

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.

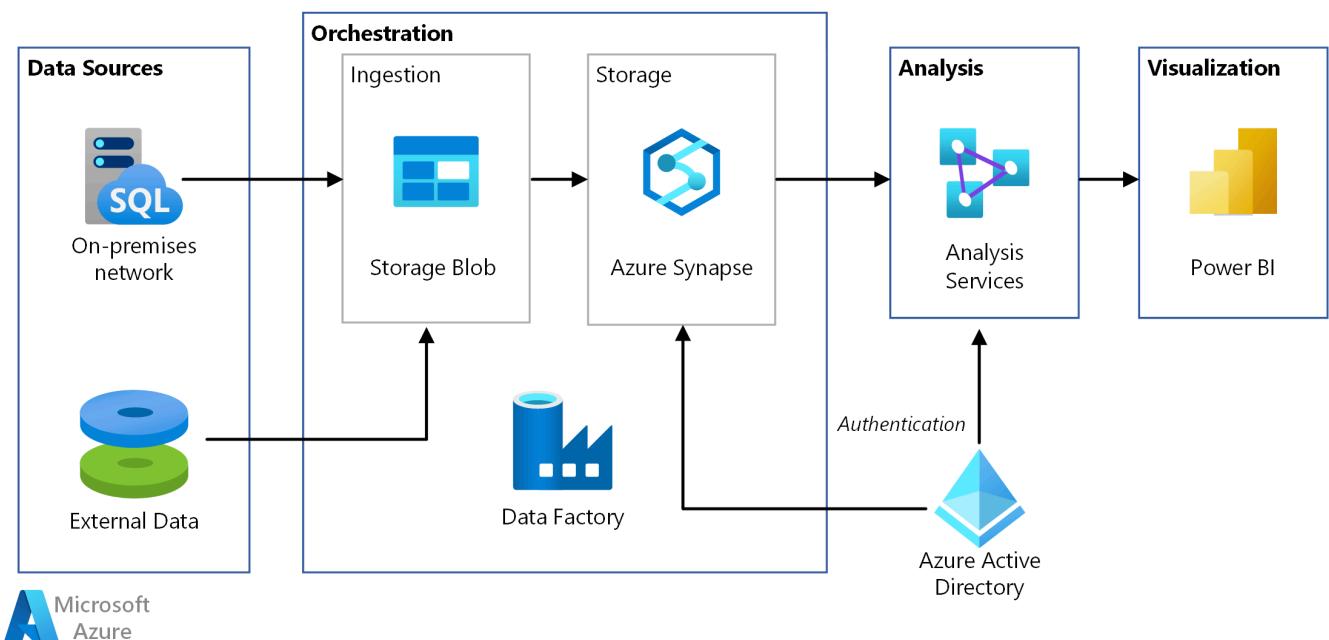
# I. Introduction :

## 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.

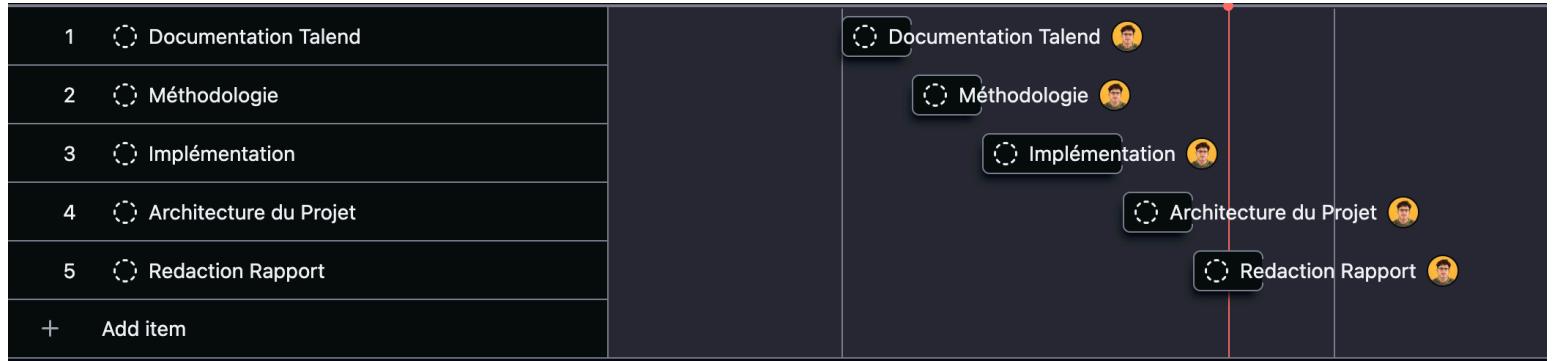


Figure : Diagramme de Gant

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 .

