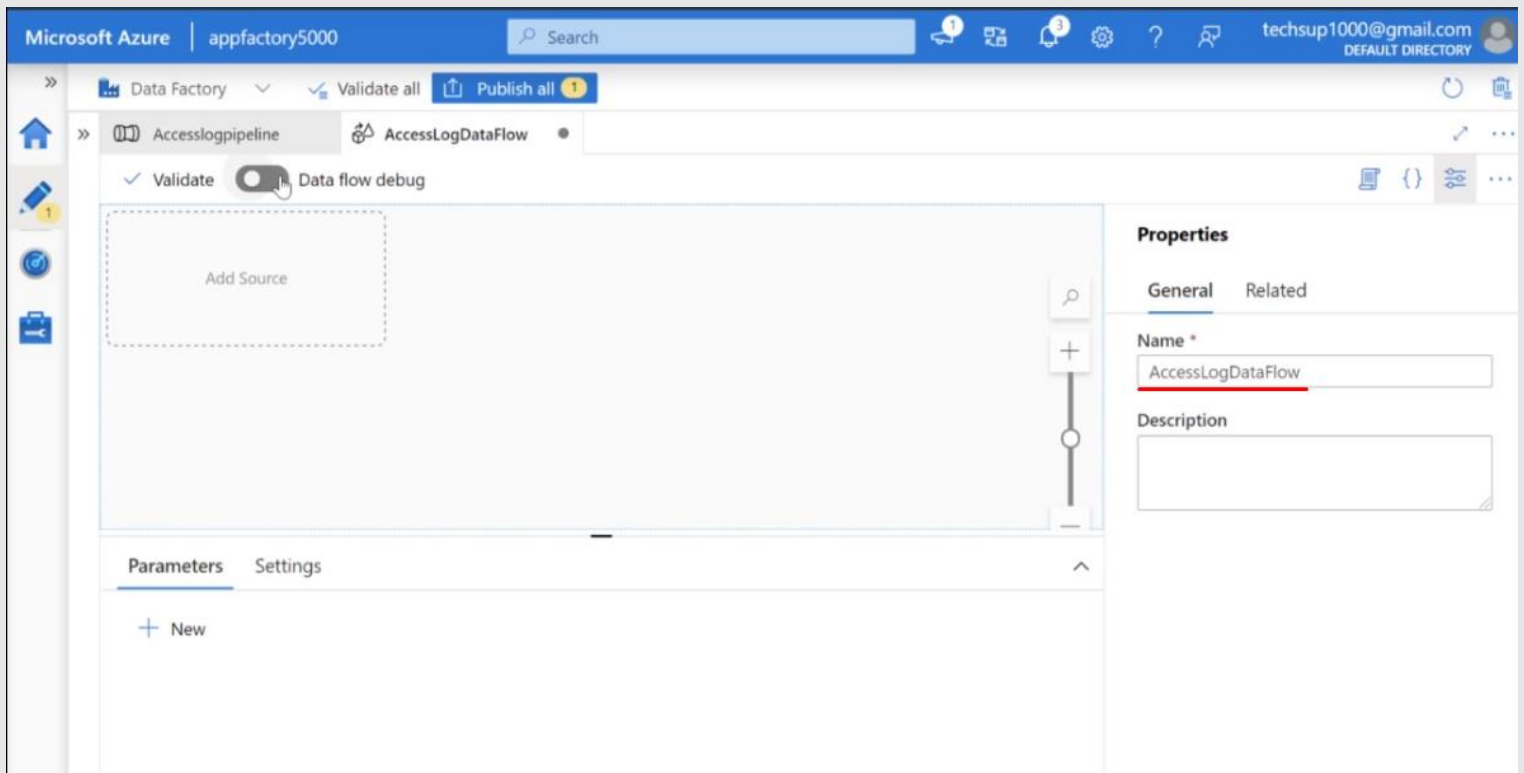


Azure Data Factory Mapping Data Flow para modificar y estructurar un archivo log complejo

Comenzamos creando un nuevo Data Flow.



