



# Solution Cloud Azure

SYNAPSE ANALYTICS : DATA FACTORY

IA-DS | Big Data | 06/06/2021

**Par : Benidiri Lahlou**

# Introduction

In this work, we will present the approach followed to set up a tool that serves to make a capture on a source of covid data in every 15 minutes according to the request of the customer (Streaming). Then we thought of putting the data in a Power BI Dashboard. The purpose of which is to share a report with relevant analysis before the completion of this interval to meet the expectations of the client.

## Data source:

<https://www.data.gouv.fr/fr/datasets/r/d3a98a30-893f-47f7-96c5-2f4bcaaaod71>

# Architecture of the solution

## SOLUTION OPTED

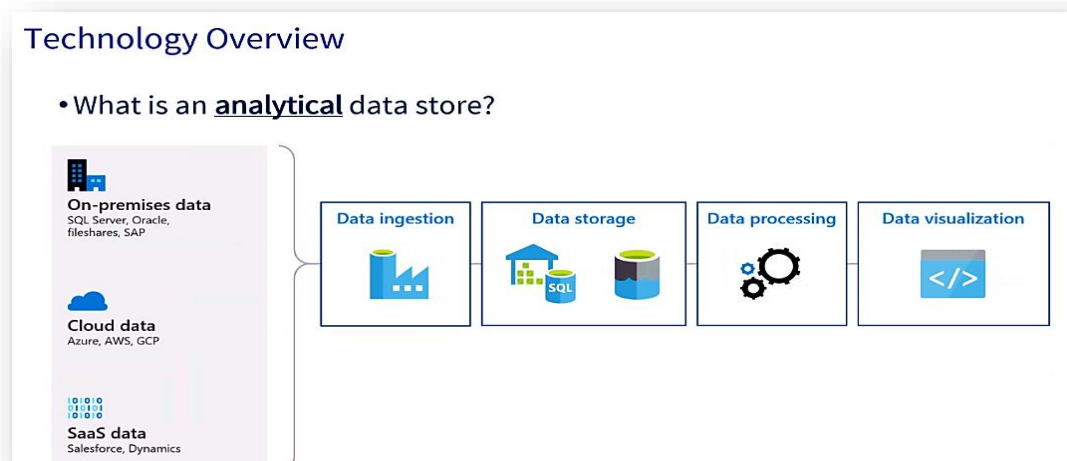
We explored the Data Factory resource of the Synapse Analytics Azure environment to best address the problem.

Before presenting the approach followed, we prefer to first present the associated mechanism.

We thought of creating a Data Pipeline, as shown above:

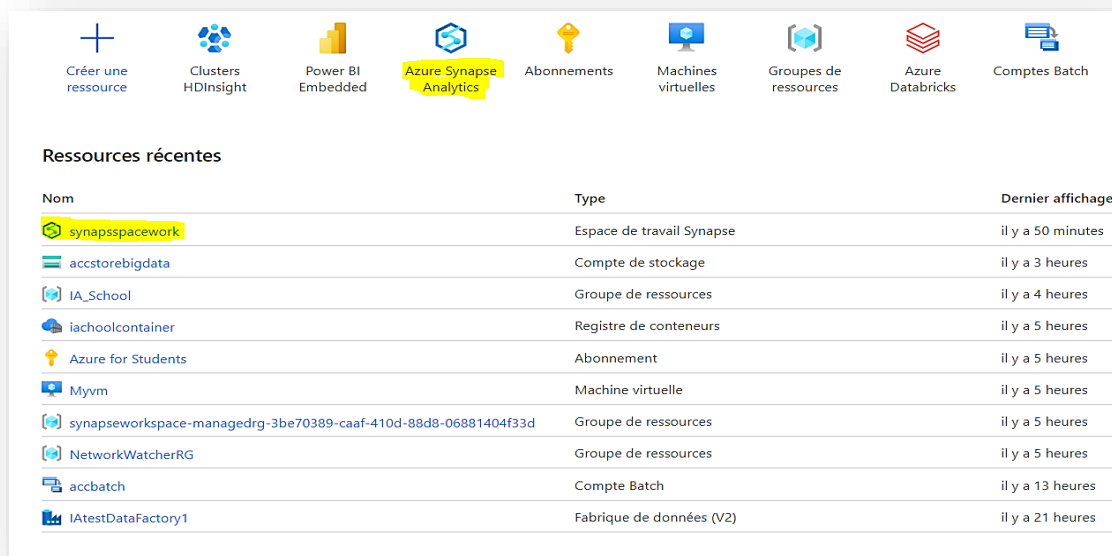












Concerning data ingestion, we are inspired by the following architecture:



