

THURSDAY, SEPTEMBER 21, 2023

AZURE

ELT

REPORT

M. ABDERRAHIM
ELOUTMADI

AYMANE SABRI

DATA DEVELOPER

Objectif du projet :

L'objectif principal de ce projet est de développer une compréhension approfondie des concepts fondamentaux de l'[ETL](#) en utilisant les services [Microsoft Azure](#). Nous nous sommes engagés à mettre en œuvre un scénario d'extraction, de transformation et de chargement de données et la création d'une Data Ware House.

Ce projet vise à démontrer notre capacité à concevoir et à mettre en œuvre un flux de travail [ETL](#) efficace, tout en traitant les défis courants tels que les nettoyages de données et l'optimisation du data ware house.

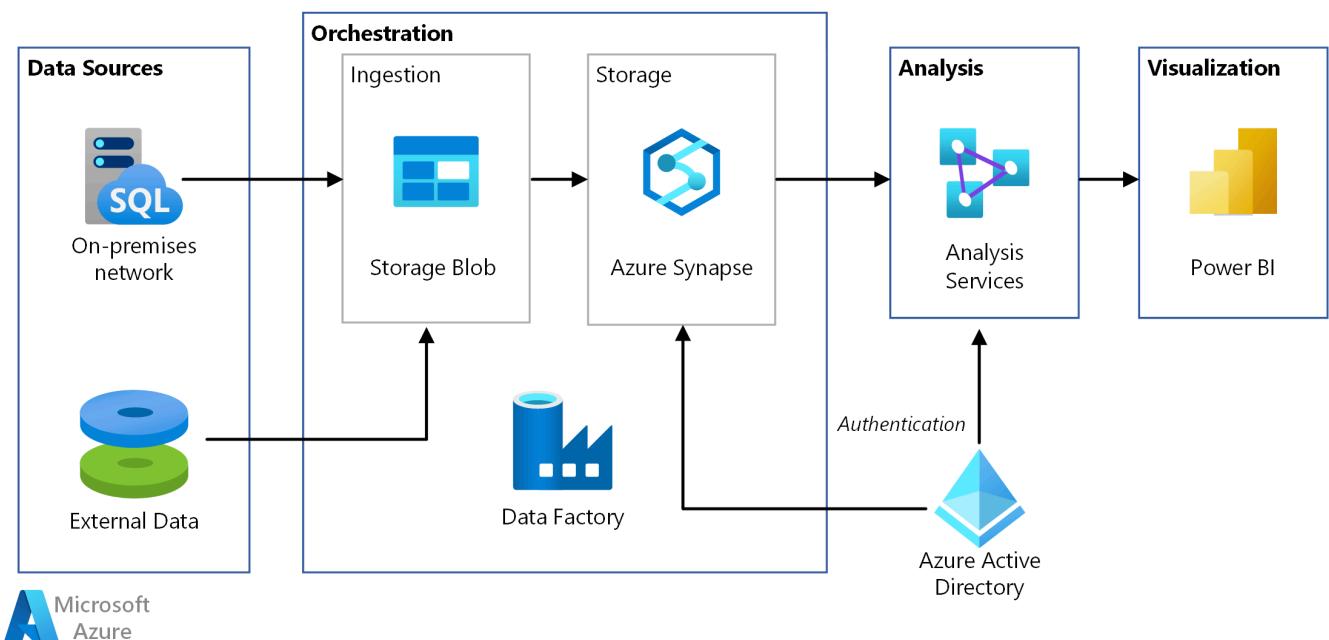
I. Introduction :

1. Contexte de projet :

L'explosion de la quantité de données disponibles, qu'elles soient structurées ou non, a mis en évidence la nécessité d'adopter des solutions de stockage et d'intégration de données sophistiquées. Dans ce contexte, notre projet vise à explorer les fonctionnalités d'intégration et de stockage ([Cloud](#)) de données en [Azure](#) en concevant un flux de travail [ETL](#) simple mais représentatif.

L'objectif principal de ce travail se résume à “ [Microsoft Azure](#) ”

- Extraction Données .
- Integration Des Données .
- Chargement Des Données .
- Optimization Des Performances .



