

**Azure DataFactory Project**

**Airlines Delays and Cancellations 2015**

**By: Brahim Fakri**

**College Rosemont - Montréal**

## Contents

Description du problème à résoudre .....	2
Le choix des données massives.....	2
Présentation du système infonuagique Big data, de l'architecture générale et des outils utilisés et qui seront utilisés dans les prochaines étapes de la consultation.....	6
Justification des outils utilisés.....	6
Implémentation : Structure, Extraction, Ingestion et Pipelines .....	8
Définition des éléments projet final .....	8
Extraction et ingestion des données.....	10
Programmer des traitements automatiques (Datafactory) et transmettre les données résultantes vers un système distribué :.....	32
Visualisation des données.....	36
Partage dans GITHUB.....	37
Références .....	38

## Description du problème à résoudre

Le Département américain des transports (DOT) suit la ponctualité des vols intérieurs opérés par les grands transporteurs aériens. Des informations récapitulatives sur le nombre de vols à l'heure, retardés, annulés et détournés sont publiées dans le rapport mensuel Air Travel Consumer Report du DOT.

En tant qu'équipe de consultants pour le DOT, on veut répondre aux questions suivantes :

- Les raisons les plus fréquentes pour les délais de vols
- Les compagnies aériennes qui ont eu le plus de retards et d'annulations
- Les aéroports avec le plus de retards et d'annulations
- Les jours de semaines avec le plus de retards et d'annulations
- Les mois avec le plus de retards et d'annulations

Le client souhaite avoir un tableau de bord incluant des visuels qui simplifient la compréhension de l'analyse des données fournies.

## Le choix des données massives

Notre équipe a décidé d'utiliser le dataset Kaggle sur les retards et les annulations de vols en 2015. Ce dataset inclut le premier trimestre de 2015.

[2015 Flight Delays and Cancellations | Kaggle](#)

<https://www.kaggle.com/datasets/usdot/flight-delays>

The screenshot shows the Kaggle dataset page for '2015 Flight Delays and Cancellations'. The page has a sidebar on the left with links like 'Create', 'Home', 'Competitions', 'Datasets', 'Code', 'Discussions', 'Learn', 'More', 'Your Work', and 'RECENTLY VIEWED' (Flight Delay Predictions, 2015 Flight Delays and...). The main content area features a title '2015 Flight Delays and Cancellations' with a subtitle 'Which airline should you fly on to avoid significant delays?'. It includes a preview image showing flight information for Frankfurt, Budapest, and Paris. Below the title are tabs for 'Data', 'Code (173)', 'Discussion (16)', and 'Metadata'. The 'About Dataset' section contains 'Context' and 'Acknowledgements' with a note about the data source being the DOT's Bureau of Transportation Statistics. On the right, there are sections for 'Usability' (8.82), 'License' (CC0: Public Domain), and 'Expected update frequency' (Not specified).

## 2015 Flight Delays and Cancellations

Data Code (173) Discussion (16) Metadata

899

New Notebook

Download (200 MB)



**flights.csv** (592.41 MB)



Detail Compact Column

10 of 31 columns ▾

### About this file

IATA airline codes and names

# YEAR	# MONTH	# DAY	# DAY_OF_WEEK	▲ AIRLINE	# FLIGHT
Year of the Flight Trip	Month of the Flight Trip	Day of the Flight Trip	Day of week of the Flight Trip	Airline Identifier	Flight Id
2015	2015	1	1	WN	22%
		12	31	DL	15%
			7	Other (3681343)	63%
				1	1
2015	1	1	4	AS	98
2015	1	1	4	AA	2336
2015	1	1	4	US	840
2015	1	1	4	AA	258
2015	1	1	4	AS	135
2015	1	1	4	DL	886
2015	1	1	4	NK	612
2015	1	1	4	US	2813

### Data Explorer

Version 1 (592.43 MB)

airlines.csv

airports.csv

flights.csv

EVALUATION FINAL > data airlines



Search data airlines



airlines.csv



airports.csv



flights.csv

### Summary

3 files

.csv

3

40 columns

# Integer

26

▲ String

12

Dans le fichier **airlines**, le nom de la clé est le code IATA : **IATA\_CODE**

A	B
1	<b>IATA_CODE</b> <b>AIRLINE</b>
2	UA United Air Lines Inc.
3	AA American Airlines Inc.
4	US US Airways Inc.
5	F9 Frontier Airlines Inc.
6	B6 JetBlue Airways
7	OO Skywest Airlines Inc.
8	AS Alaska Airlines Inc.
9	NK Spirit Air Lines
10	WN Southwest Airlines Co.
11	DL Delta Air Lines Inc.
12	EV Atlantic Southeast Airlines
13	HA Hawaiian Airlines Inc.
14	MQ American Eagle Airlines Inc.
15	VX Virgin America
16	
17	
18	
19	

Aussi, dans le fichier **airports**, le nom de la clé est le code IATA : **IATA\_CODE**

A	B	CITY	STATE	COUNTRY	LATITUDE	LONGITUDE
1	<b>IATA_CODE</b> <b>AIRPORT</b>	Allentown	PA	USA	40.65236	-75.4404
2	ABE Lehigh Valley International Airport	Abilene	TX	USA	32.41132	-99.6819
3	ABI Abilene Regional Airport	Albuquerque	NM	USA	35.04022	-106.60919
4	ABQ Albuquerque International Sunport	Aberdeen	SD	USA	45.44906	-98.42183
5	ABR Aberdeen Regional Airport	Albany	GA	USA	31.53552	-84.19447
6	ABY Southwest Georgia Regional Airport	Nantucket	MA	USA	41.25305	-70.06018
7	ACK Nantucket Memorial Airport	Waco	TX	USA	31.61129	-97.23052
8	ACT Waco Regional Airport	Arcata/Eureka	CA	USA	40.97812	-124.10862
9	ACV Arcata Airport	Atlantic City	NJ	USA	39.45758	-74.57717
10	ACY Atlantic City International Airport	Adak	AK	USA	51.87796	-176.64603
11	ADK Adak Airport	Kodiak	AK	USA	57.74997	-152.49386
12	ADQ Kodiak Airport	Alexandria	LA	USA	31.32737	-92.54856
13	AEX Alexandria International Airport	Augusta	GA	USA	33.36996	-81.9645
14	AGS Augusta Regional Airport (Bush Field)	King Salmon	AK	USA	58.6768	-156.64922
15	AKN King Salmon Airport	Albany	NY	USA	42.74812	-73.80298
16	ALB Albany International Airport	Waterloo	IA	USA	42.55708	-92.40034
17	ALO Waterloo Regional Airport	Amarillo	TX	USA	35.21937	-101.70593
18	AMA Rick Husband Amarillo International Airport	Anchorage	AK	USA	61.17432	-149.99619
19	ANC Ted Stevens Anchorage International Airport	Alpena	MI	USA	45.07807	-83.56029
20	APN Alpena County Regional Airport	Aspen	CO	USA	39.22316	-106.86885
21	ASE Aspen-Pitkin County Airport	Atlanta	GA	USA	33.64044	-84.42694
22	ATL Hartsfield-Jackson Atlanta International Airport	Appleton	WI	USA	44.25741	-88.51948
23	ATW Appleton International Airport	Austin	TX	USA	30.19453	-97.66987
24	AUS Austin-Bergstrom International Airport					

Ces mêmes clés primaires apparaissent dans le fichier flights comme clés secondaires, mais avec des noms différents : **AIRLINE**, **ORIGIN\_AIRPORT**, **DESTINATION\_AIRPORT**

YEAR	MONTH	DAY	DAY_OF_WEEK	AIRLINE	FLIGHT_NUMBER	TAIL_NUMBER	ORIGIN_AIRPORT	DESTINATION_AIRPORT	SCHEDULED_DEPARTURE	DEPARTURE_TIME	DEPARTURE_DELAY	TAXI_OUT	WHEELS
2015	1	1	4	AS	98 N407AS	ANC	SEA		5	2354	-11	21	
2015	1	1	4	AA	2336 N3KUAA	LAX	PBI		10	2	-8	12	
2015	1	1	4	US	840 N171US	SFO	CLT		20	18	-2	16	
2015	1	1	4	AA	258 N3HYAA	LAX	MIA		20	15	-5	15	
2015	1	1	4	AS	135 N527AS	SEA	ANC		25	24	-1	11	
2015	1	1	4	DL	808 N3730B	SFO	MSP		25	20	-5	18	
2015	1	1	4	NK	612 N635NK	LAS	MSP		25	19	-6	11	
2015	1	1	4	US	2013 N584UW	LAX	CLT		30	44	14	13	
2015	1	1	4	AA	1112 N3LAAA	SFO	DFW		30	19	-11	17	
2015	1	1	4	DL	1173 N826DN	LAS	ATL		30	33	3	12	
2015	1	1	4	DL	2336 N958DN	DEN	ATL		30	24	-6	12	
2015	1	1	4	AA	1674 N853AA	LAS	MIA		35	27	-8	21	
2015	1	1	4	DL	1434 N547US	LAX	MSP		35	35	0	18	
2015	1	1	4	DL	2324 N3751B	SLC	ATL		40	34	-6	18	
2015	1	1	4	DL	2440 N651DL	SEA	MSP		40	39	-1	28	
2015	1	1	4	AS	108 N309AS	ANC	SEA		45	41	-4	17	
2015	1	1	4	DL	1560 N3743H	ANC	SEA		45	31	-14	25	
2015	1	1	4	UA	1197 N7844B	SFO	IAH		48	42	-6	11	
2015	1	1	4	AS	122 N413AS	ANC	PDX		50	46	-4	11	
2015	1	1	4	DL	1670 N806DN	PDX	MSP		50	45	-5	9	
2015	1	1	4	NK	520 N525NK	LAS	MCI		55	120	25	11	
2015	1	1	4	AA	371 N3GXAA	SEA	MIA		100	52	-8	30	
2015	1	1	4	NK	214 N632NK	LAS	DFW		103	102	-1	13	
2015	1	1	4	AA	115 N3CTAA	LAX	MIA		105	103	-2	14	
2015	1	1	4	DL	1450 N671DN	LAS	MSP		105	102	-3	11	

DATA_CD	AIRLINE
1	AA American Airlines Inc.
2	DL Delta Air Lines Inc.
3	US United Airlines Inc.
4	US Airways Inc.
5	US Airways Inc.
6	UA United Airlines Inc.
7	AA American Airlines Inc.
8	DL Delta Air Lines Inc.
9	US United Airlines Inc.
10	US Airways Inc.
11	UA United Airlines Inc.
12	AA American Airlines Inc.
13	DL Delta Air Lines Inc.
14	US United Airlines Inc.
15	US Airways Inc.
16	UA United Airlines Inc.
17	AA American Airlines Inc.
18	DL Delta Air Lines Inc.
19	US United Airlines Inc.
20	US Airways Inc.
21	UA United Airlines Inc.

