



Présentation de la construction de pipelines de données en lots

Que sont les pipelines de traitement par lots ?

Il s'agit de pipelines qui traitent une quantité *limitée* de données et qui arrêtent ensuite leur exécution.

Il peut s'agir, par exemple, d'un pipeline de traitement par lots qui fonctionne une fois par jour. Il gère toutes les opérations de crédit, de débit et de transfert d'argent de cette journée, équilibre les livres et enregistre les données rapprochées dans le data warehouse.

Programme

EL, ELT, ETL

Considérations sur la qualité

Comment exécuter des opérations
dans BigQuery

Inconvénients

ETL pour résoudre les problèmes
liés à la qualité des données



Si vous devez écrire un tel pipeline, pour équilibrer les livres, devez-vous utiliser EL, ELT ou ETL ?

Rappelez-vous que EL extrait et charge.

ELT charge les données telles quelles, puis les transforme à la volée.

ETL extrait les données, les transforme, puis les charge dans un e-data warehouse.

Nous allons voir quand utiliser chacun d'eux.

Comment cela dépend des types de transformations dont vous avez besoin, et comment celles-ci dépendent à leur tour de considérations de qualité.

Nous examinerons comment construire des pipelines EL et ELT dans BigQuery, les exemples où EL et ELT ne suffisent pas et pourquoi vous pourriez vouloir ETL.

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



Extraire et charger



Extraire, charger, transformer



Extraire, transformer, charger



EL signifie extraire et charger. Cela fait référence au moment où 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.

L'idée consiste à regrouper toutes les transactions, à les stocker dans une table, sans qu'il soit nécessaire de les transformer.

Cela fonctionnera-t-il dans le cas d'un rapprochement des livres tous les soirs ? Le risque est de vous retrouver avec des transactions non réconciliées qui traînent dans le data warehouse.

Vos rapports d'analyses pourraient alors être tout à fait erronés et déclencher de nombreuses alarmes.

L'ELT permet le chargement direct des données brutes dans la cible et leur transformation chaque fois que cela est nécessaire. Par exemple, vous pouvez donner accès aux données brutes par le biais d'une vue qui détermine si l'utilisateur veut toutes les transactions ou seulement celles qui sont rapprochées.

Cette vue va vous donner beaucoup de travail si vous la choisissez. Lorsque l'ampleur de transformation nécessaire est importante, vous pouvez faire appel à des machines lourdes. C'est ETL.

Extraire, transformer, charger (ETL) forme un processus d'intégration de données dans lequel la transformation a lieu dans un service intermédiaire avant d'être chargée dans la cible. Par exemple, les données peuvent être transformées dans Cloud Dataflow avant d'être chargées dans BigQuery.

ETL est l'outil le plus approprié dans ce cas. Nous allons extraire toutes les données des transactions et effectuer le traitement pour les rapprocher, puis enregistrer les transactions ainsi rapprochées dans le data warehouse, en laissant les transactions non rapprochées pour la prochaine exécution du travail par lots. Bien entendu, si une transaction n'a pas été rapprochée depuis environ 15 jours, il est possible d'intervenir...

<https://pixabay.com/fr/illustrations/warehouse-shipping-box-business-3688280/>

<https://pixabay.com/fr/vectors/moving-box-relocation-people-new-312082/>

<https://pixabay.com/fr/vectors/box-car-forklift-loader-vehicle-159302/>

Quand utiliser EL ?

Architecture	Quand exécuter la tâche
Extraire les données des fichiers sur Google Cloud Storage Les charger dans le stockage natif de BigQuery Vous pouvez déclencher cette action à partir de Cloud Composer, Cloud Functions ou de requêtes programmées	Chargement par lots de données historiques Chargements périodiques programmés des fichiers journaux (par exemple, une fois par jour) Mais uniquement si les données sont déjà propres et correctes !



Quand utiliser EL ?

Résumé : vous devriez utiliser EL uniquement si les données sont déjà propres et correctes.

Vous disposez peut-être de fichiers journaux dans Google Cloud Storage.

Vous pouvez donc

- Extraire les données des fichiers sur Google Cloud Storage
- Les charger dans le stockage natif de BigQuery

Il suffit simplement d'appeler l'API REST.

Vous pouvez déclencher ce pipeline à partir de Cloud Composer, Cloud Functions ou de requêtes programmées

Vous pouvez même le configurer pour qu'il fonctionne en micro lots - pas tout à fait en flux de données, mais presque en temps réel : chaque fois qu'un nouveau fichier arrive sur Cloud Storage, la fonction cloud s'exécute, et la fonction invoque une tâche Bigquery.

Le service de transfert de données de BigQuery fonctionne également dans ce cas.

Utilisez EL pour le chargement par lots de données historiques, ou exécutez des chargements programmés de fichiers journaux.

Mais j'insiste : vous devriez utiliser EL uniquement si les données sont déjà propres et

correctes.

Quand utiliser EL ?

Architecture	Quand exécuter la tâche
Extraire les données des fichiers dans Google Cloud Storage vers BigQuery Transformer les données à la volée via les vues BigQuery ou les stocker des de nouvelles tables	Des ensembles de données expérimentales pour lesquels vous ne savez pas encore quels types de transformations sont nécessaires pour rendre les données utilisables. Tout ensemble de données de production où la transformation peut être exprimée en SQL



ELT commence par EL.

