

Azure Data Factory Cargar datos y metadata de dos JSON correlativos en un Dedicated SQL Pool

Ambos se encuentran en Azure Data Lake Storage, en la ruta **data/customer**

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

Dashboard > Storage accounts > datalake2000 >

data Container

Search (Ctrl+ /)

Upload Add Directory Refresh Rename Delete Change tier

Authentication method: Access key (Switch to Azure AD User Account)
Location: data / customer

Search blobs by prefix (case-sensitive)

Name	Modified	Access tier	Blob type
[.]			
customer1.json	7/24/2021, 12:25:36 ...	Hot (Inferred)	Block blob
customer2.json	7/24/2021, 12:25:36 ...	Hot (Inferred)	Block blob

Los archivos tienen la siguiente estructura

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

Dashboard > Storage accounts > datalake2000 > data >

customer/customer1.json

Blob

Save Discard Download Refresh Delete

Overview Versions Edit Generate SAS

```
2 {
3   "customerid":1,
4   "customername":"UserA",
5   "registered":true
6 },
7 {
8   "customerid":2,
9   "customername":"UserB",
10  "registered":true
11 }
12 ]
13
```

Json Preview

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

Dashboard > Storage accounts > datalake2000 > data >

customer/customer2.json

Blob

Save Discard Download Refresh Delete

Overview Versions Edit Generate SAS

```
1 [
2 {
3   "customerid":3,
4   "customername":"UserC",
5   "registered":true
6 },
7 {
8   "customerid":4,
9   "customername":"UserD",
10  "registered":true
11 }
12 ]
```

Json Preview

Se busca cargar sus datos en la siguiente tabla. Además queremos cargar la metadata correspondiente al nombre del archivo y la fecha en la que se realiza esta operación.

```
04.sql
1 CREATE TABLE [dbo].[Customer]
2 (
3     [customerid] int,
4     [customername] varchar(20),
5     [registered] bit,
6     [FileDate] datetime,
7     [FileName] varchar(200)
8 )
```

Para comenzar creamos un Pipeline y añadimos la actividad **Get Metadata**

The screenshot displays the Microsoft Azure Data Factory portal. The top navigation bar shows the user 'techsup1000@gmail.com' and the 'appfactory5000' workspace. The main interface is divided into three sections: 'Activities', 'Properties', and 'General'. In the 'Activities' section, the 'Get Metadata' activity is selected, and its configuration is shown in the 'Properties' section. The 'Name' field is set to 'GetMetadataFiles'. The 'Description' field is empty. The 'Timeout' is set to '7.00:00:00'. The 'General' tab is active, showing the 'Name' and 'Description' fields. The 'Properties' section on the right shows the 'Name' field set to 'customer' and the 'Description' field empty. A red arrow points from the 'Get Metadata' activity in the 'Activities' list to the 'GetMetadataFiles' activity in the 'Properties' section.

Microsoft Azure | appfactory5000

Search

techsup1000@gmail.com

DATA FACTORY

customer

Activities

Search activities

General

Append variable

Delete

Execute Pipeline

Execute SSIS package

Get Metadata

Lookup

Stored procedure

Set variable

Validation

Web

WebHook

Get Metadata

GetMetadataFiles

Properties

General

Name *

customer

Description

Concurrency has moved to the [Settings tab.](#)

