BigData : intégration SQL, Hive, SparkDataFrames

CB045

Durée: 2 jours

1 570 €

22 au 23 mai
11 au 12 septembre
6 au 7 février
27 au 28 novembre

Public :

Experts en bases de données relationnelles, chefs de projet.

Objectifs :

Comprendre les connexions existantes entre les mondes relationnels et NoSQL en environnement Big Data. Savoir mettre en oeuvre Hive et Pig, Impala, les Spark Dataframes.

Connaissances préalables nécessaires :

Connaissance générale des systèmes d'informations et des bases de données.

Programme :

Présentation

Besoin. Adéquation entre les objectifs et les outils.
Faciliter la manipulation de gros volumes de données en conservant une approche utilisateurs.
Rappels sur le stockage : HDFS, Cassandra, HBase
et les formats de données : parquet, orc, raw, clés/valeurs
Les outils : Hive, Impala, Tez, Presto, Drill, Pig, SparkQL

Hive et Pig

Présentation. Mode de fonctionnement. Rappel sur map/reduce.
Hive : le langage HiveQL. Exemples.
Pig : le langage pig/latin. Exemples.

Impala

Présentation. Cadre d'utilisation. Contraintes. Liaison avec le metastore Hive.

Atelier : mise en évidence des performances.

Presto

Cadre d'utilisation. Sources de données utilisables.

Atelier : mise en oeuvre d'une requête s'appuyant sur Cassandra et PostgreSQL.

Spark DataFrame

Les différentes approches. Syntaxe SparkQL. APIs QL.
Compilation catalyst. Syntaxe, opérateurs.

Atelier : mise en oeuvre d'une requête s'appuyant sur HBase et HDFS.

Drill

Utilisation d'APIs JDBC, ODBC. Indépendance Hadoop. Contraintes d'utilisation. Performances.

Comparatifs

Compatibilité ANSI/SQL. Approches des différents produits.
Critères de choix.