

# RG\_PROYECTO\_FINAL\_DE\_AZURE\_ESTEBAN\_SAENZ

- Proyecto final curso Azure Datapath.
- PF. ESTEBAN\_SAENZ\_20230827
- Esteban Sáenz V. (esaenz7@gmail.com).
- Repositorio en github.

## Antecedentes

Este proyecto consiste en un flujo de datos ETL, el cual permite obtener las fuentes de un repositorio local, almacenándolas en la primera capa (bronce) de un datalake. Seguidamente, los datos son unidos en un único conjunto en la segunda capa (silver), para que finalmente sean transformados, preparados - a través de un notebook y un job de spark - y almacenados en la tercera capa (gold). A partir de esta capa los datos pueden ser utilizados como insumos para modelos analíticos, tableros de información, o modelos de IA.

Contexto y alcance:

El proyecto tiene como contexto los vuelos domésticos registrados en EEUU durante el año 2018, al cual se adjuntan los registros de eventos meteorológicos y su severidad ocurridos para cada día del año y para cada aeropuerto en particular. Ambos datasets son unidos gracias al conjunto de datos proporcionado por la lista extensa de aeropuertos del repositorio "openflights.org" el cual aporta tanto el código IATA (International Air Transport Association) utilizado coloquialmente para el manejo de vuelos comerciales, como el código oficial ICAO (International Civil Aviation Organization) utilizado para identificar las estaciones de medición meteorológica.

Ruta:

- a) Realizar la carga, exploración, análisis, filtrado y limpieza de datos correspondiente para cada dataset.
- b) Determinar a partir de la columna de demora en el arribo "arr\_delay", el valor de tiempo que permitirá realizar la clasificación binaria entre un vuelo demorado y un vuelo a tiempo.
- c) Definir la columna correspondiente para dichas clases. La columna objetivo o "target" se llamará "delayed" y contendrá el valor booleano para las 2 clases posibles de resultados: "demorado" o "a tiempo".
- d) Realizar la unión de los datasets a partir de las columnas correspondientes según se detalla con el siguiente pseudocódigo (algunos nombres de columnas son modificados durante la fase de preparación para facilitar su uso):

```
dataframe -> join(flights.origin == airports.iata) dataframe -> join(airports.icao == weather.airportcode & flights.date == weather.date)
```

En este caso se realizará la unión de los conjuntos de datos principales de vuelos ("flights") y eventos meteorológicos ("weather") por medio de un dataset intermedio con la lista de aeropuertos ("airports"), el cual posee los códigos IATA e ICAO que utilizan respectivamente los datasets principales. Además se incorpora una llave adicional entre fechas, para mantener la relación del vuelo con el evento meteorológico reportado según el día en particular.

- e) Filtrar y preparar las columnas y sub-conjuntos para el proceso de aprendizaje automático.

## Arquitectura de datos

### Azure Storage Account

- Contenedor y capas:

Home > Storage accounts > esaenz7azsa101 | Containers >

esaenzazdl101

Container

Search

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: esaenzazdl101

Search blobs by prefix (case-sensitive)

Show deleted objects

	Name	Modified	Access tier	Archive status	Blob type	Size
<input type="checkbox"/>	📁 bronze					
<input type="checkbox"/>	📁 gold					
<input type="checkbox"/>	📁 silver					
<input type="checkbox"/>	📁 synapse					

- Capa bronce luego de la ejecución del pipeline:

Home > Storage accounts > esaenz7azsa101 | Containers >

esaenzazdl101

Container

Search

«

Overview

Diagnose and solve problems

Access Control (IAM)

Settings

Shared access tokens

Manage ACL

Access policy

Properties

Metadata

Upload

Add Directory

Refresh

Rename

Delete

Change tier

Acquire lease

Break lease

Give feedback

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

Location: esaenzazdl101 / bronze

Search blobs by prefix (case-sensitive)

Show deleted objects

Name	Modified	Access tier	Archive status	Blob type	Size
<input type="checkbox"/> [-]					
<input type="checkbox"/> data1.csv	8/27/2023, 5:03:53 PM	Hot (Inferred)		Block blob	16.53 KiB
<input type="checkbox"/> data2.csv	8/27/2023, 5:03:57 PM	Hot (Inferred)		Block blob	1.16 MiB
<input type="checkbox"/> data3.csv	8/27/2023, 5:08:34 PM	Hot (Inferred)		Block blob	93.88 MiB

• Capa silver luego de la ejecución del pipeline:

Home > Storage accounts > esaenz7azsa101 | Containers >

esaenzazdl101

Container

Search

«

Overview

Diagnose and solve problems

Access Control (IAM)

Settings

Shared access tokens

Manage ACL

Access policy

Properties

Metadata

Upload

Add Directory

Refresh

Rename

Delete

Change tier

Acquire lease

Break lease

Give feedback

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

Location: esaenzazdl101 / silver

Search blobs by prefix (case-sensitive)

Show deleted objects

Name	Modified	Access tier	Archive status	Blob type	Size
<input type="checkbox"/> [-]					
<input type="checkbox"/> data1.csv	8/27/2023, 5:09:32 PM	Hot (Inferred)		Block blob	10.18 MiB

• Capa gold luego de la ejecución del pipeline:

Home > Storage accounts > esaenz7azsa101 | Containers >

esaenzazdl101

Container

Search

«

Upload

Add Directory

Refresh

Rename

Delete

Change tier

Acquire lease

Break lease

Give feedback

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: esaenzazdl101 / gold

Search blobs by prefix (case-sensitive)

