

Azure Data Factory Mapping Data Flow para convertir dos archivos CSV con schema diferente a JSON

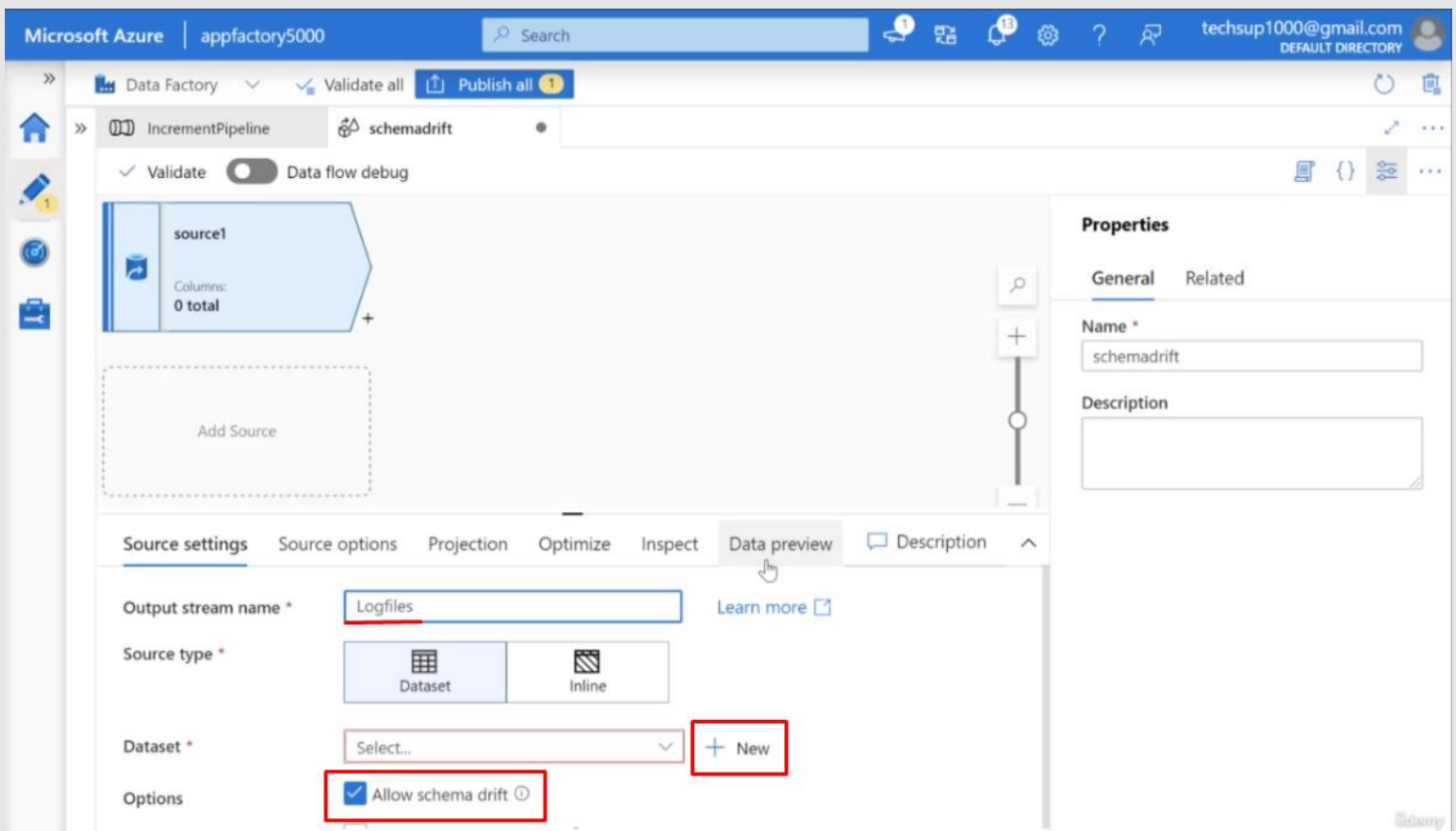
El primer archivo tiene 10 columnas y no tendrá la columna ResourceGroup

```
Log-withonecolumnless.csv U X
2-DP-203_-_Data_Engineering_on_Microsoft_Azure > 04 - Design and Develop Data Processing - Azure Data Factory > material > ejemplo_schema_drift > Log-withonecolumnless.csv
1 Id,Correlationid,Operationname,Status,Eventcategory,Level,Time,Subscription,Eventinitiatedby,ResourceType
2 1,66641e13-d19f-4ce5-aafd-9d5d7bfa557,Delete SQL database,Succeeded,Administrative,Informational,2021-06-15T04:44:38.223Z,20c6eec9-2
3 2,66641e13-d19f-4ce5-aafd-9d5d7bfa557,Delete SQL database,Started,Administrative,Informational,2021-06-15T04:44:21.547Z,20c6eec9-2d8
4 3,66641e13-d19f-4ce5-aafd-9d5d7bfa557,Delete SQL database,Accepted,Administrative,Informational,2021-06-15T04:44:21.702Z,20c6eec9-2d
5 4,e2958162-93d9-4643-a847-82cf25c49930,Delete SqlPools,Succeeded,Administrative,Informational,2021-06-15T04:44:31.332Z,20c6eec9-2d80-
6 5,e2958162-93d9-4643-a847-82cf25c49930,Delete SqlPools,Started,Administrative,Informational,2021-06-15T04:44:12.533Z,20c6eec9-2d80-47
7 6,e2958162-93d9-4643-a847-82cf25c49930,Delete SqlPools,Accepted,Administrative,Informational,2021-06-15T04:44:16.038Z,20c6eec9-2d80-4
8 7,08cd2e19-477c-4ecc-83a6-575b9ce265e3,Pause SQL Analytics pools.,Succeeded,Administrative,Informational,2021-06-14T17:57:02.240Z,20c
9 8,08cd2e19-477c-4ecc-83a6-575b9ce265e3,Pause SQL Analytics pools.,Started,Administrative,Informational,2021-06-14T17:55:17.612Z,20c6e