Annotations

+ New

General

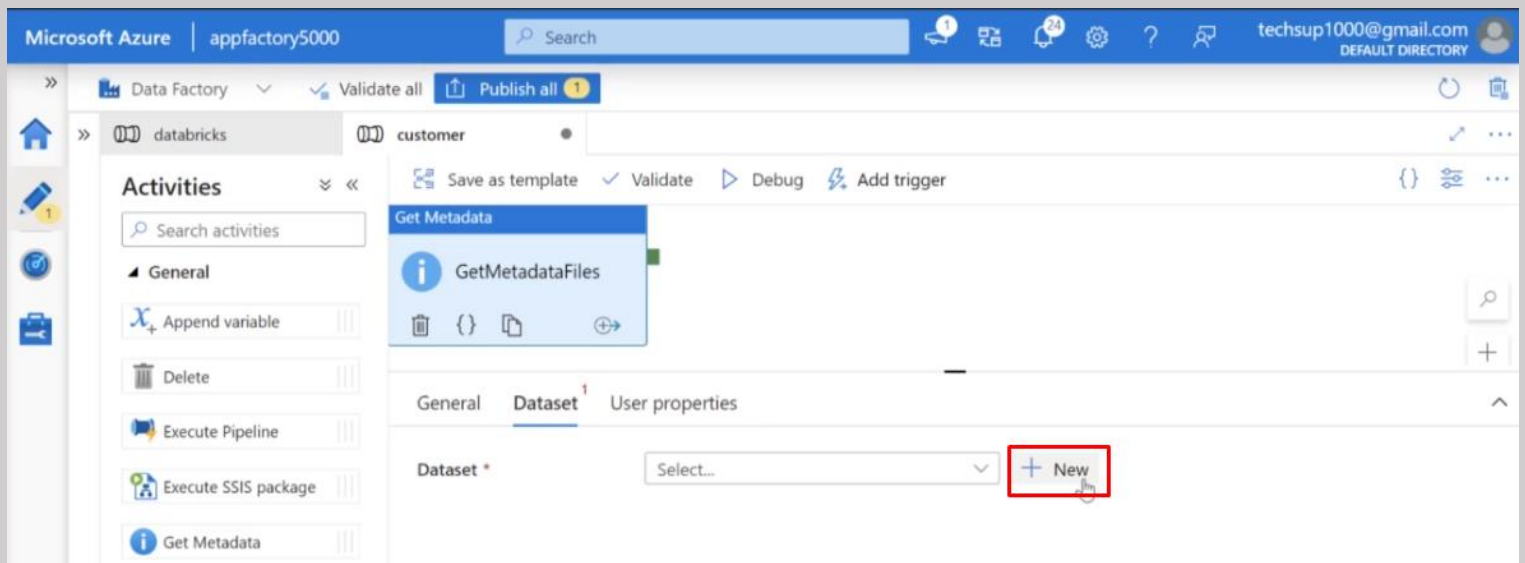
Name *

GetMetadataFiles

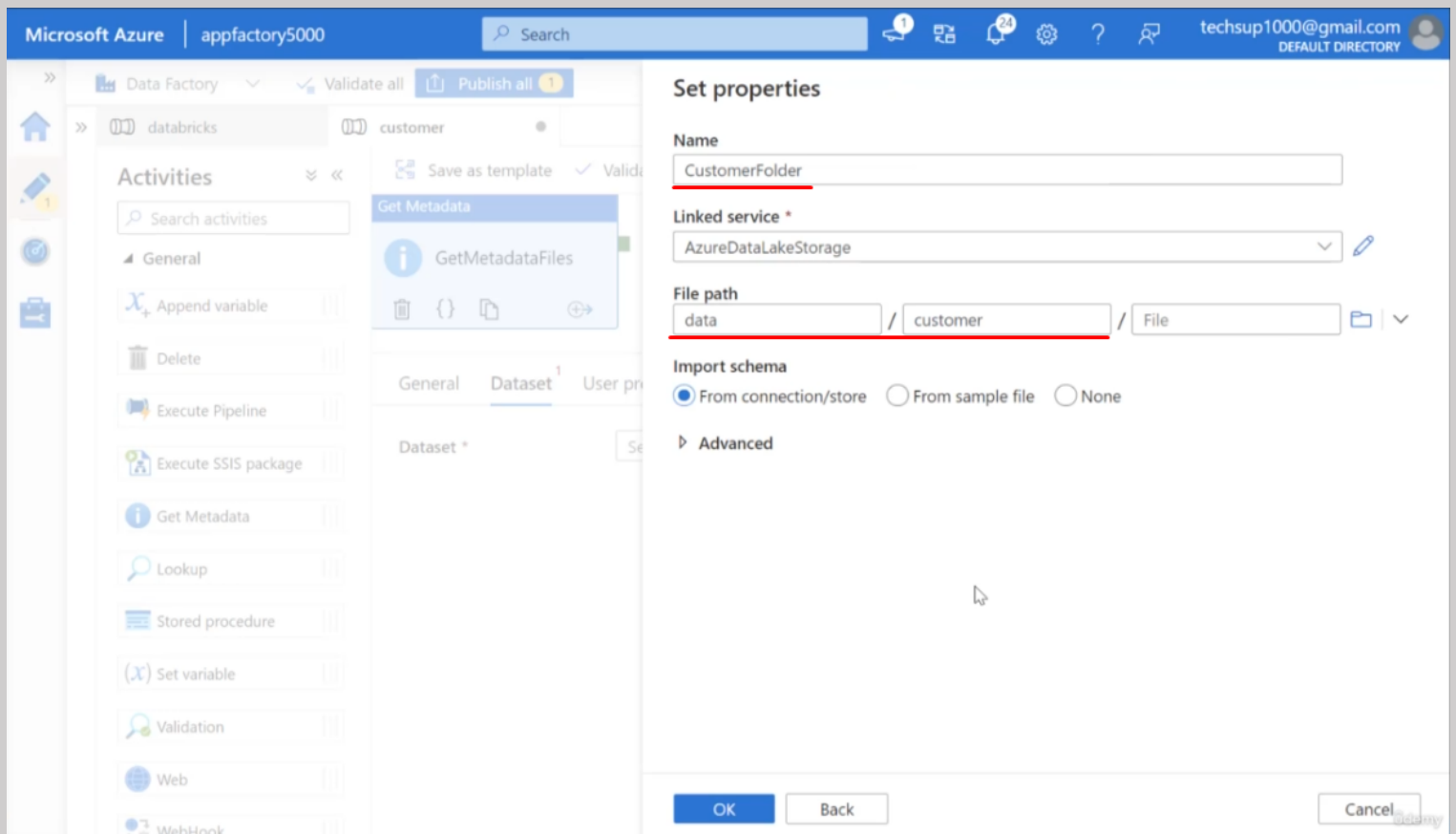