Show deleted objects

Name	Modified	Access tier	Archive status	Blob type	Size
<input type="checkbox"/> [-]					
<input type="checkbox"/> data1.csv	8/27/2023, 5:00:16 PM	Hot (Inferred)		Block blob	2.4 MiB

Azude Data Factory

- Configuración GIT
  - Main branch (Data Factory):

main branch

Validate all

Save all

Publish

Auto Save

General

Factory settings

Connections

Linked services

Integration runtimes

Microsoft Purview

Source control

Git configuration

ARM template

Author

Triggers

Global parameters

Data flow libraries

Security

Credentials

Customer managed key

Outbound rules

Managed private endpoints

Workflow orchestration manager

Apache Airflow

Git repository

Git repository information associated with your data factory. [CI/CD best practices](#)

Edit

Overwrite live mode

Disconnect

Import resources

Repository type	Git-Hub
GitHub account	esaenz7
Repository name	RG_PROYECTO_FINAL_DE_AZURE_ESTEBAN_SAEENZ
Collaboration branch	main
Publish branch	main
Root folder	/
Last published commit	c0008f66f24a586d98ad334299fd9ffa5bf777ee
Publish (from ADF Studio)	Enabled
Custom comment	Disabled

- Synapse branch (Synapse):

synapse branch

Validate all

Commit all

Publish

«

Analytics pools

SQL pools

Apache Spark pools

Data Explorer pools (preview)

External connections

Linked services

Microsoft Purview

Integration

Triggers

Integration runtimes

Security

Access control

Credentials

Managed private endpoints

Configurations + libraries

Workspace packages

Data flow libraries

Apache Spark configurations

Source control

Git configuration

Configure a repository

Connect your workspace with your Git repository just within few clicks. To learn more about best practices about CI/CD please view document here. [Learn more](#)

Edit

Overwrite live mode

Disconnect

Import resources

Repository type

GitHub

GitHub account

esaenz7

Repository name

RG\_PROYECTO\_FINAL\_DE\_AZURE\_ESTEBAN\_SAEENZ

Collaboration branch

synapse

Publish branch

synapse

Root folder

/

Last published commit

60d1ccfd782698e0c2670b92fc1aa2e7f05c4f1b

Custom comment

Enabled

• Linked Services:

main branch

Validate all

Save all

Publish

Auto Save

«

General

Factory settings

Connections

Linked services

Integration runtimes

Microsoft Purview

Source control

Git configuration

ARM template

Author

Triggers

Global parameters

Data flow libraries

Security

Credentials

Customer managed key

Outbound rules

Managed private endpoints

Workflow orchestration manager

Apache Airflow

Linked services

Linked service defines the connection information to a data store or compute. [Learn more](#)

+ New

Filter by name

Annotations : Any

Showing 1 - 4 of 4 items

Name	Type	Related
AzureDataLakeStorage1	Azure Data Lake Storage Gen2	3
esaenz7azfs101	File system	3
esaenz7azkv101	Azure Key Vault	0
esaenz7azsyaf101	Azure Synapse Analytics (Artifacts)	1

• Integration runtime:

main branch

Validate all

Save all

Publish

Auto Save

General

Factory settings

Connections

Linked services

Integration runtimes

Microsoft Purview

Source control

Git configuration

ARM template

Author

Triggers

Global parameters

Data flow libraries

Security

Credentials

Customer managed key

Outbound rules

Managed private endpoints

Workflow orchestration manager

Apache Airflow

Integration runtimes

The integration runtime (IR) is the compute infrastructure to provide the following data integration capabilities across different network environment. [Learn more](#)

+ New

Refresh

Filter by name

Showing 1 - 2 of 2 items

Name	Type	Sub-type	Status	Related	Region
AutoResolveIntegrationRuntime	Azure	Public	Running	0	Auto Resolve
esaenz7azir101	Self-Hosted	---	Running	1	---

Key vaults:

Home > Resource groups > RG\_PROYECTO\_FINAL\_DE\_AZURE\_ESTEBAN\_SAENZ > esenz7azkv101

esenz7azkv101 | Secrets

Key vault

Search

+ Generate/Import

Refresh

Restore Backup

View sample code

Manage deleted secrets

Overview

Activity log

Access control (IAM)

Tags

Diagnose and solve problems

Access policies

Events

Objects

Keys

Secrets

Certificates

Settings

Access configuration

Networking

Microsoft Defender for Cloud

Properties

Locks

Monitoring

Name	Type	Status	Expiration date
esaenz7azfs101esaen		Enabled	

Data sets:

»

main branch

Validate all

Save all

Publish

Auto Save

Factory Resources

Filter resources by name

Pipelines

esaenz7azpl101

Change Data Capture (preview)

Datasets

esaenz7azdl101

esaenz7azds101

esaenz7azds102

esaenz7azds103

esaenz7azds201

esaenz7azds301

Data flows

esaenz7azdf101

Power Query

Templates

esaenz7azpl101

esaenz7azdl101

esaenz7azds101

esaenz7azds102

esaenz7azds103

esaenz7azds201

esaenz7azds301

Saved

DelimitedText

esaenz7azdf101

CSV

Connection

Schema

Parameters

Linked service

AzureDataLakeStorage1

Test connection

Edit

New

Learn more

File path

@dataset().filesystem

@dataset().path

@dataset().filename

Browse

Compression type

Select...

