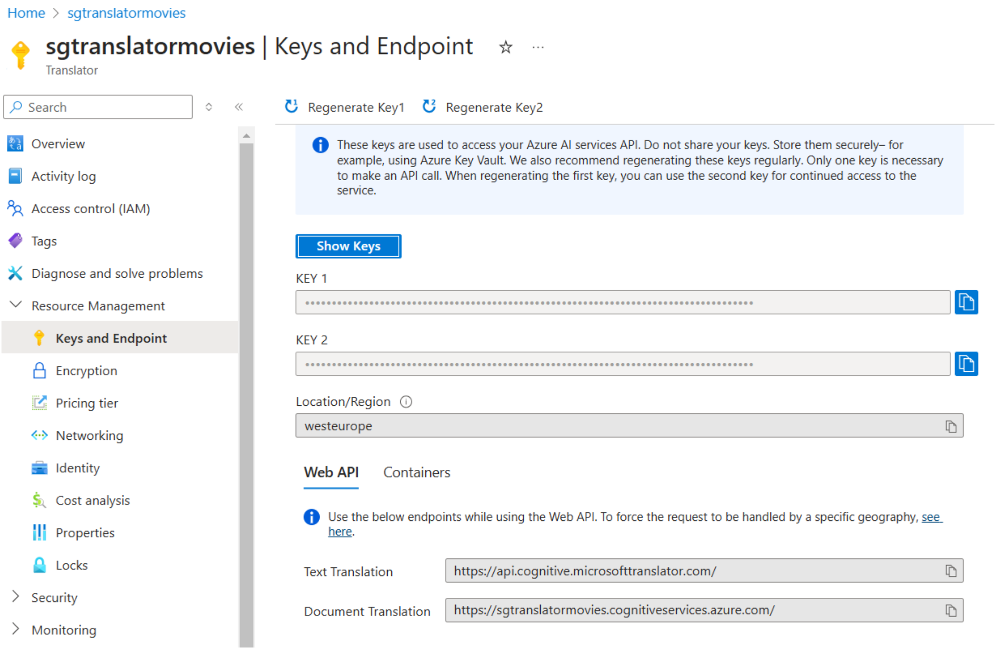
Selected identity:

SgIdGes
Resource Group: SgProfAI
Subscription: Azure subscription 1
Remove

Fig. 5 - Encryption' settings for storage account creation.

3. Creating the Azure Translator

A translator was created, keeping the keys and URL for use in the pipeline web activity (Fig. 6).



Home > sgtranslatormovies

sgtranslatormovies | Keys and Endpoint
Translator

These keys are used to access your Azure AI services API. Do not share your keys. Store them securely– for example, using Azure Key Vault. We also recommend regenerating these keys regularly. Only one key is necessary to make an API call. When regenerating the first key, you can use the second key for continued access to the service.

KEY 1: ••

KEY 2: ••

Location/Region: westeurope

Web API | Containers

Use the below endpoints while using the Web API. To force the request to be handled by a specific geography, see here.

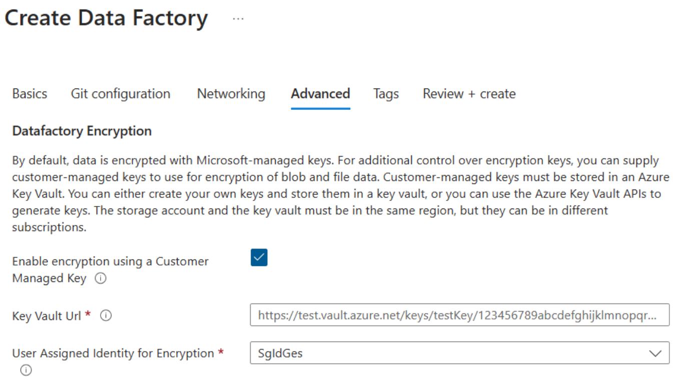
Text Translation: https://api.cognitive.microsofttranslator.com/

Document Translation: https://sgtranslatormovies.cognitiveservices.azure.com/

Fig. 6 – Page containing keys and URLs to be passed to the pipeline web activity.

4. Creation of the Azure Data Factory

A Data Factory has been created. All previously configured security settings have been entered in the 'Advanced' section (Fig. 7).



Create Data Factory

Datafactory Encryption

By default, data is encrypted with Microsoft-managed keys. For additional control over encryption keys, you can supply customer-managed keys to use for encryption of blob and file data. Customer-managed keys must be stored in an Azure Key Vault. You can either create your own keys and store them in a key vault, or you can use the Azure Key Vault APIs to generate keys. The storage account and the key vault must be in the same region, but they can be in different subscriptions.

Enable encryption using a Customer Managed Key: ☑

Key Vault Url *: https://test.vault.azure.net/keys/testKey/123456789abcdefghijklmnopqr...

User Assigned Identity for Encryption *: SgIdGes

Fig. 7 – Data Factory Encryption Settings

5. Creation of the pipeline

The pipeline created (Fig. 8) takes the original dataset as input via a data flow - 'DataWrangling' (Fig. 9) - and selects only the columns 'Movies'[1], 'Genres' and 'Ratings'. Genres' and 'Ratings'[2] columns, filters the films, keeping only those with a rating higher than 7[3] (Fig. 11), and sorts them in descending order (optional). The resulting file is saved in an intermediate file for subsequent pipeline activities.

