

WEDNESDAY, SEPTEMBER 27,

AZURE DATA BRICKS

REPORT

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.

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 Intégration de l'Infrastructure Data Lake .

I. Introduction :

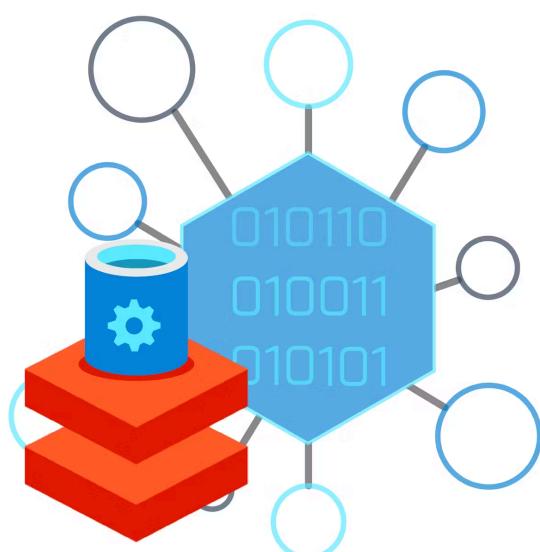
1. Contexte de projet :

Avec l'urbanisation croissante et la demande de mobilité, il est essentiel d'avoir une vue claire des opérations de transport. Les données, allant des horaires des bus aux retards, peuvent offrir des perspectives pour améliorer les services et répondre aux besoins des citoyens.

En tant que développeur Data, le professionnel en charge de cette situation est sollicité pour mettre en place des solutions basées sur les données pour répondre à ces défis. Cela implique:

L'objectif principal de ce travail se résume à “ ETL , Azure Data Bricks ”

- Conception de l'Architecture Data Lake .
- Intégration de l'Infrastructure Data Lake .
- Processus ETL avec Azure Data bricks .
- Automatisation des Politiques de Conservation.
- Génération de données à intervalles de lots (Batch Intervals).
- Batch Processing.



Azure Databricks



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 :

- **Conception** de l'Architecture Data Lake .
- **Intégration** de l'Infrastructure Data Lake .
- Processus **ETL** avec **Azure Data Bricks** .
- Automatisation des Politiques de Conservation.
- **Génération** de données à intervalles de lots (Batch Intervals).
- Batch Processing.

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.

Title	Assignees	Status
1 Conception de l'Architecture Data Lake	AymaneSab	Done
2 Intégration de l'Infrastructure Data Lake	AymaneSab	Done
3 Processus ETL avec Azure Data Bricks	AymaneSab	Done
4 Automatisation des Politiques de Conservation	AymaneSab	Done
5 Génération de données à intervalles de lots (Batch Intervals)	AymaneSab	Done
6 Batch Processing	AymaneSab	Done

Conception de l'Architecture Data Lake

Draft AymaneSab opened 2 minutes ago

AymaneSab now (edited)

This item hasn't been started

1. Creation Du azure data lake Gen2 .
2. Creation des container nécessaires .
3. importation du fichier source .

Assignees: AymaneSab
Status: Todo

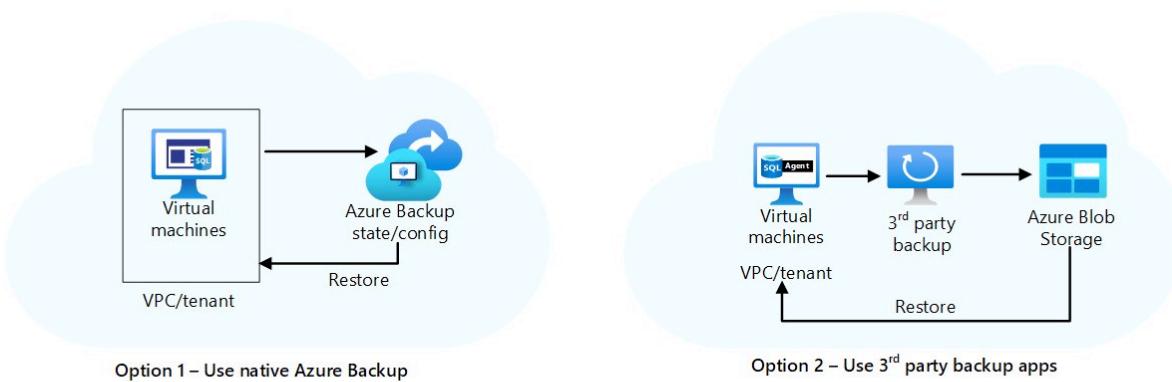
Actions: Convert to issue, Copy link in project, Archive, Delete from project