10 9,08cd2e19-477c-4ecc-83a6-575b9ce265e3,Pause SQL Analytics pools.,Accepted,Administrative,Informational,2021-06-14T17:55:18.577Z,20c6
11 10,d2d9d7c4-2766-4e7d-83fd-c4f07b3d6825,Pause a Datawarehouse database.,Succeeded,Administrative,Informational,2021-06-14T17:56:51.94
12 11,d2d9d7c4-2766-4e7d-83fd-c4f07b3d6825,Pause a Datawarehouse database.,Started,Administrative,Informational,2021-06-14T17:55:18.393Z
13 12,d2d9d7c4-2766-4e7d-83fd-c4f07b3d6825,Pause a Datawarehouse database.,Accepted,Administrative,Informational,2021-06-14T17:55:18.478
14 13,1c735927-517e-470f-bccf-239cdf3b97f8,Create Pipeline Run,Succeeded,Administrative,Informational,2021-06-14T14:24:48.367Z,20c6eec9-
15 14,1c735927-517e-470f-bccf-239cdf3b97f8,Create Pipeline Run,Started,Administrative,Informational,2021-06-14T14:24:45.734Z,20c6eec9-2d
16 15,0ae558f0-ebbd-4e9e-b3de-406cf39ae0ad,Create or Update any Pipeline.,Succeeded,Administrative,Informational,2021-06-14T14:24:45.429
17 16,0ae558f0-ebbd-4e9e-b3de-406cf39ae0ad,Create or Update any Pipeline.,Started,Administrative,Informational,2021-06-14T14:24:44.499Z,
18 17,072d5d31-b4b0-4bd1-bfef-088c7faf11eb,Create or Update Dataset,Succeeded,Administrative,Informational,2021-06-14T14:24:44.209Z,20c6
19 18,072d5d31-b4b0-4bd1-bfef-088c7faf11eb,Create or Update Dataset,Started,Administrative,Informational,2021-06-14T14:24:43.264Z,20c6ee
20 19,77152ae0-297f-4d10-9179-90d50bec8dac,Create or Update Dataset,Succeeded,Administrative,Informational,2021-06-14T14:24:44.189Z,20c6
21 20,77152ae0-297f-4d10-9179-90d50bec8dac,Create or Update Dataset,Succeeded,Administrative,Informational,2021-06-14T14:24:44.189Z,20c6
22
```

