Google BigQuery et Tableau : les meilleures pratiques

Introduction

Tableau et Google BigQuery permettent d'analyser de très grandes quantités de données et d'obtenir des réponses rapidement dans une interface visuelle conviviale. L'utilisation conjointe de ces deux outils présente plusieurs avantages.

- Les utilisateurs bénéficient au quotidien de la puissance de Google BigQuery pour des analyses rapides et interactives.
- Des milliards de lignes peuvent être analysées en quelques secondes avec des outils d'analyse visuelle, sans nécessiter la moindre ligne de code et sans aucune gestion côté serveur.
- Vous pouvez créer en quelques minutes de superbes tableaux de bord qui se connectent à vos données Google BigQuery et permettent à votre entreprise de garder le rythme.
- Vous avez la possibilité de partager vos rapports et vos découvertes sur le Web grâce à Tableau Server et Tableau Online afin que quiconque puisse y accéder à partir de n'importe quel appareil.
- Vous pouvez combiner l'agilité du cloud que procure Google BigQuery et la vitesse fulgurante de Tableau afin de reconnaître plus rapidement la valeur d'un projet.

Optimiser les deux technologies ensemble permet d'obtenir des gains de performances significatifs, de raccourcir les cycles de conception et d'aider les utilisateurs et structures à mieux réussir. Dans ce livre blanc, nous étudierons des techniques permettant d'optimiser la modélisation des données et la formation des requêtes afin de maximiser la réactivité des visualisations. Nous aborderons également des techniques permettant d'obtenir la meilleure rentabilité lors de l'utilisation simultanée de Tableau et de BigQuery.

Les auteurs

Pierce Young, chef de produit, Tableau

Vaidy Krishnan, chef de produit senior, Tableau

Riley Maris, spécialiste senior du marketing produit, Tableau

Babu Prasad Elumala, ingénieur en solutions, Google

Seth Hollyman, responsable de programme technique, Google

Tino Tereshko, ingénieur en solutions d'entreprise, Google

Mike Graboski, ingénieur en solutions, Google

Sommaire

Présentation des technologies

Google BigQuery

BigQuery traite en quelques secondes plusieurs pétaoctets de données en SQL brut, sans nécessiter de configuration avancée ni de compétences particulières. S'appuyant sur Dremel, BigQuery, la technologie révolutionnaire de Google destinée à l'analyse d'ensembles de données volumineux, offre un niveau de performances pour lequel les grandes entreprises devaient précédemment payer des millions de dollars (au coût de quelques centimes par gigaoctet).

BigQuery est un entrepôt de données parfaitement adapté à l'exécution de requêtes SQL sur des ensembles de données très volumineux, structurés et semi-structurés. Voici quelques exemples de cas d'utilisation et d'ensembles de données :

- Analytique ad hoc
- Journaux Web
- Journaux de machines/serveurs
- Ensembles de données produits par l'IoT (Internet des objets)
- Comportement des clients dans l'e-commerce
- Données d'applications mobiles
- Analytique dans la vente au détail
- Télémétrie dans les jeux
- Données Google Analytics Premium
- Tout ensemble de données pour lequel l'exécution d'une requête par lots prend plusieurs minutes (ou heures) à un RDBMS normal

BigQuery ne nécessite aucun personnel ni aucune maintenance et est intégré à Google Cloud Platform. Contrairement à d'autres solutions analytiques basées sur le cloud, BigQuery ne nécessite pas la mise en service préalable d'un cluster de serveurs. Les clusters de traitement sont dimensionnés et mis en service par BigQuery au moment de l'exécution.

BigQuery ajoute automatiquement de la puissance de traitement au fur et à mesure que le volume de données augmente, mais le tarif par gigaoctet reste le même.