II. Realisation :

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

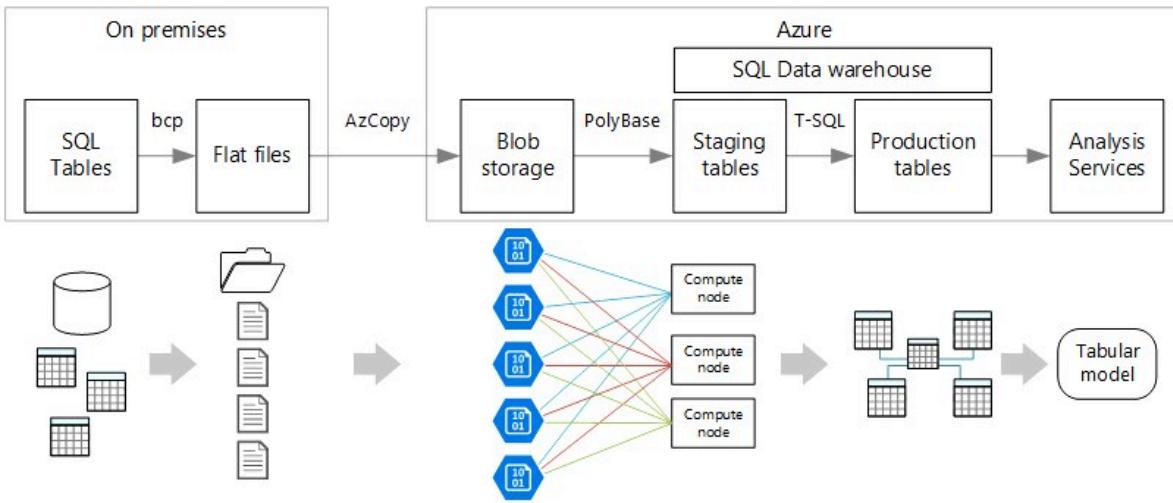
Nous avons utilisé une variété de composants et services azure pour mettre en œuvre notre flux de travail **ETL**. Parmi les ressources clés, citons :

- Azure Blob Storage :



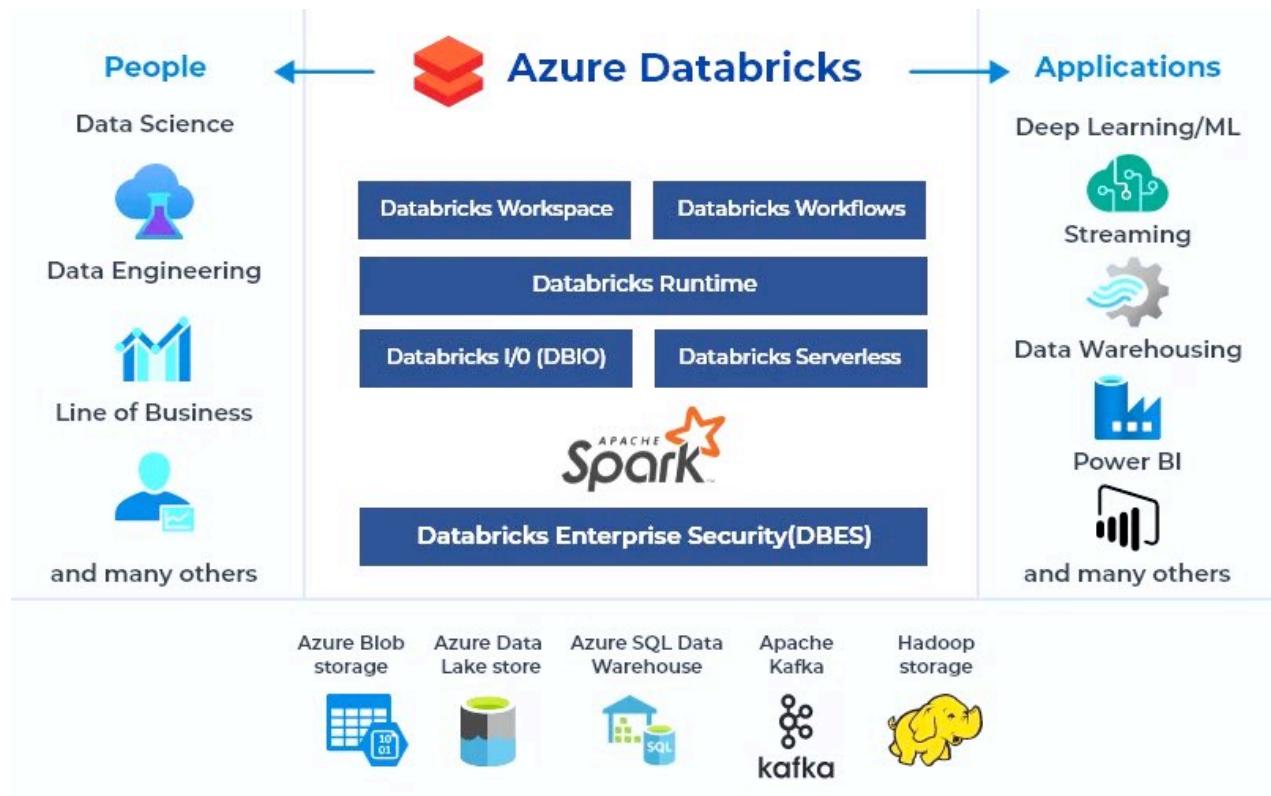
Azure Blob Storage is a highly **scalable** and **cost-effective** cloud storage service provided by **Microsoft Azure**. It is designed for the storage of **unstructured data**, also known as "**blobs**," which can include anything from **documents**, **images**, **videos**, and **backups** to **log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



