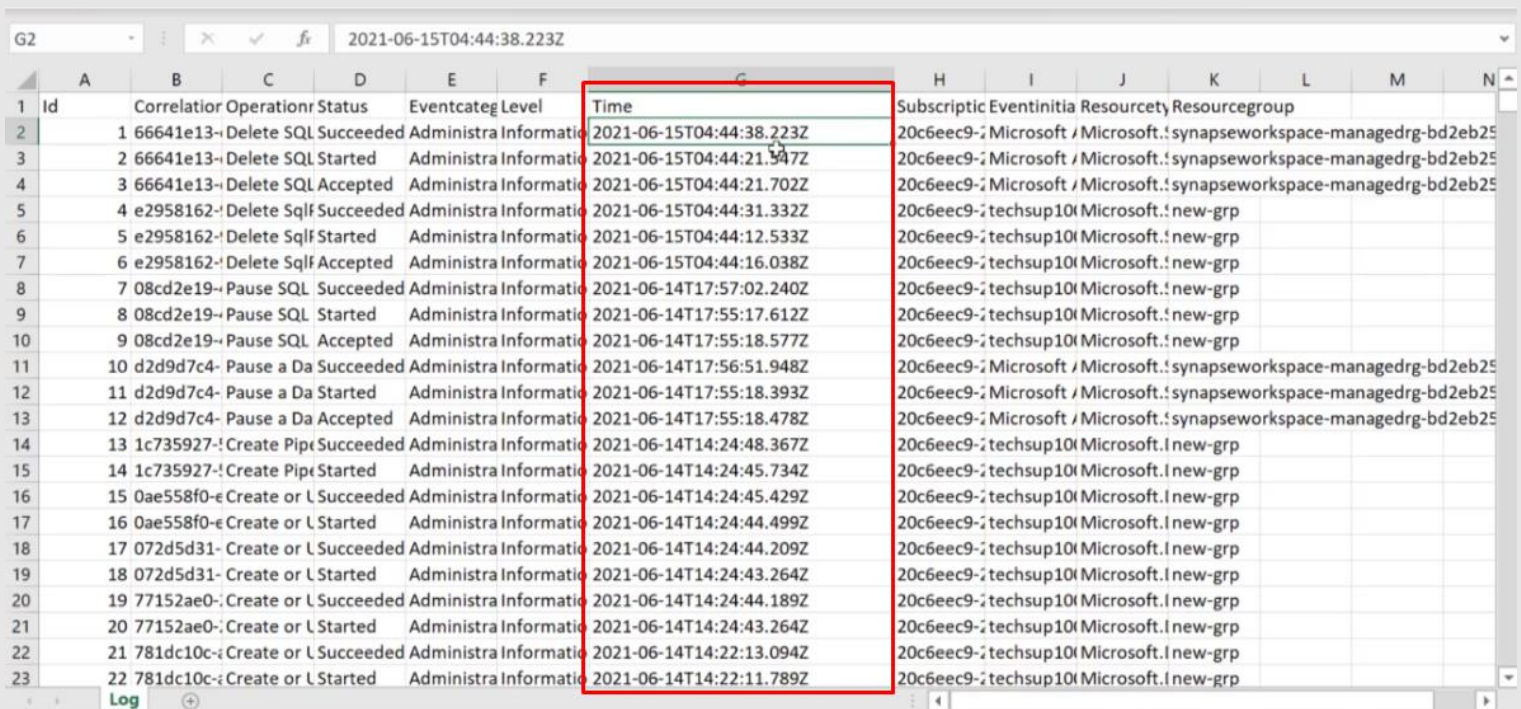


Azure Data Factory

Modificar una columna de fecha con Data Flow

Buscamos modificar la columna "Time" del archivo 'Log.csv' y cambiar su estructura para que pueda ser utilizado y no devuelva errores.