Column delimiter

Comma (,)

Row delimiter

Default (\r\n, or \r\n)

Encoding

Default(UTF-8)

Quote character

Double quote (")

Escape character

Backslash (\)

First row as header

✓

Null value

main branch

Validate all

Save all

Publish

Auto Save

Factory Resources

Filter resources by name

Pipelines

esaenz7azpl101

Change Data Capture (preview)

Datasets

esaenz7azdl101

esaenz7azds101

esaenz7azds102

esaenz7azds103

esaenz7azds201

esaenz7azds301

Data flows

esaenz7azdf101

Power Query

Templates

esaenz7azpl101

esaenz7azdl101

esaenz7azds101

esaenz7azds102

esaenz7azds103

esaenz7azds201

esaenz7azds301

Saved

✓ Validate

Data flow debug

Debug Settings

source1

Import data from esaenz7azds101

source2

Import data from esaenz7azds101

join1

Inner join on 'source2' and 'source1'

source3

Import data from esaenz7azds101

join2

Inner join on 'join1' and 'source3'

sink101

Export data to esaenz7azds101

Add Source

Parameters

Settings

+ New

Delete

Name	Type	Default value
column101	abc string	'Code1'
column102	abc string	'DEST'
column201	abc string	'Code2'
column202	abc string	'AirportCode'

Notebook de transformación y preparación en Synapse:

Synapse live Validate all Publish all

Develop

Filter resources by name

Notebooks

esaenz7azipymb101

Run all Undo Publish Outline Attach to esaenz7azsp101 Language PySpark (Python) Variables

Ready

```
1 #librerías
2 from IPython.display import Javascript
3 import sys, os, glob, datetime as dt, numpy as np, random, collections as coll
4 import pandas as pd, seaborn as sns, matplotlib.pyplot as plt
5 from sklearn.metrics import roc_curve, auc, classification_report, confusion_matrix
6 from pyspark.sql import SparkSession, functions as F, window as W, DataFrame as DF
7 from pyspark.sql.types import (DateType, IntegerType, FloatType, DoubleType, LongType, StringType, StructField, StructType, TimestampType)
8 from pyspark.ml import functions as mlF, Pipeline as pipe
9 from pyspark.ml.stat import Correlation
10 from pyspark.ml.linalg import Vectors
11 from pyspark.ml.feature import Imputer, StandardScaler, MinMaxScaler, Normalizer, PCA, StringIndexer, OneHotEncoder, VectorAssembler
12 from pyspark.ml.regression import LinearRegression
13 from pyspark.ml.classification import LogisticRegression, DecisionTreeClassifier, DecisionTreeClassificationModel, RandomForestClassifier, GBTClassifier
14 from pyspark.ml.evaluation import BinaryClassificationEvaluator
15 from pyspark.mllib.evaluation import BinaryClassificationMetrics, MulticlassMetrics
16 from pyspark.ml.tuning import CrossValidator, CrossValidatorModel, ParamGridBuilder
17 from functools import reduce
18 from difflib import SequenceMatcher as seqmatch
19 # import findspark
20 # findspark.init('/usr/lib/python3.7/site-packages/pyspark')
21 # !pip install -q handyspark
22 # from handyspark import *
23
24 #variables postgres
25 # args = sys.argv
26 # print(args)
27 #estos parámetros corresponden a la instancia de postgres dentro del ambiente de docker que se adjunta al trabajo
28 host = '10.7.84.102'
29 port = '5432'
30 user = 'postgres'
31 password = 'testPassword'
32
33 #sesión de spark
34 spark = SparkSession.builder\
35     .master("local")\
36     .appName("Main")\
37     .config("spark.ui.port", '4050')\
38     .config("spark.driver.extraClassPath", "postgresql-42.2.14.jar") \
39     .config("spark.executor.extraClassPath", "postgresql-42.2.14.jar") \
40     .config("spark.jars", "postgresql-42.2.14.jar") \
41     .getOrCreate()
42 spark.sparkContext.setLogLevel("ERROR")
43
44 #funciones
45 #función para almacenar en base de datos
46 def escribir_df(df, host=host, port=port, user=user, password=password, table='table'):
47     try:
48         #almacenamiento en base de datos
49         # .option("driver", "postgresql-42.2.14.jar") \
50         df \
51             .write \
```

Run

df\_1 = crear\_df(paths=['abfss://esaenzadl101@esaenz7azsa101.dfs.core.windows.net/silver/data1.csv'],  
formats=['csv'], headers=[True], samples\_fr=[1.], rand\_st=999, print=True)

[86] ✓ 50 sec - Command executed in 50 sec 200 ms by esaenz7az01 on 5:59:59 PM, 8/27/23

> Job execution Succeeded Spark 1 executors 4 cores

View in monitoring

...

Dataframe 1 ( abfss://esaenzadl101@esaenz7azsa101.dfs.core.windows.net/silver/data1.csv )

