PUILLANDRE Valentin SAULNIER Guillaume

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 plateforme 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

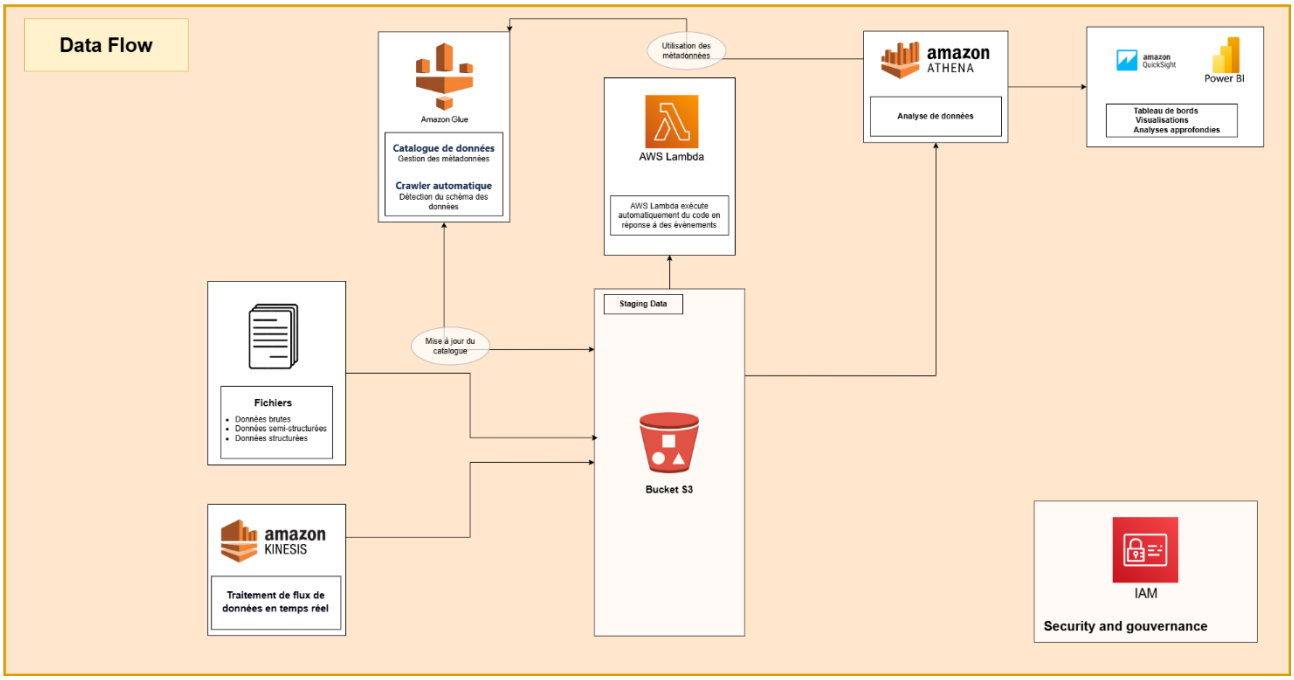
ETL (**E**xtract, **T**ransform, **L**oad) approche traditionnelle où la donnée est **transformée avant d'être chargée** dans un Data Warehouses par exemple.

ELT (**E**xtract, **L**oad, **T**ransform) approche plutôt utilisée pour les Data Lake où les données sont d'abord **chargées dans leurs formats d'origines et ensuite transformés**.

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)



PUILLANDRE Valentin SAULNIER Guillaume

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 donnée
- 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 continue)

- 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.

PUILLANDRE Valentin SAULNIER Guillaume

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

PUILLANDRE Valentin SAULNIER Guillaume

Partie 2 : Ingestion et Transformation des Données

Ingestion de Données Brutes

Description :

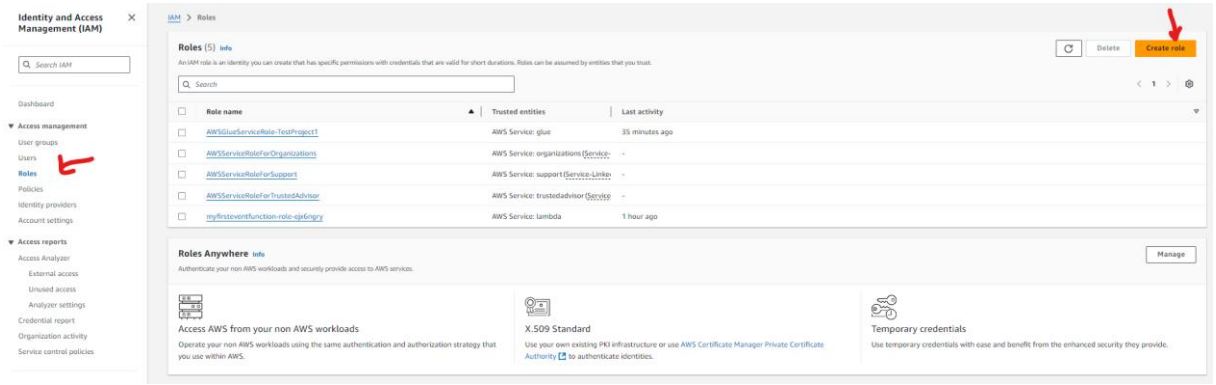
- Mise en place de l'ingestion des données provenant de différentes sources : bases de données SQL, logs, ux de données en temps réel (via Apache Kafka, AWS Kinesis).
- Mise en place d'un pipeline d'ingestion pour des données structurées, semistructurées et non structurées.
- Pour chaque type d données,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

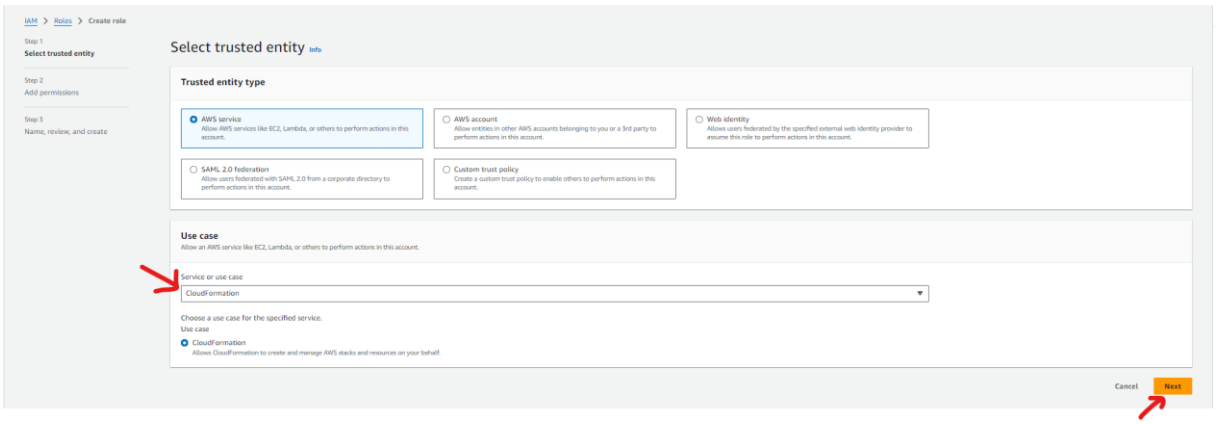
Livrables :

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

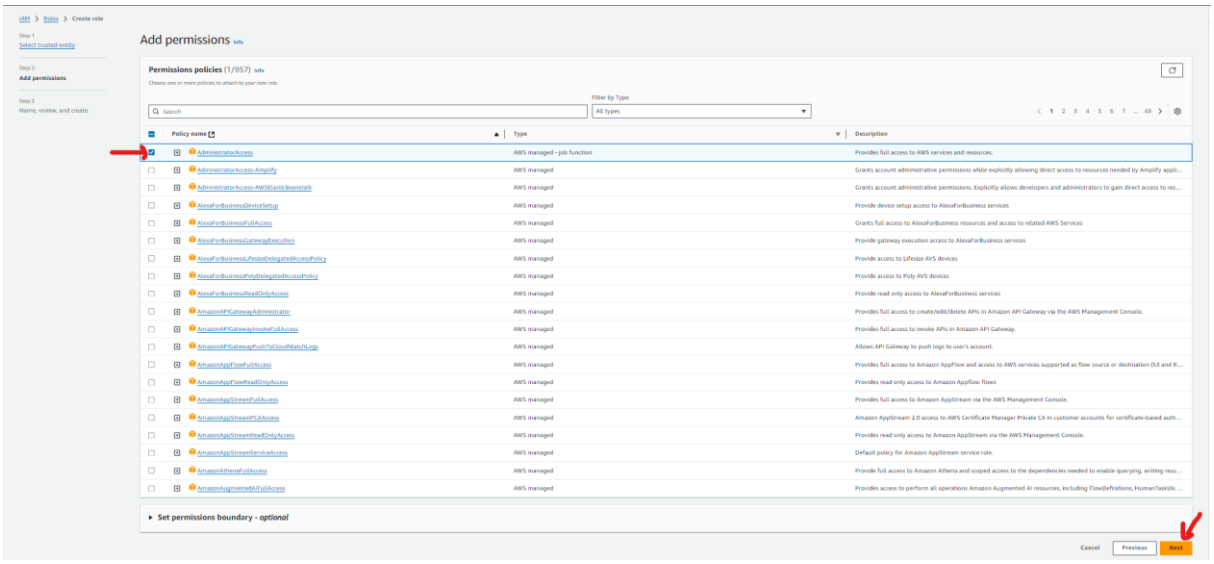
Utilisation de AWS Kinesis Streaming

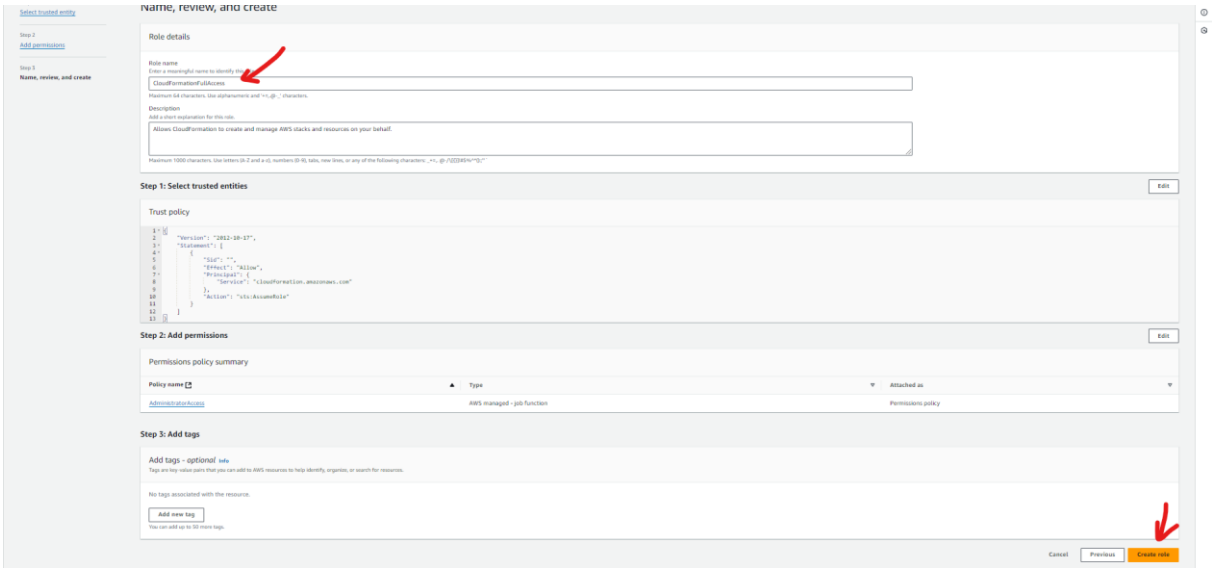
Nous allons générer un stream, mais d'abord nous avons besoin de créer un IAM role





