www.eazytraining.fr

Services d'ingénierie des données AWS

Par Lahda Biassou Alphonsine

Lahda Biassou Alphonsine

Ingénieure cloud et formatrice

Plan

- **Amazon Athena**
- **Amazon RedShift**
- **Amazon Kinesis**
- **AWS Glue**
- **Amazon QuickSight**
- **Amazon Elastic MapReduce (EMR)**
- **AWS Lake Formation**
- **Amazon ElasticSearch Service**
- **Amazon Managed Streaming for Apache Kafka**
- **AWS Batch**

Amazon Athena

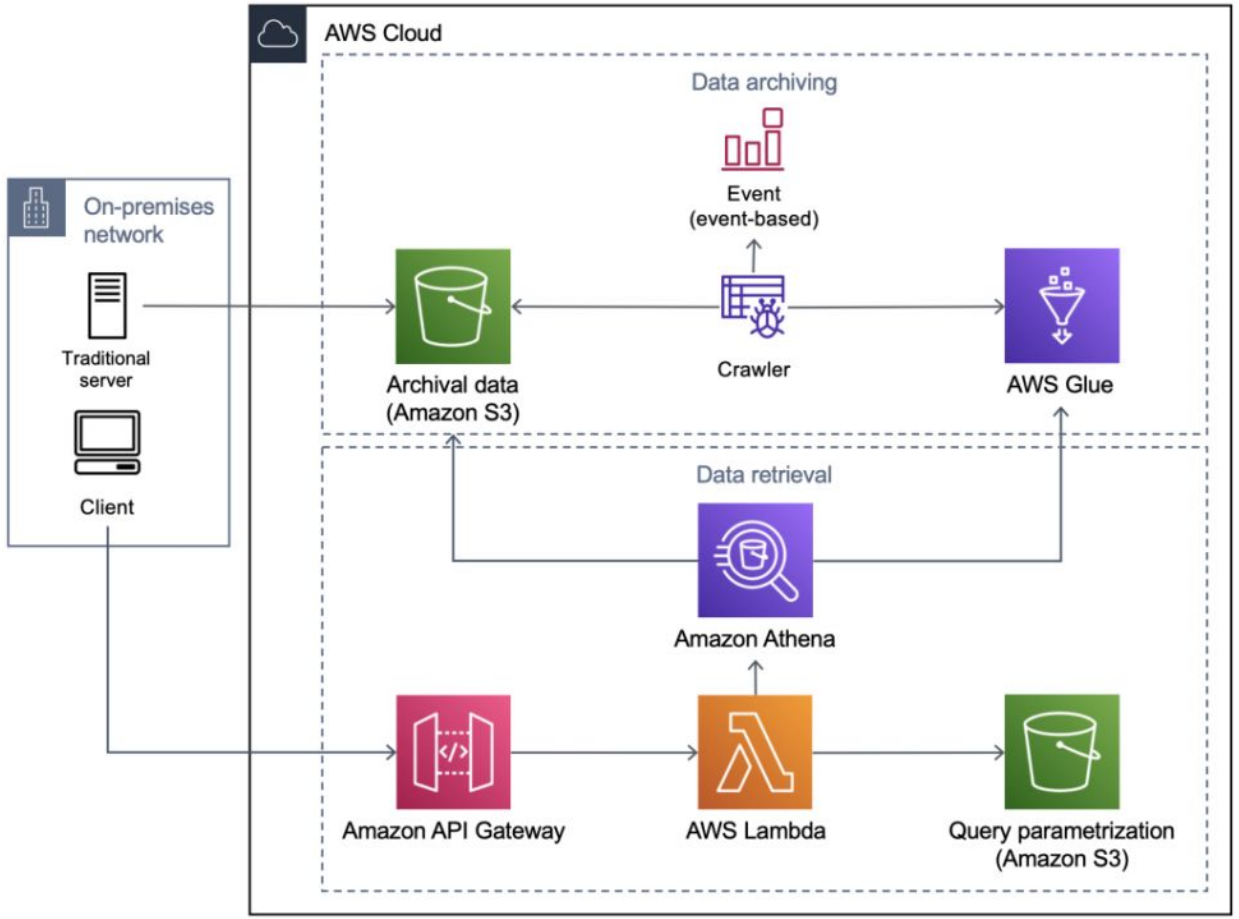
Définition

Amazon Athena est un service interactif sans serveur utilisé pour analyser les données directement dans Amazon Simple Storage Service à l'aide des requêtes SQL standard ad-hoc.

- Il exécute plusieurs requêtes en parallèle, sans avoir à se soucier des ressources de calcul.
- Il prend en charge divers formats de données standard, tels que CSV, JSON, ORC, Avro et Parquet.
- Il permet d'analyser différents types de données (non structurées, semi-structurées et structurées) stockées dans Amazon S3, semi-structurées et structurées) stockées dans Amazon S3.
- Athena permet d'exécuter des requêtes ad hoc en utilisant ANSI SQL sans avoir à charger les données dans Athena.
- Il peut être intégré à Amazon Quick Sight pour la visualisation des données et permet de générer des rapports à l'aide d'outils de veille stratégique.
- Il peut être intégré au catalogue de données AWS Glue pour stocker des métadonnées dans Amazon S3 et offre les fonctionnalités de découverte de données d'AWS Glue.
- Il permet de connecter les clients SQL avec un pilote JDBC ou ODBC.
- Prix : 5,00 $ par TB de données analysées

Amazon Athena

Amazon Athena -composants



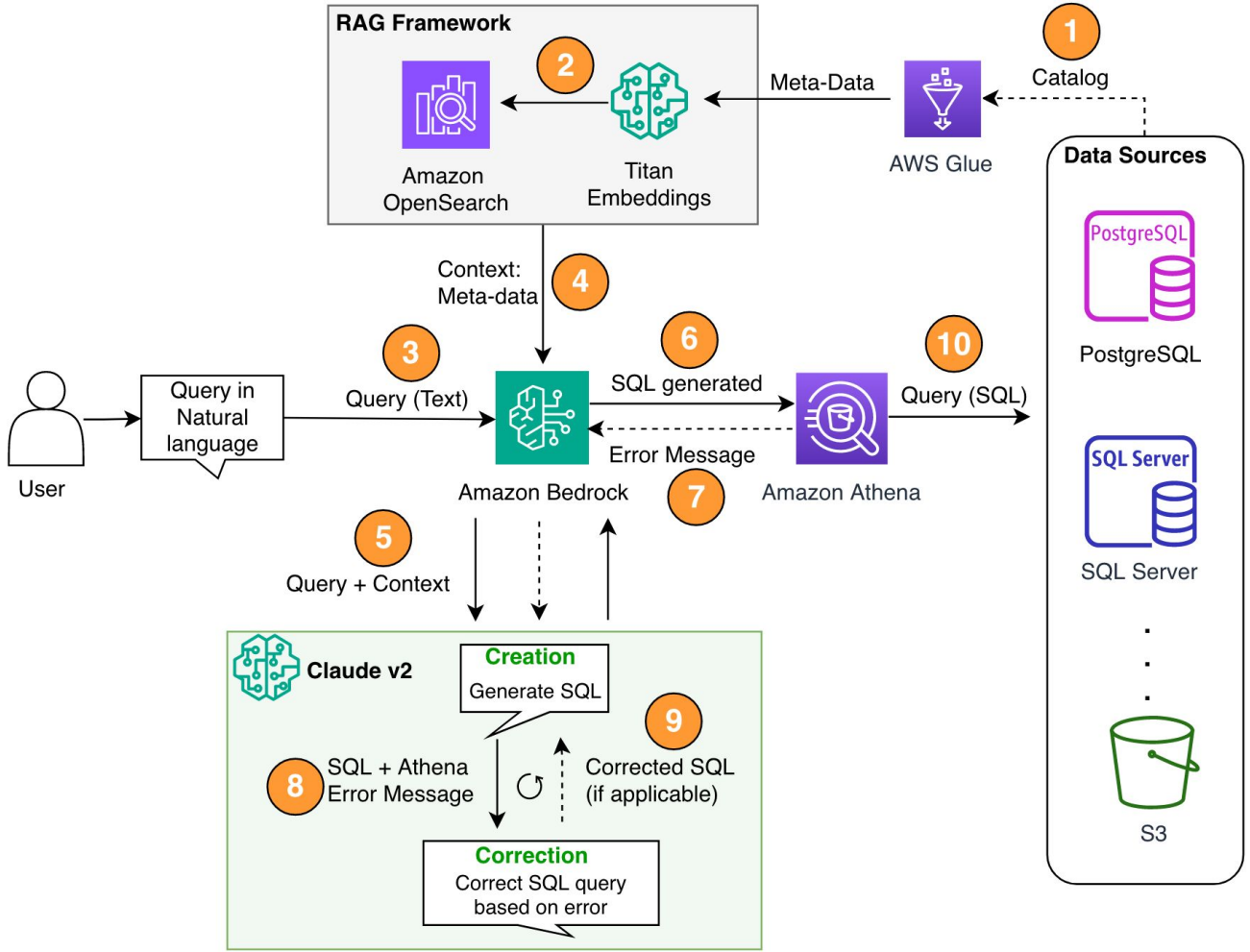
- **Éditeur de requêtes :** L'interface utilisateur fournie par la console de gestion AWS et d'autres applications clientes qui permet aux utilisateurs de soumettre des requêtes et de visualiser les résultats.

- **Moteur de requête :** c'est le cœur d'Athena, qui prend les requêtes SQL et les convertit en plans d'exécution distribués.

- **Magasin de métadonnées :** Athena maintient un magasin de métadonnées qui contient des informations sur les bases de données, les tables, les partitions et le schéma des données stockées dans S3. Ces métadonnées sont essentielles pour la planification et l'optimisation des requêtes.

- **Stockage de l'ensemble des résultats :** Les résultats temporaires de la requête sont stockés dans un emplacement sécurisé et durable dans Amazon S3, ce qui garantit que les résultats sont toujours disponibles et peuvent être récupérés ultérieurement.

