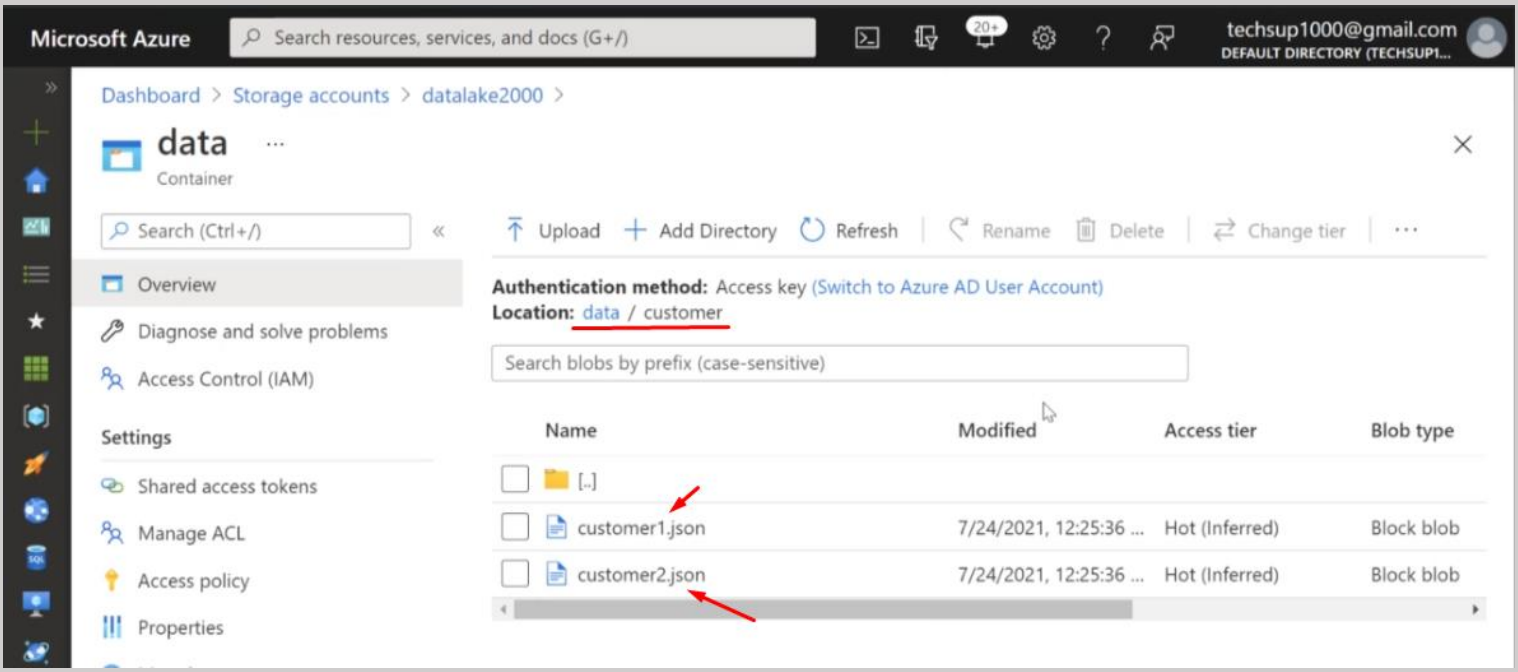
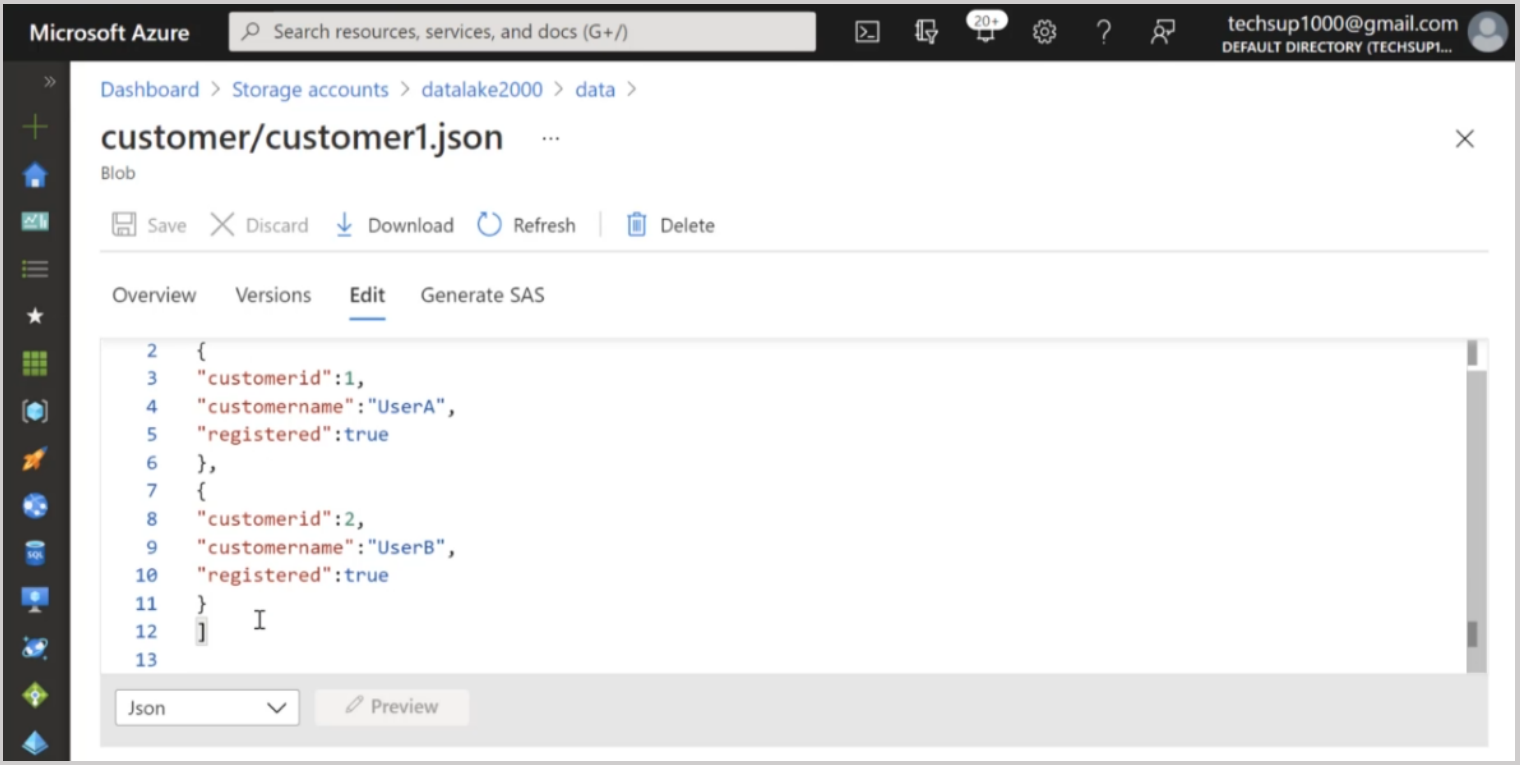
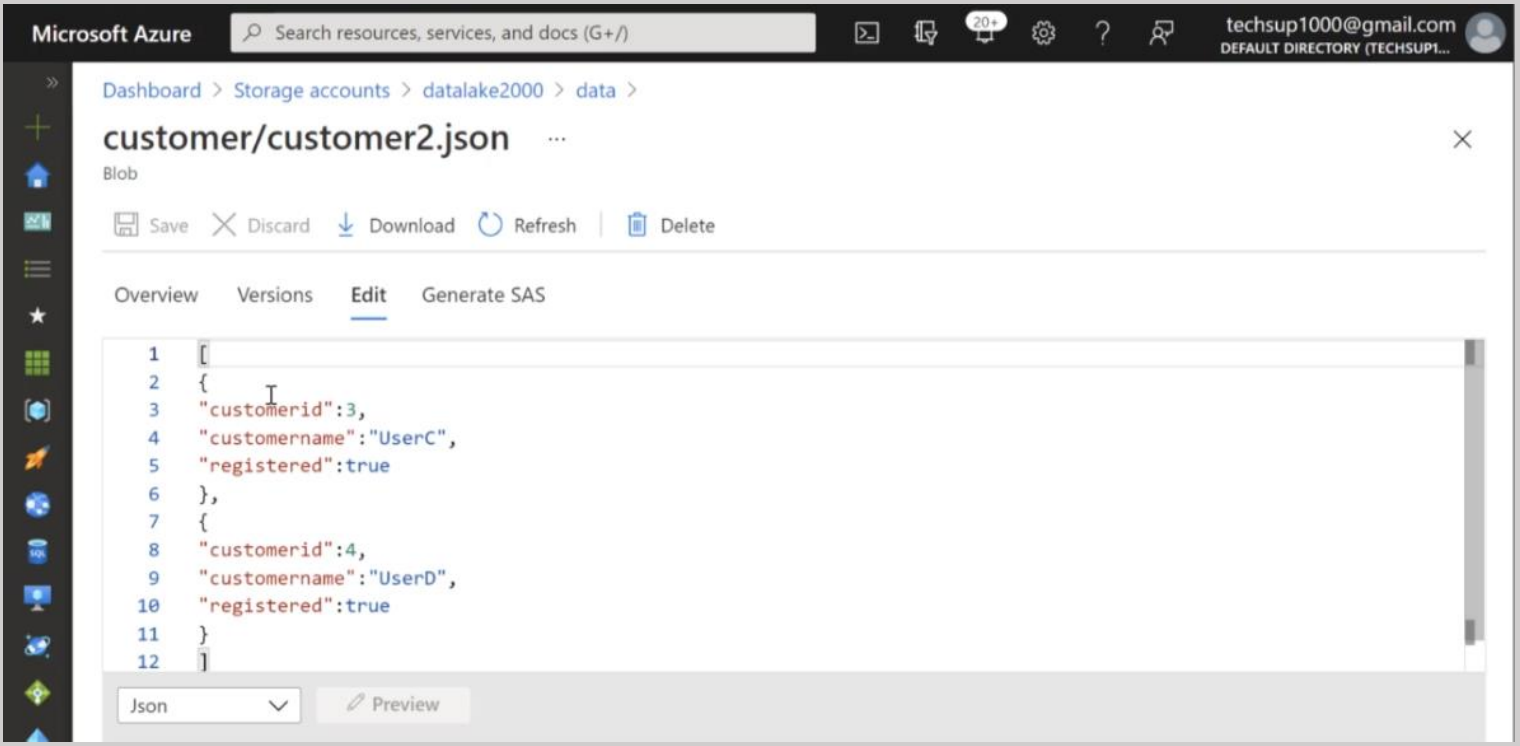
Azure Data Factory Cargar datos y metadata de dos JSON correlativos en un Dedicated SQL Pool

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



Los archivos tienen la siguiente estructura

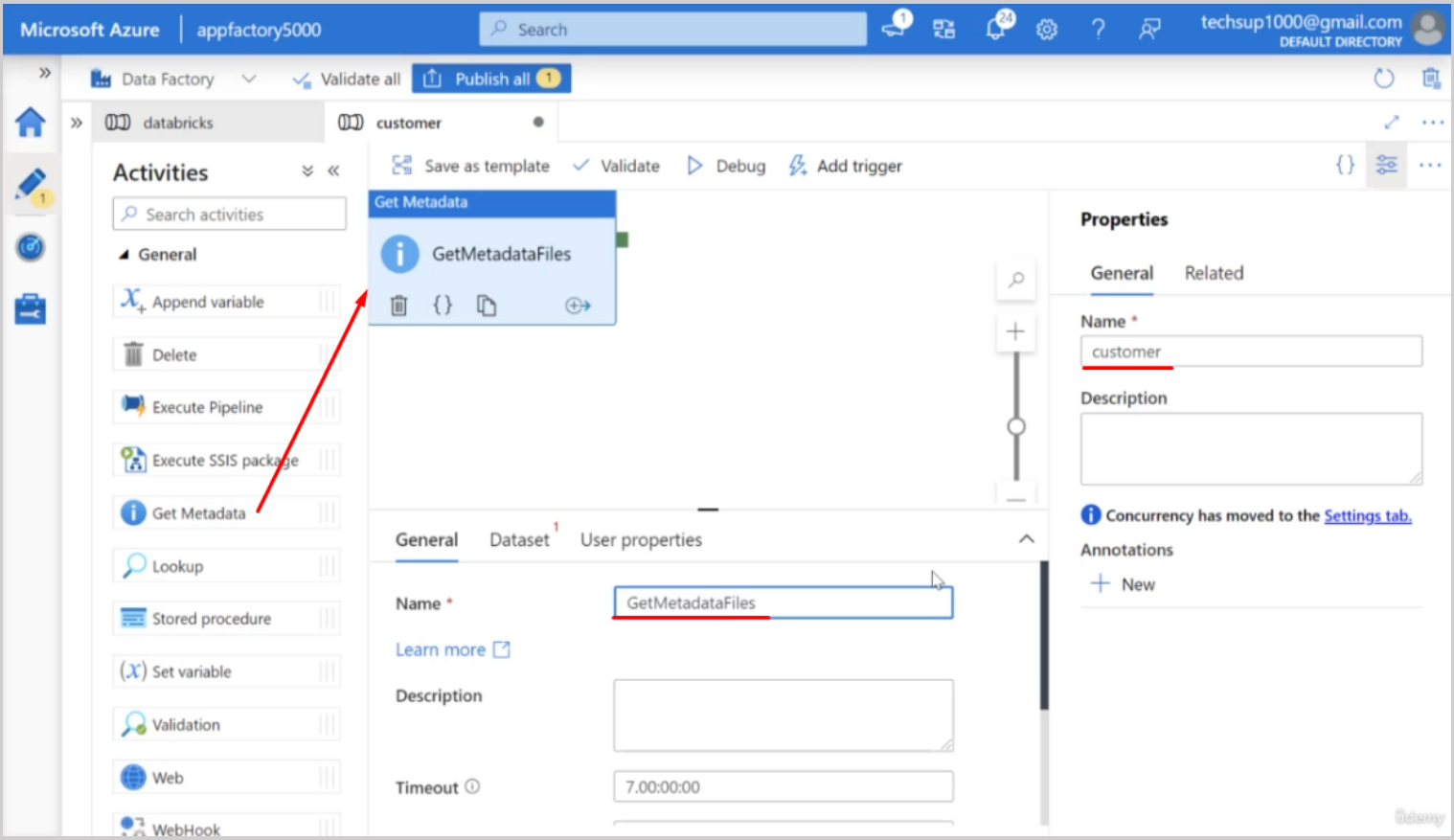


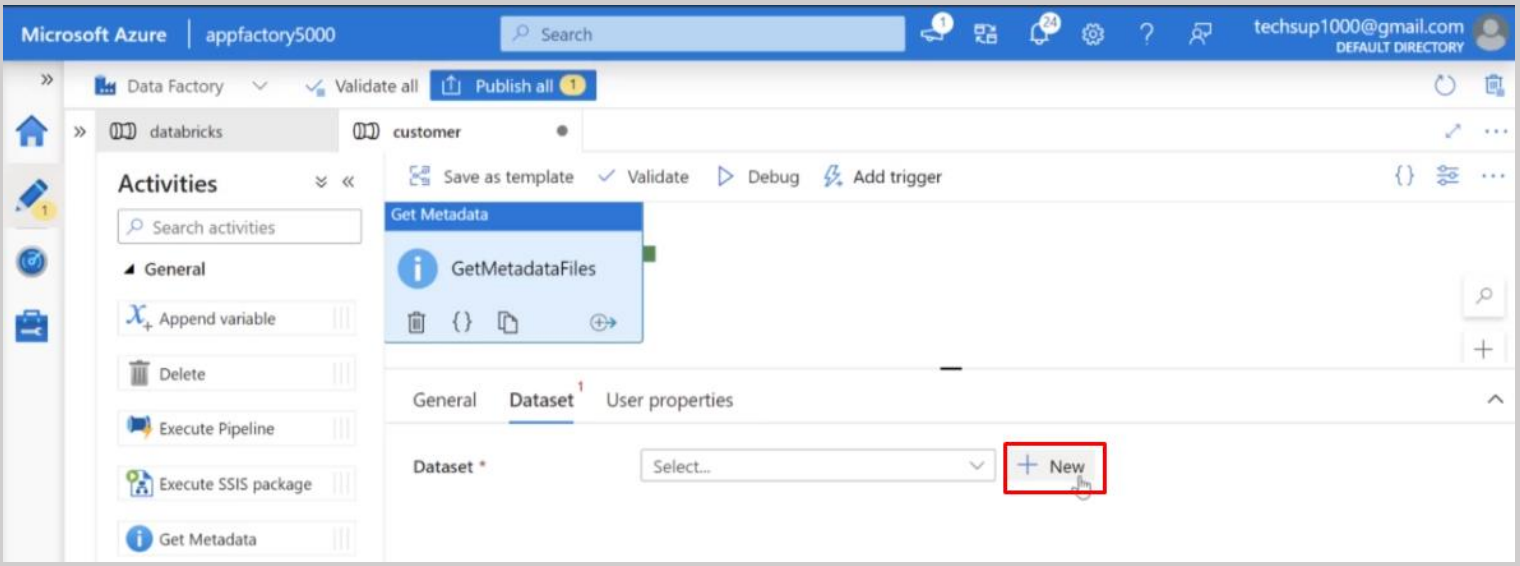


Se busca cargar sus datos en la siguiente tabla. Además queremos cargar la metadata correspondiente al nombre del archivo y la fecha en la que se realiza esta operación.