Id	Airport	City	Country	Code1 Code2	LocationLat	LocationLng	c1	c2	c3	c4	c5	c6	FL_DATE
OP_CARRIER OP_CARRIER_FL_NUM	ORIGIN DEST	CRS_DEP_TIME DEP_TIME DEP_DELAY	TAXI_OUT WHEELS_OFF WHEELS_ON	TAXI_IN CRS_ARR_TIME ARR_TIME ARR_DELAY	CANCELLED	CANCELLATION_CODE	DIVERTED	CRS_EL					
ME ACTUAL_ELAPSED_TIME AIR_TIME DISTANCE	CARRIER_DELAY WEATHER_DELAY NAS_DELAY SECURITY_DELAY	LATE_AIRCRAFT_DELAY	Unnamed: 27	EventId	Type	Severity	StartTime(UTC)	EndTime(UTC)					
TimeZone	AirportCode	County	State	ZipCode									
3876	Charlotte Douglas International Airport	Charlotte	United States	CLT  KCLT	35.2140007019043 -80.94309997558594	748 -5.0	A  America/New_York airport OurAirports	2018-01-01 00:00:00					
2158	ORD  CLT	2101	2127.0  26.0  16.0	2143.0  2359.0	4.0  3	3.0  0.0	0.0  null	0.0  122.0					
76.0	599.0	null	null	null	null	W-5505728	Rain	Light	2018-01-11 07:52:00	2018-01-11 08:52:00	US/Eastern	KCLT	
Mecklenburg	NC	28278											
3876	Charlotte Douglas International Airport	Charlotte	United States	CLT  KCLT	35.2140007019043 -80.94309997558594	748 -5.0	A  America/New_York airport OurAirports	2018-01-01 00:00:00					
2158	ORD  CLT	2101	2127.0  26.0  16.0	2143.0  2359.0	4.0  3	3.0  0.0	0.0  null	0.0  122.0					
76.0	599.0	null	null	null	null	W-5505733	Rain	Light	2018-01-12 00:42:00	2018-01-12 03:52:00	US/Eastern	KCLT	
Mecklenburg	NC	28278											
3876	Charlotte Douglas International Airport	Charlotte	United States	CLT  KCLT	35.2140007019043 -80.94309997558594	748 -5.0	A  America/New_York airport OurAirports	2018-01-01 00:00:00					
2158	ORD  CLT	2101	2127.0  26.0  16.0	2143.0  2359.0	4.0  3	3.0  0.0	0.0  null	0.0  122.0					
76.0	599.0	null	null	null	null	W-5505737	Rain	Light	2018-01-12 17:17:00	2018-01-12 17:52:00	US/Eastern	KCLT	
Mecklenburg	NC	28278											
3876	Charlotte Douglas International Airport	Charlotte	United States	CLT  KCLT	35.2140007019043 -80.94309997558594	748 -5.0	A  America/New_York airport OurAirports	2018-01-01 00:00:00					
2158	ORD  CLT	2101	2127.0  26.0  16.0	2143.0  2359.0	4.0  3	3.0  0.0	0.0  null	0.0  122.0					

Synapse liveValidate allPublish all

Develop

Filter resources by name

Notebooks1

esaenz7azipyb101

esaenz7azipyb101

Run allUndoPublishOutlineAttach toesaenz7azsp101LanguagePySpark (Python)Variables

Ready

```
14      F.col('DEST').alias('dest'),
15      F.col('OP_CARRIER').alias('carrier'),
16      F.col('CRS_DEP_TIME').cast('int').alias('sdeptim'),
17      F.col('DEP_TIME').cast('int').alias('deptim'),
18      F.col('DEP_DELAY').cast('int').alias('depdel'),
19      F.col('TAXI_OUT').cast('int').alias('txout'),
20      F.col('WHEELS_OFF').cast('int').alias('woffftim'),
21      F.col('WHEELS_ON').cast('int').alias('wontim'),
22      F.col('TAXI_IN').cast('int').alias('txin'),
23      F.col('ARR_DELAY').cast('int').alias('arrdel'),
24      F.col('ARR_TIME').cast('int').alias('arrtim'),
25      F.col('CRS_ARR_TIME').cast('int').alias('sarrtim'),
26      F.col('CRS_ELAPSED_TIME').cast('int').alias('selap'),
27      F.col('ACTUAL_ELAPSED_TIME').cast('int').alias('aelap'),
28      F.col('AIR_TIME').cast('int').alias('airtim'),
29      F.col('DISTANCE').cast('int').alias('dist'))\
30      .withColumn('daywk', F.dayofweek(F.col('date2')).cast('int'))\
31      .withColumn('wkday', F.when(F.col('daywk')<5,1).otherwise(0))\
32      .withColumn('month', F.month(F.col('date2')).cast('int'))\
33      .withColumn('sdephr', F.expr('substring(sdeptim, 1, length(sdeptim)-2)').cast('int'))\
34      .withColumn('sarhrh', F.expr('substring(sarrtim, 1, length(sarrtim)-2)').cast('int'))\
35      .withColumn('morning', F.when(F.col('sdephr')<12,1).otherwise(0))\
36      .withColumn('label', F.when(F.col('arrdel')>0,1).otherwise(0))\
37      .withColumn('carrier_cnt', F.count('carrier').over(W.Window.partitionBy('carrier')))\
38      .withColumn('carrier_rnk', F.dense_rank().over(W.Window.orderBy(F.desc('carrier_cnt'))))\
39      .withColumn('carrier', F.when(F.col('carrier_rnk')>9,'00').otherwise(F.col('carrier')))\
40      .drop('daywk','carrier_cnt','carrier_rnk','sdephr','sdeptim','deptim','woffftim','wontim','txin','arrdel','arrtim','sarrtim','sarhrh','aelap','airtim')\
41
```

[72] ✓ <1 sec - Command executed in 553 ms by esaenz7az01 on 4:57:21 PM, 8/27/23

1 df\_3 = df\_2.toPandas()

[75] ✓ 4 sec - Command executed in 4 sec 91 ms by esaenz7az01 on 4:59:29 PM, 8/27/23

> Job execution Succeeded Spark 1 executors 4 cores

1 df\_3.to\_csv('abfss://esaenzazd1101@esaenz7azsa101.dfs.core.windows.net/gold/data1.csv')

[77] ✓ 1 sec - Command executed in 1 sec 169 ms by esaenz7az01 on 5:00:16 PM, 8/27/23

...