2. Planification du plan du brief :

Dans l'objectif de bien mener le Brief, j'ai commencé par établir le **planning** à suivre durant la période de brief .

Pour ce faire, j'ai d'abords décomposé mon projet en **phases**, où chaque phase est définie par un certain nombre de tâches. Ensuite, j'ai élaboré une planification de ces phases sur la durée du projet, à l'aide d'un diagramme de Gantt.

2.1. Etapes Suivie :

Les **étapes** suivies sont :

- **Comprendre** la nature et l'étendue du travail demandé.
- **Extraction** de données .
- **Creation** d'un storage account .
- **Creation** du data factory .
- **Creation** du service azure synapse analysis .
- **Creation** du data warehouse .

Suite à ces étapes j'ai identifié les besoins à satisfaire, définit l'aspect fonctionnel de projet et sa conception, réalisé le système et finalement je l'ai soumis à plusieurs tests pour s'assurer de son adaptation aux **besoins** exprimés précédemment.



2.2. Diagramme de Gant :

Ce diagramme représente la durée de chaque tâche effectué dans mon projet.

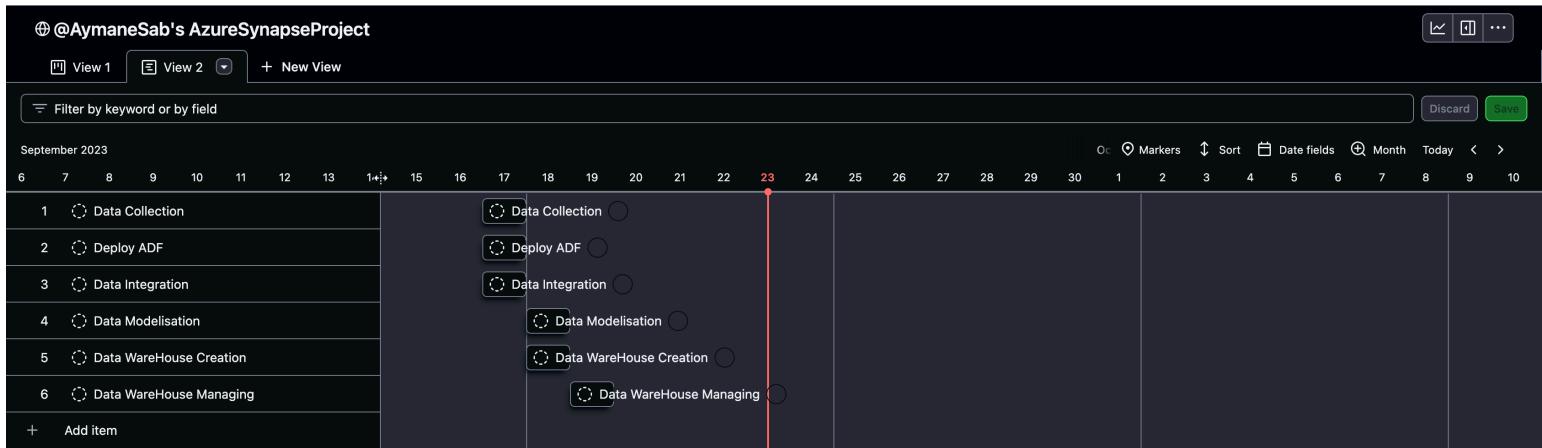


Figure : Diagramme de Gant

La portée du projet couvre les étapes fondamentales de la configuration de la source de données à la conception des transformations, en passant par l'utilisation des différents ressources azure afin de créer une data warehouse bien optimisé .

Nous abordons également les considérations liées à la qualité des données et aux différents normes de RGPD .