El primer archivo tiene 11 columnas, dado que tiene la columna ResourceGroup

```
Log.csv
2-DP-203_-_Data_Engineering_on_Microsoft_Azure > 04 - Design and Develop Data Processing - Azure Data Factory > material > ejemplo_schema_drift > Log.csv
1 Id,Correlationid,Operationname,Status,Eventcategory,Level,Time,Subscription,Eventinitiatedby,Resourcetype,Resourcegroup
2 21,66641e13-d19f-4ce5-aafd-9d5d7bfa557,Delete SQL database,Succeeded,Administrative,Informational,2021-06-15T04:44:38.223Z,20c6
3 22,66641e13-d19f-4ce5-aafd-9d5d7bfa557,Delete SQL database,Started,Administrative,Informational,2021-06-15T04:44:21.547Z,20c6
4 23,66641e13-d19f-4ce5-aafd-9d5d7bfa557,Delete SQL database,Accepted,Administrative,Informational,2021-06-15T04:44:21.702Z,20c6
5 24,e2958162-93d9-4643-a847-82cf25c49930,Delete SqlPools,Succeeded,Administrative,Informational,2021-06-15T04:44:31.332Z,20c6ee
6 25,e2958162-93d9-4643-a847-82cf25c49930,Delete SqlPools,Started,Administrative,Informational,2021-06-15T04:44:12.533Z,20c6ee
7 26,e2958162-93d9-4643-a847-82cf25c49930,Delete SqlPools,Accepted,Administrative,Informational,2021-06-15T04:44:16.038Z,20c6ee
8 27,08cd2e19-477c-4ecc-83a6-575b9ce265e3,Pause SQL Analytics pools.,Succeeded,Administrative,Informational,2021-06-14T17:57:02.
9 28,08cd2e19-477c-4ecc-83a6-575b9ce265e3,Pause SQL Analytics pools.,Started,Administrative,Informational,2021-06-14T17:55:17.61
10 29,08cd2e19-477c-4ecc-83a6-575b9ce265e3,Pause SQL Analytics pools.,Accepted,Administrative,Informational,2021-06-14T17:55:18.5
11 30,d2d9d7c4-2766-4e7d-83fd-c4f07b3d6825,Pause a Datawarehouse database.,Succeeded,Administrative,Informational,2021-06-14T17:5
12 31,d2d9d7c4-2766-4e7d-83fd-c4f07b3d6825,Pause a Datawarehouse database.,Started,Administrative,Informational,2021-06-14T17:55:
13 32,d2d9d7c4-2766-4e7d-83fd-c4f07b3d6825,Pause a Datawarehouse database.,Accepted,Administrative,Informational,2021-06-14T17:55:
14 33,1c735927-517e-470f-bccf-239cdf3b97f8,Create Pipeline Run,Succeeded,Administrative,Informational,2021-06-14T14:24:48.367Z,20c6
15 34,1c735927-517e-470f-bccf-239cdf3b97f8,Create Pipeline Run,Started,Administrative,Informational,2021-06-14T14:24:45.734Z,20c6
16 35,0ae558f0-ebbd-4e9e-b3de-406cf39ae0ad,Create or Update any Pipeline.,Succeeded,Administrative,Informational,2021-06-14T14:24:
17 36,0ae558f0-ebbd-4e9e-b3de-406cf39ae0ad,Create or Update any Pipeline.,Started,Administrative,Informational,2021-06-14T14:24:4
18 37,072d5d31-b4b0-4bd1-bfef-088c7f11eb,Create or Update Dataset,Succeeded,Administrative,Informational,2021-06-14T14:24:44.26
19 38,072d5d31-b4b0-4bd1-bfef-088c7f11eb,Create or Update Dataset,Started,Administrative,Informational,2021-06-14T14:24:43.264Z
20 39,77152ae0-297f-4d10-9179-90d50bec8dac,Create or Update Dataset,Succeeded,Administrative,Informational,2021-06-14T14:24:44.18
21 40,77152ae0-297f-4d10-9179-90d50bec8dac,Create or Update Dataset,Succeeded,Administrative,Informational,2021-06-14T14:24:44.18
22
```

En Azure Data Factory comenzamos creando nuestro Data Flow



Microsoft Azure | appfactory5000

Search

techsup1000@gmail.com
DEFAULT DIRECTORY

>>

Data Factory

Validate all

Publish all 1

IncrementPipeline

schemadrift

✓ Validate

⏻ Data flow debug

Logfiles

Columns: 0 total

Add Source

Source settings

Source options

Projection

Optimize

Output stream name *

Logfiles

Source type *

Dataset

Inline

Dataset *

Select...