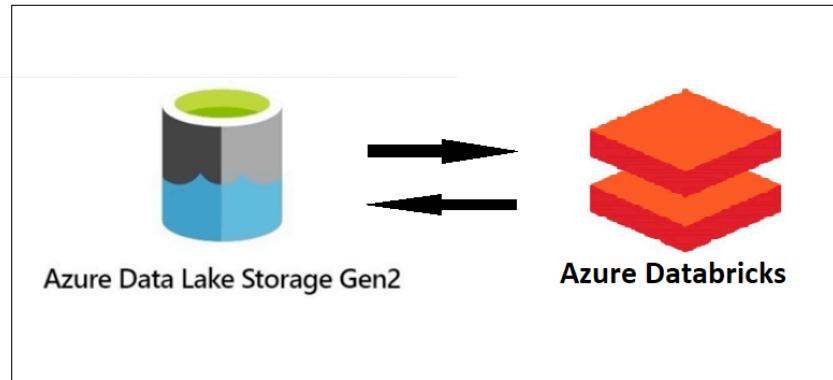
- Azure Synapse Analytics, formerly known as **SQL Data Warehouse**, is a cloud-based analytics service provided by Microsoft Azure. It is designed to enable organizations to **analyze** and gain insights from **large volumes** of data in a highly **scalable** and **performance-oriented** manner. Azure 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. Creation du data lake :

Azure Data Lake Storage Gen2 is a cloud-based data lake solution provided by Microsoft Azure. It's designed to store and manage large amounts of data for analytics and big data processing.