Description

Timeout

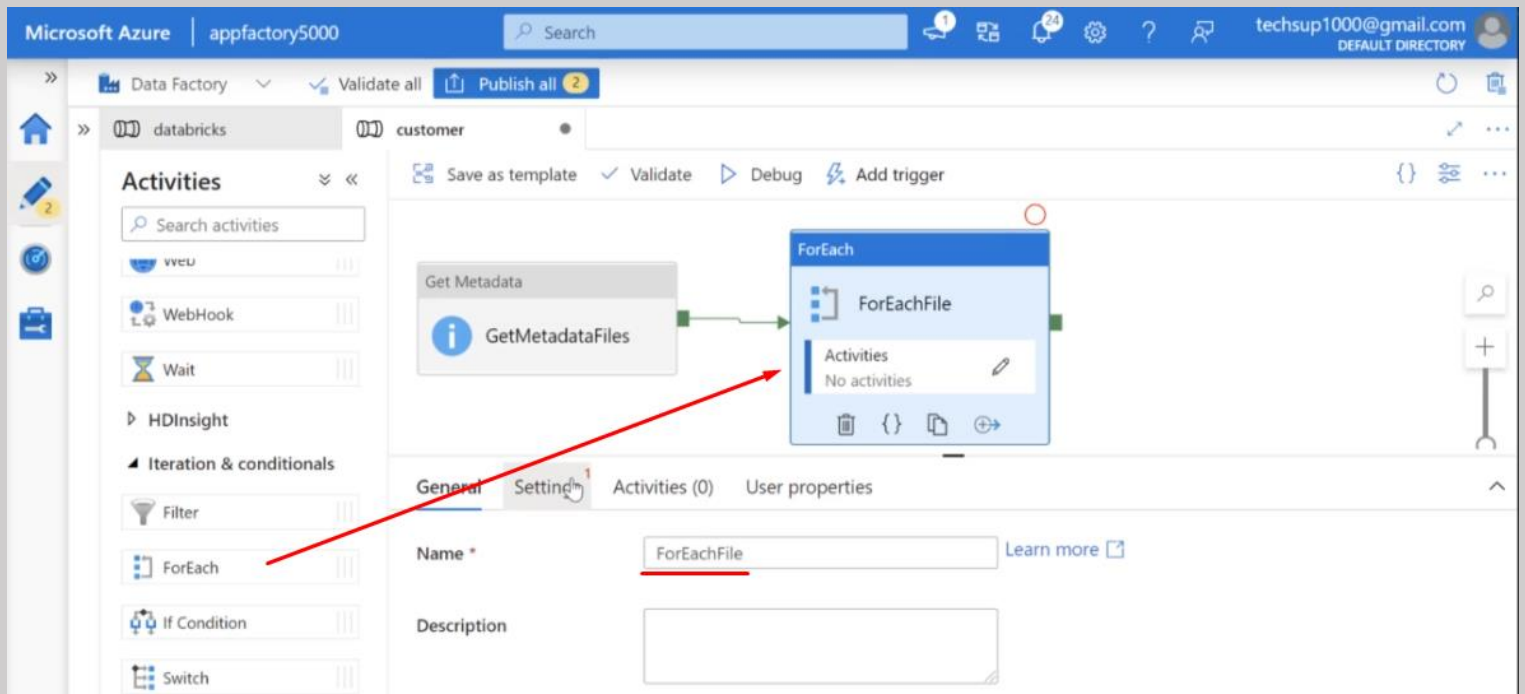
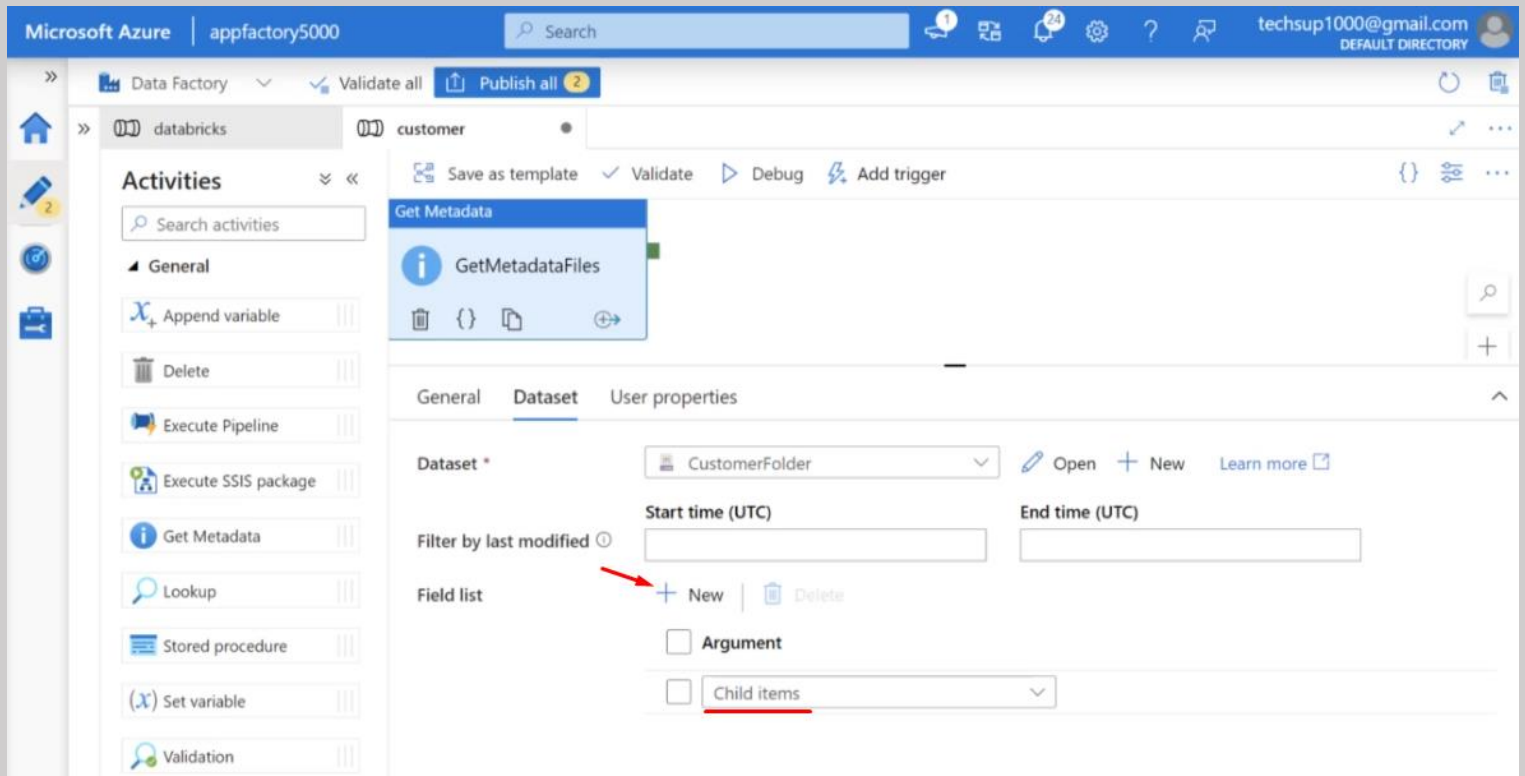
7.00:00:00



