+33 (0)1 48 10 09 46
formation@ambient-it.net
140 Avenue Jean Lolive 93500 Pantin - France

Mis à jour le 29/07/2025

S'inscrire

Formation Greenplum Parallel SQL

3 jours (21 heures)

Présentation

Maîtrisez Greenplum Parallel SQL dans toute sa puissance avec cette formation experte, conçue pour les ingénieurs data, développeurs SQL et architectes souhaitant exploiter pleinement les capacités de traitement massivement parallèle de Greenplum, au service de l'analyse à grande échelle.

La formation s'ouvre sur les fondamentaux de Greenplum et l'architecture MPP, avec un focus sur la distribution des données, le partitionnement et les stratégies de modélisation conçues pour maximiser le parallélisme natif du moteur.

Vous apprendrez à écrire des requêtes SQL performantes dans un contexte distribué : jointures optimisées, agrégats parallèles, fonctions analytiques avancées, et exploitation fine des plans d'exécution à travers des cas concrets de tuning.

Les modules pratiques couvrent également l'ingestion massive (COPY, gpload, external tables), les mises à jour distribuées, l'usage d'UDF, et l'intégration de Greenplum dans un écosystème BI ou data engineering plus large.

Comme pour toutes nos formations, celle-ci vous sera présentée avec les toutes dernières actualisations de Greenplum Parallel.

Objectifs

- Comprendre l'architecture massivement parallèle de Greenplum et ses spécificités de distribution et partitionnement
- Savoir modéliser, distribuer et partitionner des schémas de données optimisés pour le traitement SQL
- Maîtriser les méthodes d'écriture de requêtes SQL performantes en environnement parallèle
- Être capable d'ingérer des volumes massifs de données via COPY, gpload ou tables externes
- Exploiter tout le potentiel du moteur Greenplum dans des cas concrets BI et data engineering

Public visé

- Architectes data
- Développeurs SQL

Pré-requis

- Maîtrise des bases du langage SQL

Programme de la formation Greenplum Parallel SQL

Introduction à Greenplum et au SQL Massivement Parallèle

- Présentation de Greenplum
- PostgreSQL comme socle
- Cas d’usage typiques
- Segments, master et interconnexions
- Mirroring et tolérance aux pannes
- Différences avec SMP et Hadoop
- Concepts clés du SQL parallèle
- Partitionnement des données
- Distribution vs réplication
- Traitement distribué de requêtes SQL

Modélisation des données pour le parallélisme

- Création de tables parallèles
- Syntaxe SQL spécifique à Greenplum
- Paramètres de distribution
- Comparaison avec les tables locales
- Choix des clés de distribution
- Clé unique vs clé fréquente
- Impact sur les performances
- Cas pratiques de mauvaise distribution
- Partitionnement logique
- Syntaxe de partitionnement
- Partitionnement par plage ou liste
- Exemples de requêtes optimisées avec partitions

Requêtes SQL parallèles

- SELECT et agrégation parallèle
- Agrégats distribués
- Fonctions d’agrégation spécifiques
- Group By sur données distribuées
- Jointures parallèles
- Broadcast vs Hash Join
- Optimisation des jointures sur clés distribuées
- Cas de Shuffle cost élevé
- Analyse de plans d’exécution
- Utilisation de EXPLAIN et EXPLAIN ANALYZE
- Lecture des plans distribués
- Interprétation des coûts intersegments

Écriture et insertion de données parallèles

- INSERT, UPDATE, DELETE
- Comportement distribué de l’écriture
- Restrictions sur les UPDATE avec jointures
- Tris et insertions en bulk
- COPY et chargement massif
- Utilisation de COPY en parallèle
- Gestion des erreurs et des logs
- Chargement de fichiers externes
- Contraintes et indexation
- Contraintes supportées en parallèle
- Index bitmap vs BTree
- Impact des index sur les performances distribuées

Fonctions analytiques et avancées

- Fonctions de fenêtrage
- Utilisation de ROW_NUMBER, RANK, LEAD, LAG
- Partitionnement des fenêtres sur segments
- Cas d’usage typiques en BI
- Expressions et types complexes
- Tables dérivées, CTE, sous-requêtes parallèles
- Types ARRAY et JSON dans Greenplum
- Fonctions utilisateur (UDF/UDT)
- Création de fonctions en SQL ou PL/pgSQL
- Portabilité depuis PostgreSQL
- Considérations de parallélisme

Optimisation et tuning de requêtes SQL parallèles

- Statistiques et analyse
- Utilisation de ANALYZE sur tables distribuées
- Statistiques globales vs locales
- Outils d’analyse de performance Greenplum
- Rewriting et indexation intelligente
- Refonte des requêtes coûteuses
- Utilisation d’index sur colonnes de jointure
- Cas pratiques de réécriture
- Optimisation des transactions
- Isolation et gestion des locks
- Verrous en environnement parallèle
- Meilleures pratiques pour les écritures concurrentes

Intégration, import/export et accès externe

- Tables externes
- Déclaration de fichiers plats ou HDFS
- Intégration avec gpload et gpfdist
- Chargement parallèle de gros volumes
- gpload et gpexpand
- Configuration de fichiers YAML pour gpload
- Ajout de nouveaux segments dynamiquement
- Étendre un cluster sans redéploiement
- Accès depuis outils externes
- Connexion via ODBC/JDBC
- Requêtes SQL depuis Python, R, Java
- Compatibilité BI (Tableau, Power BI)

Cas pratiques et projet final

- Cas pratique BI (reporting sur gros volumes)
- Création de schémas distribués
- Requêtes analytiques optimisées
- Tableau de bord connecté
- Cas pratique Data Engineering
- Ingestion massive avec partitionnement
- Pipeline SQL avec jointures complexes
- Monitoring des performances
- Projet de synthèse
- Choix de la stratégie de distribution
- Requêtage, chargement, analyse
- Rapport de performances

Sociétés concernées

Cette formation s’adresse à la fois aux particuliers ainsi qu’aux entreprises, petites ou grandes, souhaitant former ses équipes à une nouvelle technologie informatique avancée ou bien à

acquérir des connaissances métiers spécifiques ou des méthodes modernes.

Positionnement à l'entrée en formation

Le positionnement à l'entrée en formation respecte les critères qualité Qualiopi. Dès son inscription définitive, l'apprenant reçoit un questionnaire d'auto-évaluation nous permettant d'apprécier son niveau estimé sur différents types de technologies, ses attentes et objectifs personnels quant à la formation à venir, dans les limites imposées par le format sélectionné. Ce questionnaire nous permet également d'anticiper certaines difficultés de connexion ou de sécurité interne en entreprise (intraentreprise ou classe virtuelle) qui pourraient être problématiques pour le suivi et le bon déroulement de la session de formation.

Méthodes pédagogiques

Stage Pratique : 60% Pratique, 40% Théorie. Support de la formation distribué au format numérique à tous les participants.

Organisation

Le cours alterne les apports théoriques du formateur soutenus par des exemples et des séances de réflexions, et de travail en groupe.

Validation

À la fin de la session, un questionnaire à choix multiples permet de vérifier l'acquisition correcte des compétences.

Sanction

Une attestation sera remise à chaque stagiaire qui aura suivi la totalité de la formation.