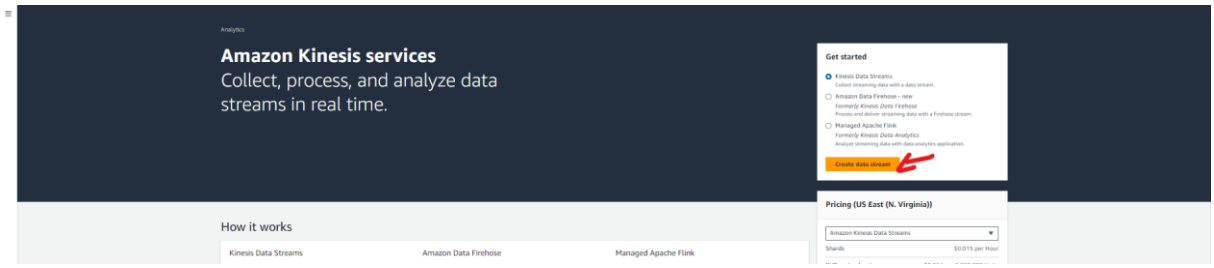
PUILLANDRE Valentin SAULNIER Guillaume

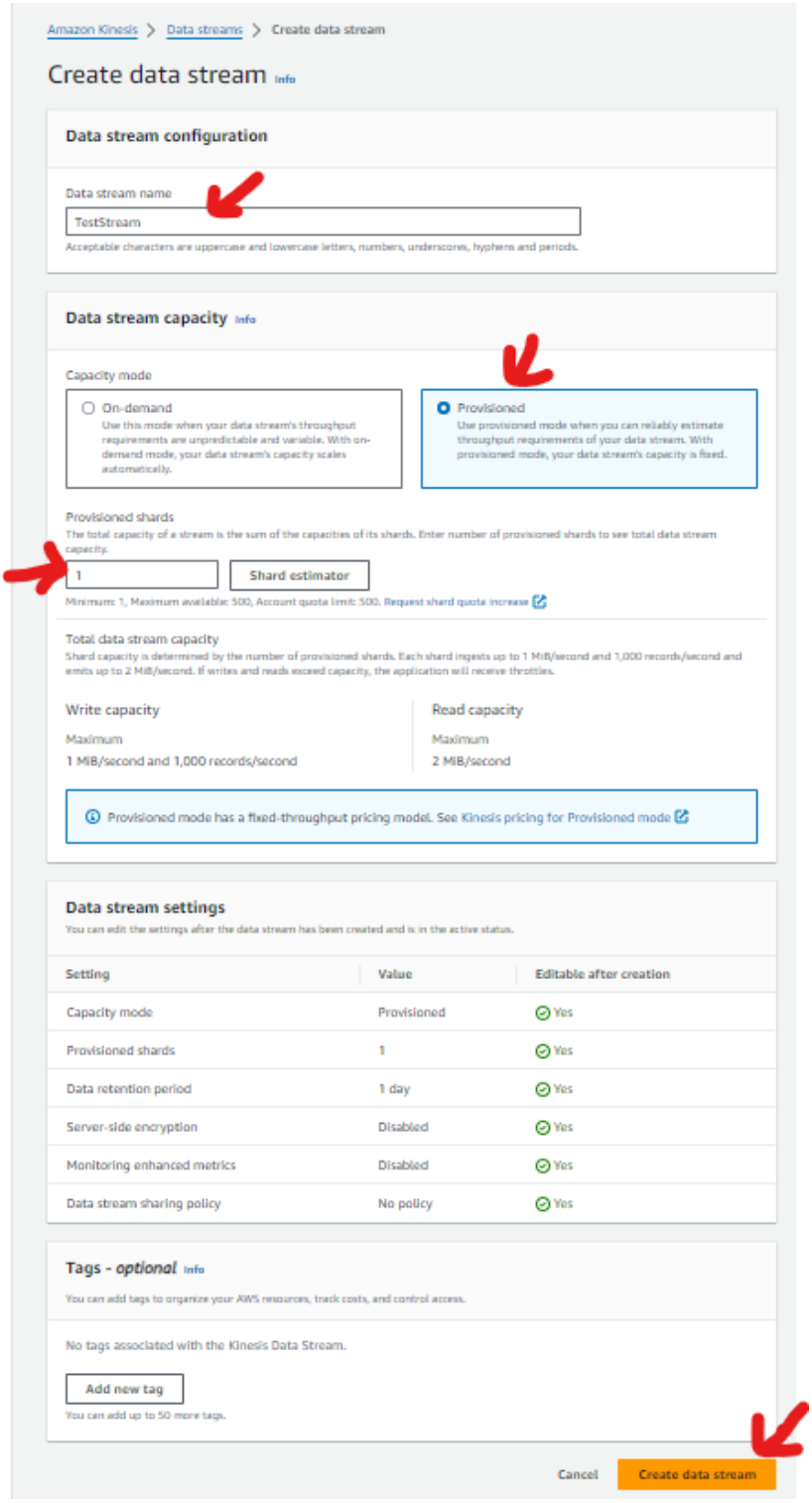




PUILLANDRE Valentin SAULNIER Guillaume

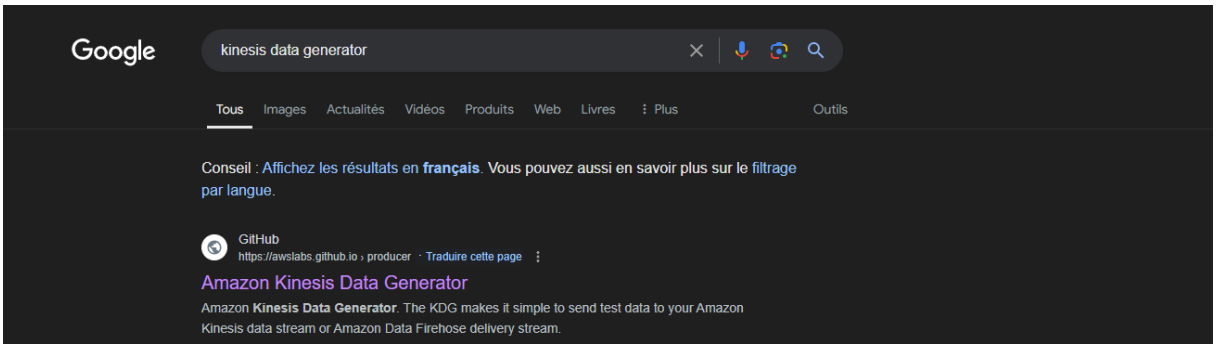
Créer un data stream avec Kinesis

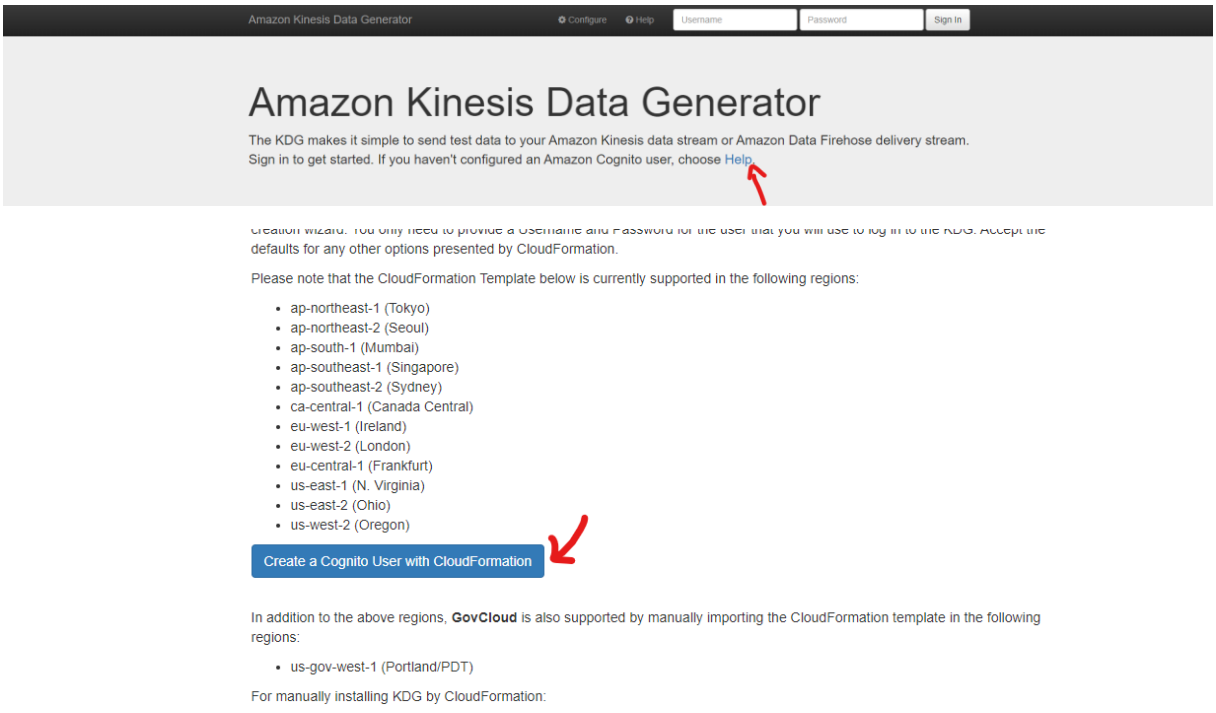


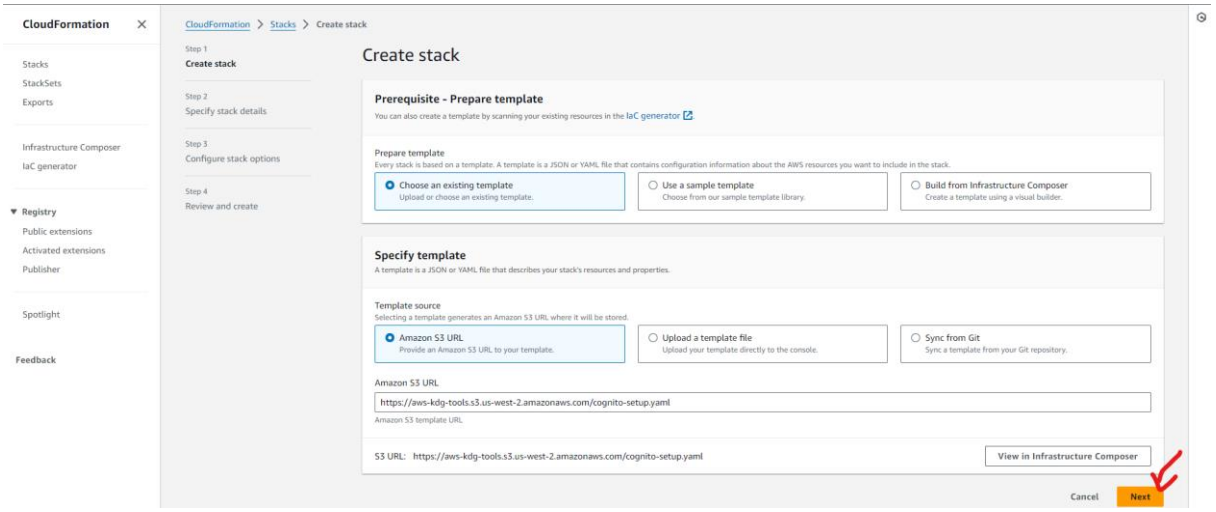


PUILLANDRE Valentin SAULNIER Guillaume

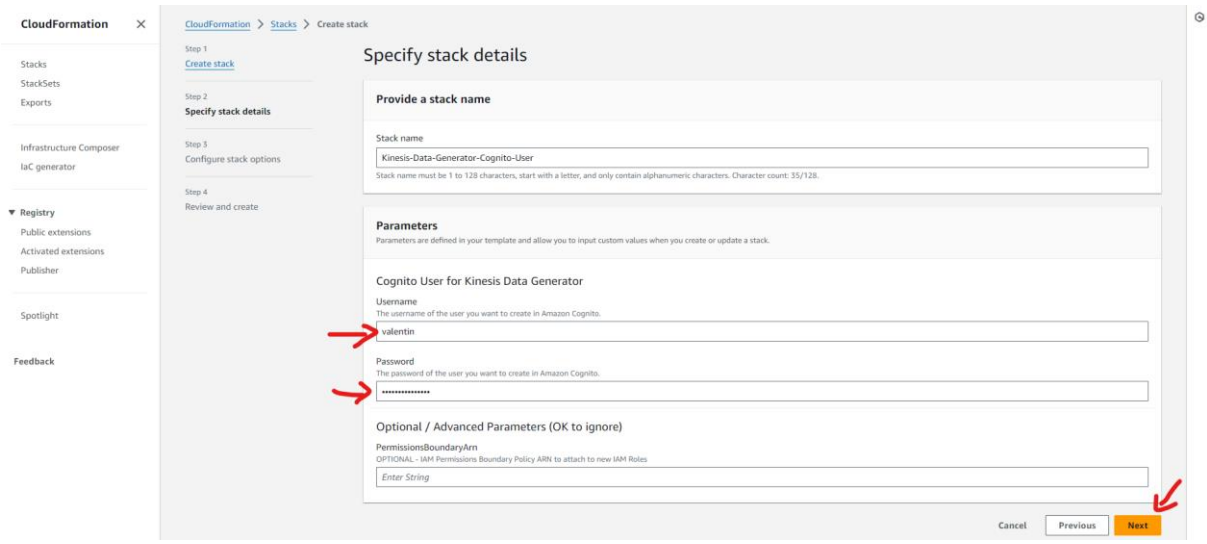
Utiliser un kinesis Data generator

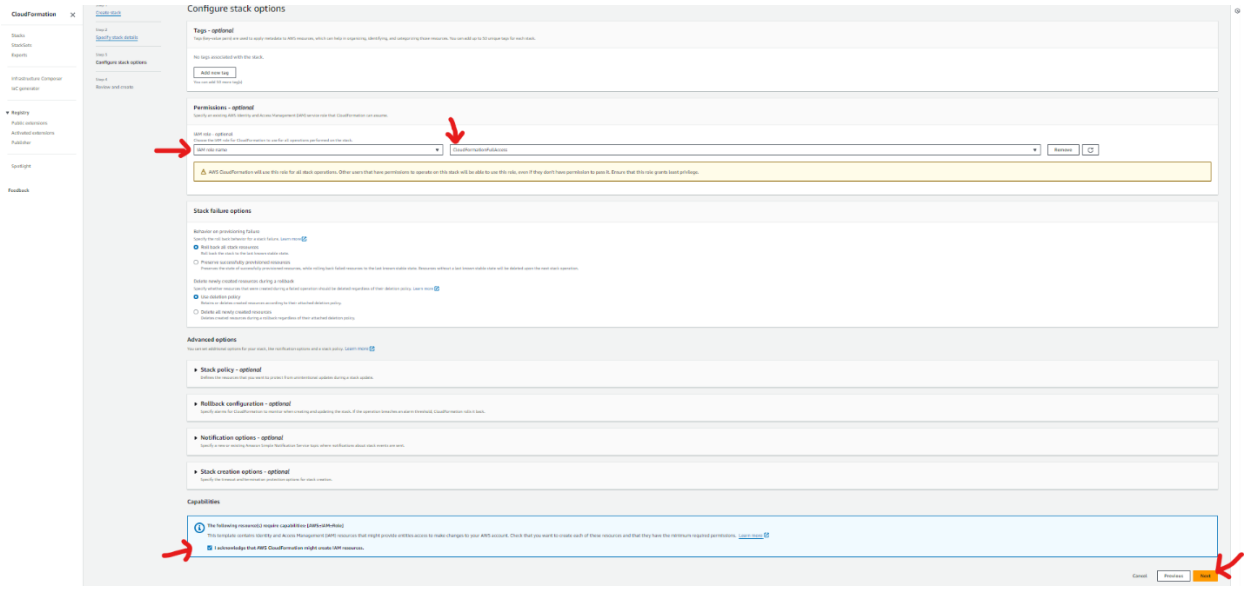


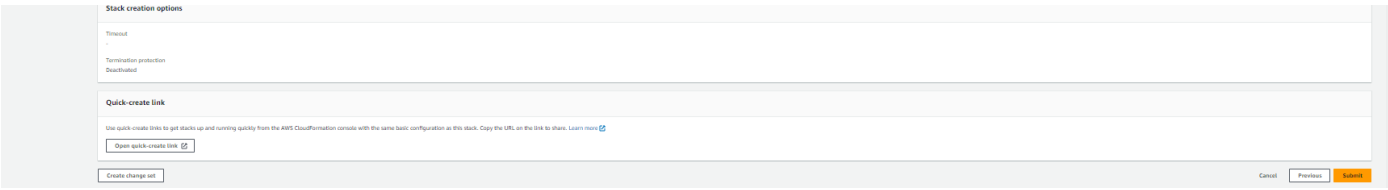




PUILLANDRE Valentin SAULNIER Guillaume

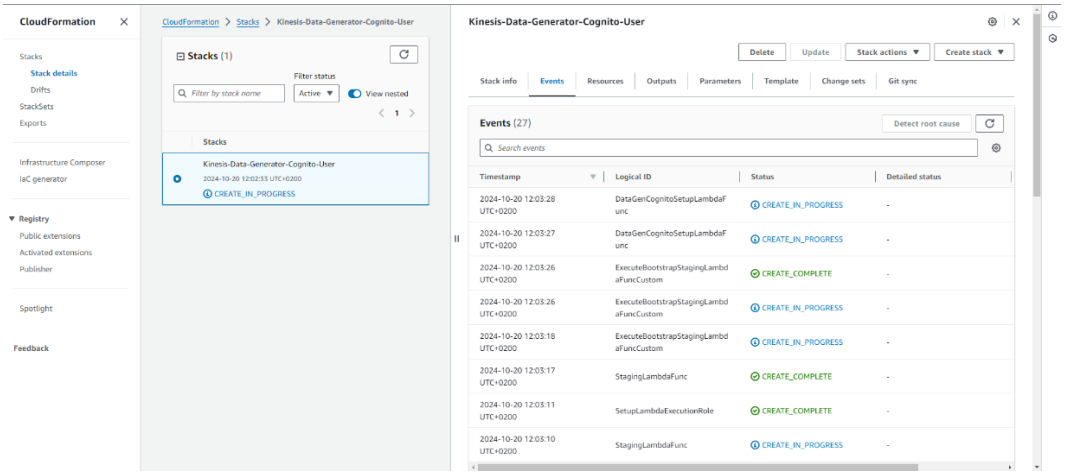




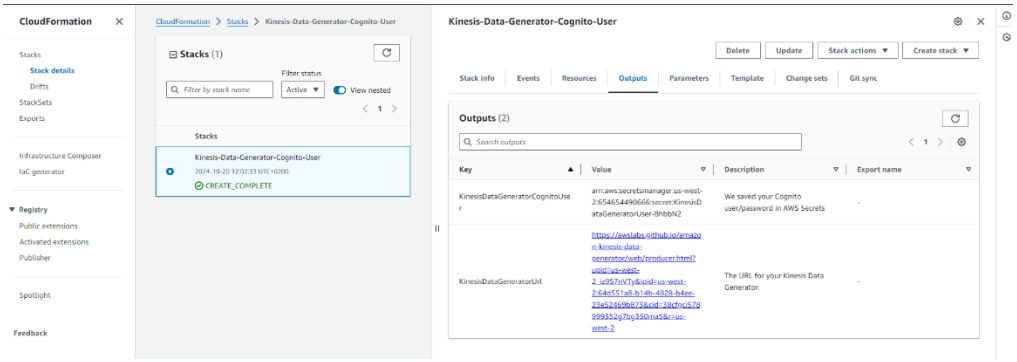


PUILLANDRE Valentin SAULNIER Guillaume

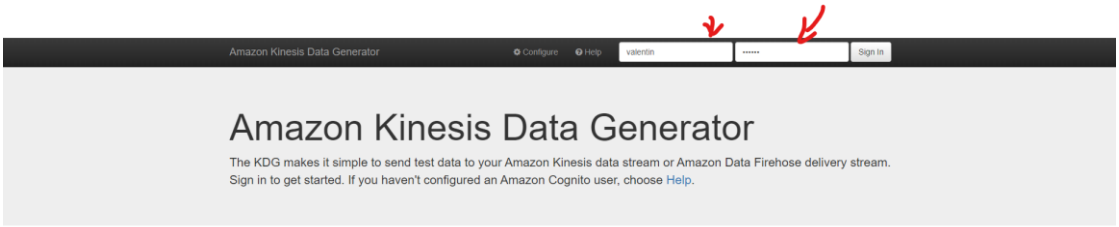
Tout a maintenant été créé



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

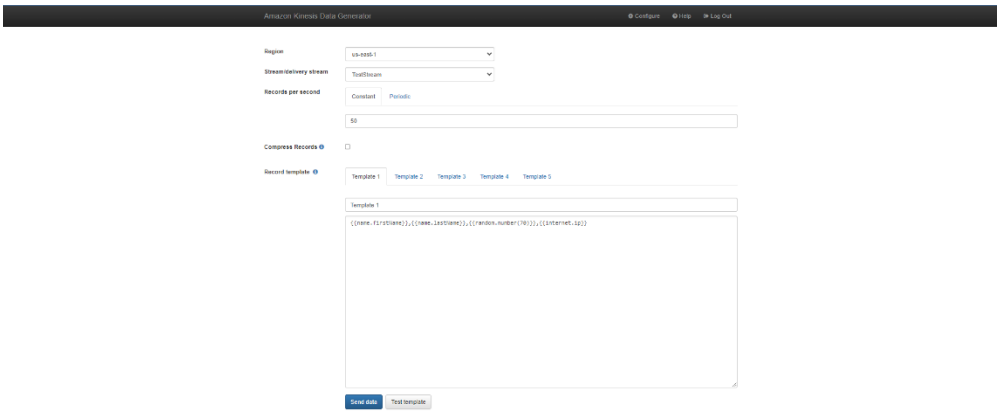


Se connecter avec les identifiants précédemment créés : valentinpuillandre@gmail.com



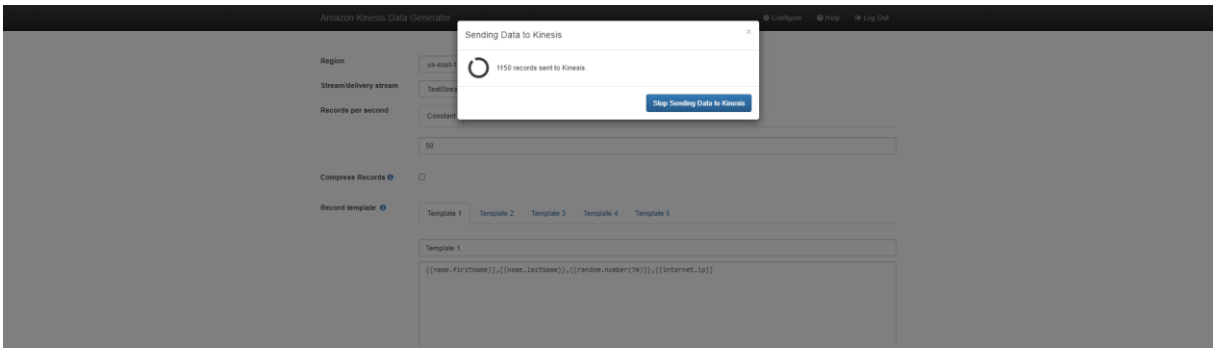
Données pour le template :

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

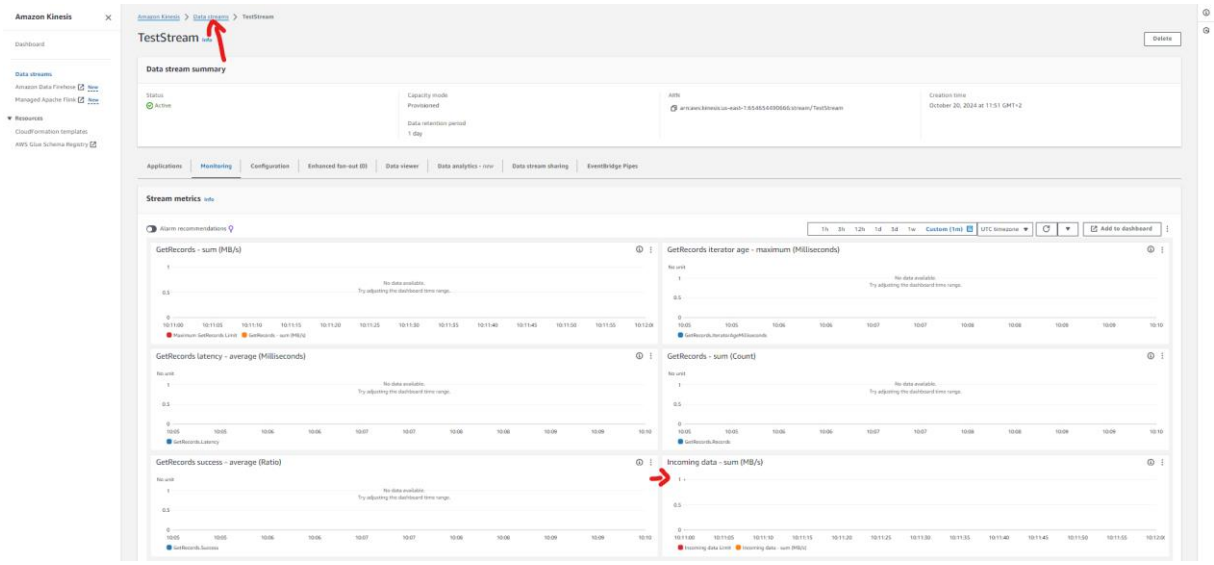


PUILLANDRE Valentin SAULNIER Guillaume

Les datas sont maintenant envoyées



En laissant un peu tourner on peut voir des données apparaitre