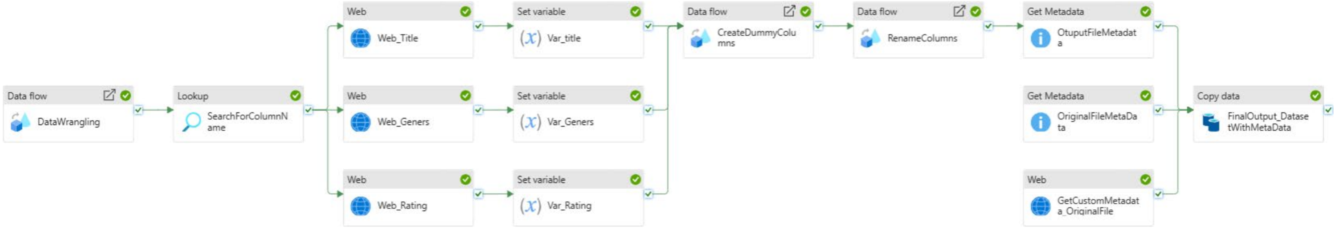


Fig. 8 - Complete Pipeline.

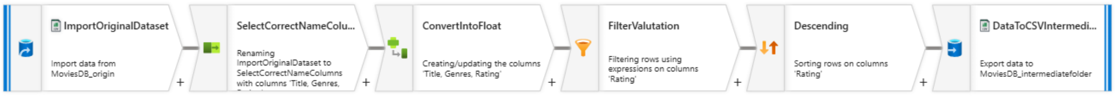
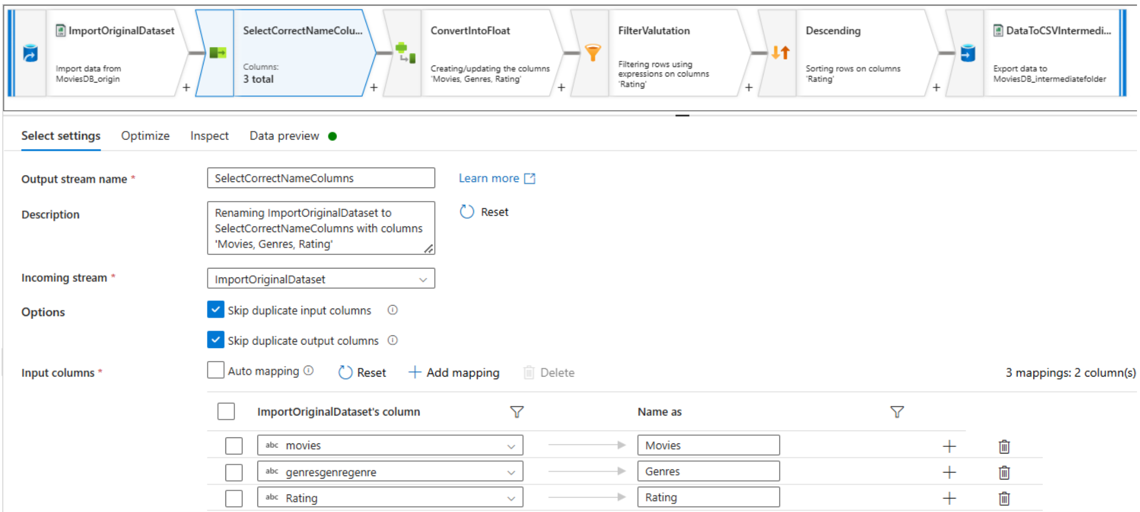

ImportOriginalDataset
Import data from MoviesDB_origin
SelectCorrectNameColu...
Renaming ImportOriginalDataset to SelectCorrectNameColumns with columns 'Title, Genres,
ConvertIntoFloat
Creating/updating the columns 'Title, Genres, Rating'
FilterValutation
Filtering rows using expressions on columns 'Rating'
Descending
Sorting rows on columns 'Rating'
DataToCSVIntermedi...
Export data to MoviesDB_intermediatefolder

Fig. 9 - Data Flow 'DataWrangling'.


ImportOriginalDataset
Import data from MoviesDB_origin
SelectCorrectNameColu...
Columns:
3 total
ConvertIntoFloat
Creating/updating the columns 'Movies, Genres, Rating'
FilterValutation
Filtering rows using expressions on columns 'Rating'
Descending
Sorting rows on columns 'Rating'
DataToCSVIntermedi...
Export data to MoviesDB_intermediatefolder
Select settings
Optimize
Inspect
Data preview
Learn more
Reset

Output stream name * SelectCorrectNameColumns

Description: Renaming ImportOriginalDataset to SelectCorrectNameColumns with columns 'Movies, Genres, Rating'

Incoming stream * ImportOriginalDataset

Options: ☑ Skip duplicate input columns ☑ Skip duplicate output columns

Input columns * □ Auto mapping · Reset · + Add mapping · Delete · 3 mappings: 2 column(s)