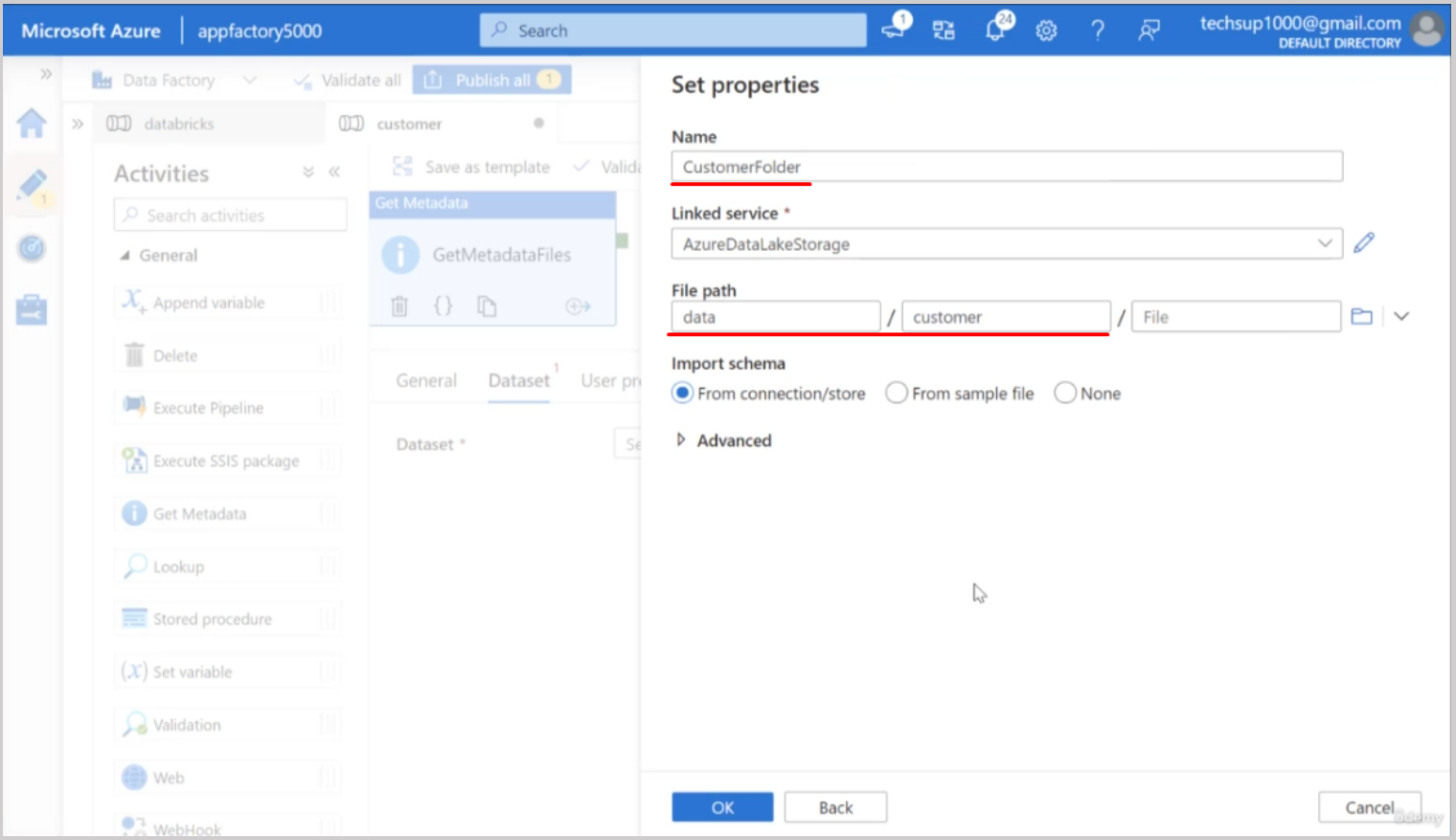
```sql
CREATE TABLE [dbo].[Customer]
(
  [customerid] int,
  [customername] varchar(20),
  [registered] bit,
  [FileDate] datetime,
  [FileName] varchar(200)
)
```

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

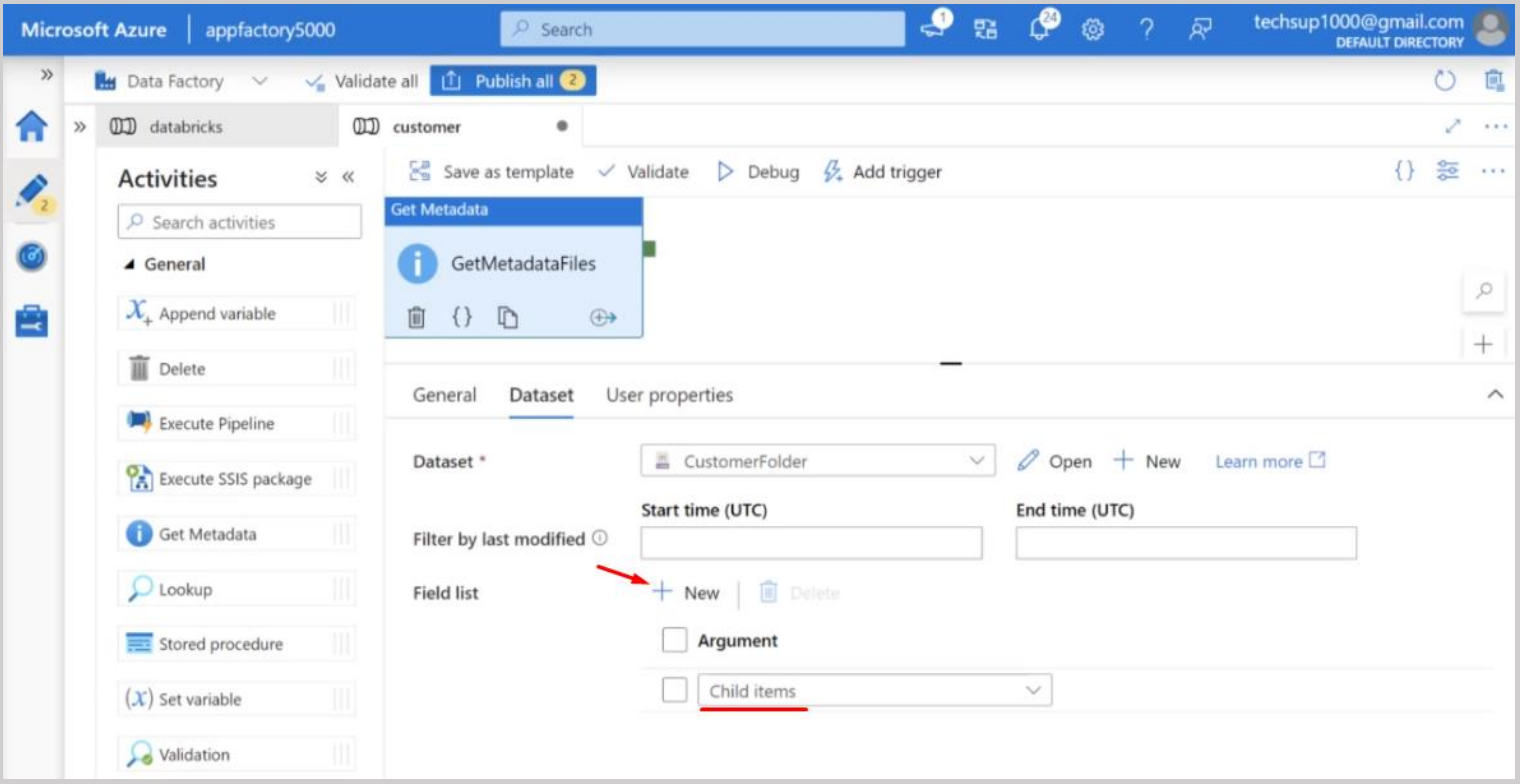


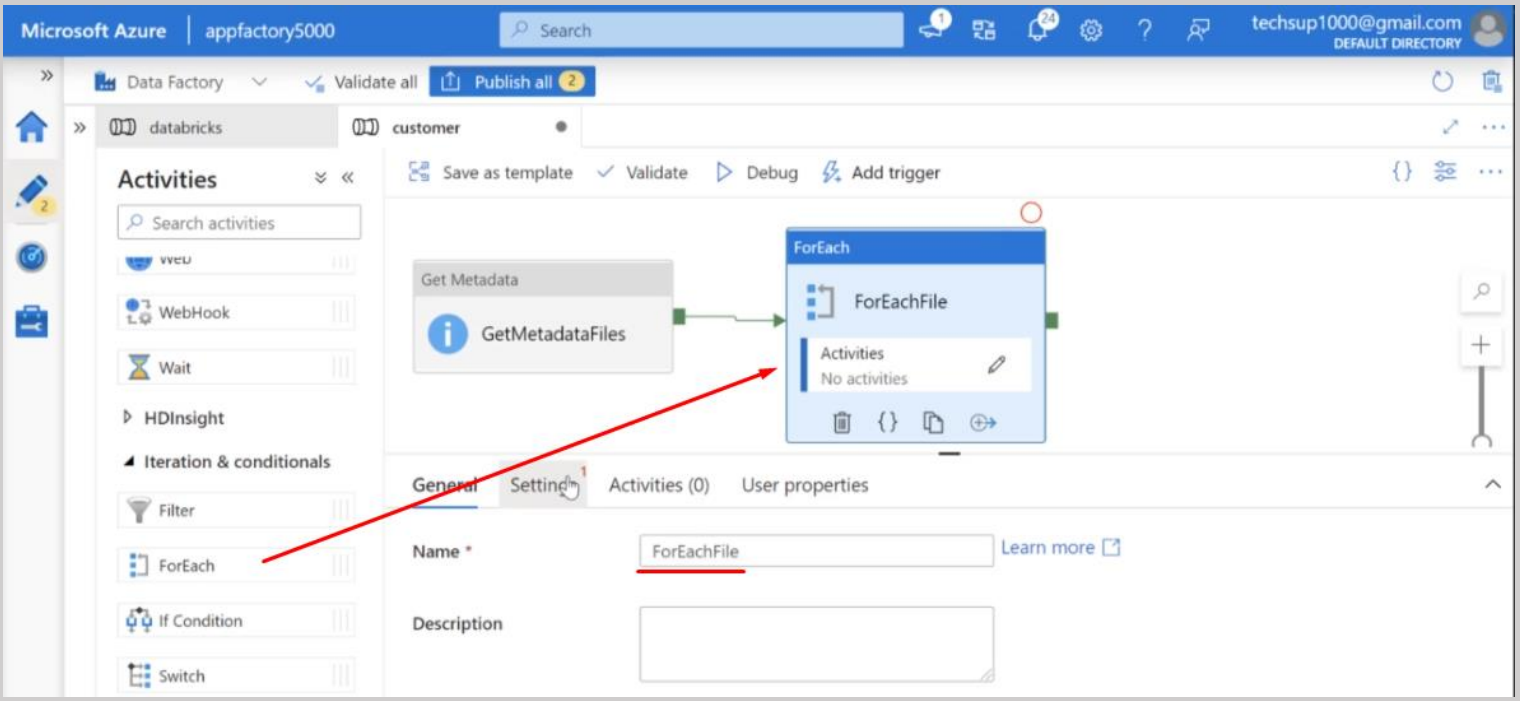


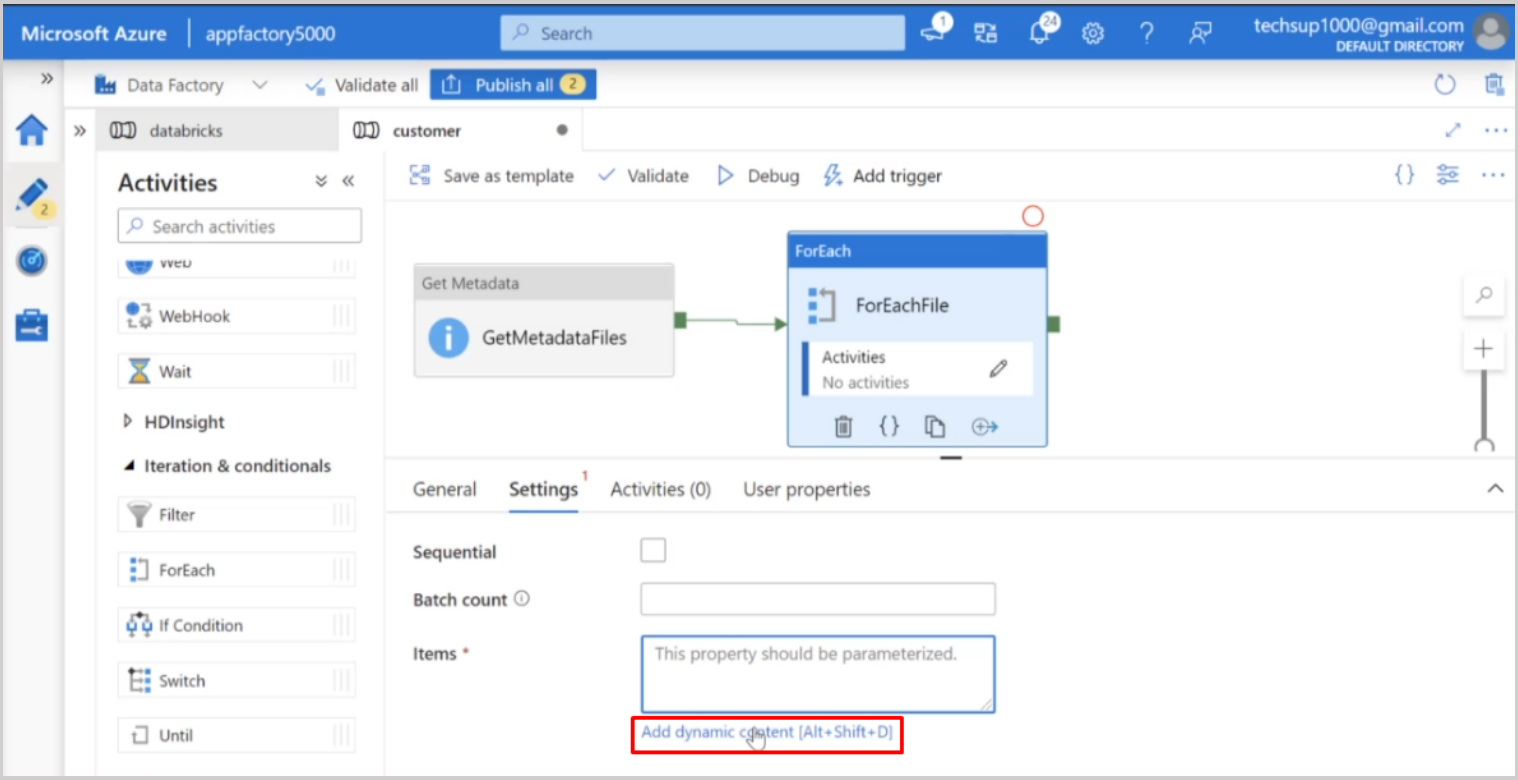
El dataset de la fuente hará referencia hacia Azure Data Lake Storage y el tipo de dato será JSON