PUILLANDRE Valentin SAULNIER Guillaume

Pipelines de Transformation des Données

Description :

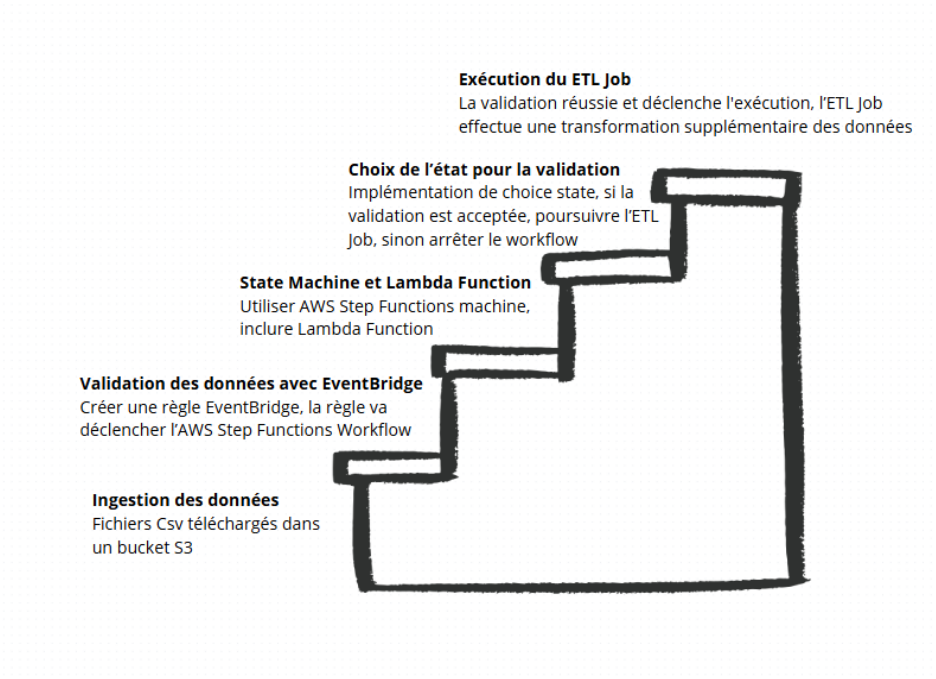
- Conception et développement de pipelines ETL (ou ELT) pour transformer et enrichir les données dans le Data Lake.
- Utilisation d'Apache Spark ou une technologie similaire pour effectuer ces transformations.

Livrables :

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

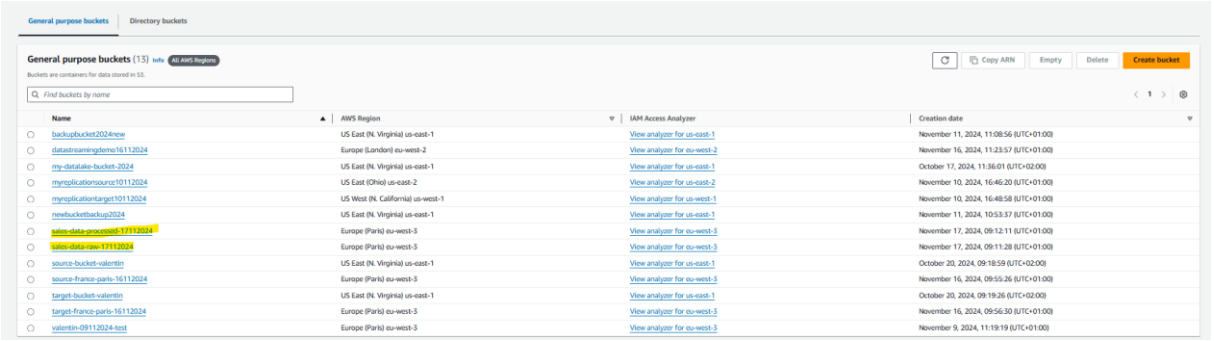
L'objectif ici va être d'automatiser le data Workflow, automatiser l'ingestion et le traitement de la donnée.

Schéma du workflow



Créer Deux buckets

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



PUILLANDRE Valentin SAULNIER Guillaume

Etape 2: Préparer les fichiers

Fichiers présents dans mes dossiers locaux

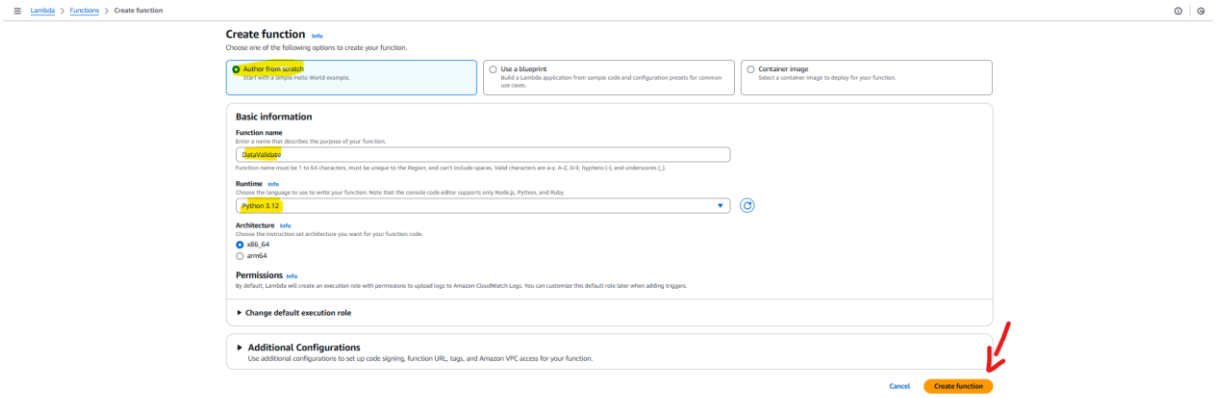
sales_data_london.csv

sales_data_london_different_schema.csv

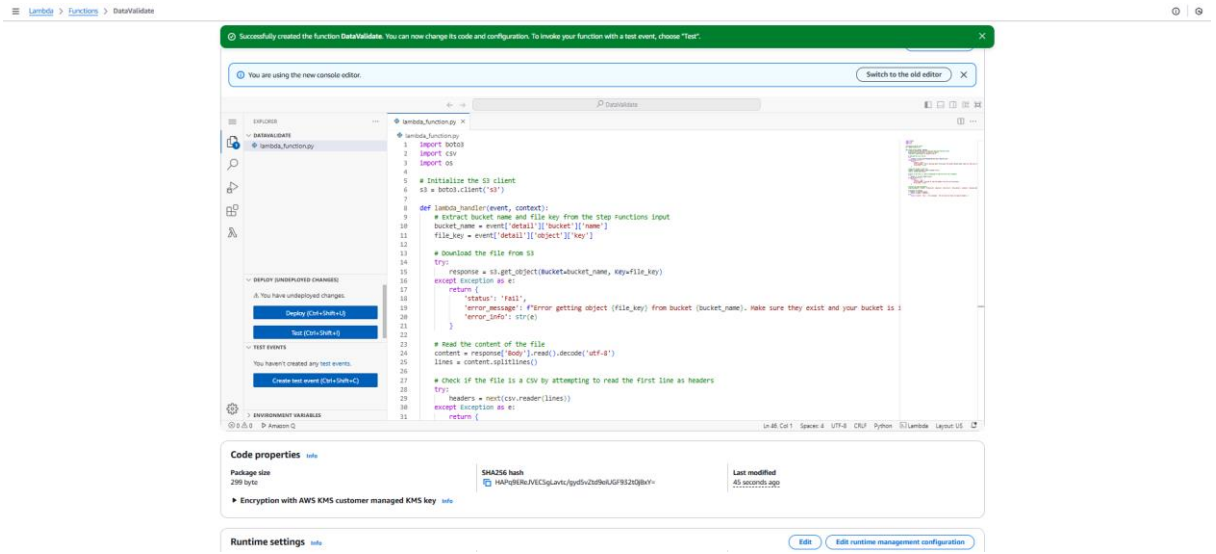
Etapes 3: Créer une Lambda function

- Validation des données
- IAM rôle : S3 access needed

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

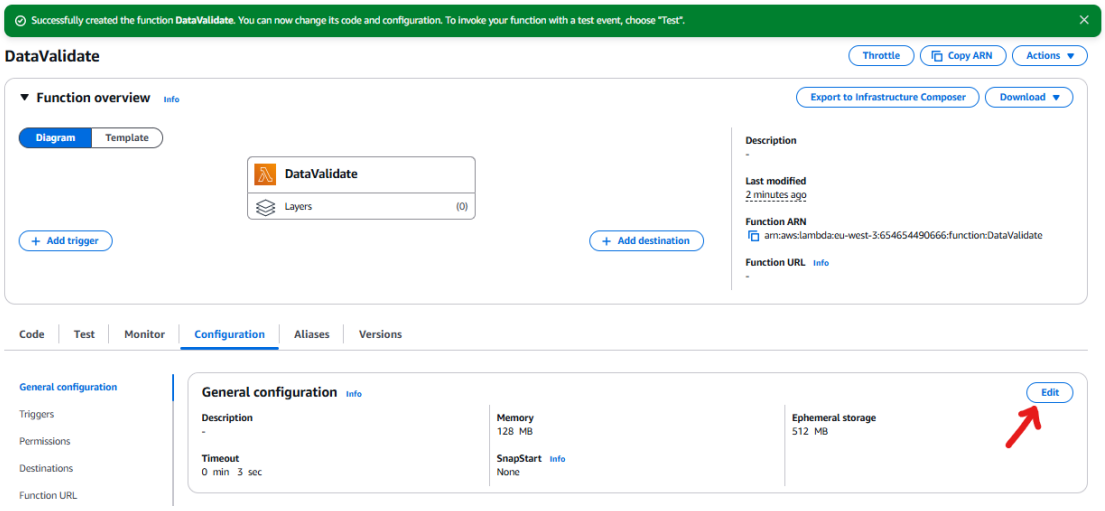


Mettre le code qui fait la validation des données (cf fichier ValidateDataFunction.py)

