

Vidéo : 1_l1

Découvrez le rôle de l'ingénieur données

Présentateur : Lak

[FIN DE LA VIDÉO]



Introduction à l'ingénierie des données

Bonjour et bienvenue dans notre cours sur l'ingénierie des données.
Je m'appelle Lak et je dirige les solutions Analytics et IA sur Google Cloud.

Programme

Découvrez le rôle de l'ingénieur données

Analysez les défis de l'ingénierie des données

Introduction à BigQuery

Data Lakes et Data Warehouses

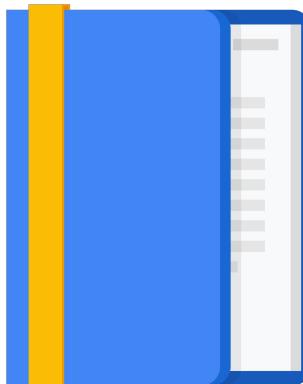
Bases de données transactionnelles vs Data Warehouses

Travailler efficacement avec les autres équipes chargées des données

- Gérer l'accès et la gouvernance des données
- Créer des pipelines prêts à l'emploi

Passer en revue les études de cas des clients GCP

Atelier : Analyser les données à l'aide de BigQuery



Au cours de ce module, nous allons parler du travail d'un ingénieur données.

L'objectif d'un ingénieur données est de créer des pipelines de données. Nous allons découvrir ce que cela signifie : quels types de pipelines crée un ingénieur données et quel est le but de ces pipelines.

Après avoir découvert la tâche d'un ingénieur données, nous allons expliquer pourquoi nous vous recommandons d'appliquer l'ingénierie des données sur le cloud, et sur Google Cloud Platform en particulier.

Nous allons nous pencher sur les défis associés à la pratique de l'ingénierie des données et voir combien de ces défis sont plus faciles à traiter lorsqu'on crée des pipelines de données sur le cloud.

Nous verrons ensuite quel est l'objectif des pipelines de données que vous créez. Lorsque vous les créez correctement, qu'est-ce que vous rendez possible dans les organisations pour lesquelles vous créez des pipelines de données ?

Enfin, nous regarderons quelques exemples d'architectures de référence. Les architectures de référence sont des architectures que vous adaptez. Pour ce faire, vous devez connaître le but de chaque partie de l'architecture de données. Nous allons examiner plus en détail les produits des composants de cette spécialisation, mais nous commençons ici avec une vue d'ensemble de l'objectif des divers produits.

Un ingénieur données crée des pipelines de données pour permettre la prise de décisions basées sur les données



Que fait un ingénieur données ?

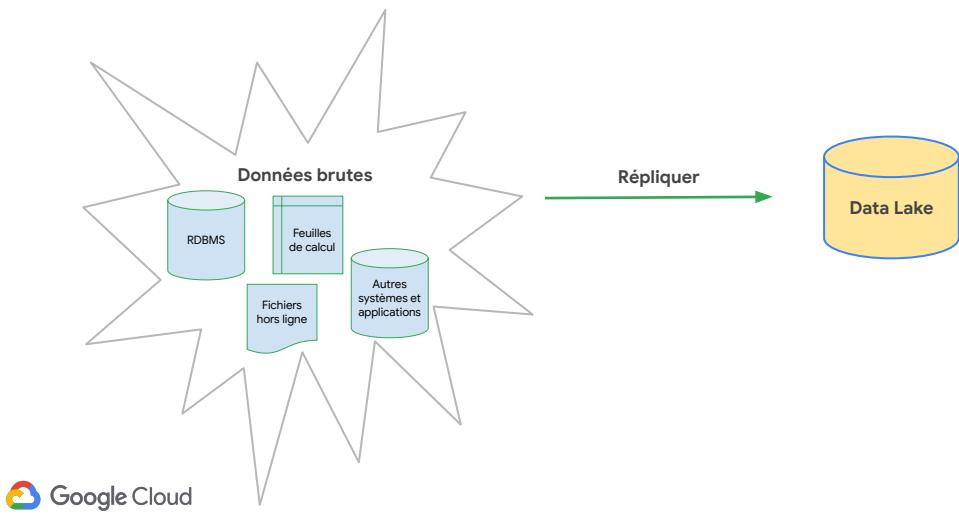
Un ingénieur données crée des pipelines de données.

Pourquoi l'ingénieur données crée-t-il des pipelines de données ?

Il veut placer ses données dans un lieu tel qu'un tableau de bord ou un rapport ou un modèle de machine learning, à partir duquel l'entreprise peut prendre des décisions en fonction des données.

Les données doivent être utilisables afin qu'une personne puisse les utiliser pour prendre des décisions. Bien souvent, les données brutes mêmes ne sont pas très **utiles**.

Un data lake réunit les données de toute l'entreprise à un seul endroit



Un terme que vous entendrez souvent dans l'ingénierie des données est le concept de data lake.

Un data lake réunit les données de toute l'entreprise à un seul endroit.

Ainsi, vous pouvez obtenir les données d'une base de données relationnelle ou d'une feuille de calcul, et stocker les données brutes dans un data lake.

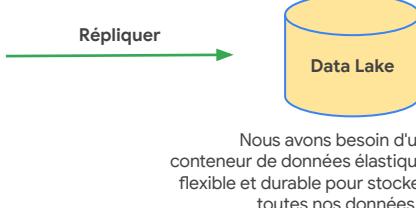
Une des options pour ce lieu unique permettant de stocker les données brutes est de les stocker dans un bucket Cloud Storage.

Quels sont les éléments clés à prendre en considération avant de décider parmi les options de data lakes ? Qu'en pensez-vous ?

<https://cloud.google.com/solutions/build-a-data-lake-on-gcp#cloud-storage-as-data-lake>

Éléments clés à prendre en considération lors de la création d'un Data Lake

1. Votre data lake peut-il prendre en charge tous les types de données que vous avez ?
2. Peut-il évoluer pour répondre à la demande ?
3. Supporte-t-il l'ingestion à haut débit ?
4. Offre-t-il un contrôle d'accès précis aux objets ?
5. Est-il possible de connecter facilement d'autres outils ?



Quels éléments clés à prendre en considération devez-vous garder à l'esprit lorsque vous créez un data lake ?

Votre data lake prend-il en charge tous les types de données que vous avez ? Est-il possible d'en intégrer la totalité dans un bucket Cloud Storage ? Si vous disposez d'un RDBMS, vous pourrez avoir besoin de mettre les données dans Cloud SQL, une base de données gérée, plutôt que dans Cloud Storage.

Peut-il évoluer de manière élastique pour répondre à la demande ? Au fur et à mesure que vos données collectées augmentent, allez-vous manquer d'espace disque ? Ce problème concerne davantage les systèmes sur site que sur le cloud.

Supporte-t-il l'ingestion à haut débit ? Quelle est la bande passante du réseau ? Disposez-vous de points de présence périphériques ?

Offre-t-il un contrôle d'accès précis aux objets ? Les utilisateurs ont-ils besoin de rechercher dans un fichier ? Ou l'obtention d'un fichier dans son ensemble est-elle suffisante ? Cloud Storage étant un stockage en blob, vous devrez penser à la granularité des éléments que vous stockez.

Est-il possible de connecter facilement d'autres outils ? Comment accèdent-ils au magasin ? Ne perdez pas de vue que l'objectif du data lake est de rendre les données accessibles pour analyse.

<https://cloud.google.com/solutions/build-a-data-lake-on-gcp#cloud-storage-as-data-lake?hl=fr>

Cloud Storage est conçu pour une durabilité annuelle de 99,99999999 %



Sauvegarde

Remplacement/décommissionnement d'infrastructure

Analyses et machine learning

Stockage et fourniture de contenu

Créez rapidement des buckets avec Cloud Shell `gsutil mb gs://your-project-name`



Nous avons parlé de notre premier produit Google Cloud, le bucket Cloud Storage, qui représente une bonne option pour stocker toutes vos données brutes en un lieu avant de créer des pipelines de transformation dans votre entrepôt de données.

Pourquoi choisir Google Cloud Storage ? En règle générale, les entreprises utilisent Cloud Storage comme outil de sauvegarde et d'archivage pour leurs activités. Grâce aux nombreux lieux de centre de données et à la grande disponibilité de réseau que propose Google, le stockage des données dans un bucket GCS est durable et performant.

En tant qu'ingénieur données, vous utiliserez souvent un bucket Cloud Storage dans le cadre de votre data lake pour stocker divers fichiers de données brutes (CSV, JSON, Avro). Vous pourrez ensuite les charger ou les interroger directement depuis BigQuery en tant que data warehouse. Ultérieurement dans le cours, vous créerez des buckets Cloud Shell en utilisant la console et la ligne de commande comme vous le voyez ici. D'autres produits et services Google Cloud Platform peuvent facilement être interrogés et intégrés à votre bucket une fois que vous l'avez configuré et chargé de données.

Et si vos données ne sont pas utilisables dans leur format d'origine ?



Extraire, transformer
et charger

Traitement des données



Cloud Dataproc



Cloud Dataflow



En parlant de chargement de données, que se passe-t-il si vos données brutes nécessitent un traitement supplémentaire ?

Vous pourrez avoir besoin d'extraire les données de son emplacement d'origine, de les transformer et de les charger.

L'une des options est d'effectuer un traitement des données. Cette action est souvent réalisée à l'aide de Cloud Dataproc ou Cloud Dataflow. Nous aborderons l'utilisation de ces produits pour traiter les pipelines par lots ultérieurement dans ce cours.

Et si vos données arrivent continuellement, sans fin ?



Traitement des flux de données



Cloud Pub/Sub



Cloud Dataflow



BigQuery



Mais que se passe-t-il si les pipelines par lots ne sont pas suffisants ? Que se passe-t-il si vous avez besoin d'une analyse des données en temps réel continue et infinie ?

Dans ce cas, vous pouvez recevoir des données dans Cloud Pub/Sub, les transformer à l'aide de Cloud Pub/Sub et les transmettre dans BigQuery. Nous aborderons la transmission des pipelines ultérieurement dans ce cours.

<https://pixabay.com/fr/photos/chutes-d-eau-cours-d-eau-ruisseaux-1149994/>

Vidéo : 1_I2

Analysez les défis de l'ingénierie des données

Présentateur : Lak

[FIN DE LA VIDÉO]

Programme

Découvrez le rôle de l'ingénieur données

Analysez les défis de l'ingénierie des données

Introduction à BigQuery

Data Lakes et Data Warehouses

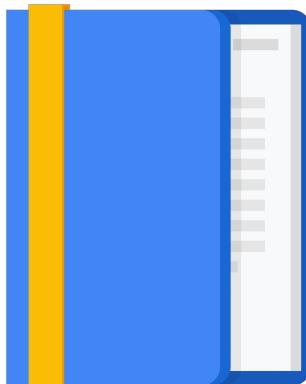
Bases de données transactionnelles vs Data Warehouses

Travailler efficacement avec les autres équipes chargées des données

- Gérer l'accès et la gouvernance des données
- Créer des pipelines prêts à l'emploi

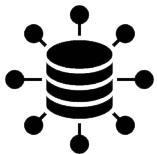
Passer en revue les études de cas des clients GCP

Atelier : Analyser les données à l'aide de BigQuery



Regardons les défis de l'ingénierie des données.