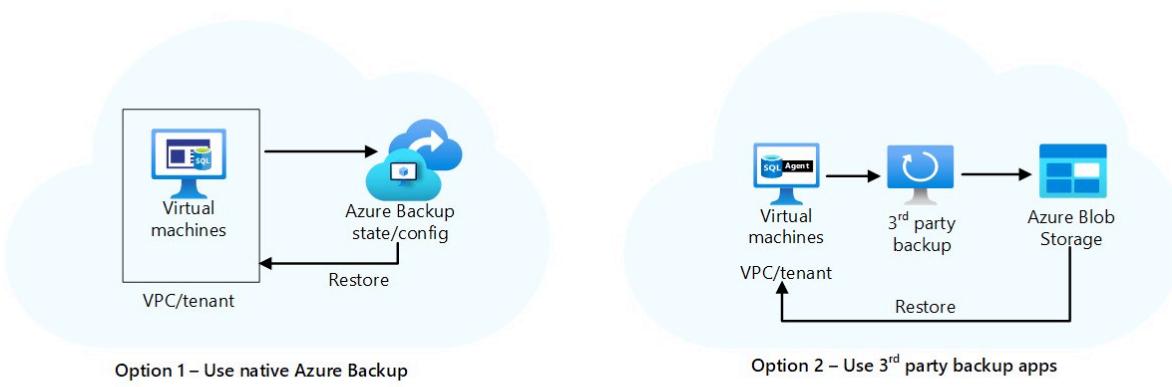


## II. Realisation :

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

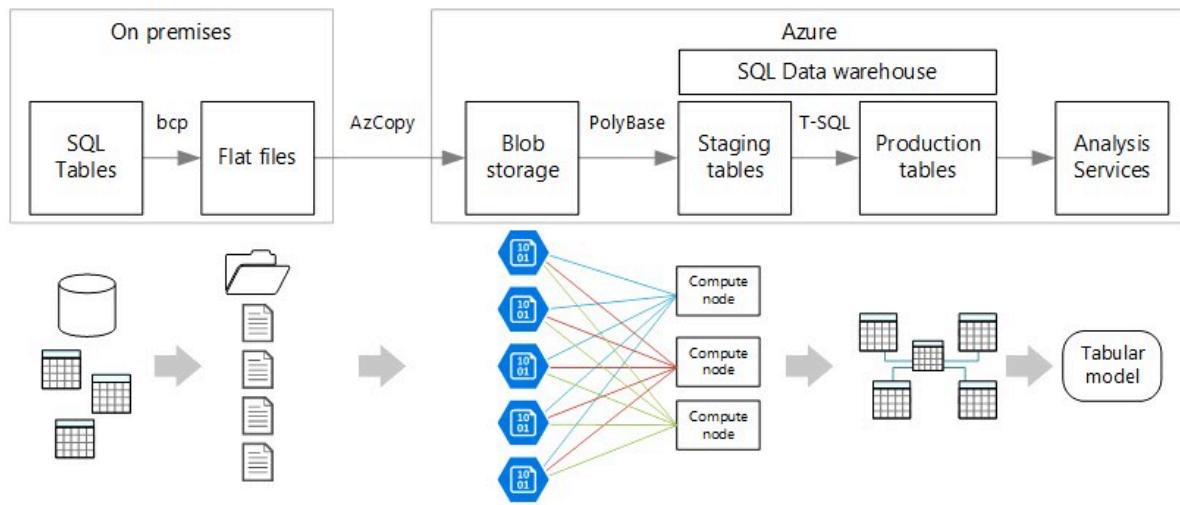
Nous avons utilisé une variété de composants et services azure pour mettre en œuvre notre flux de travail **ETL**. Parmi les ressources clés, citons :

- Azure Blob Storage :