L'ancien SQL et le SQL standard

Google BigQuery a modernisé ses API qui utilisent désormais le SQL standard en plus du SQL BigQuery (également appelé l'ancien SQL). Tableau a donc mis à niveau le connecteur Google BigQuery afin de prendre en charge cette adoption du SQL standard. Le SQL standard présente plusieurs avantages pour les utilisateurs de BigQuery, dont les expressions LOD, la validation plus rapide des métadonnées et la possibilité de choisir un projet de facturation avec leur connexion. Dans ce guide, nous partons du principe que vous utilisez le SQL standard.

Pour savoir comment passer de l'ancien SQL au SQL standard, consultez le guide de l'aide en ligne sur la migration vers le SQL standard (en anglais) sur le site Web de Google Cloud Platform.

Tableau

Tableau aide les utilisateurs à voir et à comprendre leurs données. Notre plate-forme analytique moderne, basée sur une technologie développée à l'université de Stanford, met la puissance des données à la portée de tous. Cela permet à de nombreux utilisateurs d'interagir avec leurs données, de poser des questions, de résoudre des problèmes, de partager des insights et de créer de la valeur. Qu'ils soient à l'aise ou non avec les outils BI traditionnels ou non, ils peuvent rapidement apprendre à tirer parti de Tableau pour créer et explorer des visualisations interactives très complètes et des tableaux de bord puissants dans une interface utilisateur intuitive en glisser-déposer.

Nous avons récemment repoussé les limites de notre plate-forme et inclus des outils de préparation de données visuels, directs et intelligents, ainsi que la possibilité d'interroger des sources de données publiées en utilisant le langage naturel.

Optimisations natives de Tableau

Connecteur de source de données : Tableau comporte un connecteur natif et optimisé pour Google BigQuery, qui prend en charge aussi bien la connectivité des données en direct que des extraits en mémoire. La fonction de fusion de données de Tableau permet de fusionner des données provenant de BigQuery avec des données d'une autre source. Notez que plus de 67 sources de données sont prises en charge dans Tableau. Pour les visualisations publiées dans le cloud à l'aide de Tableau Server ou Tableau Online, une connectivité directe vers Google BigQuery peut être maintenue.

Requêtes parallèles : Tableau tire parti de la capacité de Google BigQuery et d'autres sources de données à exécuter simultanément jusqu'à 16 requêtes au total. Des lots de requêtes indépendantes et dédupliquées sont regroupés et envoyés à BigQuery si le résultat ne figure pas déjà dans le cache. Les utilisateurs peuvent s'attendre à d'importants gains de performances liés aux requêtes parallèles, grâce à l'architecture scale-out de BigQuery.

Fusion des requêtes : Tableau rassemble plusieurs requêtes provenant de classeurs et de tableaux de bord, puis les fusionne dans la mesure du possible afin de réduire le nombre de requêtes envoyées à BigQuery. Tableau identifie d'abord les requêtes similaires, en excluant les différences dans les colonnes renvoyées. Il combine ensuite les requêtes entre lesquelles les seules différences sont le niveau d'agrégation ou un calcul de l'utilisateur.

Cache de requêtes externe : si la source de données sous-jacente n'a pas changé depuis la dernière exécution de la même requête, Tableau lit automatiquement les résultats depuis le cache de requêtes enregistré précédemment afin de fournir des temps de chargement presque instantanés.

Connexions à la demande dans Tableau Desktop : lorsque vous ouvrez un classeur publié, Tableau Desktop se connecte uniquement aux sources de données nécessaires à l'affichage des données de la feuille active. Autrement dit, les données s'affichent beaucoup plus rapidement.