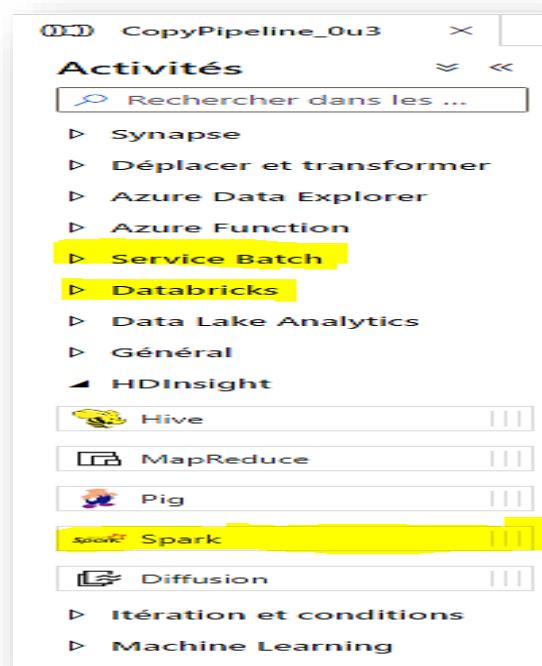
The idea is to create a trigger that will capture the data from the source, then store it in a blob of azure, (the data is updated in every 15 minutes). In order to create a report.

At first, we thought of opening Data Factory resource to create the pipeline, but we were blocked at the end because this resource does not have Power BI, then we opened synapse Analytics of Azure which has a data Factory functionality and other processing resources.



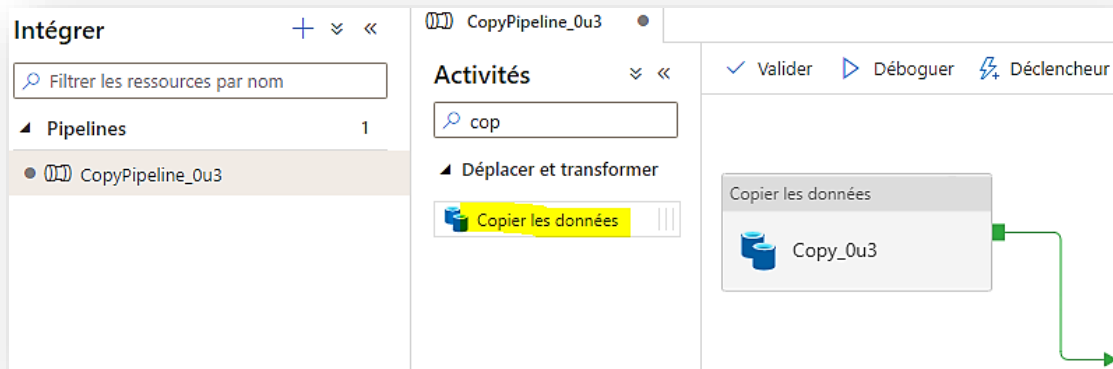
Ressources récentes		
Nom	Type	Dernier affichage
 synapsspacework	Espace de travail Synapse	il y a 50 minutes
 accstorebigdata	Compte de stockage	il y a 3 heures
 IA_School	Groupe de ressources	il y a 4 heures
 iachoolcontainer	Registre de conteneurs	il y a 5 heures
 Azure for Students	Abonnement	il y a 5 heures
 Myvm	Machine virtuelle	il y a 5 heures
 synapsspacework-managedrg-3be70389-caaf-410d-88d8-06881404f33d	Groupe de ressources	il y a 5 heures
 NetworkWatcherRG	Groupe de ressources	il y a 5 heures
 accbatch	Compte Batch	il y a 13 heures
 IAtestDataFactory1	Fabrique de données (V2)	il y a 21 heures

In the synapse studio, then in the integration section.