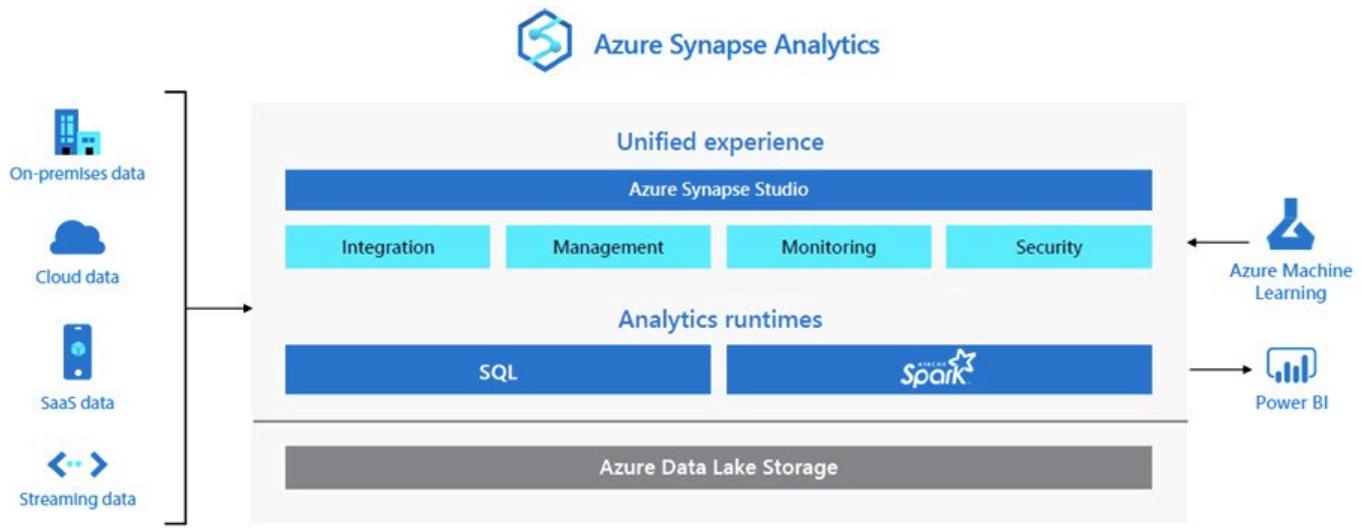
Azure Blob Storage is a highly **scalable** and **cost-effective** cloud storage service provided by **Microsoft Azure**. It is designed for the storage of **unstructured data**, also known as "**blobs**," which can include anything from **documents**, **images**, **videos**, and **backups** to **log 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](#).

