

Vidéo : 2_I1
Introduction aux data lakes
Présentateur : Evan

[FIN DE LA VIDÉO]



Créer un data lake



Bonjour à tous. Je m'appelle Evan Jones et je suis développeur de cursus techniques avec Google. Bienvenue dans le chapitre sur la création d'un data lake.

Programme

Introduction aux data lakes

Stockage de données et options ETL dans GCP

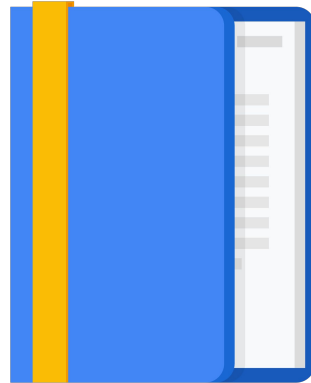
Créer un data lake avec Cloud Storage

Sécuriser Cloud Storage

Stocker des données de tous types

Cloud SQL en tant que data lake relationnel

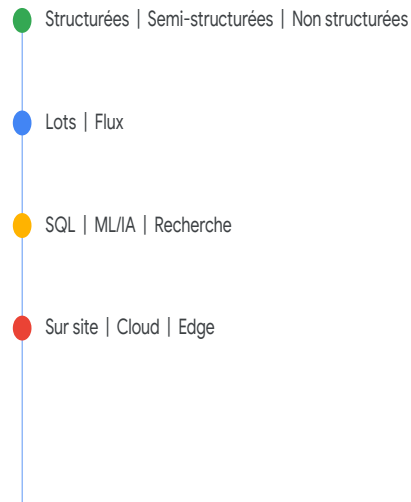
Atelier : Charger les données des taxis dans Cloud SQL



Nous allons d'abord présenter les data lakes et leur place au sein des composants essentiels de votre écosystème global d'ingénierie des données.

Qu'est-ce qu'un data lake ?

Une plate-forme de données évolutive et sécurisée qui permet aux entreprises d'ingérer, de stocker, de traiter et d'analyser tout type ou volume d'informations.



En fin de compte, qu'est-ce qu'un data lake ?

Ce terme est assez général, mais il décrit habituellement un endroit dans lequel vous pouvez stocker en toute sécurité différents types de données de toutes tailles pour les traiter et les analyser.

Les data lakes sont généralement utilisés pour les analyses de données, la science des données et les charges de travail de ML, ainsi que pour les pipelines de données par lots et par flux.

Les data lakes acceptent tous les types de données.

Et enfin, les data lakes sont mobiles, sur site ou dans le cloud.

Composants d'un écosystème d'ingénierie des données

- Sources de données
- Récepteurs de données
 - **Dépôt central de data lakes** ← L'objectif de ce module
 - Data warehouse
- Pipelines de données (lots et flux)
- Flux de travail d'orchestration de haut niveau



Voici la position qu'occupent les data lakes dans l'écosystème global d'ingénierie des données pour votre équipe.

Vous devez commencer par un ou plusieurs systèmes d'origine qui sont la source de toutes les données. Ensuite, l'ingénieur données doit élaborer des moyens fiables pour récupérer et stocker ces données, ce sont vos récepteurs de données. Le data lake constitue la première ligne de défense dans un environnement de données d'entreprise. Il s'agit encore du principe général qui nous permet de prendre en charge n'importe quelle donnée, quels que soient le volume, le format et la vitesse. Dans ce module, nous aborderons les options et les points clés pour créer un data lake.

<https://cloud.google.com/solutions/data-lake?hl=fr>

Composants d'un écosystème d'ingénierie des données


- Sources de données
- Récepteurs de données
 - Dépôt central de data lakes
 - **Data warehouse** ← **Module suivant**
- Pipelines de données (lots et flux)
- Flux de travail d'orchestration de haut niveau



Une fois que vos données sont extraites des systèmes sources et intégrées à votre environnement, vous devez généralement entreprendre un nettoyage et un traitement approfondis pour convertir ces données dans un format adapté à l'entreprise. Elles se retrouveront ensuite dans votre data warehouse. Nous nous intéressons à ce point dans le prochain module.

<https://cloud.google.com/solutions/data-lake?hl=fr>

Composants d'un écosystème d'ingénierie des données

- Sources de données
 - Récepteurs de données
 - Dépôt central de data lakes
 - Data warehouse
 - Pipelines de données (lots et flux)
 - Flux de travail d'orchestration de haut niveau
- Derniers modules + ML
- 



Dans la pratique, quel système assure le nettoyage et le traitement des données ?
Les pipelines de données. Leur rôle consiste à transformer et à traiter les données à grande échelle et à introduire des données nouvellement traitées dans l'ensemble de votre système pour les analyser.

L'ensemble de votre flux de travail désigne la couche d'abstraction supplémentaire au-dessus de vos pipelines. Vous devrez souvent coordonner les efforts de nombreux composants différents à une cadence régulière ou basée sur les événements. Lorsque votre pipeline de données traite les données de votre data lake à data warehouse, il se peut que votre flux de travail d'orchestration soit responsable du lancement de ce pipeline de données après avoir remarqué que de nouvelles données brutes étaient disponibles depuis une source.

<https://cloud.google.com/solutions/data-lake?hl=fr>

L'ingénierie des données est semblable au génie civil



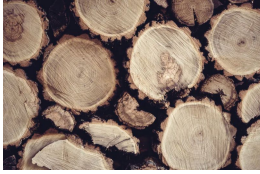
1. Les matières premières doivent être acheminées sur le site de travail (dans le data lake)



Avant de nous pencher sur les rôles des différents produits cloud, je voudrais terminer par une analogie qui m'a permis de distinguer ces composants.

Imaginez un instant que vous travaillez dans le domaine du génie civil. Vous êtes chargé de construire un gratte-ciel extraordinaire en plein centre-ville. Avant de commencer, vous devez vous assurer que vous disposez de toutes les **matières premières** nécessaires pour atteindre votre objectif final. Bien sûr, vous pouvez vous procurer certains matériaux plus tard, mais prenons cet exemple simple. Le transport vers votre chantier de l'acier, du béton, de l'eau, du bois, du sable et du verre provenant d'une source située ailleurs dans la ville est comparable à celui des données provenant de systèmes sources vers votre lac.

Transformer les matières premières en matière exploitable



1. Les matières premières doivent être acheminées sur le site de travail (dans le data lake)
2. Les matériaux doivent être coupés, transformés en fonction des besoins et stockés (des pipelines aux récepteurs de données)



Parfait ! Vous disposez désormais de toutes ces matières premières, mais vous ne pouvez pas les utiliser telles quelles pour construire votre bâtiment. Vous devez couper le bois et le métal, mesurer et ajuster la taille du verre avant de pouvoir utiliser ces matériaux pour la construction. Les produits finis, le verre coupé et le métal façonné, correspondent aux données formatées qui sont stockées dans votre data warehouse. Vous pouvez vous en servir pour ajouter directement de la valeur à votre entreprise, ce qui, dans notre analogie, revient à construire le bâtiment. Comment avez-vous transformé ces matières premières en produits exploitables ? Sur un chantier, cette tâche incombe à l'ouvrier. Comme vous le verrez plus tard lorsque nous parlerons des pipelines de données, l'unité individuelle en arrière-plan est appelée nœud de calcul (qui n'est qu'une VM) et elle prend un petit élément de données et le transforme pour vous.

Le nouveau bâtiment est le nouvel insight, le nouveau modèle de ML, etc.



 Google Cloud

1. Les matières premières doivent être acheminées sur le site de travail (dans le data lake)
2. Les matériaux doivent être coupés, transformés en fonction des besoins et stockés (des pipelines aux récepteurs de données)
3. Le bâtiment actuel est le nouvel insight ou modèle de ML, etc.

Qu'en est-il du bâtiment ? Il s'agit de l'objectif final ou des objectifs fixés pour ce projet d'ingénierie. Dans le domaine de l'ingénierie des données, le bâtiment flambant neuf pourrait être un tout nouvel insight d'analyse qui était auparavant impossible, un modèle de machine learning ou tout autre objectif que vous souhaitez atteindre depuis que vous disposez de données nettoyées.

Un système d'orchestration régit tous les aspects du flux de travail



 Google Cloud

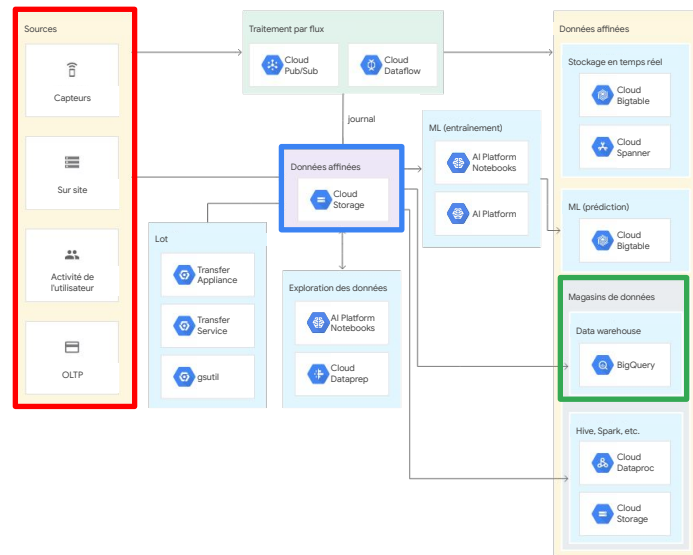
1. Les matières premières doivent être acheminées sur le site de travail (dans le data lake)
2. Les matériaux doivent être coupés, transformés en fonction des besoins et stockés (des pipelines aux récepteurs de données)
3. Le bâtiment actuel est le nouvel insight ou modèle de ML, etc.
4. Le responsable supervise tous les aspects et toutes les équipes du projet (orchestration du flux de travail)

La couche d'orchestration est la dernière partie de l'analogie. Un chantier est dirigé par un directeur ou un responsable qui détermine le déroulement des travaux et les éventuelles dépendances. Il pourrait dire : "lorsque le métal arrive, transportez-le vers cette zone du site pour le couper et le façonner, puis prévenez l'autre équipe qu'il est disponible". Dans le domaine de l'ingénierie des données, il s'agit de la couche d'orchestration ou du flux de travail global. Vous pourriez donc dire : "Chaque fois qu'un nouvel élément de données CSV arrive dans ce bucket Cloud Storage, je veux que vous le transmettiez automatiquement à notre pipeline de données pour le traiter. Lorsque le traitement est terminé, je veux que vous (le pipeline) l'envoyiez dans le data warehouse. Une fois qu'il est dans le , je signalerai au modèle de machine learning que de nouvelles données d'entraînement nettoyées sont disponibles pour l'entraînement et je lui demanderai de commencer à entraîner une nouvelle version du modèle." Pouvez-vous voir le graphique des actions de construction ? Que faire en cas d'échec d'une étape ? Que faire si vous voulez gérer cela au quotidien ? Vous commencez à voir la nécessité d'un système d'orchestration. Nous proposons donc Apache Airflow comme solution, elle fonctionnera plus tard sur Cloud Composer.

<https://cloud.google.com/solutions/data-lake?hl=fr>

Exemple d'architecture

1. Sources de données
2. Data lake
3. Pipelines de données
4. Data warehouse
5. Utilisé pour les charges de travail de ML et d'analyse



Revenons à un exemple de schéma d'architecture de solution que nous avons déjà vu dans le cours.