Ainsi, le chargement est le même et pourrait fonctionner de la même manière.

Le fichier arrive sur le Cloud Storage, la fonction invoque le chargement de BigQuery, avec la table ajoutée.

La grande différence réside dans les étapes qui suivent.

La table peut être stockée dans un ensemble de données privées.

Et tout le monde accède aux données par une vue qui impose des contrôles d'intégrité des données.

Ou vous disposez peut-être d'une tâche qui exécute une requête SQL avec une table de destination.

De cette façon, les données transformées sont stockées dans une table à laquelle tout le monde accède.

Quand utiliser ELT ?

Un cas fréquent est celui où vous ne savez pas quels types de transformations sont nécessaires pour rendre les données utilisables.

Supposons, par exemple, que quelqu'un télécharge une nouvelle image. Vous invoquez l'API Cloud Vision et vous recevez en retour un long message JSON sur toutes sortes d'éléments de l'image. Le texte dans l'image. La présence ou non d'un point de repère. Un logo. Quels objets.

Ce dont un analyste aura besoin à l'avenir. Vous ne savez pas. Vous stockez donc le JSON brut tel quel. Plus tard, si quelqu'un veut compter le nombre de fois que les

logos d'une entreprise spécifique se trouvent dans cet ensemble d'images, il peut extraire les logos de la JSON et les compter ensuite.

Bien sûr, cela ne fonctionne que si la transformation nécessaire peut être exprimée en SQL

Dans le cas de l'API Vision, le résultat est JSON, et BigQuery SQL prend en charge l'analyse JSON. Ainsi, ELT fonctionnera dans ce cas.

Programme

EL, ELT, ETL

Considérations sur la qualité

Comment exécuter des opérations dans BigQuery

Inconvénients

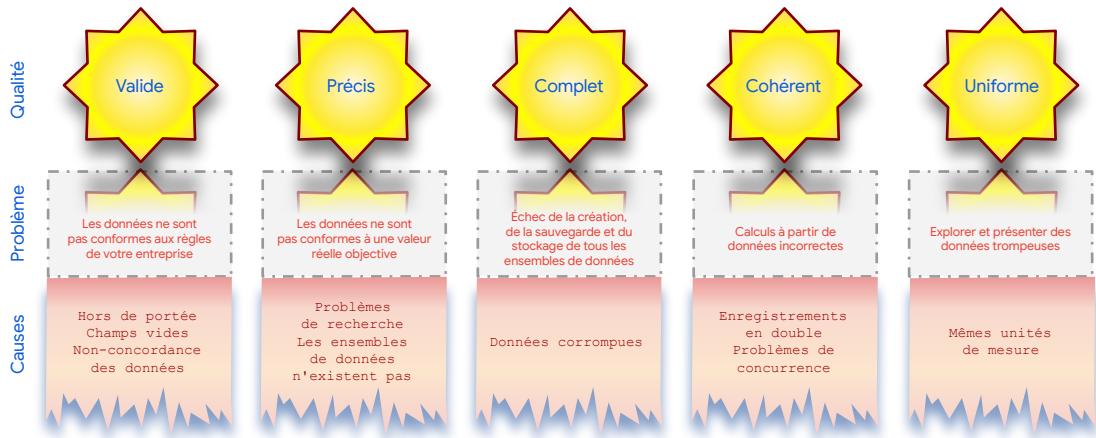
ETL pour résoudre les problèmes liés à la qualité des données



Maintenant que nous avons examiné EL et ELT, examinons certaines des transformations que vous pourriez vouloir effectuer, et comment elles peuvent être réalisées dans BigQuery.

Pour que les choses restent traçables, supposons que nos besoins en matière de traitement des données tournent tous autour de l'amélioration de la qualité.

Quels sont les objectifs du traitement de la qualité des données ?



Quelles sont certaines des raisons liées à la qualité pour lesquelles nous pourrions vouloir traiter des données ?

La ligne supérieure représente les caractéristiques de l'information -- l'information peut être valide, précise, complète, cohérente et/ou uniforme. Ces termes sont définis dans la science de la logique. Chacun est indépendant. Par exemple, les données peuvent être complètes sans être cohérentes. Elles peuvent être valides sans être uniformes. Il existe des définitions formelles pour chacun de ces termes que vous pouvez consulter en ligne. Mais la principale raison pratique de les rechercher est indiquée dans la deuxième ligne : les problèmes qu'ils posent dans l'analyse des données. C'est une chose de rechercher chacun des cinq badges pour vos données, d'avoir objectivement une bonne qualité de données. Cependant, c'en est une autre lorsque des données de mauvaise qualité interfèrent avec l'analyse des données et conduisent à des décisions commerciales incorrectes. La raison pour laquelle il faut consacrer du temps, de l'énergie et des ressources à la détection et à la résolution des problèmes de qualité est donc que cela peut avoir une incidence sur le résultat d'une entreprise.

Ainsi, si les données ne sont pas conformes à vos règles commerciales, vous avez un problème de validité. Supposons, par exemple, que vous vendiez des billets de cinéma, et que chaque billet coûte 10 \$. Si vous avez une transaction de 7 \$, alors vous avez un problème de validité.