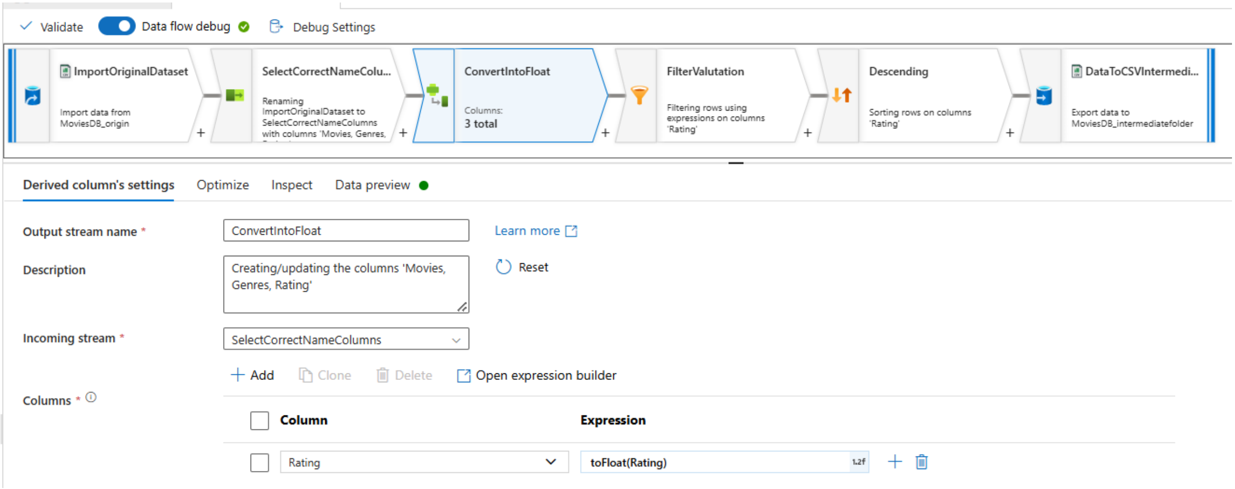
□	ImportOriginalDataset's column	Name as
□	abc movies	Movies
□	abc genresgenregenre	Genres
□	abc Rating	Rating

Fig. 10 – The node allows the selection of only the 3 required columns, instead of the original 5 columns, and is used to correct the format and errors in the names in order to facilitate subsequent machine translation activities.

[1] *It was considered to keep the original 'Movies' column, but the process works equally well with the others. Also, using Azure Translator, you can see that the translation of 'Movies' from English to Italian is 'Cinema'. Subsequently, I replaced the translation from Italian to French, resulting in 'Movies', as required by the exercise.*

[2] *The column names could have been transformed directly in this Data Flow at the 'SelectCorrectColumnName' node, but we wanted to try to automate the translation of the column names with the subsequent use of Azure Translator.*

[3] *The values of the Rating column were transformed into floats to enable the subsequent filter to function (Fig. 11)*


Validate
Data flow debug
Debug Settings
ImportOriginalDataset
Import data from MoviesDB_origin
SelectCorrectNameColu...
Renaming ImportOriginalDataset to SelectCorrectNameColumns with columns 'Movies, Genres,
ConvertIntoFloat
Columns:
3 total
FilterValutation
Filtering rows using expressions on columns 'Rating'
Descending
Sorting rows on columns 'Rating'
DataToCSVIntermedi...
Export data to MoviesDB_intermediatefolder
Derived column's settings
Optimize
Inspect
Data preview
Learn more
Reset
Add
Clone
Delete
Open expression builder

Output stream name * ConvertIntoFloat

Description: Creating/updating the columns 'Movies, Genres, Rating'

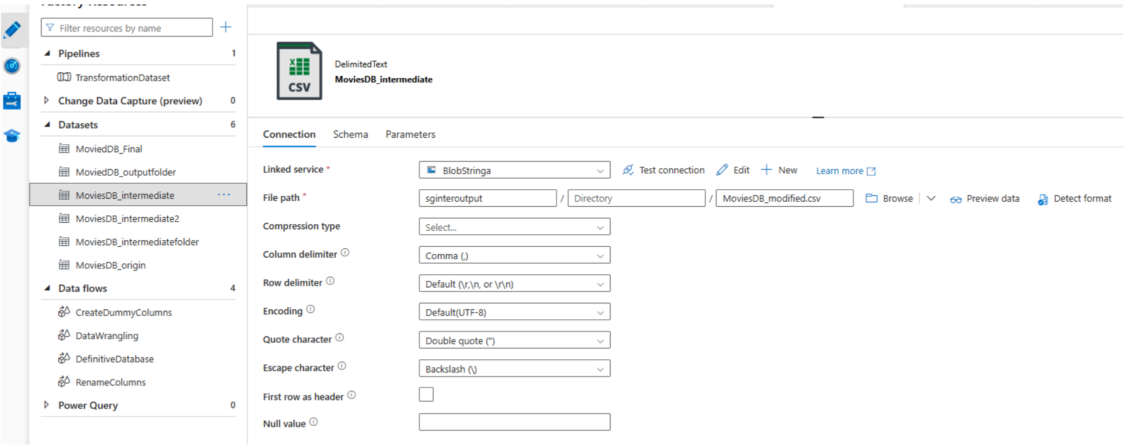
Incoming stream * SelectCorrectNameColumns

Columns *

Column	Expression
Rating	toFloat(Rating)

Fig. 11 – Casting Column Values Rating.

Next, a dataset was created from the intermediate file (Fig. 12) by disabling the 'first row as header' option to allow the Lookup – ‘SearchForColumnName’ activity to extract the column names. These were sent via a web activity (fig. 13) for automatic translation using Azure Translator (Fig. 14).



DelimitedText
MoviesDB_intermediate

Linked service * BlobStringa

File path * sginteroutput / Directory / MoviesDB_modified.csv

Compression type Select...

Column delimiter Comma (,)

Row delimiter Default (\r,\n, or \r\n)

Encoding Default(UTF-8)

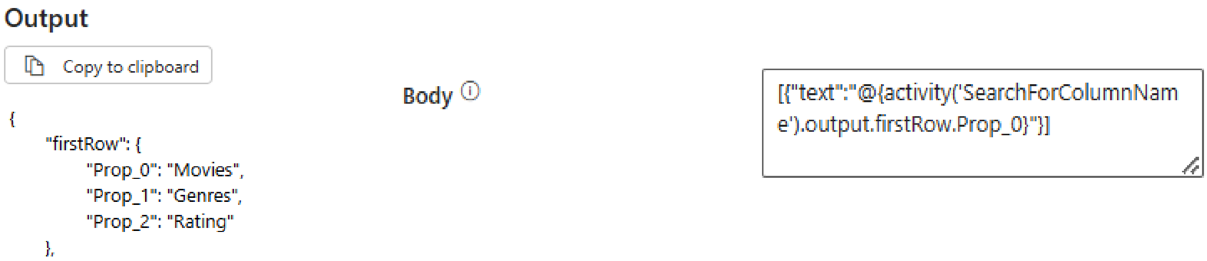
Quote character Double quote (")

Escape character Backslash (\)

First row as header ☐

Null value

Fig. 12 – Creating the dataset linked to the output file of the ‘Data Wrangling’ activity.