El dataset de la fuente hará referencia hacia Azure Data Lake Storage y el tipo de dato será JSON



Ahora podemos obtener información sobre los archivos subyacentes en nuestra carpeta particular. Lo primero que quiero hacer es obtener los elementos. Así que quiero que esta actividad primero obtenga todos los elementos que están en mi carpeta "customer". Eso es lo primero que quiero hacer. Y luego, para cada archivo, quiero realizar otra actividad.



The screenshot shows the Microsoft Azure Data Factory interface. On the left, the 'Activities' pane lists various activities including 'vveu', 'WebHook', 'Wait', 'Filter', 'ForEach', 'If Condition', 'Switch', and 'Until'. The main workspace displays a workflow with a 'GetMetadataFiles' activity followed by a 'ForEach' activity. The 'ForEach' activity is expanded, showing its 'Settings' tab. The 'Items' property is highlighted with a red box, and a tooltip message states: 'This property should be parameterized. Add dynamic content (Alt+Shift+D)'.

De la actividad anterior “GetMetadataFiles” esta rescatando los “childitems”, los elementos que capturó la actividad previa, que corresponderían a los dos archivos JSON.

The screenshot shows the 'Add dynamic content' dialog box. The text '@activity('GetMetadataFiles').output.childitems' is entered into the input field. Below the input field, there is a 'Clear contents' button and a section titled 'Add dynamic content above using any combination of expressions, functions and system variables. Click any of the available System variables or Functions below to add them directly:'. This section includes a search bar and a list of available options: 'System variables', 'Functions', and 'Activity outputs'. Under 'Activity outputs', 'GetMetadataFiles' and 'GetMetadataFiles activity output' are listed. The dialog box has 'Finish' and 'Cancel' buttons at the bottom.