The screenshot shows an Excel spreadsheet with a log file. The 'Time' column (G) is highlighted with a red box. The data in the 'Time' column is as follows:

Id	Correlator	Operation	Status	Eventcategory	Level	Time	Subscription	Eventinitial	Resource	Resourcegroup
1	66641e13-	Delete SQL	Succeeded	Administrative	Information	2021-06-15T04:44:38.223Z	20c6eec9-7	Microsoft	Microsoft	synapseworkspace-managedrg-bd2eb25
2	66641e13-	Delete SQL	Started	Administrative	Information	2021-06-15T04:44:21.547Z	20c6eec9-7	Microsoft	Microsoft	synapseworkspace-managedrg-bd2eb25
3	66641e13-	Delete SQL	Accepted	Administrative	Information	2021-06-15T04:44:21.702Z	20c6eec9-7	Microsoft	Microsoft	synapseworkspace-managedrg-bd2eb25
4	e2958162-	Delete Sqlf	Succeeded	Administrative	Information	2021-06-15T04:44:31.332Z	20c6eec9-7	techsup10f	Microsoft	new-grp
5	e2958162-	Delete Sqlf	Started	Administrative	Information	2021-06-15T04:44:12.533Z	20c6eec9-7	techsup10f	Microsoft	new-grp
6	e2958162-	Delete Sqlf	Accepted	Administrative	Information	2021-06-15T04:44:16.038Z	20c6eec9-7	techsup10f	Microsoft	new-grp
7	08cd2e19-	Pause SQL	Succeeded	Administrative	Information	2021-06-14T17:57:02.240Z	20c6eec9-7	techsup10f	Microsoft	new-grp
8	08cd2e19-	Pause SQL	Started	Administrative	Information	2021-06-14T17:55:17.612Z	20c6eec9-7	techsup10f	Microsoft	new-grp
9	08cd2e19-	Pause SQL	Accepted	Administrative	Information	2021-06-14T17:55:18.577Z	20c6eec9-7	techsup10f	Microsoft	new-grp
10	d2d9d7c4-	Pause a Da	Succeeded	Administrative	Information	2021-06-14T17:56:51.948Z	20c6eec9-7	Microsoft	Microsoft	synapseworkspace-managedrg-bd2eb25
11	d2d9d7c4-	Pause a Da	Started	Administrative	Information	2021-06-14T17:55:18.393Z	20c6eec9-7	Microsoft	Microsoft	synapseworkspace-managedrg-bd2eb25
12	d2d9d7c4-	Pause a Da	Accepted	Administrative	Information	2021-06-14T17:55:18.478Z	20c6eec9-7	Microsoft	Microsoft	synapseworkspace-managedrg-bd2eb25
13	1c735927-	Create Pipe	Succeeded	Administrative	Information	2021-06-14T14:24:48.367Z	20c6eec9-7	techsup10f	Microsoft	new-grp
14	1c735927-	Create Pipe	Started	Administrative	Information	2021-06-14T14:24:45.734Z	20c6eec9-7	techsup10f	Microsoft	new-grp
15	0ae558f0-e	Create or U	Succeeded	Administrative	Information	2021-06-14T14:24:45.429Z	20c6eec9-7	techsup10f	Microsoft	new-grp
16	0ae558f0-e	Create or U	Started	Administrative	Information	2021-06-14T14:24:44.499Z	20c6eec9-7	techsup10f	Microsoft	new-grp
17	072d5d31-	Create or U	Succeeded	Administrative	Information	2021-06-14T14:24:44.209Z	20c6eec9-7	techsup10f	Microsoft	new-grp
18	072d5d31-	Create or U	Started	Administrative	Information	2021-06-14T14:24:43.264Z	20c6eec9-7	techsup10f	Microsoft	new-grp
19	77152ae0-	Create or U	Succeeded	Administrative	Information	2021-06-14T14:24:44.189Z	20c6eec9-7	techsup10f	Microsoft	new-grp
20	77152ae0-	Create or U	Started	Administrative	Information	2021-06-14T14:24:43.264Z	20c6eec9-7	techsup10f	Microsoft	new-grp
21	781dc10c-	Create or U	Succeeded	Administrative	Information	2021-06-14T14:22:13.094Z	20c6eec9-7	techsup10f	Microsoft	new-grp
22	781dc10c-	Create or U	Started	Administrative	Information	2021-06-14T14:22:11.789Z	20c6eec9-7	techsup10f	Microsoft	new-grp