Ici, ce sont les buckets Google Cloud Storage au centre du schéma qui constituent le data lake. Il s'agit d'un emplacement consolidé, durable et hautement disponible pour les données brutes. Dans cet exemple, Google Cloud Storage est notre data lake, mais il ne s'agit pas de la seule solution en matière de data lakes. Je le répète, Cloud Storage fait partie des solutions intéressantes qui peuvent servir de data lake. Dans d'autres exemples que nous allons examiner, vous pouvez utiliser BigQuery comme data lake et comme data warehouse, sans utiliser le moindre bucket Cloud Storage. C'est pourquoi il est si important de bien comprendre ce que vous recherchez pour trouver les solutions qui répondent le mieux à vos besoins.

Quels que soient les outils et les technologies cloud que vous utilisez, votre data lake sert généralement d'emplacement unique et consolidé pour toutes vos données brutes. Je considère cela comme une zone de préproduction durable. Les données peuvent finir dans de nombreux autres emplacements, comme un pipeline de transformation qui les nettoie et les transporte vers l'entrepôt, avant qu'un modèle de machine learning ne les lise, mais tout commence par l'introduction de ces données dans votre lac.

Passons rapidement en revue quelques produits essentiels de big data de Google Cloud que vous devez connaître en tant qu'ingénieur données et que vous pourrez tester ultérieurement dans les ateliers.

Suite de produits de big data sur Google Cloud Platform



© 2018 Google LLC. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

Voici la liste complète des produits de big data et de ML, classés en fonction de l'emplacement susceptible de les contenir dans une charge de travail de traitement de données classique. Cela va du stockage des données sur la gauche à leur ingestion dans vos outils cloud natifs pour l'analyse, l'entraînement de modèles de machine learning et la diffusion d'insights.

Vous allez créer des data lakes évolutifs et durables grâce aux solutions de stockage de GCP



© 2018 Google LLC. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

Dans ce module sur les data lakes, nous nous intéresserons à deux des produits Storage essentiels qui constitueront le data lake : Google Cloud Storage et Cloud SQL pour les données relationnelles. Par la suite, vous vous entraînerez également avec Cloud Bigtable lorsque vous utiliserez des pipelines par flux à haut débit.

Comme moi, vous avez peut-être été surpris de ne pas voir BigQuery dans la colonne de stockage au début de la formation sur Google Cloud Platform. En général, BigQuery sert de data warehouse. Quelle est donc la principale différence entre un data lake et un data warehouse ?

Data lake ou data warehouse

Un data lake est une capture de chaque aspect de votre activité commerciale. Les données sont stockées dans leur format naturel/brut, généralement sous forme de blobs ou de fichiers d'objets.

- Conservation de toutes les données dans leur format natif
- Prise en charge de tous les types de données et de tous les utilisateurs
- Adaptation facile aux changements
- Généralement spécifique à une application



Un data lake est avant tout l'emplacement où vous avez enregistré tous les aspects de vos activités commerciales. Dans la mesure où vous voulez enregistrer chaque aspect, vous avez tendance à stocker les données dans leur format brut naturel, le format généré par votre application. Vous pouvez donc obtenir un fichier journal et les fichiers journaux stockés tels quels dans un data lake. Pour résumer, vous pouvez stocker tout ce que vous voulez, et puisque vous voulez tout stocker, vous avez tendance à stocker ces éléments sous forme de blobs ou de fichiers d'objets.

La flexibilité du point de collecte central que constitue le data lake est à la fois un avantage et un problème. Avec un data lake, le format des données est essentiellement déterminé par l'application qui écrit les données. Le data lake offre l'avantage de pouvoir commencer à écrire les nouvelles données dès la mise à niveau d'une application, car il ne s'agit que d'une capture des données brutes existantes. Comment exploiter cette grande quantité de données brutes et flexibles ?

Data lake ou data warehouse

En revanche, un data warehouse présente généralement les caractéristiques suivantes :

- Uniquement chargé lorsque son utilisation est définie
- Traité/Organisé/Transformé
- Accélération de la diffusion des insights
- Données actuelles/historiques pour créer des rapports
- Schéma cohérent généralement partagé entre les applications



Accédez à le data warehouse.

En revanche, le data warehouse est une solution beaucoup plus réfléchie. Vous pouvez charger les données dans un data warehouse uniquement après avoir défini un schéma et identifié le cas d'utilisation. Vous pouvez transformer, organiser, traiter et nettoyer les données brutes existantes d'un data lake, puis les stocker dans un data warehouse. Pourquoi stocker les données dans l'entrepôt ? Peut-être parce que les données contenues dans le data warehouse sont utilisées pour générer des graphiques, des rapports, des tableaux de bord, etc. Le principe est le suivant : comme le schéma est cohérent et partagé entre toutes les applications, il est possible d'analyser les données et d'en extraire des insights plus rapidement. Un data warehouse est donc généralement constitué de données structurées et semi-structurées qui sont organisées et stockées dans un format qui permet de les interroger et de les analyser.

Vidéo : 2_l2

Stockage de données et options ETL dans GCP

Présentateur : Evan

[FIN DE LA VIDÉO]

Programme

Introduction aux data lakes

Stockage de données et options ETL dans GCP

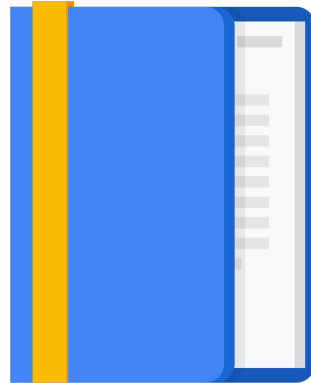
Créer un data lake avec Cloud Storage

Sécuriser Cloud Storage

Stocker des données de tous types

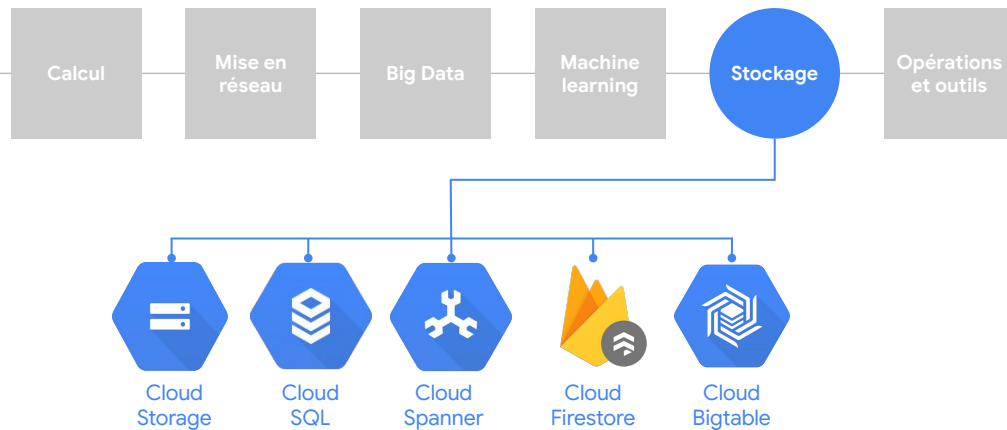
Cloud SQL en tant que data lake relationnel

Atelier : Charger les données des taxis dans Cloud SQL



Ensuite, nous allons discuter du stockage des données et des possibilités d'extraction, de transformation et de chargement sur GCP.

Options de stockage de vos données sur GCP



Les solutions de stockage de GCP que vous avez vues plus tôt vous permettent de créer un data lake. Cloud Storage est une solution extrêmement polyvalente, Cloud SQL et Cloud Spanner sont destinés aux données relationnelles, tandis que Cloud Firestore et Cloud Bigtable sont réservés aux données NoSQL. Le choix entre les différentes solutions dépend largement de votre cas d'utilisation et de ce que vous souhaitez créer. Dans ce chapitre, nous nous pencherons sur Cloud Storage et Cloud SQL, mais vous verrez les options NoSQL comme Cloud Bigtable plus tard dans le cours lorsque nous aborderons les flux à haut débit.

Comment déterminer le chemin du lac ?

Le chemin emprunté par vos données pour accéder au cloud dépend :

- de la position actuelle de vos données.



La destination finale des données sur le cloud et les chemins empruntés pour les acheminer vers le cloud dépendent de l'emplacement actuel des données,

Le chemin emprunté par vos données pour accéder au cloud dépend :

- de la position actuelle de vos données.
- de la taille de vos données.



de leur taille (il s'agit du composant Volume des 3 V de big data)

Le chemin emprunté par vos données pour accéder au cloud dépend :

- de la position actuelle de vos données.
- de la taille de vos données.
- de leur destination.



et, enfin, **de l'emplacement souhaité**. Dans les schémas d'architecture, le point d'arrivée des données est appelé le récepteur de données. Après le data lake, le récepteur de données le plus courant est le data warehouse.

Le chemin emprunté par vos données pour accéder au cloud dépend :

- de la position actuelle de vos données.
- de la taille de vos données.
- de leur destination.
- de l'ampleur de la transformation nécessaire.



Il est essentiel de tenir compte de l'ampleur du traitement et de la transformation nécessaires aux données avant que votre entreprise puisse les exploiter.

À présent, vous vous demandez peut-être si vous devriez traiter les données avant de les charger ou après les avoir chargées dans le data lake, avant qu'elles soient envoyées ailleurs. Parlons de ces modèles.

La méthode que vous utilisez pour charger les données dépend de l'ampleur de la transformation nécessaire

EL (Extract and Load)



Extraire et charger



La méthode que vous utilisez pour charger les données dans le cloud dépend du niveau de transformation nécessaire pour convertir les données brutes au format final souhaité. Dans ce chapitre, nous allons examiner quelques points à prendre en compte pour le format final souhaité.

Disposer de données dans un format permettant une ingestion facile par le produit cloud dans lequel vous souhaitez les stocker est le cas le plus simple.

Imaginons que vous disposez de données au format Avro et que vous souhaitez les stocker dans BigQuery, car il s'agit de la solution la plus adaptée à votre cas d'utilisation. Dans ce cas, nous nous contentons d'EL. Extraire et charger. BigQuery charge directement les fichiers Avro. **EL signifie extraire et charger (EL)**. Cela fait référence au moment pendant lequel les données peuvent être importées "en l'état" dans un système. Il s'agit par exemple d'importer des données à partir d'une base de données, où la source et la cible ont le même schéma.

BigQuery se distingue par une fonctionnalité qui, comme vous l'avez vu dans l'exemple de la requête fédérée, vous permet de ne pas charger les données dans BigQuery et de tout de même pouvoir les interroger. Les fichiers Avro, ORC et Parquet sont désormais tous pris en charge pour les requêtes fédérées, comme vous le verrez dans une démonstration ultérieure.

<https://pixabay.com/fr/illustrations/entrep%C3%B4t-transport-maritime-bo%C3%A0Ete-3688280/>

<https://pixabay.com/fr/vectors/d%C3%A9placement-bo%C3%A0Ete-d%C3%A0m%C3%A0C>

3%A9nagement-312082/

<https://pixabay.com/fr/vectors/bo%C3%AEte-voiture-chariot-%C3%A9l%C3%A9vat eur-159302/>

La méthode que vous utilisez pour charger les données dépend de l'ampleur de la transformation nécessaire

EL (Extract and Load)



Extraire et charger

ELT (Extract, Load, and Transform)



Extraire, charger
et transformer



Que signifie ELT ou extraire, charger et transformer les données ?