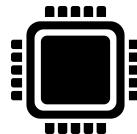
Les difficultés fréquentes que rencontrent les ingénieurs données



Accès aux données



Précision et qualité des données



Disponibilité des ressources informatiques

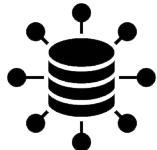


Performance de la requête



En tant qu'ingénieur données, vous rencontrerez généralement quelques problèmes lorsque vous créez des pipelines de données. Il vous sera peut-être difficile d'accéder aux données dont vous avez besoin. Vous verrez peut-être que, après **y avoir accédé**, les données n'offrent pas **la qualité** requise par l'analyse ou le modèle de machine learning. Vous prévoyez de créer un modèle, et même si la qualité des données est présente, vous verrez peut-être que les transformations nécessitent des ressources informatiques qui **ne sont pas disponibles** pour vous. Enfin, vous pouvez rencontrer des défis avec une **performance de requête** et l'exécution de toutes les requêtes et toutes les transformations dont vous avez besoin avec les ressources informatiques que vous avez.

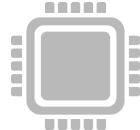
Défi : consolider des ensembles et des formats de données disparates, et gérer des accès évolutifs



Accès aux données



Précision et qualité des données



Disponibilité des ressources informatiques



Performance de la requête



Commençons par le premier défi qui consiste à consolider des ensembles et des formats de données disparates, et à gérer des accès évolutifs.

Il est difficile d'obtenir des insights pour plusieurs ensembles de données sans data lake

Les données sont réparties dans les produits Google Analytics 360, CRM et Campaign Manager, parmi d'autres ressources.

Les données clients et ventes sont stockées dans un système CRM.

Aucun outil commun n'existe pour analyser des données et partager les résultats avec le reste de l'organisation.

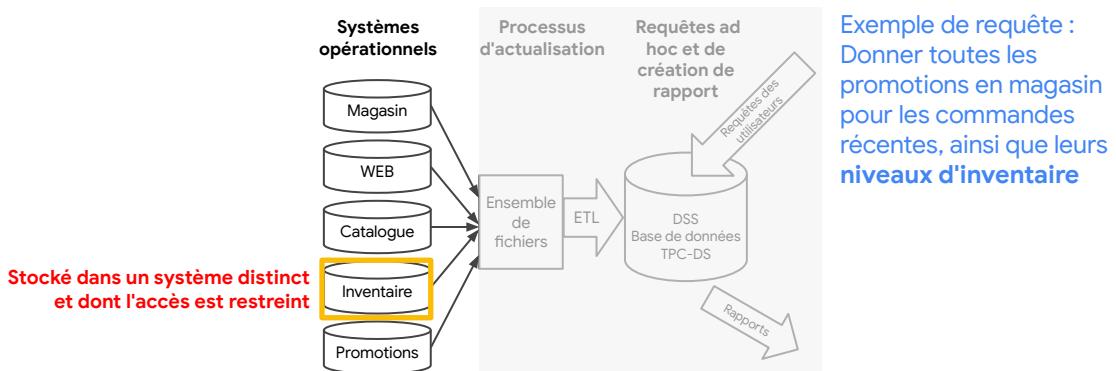
Certaines données ne sont pas dans un format pouvant être interrogé.



Par exemple, vous souhaitez calculer les coûts d'acquisition du client : combien cela coûte-t-il en termes de marketing, promotions et réductions d'acquérir un client ? Ces données peuvent être réparties sur plusieurs produits marketing et divers logiciels de gestion de la relation client. Il peut être difficile de trouver un outil pouvant analyser toutes ces données, car elles peuvent provenir de différentes organisations, différents outils et différents schémas. Certaines de ces données peuvent également ne pas être structurées. Donc, afin de trouver quelque chose d'aussi essentiel pour votre entreprise que le coût d'acquisition d'un nouveau client, ce qui vous permettra de décider du type de réductions à offrir pour le conserver, vous ne pouvez pas avoir vos données en silos.

<https://pixabay.com/fr/photos/alone-being-alone-archetype-513525/>

Les données sont souvent cloisonnées dans plusieurs systèmes sources en amont



Exemple de requête :
Donner toutes les promotions en magasin pour les commandes récentes, ainsi que leurs niveaux d'inventaire



Qu'est-ce qui rend l'accès aux données si difficile ? Tout d'abord, c'est parce que de nombreuses entreprises sont cloisonnées par départements et que chaque département crée ses propres systèmes de transaction pour prendre en charge ses propres processus d'entreprise.

Par exemple, vous pouvez avoir des systèmes opérationnels qui correspondent aux systèmes de magasin, avoir un différent système opérationnel entretenu par vos entrepôts de produits qui gère votre inventaire, et avoir un département marketing qui gère toutes les promotions dont vous avez besoin pour effectuer une requête analytique telle que "Donner toutes les promotions en magasin pour les commandes récentes, ainsi que leurs niveaux d'inventaire."

Vous devez savoir comment combiner les données des magasins, des promotions et des niveaux d'inventaire et, comme ils sont tous stockés dans des systèmes distincts, parfois avec un accès restreint, créer un système d'analyse utilisant ces trois ensembles de données pour répondre à une telle requête ad hoc peut être très difficile.

<https://pixabay.com/fr/photos/route-asphalte-ciel-nuages-automne-220058/>

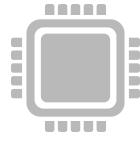
Défi : nettoyer, formater et obtenir les données afin qu'elles soient prêtes pour des insights d'entreprise utiles dans un data warehouse.



Accès aux données



Précision et qualité des données



Disponibilité des ressources informatiques



Performance de la requête



Le deuxième défi est que le nettoyage, le formatage et l'obtention des données prêtes pour des insights nécessitent de créer des pipelines ETL.

Les pipelines ETL sont généralement nécessaires pour garantir la précision et la qualité des données.

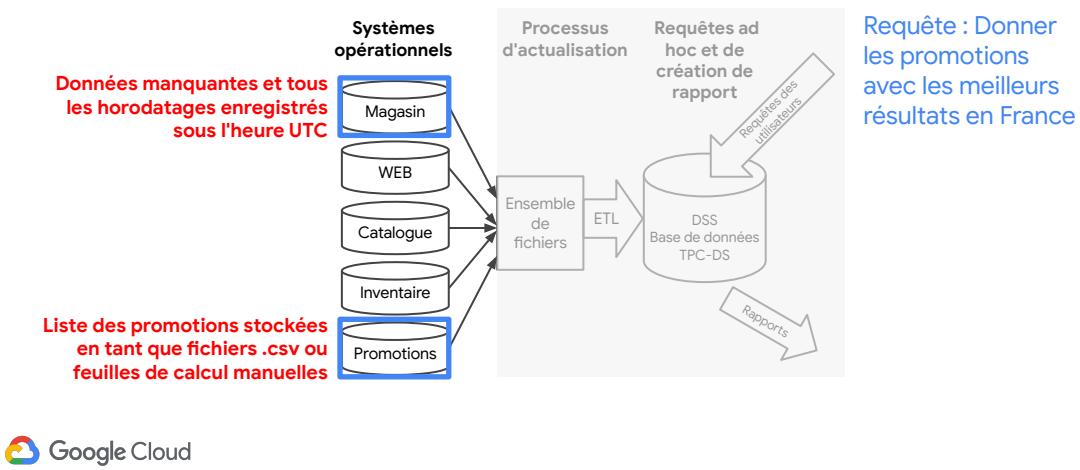
Les données nettoyées et transformées sont généralement stockées, pas dans un data lake, mais dans un data warehouse.

Un data warehouse est un lieu consolidé pour stocker les données, mais toutes les données peuvent facilement être jointes et interrogées.

Contrairement à un data lake, où les données sont au format brut.

Dans le data warehouse, les données sont stockées de manière à être interrogées efficacement.

Partir du principe que toutes les données brutes de systèmes sources doivent être nettoyées, transformées et stockées dans un data warehouse



Puisque les données ne deviennent utiles qu'une fois nettoyées, vous devez partir du principe que toutes les données brutes que vous collectez depuis des systèmes sources doivent être nettoyées et transformées.

Et si vous les transformez, vous pouvez tout autant les transformer en un format efficace à interroger.

En d'autres termes, extrayez, transformez et chargez les données et stockez-les dans un data warehouse.

Imaginons que vous êtes un détaillant et que vous devez consolider des données depuis plusieurs systèmes sources. Pensez au cas d'utilisation.

Partons du principe que le cas d'utilisation est d'obtenir les promotions en magasin avec les meilleures performances en France. Vous devrez alors obtenir les données des magasins et les données des promotions. Cependant, il est possible que les données du magasin n'aient pas toutes les informations. Peut-être que certaines transactions ont été réglées en espèce, et par conséquent, peut-être qu'il n'y a pas d'informations sur qui est le client ; ou certaines transactions peuvent être réparties sur plusieurs reçus, et vous devrez combiner ces transactions, car elles proviennent du même client. Ou peut-être que les horodatages des produits sont stockés dans l'heure locale, tandis que vous devez englober le monde entier, et vous devrez donc convertir toutes les données au format UTC avant de faire quoi que ce soit. De même, les promotions peuvent ne pas du tout être stockées dans la base de données des transactions. Elles peuvent simplement être un fichier texte chargé par un individu sur une page Web, et comporter une liste de codes utilisée par

l'application Web pour appliquer des réductions. Il peut être extrêmement difficile d'effectuer une requête telle que trouver les promotions en magasin avec les meilleures performances, car les données présentent de nombreux problèmes. Chaque fois que vous avez de telles données, vous devez obtenir les données brutes et les transformer dans un format avec lequel vous pouvez effectuer l'analyse nécessaire.

Il est évidemment préférable que vous n'effectuez ce type de nettoyage et consolidation qu'une seule fois, et que vous stockiez les données résultantes pour faciliter les analyses ultérieures.

C'est ce à quoi sert un data warehouse.

<https://pixabay.com/fr/photos/route-asphalte-ciel-nuages-automne-220058/>

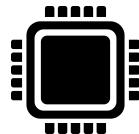
Défi : s'assurer que vous disposez de la capacité de calcul pour que votre équipe réponde à un pic de demandes



Accès aux données



Précision et qualité des données



Disponibilité des ressources informatiques

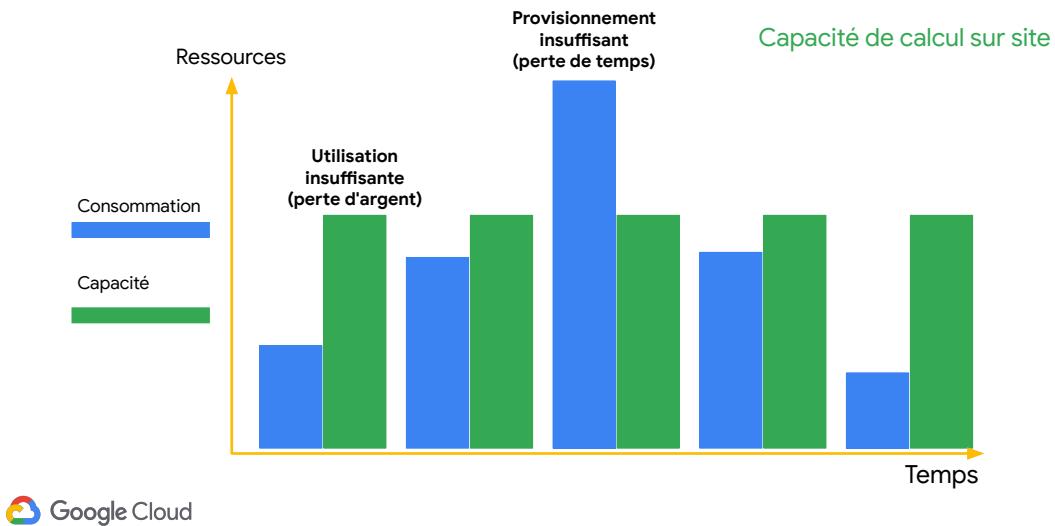


Performance de la requête



Si vous devez effectuer un grand volume de nettoyage et de consolidation, le problème récurrent est l'endroit où effectuer ces actions.
La disponibilité des ressources informatiques peut être un défi.