De même, les problèmes d'exactitude sont dus au fait que les données ne sont pas conformes à la vérité objective. L'exhaustivité est liée au fait de ne pas tout traiter.

Les problèmes de cohérence se posent si deux opérations différentes qui devraient être identiques produisent des résultats différents, et parce que vous ne savez pas à quoi vous fier, vous ne pouvez pas tirer d'informations de ces données. L'uniformité, c'est que les valeurs des données d'une même colonne dans différentes lignes signifient des choses différentes.

Les principales causes de ces problèmes sont énumérées à la troisième ligne. Je vais marquer une petite pause pour vous donner le temps de les lire. Dans les prochaines diapositives, nous allons explorer les méthodes permettant de détecter chacun de ces problèmes dans les données.

BigQuery peut résoudre de nombreux problèmes de qualité des données grâce à SQL et aux vues



Vous avez maintenant détecté les problèmes. Que faites-vous pour les résoudre ?

ELT dans BigQuery peut résoudre de nombreux problèmes de qualité des données. Voici un exemple. Imaginez que vous ayez l'intention d'analyser des données mais qu'elles comportent des enregistrements en double, ce qui donne l'impression qu'un seul type d'événement est plus courant, alors qu'en fait il s'agit simplement d'un problème de qualité des données.

Vous ne pouvez pas extraire des informations des données tant que les doublons n'ont pas été supprimés.

Donc, avez-vous besoin d'une étape de transformation pour supprimer les doublons avant de stocker les données ?

Peut-être...

Mais il existe une solution plus simple, qui consiste à compter les enregistrements uniques. Vous disposez bien sûr de la fonction COUNT DISTINCT dans BigQuery et vous pouvez l'utiliser à la place.

De même, BigQuery permet de résoudre un problème comme le fait que les données soient en dehors de l'intervalle, sans étape de transformation intermédiaire. En effet, il est possible de filtrer les données non valides à l'aide d'une vue BigQuery, et tout le monde peut accéder à la vue plutôt qu'aux données brutes.

Programme

EL, ELT, ETL

Considérations sur la qualité

Comment exécuter des opérations
dans BigQuery

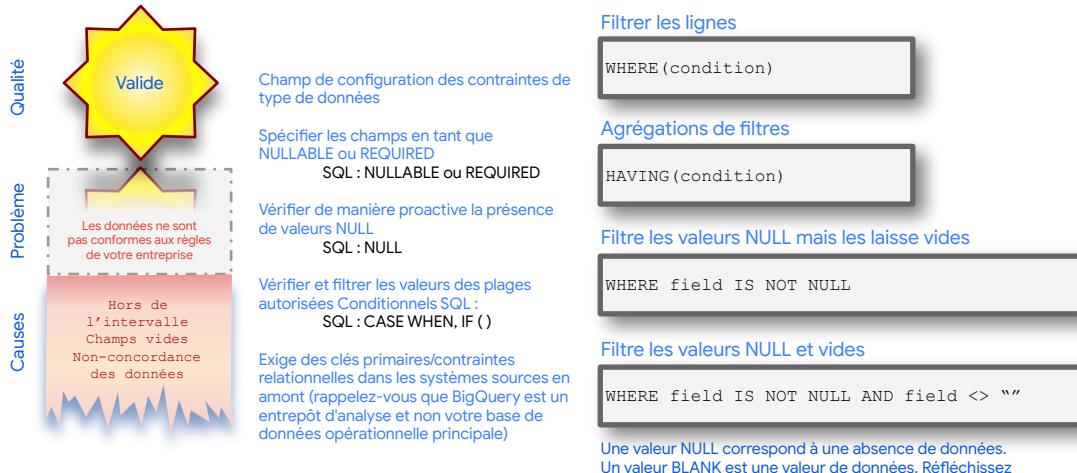
Inconvénients

ETL pour résoudre les problèmes
liés à la qualité des données



Dans cette section, nous allons examiner divers problèmes de qualité et découvrir certaines fonctionnalités de BigQuery qui peuvent vous aider à résoudre ces problèmes de qualité.

Filtrer pour identifier et isoler les données non valides



Nous pouvons utiliser les vues pour filtrer les valeurs qui présentent des problèmes de qualité.

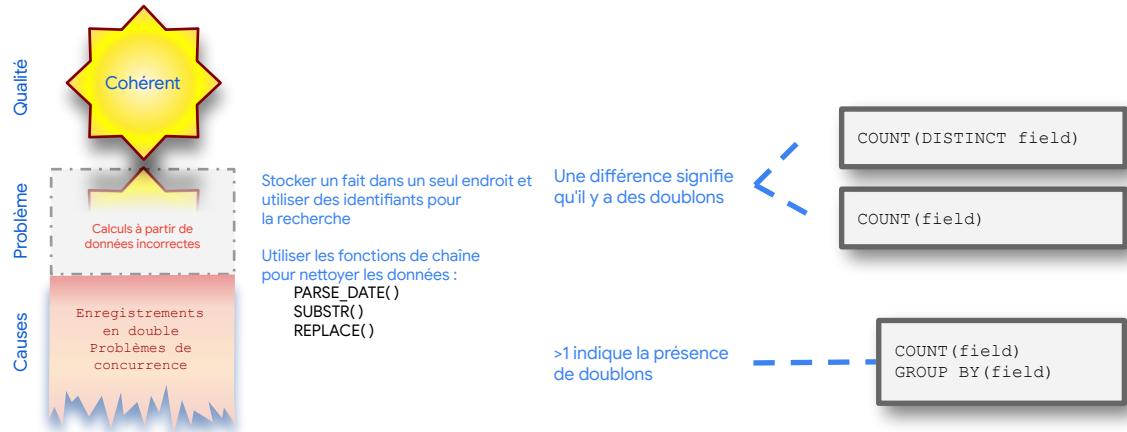
Par exemple, supprimer les quantités inférieures à zéro en utilisant une clause WHERE.

