

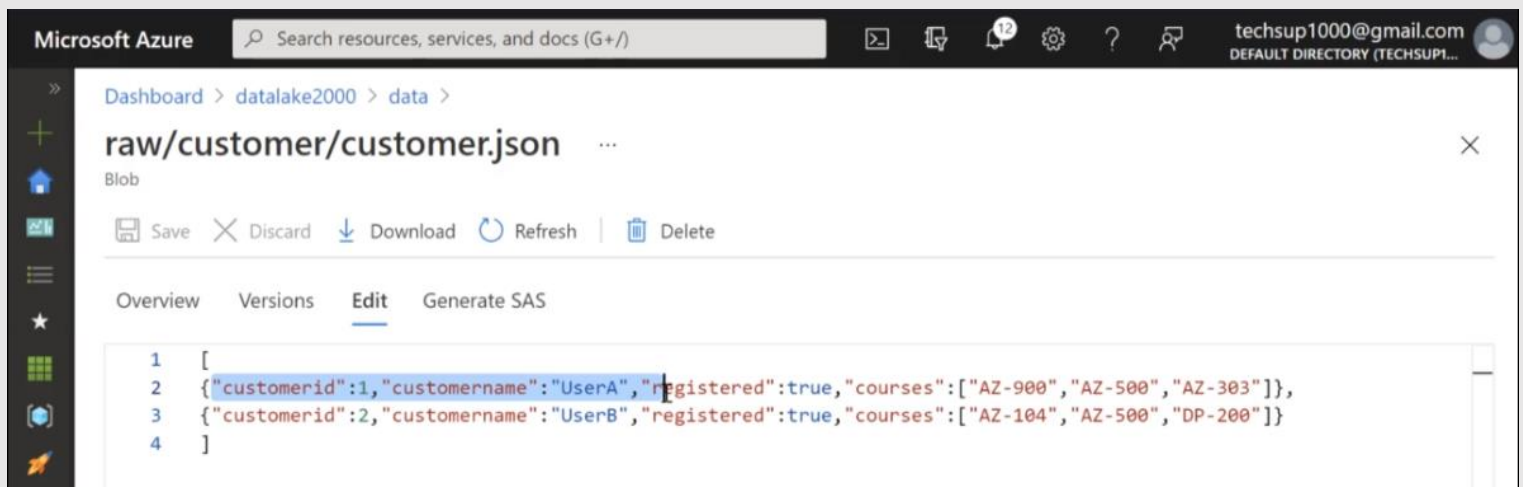
Azure Data Factory

Cargar un archivo

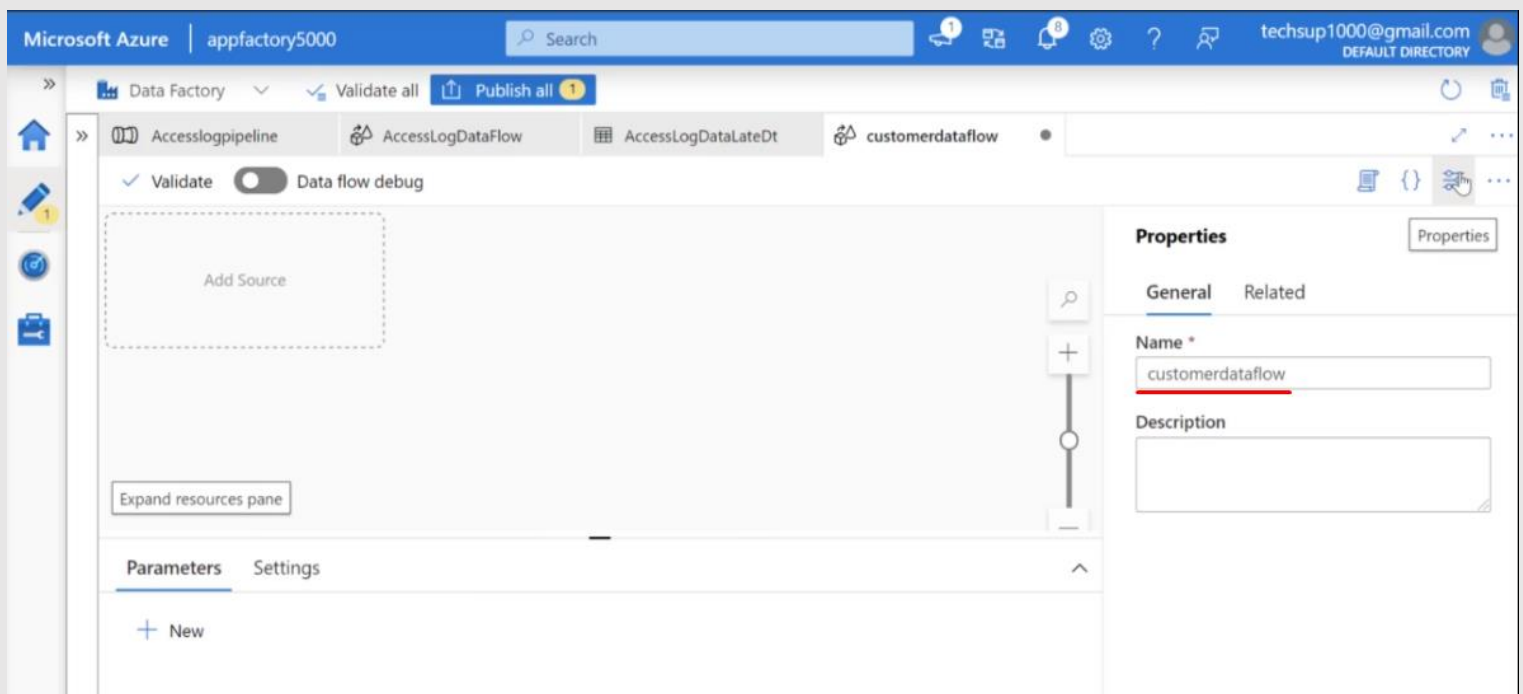
JSON Array a un

Dedicated SQL Pool

Se busca procesar el siguiente archivo JSON Array



Comenzamos creando nuestro Dataflow



Microsoft Azure | appfactory5000

Search

techsup1000@gmail.com
DEFAULT DIRECTORY

Data Factory | Validate all | Publish all 1

AccessLogpipeline | AccessLogDataFlow | AccessLogDataLateDt | customerdataflow

Validate | Data flow debug

source1
Columns:
0 total

Add Source

Source settings | Source options | Projection | Optimize | Inspect | Data preview | Description

Output stream name * | CustomerStream | Learn more

Source type * | Dataset | Inline

Dataset * | Select... | + New

Options | ☒ Allow schema drift

Cargaremos el archivo desde Azure Data Lake Storage y será un archivo JSON:

Microsoft Azure | appfactory5000

Search

techsup1000@gmail.com
DEFAULT DIRECTORY

Data Factory | Validate all | Publish all 1

AccessLogpipeline | AccessLogDataFlow | AccessLogDataLateDt | AccessLogDataLateDt

Validate | Data flow debug

CustomerStream
Columns:
0 total

Add Source

Source settings | Source options | Projection | Optimize | Inspect | Data preview | Description

Output stream name * | CustomerStream

Source type * | Dataset | Inline

Dataset * | Select...

Options | ☒ Allow schema drift

Set properties

Name | CustomerCourseDt

Linked service * | AzureDataLakeStorage

File path | data | / raw/customer | / customer.json

Import schema | ☒ From connection/store | ☐ From sample file | ☐ None

Advanced

OK | Back | Cancel

Microsoft Azure | appfactory5000