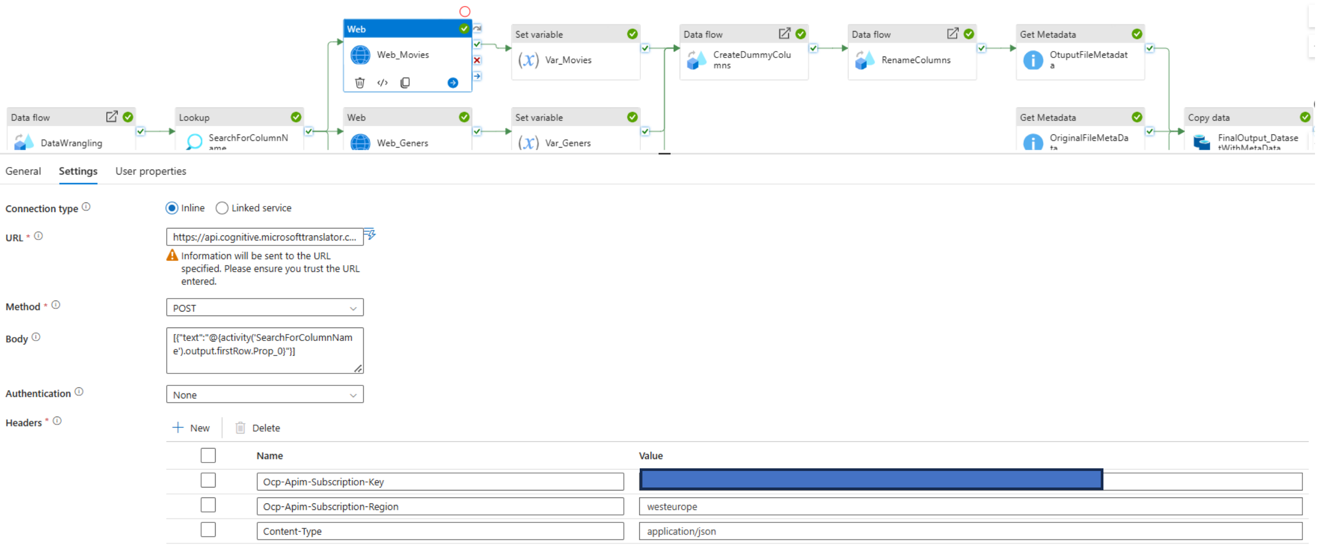


```
{
    "firstRow": {
        "Prop_0": "Movies",
        "Prop_1": "Genres",
        "Prop_2": "Rating"
    },
```

Body

```
[{"text":"@{activity('SearchForColumnName').output.firstRow.Prop_0}"}]
```

Fig. 13 – On the left the output of the Look up activity, on the right the body of the Web activity sent for translation via Azure Translator.



Connection type ● Inline ○ Linked service

URL * https://api.cognitive.microsofttranslator.c...

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

Method * POST

Body [{"text":"@{activity('SearchForColumnName').output.firstRow.Prop_0}"}]

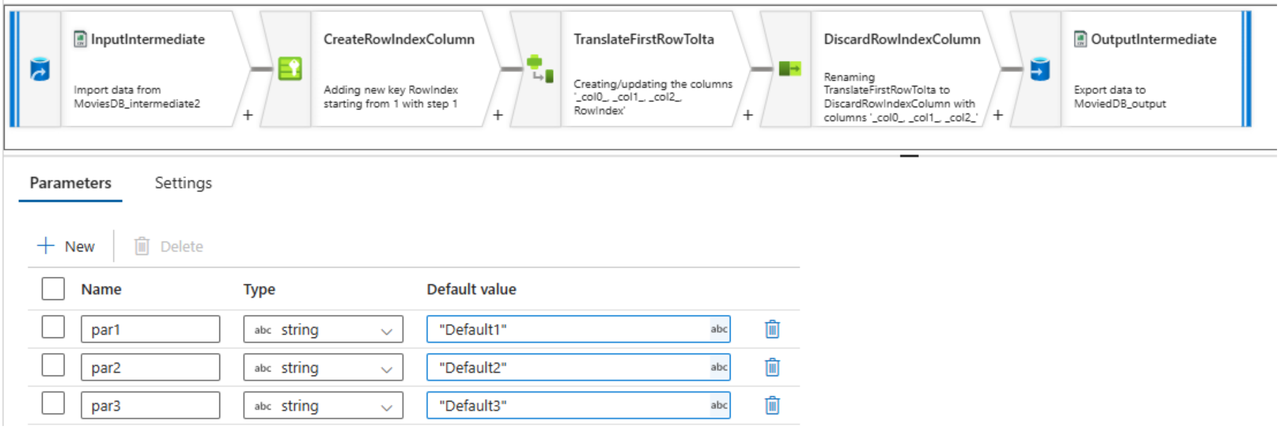
Authentication None

Headers *

	Name	Value
□	Ocp-Apim-Subscription-Key	[redacted]
□	Ocp-Apim-Subscription-Region	westeurope
□	Content-Type	application/json