Après avoir exécuté une commande group by, vous pouvez éliminer les groupes dont le nombre total d'enregistrements est < 10 en utilisant la clause HAVING.

Réfléchissez bien à la manière dont vous souhaitez traiter les zéros et les blancs. Une valeur **NULL** correspond à une absence de données. Un **BLANK** est une chaîne vide. Réfléchissez si vous essayez de filtrer à la fois les **NULS** et les **BLANCS** ou seulement les **NULS** ou seulement les **BLANCS**.

Vous pouvez facilement compter les valeurs non nulles en utilisant COUNTIF Et utiliser l'instruction IF pour éviter d'utiliser des valeurs spécifiques dans les calculs

Déetecter les doublons, faire respecter l'unicité pour la cohérence



Les problèmes de cohérence sont souvent dus à des doublons. Vous vous attendez à ce que quelque chose soit unique, et ce n'est pas le cas, alors des résultats comme les totaux sont faux.

COUNT indique le nombre de lignes d'un tableau qui contiennent une valeur non nulle

COUNT DISTINCT indique le nombre de valeurs uniques

S'ils sont différents, cela signifie que vous avez des valeurs en double.

De même, si vous exécutez une instruction group by, et qu'un groupe contient plus d'une ligne, alors vous savez que vous avez deux occurrences ou plus de cette valeur.

Une autre raison pour laquelle vous pourriez avoir des problèmes de cohérence est que des caractères supplémentaires sont ajoutés aux champs. Par exemple, il se peut que vous obteniez des horodatages, dont certains peuvent inclure un fuseau horaire. Ou certaines de vos chaînes sont complétées. Utilisez les fonctions de chaîne pour nettoyer ces données avant de les transmettre.

Tester la précision des données par rapport aux valeurs valides connues



Pour vérifier l'exactitude des données, les tester par rapport aux valeurs valides connues.

Par exemple, si vous avez une commande, vous pouvez calculer le sous-total à partir de la quantité commandée et du prix de l'article et vous assurer que le calcul est correct.

De même, vous pouvez vérifier si une valeur qui est insérée appartient à une liste canonique de valeurs acceptables. Pour cela, vous pouvez utiliser une requête SQL IN.

Identifier et compléter les valeurs manquantes pour assurer l'exhaustivité



A des fins d'exhaustivité, identifiez toute valeur manquante et filtrez-la ou remplacez-la par une valeur raisonnable.

Si la valeur manquante est NULL, SQL fournit des fonctions comme NULLIF, COUNTIF, COALESCE, etc. Pour les exclure des calculs.

Vous pouvez peut-être exécuter une UNION à partir d'une autre source pour remplir les mois de données manquantes.

Le processus automatique de détection des pertes de données et de demande d'éléments de données pour combler les lacunes est appelé « remblaiement ». Il s'agit d'une caractéristique de certains services de transfert de données.

Lors du chargement des données, vérifier l'intégrité des fichiers avec des valeurs de contrôle (hash, MD5)

Rendre les types et les formats de données explicites dans un souci d'uniformité



Que se passe-t-il si vous stockez une valeur en centimètres, et que soudain, vous commencez à obtenir la valeur en millimètres ?

Votre data warehouse se retrouvera avec des données non uniformes. Vous devez vous protéger contre cela.

Utilisez SQL cast pour éviter les problèmes liés à la modification des types de données dans une table.

Utilisez la fonction SQL FORMAT() pour indiquer clairement les unités. Et en général, documentez-les très clairement.

J'espère que vous sortirez avec l'idée que BigQuery SQL est très puissant et que vous pourrez en tirer profit.

Démons- tration

ELT pour améliorer la qualité
des données dans BigQuery

Instructions de la démonstration :

https://github.com/GoogleCloudPlatform/training-data-analyst/blob/master/courses/data-engineering/demos/simple_healthcheck.md