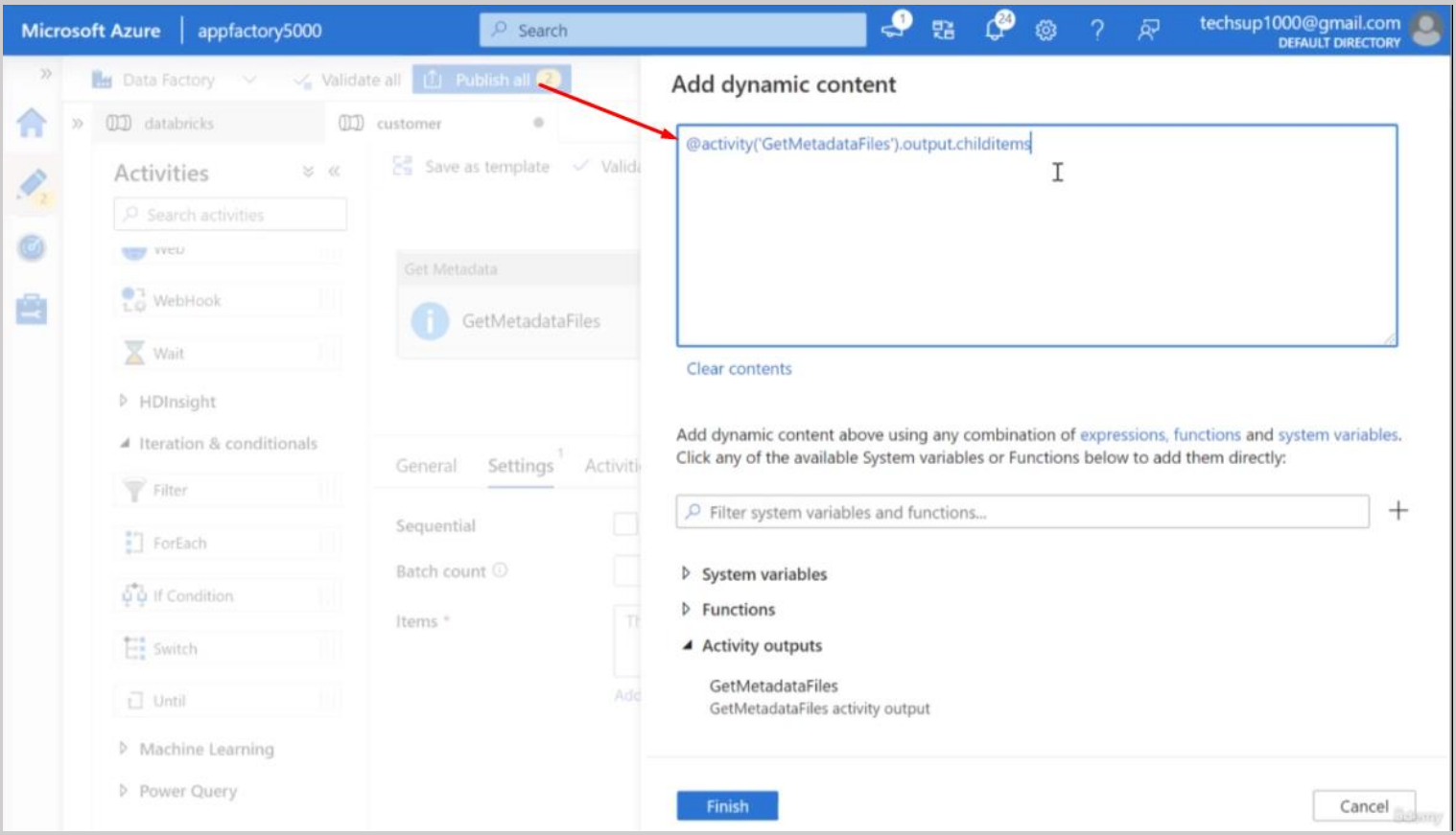
Ahora podemos obtener información sobre los archivos subyacentes en nuestra carpeta particular. Lo primero que quiero hacer es obtener los elementos. Así que quiero que esta actividad primero obtenga todos los elementos que están en mi carpeta "customer". Eso es lo primero que quiero hacer. Y luego,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

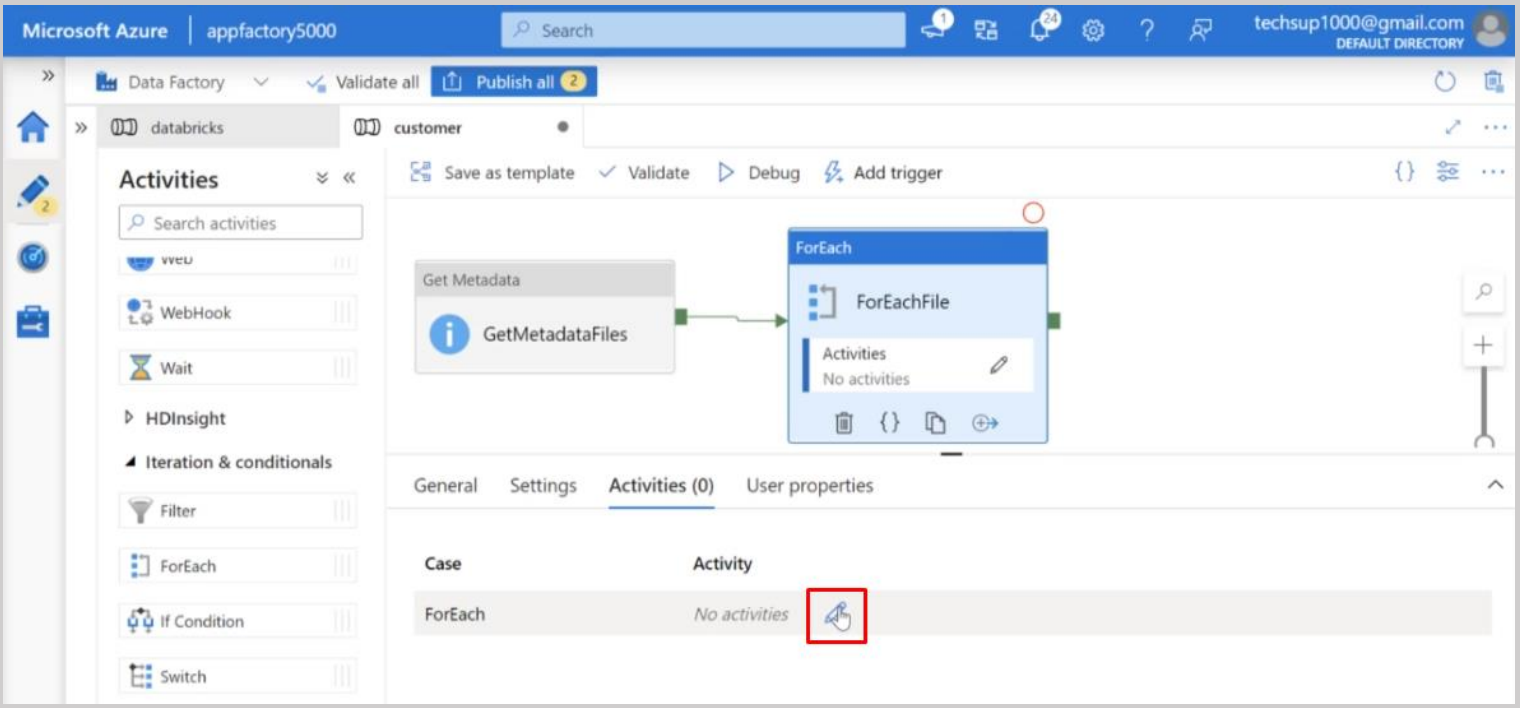


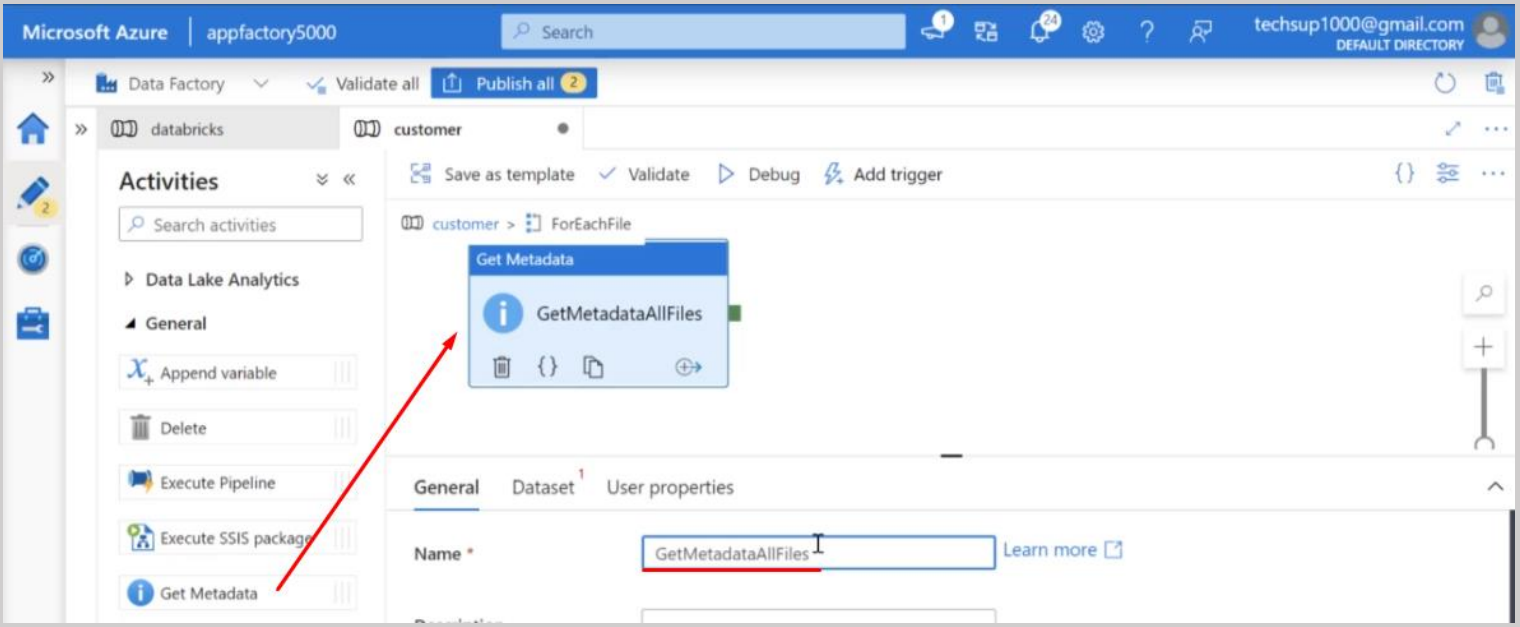


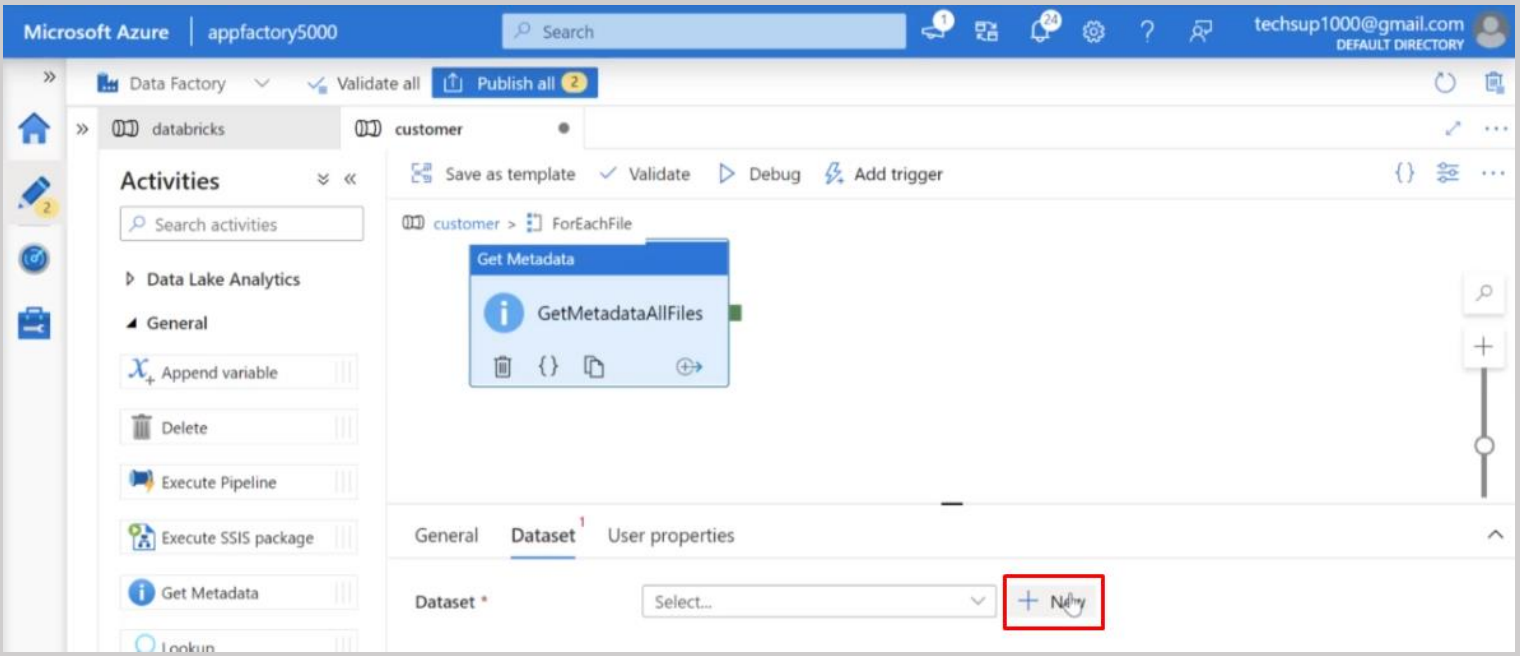


De la actividad anterior “GetMetadataFiles” esta rescatando los “childitems”, los elementos que capturó la actividad previa, que corresponderían a los dos archivos JSON.