Cela signifie que les données chargées dans le produit cloud ne sont pas dans le format désiré. Il se peut que vous vouliez les nettoyer. Vous pourriez également envisager de transformer les données afin de les corriger, par exemple. En d'autres termes, vous procéderiez à l'extraction depuis le système sur site, au chargement dans le produit cloud, puis à la transformation.

Il s'agit d'extraire, de charger et de transformer, ou ELT. Vous avez tendance à opter pour cette solution lorsque le niveau de transformation nécessaire n'est pas très élevé et que la transformation ne réduit pas énormément la quantité de données à votre disposition. L'ELT permet de charger des données brutes directement dans la cible et de les y transformer. Par exemple, dans BigQuery, vous pouvez utiliser SQL pour transformer les données et écrire une nouvelle table.

<https://pixabay.com/fr/illustrations/entrep%C3%B4t-transport-maritime-bo%C3%A0Ete-3688280/>

<https://pixabay.com/fr/vectors/d%C3%A9placement-bo%C3%A0Ete-d%C3%A0m%C3%A0nagement-312082/>

<https://pixabay.com/fr/vectors/bo%C3%A0Ete-voiture-chariot-%C3%A0l%C3%A0vat eur-159302/>

La méthode que vous utilisez pour charger les données dépend de l'ampleur de la transformation nécessaire

EL (Extract and Load)



Extraire et charger

ELT (Extract, Load, and Transform)



Extraire, charger
et transformer

ETL (Extract, Transform, and Load)



Extraire, transformer
et charger



La troisième option consiste à extraire, transformer et charger, ou ETL. C'est le cas lorsque vous voulez extraire les données, les soumettre à un certain nombre de traitements, puis les charger dans le produit cloud. En général, vous choisissez cette option lorsque cette transformation est essentielle ou si cette transformation réduit considérablement la taille des données. Ainsi, en transformant les données avant de les charger dans le cloud, il est possible de réduire considérablement la bande passante réseau dont vous avez besoin. Si les données sont dans un format binaire propriétaire et que vous devez les convertir avant de les charger, vous avez également besoin de l'ETL.

L'extraction, la transformation, le chargement (ETL) forment un processus d'intégration de données dans lequel la transformation a lieu dans un service intermédiaire avant le chargement dans la cible. Par exemple, les données peuvent être transformées dans un pipeline de données comme Cloud Dataflow avant d'être chargées dans BigQuery.

<https://pixabay.com/fr/illustrations/entrep%C3%B4t-transport-maritime-bo%C3%A0Ete-3688280/>

<https://pixabay.com/fr/vectors/d%C3%A9placement-bo%C3%A0Ete-d%C3%A9m%C3%A9nagement-312082/>

<https://pixabay.com/fr/vectors/bo%C3%A0Ete-voiture-chariot-%C3%A9l%C3%A9vateur-159302/>

Vidéo : 2_I3

Créer un data lake avec Cloud Storage

Présentateur : Evan

[FIN DE LA VIDÉO]

Programme

Introduction aux data lakes

Stockage de données et options ETL dans GCP

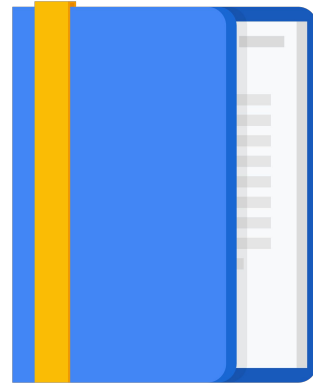
Créer un data lake avec Cloud Storage

Sécuriser Cloud Storage

Stocker des données de tous types

Cloud SQL en tant que data lake relationnel

Atelier : Charger les données des taxis dans
Cloud SQL

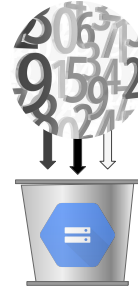


Cloud Storage est le service de stockage par excellence pour le traitement des données, en particulier les données non structurées dans le cloud. Découvrons pourquoi la solution de data lake la plus populaire est Google Cloud Storage.

Cloud Storage



Cloud Storage



Qualités de Cloud Storage
qui contribuent au succès des
solutions Data Engineering :

Persistance Durabilité Forte cohérence
Disponibilité Débit élevé



Les données dans Cloud Storage sont conservées au-delà de la durée de vie des VM ou des clusters (autrement dit, elles sont persistantes). Cette solution est également relativement économique par rapport au coût du calcul.

Il peut donc être plus avantageux, par exemple, de mettre en cache les résultats des calculs précédents dans Cloud Storage.

Si vous n'avez pas besoin qu'une application fonctionne en permanence, il peut être utile d'enregistrer l'état de votre application dans Cloud Storage et d'éteindre la machine utilisée lorsque vous n'en avez pas besoin.

Cloud Storage est un magasin d'objets qui stocke et récupère des objets binaires sans tenir compte des données que ces objets contiennent. Toutefois, il assure également, dans une certaine mesure, la compatibilité des systèmes de fichiers et permet aux objets de ressembler à des fichiers et de fonctionner comme tels, afin que vous puissiez copier des fichiers à l'intérieur et à l'extérieur.

Les données stockées dans Cloud Storage y sont conservées à jamais (autrement dit, elles sont durables), mais sont immédiatement disponibles, offrant une cohérence forte.

Vous pouvez partager des données à l'échelle mondiale, mais elles sont chiffrées, entièrement contrôlées et privées si vous le souhaitez.

Il s'agit d'un service international et vous pouvez accéder aux données de n'importe où (en d'autres termes, il offre une disponibilité mondiale). Cependant, les données

peuvent aussi être conservées dans un seul emplacement géographique si nécessaire.

Les données sont diffusées avec une latence modérée et un débit élevé.

En tant qu'ingénieur données, vous devez comprendre la façon dont Cloud Storage réunit ces qualités apparemment contradictoires, ainsi que le moment et la manière de les utiliser dans les solutions.

<https://pixabay.com/fr/illustrations/%C3%A9chechs-noir-et-blanc-de-la-damier-3413429/https://pixabay.com/fr/illustrations/payer-chiffres-nombre-remplissage-1036469/>

Comment fonctionne Cloud Storage ?

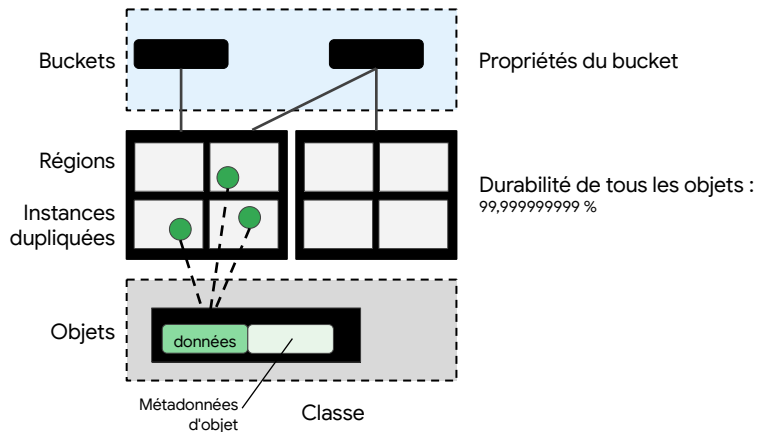


Un espace de noms global unique simplifie la localisation des buckets et des objets

Emplacement pour contrôler la latence

Durabilité et disponibilité

Les noms d'objets longs simulent la structure



De nombreuses propriétés exceptionnelles de Cloud Storage sont dues au fait qu'il s'agit d'un magasin d'objets. De plus, d'autres fonctionnalités sont ajoutées à cette base.

Les buckets et les objets sont les deux principales entités de Cloud Storage. Les buckets sont des conteneurs d'objets. Et les objets existent uniquement au sein des buckets. Les buckets sont donc des conteneurs de données.

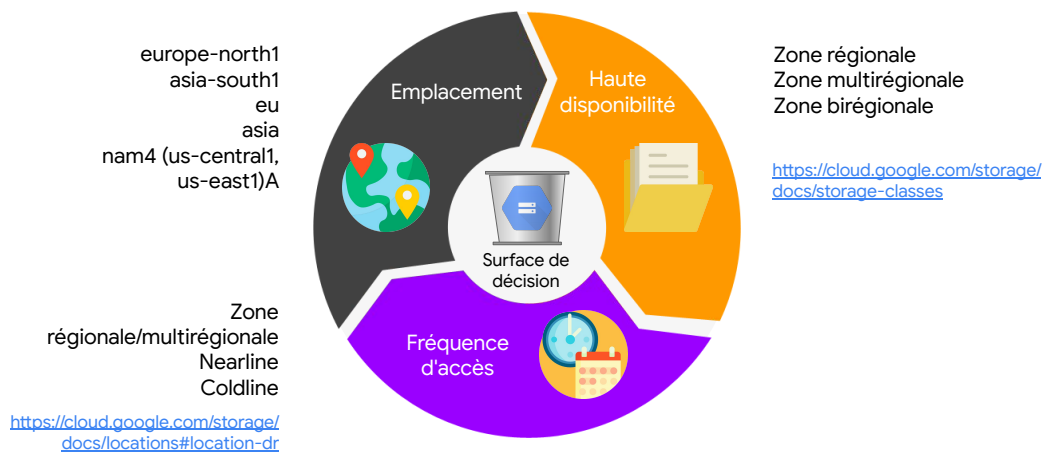
Les **buckets** sont identifiés dans un seul espace de noms global et unique. Par conséquent, lorsque le bucket est nommé, le nom ne peut plus être utilisé tant que le bucket n'a pas été supprimé et que le nom n'est pas disponible. Un espace de noms global pour les buckets simplifie la localisation d'un bucket particulier. Lorsqu'un bucket est créé, il est associé à une région particulière ou à plusieurs régions. Choisir une région proche de l'emplacement où les données seront traitées permet de réduire la latence, et utiliser des services cloud dans la région pour les traiter permet d'économiser les frais de sortie réseau.

Cloud Storage réplique un objet **lorsqu'il est stocké**. Il surveille les instances dupliquées, et si l'une d'entre elles est perdue ou corrompue, il la remplace par une nouvelle copie. Cette méthode permet à Cloud Storage d'obtenir une durabilité élevée (de plusieurs fois le chiffre 9). Dans le cas d'un bucket multirégional, les objets sont répliqués entre les régions, et dans le cas d'un bucket à région unique, les objets sont répliqués entre les zones.

Dans tous les cas, lorsque l'objet est récupéré, il est diffusé à partir de l'instance dupliquée la plus proche du demandeur. C'est ainsi que nous obtenons une faible latence. Plusieurs demandeurs peuvent récupérer les objets simultanément à partir de différentes instances dupliquées. C'est ainsi que nous obtenons un débit élevé.

Enfin, les objets sont stockés avec des métadonnées. Les métadonnées sont des informations sur l'objet. Certaines fonctionnalités supplémentaires de Cloud Storage utilisent les métadonnées, notamment pour le contrôle d'accès, la compression, le chiffrement et la gestion du cycle de vie. Par exemple, Cloud Storage connaît la date de stockage d'un objet et peut être configuré pour le supprimer automatiquement après un certain temps. Cette fonctionnalité utilise les métadonnées de l'objet pour déterminer le moment de sa suppression.

Les propriétés du bucket dépendent de vos besoins



Vous devez prendre plusieurs décisions lorsque vous créez un bucket.