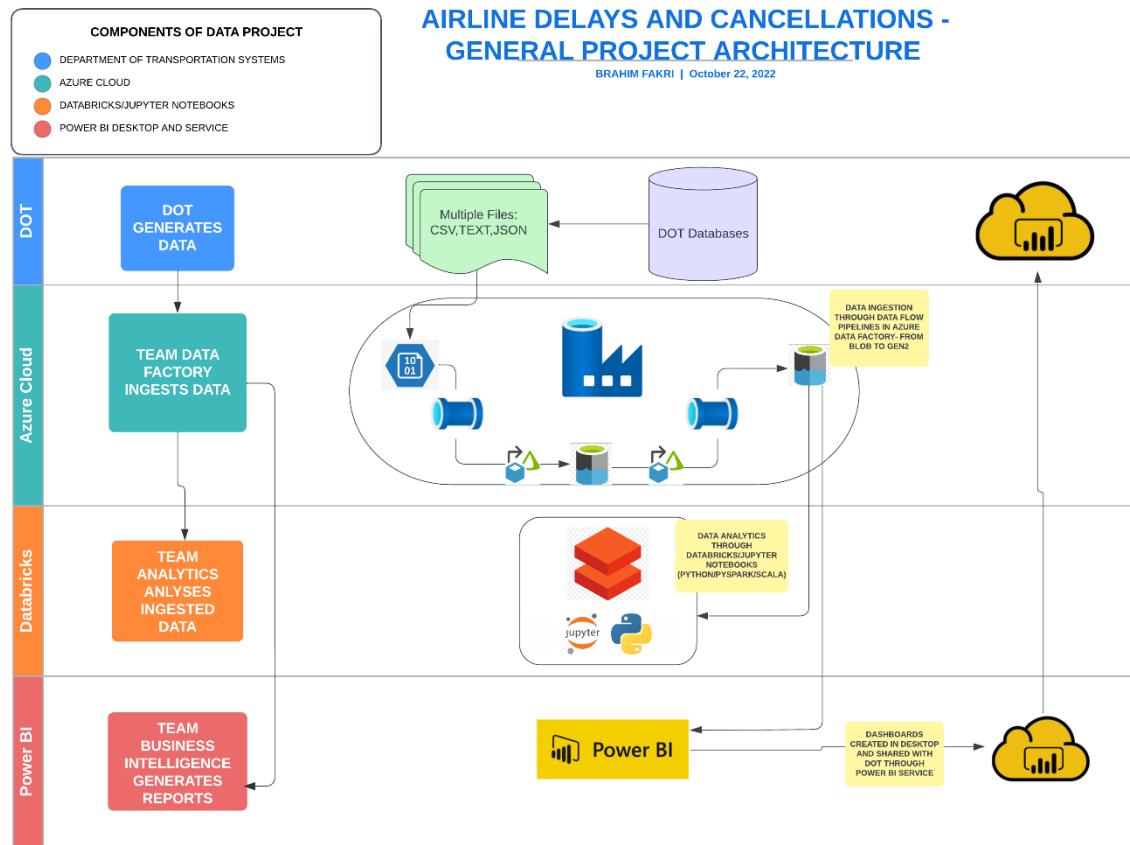
  

YEAR	MONTH	DAY	DAY_OF_WEEK	AIRLINE	FLIGHT_NUMBER	TAIL_NUMBER	ORIGIN_AIRPORT	DESTINATION_AIRPORT	SCHEDULED_DEPARTURE	DEPARTURE_TIME	DEPARTURE_DELAY	TAXI_OUT	WHEELS
2015	1	1	4	AS	98 N407AS	ANC	SEA		5	2354	-11	21	
2015	1	1	4	AA	2336 N3KUAA	LAX	PBI		10	2	-8	12	
2015	1	1	4	US	840 N171US	SFO	CLT		20	18	-2	16	
2015	1	1	4	AA	258 N3HYAA	LAX	MIA		20	15	-5	15	
2015	1	1	4	AS	135 N527AS	SEA	ANC		25	24	-1	11	
2015	1	1	4	DL	808 N3730B	SFO	MSP		25	20	-5	18	
2015	1	1	4	NK	612 N635NK	LAS	MSP		25	19	-6	11	
2015	1	1	4	US	2013 N584UW	LAX	CLT		30	44	14	13	
2015	1	1	4	AA	1112 N3LAAA	SFO	DFW		30	19	-11	17	
2015	1	1	4	DL	1173 N826DN	LAS	ATL		30	33	3	12	
2015	1	1	4	DL	2336 N958DN	DEN	ATL		30	24	-6	12	
2015	1	1	4	AA	1674 N853AA	LAS	MIA		35	27	-8	21	
2015	1	1	4	DL	1434 N547US	LAX	MSP		35	35	0	18	
2015	1	1	4	DL	2324 N3751B	SLC	ATL		40	34	-6	18	
2015	1	1	4	DL	2440 N651DL	SEA	MSP		40	39	-1	28	
2015	1	1	4	AS	108 N309AS	ANC	SEA		45	41	-4	17	
2015	1	1	4	DL	1560 N3743H	ANC	SEA		45	31	-14	25	
2015	1	1	4	UA	1197 N7844B	SFO	IAH		48	42	-6	11	
2015	1	1	4	AS	122 N413AS	ANC	PDX		50	46	-4	11	
2015	1	1	4	DL	1670 N806DN	PDX	MSP		50	45	-5	9	
2015	1	1	4	NK	520 N525NK	LAS	MCI		55	120	25	11	
2015	1	1	4	AA	371 N3GXAA	SEA	MIA		100	52	-8	30	
2015	1	1	4	NK	214 N632NK	LAS	DFW		103	102	-1	13	
2015	1	1	4	AA	115 N3CTAA	LAX	MIA		105	103	-2	14	
2015	1	1	4	DL	1450 N671DN	LAS	MSP		105	102	-3	11	

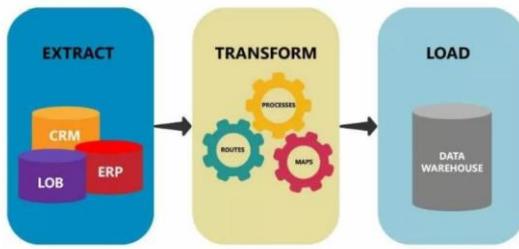
DATA_CD	AIRPORT	CITY	CTRY
1	Lehigh Valley International Airport	Allentown	USA
2	Albuquerque International Sunport	Albuquerque	USA
3	Southwest Georgia Regional Airport	Albany	USA
4	Hanover Regional Airport	Nashua	USA
5	Acurea Airport	Acurea	USA
6	Adak Airport	Adak	USA
7	Alexandria International Airport	Alexandria	USA
8	Alaska Anchorage International Airport	Anchorage	USA
9	Alaska Juneau International Airport	Juneau	USA
10	Alaska McGrath Airport	McGrath	USA
11	Alaska Sitka Airport	Sitka	USA
12	Alaska Unalaska International Airport	Unalaska	USA
13	Alaska Wrangell-St. Elias International Airport	Wrangell	USA
14	Austin-Bergstrom International Airport	Austin	USA
15	Baker City Municipal Airport	Baker City	USA
16	Barataria-Belle Chasse Regional Airport	King Salmon	USA
17	Bethel Airport	Bethel	USA
18	Benton County Regional Airport	McCall	USA
19	Bentonville-Benton County Regional Airport	Bentonville	USA
20	Bethel Airport	Bethel	USA
21	Big Bear Mountain Airport	Big Bear Lake	USA
22	Greater Bingham Airport	Bingham	USA
23	Binghamton-Syracuse International Airport	Binghamton	USA
24	Bismarck-Mandan Airport	Bismarck	USA
25	Blind Bay Airport	Blind Bay	USA
26	Boise Air Terminal	Boise	USA
27	Bonanza Field	Bonanza	USA
28	Brownsville-Harlingen International Airport	Brownsville	USA
29	Brownwood Municipal Airport	Brownwood	USA
30	Brownwood Regional Airport	Brownwood	USA
31	Brownwood Regional Airport at Broomfield-Normal	Broomfield	USA
32	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
33	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
34	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
35	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
36	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
37	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
38	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
39	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
40	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
41	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
42	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
43	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
44	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
45	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
46	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
47	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
48	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
49	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
50	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
51	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
52	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
53	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
54	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
55	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
56	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
57	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
58	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
59	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
60	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
61	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
62	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
63	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
64	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
65	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
66	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
67	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
68	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
69	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
70	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
71	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
72	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
73	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
74	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
75	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
76	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
77	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
78	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
79	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
80	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
81	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
82	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
83	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
84	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
85	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
86	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
87	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
88	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
89	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
90	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
91	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
92	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
93	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
94	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
95	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
96	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
97	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
98	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
99	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
100	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
101	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
102	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
103	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
104	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
105	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
106	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
107	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
108	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
109	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
110	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
111	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
112	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
113	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
114	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
115	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
116	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
117	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
118	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
119	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
120	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
121	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
122	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
123	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
124	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
125	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
126	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
127	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
128	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
129	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
130	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
131	Brownwood Regional Airport at Broomfield-Normal	Brownfield	USA
132			

Présentation du système infonuagique Big data, de l'architecture générale et des outils utilisés et qui seront utilisés dans les prochaines étapes de la consultation



### Justification des outils utilisés

Dans le monde des grandes organisations (entreprises, gouvernements, grandes ONG, etc.), les données peuvent être extraites des CRM, des ERP, des sites web, des fichiers CSV, des logiciels utilisés et de différentes sources. Ensuite, on fait une transformation pour préparer les données à être stockées dans un entrepôt de données par exemple. Une fois qu'elles sont stockées dans les entrepôts de données, on peut alors faire des analyses via OLAP ou autres outils.

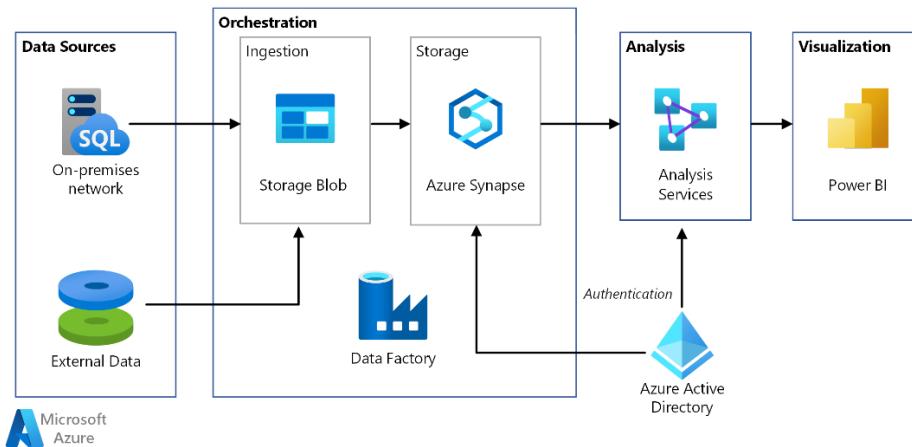


Les entrepôts de données en étant la destination des données dans l'architecture ETL, jouent un rôle primordial dans ce processus. Sans les entrepôts de données, l'accès aux données direct dans les bases de données traditionnel n'est ni productif ni sécuritaire. En effet travailler directement sur les bases de données de production peut s'avérer fatal pour n'importe quelle organisation. Les entrepôts des données permettent d'isoler et sécuriser les données de production pour créer des répliques transformés sur lesquelles les analystes peuvent travailler en toute sécurité.

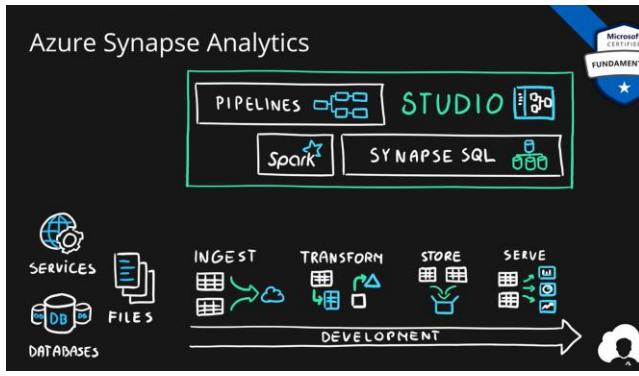
L'exemple le plus évident est le domaine de l'intelligence d'affaire. Les entrepôts de données permettent aux analystes BI de faire des analyses qui permettent une prise de décision bien informée et plus rapide, car ils ont entre les mains les données les plus pertinentes qu'ils peuvent manipuler et exploiter en toute sécurité.

L'automatisation et L'intelligence artificielle sont aussi des domaines qui reposent sur les entrepôts de données car ils ont essentiellement besoin de données massives et structurées. Plusieurs outils sont disponibles pour tirer profit de ces nouveaux paradigmes et architectures.

Par exemple, dans l'architecture ETL d'Azure DataFactory, le service Azure Synapse Analytics fournit la fonctionnalité Synapse Pipeline qui permet d'appliquer les étapes d'extraction et de transformation des données à l'aide d'une interface graphique de workflow visuels, gérée dans Synapse Studio qui permet d'utiliser tous ces outils et fonctionnalités d'ingestion et de transformation dans un endroit unique :



Le service Synapse Analytics est fourni avec Apache Spark et Synapse SQL déjà intégrés, ce qui permet des traitements en mode big data et en mode base de données relationnelles (SQL Server) :



En plus, Azure HDInsights et Azure Databricks fournissent un support puissant pour le processus ETL en permettant d'utiliser les technologies de clusters Big Data comme Hadoop et Spark.

On voit donc l'importance du service Azure DataFactory comme étant une plateforme avec des capacités d'ETL, de big data et d'intégration avec les outils d'analyse et apprentissage machine et finalement, de visualisation. Ces différentes capacités permettent également d'obtenir de meilleures automatisations et de meilleurs taux de précision dans les scénarios d'apprentissage machine, vu le grand volume et la structuration optimisée des données exploitées.

