Architecture Streaming Data

Architecture de stockage et de sortie des données en streaming

Amina MARIE

Ordre du jour

Options de stockage pour les données en streaming

Mécanismes de sortie des données pour alimenter les systèmes en aval

Exercice 1 : Stockage des données en streaming dans un data lake à l'aide d'Apache Hadoop et Apache Parquet.

Exercice 2 : Visualisation des données en streaming à l'aide d'un outil de business intelligence tel que Tableau.

Options de stockage pour les données en streaming

Options de stockage pour les données en streaming

Il n’y a pas de bonnes ou mauvaises options de stockage. Tout dépend des besoins du projet.

- Systèmes de fichiers distribués (DFS)
- Bases de données NoSQL
- Data Lakes
- Bases de données In-Memory
- Services de streaming gérés

Options de stockage pour les données en streaming

- Systèmes de fichiers distribués (DFS)

Infrastructure informatique qui permet à plusieurs ordinateurs de travailler ensemble pour stocker et gérer des fichiers de manière distribuée sur un réseau.

Nous avons des DFS comme que Hadoop Distributed File System (HDFS) ou Amazon S3, qui sont conçus pour stocker de grandes quantités de données à l'échelle.

Ils offrent une redondance intégrée et une tolérance aux pannes, ce qui les rend robustes pour le stockage de données en streaming.

Options de stockage pour les données en streaming

- Bases de données NoSQL

Les bases de données NoSQL, telles que MongoDB, Cassandra ou Apache CouchDB, sont bien adaptées au stockage de données en streaming en raison de leur capacité à gérer des flux de données volumineux avec une latence minimale.

Elles permettent également une évolutivité horizontale, ce qui signifie qu'elles peuvent s'adapter à une augmentation de la charge de travail.

Options de stockage pour les données en streaming

- Data Lakes

Les Data Lakes, tels que Apache Hadoop ou Amazon S3, sont des dépôts de données qui stockent des données brutes dans leur format natif jusqu'à ce qu'elles soient nécessaires pour l'analyse.

Ils sont populaires pour le stockage de données en streaming car ils peuvent gérer une variété de données non structurées, semi-structurées et structurées.

Options de stockage pour les données en streaming

- Bases de données In-Memory

Ces bases de données, telles que Redis ou Apache Ignite, stockent les données en mémoire vive plutôt que sur un disque, ce qui permet des temps de réponse très rapides.

Elles sont idéales pour les cas d'utilisation nécessitant un accès ultra-rapide aux données en streaming.

Options de stockage pour les données en streaming

- Services de streaming gérés

Des services comme Amazon Kinesis ou Apache Kafka proposent des solutions intégrées pour la collecte, le traitement et le stockage des données en streaming.

Ils simplifient souvent la configuration et la gestion, mais peuvent être moins flexibles que les solutions auto-hébergées.

Mécanismes de sortie des données pour alimenter les systèmes en aval

Mécanismes de sortie des données pour alimenter les systèmes en aval

- Quels sont les systèmes en aval?
  - Visualisation
  - Tableaux de bord
  - Base de données...

Les mécanismes:

- Kafka Connect
- API REST
- Intégrations natives
- Protocoles de streaming
- Connecteurs personnalisés

Mécanismes de sortie des données pour alimenter les systèmes en aval

- Kafka Connect

Kafka Connect est une plateforme open source pour connecter Kafka à des systèmes externes tels que des bases de données, des entrepôts de données, etc. Il permet d'importer et d'exporter des données en temps réel vers et depuis Kafka.

- API REST

De nombreuses plateformes de streaming offrent des API REST pour permettre aux utilisateurs d'envoyer et de recevoir des données en temps réel. Ces API sont souvent utilisées pour intégrer des applications tierces avec des systèmes de streaming.

Mécanismes de sortie des données pour alimenter les systèmes en aval

- Intégrations natives

Certains systèmes de streaming ont des intégrations natives avec d'autres technologies courantes telles que Hadoop, Spark, Elasticsearch, etc. Ces intégrations facilitent le transfert de données entre différents systèmes sans nécessiter de développements supplémentaires.

- Protocoles de streaming

Des protocoles comme MQTT, AMQP, et d'autres sont souvent utilisés pour le streaming de données en temps réel. Ils permettent de transférer des données de manière efficace et fiable entre les différents composants du système.