Programme

EL, ELT, ETL

Considérations sur la qualité

Comment exécuter des opérations
dans BigQuery

Inconvénients

ETL pour résoudre les problèmes
liés à la qualité des données



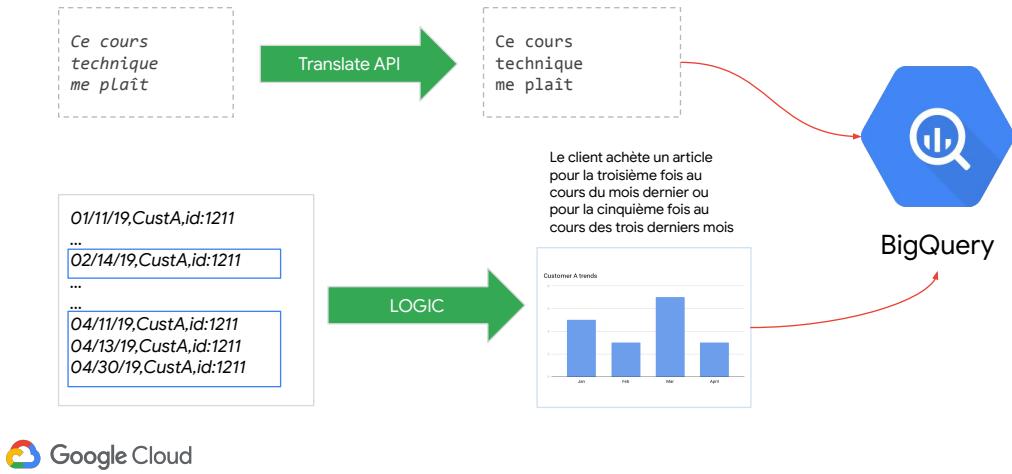
Dans la section précédente, nous vous avons montré quelques-unes des façons dont vous pouvez utiliser SQL dans un pipeline ELT pour vous prémunir contre les problèmes de qualité.

Le fait est que vous n'avez pas toujours besoin d'ETL. 'ELT peut être une option même si vous avez besoin d'une transformation.

Cependant, il y a des situations où ELT ne sera pas suffisant. Dans ce cas, ETL pourrait être ce qu'il vous faut.

Dans quelles situations ?

Que faire si les transformations ne peuvent pas être exprimées en SQL ? Ou si elles sont trop complexes pour être réalisées dans SQL ?



Le premier exemple (traduction de l'espagnol vers l'anglais) nécessite une API externe. Il ne peut pas être effectué en SQL.

Le deuxième exemple (examiner un flux d'actions de clients sur une fenêtre de temps) est complexe. Vous pouvez le faire avec des agrégations de fenêtres, mais c'est beaucoup plus simple avec la logique.

Ainsi, si les transformations ne peuvent pas être exprimées en SQL ou sont trop complexes pour être réalisées en SQL, vous pouvez transformer les données avant de les charger dans BigQuery.

Construire des pipelines de données ETL dans Dataflow et déposer les données dans BigQuery

Architecture	Quand exécuter la tâche
Extraire des données de Pub/Sub, Google Cloud Storage, Cloud Spanner, Cloud SQL, etc.	Lorsque les données brutes doivent être contrôlées, transformées ou enrichies avant d'être chargées dans BigQuery.
Transformer les données à l'aide de Cloud Dataflow	Lorsque le chargement des données doit se faire en continu, c'est-à-dire si le cas d'utilisation nécessite un flux de données.
Faire en sorte que le flux de données Dataflow d'écrire sur BigQuery	Lorsque vous souhaitez intégrer des systèmes d'intégration continue / livraison continue (CI/CD) et effectuer des tests unitaires sur tous les composants.



L'architecture de référence GCP suggère Cloud Dataflow comme outil ETL. Nous vous recommandons de construire des pipelines de données ETL dans Dataflow et de déposer les données dans BigQuery

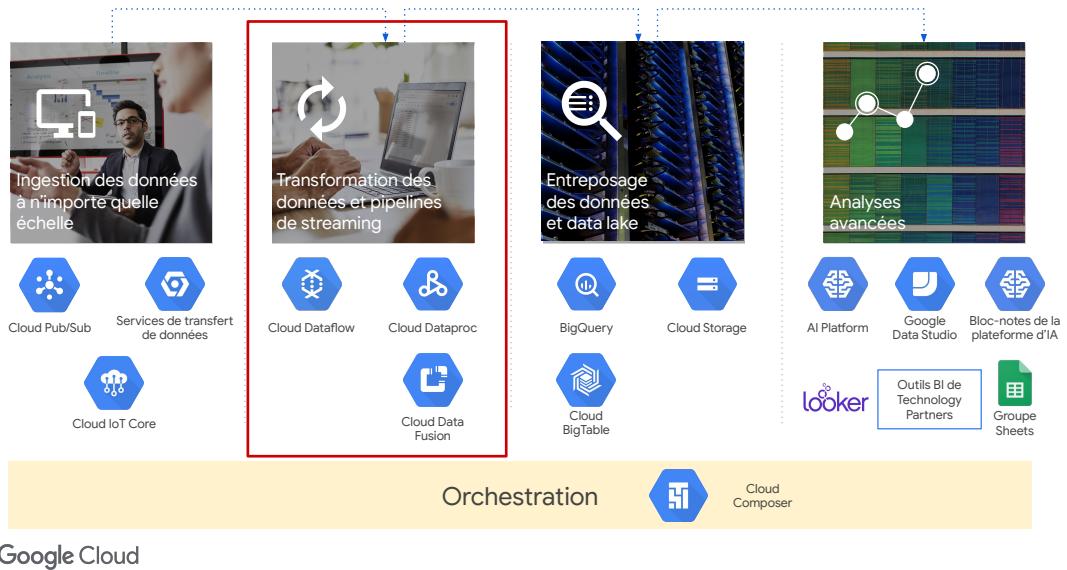
Voici à quoi ressemble l'architecture :

- Extraire des données de Pub/Sub, Google Cloud Storage, Cloud Spanner, Cloud SQL, etc.
- Transformer les données à l'aide de Cloud Dataflow
- Faire en sorte que le flux de données Dataflow d'écrire sur BigQuery

Quand exécuter ces opérations ?