+ Code + Markdown

Apache Spark applications > Livy ID 3

esaenz7azipyb101\_esaenz7azsp101\_1693180637

Completed tasks 14 of 14 Status Running Total duration 10m 34s

Cancel Refresh Spark UI

Attempts 0 of 0

All job IDs View Progress Playback 0 ms / 51 sec 356 ms

Job 0

Tasks: 1

Duration: 10 sec 35 ms

Rows: 1

Data read: 64.0 KB

Data written: 0 bytes

1 Stage

Job 1

Tasks: 3

Duration: 4 sec 255 ms

Rows: 29,100

Data read: 10.3 MB

Data written: 0 bytes

1 Stage

Job 2

Tasks: 1

Duration: 843 ms

Rows: 11

Data read: 64.0 KB

Data written: 0 bytes

1 Stage

Job 3

Tasks: 3

Duration: 21 sec 218 ms

Rows: 29,102

Data read: 10.3 MB

Data written: 4.6 KB

1 Stage

Job 4

Tasks: 4

Duration: 2 sec 105 ms

Rows: 3

Data read: 4.6 KB

Data written: 0 bytes

2 Stages

Job 5

Tasks: 3

Duration: 1 sec 410 ms

Rows: 58,198

Data read: 10.3 MB

Data written: 198.6 KB

1 Stage

Job 6

Tasks: 4

Duration: 1 sec 70 ms

Rows: 58,198

Data read: 198.6 KB

Data written: 198.4 KB

2 Stages

Diagnostics Logs Input data Output data

Copy input Export as CSV Filter by name, read

Showing 1 - 1 of 1 items

Name ↑↓	Read format ↑↓	Size	Path ↑↓
data1.csv	csv	10.18 MB	abfss://esaenzazd1101@esaenz7azsa101

• Pipeline:



»

main branch

Validate all

Save all

Publish

Auto Save

Factory Resources

Filter resources by name

Pipelines

esaenz7azpl101

Change Data Capture (preview)

Datasets

esaenz7azd101

esaenz7azds101

esaenz7azds102

esaenz7azds103

esaenz7azds201

esaenz7azds301

Data flows

esaenz7azdf101

Power Query

Templates

Activities

Search activities

Move and transform

Synapse

Azure Data Explorer

Azure Function

Batch Service

Databricks

Data Lake Analytics

General

HDInsight

Iteration & conditionals

Machine Learning

Power Query

Save

Save as template

Validate

Cancel options

Add trigger

Data flow debug

Copy data

esaenz7azcd101

Copy data

esaenz7azcd102

Copy data

esaenz7azcd103

Get Metadata

esaenz7azgm101

Data flow

esaenz7azdf101

Notebook

esaenz7azsynb101

Get Metadata

esaenz7azgm102

Get Metadata

esaenz7azgm103

Parameters

Variables

Settings

Output

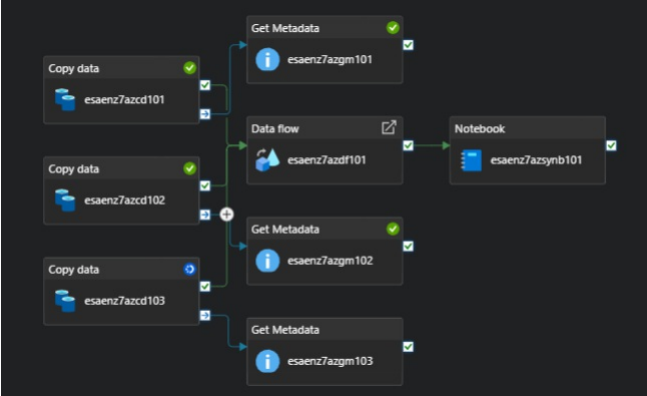
Pipeline run ID: c131966a-9286-498f-9de1-2a12a3bb83eb

Pipeline status: In progress

All status

Showing 1 - 8 of 8 items

Activity name	Activity status	Activity type	Run start	Duration	Log	Integrations
esaenz7azsynb101	In progress	Notebook	8/27/2023, 5:10:33 PM	10m 1s		
esaenz7azdf101	Succeeded	Data flow	8/27/2023, 5:08:37 PM	1m 54s		debug
esaenz7azgm103	Succeeded	Get Metadata	8/27/2023, 5:08:37 PM	3s		AutoR
esaenz7azgm102	Succeeded	Get Metadata	8/27/2023, 5:04:00 PM	3s		AutoR
esaenz7azgm101	Succeeded	Get Metadata	8/27/2023, 5:03:57 PM	3s		AutoR
esaenz7azcd102	Succeeded	Copy data	8/27/2023, 5:03:41 PM	18s		esaen
esaenz7azcd101	Succeeded	Copy data	8/27/2023, 5:03:40 PM	16s		esaen
esaenz7azcd103	Succeeded	Copy data	8/27/2023, 5:03:40 PM	4m 55s		esaen



Edit trigger

Name \*

esaenz7azdfst101

Description

Type \*

ScheduleTrigger

Start date \*

8/28/2023, 12:00:00 AM

Time zone \*

Coordinated Universal Time (UTC)

Recurrence \*

Every 7 Day(s)

Advanced recurrence options

Execute at these times

Hours

Minutes

Schedule execution times

00:00

Specify an end date

End On \*

12/31/2023, 12:00:00 AM

Annotations

+ New

Status

Started

Stopped

OK

Cancel

- Trigger:
- Vista de las fuentes de datos iniciales (locales):