## Implémentation : Structure, Extraction, Ingestion et Pipelines

### Définition des éléments projet final

Ressources	Noms
<b>Groupe de ressources</b>	projet_final_groupe1_rg
<b>Compte de stockage BLOB</b>	projefinalgroupe1sa
<b>Containers BLOB</b>	projefinalgroupe1blob
<b>Compte de stockage GEN2</b>	projefinalgroupe1dl
<b>Containers GEN2</b>	containerprojefinalgroupe1dl caontainerfinalcleaned powerbicontainer
<b>Fabrique de données</b>	projefinalgroup1-adf

<b>Datasets</b>	<ul style="list-style-type: none"> <li>▲ Datasets</li> <li>ds_airlines_destination_projetfinalgroupe1adls</li> <li>ds_airlines_powerbi</li> <li>ds_airlines_projetfinalgroupe1blb</li> <li>ds_airports_destination_projetfinalgroupe1adls</li> <li>ds_airports_projetfinalgroupe1blb</li> <li>ds_flights_destination_projetfinalgroupe1adls</li> <li>ds_flights_projetfinalgroupe1blb</li> <li>ds_projetfinal_cleaned</li> </ul>	8
<b>Linked Services</b>	ls_projetfinalgroupe1_adls ls_projetfinalgroupe1_blb	
<b>Copy Activities</b>	<div style="border: 1px solid #ccc; padding: 5px; margin-bottom: 10px;"> <span>Copier les données</span>   Copy data airlines         </div> <div style="border: 1px solid #ccc; padding: 5px; margin-bottom: 10px;"> <span>Copier les données</span>   Copy data airports         </div> <div style="border: 1px solid #ccc; padding: 5px; margin-bottom: 10px;"> <span>Copier les données</span>   Copy data flights         </div>	
<b>Dataflows</b>	<ul style="list-style-type: none"> <li>▲ Data flows</li> <li>df_airlines_powerbi</li> <li>df_projetfinal_airlines</li> </ul>	2
<b>Pipelines</b>	<ul style="list-style-type: none"> <li>▲ Pipelines</li> <li>plp2-projefinalgroupe1</li> <li>plp3_projetfinal_powerbi</li> <li>plp-projetfinalgroup1</li> </ul>	3

Tableau de bord des ressources utilisées :

**Ressources**

projet\_final\_groupe1\_rg

Actualiser

projetfinalgroupe1dl	Compte de stockage	East US
projetfinalgroup1adf	Fabrique de données (V2)	
projetfinalgroupe1sa	Compte de stockage	

**projetfinalgroupe1dl**  
Compte de stockage

Genre StorageV2  
ID de l'abonnement a269cc7b-ece1-4f11-b873-d9a970b9f5...  
Groupe de ressources projet\_final\_groupe1\_rg  
Type de réplication Stockage géo-redondant avec accès en...  
Niveau d'accès

Emplacement East US  
Abonnement Azure for Students  
RÉFÉRENCE (SKU) Standard\_RAGRS  
Créé 14/10/2022 5:06:22 PM  
Emplacement principal

**projetfinalgroupe1sa**  
Compte de stockage

Genre StorageV2  
ID de l'abonnement a269cc7b-ece1-4f11-b873-d9a970b9f5...  
Groupe de ressources projet\_final\_groupe1\_rg  
Type de réplication Stockage géo-redondant avec accès en...  
Niveau d'accès

Emplacement East US  
Abonnement Azure for Students  
RÉFÉRENCE (SKU) Standard\_RAGRS  
Créé 14/10/2022 4:45:28 PM  
Emplacement principal

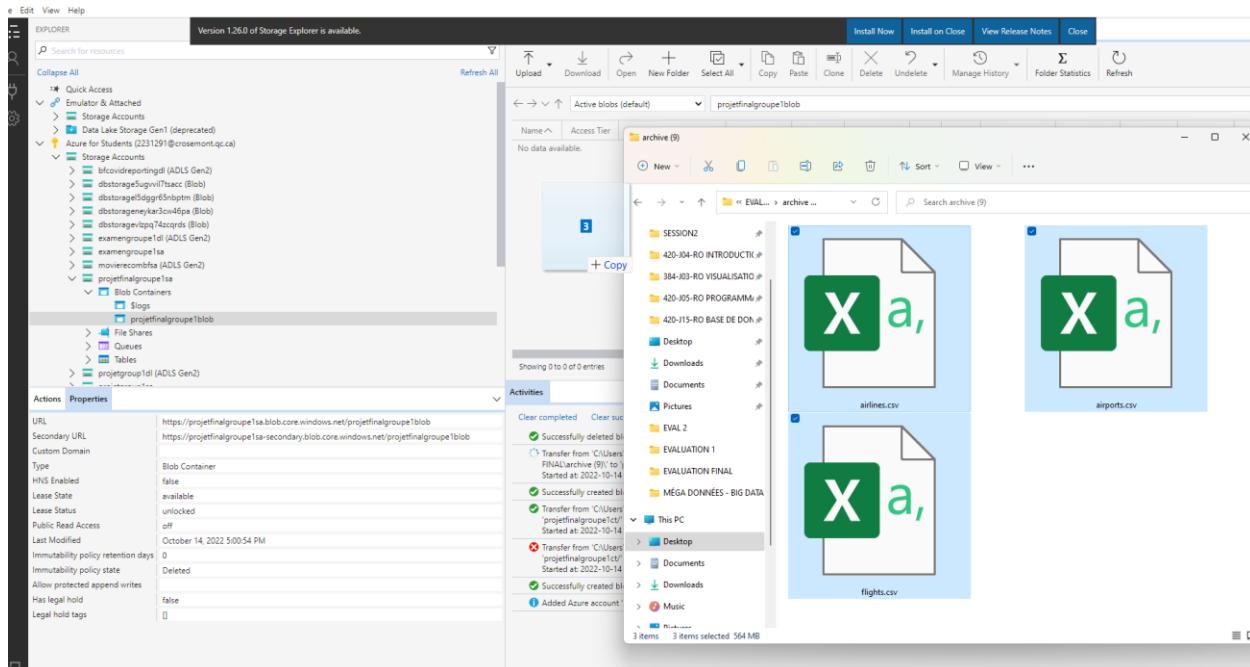
## Extraction et ingestion des données

Transformations envisagées

Transformations et opérations sur les colonnes :

- SCHEDULED\_DEPARTURE : mettre au format string, puis au format heure
- DEPARTURE\_TIME : mettre au format string, puis au format heure
- IATA\_CODE : utiliser pour Lookup avec flights. On utilisera juste le IATA\_CODE de l'origine du vol
- MONTH : transformer le nombre à un nom de mois
- DAY\_OF\_WEEK : transformer le nombre à un nom de jour

## Création des éléments du projet :



**Microsoft Azure** Rechercher dans les ressources, services et documents (G+)

Accueil > projetfinalgroupe1dl\_1665781574190 | Vue d'ensemble >

### projetfinalgroupe1dl

Compte de stockage

Rechercher

Vue d'ensemble

- Journal d'activité
- Étiquettes
- Diagnostiquer et résoudre les problèmes
- Contrôle d'accès (IAM)
- Migration des données
- Événements
- Navigateur de stockage (préversion)

Bases

Groupe de ressources (déplacé) : projet_final_groupe1_rg	Performances : Standard
Emplacement : East US	Réplication : Stockage géo-redondant avec accès en lecture (RA-GRS)
Emplacement principal/secondaire : Principal : East US, secondaire : West US	Type de compte : StorageV2 (v2 à usage général)
Abonnement (déplacer) : Azure for Students	État de provisionnement : Réussite
ID d'abonnement : a269cc7b-bce1-4f11-b873-d9a970b9f5d4	Créé : 10/14/2022, 5:06:22 PM
État du disque : Principal : Disponible, secondaire : Disponible	

Étiquettes (modifier) : Cliquez ici pour ajouter des étiquettes

Propriétés Supervision Fonctionnalités (5) Recommandations Tutoriels Outils de développement

**Data Lake Storage**

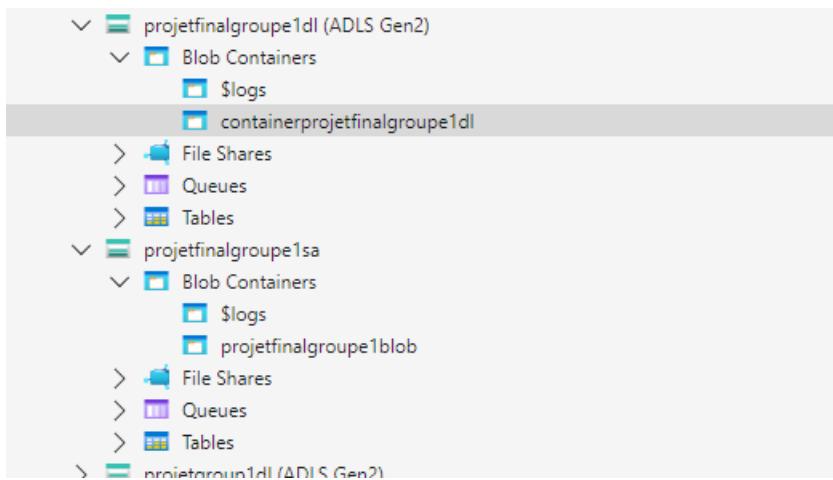
Espace de noms hiérarchique	Activé
Niveau d'accès par défaut	Hot
Accès public aux objets blob	Activé
Suppression réversible d'objet blob	Activé (7 jours)
Suppression réversible de conteneur	Activé (7 jours)
Gestion des versions	Désactivé
Flux de modification	Désactivé
NFS v3	Désactivé
SFTP (préversion)	Désactivé

**Sécurité**

Exiger un transfert sécurisé pour les opérations d'API REST	Activé
Accès de clé de compte de stockage	Activé
Version TLS minimale	Version 1.2
Chiffrement d'infrastructure	Désactivé

**Réseau**

Autoriser l'accès à partir de	Tous les réseaux
Nombre de connexions de point de terminaison privé	0



## Créer un nouvel ADF (Azure Data Factory) : projetfinalgroup1-adf

The screenshot shows the Azure portal interface for creating a new Azure Data Factory. The top navigation bar includes 'Microsoft Azure', a search bar, and user information '2231291@crosemont.qc... COLLEGE DE ROSEMONT (CROSE...)'.

The main content area shows the creation details for 'projetfinalgroup1-adf':

- Bases**:
  - Groupe de res... ([déplacer](#)) : `projet_final_groupe1_rg`
  - Status : Succeeded
  - Emplacement : East US
  - Abonnement ([déplacer](#)) : `Azure for Students`
  - ID d'abonnement : `a269cc7b-ece1-4f11-b873-d9a970b95d4`
- Démarrer**:
  - [Ouvrir Azure Data Factory Studio](#) (Icone d'usine)
  - [Lire la documentation](#) (Icone de livre)
- Supervision**:
  - PipelineRuns**: 100, 80, 60
  - ActivityRuns**: 100, 80, 60

Microsoft Azure | Data Factory > projetfinalgroup1-adf

Azure Data Factory vous permet de configurer un dépôt Git avec Azure DevOps ou GitHub. Git est un système de contrôle de version qui facilite le suivi des changements et la collaboration. En savoir plus

Data factory  
projetfinalgroup1-adf

Nouveau

Ingérer Copier les données à grande échelle une fois ou selon une planification.

Orchestrer Pipelines de données sans code.

Transformer les données Transformer vos données à l'aide des flux de données.

Configurer SSIS Gérez et exécutez vos packages SSIS dans le cloud.

En savoir plus

Parcourir les partenaires (préversion)

Modèles de pipeline

Modèles de pipeline SAP

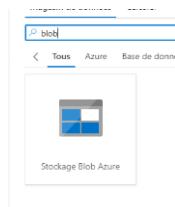
Ressources récentes

Aucun élément à afficher

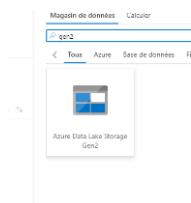
Vos dernières ressources ouvertes s'affichent ici.

Créer deux services liés (Linked services) :

Le premier correspond au compte de stockage source : ls\_projetfinalgroupe1\_bblob



Le deuxième correspond au compte de stockage destination : ls\_projetfinalgroupe1\_adls



Créer six jeux de données:

(côté source correspond au Blob storage)

**ds\_airlines\_projetfinalgroupe1blob**

**ds\_airports\_projetfinalgroupe1blob**

**ds\_flights\_projetfinalgroupe1blob**

Microsoft Azure | Data Factory > projetfinalgroup1-adf

Tout valider Rechercher Tout publier ?

**Ressources de fabrique**

- PIPES
- Datasets
- Data flows
- Power Query

Filtrer les ressources par nom : +

**ds\_airlines\_projetfinalgroup1adfs** (DelimitedText)

Connexion Schéma Réglages

Service lié : ls\_projetfinalgroup1\_adfs

Chemin d'accès au fichier : containerprojetfinalgroup1adfs\annuaire\Nom de fichier

Type de compression : Aucun

Séparateur de colonne : Vierge (,)

Délimiteur de ligne : Par défaut (\r\n ou \n\r)

Encodage : Par défaut(UTF-8)

Caractère d'échappement : Barre oblique inverse (\)

Guillemet : Guillemet double (")

Première ligne comme en-tête :

**Tout publier**

Vous êtes sur le point de publier tous les changements en attente de l'environnement en direct.

En savoir plus

**Modifications en attente (3)**

NOM	CHANGER	EXISTANT
Jeux de données		
ds_airlines_destination_pr... (Nouveau)	-	
ds_airports_destination_pr... (Nouveau)	-	
ds_flights_destination_pr... (Nouveau)	-	

**Publier** **Annuler**

Microsoft Azure | Data Factory > projetfinalgroup1-adf

Tout valider Rechercher Tout publier ?

**Ressources de fabrique**

- PIPES
- Datasets
- Data flows
- Power Query

Filtrer les ressources par nom : +

**ds\_airlines\_projetfinalgroup1adfs** (Dataset)

**Aperçu des données**

Service lié : ls\_projetfinalgroup1\_adfs

Objet : airlines.csv

IATA_CODE	ALINE
UA	United Air Lines Inc.
AA	American Airlines Inc.
US	US Airways Inc.
F9	Frontier Airlines Inc.
B6	JetBlue Airways
OO	Skywest Airlines Inc.
AS	Alaska Airlines Inc.
NK	Spirit Air Lines
WN	Southwest Airlines Co.
DL	Delta Air Lines Inc.

