Talend Open Studio for Big Data, exploiter ses données massives

Cours Pratique de 2 jours - 14h
Réf : TAD - Prix 2024 : 1 620CHF HT

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 cours vous apportera les bases pour bien utiliser les composants Talend crées pour communiquer avec les systèmes Big Data.

OBJECTIFS PÉDAGOGIQUES

À l'issue de la formation l'apprenant sera en mesure de :

Ecrire des données sur HDFS et dans des BDD NoSQL avec des jobs Talend

Réaliser des jobs de transformation à l'aide de Pig et Hive

Utiliser Sqoop pour faciliter la migration de bases de données relationnelles dans Hadoop

Adopter des bonnes pratiques et concevoir des Systèmes d'informations flexibles et robustes

TRAVAUX PRATIQUES

Succession de mini-projets donnant lieu à la conception de jobs Talend Big Data de difficulté croissante.

LE PROGRAMME

dernière mise à jour : 03/2022

1) 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.

2) 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.

3) 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.

PARTICIPANTS

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", Réf. TOT.

COMPÉTENCES DU FORMATEUR

Les experts qui animent la formation sont des spécialistes des matières abordées. Ils ont été validés par nos équipes pédagogiques tant sur le plan des connaissances métiers que sur celui de la pédagogie, et ce pour chaque cours qu'ils enseignent. Ils ont au minimum cinq à dix années d'expérience dans leur domaine et occupent ou ont occupé des postes à responsabilité en entreprise.

MODALITÉS D'ÉVALUATION

Le formateur évalue la progression pédagogique du participant tout au long de la formation au moyen de QCM, mises en situation, travaux pratiques…
Le participant complète également un test de positionnement en amont et en aval pour valider les compétences acquises.

MOYENS PÉDAGOGIQUES ET TECHNIQUES

• Les moyens pédagogiques et les méthodes d'enseignement utilisés sont principalement : aides audiovisuelles, documentation et support de cours, exercices pratiques d'application et corrigés des exercices pour les stages pratiques, études de cas ou présentation de cas réels pour les séminaires de formation.
• À l'issue de chaque stage ou séminaire, ORSYS fournit aux participants un questionnaire d'évaluation du cours qui est ensuite analysé par nos équipes pédagogiques.
• Une feuille d'émargement par demi-journée de présence est fournie en fin de formation ainsi qu'une attestation de fin de formation si le stagiaire a bien assisté à la totalité de la session.

MODALITÉS ET DÉLAIS D'ACCÈS

L'inscription doit être finalisée 24 heures avant le début de la formation.

ACCESSIBILITÉ AUX PERSONNES HANDICAPÉES

Vous avez un besoin spécifique d'accessibilité ? Contactez Mme FOSSE, référente handicap, à l'adresse suivante psh-accueil@orsys.fr pour étudier au mieux votre demande et sa faisabilité.

4) 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.

5) 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.

6) 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.

LES DATES

CLASSE À DISTANCE
2024 : 21 oct.