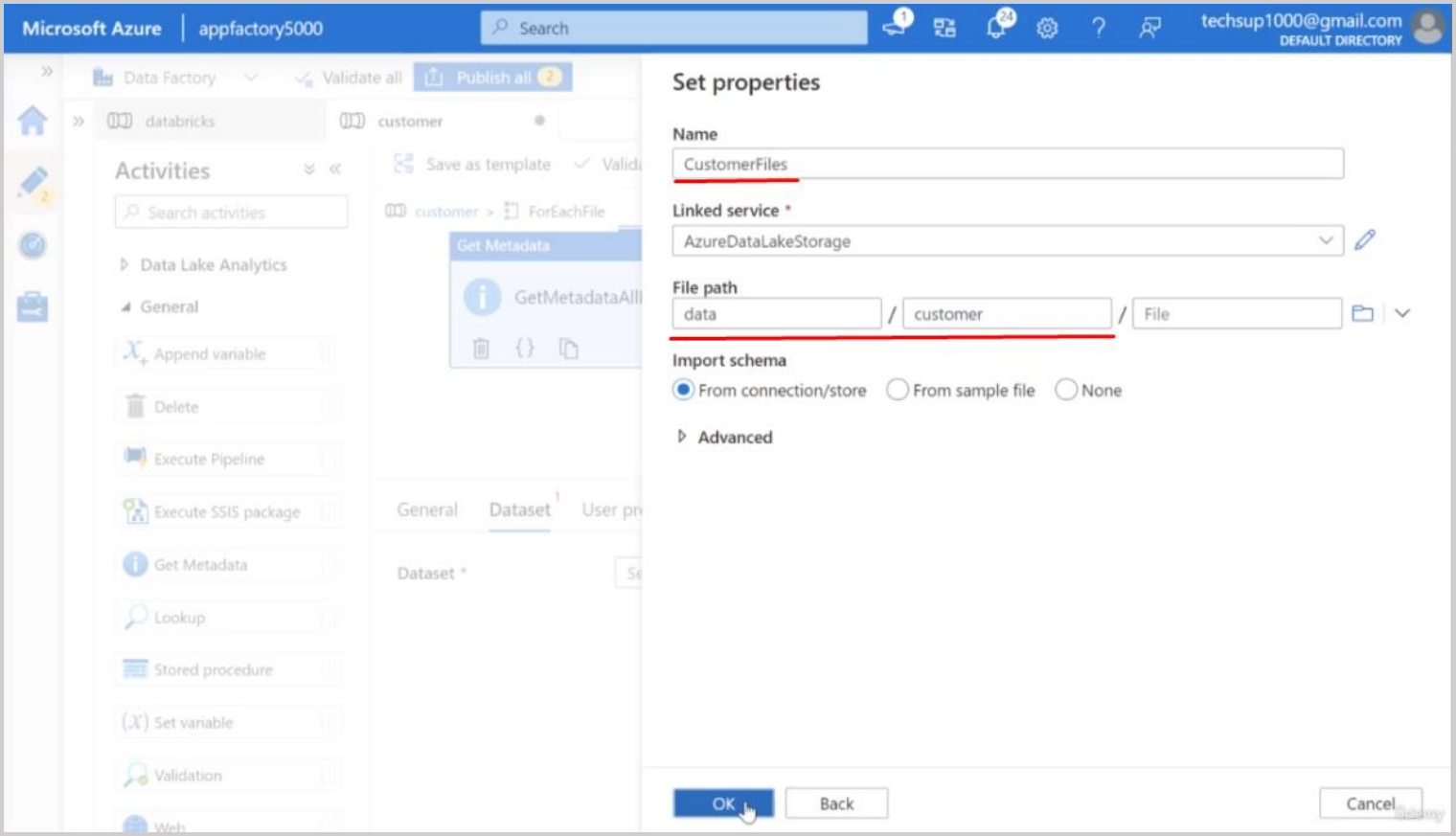


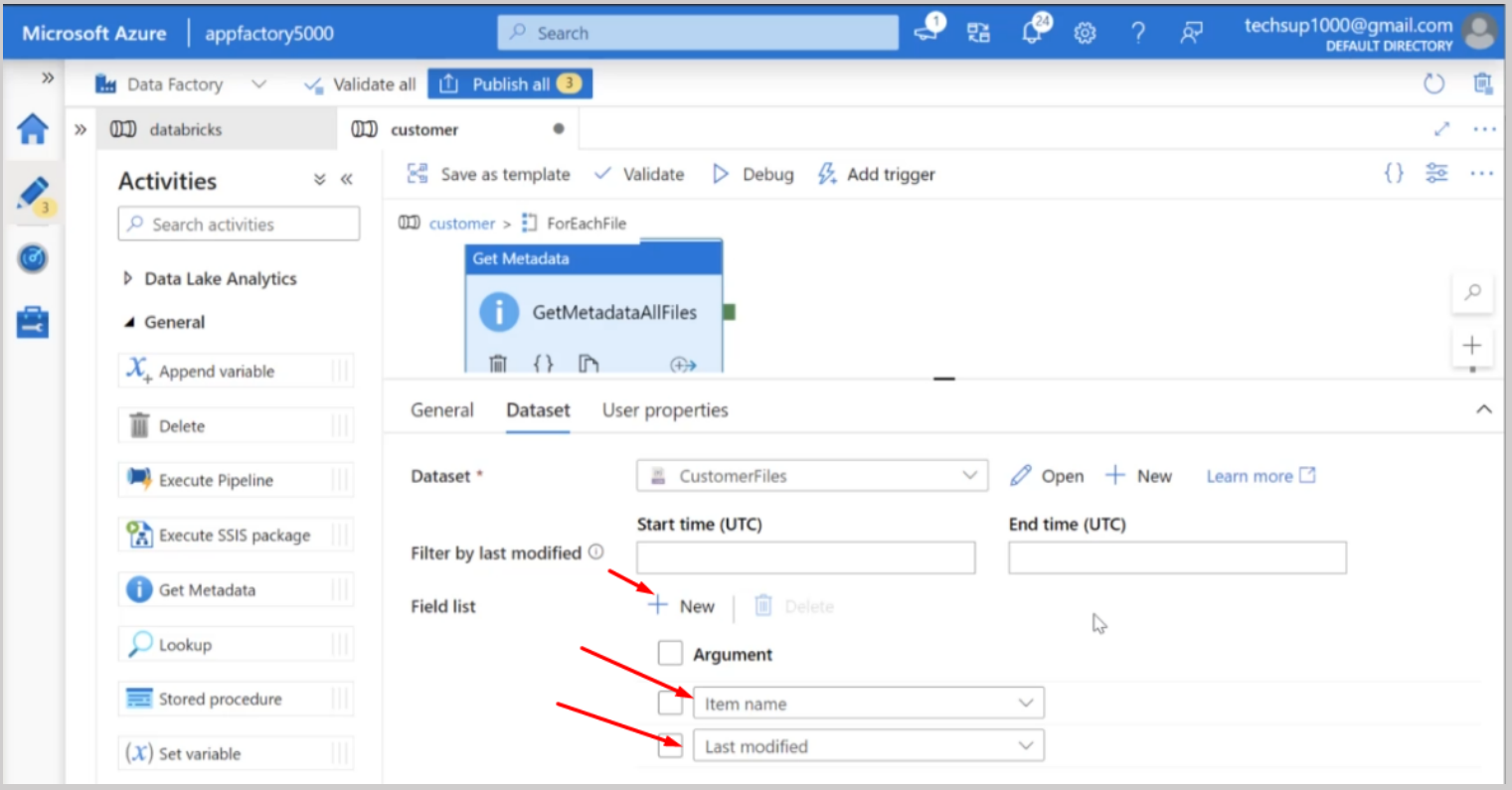

Microsoft Azure
appfactory5000
Search
techsup1000@gmail.com
DEFAULT DIRECTORY
Data Factory
Validate all
Publish all 2
databricks
customer
Activities
Search activities
Save as template
Validate
Debug
Add trigger
Get Metadata
GetMetadataFiles
ForEach
ForEachFile
Activities
No activities
WebHook
Wait
HDInsight
Iteration & conditionals
Filter
ForEach
If Condition
Switch
General
Settings
Activities (0)
User properties
Case
Activity
ForEach
No activities

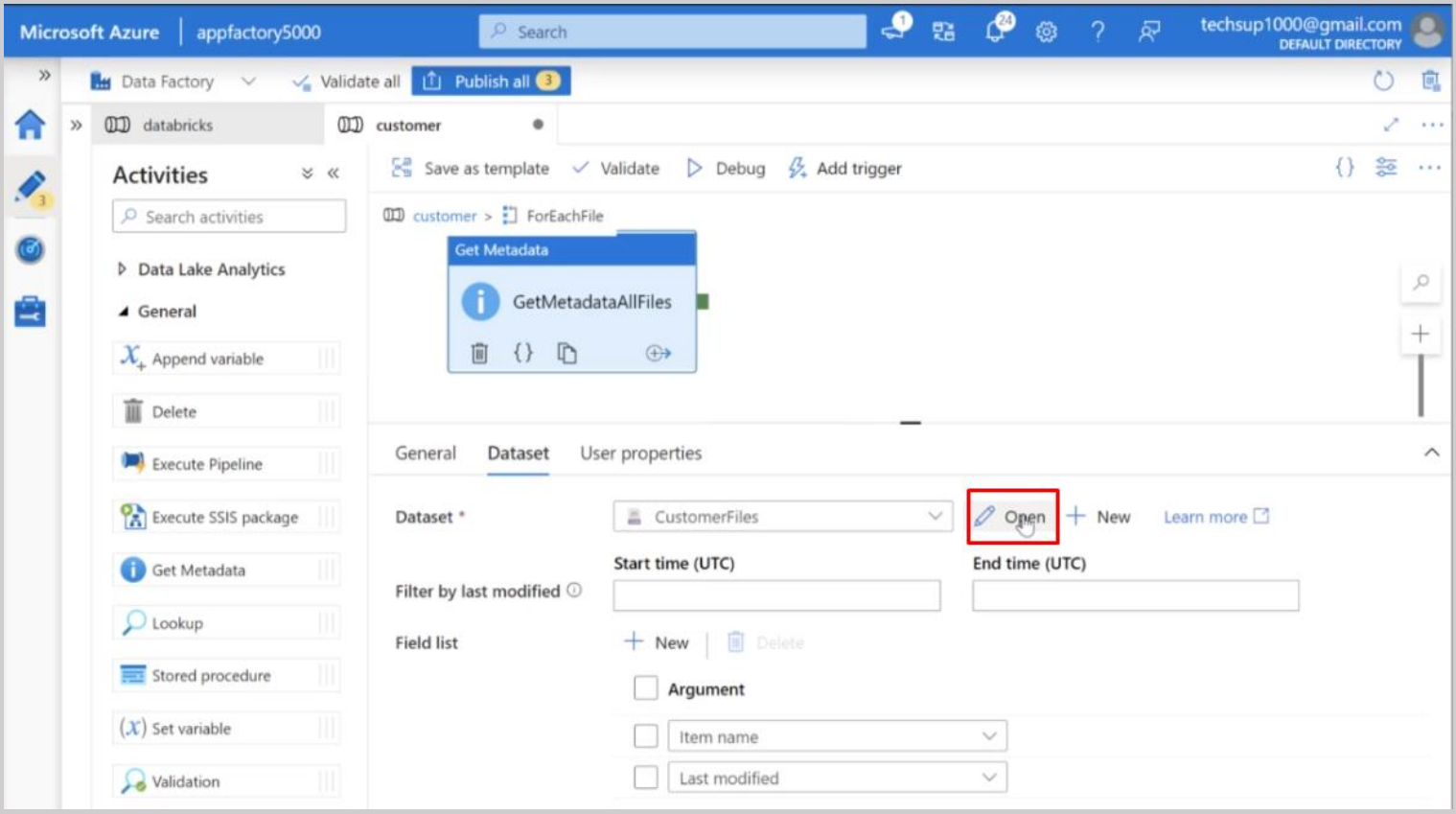

Microsoft Azure
appfactory5000
Search
techsup1000@gmail.com
DEFAULT DIRECTORY
Data Factory
Validate all
Publish all 2
databricks
customer
Activities
Search activities
Save as template
Validate
Debug
Add trigger
customer > ForEachFile
Get Metadata
GetMetadataAllFiles
Data Lake Analytics
General
Append variable
Delete
Execute Pipeline
Execute SSIS package
Get Metadata
General
Dataset 1
User properties
Name *
GetMetadataAllFiles
Learn more

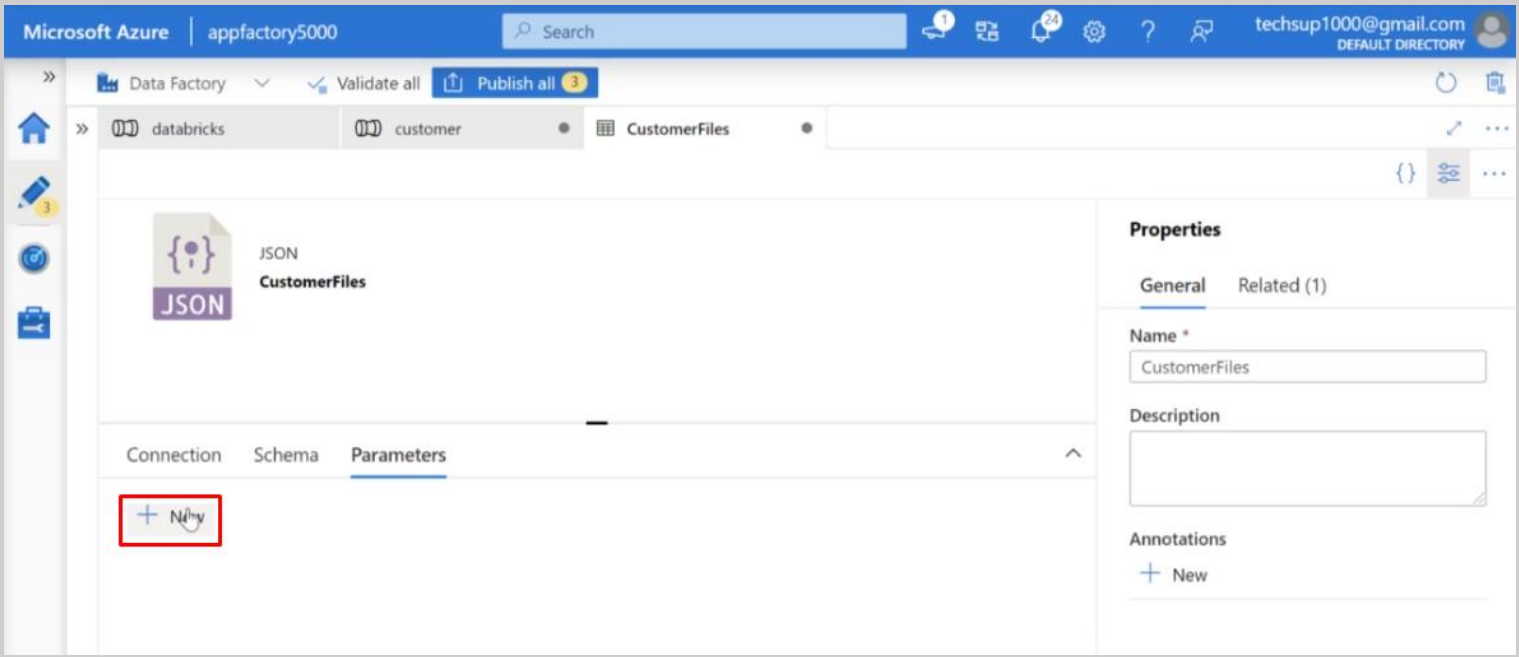

Microsoft Azure
appfactory5000
Search
techsup1000@gmail.com
DEFAULT DIRECTORY
Data Factory
Validate all
Publish all 2
databricks
customer
Activities
Search activities
Save as template
Validate
Debug
Add trigger
customer > ForEachFile
Get Metadata
GetMetadataAllFiles
Data Lake Analytics
General
Append variable
Delete
Execute Pipeline
Execute SSIS package
Get Metadata
Lookup
General
Dataset 1
User properties
Dataset *
Select...
+ New

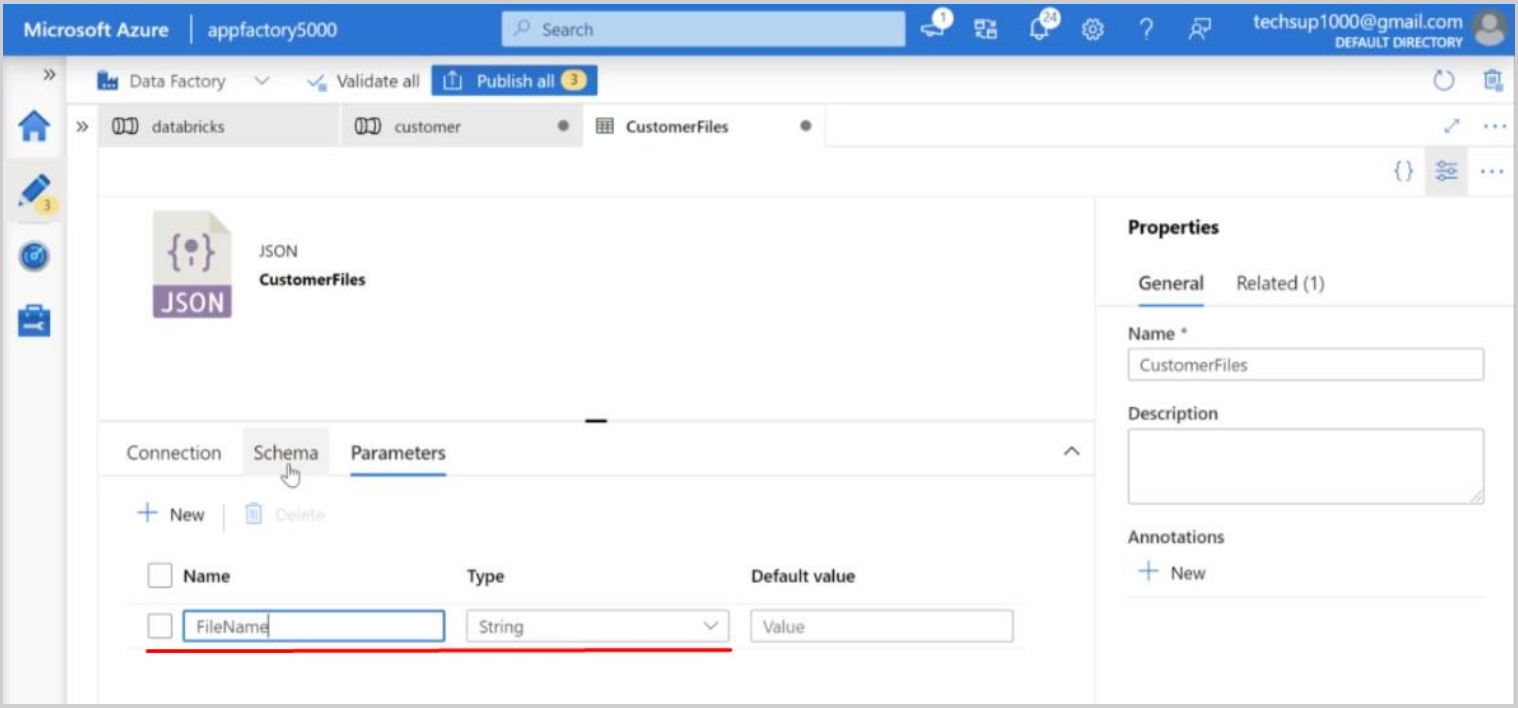
Creamos un nuevo dataset con referencia a Azure Data Lake Storage y seleccionamos como tipo de dato JSON