Défi : sur site, les ingénieurs données doivent gérer la capacité des clusters et des serveurs



Si vous utilisez un système sur site, les ingénieurs données devront gérer la capacité des serveurs et clusters et s'assurer qu'il y a suffisamment de capacité pour effectuer des tâches ETL.

Le problème est que les ressources informatiques nécessaires aux tâches ETL ne sont pas constantes dans le temps.

Très souvent, elles varient de semaine en semaine, et en fonction de facteurs tels que les vacances et les ventes promotionnelles.

Cela signifie que lorsque le trafic est lent, vous perdez de l'argent et lorsque le trafic est élevé, vos tâches prennent trop de temps.

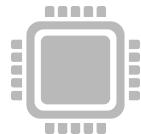
Défi : les requêtes doivent être optimisées pour être performantes (mise en cache, exécution en parallèle)



Accès aux données



Précision et qualité des données



Disponibilité des ressources informatiques



Performance de la requête



Une fois vos données dans votre data warehouse, vous devez optimiser les requêtes qu'exécutent vos utilisateurs pour une utilisation efficace de vos ressources informatiques.

Défi : gérer les performances de requête sur site avec des charges supplémentaires ;

- choisir un moteur de requête ;
- corriger et mettre à jour continuellement le logiciel du moteur de requête ;
- gérer les clusters et décider quand reclusteriser ;
- optimiser pour des requêtes simultanées et des quotas/demandes entre équipes.

Y a-t-il une meilleure façon de gérer les charges de serveur pour pouvoir se concentrer sur les insights ?



Si vous gérez un cluster d'analyse de données sur site, vous devrez choisir un moteur de requête, installer le logiciel du moteur de requête et le maintenir à jour, ainsi qu'alimenter d'autres serveurs pour une capacité supplémentaire.

N'y a-t-il pas une meilleure façon de gérer les charges de serveur pour pouvoir se concentrer sur les insights ?

Vidéo : 1_I3
Introduction à BigQuery
Présentateur : Lak

[FIN DE LA VIDÉO]

Programme

Découvrez le rôle de l'ingénieur données

Analysez les défis de l'ingénierie des données

Introduction à BigQuery

Data Lakes et Data Warehouses

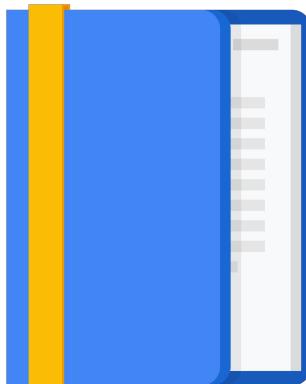
Bases de données transactionnelles vs Data Warehouses

Travailler efficacement avec les autres équipes chargées des données

- Gérer l'accès et la gouvernance des données
- Créer des pipelines prêts à l'emploi

Passer en revue les études de cas des clients GCP

Atelier : Analyser les données à l'aide de BigQuery



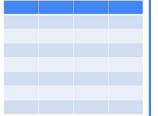
Il existe une meilleure façon de gérer les charges de serveur pour pouvoir se concentrer sur les insights.

C'est d'utiliser un data warehouse sans serveur.

BigQuery est le data warehouse sans serveur à l'échelle du pétaoctet de Google Cloud.

Vous n'avez pas besoin de gérer de clusters. Concentrez-vous seulement sur les insights.

BigQuery est la solution de data warehouse de Google

				
Data Warehouse	Magasin de données	Data Lake	Tables et vues	Octrois
BigQuery remplace la configuration matérielle type d'un data warehouse	BigQuery organise les tables de données en unités appelées ensembles de données	BigQuery définit les schémas et émet directement les requêtes sur des sources de données externes	Fonctionne de la même manière qu'un data warehouse classique	Cloud IAM octroie des autorisations pour effectuer des actions spécifiques



Le service BigQuery remplace la configuration matérielle type d'un data warehouse classique. C'est-à-dire qu'il sert de foyer collectif pour toutes les données d'analyse dans une organisation.

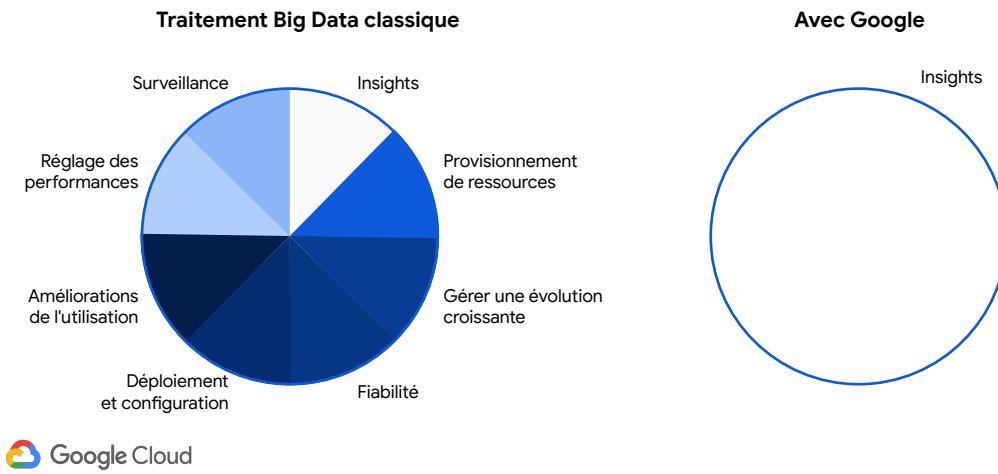
Les ensembles de données sont des collections de tables pouvant être divisées selon les lignes de l'entreprise ou un domaine d'analyse donné. Chaque ensemble de données est lié à un projet GCP.

Un data lake peut contenir des fichiers dans Cloud Storage ou Google Drive, ou des données transactionnelles dans Cloud Bigtable. BigQuery peut définir un schéma et émettre des requêtes directes sur des données externes comme sources de données fédérées.

Les tables et vues de base de données fonctionnent de la même manière dans BigQuery que dans un data warehouse classique, permettant à BigQuery de prendre en charge les requêtes rédigées dans un dialecte SQL standard, conforme à ANSI : 2011.

Cloud Identity and Access Management est utilisé pour octroyer des autorisations permettant d'effectuer des actions spécifiques dans BigQuery. Cela remplace les déclarations SQL OCTROYER et RÉVOQUER utilisées pour gérer les autorisations d'accès dans les bases de données SQL classiques.

Cloud permet aux ingénieurs données de passer moins de temps à gérer le matériel et l'évolutivité ; laissez Google s'en charger pour vous



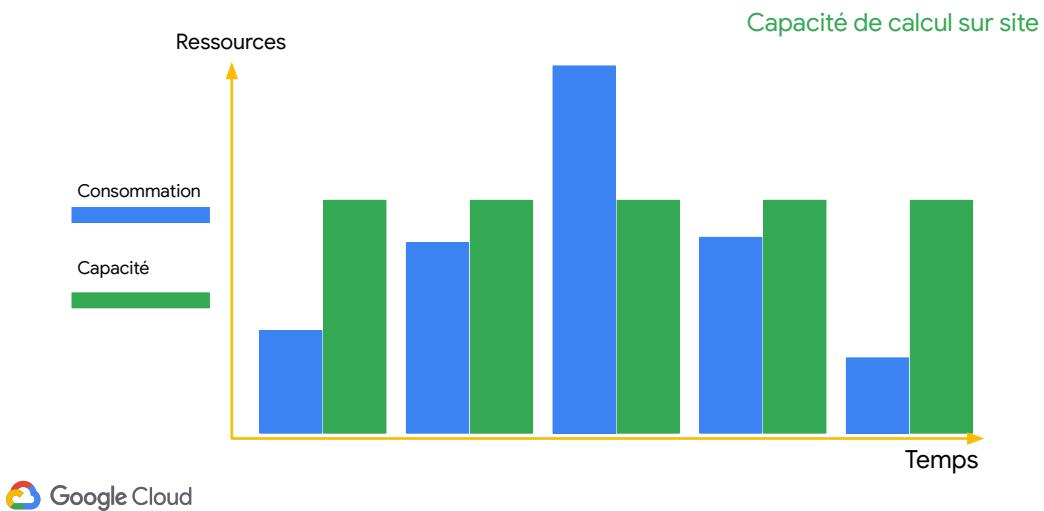
Un élément clé à prendre en considération en plus de l'agilité est la capacité à en faire plus avec moins. Il est important de vous assurer que vous n'effectuez rien qui n'apporte pas de la valeur.

Si vous effectuez des tâches communes à plusieurs secteurs, ce n'est probablement pas quelque chose pour lequel votre entreprise souhaite payer.

Le cloud vous permet, en tant qu'ingénieur données, de passer moins de temps à gérer le matériel et plus de temps à réaliser des actions plus personnalisées et spécifiques à l'entreprise.

Inutile de vous préoccuper de l'approvisionnement et des améliorations en termes de fiabilité et d'utilisation des performances ou de l'ajustement du cloud. Vous pouvez donc vous concentrer sur la manière d'obtenir de meilleurs insights à partir de vos données.

Vous n'avez pas besoin de provisionner des ressources avant d'utiliser BigQuery



Vous n'avez pas besoin de provisionner des ressources avant d'utiliser BigQuery, contrairement à de nombreux systèmes RDBMS. BigQuery alloue des ressources de stockage et de requête de manière dynamique en fonction de vos habitudes d'utilisation.

Les ressources de stockage sont allouées lorsque vous les utilisez et désallouées lorsque vous retirez des données ou déposez des tables.

Les ressources de requête sont allouées conformément au type et à la complexité de la requête. Chaque requête utilise un certain nombre d'emplacements, qui sont des unités informatiques comprenant un certain volume de processeur et de mémoire RAM.

Vidéo : 1_14

Data Lakes et Data Warehouses

Présentateur : Lak

[FIN DE LA VIDÉO]

Programme

Découvrez le rôle de l'ingénieur données

Analysez les défis de l'ingénierie des données

Introduction à BigQuery

Data Lakes et Data Warehouses

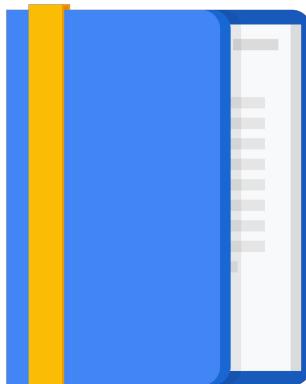
Bases de données transactionnelles vs Data Warehouses

Travailler efficacement avec les autres équipes chargées des données

- Gérer l'accès et la gouvernance des données
- Créer des pipelines prêts à l'emploi

Passer en revue les études de cas des clients GCP

Atelier : Analyser les données à l'aide de BigQuery



Nous avons défini ce qu'était un data lake et ce qu'était un data warehouse.
Examinons-les plus en détail.

