

Conception, Développement, et Exploitation d'un Data Lake pour une Entreprise Digitale

Partie 1 : Conception du Data Lake

Analyse et Conception

Description :

- Faire l'analyse des besoins en données de l'entreprise et faites-en la conception logique et physique d'un data Lake

Livrables :

- Un document de conception détaillé décrivant l'architecture du Data Lake.
- Un diagramme des flux de données (comment les données sont collectées, stockées et traitées).
- Choix des technologies pour le stockage (HDFS, Amazon S3 ou le system si le travail se fait sur une machine locale).

C'est quoi un Data Lake?

Un **data Lake** est un grand espace de stockage où l'on dépose toutes sortes de données, qu'elles soient **brutes** ou **structurées**, dans leur **format d'origine**. Contrairement aux bases de données traditionnelles, il n'impose pas de structure préalable aux informations, offrant ainsi une **flexibilité** maximale pour **analyser** ou **exploiter** ces données plus tard. Il permet donc aux entreprises de **centraliser** leurs données, de les **conserver à moindre coût**, et de les préparer pour des **analyses** futures, tout en gardant une grande capacité **d'adaptation** aux besoins changeants.

- Vaste référentiel centralisé
- Il stocke des données structurées et non structurées
- Données stockées dans un format natif

Bénéfices d'un Data Lake

Les Data Lake permet une solution :

- **Centralisée / Accessible**

Permet à différentes équipes d'accéder aux mêmes données dans un cas d'usage différent

- **Flexible**

La data peut être stockée sous sa forme d'origine (CSV, JSON, Binaire etc...)

Il sera donc possible de récupérer des données et effectuer des transformations à la volée.

- **Évolutive**

Le data Lake est conçu pour gérer énormément de données très facilement

- **Rentable**

Les plateformes cloud offrent des solutions de stockage efficaces. Le **Data Lake** est une solution pour **stocker** et **analyser** d'énormes quantités de données.

ETL VS ELT

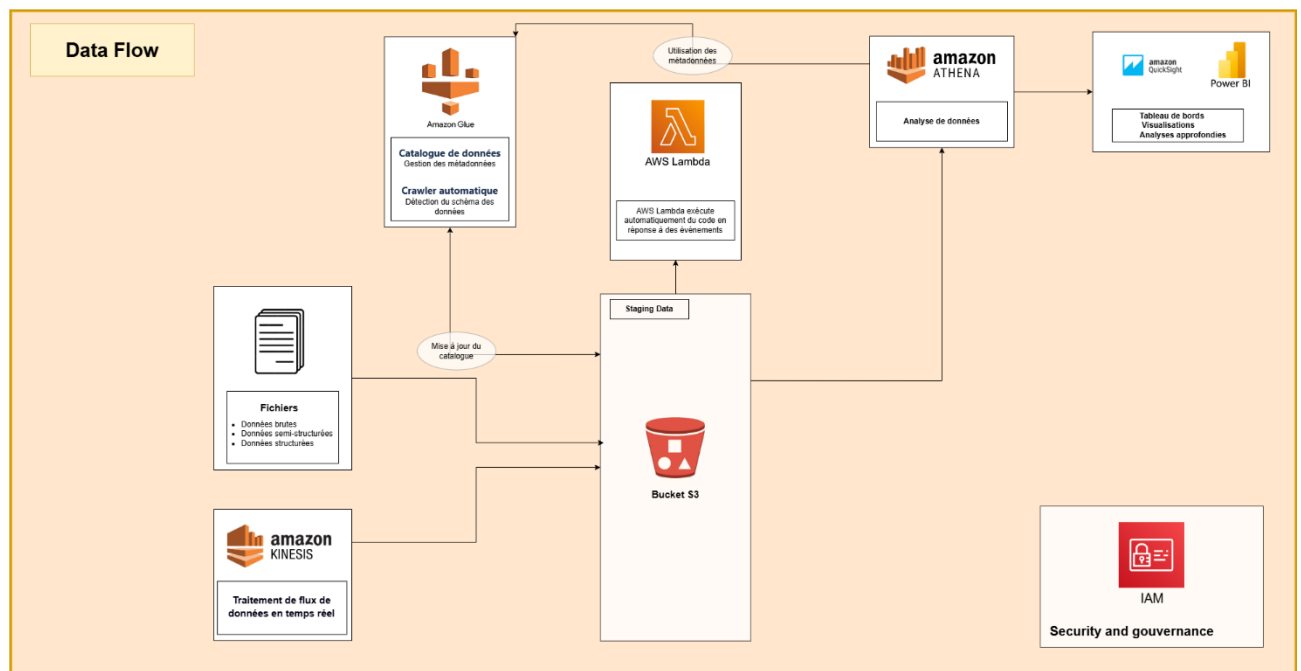
ETL (Extract, Transform, Load) approche traditionnelle où la donnée est **transformée avant d'être chargée** dans un Data Warehouse par exemple.

ELT (Extract, Load, Transform) approche plutôt utilisée pour les Data Lake où les données sont d'abord **chargées dans leurs formats d'origines et ensuite transformées**.

Schema-on-read

L'approche typique des data lakes, où la structure des données n'est définie qu'au moment de la lecture. Les données sont stockées sous leur forme brute, et leur schéma est appliqué uniquement lors de l'extraction et de l'analyse.

Diagramme explicatif (selon nous)



Les outils utilisés

AWS S3 Buckets

- Stockage de base pour un Data Lake
- Différentes zones peuvent être implémentées en utilisant des buckets S3 distincts

AWS Glue

- Extrait les données de différentes sources
- Catalogue et classe les données entrant dans le Data Lake
- Centralise le répertoire des métadonnées
- Peut appliquer des transformations aux données

Création de l'infrastructure

Description :

- Implémentation de l'infrastructure du Data Lake sur une machine locale ou sur une plateforme Cloud (AWS, Azure, ou Google Cloud).
- Utilisation de clusters Hadoop ou de solutions similaires pour le stockage.
- Utilisation de Spark/Kafka/Kafka stream pour le traitement distribué ou une solution similaire.

Livrables :

- Scripts pour déployer l'infrastructure.
- Documentation technique sur les choix de technologies et de solutions Cloud.

Choix des technologies :

AWS Lambda

- Service sans serveur pour l'exécution de code sans provisionner de serveurs
- Utilisable pour des tâches d'ingestion et de traitement de données spécifiques

Amazon Athena

- Permet de requêter et analyser directement les données depuis des buckets S3
- Idéal pour les requêtes ad hoc, l'exploration et l'analyse des données
- Offre des requêtes SQL performantes
- Évite le besoin de charger les données dans des bases de données ou entrepôts de données distincts

Kinesis (Ingestion en continu)

- Permet une ingestion des données en temps réel
- Idéal pour les données sensibles au temps
- Mise en œuvre à l'aide de services tels qu'Amazon Kinesis pour les données en continu.

AWS IAM (Identity and Access Management)

- Gestion centralisée des permissions et de la sécurité
- Contrôle l'accès aux ressources AWS
- Permet la mise en œuvre de politiques spécifiques pour sécuriser les données sensibles

AWS Step Functions

- Orchestration des workflows
- Gère l'exécution séquentielle ou parallèle des services AWS comme Lambda, Glue ou SageMaker
- Simplifie les pipelines de données complexes avec des états bien définis

Amazon SageMaker

- Permet de créer, former et déployer des modèles de machine learning
- Peut utiliser les données du S3 bucket pour entraîner des modèles
- S'intègre avec Kinesis pour les prédictions en temps réel

Power BI

- Plateforme d'analyse et de visualisation des données.
- Peut se connecter directement à S3 ou à des bases de données compatibles AWS via Athena comme nous allons le présenter
- Peut être intégré avec les modèles produits par SageMaker pour des visualisations avancées

Partie 2 : Ingestion et Transformation des Données

Ingestion de Données Brutes

Description :

- Mise en place de l'ingestion des données provenant de différentes sources : bases de données SQL, logs, ux de données en temps réel (via Apache Kafka, AWS Kinesis).
- Mise en place d'un pipeline d'ingestion pour des données structurées, semistructurées et non structurées.
- Pour chaque type de données, générez un exemple de jet de données pour illustrer l'exemple

Livrables :

- Pipeline d'ingestion en temps réel ou batch pour les différentes sources de données.
- Exemple de datasets ingérés dans le Data Lake.

Utilisation de AWS Kinesis Streaming

Nous allons générer un stream, mais d'abord nous avons besoin de créer un IAM role

The image shows two screenshots from the AWS IAM console. The top screenshot displays the 'Roles' page with a list of existing roles and a 'Create role' button highlighted with a red arrow. The bottom screenshot shows the 'Create role' wizard, specifically the 'Select trusted entity' step. In this step, the 'AWS service' option is selected, and the 'Service or use case' dropdown is set to 'CloudFormation', both highlighted with red arrows. The 'Next' button at the bottom right is also highlighted with a red arrow.

Roles (5) Info

An IAM role is an identity you can create that has specific permissions with credentials that are valid for short durations. Roles can be assumed by entities that you trust.

Role name	Trusted entities	Last activity
AWSGlueServiceRole-TestProject1	AWS Service: glue	35 minutes ago
AWSServiceRoleForOrganizations	AWS Service: organizations (Service-Linked Role)	-
AWSServiceRoleForSupport	AWS Service: support (Service-Linked Role)	-
AWSServiceRoleForTrustedAdvisor	AWS Service: trustedadvisor (Service-Linked Role)	-
myfirsteventfunction-role-qlfsgy	AWS Service: lambda	1 hour ago

Roles Anywhere Info

Authenticate your non-AWS workloads and securely provide access to AWS services.

Access AWS from your non-AWS workloads

Operate your non-AWS workloads using the same authentication and authorization strategy that you use within AWS.

X.509 Standard

Use your own existing PKI infrastructure or use AWS Certificate Manager Private Certificate Authority to authenticate identities.

Temporary credentials

Use temporary credentials with ease and benefit from the enhanced security they provide.

Step 1: Select trusted entity

Select trusted entity Info

Trusted entity type

- ☒ **AWS service**
Allow AWS services like EC2, Lambda, or others to perform actions in this account.
- ☐ **AWS account**
Allow entities in other AWS accounts belonging to you or a 3rd party to perform actions in this account.
- ☐ **Web identity**
Allow users federated by the specified external web identity provider to assume this role to perform actions in this account.
- ☐ **SAML 2.0 federation**
Allow users federated with SAML 2.0 from a corporate directory to perform actions in this account.
- ☐ **Custom trust policy**
Create a custom trust policy to enable others to perform actions in this account.

Use case

Allow an AWS service like EC2, Lambda, or others to perform actions in this account.

Service or use case

CloudFormation

Choose a use case for the specified service.

Use case

- ☒ **CloudFormation**
Allows CloudFormation to create and manage AWS stacks and resources on your behalf.

Cancel Next

[IAM](#) > [Roles](#) > Create role

Step 1: [Select trusted entity](#)

Step 2: **Add permissions**

Step 3: Name, review, and create

Add permissions 44

Permissions policies (1/577) 44

Choose one or more policies to attach to your new role.

Filter by Type: All types

Policy name	Type	Description
<input checked="" type="checkbox"/> AdministratorAccess	AWS managed - job function	Provides full access to AWS services and resources.
<input type="checkbox"/> AdministratorAccess-Ampify	AWS managed	Grants account administrative permissions while explicitly allowing direct access to resources needed by Ampify appl...
<input type="checkbox"/> AdministratorAccess-AmazonElasticBeanstalk	AWS managed	Grants account administrative permissions. Explicitly allows developers and administrators to gain direct access to res...
<input type="checkbox"/> AlexaForBusinessDeviceSetup	AWS managed	Provides device setup access to AlexaForBusiness services.
<input type="checkbox"/> AlexaForBusinessFullAccess	AWS managed	Grants full access to AlexaForBusiness resources and access to related AWS Services.
<input type="checkbox"/> AlexaForBusinessGatewayExecution	AWS managed	Provides gateway execution access to AlexaForBusiness services.
<input type="checkbox"/> AlexaForBusinessProfileOutgoingAccountPolicy	AWS managed	Provides access to Lifesize AWS devices.
<input type="checkbox"/> AlexaForBusinessProfileOutgoingAccountPolicy	AWS managed	Provides access to Poly AWS devices.
<input type="checkbox"/> AlexaForBusinessProfileOutgoingAccountPolicy	AWS managed	Provides read only access to AlexaForBusiness services.
<input type="checkbox"/> AmazonAPIGatewayAdminister	AWS managed	Provides full access to create/update APIs in Amazon API Gateway via the AWS Management Console.
<input type="checkbox"/> AmazonAPIGatewayInvokeFullAccess	AWS managed	Provides full access to invoke APIs in Amazon API Gateway.
<input type="checkbox"/> AmazonAPIGatewayPushToCloudWatchLogs	AWS managed	Allows API Gateway to push logs to user's account.
<input type="checkbox"/> AmazonAppFlowFullAccess	AWS managed	Provides full access to Amazon AppFlow and access to AWS services supported as flow source or destination (S3 and R...
<input type="checkbox"/> AmazonAppFlowReadOnlyAccess	AWS managed	Provides read only access to Amazon AppFlow flows.
<input type="checkbox"/> AmazonAppStreamFullAccess	AWS managed	Provides full access to Amazon AppStream via the AWS Management Console.
<input type="checkbox"/> AmazonAppStreamFullAccess	AWS managed	Amazon AppStream 2.0 access to AWS Certificate Manager Private CA in customer accounts for certificate-based auth...
<input type="checkbox"/> AmazonAppStreamReadOnlyAccess	AWS managed	Provides read only access to Amazon AppStream via the AWS Management Console.
<input type="checkbox"/> AmazonAppStreamServiceRole	AWS managed	Default policy for Amazon AppStream service role.
<input type="checkbox"/> AmazonAthenaFullAccess	AWS managed	Provides full access to Amazon Athena and scoped access to the dependencies needed to enable querying, writing, and...
<input type="checkbox"/> AmazonAugmentedAIReadOnlyAccess	AWS managed	Provides access to perform all operations Amazon Augmented AI resources, including FlowDefinitions, HumanTaskDefi...