Data Integration
Draft AymaneSab opened 4 days ago

AymaneSab now (edited)

1. Data Flow (Mapping & Transformation) ;
2. She-dulling Activities ;
3. Set Triggers ;

Data Warehouse Creation
Draft AymaneSab opened 4 days ago

AymaneSab 1 minute ago (edited)

1. Create users ;
2. Create and insert dimensions tables ;
3. Create fact table ;
4. Set the triggers ;

Data Warehouse Managing
Draft AymaneSab opened 4 days ago

AymaneSab 1 minute ago (edited)

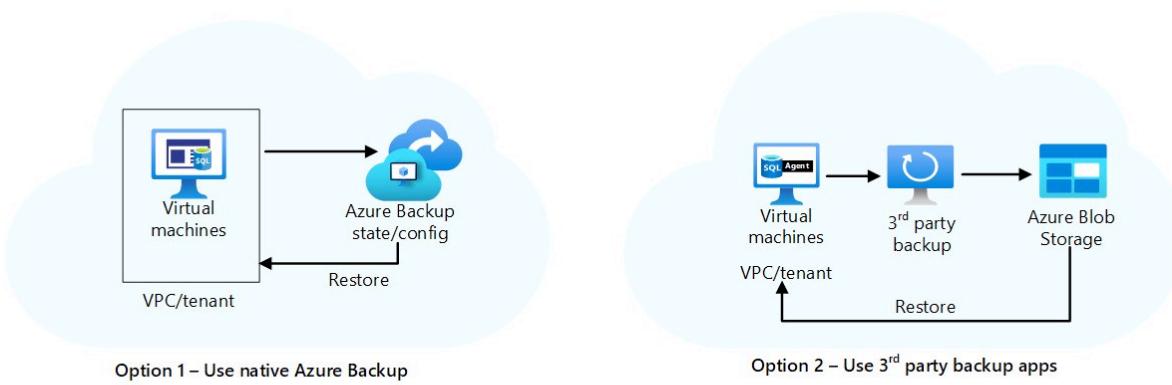
1. Sql unit testing ;
2. triggers testing ;
3. sauvegarde mechanisms ;

II. Realisation :

1. Vue D'ensemble sur les ressource azure utilisées .

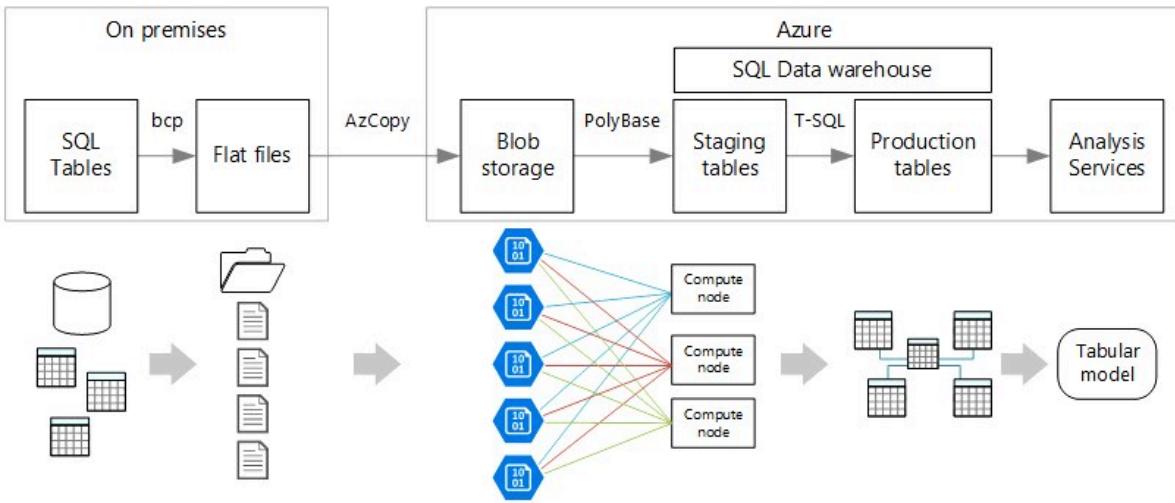
Nous avons utilisé une variété de composants et services azure pour mettre en œuvre notre flux de travail **ETL**. Parmi les ressources clés, citons :

- Azure Blob Storage :