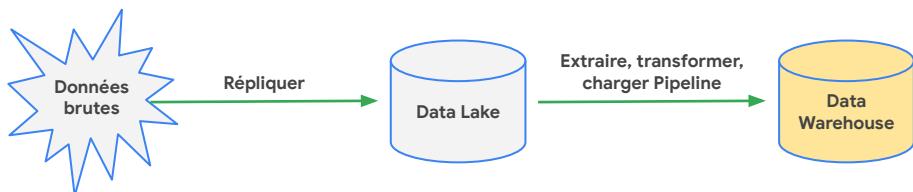
Un ingénieur données obtient des données utilisables



Rappelez-vous : nous avons mis l'accent sur le fait que les données doivent être utilisables afin qu'une personne puisse les utiliser pour prendre des décisions. Bien souvent, les données brutes mêmes ne sont pas **très utiles**.

Un data warehouse stocke des données transformées en un état utilisable pour obtenir des insights d'entreprise

Quels sont les éléments clés à prendre en considération avant de décider parmi les options de data warehouse ?



Nous avons dit que les données brutes sont répliquées et stockées dans un data lake.

Vous utiliserez des pipelines Extraire-Transformer-Charger ou ETL (Extract-Transform-Load) pour rendre ces données utilisables et les stocker dans un data warehouse.

Voyons quels sont les éléments clés à prendre en considération avant de décider parmi les options de data warehouse...

Éléments clés à prendre en considération lorsque vous choisissez un data warehouse

- Les éléments clés à prendre en considération lorsque vous choisissez un data warehouse sont les suivants :
- Peut-il servir de récepteur pour les pipelines de données par lots et par flux ?
- Le data warehouse peut-il évoluer pour répondre à mes besoins ?
- Comment sont organisées et cataloguées les données ? L'accès est-il contrôlé ?
- Le warehouse est-il conçu pour garantir un niveau élevé de performance ?
- Quel niveau de maintenance devra fournir notre équipe d'ingénieurs ?



Nous devons nous poser les questions suivantes :

Le data warehouse va servir de récepteur, vous allez y stocker des données. Mais va-t-il être alimenté par un pipeline par lots ou par un pipeline par flux ? L'entrepôt doit-il être correct à la minute près ? Ou est-ce suffisant d'y charger des données une fois par jour ou une fois par semaine ?

Le data warehouse évoluera-t-il pour répondre à mes besoins ? De nombreux data warehouses basés sur des clusters définiront des limites de requêtes simultanées par cluster. Ces limites de requêtes entraîneront-elles un problème ? Le cluster sera-t-il suffisamment grand pour stocker et transporter vos données ?

Comment sont organisées et cataloguées les données ? L'accès est-il contrôlé ? Pourrez-vous partager l'accès aux données avec toutes les personnes concernées ? Que se passe-t-il si elles souhaitent interroger les données ? Qui paiera pour cette requête ?

L'entrepôt est-il conçu pour garantir un niveau élevé de performance ? Là encore, pensez soigneusement aux performances de requêtes simultanées. Et si ces performances sont directes ou si vous devez les personnaliser en créant des indexées et en ajustant le data warehouse.

Enfin, quel niveau de maintenance devra fournir notre équipe d'ingénieurs ?

BigQuery est un data warehouse moderne qui bouscule le mode conventionnel d'entreposage des données



Les data warehouses classiques sont difficiles à gérer et à faire fonctionner. Ils sont conçus pour un paradigme en lots d'analyses de données et pour les besoins en rapports opérationnels. Les données dans le data warehouse sont destinées à être utilisées par quelques gestionnaires à des fins de création de rapports. BigQuery est un data warehouse moderne qui bouscule le mode conventionnel d'entreposage des données. Ici, nous pouvons voir certains des éléments clés de comparaison entre un data warehouse classique et BigQuery.

BigQuery offre des mécanismes pour un transfert automatisé des données et alimente les applications d'entreprise à l'aide de technologies que les équipes connaissent et utilisent déjà. Ainsi, tout le monde a accès aux insights de données. Vous pouvez créer des sources de données partagées en lecture seule que les utilisateurs internes et externes peuvent interroger, et rendre les résultats accessibles à tout le monde grâce à des outils conviviaux tels que Google Sheets, Tableau, Qlik ou Google Data Studio.

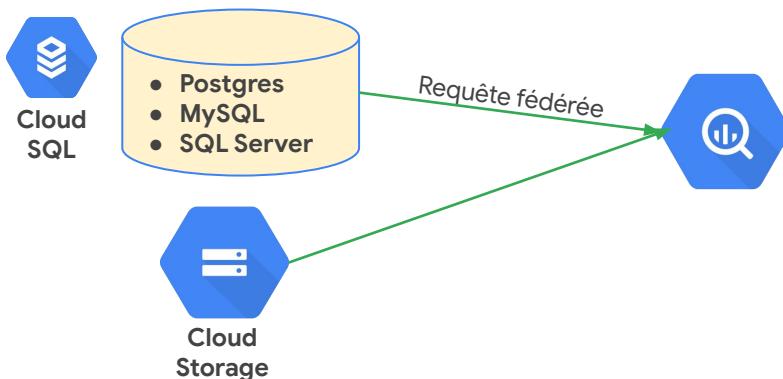
BigQuery établit la base pour l'IA. Il est possible d'entraîner les modèles Tensorflow et Google machine learning directement à l'aide des data warehouses stockés dans BigQuery, et BigQuery ML peut être utilisé pour créer et entraîner des modèles de machine learning à l'aide d'un SQL simple. Autre capacité avancée : BigQuery GIS. Elle permet aux organisations d'analyser des données géographiques dans BigQuery, essentielles à de nombreuses décisions d'entreprise importantes qui reposent sur les données de lieu.

BigQuery permet également aux organisations d'analyser les événements d'entreprise en temps réel, à mesure qu'ils se déroulent, en ingérant automatiquement des données et en les rendant immédiatement disponibles pour requête dans leur data warehouse. Cette fonctionnalité est possible grâce à la capacité de BigQuery à ingérer jusqu'à 100 000 lignes de données par secondes et à rechercher des pétaoctets de données à une vitesse fulgurante.

Grâce à l'infrastructure sans serveur entièrement gérée de Google et au réseau mondiallement disponible, BigQuery élimine le travail associé au provisionnement et à la maintenance d'une infrastructure d'entreposage de données classique.

BigQuery simplifie également les opérations de données grâce à l'utilisation d'Identity and Access Management pour contrôler l'accès aux ressources par les utilisateurs, en créant des rôles et des groupes et en assignant des autorisations pour les tâches et les requêtes en cours dans un projet, et en fournissant également des sauvegardes et des réplications des données automatiques.

Vous pouvez simplifier les pipelines ETL du data warehouse grâce aux connexions externes à Cloud Storage et Cloud SQL



Bien que nous ayons parlé de l'intégration de données dans BigQuery en exécutant des pipelines ETL, il existe une autre option.

Il s'agit de traiter BigQuery comme un moteur de requêtes et de lui permettre de rechercher les données sur place.

Par exemple, vous pouvez utiliser BigQuery pour interroger directement les données d'une base de données dans Cloud SQL, c'est-à-dire des bases de données relationnelles gérées telles que Postgres, MySQL et SQL Server.

Vous pouvez également utiliser BigQuery pour interroger directement des fichiers sur Cloud Storage à condition que ces fichiers soient dans un format tel que CSV ou Parquet.

Le véritable atout est de laisser des données sur place et de parvenir quand même à les lier à d'autres données dans votre data warehouse.

Regardons une vidéo pour en savoir plus.

Vidéo : 1_15

Démonstration : Requêtes fédérées avec BigQuery

Présentateur : Lak

[FIN DE LA VIDÉO]

Démons- tration

Requêtes fédérées avec BigQuery

Instructions de la démonstration :

<https://github.com/GoogleCloudPlatform/training-data-analyst/blob/master/courses/data-to-insights/demos/external-data-query.sql>

Vidéo : 1_16

Bases de données transactionnelles vs Data Warehouses

Présentateur : Lak

[FIN DE LA VIDÉO]

Programme

Découvrez le rôle de l'ingénieur données

Analysez les défis de l'ingénierie des données

Introduction à BigQuery

Data Lakes et Data Warehouses

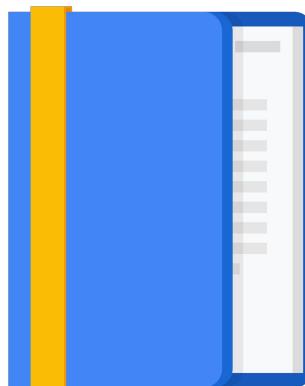
Bases de données transactionnelles vs Data Warehouses

Travailler efficacement avec les autres équipes chargées des données

- Gérer l'accès et la gouvernance des données
- Créer des pipelines prêts à l'emploi

Passer en revue les études de cas des clients GCP

Atelier : Analyser les données à l'aide de BigQuery



Les ingénieurs données sont responsables des systèmes de base de données transactionnels backend qui prennent en charge les applications de votre entreprise ET les data warehouses qui prennent en charge vos charges de travail analytiques. Dans ce module, nous verrons la différence entre les bases de données et les data warehouses et les solutions Google Cloud pour chaque charge de travail.

Cloud SQL est entièrement géré SQL Server, Postgres ou MySQL pour votre base de données relationnelle (RDBMS transactionnelle)

- Chiffrement automatique
- Capacité de stockage de 30 To
- 60 000 IOPS (lecture/écriture par seconde)
- Auto-scale et sauvegarde automatiques

Pourquoi ne pas utiliser Cloud SQL pour créer les rapports des flux de travail ?



Si vous utilisez SQL Server, MySQL ou Postgres comme base de données relationnelle, vous pouvez migrer vers Cloud SQL qui est la solution de base de données relationnelle entièrement gérée de Google Cloud.

Cloud SQL est une solution hautes performances et évolutive qui offre jusqu'à 30 To d'espace de stockage, 60 000 IOPS et 416 Go de mémoire RAM par instance. Vous pouvez profiter de l'évolutivité automatique du stockage pour gérer les besoins grandissants de votre base de données sans aucune interruption.

Vous pouvez vous demander : "Pourquoi ne pas utiliser Cloud SQL pour créer les rapports des flux de travail ? Il est possible d'exécuter SQL directement sur la base de données, n'est-ce pas ?"

<https://cloud.google.com/products/databases/>

Les RDBMS sont optimisés pour les données provenant d'une seule source et les entrepôts avec écriture haut-débit/lecture fréquente des données



Cloud SQL

Votre architecture finale contiendra généralement une base de données et un data warehouse



BigQuery

- Évolutif en Go et To
- Idéal pour les applications de base de données back-end
- Stockage en dossiers
- Évolutif en Po
- Se connecte facilement aux sources de données externes pour ingestion
- Stockage en colonnes

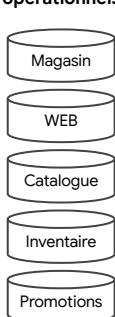


C'est une très bonne question et nous y reviendrons plus en détail dans notre module sur les data warehouses. Pour l'instant, sachez que Cloud SQL est optimisé pour être une base de données destinée aux transactions (écritures) et BigQuery est un data warehouse optimisé pour la création de rapport de charges de travail (principalement lectures). L'architecture fondamentale de ces options de stockage de données est assez différente. Les bases de données Cloud SQL sont un stockage organisé en dossiers, ce qui signifie que le dossier doit être ouvert dans son intégralité sur le disque, même si vous avez sélectionné une seule colonne dans votre requête. BigQuery est un stockage en colonne qui, comme vous le devinez, offre des schémas de création de rapports très larges puisque vous pouvez simplement lire des colonnes individuelles depuis le disque.