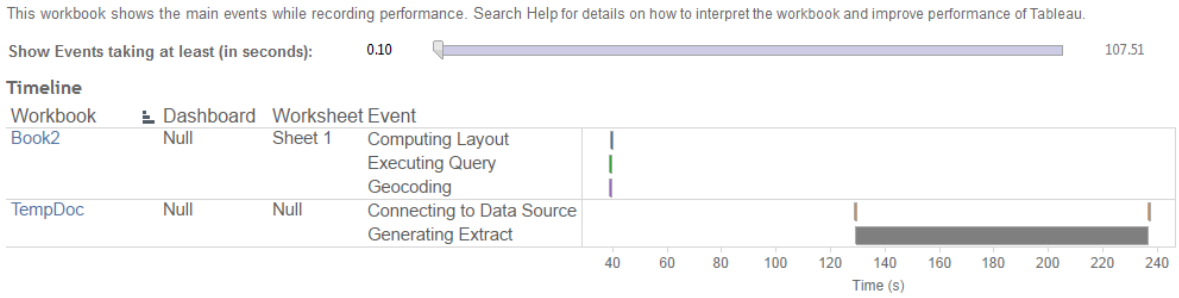
Meilleures pratiques en matière de performances :

Tableau

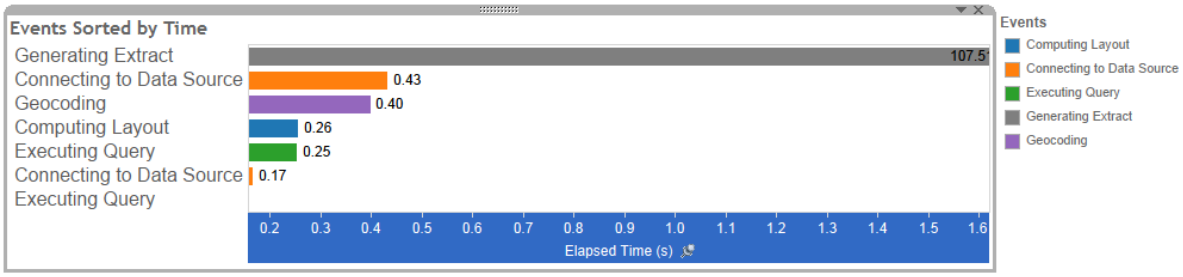
Avant de parler des outils supplémentaires et des paramètres personnalisés, voici notre première recommandation : pensez à mettre votre déploiement Tableau à jour dès que vous le pouvez. Cela permet de tirer parti de toutes les améliorations de performances que nous intégrons à chaque nouvelle version.

Enregistrement des performances

L'enregistrement des performances est un puissant outil intégré qui vous permet de détecter précisément les requêtes lentes et d'optimiser vos classeurs afin d'obtenir des performances maximales. Cette fonction surveille le délai nécessaire à un classeur individuel pour exécuter une requête et calculer la disposition. L'utilisateur peut savoir quelle requête est générée sur BigQuery en survolant avec la souris l'une des barres vertes ci-dessous. Après avoir identifié une requête lente, vous pouvez souvent résoudre le problème de performances en revoyant votre modèle de données.



Dans la vue Timeline, les colonnes Workbook, Dashboard et Worksheet donnent du contexte pour les évènements.



S'ils durent longtemps, vous savez où intervenir pour accélérer les classeurs.

Cliquez sur l'un des liens suivants pour plus de détails sur la création ou l'interprétation d'un enregistrement de performances :

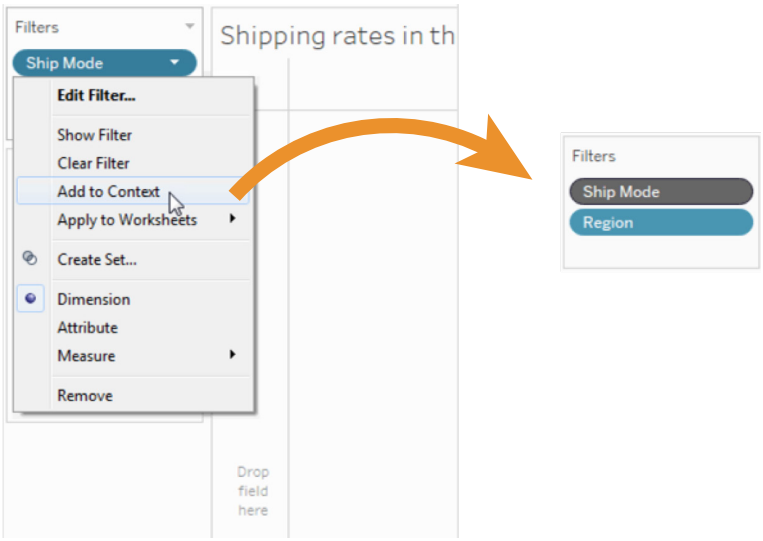
Enregistrement des performances dans Tableau Desktop (création)
Enregistrement des performances dans Tableau Server (interprétation)