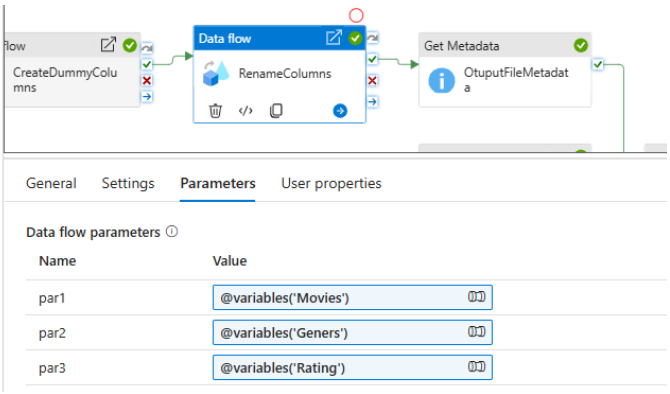
Fig. 14 – Web Activity Settings. In the URL, in addition to the 'Text Translation' link in figure 6, 'translate?api version=3.0&from=en&to=en' must also be entered to specify the version of the API you are using and that you want the text to be translated from English into Italian. In the first header, the 'key1' in figure 6 should be entered.

The outputs of the web activities were stored as pipeline variables, which were crucial for the subsequent activities; in fact, parameters were created in the Data Flow – 'RenameColumns' (Fig. 15 and Fig. 16) to link the pipeline variables to the column names (Fig. 17).


InputIntermediate
Import data from MoviesDB_intermediate2
CreateRowIndexColumn
Adding new key RowIndex starting from 1 with step 1
TranslateFirstRowToIta
Creating/updating the columns '_col0_, _col1_, _col2_, RowIndex'
DiscardRowIndexColumn
Renaming TranslateFirstRowToIta to DiscardRowIndexColumn with columns '_col0_, _col1_, _col2_'
OutputIntermediate
Export data to MoviedDB_output
Parameters
Settings
+ New
Delete

□	Name	Type	Default value
□	par1	abc string	"Default1"
□	par2	abc string	"Default2"
□	par3	abc string	"Default3"

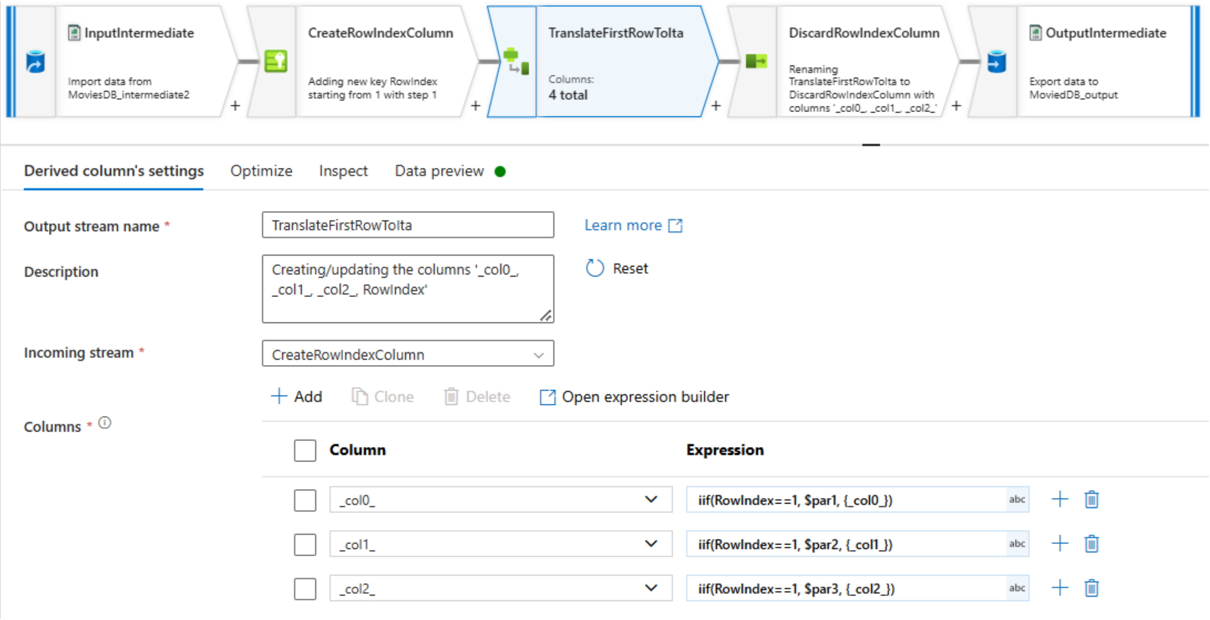
Fig. 15 – Creation of Data Flow parameters.



Data flow parameters

Name	Value
par1	@variables('Movies')
par2	@variables('Geners')
par3	@variables('Rating')

Fig. 16 – Linking in the pipeline of parameters to data flow variables.