La première est l'emplacement du bucket. L'emplacement est défini lors de la création du bucket et ne peut jamais être modifié. Pour déplacer un bucket ultérieurement, vous devrez copier l'ensemble du contenu dans le nouvel emplacement et payer les frais de sortie réseau. Choisissez donc l'emplacement avec soin. L'emplacement peut être une région unique, comme europe-north1 ou asia-south1. Il peut être multirégional, comme eu ou asia, ce qui signifie que l'objet est répliqué dans plusieurs régions de l'Union européenne ou de l'Asie. La troisième option consiste à choisir un bucket birégional comme emplacement. Par exemple, north-america4 signifie que l'objet est répliqué en us-central1 et us-east1.

Comment choisir la région ? Imaginons que tous vos calculs et utilisateurs se trouvent en Asie. Vous choisissez donc une région asiatique pour réduire la latence du réseau. En dehors de cela, comment choisir entre asia-south1 et l'Asie multirégionale ?

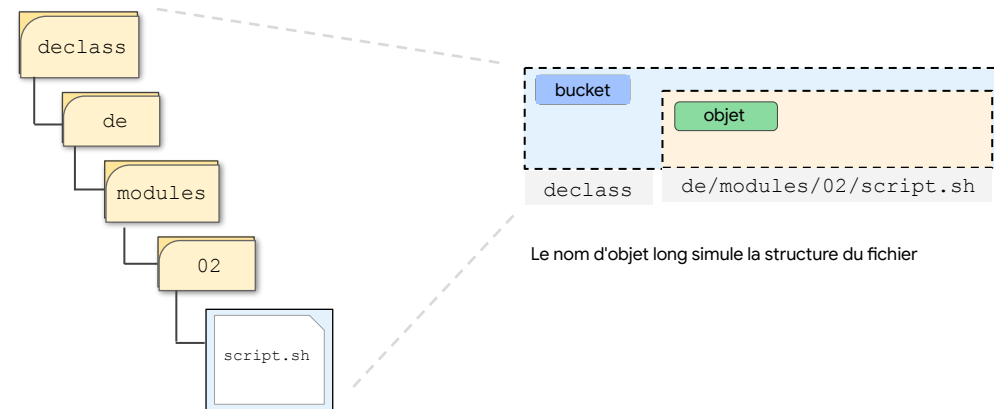
Vous pouvez sélectionner une région et les données seront répliquées dans plusieurs zones de la région. De cette façon, il est possible d'accéder aux données même si une zone est indisponible. Les différentes zones d'une même région offrent une protection contre la plupart des types de pannes : pannes d'infrastructure, pannes matérielles et pannes logicielles. Toutefois, si toute la région est touchée, par exemple en cas d'inondation, vous ne pourrez pas accéder aux données.

Pour vous assurer que les données sont disponibles lors d'une catastrophe naturelle, vous pouvez choisir une solution multirégionale ou birégionale, auquel cas les instances dupliquées sont stockées dans des centres de données séparés physiquement. Les contrats de niveau de service actuels sont disponibles en ligne.

Troisièmement, vous devez déterminer à quelle fréquence vous devez accéder aux données ou les modifier. Vous pouvez bénéficier de remises importantes sur le stockage des données si vous êtes prêt à déboursier davantage lorsque vous devez y accéder. La remise est intéressante si vous accédez à vos données au maximum une fois par mois ou une fois par trimestre. Le stockage d'archives, les sauvegardes ou la reprise après sinistre sont de bons exemples. De plus, la remise est vraiment avantageuse si vous n'accédez aux données qu'une fois par trimestre ou une fois par an. Ce sont les classes de stockage. Consultez le lien pour découvrir les contrats de niveau de service et les coûts associés à ces classes de stockage.

<https://pixabay.com/fr/vectors/dossier-ouverte-fichier-d-affaires-27857/>
<https://pixabay.com/fr/illustrations/horloge-temps-calendrier-163202/>

Cloud Storage simule un système de fichiers



Accès aux fichiers `gs://declass/de/modules/02/script.sh`

Accès Web `https://storage.cloud.google.com/declass/de/modules/02/script.sh`



Cloud Storage utilise le nom du bucket et le nom de l'objet pour simuler un système de fichiers. Voici le fonctionnement. Le nom du bucket est le premier terme de l'URI. Une barre oblique y est ajoutée, elle est ensuite concaténée avec le nom de l'objet. La barre oblique est un caractère autorisé dans le nom de l'objet. Le nom d'objet très long contenant des barres obliques ressemble à un chemin de système de fichiers, même s'il ne s'agit que d'un nom unique.

Dans cet exemple, le nom du bucket est "**declass**". Le nom de l'objet est "**de/modules/02/script.sh**". Les barres obliques ne sont que des caractères du nom. Si ce chemin se trouvait dans un système de fichiers, il apparaîtrait comme illustré sur la gauche, avec un ensemble de répertoires imbriqués, commençant par "declass".

Dans la pratique, il fonctionne comme un système de fichiers. Néanmoins, il existe des différences. Par exemple, imaginons que vous vouliez déplacer tous les fichiers du répertoire 02 vers le répertoire 03 à l'intérieur du répertoire des modules. Un système de fichiers contient de véritables structures de répertoire et il suffit de modifier les métadonnées du système de fichiers pour que le mouvement soit entièrement atomique. En revanche, la simulation d'un système de fichiers dans un magasin d'objets nécessite de rechercher les noms contenant "02" correctement positionné dans le nom parmi tous les objets contenus dans le bucket. Il faut ensuite renommer chaque objet en utilisant 03. Apparemment, cela donne le même résultat : le déplacement des fichiers d'un répertoire à l'autre. Cependant, au lieu d'une douzaine de fichiers dans un répertoire, le système a dû rechercher des milliers

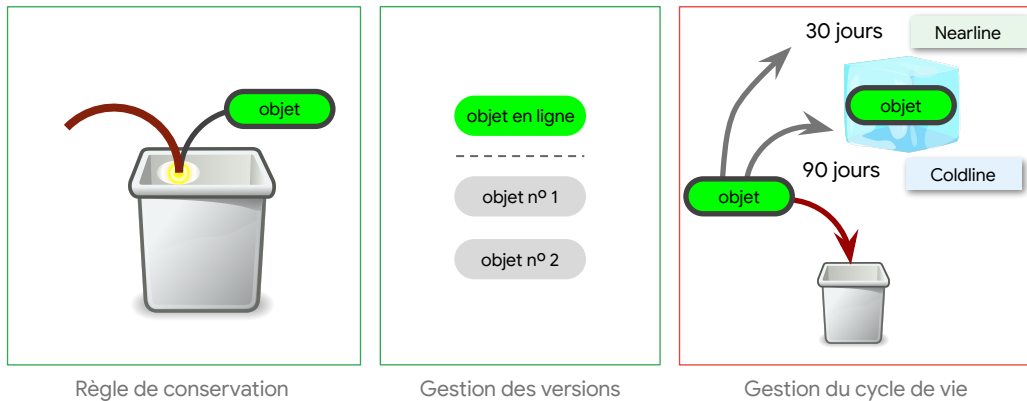
d'objets dans le bucket pour trouver ceux portant les noms recherchés et modifier chacun d'entre eux. Les caractéristiques de performances sont donc différentes. Déplacer une douzaine d'objets du répertoire 02 vers le répertoire 03 peut prendre plus de temps en fonction du nombre des autres objets stockés dans le bucket. Pendant le déplacement, vous obtenez une liste d'incohérences. Certains fichiers se trouvent dans l'ancien répertoire et d'autres dans le nouveau.

Il est recommandé d'éviter l'utilisation d'informations sensibles dans les noms de bucket, car ces derniers se trouvent dans un espace de noms global. Les noms de buckets, pas les données des buckets. La confidentialité des données contenues dans les buckets peut être préservée si nécessaire.

Vous pouvez utiliser une méthode d'accès aux fichiers pour accéder à Cloud Storage. Cela vous permet, par exemple, d'utiliser une commande de copie d'un fichier local directement vers Cloud Storage. Pour ce faire, utilisez l'outil gsutil.

Vous pouvez également accéder à Cloud Storage sur le Web. Le site (<https://storage.cloud.google.com>) utilise TLS (HTTPS) pour transporter vos données, protégeant ainsi les identifiants et les données en transit.

Cloud Storage offre de nombreuses fonctionnalités de gestion d'objets



Cloud Storage offre de nombreuses fonctionnalités de gestion d'objets. Par exemple, vous pouvez définir une règle de conservation pour tous les objets se trouvant dans le bucket. Par exemple, les objets doivent être arrivés à expiration au bout de 30 jours.

Vous pouvez aussi utiliser la gestion des versions, afin de suivre les différentes versions d'un objet et de les rendre disponibles si nécessaire. Vous pouvez même configurer la gestion du cycle de vie, pour déplacer automatiquement les objets qui n'ont pas été consultés après 30 jours vers Nearline et après 90 jours vers Coldline.

Examinons la façon dont vous pouvez gérer les cycles de vie de ces objets de manière un peu plus automatisée, afin d'optimiser l'utilisation de Cloud Storage et de réduire les coûts de stockage.

<https://pixabay.com/fr/vectors/corbeille-corbeille-%C3%A0-papier-155907/https://pixabay.com/fr/vectors/la-glace-cube-bleu-l-eau-bloc-34075/>

Vidéo : 2_l4

Démonstration : Optimiser les coûts avec les classes
Google Cloud Storage et Cloud Functions

Présentateur : Evan

[FIN DE LA VIDÉO]

Démonstration

Optimiser les coûts avec
les classes Google Cloud
Storage et Cloud Functions

[DÉMONSTRATION **FACULTATIVE** - enregistrement disponible dans la v2.0.1]

<https://www.qwiklabs.com/focuses/7830?parent=catalog>

GCP propose des règles de cycle de vie d'objets de stockage que vous pouvez utiliser pour déplacer automatiquement les objets vers différentes classes de stockage. Ces règles peuvent être basées sur un ensemble d'attributs, comme la date de création ou l'état actuel. Cependant, elles ignorent si les objets ont été consultés. Vous pouvez réduire vos frais en déplaçant les objets les plus récents vers le stockage Nearline s'ils n'ont pas été consultés pendant un certain temps.

Dans cette démonstration, nous allons créer deux buckets de stockage et générer du trafic sur l'un d'entre eux.

Ensuite, nous allons déployer une fonction Cloud pour migrer le bucket inactif vers une classe de stockage plus économique. Vous pouvez tester cela en utilisant une notification fictive de Stackdriver pour déclencher la fonction.