Set permissions boundary - optional

Cancel Previous **Next**

[IAM](#) > [Roles](#) > Create role

Step 1: [Select trusted entity](#)

Step 2: [Add permissions](#)

Step 3: **Name, review, and create**

Name, review, and create

Role details

Role name: Enter a meaningful name to identify this role.

Description: Add more explanation for this role.

Maximum 64 characters. Use alphanumeric and hyphen (-) characters.

Maximum 1000 characters. Use letters (a-z and A-Z), numbers (0-9), tabs, new lines, or any of the following characters: ~, !, @, #, \$, %, ^, &, *, (,), =, +, ~, {, }, [,], |, \, ;, :, ", '.

Step 1: Select trusted entities Edit

Trust policy

```

1 {
2   "Version": "2012-10-17",
3   "Statement": [
4     {
5       "Effect": "Allow",
6       "Principal": {
7         "Service": "cloudformation.amazonaws.com"
8       },
9       "Action": "sts:AssumeRole"
10    }
11  ]
12 }
13

```

Step 2: Add permissions Edit

Permissions policy summary

Policy name	Type	Attached as
AdministratorAccess	AWS managed - job function	Permissions policy

Step 3: Add tags

Add tags - optional 44

Tags are key-value pairs that you can add to AWS resources to help identify, organize, or search for resources.

No tags associated with the resource.

[Add new tag](#)

You can add up to 50 more tags.

Cancel Previous **Create role**

Créer un data stream avec Kinesis

Amazon Kinesis services

Collect, process, and analyze data streams in real time.

How it works

Kinesis Data Streams

Amazon Data Firehose

Managed Apache Flink

Get started

Kinesis Data Streams

Collect streaming data with a data stream.

Amazon Data Firehose - new

Formerly Kinesis Data Firehose

Process and deliver streaming data with a Firehose stream.

Managed Apache Flink

Formerly Kinesis Data Analytics

Analyze streaming data with data science applications.

Create data stream

Pricing (US East (N. Virginia))

Amazon Kinesis Data Streams

Shards

\$0.015 per Hour

Amazon Kinesis > Data streams > Create data stream

Create data stream [info](#)

Data stream configuration

Data stream name

TestStream

Acceptable characters are uppercase and lowercase letters, numbers, underscores, hyphens and periods.

Data stream capacity [info](#)

Capacity mode

☐ On-demand

Use this mode when your data stream's throughput requirements are unpredictable and variable. With on-demand mode, your data stream's capacity scales automatically.

☒ Provisioned

Use provisioned mode when you can reliably estimate throughput requirements of your data stream. With provisioned mode, your data stream's capacity is fixed.

Provisioned shards

The total capacity of a stream is the sum of the capacities of its shards. Enter number of provisioned shards to see total data stream capacity.

1

Shard estimator

Minimum: 1, Maximum available: 500, Account quota limit: 500, Request shard quote increase

Total data stream capacity

Shard capacity is determined by the number of provisioned shards. Each shard ingests up to 1 MiB/second and 1,000 records/second and emits up to 2 MiB/second. If writes and reads exceed capacity, the application will receive throttles.

Write capacity

Maximum

1 MiB/second and 1,000 records/second

Read capacity

Maximum

2 MiB/second

Provisioned mode has a fixed-throughput pricing model. See Kinesis pricing for Provisioned mode

Data stream settings

You can edit the settings after the data stream has been created and is in the active status.

Setting	Value	Editable after creation
Capacity mode	Provisioned	Yes
Provisioned shards	1	Yes
Data retention period	1 day	Yes
Server-side encryption	Disabled	Yes
Monitoring enhanced metrics	Disabled	Yes
Data stream sharing policy	No policy	Yes

Tags - optional [info](#)

You can add tags to organize your AWS resources, track costs, and control access.

No tags associated with the Kinesis Data Stream.

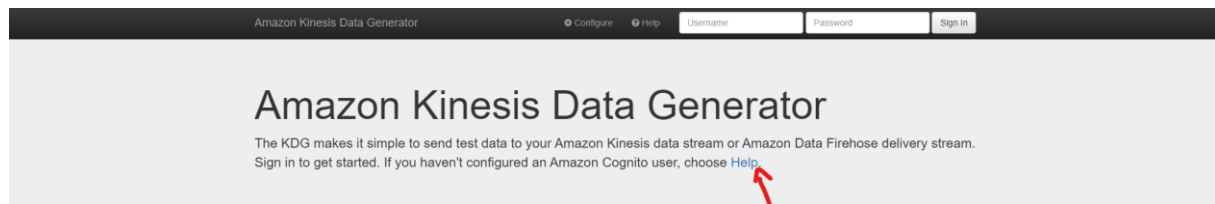
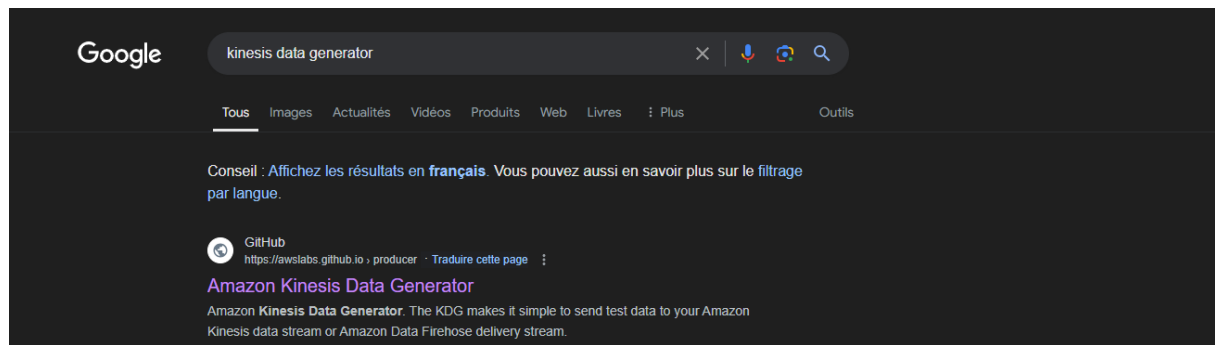
Add new tag

You can add up to 50 more tags.

Cancel

Create data stream

Utiliser un kinesis Data generator



creation wizard. You only need to provide a Username and Password for the user that you will use to log in to the KDG. Accept the defaults for any other options presented by CloudFormation.

Please note that the CloudFormation Template below is currently supported in the following regions:

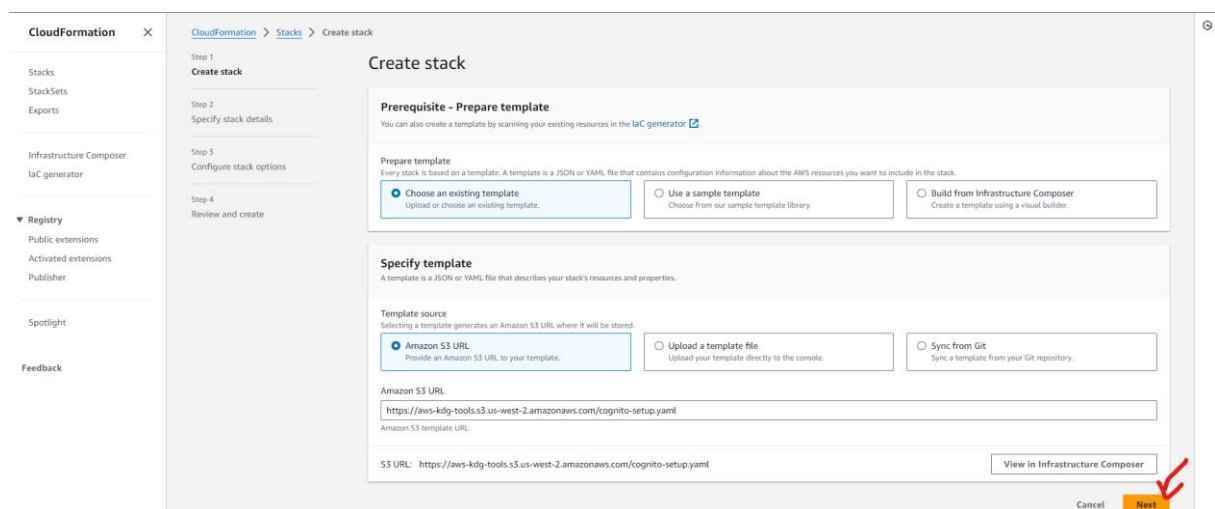
- ap-northeast-1 (Tokyo)
- ap-northeast-2 (Seoul)
- ap-south-1 (Mumbai)
- ap-southeast-1 (Singapore)
- ap-southeast-2 (Sydney)
- ca-central-1 (Canada Central)
- eu-west-1 (Ireland)
- eu-west-2 (London)
- eu-central-1 (Frankfurt)
- us-east-1 (N. Virginia)
- us-east-2 (Ohio)
- us-west-2 (Oregon)

[Create a Cognito User with CloudFormation](#)

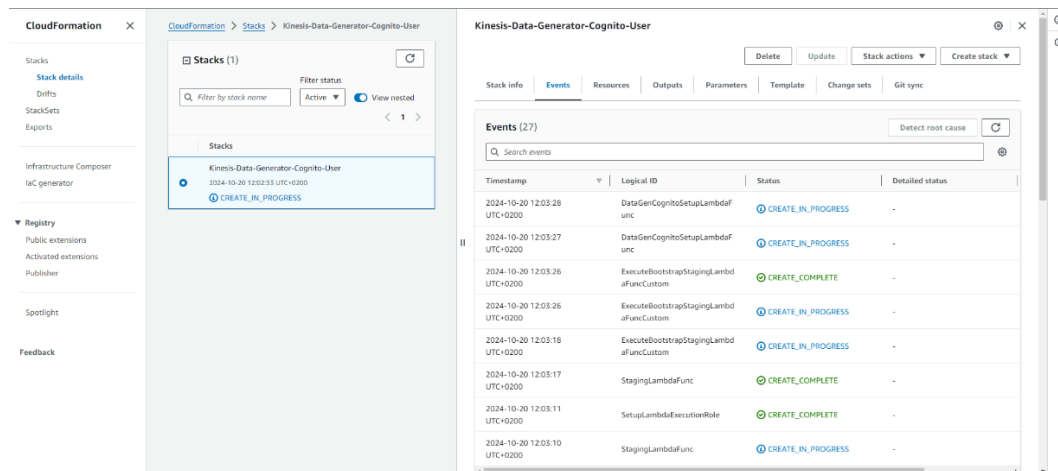
In addition to the above regions, **GovCloud** is also supported by manually importing the CloudFormation template in the following regions:

- us-gov-west-1 (Portland/PDT)

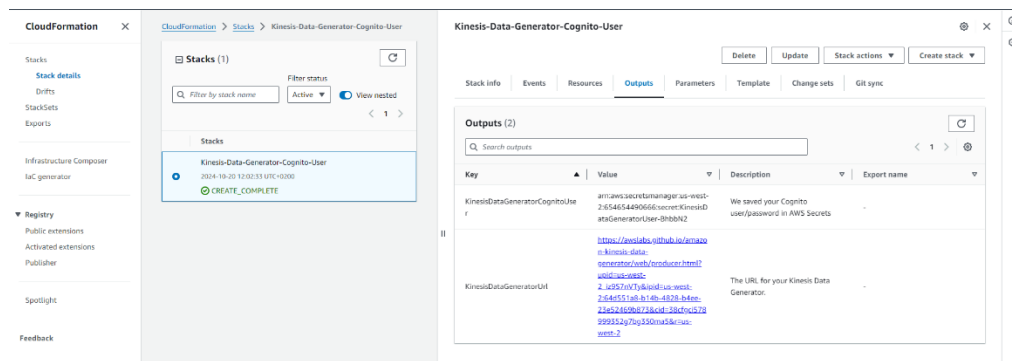
For manually installing KDG by CloudFormation:



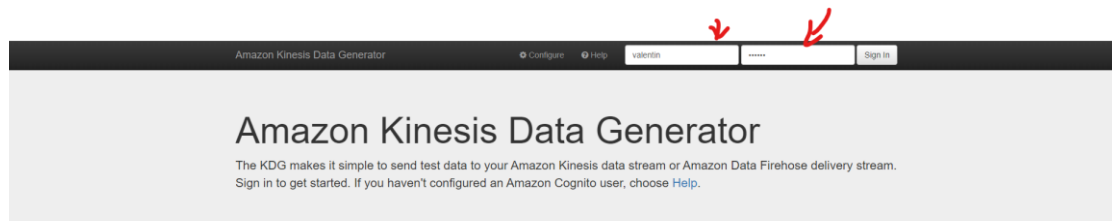
Tout a maintenant été créé



Nous pouvons maintenant aller dans l'onglet "Outputs", et cliquer sur le lien

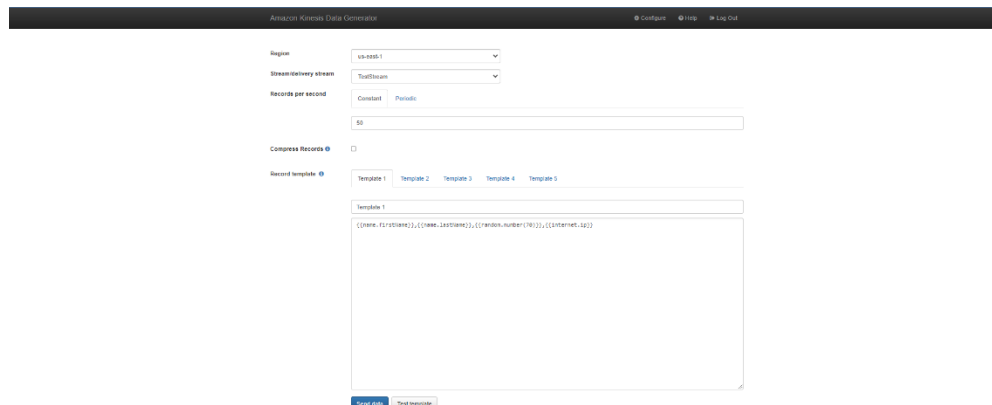


Se connecter avec les identifiants précédemment créés : valentinpuillandre@gmail.com

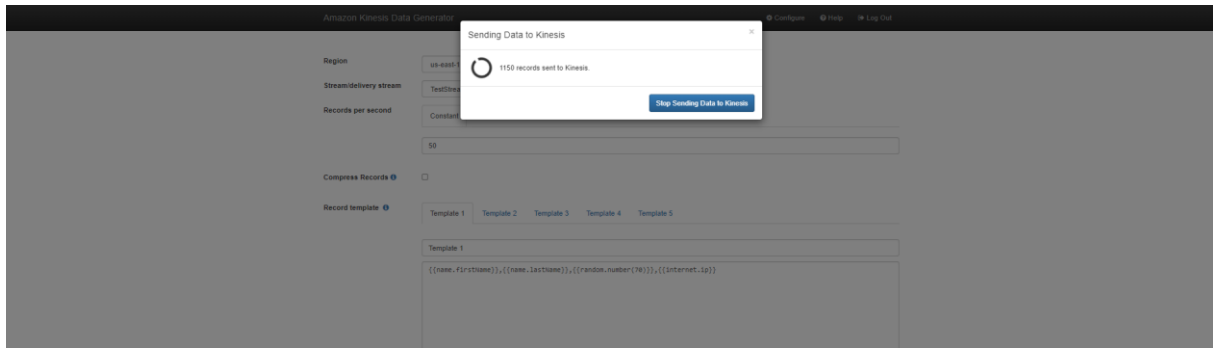


Données pour le template :

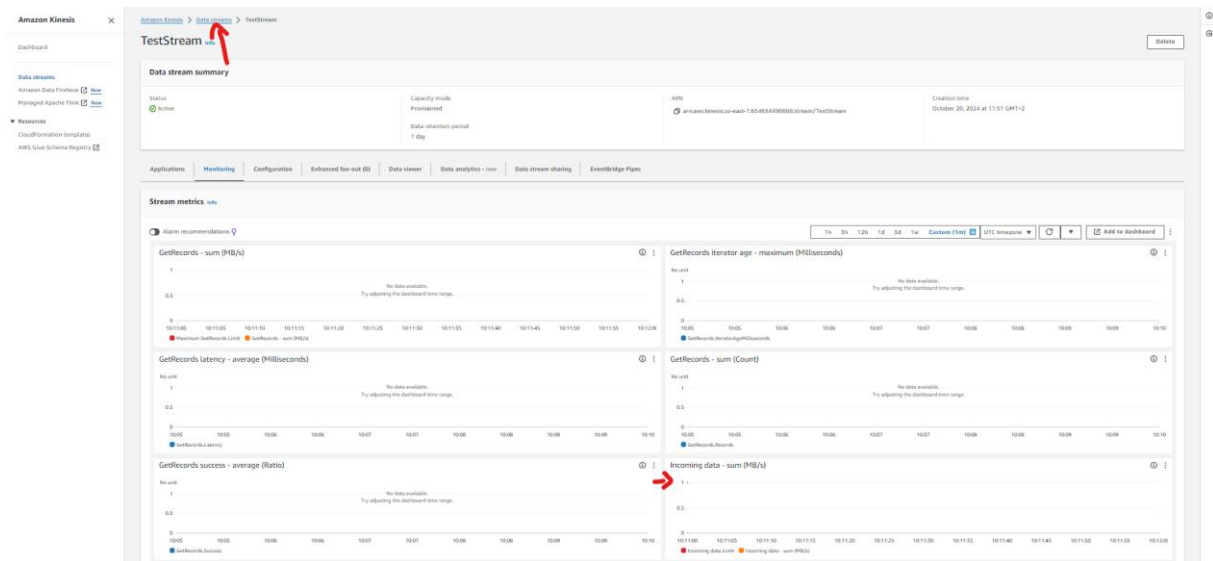
{{name.firstName}},{{name.lastName}},{{random.number(70)}},{{internet.ip}}



Les datas sont maintenant envoyées



En laissant un peu tourner on peut voir des données apparaître



Pipelines de Transformation des Données

Description :

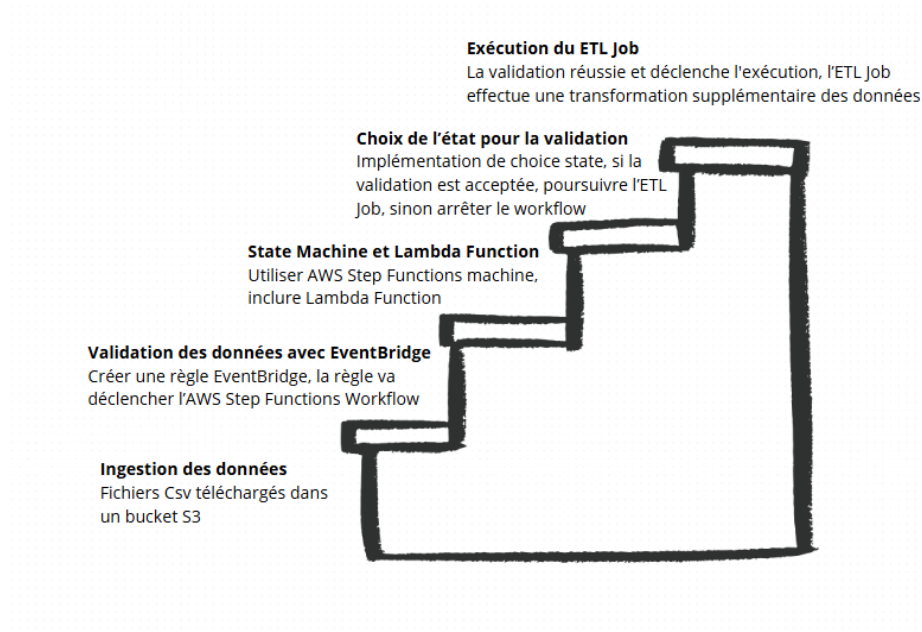
- Conception et développement de pipelines ETL (ou ELT) pour transformer et enrichir les données dans le Data Lake.
- Utilisation d'Apache Spark ou une technologie similaire pour effectuer ces transformations.

Livrables :

- Code Spark pour nettoyer et transformer les données, ou avec une autre technologie.
- Documentation sur les différentes transformations appliquées aux données.

L'objectif ici va être d'automatiser le data Workflow, automatiser l'ingestion et le traitement de la donnée.

Schéma du workflow



Créer Deux buckets

- sales-data-raw-17112024
- sales-data-processed-17112024

General purpose buckets			
Directory buckets			
General purpose buckets (13) info all aws regions			
Buckets are containers for data stored in S3.			
<input type="text" value="Find buckets by name"/>			
Name	AWS Region	IAM Access Analyzer	Creation date
backupperbucket2024new	US East (N. Virginia) us-east-1	View analyzer for us-east-1	November 11, 2024, 11:08:56 (UTC+01:00)
datastreamingdemo16112024	Europe (London) eu-west-2	View analyzer for eu-west-2	November 16, 2024, 11:23:57 (UTC+01:00)
my-datalake-bucket-2024	US East (N. Virginia) us-east-1	View analyzer for us-east-1	October 17, 2024, 11:36:01 (UTC+02:00)
myreplicationsource10112024	US East (Ohio) us-east-2	View analyzer for us-east-2	November 10, 2024, 16:46:20 (UTC+01:00)
myreplicationtarget10112024	US West (N. California) us-west-1	View analyzer for us-west-1	November 10, 2024, 16:48:58 (UTC+01:00)
newbucketbackuper2024	US East (N. Virginia) us-east-1	View analyzer for us-east-1	November 11, 2024, 10:53:37 (UTC+01:00)
sales-data-processed-17112024	Europe (Paris) eu-west-3	View analyzer for eu-west-3	November 17, 2024, 09:12:11 (UTC+01:00)
sales-data-raw-17112024	Europe (Paris) eu-west-3	View analyzer for eu-west-3	November 17, 2024, 09:11:28 (UTC+01:00)
source-bucket-valentin	US East (N. Virginia) us-east-1	View analyzer for us-east-1	October 20, 2024, 09:18:59 (UTC+02:00)
source-france-paris-16112024	Europe (Paris) eu-west-3	View analyzer for eu-west-3	November 16, 2024, 09:55:26 (UTC+01:00)
target-bucket-valentin	US East (N. Virginia) us-east-1	View analyzer for us-east-1	October 20, 2024, 09:19:26 (UTC+02:00)
target-france-paris-16112024	Europe (Paris) eu-west-3	View analyzer for eu-west-3	November 16, 2024, 09:56:30 (UTC+01:00)
valentin-09112024-test	Europe (Paris) eu-west-3	View analyzer for eu-west-3	November 9, 2024, 11:19:19 (UTC+01:00)

Etape 2: Préparer les fichiers

Fichiers présents dans mes dossiers locaux

sales_data_london.csv

sales_data_london_different_schema.csv

Etapes 3: Créer une Lambda function

- Validation des données
- IAM rôle : S3 access needed

Aller dans la partie **Lambda Function** et créer une fonction

Create function

Choose one of the following options to create your function.

☒ **Author from scratch**
Create a new function from scratch. World example.

☐ **Use a blueprint**
Build a Lambda application from sample code and configuration presets for common use cases.

☐ **Container image**
Select a container image to deploy for your function.

Basic information

Function name
Enter a name that describes the purpose of your function.

Function name must be 1 to 64 characters, must be unique to the Region, and can't include spaces. Valid characters are a-z, A-Z, 0-9, hyphens (-), and underscores (_).

Runtime
Choose the language to use to write your function. Note that the console code editor supports only Node.js, Python, and Ruby.
☒ **Python 3.12**

Architecture
Choose the instruction set architecture you want for your function code.
☒ **x86_64**

Permissions
By default, Lambda will create an execution role with permissions to upload logs to Amazon CloudWatch Logs. You can customize this default role later when adding triggers.

[Change default execution role](#)

Additional Configurations
Use additional configurations to set up code signing, function URL, tags, and Amazon VPC access for your function.

[Cancel](#) [Create function](#)

Mettre le code qui fait la validation des données (cf fichier ValidateDataFunction.py)

Successfully created the function data_validator. You can now change its code and configuration. To invoke your function with a test event, choose "test".

You are using the new console editor.