Options

✓ Allow schema drift ⓘ

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 1

IncrementPipeline

schemadrift

✓ Validate

⏻ Data flow debug

Logfiles

Columns: 0 total

Add Source

Source settings

Source options

Projection

Optimize

Output stream name *

Logfiles

Source type *

Dataset

Inline

Dataset *

Select...

Options

✓ Allow schema drift ⓘ

Select format

Choose the format type of your data

Avro

DelimitedText

Excel

JSON

ORC

Parquet

XML

01

Continue

Back

Cancel

Escogemos la ruta donde se encuentran ambos archivos CSV, en este caso, dentro del contenedor schema. Por otro lado, no importaremos el schema. Ahora, la razón de esto es cuando estamos permitiendo **Schema Drift**, que como puedes ver tiene una opción que se activa automáticamente en estos ajustes. Esto permite que tu esquema sea diferente cuando se trata de archivos fuente. Ahora, si importo el esquema, mirará el archivo Log.csv, que tenemos en el contenedor "schema". Así que, si recuerdas, si vas al contenedor "schema", mirará el archivo Log.csv y tomará o importará el esquema de ahí. ¿Pero qué pasa con el schema del otro archivo? Así que cuando se desea implementar el Schema Drift, es bueno que Azure Data Factory determine el schema en runtime cuando está copiando cada archivo. En lugar de importar el schema y decirle a Azure Data Factory, como debería ser el schema (eso lo haríamos escogiendo la opción "From connection/store"), en lugar de eso, deja que Azure Data Factory, determine el schema en runtime por sí mismo. Es por eso que estoy eligiendo el schema de importación "none".

Microsoft Azure | appfactory5000

Search

techsup1000@gmail.com

Set properties

Name: LogFilesDt

Linked service *: AzureDataLakeStorage

File path: schema / Directory / File

First row as header: ☒

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

Advanced

OK Back Cancel

Jerarquía en Data Lake Storage:

- Contenedor: schema
 - Archivo: Log.csv
 - Directorio: newfiles
 - Archivo: Log-withcolumnless.csv

Source settings Source options Projection Optimize

Output stream name *: LogFiles

Source type *: Dataset Inline

Dataset *: Select...

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

Sampling *: ☐ Enable ☒ Disable

Ahora, antes de que pueda añadir un fregadero, voy a añadir una Derived column. Así que aquí en la Derived column, quiero decir que, cualesquiera que sean las columnas que están aquí (en el paso anterior, en Source), porque se puede ver que no hay columnas que están siendo detectadas por Azure data factory porque le indicamos que no detectara esas columnas. Se quería inferir las columnas en tiempo de ejecución (en runtime).

The screenshot shows the Microsoft Azure Data Factory interface for a pipeline named 'IncrementPipeline'. The 'Derived column's settings' panel is open, showing the 'Columns' section with a red circle around the 'Add or select a column...' dropdown and an arrow pointing to the 'Expression' field. The 'Output stream name' is 'DerivedColumn' and the 'Incoming stream' is 'LogFiles'.

Así que ahora quiero añadir también una Derived column para decirle a Azure Data Factory, que cualquier columna que encuentre en el paso anterior "Source", la derive tal cual. Así que aquí en el mapeo de columna, voy a añadir un patrón de columna.

The screenshot shows the Microsoft Azure Data Factory interface for a pipeline named 'IncrementPipeline'. The 'Derived column's settings' panel is open, showing the 'Columns' section. The 'Add column pattern' option is highlighted with a red box. The 'Output stream name' is 'DerivedColumn' and the 'Incoming stream' is 'LogFiles'.