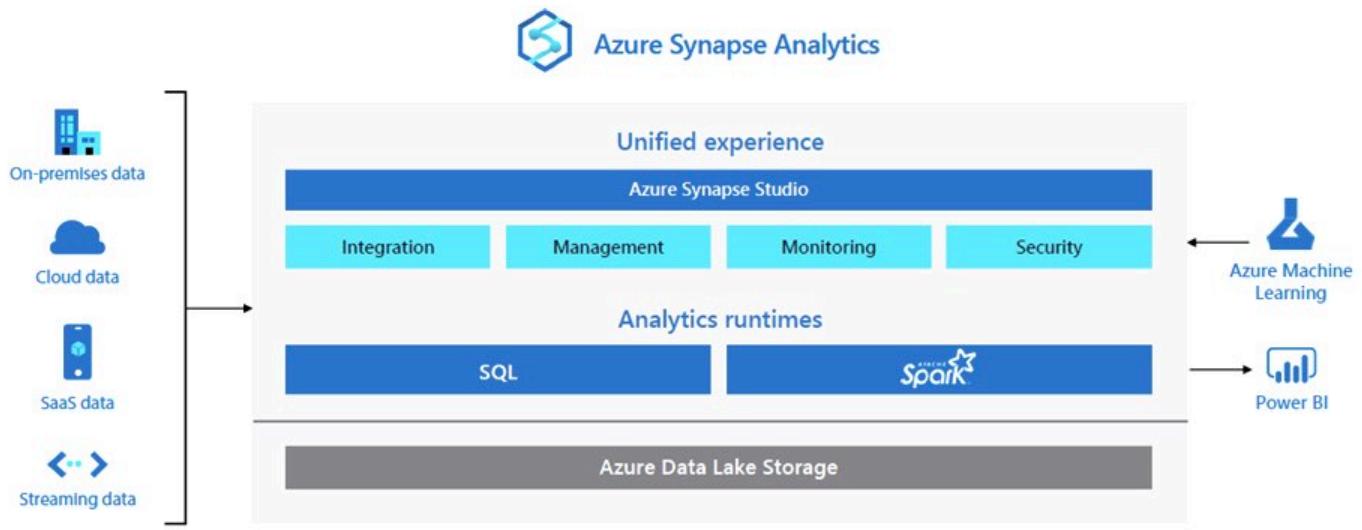
Azure Blob Storage is a highly **scalable** and **cost-effective** cloud storage service provided by **Microsoft Azure**. It is designed for the storage of **unstructured data**, also known as "**blobs**," which can include anything from **documents**, **images**, **videos**, and **backups** to **log files**, **datasets**, and more. **Blob Storage** offers a secure and reliable way to store and manage vast amounts of data in the **cloud**.

- Azure Data Factory :



Azure Data Factory ([ADF](#)) is a cloud-based data integration service provided by Microsoft Azure. It serves as a platform for [creating](#), [scheduling](#), and [managing](#) data- driven workflows, often referred to as data pipelines. [ADF](#) allows organizations to efficiently [move](#), [transform](#), and [process](#) data from various sources to destinations, making it an essential tool for modern [data management](#) and [processing](#).