Les systèmes de gestion de base de données relationnelle (RDBMS) sont essentiels pour gérer de nouvelles transactions

Les RDBMS sont optimisés pour l'**ÉCRITURE** haut débit dans des DOSSIERS

Systèmes opérationnels



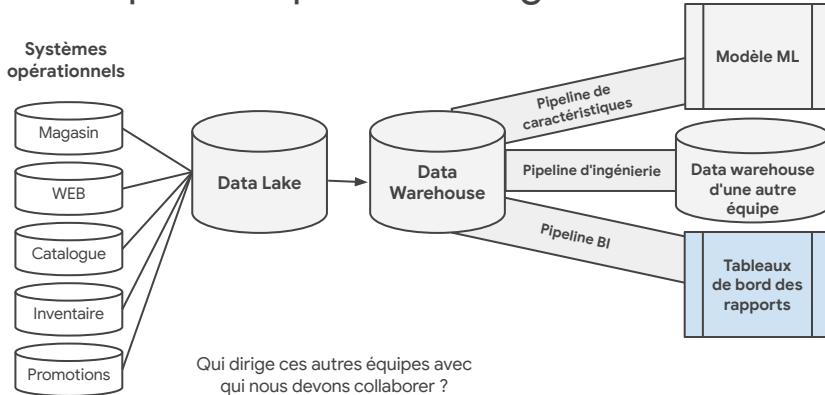
 Google Cloud

Nous ne disons pas que les RDBMS ne sont pas aussi performants que les data warehouses, simplement qu'ils servent des fins différentes. Les RDBMS aident votre entreprise à gérer de nouvelles transactions. Prenons l'exemple de ce terminal de point de vente au niveau d'une vitrine. Les commandes et produits commandés sont généralement inscrits dans un nouveau dossier quelque part dans une base de données relationnelle. Cette base de données peut stocker toutes les commandes reçues depuis leur site Web, tous les produits répertoriés dans le catalogue, ou le nombre d'articles dans leur inventaire.

Ainsi, lorsqu'une commande existante est modifiée, elle peut être rapidement mise à jour dans la base de données. Les systèmes transactionnels permettent de modifier une ligne individuelle dans la table de base de données de manière cohérente. Ils sont également conçus selon certains principes de base de données relationnelle, comme l'intégrité référentielle, pour assurer une protection dans certaines situations (un client commande un produit qui n'existe pas dans la table de produits, par exemple).

Où finissent donc toutes ces données brutes dans notre discussion sur les data lakes et les data warehouses ? Quelle est la vision d'ensemble ?

Tableau complet : Les données sources entrent dans le data lake, sont traitées dans le data warehouse et mises à disposition pour les insights



La voici. Nos systèmes opérationnels, tels que nos bases de données relationnelles qui stockent les commandes en ligne, l'inventaire et les promotions, sont nos sources de données brutes sur la gauche. À noter qu'ils ne sont pas exhaustifs, vous pouvez avoir d'autres systèmes sources manuels tels que des fichiers CSV ou des feuilles de calcul.

Ces sources de données en amont sont regroupées à un seul endroit consolidé dans notre data lake conçu pour être durable et hautement disponible.

Une fois dans le lac, les données doivent souvent être traitées via des transformations qui les placent dans notre data warehouse où elles sont prêtes à être utilisées par des équipes en aval. Voici trois exemples rapides d'autres équipes qui créent souvent des pipelines sur notre data warehouse :

- Une équipe ML qui crée un pipeline dans le but d'obtenir des caractéristiques pour ses modèles.
- Une équipe d'ingénierie qui peut utiliser nos données dans le cadre de son data warehouse.
- Et une équipe BI qui peut vouloir créer un tableau de bord à l'aide de certaines de nos données.

Qui travaille dans ces équipes et comment collaborent-elles avec notre équipe d'ingénierie des données ?

Vidéo : 1_I7

Travailler efficacement avec les autres équipes
chargées des données

Présentateur : Lak

[FIN DE LA VIDÉO]

Programme

Découvrez le rôle de l'ingénieur données

Analysez les défis de l'ingénierie des données

Introduction à BigQuery

Data Lakes et Data Warehouses

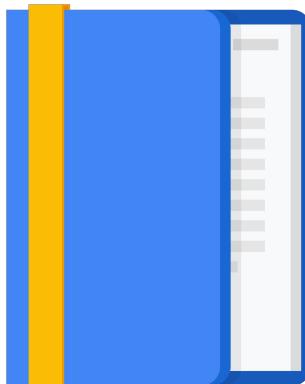
Bases de données transactionnelles vs Data Warehouses

Travailler efficacement avec les autres équipes chargées des données

- Gérer l'accès et la gouvernance des données
- Créer des pipelines prêts à l'emploi

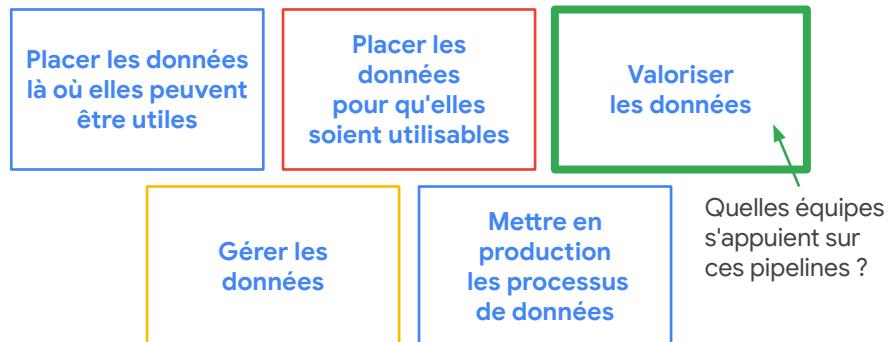
Passer en revue les études de cas des clients GCP

Atelier : Analyser les données à l'aide de BigQuery



Puisqu'un data warehouse sert également à d'autres équipes, il est essentiel d'apprendre comment donner accès au data warehouse tout en respectant les bonnes pratiques de gouvernance des données.

Un ingénieur données crée des pipelines de données pour permettre la prise de décisions basées sur les données



Souvenez-vous qu'une fois que les données se trouvent là où elles peuvent être utiles et dans un format utilisable, nous devons apporter de la valeur aux données à travers les analyses et le machine learning. Quelles équipes peuvent s'appuyer sur vos données ?

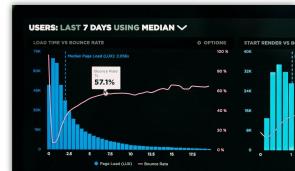
De nombreuses équipes s'appuient sur ces partenariats avec une ingénierie des données pour tirer profit de leurs données



Ingénieur machine learning



Analyste de données



Ingénieur données

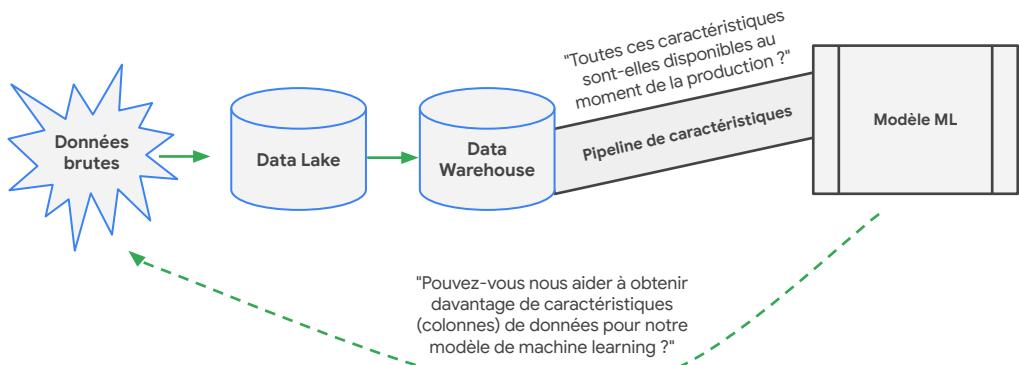
Comment chacune de ces équipes s'appuie-t-elle sur l'ingénierie des données ?



De nombreuses équipes de données peuvent s'appuyer sur votre data warehouse et collaborer avec des ingénieurs données pour créer et maintenir de nouveaux pipelines de données. Les trois clients les plus courants sont (1) l'ingénieur machine learning (2) l'analyste de données ou BI et (3) d'autres ingénieurs données.

Voyons comment chacun de ces rôles interagit avec votre nouveau data warehouse et comment les ingénieurs données peuvent collaborer au mieux avec vous.

Les équipes de machine learning ont besoin des ingénieurs données pour les aider à capturer de nouvelles caractéristiques dans un pipeline stable

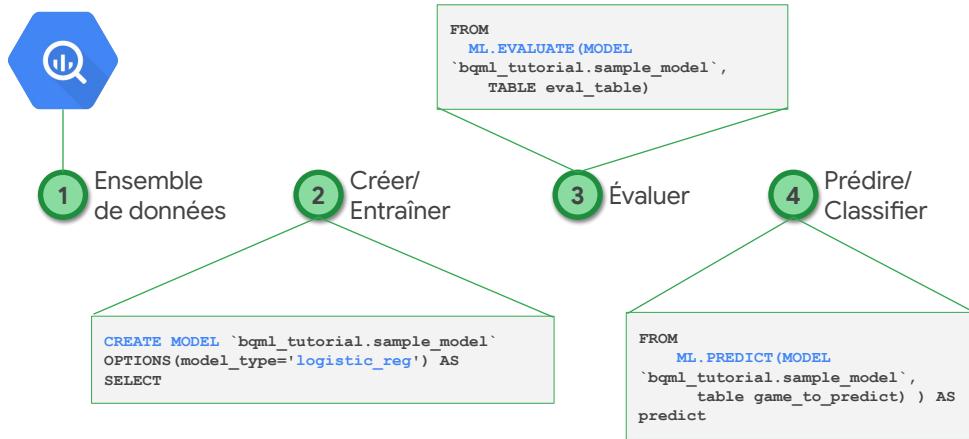


Comme vous le verrez dans notre cours sur le machine learning, les modèles d'une équipe de machine learning s'appuient sur l'obtention de nombreuses données d'excellente qualité pour créer, entraîner, tester et évaluer leurs modèles. L'équipe de machine learning collabore souvent avec des équipes d'ingénierie des données pour créer des pipelines et des ensembles de données à utiliser dans ses modèles.

Une question que l'on peut souvent vous poser est : "combien de temps faut-il à une transaction pour passer des données brutes au data warehouse ?" Cette question vient du fait que toute donnée utilisée pour entraîner des modèles doit également être disponible au moment de la prédiction. Si le délai de collecte et d'agrégation des données brutes est long, cela aura un impact sur la capacité de l'équipe ML à créer des modèles utiles.