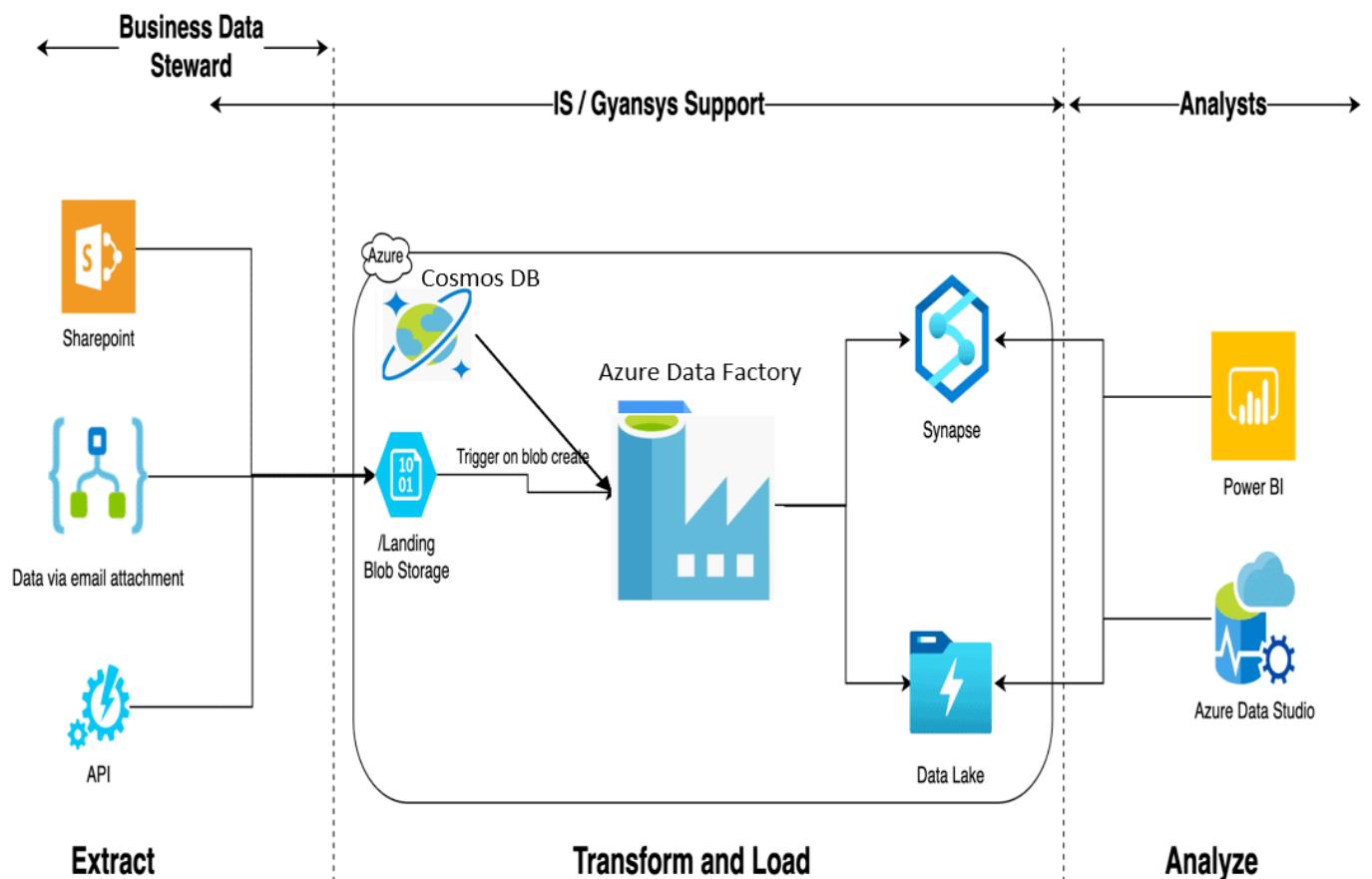
- Azure Synapse Analytics :



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' page in the Microsoft Azure portal. The 'Review' tab is selected. The configuration includes:

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

**Advanced** settings include:

- 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** settings include:

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

**Security** settings include:

- Secure transfer: Enabled

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

The screenshot shows the 'Containers' page for the 'input' container. The 'Overview' tab is selected. A success message indicates 'Successfully uploaded blob(s)' and 'Successfully uploaded 3 blob(s)'. The table lists three blobs:

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 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 .

## 2.2. 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](#)

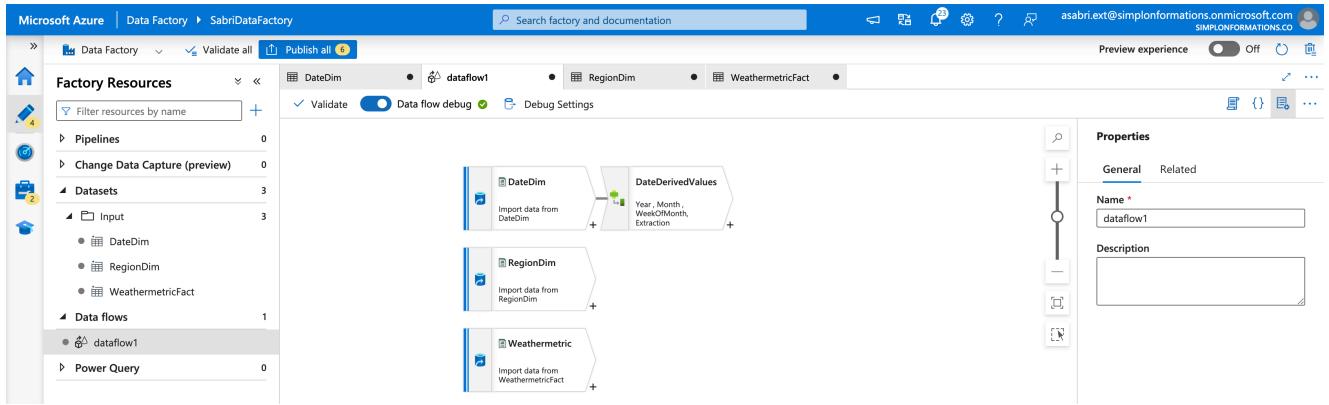
**Parameters** [New](#)

