

Building Batch Data Analytics Solutions on AWS

Durée: 1 Jour **Réf de cours: GK7378** **Méthodes d'apprentissage: Intra-entreprise & sur-mesure**

Résumé:

Cette formation permet aux participants d'acquérir les compétences et connaissances nécessaires pour construire des solutions d'analyse de données par lots à l'aide d'Amazon EMR, un service géré Apache Spark et Apache Hadoop de niveau entreprise. Ils apprendront comment Amazon EMR s'intègre aux projets open-source tels que Apache Hive, Hue et HBase, ainsi qu'aux services AWS tels que AWS Glue et AWS Lake Formation. La formation aborde les composants de collecte, d'ingestion, de catalogage, de stockage et de traitement des données dans le contexte de Spark et Hadoop. Ils apprendront à utiliser EMR Notebooks pour prendre en charge les charges de travail d'analyse et d'apprentissage automatique et à appliquer les meilleures pratiques de sécurité, de performance et de gestion des coûts au fonctionnement d'Amazon EMR.

Financement :

-Formation éligible au Compte Personnel de Formation (CPF), permettant de préparer la Certification associée inscrite au répertoire de France Compétences.

La Certification professionnelle inscrite au Répertoire Spécifique de France Compétences RS6222 "Créer et maintenir des solutions d'analyse de données efficaces et sécurisées sur le cloud AWS" valable jusqu'au 24-11-2026.

Nous vous invitons à consulter les prérequis, les objectifs, le contexte de la certification ainsi que les détails de la certification sur :

<https://www.francecompetences.fr/recherche/rs/6622>

Mise à jour : 08.07.2024

Public visé:

Cette formation s'adresse aux Ingénieurs de plateforme de données et Architectes et opérateurs qui construisent et gèrent des pipelines d'analyse de données

Objectifs pédagogiques:

- A l'issue de la formation, les participants seront capables de :
- Comparer les caractéristiques et les avantages des entrepôts de données, des lacs de données et des architectures de données modernes
- Concevoir et mettre en œuvre une solution d'analyse de données par lots
- Identifier et appliquer les techniques appropriées, notamment la compression, pour optimiser le stockage des données
- Sélectionner et déployer les options appropriées pour ingérer, transformer et stocker les données
- Sélectionner les types d'instances et de nœuds, les clusters, la mise à l'échelle automatique et la topologie réseau appropriés pour un cas d'utilisation particulier
- Comprendre comment le stockage et le traitement des données affectent les mécanismes d'analyse et de visualisation nécessaires pour obtenir des informations commerciales exploitables
- Décrire comment sécuriser les données au repos et en transit
- Surveiller les charges de travail analytiques afin d'identifier et de résoudre les problèmes
- Identifier les meilleures pratiques de gestion des coûts

Pré-requis:

Posséder au moins un an d'expérience dans la gestion de frameworks de données open-source tels que Apache Spark ou Apache Hadoop.

- GK7377 - Building Data Lakes on AWS

Test et certification

Afin de se préparer efficacement à l'obtention de la certification **AWS Certified Data Analytics Speciality**, nous vous recommandons de suivre également les formations suivantes :

[GK7377](#) - Building Data Lakes on AWS

[GK7379](#) - Building Data Analytics Solutions using Amazon Redshift

[GK7400](#) - Building Streaming Data Analytics Solutions on AWS

**

Cette formation mène également à la Certification professionnelle inscrite au Répertoire Spécifique de France Compétences RS6222 "Créer et maintenir des solutions d'analyse de données efficaces et sécurisées sur le cloud AWS". Plus de détails sur : <https://www.francecompetences.fr/recherche/rs/6222/>

Après cette formation, nous vous conseillons le(s) module(s) suivant(s):

■ GK7379 - Building Data Analytics Solutions Using Amazon Redshift

Contenu:

Module A : Aperçu de l'analyse des données et du pipeline de données

- Cas d'utilisation de l'analyse de données
- Utilisation du pipeline de données pour l'analyse