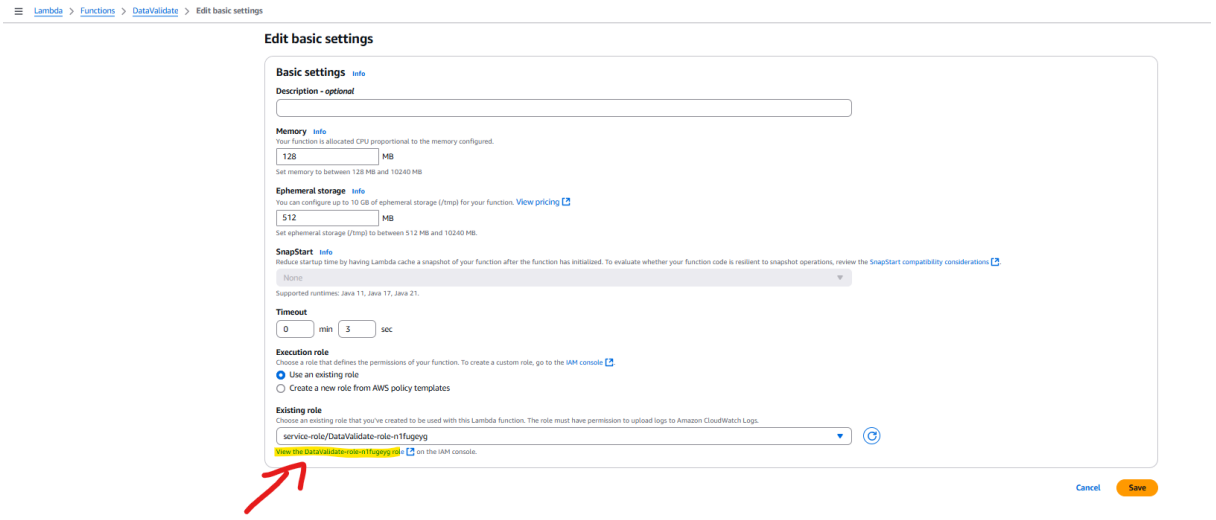


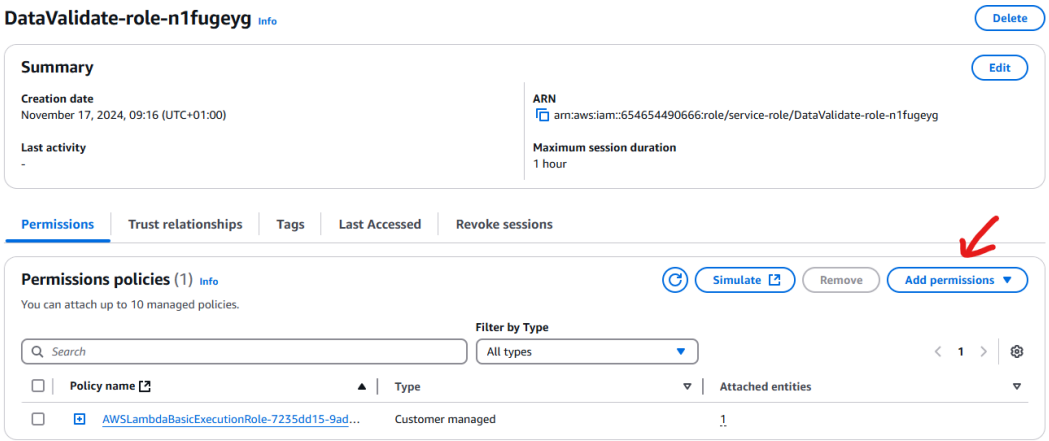
PUILLANDRE Valentin SAULNIER Guillaume

Configurer les bons droits pour utiliser cette Lambda Function



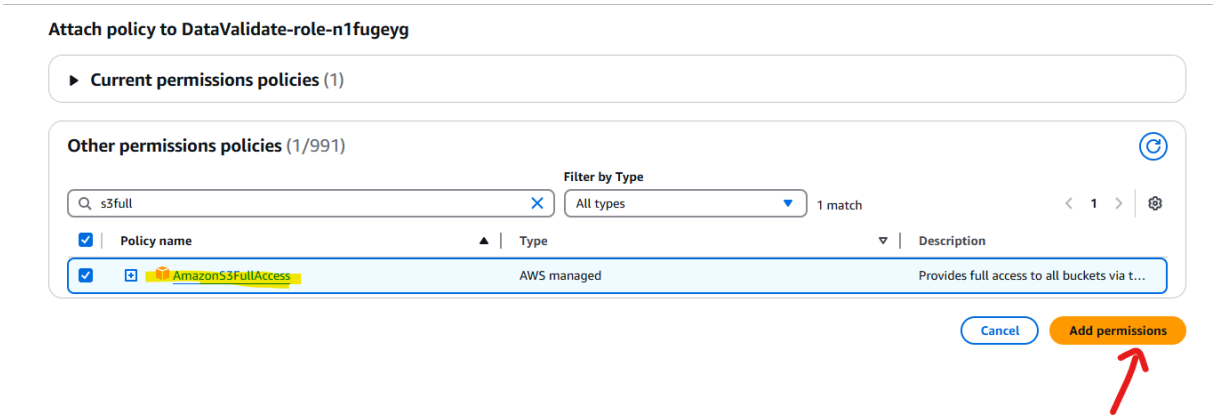
Nous devons ajouter une nouvelle permission pour notre bucket





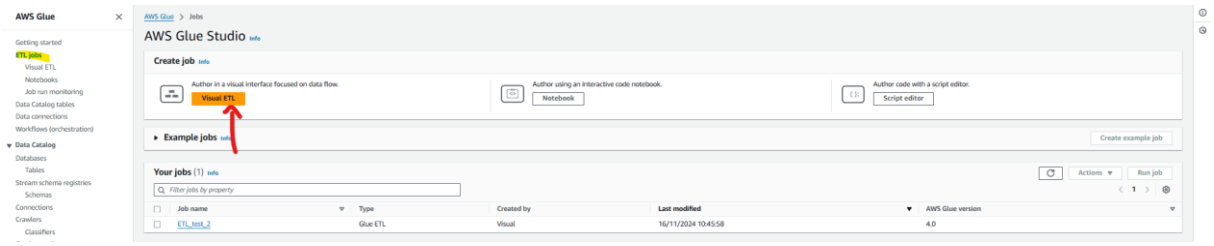
PUILLANDRE Valentin SAULNIER Guillaume

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

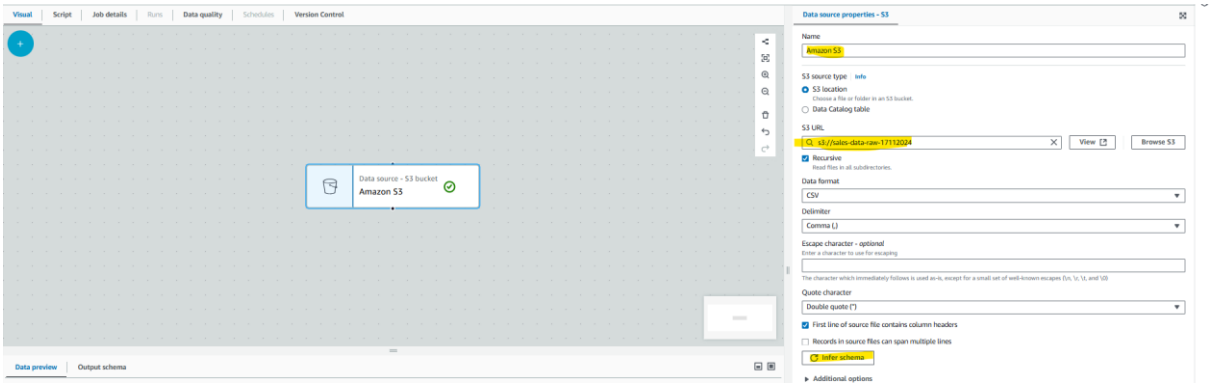


Etape 4: Créer un Job ETL

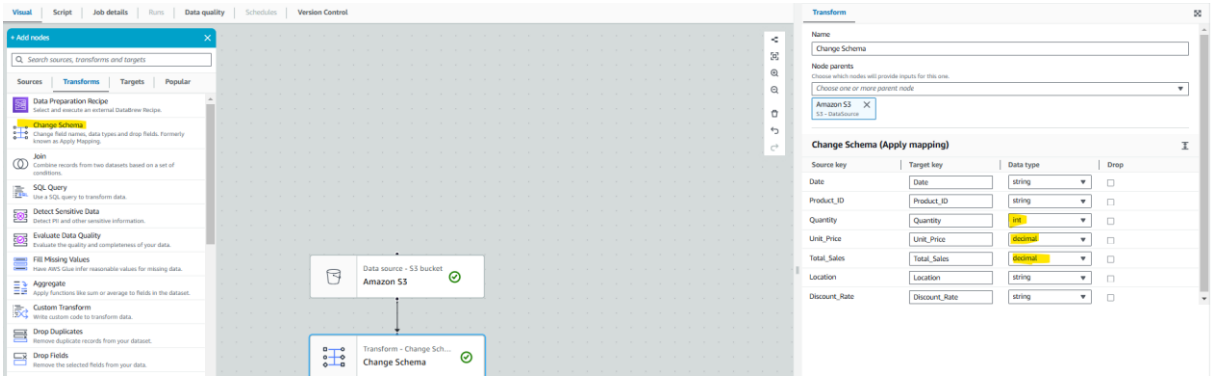
Types des données, Agréger et convertir au format Parquet



Nous allons maintenant créer notre job ETL par la visualisation

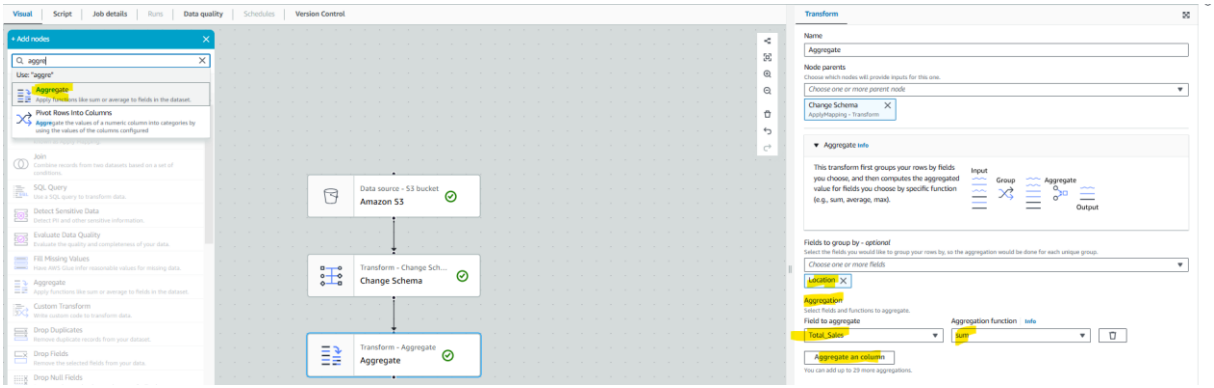


Ajouter un composant “Change schema”