Así que cada vez que hay una columna, del tipo "string" (y en nuestro archivo de ejemplo, todas las columnas son de tipo "string"), a continuación, que mapee cualquiera sea el nombre de la llamada entrante, que en realidad se puede abordar con dos símbolos de dólar "\$\$". Y cuando se copie en el destino, que tome el mismo nombre de columna. Así que, si un nombre de columna es "Id", entonces copiará en el destino con el nombre "Id". Entonces en el Derived Column sólo estamos diciendo "por favor, mapea las columnas como están". Estamos haciendo esto porque recuerda, nuestras columnas no están definidas. Necesitamos que se definan automáticamente en runtime.

Microsoft Azure | appfactory5000

Search

techsup1000@gmail.com
DEFAULT DIRECTORY

Data Factory | IncrementPipeline | schemadrift

Validate | Data flow debug

LogFiles
Import data from LogFilesDt

DerivedColumn
Columns: 0 total

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

Output stream name * | DerivedColumn | Learn more

Incoming stream * | LogFiles

+ Add | Clone | Delete | Open expression builder

Columns * 1

Column	Expression
<input type="checkbox"/> Each column that matches	<input type="text" value="type == 'string'"/> creates 1 column(s)
<input type="text" value="\$"/>	<input type="text" value="ANY"/>

Open expression builder

Microsoft Azure | appfactory5000

Search

techsup1000@gmail.com
DEFAULT DIRECTORY

Data Factory | IncrementPipeline | schemadrift

Validate | Data flow debug

LogFiles | DerivedColumn | sink1

Import data from LogFilesDt | Creating/updating the columns | Columns: 0 total

Sink Settings Mapping Optimize Inspect Data preview

Output stream name * JsonSink [Learn more](#)

Incoming stream * DerivedColumn

Sink type * Dataset Inline Cache

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

Options ☒ Allow schema drift ☐ Validate schema

En el dataset de destino escogemos el Azure Data Lake Storage y como tipo de dato JSON

Microsoft Azure | appfactory5000

Search

techsup1000@gmail.com
DEFAULT DIRECTORY

Data Factory | IncrementPipeline | schemadrift

Validate | Data flow debug

LogFiles | DerivedColumn

Import data from LogFilesDt | Creating/updating the columns

Sink Settings Mapping Optimize Inspect Data preview

Output stream name * JsonSink

Incoming stream * DerivedColumn

Sink type * Dataset Inline

Dataset * Select...