Microsoft Azure | appfactory5000

Search

techsup1000@gmail.com
DEFAULT DIRECTORY

Data Factory | Validate all | Publish all 1

AccessLogpipeline | AccessLogDataFlow

Validate | **Data flow debug** | Debug Settings

source1
Columns: 0 total

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

Output stream name * | AccessLogStream | Learn more

Source type * | Dataset | Inline

Dataset * | Select... | **+ New**

Options

- ☒ Allow schema drift
- ☐ Infer drifted column types
- ☐ Validate schema

Sampling * | ☐ Enable | ☒ Disable

Properties

General | Related

Name * | AccessLogDataFlow

Description

Microsoft Azure | appfactory5000

Search

techsup1000@gmail.com
DEFAULT DIRECTORY

Data Factory | Validate all | Publish all 1

AccessLogpipeline | AccessLogDataFlow

Validate | Data flow debug | Debug Settings

AccessLogStream
Columns: 0 total

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

Output stream name * | AccessLogStream

Source type * | Dataset | Inline

Dataset * | Select...

Options

- ☒ Allow schema drift
- ☐ Infer drifted column types
- ☐ Validate schema

Sampling * | ☐ Enable | ☒ Disable

New dataset

In pipeline activities and data flows, reference a dataset to specify the location and structure of your data within a data store. [Learn more](#)

Select a data store

Search

All | Azure | Database | File | Generic protocol | NoSQL | Services and apps

Azure Blob Storage	Azure Cosmos DB (SQL API)	Azure Data Lake Storage Gen1
Azure Data Lake Storage Gen2	Azure Database for MySQL	Azure Database for PostgreSQL

Continue | Cancel

Microsoft Azure | appfactory5000

Search

techsup1000@gmail.com
DEFAULT DIRECTORY

Data Factory | Validate all | Publish all

Accesslogpipeline | AccessLogDataFlow

✓ Validate | Data flow debug | Debug Settings

AccessLogStream
Columns: 0 total

Source settings | Source options | Projection | Optimize

Output stream name * | AccessLogStream

Source type * | Dataset | Inline

Dataset * | Select...

Options | ☒ Allow schema drift | ☐ Infer drifted column types | ☐ Validate schema

Sampling * | ☐ Enable | ☒ Disable

Select format

Choose the format type of your data

Avro | **DelimitedText** | Excel

JSON | ORC | Parquet

XML | 01

Continue | Back | Cancel

Microsoft Azure | appfactory5000

Search

techsup1000@gmail.com
DEFAULT DIRECTORY

Data Factory | Validate all | Publish all

Accesslogpipeline | AccessLogDataFlow

✓ Validate | Data flow debug | Debug Settings

AccessLogStream
Columns: 0 total

Source settings | Source options | Projection | Optimize

Output stream name * | AccessLogStream

Source type * | Dataset | Inline

Dataset * | Select...

Options | ☒ Allow schema drift | ☐ Infer drifted column types | ☐ Validate schema

Sampling * | ☐ Enable | ☒ Disable

Set properties

Name | AccessLogDataLateDt

Linked service * | AzureDataLakeStorage

File path | data | / raw/nginx | / access.log

First row as header | ☐

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

Advanced

OK | Back | Cancel

El archivo tiene una columna y dos filas. Si nos fijamos no tiene delimitadores claros.

The screenshot shows the Microsoft Azure Data Factory interface. The 'Data preview' tab is selected, showing two rows of data. The first row is: 127.0.0.1 - - [09/Jul/2021:12:36:33 +0000] "GET / HTTP/1.1" 200 612 "-" "Mozilla/5.0 (Windows ...". The second row is: 127.0.0.1 - - [09/Jul/2021:12:36:34 +0000] "GET /favicon.ico HTTP/1.1" 404 153 "-" "Mozilla/5.0 (...". Red annotations highlight the 'Data preview' tab and the first column of data.