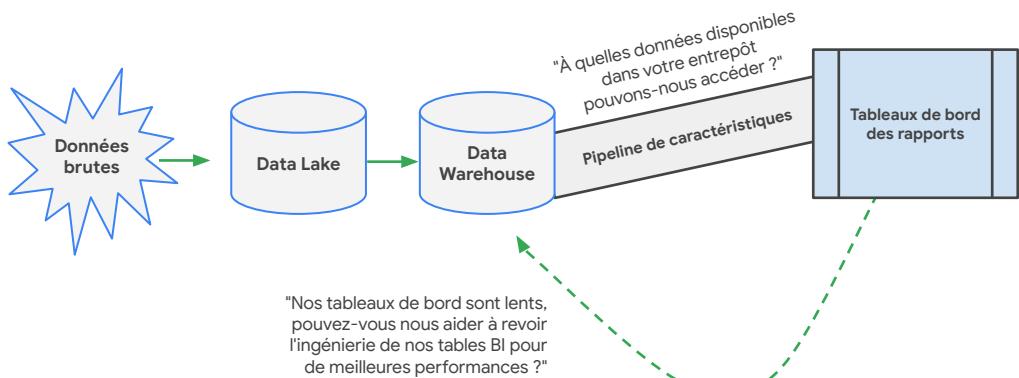
Une seconde question que vous entendrez forcément est : "à quel point serait-ce difficile d'ajouter davantage de colonnes ou de lignes à certains ensembles de données ?" Là encore, l'équipe ML s'appuie sur l'établissement de relations entre les colonnes de données et un riche historique pour entraîner les modèles. Vous gagnerez la confiance des équipes ML partenaires en rendant vos ensembles de données facilement trouvables, documentés et disponibles pour que les équipes ML puissent rapidement faire leurs expériences.

Valeur ajoutée : le machine learning directement dans BigQuery



Une caractéristique unique de BigQuery est la possibilité de créer des modèles de machine learning haute performance directement dans BigQuery en utilisant uniquement SQL grâce à BigQuery ML. Voici le code de modèle actuel pour CRÉER un modèle, l'ÉVALUER et FAIRE des prédictions. Vous le verrez également dans nos cours sur le machine learning ultérieurement.

Les équipes d'analyse des données et de veille stratégique s'appuient sur l'ingénierie des données pour mettre en avant les derniers insights

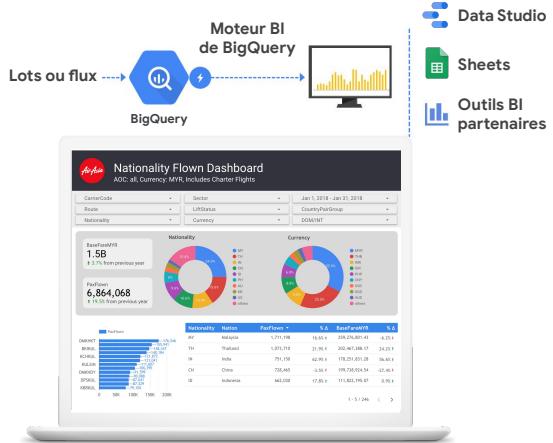


Autre partenaire important : les équipes de veille stratégique et d'analystes de données qui s'appuient sur de bonnes données propres pour interroger des insights et créer des tableaux de bord.

Ces équipes ont besoin d'ensembles de données dotés de schémas clairement définis, de la possibilité d'afficher rapidement les lignes et d'un haut niveau de performance pour s'adapter à plusieurs utilisateurs de tableaux de bord simultanés.

Valeur ajoutée : moteur BI pour les performances du tableau de bord

- Nul besoin de gérer des cubes OLAP ou des serveurs BI distincts pour les performances du tableau de bord
- Intégration native au flux BigQuery pour une actualisation des données en temps réel
- Moteur d'exécution BI en mémoire organisé en colonnes

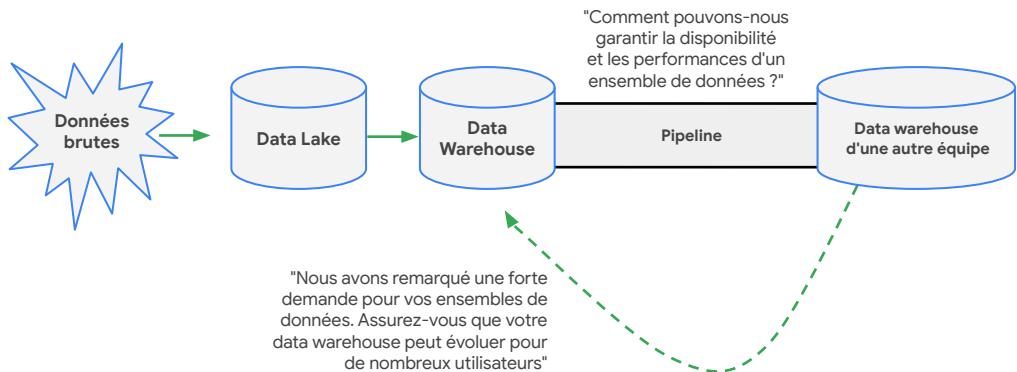


L'un des produits Google Cloud qui permet de gérer les performances des tableaux de bord est BigQuery BI Engine, actuellement en bêta au moment de l'enregistrement. [BI Engine](#) est un service d'analyse rapide et en mémoire, créé directement dans BigQuery et disponible pour accélérer vos applications de veille stratégique.

Par le passé, les équipes BI devaient créer, gérer et optimiser leurs propres serveurs BI et cubes OLAP pour prendre en charge les applications de rapports. À présent, BI Engine vous permet d'obtenir un temps de réponse très court sur vos ensembles de données BigQuery sans avoir à créer vos propres cubes. BI Engine repose sur le même stockage que BigQuery et calcule l'architecture et les serveurs en tant que service de cache intelligent en mémoire rapide qui maintient l'état.

<https://www.youtube.com/watch?v=TqlrlcmqPgo>

D'autres équipes d'ingénierie des données peuvent s'appuyer sur le fait que vos pipelines sont actuels et exempts d'erreur



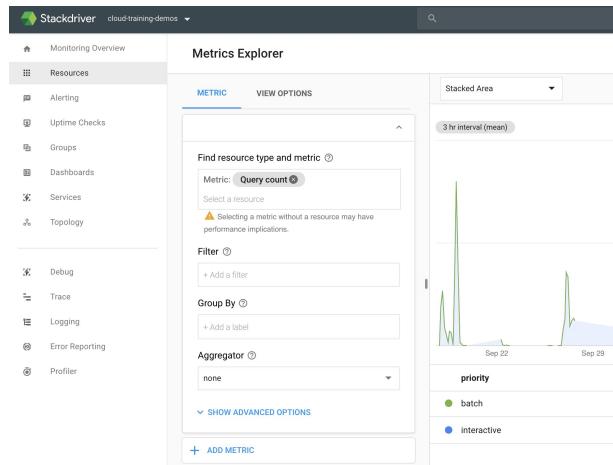
Les derniers partenaires de votre équipe d'ingénierie des données sont les autres ingénieurs données qui s'appuient sur la disponibilité et les performances de votre data warehouse et vos pipelines pour leurs data lakes et data warehouses en aval.

Ils demanderont souvent : "comment pouvons-nous garantir que vos pipelines de données dont nous dépendons seront toujours disponibles quand nous en aurons besoin ?"

Ou "nous avons remarqué une demande élevée pour certains ensembles de données populaires, comment pouvez-vous surveiller et faire évoluer la santé de votre écosystème de données complet ?"

Valeur ajoutée : surveillance des performances Stackdriver

- Afficher les requêtes en cours et terminées
- Suivre les dépenses sur les ressources BigQuery
- Utiliser les journaux d'audit Cloud pour consulter les informations de la tâche en cours (qui l'a exécutée, quelle requête a été effectuée)
- Créer des alertes et envoyer des notifications



Une manière courante est d'utiliser la fonction Stackdriver Monitoring intégrée pour toutes les ressources sur Google Cloud Platform. Puisque Google Cloud Platform et BigQuery sont des ressources, vous pouvez configurer des alertes et notifications pour les métriques telles que "Nombre de requêtes" ou "Octets de données traités" afin de mieux suivre l'utilisation et les performances.

Stackdriver est utilisé pour deux autres raisons : suivre les dépenses de toutes les différentes ressources utilisées et connaître les tendances de facturation de votre équipe ou organisation. Enfin, vous pouvez utiliser les journaux d'audit Cloud pour afficher les informations de tâche de requête actuelle afin de voir les détails précis concernant le type de requêtes exécutées et les personnes qui les ont lancées. Cette fonction est utile si vous détenez des ensembles de données sensibles qui doivent être surveillés minutieusement. Nous approfondirons le sujet une prochaine fois.

<https://cloud.google.com/bigquery/docs/monitoring?hl=fr>

Vidéo : 1_I8

Gérer l'accès et la gouvernance des données

Présentateur : Lak

[FIN DE LA VIDÉO]

Programme

Découvrez le rôle de l'ingénieur données

Analysez les défis de l'ingénierie des données

Introduction à BigQuery

Data Lakes et Data Warehouses

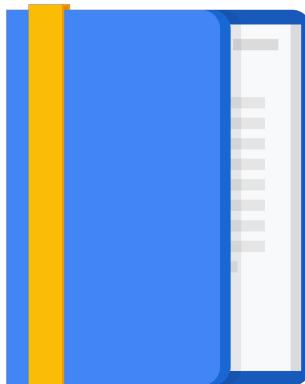
Bases de données transactionnelles vs Data Warehouses

Travailler efficacement avec les autres équipes chargées des données

- Gérer l'accès et la gouvernance des données
- Créer des pipelines prêts à l'emploi

Passer en revue les études de cas des clients GCP

Atelier : Analyser les données à l'aide de BigQuery



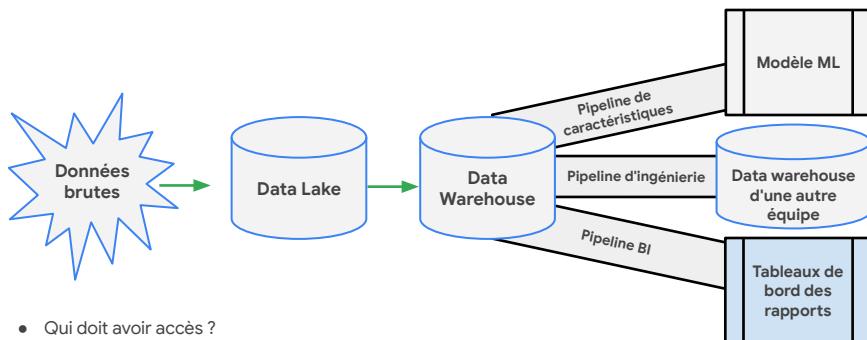
Afin d'être un partenaire efficace, votre équipe d'ingénierie devra configurer des stratégies d'accès aux données et une gouvernance globale sur la manière d'utiliser et de ne PAS utiliser les données par vos utilisateurs.

Un ingénieur données gère l'accès aux données et la gouvernance



C'est ce que nous voulons dire quand nous disons qu'un ingénieur données doit gérer les données. Cela inclut des sujets essentiels tels que la confidentialité et la sécurité. Quels sont les éléments clés à prendre en considération lors de la gestion de certains ensembles de données ?

L'ingénierie des données doit définir et communiquer un modèle de gouvernance des données responsable



- Qui doit avoir accès ?
- Comment sont traitées les informations personnelles ?
- Comment informer les utilisateurs finaux à propos de notre catalogue de données ?



Communiquer clairement un modèle de gouvernance des données :

Qui peut avoir accès et qui ne doit pas avoir accès ?

Comment sont traitées les informations personnellement identifiables (telles que les numéros de téléphone ou les adresses électroniques) ?

Et des tâches encore plus basiques : de quelle manière nos utilisateurs finaux découvrent-ils les différents ensembles de données que nous avons pour analyse ?