Microsoft Azure | appfactory5000

Search

techsup1000@gmail.com
DEFAULT DIRECTORY

Data Factory | Validate all | Publish all 2

databricks customer

Activities

Search activities

WebHook

Wait

HDInsight

Iteration & conditionals

Filter

ForEach

If Condition

Switch

Save as template | Validate | Debug | Add trigger

Get Metadata

GetMetadataFiles

ForEach

ForEachFile

Activities

No activities

General | Settings | Activities (0) | User properties

Case

Activity

ForEach

No activities

Microsoft Azure | appfactory5000

Search

techsup1000@gmail.com
DEFAULT DIRECTORY

Data Factory | Validate all | Publish all 2

databricks customer

Activities

Search activities

Data Lake Analytics

General

Append variable

Delete

Execute Pipeline

Execute SSIS package

Get Metadata

Save as template | Validate | Debug | Add trigger

customer > ForEachFile

Get Metadata

GetMetadataAllFiles

General | Dataset 1 | User properties

Name *

GetMetadataAllFiles

Learn more

Microsoft Azure | appfactory5000

Search

techsup1000@gmail.com
DEFAULT DIRECTORY

Data Factory | Validate all | Publish all 2

databricks customer

Activities

Search activities

Data Lake Analytics

General

Append variable

Delete

Execute Pipeline

Execute SSIS package

Get Metadata

Look up

Save as template | Validate | Debug | Add trigger

customer > ForEachFile

Get Metadata

GetMetadataAllFiles

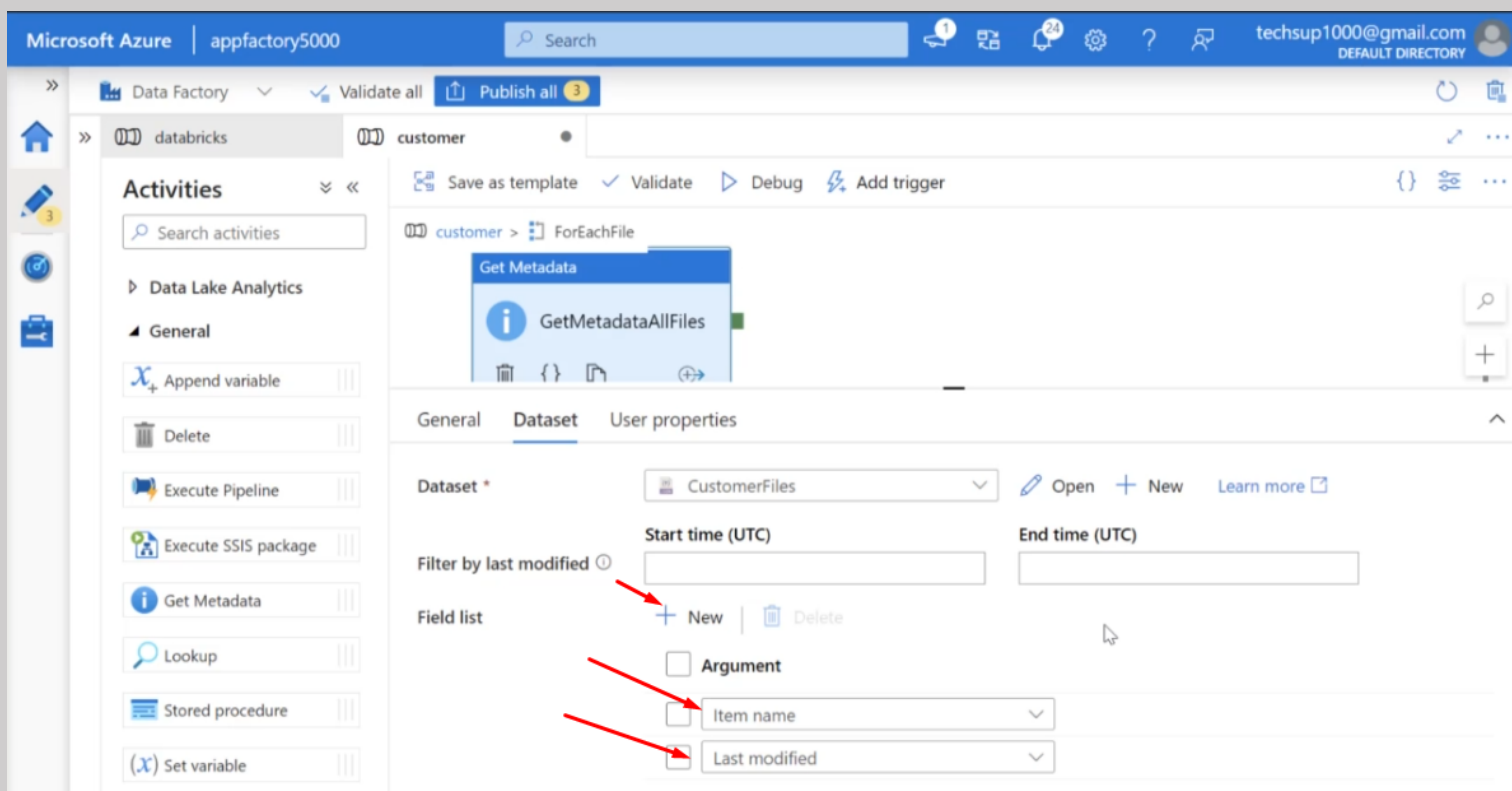
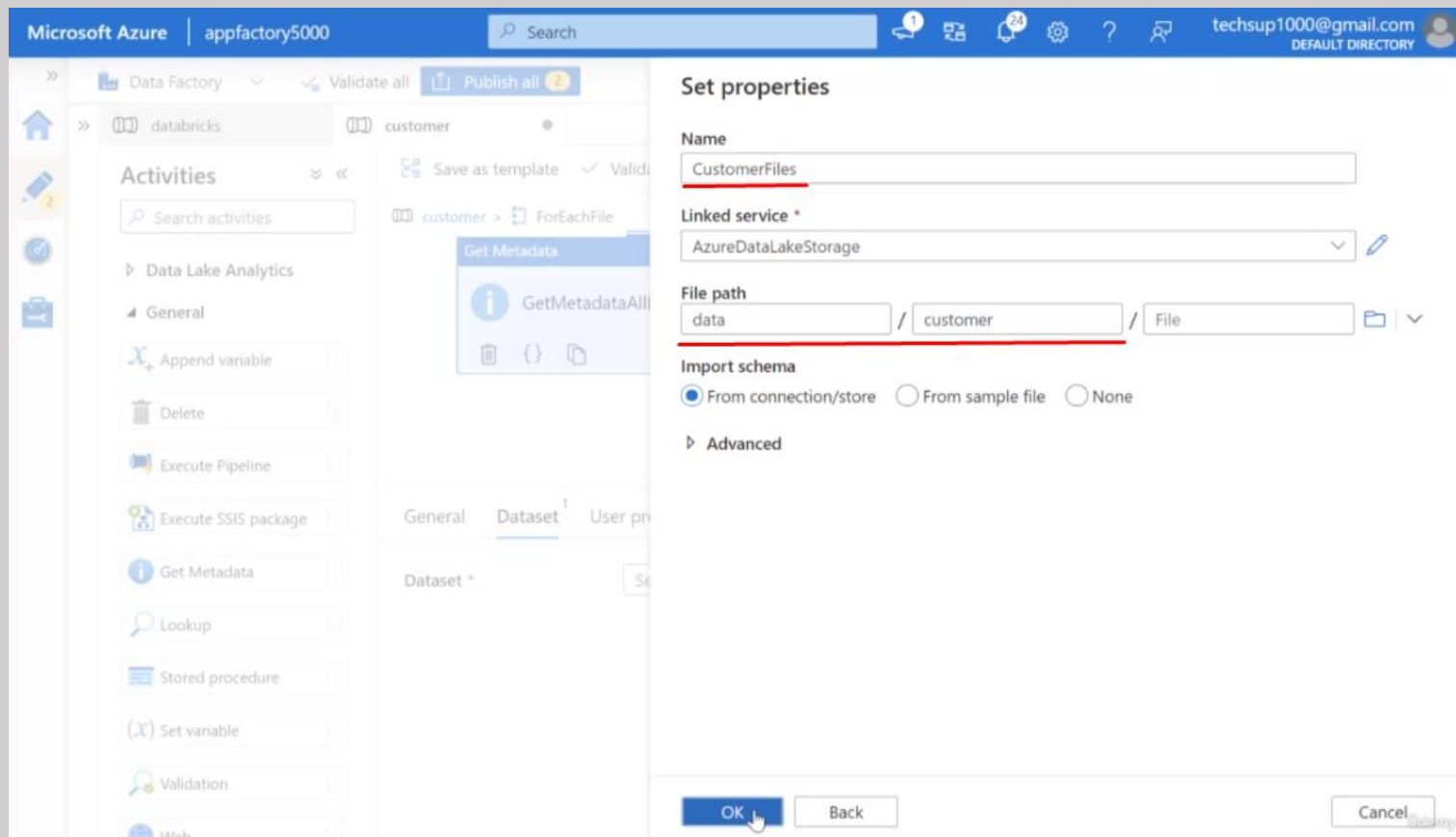
General | Dataset 1 | User properties

