

Azure DataFactory Project

Airlines Delays and Cancellation 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	37

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 navigation links like Home, Competitions, Datasets, and Learn. 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 a 'Context' paragraph explaining the dataset's purpose and source. The 'Acknowledgements' section notes that the data was collected by the DOT's Bureau of Transportation Statistics. On the right side, 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 Ide
2015	2015	1	1	WN	22%
2015	1	12	12	DL	15%
2015	1	31	31	Other (3681343)	63%
2015	1	7	7	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



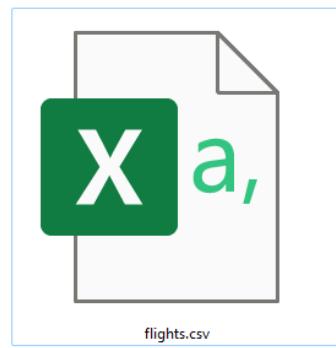
Search data airlines



airlines.csv



airports.csv



flights.csv

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

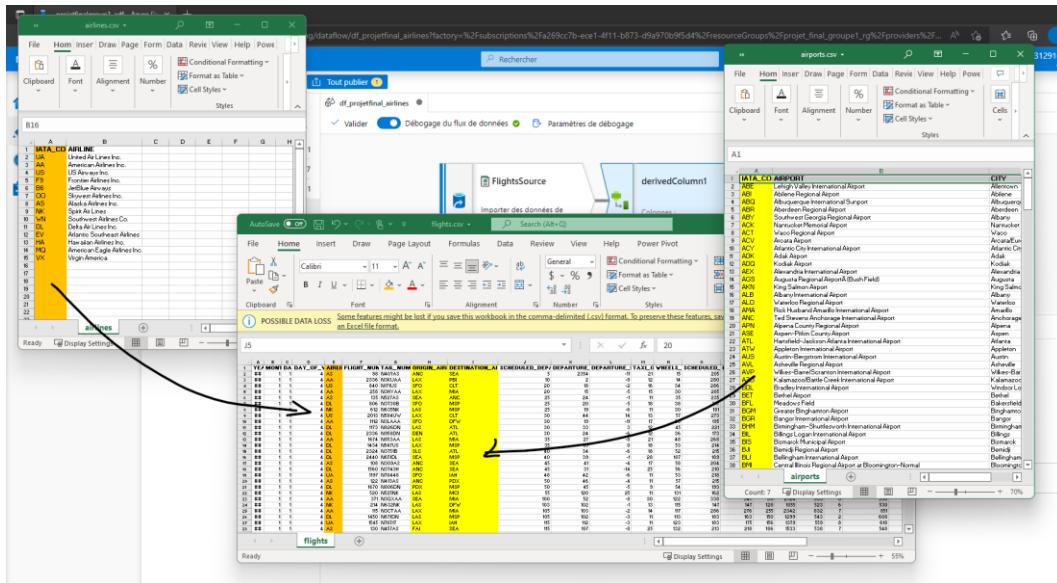
	AIRLINE
1	UA United Air Lines Inc.
2	AA American Airlines Inc.
3	US US Airways Inc.
4	F9 Frontier Airlines Inc.
5	B6 JetBlue Airways
6	OO Skywest Airlines Inc.
7	AS Alaska Airlines Inc.
8	NK Spirit Air Lines
9	WN Southwest Airlines Co.
10	DL Delta Air Lines Inc.
11	EV Atlantic Southeast Airlines
12	HA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

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

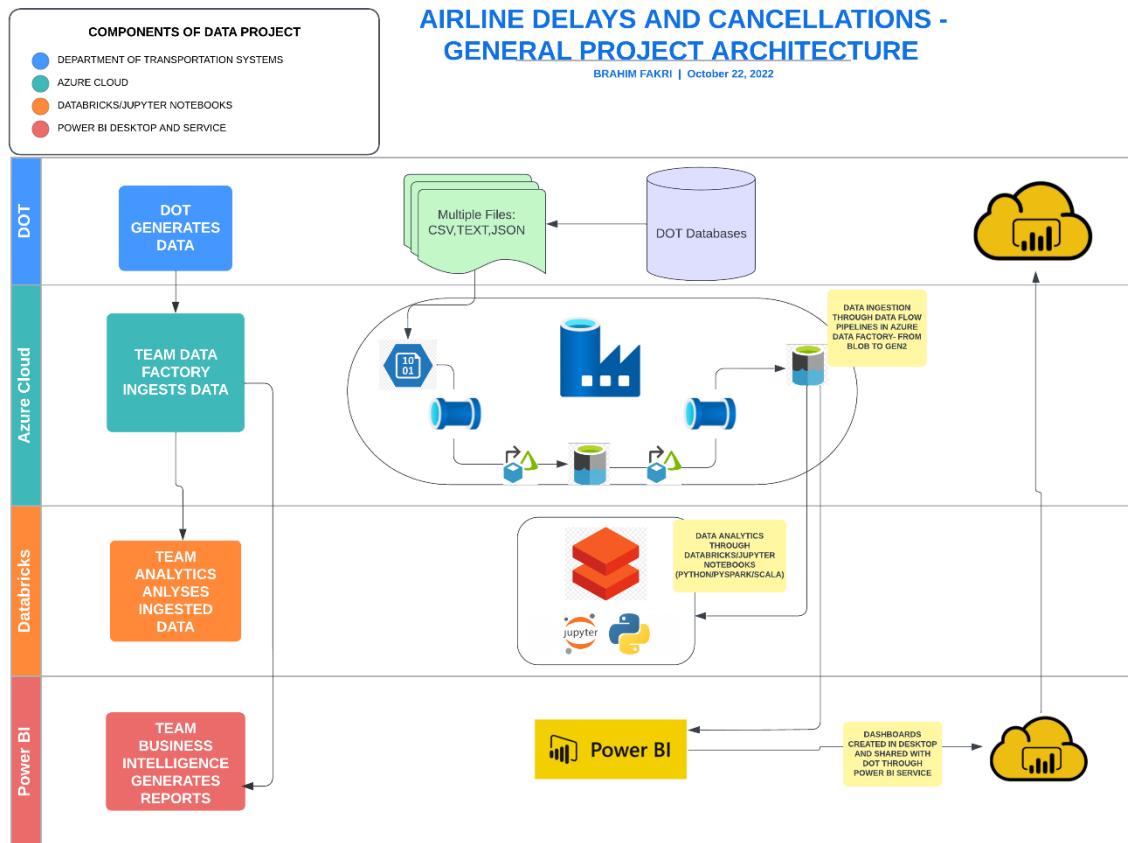
	AIRPORT	CITY	STATE	COUNTRY	LATITUDE	LONGITUDE
1	ABE Lehigh Valley International Airport	Allentown	PA	USA	40.65236	-75.4404
2	ABI Abilene Regional Airport	Abilene	TX	USA	32.41132	-99.6819
3	ABQ Albuquerque International Sunport	Albuquerque	NM	USA	35.04022	-106.60919
4	ABR Aberdeen Regional Airport	Aberdeen	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
10	ADK Adak Airport	Adak	AK	USA	51.87796	-176.64603
11	ADQ Kodiak Airport	Kodiak	AK	USA	57.74997	-152.49386
12	AEX Alexandria International Airport	Alexandria	LA	USA	31.32737	-92.54856
13	AGS Augusta Regional Airport (Bush Field)	Augusta	GA	USA	33.36996	-81.9645
14	AKN King Salmon Airport	King Salmon	AK	USA	58.6768	-156.64922
15	ALB Albany International Airport	Albany	NY	USA	42.74812	-73.80298
16	ALO Waterloo Regional Airport	Waterloo	IA	USA	42.55708	-92.40034
17	AMA Rick Husband Amarillo International Airport	Amarillo	TX	USA	35.21937	-101.70593
18	ANC Ted Stevens Anchorage International Airport	Anchorage	AK	USA	61.17432	-149.99619
19	APN Alpena County Regional Airport	Alpena	MI	USA	45.07807	-83.56029
20	ASE Aspen-Pitkin County Airport	Aspen	CO	USA	39.22316	-106.86885
21	ATL Hartsfield-Jackson Atlanta International Airport	Atlanta	GA	USA	33.64044	-84.42694
22	ATW Appleton International Airport	Appleton	WI	USA	44.25741	-88.51948
23	AUS Austin-Bergstrom International Airport	Austin	TX	USA	30.19453	-97.66987

