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

Mis à jour le 29/10/2024

S'inscrire

Formation Apache Parquet

2 jours (14 heures)

Présentation

Apache Parquet est la technologie qu'il vous faut ! Il s'agit d'un format de fichier open source, optimisé pour le stockage et le traitement de grandes quantités de données analytiques. Il utilise un stockage en colonnes, qui permet de lire rapidement des ensembles de données spécifiques sans charger le fichier complet.

Pendant cette formation, vous explorerez les fonctionnalités d'Apache Parquet, notamment sa structure interne et l'organisation de ses métadonnées, qui optimisent le traitement des données. Vous apprendrez également à configurer vos fichiers pour une performance optimale, à définir des schémas et à utiliser les filtres de Bloom, ce qui permet de restreindre les résultats et d'accélérer les requêtes.

Le programme couvrira également des fonctionnalités avancées, telles que les options de compression disponibles (par exemple, Snappy, Gzip) pour réduire le poids des données tout en maintenant leur intégrité, et les index de page, qui facilitent la récupération efficace des données. Vous découvrirez aussi comment gérer les erreurs et les corruptions de données grâce aux sommes de contrôle, essentielles pour garantir la fiabilité des analyses.

Au terme de cette formation, vous serez capable de configurer, optimiser et maintenir des fichiers Parquet adaptés à divers cas d'utilisation. Vous maîtriserez l'encodage des données et l'utilisation des index, des compétences essentielles pour traiter de grandes bases de données avec rapidité et précision. Vous aurez aussi les bases pour contribuer au projet Apache Parquet si vous le souhaitez, enrichissant ainsi l'écosystème open source.

Comme pour toutes, la formation Apache Parquet vous sera présentée avec sa dernière version : Apache Parquet 1.14.

Objectifs

- Maîtriser les avantages d'Apache Parquet pour le traitement de grandes données
- Configurer efficacement les fichiers Parquet pour une performance optimale

- Utiliser les encodages et filtres pour optimiser les requêtes
- Choisir et appliquer les méthodes de compression adaptées
- Contribuer au développement et à l'évolution du projet Apache Parquet

Public visé

- Data analysts
- Data engineers
- Architectes
- Développeurs

Pré-requis

- Bonne compréhension des bases de données et de la manipulation de fichiers structurés
- Expérience avec un langage de programmation, tel que Python ou Java
- Connaissances de base en traitement de données massives (Hadoop, Spark, etc.)
- Familiarité avec les concepts de stockage en colonnes et de compression

PROGRAMME DE NOTRE FORMATION APACHE PARQUET

INTRODUCTION À APACHE PARQUET

- Présentation d'Apache Parquet
- Avantages de l'utilisation de Parquet pour le traitement de grandes quantités de données
- Comparaison avec d'autres formats de fichiers comme CSV, JSON, et ORC
- Cas d'utilisation typiques

CONFIGURATION ET MÉTADONNÉES

- Comprendre la structure interne d'un fichier Parquet
- Configuration optimale pour des performances améliorées
- Gestion et utilisation des métadonnées
- Extensibilité de Parquet pour des besoins spécifiques
- Exemples pratiques de configuration de fichiers Parquet

TYPES DE DONNÉES ET ENCODAGE

- Types de données supportés
- Encodage imbriqué pour les données structurées
- Utilisation des filtres de Bloom pour optimiser les requêtes
- Gestion des valeurs nulles dans les ensembles de données
- Ateliers pratiques sur la définition des schémas et l'encodage des données

COMPRESSION ET GESTION DES PAGES DE DONNÉES

- Méthodes de compression disponibles et comment les choisir
- Importance des sommes de contrôle pour l'intégrité des données
- Segmentation des données en morceaux de colonnes pour une récupération efficace
- Utilisation des index de page pour améliorer la performance des requêtes
- Scénarios de récupération d'erreurs et de corruption de données

GUIDE DU DÉVELOPPEUR ET CONTRIBUTION AU PROJET

- Exploration des sous-projets liés à Parquet
- Processus pour compiler et construire Parquet à partir des sources
- Comment contribuer à Parquet-Java et au développement général de Parquet
- Processus de publication et de mise à jour des versions de Parquet
- Création d'un environnement de développement pour les contributions

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.