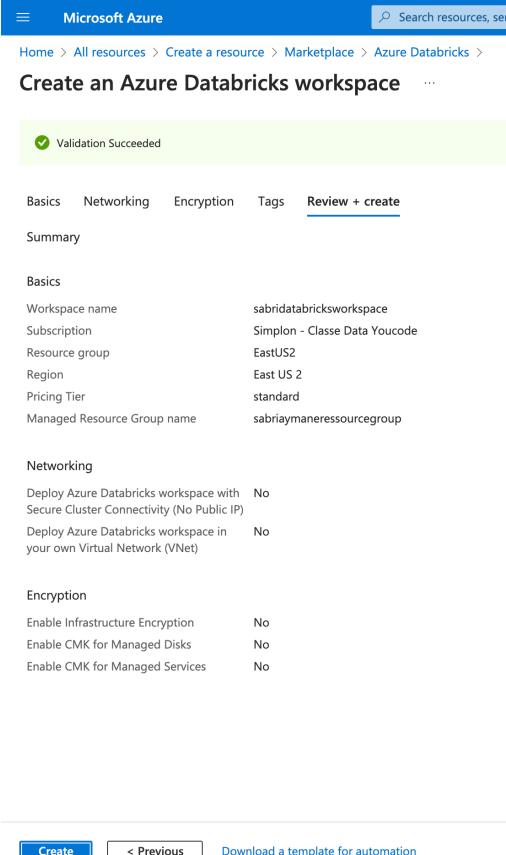
The screenshot shows the Microsoft Azure portal interface. At the top, there is a search bar with the placeholder "Search resources, services, and docs (G+/" and a user profile icon. Below the header, the URL "sabrilakestorageaccount_1695392153025 | Overview" is displayed. On the left, a navigation sidebar lists "Overview", "Inputs", "Outputs", and "Template". The main content area displays a green checkmark icon and the message "Your deployment is complete". It provides deployment details: Deployment name: sabrilakestorageaccount_1695392153025, Subscription: Simplon - Classe Data Youcode, Resource group: EastUS2. It also shows the start time as 9/22/2023, 3:15:58 PM and a Correlation ID: 35ae0b03-6482-4227-92e6-5f2f3ebbc121. A "Go to resource" button is present. To the right, there are two cards: "Cost Management" (with a dollar sign icon) and "Microsoft Defender for Cloud" (with a shield icon). Both cards have a "Set up cost alerts >" and "Go to Microsoft Defender for Cloud >" link respectively.

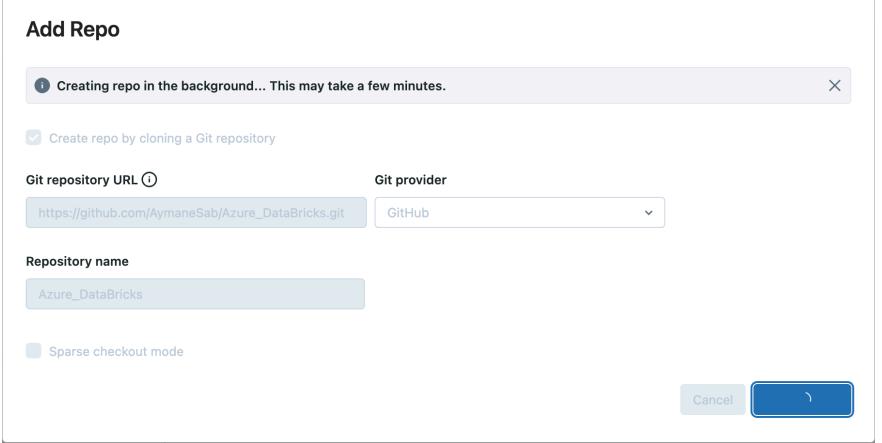
The screenshot shows the Microsoft Azure portal interface. At the top, there is a search bar with the placeholder "Search resources, services, and docs (G+/" and a user profile icon. Below the header, the URL "publictransportdata | Container" is displayed. On the left, a navigation sidebar lists "Overview", "Diagnose and solve problems", "Access Control (IAM)", "Settings" (with sub-options: Shared access tokens, Manage ACL, Access policy, Properties, Metadata), and "Upload" (with sub-options: Add Directory, Refresh, Rename, Delete, Change tier, Acquire lease, Break lease, Give feedback). The main content area displays a table of blobs. The table has columns: Name, Modified, Access tier, Archive status, Blob type, Size, and Lease state. Two blobs are listed: "processed" and "raw". A success message "Successfully added directory" and "Successfully added directory 'processed'" is shown in a toast notification. A search bar at the bottom allows searching by prefix (case-sensitive).