**Propriétés**

Général Associé

Nom : ds\_airlines\_projetfinalgroup1adfs

Description :

Annotations

**Guillemet :**  Guillemet double (")

Première ligne comme en-tête :

Microsoft Azure | Data Factory > projetfinalgroup1-adf

Home Author Monitor Manage

Ressources de fabrique

Aperçu des données

Service lié: ls\_projetfinalgroupe1\_bib  
Objet: airports.csv

	IATA_CODE	AIRPORT	CITY	STATE	COUNTRY	LATITUDE	LONGITUDE
1	ABE	Lehigh Valley International Airport	Allentown	PA	USA	40.65236	-75.44040
2	ABI	Abilene Regional Airport	Abilene	TX	USA	32.41132	-99.68190
3	ABQ	Albuquerque International Sunport	Albuquerque	NM	USA	35.04022	-106.60919
4	ABR	Aberdeen Regional Airport	Aberdeen	SD	USA	45.44906	-98.42183
5	ABY	Southwest Georgia Regional Airport	Albany	GA	USA	31.53552	-84.19447
6	ACK	Nantucket Memorial Airport	Nantucket	MA	USA	41.25305	-70.06018
7	ACT	Waco Regional Airport	Waco	TX	USA	31.61129	-97.23052
8	ACV	Arcata Airport	Arcata/Eureka	CA	USA	40.97812	-124.10862
9	ACY	Atlantic City International Airport	Atlantic City	NJ	USA	39.45758	-74.57717

Guillemet  Guillemet double  Modifier

Première ligne comme en-tête

Propriétés

Nom: ds\_airports\_projetfinalgroupe1bib

Description:

Annotations

Microsoft Azure | Data Factory > projetfinalgroup1-adf

Home Author Monitor Manage

Ressources de fabrique

Aperçu des données

Service lié: ls\_projetfinalgroupe1\_bib  
Objet: flights.csv

	YEAR	MONTH	DAY	DAY_OF_WEEK	AIRLINE	FLIGHT_NUMBER	TAIL_NUMBER	ORIGIN_AIRPORT
1	2015	1	1	4	AS	98	N407AS	ANC
2	2015	1	1	4	AA	2336	N3KUAA	LAX
3	2015	1	1	4	US	840	N171US	SFO
4	2015	1	1	4	AA	258	N3HYAA	LAX
5	2015	1	1	4	AS	135	N527AS	SEA
6	2015	1	1	4	DL	806	N3730B	SFO
7	2015	1	1	4	NK	612	N635NK	LAS
8	2015	1	1	4	US	2013	N584UW	LAX
9	2015	1	1	4	AA	1112	N3LAAA	SFO
10	2015	1	1	4	DL	1173	N826DN	LAS

Guillemet  Guillemet double  Modifier

Première ligne comme en-tête

Propriétés

Nom: ds\_flights\_projetfinalgroupe1bib

Description:

Annotations

(côté réception correspond au Data storage Lake Gen 2)

ds\_airlines\_destination\_projetfinalgroupe1adls

ds\_airports\_destination\_projetfinalgroupe1adls

ds\_flights\_destination\_projetfinalgroupe1adls

**Ressources de fabrique**

**Propriétés**

**Connexion**

- Service lié : ls\_projetfinalgroupe1\_ads
- Chemin d'accès au fichier : containerprojetfinalgro... / Annuaire
- Nom de fichier : Nom de fichier
- Parcourir
- Aperçu des données

**Schéma**

**Règles**

Créer un pipeline de données plp-projetfinalgroup1 et ajouter une activité pour copier les données (Copy Activity) du conteneur source vers le conteneur destination

**Ressources de fabrique**

**Propriétés**

**Source**

Jeu de données source : ds\_airlines\_projetfinalgroupe1bbs

Type de chemin de fichier : Chemin de fichier dans le jeu de données

Heure de début (UTC)

Heure de fin (UTC)

Filtrer par heure de dernière modification

De manière récursive

Activer la découverte de partition

Nombre maximal de connexions simultanées

Ignorer le nombre de lignes

The screenshot shows the Microsoft Azure Data Factory pipeline editor. On the left, the 'Ressources de fabrique' sidebar lists Pipelines, Datasets, Data flows, and Power Query. The main workspace displays three parallel 'Copy data' operations within a single pipeline:

- Top: 'Copy data airlines' (green checkmark)
- Middle: 'Copy data airports' (green checkmark)
- Bottom: 'Copy data flights' (green checkmark)

The 'Propriétés' panel on the right shows the pipeline's properties: Name = 'plp-projetfinalgroup1'. The 'Sortie' tab is selected.

On lance le débogage :

The screenshot shows the Microsoft Azure Data Factory pipeline editor in debug mode. The 'Sortie' tab is selected. The 'ID d'exécution de pipeline' is listed as 'e43bSafe-d528-4961-8fd9-10829716af88'. Below it, a table details the execution results for each step:

Nom	Type	Début de l'exécution	Durée	Etat	Runtime
Copy data airports	Copier les données	2022-10-15T14:50:09.831677	00:00:08	Opération réussie	AutoResc
Copy data flights	Copier les données	2022-10-15T14:50:09.831677	00:00:18	Opération réussie	AutoResc
Copy data airlines	Copier les données	2022-10-15T14:50:09.831677	00:00:08	Opération réussie	AutoResc

Microsoft Azure | Data Factory > projetfinalgroup1-adf

Détails    Actualiser

Explorez les détails des performances de copie ici.

ID d'exécution d'activité : 8a119965-c427-416f-9771-d08ee9313219

 Stockage Blob Azure  
Région: East US      Opération réussie       Azure Data Lake Storage Gen2  
Région: East US

Données lues : ①	564.963 Mo	Données écrites : ①	564.963 Mo
Fichiers lus : ①	1	Fichiers écrits : ①	1
Nombre maximal de connexions : ①	10	Nombre maximal de connexions : ①	16

Durée de copie      00:00:15  
Débit : ①      37.664 Mo/s

Stockage Blob Azure → Azure Data Lake Storage Gen2

Heure de début	Oct 15, 2022, 10:50:10 am	
DIU utilisées ①	4	
Copies parallèles utilisées ①	1	
Durée	00:00:15	
Détails	Durée de travail	Durée totale
File d'attente ①	[ Source de liste ① 00:00:00 Lecture à partir de la source ① 00:00:04 Écriture dans le récepteur ① 00:00:07 ]	00:00:04
Transfert ①		00:00:09

Vérification de la cohérence des données      ① Non vérifié

Étes-vous satisfait ou mécontent des performances de cette activité Copy ? 

Microsoft Azure | Data Factory > projetfinalgroup1-adf

Détails    Actualiser

Explorez les détails des performances de copie ici.

ID d'exécution d'activité : cb92b778-7503-411c-ab35-ff011c619b45

 Stockage Blob Azure  
Région: East US      Opération réussie       Azure Data Lake Storage Gen2  
Région: East US

Données lues : ①	23.308 Ko	Données écrites : ①	23.308 Ko
Fichiers lus : ①	1	Fichiers écrits : ①	1
Nombre maximal de connexions : ①	1	Nombre maximal de connexions : ①	1

Durée de copie      00:00:06  
Débit : ①      3.885 Ko/s

Stockage Blob Azure → Azure Data Lake Storage Gen2

Heure de début	Oct 15, 2022, 10:50:10 am	
DIU utilisées ①	4	
Copies parallèles utilisées ①	1	
Durée	00:00:06	
Détails	Durée de travail	Durée totale
File d'attente ①	[ Source de liste ① 00:00:00 ]	00:00:03
Transfert ①	[ Lecture à partir de la source ① 00:00:00 Écriture dans le récepteur ① 00:00:00 ]	00:00:01

Vérification de la cohérence des données      ① Non vérifié

Étes-vous satisfait ou mécontent des performances de cette activité Copy ? 

Microsoft Azure | Data Factory > projetfinalgroup1-adf

Détails    Actualiser

Explorez les détails des performances de copie ici.

ID d'exécution d'activité : 693bf523-7e5e-46ea-bf36-388e6ec71df2

**Opération réussie**

Stockage Blob Azure → Azure Data Lake Storage Gen2

Région: East US      Région: East US

Données lues : 359 octets      Données écrites : 359 octets

Fichiers lus : 1      Fichiers écrits : 1

Nombre maximal de connexions : 1      Nombre maximal de connexions : 1

Durée de copie : 00:00:06

Débit : 59.392 octets/s

Vérification de la cohérence des données : Non vérifié

Étes-vous satisfait ou mécontent des performances de cette activité Copy ? ★★★★☆

## On vérifie le container destination :

Microsoft Azure Storage Explorer

File Edit View Help

EXPLORER Version 1.26.0 of Storage Explorer is available.

Quick Access

- Emulator & Attached
- Storage Accounts
- Data Lake Storage Gen1 (deprecated)
- Azure for Students (2331291@crosemont.qc.ca)
  - Storage Accounts
    - bfcovreportingndl (ADLS Gen2)
    - dbstorageSugvnl7tacc (Blob)
    - dbstoragef5dgr5n1ptpm (Blob)
    - dbstoragefeneckaricw4dp6 (Blob)
    - dbstoragef7p74ccordrs (Blob)
    - examengroupe1dl (ADLS Gen2)
    - examengroupe1sa
    - movierecombs (ADLS Gen2)
  - blob Containers
    - Slogs
    - containerprojetfinalgroup1dl
    - File Shares
    - Queues
    - Tables
  - projectfinalgroup1sa
  - projectgroup1dl (ADLS Gen2)
  - projectgroup1sa
  - rovcoideReportingbfsa
- Disks
- CovidReporting-rg
- databricks-rg-db-examen-groupe1-q4kay4dwxmduy

Actions Properties

URL: https://projetfinalgroup1dl.blob.core.windows.net/

Secondary URL: https://projetfinalgroup1dl-secondary.blob.core.windows.net/

Custom Domain:

Type: Blob Container (ADLS Gen2)

HNS Enabled: true

DFS Endpoint: https://projetfinalgroup1dl.dfs.core.windows.net/

Activities

- Added Azure account 2231291@crosemont.qc.ca'
- Deletion of 'New Folder/' from 'projetfinalgroup1blob/' completed: 2 completed (used SAS discovery completed)
- Transfer from 'C:\Users\B\Desktop\COURS\SESSION3\PROJET VALORISATION DE DONNÉES\EVALUATION FINAL\archive ([9])' to 'projetfinalgroup1blob/New Folder/' complete: 2 items transferred (used SAS discovery completed)
- Successfully created blob container 'containerprojetfinalgroup1dl'
- Transfer from 'C:\Users\B\Desktop\COURS\SESSION3\PROJET VALORISATION DE DONNÉES\EVALUATION FINAL\archive ([9])' to 'projetfinalgroup1blob/' complete: 3 items transferred (used SAS discovery completed)
- Successfully deleted blob container 'projetfinalgroup1ct'
- Transfer from 'C:\Users\B\Desktop\COURS\SESSION3\PROJET VALORISATION DE DONNÉES\EVALUATION FINAL\archive ([9])' to 'projetfinalgroup1blob/' complete: 3 items transferred (used SAS discovery completed)
- Successfully created blob container 'projetfinalgroup1blob'
- Transfer from 'C:\Users\B\Desktop\COURS\SESSION3\PROJET VALORISATION DE DONNÉES\EVALUATION FINAL\archive ([9])' to 'projetfinalgroup1ct/' complete: 3 items transferred (used SAS discovery completed)

## On crée un DataFlow pour la transformation des données ingérées :

**Ressources de fabrique**

- Pipelines
- Datasets
- Data flows
- Power Query

**Flux entrant**

**Récepteur**

**Définir les propriétés**

Nom : ds\_projetfinal\_cleaned

Service lié : Is\_projetfinalgroup1\_adls

Chemin d'accès au fichier : caontainerfinalcleaned / Annuaire / Nom de fichier

Première ligne comme en-tête :

Importer un schéma :  À partir d'une connexion/un magasin  À partir d'un exemple de fichier  
Aucun