Code editor

```
1 import boto3
2 import csv
3 import os
4
5 # Initialize the S3 client
6 s3 = boto3.client('s3')
7
8 def lambda_handler(event, context):
9     # Extract bucket name and file key from the step functions input
10    bucket_name = event['detail']['bucket']['name']
11    file_key = event['detail']['object']['key']
12
13    # Download the file from S3
14    try:
15        response = s3.get_object(Bucket=bucket_name, Key=file_key)
16    except Exception as e:
17        return {
18            'statusCode': 404,
19            'error_message': f'Error getting object {file_key} from bucket {bucket_name}. Make sure they exist and your bucket is in the same region as the function.'
20        }
21
22    # Read the content of the file
23    content = response['Body'].read().decode('utf-8')
24    lines = content.splitlines()
25
26    # Check if the file is a CSV by attempting to read the first line as headers
27    try:
28        headers = next(csv.reader(lines))
29    except Exception as e:
30        return {
31            'statusCode': 400,
32            'error_message': f'Invalid CSV file: {str(e)}'
33        }
```

Code properties

Package size: 299 bytes

SHA256 hash: WAPuRtUJVEGjLwVtYgdvZbNtAGF912QbVv

Last modified: 45 seconds ago

Runtime settings

[Edit](#) [Edit runtime management configuration](#)

Configurer les bons droits pour utiliser cette Lambda Function

Successfully created the function DataValidate. You can now change its code and configuration. To invoke your function with a test event, choose "Test".

DataValidate

Throttle Copy ARN Actions

Export to Infrastructure Composer Download

Function overview

Diagram Template

+ Add trigger

+ Add destination

Description

Last modified: 2 minutes ago

Function ARN: arn:aws:lambda:eu-west-3:654654490666:function:DataValidate

Function URL

Code Test Monitor **Configuration** Aliases Versions

General configuration

Triggers

Permissions

Destinations

Function URL

General configuration

Description

Timeout: 0 min 3 sec

Memory: 128 MB

SnapStart

Ephemeral storage: 512 MB

Edit

Nous devons ajouter une nouvelle permission pour notre bucket

Lambda > Functions > DataValidate > Edit basic settings

Edit basic settings

Basic settings

Description - optional

Memory

128 MB

Ephemeral storage

512 MB

SnapStart

None

Timeout

0 min 3 sec

Execution role

Use an existing role

Existing role

service-role/DataValidate-role-n1fugeyg

Cancel Save

DataValidate-role-n1fugeyg

Summary

Creation date: November 17, 2024, 09:16 (UTC+01:00)

Last activity

ARN: arn:aws:iam::654654490666:role/service-role/DataValidate-role-n1fugeyg

Maximum session duration: 1 hour

Permissions Trust relationships Tags Last Accessed Revoke sessions

Permissions policies (1)

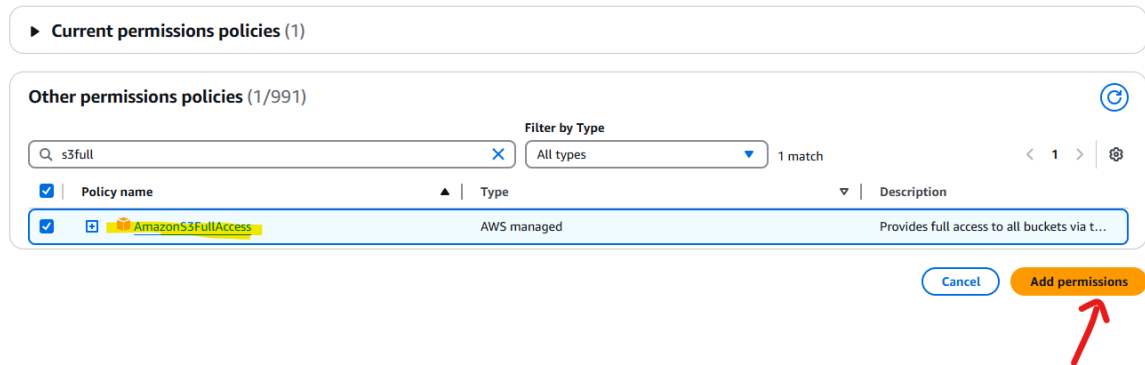
You can attach up to 10 managed policies.

Filter by Type: All types

Policy name	Type	Attached entities
AWSLambdaBasicExecutionRole-7235dd15-9ad...	Customer managed	1

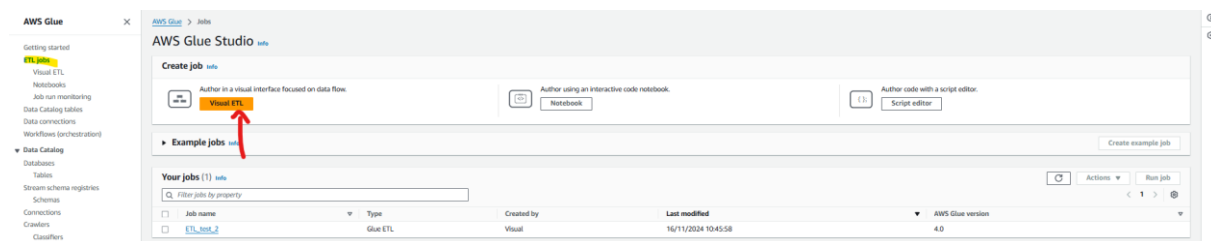
Donner les droits en full access au S3 (ne pas faire en production)

Attach policy to DataValidate-role-n1fugeyg

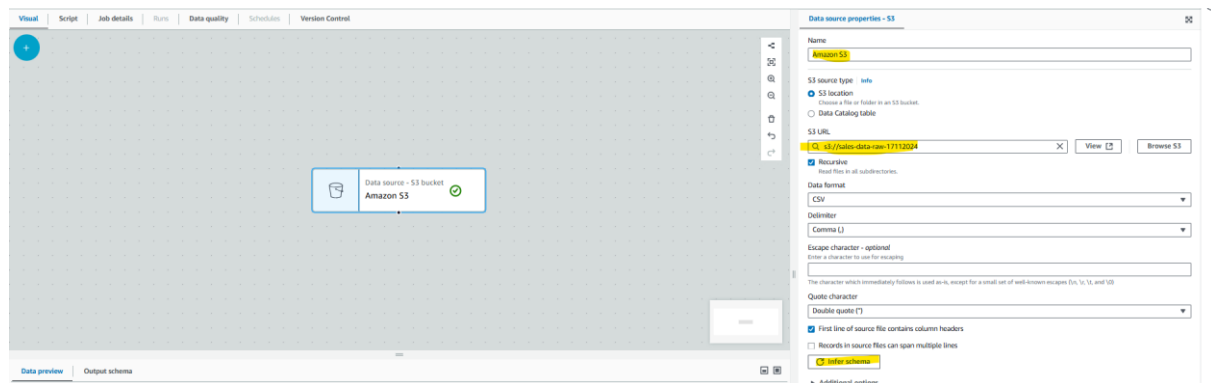


Etape 4: Créer un Job ETL

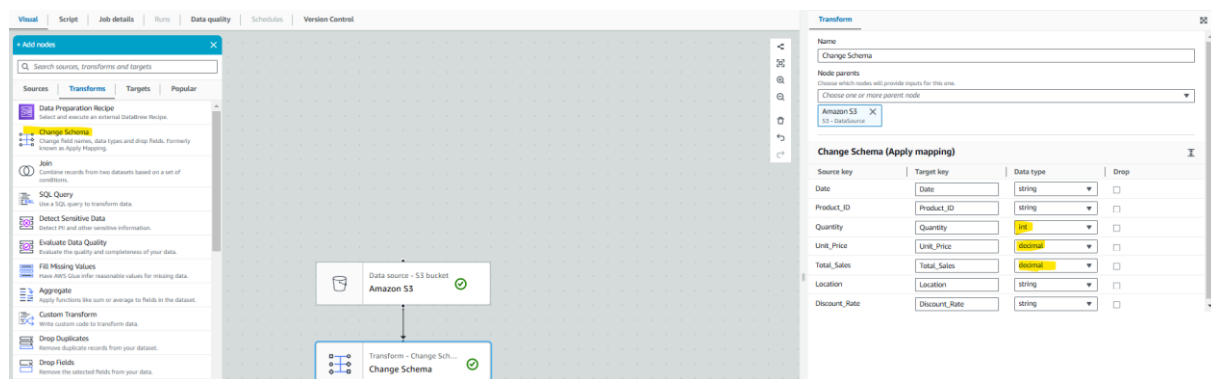
Types des données, Agréger et convertir au format Parquet



Nous allons maintenant créer notre job ETL par la visualisation



Ajouter un composant "Change schema"



Ajouter le composant aggregate pour agir sur les données en l'occurrence ici on fait un group by sur la colonne "location" qui va agréger la colonne "Total_Sales" par somme en fonction de la location

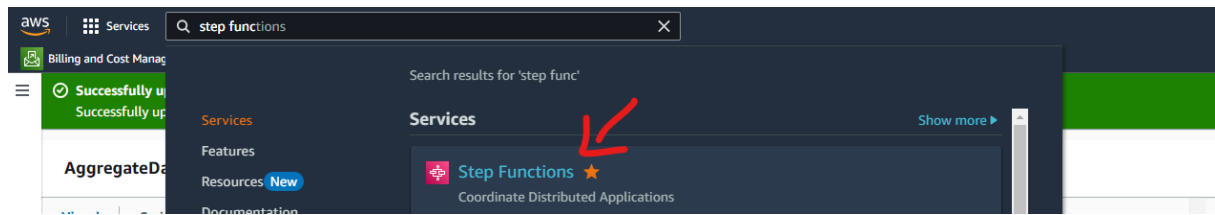
The screenshot displays the AWS Glue Studio interface. On the left, the 'Add nodes' panel is open, showing a search for 'aggr' which results in the 'Aggregate' transform being highlighted. The main workspace shows a workflow diagram with three nodes: 'Data source - S3 bucket Amazon S3', 'Transform - Change Schema', and 'Transform - Aggregate'. The 'Aggregate' node is selected, and its configuration panel on the right is visible. The 'Name' field is set to 'Aggregate'. Under 'Fields to group by - optional', the 'location' field is selected. Under 'Field to aggregate', 'Total_Sales' is selected, and the 'Aggregation function' is set to 'sum'.

Ajouter le composant bucket S3 target

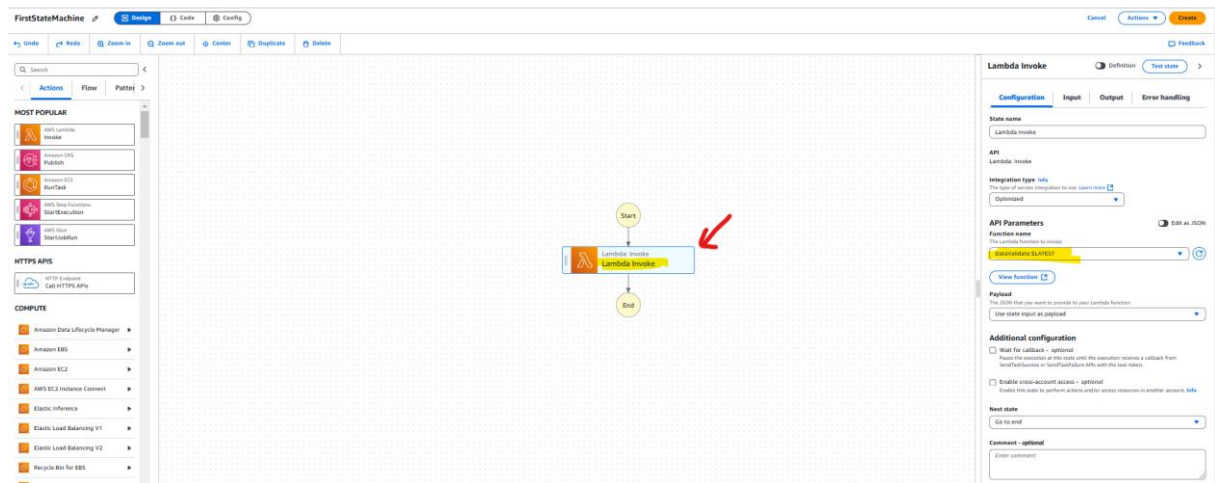
The screenshot shows the 'Data target properties - S3' configuration panel in AWS Glue Studio. The 'Name' field is set to 'Amazon S3'. The 'Format' is set to 'Parquet'. The 'S3 Target Location' is set to 's3://auto-data-catalog-7711302/'. The 'Data Catalog update options' section is expanded, showing the 'Do not update the Data Catalog' option selected. The 'Partition keys - optional' section is also visible, with an 'Add a partition key' button.