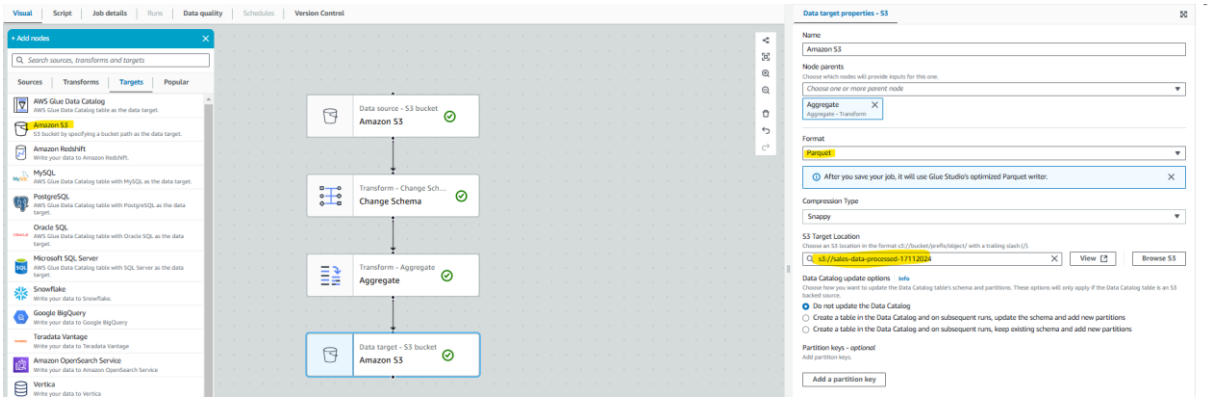


PUILLANDRE Valentin SAULNIER Guillaume

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

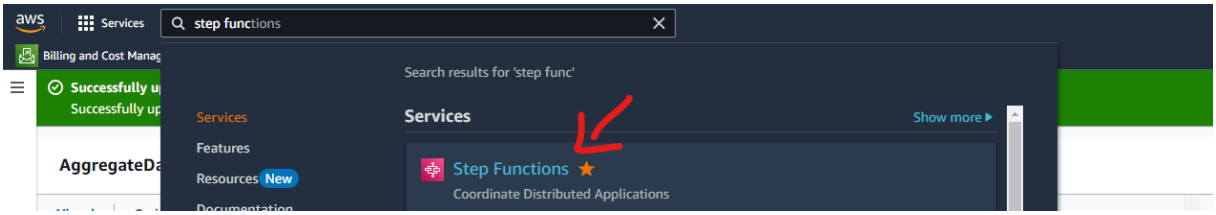


Ajouter le composant bucket S3 target

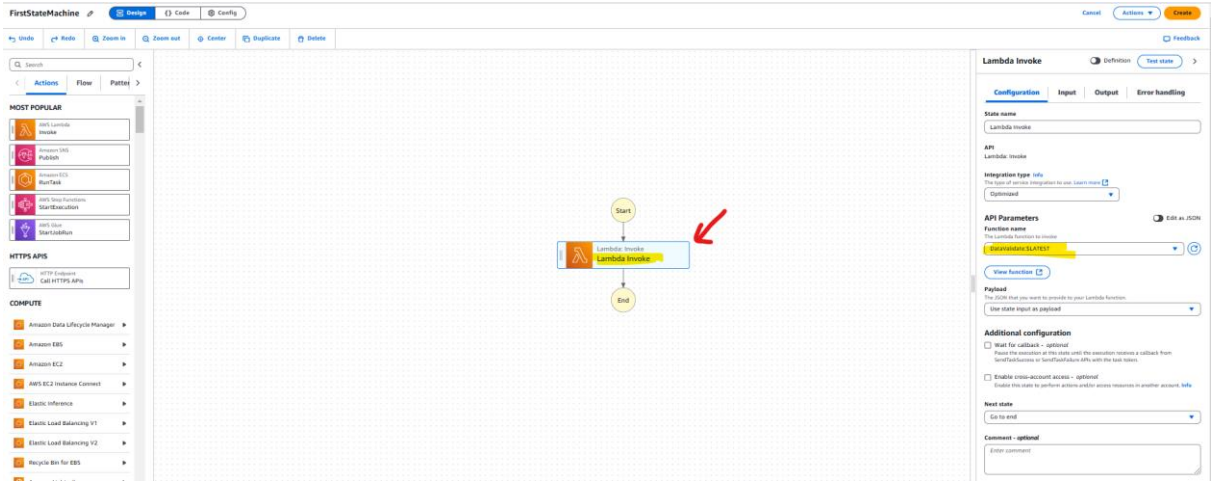


Notre Job ETL est maintenant terminé

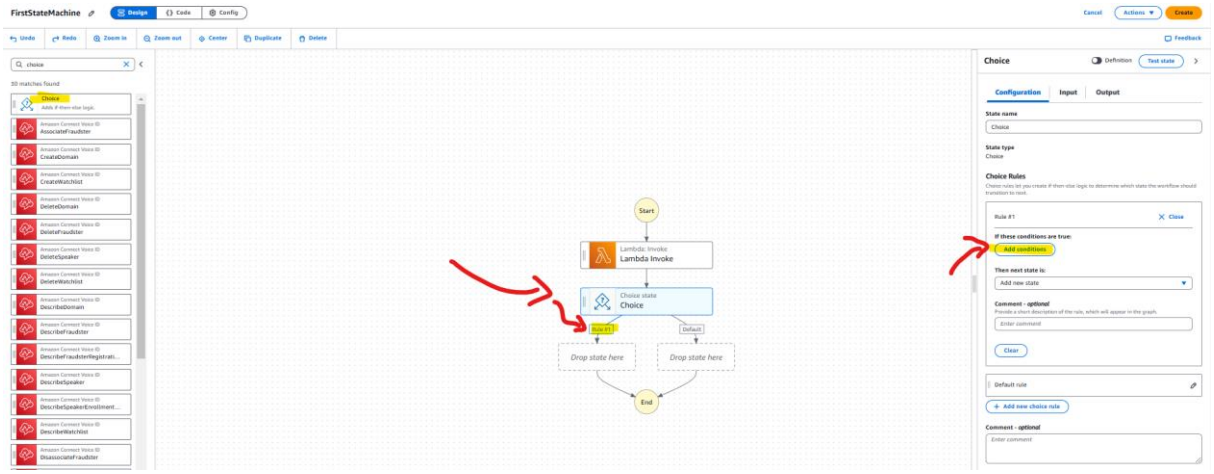
Etape 5: Créer notre State Machine



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

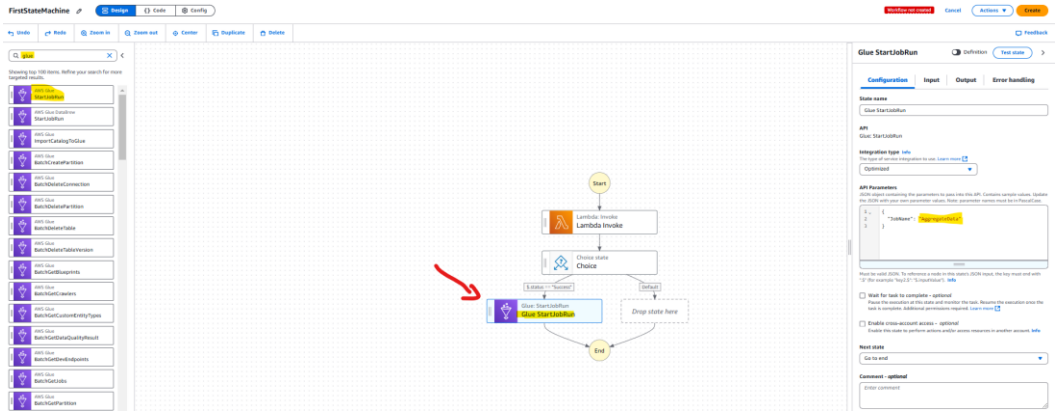


Ajouter le composant “Choice” qui va nous permettre de créer une condition, cliquer sur “add Rule” de l'élement puis “add condition”

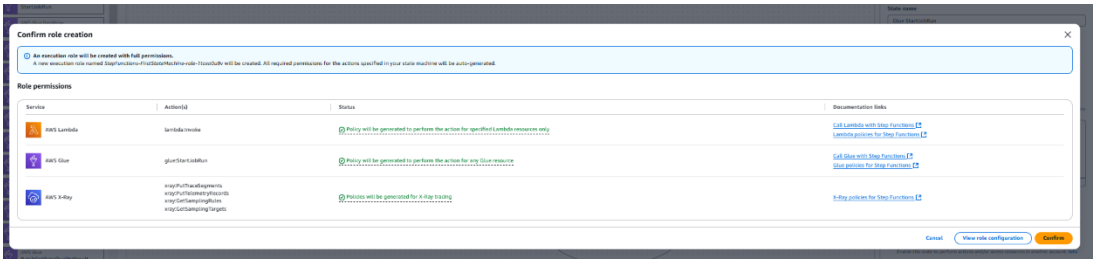


La condition va récupérer le résultat de notre Lamda function créé qui est soit {“status” : “Fail”} soit {“status” : “Success”}, ici on va récupérer en cas de Success

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”

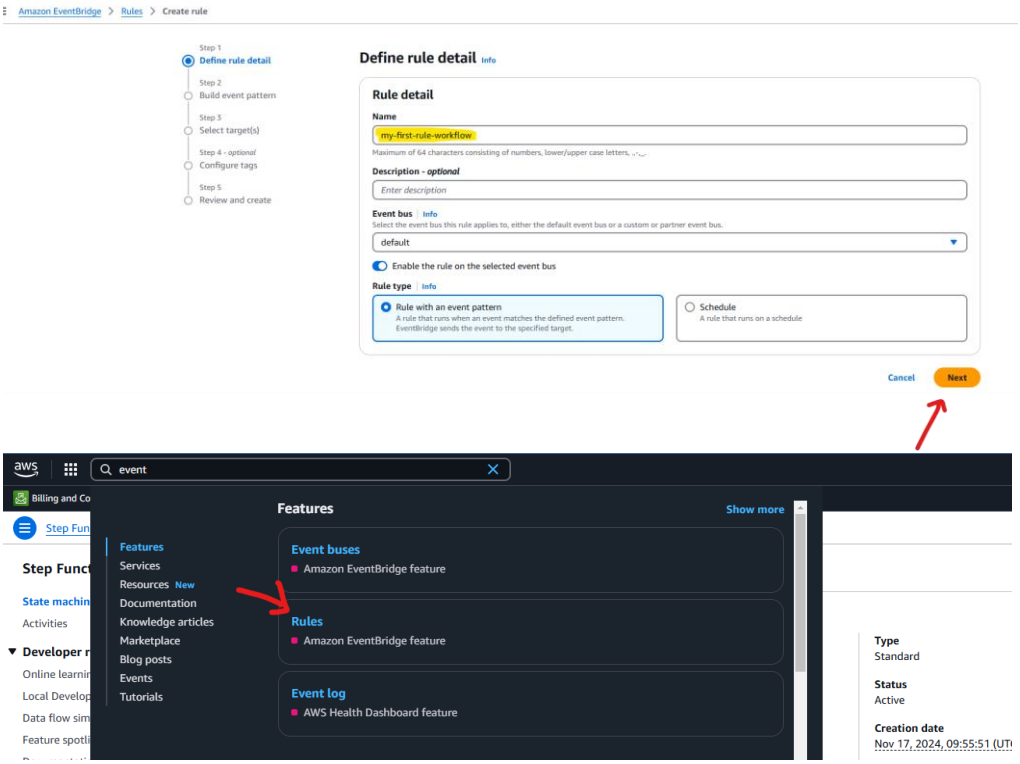


Appuyez sur le boutton “Create” en haut à droite qui va nous permettre d’avoir une vue d’ensemble de la State machine que nous créons, appuyez sur le boutton “confirm”

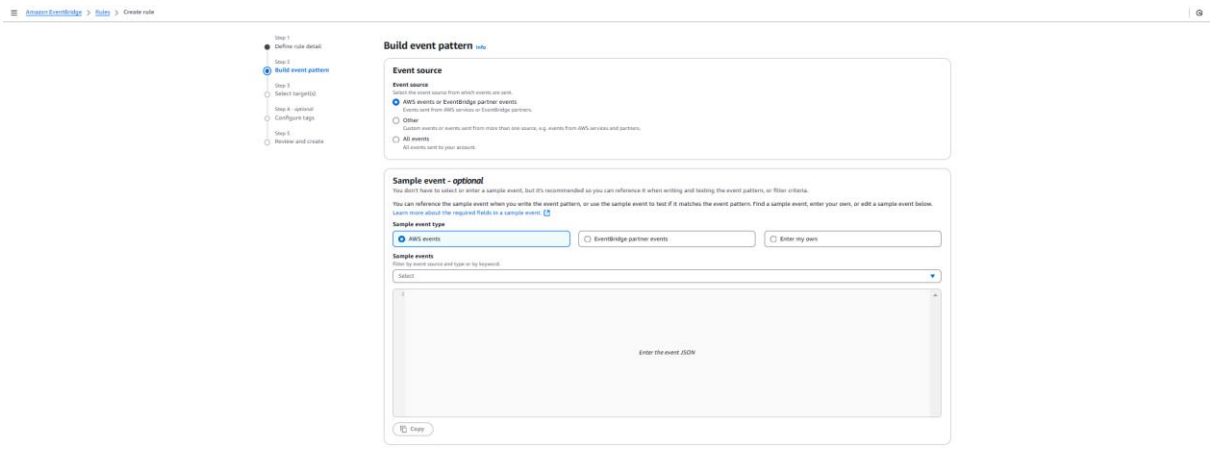


Etape 6: Déclencher le workflow créé

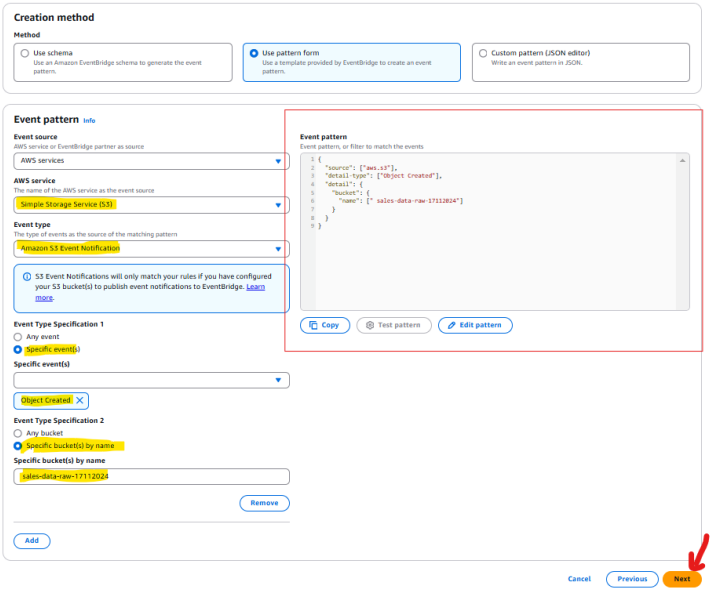
Nous allons ici créer une règle avec Amazon EventBridge