Vidéo : 2_I5
Sécuriser Cloud Storage
Présentateur : Evan

[FIN DE LA VIDÉO]

Programme

Introduction aux data lakes

Stockage de données et options ETL dans GCP

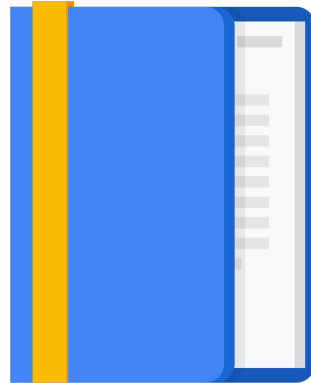
Créer un data lake avec Cloud Storage

Sécuriser Cloud Storage

Stocker des données de tous types

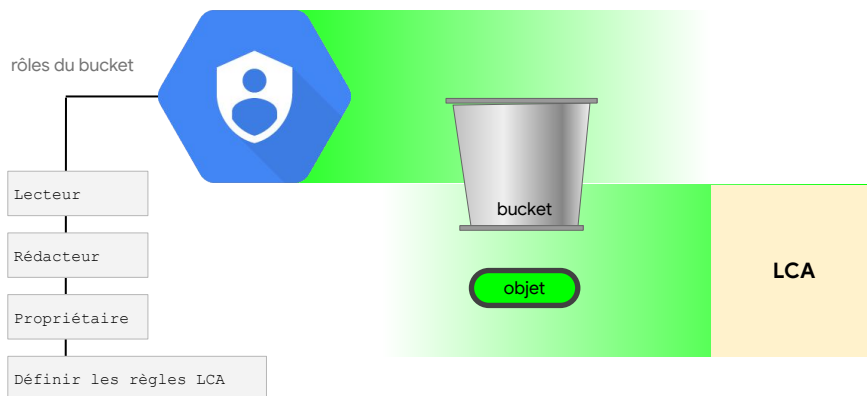
Cloud SQL en tant que data lake relationnel

Atelier : Charger les données des taxis dans
Cloud SQL



La sécurisation de votre data lake fonctionnant sur Cloud Storage est d'une importance capitale. Nous discuterons des principales fonctionnalités de sécurité que vous devez connaître en tant qu'ingénieur données pour contrôler l'accès à vos objets.

Contrôler l'accès avec Cloud IAM et les listes d'accès



Cloud Storage met en œuvre deux méthodes complètement distinctes, mais qui se chevauchent pour contrôler l'accès aux objets : la stratégie Cloud IAM et les listes de contrôle d'accès. Cloud IAM est une solution standard de Google Cloud Platform. Elle est définie au niveau du bucket et applique des règles d'accès uniformes à tous les objets contenus dans un bucket. Les listes de contrôle d'accès peuvent être appliquées au niveau du bucket ou sur des objets individuels. Elles offrent donc un contrôle d'accès plus précis

Les contrôles Cloud IAM sont tels que vous pouvez les imaginer. Cloud IAM offre des rôles de projet et des rôles de bucket. Notamment les suivants : lecteur de bucket, rédacteur de bucket et propriétaire de bucket. Le rôle IAM au niveau du bucket permet de créer ou de modifier les listes de contrôle d'accès. Le rôle au niveau du projet permet de créer et supprimer des buckets et de définir la stratégie IAM. Des rôles personnalisés sont disponibles. Les rôles lecteur, éditeur et propriétaire au niveau du projet permettent aux utilisateurs d'être membres de groupes internes spéciaux qui leur donnent accès aux rôles du bucket. Consultez la documentation en ligne pour plus d'informations.

Vous pouvez désactiver les listes d'accès et n'utiliser que Cloud IAM lorsque vous créez un bucket. Les listes d'accès sont actuellement activées par défaut. Avant, ce choix était immuable. Désormais, vous pouvez désactiver les listes d'accès même si elles étaient en vigueur auparavant.

Par exemple, vous pouvez accorder à un certain bob@example.com l'accès en

lecture à un bucket grâce à IAM. Vous pouvez aussi lui donner l'accès en écriture à un fichier spécifique dans ce bucket grâce aux listes de contrôle d'accès. Vous pouvez également accorder de telles autorisations à des comptes de service associés à des applications individuelles.

Options de chiffrement des données pour les différentes exigences



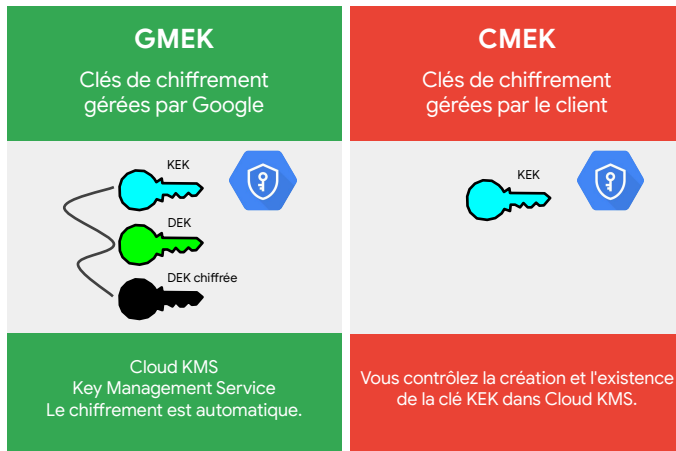
Toutes les données dans Google Cloud sont chiffrées au repos et en transit. Il est impossible de désactiver le chiffrement.

Le chiffrement est effectué par Google à l'aide de clés de chiffrement que nous gérons : **clés de chiffrement gérées par Google (GMEK, Google Managed Encryption Keys)**.

Nous utilisons deux niveaux de chiffrement : Les données sont d'abord chiffrées à l'aide d'une clé de chiffrement de données. Ensuite, la clé de chiffrement des données est elle-même chiffrée à l'aide d'une clé de chiffrement (KEK, Key encryption key).

Les KEK font l'objet d'une rotation automatique selon un calendrier et la KEK actuelle est stockée dans le service de gestion des clés Cloud KMS (Key Management Service). Aucune intervention n'est nécessaire. Il s'agit d'un comportement automatique.

Options de chiffrement des données pour les différentes exigences



Vous pouvez aussi gérer la KEK par vous-même. Il est possible de contrôler la création et l'existence de la KEK utilisée, au lieu de confier la gestion de la clé de chiffrement à Google.

Dans ce cas, nous parlons de **clés de chiffrement gérées par le client (CMEK, Customer Managed Encryption Keys)**.

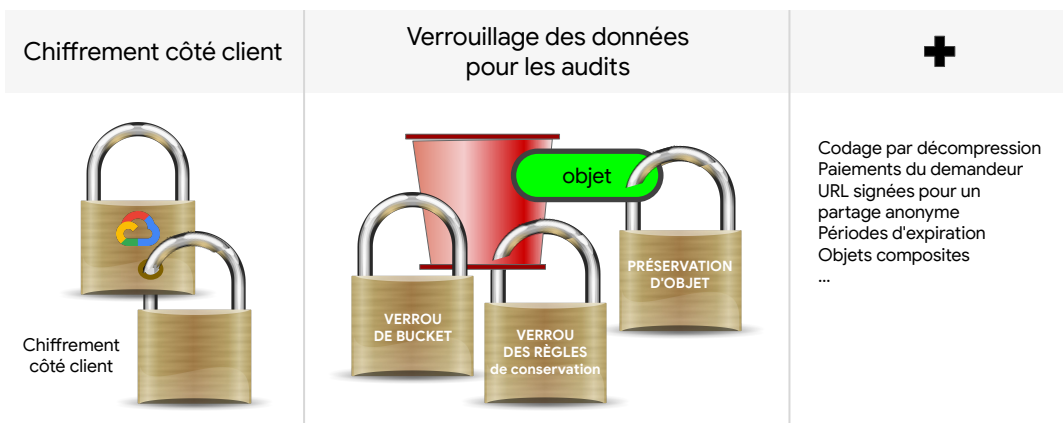
Options de chiffrement des données pour les différentes exigences



Vous pouvez complètement éviter Cloud KMS et fournir votre propre mécanisme de chiffrement et de rotation. Dans ce cas, nous parlons de **clés de chiffrement fournies par le client (CSEK, Customer Supplied Encryption Keys)**.

L'option de chiffrement des données est choisie en fonction des exigences commerciales, légales et réglementaires. Veuillez vous adresser au conseiller juridique de votre entreprise.

Cloud Storage prend en charge de nombreux cas d'utilisation particuliers



La quatrième option de chiffrement est le chiffrement côté client.

Le chiffrement côté client signifie simplement que vous avez chiffré les données avant qu'elles ne soient chargées et que vous devez les déchiffrer vous-même avant qu'elles ne soient utilisées.

Cloud Storage continue de procéder au chiffrement GMEK, CMEK ou CSEK sur l'objet. Il ignore la présence de la couche de chiffrement supplémentaire que vous avez peut-être ajoutée.

Cloud Storage prend en charge la journalisation de l'accès aux données et ces journaux sont immuables.

Outre les journaux d'audit Cloud et les journaux d'accès Cloud Storage, vous pouvez placer des préservations à titre conservatoire et des verrous sur les données elles-mêmes.

À des fins d'audit, vous pouvez placer une préservation à titre conservatoire sur un objet, et toutes les opérations susceptibles de modifier ou de supprimer l'objet sont suspendues jusqu'à ce que la préservation soit levée. Vous pouvez également verrouiller un bucket et aucune modification ou suppression ne peut survenir tant que le verrou n'est pas retiré. Enfin, il reste l'option de la règle de conservation verrouillée mentionnée précédemment. Elle reste en vigueur et empêche la suppression, qu'un verrou de bucket ou une préservation d'objet soient en vigueur ou non.

Le verrouillage des données et le chiffrement sont deux choses différentes. Le chiffrement empêche de lire les données, tandis que le verrouillage empêche de les modifier.

Cloud Storage prend en charge toute une série de cas d'utilisation spécifiques. Par exemple, le codage par décompression.

Par défaut, les données que vous chargez et celles que vous récupérez de Cloud Storage sont les mêmes. Cela comprend les archives gzip, qui sont généralement renvoyées au même format. Cependant, si vous ajoutez un libellé adéquat à un objet dans les métadonnées, Cloud Storage peut décompresser le fichier à mesure qu'il est diffusé. La réduction du temps de chargement et des coûts de stockage est un avantage des fichiers compressés de plus petite taille par rapport aux fichiers non compressés.

Vous pouvez définir un bucket en tant que paiements du demandeur. Normalement, si les données sont accessibles à partir d'une autre région, vous devrez payer des frais de sortie réseau. Par contre, il est possible de faire payer le demandeur, de sorte que vous ne payez que pour le stockage des données.

Vous pouvez créer une URL signée pour partager anonymement un objet dans Cloud Storage, et même faire en sorte que l'URL expire après un certain temps.

Il est possible de charger un objet en plusieurs éléments et de créer un objet composite sans avoir à concaténer les éléments après le chargement.

Cloud Storage offre de nombreuses fonctionnalités utiles, mais nous devons passer à la suite.