Filtres contextuels

Si vous appliquez des filtres à une source de données volumineuse, vous pouvez améliorer les performances en définissant des filtres contextuels. Un filtre contextuel est d'abord appliqué à la source de données, afin que des filtres supplémentaires puissent être appliqués uniquement aux enregistrements obtenus. Cette manière de procéder vous permet d'éviter d'appliquer chaque filtre à chaque enregistrement de la source de données.

Si vous définissez des filtres qui réduisent considérablement la taille de l'ensemble de données et envisagez de les utiliser pour un grand nombre de vues, vous devriez les définir en tant que filtres contextuels.

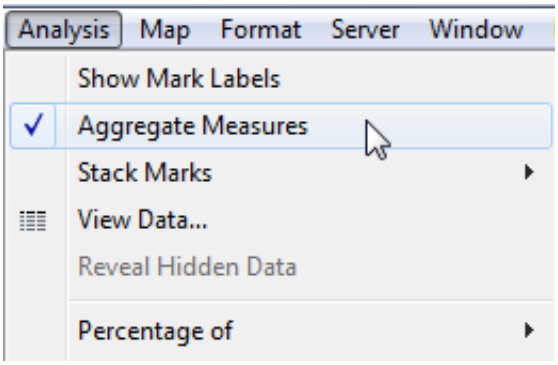
Pour en savoir plus, lisez l'article Améliorer les performances de la vue avec les filtres contextuels dans l'aide en ligne.



Vous pouvez définir un ou plusieurs filtres contextuels pour améliorer les performances.

Agrégation des mesures

Si les vues que vous avez créées sont lentes, utilisez des mesures agrégées plutôt que des mesures non agrégées. Lorsqu'une vue est lente, cela signifie généralement que vous essayez d'afficher trop de lignes de données en même temps. Vous pouvez réduire le nombre de lignes en agrégeant les données.



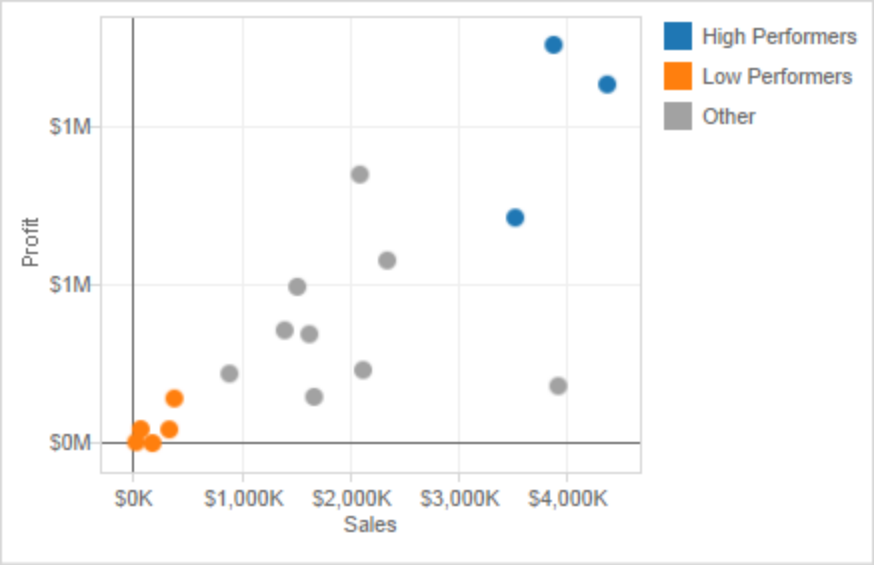
Vérifiez si les mesures sont agrégées dans le menu Analyse. Vous pouvez également définir des agrégations par défaut pour les mesures.