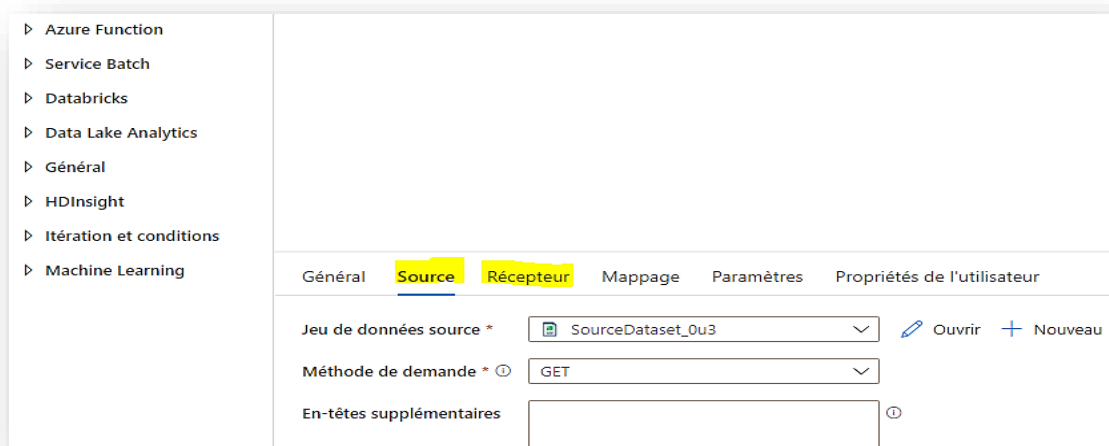
We clicked on create Pipeline, writing on the research box: Copy data.

Switching the resource to the workspace. We have created the pipeline.



In the next chapter we will explain the data ingestion part.

## INGESTION DE DONNEES



First, we created a connection with the source that we named HttpServer2 via source. Then we configured the receiver in the Blob via receiver. Knowing that we have defined Blob Azure as the source of storage.

The figure below shows the details of this configuration.

Connexion	
Service lié *	HttpServer2 [Tester la connexion] [Modifier]
Runtime d'intégration *	AutoResolveIntegrationRuntime [Modifier]
URL de base	https://www.data.gouv.fr/fr/datasets/r/d3a! [Aperçu des données]
URL relative ⓘ	
Type de compression	Aucun
Séparateur de colonne ⓘ	Virgule (,) [Modifier]
Délimiteur de ligne ⓘ	Par défaut (\r,\n ou \r\n) [Modifier]
Encodage	Par défaut(UTF-8)
Caractère d'échappement	Barre oblique inverse (\) [Modifier]
Guillemet	Guillemet double (") [Modifier]

This configuration is based on a runtime (tool responsible for executions) that we have created before.

You can see how to create httpServer2 on the figure below:

Microsoft Azure | Synapse Analytics | synapsspacework

Outil Copier des données

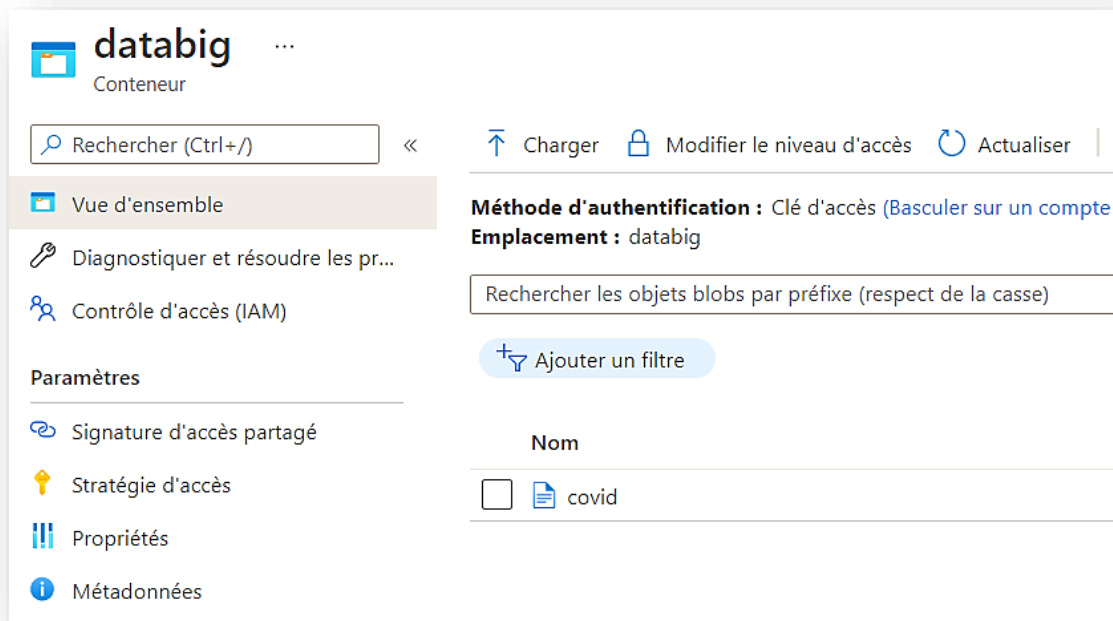
Source → HTTP → Stockage Blob Azure

Déploiement ...

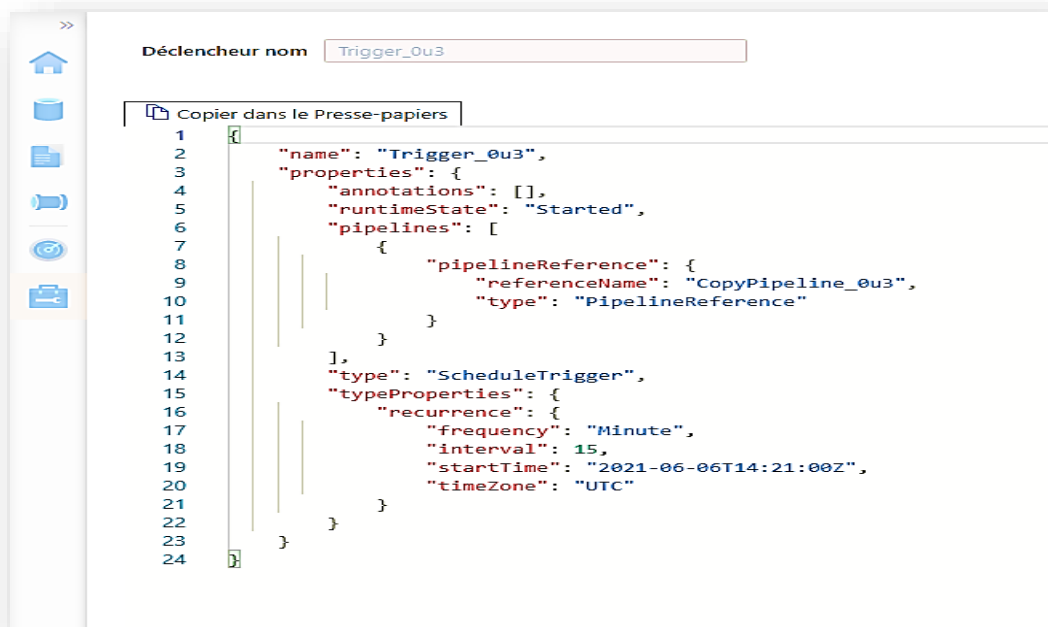
▸ Valider l'environnement du runtime de copie ✓

Étape de déploiement	État
> Création de jeux de données	Opération réussie ✓
> Création de pipelines	En cours ⌚
> Création de déclencheurs	En attente ⌚
> Démarrage des déclencheurs	En attente ⌚

The covid data will be stored in a Blob container in Azure.