Ces mêmes clés primaires apparaissent dans le fichier flights comme clés secondaires, mais avec des noms différents : **AIRLINE**, **ORIGIN_AIRPORT**, **DESTINATION_AIRPORT**

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	YEAR	MONTH	DAY	DAY_OF_WEEK	AIRLINE	FLIGHT_NUMBER	TAIL_NUMBER	ORIGIN_AIRPORT	DESTINATION_AIRPORT	SCHEDULED_DEPARTURE	DEPARTURE_TIME	DEPARTURE_DELAY	TAXI_OUT	WHEELS	CRS_ELAPSED_TIME
2	2015	1	1	4	AS	98	N407AS	ANC	SEA		2354	-11	21		145
3	2015	1	1	4	AA	2336	N3KUAA	LAX	PBI		10	-8	12		145
4	2015	1	1	4	US	840	N171US	SFO	CLT		20	-2	16		145
5	2015	1	1	4	AA	258	N3HYAA	LAX	MIA		20	-5	15		145
6	2015	1	1	4	AS	135	N527AS	SEA	ANC		25	-1	11		145
7	2015	1	1	4	DL	806	N3730B	SFO	MSP		25	-5	18		145
8	2015	1	1	4	NK	612	N635NK	LAS	MSP		25	-6	11		145
9	2015	1	1	4	US	2013	N584UW	LAX	CLT		30	14	13		145
10	2015	1	1	4	AA	1112	N3LAAA	SFO	DFW		30	-11	17		145
11	2015	1	1	4	DL	1173	N826DN	LAS	ATL		30	3	12		145
12	2015	1	1	4	DL	2336	N958DN	DEN	ATL		30	-6	12		145
13	2015	1	1	4	AA	1674	N853AA	LAS	MIA		35	-8	21		145
14	2015	1	1	4	DL	1434	N547US	LAX	MSP		35	0	18		145
15	2015	1	1	4	DL	2324	N3751B	SLC	ATL		40	-6	18		145
16	2015	1	1	4	DL	2440	N651DL	SEA	MSP		40	-1	28		145
17	2015	1	1	4	AS	108	N309AS	ANC	SEA		45	-4	17		145
18	2015	1	1	4	DL	1560	N3743H	ANC	SEA		45	-14	25		145
19	2015	1	1	4	UA	1197	N7844B	SFO	IAH		48	-6	11		145
20	2015	1	1	4	AS	122	N413AS	ANC	PDX		50	-4	11		145
21	2015	1	1	4	DL	1670	N806DN	PDX	MSP		50	-5	9		145
22	2015	1	1	4	NK	520	N525NK	LAS	MCI		55	25	11		145
23	2015	1	1	4	AA	371	N3GKAA	SEA	MIA		100	-8	30		145
24	2015	1	1	4	NK	214	N632NK	LAS	DFW		103	-1	13		145
25	2015	1	1	4	AA	115	N3CTAA	LAX	MIA		105	-2	14		145
26	2015	1	1	4	DL	1450	N671DN	LAS	MSP		105	-3	11		145

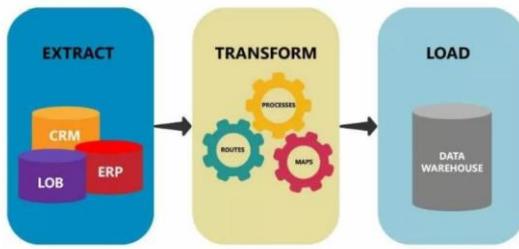


Présentation du système infonuagique Big data, de l'architecture générale et des outils utilisés et qui seront utilisés dans les prochaines étapes de la consultation



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.

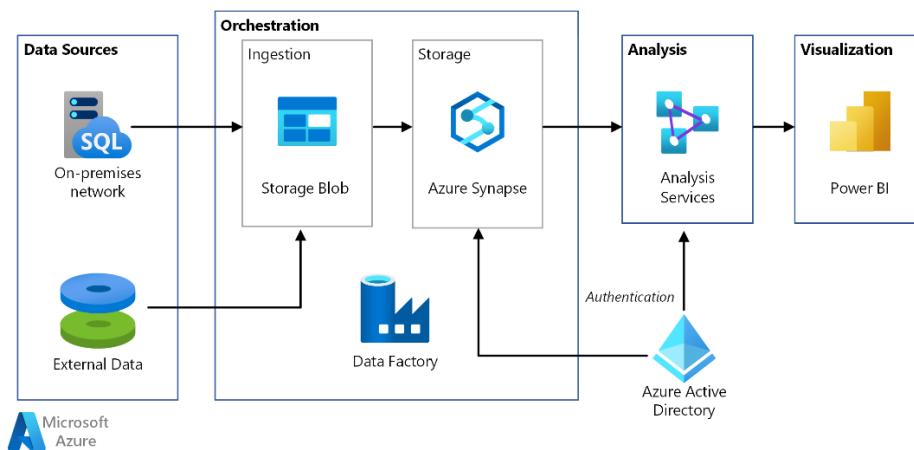


Les entrepôts de données en étant la destination des données dans l'architecture ETL, jouent un rôle primordial dans ce processus. Sans les entrepôts de données, l'accès aux données direct dans les bases de données traditionnel n'est ni productif ni sécuritaire. En effet travailler directement sur les 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.

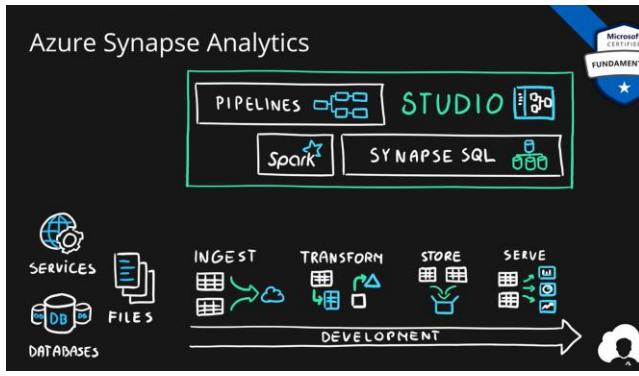
L'exemple le plus évident est le domaine de l'intelligence d'affaire. Les entrepôts de données permettent aux analystes BI de faire des analyses qui permettent une prise de décision bien informée et plus rapide, car ils ont entre les mains les données les plus pertinentes qu'ils peuvent manipuler et 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
Groupe de ressources	projet_final_groupe1_rg
Compte de stockage BLOB	projefinalgroupe1sa
Containers BLOB	projefinalgroupe1blob
Compte de stockage GEN2	projefinalgroupe1dl
Containers GEN2	containerprojefinalgroupe1dl caontainerfinalcleaned powerbicontainer
Fabrique de données	projefinalgroup1-adf