Microsoft Azure | appfactory5000

Search

techsup1000@gmail.com

Data Factory | Validate all | Publish all 1

Timeflow

Validate | Data flow debug | Debug Settings

source1

Columns: 0 total

Add Source

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

Output stream name * | Logsource | Learn more

Source type * | Dataset | Inline

Dataset * | Select... | + New

Options | ☒ Allow schema drift

Properties

General | Related

Name * | Timeflow

Description

El dataset apuntará a Azure Data Lake Storage.

Microsoft Azure | appfactory5000

Search

techsup1000@gmail.com

Data Factory | Validate all | Publish all 1

Timeflow

Validate | Data flow debug | Debug Settings

Logsource

Columns: 0 total

Add Source

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

Output stream name * | Logsource

Source type * | Dataset | Inline

Dataset * | Select...

Options | ☒ Allow schema drift

Set properties

Name | LogSourceTime

Linked service * | AzureDataLakeStorage

File path | data / raw / Log.csv

First row as header | ☒

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

Timeflow

Validate | Data flow debug | Debug Settings

Logsource
Import data from LogSourceTime

ChangeTime
Columns: 11 total

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

Output stream name *
ChangeTime

Incoming stream *
Logsource

+ Add | Clone | Delete | Open expression builder

Columns * 1

Column	Expression
Time	Enter expression... Open expression builder

Microsoft Azure | appfactory5000

Search

techsup1000@gmail.com
DEFAULT DIRECTORY

Visual expression builder

ChangeTime

Derived Columns
+ Create new

ANY Time

Column name *
Time

Expression
`toTimestamp(concat(substring(Time,0,10),' ',substring(Time,12,8)), 'yyyy-MM-dd HH:mm:ss')`

Save

Expression elements
All
Functions
Input schema
Parameters
Cached lookup

Expression values
Filter by keyword
+ Create new
Id
Correlationid
Operationname

Data preview

Save and finish | Cancel | Clear contents

Microsoft Azure | appfactory5000

Visual expression builder

ChangeTime

Derived Columns

+ Create new

Time

Column name *

Time

Expression

toTimestamp(concat(substring(Time,0,10),' ',substring(Time,12,8)), 'yyyy-MM-dd HH:mm:ss')

Save

Expression values

Data preview Refresh

Output: Time	Time
2021-06-15 04:44:38	2021-06-15T04:44:38.223Z
2021-06-15 04:44:21	2021-06-15T04:44:21.547Z
2021-06-15 04:44:21	2021-06-15T04:44:21.702Z
2021-06-15 04:44:31	2021-06-15T04:44:31.337Z

Save and finish Cancel Clear contents

Sin embargo, los tipos de datos en el archivo los dejaremos tal cual estaban, es decir, en el tipo STRING. Aunque también es posible 'proyectar' el tipo de dato desde el paso fuente, desde la caja "Logsource" en la pestaña "Projection".

Microsoft Azure | appfactory5000

Data Factory

Validate all Publish all

Timeflow

Validate Data flow debug Debug Settings

Logsource

Columns: 11 total

ChangeTime

Creating/updating the columns
Id, Correlationid, Operationname, Status, Eventcategory, Level, Time,

Source settings Source options Projection Optimize Inspect Data preview Description