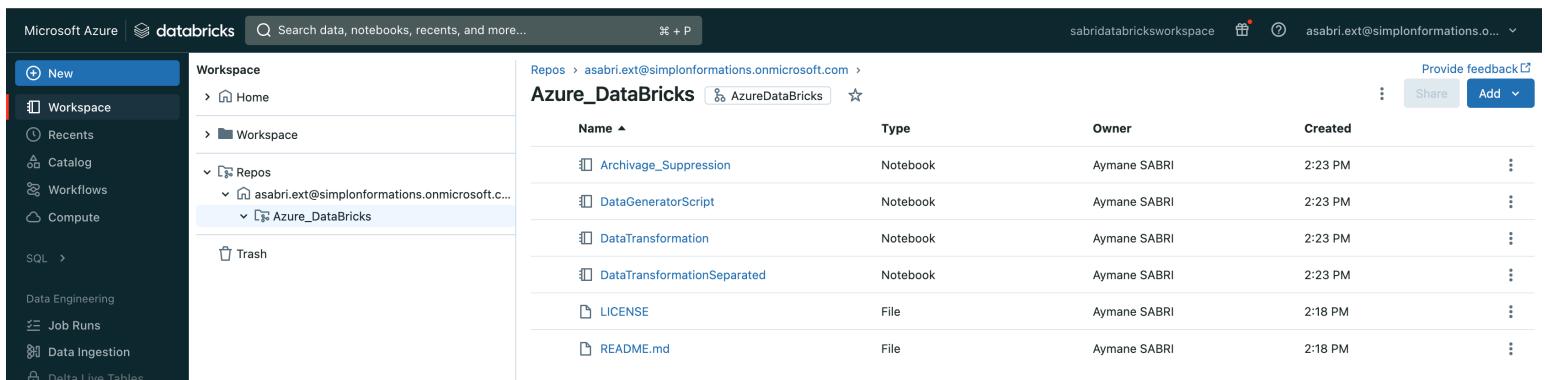


3. Processus ETL avec Azure Data Bricks :

Azure Databricks est une plateforme de traitement des données cloud qui simplifie et accélère le processus ETL (Extract, Transform, Load).







Microsoft Azure | Search resources, services, and docs (G+)

Home > Storage accounts > sabrilakestorage | Containers >

public-transport-data

Container

Search

Upload Add Directory Refresh Rename Delete Change tier Acquire lease Break lease Give feedback

Overview Diagnose and solve problems Access Control (IAM)

Authentication method: Access key (Switch to Azure AD User Account)
Location: public-transport-data / raw

Search blobs by prefix (case-sensitive) Show deleted objects

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
[...]						...
Archive						...
publicTransportData.csv	9/27/2023, 8:46:43 AM	Hot (Inferred)		Block blob	58.96 KiB	Available
SourceRaw-2020-01-01.csv	9/27/2023, 9:02:34 AM	Hot (Inferred)		Block blob	776.83 KiB	Available
SourceRaw-2021-01-01.csv	9/27/2023, 9:02:48 AM	Hot (Inferred)		Block blob	774.88 KiB	Available
SourceRaw-2022-01-01.csv	9/27/2023, 9:02:59 AM	Hot (Inferred)		Block blob	774.56 KiB	Available
SourceRaw-2023-01-01.csv	9/27/2023, 9:03:10 AM	Hot (Inferred)		Block blob	774.68 KiB	Available

Job name: DataGeneratorScript

Schedule: Scheduled (Every Day at 23:59 (UTC+01:00) Casablan...)

Cluster: Aymane SABRI's Cluster (8 GB · 4 Cores · DBR 13.3 LTS · Spark 3.4.1 · Scala 2.12)

Parameters: + Add

Alerts: asabri.ext@simplonform... Start Success Failure Add

Create

Microsoft Azure | databricks | Search data, notebooks, recents, and more... + P

sabridatabricksworkspace asabri.ext@simplonformations.o... v

Workflows

+ New

Workspace Recents Catalog Workflows Compute SQL > Data Engineering

Jobs Job runs

Filter jobs Only jobs owned by me Create job

Name	Tags	Created by	Trigger	Last run	⋮
DataTransformation		Aymane SABRI	Scheduled	Failed	⋮
DataGeneratorScript		Aymane SABRI	Scheduled		⋮



Job details

Job ID 870954637039866

Creator Aymane SABRI

Run as Aymane SABRI

Tags

Git

Repository URL github.com/AymaneSab/Azure_DataBricks

Branch AzureDataBricks

Edit Git settings

Remove Git settings

Schedule

Every hour (UTC+01:00 — Casablanca, Monrovia)

Edit schedule

Pause

Delete

Compute

Aymane SABRI's Cluster

Driver: Standard_F4 · Workers: Standard_F4 · 0 workers · 13.3 LTS
(includes Apache Spark 3.4.1, Scala 2.12)

View details

Swap

Spark UI

Logs

Metrics

Job run settings

Maximum concurrent runs

1

Edit concurrent runs

Duration thresholds