<https://pixabay.com/fr/vectors/cadenas-de-s%C3%A9curit%C3%A9-verrouillage-308589/>

<https://pixabay.com/fr/illustrations/bo%C3%A9te-en-carton-bo%C3%A9te-en-carton-1536798/>

Vidéo : 2_l6
Stocker des données de tous types
Présentateur : Evan

[FIN DE LA VIDÉO]

Programme

Introduction aux data lakes

Stockage de données et options ETL dans GCP

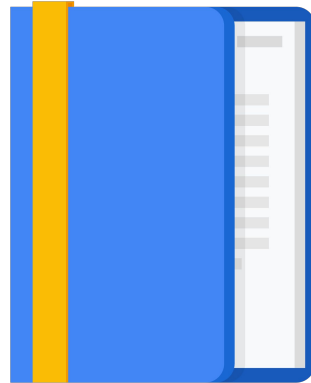
Créer un data lake avec Cloud Storage

Sécuriser Cloud Storage

Stocker des données de tous types

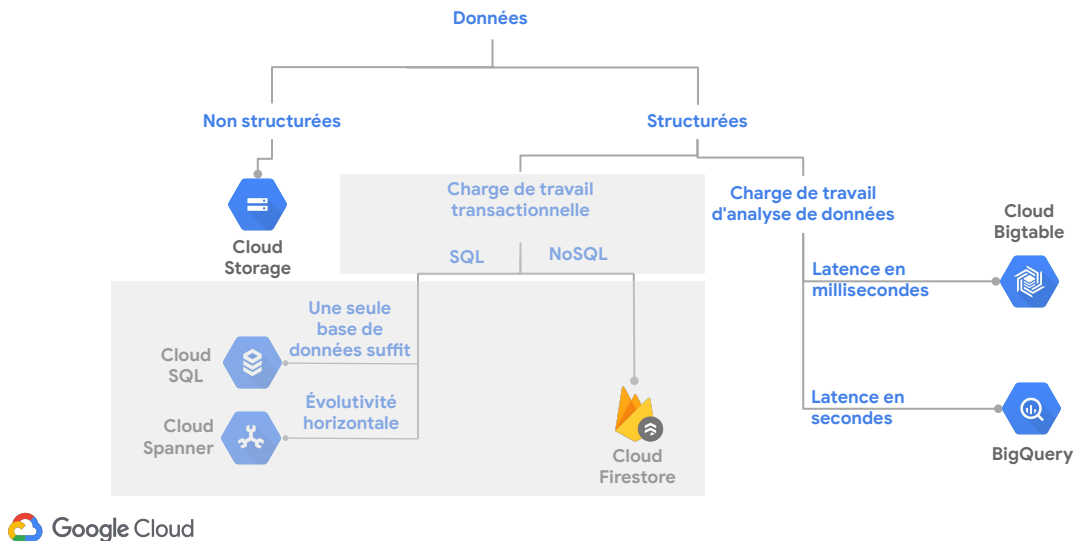
Cloud SQL en tant que data lake relationnel

Atelier : Charger les données des taxis dans
Cloud SQL



Comme indiqué précédemment, Cloud Storage n'est pas votre seul choix en matière de stockage de données sur Google Cloud.

Différentes considérations concernant les charges de travail transactionnelles



Il n'est pas conseillé d'utiliser Cloud Storage pour les charges de travail transactionnelles. Bien que la latence de Cloud Storage soit faible, elle n'est pas suffisamment faible pour prendre en charge les écritures à haute fréquence. Pour les charges de travail transactionnelles, utilisez Cloud SQL ou Firestore respectivement selon que vous souhaitez utiliser SQL ou NoSQL.

Il n'est également pas recommandé d'utiliser Cloud Storage pour l'analyse de données structurées. Autrement, vous risquez de consacrer une quantité importante de données à l'analyse des calculs. Il est préférable d'utiliser Cloud Bigtable ou BigQuery pour les charges de travail d'analyse des données structurées, en fonction de la latence requise.

Transactions ou analyses

	Transactions	Analyses
Source de données	Données opérationnelles. Les OLTP sont la source d'origine des données.	Données de consolidation. Les données OLAP proviennent des différentes bases de données OLTP.
Objectif des données	Contrôler et exécuter les tâches de gestion fondamentales	Simplifier la planification, la résolution des problèmes et la prise de décision
Informations fournies par les données	Affiche un instantané des processus métier en cours	Affichages multidimensionnels de différents types d'activités commerciales
Insertions et mises à jour	Insertions et mises à jour courtes et rapides réalisées par les utilisateurs finaux	Des tâches par lots périodiques et de longue durée actualisent les données
Requêtes	Des requêtes relativement standardisées et simples renvoyant très peu d'enregistrements	Des requêtes souvent complexes comprenant] des agrégations
Vitesse de traitement	Généralement très rapide	Dépend de la quantité de données concernées. Amélioration de la vitesse des requêtes au moyen d'index.
Espace requis	Peut être assez faible si les données historiques sont archivées	Plus grand, plus d'index qu'OLTP



Nous n'arrêtons pas de parler des charges de travail transactionnelles par opposition aux charges de travail d'analyse.

Qu'est-ce que cela signifie exactement ?

Les charges de travail transactionnelles sont celles pour lesquelles des insertions et des mises à jour rapides sont nécessaires.

Il est recommandé de conserver un instantané de l'état actuel du système.

En contrepartie, les requêtes sont généralement assez simples et ne concernent que quelques enregistrements.

Par exemple, dans un système bancaire, verser votre salaire sur un compte est une transaction. Elle met à jour le solde.

La banque procède au traitement transactionnel en ligne (OLTP, online transaction processing).

Transactions ou analyses

	Transactions	Analyses
Source de données	Données opérationnelles. Les OLTP sont la source d'origine des données.	Données de consolidation. Les données OLAP proviennent des différentes bases de données OLTP.
Objectif des données	Contrôler et exécuter les tâches de gestion fondamentales	Simplifier la planification, la résolution des problèmes et la prise de décision
Informations fournies par les données	Affiche un instantané des processus métier en cours	Affichages multidimensionnels de différents types d'activités commerciales
Insertions et mises à jour	Insertions et mises à jour courtes et rapides réalisées par les utilisateurs finaux	Des tâches par lots périodiques et de longue durée actualisent les données
Requêtes	Des requêtes relativement standardisées et simples renvoyant très peu d'enregistrements	Des requêtes souvent complexes comprenant] des agrégations
Vitesse de traitement	Généralement très rapide	Dépend de la quantité de données concernées. Amélioration de la vitesse des requêtes au moyen d'index.
Espace requis	Peut être assez faible si les données historiques sont archivées	Plus grand, plus d'index qu'OLTP



Une charge de travail d'analyse, d'autre part, lit généralement l'intégralité de l'ensemble de données et est souvent utilisée pour la planification ou la prise de décision.

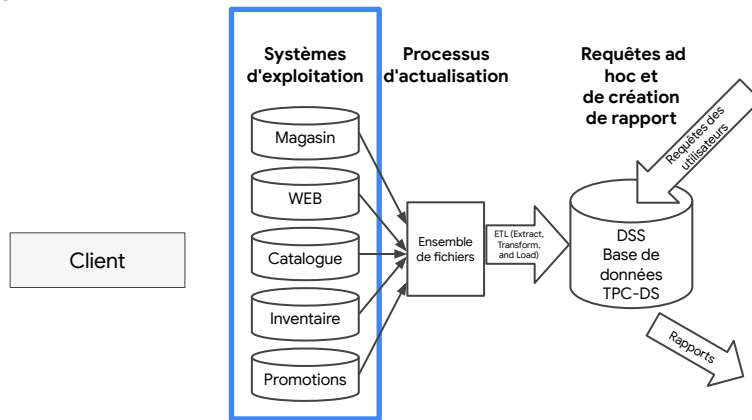
Les données peuvent provenir d'un système de traitement des transactions, mais elles sont souvent consolidées à partir de nombreux systèmes OLTP.

Par exemple, un régulateur bancaire peut nous demander de fournir un rapport sur chaque client ayant transféré plus de 10 000 \$ sur un compte à l'étranger. Il peut demander à la banque d'inclure les clients qui tentent de transférer les 10 000 \$ en plusieurs petits versements pendant une semaine.

Un rapport comme celui-ci implique la consultation d'un ensemble de données très important et nécessite une requête complexe comportant une agrégation sur des fenêtres temporelles mobiles.

Il s'agit d'un exemple de traitement analytique en ligne (OLAP, online analytical processing).

Les opérations des systèmes transactionnels se composent à 80 % d'écriture et à 20 % de lecture*



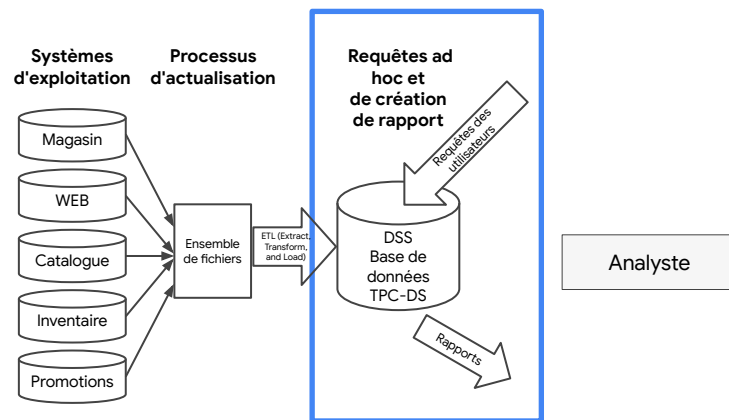
Nous traitons ces cas d'utilisation différemment à cause de l'écriture lourde des systèmes transactionnels.

Il s'agit généralement de systèmes opérationnels. Par exemple, les données du catalogue d'un commerçant doivent être mises à jour chaque fois qu'il ajoute un nouvel article ou modifie le prix.

Les données d'inventaire doivent être mises à jour chaque fois que le commerçant vend un article.

En effet, les systèmes de catalogue et d'inventaire doivent conserver un instantané actualisé de l'activité.

Les opérations des systèmes d'analyse se composent à 20 % d'écriture et à 80 % de lecture*

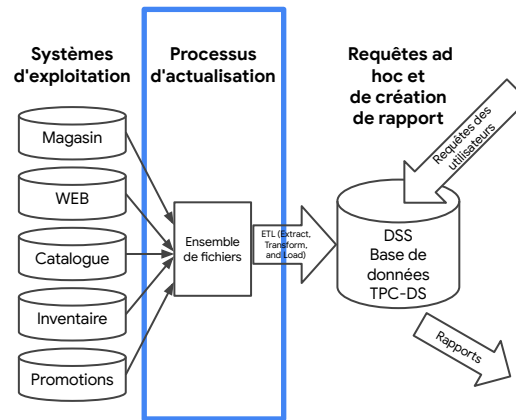


Les systèmes dédiés à l'analyse peuvent être renseignés régulièrement à partir des systèmes opérationnels.

Nous pouvons nous en servir une fois par jour pour générer un rapport sur les articles de notre catalogue dont les ventes augmentent, mais dont le stock est faible. Ce rapport devra lire un certain nombre de données, sans pour autant beaucoup écrire.

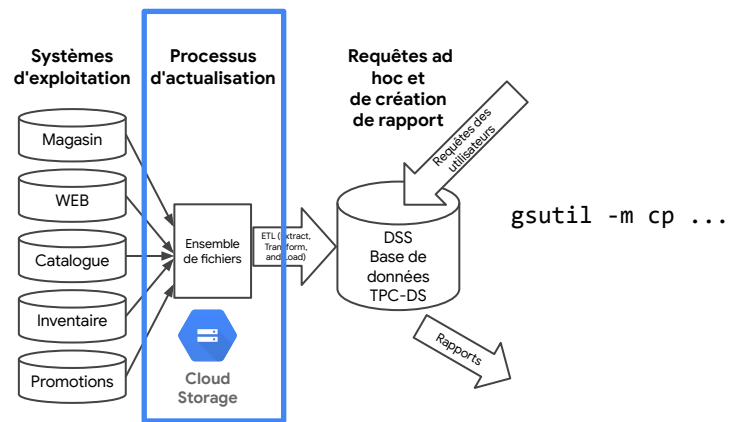
Les systèmes OLAP sont axés sur la lecture.

Les ingénieurs données créent les pipelines entre les systèmes



Souvenez-vous de ce que nous avons dit. Les systèmes dédiés à l'analyse peuvent être renseignés régulièrement à partir des systèmes opérationnels. Les ingénieurs données créent les pipelines pour renseigner le système OLAP à partir du système OLTP. Exporter la base de données sous forme de fichier et la charger dans le data warehouse peut être une solution simple. Nous appelons cela EL.