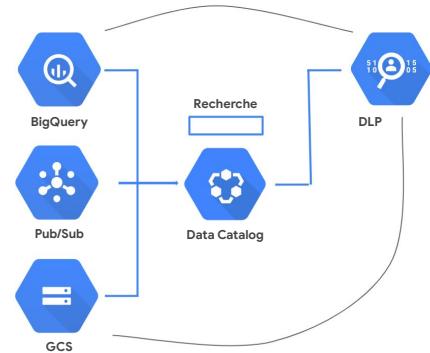
Cloud Data Catalog est une API de recherche de données gérées et de protection contre la perte de données visant à protéger les informations personnelles



La recherche de données simplifiée à n'importe quelle échelle :
Service de gestion de métadonnées entièrement géré sans aucune infrastructure à configurer ou à gérer

Une vue unifiée de tous les ensembles de données :
Catalogue de données centralisé et sécurisé dans Google Cloud gérant la capture des métadonnées et l'ajout de tags

Fondation de la gouvernance de données :
Conformité en matière de sécurité avec des contrôles du niveau d'accès et l'intégration de la protection contre la perte de données (DLP) Cloud pour gérer les données sensibles



Une solution pour la gouvernance des données est le Cloud Data Catalog et l'API de protection contre la perte de données.

Le Data Catalog met toutes les métadonnées de vos ensembles de données à disposition pour que vos utilisateurs puissent faire des recherches. Vous groupez les ensembles de données avec des balises, indiquez certaines colonnes comme sensibles, etc.

Pourquoi est-ce utile ? Si vous disposez de nombreux ensembles de données différents avec diverses tables, pour lesquels différents utilisateurs disposent de différents niveaux d'accès, le Data Catalog offre une expérience utilisateur unifiée pour découvrir rapidement ces ensembles de données. Plus besoin de d'abord chercher des noms de tables spécifiques dans SQL.

Souvent utilisée conjointement au Data Catalog, l'API de protection contre la perte de données (ou API DLP) vous aide à mieux comprendre et gérer les données sensibles. Elle offre une classification rapide et évolutive, ainsi qu'une rédaction pour les éléments de données sensibles tels que les numéros de carte de crédit, les noms, les numéros de sécurité sociale, les numéros d'identification américains ou internationaux sélectionnés, les numéros de téléphone et les identifiants GCP.

Regardons une courte démonstration de cette API.

RESSOURCES :

Protéger les informations personnelles dans de grands ensembles de données

https://www.youtube.com/watch?time_continue=461&v=BbFCbWln-as

DLP = protection contre la perte de données

Cloud DLP <https://cloud.google.com/dlp/demo/#/>

Vidéo : 1_19

Démonstration : Trouver des informations personnelles
dans votre ensemble de données avec l'API DLP

Présentateur : Lak

[FIN DE LA VIDÉO]

Démons -stration

Trouver des informations personnelles dans votre ensemble de données avec l'API DLP

Instructions de la démonstration : <https://cloud.google.com/dlp/demo/#/>

Vidéo : 1_l10

Créer des pipelines prêts à l'emploi

Présentateur : Lak

[FIN DE LA VIDÉO]

Programme

Découvrez le rôle de l'ingénieur données

Analysez les défis de l'ingénierie des données

Introduction à BigQuery

Data Lakes et Data Warehouses

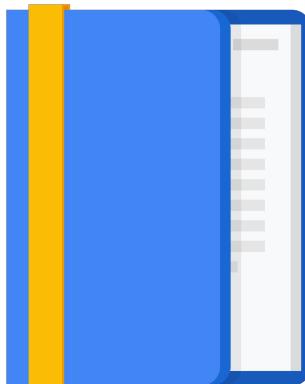
Bases de données transactionnelles vs Data Warehouses

Travailler efficacement avec les autres équipes chargées des données

- Gérer l'accès et la gouvernance des données
- Créer des pipelines prêts à l'emploi

Passer en revue les études de cas des clients GCP

Atelier : Analyser les données à l'aide de BigQuery



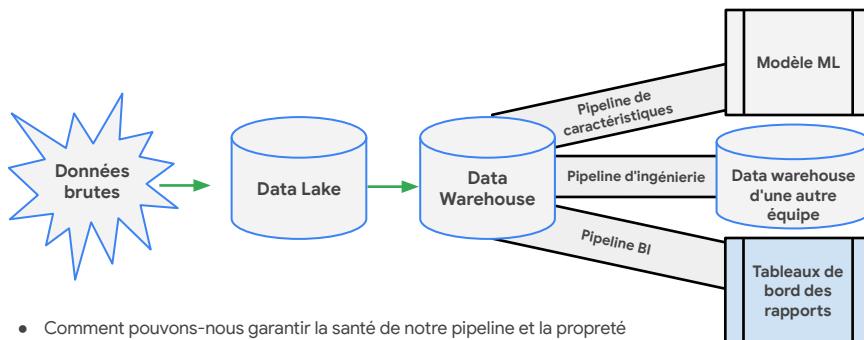
Une fois que vos data lakes et data warehouses sont configurés et que votre stratégie de gouvernance est en place, il est temps de mettre en place toute l'opération, et d'en automatiser et surveiller autant que possible.

Un ingénieur données crée des pipelines de données de production pour permettre la prise de décisions basées sur les données



Voilà ce que nous voulons dire lorsque nous parlons de mettre en place le processus des données. Il doit s'agir d'un système de traitement des données complet et évolutif.

L'ingénierie des données dirige la santé et le futur de ses pipelines de données de production



- Comment pouvons-nous garantir la santé de notre pipeline et la propreté de nos données ?
- Comment mettre en place ces pipelines pour réduire la maintenance et optimiser la disponibilité ?
- Comment répondre et s'adapter à ces schémas changeants et aux besoins de l'entreprise ?
- Utilisons-nous les outils d'ingénierie des données et les meilleures pratiques les plus récents ?

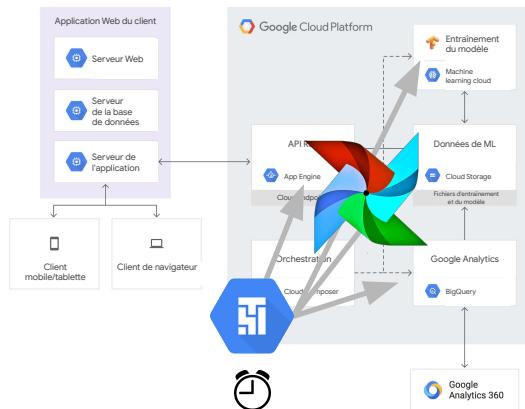


Votre équipe d'ingénierie des données est responsable de la santé des tuyaux (les pipelines) et de garantir que les données sont disponibles et à jour pour les analyses et charges de travail ML.

Les questions courantes que vous devriez poser à cette étape sont les suivantes :

- Comment pouvons-nous garantir la santé de notre pipeline et la propreté de nos données ?
- Comment mettre en place ces pipelines pour réduire la maintenance et optimiser la disponibilité ?
- Comment répondre et s'adapter à ces schémas changeants et aux besoins de l'entreprise ?
- Utilisons-nous les outils d'ingénierie des données et les meilleures pratiques les plus récents ?

Cloud Composer (Apache Airflow géré) orchestre les flux de travail de production



Google Cloud

Apache Airflow est un outil d'orchestration de flux de travail communs utilisés par les entreprises. Google Cloud dispose d'une version entièrement gérée appelée Cloud Composer.

Cloud Composer aide votre équipe d'ingénierie des données à orchestrer toutes les pièces du puzzle d'ingénierie des données dont nous avons parlé jusqu'à présent (et bien plus que vous devez encore découvrir !). Par exemple, lorsqu'un nouveau fichier CSV est déposé dans Cloud Storage, vous pouvez faire en sorte que cela déclenche automatiquement un événement qui lance un flux de travail de traitement des données et place ces données directement dans votre data warehouse.

La puissance de cet outil vient du fait que les produits et services Big Data GCP disposent de points de terminaison API que vous pouvez appeler. Une tâche Cloud Composer peut alors être exécutée toutes les nuits ou toutes les heures et amorcer la totalité de votre pipeline, des données brutes au data lake, et dans le data warehouse.

Nous détaillerons l'orchestration du flux de travaux dans des modules ultérieurs et vous participerez également à un atelier sur Cloud Composer.

Vidéo : 1_l11

Passer en revue les études de cas des clients GCP

Présentateur : Lak

[FIN DE LA VIDÉO]

Programme

Découvrez le rôle de l'ingénieur données

Analysez les défis de l'ingénierie des données

Introduction à BigQuery

Data Lakes et Data Warehouses

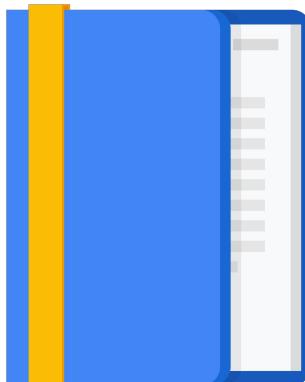
Bases de données transactionnelles vs Data Warehouses

Travailler efficacement avec les autres équipes chargées des données

- Gérer l'accès et la gouvernance des données
- Créer des pipelines prêts à l'emploi

Passer en revue les études de cas des clients GCP

Atelier : Analyser les données à l'aide de BigQuery



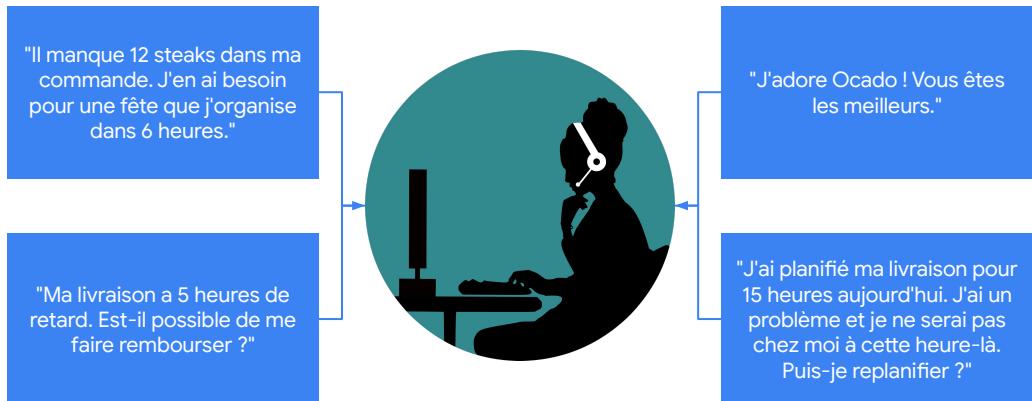
Nous avons vu de nombreux aspects du travail d'un ingénieur données.

Voyons à présent un cas d'étude sur la manière dont un client Google Cloud résout un problème d'entreprise spécifique.

Cela permettra de lier tous ces différents aspects.

Le service client d'Ocado ne cesse de recevoir des messages

Pouvons-nous utiliser le ML pour prioriser ces messages ?



Ocado est le plus grand grossiste en ligne au monde.

Le service client d'Ocado reçoit un grand nombre de messages. Il peut recevoir un message disant "Il manque 12 articles à ma commande et j'en ai besoin pour une fête que j'organise dans 6 heures" ou un autre message qui dit "J'adore Ocado, vous êtes les meilleurs" et un troisième "Ma livraison a quelques heures de retard, est-il possible de me faire rembourser ?". Ou encore un autre message qui dit "J'ai planifié une livraison pour 15 heures aujourd'hui ; je dois la replanifier". Lesquels ont la plus grande priorité ?

L'agent du service client doit lire ces messages pour le savoir, n'est-ce pas ? Mais à ce niveau, il peut tout autant traiter le message...