**Test connection** To linked service To file path

**Annotations** [New](#)

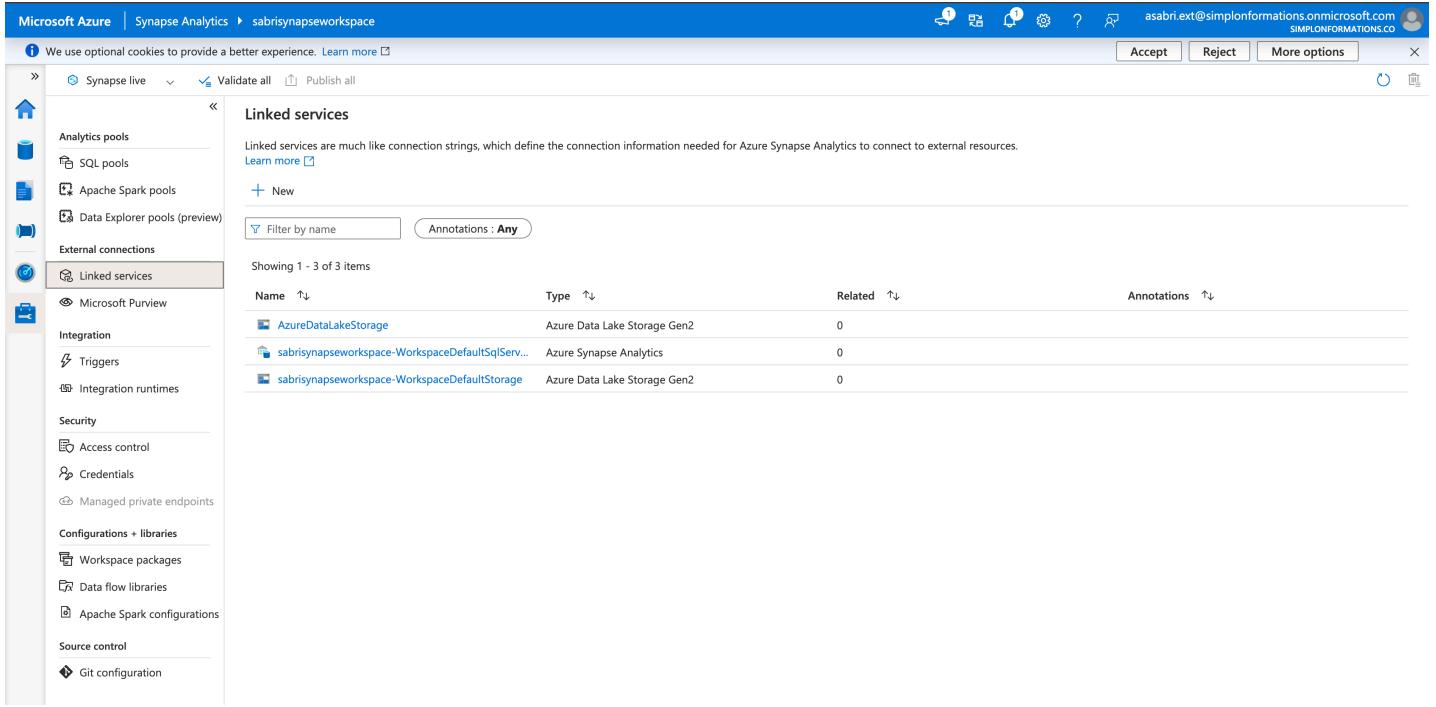
**Parameters** [New](#)