2018.csv	
1	FL DATE,OP CARRIER,OP CARRIER_FL_NUM,ORIGIN,DEST,CRS DEP TIME,DEP TIME
2	2018-01-01,UA,2429,EWR,DEN,1517,1512.0,-5.0,15.0,1527.0,1712.0,10.0,17
3	2018-01-01,UA,2427,LAS,SFO,1115,1107.0,-8.0,11.0,1118.0,1223.0,7.0,125
4	2018-01-01,UA,2426,SNA,DEN,1335,1330.0,-5.0,15.0,1345.0,1631.0,5.0,164
5	2018-01-01,UA,2425,RSW,ORD,1546,1552.0,6.0,19.0,1611.0,1748.0,6.0,1756
6	2018-01-01,UA,2424,ORD,ALB,630,650.0,20.0,13.0,703.0,926.0,10.0,922,93
7	2018-01-01,UA,2422,ORD,OMA,2241,2244.0,3.0,15.0,2259.0,1.0,2.0,14,3.0,
8	2018-01-01,UA,2421,IAH,LAS,750,747.0,-3.0,14.0,801.0,854.0,6.0,916,900
9	2018-01-01,UA,2420,DEN,CID,1324,1318.0,-6.0,11.0,1329.0,1554.0,6.0,161
10	2018-01-01,UA,2419,SMF,EWR,2224,2237.0,13.0,10.0,2247.0,627.0,9.0,638,
11	2018-01-01,UA,2418,RTC,DEN,1601,1559.0,-2.0,12.0,1611.0,1748.0,8.0,181
12	2018-01-01,UA,2417,PDX,EWR,2240,2235.0,-5.0,9.0,2244.0,624.0,7.0,647,6
13	2018-01-01,UA,2416,ORD,CLE,2059,2300.0,121.0,24.0,2324.0,112.0,8.0,231
14	2018-01-01,UA,2415,EWR,PDX,825,822.0,-3.0,15.0,837.0,1104.0,5.0,1135,1
15	2018-01-01,UA,2414,EWR,ATL,1044,1055.0,11.0,11.0,1106.0,1310.0,5.0,131
16	2018-01-01,UA,2413,ORD,BTV,2114,2230.0,76.0,14.0,2244.0,123.0,5.0,15,1
17	2018-01-01,UA,2412,MCO,LAX,653,747.0,54.0,14.0,801.0,1003.0,22.0,930,1
18	2018-01-01,UA,2411,EWR,SMF,1810,1922.0,72.0,16.0,1938.0,2157.0,4.0,213
19	2018-01-01,UA,2410,RSW,EWR,1250,1337.0,47.0,12.0,1349.0,1600.0,6.0,153
20	2018-01-01,UA,2409,IAH,JAC,940,934.0,-6.0,18.0,952.0,1156.0,4.0,1218,1
21	2018-01-01,UA,2408,TYS,EWR,1131,1140.0,9.0,9.0,1149.0,1307.0,5.0,1333,
22	2018-01-01,UA,2406,EWR,TYS,830,844.0,14.0,20.0,904.0,1052.0,3.0,1049,1
23	2018-01-01,UA,2405,SFO,IAH,530,521.0,-9.0,13.0,534.0,1050.0,4.0,1122,1
24	2018-01-01,UA,2404,RSW,ORD,1220,1213.0,-7.0,25.0,1238.0,1419.0,3.0,143
25	2018-01-01,UA,2403,SFO,HNL,1927,1925.0,-2.0,15.0,1940.0,2243.0,7.0,232
26	2018-01-01,UA,2402,JAC,EWR,1343,1351.0,8.0,38.0,1429.0,2003.0,10.0,195
27	2018-01-01,UA,2400,BOS,SFO,1820,1826.0,6.0,24.0,1850.0,2122.0,15.0,214
28	2018-01-01,UA,2399,ORD,MIA,1405,1402.0,-3.0,18.0,1420.0,1757.0,9.0,181
29	2018-01-01,UA,2398,MSY,EWR,2043,2131.0,48.0,10.0,2141.0,54.0,7.0,34,10
30	2018-01-01,UA,2397,MIA,IAH,900,859.0,-1.0,17.0,916.0,1048.0,6.0,1056,1
31	2018-01-01,UA,2394,EWR,LAX,1500,1515.0,15.0,19.0,1534.0,1804.0,8.0,182
32	2018-01-01,UA,2393,SEA,IAH,27,15.0,-12.0,15.0,30.0,601.0,11.0,644,612,
33	2018-01-01,UA,2392,SAT,IAH,730,729.0,-1.0,13.0,742.0,816.0,4.0,834,820
34	2018-01-01,UA,2391,SLC,EWR,1318,1314.0,-4.0,10.0,1324.0,1905.0,7.0,193
35	2018-01-01,UA,2390,MCO,SFO,852,854.0,2.0,10.0,904.0,1142.0,4.0,1203,11
36	2018-01-01,UA,2389,MSY,IAD,1758,1751.0,-7.0,13.0,1804.0,2056.0,3.0,212
37	2018-01-01,UA,2388,DEN,ICT,950,948.0,-2.0,13.0,1001.0,1454.0,6.0,1516,
38	2018-01-01,UA,2386,ORD,RSW,1350,1348.0,-2.0,23.0,1411.0,1733.0,4.0,174
39	2018-01-01,UA,2386,RDU,ORD,955,951.0,-4.0,14.0,1005.0,1055.0,3.0,1115,
40	2018-01-01,UA,2385,DEN,ORD,915,909.0,-6.0,12.0,921.0,1216.0,19.0,1242,
41	2018-01-01,UA,2384,EWR,FLL,1627,1624.0,-3.0,38.0,1702.0,1938.0,6.0,193
42	2018-01-01,UA,2381,MIA,ORD,729,727.0,-2.0,14.0,741.0,925.0,11.0,950,93
43	2018-01-01,UA,2381,ORD,SAN,1215,1213.0,-2.0,18.0,1231.0,1421.0,1.0,144
44	2018-01-01,UA,2380,FLL,EWR,2025,2029.0,4.0,13.0,2042.0,2301.0,5.0,2325
45	2018-01-01,UA,2173,IAD,DEN,2217,2235.0,18.0,14.0,2249.0,3.0,5.0,14,8.0
46	2018-01-01,UA,2172,EWR,IAH,1959,2000.0,1.0,22.0,2022.0,2248.0,5.0,2308
47	2018-01-01,UA,2171,DFW,IAD,1710,1703.0,-7.0,18.0,1721.0,2034.0,7.0,205
48	2018-01-01,UA,2170,DEN,SLC,2220,2216.0,-4.0,24.0,2240.0,2345.0,3.0,235
49	2018-01-01,UA,2169,DEN,PDX,2220,2216.0,-4.0,11.0,2227.0,2356.0,5.0,12,
50	2018-01-01,UA,2168,DEN,ICT,2018,2017.0,-1.0,13.0,2030.0,2221.0,4.0,224
51	2018-01-01,UA,2167,DEN,EWR,2025,2022.0,-3.0,11.0,2033.0,135.0,6.0,202,
52	2018-01-01,UA,2166,ANC,DEN,2340,2336.0,-4.0,13.0,2349.0,623.0,6.0,658,
53	2018-01-01,UA,2165,MSP,ORD,500,502.0,2.0,13.0,515.0,611.0,12.0,634,623
54	2018-01-01,UA,2164,ALB,ORD,1004,1022.0,18.0,8.0,1030.0,1116.0,8.0,1134
55	2018-01-01,UA,2163,SFO,RNO,1048,1044.0,-4.0,22.0,1106.0,1138.0,5.0,115
56	2018-01-01,UA,2162,SFO,EUG,2013,2003.0,-10.0,19.0,2022.0,2132.0,9.0,21
57	2018-01-01,UA,2161,ORD,MCO,2113,2123.0,10.0,18.0,2141.0,47.0,6.0,57,53
58	2018-01-01,UA,2160,ORD,MCT,2030,2111.0,41.0,21.0,2132.0,2230.0,5.0,220