InputIntermediate
Import data from MoviesDB_intermediate2
CreateRowIndexColumn
Adding new key RowIndex starting from 1 with step 1
TranslateFirstRowTolta
Columns:
4 total
DiscardRowIndexColumn
Renaming TranslateFirstRowTolta to DiscardRowIndexColumn with columns '_col0_, _col1_, _col2_'
OutputIntermediate
Export data to MoviedDB_output
Derived column's settings
Optimize
Inspect
Data preview
Learn more
Reset
Add
Clone
Delete
Open expression builder

Output stream name * TranslateFirstRowTolta

Description: Creating/updating the columns '_col0_, _col1_, _col2_, RowIndex'

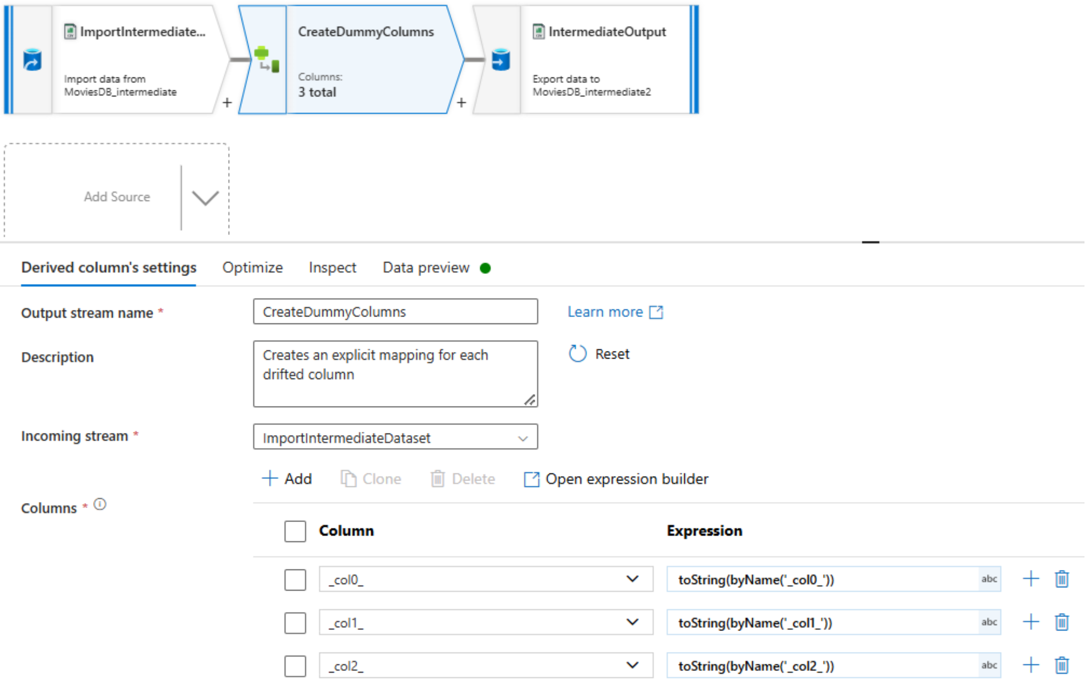
Incoming stream * CreateRowIndexColumn

Columns *

☐ Column	Expression
☐ _col0_	iif(RowIndex==1, $par1, {_col0_})
☐ _col1_	iif(RowIndex==1, $par2, {_col1_})
☐ _col2_	iif(RowIndex==1, $par3, {_col2_})

Fig. 17 – Replacement of the dummy columns with the values of the Data Flow parameters, i.e. the pipeline variables.

As there is currently no activity to dynamically change the column names, a trick was found to create dummy columns using an additional dataflow – 'Create DummyColumns' (Fig. 18). In this way, the column names are values that are inserted into a row of the dataset, making them replaceable with the parameters created via a 'Derived Column'.



Output stream name * CreateDummyColumns

Description: Creates an explicit mapping for each drifted column

Incoming stream * ImportIntermediateDataset

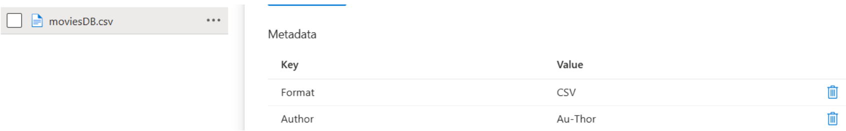
Columns *

☐ Column	Expression
☐ _col0_	toString(byName('_col0_'))
☐ _col1_	toString(byName('_col1_'))
☐ _col2_	toString(byName('_col2_'))

Fig. 18 – Use of a Data Flow to create dummy columns for the purposes described..

In parallel, a Get Metadata activity was used to extract information from the original dataset.

A Shared Access Signature (SAS) web activity was used to obtain custom metadata (Fig. 19), using the SAS URL blob (Fig. 20).