Amazon Athena -exemple d'architecture de solution de requêtes complexe



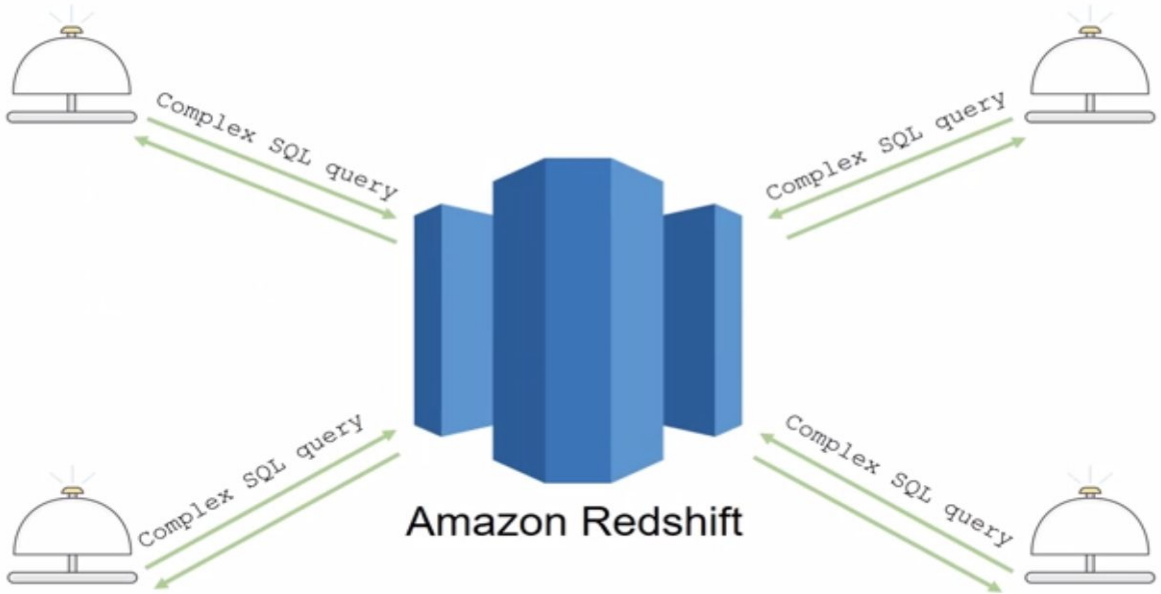
Construire une solution robuste de conversion de texte en SQL générant des requêtes complexes, s'auto-corrigeant et interrogeant diverses sources de données.

Plan

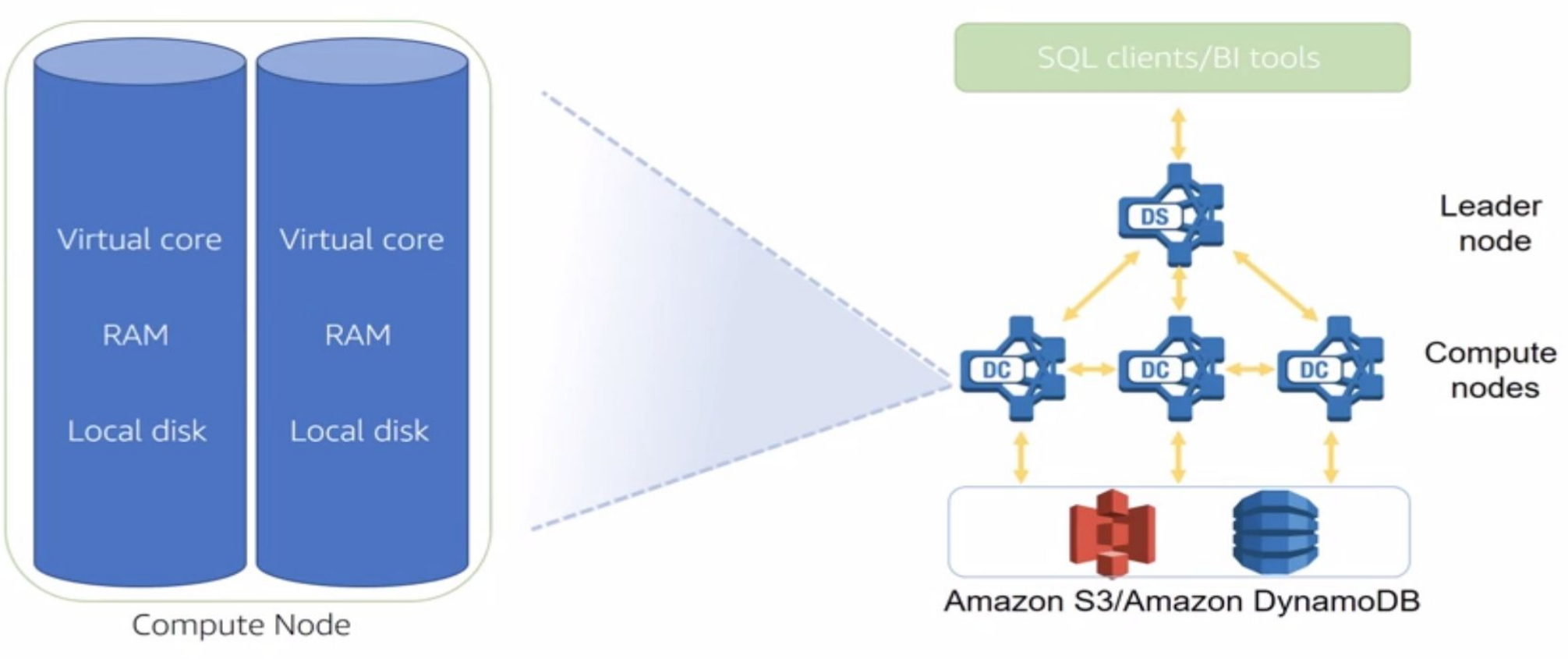
- Amazon Athena
- Amazon RedShift
- Amazon Kinesis
- AWS Glue
- Amazon QuickSight
- Amazon Elastic MapReduce (EMR)
- AWS Lake Formation
- Amazon ElasticSearch Service
- Amazon Managed Streaming for Apache Kafka
- AWS Batch

Amazon RedShift -vue globale

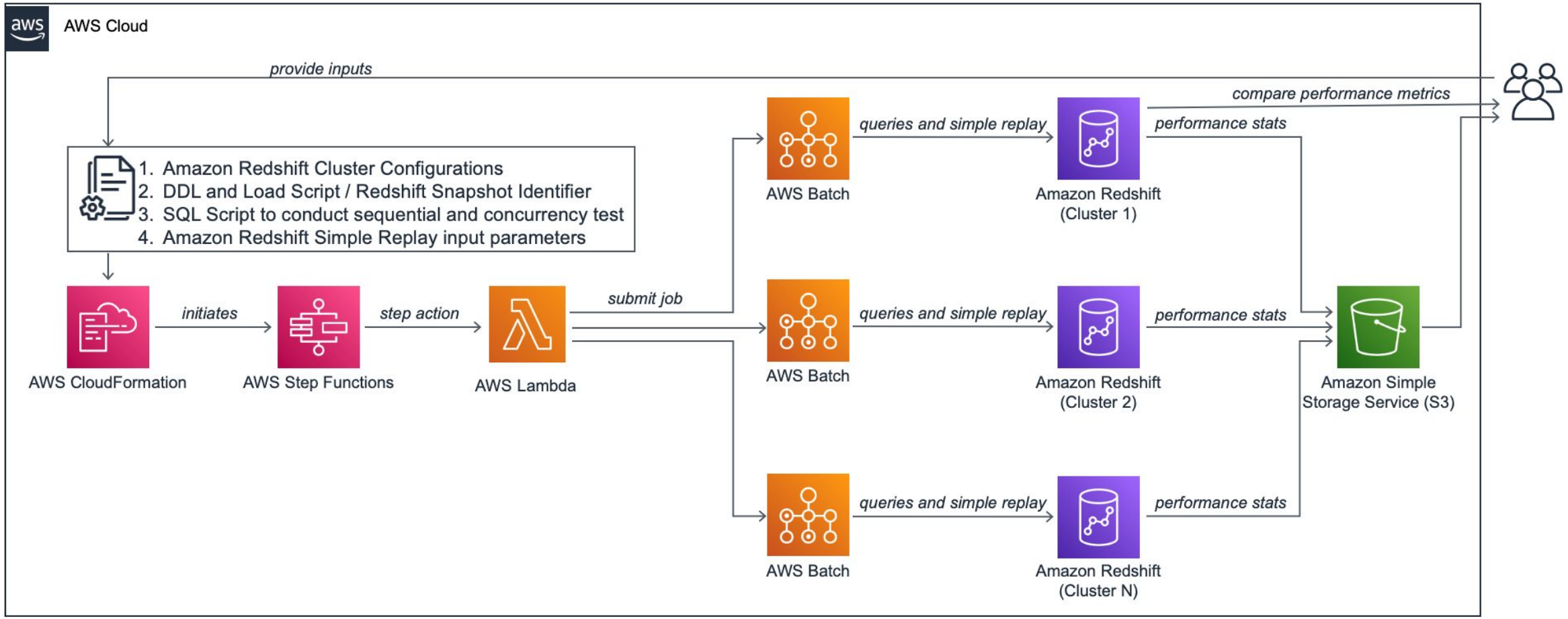
- Redshift est basé sur PostgreSQL, mais il n'est pas utilisé pour l'OLTP
- Il s'agit d'OLAP - traitement analytique en ligne (analyse et entreposage de données).
- Performances 10 fois supérieures à celles des autres entrepôts de données, mise à l'échelle jusqu'à des Po de données
- Stockage des données en colonnes (au lieu de lignes)
- Exécution massivement parallèle des requêtes (MPP)
- Payez au fur et à mesure en fonction des instances provisionnées
- Dispose d'une interface SQL pour l'exécution des requêtes
- Les outils de BI tels que AWS Quicksight ou Tableau s'y intègrent.