Si visualizamos el archivo podemos determinar que quizás un delimitador podría ser el "ESPACIO"

The screenshot shows the Microsoft Azure portal interface. The 'raw/nginx/access.log' file is displayed in a blob storage account. The file content is shown as a text view, displaying two lines of log data. The first line is: 1 127.0.0.1 - - [09/Jul/2021:12:36:33 +0000] "GET / HTTP/1.1" 200 612 "-" "Mozilla/5.0 (Windows NT 10.0; WOW64; Tride. The second line is: 2 127.0.0.1 - - [09/Jul/2021:12:36:34 +0000] "GET /favicon.ico HTTP/1.1" 404 153 "-" "Mozilla/5.0 (Windows NT 10.0; V. Red annotations highlight the first column of data.

Microsoft Azure | appfactory5000

Search

techsup1000@gmail.com
DEFAULT DIRECTORY

Data Factory > Validate all > Publish all 2

Accesslogpipeline > AccessLogDataFlow

✓ Validate Data flow debug Debug Settings

AccessLogStream
Columns: 0 total

Source settings Source options Projection Optimize Inspect Data preview Description

Output stream name * AccessLogStream Learn more

Source type * Dataset Inline

Dataset * AccessLogDataLateDt

Test connection **Open** + New

Options
☒ Allow schema drift
☐ Infer drifted column types
☐ Validate schema

Properties
General Related
Name * AccessLogDataFlow
Description

En el dataset como delimitador de columnas coloco un espacio.

Microsoft Azure | appfactory5000

Search

techsup1000@gmail.com
DEFAULT DIRECTORY

Data Factory > Validate all > Publish all 2

Accesslogpipeline > AccessLogDataFlow > AccessLogDataLateDt

Connection Schema Parameters

Linked service * AzureDataLakeStorage

Test connection Edit + New Learn more

File path * data / raw/nginx / access.log Browse

Compression type None

Column delimiter |
☒ Edit
Add dynamic content [Alt+Shift+D]

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

Encoding Default(UTF-8)

Escape character Backslash (\) Edit

Quote character Double quote (")

Properties
General Related (1)
Name * AccessLogDataLateDt
Description
Annotations
+ New

Microsoft Azure | appfactory5000

Search

techsup1000@gmail.com
DEFAULT DIRECTORY

Data Factory > Validate all > Publish all 2

AccessLogpipeline > AccessLogDataFlow > AccessLogDataLateDt

✓ Validate Data flow debug Data flow debug Debug Settings

AccessLogStream
Columns: 0 total

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

Number of rows INSERT 2 UPDATE 0 DELETE 0 UPSERT 0 LOOKUP 0 TOTAL N/A

Refresh Refresh the data preview with any changes from the data flow

	col0 abc
+	127.0.0.1 - - [09/Jul/2021:12:36:33 +0000] "GET / HTTP/1.1" 200 612 "-" "Mozilla/5.0 (Windows ...
+	127.0.0.1 - - [09/Jul/2021:12:36:34 +0000] "GET /favicon.ico HTTP/1.1" 404 153 "-" "Mozilla/5.0 (...

Properties

General Related

Name *
AccessLogDataFlow

Description

Microsoft Azure | appfactory5000

Search

techsup1000@gmail.com
DEFAULT DIRECTORY

Data Factory > Validate all > Publish all 2

AccessLogpipeline > AccessLogDataFlow > AccessLogDataLateDt

✓ Validate Data flow debug Data flow debug Debug Settings

AccessLogStream
Columns: 0 total

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

Number of rows INSERT 2 UPDATE 0 DELETE 0 UPSERT 0 LOOKUP 0 TOTAL 2

Refresh Typecast Modify Map drifted Statistics Remove

	col0 abc	_col1_ abc	_col2_ abc	_col3_
+	127.0.0.1	-	-	[09/Jul/
+	127.0.0.1	-	-	[09/Jul/