Alors, comment optimiser le traitement des messages au service client pour savoir quels sont les messages ayant la plus grande priorité ? C'est le problème que rencontrait l'équipe d'ingénierie des données d'Ocado.

<https://pixabay.com/fr/vectors/call-customer-support-woman-3613071/>

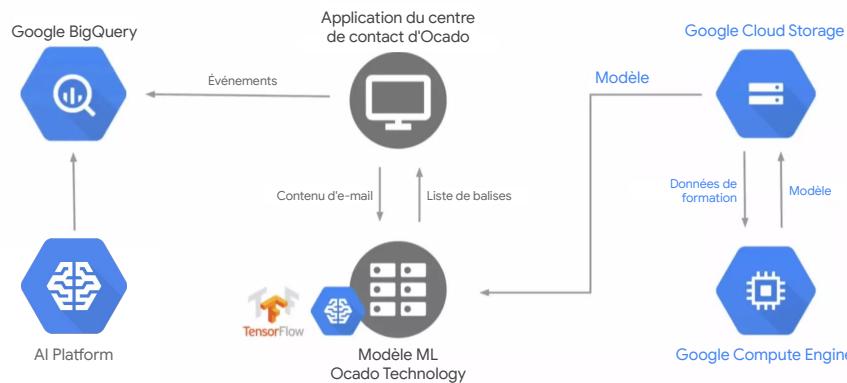
<formateur>

<https://youtu.be/kGVDFzlbhco?t=1451>

</formateur>

La solution GCP d'Ocado leur permet de répondre aux e-mails urgents des clients 4 fois plus vite grâce au ML

L'amélioration de l'efficacité du centre de contact permet aux représentants de passer plus de temps sur des tâches hautement prioritaires



<http://www.multichannel-blog.co.uk/2017/05/03/google-the-future-of-cloud-conference-in-london-3-4th-may/>

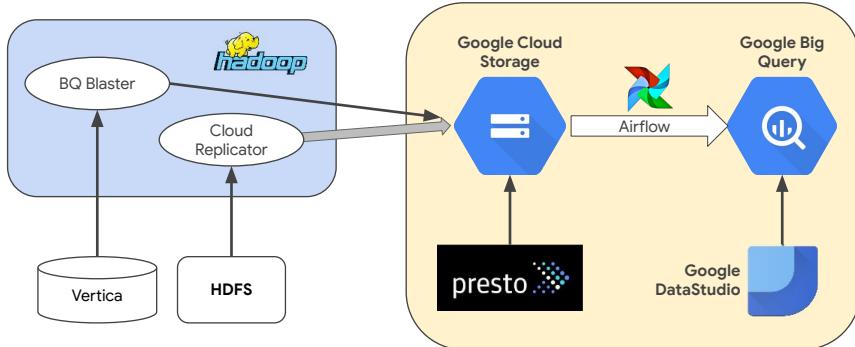
L'équipe d'ingénierie a utilisé Google Cloud Platform pour apprendre à répondre aux e-mails clients urgents quatre fois plus vite. Comment ont-ils fait ? Ils ont utilisé la capacité de traitement de texte de google et créé plusieurs modèles de machine learning personnalisés dans le principal but de lire ces e-mails et de déterminer quel était le message, de quoi il s'agissait et si le message était urgent. Ils ont créé un modèle de machine learning qui classe les messages en termes de priorité et les place dans différentes files d'attente afin que les messages urgents soient traités plus rapidement. Une demande de remboursement n'est pas un message urgent et peut probablement attendre quelques heures pour être traité, mais si quelqu'un écrit "j'ai une fête ce soir et mes articles ne sont pas encore arrivés" ou "j'ai planifié une livraison pour 15 heures aujourd'hui et j'aimerais replanifier", ce sont des messages bien plus urgents, qui doivent être traités en priorité. Un modèle de machine learning pouvant comprendre le langage naturel est en mesure de le faire. Ocado a pu créer un tel modèle de machine learning à l'aide des technologies Google Cloud. Dans ce cours, vous n'apprendrez pas à créer des modèles, mais vous apprendrez à les déployer et les mettre en place, en partant du principe que vous disposez d'une équipe de machine learning capable de créer ces modèles.

<http://i2.wp.com/www.multichannel-blog.co.uk/wp-content/uploads/2017/05/Arch-Diag-Ocado-2017-05-03.png>

Twitter démocratise l'analyse des données avec BigQuery



"Nous pensons que les utilisateurs possédant un large éventail de capacités techniques doivent pouvoir découvrir les données et avoir accès à des analyses SQL et à des outils d'affichage qui fonctionnent"
-- Twitter



https://blog.twitter.com/engineering/en_us/topics/infrastructure/2019/democratizing-data-analysis-with-google-bigquery.html

Un second exemple d'entreprise qui peut démocratiser l'accès aux données via l'utilisation de Google Cloud est Twitter. Twitter a de grandes quantités de données, et a également des équipes de ventes et marketing de haut vol qui, pendant longtemps, n'avaient pas accès aux données et ne pouvaient pas utiliser les données pour effectuer les analyses qu'elles voulaient faire. La plupart des données étaient stockées dans des clusters Hadoop qui étaient vraiment surtaxés. Alors, Twitter a répliqué certaines de ces données de hdfs sur Cloud Storage, les a chargées dans BigQuery et a mis BigQuery à disposition du reste de l'organisation. Il s'agissait des ensembles de données les plus fréquemment requis au sein de Twitter. Ils ont découvert qu'avec un accès direct aux données, de nombreuses personnes qui n'étaient pas des analystes de données analysaient désormais les données et, par conséquent, prenaient de meilleures décisions.

<https://pixabay.com/fr/vectors/twitter-tweet-oiseau-de-twitter-312464/>

Vidéo : 1_l12

Récapitulatif :

Présentateur : Lak

[FIN DE LA VIDÉO]

Récapitulatif :

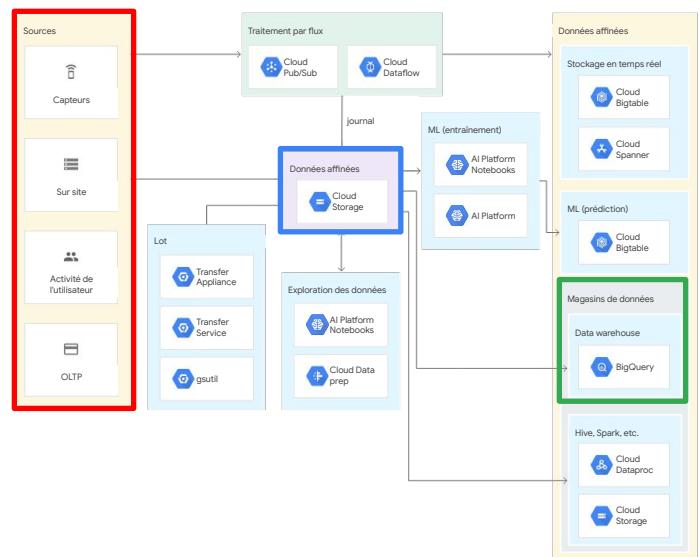
- Sources de données
- Data lakes
- Data warehouses
- Solutions Google Cloud pour l'ingénierie des données



Résumons les principaux sujets couverts dans cette introduction.

Récapitulatif du concept :

Les **sources de données** alimentent un **Data Lake** et sont traitées dans votre **Data Warehouse** pour analyse



Souvenez-vous que vos sources de données (à gauche) sont vos systèmes en amont (comme les RDBMS) et que les autres données brutes proviennent de votre entreprise, sous différents formats.

Les data lakes, qui sont votre emplacement consolidé, durable et hautement disponible pour les données brutes. Dans cet exemple, notre data lake est un Google Cloud Storage.

Et les data warehouses, qui sont le résultat final du pré-traitement des données brutes dans votre data lake et de la préparation pour l'analyse et les charges de travail ML

Vous verrez de nombreuses autres icônes de produits GCP ici, tels que le traitement de données par lots et par flux dans votre lac et l'exécution de ML sur vos données. Nous détaillerons ces sujets ultérieurement dans ce cours.

Voici un guide utile pour les "Produits GCP en 4 mots ou moins"

<https://github.com/gregsranglings/google-cloud-4-words>

Continuellement mis à jour par Greg Wilson, directeur des relations développeurs chez Google

DATABASES	
Cloud Bigtable	Petabyte-scale, low-latency, non-relational
Cloud Datastore	Horizontally scalable document DB
Cloud Firestore	Strongly-consistent serverless document DB
Cloud Memorystore	Managed Redis
Cloud Spanner	Horizontally scalable relational DB
Cloud SQL	Managed MySQL and PostgreSQL
DATA AND ANALYTICS	
BigQuery	Data warehouse/analytics
BigQuery BI Engine	In-memory analytics engine
BigQuery ML	BigQuery model training/serving
Cloud Composer	Managed workflow orchestration service
Cloud Data Fusion	Graphically manage data pipelines
Cloud Dataflow	Stream/batch data processing
Cloud Datalab	Managed Jupyter notebook
Cloud Dataprep	Visual data wrangling
Cloud Dataproc	Managed Spark and Hadoop
Cloud Pub/Sub	Global real-time messaging
Data Catalog	Metadata management service
Data Studio	Collaborative data exploration/dashboarding
Genomics	Managed genomics platform
AI/ML	
AI Hub	Hosted AI component sharing
AI Platform	Managed platform for ML
AI Platform Data Labeling	Data labeling by humans
AI Platform Deep Learning VMs	Preconfigured VMs for deep learning
AI Platform Notebooks	Managed JupyterLab notebook instances
AI Platform Training	Parallel and distributed training



Un aide-mémoire utile à conserver pour référence est le guide "Produits GCP en 4 mots ou moins", qui est constamment maintenu à jour sur github par notre équipe Google Developer Relations. C'est également une excellente manière de se tenir au courant des nouveaux produits et services qui sont publiés en vous abonnant aux commits github !

Vidéo : 1_l13

Atelier : Analyser les données à l'aide de BigQuery

Présentateur : Lak

[FIN DE LA VIDÉO]

Programme

Découvrez le rôle de l'ingénieur données

Analysez les défis de l'ingénierie des données

Introduction à BigQuery

Data Lakes et Data Warehouses

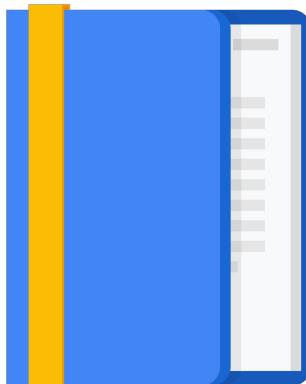
Bases de données transactionnelles vs Data Warehouses

Travailler efficacement avec les autres équipes chargées des données

- Gérer l'accès et la gouvernance des données
- Créer des pipelines prêts à l'emploi

Passer en revue les études de cas des clients GCP

Atelier : Analyser les données à l'aide de BigQuery



Il est temps de mettre en pratique l'analyse des données avec BigQuery dans votre atelier.



Atelier

Analyse des données avec BigQuery

Objectifs

- Exécuter des requêtes interactives dans la console BigQuery
- Combiner et exécuter des analyses sur plusieurs ensembles de données

Dans cet atelier, vous allez :

- Exécuter des requêtes interactives dans la console BigQuery
- Combiner et exécuter des analyses sur plusieurs ensembles de données

Essayez, puis passez à la vidéo de solution guidée.

Vidéo : 1_l14

Solution : Analyser les données à l'aide de BigQuery

Présentateur : Lak

[FIN DE LA VIDÉO]