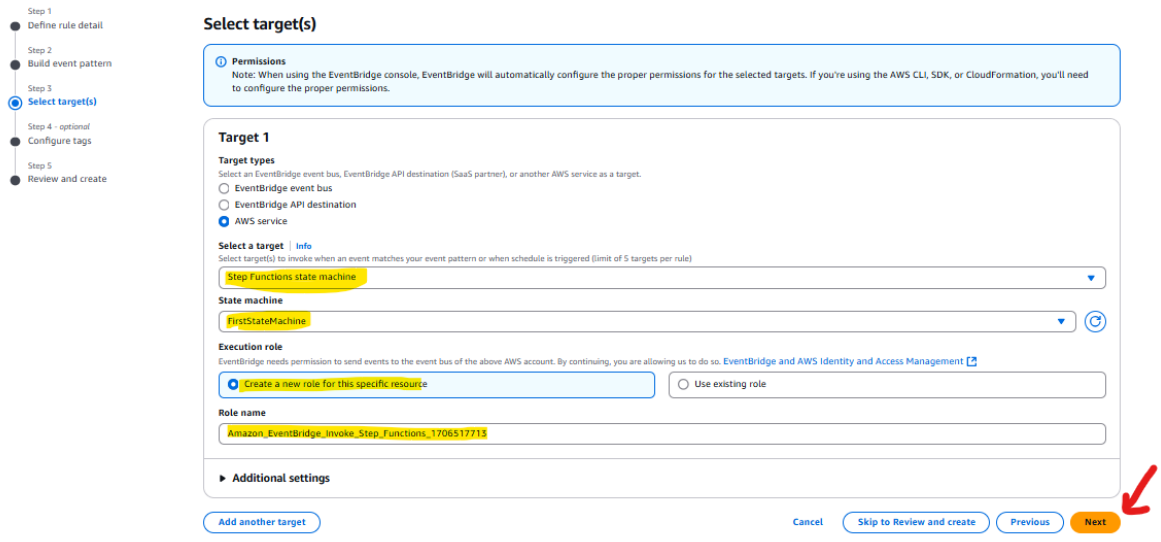
PUILLANDRE Valentin SAULNIER Guillaume



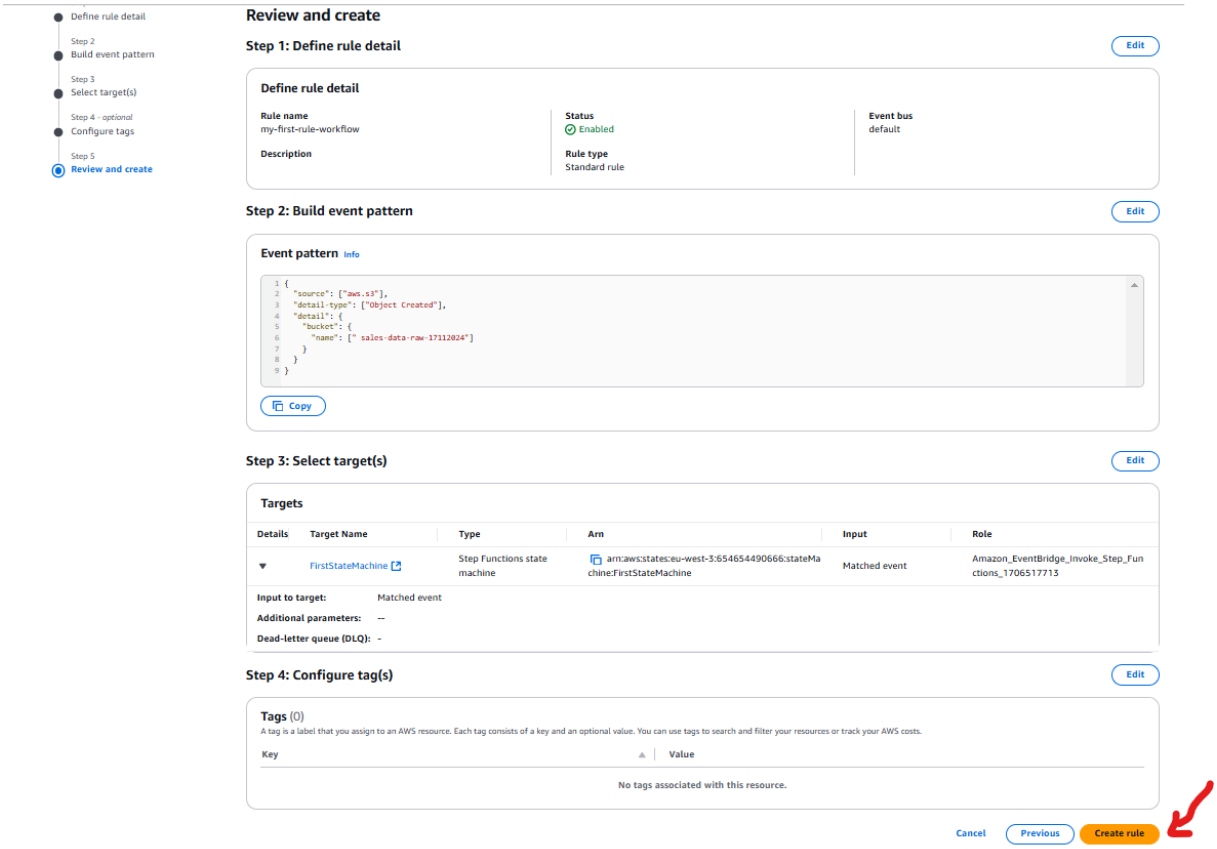
Nous allons maintenant choisir l'Event pattern, sur un évènement spécifique de création d'objet dans notre bucket "sales-data-raw-17112024"



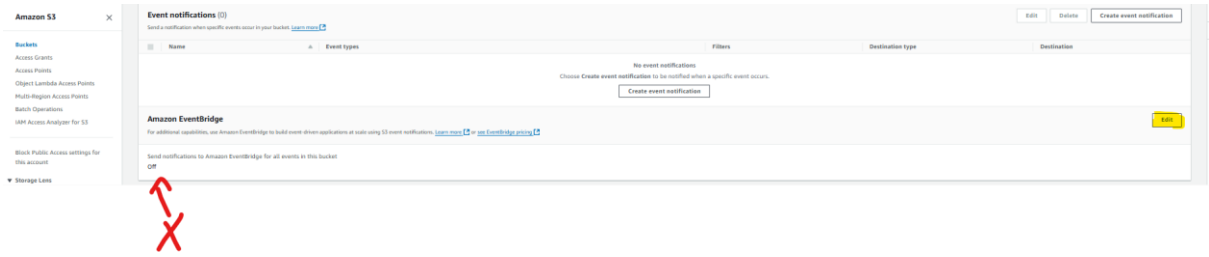
Sélectionner la target qui va être utilisée pour notre règle, notre State machine "FirstStateMachine"



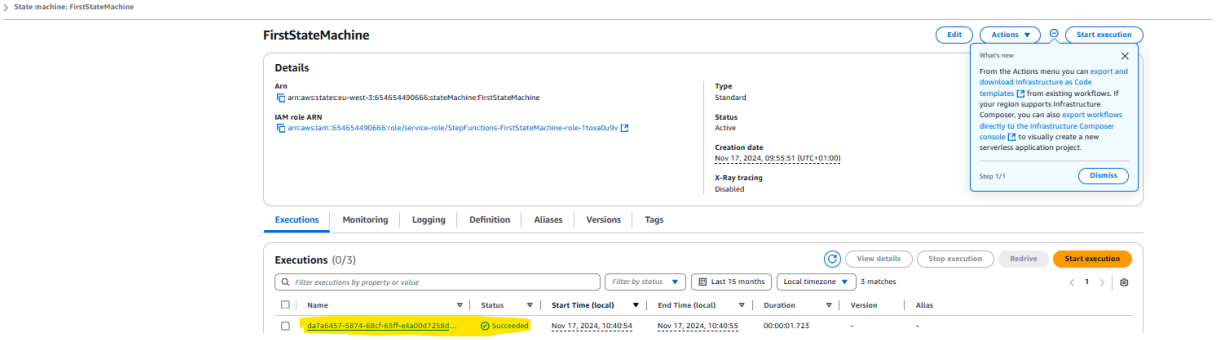
PUILLANDRE Valentin SAULNIER Guillaume



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



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



Partie 3 : Analyse et Exploitation des Données

Description :

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

- L'API devra exécuter des repûtes 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.

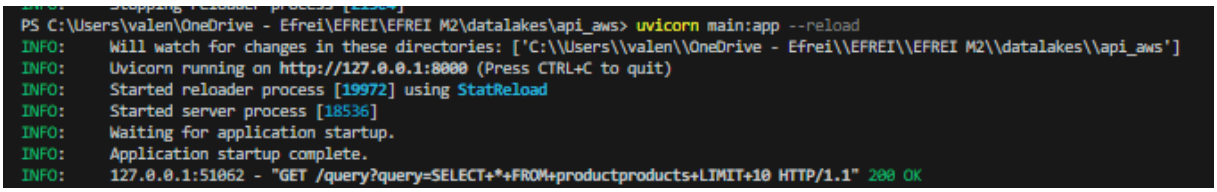
Livrables :

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

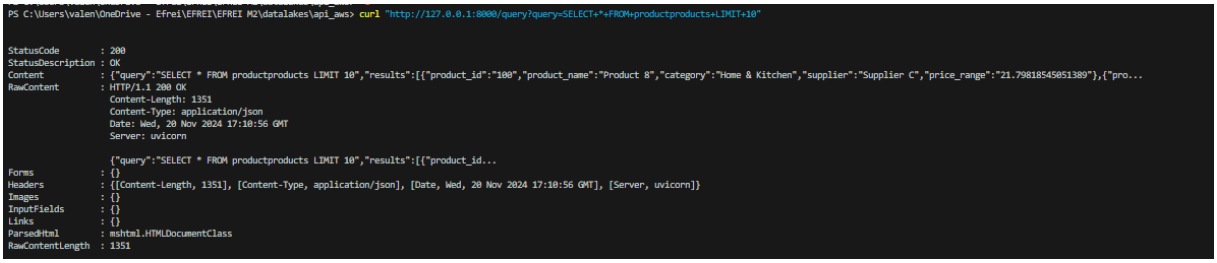
API

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.



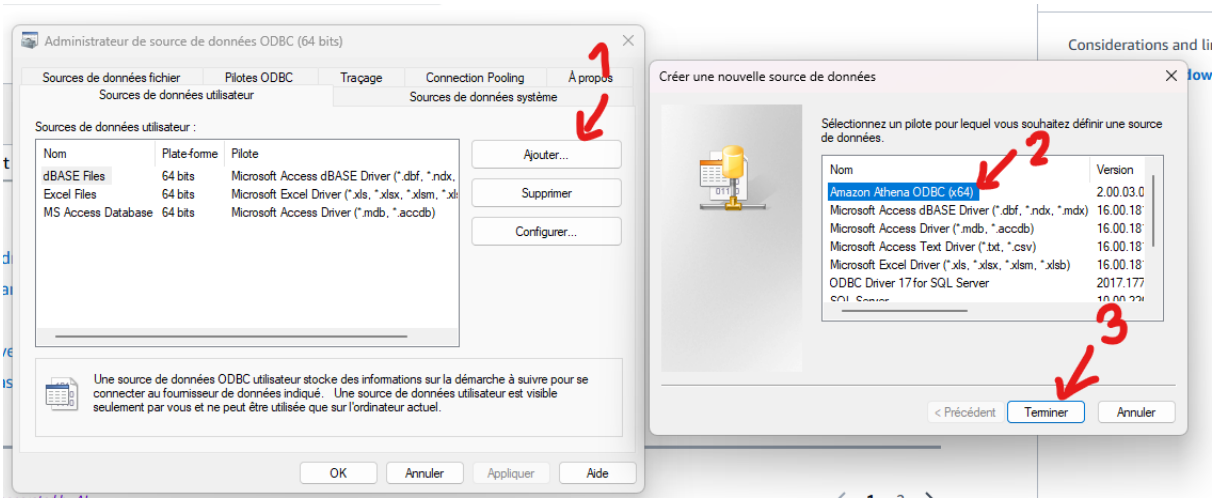
Nous pouvons voir ici le résultat de notre requête précédemment envoyée au format JSON.



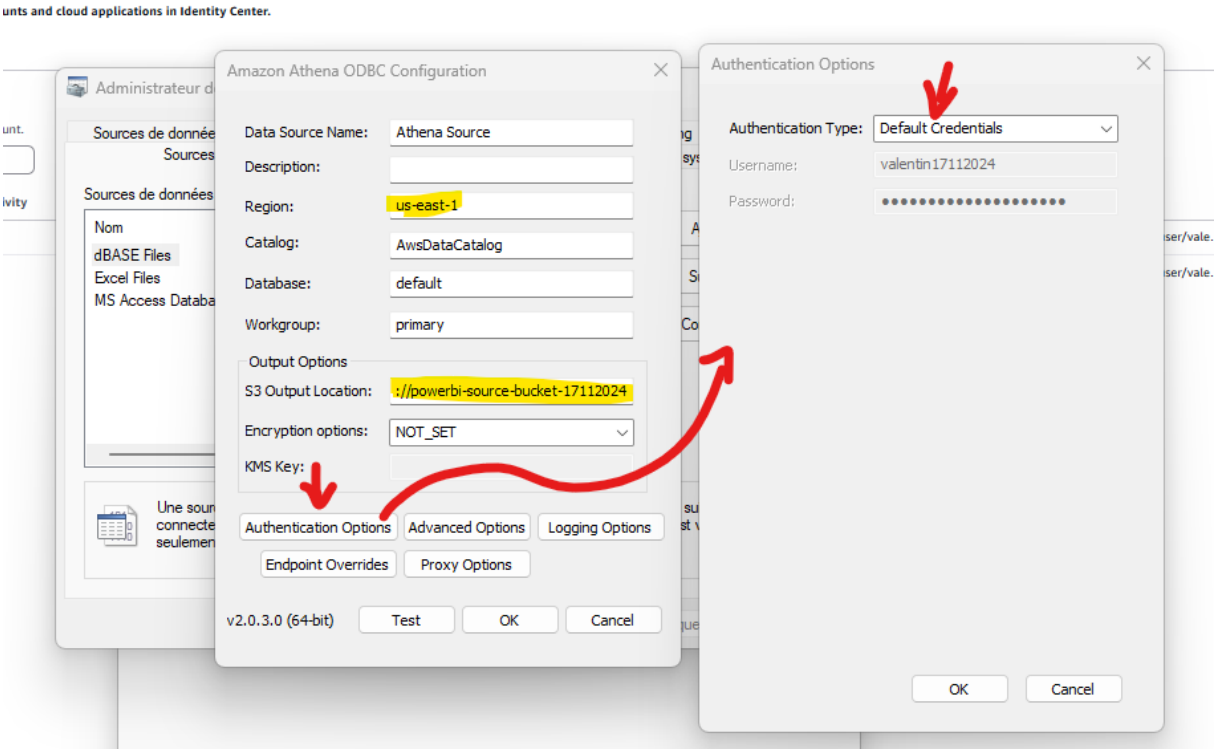
POWER BI

Pour cela nous allons devoir téléchargez 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

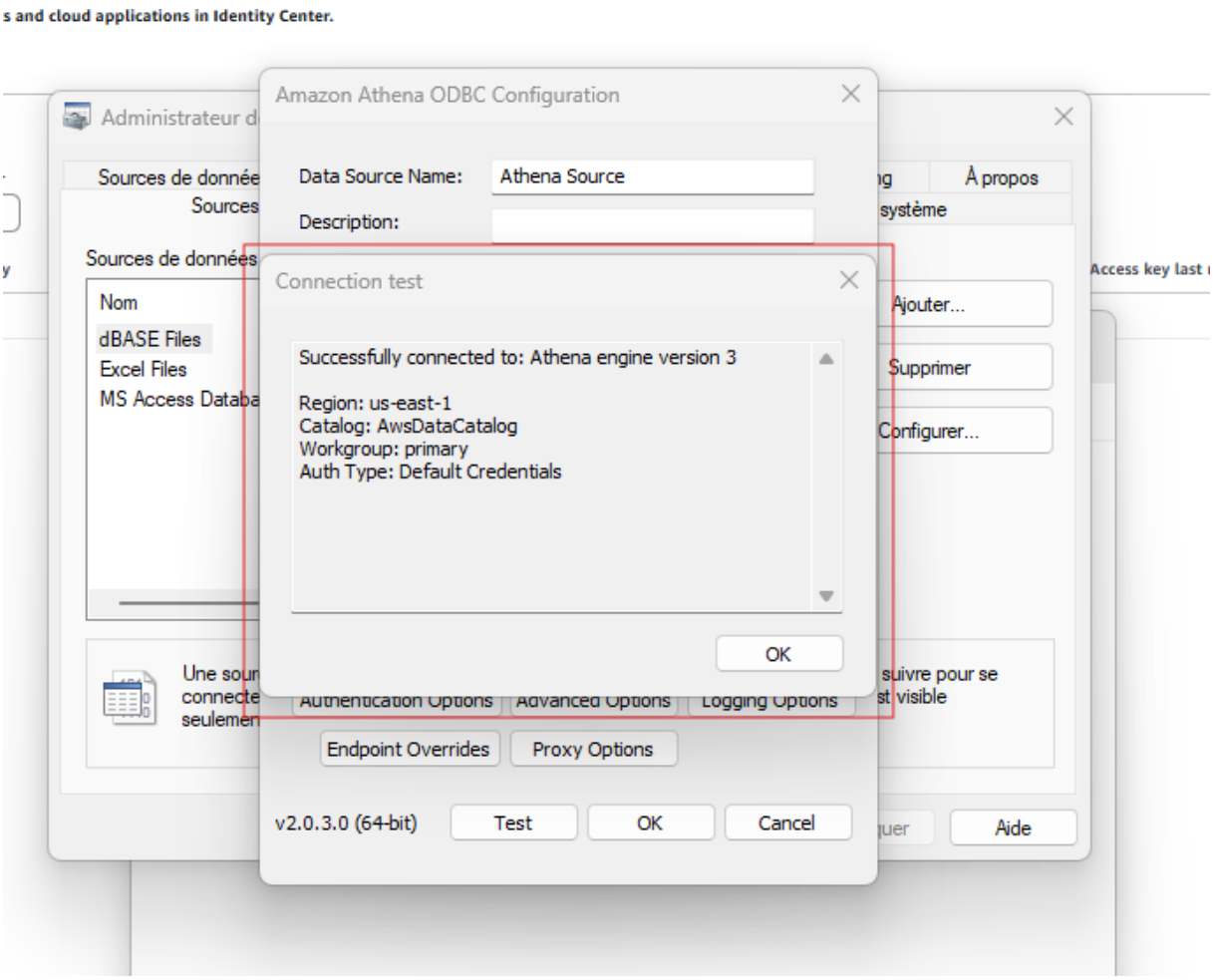


Créer un bucket source pour recevoir les données d'ODBC, appelé "powerbi-source-bucket-17112024", ne pas oublier de mettre "s3://" avant le nom du bucket qui sera utilisé "s3://powerbi-source-bucket-17112024". Terminez la configuration d'odbc source Athena.