Se busca mover los datos a una tabla en el Dedicated SQL Pool

```
1 CREATE TABLE [serverlogs]
2 (
3 [remote_addr] varchar(20),
4 [time_local] varchar(100),
5 [request] varchar(200),
6 [status] int,
7 [bytes] int,
8 [remote_user] varchar(100),
9 [http_user_agent] varchar(500)
10 )
11
```

Vamos a seleccionar las columnas y renombrarlas

The screenshot displays the Microsoft Azure Data Factory console for the workspace 'appfactory5000'. The 'AccessLogpipeline' is selected, and the 'AccessLogDataFlow' activity is in focus. The 'Select1' activity is highlighted with a red box. The 'Select settings' tab is active, showing the 'Output stream name' as 'SelectColumns' and the 'Incoming stream' as 'AccessLogStream'. Under 'Options', both 'Skip duplicate input columns' and 'Skip duplicate output columns' are checked. The 'Input columns' section shows a table with columns from 'AccessLogStream' being mapped to new names.

AccessLogStream's column	Name as
abc_col0_	_col0_
abc_col1_	_col1_
abc_col2_	_col2_
abc_col3_	_col3_

The 'Properties' pane on the right shows the 'General' tab with the 'Name' set to 'AccessLogDataFlow' and a blank 'Description' field.

Microsoft Azure | appfactory5000

Search

techsup1000@gmail.com
DEFAULT DIRECTORY

>> Data Factory > Validate all > Publish all 2

AccessLogpipeline > AccessLogDataFlow > AccessLogDataLateDt

✓ Validate > Data flow debug > Debug Settings

AccessLogStream > SelectColumns

Select settings > Optimize > Inspect > Data preview ●

Description

Options

- ✓ Skip duplicate input columns ⓘ
- ✓ Skip duplicate output columns ⓘ

Input columns

Auto mapping ⓘ > Reset > + Add mapping > Delete

7 mappings: 3 column(s) from the inputs left unmapped ⓘ

<input type="checkbox"/>	AccessLogStream's column		Name as	
<input type="checkbox"/>	abc_col0_	→	remote_addr	+ ⓧ
<input type="checkbox"/>	abc_col3_	→	time_local	+ ⓧ
<input type="checkbox"/>	abc_col5_	→	request	+ ⓧ
<input type="checkbox"/>	12s_col6_	→	status	+ ⓧ
<input type="checkbox"/>	12s_col7_	→	bytes	+ ⓧ
<input type="checkbox"/>	abc_col8_	→	remote_user	+ ⓧ
<input type="checkbox"/>	abc_col9_	→	http_user_agent	+ ⓧ