Output stream name * Logsource

Source type * Dataset Inline

Dataset * LogSourceTime

Test connection Open New

Options

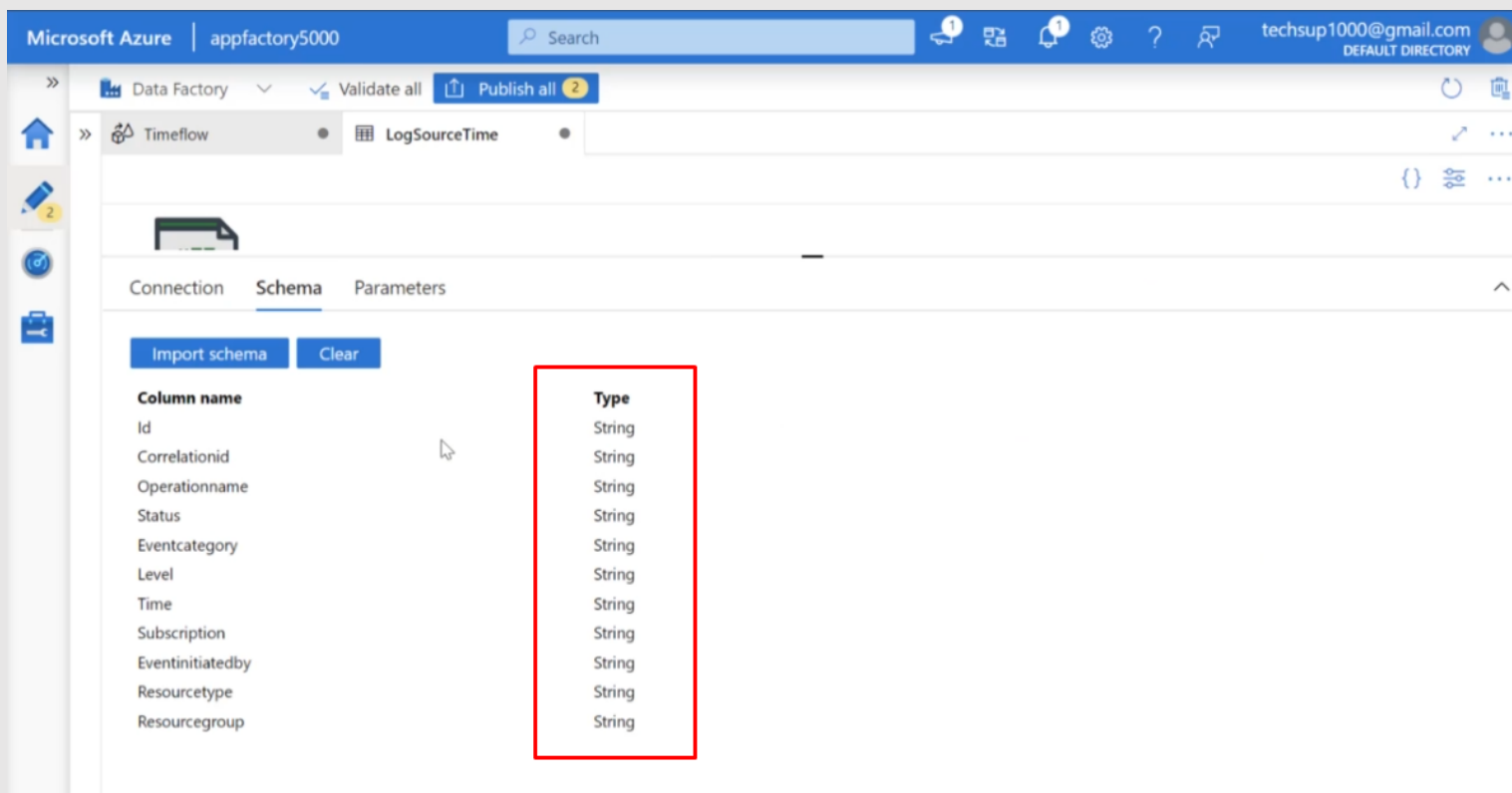
☒ Allow schema drift

☐ Infer drifted column types