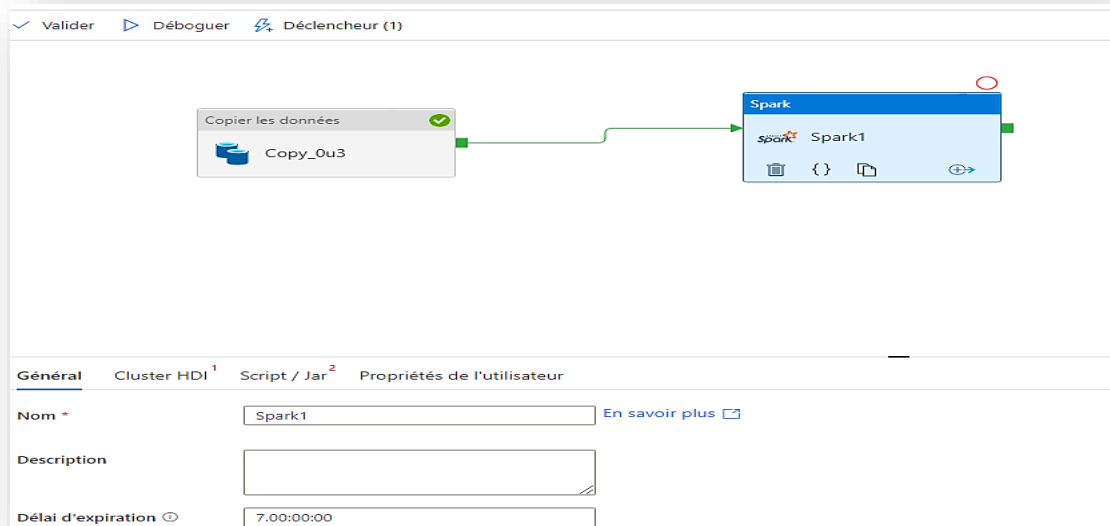


Below is the code of the Trigger that we have set for intervals of 15 minutes.



## DATA PROCESSING

To process data in real time, we can use Spark or Azure functions to include custom codes and even python scripts to process it before it is consumed.



## DISPLAY OF RESULTS

Output of pipeline executions.

Exécutions de pipeline

Déclenché ... Réexécuter Annuler Actualiser Modif Liste Gantt

Rechercher par ID ou no... Heure locale : Dernières 24 heures Nom du pipeline : Tous État : Tous Exécutions : Dernières exécutions

Affichage des éléments 1-33

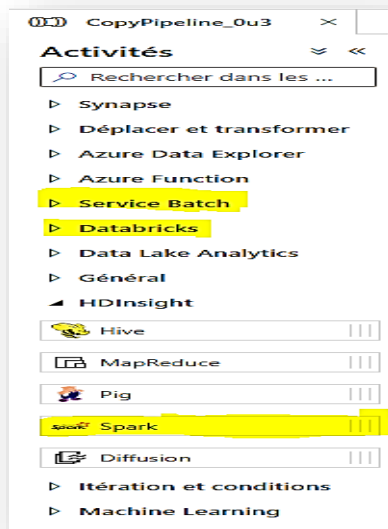
<input type="checkbox"/> Nom du pipeline	Début de l'exécution ↑↓	Fin de l'exécution	Durée	Déclenchement efficace	État
<input type="checkbox"/> CopyPipeline_Ou3	6/6/21, 21:06:00	6/6/21, 21:06:13	00:00:12	Trigger_Ou3	Opération réussie
<input type="checkbox"/> CopyPipeline_Ou3	6/6/21, 20:51:00	6/6/21, 20:51:11	00:00:11	Trigger_Ou3	Opération réussie
<input type="checkbox"/> CopyPipeline_Ou3	6/6/21, 20:36:00	6/6/21, 20:37:09	00:01:08	Trigger_Ou3	Opération réussie
<input type="checkbox"/> CopyPipeline_Ou3	6/6/21, 20:21:00	6/6/21, 20:21:10	00:00:10	Trigger_Ou3	Opération réussie
<input type="checkbox"/> CopyPipeline_Ou3	6/6/21, 20:06:01	6/6/21, 20:06:12	00:00:11	Trigger_Ou3	Opération réussie
<input type="checkbox"/> CopyPipeline_Ou3	6/6/21, 19:51:00	6/6/21, 19:51:12	00:00:11	Trigger_Ou3	Opération réussie
<input type="checkbox"/> CopyPipeline_Ou3	6/6/21, 19:35:59	6/6/21, 19:36:12	00:00:12	Trigger_Ou3	Opération réussie
<input type="checkbox"/> CopyPipeline_Ou3	6/6/21, 19:21:00	6/6/21, 19:21:11	00:00:11	Trigger_Ou3	Opération réussie
<input type="checkbox"/> CopyPipeline_Ou3	6/6/21, 19:06:01	6/6/21, 19:06:12	00:00:11	Trigger_Ou3	Opération réussie
<input type="checkbox"/> CopyPipeline_Ou3	6/6/21, 18:51:00	6/6/21, 18:51:10	00:00:10	Trigger_Ou3	Opération réussie

As you can see, the data is updated every 15 minutes in the Blob.