Pour en savoir plus, lisez l'article Agrégation de données dans Tableau dans l'aide en ligne.

Ensembles et groupes

Pour filtrer une dimension afin de supprimer des membres en fonction d'une plage de valeurs de mesures, il est préférable de créer un ensemble plutôt que d'utiliser un filtre quantitatif. Par exemple, vous pouvez créer un ensemble qui ne renvoie que les 50 premiers éléments d'une dimension au lieu de les renvoyer tous.

Lors de la création d'un groupe à partir d'une sélection, assurez-vous d'inclure uniquement les colonnes qui vous intéressent. Chaque colonne supplémentaire incluse dans l'ensemble entraînera une réduction des performances.



Lorsque vous créez des groupes dans Tableau, vous avez la possibilité de regrouper tous les membres restants dans un autre groupe.

Pour en savoir plus, lisez les articles Créer des ensembles *et* Réunir vos données dans l'aide en ligne.

Ajout de filtres au préalable

Si vous utilisez une source de données volumineuse et avez désactivé les mises à jour automatiques, l'ajout de filtres à la vue peut produire une requête vraiment très lente. Plutôt que de créer la vue et de spécifier les filtres ensuite, commencez par spécifier les filtres, puis faites glisser des champs dans la vue. De cette manière, les filtres sont évalués en premier lorsque vous lancez l'actualisation ou activez les mises à jour automatiques.

Désactivation des mises à jour automatiques

Lorsque vous placez un champ sur une étagère, Tableau génère la vue en envoyant automatiquement une requête à la source de données. Si vous créez une vue de données dense, ces requêtes peuvent prendre du temps et dégrader notablement les performances du système. Dans ce cas, vous pouvez demander à Tableau de désactiver les requêtes pendant que vous créez la vue. Vous pouvez ensuite réactiver les requêtes une fois que vous êtes prêt à afficher le résultat.

Pour en savoir plus, lisez l'article Désactiver les mises à jour automatiques pour améliorer les performances *dans l'aide en ligne.*

Observation des avertissements

Tableau affiche une boîte de dialogue d'avertissement sur les performances lorsque vous placez une dimension volumineuse (comptant de nombreux membres) sur une étagère. La boîte de dialogue offre quatre choix, comme l'illustre la figure ci-dessous. Si vous choisissez d'ajouter tous les membres, vous pourriez constater une dégradation significative des performances.

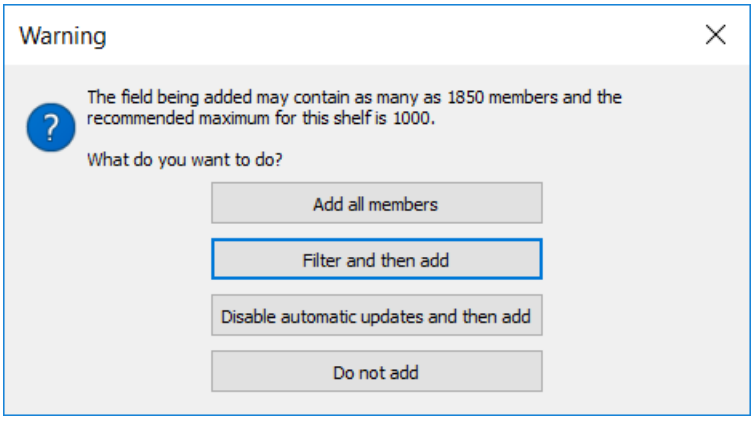


Tableau vous avertit si le fait de placer une dimension volumineuse sur une étagère met les performances en péril.