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

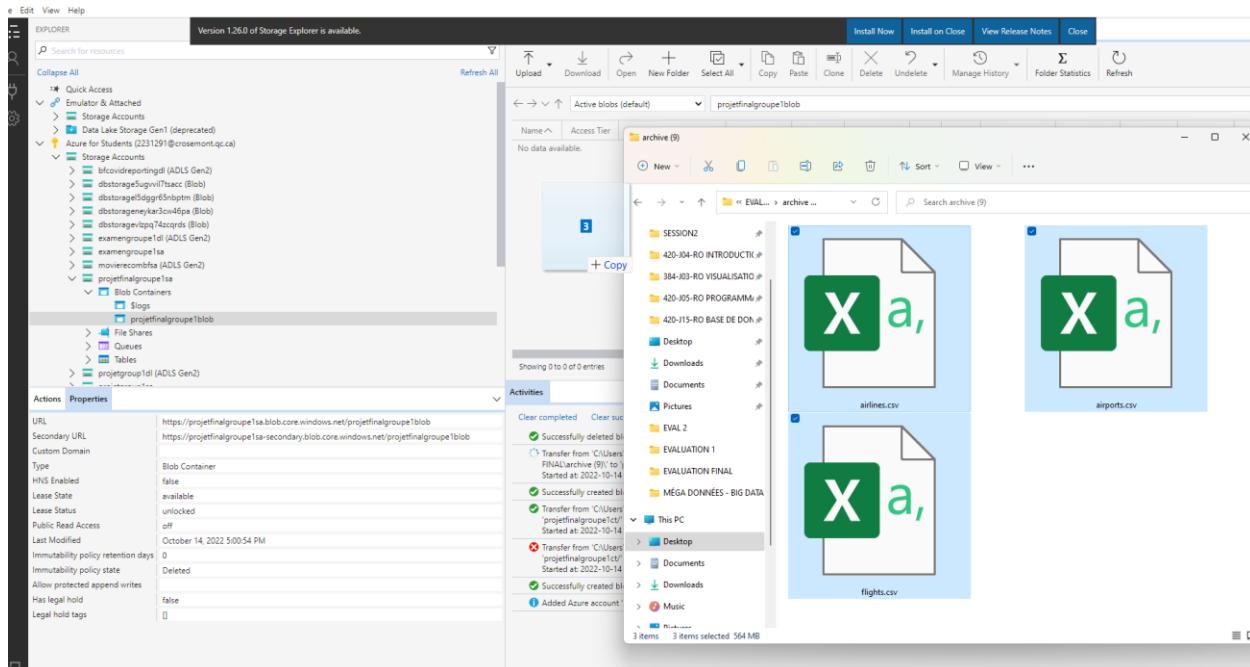
Extraction et ingestion des données

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

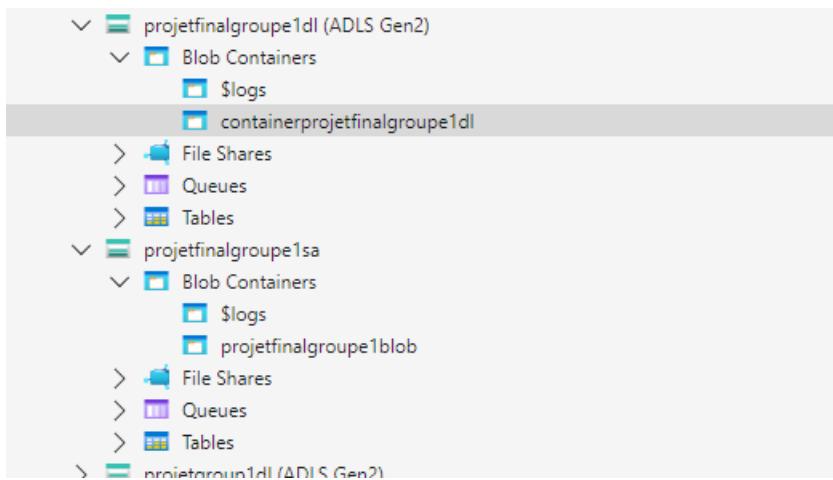
Espace de noms hiérarchique	Activé
Niveau d'accès par défaut	Hot
Accès public aux objets blob	Activé
Suppression réversible d'objet blob	Activé (7 jours)
Suppression réversible de conteneur	Activé (7 jours)
Gestion des versions	Désactivé
Flux de modification	Désactivé
NFS v3	Désactivé
SFTP (préversion)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

- Vue d'ensemble**: Shows the factory's status as 'Succeeded'.
- Bases**: Details about the resource group ('projet_final_groupe1_rg'), location ('East US'), and subscription ('Azure for Students').
- Démarrer**: Buttons to 'Ouvrir Azure Data Factory Studio' and 'Lire la documentation'.
- Supervision**: Two charts showing 'PipelineRuns' and 'ActivityRuns' counts over time.

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

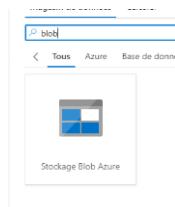
Ressources récentes

Aucun élément à afficher

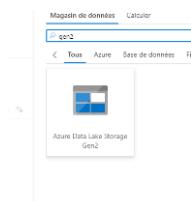
Vos dernières ressources ouvertes s'affichent ici.

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



Microsoft Azure | Data Factory > projetfinalgroupe1-adf

Services liés

Le service lié définit les informations de connexion à un magasin de données ou un calcul. [En savoir plus](#)

+ Nouveau

Filtrer par nom Annotations : Tout

Nom	Type	Associé	Annotations
ls_projetfinalgroupe1_adls	Azure Data Lake Storage Gen2	0	
ls_projetfinalgroupe1_bblob	Stockage Blob Azure	0	

Créer six jeux de données:

(côté source correspond au Blob storage)

ds_airlines_projetfinalgroupe1blob

ds_airports_projetfinalgroupe1blob

ds_flights_projetfinalgroupe1blob

Microsoft Azure | Data Factory > projetfinalgroupe1-adf

Ressources de fabrique

Filter les ressources par nom

- Pipelines 0
- Datasets 3
 - ds_airlines_projetfinalgroupe1blob
 - ds_airports_projetfinalgroupe1blob
 - ds_flights_projetfinalgroupe1blob
- Data flows 0
- Power Query 0

Propriétés

Général Associé

Nom * ds_flights_projetfinalgroupe1blob

Description

Annotations

+ Nouveau

Connexion Schéma Réglages

Service lié * ls_projetfinalgroupe1_bblob

Chemin d'accès au fichier * projetfinalgroupe1blob / / flights.csv

Type de compression Aucun

Séparateur de colonne Virgule (,)

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

Encodage Par défaut(UTF-8)

Caractère d'échappement Barre oblique inverse (\)

Guillemet Guillemet double ("")

Première ligne comme en-tête

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	AIRLINE
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	2338	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. The top operation is 'Copy data airlines', the middle is 'Copy data airports', and the bottom is 'Copy data flights'. Each operation has its own configuration pane on the right, showing settings like 'Jeu de données de récepteur' (Target dataset), 'Nombre maximal de connexions simultanées' (Max concurrent connections), and 'Extension du fichier' (File extension). The pipeline editor interface includes tabs for General, Source, Receiver, Mapping, Parameters, and User Properties.