1. Lorsque les données brutes doivent être contrôlées, transformées ou enrichies avant d'être chargées dans BigQuery. Et que les transformations sont difficiles à effectuer en SQL.
2. Lorsque le chargement des données doit se faire en continu, c'est-à-dire si le cas d'utilisation nécessite un flux de données. Dataflow prend en charge le streaming. Nous examinerons plus en détail le streaming dans le prochain cours.
3. Lorsque vous souhaitez intégrer des systèmes d'intégration continue / livraison continue (CI/CD) et effectuer des tests unitaires sur tous les composants. Il est facile de programmer le lancement d'un pipeline de flux de données.

Google Cloud propose une gamme d'outils ETL



Dataflow n'est pas la seule option dont vous disposez sur GCP pour exécuter ETL.

Dans ce cours, nous examinerons plusieurs services de traitement et de transformation des données fournis par les GCP : Cloud Data Proc, Cloud Dataflow et Cloud Data Fusion.

S'il est logique d'utiliser BigQuery pour des transformations ELT plus simples, et qu'il est plus facile à configurer, ce ne devrait pas être votre seul outil ETL. Cloud Dataproc et Cloud Dataflow peuvent être utilisés pour des pipelines ETL plus complexes.

Cloud Dataproc repose sur Apache Hadoop et nécessite une grande expertise d'Hadoop pour être directement exploité. Cloud Data Fusion fournit une interface graphique simple à utiliser pour construire des pipelines ETL qui peuvent ensuite être facilement déployés à l'échelle des clusters Cloud Dataproc.

Cloud Dataflow est un service de traitement de données entièrement géré et sans serveur, basé sur Apache Beam, qui prend en charge les pipelines de traitement de données par lots et en continu. Bien qu'une expertise significative d'Apache Beam soit souhaitable pour exploiter toute la puissance de Cloud Dataflow, Google fournit également des modèles de démarrage rapide pour Cloud Dataflow afin de vous permettre de déployer rapidement un certain nombre de pipelines de données utiles.

Vous pouvez utiliser n'importe lequel de ces trois produits pour effectuer une transformation de données, puis stocker les données dans un data lake ou un data warehouse pour prendre en charge des analyses avancées.

Programme

EL, ELT, ETL

Considérations sur la qualité

Comment exécuter des opérations
dans BigQuery

Inconvénients

ETL pour résoudre les problèmes
liés à la qualité des données



Cas où vous regardez au-delà de Dataflow et BigQuery

Problème	Solution
Latence, débit	Dataflow vers Bigtable
Réutiliser des flux de données Spark	Cloud Dataproc
Besoin de construire des flux de données visuel	Cloud Data Fusion



À moins que vous ayez des besoins spécifiques, nous vous recommandons d'utiliser Dataflow et BigQuery.

Quels pourraient être ces besoins ?

1. Latence et débit. Les requêtes BigQuery sont soumises à un temps de latence de l'ordre de quelques centaines de millisecondes et vous pouvez transmettre en continu un million de lignes par seconde dans une table BigQuery - il y avait auparavant 100 000 lignes, mais ce chiffre est récemment passé à 1 million par projet (si vous pouvez vous accommoder de l'absence de déduplication au mieux). Le nombre de latence typique cité pour BigQuery est de l'ordre de la seconde, mais avec le moteur BI il est possible d'obtenir une latence de l'ordre de 100 millisecondes -- vous devriez toujours vérifier les dernières valeurs dans la documentation et les pages de solutions. Si vos considérations de latence et de débit sont plus strictes, alors Cloud Bigtable pourrait être un meilleur choix pour vos pipelines de traitement de données.
2. Réutilisation des flux de données Spark. Vous avez peut-être déjà investi dans Hadoop et Spark. Dans ce cas, vous pourriez être beaucoup plus productif avec une technologie qui vous est familière. Utilisez Spark si c'est ce que vous connaissez vraiment bien.
3. Besoin d'une construction visuelle des flux de données. Dataflow vous oblige à coder les pipelines de données en Java ou en Python. Si vous voulez que des analystes de données et des utilisateurs non techniques créent des pipelines de données, utilisez Cloud Data Fusion. Ils peuvent faire un glisser-déposer et construire visuellement des pipelines.

Nous allons examiner brièvement toutes ces options maintenant et plus en détail dans la suite de ce cours.