**Test connection** **Connection successful** [Cancel](#)

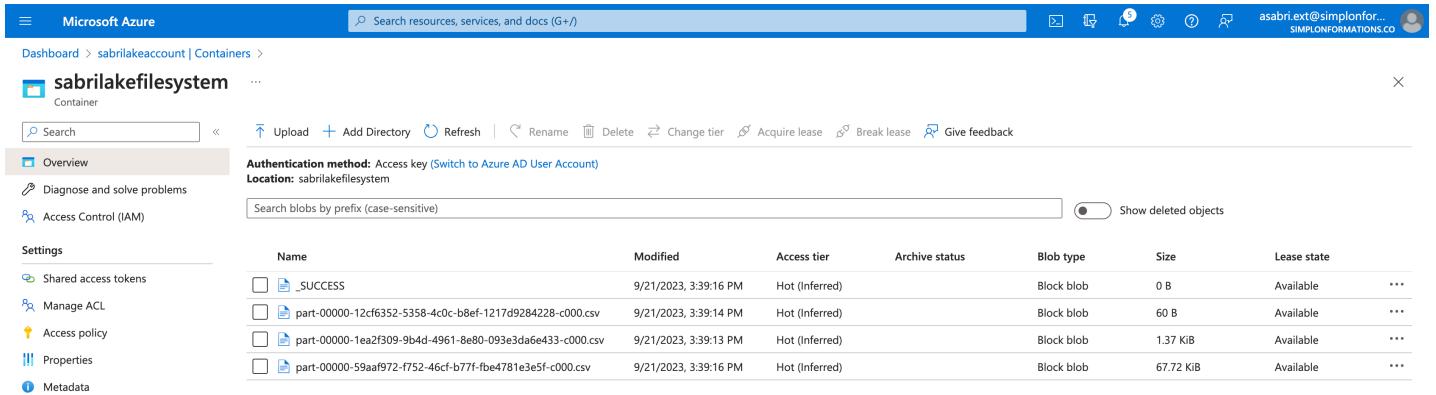


Le cœur du projet résidait dans le processus **ETL**, où nous avons mis en œuvre les étapes précédemment définies de manière cohérente et efficace. À l'aide des fonctionnalités de conception visuelle de **Azure**, nous avons créé des flux de travail **ETL** en reliant les différentes étapes, de la lecture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

## 2.3. Transformation De Données :



The screenshot shows the Microsoft Azure Synapse Analytics workspace interface. The left sidebar navigation includes options like Analytics pools, SQL pools, Apache Spark pools, Data Explorer pools (preview), External connections, Linked services (selected), Microsoft Purview, Integration, Triggers, Integration runtimes, Security, Access control, Credentials, Managed private endpoints, Configurations + libraries, Workspace packages, Data flow libraries, Apache Spark configurations, Source control, and Git configuration. The main content area is titled "Linked services" and displays a list of three items: "AzureDataLakeStorage" (Azure Data Lake Storage Gen2), "sabrisynapseworkspace-WorkspaceDefaultSqlServ..." (Azure Synapse Analytics), and "sabrisynapseworkspace-WorkspaceDefaultStorage" (Azure Data Lake Storage Gen2). A "New" button is available for creating new linked services.



The screenshot shows the Microsoft Azure Blob storage container details for "sabrilakefilesystem". The left sidebar navigation includes Overview, Diagnose and solve problems, Access Control (IAM), Settings, Shared access tokens, Manage ACL, Access policy, Properties, and Metadata. The main content area shows the container's authentication method as "Access key (Switch to Azure AD User Account)" and its location as "sabrilakefilesystem". It features a search bar for blobs by prefix and a table of blob objects. The table columns are Name, Modified, Access tier, Archive status, Blob type, Size, and Lease state. The table lists four blob objects:

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
_SUCCESS	9/21/2023, 3:39:16 PM	Hot (Inferred)		Block blob	0 B	Available
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-00000-59aaf972-f752-46cf-b77f-fbe4781e3e5f-c000.csv	9/21/2023, 3:39:16 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 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 .

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

**Sink**

Output stream name \* sink1

Description Add sink dataset

Incoming stream \* RegionDim

Sink type \* Integration dataset

Dataset \* Select... New

Options Allow schema drift:  Validate schema:

**Create** **Cancel** **Test connection**

**Develop**

**SQL scripts** 1

**Data flows** 1

**Dataflow1**

**SQL script 1**

**Pipeline 1**

**Dataflow1**

**Committed** **Validate** **Data flow debug** **Debug Settings**

**Parameters** **Settings**

**New**

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