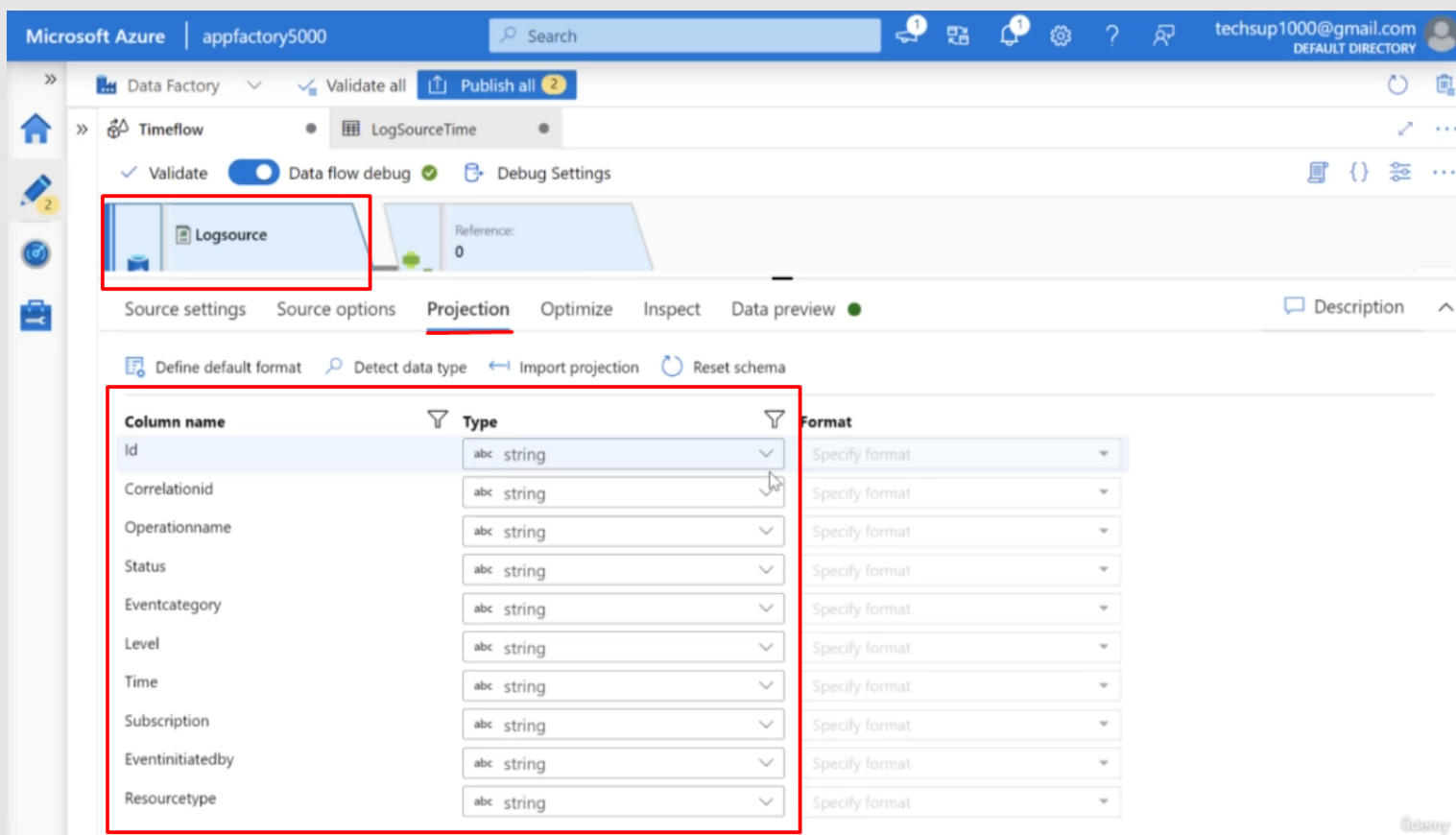
☐ Validate schema

Skip line count

Sampling * ☐ Enable ☒ Disable



Es aquí donde podríamos modificar el tipo de dato. Pero no lo haremos.



Luego, creamos el paso de destino.