Notre Job ETL est maintenant terminé

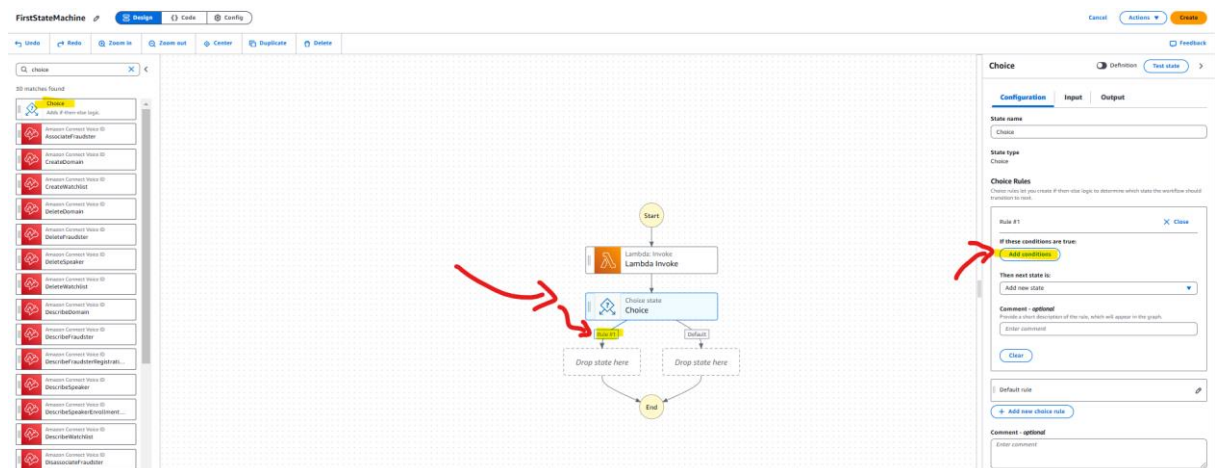
Etape 5: Créer notre State Machine



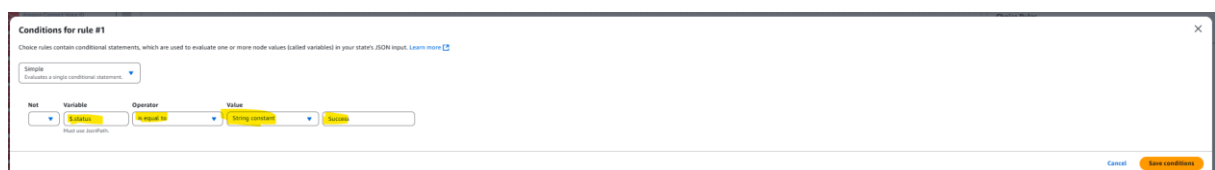
Ajouter le composant Invoke AWS Lambda et drag and drop la fonction précédemment créé



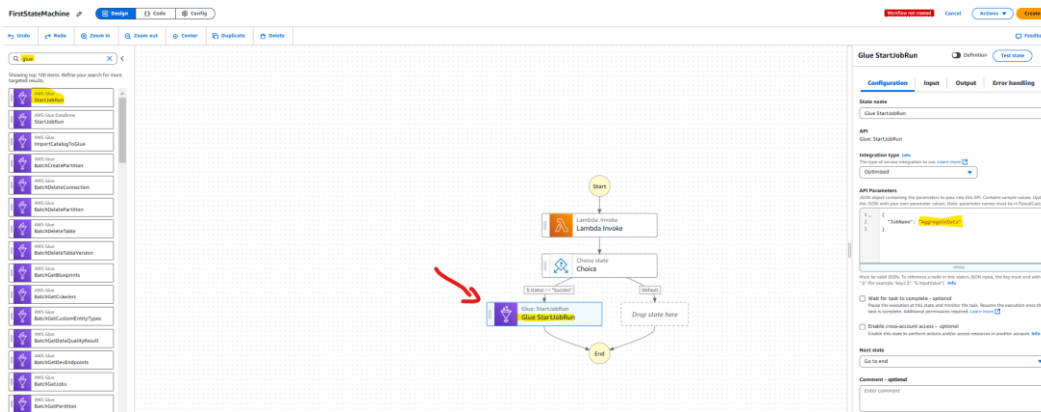
Ajouter le composant “Choice” qui va nous permettre de créer une condition, cliquer sur “add Rule” de l’élément puis “add condition”



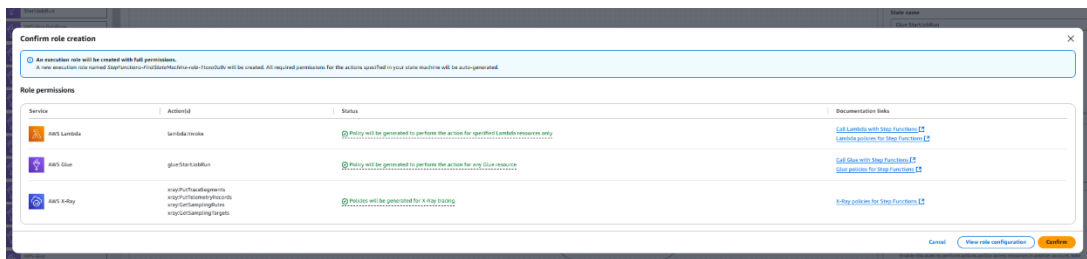
La condition va récupérer le résultat de notre Lambda fonction créé qui est soit {“status” : “Fail”} soit {“status” : “Success”}, ici on va récupérer en cas de Success



En cas de succès, nous ajoutons un composant “Glue StartJobRun”, dans lequel on précise le nom de notre job ETL créé par la visualisation juste avant “AggregateData”

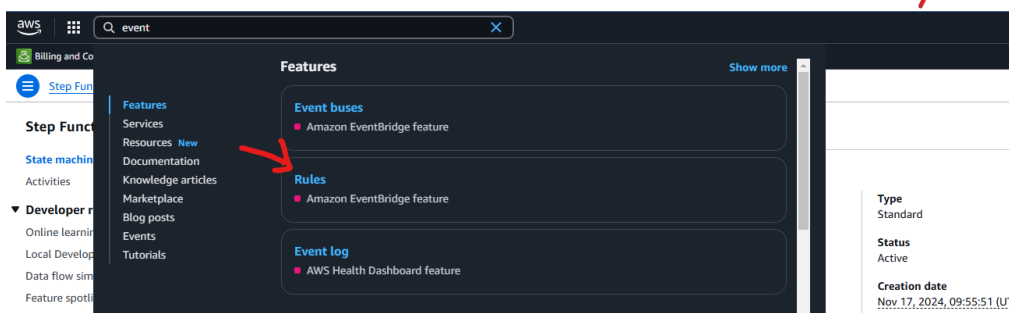
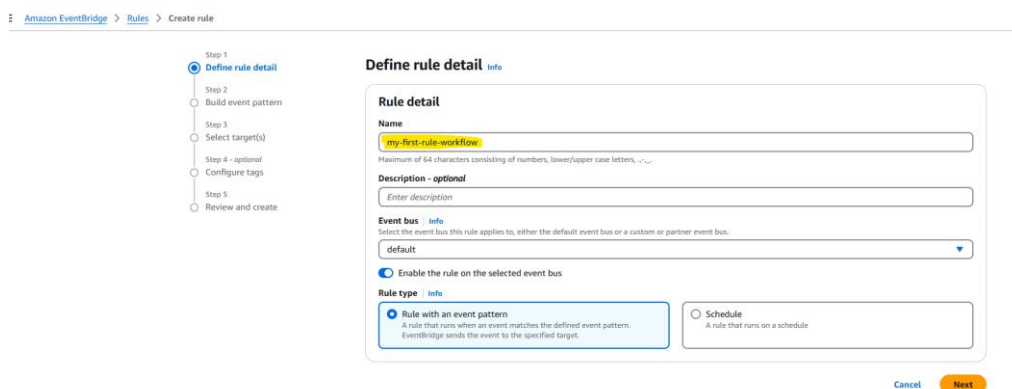


Appuyez sur le bouton “Create” en haut à droite qui va nous permettre d’avoir une vue d’ensemble de la State machine que nous créons, appuyez sur le bouton “confirm”



Etape 6: Déclencher le workflow créé

Nous allons ici créer une règle avec Amazon EventBridge



Step 1: Define rule detail
Step 2: Build event pattern
Step 3: Select target(s)
Step 4: optional: Configure tags
Step 5: Review and create

Build event pattern

Event source
Select the event source from which events are sent.
☒ AWS services or EventBridge partner events
 Events sent from AWS services or EventBridge partners.
☐ Other
 Custom events or events sent from more than one source, e.g. events from AWS services and partners.
☐ All events
 All events sent to your account.

Sample event - optional
You don't have to select or enter a sample event, but it's recommended so you can reference it when writing and testing the event pattern, or filter criteria.
 You can reference the sample event when you write the event pattern, or use the sample event to test if it matches the event pattern. Find a sample event, enter your own, or edit a sample event below.
 Learn more about the required fields in a sample event.

Sample event type
☒ AWS events
☐ EventBridge partner events
☐ Enter my own

Sample events
 Select an event source and type or by keyword.
 Select

Enter the event JSON

Copy

Nous allons maintenant choisir l'Event pattern, sur un évènement spécifique de création d'objet dans notre bucket "sales-data-raw-17112024"

Creation method
 Method
☐ Use schema
 Use an Amazon EventBridge schema to generate the event pattern.
☒ Use pattern form
 Use a template provided by EventBridge to create an event pattern.
☐ Custom pattern (JSON editor)
 Write an event pattern in JSON.

Event pattern
 Event source
 AWS service or EventBridge partner as source
 AWS services
 AWS service
 The name of the AWS service as the event source
 Amazon S3
 Event type
 The type of events as the source of the matching pattern
 Amazon S3 Event Notification
 S3 Event Notifications will only match your rules if you have configured your S3 bucket(s) to publish event notifications to EventBridge. Learn more.
 Event pattern
 Event pattern, or filter to match the events

```
{
  "source": ["aws.s3"],
  "detail-type": ["Object Created"],
  "detail": {
    "bucket": {
      "name": ["sales-data-raw-17112024"]
    }
  }
}
```

 Copy Text pattern Edit pattern
 Event Type Specification 1
☐ Any event
☒ Specific event(s)
 Specific event(s)
 Object Created
 Event Type Specification 2
☐ Any bucket
☒ Specific bucket(s) by name
 Specific bucket(s) by name
 sales-data-raw-17112024
 Remove
 Add
 Cancel Previous Next

Sélectionner la target qui va être utilisée pour notre règle, notre State machine "FirstStateMachine"

Step 1: Define rule detail
Step 2: Build event pattern
Step 3: Select target(s)
Step 4: optional: Configure tags
Step 5: Review and create

Select target(s)

Permissions
 Note: When using the EventBridge console, EventBridge will automatically configure the proper permissions for the selected targets. If you're using the AWS CLI, SDK, or CloudFormation, you'll need to configure the proper permissions.

Target 1
Target types
 Select an EventBridge event bus, EventBridge API destination (SaaS partner), or another AWS service as a target.
☐ EventBridge event bus
☐ EventBridge API destination
☒ AWS service
Select a target
 Select target(s) to invoke when an event matches your event pattern or when schedule is triggered (limit of 5 targets per rule)
 Step Functions state machine
State machine
 FirstStateMachine
Execution role
 EventBridge needs permission to send events to the event bus of the above AWS account. By continuing, you are allowing us to do so. EventBridge and AWS Identity and Access Management
☒ Create a new role for this specific resource
☐ Use existing role
Role name
 Amazon.EventBridge_Invoke_Step_Functions_1706517713
Additional settings
 Add another target
 Cancel Skip to Review and create Previous Next

- Define rule detail
- Step 2
- Build event pattern
- Step 3
- Select target(s)
- Step 4 - optional
- Configure tags
- Step 5
- Review and create**

Review and create

Step 1: Define rule detail

Define rule detail

Rule name
my-first-rule-workflow

Description

Status
Enabled

Rule type
Standard rule

Event bus
default

Step 2: Build event pattern

Event pattern

```

1 {
2   "source": ["aws.s3"],
3   "detail-type": ["Object Created"],
4   "detail": {
5     "bucket": {
6       "name": ["sales-data-rw-17118024"]
7     }
8   }
9 }

```

Step 3: Select target(s)

Details	Target Name	Type	Arn	Input	Role
▼	FirstStateMachine	Step Functions state machine	arn:aws:states:::west-3:54654490666:stateMachine:FirstStateMachine	Matched event	Amazon_EventBridge_Invoke_Step_Functions_1706517713

Input to target: Matched event

Additional parameters: --

Dead-letter queue (DLQ): -

Step 4: Configure tag(s)

Tags (0)

A tag is a label that you assign to an AWS resource. Each tag consists of a key and an optional value. You can use tags to search and filter your resources or track your AWS costs.

Key	Value
No tags associated with this resource.	

[Cancel](#)
[Previous](#)
[Create rule](#)

Nous devons maintenant configurer notre bucket, pour cela nous devons activer l'Amazon EventBridge

Amazon S3

- Buckets
- Access Grants
- Access Points
- Object Lambda Access Points
- Multi-Region Access Points
- Batch Operations
- VM Access Analyzer for S3
- Block Public Access settings for this account
- Storage Lens

Event notifications (0)

Select a notification when specific events occur in your bucket.

Name	Event types	Filters	Destination type	Destination
No event notifications				

Choose Create event notification to be notified when a specific event occurs.

Create event notification

Amazon EventBridge

For additional capabilities, use Amazon EventBridge to build event-driven applications or rules using S3 event notifications.