Dataset *

Select...

+ New

Creamos un nuevo dataset con referencia a Azure Data Lake Storage y seleccionamos como tipo de dato JSON



Microsoft Azure | appfactory5000

Search

techsup1000@gmail.com
DEFAULT DIRECTORY

Data Factory | Validate all | Publish all 3

databricks | customer

Activities

Search activities

Data Lake Analytics

General

Append variable

Delete

Execute Pipeline

Execute SSIS package

Get Metadata

Lookup

Stored procedure

Set variable

Validation

Get Metadata

GetMetadataAllFiles

Save as template | Validate | Debug | Add trigger

customer > ForEachFile

General | Dataset | User properties

Dataset * | CustomerFiles | Open | New | Learn more

Filter by last modified

Start time (UTC) | End time (UTC)

Field list

+ New | Delete

Argument

Item name

Last modified

Microsoft Azure | appfactory5000

Search

techsup1000@gmail.com
DEFAULT DIRECTORY

Data Factory | Validate all | Publish all 3

databricks | customer | CustomerFiles

JSON

CustomerFiles

Connection | Schema | Parameters

+ New

Properties

General | Related (1)

Name * | CustomerFiles

Description

Annotations

+ New

Microsoft Azure | appfactory5000

Search

techsup1000@gmail.com
DEFAULT DIRECTORY

Data Factory | Validate all | Publish all 3

databricks | customer | CustomerFiles

JSON
CustomerFiles

Connection | **Schema** | Parameters

+ New | Delete

Name	Type	Default value
FileName	String	Value

Properties

General | Related (1)

Name *
CustomerFiles

Description

Annotations
+ New

Microsoft Azure | appfactory5000

Search

techsup1000@gmail.com
DEFAULT DIRECTORY

Data Factory | Validate all | Publish all 3

databricks | customer | CustomerFiles

JSON
CustomerFiles

Connection | Schema | Parameters

Linked service *
AzureDataLakeStorage

Test connection | Edit | + New | Learn more

File path *
data / customer / File

Compression type
None

Encoding
Default(UTF-8)

Add dynamic content
[Alt+Shift+D]

Properties

General | Related (1)

Name *
CustomerFiles

Description

Annotations
+ New

Microsoft Azure | appfactory5000

Search

techsup1000@gmail.com
DEFAULT DIRECTORY

Data Factory | Validate all | Publish all (3)

databricks | customer | CustomerFiles

JSON
CustomerFiles

Connection | Schema | Parameters

Linked service * AzureDataLakeStorage

Test connection | Edit | + New | Learn more

File path * data / customer

Compression type None

Encoding Default(UTF-8)

Add dynamic content

@dataset().FileName

Clear contents

Add dynamic content above using any combination of [expressions](#), [functions](#) and [system variables](#). Click any of the available System variables or Functions below to add them directly:

Filter system variables and functions...

Functions

Parameters

FileName

Finish Cancel

Así que aquí estamos tratando de hacer nuestro dataset más dinámico. Así que para cada archivo obtendrá automáticamente el nombre del archivo. Ahora bien, ¿cómo se obtiene realmente el nombre del archivo?

Microsoft Azure | appfactory5000

Search

techsup1000@gmail.com
DEFAULT DIRECTORY

Data Factory | Validate all | Publish all (3)

databricks | customer | CustomerFiles

JSON
CustomerFiles

Connection | Schema | Parameters

Linked service * AzureDataLakeStorage

Test connection | Edit | + New | Learn more

File path * data / customer / @dataset().FileName

Compression type None

Encoding Default(UTF-8)

Browse Preview data

Si vuelvo ahora a mi pipeline "customer", puedes ver nuestro dataset customer. Puedes ver que se añade ahora una dataset property, por lo que ahora queremos pasar el nombre de archivo, tenemos una propiedad que luego será pasado a CustomerFiles y CustomerFiles puede ahora tratar sobre cada archivo en nuestra carpeta. Y aquí, ¿cómo obtenemos el valor? Así que, de nuevo, podemos añadir contenido dinámico

The screenshot shows the Microsoft Azure Data Factory portal interface. The top navigation bar includes the Microsoft Azure logo, the workspace name 'appfactory5000', a search bar, and user information 'techsup1000@gmail.com'. The main content area displays the 'customer' pipeline with the 'CustomerFiles' dataset selected. The 'Dataset' tab is active, showing the 'CustomerFiles' dataset. A red arrow points to the 'FileName' property in the 'Dataset properties' table. A red box highlights the 'Add dynamic content (Alt+Shift+D)' button. The table has columns for Name, Value, and Type. The 'FileName' property has a value of 'Value' and a type of 'string'. Below the table, there are fields for 'Start time (UTC)' and 'End time (UTC)', and a 'Field list' section with a '+ New' button and a 'Delete' button.