On lance le débogage :

The screenshot shows the Microsoft Azure Data Factory pipeline editor in debug mode. The interface is similar to the previous one, but the pipeline status indicates that all three parallel 'Copy data' operations have completed successfully. The 'Sortie' (Output) tab is selected, displaying a table of execution details. The table includes columns for Name, Type, Start time, Duration, Status, and Runtime. All three operations show a green checkmark indicating success. The table also includes a link 'Voir la consommation de l'exécution de débogage' (View debug execution consumption).

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)
 - dbstoragefeykaricw4dp6 (Blob)
 - dbstoragevpq74ccordrs (Blob)
 - examengroupe1dl (ADLS Gen2)
 - examengroupe1sa
 - movierecombs (ADLS Gen2)
 - blob Containers
 - Slogs
 - containerprojetfinalgroup1dl
 - File Shares
 - Queues
 - Tables
 - projectfinalgroup1sa
 - projectgroup1dl (ADLS Gen2)
 - projectgroup1sa
 - rovcoideReportingblob
- 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 :

Microsoft Azure | Data Factory > projetfinalgroup1-adf

Ressources de fabrique

- Pipelines
- Datasets
- Data flows
- Power Query

Tout publier | Valider | Débogage du flux de données | Paramètres de débogage

Definir les propriétés

Nom: ds_projetfinal_cleaned

Service lié: ls_projetfinalgroup1_adls

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

Première ligne comme en-tête: checked

Importer un schéma

- À partir d'une connexion/un magasin
- À partir d'un exemple de fichier
- Aucun

Avancé

Récepteur

Paramètres

Erreurs

Mappage

Optimiser

Inspecter

Aperçu des données

Flux entrant: FlightSource

Type de récepteur: Jeu de données

Jeu de données: Sélectionner... | Nouveau

Options

Autoriser la dérive de schéma

Valider le schéma

OK | **Précédent** | **Annuler**

Microsoft Azure | Data Factory > projetfinalgroup1-adf

Ressources de fabrique

- Pipelines
- Datasets
- Data flows
- ds_projetfinal_cleaned
- df_projetfinal_airlines
- Power Query

Tout publier | Valider | Débogage du flux de données | Paramètres de débogage

Propriétés

Général

Nom: df_projetfinal_airlines

Description

Paramètres de la source

Nom du flux de sortie: FlightSource

Description: Importer des données de ds_flights_destination_projetfinalgroup1adls

Type de source: Jeu de données

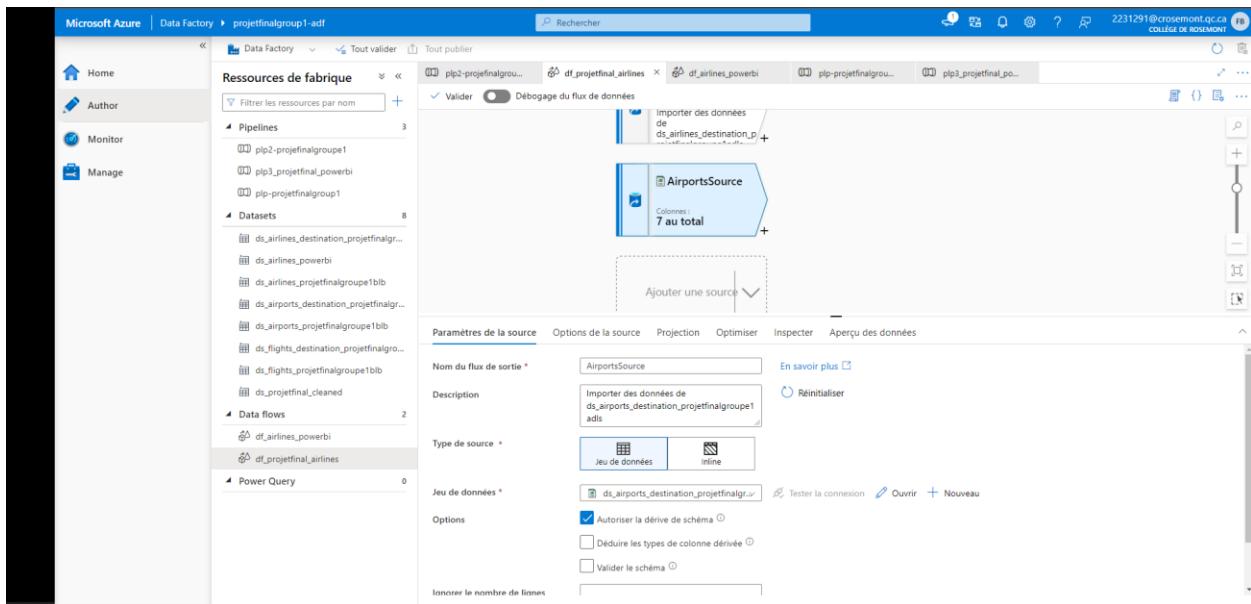
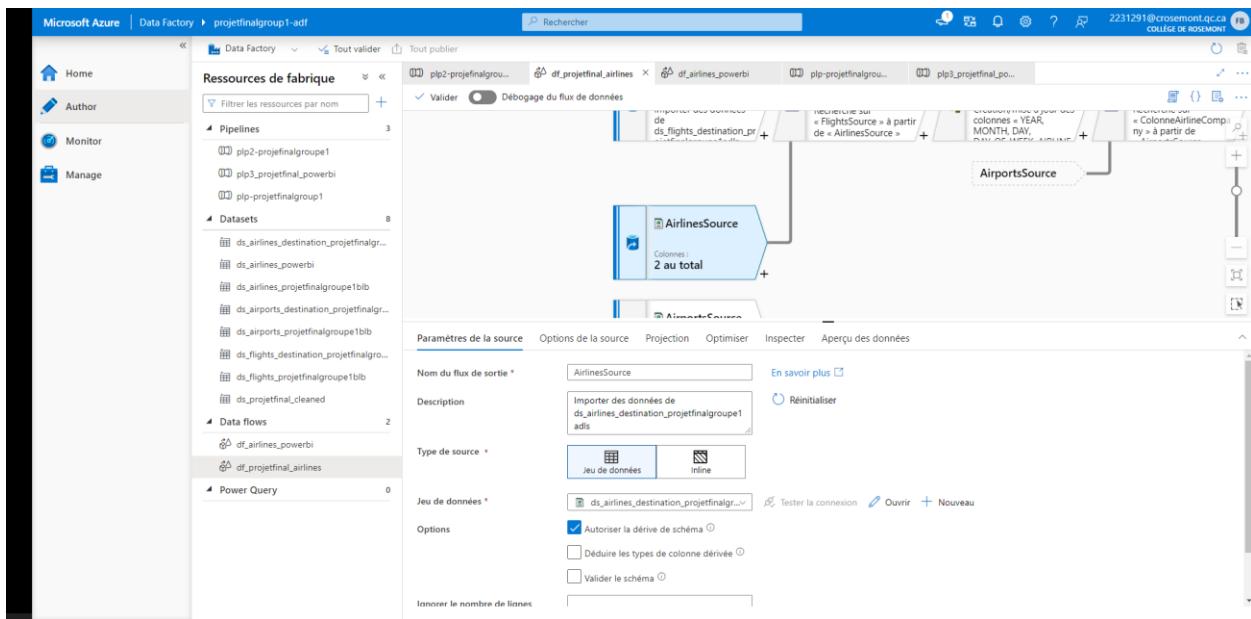
Jeu de données: ds_flight_destination_projetfinalgroup1adls