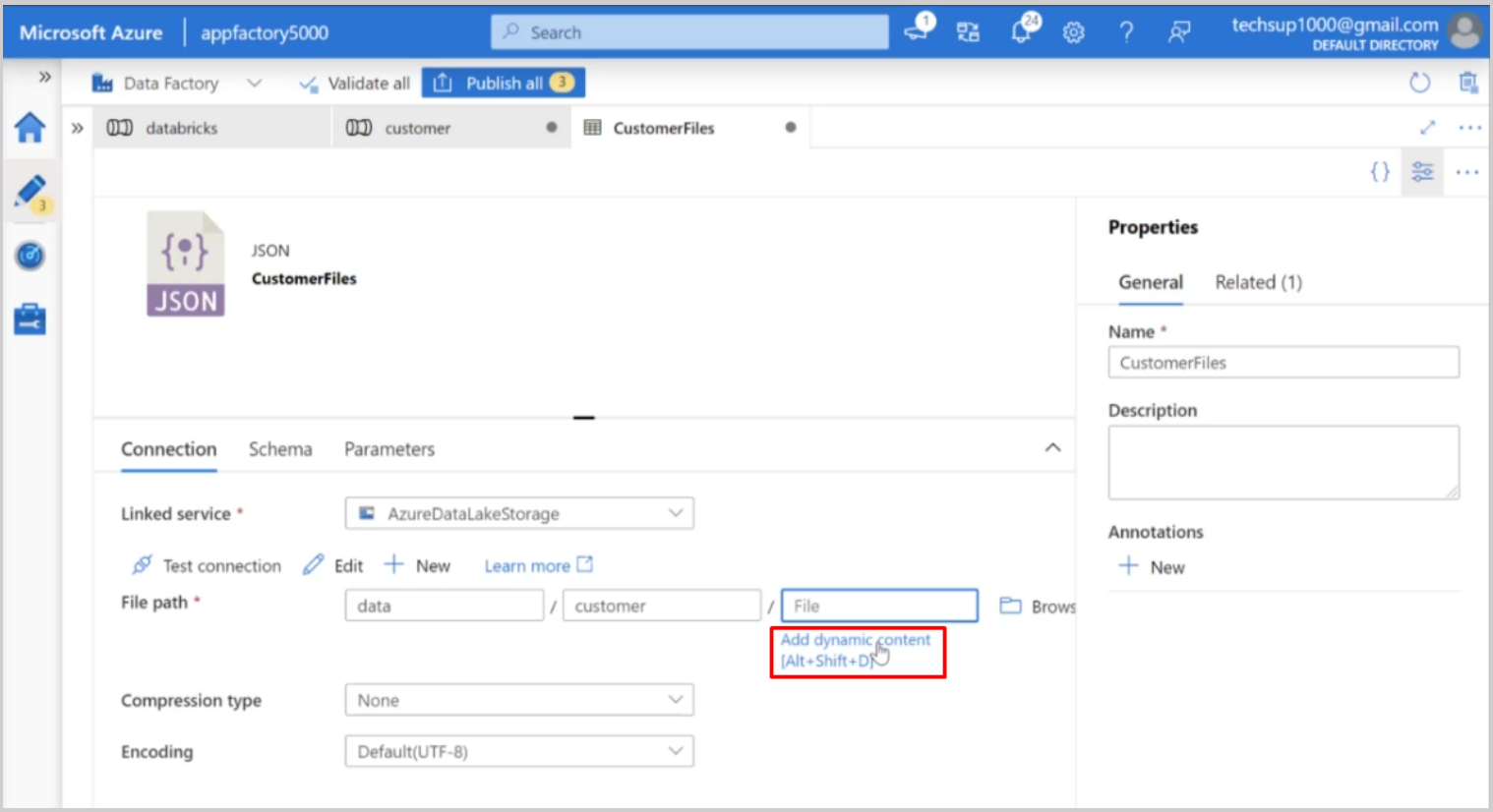


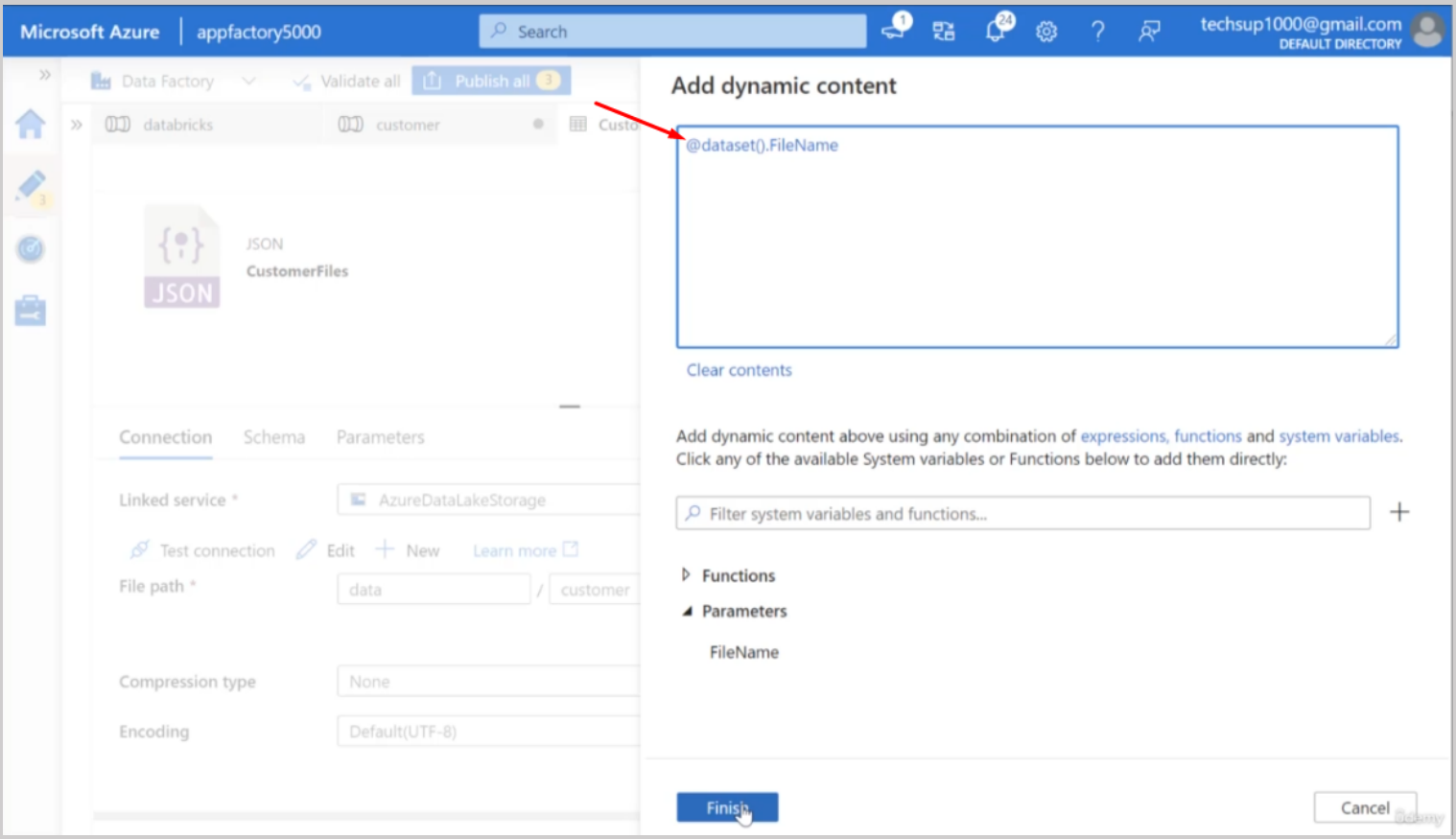



Microsoft Azure | appfactory5000
Search
techsup1000@gmail.com
DEFAULT DIRECTORY
Data Factory
Validate all
Publish all 3
databricks
customer
Activities
Search activities
Data Lake Analytics
General
Append variable
Delete
Execute Pipeline
Execute SSIS package
Get Metadata
Lookup
Stored procedure
Set variable
Validation
Save as template
Validate
Debug
Add trigger
customer > ForEachFile
Get Metadata
GetMetadataAllFiles
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

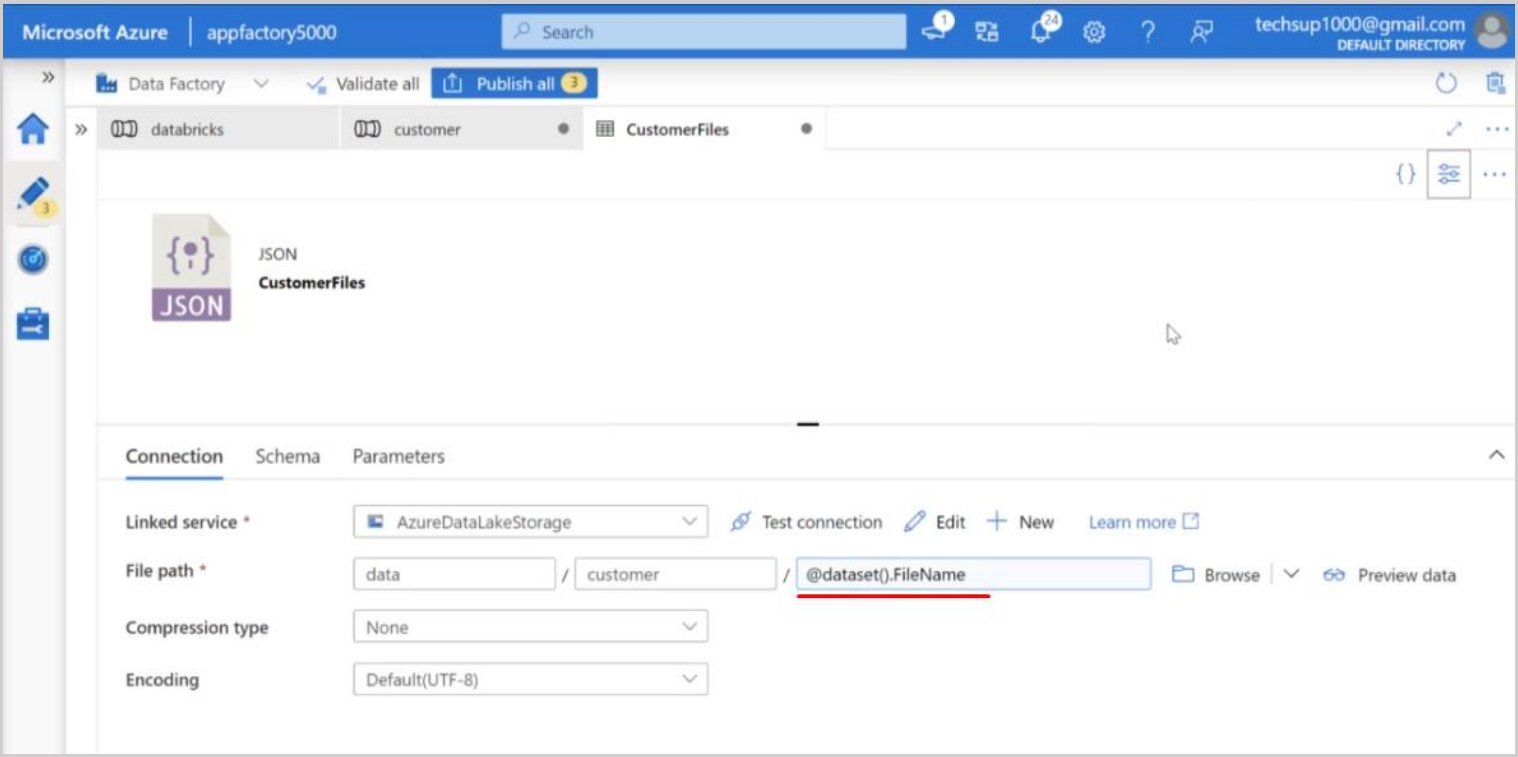

Microsoft Azure | appfactory5000
Search
techsup1000@gmail.com
DEFAULT DIRECTORY
Data Factory
Validate all
Publish all 3
databricks
customer
CustomerFiles
JSON
CustomerFiles
Connection
Schema
Parameters
New
Properties
General
Related (1)
Name *
CustomerFiles
Description
Annotations
New


Microsoft Azure
appfactory5000
Search
techsup1000@gmail.com
DEFAULT DIRECTORY
Data Factory
Validate all
Publish all 3
databricks
customer
CustomerFiles
JSON
CustomerFiles
Properties
General
Related (1)
Name *
CustomerFiles
Description
Annotations
New
Connection
Schema
Parameters
New
Delete
Name
Type
Default value
FileName
String
Value

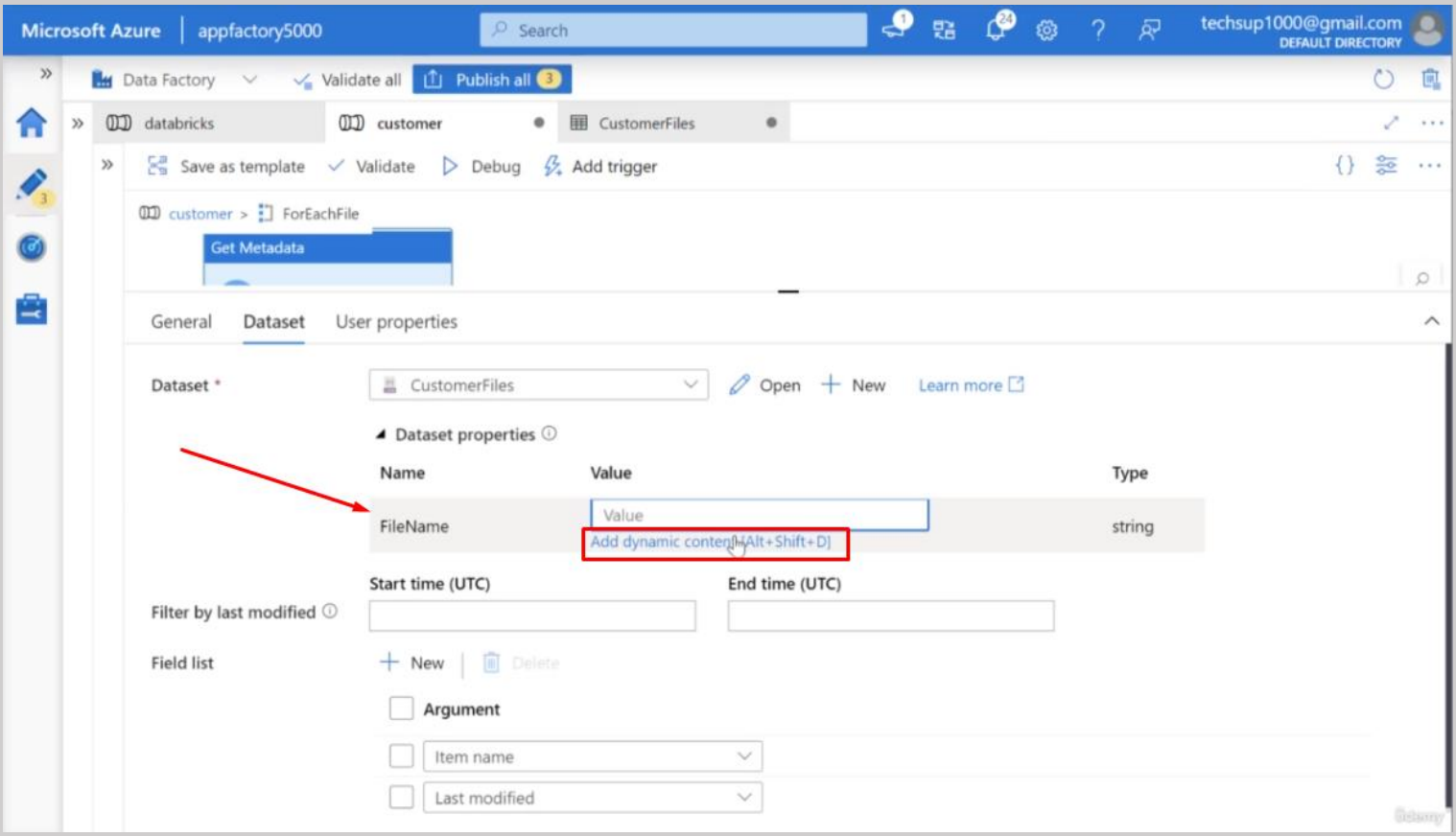

Microsoft Azure
appfactory5000
Search
techsup1000@gmail.com
DEFAULT DIRECTORY
Data Factory
Validate all
Publish all 3
databricks
customer
CustomerFiles
JSON
CustomerFiles
Properties
General
Related (1)
Name *
CustomerFiles
Description
Annotations
New
Connection
Schema
Parameters
Linked service *
AzureDataLakeStorage
Test connection
Edit
New
Learn more
File path *
data
customer
File
Brows
Add dynamic content
[Alt+Shift+D]
Compression type
None
Encoding
Default(UTF-8)