Search

techsup1000@gmail.com
DEFAULT DIRECTORY

Data Factory | Validate all | Publish all 2

AccessLogpipeline | AccessLogDataFlow | AccessLogDataLateDt | customerdataflow

Validate | Data flow debug

CustomerStream
Import data from CustomerCourseDt

FlattenStream
Columns: 4 total

Flatten settings | Optimize | Inspect | Data preview | Description

Output stream name * FlattenStream ? Help Learn more

Incoming stream * CustomerStream

Unroll by * [] courses

Unroll root [] courses

Options
☐ Skip duplicate input columns
☐ Skip duplicate output columns

Input columns *
Reset + Add mapping Delete 4 mappings: All inputs mapped

<input type="checkbox"/> CustomerStream's column		Name as	
--	--	---------	--

Microsoft Azure | appfactory5000

Search

techsup1000@gmail.com
DEFAULT DIRECTORY

Data Factory | Validate all | Publish all 2

AccessLogpipeline | AccessLogDataFlow | AccessLogDataLateDt | customerdataflow

Validate | Data flow debug

Reference: 1
Columns: 4 total

FlattenStream
Columns: 4 total

Flatten settings | Optimize | Inspect | Data preview | Description

Unroll root [] courses Add dynamic content [Alt+Shift+D]

Options
☐ Skip duplicate input columns Add dynamic content [Alt+Shift+D]
☐ Skip duplicate output columns

Input columns *
Reset + Add mapping Delete 4 mappings: All inputs mapped

<input type="checkbox"/> CustomerStream's column		Name as	
<input type="checkbox"/> abc customerid	→	customerid	+
<input type="checkbox"/> abc customername	→	customername	+
<input type="checkbox"/> <input checked="" type="checkbox"/> registered	→	registered	+
<input type="checkbox"/> courses	→	courses	+