Avancé

**Ressources de fabrique**

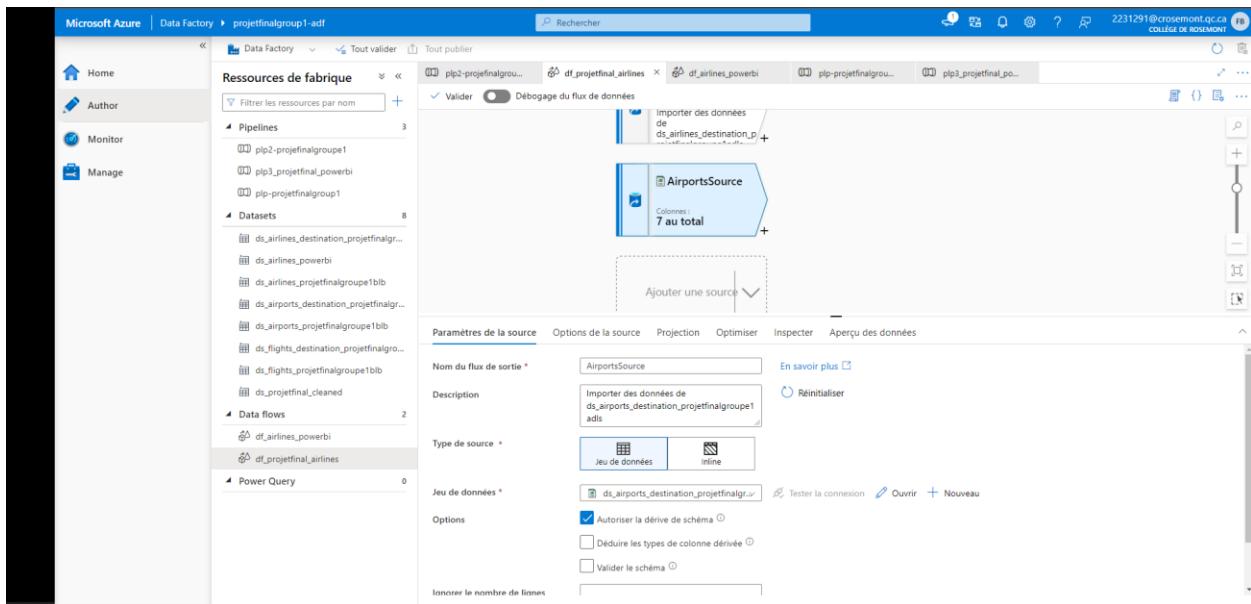
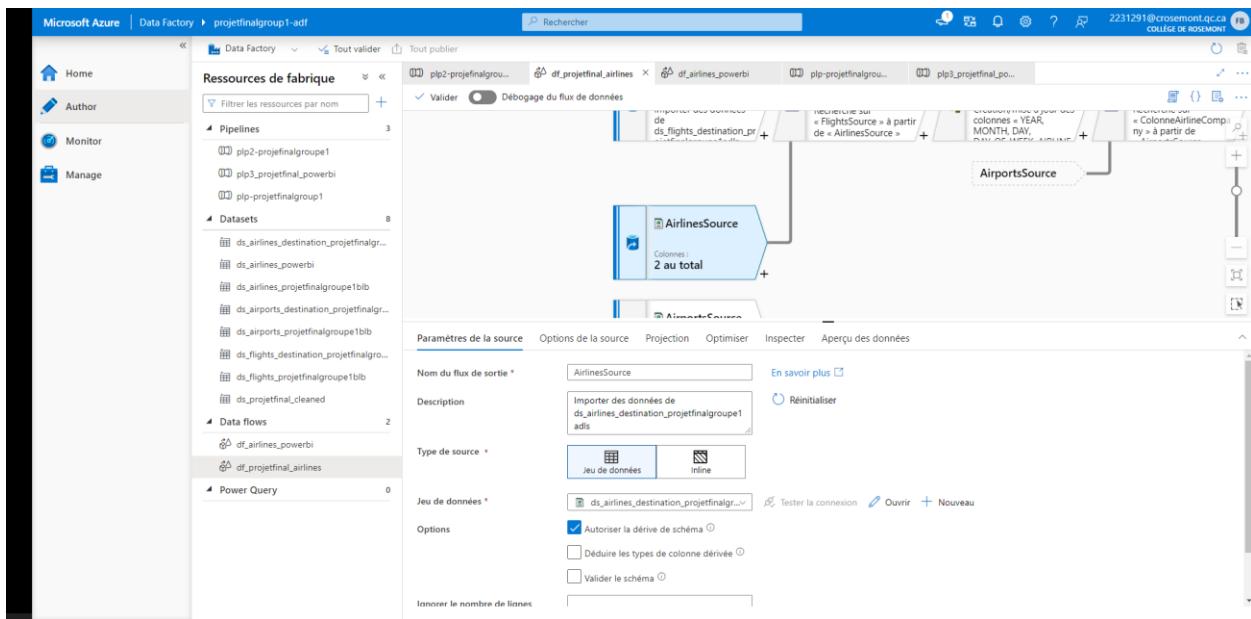
- Pipelines
- Datasets
- Data flows
- Power Query

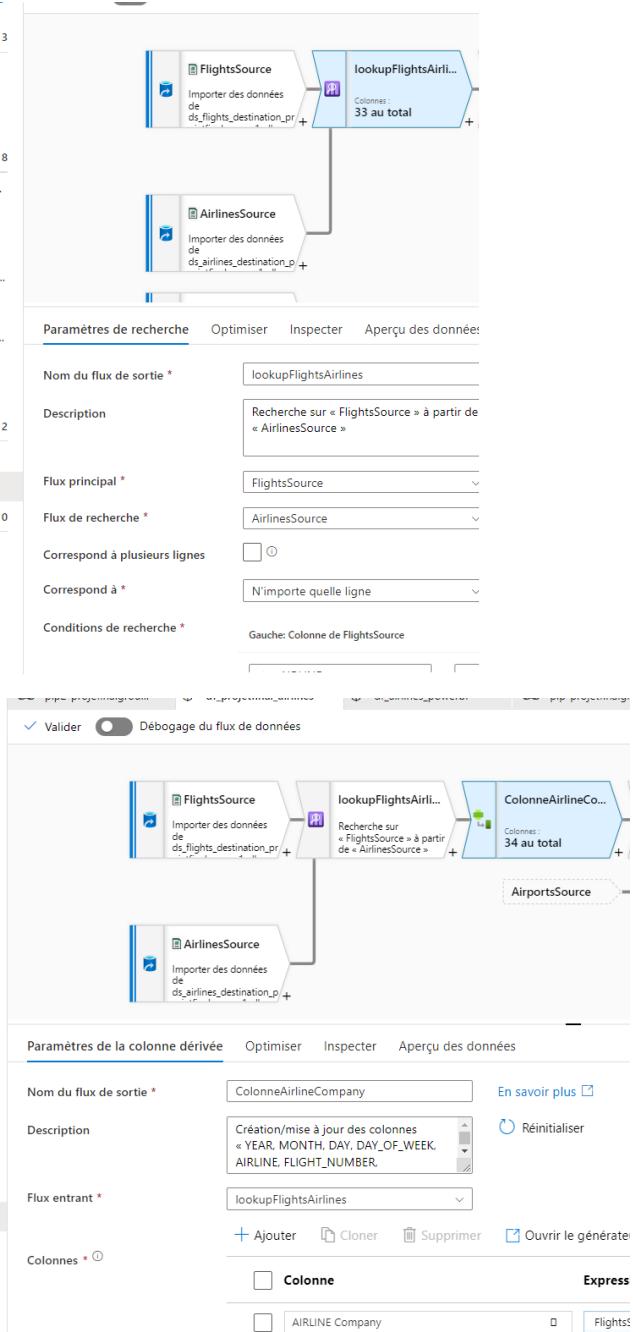
**Flux entrant**

**Propriétés**

Nom : df\_projetfinal\_airlines

Description :