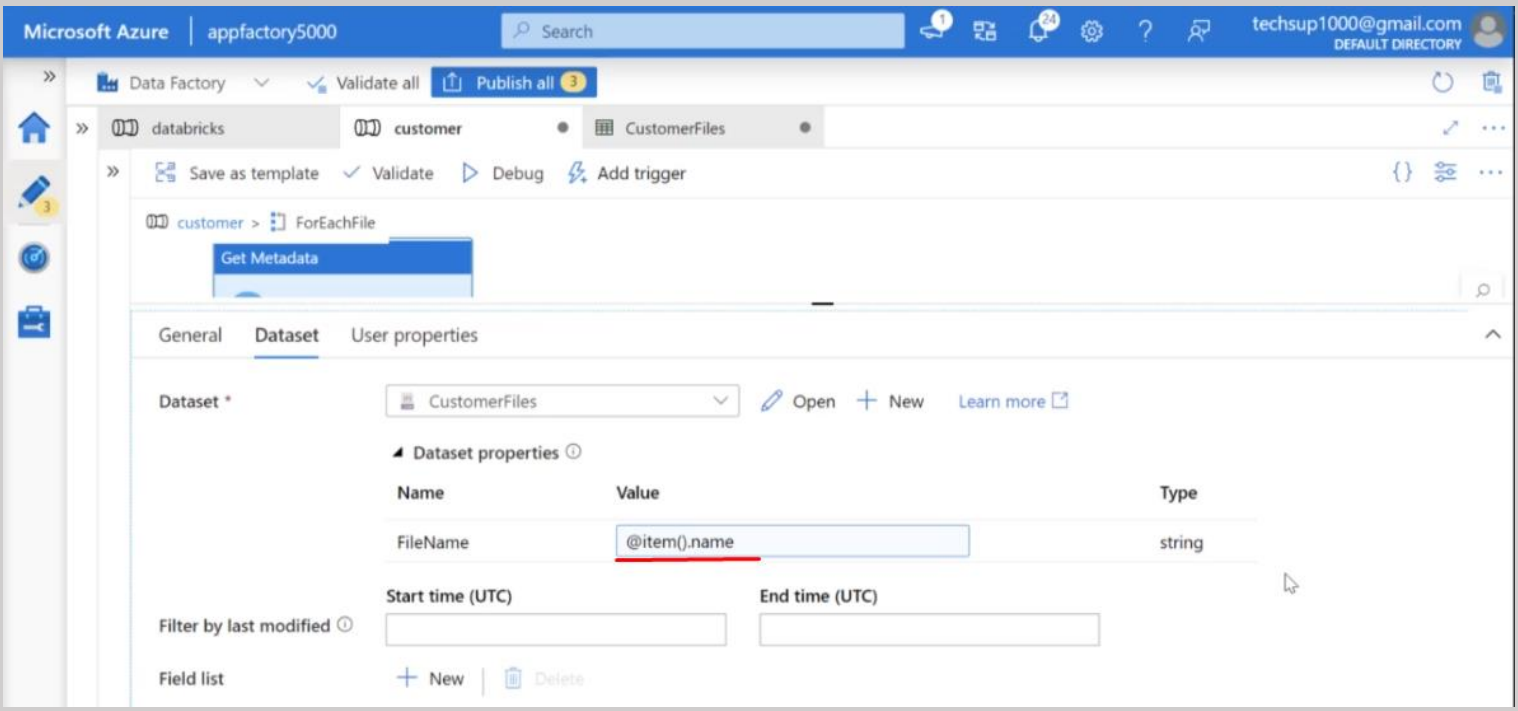


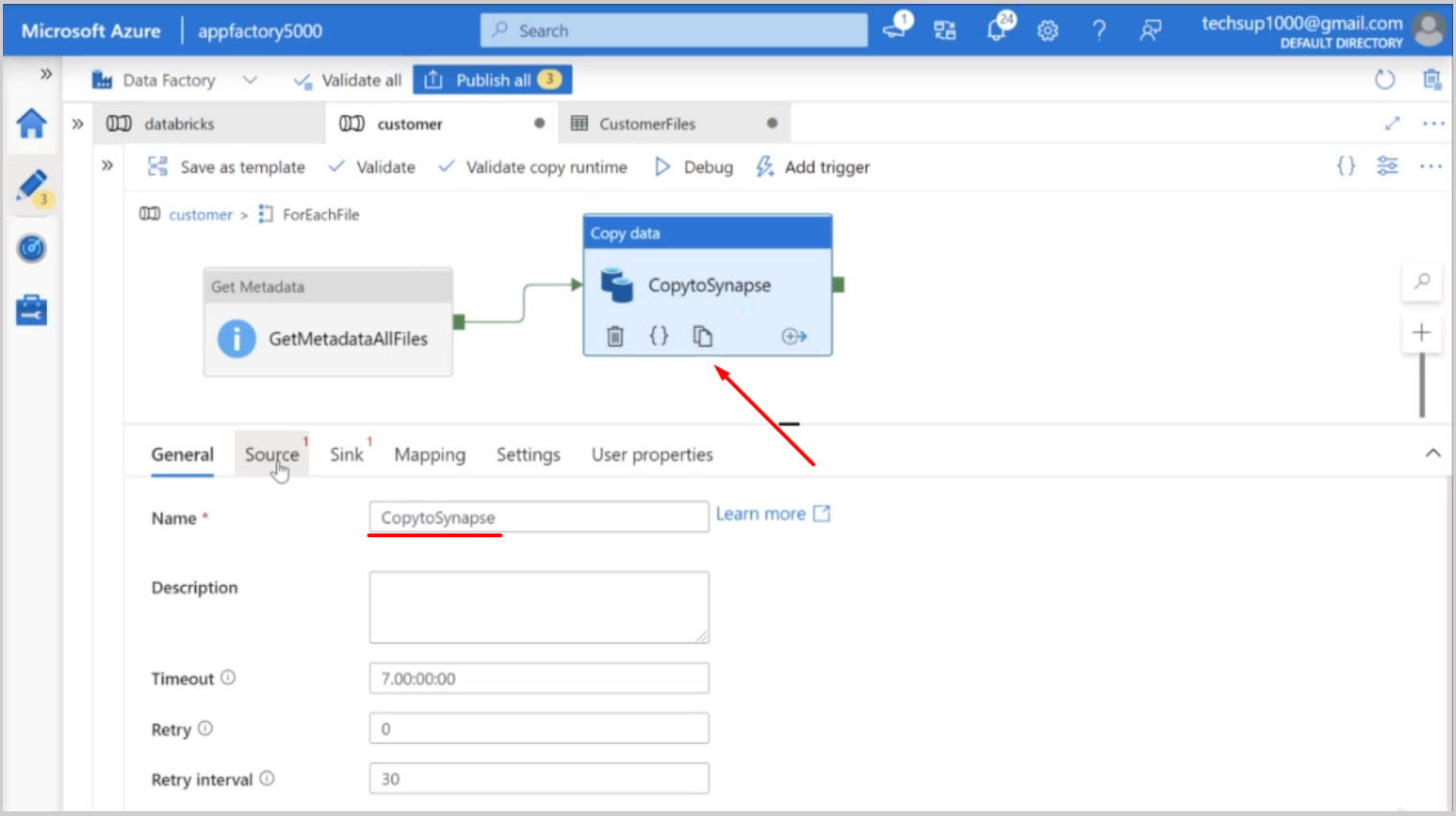
Así que aquí estamos tratando de hacer nuestro dataset más dinámico. Así que para cada archivo obtendrá automáticamente el nombre del archivo. Ahora bien, ¿cómo se obtiene realmente el nombre del archivo?

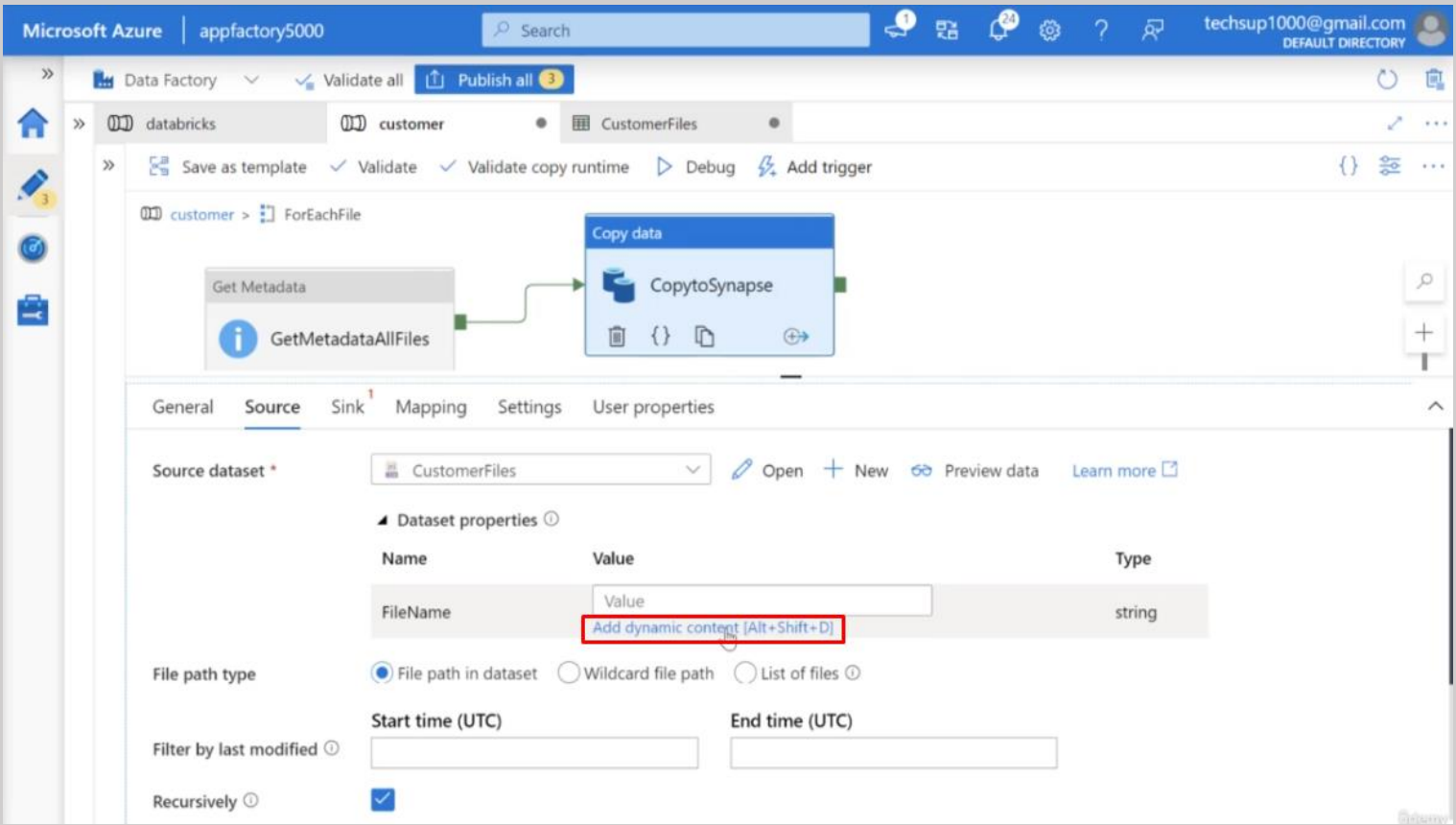


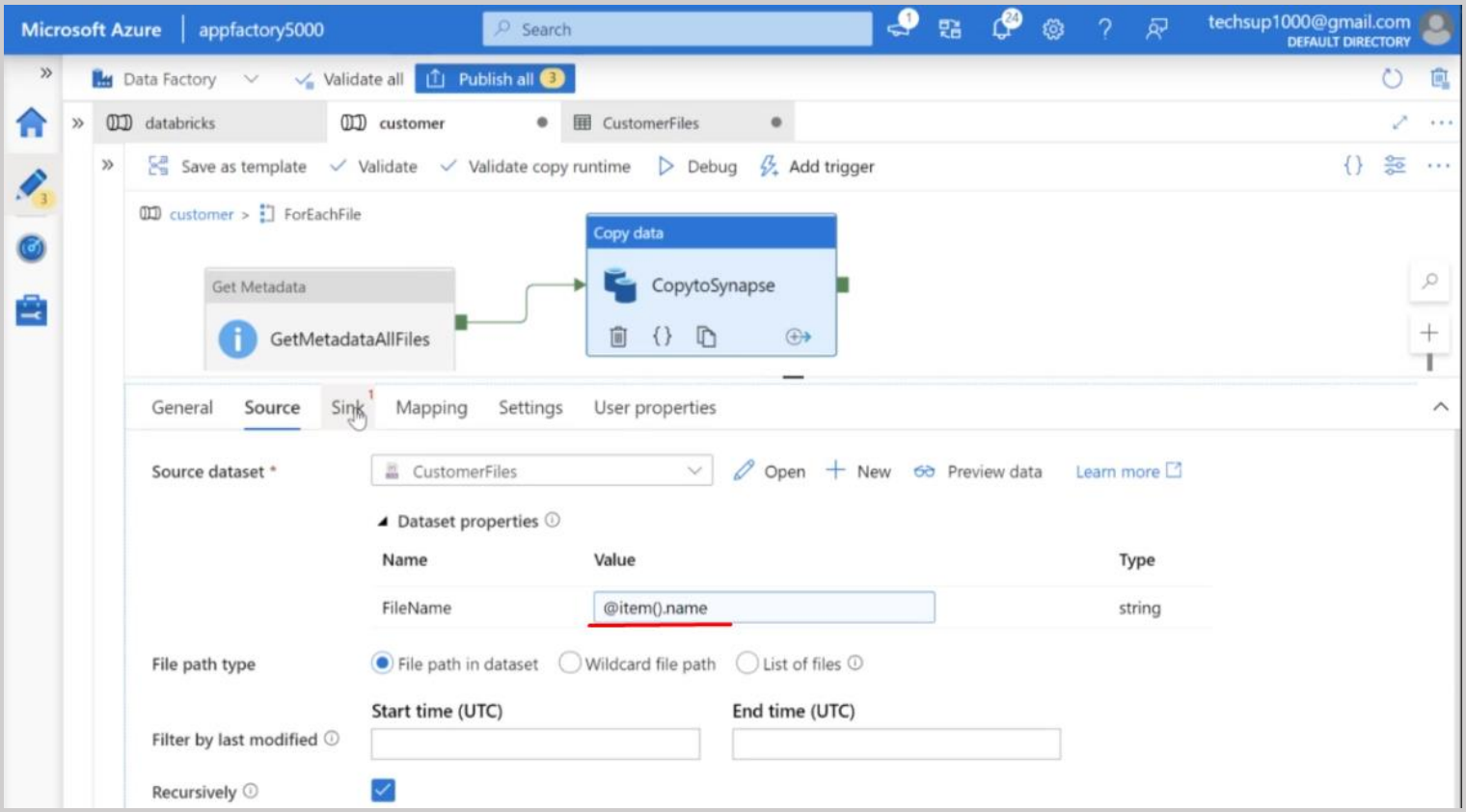
Si vuelvo ahora a mi pipeline "customer", puedes ver nuestro dataset customer. Puedes ver que se añade ahora una dataset property, por lo que ahora queremos pasar el nombre de archivo, tenemos una propiedad que luego será pasado a CustomerFiles y CustomerFiles puede ahora tratar sobre cada archivo en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