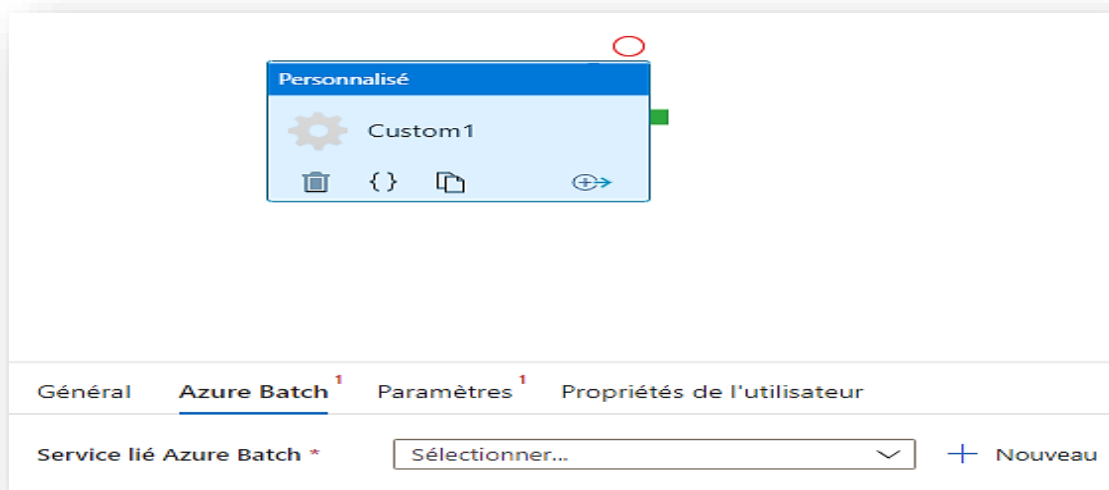
## ORCHESTRATION

We also tried to create an azure batch, we also created an azure batch and synapse link, from the Data Factory workspace, via the Service Batch section. But our choice fell on the Data Copy tool.



It is also possible to use data bricks in addition to Spark and Hadoop. To process the data. In our case the dataset is not large, it does not require l'utilisation de ce genre d'outil.

## OTHER ALTERNATIVE: AZURE BATCH

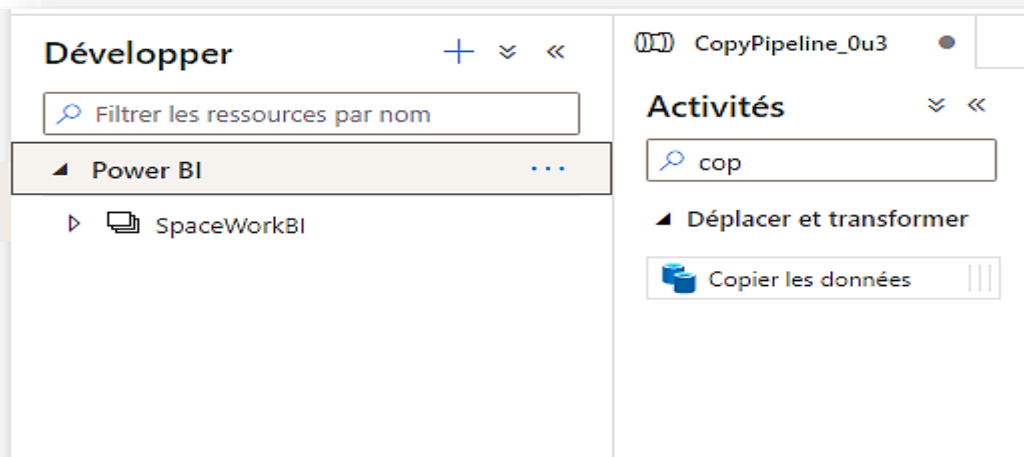


There are other tools that are able to do the same job like MongoDB, AirFlow...