Le Cloud Bigtable est un service de base de données NoSQL entièrement géré à l'échelle du pétaoctet pour les cas d'utilisation où un accès aléatoire aux données à faible latence, l'évolutivité et la fiabilité sont essentiels

	 Cloud Storage	 Cloud Firestore	 Cloud Bigtable	 Cloud SQL	 Cloud Spanner	 BigQuery
Taille typique	Tous	< 200 To	2 - 10 Po	< 10 To	Tous	Tous
Type de stockage	Objets	Document	Clé-Valeur	Relationnel	Relationnel	En colonnes
Latence L/E	Moyen (100s de ms)	Moyen (10s de ms)	Faible (ms)	Faible (ms)	Faible (ms)	Élevée (s)



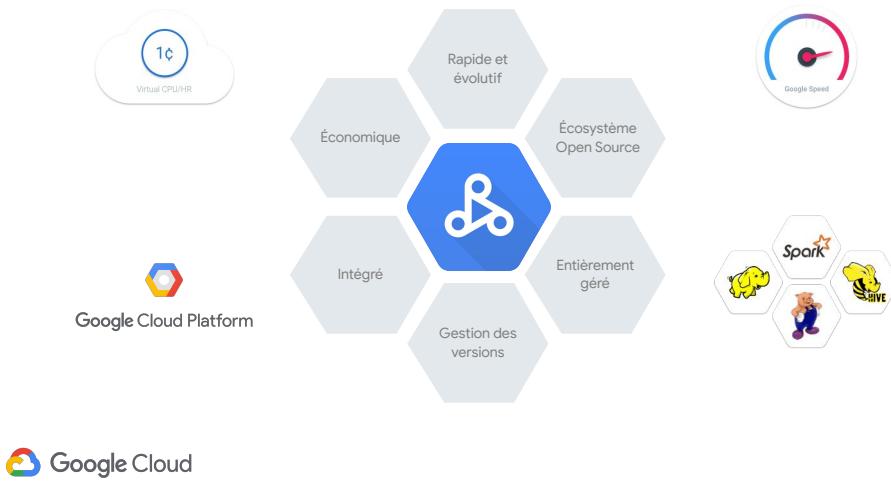
Cloud Bigtable est un service de base de données entièrement géré à l'échelle du pétaoctet pour les cas d'utilisation où un accès aléatoire aux données à faible latence, l'évolutivité et la fiabilité sont essentiels

Vous pouvez stocker et rechercher des pétaoctets de données avec une latence de quelques millisecondes.

Contrairement à BigQuery, Cloud Bigtable est un service de base de données NoSQL. Vous stockez des paires clé-valeur et recherchez des valeurs basées sur des clés.

Vous ne pouvez pas interroger Bigtable à l'aide de SQL ; vous devez plutôt y accéder à l'aide de l'API HBase.

Cloud Dataproc est un service géré pour le traitement par lots, l'interrogation, le streaming et le ML

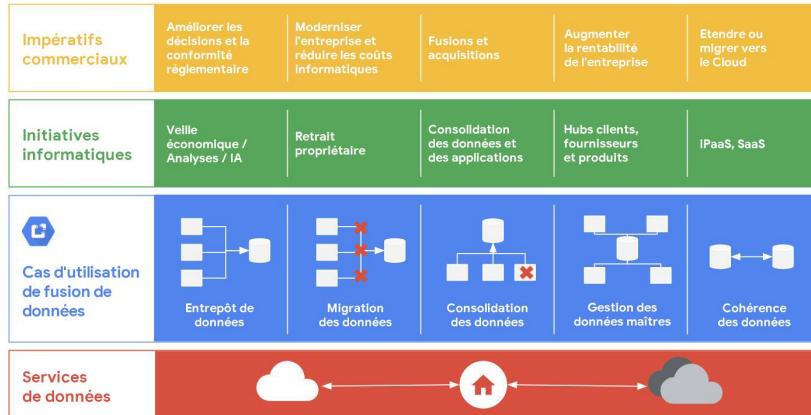


Cloud Dataproc est un service géré pour le traitement par lots, l'interrogation, le streaming et le ML

Il fournit un service géré pour les charges de travail de Hadoop et est très rentable - environ 1c de plus que le coût de son fonctionnement à l'état brut et de la prise en charge de toutes les activités de maintenance de Hadoop.

Il propose également quelques fonctions intéressantes comme la mise à l'échelle automatique et l'intégration prête à l'emploi avec des produits GCP comme BigQuery.

Cloud Data Fusion est un service d'intégration de données d'entreprise entièrement géré, natif Cloud, permettant de créer et de gérer rapidement des pipelines de données



Cloud Data Fusion est un service d'intégration de données d'entreprise entièrement géré, natif Cloud, permettant de créer et de gérer rapidement des pipelines de données

Vous pouvez l'utiliser pour alimenter un data warehouse, mais vous pouvez aussi l'utiliser pour les transformations et le nettoyage et pour assurer la cohérence des données.

Les utilisateurs, qui peuvent faire partie de l'entreprise, peuvent construire des pipelines visuels pour répondre aux impératifs commerciaux tels que la conformité réglementaire sans avoir à attendre qu'une équipe informatique code un pipeline Dataflow.

Data Fusion dispose également d'une API de codage ; les informaticiens peuvent l'utiliser pour créer des scripts et automatiser.

Le suivi de la lignée dans les pipelines ETL peut être important



Leur source



Le processus qu'elles ont traversé



Leur emplacement et leur état actuel

Leur format
Leurs qualités
Leur adéquation à leur utilisation prévue
S'il est possible de les transformer ou de les traiter pour les adapter à leurs utilisations prévues

Lignée : Métadonnées à propos des données



Quelle que soit l'utilisation d'ETL - Dataflow, Dataproc, Data Fusion - il y a des aspects cruciaux à garder à l'esprit.