Amazon RedShift -architecture de traitement parallèle



Amazon RedShift -exemple d'architecture de solution



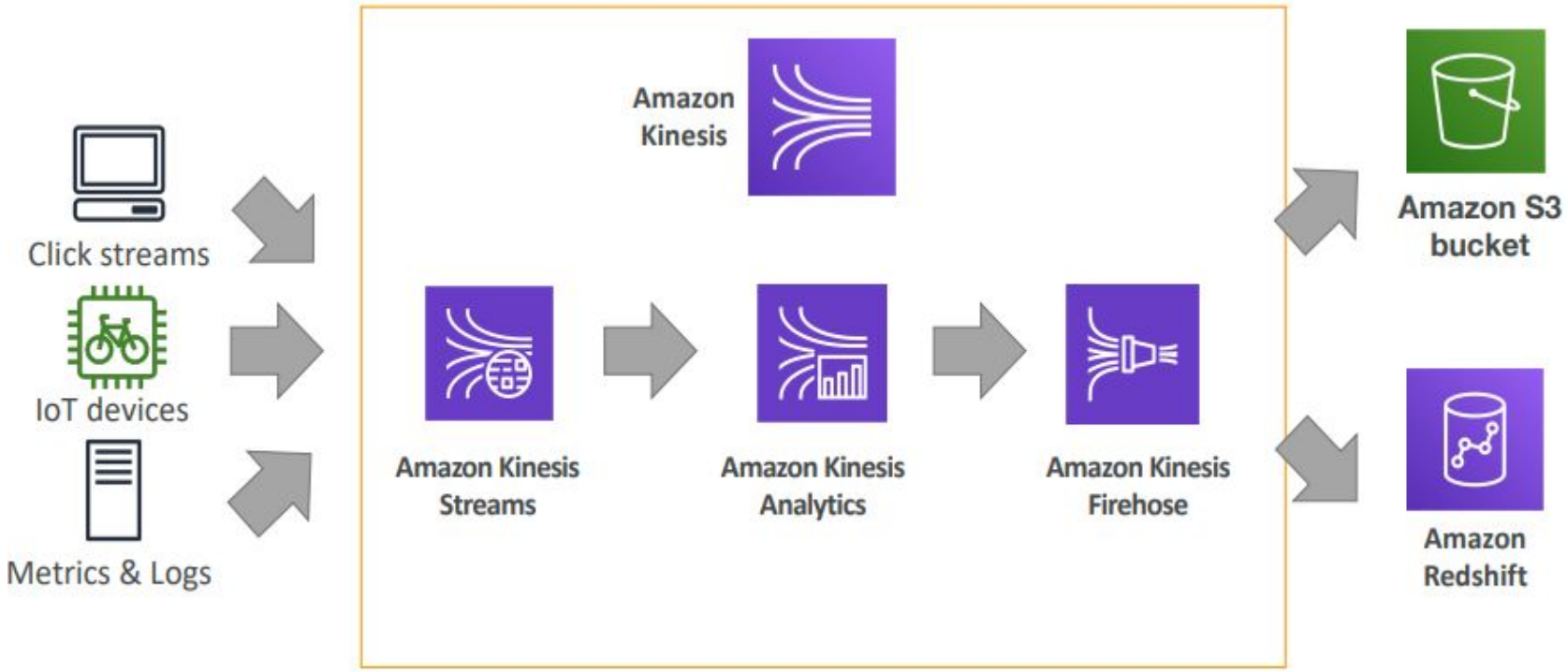
Plan

- Amazon Athena
- Amazon RedShift
- Amazon Kinesis
- AWS Glue
- Amazon QuickSight
- Amazon Elastic MapReduce (EMR)
- AWS Lake Formation
- Amazon ElasticSearch Service
- Amazon Managed Streaming for Apache Kafka
- AWS Batch

Amazon Kinesis -vue globale

- Kinesis est un service géré de "flux de données".
- Idéal pour les journaux d'application, les mesures, l'IdO, les flux de clics
- Idéal pour les big data "en temps réel
- Idéal pour les frameworks de traitement en streaming (Spark, NiFi, etc...)
- Les données sont automatiquement répliquées de manière synchrone vers 3 AZ
- Kinesis Streams : ingestion de flux à faible latence à l'échelle
- Kinesis Analytics : analyse en temps réel des flux à l'aide de SQL
- Kinesis Firehose : chargement des flux dans S3, Redshift, ElasticSearch & Splunk

Amazon Kinesis



Amazon Kinesis - Data Streams

Définition

Amazon Kinesis Data Streams (KDS) est un service évolutif de diffusion de données en temps réel.
Ce service Il capture des gigaoctets de données à partir de sources telles que les flux de clics sur les sites web, les flux d'événements (base de données et suivi de l'emplacement) et les des médias sociaux.

Amazon Kinesis est un service utilisé pour collecter, traiter, et d'analyser des données en temps réel. Il peut être une alternative à Apache Kafka.

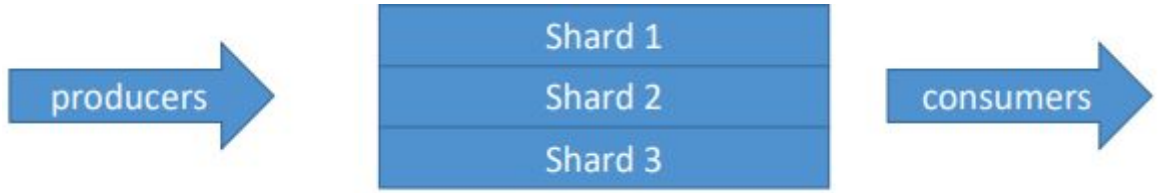
La famille Kinesis se compose de Kinesis Data Streams, Kinesis Data Analytics, Kinesis Data Firehose et Kinesis Video.

Les données en temps réel peuvent être récupérées à partir de producteurs qui sont Kinesis Streams API, Kinesis Producer Library (KPL), et l'agent Kinesis.

Il permet de créer des applications personnalisées connues sous le nom d'applications kinesis data streams(Consommateurs), qui lit les données d'un flux de les données d'un flux de données sous forme d'enregistrements.

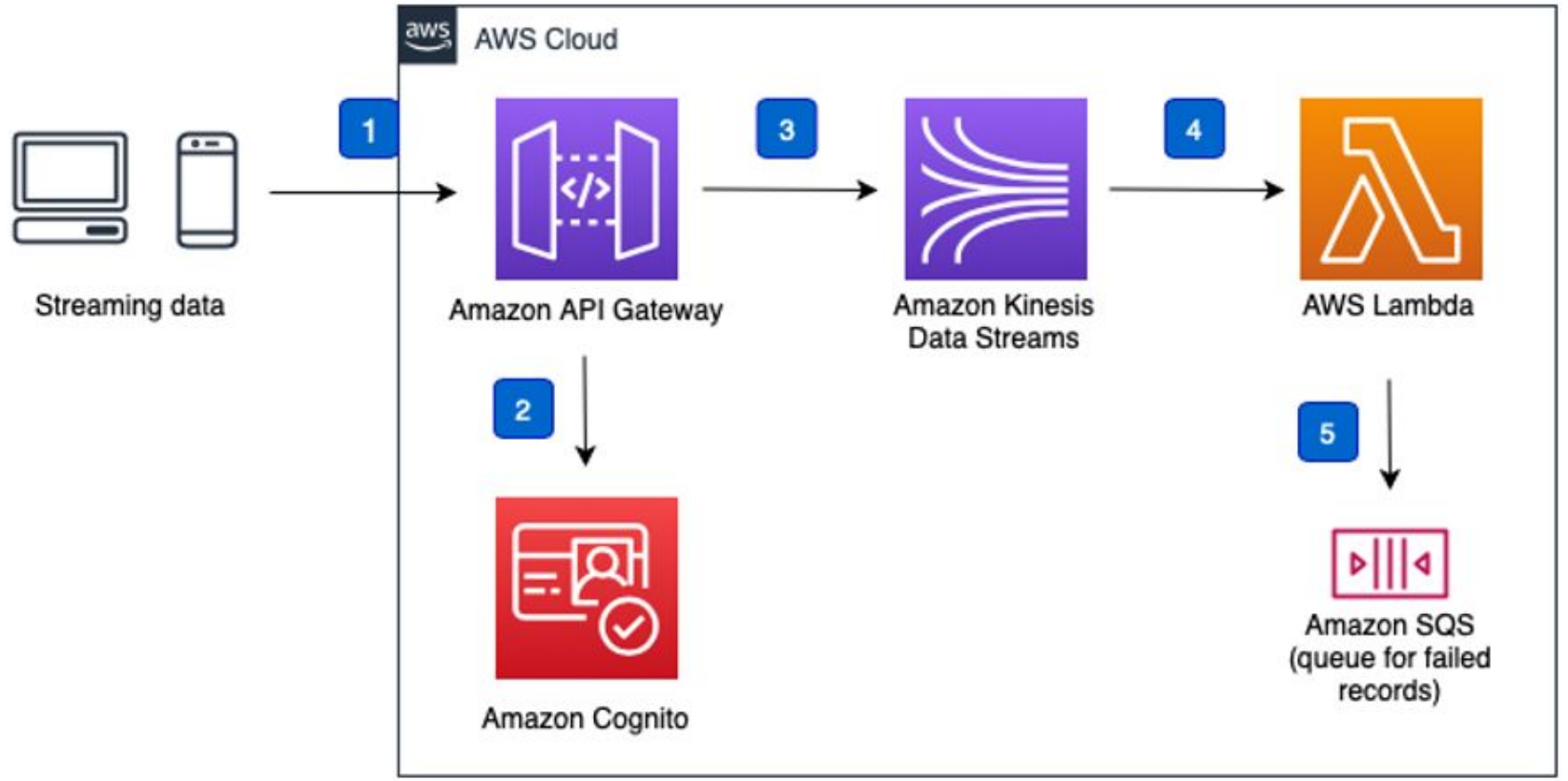
Amazon Kinesis - Data Streams

- Les flux de données sont divisés en groupes ordonnés (Shards / Partitions).



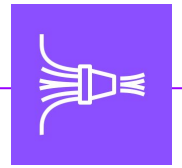
- La durée de conservation des données est de 24 heures par défaut, mais peut aller jusqu'à 365 jours.
- Possibilité de retraiter / rejouer les données
- Plusieurs applications peuvent consommer le même flux
- Traitement en temps réel avec une échelle de débit
- Une fois que les données sont insérées dans Kinesis, elles ne peuvent pas être supprimées (immuabilité).
- **Deux modes pour la capacité :**
  - **A la demande :** pas de planification de la capacité, Kinesis dimensionne les shards automatiquement.
  - **Provisionné :** vous gérez les unités au fil du temps.
- Batching disponible ou par appel de message.
- Le nombre de shards peut évoluer dans le temps (reshard / merge).
- **Les enregistrements sont ordonnés par shard**
-

Amazon Kinesis - exemple d’architecture de solution



Amazon Kinesis - Data Firehose

Définition



Amazon Kinesis Data Firehose est un service sans serveur utilisé pour capturer, transformer et charger dans les magasins de données et les services d'analyse.

Kinesis Data Streams, les événements CloudWatch peuvent être considérés comme source(s) de Kinesis Data Firehose.