Optimisation des requêtes parallèles

Vous pouvez utiliser des attributs de personnalisation pour améliorer les performances d'ensembles de résultats volumineux renvoyés de BigQuery vers Tableau Online et Tableau Server, et dans Tableau Desktop, en créant des requêtes parallèles. Vous pouvez inclure ces attributs de personnalisation dans vos classeurs ou sources de données publiés, dès lors que vous les spécifiez avant de publier le classeur ou la source de données sur Tableau Online ou Tableau Server.

Pour en savoir plus, lisez « Utiliser les attributs de personnalisation pour améliorer les performances de requête » dans le guide de l'aide en ligne sur Google BigQuery.

Meilleures pratiques en matière de coûts et de performances : Google BigQuery

Pour optimiser les performances des requêtes et réduire les coûts, il est généralement conseillé d'éviter les tables fédérées dont les données se trouvent dans une source externe telle que Google Cloud Storage. Dans ce cas, si vous souhaitez utiliser des requêtes itératives sur l'ensemble de données, utilisez l'API de requête pour matérialiser les données dans BigQuery (indépendamment de Tableau) afin d'optimiser les performances des requêtes au niveau de l'ensemble de données avec Tableau.

Dénormalisation et opération de jointure préalable

BigQuery prend en charge les opérations de jointure très volumineuses et leurs performances sont excellentes. Cela dit, BigQuery est un magasin de données en colonnes et les performances maximales sont atteintes sur les ensembles de données dénormalisés.

L'un des avantages du cloud est sa capacité à séparer les ressources de stockage des ressources de calcul, permettant ainsi aux utilisateurs de dimensionner (et payer) séparément chaque type de ressources. Le stockage BigQuery étant extrêmement abordable et évolutif, il est souvent plus prudent de dénormaliser les ensembles de données et de les réunir au préalable afin de disposer de tables homogènes. Cela signifie essentiellement que vous utiliserez moins de ressources de calcul, mais davantage de ressources de stockage (ces dernières étant plus performantes et plus économiques). BigQuery étant un magasin de données en colonnes et pouvant compresser les données efficacement, il est plus judicieux, et probablement plus rentable, de réduire les ressources de calcul pour augmenter les ressources de stockage.

BigQuery est un excellent outil ETL qui vous permet d'exécuter rapidement et efficacement des transformations et des pipelines de grande ampleur. N'oubliez pas d'activer l'option Allow Large Results (Autoriser les résultats volumineux) lors de la matérialisation des ensembles de données de plus de 128 Mo.

Pour découvrir comment préparer les données au chargement et comment interroger les données avec le langage BigQuery SQL, consultez les documents ci-dessous.

Charger des données dénormalisées, imbriquées et répétées
Écrire des résultats de requête volumineux

Fractionnement des tables par date

Le fractionnement des tables en plus petites partitions permet de simplifier la gestion des données et d'améliorer les performances des requêtes et la rentabilité. Par ailleurs, BigQuery prend en charge le clustering sur des tables fractionnées, ce qui est utile si les données sont déjà partitionnées en fonction d'une colonne de date ou d'horodatage, ou si vous utilisez des filtres ou des agrégations sur des colonnes particulières dans vos requêtes.

Certaines données se prêtent naturellement à un partitionnement par date : par exemple, les données de journaux, ou les données dont les enregistrements comportent un horodatage incrémenté de manière régulière. Dans ce cas, fractionnez les tables BigQuery par date et incluez la date dans le nom de la table. Pour en tirer profit, exploitez le SQL personnalisé dans Tableau.

Pour en savoir plus, lisez l'article Se connecter à une requête SQL personnalisée dans l'aide en ligne.