Send notifications to Amazon EventBridge for all events in this bucket

Off

Supprimons le fichier précédemment uploadé dans notre bucket Source et remettons le pour tester si tout ce qu'on a créé auparavant fonctionne, lors du téléchargement du fichier, la State machine se mets en marche et réussi

State machine: FirstStateMachine

FirstStateMachine

Details

Arn
arn:aws:states:::west-3:54654490666:stateMachine:FirstStateMachine

IAM role ARN
arn:aws:iam::54654490666:role/service-role/StepFunctions-FirstStateMachine-role-1t0u0u0v

Type
Standard

Status
Active

Creation date
Nov 17, 2024, 09:55:51 (UTC+01:00)

X-Ray tracing
Disabled

[Edit](#)
[Actions](#)
[Start execution](#)

Executions (0/3)

Filter executions by property or value

Filter by status

Last 15 months

Local timezone

3 matches

Name	Status	Start Time (local)	End Time (local)	Duration	Version	Alias
6a726457-5874-68b1-655f-4a00d7218d...	Succeeded	Nov 17, 2024, 10:40:54	Nov 17, 2024, 10:40:55	00:00:01.723	-	-

Partie 3 : Analyse et Exploitation des Données

Description :

Implémentation d'une API pour exposer les données du data lake.

- L'API devra exécuter des repêtes pour extraire des données
- Création de rapports et de tableaux de bord sur les données collectées.
- Bonus: Utilisation d'outils BI (comme Tableau ou Power BI) pour visualiser les résultats.

Livrables :

- Code API
- Documentation des endpoints implémentés et les résultats qu'ils renvoient

API

Le code API se trouve dans le dossier sous le nom « **api.py** » . Cette API permet de faire n'importe quelle requête SQL afin d'échanger avec notre base de données dans Amazon Athena.

Un **GET** est fait grâce au endpoint **query**, où on y passe une requête SQL, on retrouve un Status Code « 200 OK », qui montre que la requête s'est bien déroulée.

```
PS C:\Users\valen\OneDrive - Efrei\EFREI\EFREI M2\data lakes\api_aws> uvicorn main:app --reload
INFO: Will watch for changes in these directories: ['C:\\Users\\valen\\OneDrive - Efrei\\EFREI\\EFREI M2\\data lakes\\api_aws']
INFO: Uvicorn running on http://127.0.0.1:8000 (Press CTRL+C to quit)
INFO: Started reloader process [19972] using StatReload
INFO: Started server process [18536]
INFO: Waiting for application startup.
INFO: Application startup complete.
INFO: 127.0.0.1:51062 - "GET /query?query=SELECT+*+FROM+productproducts+LIMIT+10 HTTP/1.1" 200 OK
```

Nous pouvons voir ici le résultat de notre requête précédemment envoyée au format JSON.

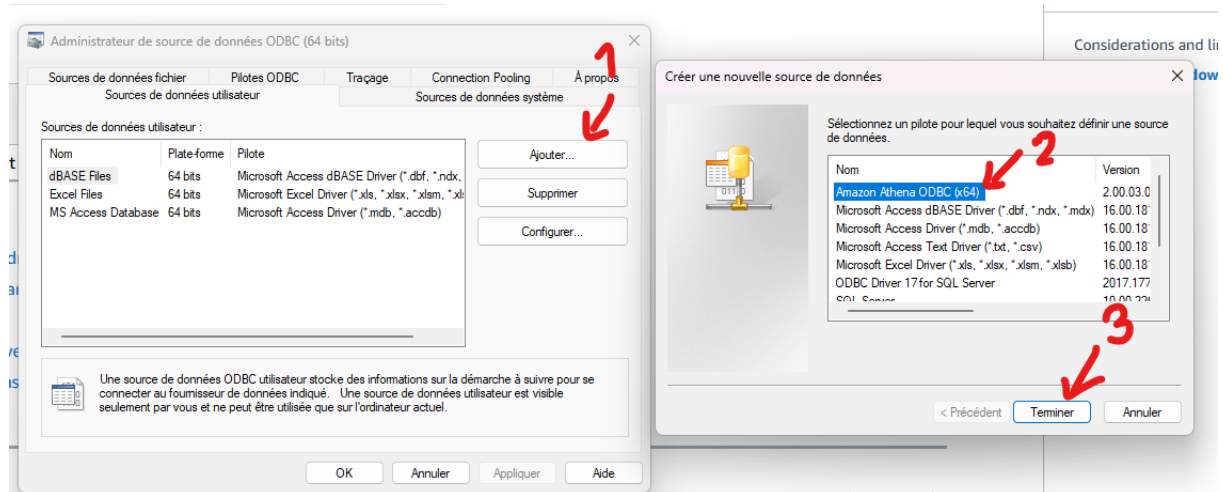
```
PS C:\Users\valen\OneDrive - Efrei\EFREI\EFREI M2\data lakes\api_aws> curl "http://127.0.0.1:8000/query?query=SELECT+*+FROM+productproducts+LIMIT+10"
StatusCode      : 200
StatusDescription : OK
Content         : {"query": "SELECT * FROM productproducts LIMIT 10", "results": [{"product_id": "100", "product_name": "Product 8", "category": "Home & Kitchen", "supplier": "Supplier C", "price_range": "21.79818545851389"}, {"pro...
RawContent      : HTTP/1.1 200 OK
                  Content-Length: 1351
                  Content-Type: application/json
                  Date: Wed, 20 Nov 2024 17:10:56 GMT
                  Server: uvicorn

{"query": "SELECT * FROM productproducts LIMIT 10", "results": [{"product_id...
Forms          : {}
Headers        : [{"Content-Length", 1351}, {"Content-Type", application/json}, {"Date", Wed, 20 Nov 2024 17:10:56 GMT}, {"Server", uvicorn}]
Images         : {}
InputFields    : {}
Links          : {}
ParsedHtml     : mshtml.HTMLDocumentClass
RawContentLength : 1351
```

POWER BI

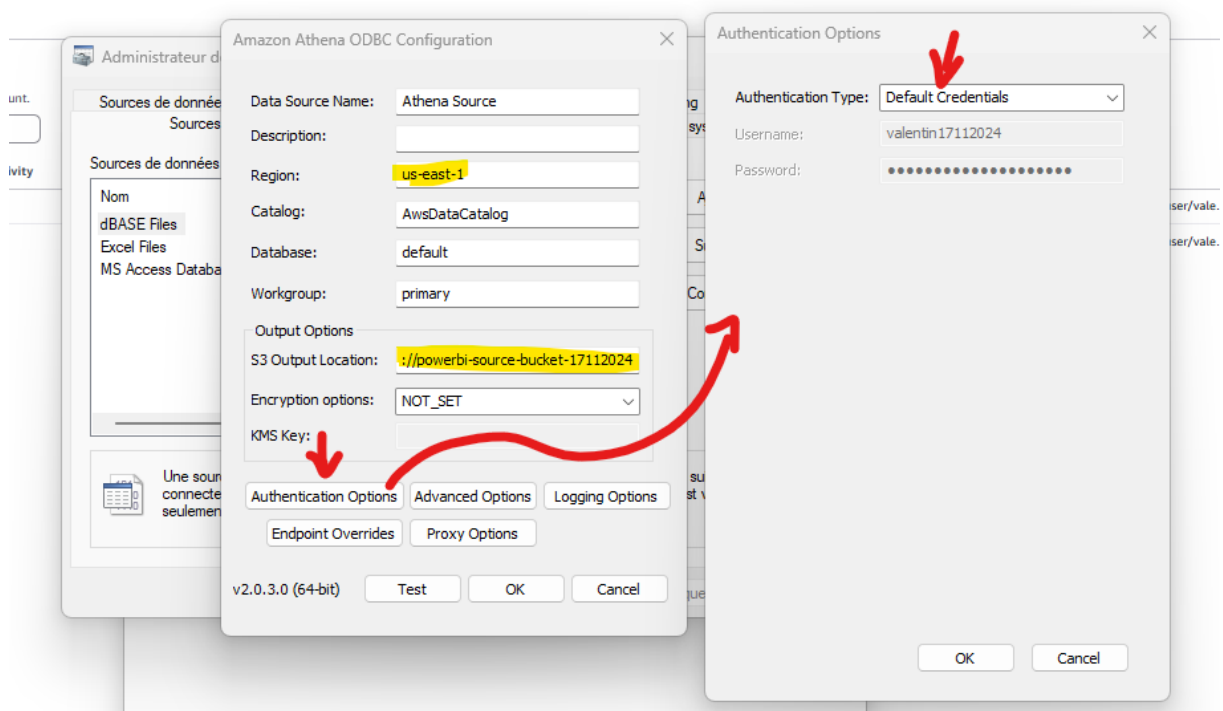
Pour cela nous allons devoir télécharger power BI, téléchargez ODBC Driver. Nous allons aussi devoir faire la configuration utilisateur avec une clé qui va être stockée dans ODBC source. Pour finir nous allons connecter Power BI.

Ouvrons maintenant l'ODBC en 64 bits, nous arrivons sur cette maintenant où nous devons ajouter ODBC Athena



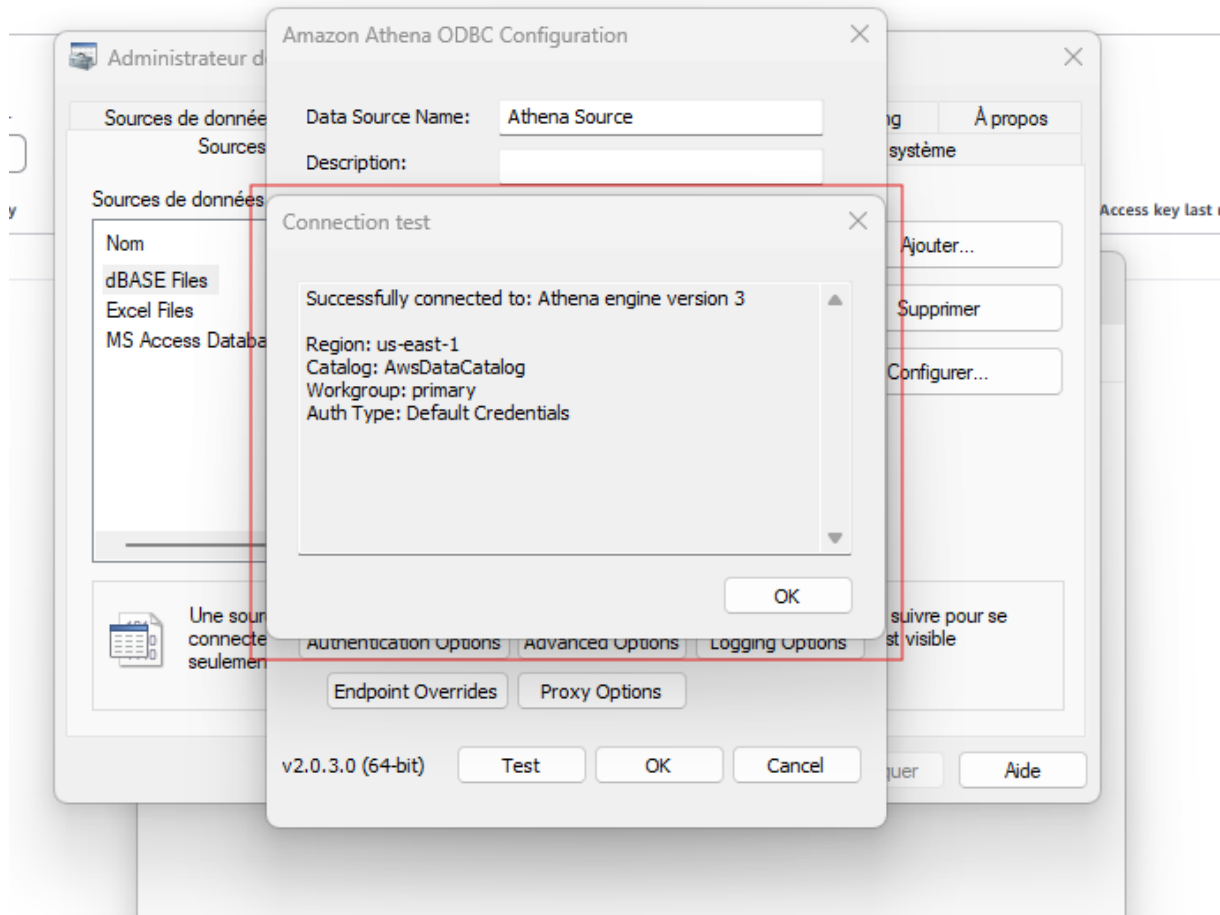
Créer un bucket source pour recevoir les données d'ODBC, appelé "powerbi-source-bucket-17112024", ne pas oublier de mettre "s3://" avant le nom du bucket qui sera utilisé "s3://powerbi-source-bucket-17112024". Terminez la configuration d'odbc source Athena.

units and cloud applications in Identity Center.