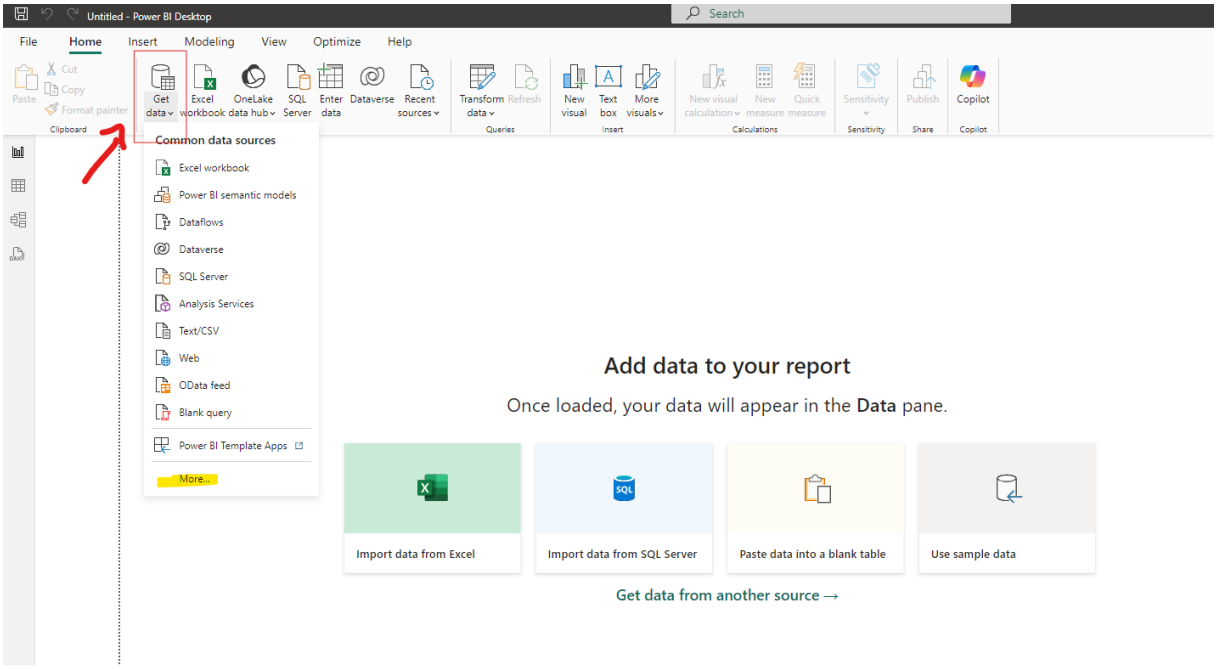


PUILLANDRE Valentin SAULNIER Guillaume

La connection a bien été autorisée

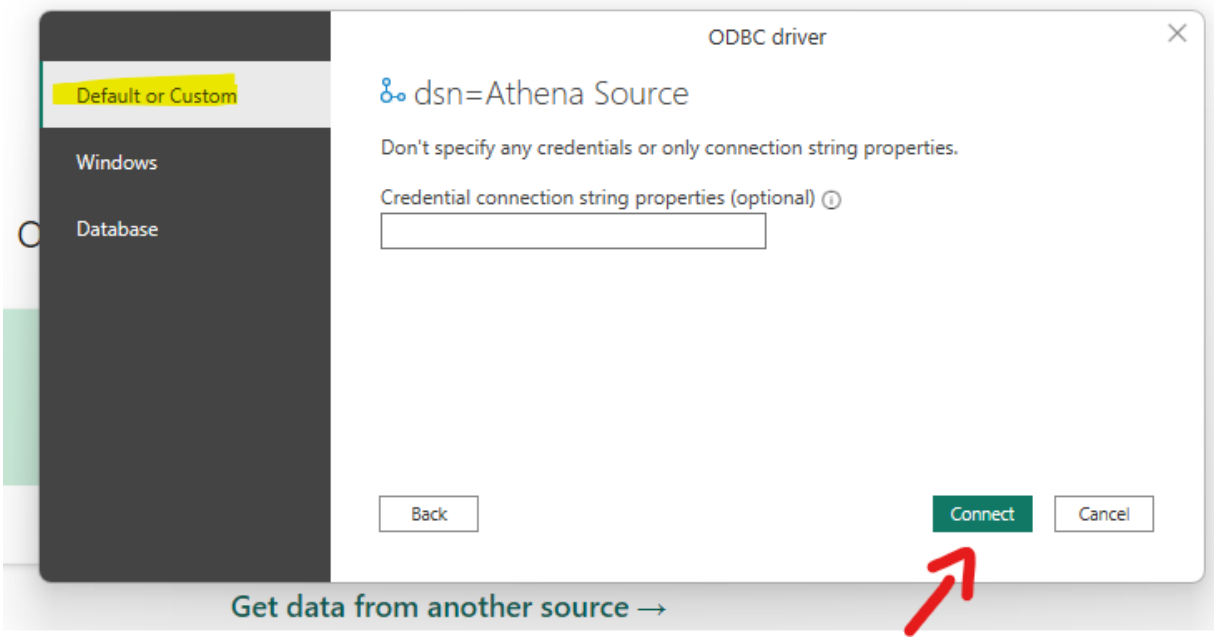


Il ne nous reste plus qu'à connecter power bi

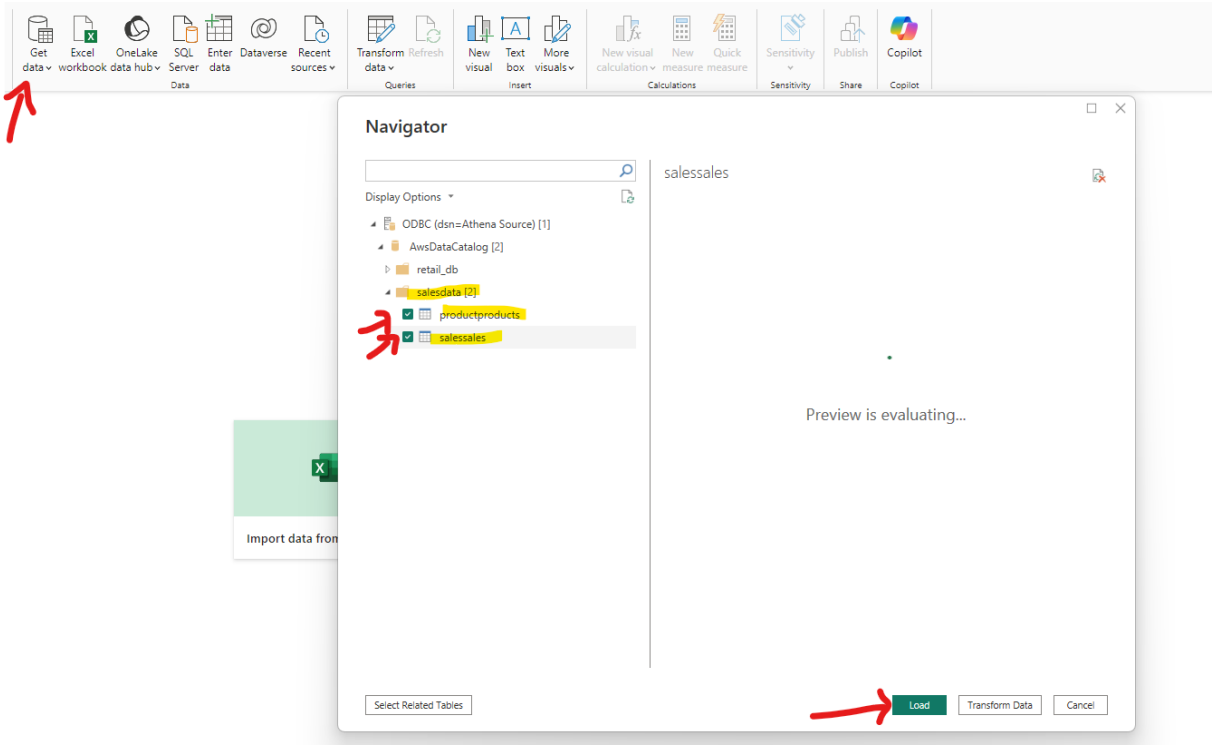


PUILLANDRE Valentin SAULNIER Guillaume

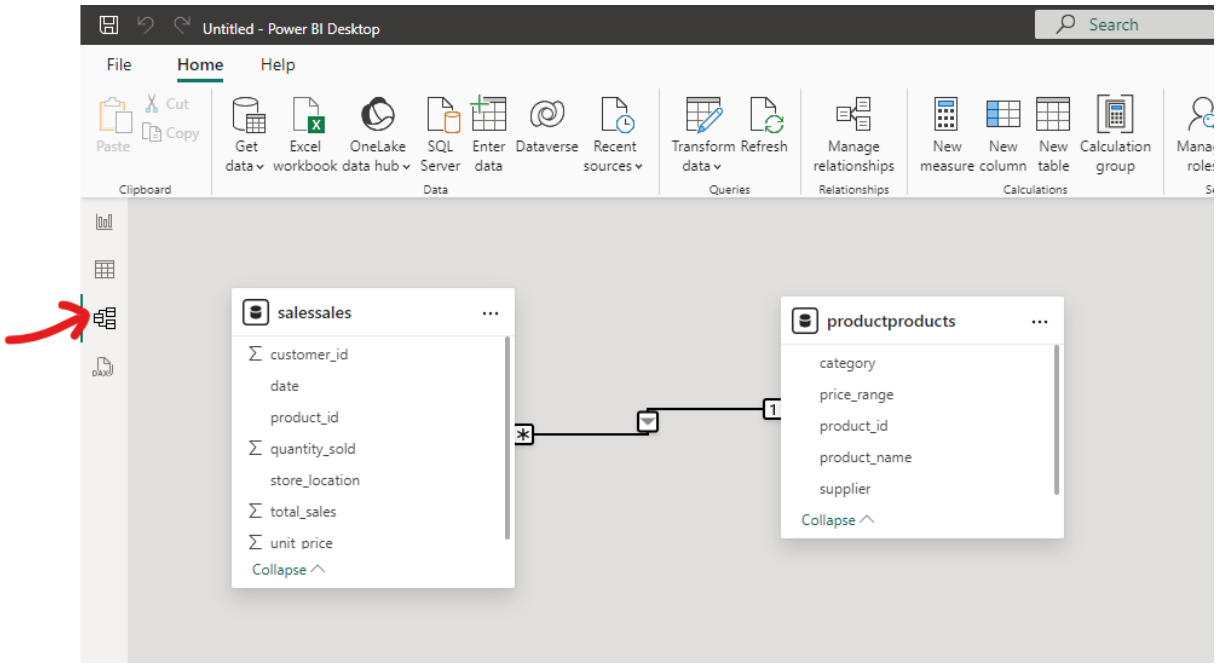
Trouver l'odbc et établir la connection, choisir la "Athena Source", choisir la "Default or Custom" et appuyez sur le boutton "Connect"



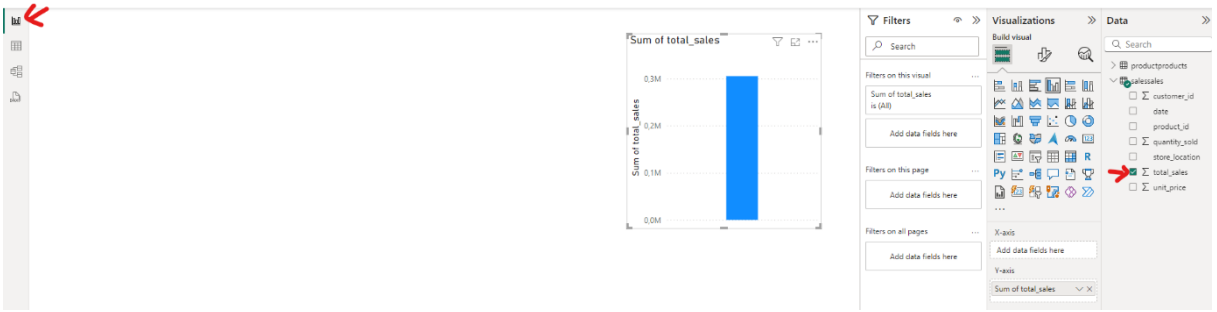
Nous avons maintenant accès à nos différentes tables présentes sur Amazon Athena, les sélectionner et les charger "load"



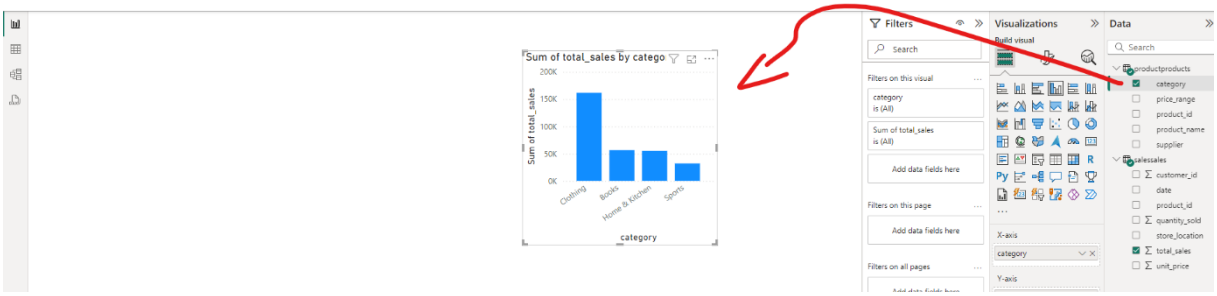
Nous avons maintenant un aperçu de nos tables dans power BI !



