

Talend Open Studio for Big Data, exploiter vos données massives



SII-298 2 Jours (14 Heures)

Description

La plateforme d'intégration de données de Talend étend ses possibilités aux technologies Big Data que sont Hadoop (HDFS, HBase, HCatalog, Hive et Pig) et les bases NoSQL Cassandra et MongoDB. Ce stage vous apportera les bases pour bien utiliser les composants Talend créés pour communiquer avec les systèmes Big Data.

À qui s'adresse cette formation ?

Pour qui

Gestionnaires de données, architectes, consultants en informatique décisionnelle.

Prérequis

Expérience dans l'utilisation de l'outil Talend Open Studio For Data Integration ou compétences acquises durant la formation "Talend Open Studio, mettre en œuvre l'intégration de données"

Les objectifs de la formation

- Lire/écrire des données sur HDFS/HBase/HCatalog Réaliser des jobs de transformation à l'aide de Pig et Hive Utiliser Scoop pour faciliter la migration de bases de données relationnelles dans Hadoop Adopter des bonnes pratiques et concevoir des Systèmes d'informations flexibles et robustes

Programme de la formation

Présentation de Talend Open Studio for Big Data

- Problématique du Big Data : le modèle de 3V, les cas d'usage.
- L'écosystème Hadoop (HDFS, MapReduce, HBase, Hive, Pig...).
- Données non structurées et bases de données NoSQL.
- TOS for Big Data versus TOS for Data Integration.
- Travaux pratiques : Installation/configuration de TOS for Big Data et d'un cluster Hadoop (Cloudera ou Hortonworks), vérification du bon fonctionnement.

Intégration de données dans un cluster et des bases de données NoSQL

- Définition des métadonnées de connexion du cluster Hadoop.
- Connexion à une base de MongoDB, Neo4j, Cassandra ou Hbase et export de données.
- Intégration simple de données avec un cluster Hadoop.
- Capture de tweets (composants d'extension) et importation directe dans HDFS.
- Travaux pratiques : Lire des tweets et les stocker sous forme de fichiers dans HDFS, analyser la fréquence des thèmes abordés et mémorisation du résultat dans HBase.
-

Import / Export avec SQOOP

- Utiliser Sqoop pour importer, exporter, mettre à jour des données entre systèmes RDBMS et HDFS.
- Importer/exporter partiellement, de façon incrémentale de tables.
- Importer/Exporter une base SQL depuis et vers HDFS.
- Les formats de stockage dans le Big Data (AVRO, Parquet, ORC...).
- Travaux pratiques : Réaliser une migration de tables relationnelles sur HDFS et réciproquement.
-

Effectuer des manipulations sur les données

- Présentation de la brique PIG et de son langage PigLatin.
- Principaux composants Pig de Talend, conception de flux Pig.
- Développement de routines UDF.
- Travaux pratiques: Dégager les tendances d'utilisation d'un site Web à partir de l'analyse de ses logs.
-

Architecture et bonnes pratiques dans un cluster Hadoop

- Concevoir un stockage efficient dans HADOOP.
- Datalake versus Datawarehouse, doit-on choisir ?
- HADOOP et le Plan de Retour d'Activité (PRA) en cas d'incident majeur.
- Automatiser ses workflows.
- Travaux pratiques: Créer son datalake et automatiser son fonctionnement.
-

Analyser et entreposer vos données avec Hive

- Métadonnées de connexion et de schéma Hive.
- Le langage HiveQL.
- Conception de flux Hive, exécution de requêtes.
- Mettre en œuvre les composants ELT de Hive.
- Travaux pratiques: Stocker dans HBase l'évolution du cours d'une action, consolider ce flux avec Hive de manière à matérialiser son évolution heure par heure pour une journée donnée.
-