Microsoft Azure | Data Factory > projetfinalgroup1-adf

Rechercher

Générateur d'expressions de flux de données

ColonneAirlineCompany

Colonnes Dérivée

+ Créer

AIRLINE Company

Nom de la colonne \* AIRLINE Company

Expression

FlightsSource@AIRLINE+->AirlinesSource@AIRLINE

Enregistrer

Éléments d'expression

Valeurs d'expression

Tous

Fonctions

Schéma d'entrée

Réglages

Recherche en cache

Fonctions de bibliothèque de flux de données

Variables locales

YEAR

MONTH

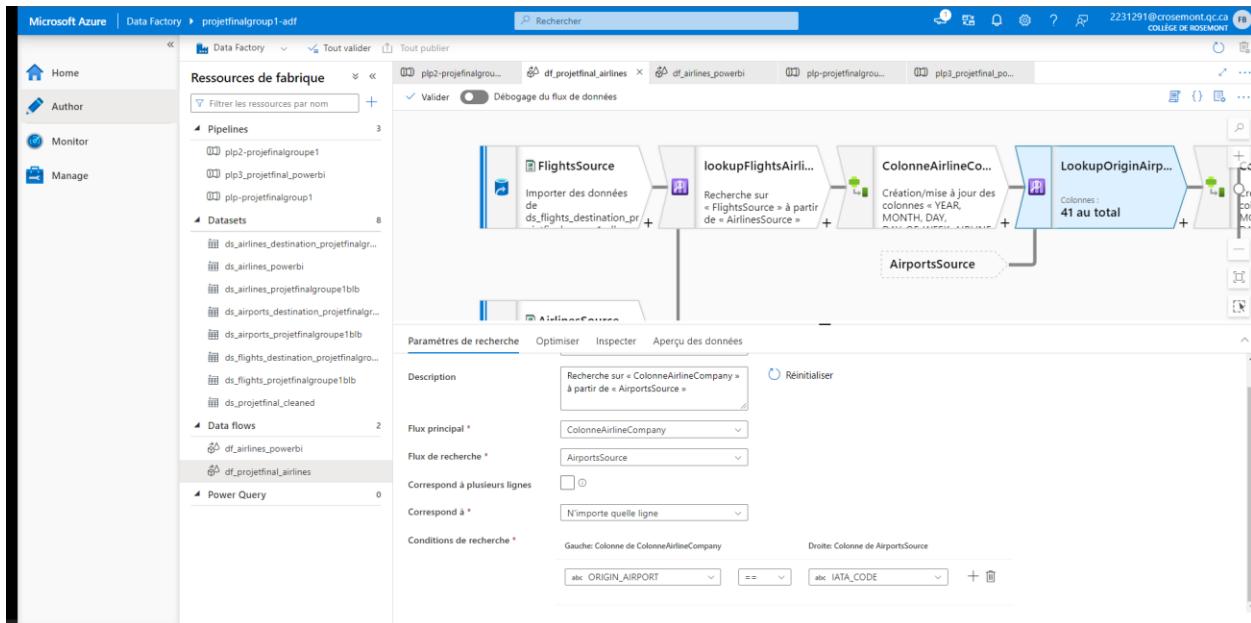
DAY

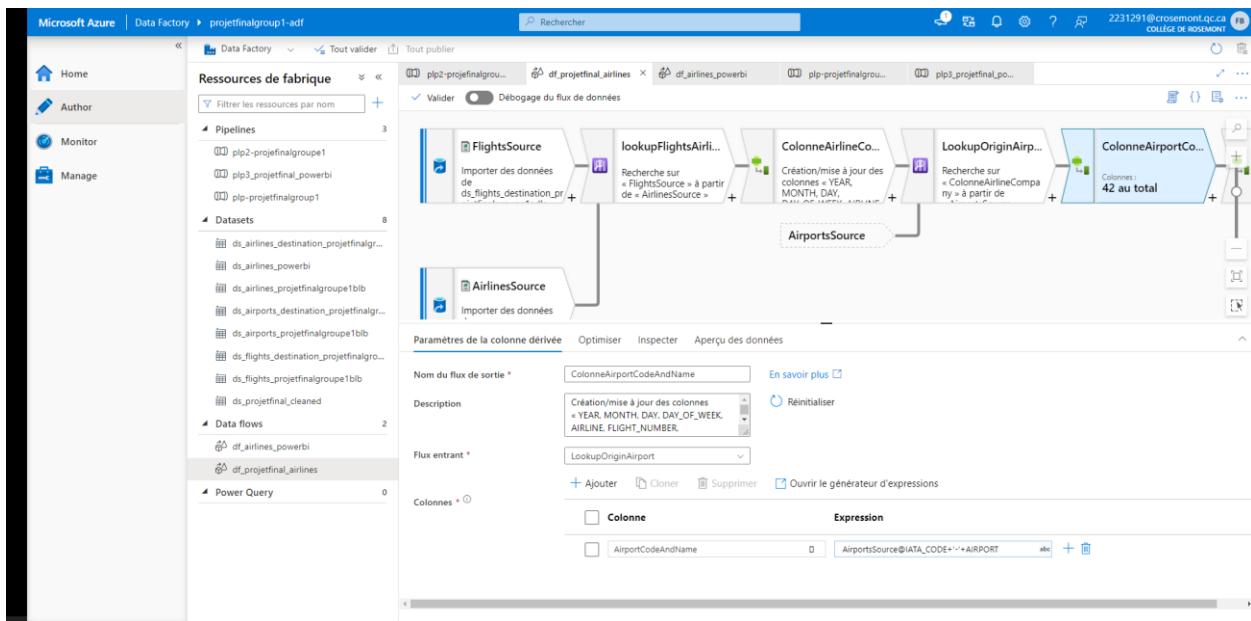
DAY\_OF\_WEEK

FlightsSource@AIRLINE

FLIGHT\_NUMBER

Enregistrer et terminer Annuler Effacer le contenu





Microsoft Azure | Data Factory > projetfinalgroup1-adf

Générateur d'expressions de flux de données

Colonnes Dérivée

+ Créer

Colonnes Dérivée

Nom de la colonne \* AirportCodeAndName

Expression

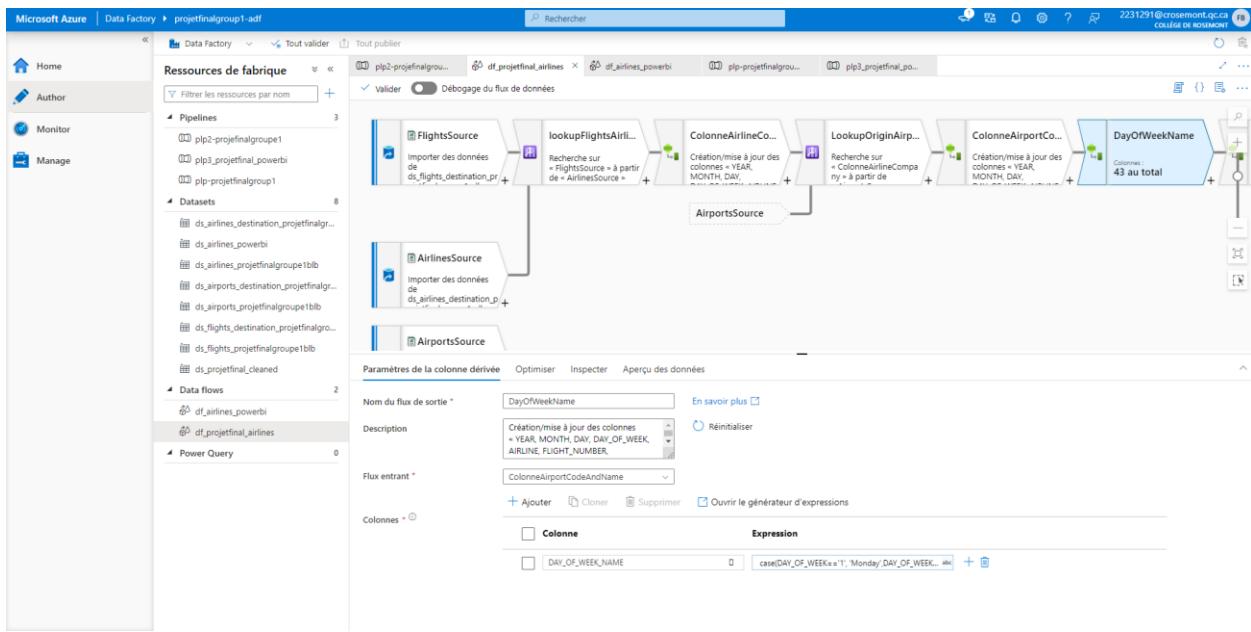
AirportsSource@IATA\_CODE+''+AIRPORT

Éléments d'expression

	Valeurs d'expression
Tous	Filtrer par mot clé
Fonctions	<input type="button" value="Créer"/>
Schéma d'entrée	YEAR
Réglages	MONTH
Recherche en cache	DAY
Fonctions de bibliothèque de flux de données	DAY_OF_WEEK
Variables locales	FlightsSource@AIRLINE
	FLIGHT_NUMBER

Aperçu des données

Enregistrer et terminer Annuler Effacer le contenu



Générateur d'expressions de flux de données

**Colonnes Dérivée**

- + Créer
- DAY\_OF\_WEEK\_NAME

**Nom de la colonne \***: DAY\_OF\_WEEK\_NAME

**Expression**

```
case(
    DAY_OF_WEEK<=1, "Monday",
    DAY_OF_WEEK<=2, "Tuesday",
    DAY_OF_WEEK<=3, "Wednesday",
    DAY_OF_WEEK<=4, "Thursday",
    DAY_OF_WEEK<=5, "Friday",
    DAY_OF_WEEK<=6, "Saturday",
    DAY_OF_WEEK<=7, "Sunday"
)
```

**Éléments d'expression**

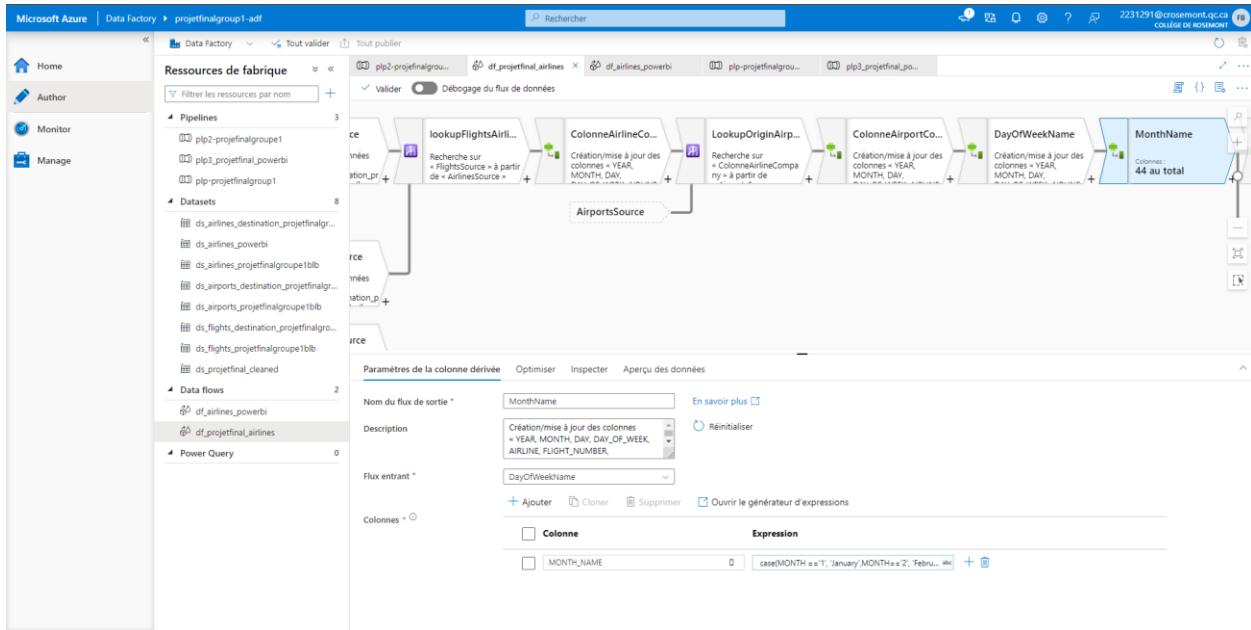
- Tous
- Fonctions
- Schéma d'entrée
- Réglages
- Recherche en cache
- Fonctions de bibliothèque de flux de données
- Variables locales

**Valeurs d'expression**

- YEAR
- MONTH
- DAY
- DAY\_OF\_WEEK
- FlightsSource@AIRLINE
- FLIGHT\_NUMBER
- TAIL NUMBER

Aperçu des données

Enregistrer et terminer Annuler Effacer le contenu



Microsoft Azure | Data Factory | projetfinalgroup1-adf

**Générateur d'expressions de flux de données**

**Colonnes Dérivée**

**MonthName**

**Colonnes**

**MONTH\_NAME**

**Nom de la colonne \***: MONTH\_NAME

**Expression**

```

case(
    MONTH-->1, 'January',
    MONTH-->2, 'February',
    MONTH-->3, 'March',
    MONTH-->4, 'April',
    MONTH-->5, 'May',
    MONTH-->6, 'June',
    MONTH-->7, 'July',
    MONTH-->8, 'August',
    MONTH-->9, 'September',
    MONTH-->10, 'October',
    MONTH-->11, 'November',
    MONTH-->12, 'December'
)

```

**Éléments d'expression**

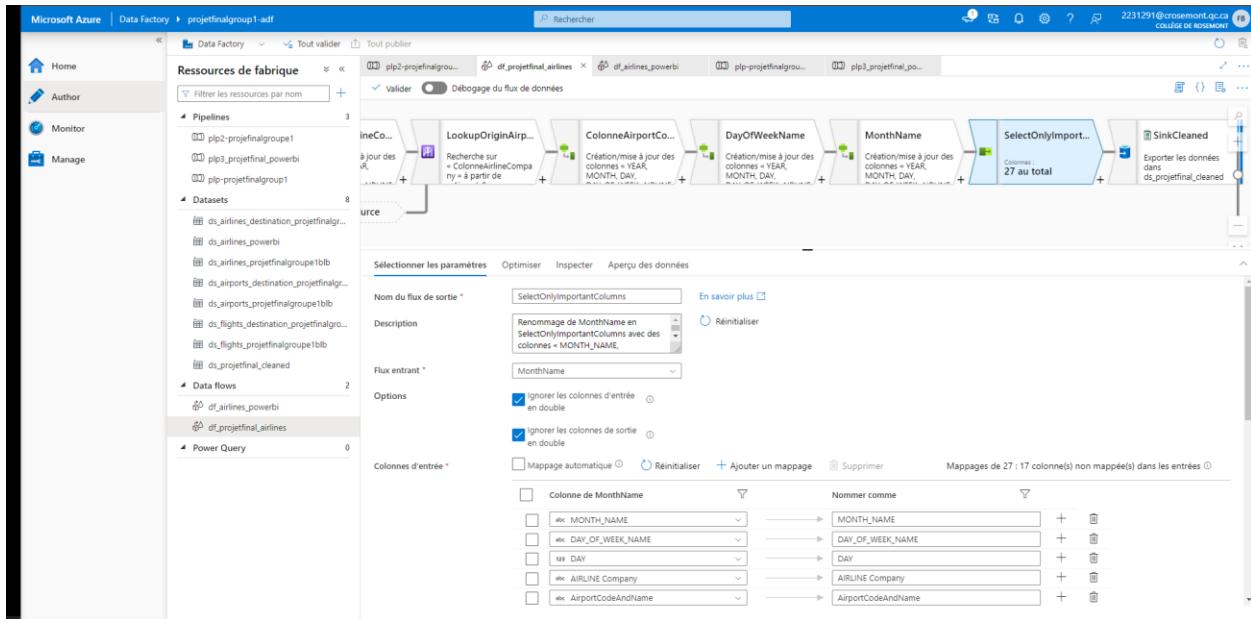
- Tous
- Fonctions
- Schéma d'entrée
- Réglages
- Recherche en cache
- Fonctions de bibliothèque de flux de données
- Variables locales

**Valeurs d'expression**

- YEAR
- MONTH
- DAY
- DAY\_OF\_WEEK
- FlightSource@AIRLINE

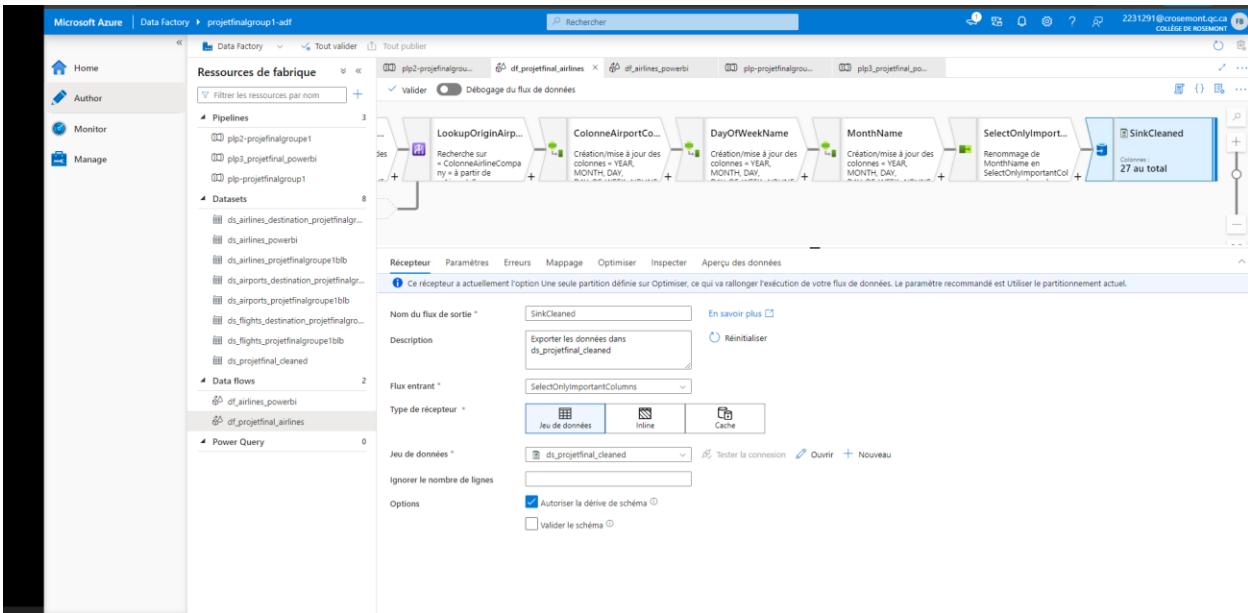
**Aperçu des données**

**Enregistrer et terminer** **Annuler** **Effacer le contenu**



On garde uniquement les colonnes nécessaires à notre analyse

Colonne de MonthName	Nommer comme
12s AIR_TIME	AIR_TIME
12s DISTANCE	DISTANCE
12s WHEELS_ON	WHEELS_ON
12s SCHEDULED_ARRIVAL	SCHEDULED_ARRIVAL
12s ARRIVAL_TIME	ARRIVAL_TIME
12s ARRIVAL_DELAY	ARRIVAL_DELAY
✗ DIVERTED	DIVERTED
✗ CANCELLED	CANCELLED
abc CANCELLATION_REASON	CANCELLATION_REASON
12s AIR_SYSTEM_DELAY	AIR_SYSTEM_DELAY
12s SECURITY_DELAY	SECURITY_DELAY
12s AIRLINE_DELAY	AIRLINE_DELAY
12s LATE_AIRCRAFT_DELAY	LATE_AIRCRAFT_DELAY
12s WEATHER_DELAY	WEATHER_DELAY
abc STATE	STATE



Après, on crée les pipelines pour transférer les données :

The screenshot shows the Microsoft Azure Data Factory pipeline editor. The left sidebar shows the pipeline 'df\_projfinal\_airlines' selected under 'Pipelines'. The main area shows the pipeline's configuration. The properties panel on the right shows the pipeline name as 'df\_projfinal\_airlines' and a description. The 'General' tab is selected, showing fields for 'Nom' (df\_projfinal\_airlines), 'Description', 'Délai d'expiration' (0:12:00:00), 'Réessayer' (0), 'Intervalle de nouvelle tentative(s)' (30), 'Sortie sécurisée' (unchecked), and 'Entrée sécurisée' (unchecked).