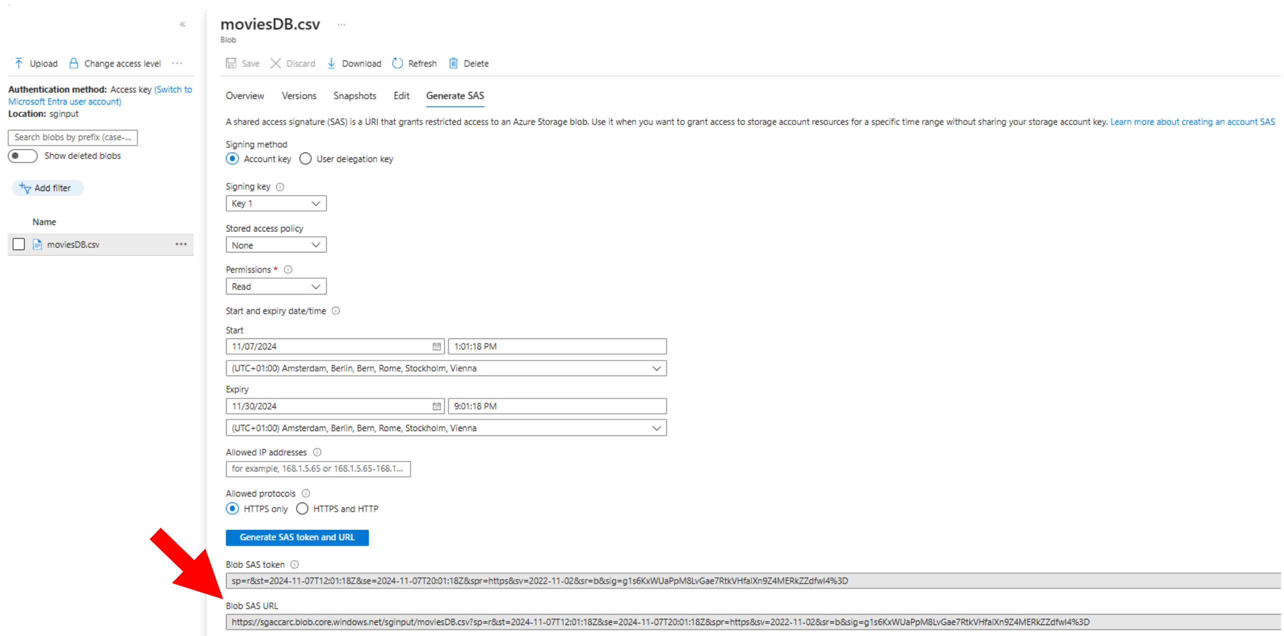


□ moviesDB.csv ···

Metadata

Key	Value
Format	CSV
Author	Au-Thor

Fig. 19 – Custom Metadata.



Authentication method: Access key (Switch to Microsoft Entra user account)
Location: sginput

Name
□ moviesDB.csv ···

moviesDB.csv ···
Blob

A shared access signature (SAS) is a URI that grants restricted access to an Azure Storage blob. Use it when you want to grant access to storage account resources for a specific time range without sharing your storage account key. Learn more about creating an account SAS

Signing method
● Account key ○ User delegation key

Signing key: Key 1

Stored access policy: None

Permissions *: Read

Start and expiry date/time

Start: 11/07/2024 1:01:18 PM
(UTC+01:00) Amsterdam, Berlin, Bern, Rome, Stockholm, Vienna

Expiry: 11/30/2024 9:01:18 PM
(UTC+01:00) Amsterdam, Berlin, Bern, Rome, Stockholm, Vienna

Allowed IP addresses: for example, 168.1.5.65 or 168.1.5.65-168.1...

Allowed protocols
● HTTPS only ○ HTTPS and HTTP

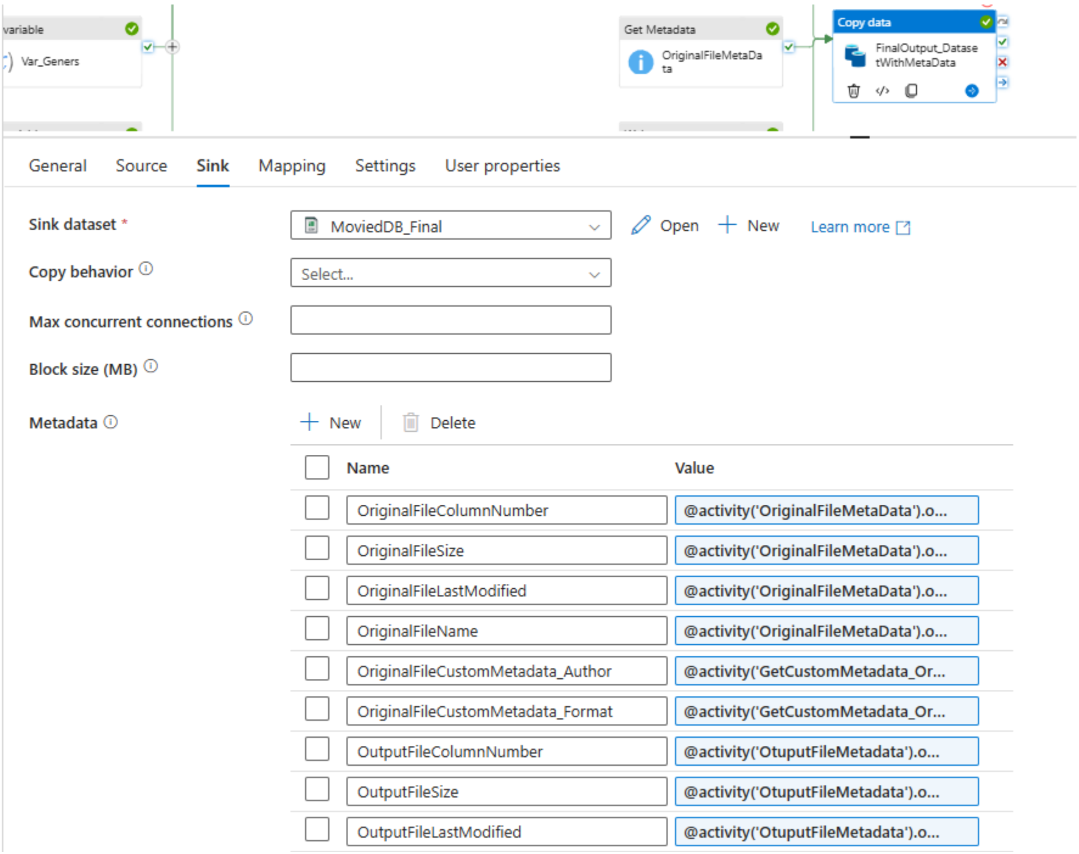
Generate SAS token and URL

Blob SAS token
```
sp=r&st=2024-11-07T12:01:18Z&se=2024-11-07T20:01:18Z&spr=https&sv=2022-11-02&sr=b&sig=g1s6KxWUaPpM8LvGae7RtkVHfalXn9Z4MERkZZdfwl4%3D
```

Blob SAS URL
```
https://sgaccarc.blob.core.windows.net/sginput/moviesDB.csv?sp=r&st=2024-11-07T12:01:18Z&se=2024-11-07T20:01:18Z&spr=https&sv=2022-11-02&sr=b&sig=g1s6KxWUaPpM8LvGae7RtkVHfalXn9Z4MERkZZdfwl4%3D
```

Fig. 20 – URL to be inserted in the web activity to allow temporary access to the dataset.