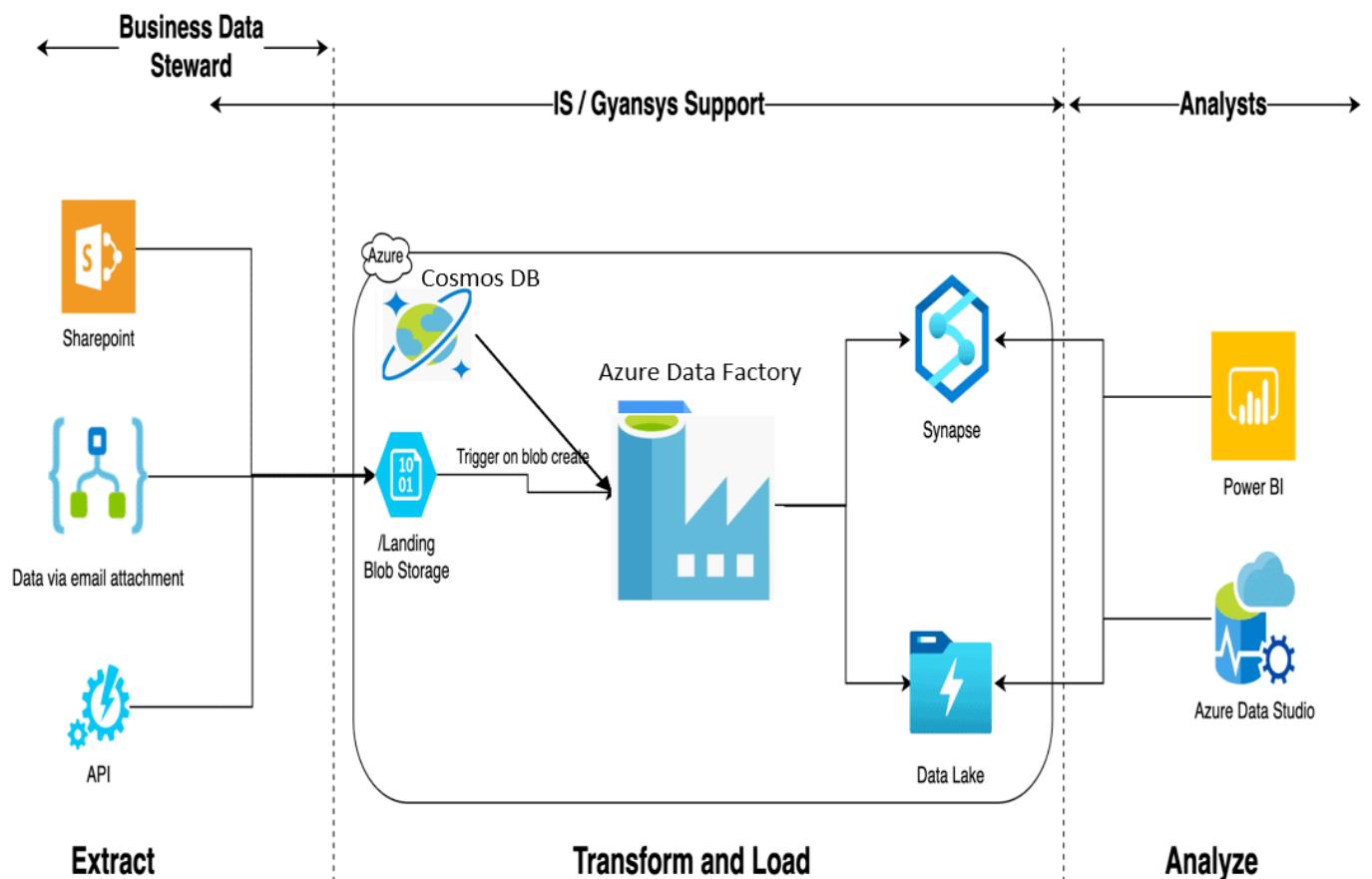
- Azure Synapse Analytics :



Azure Synapse Analytics, formerly known as SQL Data Warehouse, is a cloud-based analytics service provided by Microsoft Azure. It is designed to enable organizations to analyze and gain insights from large volumes of data in a highly scalable and performance-oriented manner. Azure Synapse Analytics brings together big data and data warehousing capabilities into a unified platform, making it easier for businesses to ingest, prepare, manage, and analyze data for various data-driven applications.

2. ETL :

Le cœur du projet résidait dans le processus **ETL**, où nous avons mis en œuvre les étapes précédemment définies de manière cohérente et efficace. À l'aide des fonctionnalités de conception visuelle de **Azure**, nous avons créé des flux de travail **ETL** en reliant les différentes étapes, de la lecture initiale des données à la chargement final dans la data ware house cible .



2.1. Collecte De Données :

The screenshot shows the 'Create a storage account' review step in the Azure portal. It displays configuration details across several sections: Basics, Advanced, Networking, and Security. The 'Review' tab is selected at the top.