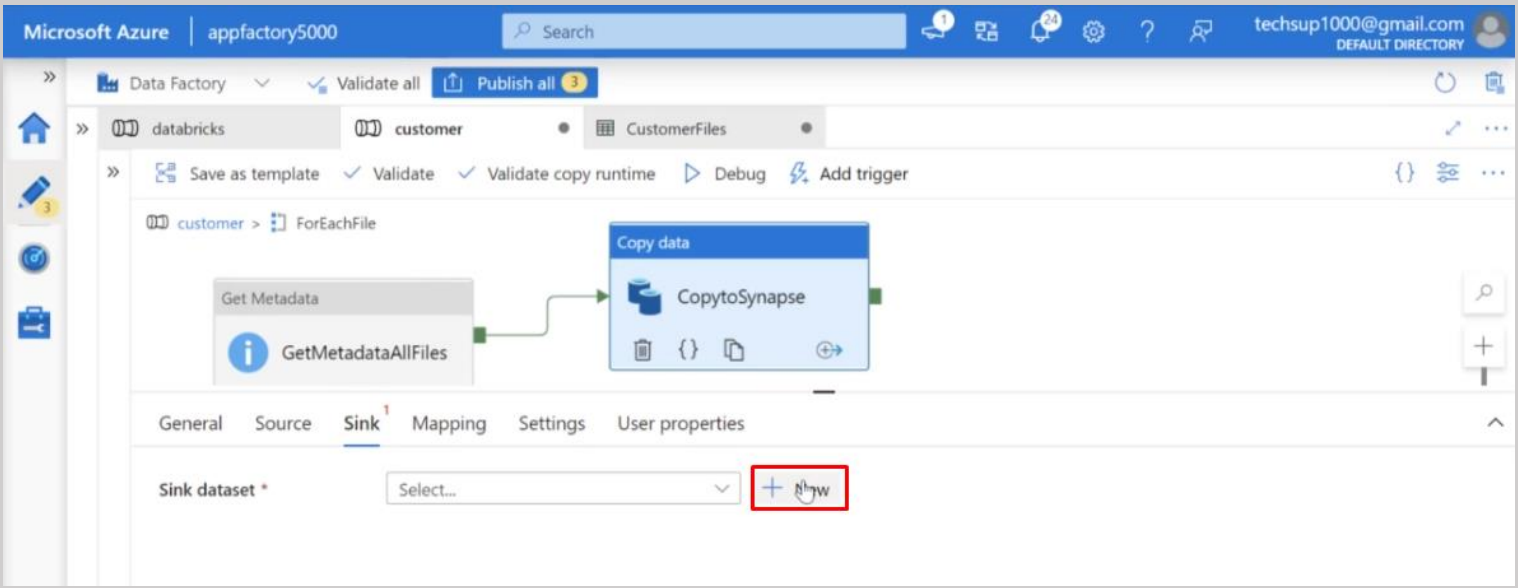


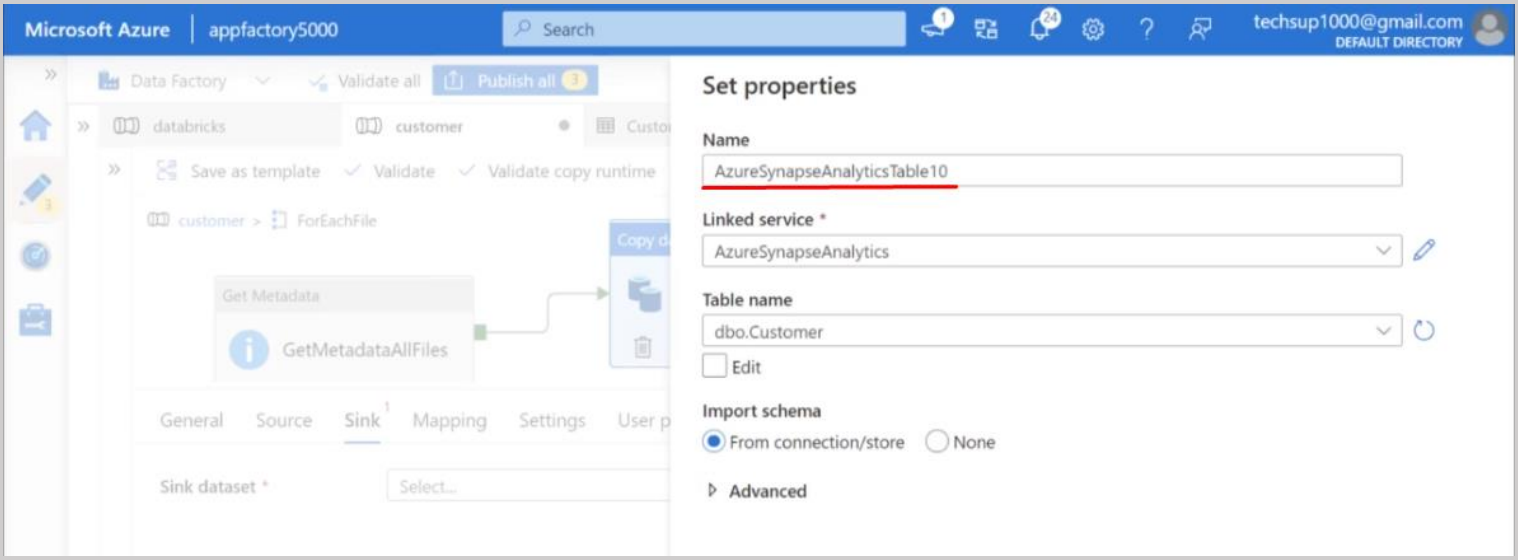


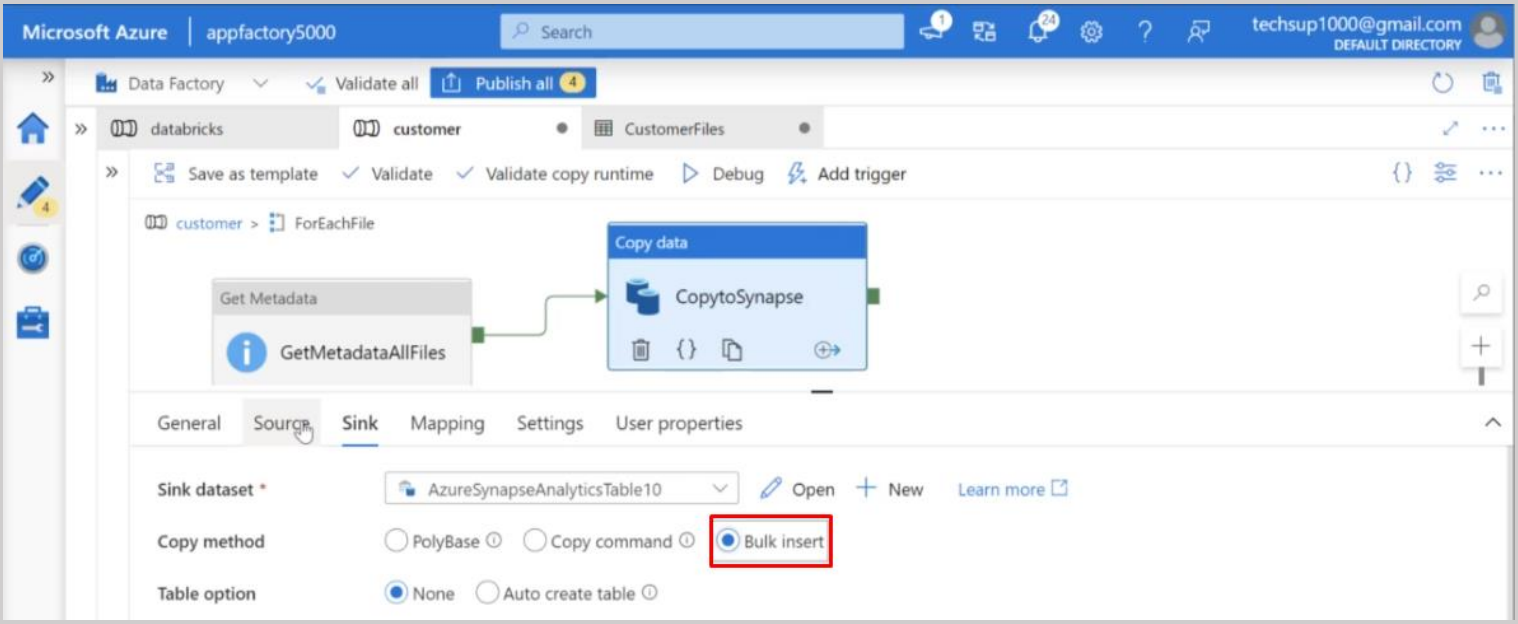

Microsoft Azure
appfactory5000
Search
techsup1000@gmail.com
DEFAULT DIRECTORY
Data Factory
Validate all
Publish all 3
databricks
customer
CustomerFiles
Save as template
Validate
Validate copy runtime
Debug
Add trigger
customer > ForEachFile
Get Metadata
GetMetadataAllFiles
Copy data
CopytoSynapse
General
Source 1
Sink 1
Mapping
Settings
User properties
Name *
CopytoSynapse
Learn more
Description
Timeout
7.00:00:00
Retry
0
Retry interval
30

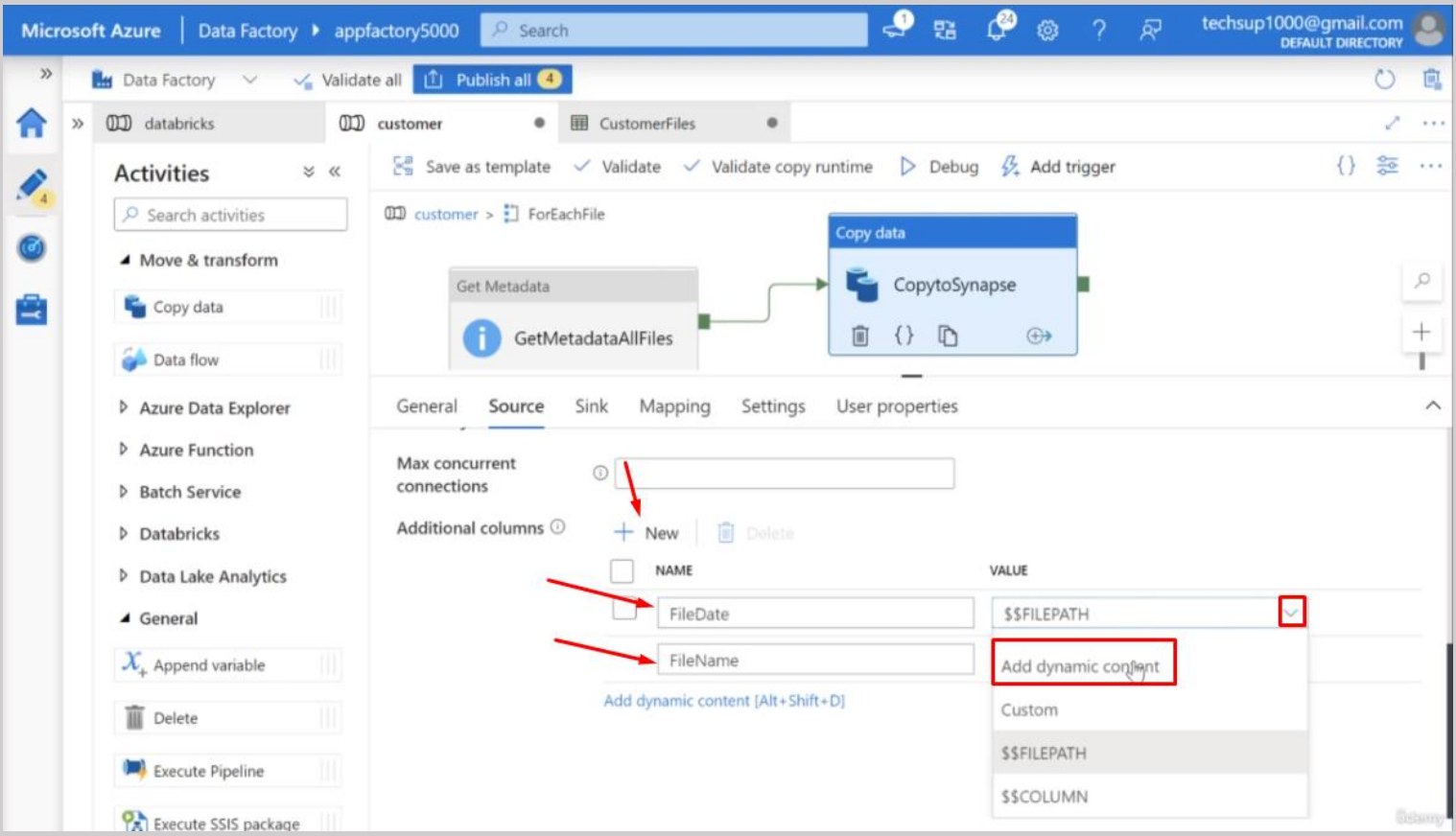

Microsoft Azure
appfactory5000
Search
techsup1000@gmail.com
DEFAULT DIRECTORY
Data Factory
Validate all
Publish all 3
databricks
customer
CustomerFiles
Save as template
Validate
Validate copy runtime
Debug
Add trigger
customer > ForEachFile
Get Metadata
GetMetadataAllFiles
Copy data
CopytoSynapse
General
Source
Sink 1
Mapping
Settings
User properties
Source dataset *
CustomerFiles
Open
New
Preview data
Learn more
Dataset properties
Name
Value
Type
FileName
Value
string
Add dynamic content [Alt+Shift+D]
File path type
File path in dataset
Wildcard file path
List of files
Start time (UTC)
End time (UTC)
Filter by last modified
Recursively

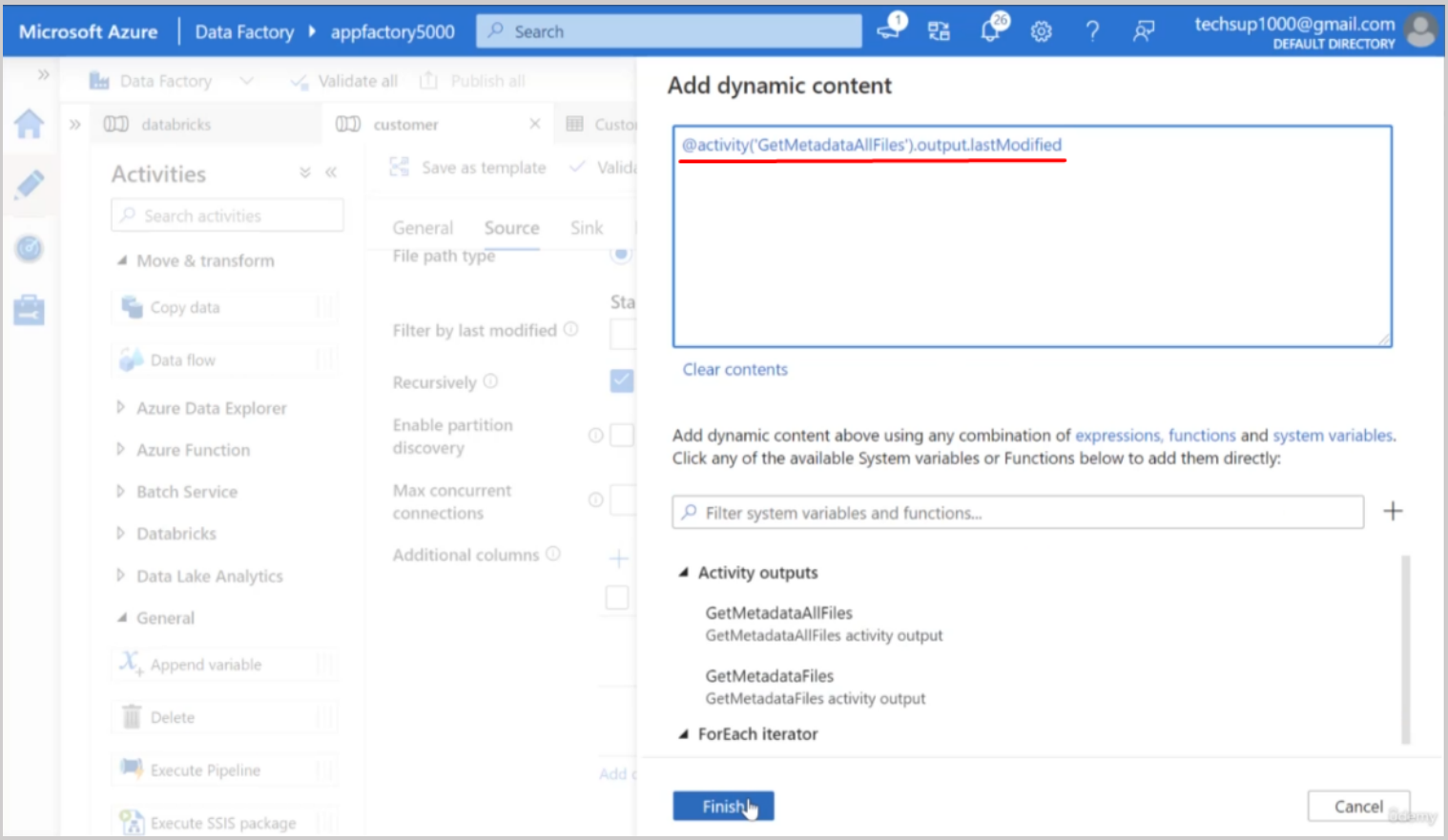

Microsoft Azure
appfactory5000
Search
techsup1000@gmail.com
DEFAULT DIRECTORY
Data Factory
Validate all
Publish all 3
databricks
customer
CustomerFiles
Save as template
Validate
Validate copy runtime
Debug
Add trigger
customer > ForEachFile
Get Metadata
GetMetadataAllFiles
Copy data
CopytoSynapse
General
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
Start time (UTC)
End time (UTC)
Filter by last modified
Recursively

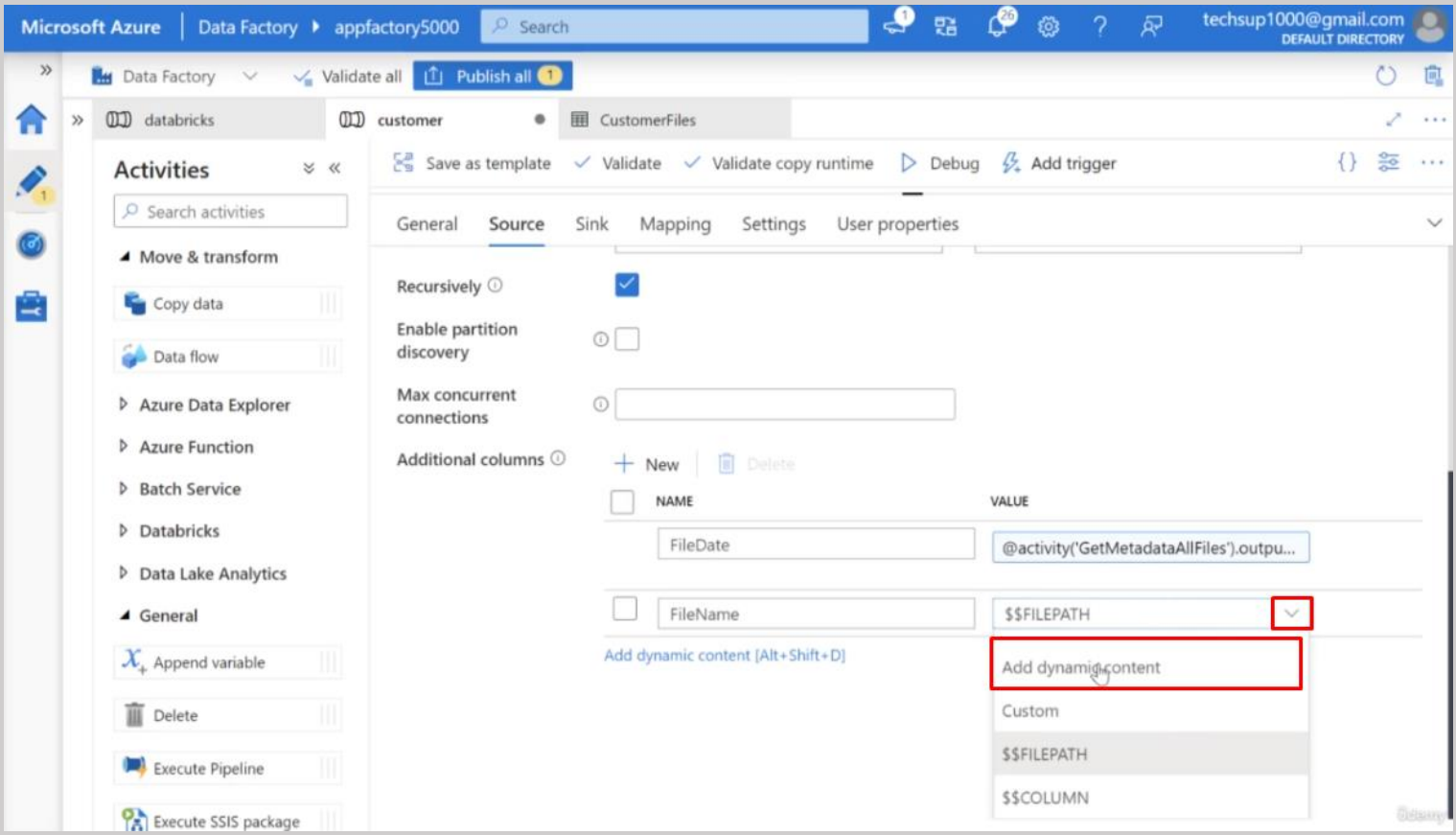

Microsoft Azure
appfactory5000
Search
techsup1000@gmail.com
DEFAULT DIRECTORY
Data Factory
Validate all
Publish all 3
databricks
customer
CustomerFiles
Save as template
Validate
Validate copy runtime
Debug
Add trigger
customer > ForEachFile
Get Metadata
GetMetadataAllFiles
Copy data
CopytoSynapse
General
Source
Sink
Mapping
Settings
User properties
Sink dataset *
Select...
New