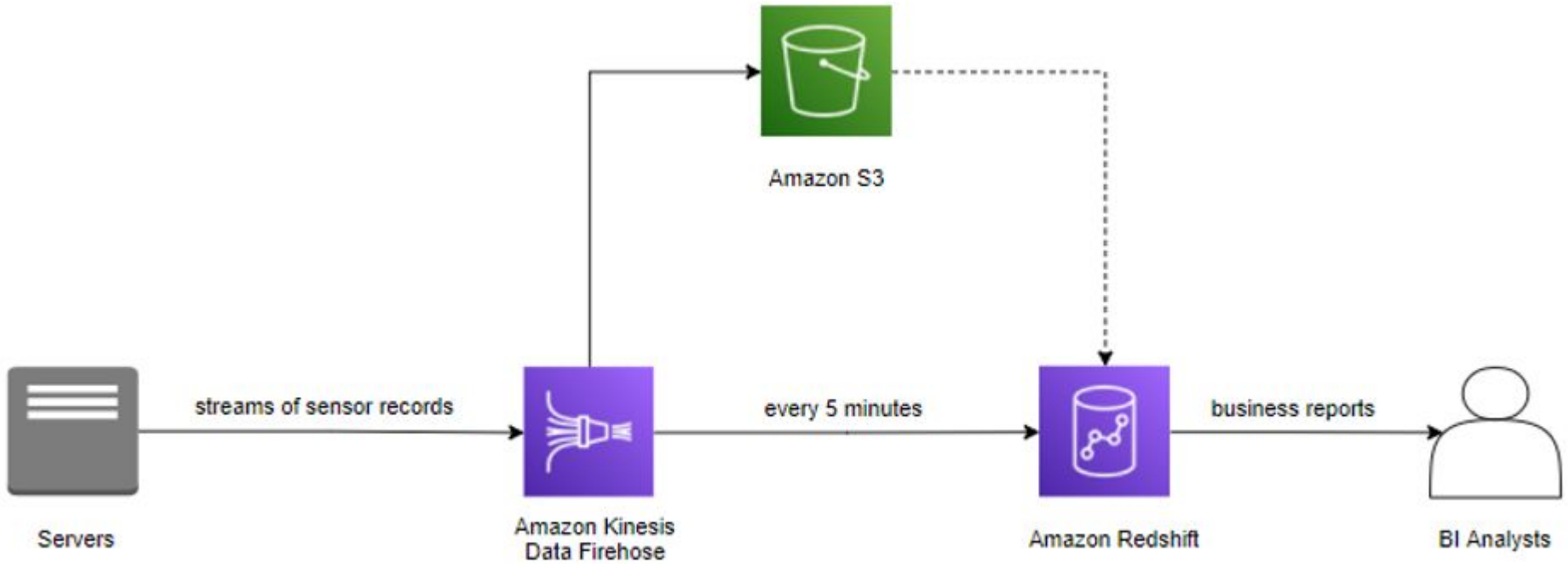
Il réplique les données de manière synchrone sur trois AZ tout en les livrant aux destinataires.

Il permet une analyse en temps réel avec les outils de veille stratégique existants et aide à transformer, compresser et crypter les données en lots avant de les livrer.

Il crée un flux de livraison Kinesis Data Firehose pour envoyer les données. Chaque flux de livraison conserve les enregistrements de données pendant un jour.

Il a un temps de latence minimum de 60 secondes ou un minimum de 32 MB de données à la fois.

Amazon Kinesis - Data Firehose diagramme



Amazon Kinesis - Data Firehose buffer sizing

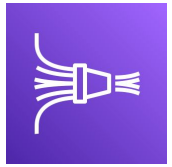
- La mémoire tampon est vidée en fonction de règles de temps et de taille
- Taille de la mémoire tampon (ex : 32MB) : si cette taille est atteinte, la mémoire tampon est vidée.
- Durée de la mémoire tampon (ex : 1 minute) : si cette durée est atteinte, la mémoire tampon est vidée.
- Firehose peut automatiquement augmenter la taille du tampon pour augmenter le débit
- Débit élevé => la taille de la mémoire tampon sera atteinte
- Faible débit => la durée de la mémoire tampon sera atteinte

Si une vidange en temps réel des flux de données Kinesis vers S3 est nécessaire, utilisez Lambda.

Comparaison entre Kinesis Data Streams et Firehose

Kinesis Data Streams

- Service de streaming pour l'ingestion à l'échelle
- Écrire du code personnalisé (producteur / consommateur)
- Temps réel (~200 ms)
- Gérer la mise à l'échelle (division / fusion)
- Stockage des données pendant 1 à 365 jours
- Prise en charge de la capacité de relecture



Kinesis Data Firehose

- Chargement de données en continu dans S3 / Redshift / ES / 3rd party / HTTP personnalisé
- Entièrement géré
- Quasi temps réel (temps de tampon min. 60 sec)
- Mise à l'échelle automatique
- Pas de stockage de données
- Ne prend pas en charge la capacité de relecture

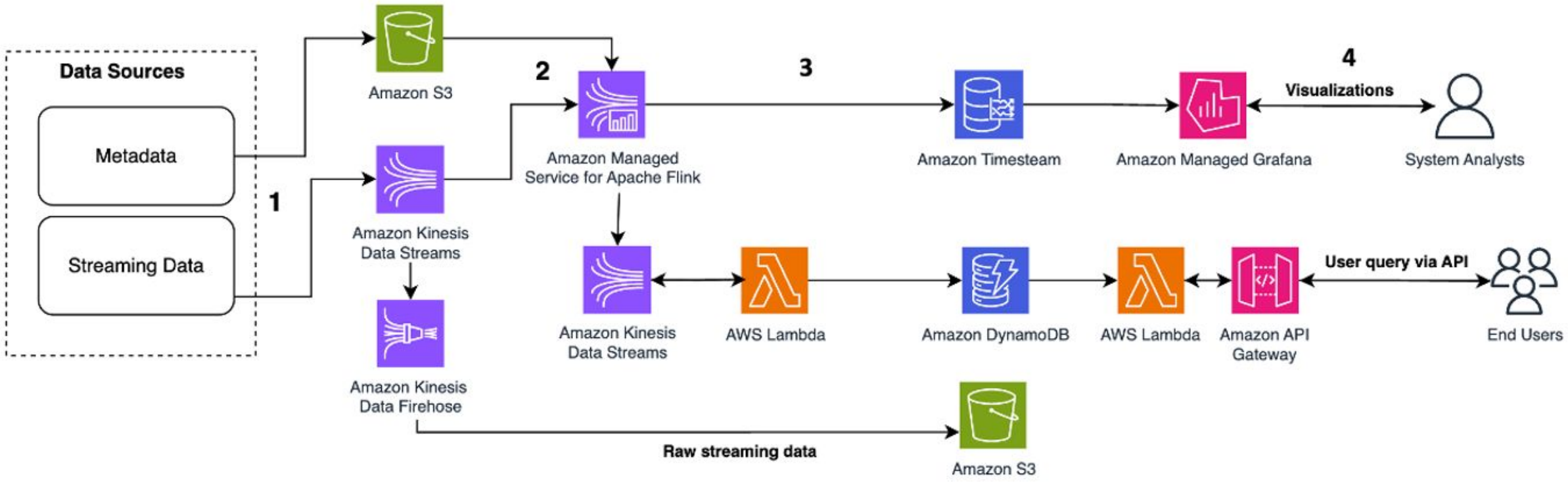
Amazon Kinesis - Data Analytics

- Cas d'utilisation:

  - ETL en continu : sélectionner des colonnes, effectuer des transformations simples, sur des données en continu
  - Génération continue de mesures : tableau de classement en direct pour un jeu mobile
  - Analyse réactive : recherche de certains critères et création d'alertes (filtrage)

- Fonctionnalités
  - Payez uniquement pour les ressources consommées (mais ce n'est pas bon marché)
  - Sans serveur ; mise à l'échelle automatique
  - Utiliser les permissions IAM pour accéder à la source et à la (aux) destination(s) du streaming
  - SQL ou Flink pour écrire le calcul
  - Découverte de schémas
  - Lambda peut être utilisé pour le prétraitement