On utilise le partitionnement unique pour avoir un seul fichier csv en sortie :

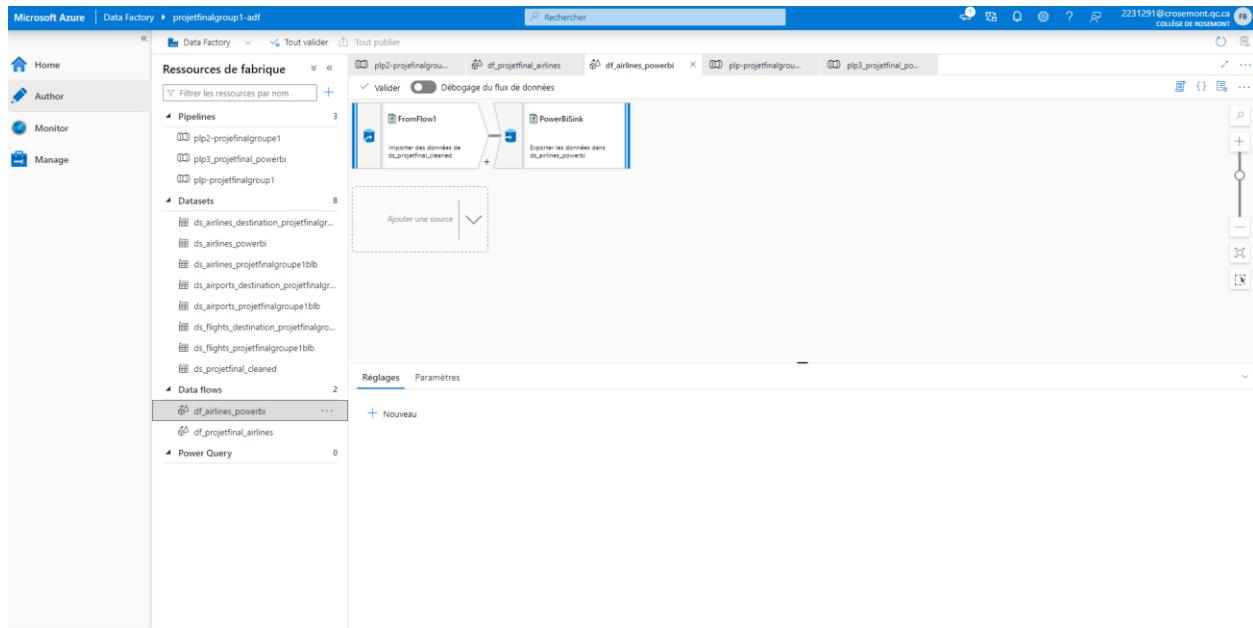
AIRPORT SOURCE

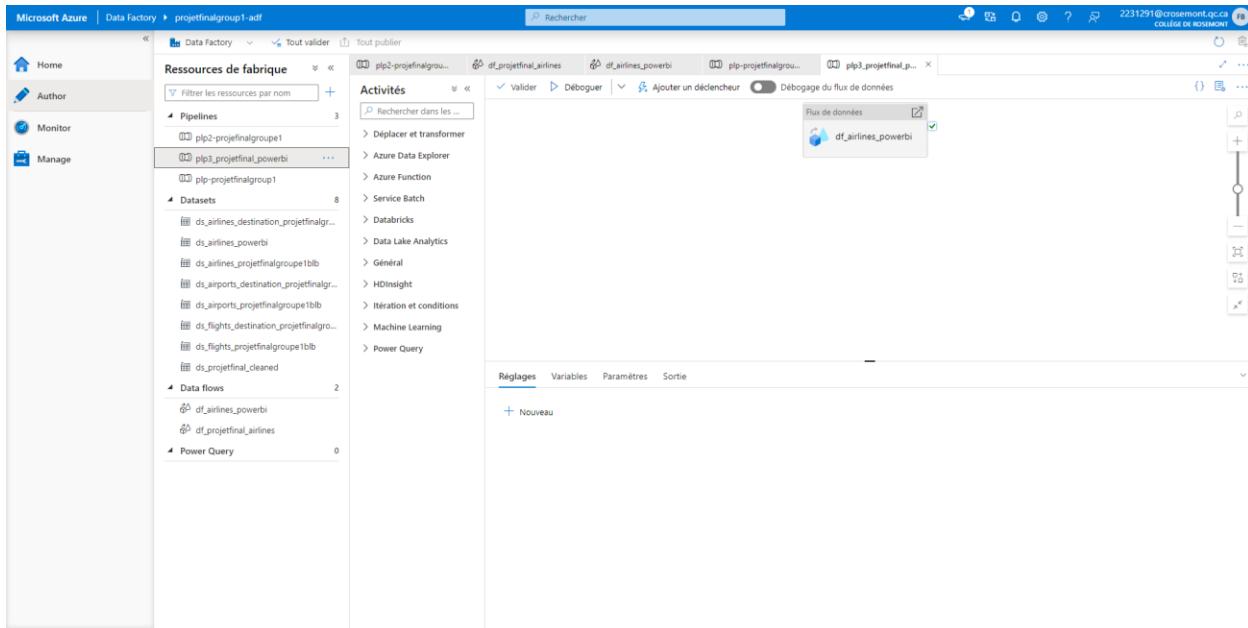
Récepteur Paramètres Erreurs Mappage Optimiser Inspector Aperçu des données ●

Ce récepteur a actuellement l'option Une seule partition définie sur Optimiser, ce qui va rallonger l'exécution de votre flux de données. Le paramètre Optimiser est recommandé pour les flux de données volumineux.

Option de partition \*  Utiliser le partitionnement actuel  Partition unique  Définir le partitionnement

Puis le pipeline final pour transférer dans le conteneur PowerBI :





Programmer des traitements automatiques (Datafactory) et transmettre les données résultantes vers un système distribué :

Création des déclencheurs sur les 3 pipelines de données

Puisque les analyses doivent être faites quotidiennement, on propose de créer des déclencheurs d'ingestions des données chaque jour à partir de 22H00 pour que les résultats soient exploitables le lendemain matin.

Mais, vu que les conteneurs dépendent les uns des autres, on doit commencer par un déclencheur d'ingestion (à 22H00), puis un déclencheur du premier dataflow (à 23H00), puis un troisième déclencheur du dernier dataflow (à 00H00), dans cet ordre :

### Modifier le déclencheur

**Nom \***  
triggerPipelineIngestion

**Description**

**Type \***  
ScheduleTrigger

**Date de début \*** ⓘ  
10/22/22 22:00:00

**Fuseau horaire \*** ⓘ  
Est (États-Unis et Canada) (UTC-5)

Ce fuseau horaire observe l'heure d'été. Le déclencheur se règle automatiquement avec une heure de différence.

**Périodicité \*** ⓘ  
Chaque 1 Jour(s)

**Option de récurrence avancée**

Exécuter à ces horaires  
Heures: \_\_\_\_\_  
Minutes: \_\_\_\_\_

Planifier les horaires d'exécution  
22:00  
 Spécifier une date de fin

**Annotations**  
+ Nouveau

**État** ⓘ  
 Démarré  Arrêté

**OK** **Annuler**

The screenshot shows the Microsoft Azure Data Factory pipeline editor interface. On the left, there's a navigation sidebar with 'Home', 'Author', 'Monitor', and 'Manage' sections. The main area displays a tree view of resources under 'Ressources de fabrique'. Under 'Pipelines', three pipelines are listed: 'plp2-projetfinalgroup1', 'df\_projetfinal\_airlines', and 'plp3\_projetfinal\_group1'. Under 'Datasets', several datasets are listed, including 'ds\_airlines\_destination\_projetfinalgr...', 'ds\_airlines\_powerbi', and 'ds\_airports\_projetfinalgroupe1bb'. Under 'Data flows', two flows are listed: 'df\_airlines\_powerbi' and 'df\_projetfinal\_airlines'. Under 'Power Query', one item is listed: 'ds\_projetfinal\_cleaned'. On the right, a detailed configuration pane for the selected pipeline 'df\_projetfinal\_airlines' is open, showing tabs for 'Général', 'Paramètres', 'Règles', and 'Propriétés de l'utilisateur'. The 'Général' tab contains fields for 'Nom' (set to 'df\_projetfinal\_airlines'), 'Description', 'Décalage d'expiration' (set to '0:12:00:00'), 'Réessayer' (set to '0'), 'Intervalle de nouvelle tentative(s)' (set to '30'), and checkboxes for 'Sortie sécurisée' and 'Entrée sécurisée'. A small 'Flux de données' (data flow) window is also visible in the top right corner.

### Modifier le déclencheur

**Nom \***  
triggerPipeline2ToGen2

**Description**

**Type \***  
ScheduleTrigger

**Date de début \*** ⓘ  
10/22/22 23:00:00

**Fuseau horaire \*** ⓘ  
Est (États-Unis et Canada) (UTC-5)

ⓘ Ce fuseau horaire observe l'heure d'été. Le déclencheur se règle automatiquement avec une heure de différence.

**Périodicité \*** ⓘ  
Chaque 1 Jour(s)

**Option de récurrence avancée**

Exécuter à ces horaires ⓘ

Heures  
Minutes

Planifier les horaires d'exécution  
23:00

Spécifier une date de fin

**Annotations**

[+ Nouveau](#)

**État** ⓘ  
 Démarré  Arrêté

**OK** **Annuler**

Nouveau déclencheur

**Nom \***  
triggerPipeline3ToPowerBi

**Description**

**Type \***  
Planifier

**Date de début \***  
10/22/22 23:59:59

**Fuseau horaire \***  
Est (États-Unis et Canada) (UTC-5)

ⓘ Ce fuseau horaire observe l'heure d'été. Le déclencheur se règle automatiquement avec une heure de différence.