Basics

Subscription	Simplon - Classe Data Youcode
Resource Group	DataResourceGRP
Location	francecentral
Storage account name	sabristorageaccount
Deployment model	Resource manager
Performance	Standard
Replication	Read-access geo-redundant storage (RA-GRS)

Advanced

Enable hierarchical namespace	Disabled
Enable network file system v3	Disabled
Allow cross-tenant replication	Disabled
Access tier	Hot
Enable SFTP	Disabled
Large file shares	Disabled

Networking

Network connectivity	Public endpoint (all networks)
Default routing tier	Microsoft network routing
Endpoint type	Standard

Security

Secure transfer	Enabled
-----------------	---------

At the bottom, there are buttons for 'Create', '< Previous' (disabled), 'Next >', 'Download a template for automation', and 'Give feedback'.

The screenshot shows the 'Containers' page for the 'input' container. It lists three blobs: 'datedim.csv', 'regiondim.csv', and 'weathermetric.csv'. A success message indicates 'Successfully uploaded blob(s)' and 'Successfully uploaded 3 blob(s)'.

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
datedim.csv	9/21/2023, 3:01:52 PM	Hot (Inferred)		Block blob	10.6 KiB	Available
regiondim.csv	9/21/2023, 3:01:52 PM	Hot (Inferred)		Block blob	60 B	Available
weathermetric.csv	9/21/2023, 3:01:53 PM	Hot (Inferred)		Block blob	67.72 KiB	Available

Le cœur du projet résidait dans le processus ETL, où nous avons mis en œuvre les étapes précédemment définies pour collecter efficacement les données depuis diverses sources vers notre compte de stockage Azure, depuis l'extraction initiale des données à partir de diverses sources jusqu'au stockage final dans notre compte Azure Blob Storage, où elles étaient prêtes à être traitées ultérieurement.

2.2. Transformation De Données :

Microsoft Azure Search resources, services, and docs (G+) asabri.ext@simplonfor... SIMPLINFORMATIONS.co

Home > Create a resource > Marketplace > Data Factory >

Create Data Factory

Basics Git configuration Networking Advanced Tags **Review + create**

[View automation template](#)

TERMS

By clicking "Create", I (a) agree to the legal terms and privacy statement(s) associated with the Marketplace offering(s) listed above; (b) authorize Microsoft to bill my current payment method for the fees associated with the offering(s), with the same billing frequency as my Azure subscription; and (c) agree that Microsoft may share my contact, usage and transactional information with the provider(s) of the offering(s) for support, billing and other transactional activities. Microsoft does not provide rights for third-party offerings. See the [Azure Marketplace Terms](#) for additional details.

Basics

Subscription	Simplon - Classe Data Youcode
Resource group	DataSourceGRP
Name	sabridatafactory
Region	France Central
Version	V2

Git configuration

Repository Type	GitHub
GitHub account	AymaneSab
Repo name	Azure_DataWarehouse
Branch name	ADF
Root folder	/

Networking

Connect via	Public endpoint
-------------	-----------------

[Previous](#) [Next](#) **Create** [Give feedback](#)

Microsoft Azure | Data Factory > sabridatafactory Search factory and documentation asabri.ext@simplonformations.onmicrosoft.com SIMPLINFORMATIONS.co

Validate all Save all Publish

Linked services

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

+ New [Filter by name](#) Annotations: Any

No linked service to show
If you expect to see results, try changing your filters or [Create linked service](#)

New linked service [Azure Blob Storage Learn more](#)

Name * BlobStorageLink

Description

Connect via integration runtime * AutoResolveIntegrationRuntime

Authentication type Account key

Connection string Azure Key Vault

Account selection method From Azure subscription

Azure subscription Select all

Storage account name * sabrstorageaccount

Additional connection properties [New](#)

Annotations [New](#)