Microsoft Azure | appfactory5000

Search

techsup1000@gmail.com
DEFAULT DIRECTORY

>> Data Factory > Validate all > Publish all 2

AccessLogpipeline > AccessLogDataFlow > AccessLogDataLateDt

✓ Validate > Data flow debug > Debug Settings

AccessLogStream > SelectColumns

Import data from AccessLogDataLateDt

Columns: 7 total

Select settings > Optimize > Inspect > Data preview ●

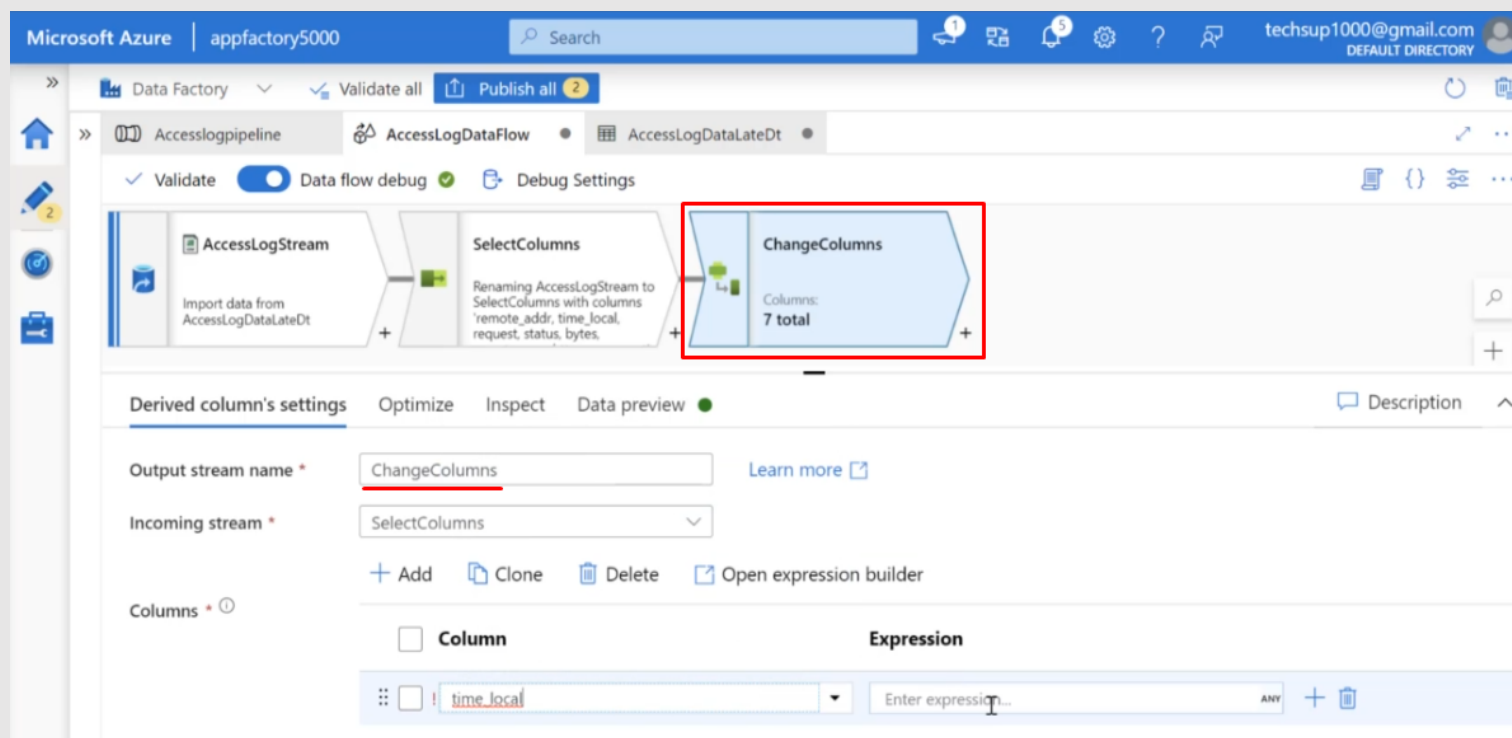
Description

Number of rows > INSERT 2 > UPDATE 0 > DELETE 0 > UPSERT 0 > LOOKUP 0 > TOTAL 2

Refresh > Typecast > Modify > Map drifted > Statistics > Remove

↑↓	remote_addr abc	time_local abc	request abc	status 12s	bytes 12s
+	127.0.0.1	[09/Jul/2021:12:36:33	GET / HTTP/1.1	200	612
+	127.0.0.1	[09/Jul/2021:12:36:34	GET /favicon.ico HTTP/1.1	404	153

Vamos a limpiar los valores de los registros



Microsoft Azure | appfactory5000

AccessLogpipeline | AccessLogDataFlow | AccessLogDataLateDt

Validate | Data flow debug | Debug Settings

AccessLogStream | SelectColumns | ChangeColumns

Import data from AccessLogDataLateDt | Renaming AccessLogStream to SelectColumns with columns 'remote_addr, time_local, request, status, bytes.' | Columns: 7 total