Utiliser Cloud Storage pour la préproduction évolutive des données brutes



Sur Google Cloud, BigQuery est généralement le data warehouse. Le volume de données que vous pouvez directement charger dans BigQuery est limité. Cela est dû au fait que votre réseau peut être un goulot d'étranglement.

Commencer par charger les données dans Cloud Storage pour ensuite les transférer de Cloud Storage vers BigQuery peut s'avérer beaucoup plus pratique que de les charger directement dans BigQuery.

Cela est dû au fait que Cloud Storage prend en charge des charges multithread avec reprise. Il suffit de fournir l'option `-m` à `gsutil`. Le chargement à partir de Cloud Storage est également plus rapide grâce au débit élevé qu'il offre.

Interroger des données dans BigQuery directement à partir de GCS

simple

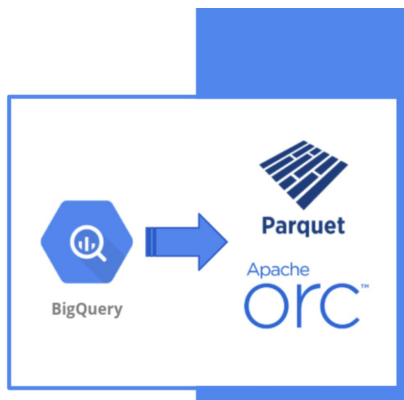
Créez des tables BigQuery avec la technologie de Parquet/ORC

rapide

Format de fichier en colonnes et partitions logiques

pratique

Les partitions Hive gèrent les requêtes et le chargement



La possibilité pour BigQuery d'interroger directement les fichiers de données se trouvant dans Google Cloud Storage sans avoir à les charger au préalable dans le stockage natif de BigQuery est une fonctionnalité qui a été récemment mise en place. Cette opération est connue sous le nom de requête fédérée ou de connexion à une source de données externe.

Les formats de fichiers CSV et Avro sont très utilisés et les formats Parquet et Apache ORC sont désormais également pris en charge. Voici une démonstration rapide du fonctionnement direct entre GCS et BigQuery avec le SQL fédéré.

Vidéo : 2_17

Démonstration : Exécution de requêtes fédérées
sur des fichiers Parquet et ORC dans BigQuery

Présentateur : Evan

[FIN DE LA VIDÉO]

Démonstration

Exécution de requêtes fédérées
sur des fichiers Parquet et ORC
dans BigQuery

[DÉMONSTRATION]

Utilisez la démonstration existante du 31/10/2019 :

<https://www.youtube.com/watch?v=I5l0knEcH4I&feature=youtu.be>

<https://cloud.google.com/blog/products/data-analytics/keep-parquet-and-orc-from-the-data-graveyard-with-new-bigquery-features>

L'équipe de BigQuery Fundamentals est ravie d'annoncer le lancement de la version bêta publique des formats fédérés sur GCS. Vous pouvez désormais interroger les formats Parquet et ORC en colonnes, mais aussi interroger et charger les tables partitionnées Hive dans GCS directement à partir de BigQuery.

Démonstration de : **Pour en savoir plus, consultez [l'article de blog intéressant de Tino Tereshko](#) ou [cette courte vidéo de démonstration](#) de Michelle Goodstein.**
Pour tester la version bêta, consultez la documentation publique suivante :
[Interroger des données Cloud Storage](#), [Interroger des données partitionnées en externe](#) et [Charger des données partitionnées externes](#).

[DÉMONSTRATION]

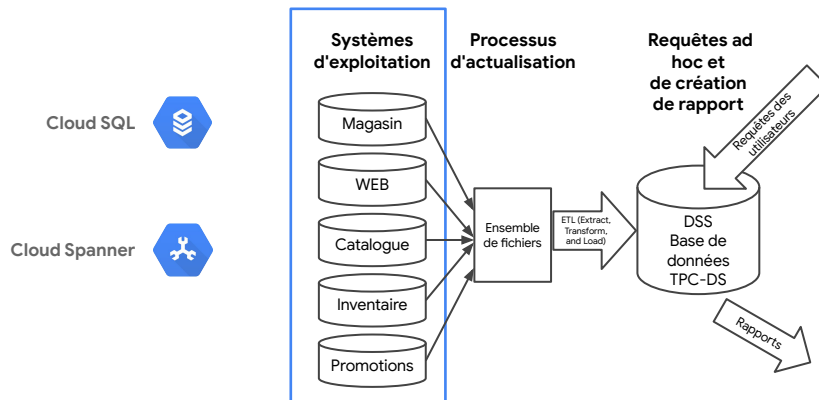
Vidéo : 2_l8

Stocker des données relationnelles dans le cloud

Présentateur : Evan

[FIN DE LA VIDÉO]

Choisir parmi les bases de données relationnelles cloud pour les charges de travail transactionnelles



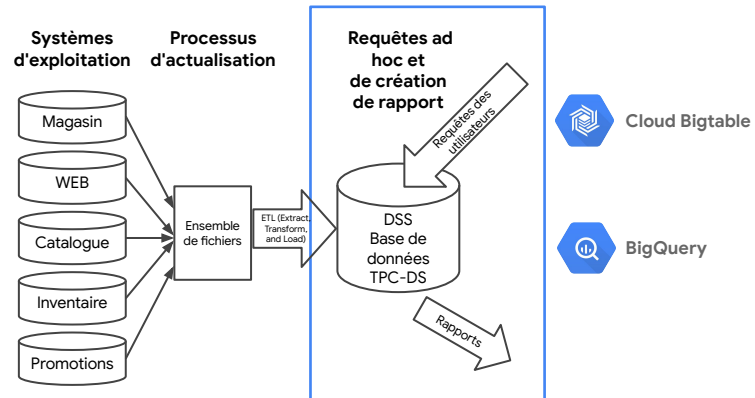
Revenons à la discussion sur les charges de travail transactionnelles. Il existe plusieurs options pour les bases de données relationnelles. Dans ce cas, Cloud SQL est le choix par défaut, mais si vous avez besoin d'une base de données distribuée dans le monde entier, utilisez Cloud Spanner.

Une base de données distribuée dans le monde entier est recommandée si votre base de données est mise à jour à partir d'applications fonctionnant dans différentes régions géographiques. La fonctionnalité TrueTime de Spanner est très intéressante pour ce type de cas d'utilisation.

Vous pourriez également choisir Spanner si votre base de données est trop volumineuse pour être contenue dans une seule instance de Cloud SQL. Si la taille de la base de données atteint plusieurs gigaoctets, vous avez besoin d'une base de données distribuée. L'évolutivité de Spanner est très intéressante pour ce type de cas d'utilisation.

Sinon, utilisez Cloud SQL, car il s'agit d'une solution plus rentable.

Choisir parmi les entrepôts de données cloud pour les charges de travail d'analyse



BigQuery est le choix par défaut pour les charges de travail d'analyse.

Cependant, si vous avez besoin d'insertions à haut débit, de plus que plusieurs millions de lignes par seconde ou d'une faible latence, de l'ordre de quelques millisecondes, utilisez Cloud Bigtable.

Sinon, utilisez BigQuery, car il s'agit d'une solution plus rentable.

Vidéo : 2_I9

Cloud SQL en tant que data lake relationnel

Présentateur : Evan

[FIN DE LA VIDÉO]

Programme

Introduction aux data lakes

Stockage de données et options ETL dans GCP

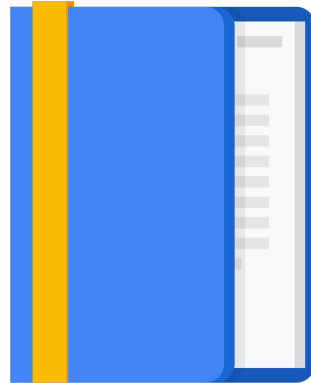
Créer un data lake avec Cloud Storage

Sécuriser Cloud Storage

Stocker des données de tous types

Cloud SQL en tant que data lake relationnel

Atelier : Charger les données des taxis dans
Cloud SQL



Nous avons expliqué que Cloud SQL est le choix par défaut pour les charges de travail OLTP (traitement transactionnel en ligne) sur Google Cloud. Examinons rapidement cela.

Cloud SQL est un service de base de données entièrement géré qui simplifie la configuration et l'administration de vos bases de données relationnelles MySQL et PostgreSQL dans le cloud



Cloud SQL est un service facile à utiliser offrant des bases de données relationnelles entièrement gérées. Cloud SQL vous permet de confier à Google les tâches banales, mais nécessaires et souvent longues, comme l'application de correctifs et de mises à jour, la gestion des sauvegardes et la configuration des répliquions, afin que vous puissiez vous consacrer à la création d'applications de pointe.

Cloud SQL est notre service géré pour les SGBDR tiers.

Il est compatible avec MySQL. Nous avons récemment ajouté Postgres à la disponibilité générale et annoncé la compatibilité prochaine avec Microsoft SQL Server.

Nous ajouterons d'autres SGBDR au fil du temps.

Cela signifie que nous fournissons une instance Compute Engine dans laquelle MySQL est déjà installé. Nous gérons l'instance en votre nom.

Nous nous chargerons des sauvegardes, des mises à jour de sécurité et de la mise à jour des versions mineures du logiciel afin que vous n'ayez pas à vous en soucier.

En d'autres termes, Google Cloud gère la base de données MySQL de sorte que vous puissiez la traiter comme un service.

Nous nous chargeons même des opérations de type DBA. Vous pouvez nous demander d'ajouter une instance dupliquée de basculement pour votre base de données. Nous la gérons à votre place et vous bénéficierez d'une garantie de disponibilité de 99,95 % dans le contrat de niveau de service.

Cloud SQL peut être utilisé avec d'autres services GCP



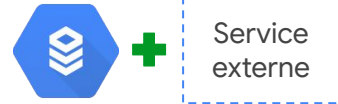
Cloud SQL est compatible avec App Engine grâce à des pilotes standard.

Vous pouvez configurer une instance Cloud SQL pour suivre une application App Engine.



Utilisez une adresse IP externe pour autoriser les instances Compute Engine à accéder aux instances Cloud SQL.

Vous pouvez configurer une zone préférée pour les instances Cloud SQL.



Vous pouvez utiliser Cloud SQL avec des applications et des clients externes.

Vous pouvez utiliser des outils standard pour administrer les bases de données.

Vous pouvez configurer des instances dupliquées externes avec accès en lecture.



Les instances Cloud SQL présentent un autre avantage : elles sont accessibles par d'autres services GCP et même par des services externes. Vous pouvez utiliser Cloud SQL avec App Engine à l'aide de pilotes standard comme Connector/J pour Java ou MySQLdb pour Python.

Vous pouvez autoriser les instances Compute Engine à accéder aux instances Cloud SQL et configurer l'instance Cloud SQL pour qu'elle se trouve dans la même zone que votre machine virtuelle.

Cloud SQL fonctionne également avec d'autres applications et outils que vous pouvez avoir l'habitude d'utiliser comme SQL Workbench, Toad et d'autres applications externes utilisant des pilotes MySQL standard.

Les opérations de sauvegarde, de récupération, de scaling et de sécurité sont gérées à votre place

- Sécurité Google
- Sauvegardes gérées
- Scaling vertical (lecture et écriture)
- Scaling horizontal (lecture)
- Réplication automatique



La gestion de votre base de données par Google vous permet notamment de bénéficier des avantages de la **sécurité de Google**. Les données client Cloud SQL sont chiffrées lorsqu'elles se trouvent sur les réseaux internes de Google et lorsqu'elles sont stockées dans des tables de base de données, des fichiers temporaires et des sauvegardes.