Tout d'abord : Le maintien de la lignée des données est importante.

Qu'entendons-nous par lignée ?

D'où viennent les données, quels sont les traitements qu'elles ont subis, dans quel état elles se trouvent.

Si vous avez la lignée, vous savez à quels types d'utilisation les données sont adaptées.

Comprendre, à partir de la lignée, l'état actuel des données et les traitements qu'elles pourraient avoir à subir pour être adaptées à l'utilisation prévue

Si vous observez que les données donnent des résultats bizarres, vous pouvez vérifier la lignée pour savoir s'il y a une cause qui peut être corrigée.

La lignée contribue également à la confiance et au respect de la réglementation.

L'autre préoccupation transversale est la nécessité de conserver les métadonnées. Vous avez besoin d'un moyen de suivre la lignée des données au sein de votre organisation pour découvrir et déterminer si elles conviennent à leurs utilisations.

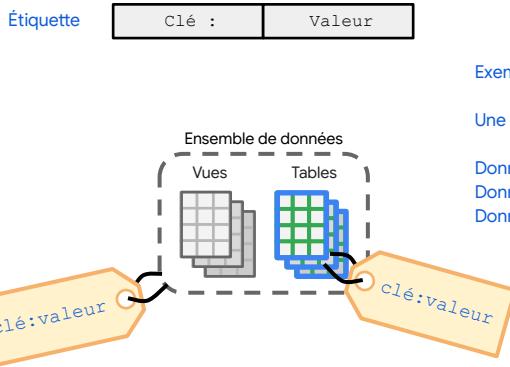
Sur Google Cloud, Cloud Data Catalog permet de les découvrir. Mais vous devez faire votre part en ajoutant des étiquettes.

<https://pixabay.com/fr/photos/barley-field-wheat-harvest-sunrise-1684052/>

<https://pixabay.com/fr/photos/yellowstone-national-park-sunset-1589616/>

<https://pixabay.com/fr/photos/trees-forest-forest-path-sunlight-3410836/>

Des étiquettes sur les ensembles de données, sur les tables et les vues peuvent aider à suivre



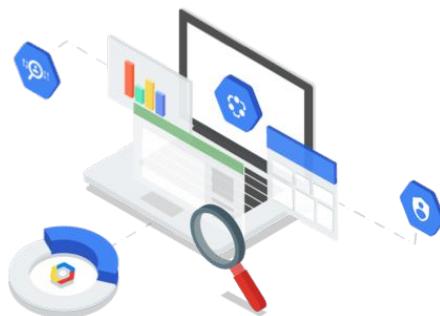
Une étiquette est une paire de valeurs clés qui vous aide à organiser vos ressources. Dans BigQuery, vous pouvez attacher des étiquettes aux ensembles de données, aux tableaux et aux vues.

Les étiquettes sont utiles pour gérer des ressources complexes, car vous pouvez les filtrer en fonction de leurs étiquettes.

Les étiquettes constituent une première étape vers un catalogue de données.

Parmi les tâches que les étiquettes facilitent, on peut citer la facturation en ligne. Si vous associez des étiquettes aux instances du Compute Engine, aux buckets et aux pipelines de Dataflow, vous disposez d'un moyen d'obtenir un aperçu précis de votre facture Cloud, car les informations sur les étiquettes sont transmises au système de facturation, et vous pouvez donc ventiler vos frais de facturation par étiquette.

Visualisez vos ensembles de données et vos étiquettes dans Data Catalog



 Google Cloud

Data Catalog est un service de découverte de données et de gestion de métadonnées entièrement géré et hautement évolutif

Il est sans serveur et ne nécessite aucune infrastructure à mettre en place ou à gérer.

Il fournit des contrôles d'accès et honore les ACL de source pour la lecture, l'écriture et la recherche des données, vous donnant un contrôle d'accès de niveau entreprise.

Considérez le Data Catalog comme une métadonnée en tant que service. Il fournit un service de gestion des métadonnées pour les actifs de catalogage des données via des API personnalisées et l'interface utilisateur, offrant ainsi une vue unifiée des données où qu'elles se trouvent.

Il prend en charge des balises schématisées (par exemple, Enum, Bool, DateTime) et pas seulement de simples balises de texte - fournissant ainsi aux organisations des métadonnées commerciales riches et organisées.

Il offre une découverte unifiée de tous les actifs de données répartis sur plusieurs projets et systèmes.

Il est fourni avec une interface de recherche simple et conviviale permettant de trouver rapidement et facilement des données, grâce à la technologie de recherche Google qui prend en charge Gmail et Drive. En tant que catalogue central, il offre un système de catalogage flexible et puissant pour la saisie de métadonnées techniques (automatiquement) et de métadonnées commerciales (balises) dans un format structuré.

L'un des aspects les plus intéressants de la découverte de données est qu'elle

s'intègre à l'API de prévention des pertes de données Cloud.

Vous pouvez l'utiliser pour découvrir et classer des données sensibles, en fournissant des renseignements et en aidant à simplifier le processus de gestion de vos données.

Data Catalog permet aux utilisateurs d'annoter les métadonnées de l'entreprise de manière collaborative et constitue la base de la gouvernance des données