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 :

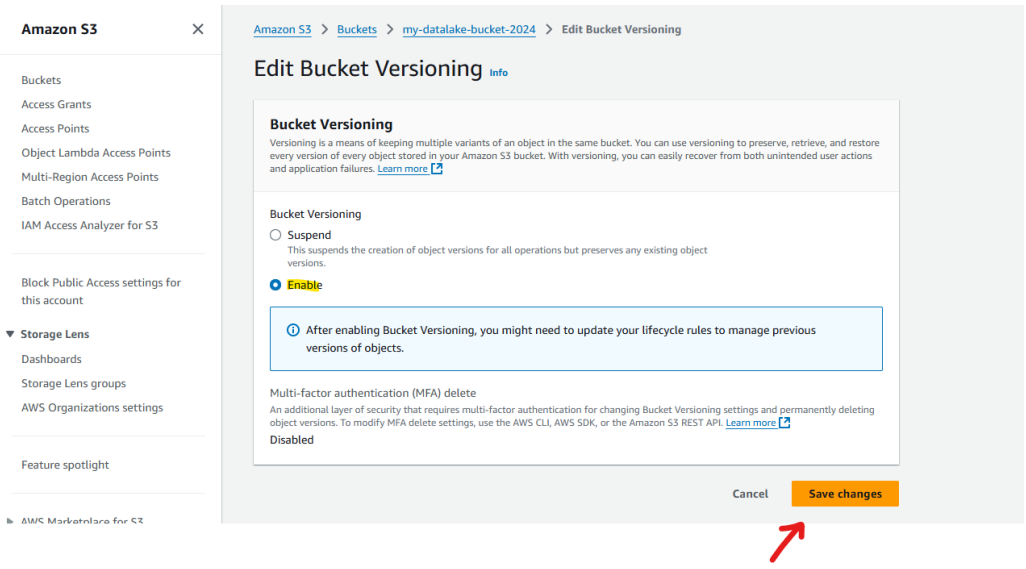
- Mise en place de solutions pour assurer la qualité des données ingérées et transformées : détection d'anomalies, gestion des valeurs manquantes.
- Surveillance des pipelines de données pour garantir leur abilité.

Livrables :

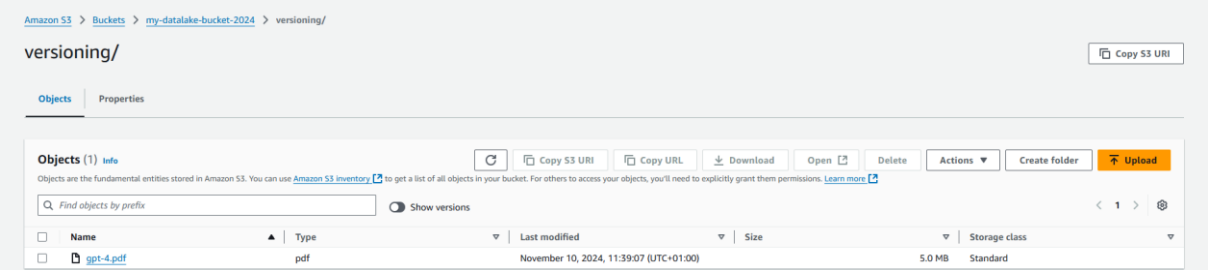
- Scripts ou outils pour la vérification de la qualité des données.
- Rapport de qualité des données avec des indicateurs clés (anomalies détectées, données manquantes, etc.).

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.

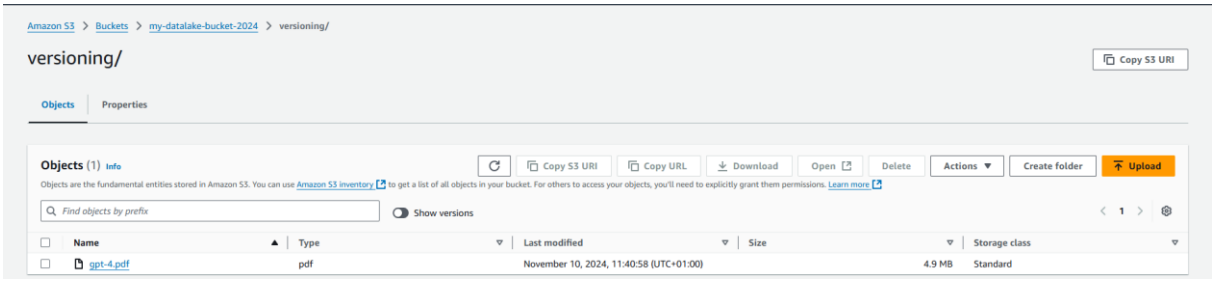
Versioning



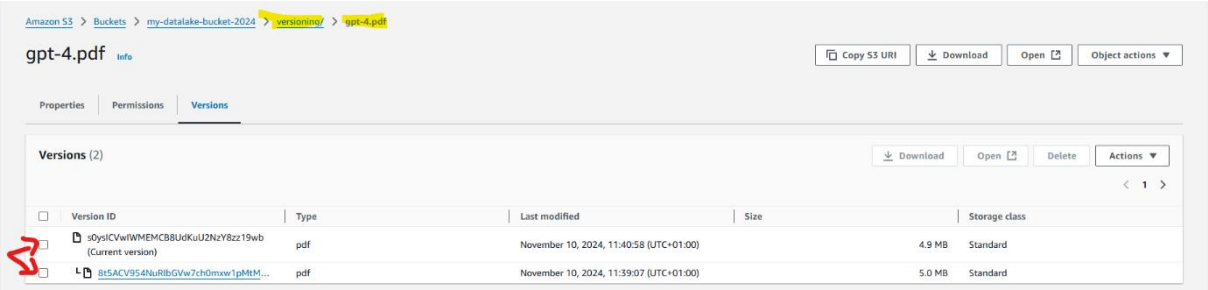
Nous allons maintenant comprendre ce qu'il se passe dans le versioning, d'abord en créant un dossier dans le bucket S3 et y ajouter un fichier



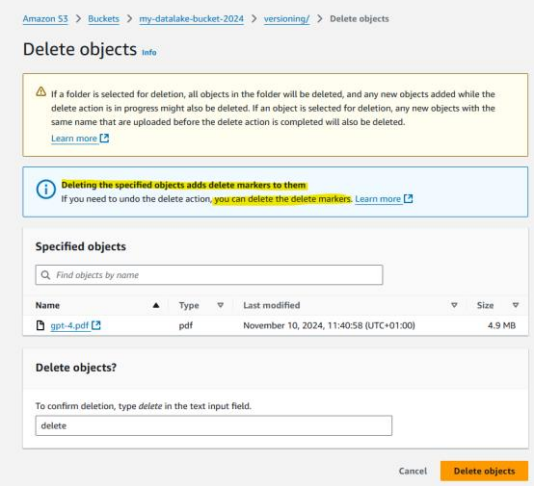
Nous allons maintenant remplacer le fichier actuel par une version plus récente modifiée, nous ne voyons pas de différence, le fichier a bien été remplacé



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.

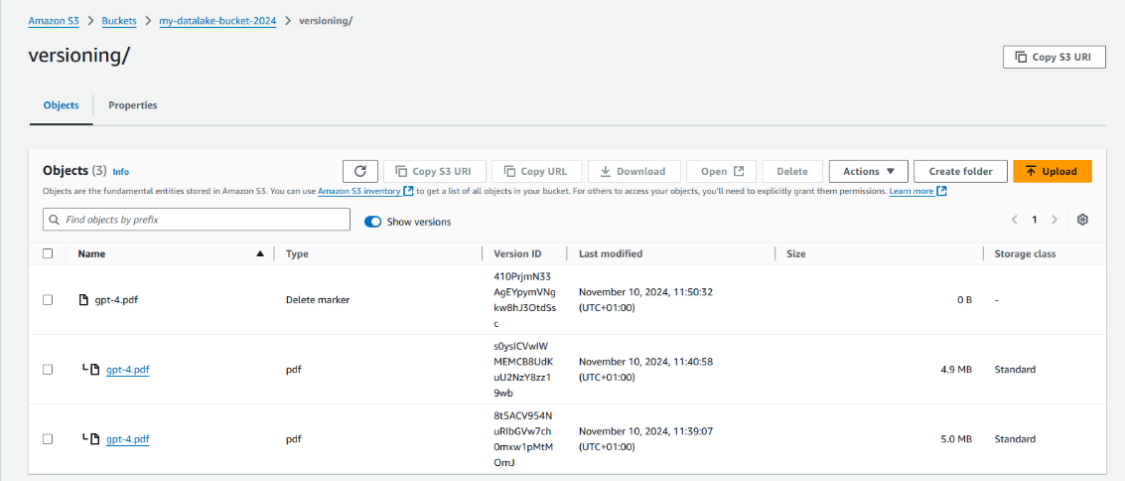


Maintenant que se passe-t-il si l'on supprime le fichier ? Cela va ajouter un "marqueur de suppression" mais ne va pas le supprimer de manière permanente.



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.



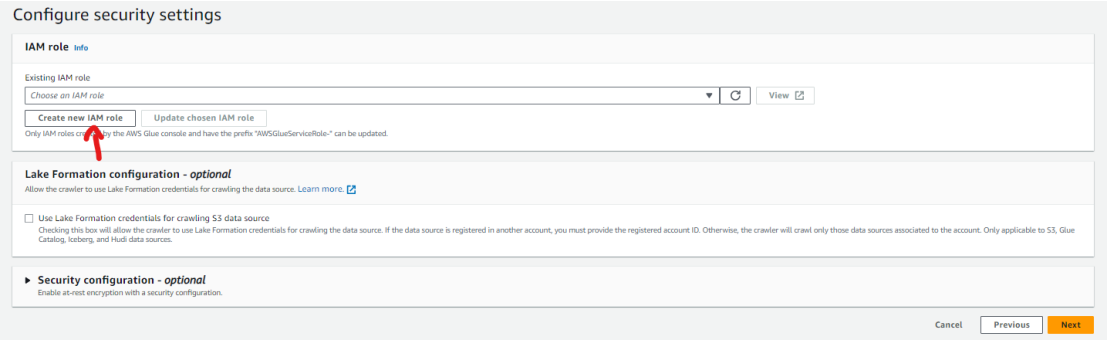
IAM

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.

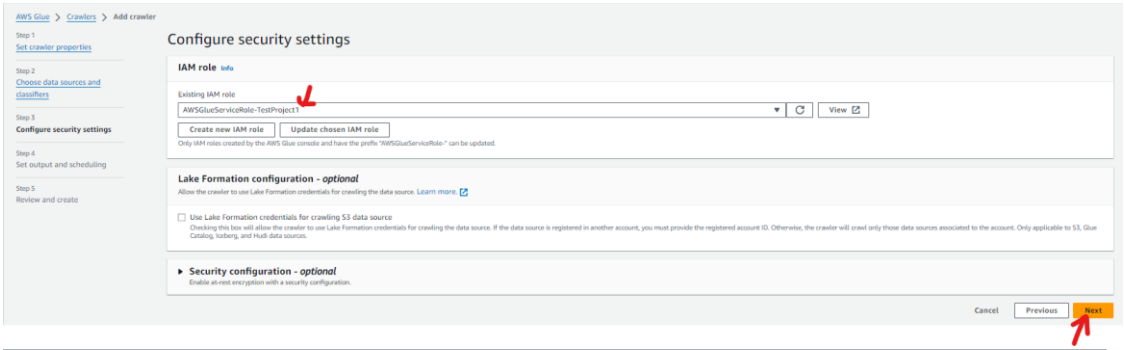
Il offre la possibilité de créer et de gérer des utilisateurs et des groupes, et de définir des autorisations précises pour contrôler l'accès aux ressources AWS.

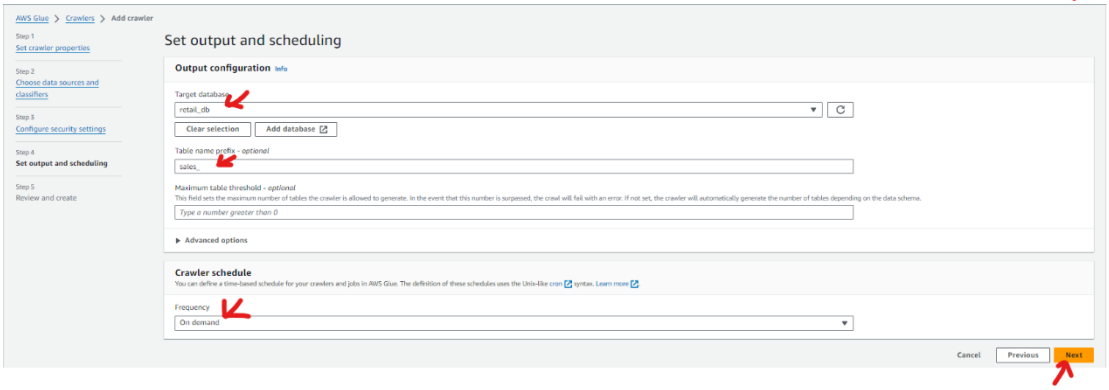
IAM assure ainsi une gestion fine des droits, garantissant que chaque utilisateur dispose uniquement des permissions nécessaires à l'exécution de ses tâches.

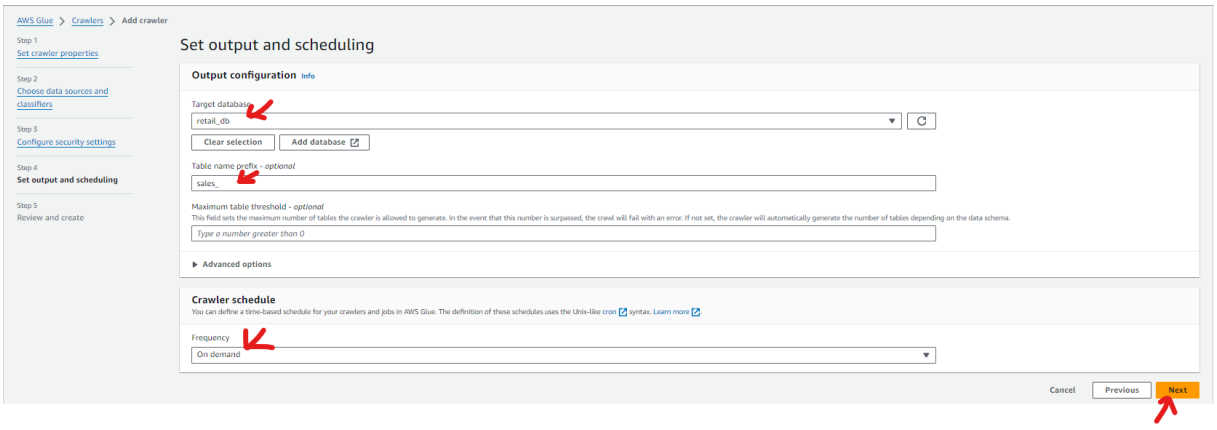
Dans la pratique, en créant un rôle IAM, ce rôle sera limité au chemin de la données indiquée



PUILLANDRE Valentin SAULNIER Guillaume







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.