Hadoop : développement

CB033

Durée: 2 jours

1 610 €

11 au 12 juillet
3 au 4 octobre
8 au 9 février
19 au 20 décembre

Public :

Développeurs, Chefs de projets, data-scientists, architectes, ...

Objectifs :

A l'issue de la formation, le stagiaire sera capable de développer des applications compatibles avec la plateforme Hadoop d'Apache pour traiter des données Big Data.

Connaissances préalables nécessaires :

Avoir la connaissance d'un langage de programmation objet comme Java et du scripting

Objectifs pédagogiques :

Comprendre l'écosystème Hadoop Cloudera/Hortonworks
Présenter les principes du Framework Hadoop
Mettre en oeuvre des tâches Hadoop pour extraire des éléments pertinents d'ensembles de données volumineux et variés
Développer des algorithmes parallèles efficaces avec MapReduce
Charger des données non structurées des systèmes HDFS et HBase

Programme :

Comprendre l'écosystème Hadoop

Les fonctionnalités du framework Hadoop. Les différentes versions.
Distributions : Apache, Cloudera, Hortonworks, EMR, MapR, DSE.
Spécificités de chaque distribution.
Architecture et principe de fonctionnement. Zoom sur la distribution Cloudera/Hortonworks
Terminologie : NameNode, DataNode, ResourceManager, NodeManager. Rôle des différents composants. Le projet et les modules : Hadoop Common, HDFS, YARN, Spark, MapReduce, Hue, Oozie, Pig, Hive, HBase, Zeppelin, ...

Atelier : Manipulations de base sur la console Hadoop

Présenter les principes du Framework Hadoop

Le projet et les modules : Hadoop Common, HDFS, YARN, Spark, MapReduce
Utilisation de yarn pour piloter les jobs map/reduce.
Infrastructure/mise en oeuvre : Avro, Ambari, Zookeeper, Pig, Tez, Oozie. Vue d'ensemble. Gestion des données. Exemple de sqoop.
Restitution : webhdfs, hive, Hawq, Mahout, ElasticSearch, ...
Outils complémentaires de traitement : Spark, SparkQL, SparkR, Spark/ML, Storm, BigTop ; outils de développement : Cascading, Scalding, Flink; outils d'analyse : RHadoop, Hama, Chukwa, kafka

Atelier : exécution de jobs sur la ferme Hadoop

Mettre en oeuvre des tâches Hadoop pour extraire des éléments pertinents d'ensembles de données volumineux et variés

Lac de données. Construction et utilisation. Exploitation des données du lac.
Les différents outils : Yarn, MapReduce, Spark, Hive, Pig
Différentes solutions : calculs en mode batch, ou en temps réel,sur des flux de données ou des données statiques.
Principe de map/reduce et exemples d'implémentations, langages et sur-couches.
Découpage des travaux (jobs) avec stockage intermédiaire. Le format parquet.

Atelier : développement d'un extracteur de données et qualification de la donnée.

Développer des algorithmes parallèles efficaces avec MapReduce

Principe et objectifs du modèle de programmation map/reduce.
Configuration des jobs, notion de configuration.
Les interfaces principales : mapper, reducer, fonctions map() et reduce(). Couples (clés, valeurs).
Implémentation par le framework Hadoop.
Etude de la collection d'exemples.

Atelier : Réduction de la donnée extraite précedemment. Recherche et scores.

Charger des données non structurées des systèmes HDFS et HBase

Format des données : texte, json, csv, parquet, ...
Format des entrées et sorties d'un job map/reduce : InputFormat et OutputFormat.

Atelier : type personnalisés : création d'un writable spécifique. Utilisation. Contraintes.

Accès à des systèmes externes : S3, hdfs, har, hbase, ...
Outils d'interfaçage entre les différents composants

Atelier : Ecriture d'un ETL HDFS vers HBase