Derived column's settings | Optimize | Inspect | Data preview

Output stream name * | ChangeColumns | Learn more

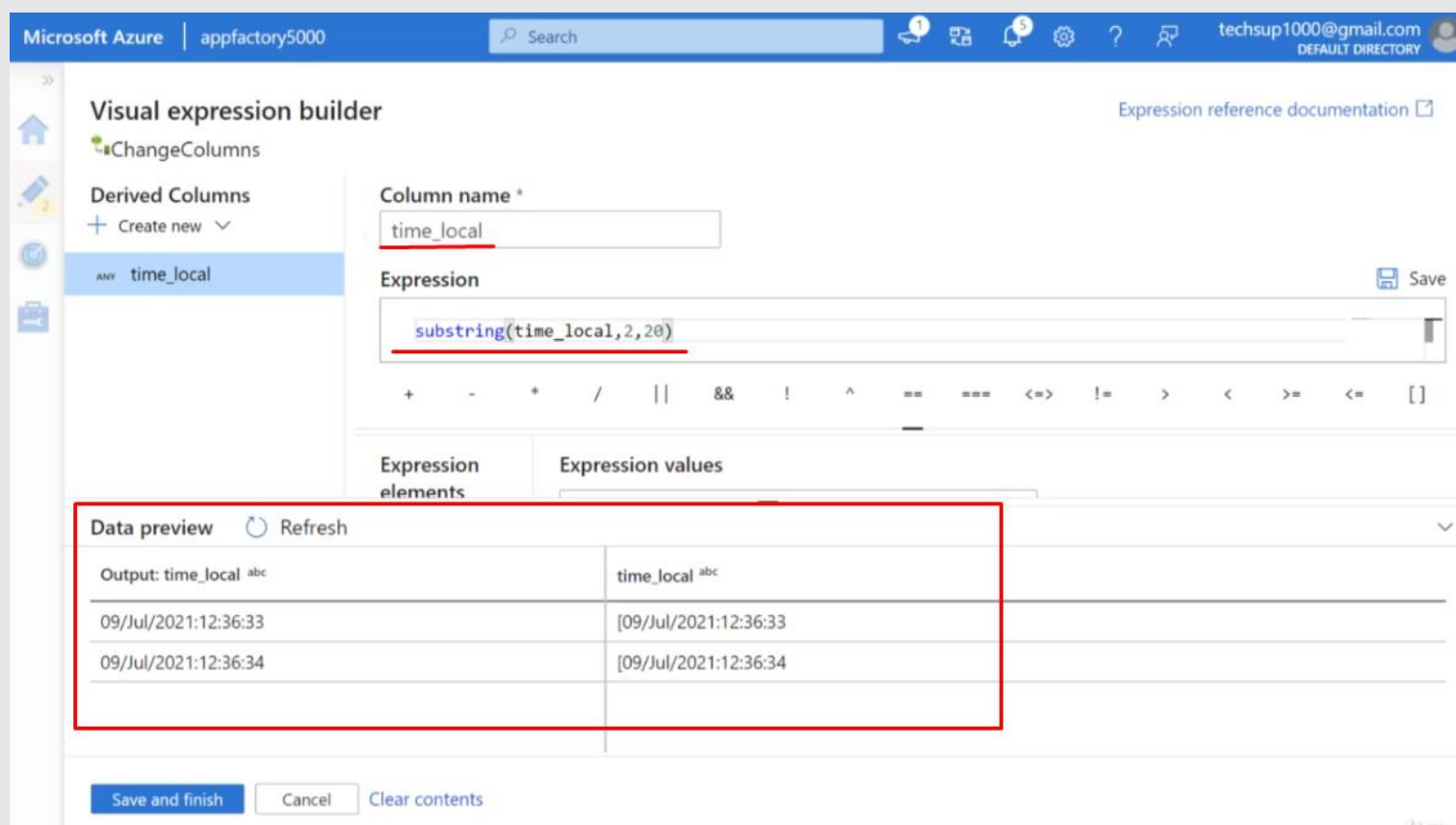
Incoming stream * | SelectColumns

+ Add | Clone | Delete | Open expression builder

Columns * 1

Column	Expression
time_local	Enter expression...