## VISUALISATION

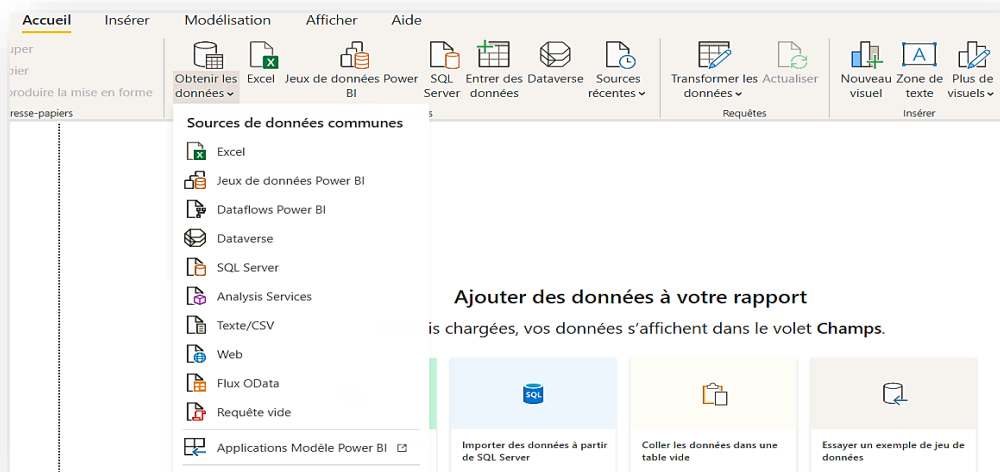
After installing Power BI desktop, and create a link with Synapse Analytics by managing resource access and roles.

We have created a connection below:



## ACCURACY:

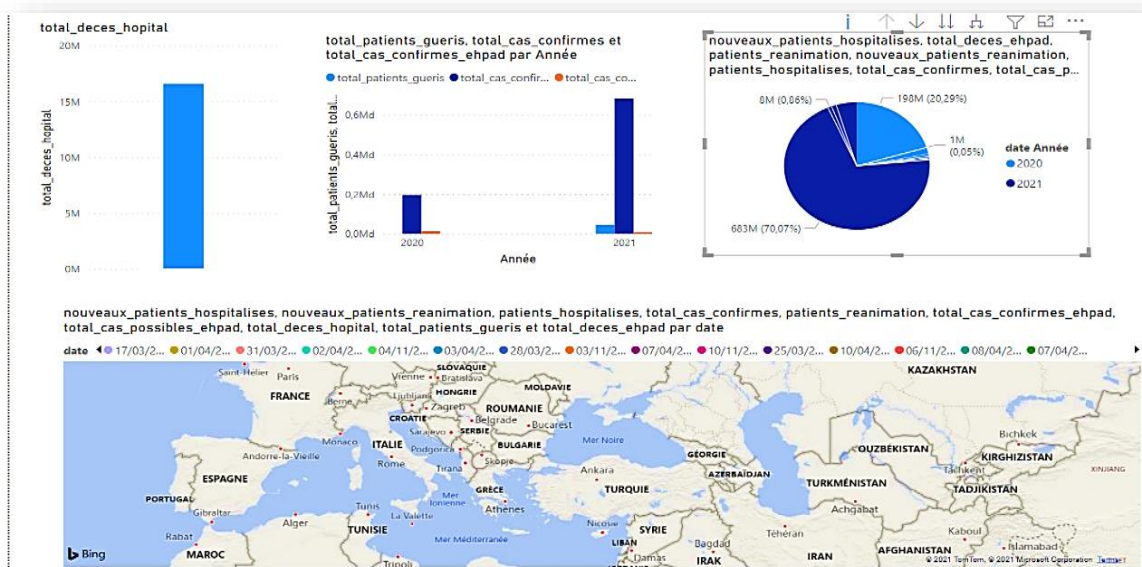
We didn't explore the Power BI resource of synapse Analytics to the end, (access problem that we don't have enough time to solve), but we found an alternative. That of connecting Power BI and the Blob store via the Get Data button.