Microsoft Azure | appfactory5000

Search

techsup1000@gmail.com
DEFAULT DIRECTORY

Data Factory | Validate all | Publish all 2

AccessLogpipeline | AccessLogDataFlow | AccessLogDataLateDt | customerdataflow

Validate | Data flow debug

CustomerStream: Import data from CustomerCourseDt

FlattenStream: Unrolling arrays from courses to courses with columns 'customerid, customername, registered, courses'

sink1: Columns: 4 total

Sink | Settings | Mapping | Optimize | Inspect | Data preview | Description

Output stream name *: CustomerSynapse

Incoming stream *: FlattenStream

Sink type *: Dataset | Inline | Cache

Dataset *: Select... | + New

Options: ☒ Allow schema drift | ☐ Validate schema

El dataset de destino apunta al Azure Synapse

Microsoft Azure | appfactory5000

Search

techsup1000@gmail.com
DEFAULT DIRECTORY

Data Factory | Validate all | Publish all 2

AccessLogpipeline | AccessLogDataFlow | AccessLogDataLateDt | customerdataflow

Validate | Data flow debug

CustomerStream: Import data from CustomerCourseDt

FlattenStream: Unrolling arrays from courses to courses with columns 'customerid, customername, registered, courses'

Sink | Settings | Mapping | Optimize | Inspect | Data preview

Output stream name *: CustomerSynapse

Incoming stream *: FlattenStream

Sink type *: Dataset | Inline

Dataset *: Select...

Options: ☒ Allow schema drift | ☐ Validate schema

Set properties

Name: CustomerSynapseDt

Linked service *: AzureSynapseAnalytics

☒ Select from existing table | ☐ Create new table

Table name: dbo.customercourse | Edit

Import schema: ☒ From connection/store | ☐ None

Advanced

OK | Back | Cancel

Microsoft Azure | appfactory5000

Search

techsup1000@gmail.com
DEFAULT DIRECTORY

Data Factory > Validate all > Publish all 3

Accesslogpipeline > AccessLogDataFlow > AccessLogDataLateDt > customerdataflow

Validate Data flow debug

Sink Settings **Mapping** Optimize Inspect Data preview Description

At least one incoming column is mapped to a column in the sink dataset schema with a conflicting type, which can cause NULL values or runtime errors.

Options

- ☒ Skip duplicate input columns
- ☒ Skip duplicate output columns
- ☐ Auto mapping
- Reset + Add mapping Delete
- Output format
- 4 mappings: All outputs mapped

Input columns	Output columns
<input type="checkbox"/> abc customerid	<input type="checkbox"/> 123 customerid
<input type="checkbox"/> abc customername	<input type="checkbox"/> abc customername
<input type="checkbox"/> % registered	<input checked="" type="checkbox"/> registered
<input type="checkbox"/> abc courses	<input type="checkbox"/> abc courses

Si revisamos los tipos de datos que toma el archivo JSON Array desde la fuente

Microsoft Azure | appfactory5000

Search

techsup1000@gmail.com
DEFAULT DIRECTORY

Data Factory > Validate all > Publish all 3

Accesslogpipeline > AccessLogDataFlow > AccessLogDataLateDt > customerdataflow

Validate Data flow debug

CustomerStream FlattenStream Reference: 0

Source settings Source options **Projection** Optimize Inspect Data preview Description

Define default format Import projection Reset schema

Column name	Type
customerid	abc string
customername	abc string
registered	% boolean
courses	[] string

Y desde el paso fuente indicamos que corresponde a un archivo JSON Array

The screenshot shows the Microsoft Azure Data Factory 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 area displays a pipeline diagram with three activities: 'CustomerStream', 'FlattenStream', and 'CustomerSynapse'. The 'CustomerStream' activity is highlighted with a red box. Below the diagram, the 'Source options' tab is selected. In the 'JSON settings' section, the 'Array of documents' option is selected and highlighted with a red box. Other options include 'Single document' and 'Document per line'. The 'After completion' section shows 'No action' selected. The 'Filter by last modified' section has input fields for 'Start time (UTC)' and 'End time (UTC)'. The 'Unquoted column names', 'Has comments', 'Single quoted', and 'Backslash escaped' options are all unchecked.

Para ejecutar nuestro Data Flow creamos un nuevo Pipeline

The screenshot shows the Microsoft Azure Data Factory 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 area displays a pipeline diagram with three activities: 'AccessLogDataFlow', 'AccessLogDataLateDt', and 'customerdataflow'. The 'AccessLogDataFlow' activity is highlighted with a red box. Below the diagram, the 'Activities' list is expanded, and the 'Data flow' activity is selected. A red arrow points from the 'Data flow' activity in the list to the 'CustomerDataFlow' activity in the center. The 'CustomerDataFlow' activity configuration is shown in the center, with the 'Name' field set to 'CustomerDataFlow'. The 'Settings' tab is selected, showing fields for 'Name', 'Description', 'Timeout', 'Retry', 'Retry interval', and 'Secure output'. The 'Properties' panel on the right shows the 'General' tab with the 'Name' field set to 'CustomerPipeline' and the 'Description' field empty. A note indicates that 'Concurrency has moved to the Settings tab'.