Ahora, ya que hemos habilitado que el "**dataflow debug**", usted está en el Data Preview y si presionamos "**Refresh**" podemos ver cuál es el impacto de nuestra expresión, que es, muy, muy importante. Si quieres ver si una expresión está realmente funcionando como debería puede utilizar la función de "**Data Preview**" que está disponible con el "**dataflow debug**". Recuerda, todo esto se está ejecutando en segundo plano en un clúster Apache Spark.



Microsoft Azure | appfactory5000

Visual expression builder | Expression reference documentation

ChangeColumns

Derived Columns | Create new

ANY | time_local

Column name * | time_local

Expression | substring(time_local,2,20) | Save

+ | - | * | / | || | && | ! | ^ | == | === | <= | >= | < | > | <= | >= | []

Expression elements | Expression values

Data preview | Refresh

Output: time_local abc	time_local abc
09/Jul/2021:12:36:33	[09/Jul/2021:12:36:33
09/Jul/2021:12:36:34	[09/Jul/2021:12:36:34

Save and finish | Cancel | Clear contents

Microsoft Azure | appfactory5000

Search

techsup1000@gmail.com
DEFAULT DIRECTORY

Data Factory | Validate all | Publish all 2

AccessLogpipeline | AccessLogDataFlow | AccessLogDataLateDt

Validate | Data flow debug | Debug Settings

AccessLogStream
Import data from AccessLogDataLateDt

SelectColumns
Renaming AccessLogStream to SelectColumns with columns 'remote_addr, time_local, request, status, bytes.'

ChangeColumns
Columns: 7 total

Derived column's settings | Optimize | Inspect | Data preview | Description

Output stream name * | ChangeColumns | Learn more

Incoming stream * | SelectColumns

+ Add | Clone | Delete | Open expression builder

Columns *

Column	Expression
<input type="checkbox"/> time_local	substring(time_local,2,20) abc +
<input type="checkbox"/> status	toInteger(status) 123 +
<input type="checkbox"/> bytes	toInteger(bytes) 123 +

Microsoft Azure | appfactory5000

Search

techsup1000@gmail.com
DEFAULT DIRECTORY

Data Factory | Validate all | Publish all 2

AccessLogpipeline | AccessLogDataFlow | AccessLogDataLateDt

Validate | Data flow debug | Debug Settings

AccessLogStream
Import data from AccessLogDataLateDt

SelectColumns
Renaming AccessLogStream to SelectColumns with columns 'remote_addr, time_local, request, status, bytes.'

Reference:
1
Columns: 7 total

ServerLogsSynapse
Columns: 7 total

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

Output stream name * | ServerLogsSynapse | Learn more

Incoming stream * | ChangeColumns

Sink type * | Dataset | Inline | Cache

Dataset * | Select... | + New

Options
☒ Allow schema drift
☐ Validate schema

Microsoft Azure | appfactory5000

Search

techsup1000@gmail.com
DEFAULT DIRECTORY

Data Factory

Validate all

Publish all 2

AccessLogpipeline

AccessLogDataFlow

AccessLogDataFlow

✓ Validate

🔍 Data flow debug

🔍 Debug Settings

AccessLogStream

Import data from AccessLogDataLateDt

+

SelectColumns

Renaming AccessLogStream to SelectColumns with columns 'remote_addr, time, local, request, status, bytes.'