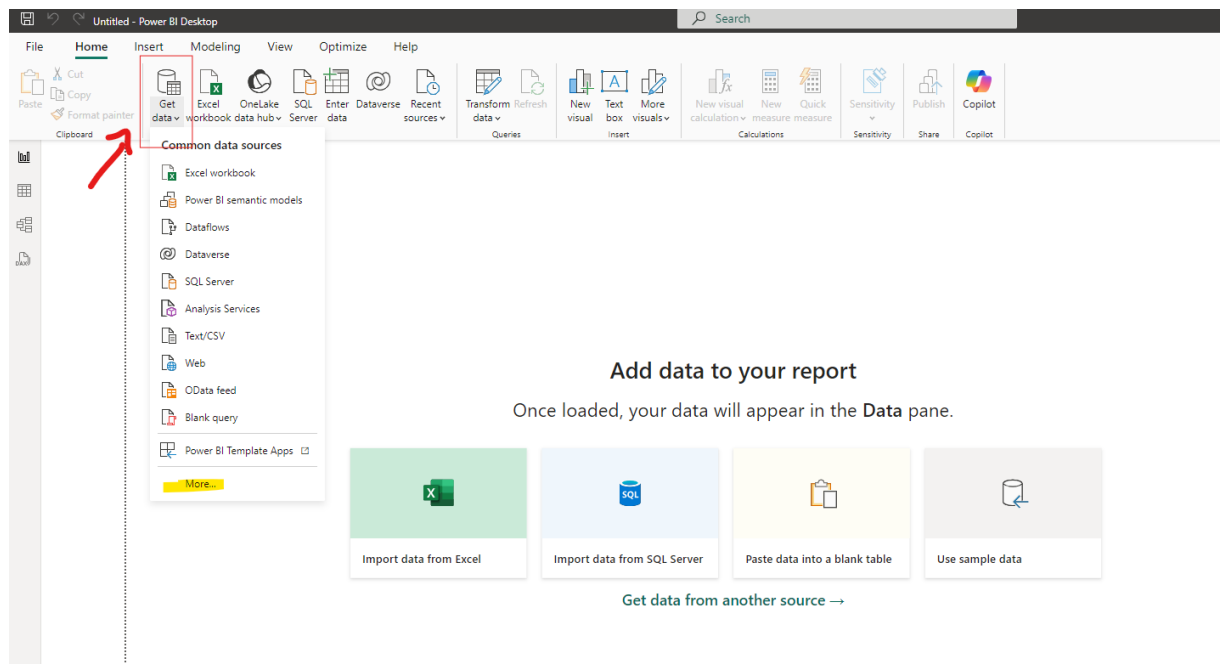


La connection a bien été autorisée

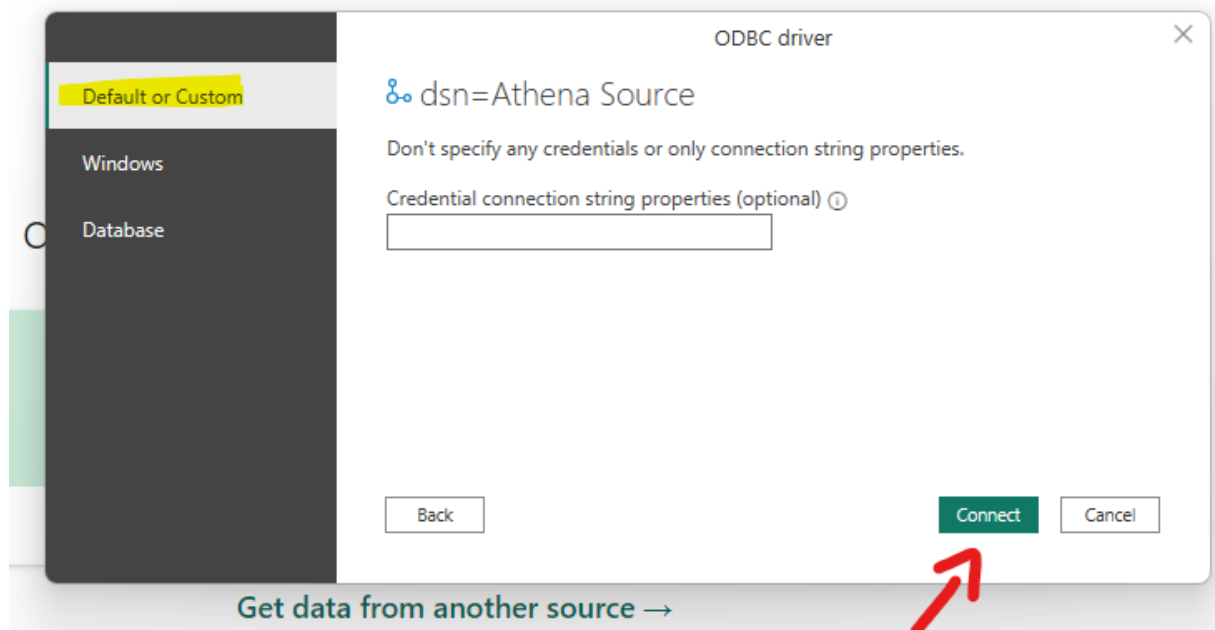
s and cloud applications in Identity Center.



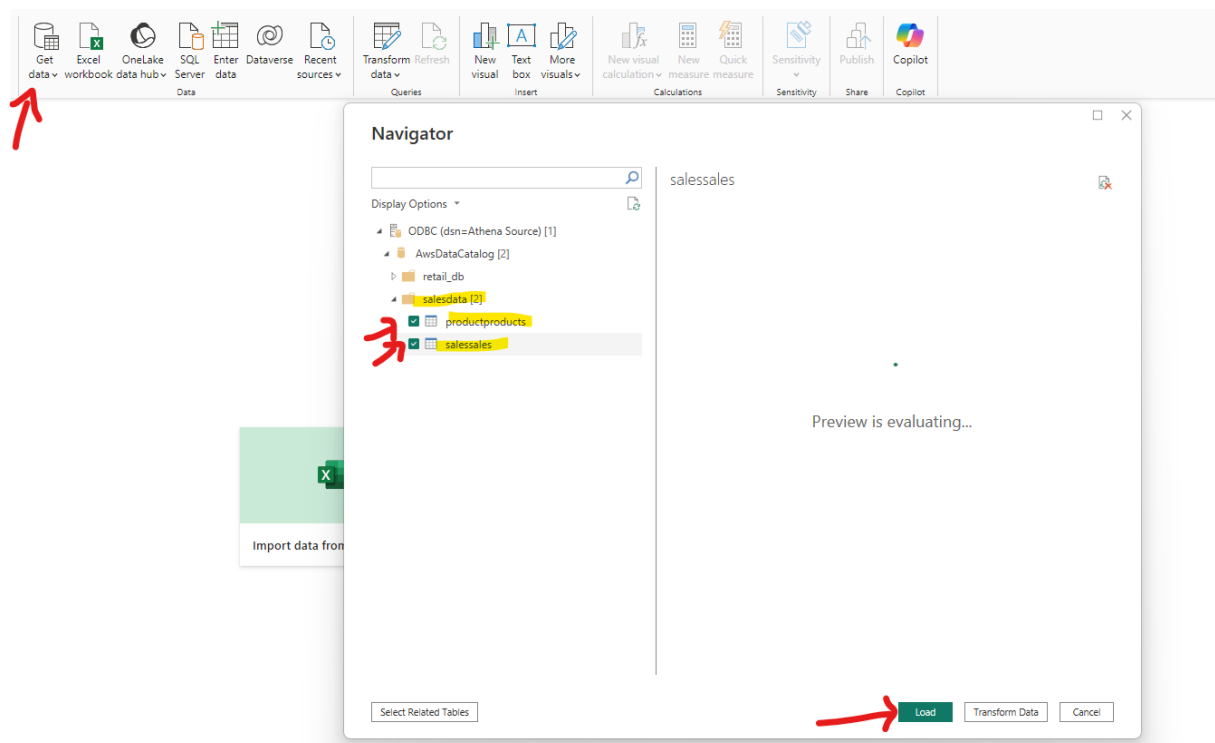
Il ne nous reste plus qu'à connecter power bi



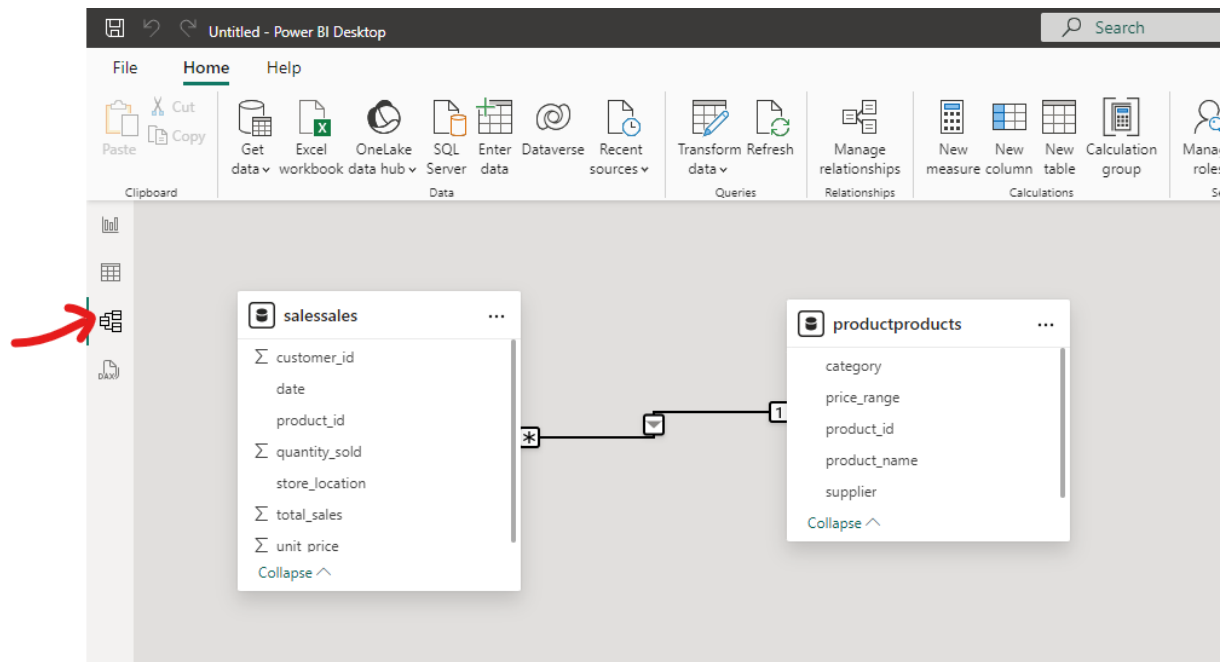
Trouver l'odbc et établir la connection, choisir la "Athena Source", choisir la "Default or Custom" et appuyez sur le bouton "Connect"



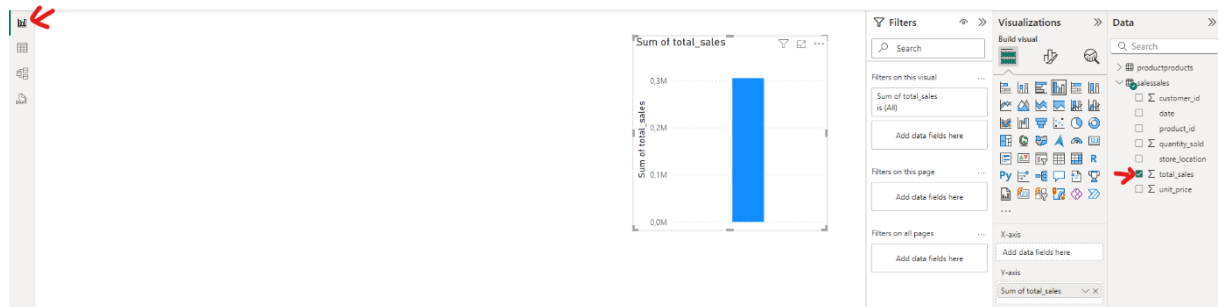
Nous avons maintenant accès à nos différentes tables présentes sur Amazon Athena, les sélectionner et les charger "load"



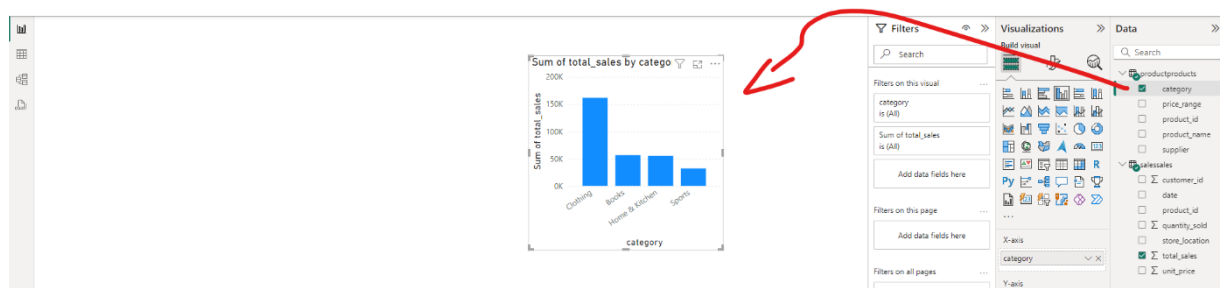
Nous avons maintenant un aperçu de nos tables dans power BI !



Dans l'onglet "Report View" sur la gauche nous pouvons maintenant sélectionner notre table "salessales" et plus précisément notre colonne "total_sales" pour le drag and drop dans le tableau vierge.