Microsoft Azure | appfactory5000

Search

techsup1000@gmail.com
DEFAULT DIRECTORY

Data Factory | Validate all | Publish all 1

AccessLogpipeline | AccessLogDataFlow | AccessLogDataLateDt | customerdataflow | CustomerPipeline

Activities

- Move & transform
- Copy data
- Data flow
- Azure Data Explorer
- Azure Function
- Batch Service
- Databricks
- Data Lake Analytics
- General
- HDInsight
- Iteration & conditionals
- Machine Learning
- Power Query

CustomerDataFlow

General Settings Parameters User properties

Data flow *
customerdataflow

Run on (Azure IR) *
AutoResolveIntegrationRuntime

Compute type *
General purpose

Core count *
4 (+ 4 Driver cores)

Logging level *
☒ Verbose ☐ Basic ☐ None

Sink properties

Properties

General Related

Name *
CustomerPipeline

Description

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

Annotations
+ New

Microsoft Azure | appfactory5000

Search

techsup1000@gmail.com
DEFAULT DIRECTORY

Data Factory | Validate all | Publish all 1

AccessLogpipeline | AccessLogDataFlow | AccessLogDataLateDt | customerdataflow | CustomerPipeline

Activities

- Move & transform
- Copy data
- Data flow
- Azure Data Explorer
- Azure Function
- Batch Service
- Databricks
- Data Lake Analytics
- General
- HDInsight
- Iteration & conditionals
- Machine Learning
- Power Query

CustomerDataFlow

General Settings Parameters User properties

Run on (Azure IR) *
AutoResolveIntegrationRuntime

Compute type *
General purpose

Core count *
4 (+ 4 Driver cores)

Logging level *
☒ Verbose ☐ Basic ☐ None

Sink properties

Staging

Staging linked service *
AzureDataLakeStorage

Staging storage folder
synapse / Directory

Test connection Edit + New

Validamos, Publicamos y Ejecutamos

The screenshot shows the Microsoft Azure Data Factory portal interface. At the top, the header displays 'Microsoft Azure | appfactory5000' and a search bar. The main navigation pane on the left lists various activities: Move & transform, Copy data, Data flow, Azure Data Explorer, Azure Function, Batch Service, Databricks, Data Lake Analytics, General, HDInsight, Iteration & conditionals, Machine Learning, and Power Query. The central workspace shows a pipeline named 'AccessLogDataFlow' with a 'CustomerDataFlow' activity. A 'Publishing completed' notification is visible in the top right corner, stating 'Successfully published'. Below this, a 'Trigger now' button is highlighted with a red box, with a tooltip indicating 'Trigger on-demand run of the last published pipeline'. The 'Settings' tab for the 'CustomerDataFlow' activity is open, showing configuration options such as 'Run on (Azure IR)' set to 'AutoResolveIntegrationRuntime', 'Compute type' set to 'General purpose', 'Core count' set to '4 (+ 4 Driver cores)', and 'Logging level' set to 'Verbose'. The 'Staging' section shows 'Staging linked service' set to 'AzureDataLakeStorage' and 'Staging storage folder' set to 'synapse / Directory'.

Esta es la estructura de la tabla a la cual serán cargados los datos

```
1 CREATE TABLE [customercourse]
2 (
3   [customerid] int,
4   [customername] varchar(200),
5   [registered] BIT,
6   [courses] varchar(200)
7 )
```


SQLQuery2.sql - appworkspace9000.sql.azuresynapse.net.newpool (sqladminuser (117))* - Microsoft SQL Server Management Studio

Quick Launch (Ctrl+Q)

File Edit View Query Project Tools Window Help

New Query

vCREATE VIEW SelectColor AS

newpool

Execute

SQLQuery2.sql - ap...sqladminuser (117))*

Object Explorer

```
CREATE TABLE [customercourse]
(
  [customerid] int,
  [customername] varchar(200),
  [registered] BIT,
  [courses] varchar(200)
)

SELECT * FROM [customercourse]
```

100 %

Results Messages

	customerid	customername	registered	courses
1	1	UserA	1	AZ-900
2	1	UserA	1	AZ-500
3	1	UserA	1	AZ-303
4	2	UserB	1	AZ-104
5	2	UserB	1	AZ-500
6	2	UserB	1	DP-200

Query executed successfully.

appworkspace9000.sql.azuresynapse.net (117) newpool 00:00:00 6 rows