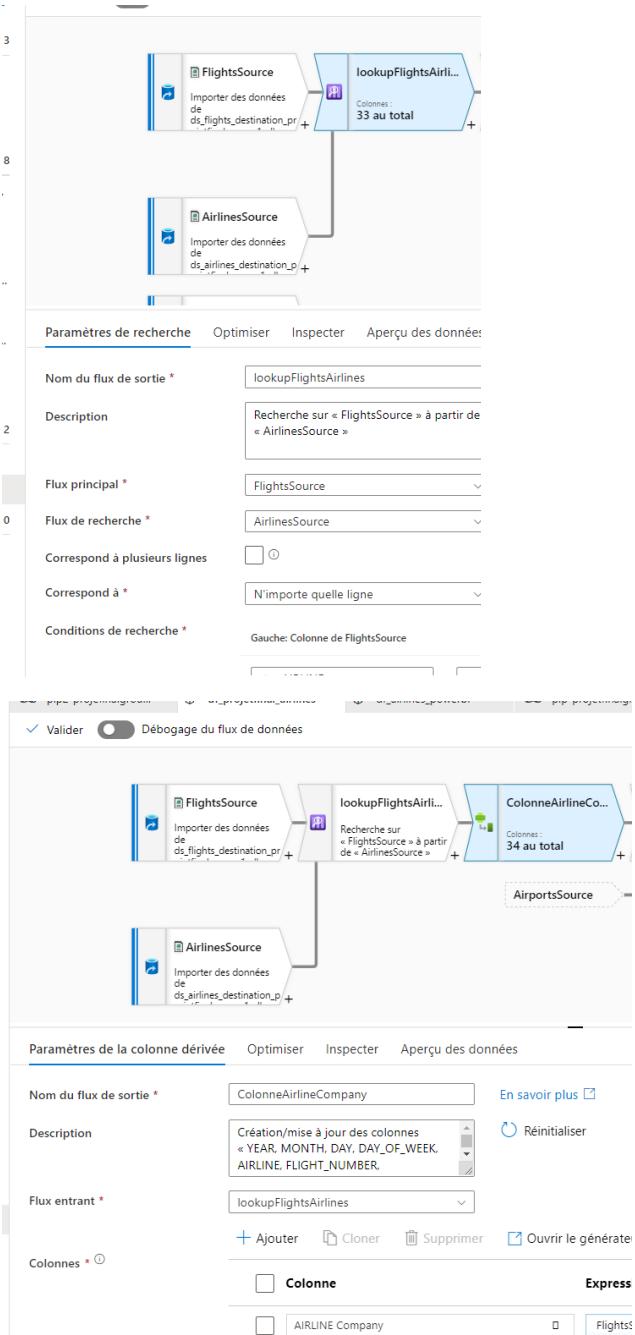
Options

Autoriser la dérive de schéma

Déduire les types de colonne dérivée

Utiliser la tâche à





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

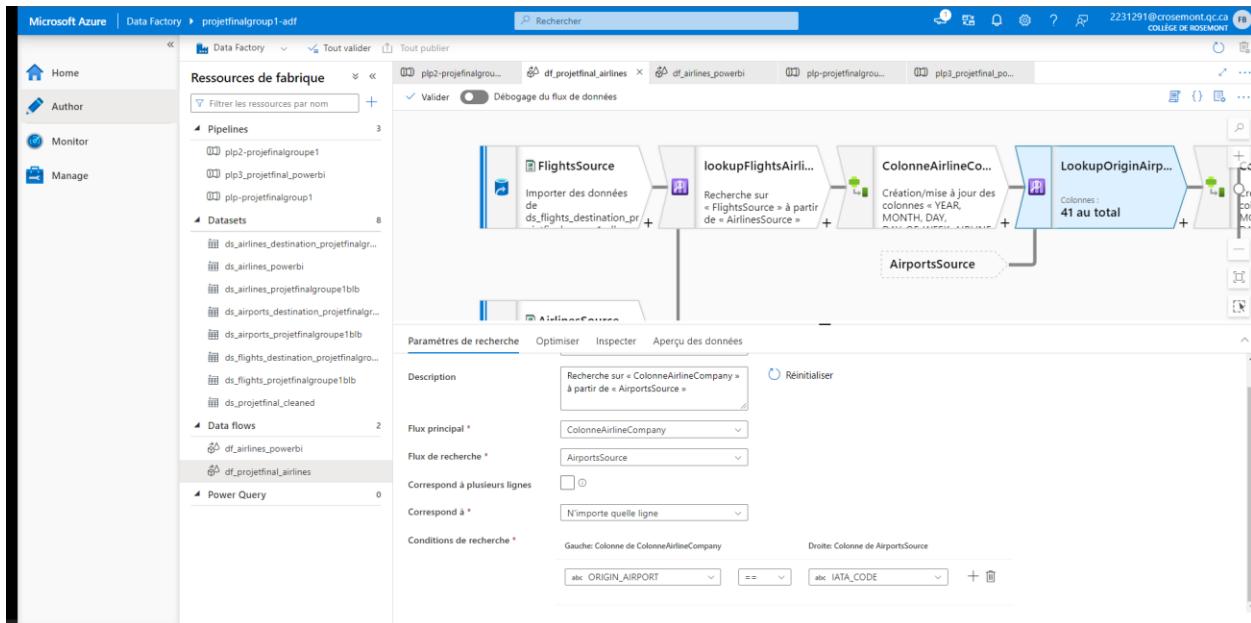
DAY

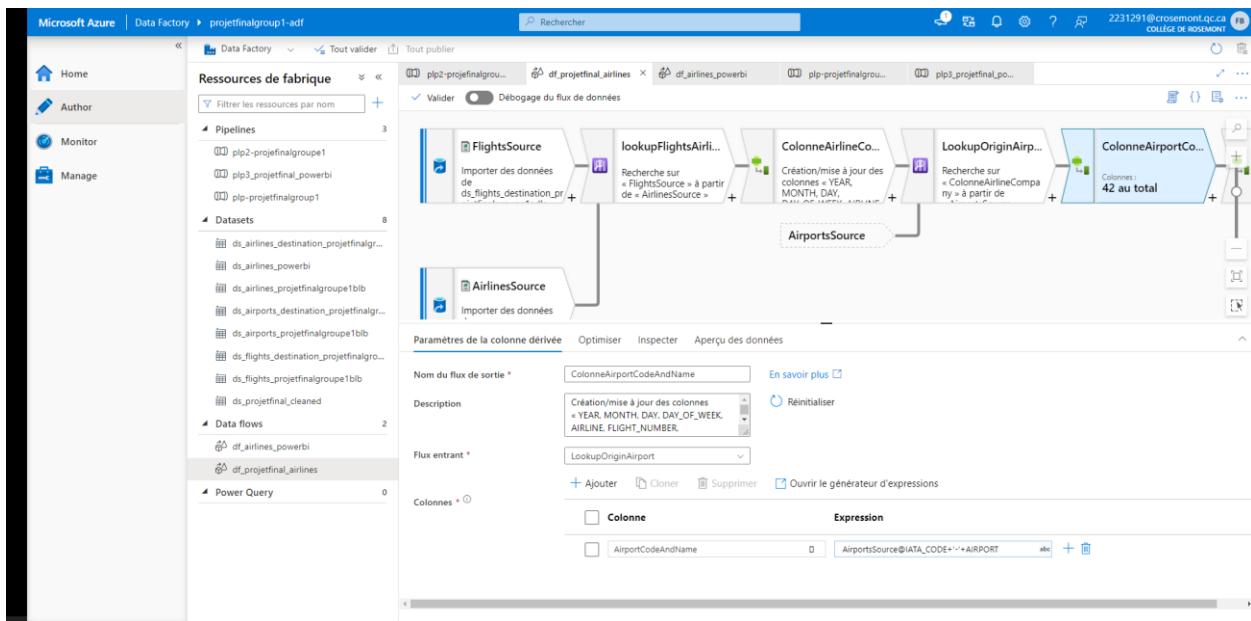
DAY_OF_WEEK

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

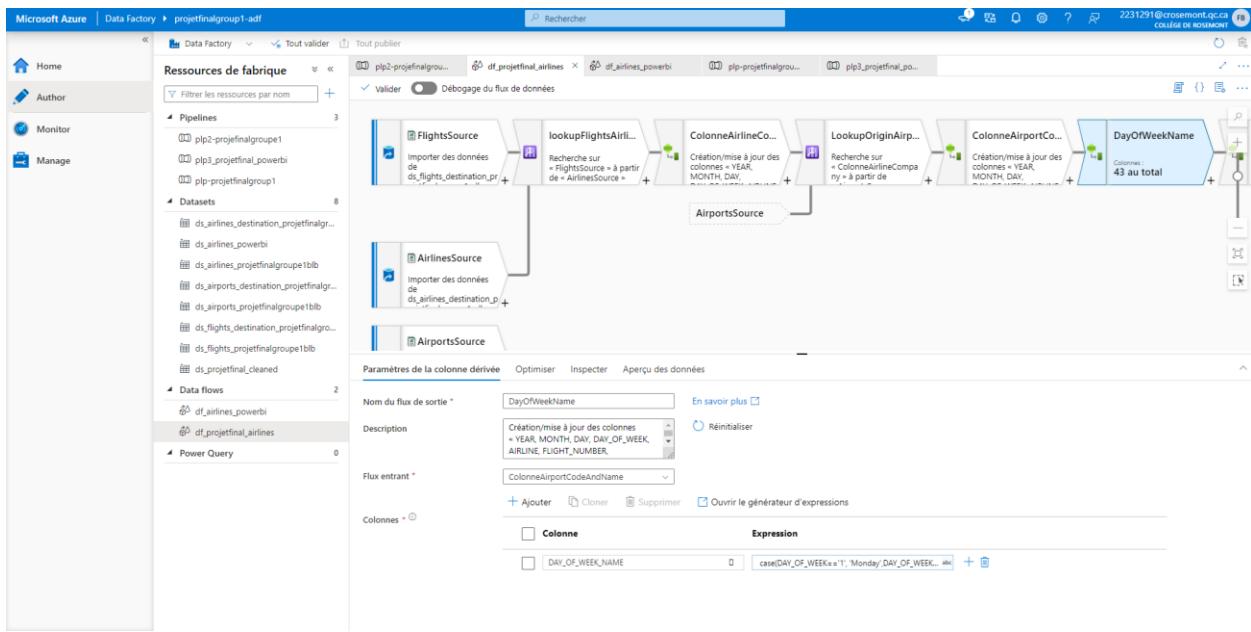
AirportsSource@IATA_CODE+''+AIRPORT

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