Amazon Kinesis Data Analytics -exemple architecture de solution



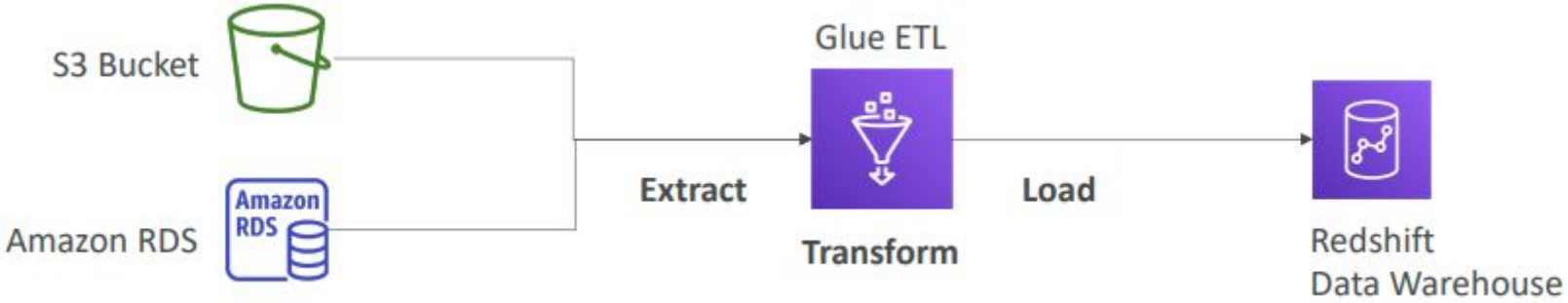
Full Data Engineering Real-time Layer

Plan

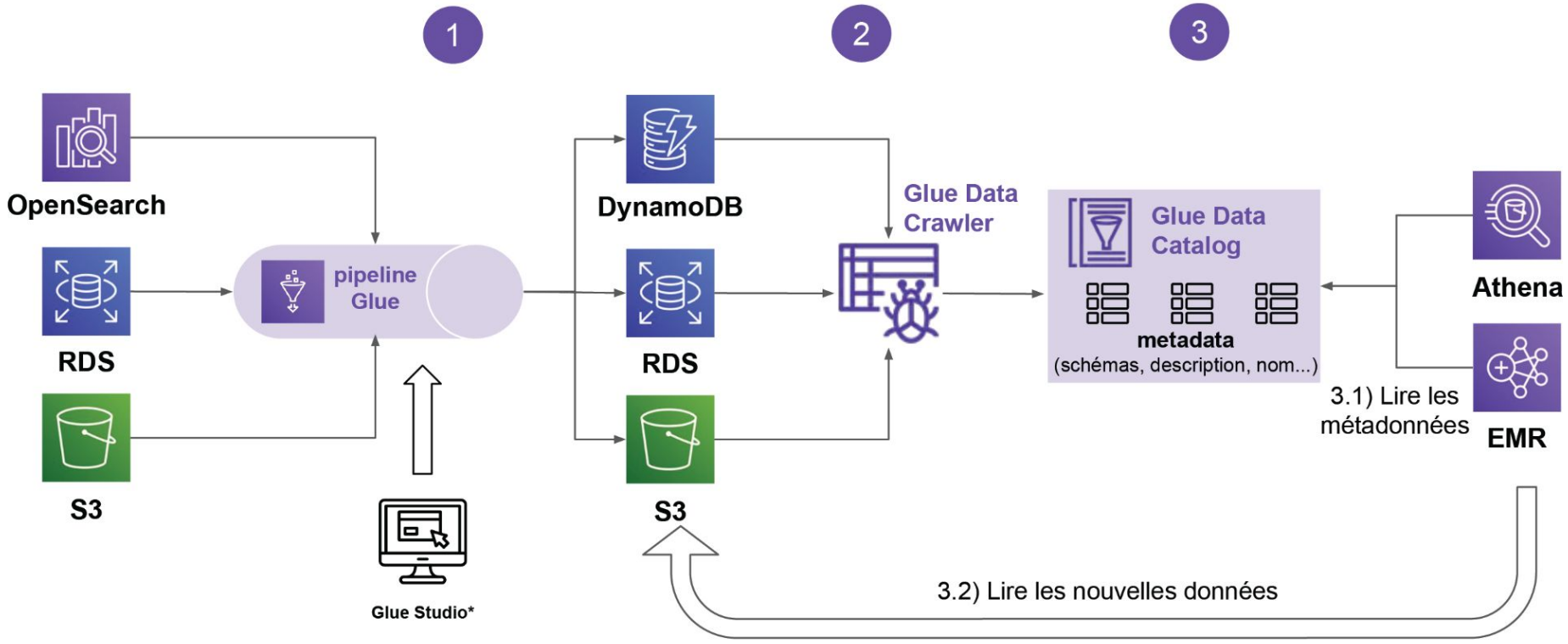
- Amazon Athena
- Amazon RedShift
- Amazon Kinesis
- AWS Glue
- Amazon QuickSight
- Amazon Elastic MapReduce (EMR)
- AWS Lake Formation
- Amazon ElasticSearch Service
- Amazon Managed Streaming for Apache Kafka
- AWS Batch

AWS GLUE

- Service géré d'extraction, de transformation et de chargement (ETL)
- Utile pour préparer et transformer les données pour l'analyse
- Service entièrement sans serveur



AWS GLUE -exemple d'architecture de solution



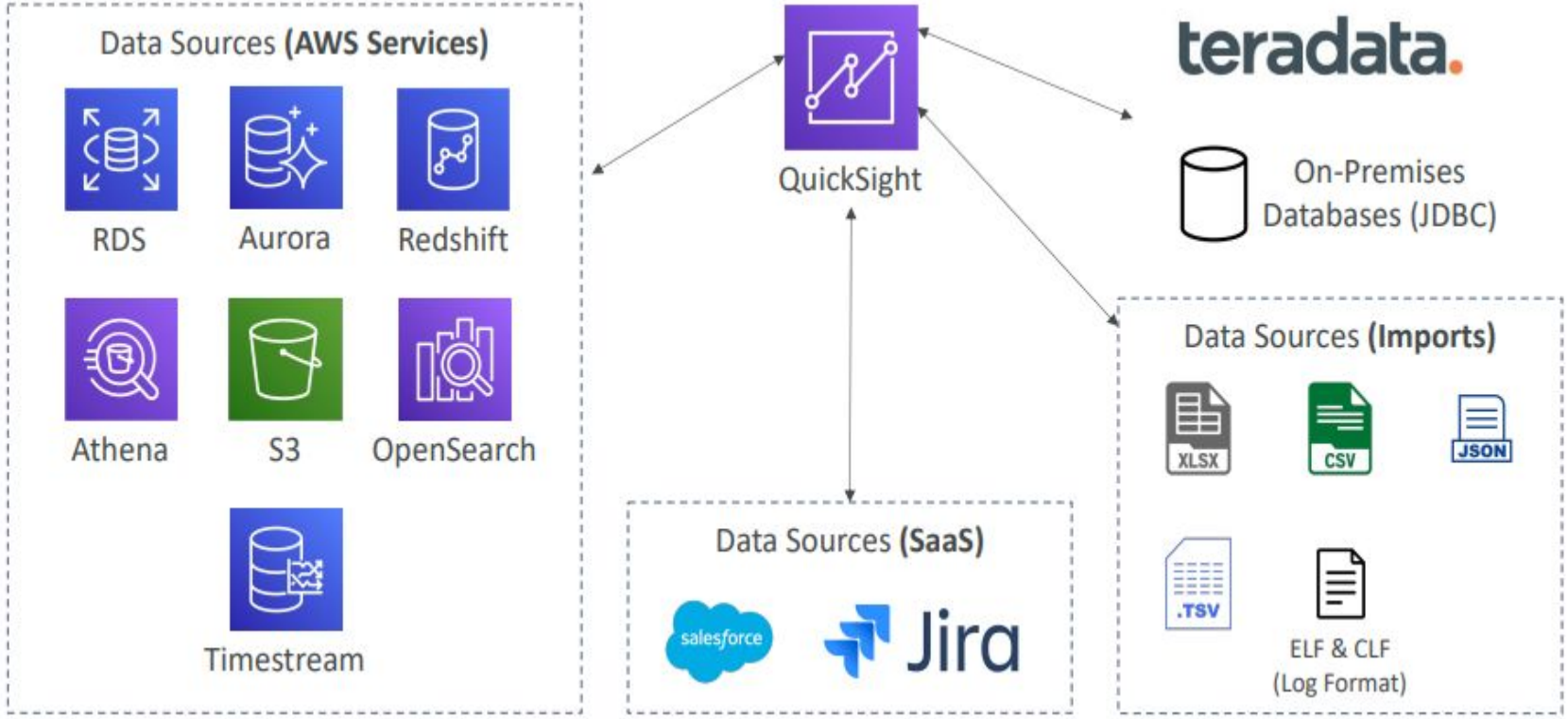
Plan

- Amazon Athena
- Amazon RedShift
- Amazon Kinesis
- AWS Glue
- Amazon QuickSight
- Amazon Elastic MapReduce (EMR)
- AWS Lake Formation
- Amazon ElasticSearch Service
- Amazon Managed Streaming for Apache Kafka
- AWS Batch

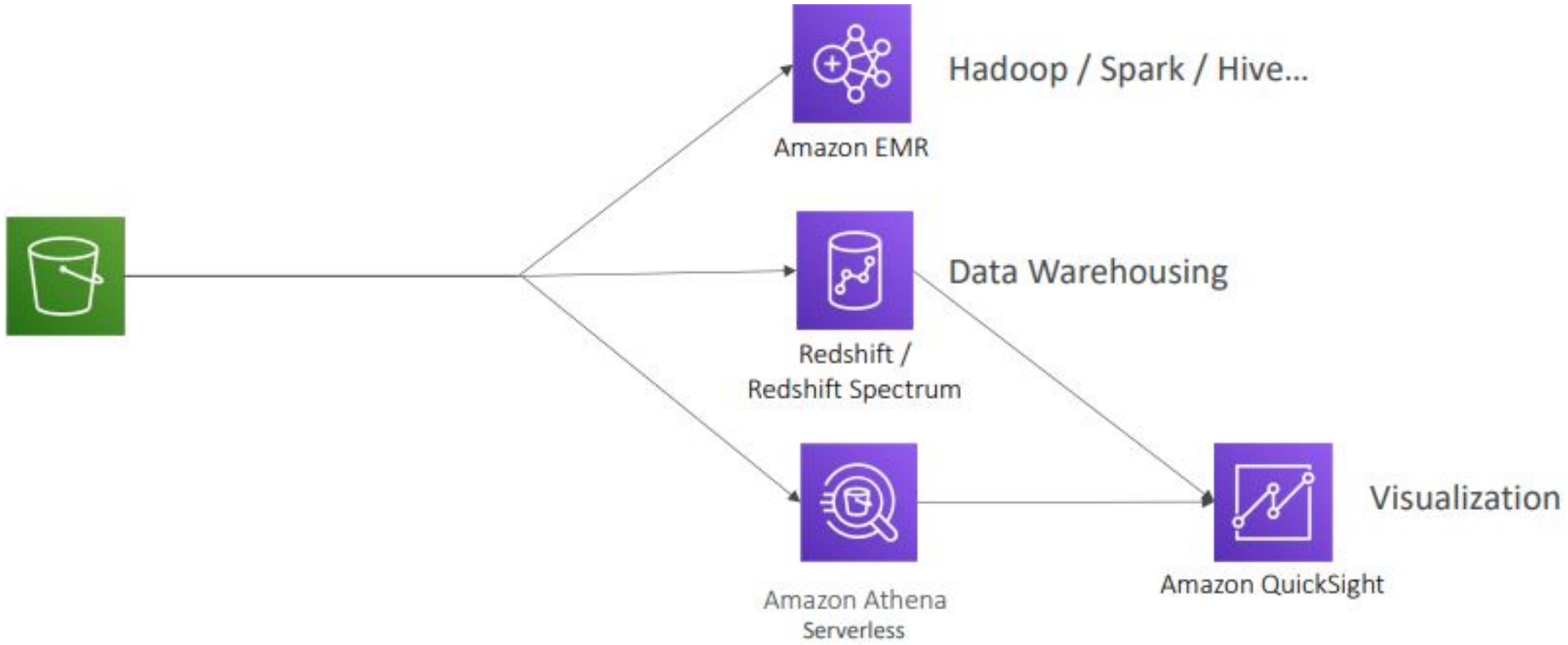
Amazon Quick Sight

- Service de veille stratégique sans serveur basé sur l'apprentissage automatique pour créer des tableaux de bord interactifs
- Rapide, automatiquement évolutif, intégrable, avec une tarification à la session.
- Cas d'utilisation :
  - Analyse d'entreprise
  - Créer des visualisations
  - Effectuer des analyses ad hoc
  - Obtenir des informations commerciales à partir des données
- Intégration avec RDS, Aurora,Athena, Redshift, S3...
- Calcul en mémoire à l'aide du moteur SPICE si les données sont importées dans QuickSight
- Édition entreprise : Possibilité de configurer Sécurité au niveau des colonnes (CLS)