Then the connection with the Blob



## DASHBOARD



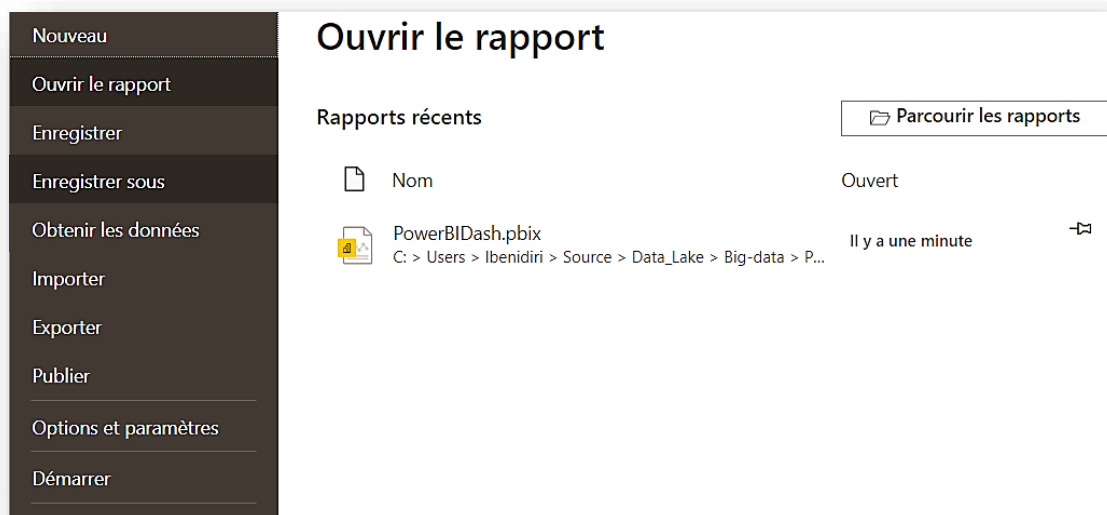
## DATA ANALYSIS

In Dashboard, we observe on pie chart that 70% of confirmed cases appeared in 2021, while in 2020 it does not exceed 30%, we also read on the histogram the total number of deaths that exceed 16 million in 2021.

On the map we can see that Asian countries and Latin America, and many countries in Europe are very affected by the pandemic.

## POWER BI (COLLABORATIVE TOOL)

When the report is created on the Power Editor, containing the analyses, you must publish it on the power BI account.

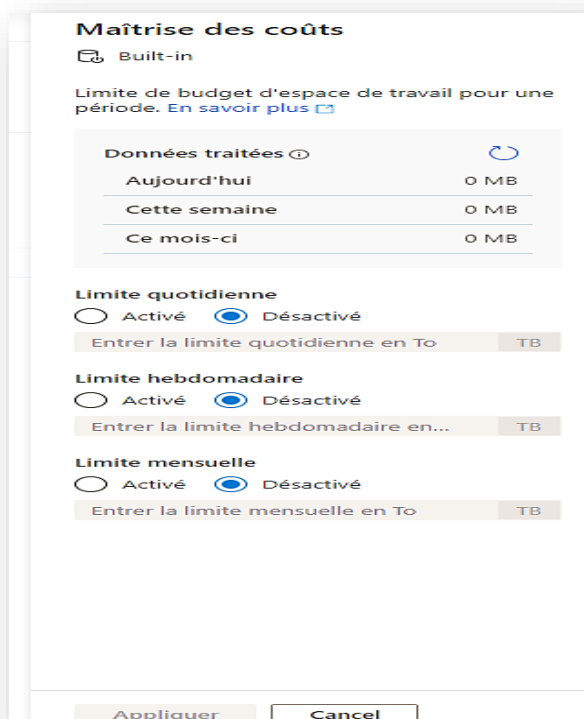


The report will be shared with the client, knowing that he has already been added as a member in read mode



Note: to manage the accesses, you have to go to shared with me, then access and roles.

## LOAD MANAGEMENT AND COST



On the synapse Analytics Cloud the user can control the costs.

You can also create a Spark cluster, Spark SQL pool for processing relational tables.

There is also a Spark notebook, and even SQL scripting, in addition to Data Factory, making this environment a complete space for data mining. Which you can associate with Power BI either integrated in this resource or even if you have it locally.

Regarding the total cost: with a free account of 100 euros offered for students we were able to finish the work without even consuming half of the amount.

## Conclusion

The Cloud solution fully meets our problem with the lowest cost if we master it, the power of calculation and agility, ease of use (without code).

The mechanism we have proposed is far from being complicated. Even if there are other architectures that rely on other tools such as Analytics streaming from Azure, which also responds to this kind of problem.

Data Factory from synapse Analytics is a space dedicated to data scientists, which offers a full range of data processing features to create execution pipelines.

Power BI allows us to interact with the data and then share it with the customer in an agile manner.

## References

<https://www.sqlshack.com/azure-data-factory-interview-questions-and-answers/>

<https://docs.microsoft.com/fr-fr/azure/data-factory/>

<https://www.hitachi-solutions.fr/blog/2021/06/transformez-votre-data-warehouse-avec-azure-synapse-analytics/>