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

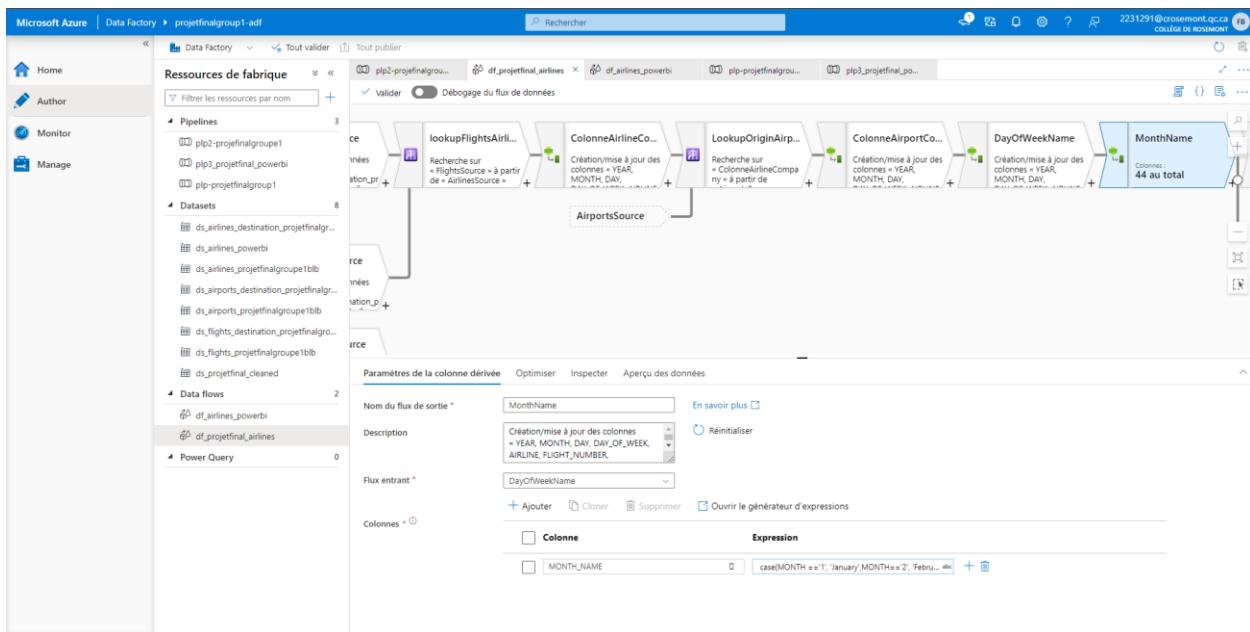
- Tous
- Fonctions
- Schéma d'entrée
- Réglages
- Recherche en cache
- Fonctions de bibliothèque de flux de données
- Variables locales

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

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

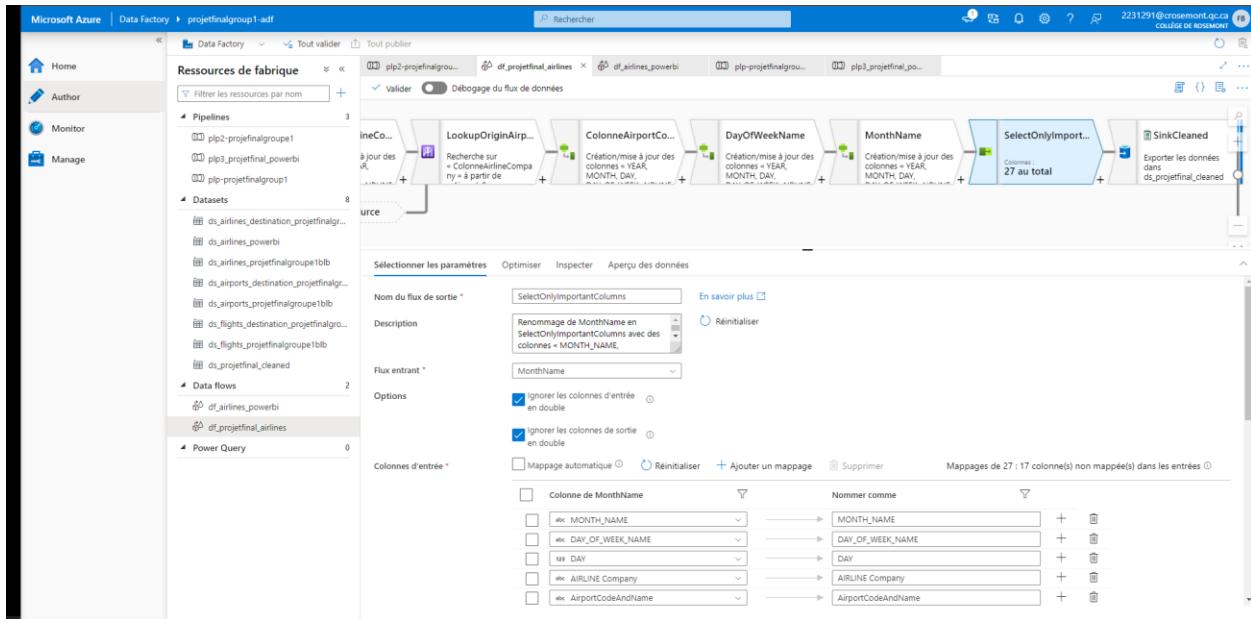
- Tous
- Fonctions
- Schéma d'entrée
- Réglages
- Recherche en cache
- Fonctions de bibliothèque de flux de données
- Variables locales