Options ☒ Allow schema drift ☐ Validate schema

Set properties

Name LogJsonNew

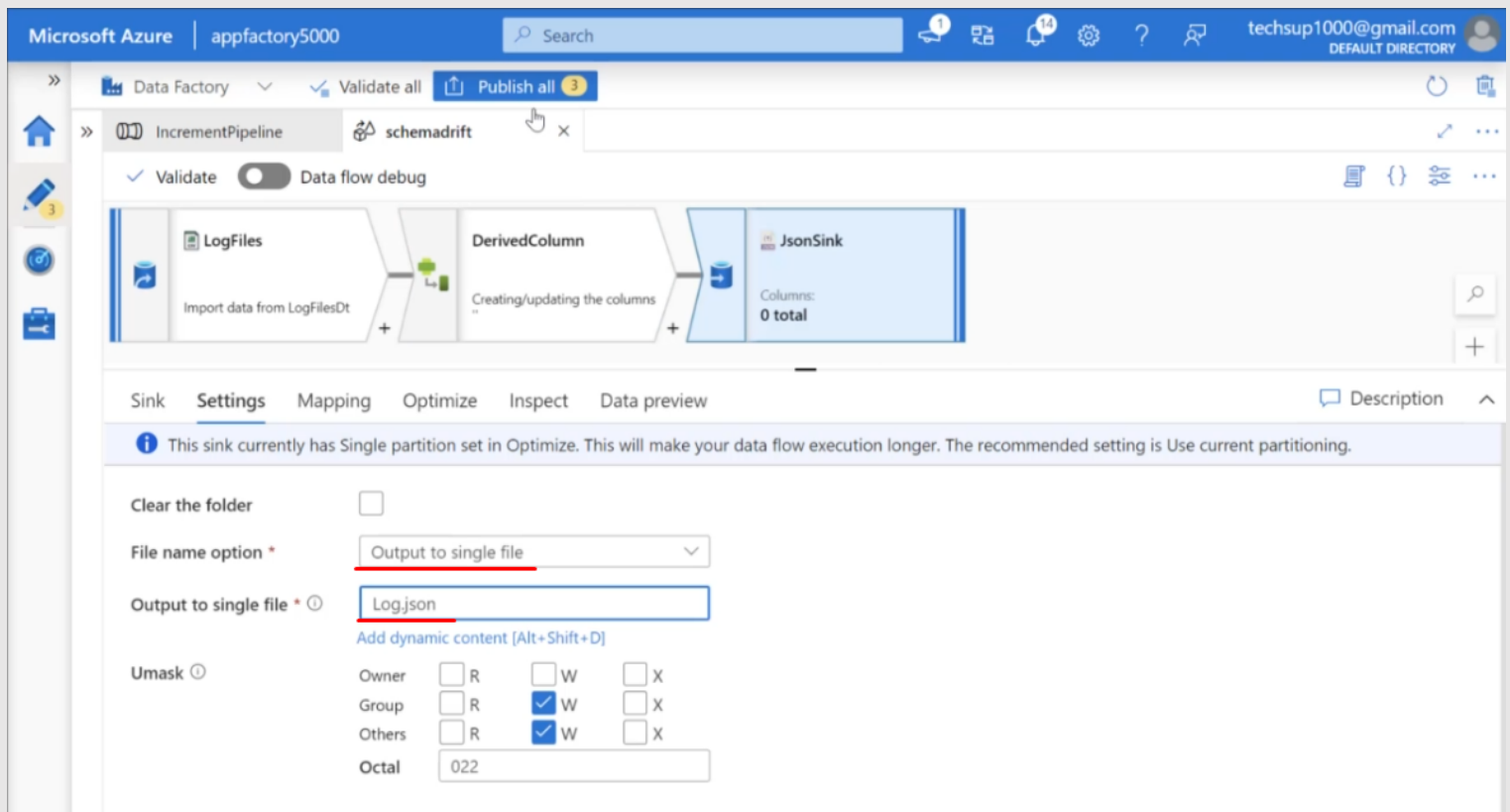
Linked service * AzureDataLakeStorage

File path schemadrift / Directory / File

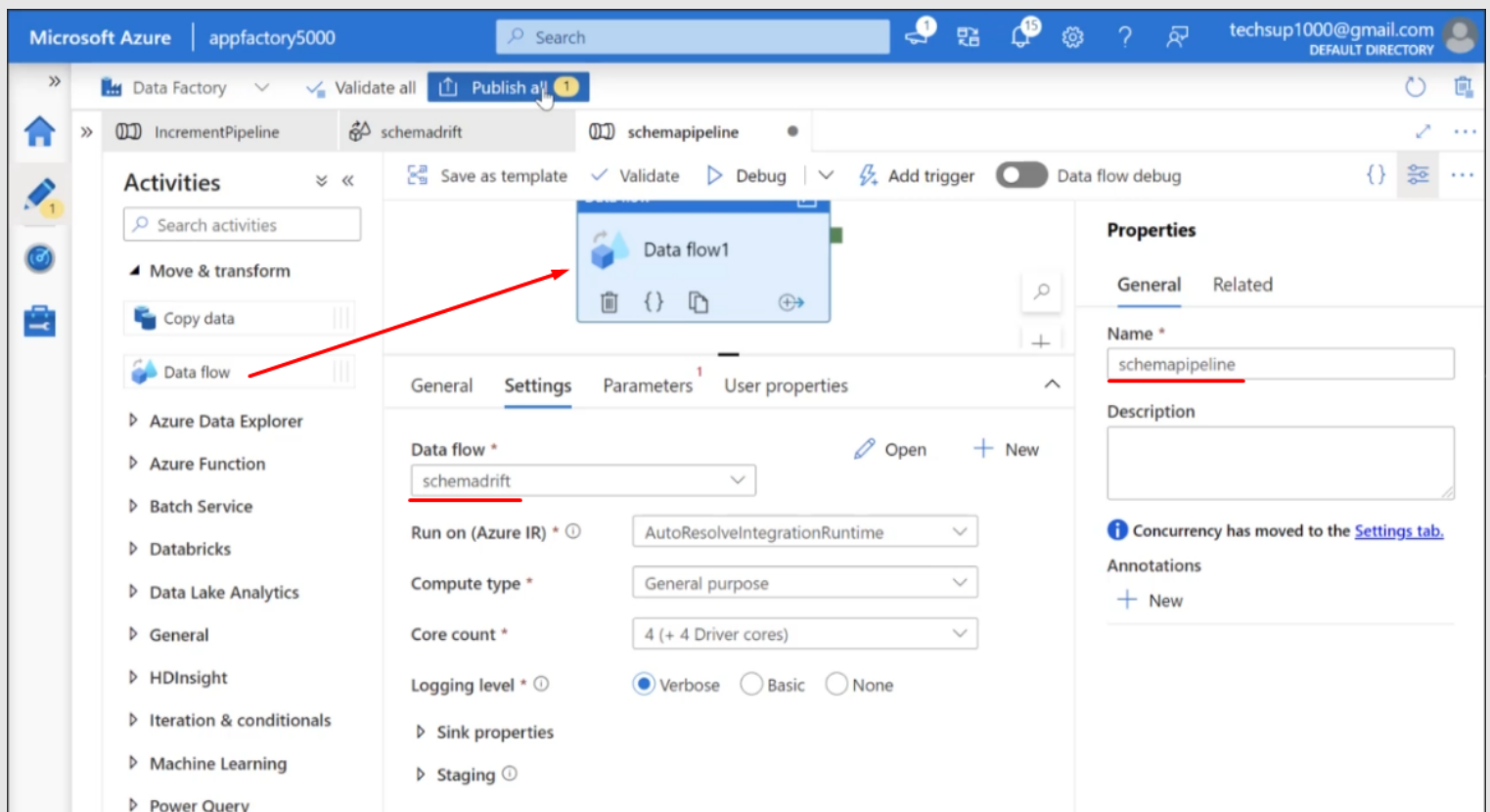
Import schema ☐ From connection/store ☐ From sample file ☒ None