AWS Quick Sight -Intégration aux services



AWS Quick Sight -exemple d’architecture de solution full data engineering pipeline analysis layer



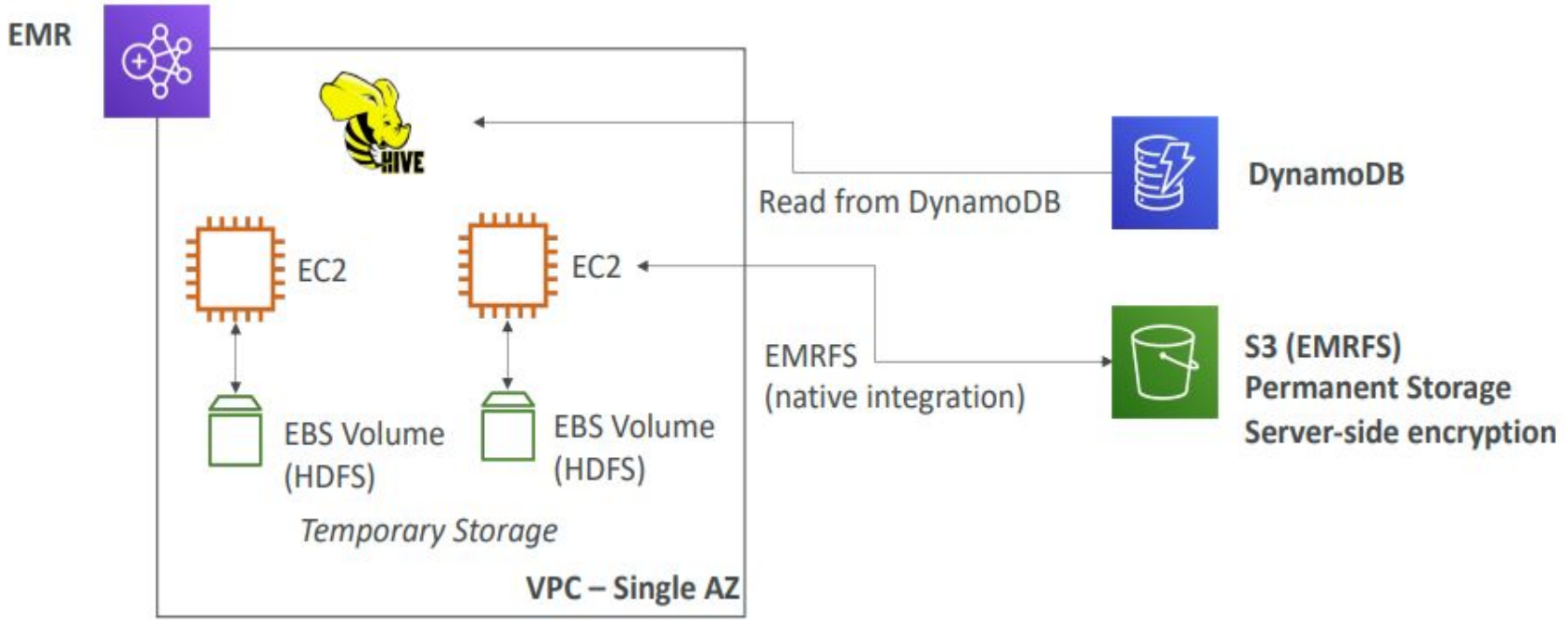
Plan

- Amazon Athena
- Amazon RedShift
- Amazon Kinesis
- AWS Glue
- Amazon QuickSight
- Amazon Elastic MapReduce (EMR)
- AWS Lake Formation
- Amazon ElasticSearch Service
- Amazon Managed Streaming for Apache Kafka
- AWS Batch

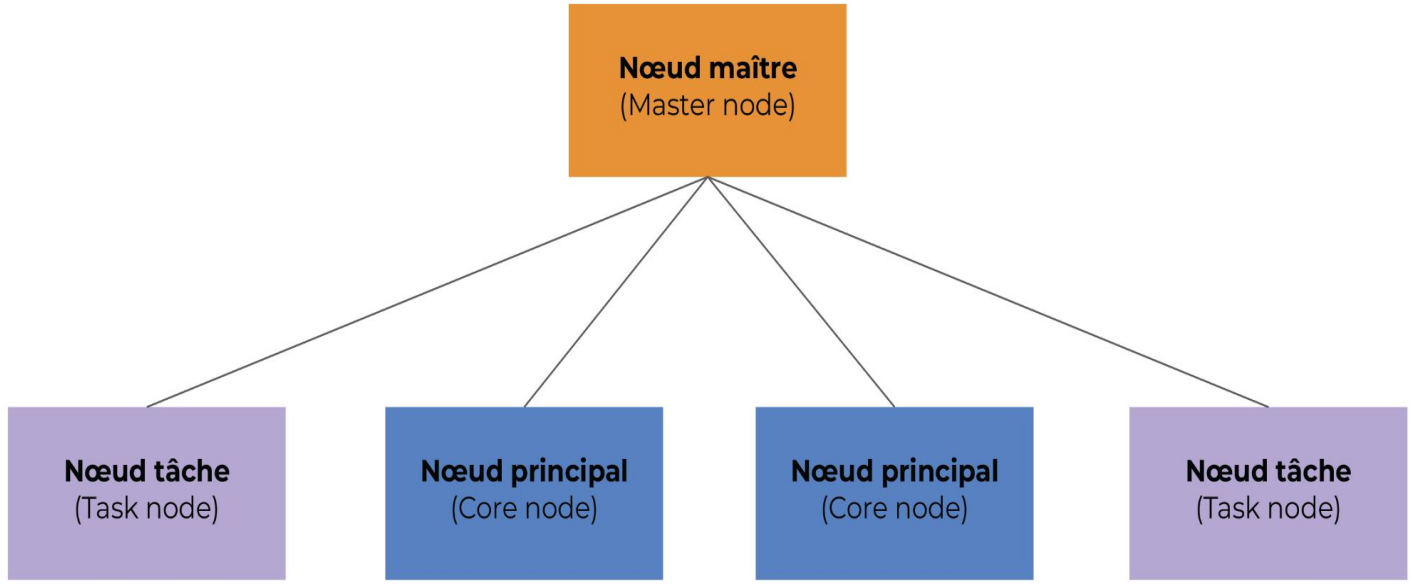
AWS Elastic MapReduce (EMR)

- EMR est l'acronyme de "Elastic MapReduce«
- EMR permet de créer des clusters Hadoop (Big Data) pour analyser et traiter de grandes quantités de données.
- Les clusters peuvent être constitués de centaines d'instances EC2.
- Supporte également Apache Spark, HBase, Presto, Flink...
- EMR prend en charge l'ensemble du provisionnement et de la configuration d’EC2
- Mise à l'échelle automatique avec CloudWatch
- Cas d'utilisation : traitement de données, apprentissage automatique, indexation web, big data...

AWS Elastic MapReduce (EMR) -intégration aux services



AWS Elastic MapReduce (EMR) -architecture cluster



Noeud maître:

- Orchestre le cluster
- Tarification à utiliser : instances réservées

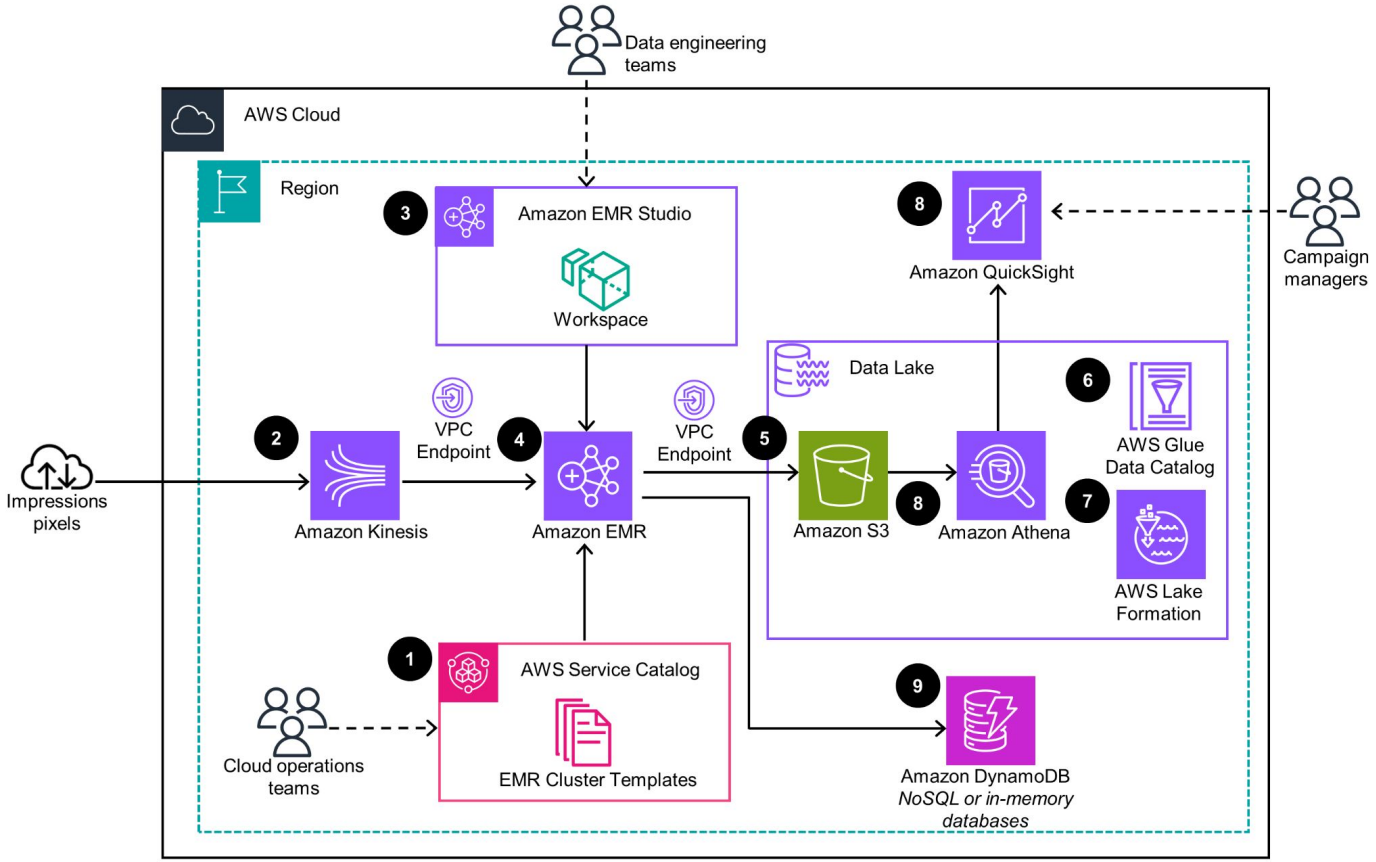
Noeud tâche:

- Exécute uniquement des tâches
- Tarification à utiliser : instances Spot

Noeud principal:

- Exécute des tâches et stocke les données
- Tarification à utiliser : instances réservées

AWS Elastic MapReduce (EMR) -exemple d'architecture de solution



Plan

- Amazon Athena
- Amazon RedShift
- Amazon Kinesis
- AWS Glue
- Amazon QuickSight
- Amazon Elastic MapReduce (EMR)
- AWS Lake Formation
- Amazon ElasticSearch Service
- Amazon Managed Streaming for Apache Kafka
- AWS Batch

AWS Lake Formation

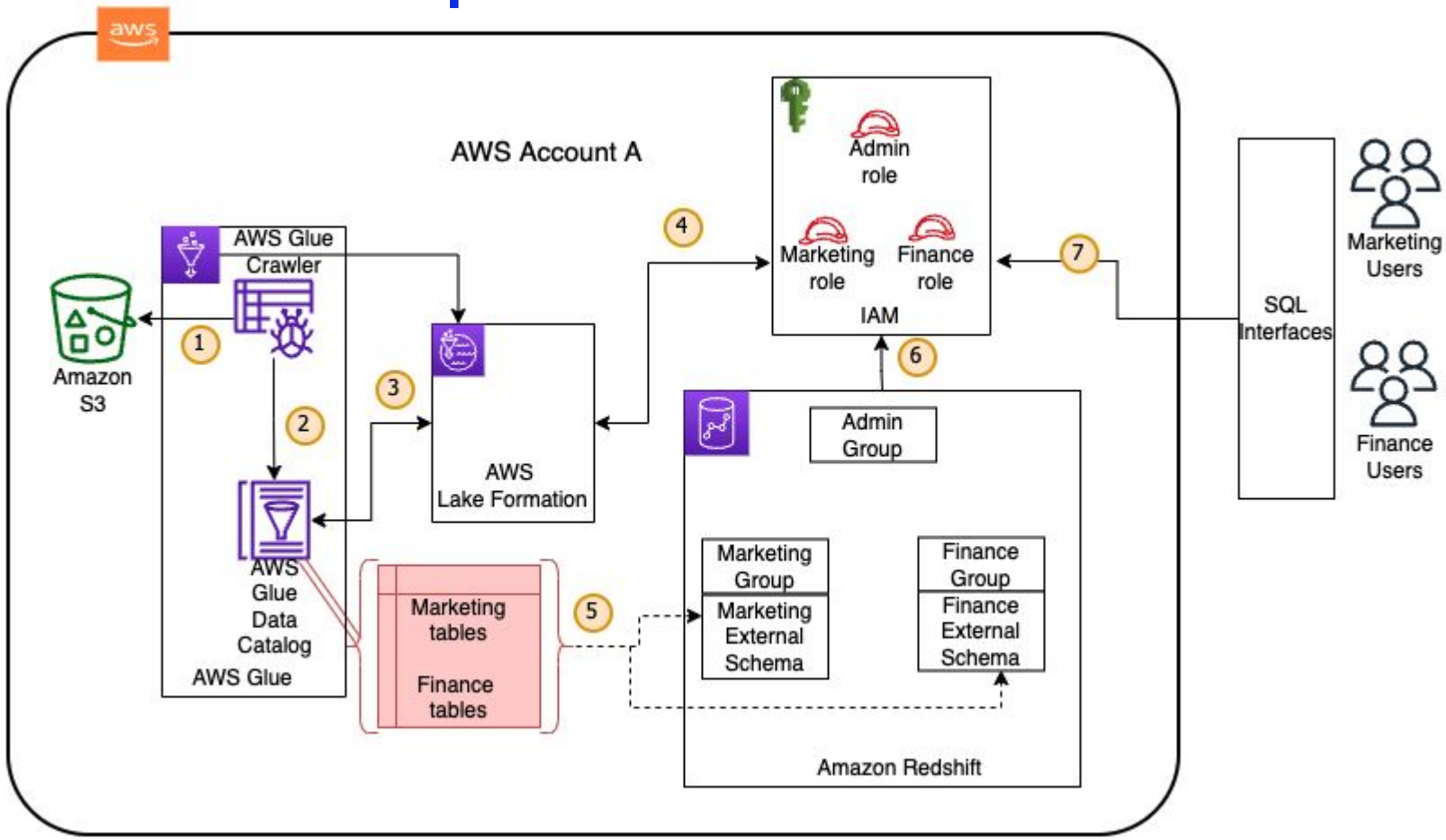
Définition

AWS Lake Formation est un service qui facilite la création, la gestion et la sécurisation d'un data lake sur AWS. Un data lake est un référentiel centralisé qui permet de stocker des données structurées et non structurées à grande échelle. Lake Formation simplifie le processus d'importation de données, de catalogage, de gestion des accès et de partage des données.

Les fonctionnalités:

- Création Rapide de Data Lakes
- Gestion de sécurité
- Catalogue des Données
- Partage de données
- Intégration a d’autres services

AWS Lake Formation -exemple d'architecture de solution

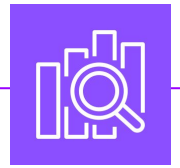


Plan

- Amazon Athena
- Amazon RedShift
- Amazon Kinesis
- AWS Glue
- Amazon QuickSight
- Amazon Elastic MapReduce (EMR)
- AWS Lake Formation
- Amazon ElasticSearch Service
- Amazon Managed Streaming for Apache Kafka
- AWS Batch

AWS ElasticSearch Service

Définition



Amazon Elasticsearch Service (Amazon ES) est un service géré qui permet aux utilisateurs de de déployer, de gérer et de mettre à l'échelle des clusters dans le nuage AWS. Amazon ES fournit un accès direct aux API Elasticsearch

Elasticsearch est un moteur de recherche libre et gratuit pour tous les types de données, qu'elles soient textuelles, numériques, géospatiales, structurées ou non structurées.

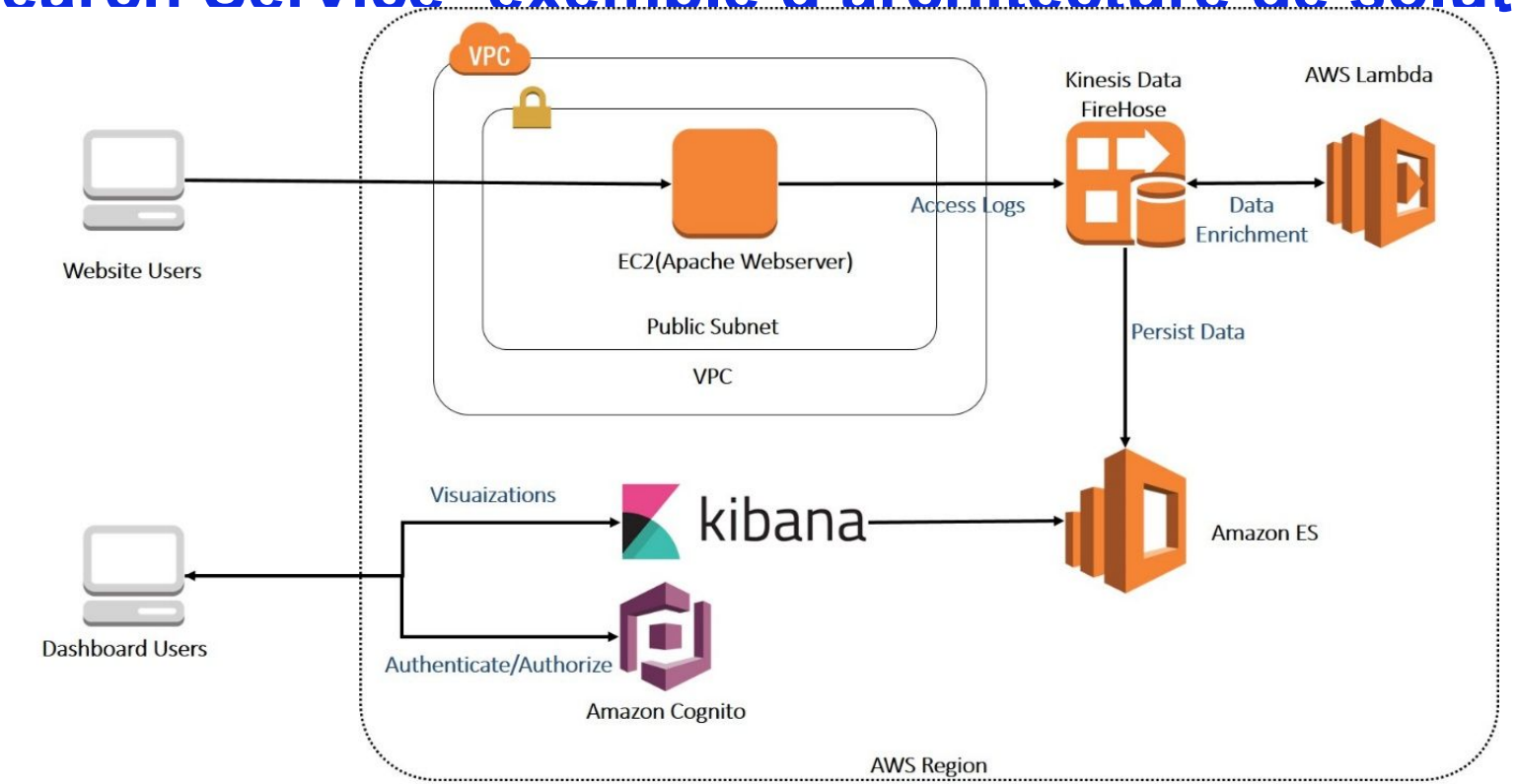
Coûts réduits

Amazon ES avec Kibana (visualisation) et Logstash (ingestion de logs) offre une expérience de recherche améliorée pour les applications et les sites web afin de trouver rapidement les données pertinentes.

Amazon ES lance les ressources du cluster Elasticsearch, détecte les nœuds Elasticsearch défaillants et les remplace.

Le cluster Elasticsearch peut être mis à l'échelle en quelques clics dans la console..