Name	Value	Type
FileName	Value	string

The screenshot shows the Microsoft Azure Data Factory portal interface, similar to the previous one. The 'Dataset' tab is active, showing the 'CustomerFiles' dataset. The 'FileName' property in the 'Dataset properties' table now has a value of '@item().name'. The rest of the interface, including the top navigation bar and the 'Field list' section, remains the same.

Name	Value	Type
FileName	@item().name	string

Microsoft Azure | appfactory5000

Search

techsup1000@gmail.com
DEFAULT DIRECTORY

Data Factory | Validate all | Publish all (3)

databricks | customer | CustomerFiles

Save as template | Validate | Validate copy runtime | Debug | Add trigger

customer > ForEachFile

Get Metadata
GetMetadataAllFiles

Copy data
CopytoSynapse

General | Source | Sink | Mapping | Settings | User properties

Name *
CopytoSynapse

Description

Timeout
7.00:00:00

Retry
0

Retry interval
30

Microsoft Azure | appfactory5000

Search

techsup1000@gmail.com
DEFAULT DIRECTORY

Data Factory | Validate all | Publish all (3)

databricks | customer | CustomerFiles

Save as template | Validate | Validate copy runtime | Debug | Add trigger

customer > ForEachFile

Get Metadata
GetMetadataAllFiles

Copy data
CopytoSynapse

General | Source | Sink | Mapping | Settings | User properties

Source dataset *
CustomerFiles

Dataset properties

Name	Value	Type
FileName	Value	string

File path type
☒ File path in dataset ☐ Wildcard file path ☐ List of files

Filter by last modified
Start time (UTC) End time (UTC)

Recursively
☒

Microsoft Azure | appfactory5000

Search

techsup1000@gmail.com
DEFAULT DIRECTORY

Data Factory | Validate all | Publish all 3

databricks | customer | CustomerFiles

Save as template | Validate | Validate copy runtime | Debug | Add trigger

customer > ForEachFile

Get Metadata

GetMetadataAllFiles

Copy data

CopytoSynapse

General | **Source** | Sink | Mapping | Settings | User properties

Source dataset * | CustomerFiles | Open | New | Preview data | Learn more

Dataset properties

Name	Value	Type
FileName	@item().name	string

File path type

☒ File path in dataset ☐ Wildcard file path ☐ List of files

Filter by last modified

Start time (UTC) | End time (UTC)

Recursively ☒

Microsoft Azure | appfactory5000

Search

techsup1000@gmail.com
DEFAULT DIRECTORY

Data Factory | Validate all | Publish all 3

databricks | customer | CustomerFiles

Save as template | Validate | Validate copy runtime | Debug | Add trigger

customer > ForEachFile

Get Metadata

GetMetadataAllFiles

Copy data

CopytoSynapse

General | Source | **Sink** | Mapping | Settings | User properties

Sink dataset * | Select... | **+ New**

Microsoft Azure | appfactory5000

Search

techsup1000@gmail.com
DEFAULT DIRECTORY

Data Factory > Validate all > Publish all (3)

databricks > customer > CustomerFiles

Save as template > Validate > Validate copy runtime

customer > ForEachFile

Get Metadata > GetMetadataAllFiles

Copy data

General Source Sink Mapping Settings User properties

Sink dataset * Select...

Set properties

Name: AzureSynapseAnalyticsTable10

Linked service *: AzureSynapseAnalytics

Table name: dbo.Customer

☐ Edit

Import schema: ☒ From connection/store ☐ None

Advanced

Microsoft Azure | appfactory5000

Search

techsup1000@gmail.com
DEFAULT DIRECTORY

Data Factory > Validate all > Publish all (4)

databricks > customer > CustomerFiles

Save as template > Validate > Validate copy runtime > Debug > Add trigger

customer > ForEachFile

Get Metadata > GetMetadataAllFiles

Copy data > CopytoSynapse

General Source Sink Mapping Settings User properties

Sink dataset * AzureSynapseAnalyticsTable10 Open + New Learn more

Copy method: ☐ PolyBase ☐ Copy command ☒ Bulk insert

Table option: ☒ None ☐ Auto create table

Microsoft Azure | Data Factory | appfactory5000

Search

techsup1000@gmail.com
DEFAULT DIRECTORY

Data Factory | Validate all | Publish all

databricks | customer | CustomerFiles

Activities

- Move & transform
 - Copy data
 - Data flow
- Azure Data Explorer
- Azure Function
- Batch Service
- Databricks
- Data Lake Analytics
- General
 - Append variable
 - Delete
 - Execute Pipeline
 - Execute SSIS package

Save as template | Validate | Validate copy runtime | Debug | Add trigger

customer > ForEachFile

Get Metadata

GetMetadataAllFiles

Copy data

CopytoSynapse

General | Source | Sink | Mapping | Settings | User properties

Max concurrent connections

Additional columns

NAME

VALUE

FileDate

FileName

\$\$FILEPATH

Add dynamic content

Custom

\$\$FILEPATH

\$\$COLUMN

Microsoft Azure | Data Factory | appfactory5000

Search

techsup1000@gmail.com
DEFAULT DIRECTORY

Data Factory | Validate all | Publish all

databricks | customer | CustomerFiles

Activities

- Move & transform
 - Copy data
 - Data flow
- Azure Data Explorer
- Azure Function
- Batch Service
- Databricks
- Data Lake Analytics
- General
 - Append variable
 - Delete
 - Execute Pipeline
 - Execute SSIS package

Save as template | Validate | Validate copy runtime | Debug | Add trigger

customer > ForEachFile

General | Source | Sink | Mapping | Settings | User properties

File path type

Filter by last modified

Recursively

Enable partition discovery

Max concurrent connections

Additional columns

Add dynamic content

@activity('GetMetadataAllFiles').output.lastModified

Clear contents

Add dynamic content above using any combination of expressions, functions and system variables. Click any of the available System variables or Functions below to add them directly:

Filter system variables and functions...