Module 1 : Introduction à Amazon EMR

- Utilisation d'Amazon EMR dans les solutions analytiques
- Architecture du cluster Amazon EMR
- Démo interactive 1 : Lancement d'un cluster Amazon EMR
- Stratégies de gestion des coûts

Module 2 : Pipeline d'analyse de données utilisant Amazon EMR : Ingestion et stockage

- Optimisation du stockage avec Amazon EMR
- Techniques d'ingestion de données

Module 3 : Analyse de données par lots haute performance avec Apache Spark sur Amazon EMR

- Cas d'utilisation d'Apache Spark sur Amazon EMR
- Pourquoi Apache Spark sur Amazon EMR ?
- Concepts de Spark interactive 2 : Connectez-vous à un cluster EMR et exécutez des commandes Scala à l'aide du shell Spark
- Transformation, traitement et analyse
- Utilisation de notebooks avec Amazon EMR

Atelier 1 : analyse de données à faible latence avec Apache Spark sur Amazon EMR

Module 4 : Traitement et analyse de données par lots avec Amazon EMR et Apache Hive

- Utilisation d'Amazon EMR avec Hive pour traiter des données par lots
- Transformation, traitement et analyse

Atelier 2 : Traitement de données par lots à l'aide d'Amazon EMR et Hive

- Introduction à Apache HBase sur Amazon EMR

Module 5 : Traitement des données sans serveur

- Traitement, transformation et analyse des données sans serveur
- Utilisation d'AWS Glue avec les charges de travail Amazon EMR

Atelier 3 : orchestrer le traitement des données dans Spark à l'aide d'AWS Step Functions

Module 6 : Sécurité et surveillance des clusters Amazon EMR

- Sécurisation des clusters EMR
- Démonstration interactive 3 : cryptage côté client avec EMRFS
- Surveillance et dépannage des clusters Amazon EMR
- Démo : Examen de l'historique du cluster Apache Spark

Module 7 : Conception de solutions d'analyse de données par lots

- Cas d'utilisation de l'analyse de données par lots

Activité 4 : Conception d'un flux de travail d'analyse de données par lots

Module B : Développement d'architectures de données modernes sur AWS

- Architectures de données modernes

Méthodes pédagogiques :

Les participants réalisent un test d'évaluation des connaissances en amont et en aval de la formation pour valider les connaissances acquises pendant la formation.
Un support de cours officiel sera remis aux stagiaires.

Autres moyens pédagogiques et de suivi:

- Compétence du formateur : Les experts qui animent la formation sont des spécialistes des matières abordées et ont au minimum cinq ans d'expérience d'animation. Nos équipes ont validé à la fois leurs connaissances techniques (certifications le cas échéant) ainsi que leur compétence pédagogique.
- Suivi d'exécution : Une feuille d'émargement par demi-journée de présence est signée par tous les participants et le formateur.
- En fin de formation, le participant est invité à s'auto-évaluer sur l'atteinte des objectifs énoncés, et à répondre à un questionnaire de satisfaction qui sera ensuite étudié par nos équipes pédagogiques en vue de maintenir et d'améliorer la qualité de nos prestations.

Délais d'inscription :

- Vous pouvez vous inscrire sur l'une de nos sessions planifiées en inter-entreprises jusqu'à 5 jours ouvrés avant le début de la formation sous réserve de disponibilité de places et de labs le cas échéant.
- Votre place sera confirmée à la réception d'un devis ou "booking form" signé. Vous recevrez ensuite la convocation et les modalités d'accès en présentiel ou distanciel.
- Attention, si cette formation est éligible au Compte Personnel de Formation, vous devrez respecter un délai minimum et non négociable fixé à 11 jours ouvrés avant le début de la session pour vous inscrire via moncompteformation.gouv.fr.

Accueil des bénéficiaires :

- En cas de handicap : plus d'info sur globalknowledge.fr/handicap
- Le Règlement intérieur est disponible sur globalknowledge.fr/reglement