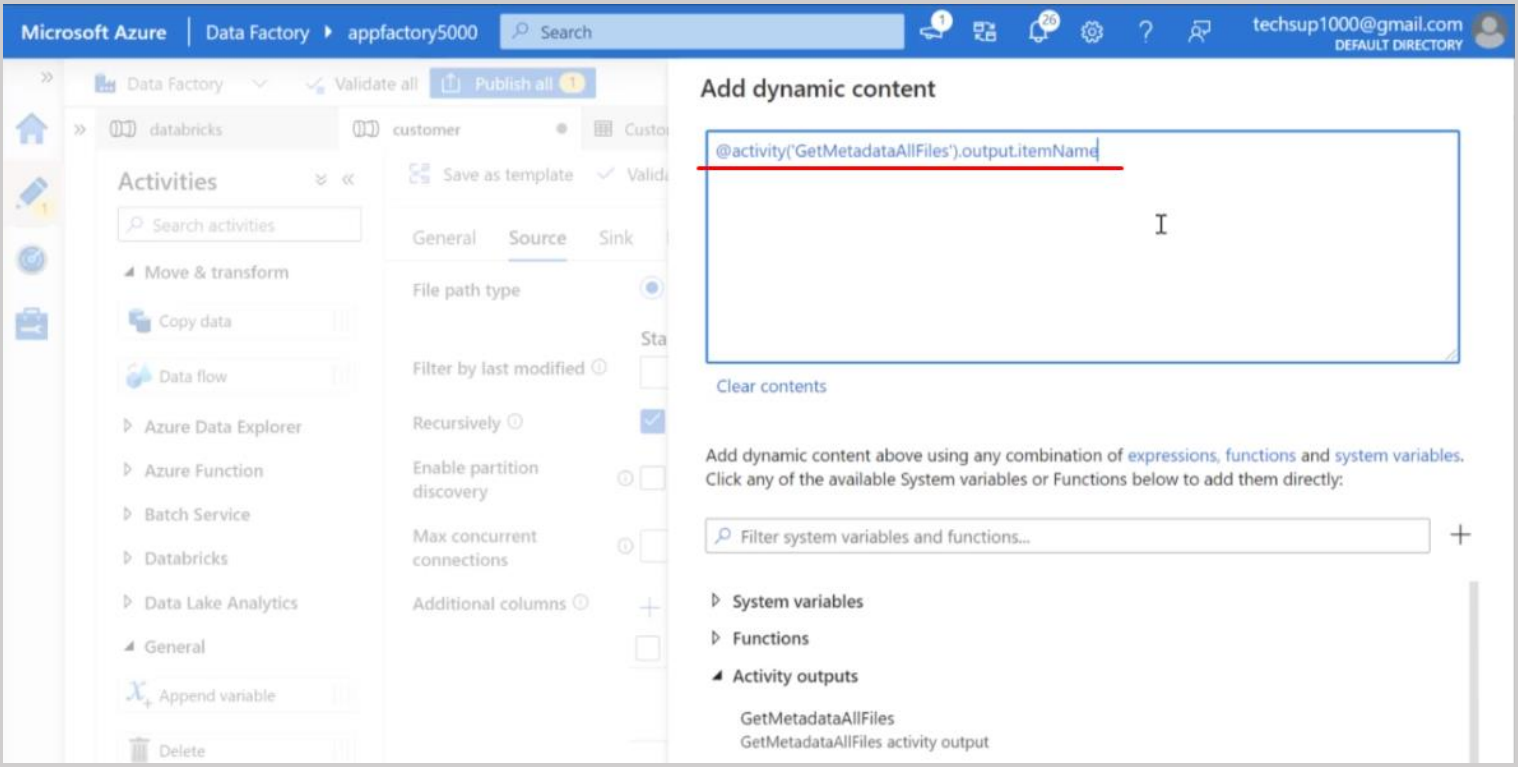

Microsoft Azure
appfactory5000
Search
techsup1000@gmail.com
DEFAULT DIRECTORY
Data Factory
Validate all
Publish all
databricks
customer
Save as template
Validate
Validate copy runtime
customer > ForEachFile
Get Metadata
GetMetadataAllFiles
General
Source
Sink
Mapping
Settings
Sink dataset *
Select...
Set properties
Name
AzureSynapseAnalyticsTable10
Linked service *
AzureSynapseAnalytics
Table name
dbo.Customer
Edit
Import schema
From connection/store
None
Advanced

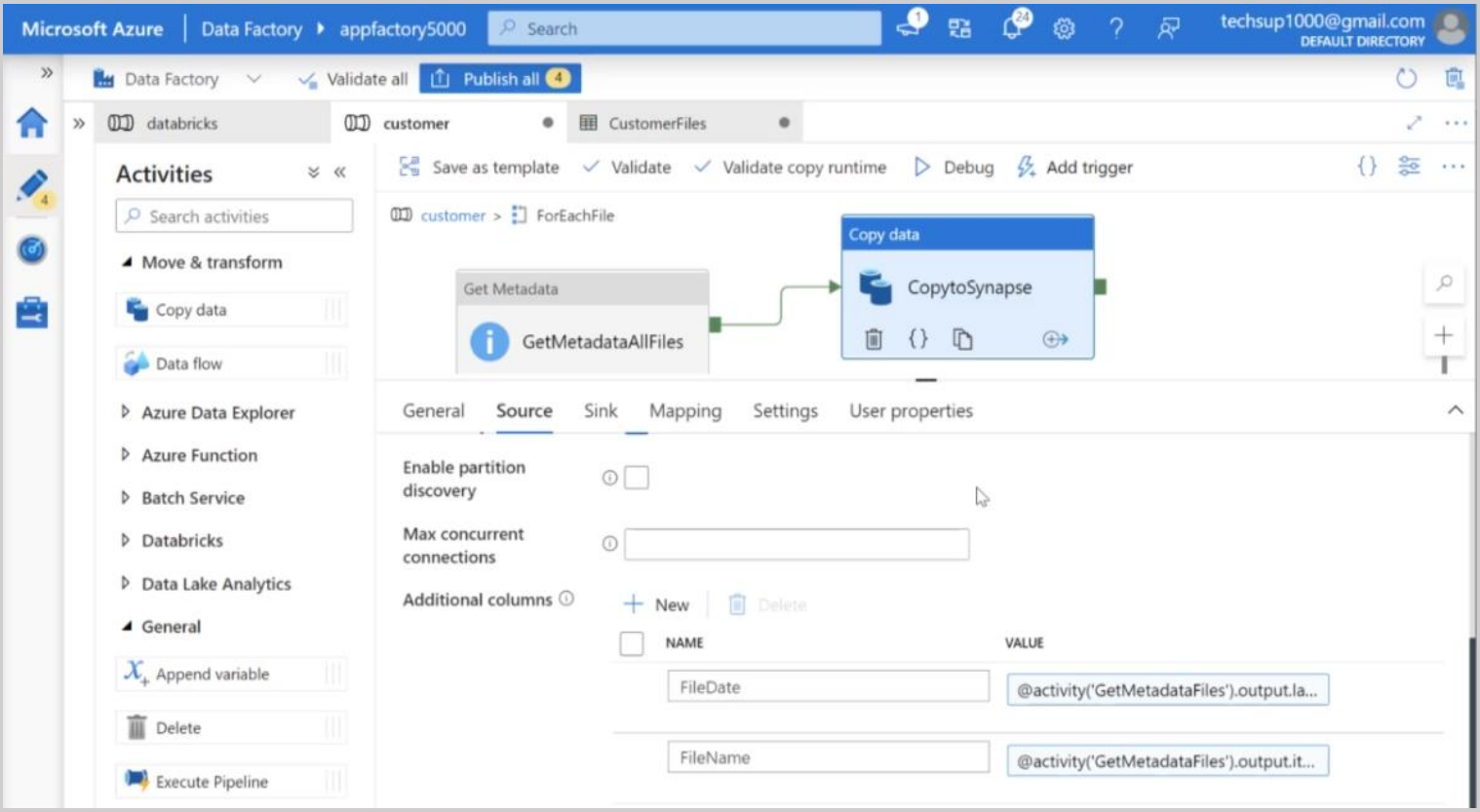

Microsoft Azure
appfactory5000
Search
techsup1000@gmail.com
DEFAULT DIRECTORY
Data Factory
Validate all
Publish all 4
databricks
customer
CustomerFiles
Save as template
Validate
Validate copy runtime
Debug
Add trigger
customer > ForEachFile
Get Metadata
GetMetadataAllFiles
Copy data
CopytoSynapse
General
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
Validate all
Publish all 4
databricks
customer
CustomerFiles
Activities
Search 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
New
Delete
NAME
VALUE
FileDate
$$FILEPATH
FileName
Add dynamic content
Add dynamic content [Alt+Shift+D]
Custom
$$FILEPATH
$$COLUMN

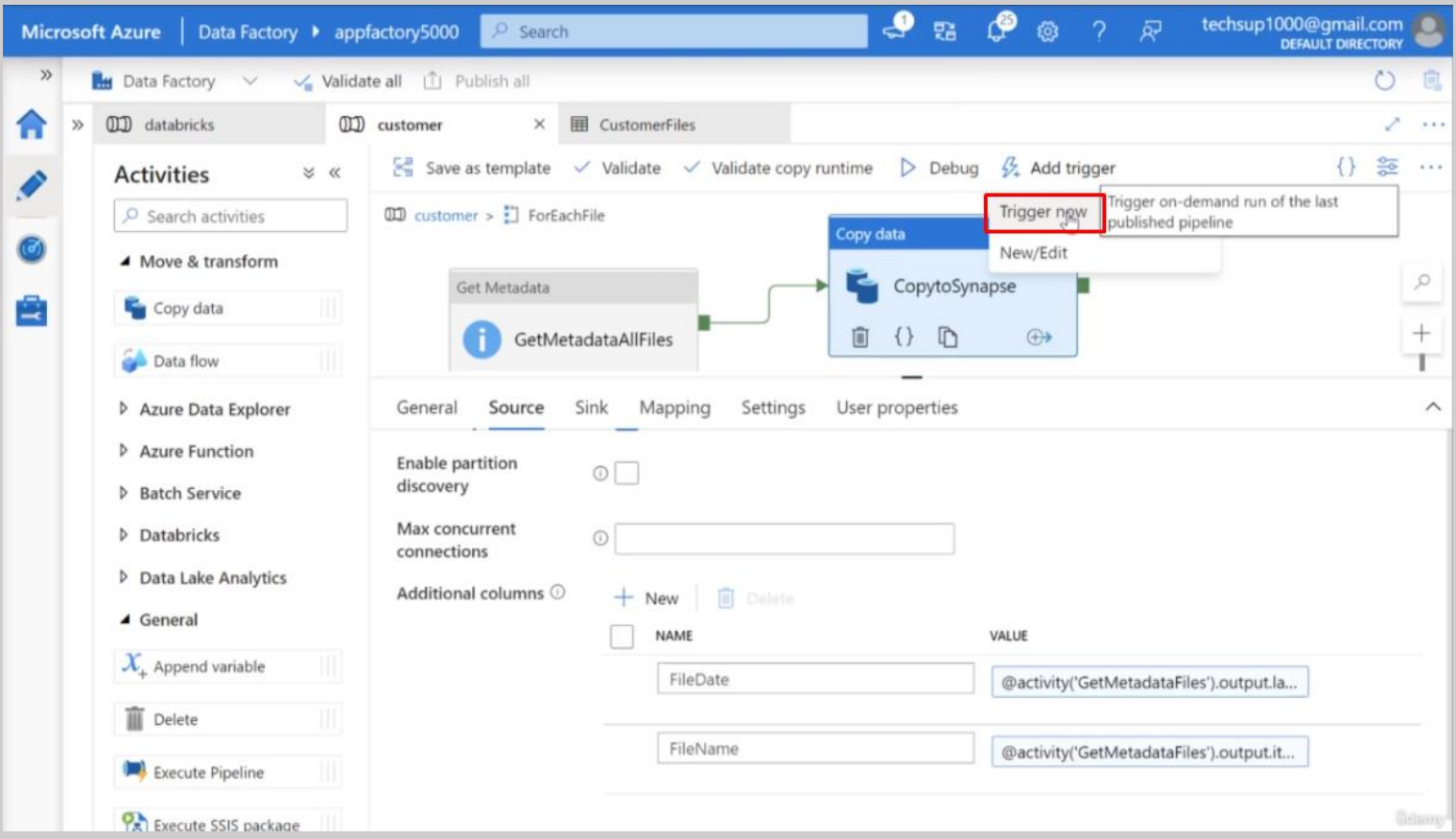

Microsoft Azure
Data Factory
appfactory5000
Search
techsup1000@gmail.com
DEFAULT DIRECTORY
Add dynamic content
@activity('GetMetadataAllFiles').output.lastModified
Clear contents
Add dynamic content above using any combination of expressions, functions and system variables.
Click any of the available 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

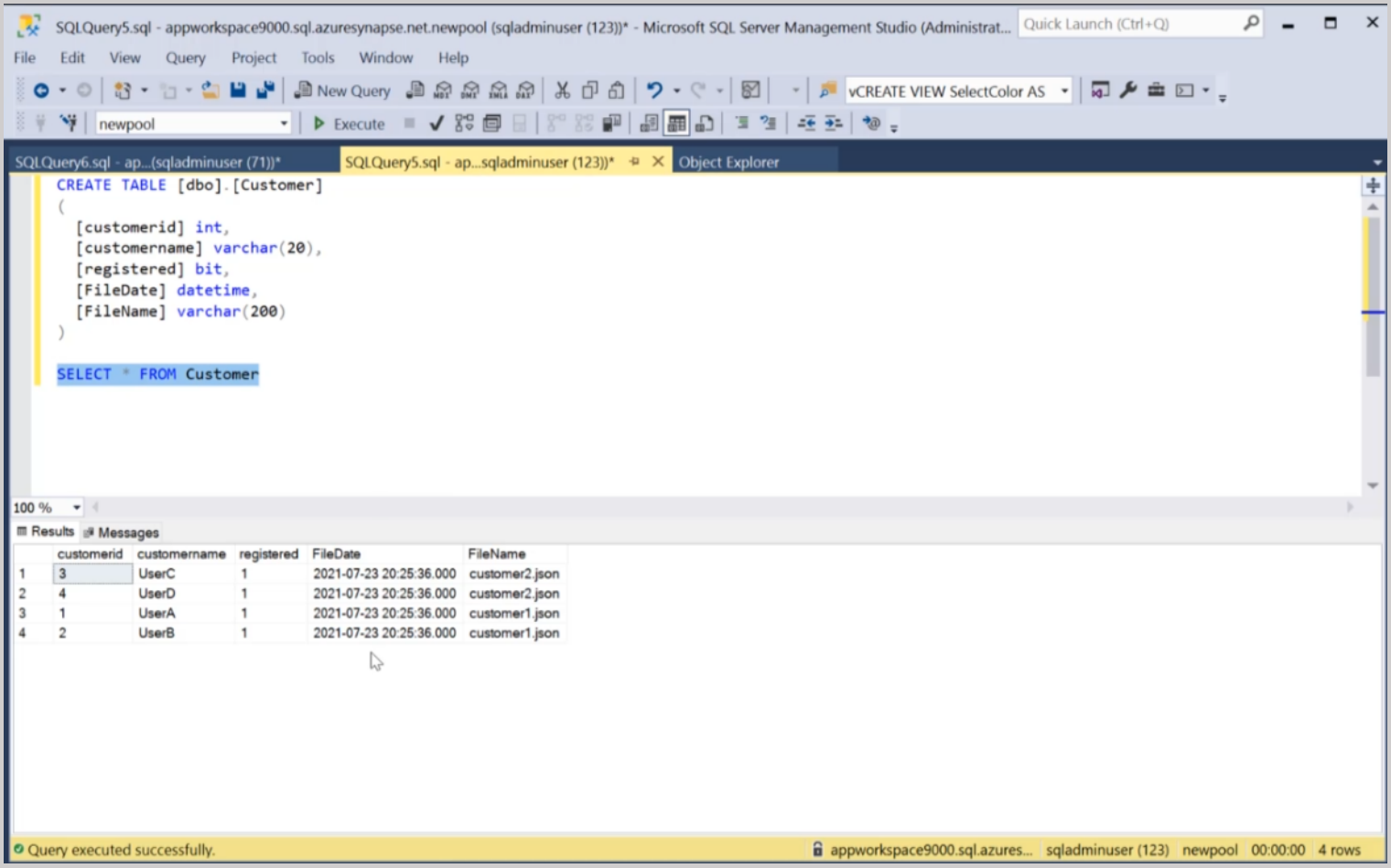

Microsoft Azure
Data Factory
appfactory5000
Search
techsup1000@gmail.com
DEFAULT DIRECTORY
Data Factory
Validate all
Publish all 1
databricks
customer
CustomerFiles
Activities
Search activities
Move & transform
Copy data
Data flow
Azure Data Explorer
Azure Function
Batch Service
Databricks
Data Lake Analytics
General
Append variable
Delete
Execute Pipeline
Execute SSIS package
Save as template
Validate
Validate copy runtime
Debug
Add trigger
General
Source
Sink
Mapping
Settings
User properties
Recursively
Enable partition discovery
Max concurrent connections
Additional columns
New
Delete
NAME
VALUE
FileDate
@activity('GetMetadataAllFiles').outpu...
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
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



Luego **Validamos**, **Publicamos** y **Ejecutamos**




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