Par exemple, donnez à vos tables des noms du type : matable_20170501, matable_20170502, etc.

Ensuite, lorsque vous voulez exécuter une requête qui filtre par date, utilisez la fonction BigQuery qui permet d'utiliser un caractère générique sur le nom de la table :

```
SELECT
 nom

FROM
 'myProject.myDataSet.mytable_*'

WHERE
 age >= 35
```

L'exemple ci-dessus inclura automatiquement toutes les tables utilisant le préfixe mytable_.

Pour qu'un caractère générique puisse être utilisé, le nom de vos tables doit suivre cette syntaxe : [préfixe arbitraire]AAAAMMJJ.

D'autres systèmes de bases de données s'appuient sur le fractionnement pour améliorer les performances. Le fractionnement par date crée en réalité une différence de performances négligeable dans BigQuery, mais le principal argument dans ce cas est le coût. Comme vous traitez une quantité moins volumineuse de données, vous payez moins par requête.

Toutefois, si vous optez pour le fractionnement par minute, le nombre élevé de partitions risque de nuire directement aux performances. Veillez donc à ne pas utiliser trop de partitions en même temps. En règle générale, nous recommandons un fractionnement minimum par jour.

Pour en savoir plus sur le fractionnement :

Présentation des tables partitionnées
Présentation des tables en cluster
Interroger plusieurs tables avec une table générique

Spécification d'une table cible en cas d'exécution de nombreuses requêtes similaires

Bien que la mise en cache des requêtes soit utile lorsque vous exécutez de nombreuses requêtes identiques, elle ne vous aidera pas si vous exécutez de nombreuses requêtes similaires mais légèrement différentes (par ex., en ne changeant que la valeur d'une clause WHERE entre deux exécutions de requête). Dans ce cas, exécutez une requête sur votre table source et stockez dans une nouvelle table cible les enregistrements sur lesquels vous effectuerez des requêtes répétées. Exécutez ensuite vos requêtes sur la nouvelle table cible que vous avez créée.

Par exemple, supposons que vous vouliez exécuter 3 requêtes avec 3 clauses WHERE différentes :

```
WHERE col1 = "a"
WHERE col1 = "b"
WHERE col1 = "c"
```

Exécutez une requête sur votre table source, puis stockez les enregistrements qui en résultent dans une table cible :

```
SELECT col1
FROM source
WHERE col1 = "a" OR col1 = "b" OR col1 = "c"
```

En liant les trois clauses WHERE par un OR, nous pouvons capturer tous les enregistrements pertinents. Notre nouvelle table cible est potentiellement bien moins volumineuse que la table source d'origine. La tarification de BigQuery étant basée sur la quantité de données traitées dans une requête, exécuter les requêtes ultérieures sur la nouvelle table cible est plus économique que les exécuter sur la table source. Pour éviter l'accumulation des coûts de stockage, il est important de nettoyer ces tables au fil du temps.

Utilisation de Tableau pour visualiser les résultats des modèles Google BigQuery ML

BigQuery ML permet d'utiliser une technologie de machine learning embarquée pour enrichir des modèles en fonction des données stockées dans BigQuery. Néanmoins, comme pour les autres types de données, l'envoi de requêtes directement vers une base de données n'est pas forcément la méthode la mieux adaptée pour explorer les résultats de votre modèle.

Grâce au connecteur BigQuery, Tableau vous permet de manipuler facilement les résultats de vos modèles prédictifs de manière à faciliter une compréhension plus intuitive de vos données. De plus, Tableau vous permet de partager facilement votre modèle et ses résultats, pour permettre à d'autres utilisateurs de tirer profit de votre travail.

Pour découvrir un exemple, lisez notre article BigQuery ML with Tableau to predict housing prices (en anglais).