Chaque instance Cloud SQL comprend un pare-feu de réseau, vous permettant de contrôler l'accès réseau à votre instance de base de données en accordant l'accès.

Cloud SQL est facile à utiliser : aucune installation de logiciel ni maintenance n'est nécessaire. De plus, **Google gère les sauvegardes**.

Cloud SQL prend en charge le stockage sécurisé de vos données sauvegardées et vous permet d'effectuer facilement une restauration à partir d'une sauvegarde, ainsi qu'une récupération à un moment précis d'un état spécifique d'une instance.

Cloud SQL conserve jusqu'à 7 sauvegardes pour chaque instance. Cela est inclus dans le coût de votre instance.

Vous pouvez **réaliser un scaling vertical de Cloud SQL**. Il suffit d'augmenter la taille de votre machine.

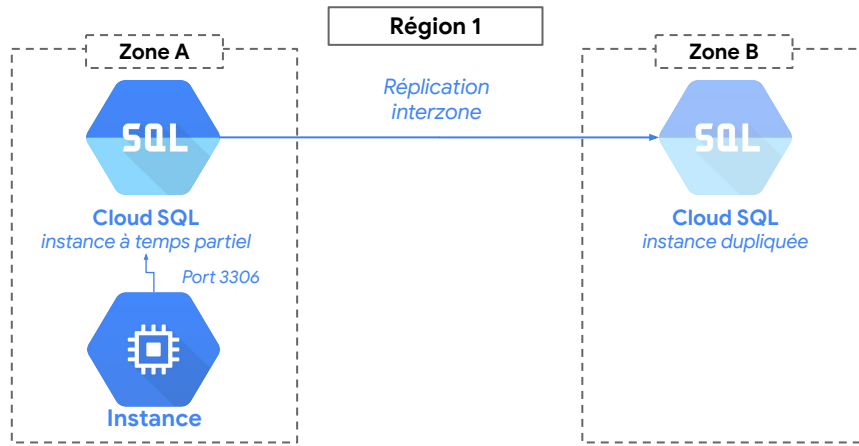
Faites évoluer votre solution en optant pour des processeurs allant jusqu'à 64 cœurs et une RAM de plus de 100 Go.

Horizontalement, vous pouvez rapidement effectuer un scaling horizontal avec des instances dupliquées en lecture. Google Cloud SQL est compatible avec trois scénarios d'instances dupliquées en lecture :

- Des instances Cloud SQL effectuant une réplication depuis une instance Cloud SQL maître.
Les instances dupliquées sont d'autres instances dans le même projet et le même emplacement que l'instance maître.
- Des instances Cloud SQL effectuant une réplication depuis une instance maître externe.
L'instance maître est externe à Google Cloud SQL. Par exemple, elle peut se trouver en dehors du réseau Google ou dans une instance Google Compute Engine. Vous pouvez vous en servir pour sauvegarder une base de données MySQL sur site.
- Des instances MySQL externes effectuant une réplication depuis une instance Cloud SQL maître.
Les instances dupliquées externes se trouvent dans des environnements d'hébergement, en dehors de Cloud SQL.

Si vous avez besoin d'un scaling horizontal en lecture/écriture, utilisez Cloud Spanner.

Réplication Cloud SQL



Cloud SQL prend en charge le cas particulier du basculement.

Les instances Cloud SQL peuvent être configurées avec une instance dupliquée de basculement dans une zone différente de la même région. Ensuite, les données de Cloud SQL sont répliquées dans les zones d'une région pour assurer la durabilité. Dans le cas peu probable d'une interruption du centre de données, une instance Cloud SQL devient automatiquement disponible dans une autre zone. Toutes les modifications apportées aux données sur l'instance maître sont répliquées sur le basculement.

En cas d'interruption de la zone de l'instance maître, Cloud SQL bascule automatiquement sur l'instance dupliquée. Le basculement n'a pas lieu si l'instance maître rencontre des problèmes qui ne sont pas liés à une interruption de la zone. Toutefois, vous pouvez lancer le basculement manuellement.

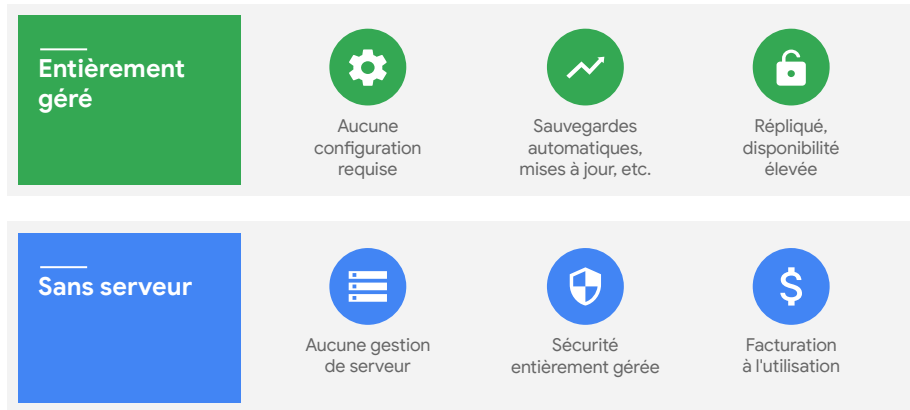
Quelques mises en garde :

1. Notez que l'instance dupliquée de basculement est facturée comme une instance distincte.
2. Lorsqu'une interruption de zone se produit et que l'instance maître bascule sur l'instance dupliquée de basculement, toutes les connexions existantes à l'instance sont interrompues. Toutefois, l'application peut se reconnecter à l'aide de la même chaîne de connexion ou adresse IP. Il est inutile de mettre à jour l'application après un basculement.
3. Après le basculement, l'instance dupliquée devient l'instance maître et Cloud SQL crée automatiquement une nouvelle instance dupliquée de

1. basculement dans une autre zone. Si vous avez localisé l'instance Cloud SQL à proximité d'autres ressources, telles qu'une instance Compute Engine, vous pouvez à nouveau déplacer l'instance Cloud SQL dans sa zone d'origine dès que celle-ci est disponible. Sinon, il n'est pas nécessaire de déplacer l'instance après un basculement.

Vous pouvez utiliser l'instance dupliquée de basculement comme une instance dupliquée en lecture pour décharger les opérations de lecture de l'instance maître. Pour plus d'informations sur les instances dupliquées de basculement, consultez <https://cloud.google.com/sql/docs/mysql/high-availability?hl=fr>

Entièrement géré ou sans serveur



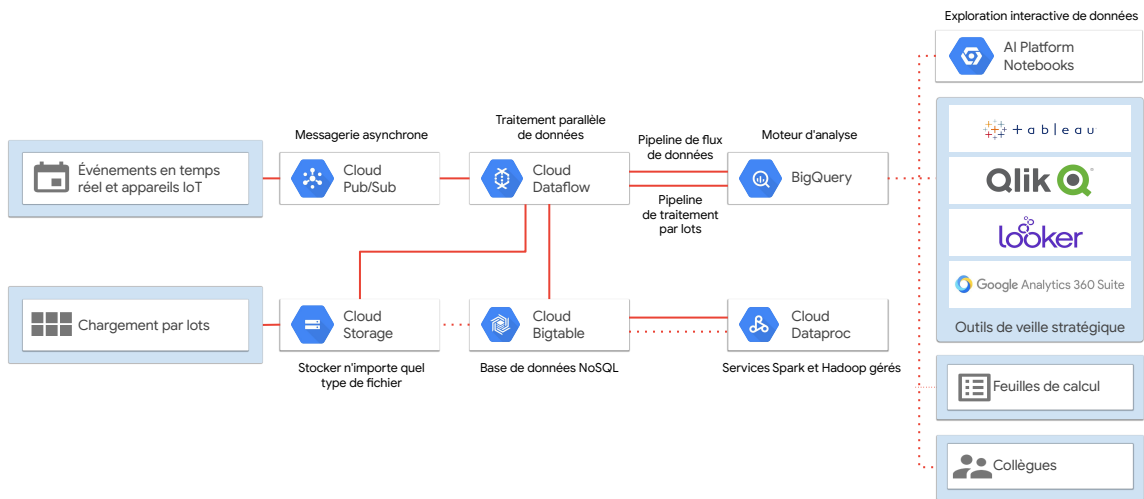
Nous n'avons cessé de répéter que Cloud SQL est entièrement géré. Nous avons également utilisé le terme sans serveur pour décrire BigQuery, par exemple. Quelle est la différence ?

Par **entièrement géré**, nous entendons que le service fonctionne sur du matériel que vous pouvez contrôler. Vous pouvez par exemple vous connecter en SSH à une instance Cloud SQL.

Cela dit, Google vous aide à gérer l'instance. Notamment en automatisant les sauvegardes et en configurant des instances de basculement, etc.

Sans serveur est la prochaine étape. Vous pouvez comparer un produit sans serveur à une simple API que vous appelez. Vous payez pour utiliser le produit, mais vous ne devez gérer aucun serveur.

Architecture moderne de gestion des données sans serveur



Google Cloud

BigQuery ne dispose d'aucun serveur. Tout comme Cloud Pub/Sub pour la messagerie asynchrone et Cloud Dataflow pour le traitement parallèle de données. Vous pouvez aussi considérer Cloud Storage comme une solution sans serveur. Bien sûr, Cloud Storage utilise des disques, mais vous n'interagissez jamais avec le matériel. L'une des particularités de Google Cloud est de pouvoir créer un pipeline de traitement de données constitué de composants bien conçus, tous entièrement dépourvus de serveur.

En revanche, Dataproc est entièrement géré. Il vous permet de gérer les charges de travail de Spark et Hadoop sans avoir à vous soucier de la configuration.

Si vous avez le choix entre un tout nouveau projet sur BigQuery ou Dataflow, qui ne disposent d'aucun serveur, et Dataproc, qui est entièrement géré, lequel devriez-vous choisir ?

Si toutes les autres caractéristiques sont équivalentes, choisissez le produit sans serveur.

Vidéo : 2_I10

Atelier : Charger les données des taxis dans Cloud SQL

Présentateur : Evan

[FIN DE LA VIDÉO]

Programme

Introduction aux data lakes

Stockage de données et options ETL dans GCP

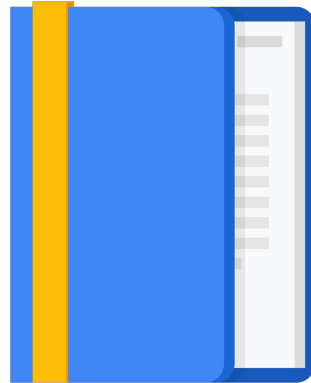
Créer un data lake avec Cloud Storage

Sécuriser Cloud Storage

Stocker des données de tous types

Cloud SQL en tant que data lake relationnel

Atelier : Charger les données des taxis dans
Cloud SQL





Charger les données des taxis dans Cloud SQL

Objectifs

- Créer une instance Cloud SQL
- Créer une base de données Cloud SQL
- Importer des données textuelles dans Cloud SQL
- Vérifier l'intégrité des données

Dans cet atelier, vous vous entraînerez à créer un data lake pour vos données relationnelles avec Cloud SQL. Vous commencez par :

créer une instance Cloud SQL pouvant contenir plusieurs bases de données ;
créer une nouvelle base de données Cloud SQL ;
Importer des données textuelles dans Cloud SQL
et terminez par vérifier l'intégrité des données.