+

Sink

Settings

Mapping

Optimize

Inspect

Data preview

Output stream name *

ServerLogsSynapse

Incoming stream *

ChangeColumns

Sink type *

Dataset

Inline

Dataset *

Select...

Options

✓ Allow schema drift

❌ Validate schema

New dataset

In pipeline activities and data flows, reference a dataset to specify the location and structure of your data within a data store. [Learn more](#)

Select a data store

Search

All

Azure

Database

File

Generic protocol

NoSQL

Services and apps

Azure Data Lake Storage Gen2

Azure Database for MySQL

Azure Database for PostgreSQL

Azure SQL Database

Azure SQL Database Managed Instance

Azure Synapse Analytics

Continue

Cancel

Microsoft Azure | appfactory5000

Search

techsup1000@gmail.com
DEFAULT DIRECTORY

Data Factory

Validate all

Publish all 2

AccessLogpipeline

AccessLogDataFlow

AccessLogDataFlow

✓ Validate

🔍 Data flow debug

🔍 Debug Settings

AccessLogStream

Import data from AccessLogDataLateDt

+

SelectColumns

Renaming AccessLogStream to SelectColumns with columns 'remote_addr, time, local, request, status, bytes.'

+

Sink

Settings

Mapping

Optimize

Inspect

Data preview

Output stream name *

ServerLogsSynapse

Incoming stream *

ChangeColumns

Sink type *

Dataset

Inline

Dataset *

Select...

Options

✓ Allow schema drift

❌ Validate schema

Set properties

Name

ServerLogsDt

Linked service *

AzureSynapseAnalytics

☒ Select from existing table

☐ Create new table

Table name

dbo.serverlogs

☐ 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

✓ Validate Data flow debug Debug Settings

AccessLogStream Import data from AccessLogDataLateDt

SelectColumns Renaming AccessLogStream to SelectColumns with columns 'remote_addr, time_local, request, status, bytes.'

ChangeColumns Creating/updating the columns 'remote_addr, time_local, request, status, bytes, remote_user, http_user_agent'

ServerLogsSynapse Columns: 7 total

Sink Settings Mapping Optimize Inspect Data preview

Options

- ✓ Skip duplicate input columns
- ✓ Skip duplicate output columns
- Auto mapping Reset Add mapping Delete

Output format 1 mappings: All outputs mapped

All inputs mapped by name including drifted columns

Microsoft Azure | appfactory5000

Search

techsup1000@gmail.com
DEFAULT DIRECTORY

Data Factory > Validate all > Publish all 3

AccessLogpipeline > AccessLogDataFlow > AccessLogDataLateDt

✓ Validate Data flow debug Debug Settings

Sink Settings Mapping Optimize Inspect Data preview

Number of rows INSERT 2 UPDATE 0 DELETE 0 UPSERT 0 LOOKUP 0 TOTAL 2

Refresh Statistics

remote_addr	time_local	request	status	bytes
127.0.0.1	09/Jul/2021:12:36:33	GET / HTTP/1.1	200	612
127.0.0.1	09/Jul/2021:12:36:34	GET /favicon.ico HTTP/1.1	404	153

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 [serverlogs]
(
    [remote_addr] varchar(20),
    [time_local] varchar(100),
    [request] varchar(200),
    [status] int,
    [bytes] int,
    [remote_user] varchar(100),
    [http_user_agent] varchar(500)
)

SELECT * FROM [serverlogs]
```

100 %

Results Messages

	remote_addr	time_local	request	status	bytes	remote_user	http_user_agent
1	127.0.0.1	09/Jul/2021:12:36:33	GET / HTTP/1.1	200	612	-	Mozilla/5.0 (Windows NT 10.0; WOW64; Trident/7...
2	127.0.0.1	09/Jul/2021:12:36:34	GET /favicon.ico HTTP/1.1	404	153	-	Mozilla/5.0 (Windows NT 10.0; WOW64; Trident/7...