- Vista del dataset final (gold):

airports.csv		weather.csv	
1	Id,Airport,City,Country,Codel,Code2,LocationLat,LocationLng,c1,c2,c3,c	1	EventI
2	1,"Goroka Airport","Goroka","Papua New Guinea","GKA","AYGA",-6.8816898	2	W-967,
3	2,"Madang Airport","Madang","Papua New Guinea","MAG","AYMD",-5.2070798	3	W-969,
4	3,"Mount Hagen Kagamuga Airport","Mount Hagen","Papua New Guinea","HGU	4	W-973,
5	4,"Nadzab Airport","Nadzab","Papua New Guinea","LAE","AYNZ",-6.569803,	5	W-974,
6	5,"Port Moresby Jacksons International Airport","Port Moresby","Papua	6	W-975,
7	6,"Wewak International Airport","Wewak","Papua New Guinea","WWK","AYWK	7	W-976,
8	7,"Narsarsuaq Airport","Narsarsuaq","Greenland","UAK","BGBW",61.1604	8	W-978,
9	8,"Godthaab / Nuuk Airport","Godthaab","Greenland","GOH","BGGH",64.190	9	W-979,
10	9,"Kangerlussuaq Airport","Sondrestrom","Greenland","SFG","BGSF",67.01	10	W-980,
11	10,"Thule Air Base","Thule","Greenland","THU","BGTL",76.5311965942,-68	11	W-982,
12	11,"Akureyri Airport","Akureyri","Iceland","AEY","BIAR",65.66008366210	12	W-983,
13	12,"Egilsstaðir Airport","Egilsstaðir","Iceland","EGS","BIEG",65.28330	13	W-985,
14	13,"Hornafjörður Airport","Hofn","Iceland","HFN","BIHN",64.295601,-15.	14	W-987,
15	14,"Husavik Airport","Husavik","Iceland","HZK","BIHU",65.952301,-17.42	15	W-990,
16	15,"Isafjörður Airport","Isafjörður","Iceland","IFJ","BISJ",66.0580978	16	W-991,
17	16,"Keflavik International Airport","Keflavik","Iceland","KEF","BIKF",	17	W-992,
18	17,"Patreksfjörður Airport","Patreksfjörður","Iceland","PFJ","BIPA",65	18	W-993,
19	18,"Reykjavík Airport","Reykjavík","Iceland","RKV","BIRK",64.129997253	19	W-995,
20	19,"Siglufjörður Airport","Siglufjörður","Iceland","SIJ","BISI",66.133	20	W-997,
21	20,"Vestmannaeyjar Airport","Vestmannaeyjar","Iceland","VEY","BIVM",63	21	W-1000
22	21,"Sault Ste Marie Airport","Sault Sainte Marie","Canada","YAM","CYAM	22	W-1001
23	22,"Winnipeg / St. Andrews Airport","Winnipeg","Canada","\N","CYAV",50.0	23	W-1004
24	23,"Halifax / CFB Shearwater Heliport","Halifax","Canada","\N","CYAW",44	24	W-1005
25	24,"St. Anthony Airport","St. Anthony","Canada","YAY","CYAY",51.391899	25	W-1012
26	25,"Tofino / Long Beach Airport","Tofino","Canada","YAZ","CYAZ",49.079	26	W-1014
27	26,"Kugaaruk Airport","Pelly Bay","Canada","YBB","CYBB",68.534401,-89.	27	W-1015
28	27,"Baie Comeau Airport","Baie Comeau","Canada","YBC","CYBC",49.132499	28	W-1017
29	28,"CFB Bagotville","Bagotville","Canada","YBG","CYBG",48.330600738525	29	W-1018
30	29,"Baker Lake Airport","Baker Lake","Canada","YBK","CYBK",64.29889678	30	W-1020
31	30,"Campbell River Airport","Campbell River","Canada","YBL","CYBL",49.	31	W-1025
32	31,"Brandon Municipal Airport","Brandon","Canada","YBR","CYBR",49.91,-	32	W-1026
33	32,"Cambridge Bay Airport","Cambridge Bay","Canada","YCB","CYCB",69.10	33	W-1027
34	33,"Nanaimo Airport","Nanaimo","Canada","YCD","CYCD",49.05497022489999	34	W-1028