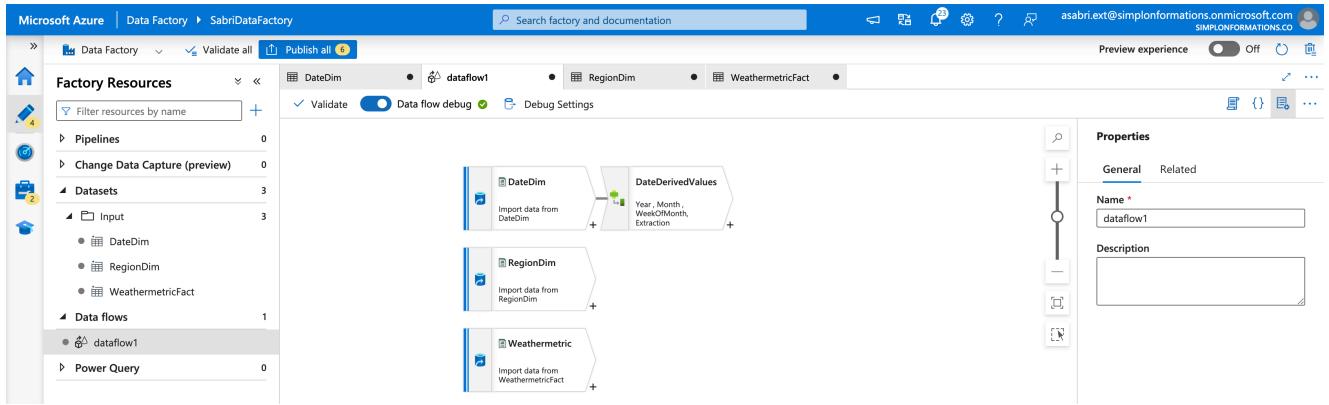
Parameters [Add](#)

Test connection To linked service To file path

Annotations [New](#)

Parameters [Add](#)

Create **Back** **Cancel** **Test connection** **Connection successful**



Une fois que les données ont été collectées depuis différentes sources vers notre compte [Azure Blob Storage](#), j'ai mis en place un processus de transformation des données en utilisant [Azure Data Factory](#). Cette étape m'a permis de appliquer le principe de granularité etc

Create Synapse workspace

Validation succeeded

Basics **Security** Networking Tags **Review + create**

Product Details

Azure Synapse Analytics workspace by Microsoft Serverless SQL est. cost/TB 4.68 EUR

Terms

By clicking Create, I (a) agree to the legal terms and privacy statement(s) associated with the Marketplace offering(s) listed above; (b) authorize Microsoft to bill my current payment method for the fees associated with the offering(s), with the same billing frequency as my Azure subscription; and (c) agree that Microsoft may share my contact, usage and transactional information with the provider(s) of the offering(s) for support, billing and other transactional activities. Microsoft does not provide rights for third-party offerings. For additional details see [Azure Marketplace Terms](#).

Basics

Description	Simplon - Classe Data Youcode
Resource group	DataResourceGRP
Region	France Central
Workspace name	(new) sabrilsynapseworkspace
Data Lake Storage Gen2 account	(new) https://sabrilaakeaccount.dfs.core.windows.net
Data Lake Storage Gen2 file system	(new) sabrilafilesystem
Managed resource group	None

Security

Authentication method	Use both local and Azure Active Directory (Azure AD) authentication
SQL Server admin login	sqladminuser
SQL Password	Auto-generated
Double encryption	No

Create < Previous Next > Download a template for automation Next >

2.3. Chargement De Données :

The screenshot shows the Microsoft Azure Synapse Analytics workspace interface. On the left, there's a navigation sidebar with various options like Analytics pools, External connections, Integration, Security, Configurations + libraries, Source control, and Git configuration. The 'Linked services' option is selected. The main area displays a list of linked services, including 'AzureDataLakeStorage', 'sabrisynapseworkspace-WorkspaceDefaultSqlServ...', and 'sabrisynapseworkspace-WorkspaceDefaultStorage'. Each entry shows its type (e.g., Azure Data Lake Storage Gen2), related count (0), and annotations.

