

AWS : BIGDATA AVEC HADOOP EMR

CODE STAGE : VC203

OBJECTIFS

Savoir mettre en oeuvre les techniques de calcul distribué avec Hadoop EMR

DURÉE

3 jours

PUBLIC

Architectes, chefs de projets souhaitant bénéficier des services offerts par la distribution Hadoop fournie par AWS

PRÉ-REQUIS

Connaissance des fondamentaux du cloud

PROGRAMME

Présentation AWS Hadoop EMR

Rappels des services de base:

EC2 (Elastic Compute Cloud), VPC (Virtual Private Cloud), S3 (Simple Storage Service), EBS (Elastic Block Storage), RDS (Relational Database Service)

Caractéristiques du calcul distribué et du service EMR (Amazon Elactis MapReduce)

La distribution Hadoop EMR

Historique du projet hadoop

Les fonctionnalités : stockage, outils d'extraction, de conversion, ETL, analyse,

Exemples de cas d'utilisation sur des grands projets.

Les principaux composants :HDFS pour le stockage et YARN pour les calculs.

Les distributions et leurs caractéristiques. Composants de la distribution Hadoop EMR

Mise en oeuvre

Démonstrations sur une architecture Hadoop multi-noeuds.

TP : mise en place d'une configuration de base avec HBase. Cycle de fonctionnement. Transferts par s3. Envoi des travaux. Visualisation des résultats.

Suivi des travaux avec Hue. Interactivité avec les Notebooks : jupyter.
Configuration des composants de la distribution.

Exploitation

Gestion des évènements avec Events. Contrôle du réseau et des VPC.

Automatisation de l'exécution sur AWS depuis un poste local. Suivi distant des travaux.

Journalisation, visualisation des logs. Utilisation de la ferme aws en débordement d'une ferme locale.

Sécurité : mise en place d'une configuration de sécurité. Liens avec IAM.

Optimisation

Analyse des performances. Déploiement d'une configuration avec Spark. Evaluation des performances par rapport à une ferme locale. Comparaison des coûts.

Mise en place d'automates d'optimisation.