Activity outputs

- GetMetadataAllFiles
 - GetMetadataAllFiles activity output
- GetMetadataFiles
 - GetMetadataFiles activity output

ForEach iterator

Finish

Cancel

Microsoft Azure | Data Factory | appfactory5000

Search

techsup1000@gmail.com
DEFAULT DIRECTORY

Data Factory | databricks | customer | CustomerFiles

Activities

- Move & transform
 - Copy data
 - Data flow
- Azure Data Explorer
- Azure Function
- Batch Service
- Databricks
- Data Lake Analytics
- General
 - Append variable
 - Delete
 - Execute Pipeline
 - Execute SSIS package

General | Source | Sink | Mapping | Settings | User properties

Recursively ☒

Enable partition discovery ☐

Max concurrent connections

Additional columns

NAME	VALUE
FileDate	@activity('GetMetadataAllFiles').output...
FileName	\$\$FILEPATH

Add dynamic content [Alt+Shift+D]

Add dynamic content

Custom

\$\$FILEPATH

\$\$COLUMN

Microsoft Azure | Data Factory | appfactory5000

Search

techsup1000@gmail.com
DEFAULT DIRECTORY

Data Factory | databricks | customer | CustomerFiles

Activities

- Move & transform
 - Copy data
 - Data flow
- Azure Data Explorer
- Azure Function
- Batch Service
- Databricks
- Data Lake Analytics
- General
 - Append variable
 - Delete

General | Source | Sink

File path type ☐

Filter by last modified ☐

Recursively ☒

Enable partition discovery ☐

Max concurrent connections

Additional columns

Add dynamic content

@activity('GetMetadataAllFiles').output.ItemName

Clear contents

Add dynamic content above using any combination of expressions, functions and system variables. Click any of the available System variables or Functions below to add them directly:

Filter system variables and functions...

- System variables
- Functions
- Activity outputs
 - GetMetadataAllFiles
 - GetMetadataAllFiles activity output

Microsoft Azure | Data Factory | appfactory5000

Search

techsup1000@gmail.com
DEFAULT DIRECTORY

Validate all Publish all 4

databricks customer CustomerFiles

Activities

- Move & transform
- Copy data
- Data flow
- Azure Data Explorer
- Azure Function
- Batch Service
- Databricks
- Data Lake Analytics
- General
- Append variable
- Delete
- Execute Pipeline

customer > ForEachFile

Get Metadata

GetMetadataAllFiles

Copy data

CopytoSynapse

General Source Sink Mapping Settings User properties

Enable partition discovery ☐

Max concurrent connections

Additional columns

NAME	VALUE
FileDate	@activity('GetMetadataFiles').output.la...
FileName	@activity('GetMetadataFiles').output.it...

Luego Validamos, Publicamos y Ejecutamos

Microsoft Azure | Data Factory | appfactory5000

Search

techsup1000@gmail.com
DEFAULT DIRECTORY

Validate all Publish all

databricks customer CustomerFiles

Activities

- Move & transform
- Copy data
- Data flow
- Azure Data Explorer
- Azure Function
- Batch Service
- Databricks
- Data Lake Analytics
- General
- Append variable
- Delete
- Execute Pipeline
- Execute SSIS package

customer > ForEachFile

Get Metadata

GetMetadataAllFiles

Copy data

CopytoSynapse

General Source Sink Mapping Settings User properties

Enable partition discovery ☐

Max concurrent connections

Additional columns

NAME	VALUE
FileDate	@activity('GetMetadataFiles').output.la...
FileName	@activity('GetMetadataFiles').output.it...

Trigger now

Trigger on-demand run of the last published pipeline

New/Edit

SQLQuery5.sql - appworkspace9000.sql.azuresynapse.net.newpool (sqladminuser (123)) - Microsoft SQL Server Management Studio (Administrat... Quick Launch (Ctrl+Q)

File Edit View Query Project Tools Window Help

New Query

vCREATE VIEW SelectColor AS

newpool

Execute

SQLQuery6.sql - ap... (sqladminuser (71))*

SQLQuery5.sql - ap...sqladminuser (123))*

Object Explorer

```
CREATE TABLE [dbo].[Customer]
(
    [customerid] int,
    [customername] varchar(20),
    [registered] bit,
    [FileDate] datetime,
    [FileName] varchar(200)
)

SELECT * FROM Customer
```

100 %

Results Messages

	customerid	customername	registered	FileDate	FileName
1	3	UserC	1	2021-07-23 20:25:36.000	customer2.json
2	4	UserD	1	2021-07-23 20:25:36.000	customer2.json
3	1	UserA	1	2021-07-23 20:25:36.000	customer1.json
4	2	UserB	1	2021-07-23 20:25:36.000	customer1.json

Query executed successfully.

appworkspace9000.sql.azures... sqladminuser (123) newpool 00:00:00 4 rows