[sfeir] Institute

Google Cloud | GCP200DWBQ

Data Warehousing with BigQuery: Storage Design, Query Optimization, and Administration

Découvrirez l'architecture de BigQuery et les bonnes pratiques de conception, de lecture, d'optimisation et d'administration de votre entrepôt de données

3 jour(s) / 21h

Objectifs pédagogiques

- Décrire les principes de base de l'architecture BigQuery.
- Implémenter des modèles de conception de stockage et de schéma pour améliorer les performances.
- Utiliser DML et planifier des transferts de données pour ingérer des données.
- Appliquer les meilleures pratiques pour améliorer l'efficacité de la lecture et optimiser les performances des requêtes.
- Gérer la capacité et automatiser les charges de travail.
- Comprendre les modèles par rapport aux anti-modèles pour optimiser les requêtes et améliorer les performances de lecture.
- Utiliser des outils de journalisation et de surveillance pour comprendre et optimiser les modèles d'utilisation.
- Appliquer les meilleures pratiques de sécurité pour gérer les données et les ressources.
- Créer et déployer plusieurs catégories de modèles de machine learning avec BigQuery ML.

Public cible

- Analystes de données, data scientists, ingénieurs de données et développeurs qui effectuent des travaux à grande échelle nécessitant des connaissances internes avancées de BigQuery pour optimiser les performances.

Prérequis

Pour tirer le meilleur parti de ce cours, les participants doivent :

- Avoir suivi le cours Big Data and Machine Learning Fundamentals ou avoir des connaissances équivalentes.

Programme

Module 1 : Principes de base de l'architecture BigQuery

Sujets

- Introduction
- Infrastructure centrale BigQuery
- Stockage BigQuery
- Traitement des requêtes BigQuery
- Shuffling des données BigQuery

Objectifs

- Expliquer les avantages du stockage en colonne.
- Comprendre comment BigQuery traite les données.
- Découvrir les principes de base du service de shuffling de BigQuery pour améliorer l'efficacité des requêtes.

Activités

- Ateliers et démos

Module 2 : Optimisations de stockage et de schéma

Sujets

- Stockage BigQuery

- Partitionnement et clustering
- Champs imbriqués et répétés
- Syntaxe ARRAY et STRUCT
- Les meilleures pratiques

Objectifs

- Comparer les performances de différents schémas (snowflake, dénormalisé, et champs imbriqués et répétés).
- Partitionner et regrouper les données pour de meilleures performances.
- Améliorer la conception du schéma à l'aide de champs imbriqués et répétés.
- Décrire les meilleures pratiques supplémentaires telles que l'expiration des tables et des partitions

Activités

- Ateliers et démos

Module 3 : Ingestion de données

Sujets

- Options d'intégration de données
- Ingestion par lots
- Ingestion de diffusion en continu
- Legacy Streaming API
- BigQuery Storage Write API
- Matérialisation des requêtes
- Interroger des sources de données externes
- Service de transfert de données

Objectifs

- Ingérer des données par lots et en continu.
- Interroger des sources de données externes.
- Planifier les transferts de données.
- Comprendre comment utiliser l'API Storage Write.

Activités

- Ateliers et démos

Module 4 : Modification des données

Sujets

- Gestion du changement dans les entrepôts de données
- Gestion des Slowly Changing Dimensions (SCD)
- Déclarations DML
- Bonnes pratiques DML et problèmes courants

Objectifs

- Écrire des instructions DML.
- Résoudre les problèmes de performances et les goulots d'étranglement courants de DML.
- Identifiez les Slowly Changing Dimensions (SCD) dans vos données et effectuez des mises à jour.

Module 5 : Améliorer les performances de lecture

Sujets

- Cache de BigQuery
- Vues matérialisées
- BI Engine
- Lectures à haut débit
- API de lecture de stockage BigQuery

Objectifs

- Explorer le cache de BigQuery.
- Créer des vues matérialisées.
- Travailler avec BI Engine pour accélérer vos requêtes SQL.
- Utiliser l'API Storage Read pour un accès rapide au stockage géré par BigQuery.
- Expliquer les écueils liés à l'utilisation de sources de données externes.

Activités

- Ateliers et démos

Module 6 : Optimisation et dépannage des requêtes

Sujets

- Exécution simple des requêtes
- SELECT et Agrégation
- JOIN et JOIN biaisés
- Filtrage et classement

- Meilleures pratiques pour les fonctions

Objectifs

- Interpréter les détails d'exécution de BigQuery et le plan de requête.
- Optimiser les performances des requêtes en utilisant les méthodes suggérées pour les instructions et les clauses SQL.
- Démontrer les meilleures pratiques pour les fonctions dans les cas d'utilisation métier.

Activités

- Ateliers et démos

Module 7 : Gestion de la charge de travail et tarification

Sujets

- Emplacements BigQuery
- Modèles de tarification et estimations
- Réservations de créneaux
- Contrôle des coûts

Objectifs

- Définir un emplacement BigQuery.
- Expliquer les modèles de tarification et les estimations de tarification (interface utilisateur BigQuery, bq dry_run, API jobs).
- Comprendre les réservations de créneaux, les engagements et les affectations.
- Identifier les meilleures pratiques pour contrôler les coûts.

Activités

- Démos

Module 8 : Journalisation et surveillance

Sujets

- Cloud Monitoring
- BigQuery Admin Panel
- Cloud Audit Logs
- INFORMATION_SCHEMA
- Chemin de requête et erreurs courantes

Objectifs

- Utiliser Cloud Monitoring pour afficher les métriques BigQuery.
- Explorez le BigQuery Admin Panel.
- Utiliser les Cloud Audit Logs.
- Utiliser les tables INFORMATION_SCHEMA pour obtenir des informations sur vos entités BigQuery.

Activités

- Ateliers et démos

Module 9 : Security in BigQuery

Sujets

- Ressources sécurisées avec IAM
- Vues autorisées
- Données sécurisées avec classification
- Chiffrement
- Découverte et gouvernance des données

Objectifs

- Explorer la découverte de données à l'aide de Data Catalog.
- Discuter de la gouvernance des données à l'aide de l'API DLP et Data Catalog.
- Créer des stratégies IAM (par exemple, des vues autorisées) pour sécuriser les ressources.
- Sécuriser les données avec des classifications (par exemple, des politiques au niveau des lignes).
- Comprendre comment BigQuery utilise le chiffrement.
- Laboratoires et démos

Activités

- Ateliers et démos

Module 10 : Automatisation des charges de travail

Sujets

- Planifier des requêtes
- Script
- Procédures stockées
- Intégration avec les produits Big Data

Objectifs

- Planifier des requêtes.
- Utiliser des scripts et des procédures stockées pour créer des transformations personnalisées.
- Décrire comment intégrer les charges de travail BigQuery à d'autres produits de big data Google Cloud.

Activités

- Démos

Module 11 : Apprentissage automatique dans BigQuery

Sujets

- Présentation de BigQuery ML
- Comment faire des prédictions avec BigQuery ML
- Comment créer et déployer un système de recommandation avec BigQuery ML
- Comment créer et déployer une solution de prévision de la demande avec BigQuery ML
- Modèles de séries temporelles avec BigQuery ML
- BigQuery ML Explainability

Objectifs

- Décrire certaines des différentes applications de BigQuery ML.
- Créer et déployer plusieurs catégories de modèles de machine learning avec BigQuery ML.
- Utiliser AutoML Tables pour résoudre des problèmes commerciaux à forte valeur.

Activités

- Ateliers et démos