Valeurs d'expression

- YEAR
- MONTH
- DAY
- DAY_OF_WEEK
- FlightSource@AIRLINE

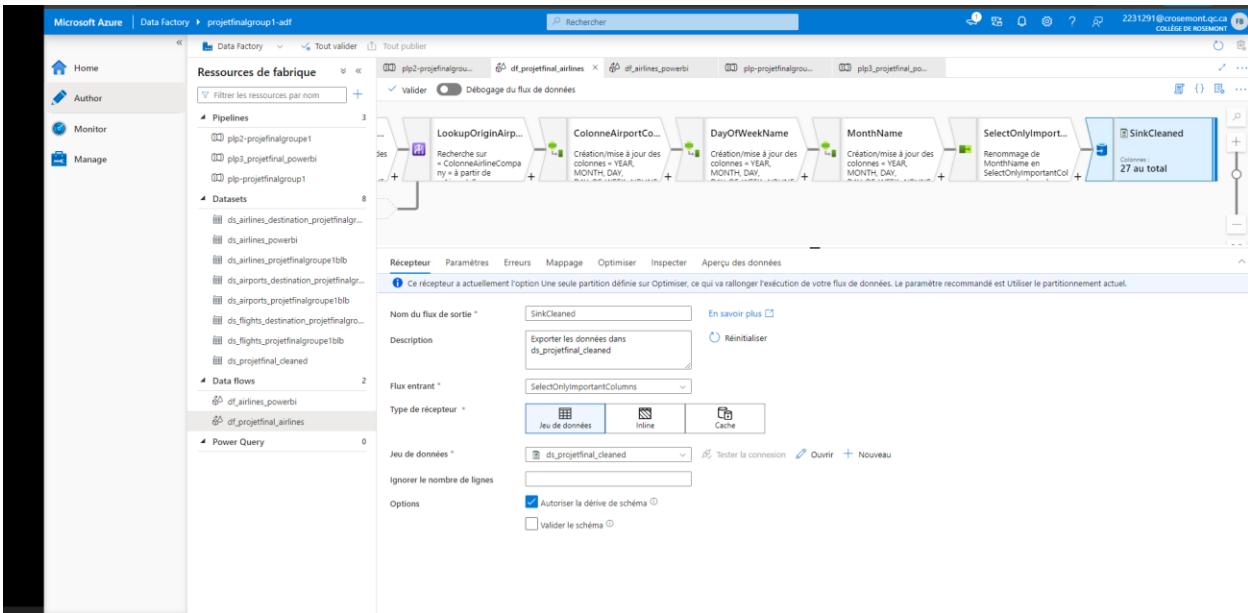
Aperçu des données

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



On garde uniquement les colonnes nécessaires à notre analyse

Colonne de MonthName	Nommer comme
12s LATE_AIRCRAFT_DELAY	LATE_AIRCRAFT_DELAY
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 :