The screenshot shows the Microsoft Azure Storage blob container for 'sabrilakefilesystem'. The container has a single item named '_SUCCESS'. The table below lists other uploaded files:

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
part-00000-12cf6352-5358-4c0c-b8ef-1217d9284228-c000.csv	9/21/2023, 3:39:14 PM	Hot (Inferred)		Block blob	60 B	Available
part-00000-1ea2f309-9b4d-4961-8e80-093e3da6e433-c000.csv	9/21/2023, 3:39:13 PM	Hot (Inferred)		Block blob	1.37 KiB	Available
part-00000-59aaef972-f752-46cf-b77f-fbe4781e3e5f-c000.csv	9/21/2023, 3:39:16 PM	Hot (Inferred)		Block blob	67.72 KiB	Available

Pour la phase de chargement des données dans notre data warehouse, nous avons opté pour Azure Synapse Analytics, un puissant service qui nous a permis de tirer pleinement parti de la capacité de stockage et de traitement à grande échelle. Après avoir transformé les données à l'aide d'Azure Data Factory, nous avons configuré des pipelines de chargement pour transférer efficacement les données nettoyées et préparées dans notre data warehouse Synapse Analytics.

Microsoft Azure | Synapse Analytics > sabrisynapseworkspace

Integrate

Dataflow1

Validate all

RegionDim Import data from DelimitedText3 + sink1 Columns: 2 total

No items to show Try creating a new item using the + button above. Learn more ↗

Sink Settings Errors Mapping Optimize Inspect Data preview

Output stream name * sink1 Learn more ↗

Description Add sink dataset Reset

Incoming stream * RegionDim

Sink type * Integration dataset Inline Workspace DB Cache

Dataset * Select... New

Allow schema drift Validate schema

Create Cancel Test connection

New linked service

Azure Synapse Analytics Learn more ↗

Connection string Azure Key Vault

Account selection method From Azure subscription Enter manually

Azure subscription Select all

Server name * sabrisynapseworkspace (Synapse workspace)

Database name * sabrededicatedpool

SQL pool * sabrededicatedpool

Authentication type * SQL authentication

User name * sqladminuser

Password Azure Key Vault

Additional connection properties

Annotations

Parameters Advanced

Microsoft Azure | Synapse Analytics > sabrisynapseworkspace

Develop

SQL scripts

Data flows

Dataflow1

SQL script 1 Pipeline 1 Dataflow1

Committed Validate Data flow debug Debug Settings

WeatherMetric WeatherMetricFact Import data from WeatherMetric Export data to sabreAzureSynapseAnalyticsTable

DateDim DateDimension Import data from DataSource Export data to DateSynapseService

RegionDim RegionDimension Import data from RegionDim Export data to RegionSynapseService

Add Source

Parameters Settings

+ New

Microsoft Azure | Synapse Analytics > sabrisynapseworkspace

Data Workspace Linked

Filter resources by name

SQL database

sabrededicatedpool (SQL)

Tables

dbo.DateDim

dbo.RegionDim

anot.17578.be77842c0694...

dbo.WeatherMetricsFact

External tables

External resources

Views

Programmability

Schemas

Security

SQL script 1

Pipeline 1

RegionDim

RegionDim

```

CREATE TABLE [dbo].[RegionDim] (
    [RegionID] Integer NOT NULL,
    [RegionName] Varchar(255)
);

CREATE TABLE [dbo].[WeatherMetricsFact] (
    [DateID] Integer NOT NULL,
    [RegionID] Integer NOT NULL,
    [AvgWindDirection] Integer,
    [AvgWindSpeed] Integer,
    [MinWindSpeed] Integer,
    [MaxWindSpeed] Integer,
    [WindDirHours] Integer
);

SELECT * FROM [RegionDim]
SELECT * FROM [DateDim]
SELECT * FROM [WeatherMetricsFact]
DROP TABLE [RegionDim]

```

Microsoft Azure | Synapse Analytics > sabrisynapseworkspace

Integrate

Pipelines

Pipeline 1

Activities

Sync Move and transform Azure Data Explorer Azure Function Batch Service Databricks Data Lake Analytics General HDInsight Iteration & conditions Machine Learning

Data Flow

Data flow

General Settings Parameters User properties

Name * Data flow 1 Learn more ↗

Description

Activity state (previous) Active Inactive

Timeout 0:12:00:00

Retry 0

Retry interval (sec) 30

Secure output

Secure input

