[sfeir] Institute

AWS | AWSBBDAS

Building Batch Data Analytics Solutions on AWS

1 jour(s) / 7h

Objectifs pédagogiques

- Comparer les fonctionnalités 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, y compris la compression, pour optimiser le stockage des données
- Sélectionner et déployer les options appropriées pour ingérer, transformer et stocker des données
- Choisir les types d'instance et de nœud, les clusters, la mise à l'échelle automatique et la topologie de réseau appropriés pour un cas d'utilisation métier 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
- Sécuriser les données au repos et en transit
- Surveiller les charges de travail analytiques pour identifier et résoudre les problèmes
- Appliquer les meilleures pratiques de gestion des coûts

Public cible

- Ingénieurs plateformes de données
- Architectes et opérateurs qui construisent et gèrent des pipelines d'analyse de données

Prérequis

Les participants ayant au moins un an d'expérience dans la gestion de frameworks de données open source tels qu'Apache Spark ou Apache Hadoop bénéficieront de ce cours. Nous suggérons le cours AWS Hadoop Fundamentals pour ceux qui ont besoin d'un rappel sur Apache Hadoop.

Nous recommandons aux participants de ce cours d'avoir suivi les cours suivants :

- AWS Technical Essentials ou Architecting sur AWS
- Building Data Lakes on AWS ou Getting Started with AWS Glue

Programme

Module A : Présentation 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 : Présentation d'Amazon EMR

- Utilisation d'Amazon EMR dans les solutions d'analyse
- Architecture de cluster Amazon EMR
- Stratégies de gestion des coûts

Module 2 : Pipeline d'analyse de données à l'aide d'Amazon EMR : ingestion et stockage

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

Module 3 : Analyse de données par lots hautes performances à l'aide d'Apache Spark sur Amazon EMR

- Cas d'utilisation d'Apache Spark sur Amazon EMR
- Pourquoi Apache Spark sur Amazon EMR
- Concepts de Spark
- Transformation, traitement et analytique
- Utilisation de blocs-notes avec Amazon EMR
- Mise en pratique 1 : Analyse de données à faible latence à l'aide d'Apache Spark sur Amazon EMR

Module 4 : Traitement et analyse des données de lot avec Amazon EMR et Apache Hive

- Utilisation d’Amazon EMR avec Hive pour traiter les données par lots
- Transformation, traitement et analytique
- Introduction à Apache HBase sur Amazon EMR
- Mise en pratique 2 : traitement de données par lots à l’aide d’Amazon EMR avec Hive