Microsoft Azure | appfactory5000

Search

techsup1000@gmail.com
DEFAULT DIRECTORY

Data Factory | Validate all | Publish all (2)

Timeflow | LogSourceTime

Validate | Data flow debug | Debug Settings

Logsource: Import data from LogSourceTime

Reference: 1
Columns: 11 total

sink1
Columns: 11 total

Sink Settings:

- Output stream name: sink1
- Incoming stream: ChangeTime
- Sink type: Dataset
- 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)

Timeflow | LogSourceTime

Validate | Data flow debug | Debug Settings

Sink Settings:

- Output stream name: sink1
- Incoming stream: ChangeTime
- Sink type: Dataset
- 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 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 2

Timeflow | LogSourceTime

Validate | Data flow debug | Debug Settings

Sink Settings Mapping Optimize Inspect Data prev

Output stream name * sink1

Incoming stream * ChangeTime

Sink type * Dataset Inline

Dataset * Select...

Options ☒ Allow schema drift ☐ Validate schema

Select format

Choose the format type of your data

Avro DelimitedText JSON

ORC Parquet Binary

Continue Back Cancel

data: contenedor ; **cleaned:** directorio

Microsoft Azure | appfactory5000

Search

techsup1000@gmail.com
DEFAULT DIRECTORY

>> Data Factory | Validate all | Publish all 2

Timeflow | LogSourceTime

Validate | Data flow debug | Debug Settings

Sink Settings Mapping Optimize Inspect Data prev

Output stream name * sink1

Incoming stream * ChangeTime

Sink type * Dataset Inline

Dataset * Select...

Options ☒ Allow schema drift ☐ Validate schema

Set properties

Name LogTimeOutput

Linked service * AzureDataLakeStorage

File path data / cleaned / File

First row as header ☒

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 (3)

Timeflow | LogSourceTime

Validate | Data flow debug | Debug Settings

Logsource: Import data from LogSourceTime

Reference: 1
Columns: 11 total

sink1: Columns: 11 total

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

This sink currently has Single partition set in Optimize. This will make your data flow execution longer. The recommended setting is Use current partitioning.

Clear the folder ☐

File name option * Output to single file

Output to single file *
Add dynamic content [Alt+Shift+D]

Quote All ☐

Headers ANY

Umask ☐ Owner ☐ R ☐ W ☐ X
Group ☐ R ☒ W ☐ X
Others ☐ R ☒ W ☐ X

File name option: vemos las opciones de partición del archivo. El output será un archivo único, aunque no se que tipo de partición se podría hacer a un archivo de texto.

Ahora creamos un nuevo Pipeline y agregamos una actividad **Dataflow** la cual va a hacer referencia al dataflow que hemos desarrollado. **Validamos** y **Publicamos**, y luego **ejecutamos**.

Microsoft Azure | appfactory5000

Search

techsup1000@gmail.com
DEFAULT DIRECTORY

Data Factory | Validate all | Publish all

Timeflow | LogSourceTime | LogConvert

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

Trigger now
Trigger on-demand run of the last published pipeline
New/Edit

Publishing completed
Successfully published

General | **Settings** | Parameters | User properties

Data flow * Timeflow Open + New

Run on (Azure IR) * AutoResolveIntegrationRuntime

Compute type * General purpose

Core count * 4 (+ 4 Driver cores)

Logging level * ☒ Verbose ☐ Basic ☐ None

Sink properties

Staging

Microsoft Azure

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

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

>>

Dashboard > All resources > datalake2000 >

data

Container

Search (Ctrl+ /)

<<

Upload

Add Directory

Refresh

Rename

Delete

Change tier

...

Overview

Diagnose and solve problems

Access Control (IAM)

Settings

Shared access tokens

Manage ACL

Access policy

Properties

Metadata

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

Location: data / cleaned

Search blobs by prefix (case-sensitive)

Name	Modified	Access tier	Blob type
<input type="checkbox"/> [..]			
<input type="checkbox"/> Log.csv	7/19/2021, 9:56:45 PM	Hot (Inferred)	Block blob