Nous pouvons aussi drag and drop dans notre même schéma la colonne "category" de la table "productproducts"



Partie 4 : Sécurité, Gouvernance et Qualité des Données

Description :

- Mise en place de la gouvernance des données : catalogage, suivi des métadonnées, versioning des datasets.
- Sécurisation des accès aux données via des solutions comme AWS IAM, Azure Active Directory, ou Kerberos pour Hadoop.

Livrables :

- Document décrivant la Politique de sécurité des données, les politiques d'accès

Gestion de la Qualité des Données

Description :

- Mise en place de solutions pour assurer la qualité des données ingérées et transformées : détection d'anomalies, gestion des valeurs manquantes.
- Surveillance des pipelines de données pour garantir leur abilité.

Livrables :

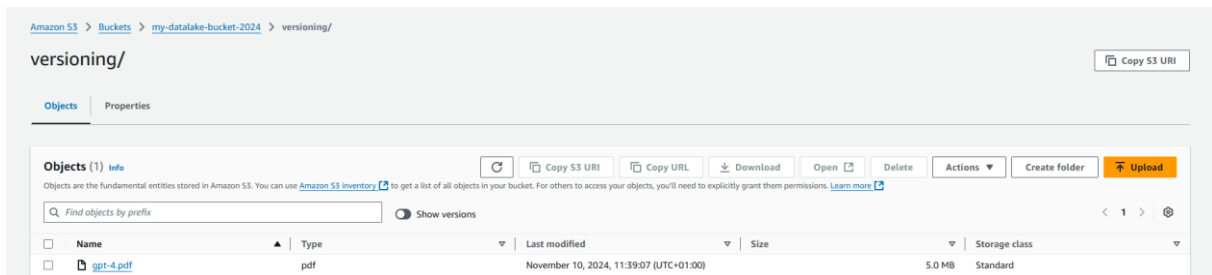
- Scripts ou outils pour la vérification de la qualité des données.
- Rapport de qualité des données avec des indicateurs clés (anomalies détectées, données manquantes, etc.).

La **data governance** dans un Data Lake désigne **l'ensemble des règles, processus et politiques mis en place pour garantir une gestion et une utilisation appropriées des données**. Cela inclut la gestion de la qualité des données, la sécurité, la conformité aux réglementations (comme le RGPD), l'accès contrôlé aux données, et la traçabilité des changements.

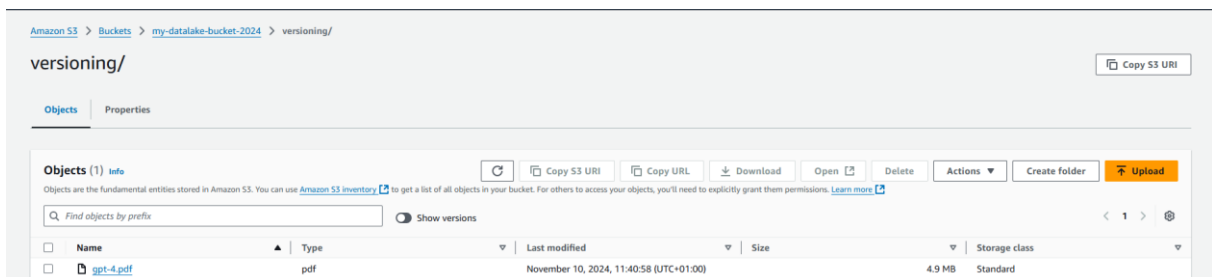
Versioning

The screenshot shows the Amazon S3 console interface. On the left is a navigation menu with options like Buckets, Access Grants, and Storage Lens. The main panel displays the 'Edit Bucket Versioning' page for the bucket 'my-datalake-bucket-2024'. Under the 'Bucket Versioning' heading, there are two radio buttons: 'Suspend' and 'Enable'. The 'Enable' option is selected. Below this, a blue information box states: 'After enabling Bucket Versioning, you might need to update your lifecycle rules to manage previous versions of objects.' At the bottom right of the main panel, there are two buttons: 'Cancel' and 'Save changes'. A red arrow points to the 'Save changes' button.

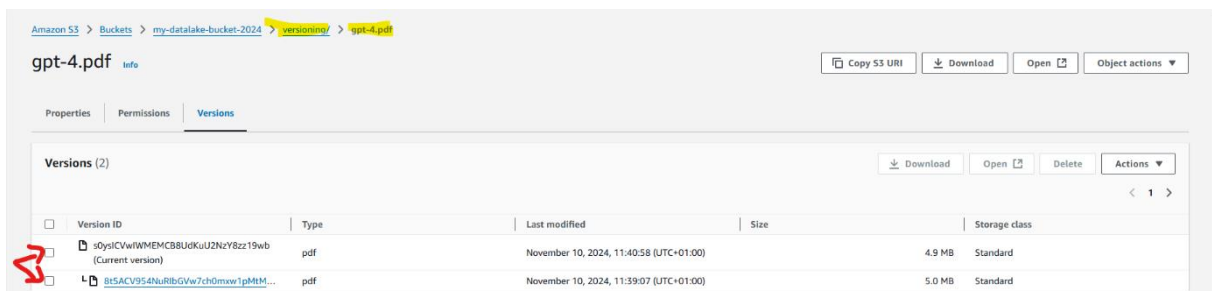
Nous allons maintenant comprendre ce qu'il se passe dans le versioning, d'abord en créant un dossier dans le bucket S3 et y ajouter un fichier



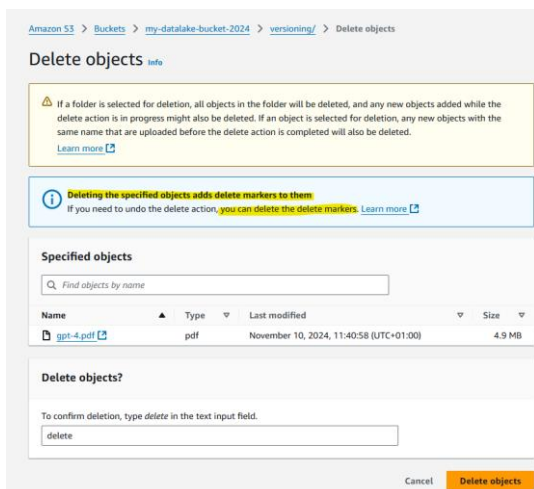
Nous allons maintenant remplacer le fichier actuel par une version plus récente modifiée, nous ne voyons pas de différence, le fichier a bien été remplacé



Comment voir les différentes versions de notre fichier ? Aller sur le fichier remplacé et accéder à l'onglet versioning, ici nous pouvons cliquer sur la version précédente qui s'affiche en bleu.

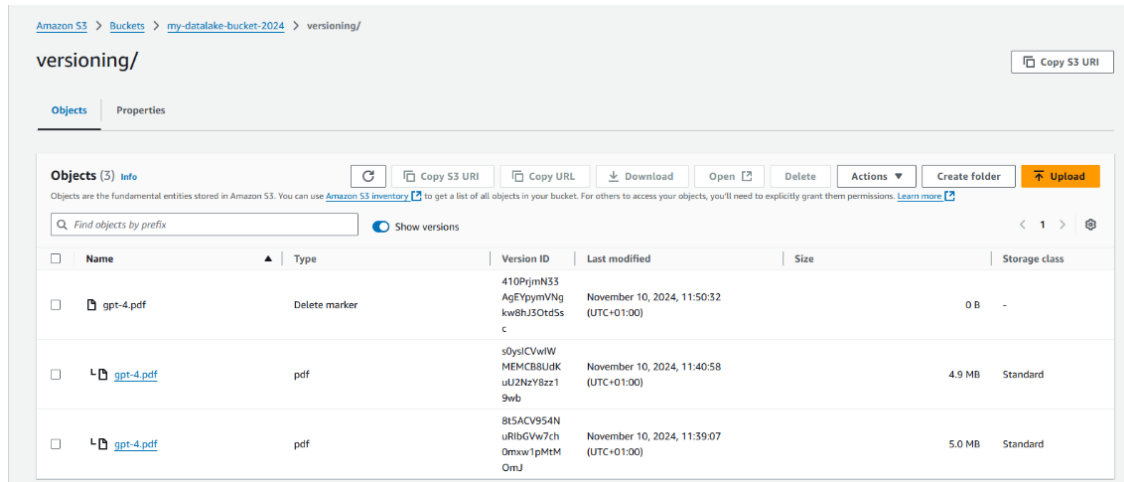


Maintenant que se passe-t-il si l'on supprime le fichier ? Cela va ajouter un "marqueur de suppression" mais ne va pas le supprimer de manière permanente.



Il est bien précisé sur l'image que si l'on veut annuler l'action de suppression il faut supprimer le marqueur de suppression.

Si l'on regarde maintenant dans le dossier les versions, on peut encore voir les différentes versions du fichier supprimé qui sont toujours disponibles.



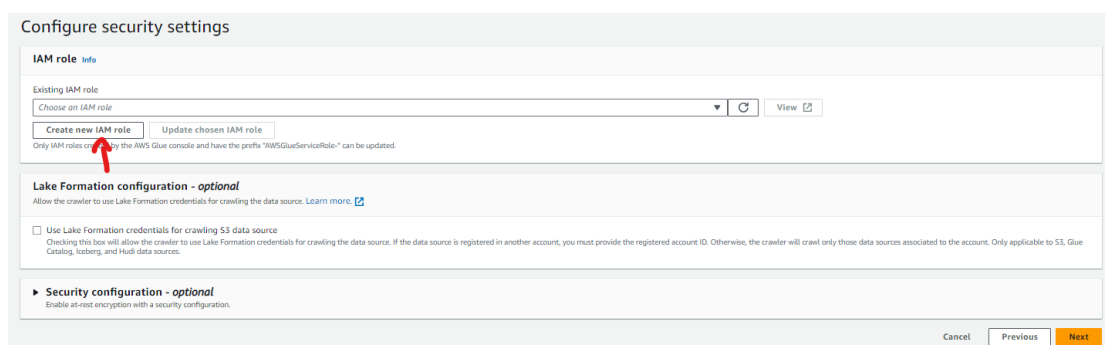
IAM

Amazon Identity and Access Management (IAM) est un service d'AWS qui permet de gérer de manière sécurisée l'accès aux services et ressources AWS.

Il offre la possibilité de créer et de gérer des utilisateurs et des groupes, et de définir des autorisations précises pour contrôler l'accès aux ressources AWS.

IAM assure ainsi une gestion fine des droits, garantissant que chaque utilisateur dispose uniquement des permissions nécessaires à l'exécution de ses tâches.

Dans la pratique, en créant un rôle IAM, ce rôle sera limité au chemin de la données indiquée



Configure security settings

IAM role [info](#)

Existing IAM role: AWSGlueServiceRole-TestProject1 [View](#)

[Create new IAM role](#) [Update chosen IAM role](#)

Only IAM roles created by the AWS Glue console and have the prefix "AWSGlueServiceRole-" can be updated.

Lake Formation configuration - optional

Allow the crawler to use Lake Formation credentials for crawling the data source. [Learn more](#)

☐ Use Lake Formation credentials for crawling S3 data source

Checking this box will allow the crawler to use Lake Formation credentials for crawling the data source. If the data source is registered in another account, you must provide the registered account ID. Otherwise, the crawler will crawl only those data sources associated to the account. Only applicable to S3, Glue Catalog, Iceberg, and Hudi data sources.

Security configuration - optional

Enable at-rest encryption with a security configuration.

[Cancel](#) [Previous](#) [Next](#)

Set output and scheduling

Output configuration [info](#)

Target database: retail_db [Clear selection](#) [Add database](#)

Table name prefix - optional: sales_

Maximum table threshold - optional:

This field sets the maximum number of tables the crawler is allowed to generate. In the event that this number is surpassed, the crawl will fail with an error. If not set, the crawler will automatically generate the number of tables depending on the data schema.

Crawler schedule

You can define a time-based schedule for your crawlers and jobs in AWS Glue. The definition of these schedules uses the Unix-like cron syntax. [Learn more](#)

Frequency: On demand

[Cancel](#) [Previous](#) [Next](#)

Set output and scheduling

Output configuration [info](#)

Target database: retail_db [Clear selection](#) [Add database](#)

Table name prefix - optional: sales_

Maximum table threshold - optional:

This field sets the maximum number of tables the crawler is allowed to generate. In the event that this number is surpassed, the crawl will fail with an error. If not set, the crawler will automatically generate the number of tables depending on the data schema.

Crawler schedule

You can define a time-based schedule for your crawlers and jobs in AWS Glue. The definition of these schedules uses the Unix-like cron syntax. [Learn more](#)

Frequency: On demand

[Cancel](#) [Previous](#) [Next](#)

Vérification de la qualité des données (State machine)

Concernant le Scripts ou outils pour la vérification de la qualité des données, nous avons utilisé AWS Glue pour détecter le schéma des données et grâce à notre State Machine précédemment créé nous avons renvoyé un message Status success ou fail si jamais le format des données n'était le bon, cela permis d'avoir le bon format de données parquet et la bonne extension (csv) dans le dossier de notre bucket S3.