AWS ElasticSearch Service -exemple d'architecture de solution



Architecture Overview

Coûts réduits

Plan

- Amazon Athena
- Amazon RedShift
- Amazon Kinesis
- AWS Glue
- Amazon QuickSight
- Amazon Elastic MapReduce (EMR)
- AWS Lake Formation
- Amazon ElasticSearch Service
- Amazon Managed Streaming for Apache Kafka
- AWS Batch

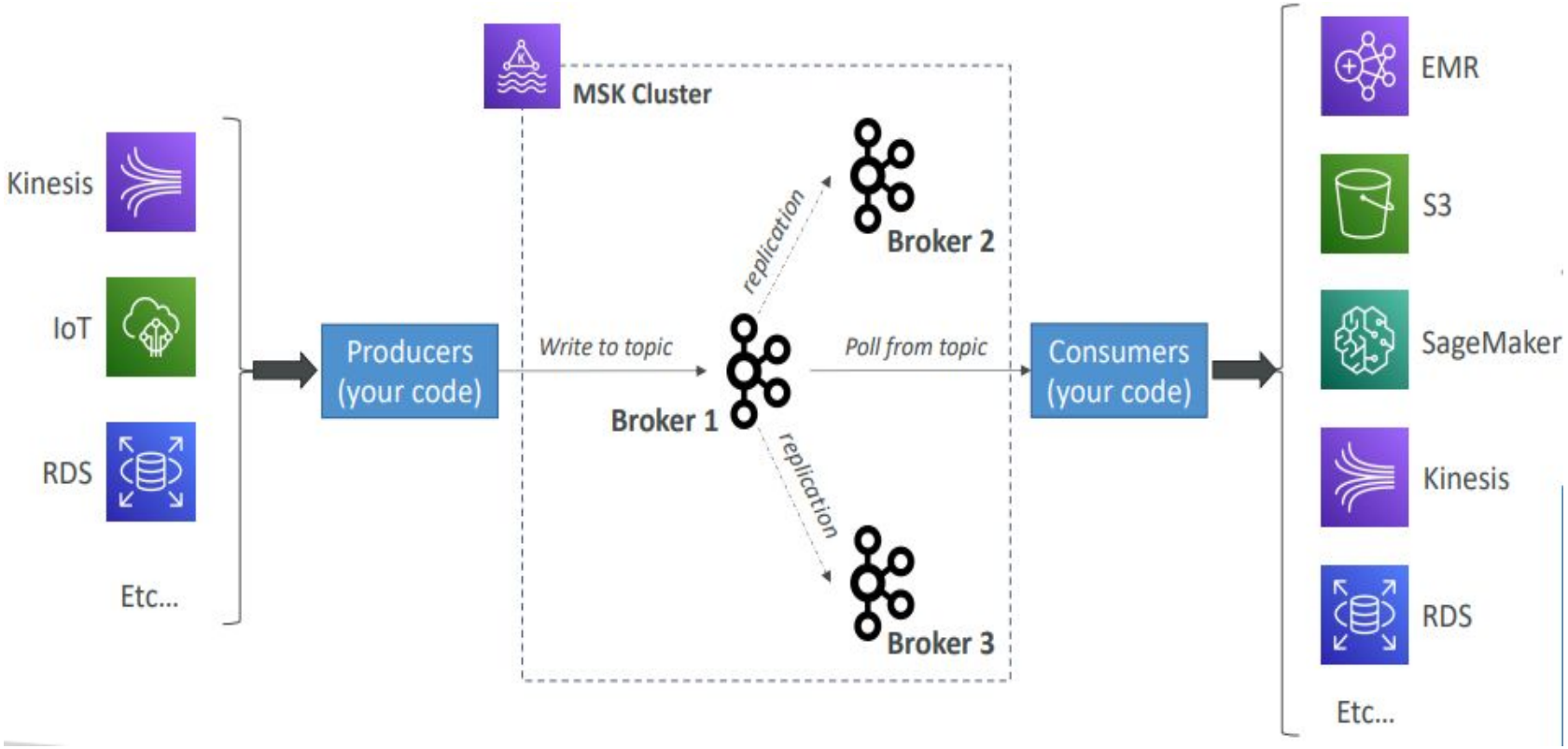
Amazon Managed Streaming for Apache Kafka (MSK)

- Alternative à Amazon Kinesis

- Apache Kafka entièrement géré sur AWS
  - Permet de créer, mettre à jour et supprimer des clusters
  - MSK crée et gère les nœuds de brokers Kafka et les nœuds de Zookeeper pour vous.
  - Déploiement du cluster MSK dans votre VPC, multi-AZ (jusqu'à 3 pour HA)
  - Récupération automatique des défaillances courantes d'Apache Kafka
  - Les données sont stockées sur des volumes EBS aussi longtemps que vous le souhaitez

- MSK sans serveur
  - Exécutez Apache Kafka sur MSK sans gérer la capacité.
  - MSK provisionne automatiquement les ressources et fait évoluer le calcul et le stockage.

Amazon Managed Streaming for Apache Kafka (MSK)



AWS MSK vs Kinesis Data Streams

Kinesis Data Streams

- Limitation de la taille des messages à 1 Mo
- Flux de données avec Shards
- Fractionnement et fusion de nuages de points (Shard Splitting & Merging)
- Cryptage TLS en vol
- Cryptage KMS au repos

Amazon MSK

- 1MB par défaut, configurer pour plus (ex : 10MB)
- Sujets Kafka avec partitions
- Peut seulement ajouter des partitions à un sujet
- Cryptage en vol PLAINTEXT ou TLS
- Cryptage KMS au repos

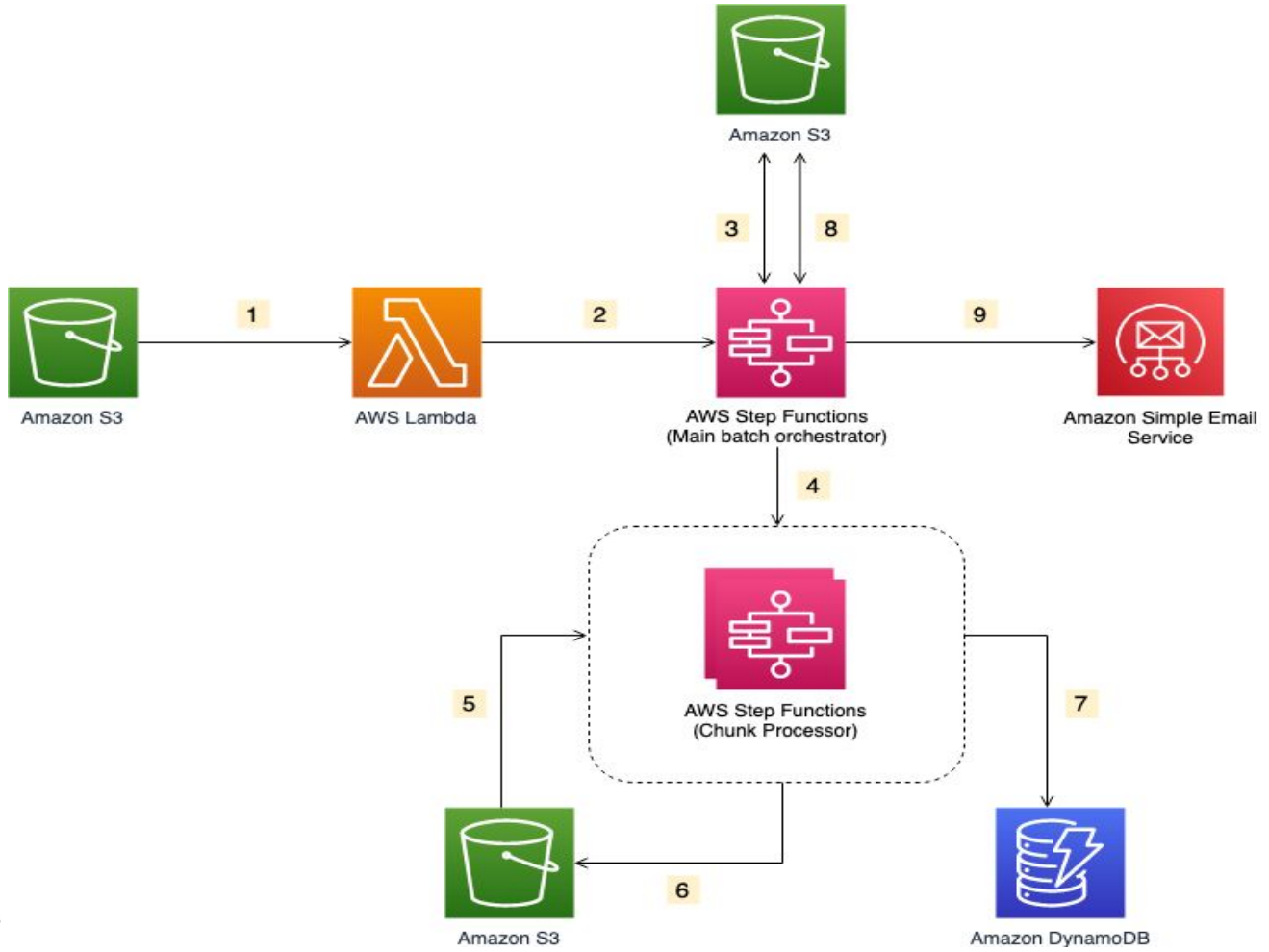
Plan

- Amazon Athena
- Amazon RedShift
- Amazon Kinesis
- AWS Glue
- Amazon QuickSight
- Amazon Elastic MapReduce (EMR)
- AWS Lake Formation
- Amazon ElasticSearch Service
- Amazon Managed Streaming for Apache Kafka
- AWS Batch

AWS Batch

- Exécuter des travaux par lots en tant qu'images Docker
- Deux options :
  1. Exécuter sur AWS Fargate (offre entièrement sans serveur).
  2. Provisionnement dynamique des instances (EC2 & Spot Instances) - dans VPC
- Quantité et type optimaux en fonction du volume et des besoins
- Pas besoin de gérer des clusters, entièrement sans serveur
- Vous ne payez que pour les ressources sous-jacentes utilisées
- **Exemple :** traitement par lots d'images, exécution de milliers de tâches simultanées
- Planification des travaux par lots à l'aide d'Amazon EventBridge
- Orchestrer les travaux par lots à l'aide d'AWS Step Functions

AWS Batch -exemple d’architecture de solution



MERCI POUR VOTRE AIMABLE ATTENTION!

EAZYTraining

Alphonsine Lahda

Lahda Biassou Alphonsine

Ingénieure cloud et Formatrice