Advanced

OK Back Cancel



Luego, creamos un Pipeline para ejecutar nuestro Data Flow. Luego **Validamos, Publicamos y Ejecutamos.**



Microsoft Azure | appfactory5000

Search

techsup1000@gmail.com
DEFAULT DIRECTORY

Data Factory | Validate all | Publish all

IncrementPipeline | schemadrift | schemapipeline

Activities

Search 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

Save as template | Validate | Debug | Add trigger | Data flow debug

Trigger now

Trigger on-demand run of the last published pipeline

New/Edit

General | Settings | Parameters | User properties

Data flow * | schemadrift

Run on (Azure IR) * | AutoResolveIntegrationRuntime

Compute type * | General purpose

Core count * | 4 (+ 4 Driver cores)

Logging level * | ☒ Verbose ☐ Basic ☐ None

Sink properties

Staging

General | Related

Name * | schemapipeline

Description

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

Annotations

+ New

Si nos dirigimos a nuestro Storage Data Lake

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

techsup1000@gmail.com
DEFAULT DIRECTORY (TECHSUP1...)

Dashboard > Storage accounts > datalake2000 >

schemadrift

Container

Search (Ctrl+ /)

Upload | Add Directory | Refresh | Rename | Delete | Change tier | ...

Authentication method: Access key (Switch to Azure AD User Account)

Location: schemadrift

Search blobs by prefix (case-sensitive)

	Name	Modified	Access tier	Blob type
<input type="checkbox"/>	Log.json	7/23/2021, 11:46:37 ...	Hot (Inferred)	Block blob