Bien que le SQL personnalisé soit nécessaire ici pour appeler un algorithme de machine learning dans Google BigQuery ML, vous devez tenir compte des performances lorsque vous utilisez le SQL personnalisé au lieu des connexions natives de Tableau dans d'autres situations. Lorsque vous le pouvez, tirez parti des connexions natives que fournit Tableau pour les sources de données afin de bénéficier de performances optimisées.

Pour en savoir plus sur la création et l'exécution des modèles de machine learning à l'aide de requêtes SQL standard, consultez l'aide en ligne de Google BigQuery ML.

Étude de cas : conseils de zulily pour l'analytique en libre-service avec Tableau et Google BigQuery

zulily est une entreprise d'e-commerce à forte croissance qui a créé une plate-forme Big Data en utilisant Google BigQuery pour son magasin de données et Tableau pour accéder aux données et les analyser visuellement. Grâce à l'intégration de BigQuery et de Tableau, l'équipe analytique est plus rapide lors de l'acquisition, de l'ingestion et de l'utilisation des données pour créer des rapports et des modèles sans l'aide de l'IT pour les tâches quotidiennes. De plus, les utilisateurs métier accèdent en temps réel aux données essentielles qui leur permettent de prendre des décisions rapides sans avoir à solliciter les analystes pour obtenir des informations basiques.

Voici quelques-unes des bonnes pratiques mises en place par zulily :

Réduction de la latence en utilisant Tableau Server sur Google Compute Engine — Au lieu d'un modèle traditionnel où les régions sont des clouds privés virtuels, vous pouvez tirer parti de l'infrastructure privée de Google sans accéder à Internet, et sans configuration supplémentaire. Vous pouvez également dimensionner vos déploiements sans mise en service superflue.

Utilisation de sources fédérées et connexion de Tableau à BigQuery — Pour les données dans Google Cloud, vous devez tirer parti de la capacité de BigQuery à envoyer des requêtes vers des sources externes et considérer BigQuery comme votre lac de données. Dans certains scénarios, vous réduisez le volume de données à transférer sur le réseau et vers Tableau pour y être analysées.

Traitement des ensembles de données volumineux avec une connexion en direct dans BigQuery — Tirez parti de la capacité de BigQuery à traiter des ensembles de données volumineux et de transférer uniquement les résultats sur le réseau. Utilisez une connexion en direct vers BigQuery dans Tableau, à moins que vous n'ayez une raison particulière d'utiliser un extrait de données.

Pour en savoir plus et pour découvrir la liste des 10 conseils de zulily, lisez notre série d'articles de blog :

Première partie : Pourquoi zulily a créé une plate-forme de marketing analytique en libre-service avec Tableau et Google BigQuery

Deuxième partie : 10 conseils de zulily pour l'analytique en libre-service avec Tableau et Google BigQuery

Conclusion

En appliquant ces meilleures pratiques, les utilisateurs métier et les analystes peuvent maximiser les performances et la réactivité des visualisations Tableau créées à partir de données Google BigQuery. Une fois ces technologies combinées, les utilisateurs peuvent réellement visualiser instantanément des milliards de lignes de données.

À propos de Tableau

Tableau est une plate-forme BI visuelle, exhaustive et facile à utiliser, conçue pour l'entreprise, qui vous permet de voir et comprendre vos données grâce à une analytique en libre-service rapide et évolutive. Sur site ou dans le cloud, sous Windows ou Linux, Tableau s'appuie sur vos investissements technologiques et évolue en fonction de vos besoins en s'adaptant à votre environnement de données. Libérez la puissance de vos ressources les plus précieuses : vos données et vos collaborateurs.

Autres ressources

Essai gratuit de Tableau

Aide pour Tableau Online : Google BigQuery

Tableau Server et Google Cloud Platform : la BI à grande vitesse dans le cloud (en anglais)

Solutions Tableau et Google

Tableau et le Big Data : tour d'horizon

Pourquoi déplacer les outils d'analyse métier dans le cloud ?

L'art de concevoir des classeurs efficaces