**PéIODICITÉ \***  
Chaque 1 Jour(s)

✓ Option de récurrence avancée

Exécuter à ces horaires  
Heures :   
Minutes :

Planifier les horaires d'exécution  
23:59  
 Spécifier une date de fin

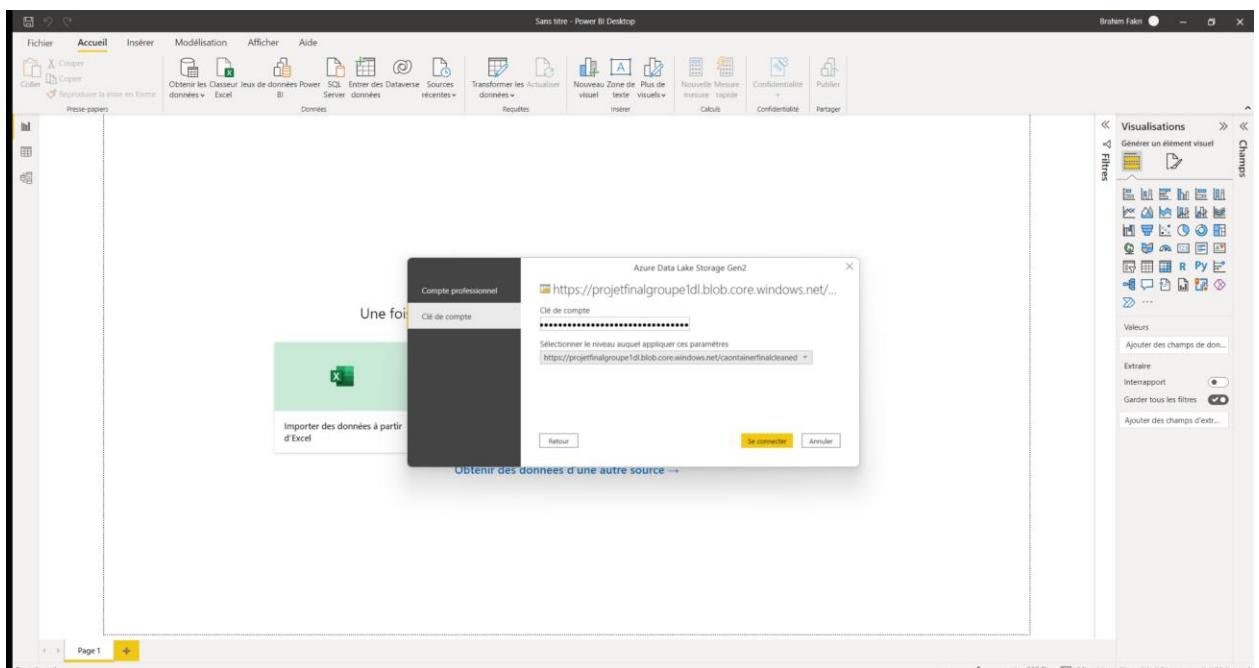
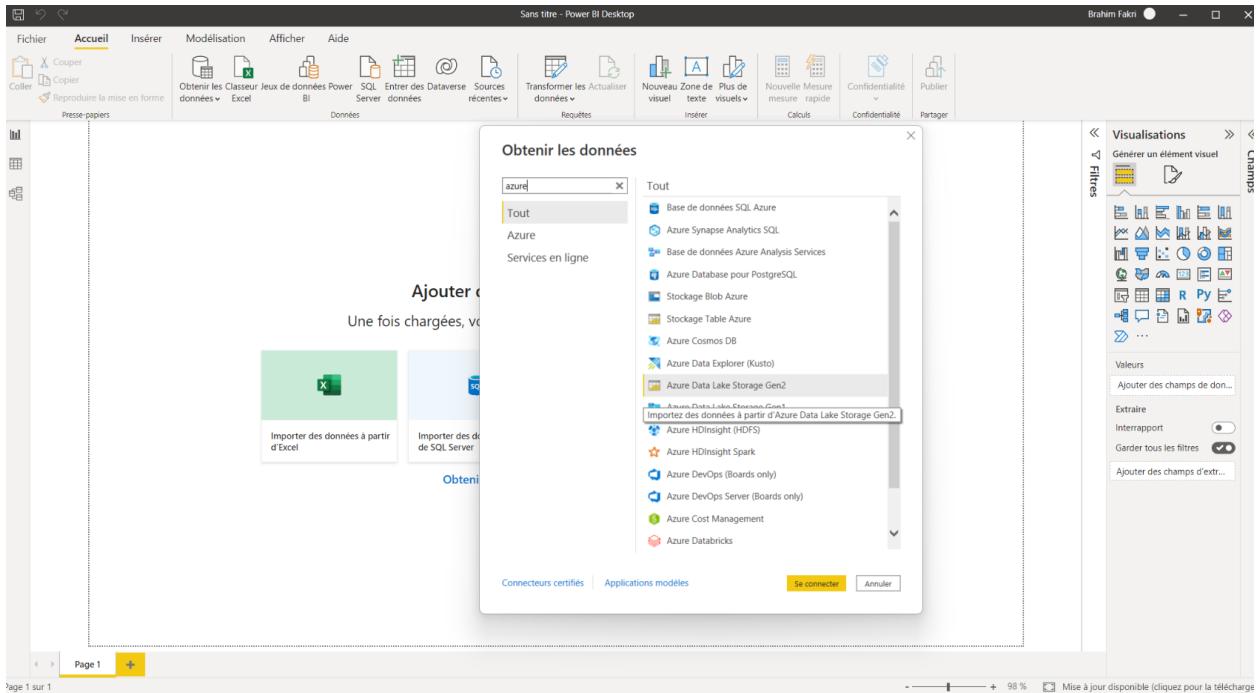
**Annotations**  
+ Nouveau

**Démarrer le déclencheur**  
 Démarrer le déclencheur lors de la création

**OK** **Annuler**

# Visualisation des données

Connexion à power BI :



## Partage dans GITHUB

```
ns 2015 (main)
$ git push -u origin main
Enumerating objects: 4, done.
Counting objects: 100% (4/4), done.
Delta compression using up to 8 threads
Compressing objects: 100% (3/3), done.
Writing objects: 100% (4/4), 179.52 MiB | 1.33 MiB/s, done.
Total 4 (delta 0), reused 0 (delta 0), pack-reused 0
remote: error: Trace: d239e5ac2417d858dd414fbb9388563dff80dc89b00982c5100a66e211b323b
remote: error: See http://git.io/iEPt8g for more information.
remote: error: File part-merged.csv is 876.16 MB; this exceeds GitHub's file size limit of 100.00 MB
remote: error: GH001: Large files detected. You may want to try Git Large File Storage - https://git-lfs.github.com.
To https://github.com/2231291/Azure-datafactory-project.git
 ! [remote rejected] main -> main (pre-receive hook declined)
error: failed to push some refs to 'https://github.com/2231291/Azure-datafactory-project.git'
```

Pas possible de commiter plus que 100 MB ! On va donc inclure juste le lien vers les données d'origine, mais pas les données csv elles même. On va aussi inclure une capture d'écran du fichier résultant : part-merged.csv :

The screenshot shows a Microsoft Excel spreadsheet titled "part-merged.csv - Read-Only". The table has 37 rows of data, starting from A1. The columns represent various flight metrics and operational status. The data covers flights from January 1st to January 31st, 2015, across different airlines and destinations. The last few rows show summary statistics for the month.

MONTH	DAY_OF_MONTH	DAY_OF_YEAR	AIRLINE	C-AirportCo	FLIGHT	DESTINATION	SCHEDULED_DEPARTURE	SCHEDULED_ARRIVAL	ELAPSED_AIR_TIME	DISTANCE_WHEELS	C_SCHEDULED_ARRIVAL	ARRIVAL	DIVERTED	CANCELLED	AIR_SYSTEM_SECURITY	AIRLINE_C	LATE_AIR	WEATHER	STATE
2	January	1	AA-Alaska	ANC-Ted I	88 SEA	5	2554	-18	205	194	169	1448	404	430	-22	FALSE	FALSE	AK	
3	January	Thursday	1 AA-Ameri	LAX-Los A	2336 PBI	10	2	-8	280	279	263	2330	737	750	741	-9	FALSE	FALSE	CA
4	January	Thursday	1 US-US Air	SFO-San F	840 CLT	20	18	-2	286	293	266	2296	800	806	811	5	FALSE	FALSE	CA
5	January	Thursday	1 AA-Ameri	LAX-Los A	258 MIA	20	15	-5	285	281	258	2342	748	805	756	-9	FALSE	FALSE	CA
6	January	Thursday	1 AS-Alaska	SEA-Seatl	135 ANC	25	24	-1	235	215	199	1448	254	320	259	-21	FALSE	FALSE	WA
7	January	Thursday	1 DL-Delta	J-SFO-San F	806 MSP	25	20	-5	217	230	206	1589	604	602	610	8	FALSE	FALSE	CA
8	January	Thursday	1 NK-Spirit	SEA-McCa	612 MSP	25	19	-6	181	170	154	1299	504	526	509	-17	FALSE	FALSE	NV
9	January	Thursday	1 US-US Air	LAX-Los A	201 CLT	30	44	14	272	245	250	2190	740	803	793	-10	FALSE	FALSE	CA
10	January	Thursday	1 AA-Ameri	SFO-San F	1112 DFW	30	19	-11	195	193	179	1464	529	545	532	-13	FALSE	FALSE	CA
11	January	Thursday	1 DL-Delta	J-LAX-McCa	1173 ATL	30	33	3	221	203	188	1747	651	711	656	-15	FALSE	FALSE	NV
12	January	Thursday	1 DL-Delta	J-DEN-Denv	2336 ATL	30	24	-6	173	149	133	1199	449	523	453	-30	FALSE	FALSE	CO
13	January	Thursday	1 AA-Ameri	LAS-McCa	1674 MIA	35	27	-8	268	266	238	2174	746	803	753	-10	FALSE	FALSE	NV
14	January	Thursday	1 DL-Delta	J-LAX-Los A	1434 MSP	35	35	0	214	210	188	1535	601	609	605	-4	FALSE	FALSE	CA
15	January	Thursday	1 DL-Delta	J-SLC-Salt Li	2324 ATL	40	34	-6	215	199	176	1590	548	615	553	-22	FALSE	FALSE	UT
16	January	Thursday	1 DL-Delta	J-SEA-Seatl	2440 MSP	40	39	-1	189	198	166	1399	553	549	557	8	FALSE	FALSE	WA
17	January	Thursday	1 AA-Ameri	SEA-Seatl	451 SEA	45	41	-4	204	172	147	1400	551	555	545	-14	FALSE	FALSE	AK
18	January	Thursday	1 DL-Delta	J-ANC-Ted I	1580 SEA	45	31	-4	210	200	171	1448	447	515	451	-24	FALSE	FALSE	AK
19	January	Thursday	1 UA-Untitec	SFO-San F	1187 IAH	48	42	-6	218	217	199	1639	612	626	619	-7	FALSE	FALSE	CA
20	January	Thursday	1 AS-Alaska	SEA-Seatl	122 FOX	50	46	-4	215	201	187	1542	504	525	507	-18	FALSE	FALSE	AK
21	January	Thursday	1 DL-Delta	J-PDX-Portl	1670 MSP	50	45	-5	193	186	171	1426	545	603	551	-12	FALSE	FALSE	OR
22	January	Thursday	1 NK-Spirit	SEA-McCa	520 MCI	55	120	25	162	143	138	1139	539	537	543	6	FALSE	FALSE	NV
23	January	Thursday	1 AA-Ameri	SEA-Seatl	371 MIA	100	52	-8	338	347	311	2724	933	938	939	1	FALSE	FALSE	WA
24	January	Thursday	1 NK-Spirit	J-LAS-McCa	214 DFW	102	102	-1	147	147	128	1055	523	530	529	-1	FALSE	FALSE	NV
25	January	Thursday	1 AA-Ameri	LAX-Los A	115 MIA	102	103	-2	286	276	255	2342	832	851	839	-12	FALSE	FALSE	CA
26	January	Thursday	1 AA-Ameri	SEA-Seatl	1450 DFW	103	20	-3	180	161	150	1400	541	550	540	-22	FALSE	FALSE	NV
27	January	Thursday	1 UA-Untitec	LAX-Los A	1345 IAH	115	112	-3	183	175	156	1379	559	618	607	-11	FALSE	FALSE	CA
28	January	Thursday	1 AS-Alaska	J-Fairbz	130 SEA	115	107	-8	213	218	186	1533	538	548	545	-3	FALSE	FALSE	AK
29	January	Thursday	1 NK-Spirit	MSP-Minr	597 FLL	115	127	12	207	220	166	1487	527	542	607	-25	FALSE	FALSE	CO
30	January	Thursday	1 US-US Air	LAS-McCa	413 CLT	120	110	-10	245	224	205	1916	747	825	754	-31	FALSE	FALSE	NV
31	January	Thursday	1 AA-Ameri	Den-Denn	2392 MIA	120	141	21	227	208	188	1709	701	707	709	2	FALSE	FALSE	CO
32	January	Thursday	1 NK-Spirit	J-PHX-Phnx	165 ORD	125	237	72	204	175	156	1440	622	549	632	-43	FALSE	FALSE	43 0 0 0 0 AZ
33	January	Thursday	1 AA-Ameri	PHL-Philadeph	2211 MIA	127	116	-11	238	234	217	1972	703	726	710	-16	FALSE	FALSE	AZ
34	January	Thursday	1 AA-Ameri	ANC-Ted I	136 SEA	135	106	-20	205	180	160	1400	600	600	600	-14	FALSE	FALSE	AK
35	January	Thursday	1 DL-Delta	J-FLC-Fatl	1471 ATL	140	134	-6	215	231	182	1599	719	751	725	-10	FALSE	FALSE	UT
36	January	Thursday	1 NK-Spirit	J-LAS-McCa	298 IAH	144	140	-4	170	148	1222	618	634	630	-4	FALSE	FALSE	NV	
37	January	Thursday	1 HA-Hawai	J-LAS-McCa	17 HNL	145	145	0	370	385	361	2762	602	555	610	15	FALSE	FALSE	0 0 NV
38	January	Thursday	1 US-US Air	AN-ANC-Ted I	617 PHX	152	143	-9	323	322	298	2552	902	915	905	-10	FALSE	FALSE	AK

## Références

- *ETL (Extract, Transform, Load) :*  
<https://www.ibm.com/cloud/learn/etl>
- *What is ETL (extract transform load):*  
<https://www.informatica.com/ca/resources/articles/what-is-etl.html>
- *Azure documentation*  
<https://docs.microsoft.com/en-us/azure>
- *SQL Server Integration Services*  
<https://docs.microsoft.com/fr-ca/sql/integration-services/sql-server-integration-services?view=sql-server-ver16>
- *Adam Marczak - Azure for Everyone*  
<https://www.youtube.com/c/Azure4Everyone/videos>