35	34,"Castlegar/West Kootenay Regional Airport","Castlegar","Canada","YC	35	W-1036
36	35,"Miramichi Airport","Chatham","Canada","YCH","CYCH",47.007801,-65.4	36	W-1039
37	36,"Charlo Airport","Charlo","Canada","YCL","CYCL",47.990799,-66.33029	37	W-1041
38	37,"Kugluktuk Airport","Coppermine","Canada","YCO","CYCO",67.816704,-1	38	W-1042
39	38,"Coronation Airport","Coronation","Canada","YCT","CYCT",52.07500076	39	W-1045
40	39,"Chilliwack Airport","Chilliwack","Canada","YCW","CYCW",49.15280151	40	W-1046
41	40,"Clyde River Airport","Clyde River","Canada","YCY","CYCY",70.486099	41	W-1048
42	41,"Coral Harbour Airport","Coral Harbour","Canada","YZS","CYZS",64.19	42	W-1049
43	42,"Dawson City Airport","Dawson","Canada","YDA","CYDA",64.04309844970	43	W-1050
44	43,"Burwash Airport","Burwash","Canada","YDB","CYDB",61.37110137939453	44	W-1052
45	44,"Princeton Airport","Princeton","Canada","\N","CYDC",49.4681015015,-1	45	W-1053
46	45,"Deer Lake Airport","Deer Lake","Canada","YDF","CYDF",49.2108001708	46	W-1054
47	46,"Dease Lake Airport","Dease Lake","Canada","YDL","CYDL",50.42210924	47	W-1055
48	47,"Dauphin Barker Airport","Dauphin","Canada","YDN","CYDN",51.1007995	48	W-1059
49	48,"Dawson Creek Airport","Dawson Creek","Canada","YDQ","CYDQ",55.7422	49	W-1060
50	49,"Edmonton International Airport","Edmonton","Canada","YEG","CYEG",5	50	W-1064
51	50,"Arviat Airport","Eskimo Point","Canada","YEK","CYEK",61.0942001343	51	W-1065
52	51,"Estevan Airport","Estevan","Canada","YEN","CYEN",49.21030004456,-10	52	W-1066
53	52,"Edson Airport","Edson","Canada","YET","CYET",53.5788993834999904,-1	53	W-1068
54	53,"Eureka Airport","Eureka","Canada","YEU","CYEU",79.9946975708,-85.8	54	W-1070
55	54,"Inuvik Mike Zubko Airport","Inuvik","Canada","YEV","CYEV",68.30419	55	W-1073
56	55,"Iqaluit Airport","Iqaluit","Canada","YFB","CYFB",63.756402,-68.555	56	W-1075
57	56,"Fredericton Airport","Fredericton","Canada","YFC","CYFC",45.868900	57	W-1076
58	57,"Forestville Airport","Forestville","Canada","YFF","CYFF",48.746101	58	W-1077

Preview data

Linked service: AzureDataLakeStorage1

Object:

	Prop_0	iata	icao2	date1	wtyp	wsev	icao6	date2	orig	dest	carrier	depdel	txout	selap	dist
1	0	CLT	KCLT	2018-01-11	Rain	Light	KCLT	2018-01-01	ORD	CLT	UA	26	16	122	599
2	1	CLT	KCLT	2018-01-12	Rain	Light	KCLT	2018-01-01	ORD	CLT	UA	26	16	122	599
3	2	CLT	KCLT	2018-01-12	Rain	Light	KCLT	2018-01-01	ORD	CLT	UA	26	16	122	599
4	3	CLT	KCLT	2018-01-12	Rain	Light	KCLT	2018-01-01	ORD	CLT	UA	26	16	122	599
5	4	CLT	KCLT	2018-01-12	Rain	Light	KCLT	2018-01-01	ORD	CLT	UA	26	16	122	599
6	5	CLT	KCLT	2018-01-17	Rain	Light	KCLT	2018-01-01	ORD	CLT	UA	26	16	122	599
7	6	CLT	KCLT	2018-01-17	Snow	Light	KCLT	2018-01-01	ORD	CLT	UA	26	16	122	599
8	7	CLT	KCLT	2018-01-17	Snow	Moderate	KCLT	2018-01-01	ORD	CLT	UA	26	16	122	599
9	8	CLT	KCLT	2018-01-17	Snow	Light	KCLT	2018-01-01	ORD	CLT	UA	26	16	122	599
10	9	CLT	KCLT	2018-01-17	Snow	Moderate	KCLT	2018-01-01	ORD	CLT	UA	26	16	122	599