Mécanismes de sortie des données pour alimenter les systèmes en aval

- Connecteurs personnalisés

Dans certains cas, des connecteurs personnalisés peuvent être développés pour intégrer des systèmes spécifiques avec des plateformes de streaming. Ces connecteurs sont généralement développés en fonction des besoins spécifiques de l'application.

L'architecture de sortie des données en streaming doit être conçue avec soin pour garantir la fiabilité, la scalabilité et les performances nécessaires pour alimenter efficacement les systèmes en aval.

Exercice 2 : Visualisation des données en streaming

Exercice 2 : Visualisation des données en streaming à l'aide de Tableau.

Tableau est un outil de business intelligence (BI) très populaire.

Il permet aux utilisateurs de visualiser et d'analyser leurs données de manière interactive.

Visualisation des données : tableaux de bord interactifs, des graphiques, des cartes, des tableaux croisés dynamiques..

Connexion à diverses sources de données : il peut se connecter à une variété de sources de données (bases de données relationnelles, NoSQL, des fichiers plats, services cloud, des données en streaming, etc.)

Analyse avancée : Tableau offre des fonctionnalités avancées d'analyse des données, telles que la prévision, le clustering, les calculs personnalisés, l'agrégation de données, etc., permettant aux utilisateurs d'extraire des insights approfondis à partir de leurs données.

Exercice 2 : Visualisation des données en streaming à l'aide de Tableau.

Interactivité : Les utilisateurs peuvent explorer les données de manière interactive en filtrant, en forant, en zoomant et en sélectionnant des parties spécifiques des visualisations. Cela permet une analyse approfondie des données sans avoir besoin de compétences techniques avancées en requêtes SQL ou en programmation.

Tableaux de bord dynamiques : Tableau permet de créer des tableaux de bord dynamiques qui regroupent plusieurs visualisations pour offrir une vue d'ensemble complète des données. Les utilisateurs peuvent interagir avec les visualisations et les filtres pour obtenir des insights en temps réel.

Partage et collaboration : Les tableaux de bord et les visualisations créés dans Tableau peuvent être partagés avec d'autres utilisateurs via des liens web, des fichiers téléchargeables ou des intégrations dans des plates-formes de collaboration telles que Slack ou Microsoft Teams.

Exercice 2 : Visualisation des données en streaming à l'aide de Tableau.

Objectif: Utiliser Tableau et créer un tableau de visualisation de données.

Les étapes:

- Aller sur le site de Tableau https://www.tableau.com/ et créer un compte d'essai.
- Activer votre compte et aller sur votre interface utilisateur.
- Créer un flux de données streaming qui représentant les ventes d'une entreprise fictive (coté producteur). Utilisez Kafka.

Les données doivent inclure: Date/heure de la vente, Montant de la vente, Produit vendu, Région de vente

- Sauvegarder les données cotés consommateur dans une base de données NoSQL. Utilisez MongoDB

Exercice 2 : Visualisation des données en streaming à l'aide de Tableau.

- Créer un projet Tableau et aller dans ce projet
- Configurer votre flux de données à Tableau en lui indiquant le serveur et le port sur lequel il doit écouter.
- Une fois connecté à vos données en streaming, commencez à créer vos visualisations.

Utilisez les champs de données appropriés pour créer des graphiques représentant l'évolution du chiffre d'affaires, la répartition géographique des ventes, les produits les plus vendus, etc.

- Explorez les différentes options de visualisation disponibles dans Tableau pour choisir les meilleures représentations pour vos données.

Exercice 2 : Visualisation des données en streaming à l'aide de Tableau.

- Ajout d'interactions et de filtres : Ajoutez des filtres interactifs pour permettre aux utilisateurs de sélectionner une période de temps spécifique, une région spécifique ou un produit spécifique.
- Personnalisation du tableau de bord : Créez un tableau de bord dans Tableau Desktop en rassemblant toutes vos visualisations créées précédemment.

Exercice 2 : Visualisation des données en streaming à l'aide de Tableau.

Ce qui est attendu:

- Un dépôt git contenant le producteur et le consommateur.
- Un compte rendu :
    - Expliquant la démarche mise en œuvre durant l'exercice
    - Contenant des screens du projet Tableau
    - Les différentes visualisations produites.
- Un partage du projet Tableau avec moi-même pour que je puisse y accéder.
- Une mise à jour en temps réel correcte des visualisations par rapport aux données.
- Des filtres interactifs fonctionnel comme prévu et des visualisations qui reflètent les sélections des utilisateurs.

Cet exercice compte pour l'évaluation finale. A rendre en binôme avant 21h. Aucun projet ne sera accepté passé ce délai.