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

Dashboard > Storage accounts > datalake2000 > schemadrift >

Log.json

Blob

Save Discard Download Refresh Delete

Overview Versions Edit Generate SAS

```
1 {"Id": "21", "Correlationid": "66641e13-d19f-4ce5-aafd-9d5d7bfa557", "Operationname": "Delete SQL database", "Status": "Succeeded"}
2 {"Id": "22", "Correlationid": "66641e13-d19f-4ce5-aafd-9d5d7bfa557", "Operationname": "Delete SQL database", "Status": "Succeeded"}
3 {"Id": "23", "Correlationid": "66641e13-d19f-4ce5-aafd-9d5d7bfa557", "Operationname": "Delete SQL database", "Status": "Succeeded"}
4 {"Id": "24", "Correlationid": "e2958162-93d9-4643-a847-82cf25c49930", "Operationname": "Delete SqlPools", "Status": "Succeeded"}
5 {"Id": "25", "Correlationid": "e2958162-93d9-4643-a847-82cf25c49930", "Operationname": "Delete SqlPools", "Status": "Succeeded"}
6 {"Id": "26", "Correlationid": "e2958162-93d9-4643-a847-82cf25c49930", "Operationname": "Delete SqlPools", "Status": "Succeeded"}
7 {"Id": "27", "Correlationid": "08cd2e19-477c-4ecc-83a6-575b9ce265e3", "Operationname": "Pause SQL Analytics pools.", "Status": "Succeeded"}
8 {"Id": "28", "Correlationid": "08cd2e19-477c-4ecc-83a6-575b9ce265e3", "Operationname": "Pause SQL Analytics pools.", "Status": "Succeeded"}
9 {"Id": "29", "Correlationid": "08cd2e19-477c-4ecc-83a6-575b9ce265e3", "Operationname": "Pause SQL Analytics pools.", "Status": "Succeeded"}
10 {"Id": "30", "Correlationid": "d2d9d7c4-2766-4e7d-83fd-c4f07b3d6825", "Operationname": "Pause a Datawarehouse database.", "Status": "Succeeded"}
11 {"Id": "31", "Correlationid": "d2d9d7c4-2766-4e7d-83fd-c4f07b3d6825", "Operationname": "Pause a Datawarehouse database.", "Status": "Succeeded"}
12 {"Id": "32", "Correlationid": "d2d9d7c4-2766-4e7d-83fd-c4f07b3d6825", "Operationname": "Pause a Datawarehouse database.", "Status": "Succeeded"}
```

Json Preview

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

Dashboard > Storage accounts > datalake2000 > schemadrift >

Log.json


Blob

Save Discard Download Refresh Delete

Overview Versions Edit Generate SAS

```
12 {"Id": "32", "Correlationid": "d2d9d7c4-2766-4e7d-83fd-c4f07b3d6825", "Operationname": "Pause a Datawarehouse database.", "Status": "Succeeded"}
13 {"Id": "33", "Correlationid": "1c735927-517e-470f-bccf-239cdf3b97f8", "Operationname": "Create Pipeline Run", "Status": "Succeeded"}
14 {"Id": "34", "Correlationid": "1c735927-517e-470f-bccf-239cdf3b97f8", "Operationname": "Create Pipeline Run", "Status": "Succeeded"}
15 {"Id": "35", "Correlationid": "0ae558f0-ebbd-4e9e-b3de-406cf39ae0ad", "Operationname": "Create or Update any Pipeline.", "Status": "Succeeded"}
16 {"Id": "36", "Correlationid": "0ae558f0-ebbd-4e9e-b3de-406cf39ae0ad", "Operationname": "Create or Update any Pipeline.", "Status": "Succeeded"}
17 {"Id": "37", "Correlationid": "072d5d31-b4b0-4bd1-bfef-088c7faf11eb", "Operationname": "Create or Update Dataset", "Status": "Succeeded"}
18 {"Id": "38", "Correlationid": "072d5d31-b4b0-4bd1-bfef-088c7faf11eb", "Operationname": "Create or Update Dataset", "Status": "Succeeded"}
19 {"Id": "39", "Correlationid": "77152ae0-297f-4d10-9179-90d50bec8dac", "Operationname": "Create or Update Dataset", "Status": "Succeeded"}
20 {"Id": "40", "Correlationid": "77152ae0-297f-4d10-9179-90d50bec8dac", "Operationname": "Create or Update Dataset", "Status": "Succeeded"}
21 {"Id": "1", "Correlationid": "66641e13-d19f-4ce5-aafd-9d5d7bfa557", "Operationname": "Delete SQL database", "Status": "Succeeded"}
22 {"Id": "2", "Correlationid": "66641e13-d19f-4ce5-aafd-9d5d7bfa557", "Operationname": "Delete SQL database", "Status": "Succeeded"}
23 {"Id": "3", "Correlationid": "66641e13-d19f-4ce5-aafd-9d5d7bfa557", "Operationname": "Delete SQL database", "Status": "Succeeded"}
```

Json Preview

 Preview