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

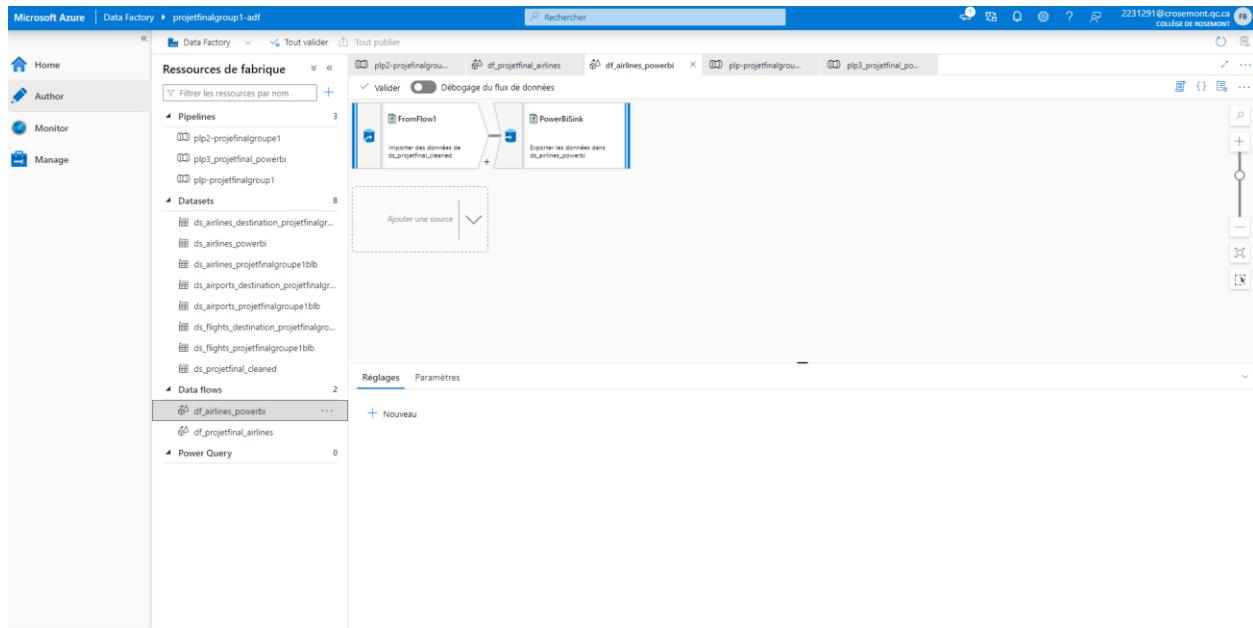
AIRPORT SOURCE

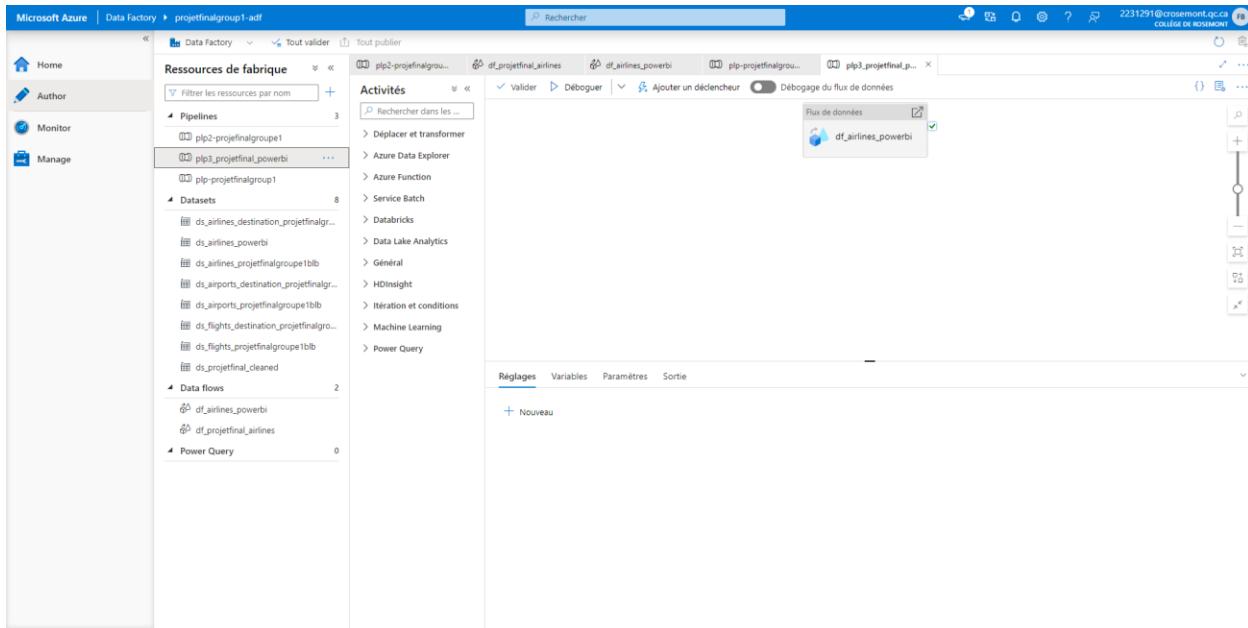
Récepteur Paramètres Erreurs Mappage Optimiser Inspector Aperçu des données ●

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. On the left, there's a navigation pane 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'.

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 * ①
10/22/22 23:59:59

Fuseau horaire * ①
Est (États-Unis et Canada) (UTC-5)

① Ce fuseau horaire observe l'heure d'été. Le déclencheur se règle automatiquement avec une heure de différence.

PéIODICITÉ * ①
Chaque 1 Jour(s)

✓ Option de récurrence avancée

Exécuter à ces horaires ①

Heures
Minutes

Planifier les horaires d'exécution
23:59

Spécifier une date de fin

Annotations
+ Nouveau

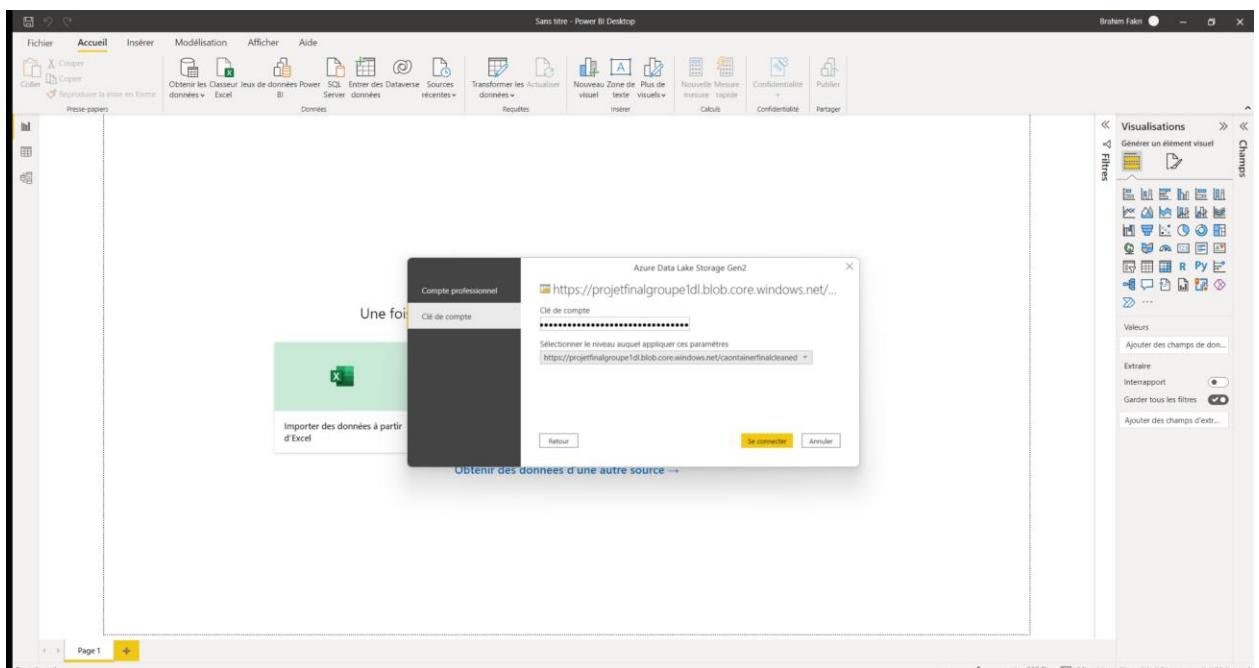
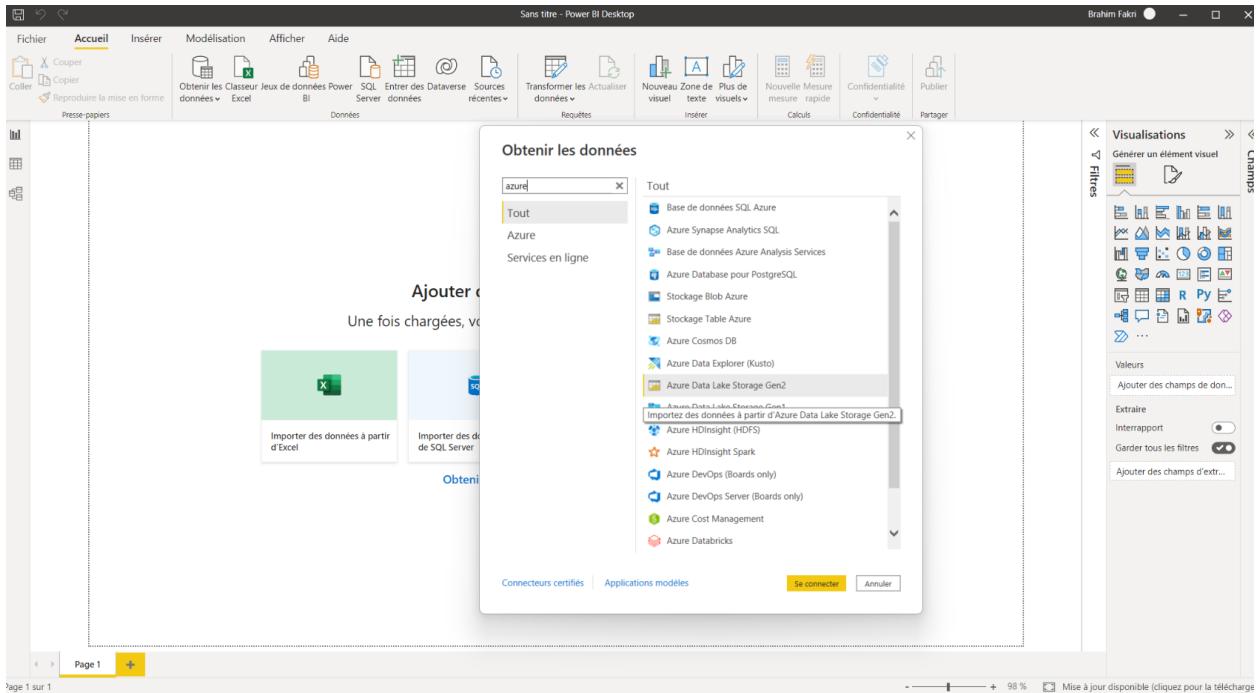
Démarrer le déclencheur ①

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

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>