All metadata, both from the original dataset and the final dataset, were stored in the final output dataset via a 'Copy data' activity (Fig.21 and Fig. 22).



Sink dataset *: MoviedDB_Final

Copy behavior: Select...

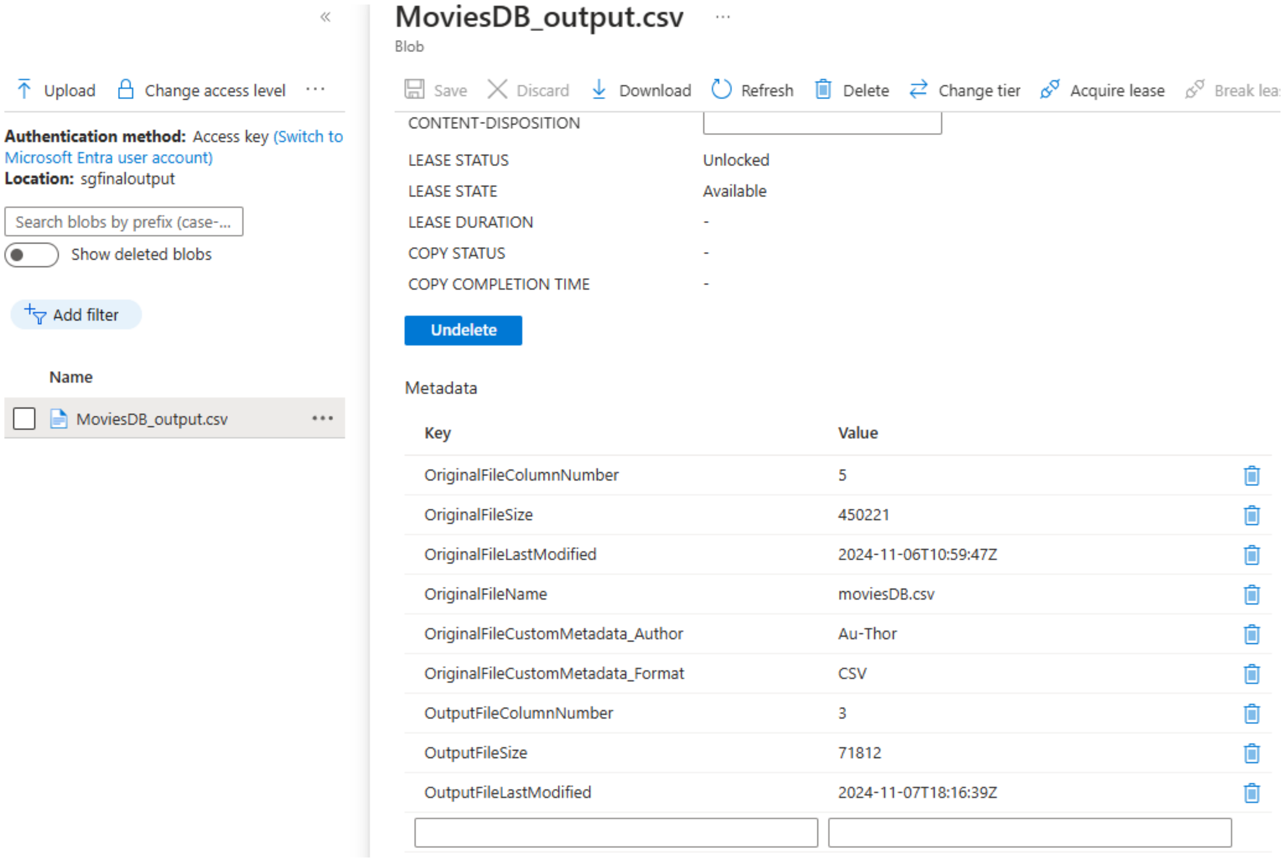
Max concurrent connections:

Block size (MB):

Metadata

Name	Value
OriginalFileColumnNumber	@activity('OriginalFileMetaData').o...
OriginalFileSize	@activity('OriginalFileMetaData').o...
OriginalFileLastModified	@activity('OriginalFileMetaData').o...
OriginalFileName	@activity('OriginalFileMetaData').o...
OriginalFileCustomMetadata_Author	@activity('GetCustomMetadata_Or...
OriginalFileCustomMetadata_Format	@activity('GetCustomMetadata_Or...
OutputFileColumnNumber	@activity('OtupputFileMetadata').o...
OutputFileSize	@activity('OtupputFileMetadata').o...
OutputFileLastModified	@activity('OtupputFileMetadata').o...

Fig. 21 – Metadata saved in the final output file.



Authentication method: Access key (Switch to Microsoft Entra user account)
Location: sgfinaloutput

Search blobs by prefix (case-...

Show deleted blobs

Add filter

Name
□ MoviesDB_output.csv

CONTENT-DISPOSITION	
LEASE STATUS	Unlocked
LEASE STATE	Available
LEASE DURATION	-
COPY STATUS	-
COPY COMPLETION TIME	-

Metadata

Key	Value
OriginalFileColumnNumber	5
OriginalFileSize	450221
OriginalFileLastModified	2024-11-06T10:59:47Z
OriginalFileName	moviesDB.csv
OriginalFileCustomMetadata_Author	Au-Thor
OriginalFileCustomMetadata_Format	CSV
OutputFileColumnNumber	3
OutputFileSize	71812
OutputFileLastModified	2024-11-07T18:16:39Z

Fig. 22 – Output File Metadata.