Talend, intégration de données pour le Big Data

Cours Pratique de 3 jours - 21h
Réf : IDB - Prix 2024 : 2 280CHF 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 :

Maîtriser Talend dans un environnement Big Data

Se servir de Talend comme lien entre les fichiers, applications et bases de données

Acquérir la philosophie de l'outil

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

Être capable d'implémenter ses Jobs

Lire et écrire des données sur HDFS et dans des bases de données NoSQL avec des Jobs Talend

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

Gérer la qualité de la donnée avec Talend

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

Maîtriser l'utilisation de la bibliothèque de composants

Effectuer des traitements ETL (Extract, Transform and Load) simple et complexes de bout en bout

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

LE PROGRAMME

dernière mise à jour : 11/2022

1) Présentation de Talend Open Studio

- L'intégration de données. Les solutions ETL.
- Le Big Data. Données non structurées. Bases de données NoSQL.
- L'écosystème Hadoop (HDFS, MapReduce, HBase, Hive, Pig...).
- TOS for Data Integration : intégration des données.
- TOS for Data Quality : gestion de la qualité de la donnée.
- TOS for Big Data.
- Philosophie du produit.

Travaux pratiques : Installation/configuration de TOS for Big Data. Prise en main.

2) Concevoir des Jobs

- Présentation de Business Modeler, de Job Designer.
- Composants de transformation simples.
- Visualiser du code généré, exécuter un job.
- Paramétrer les jobs.

PARTICIPANTS

Consultants BI, Architectes, chefs de projets, gestionnaires de données ou toute personne devant gérer des flux de données.

PRÉREQUIS

Avoir des connaissances en Hadoop, Spark et Kafka.

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.

- Créer et gérer ses propres variables.
- Bonnes pratiques de conception.

Travaux pratiques : Développement d'un job se connectant à une source de données, filtrage, transformation et stockage du résultat dans un fichier.

3) 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.
- Présentation de composants d'extension.
- Utilisation du composant d'extension : capture de tweets 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.

4) 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 des 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.

5) 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.

6) 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 Reprise d'Activité (PRA) en cas d'incident majeur.
- Automatiser ses workflows.

Travaux pratiques : Créer son datalake et automatiser son fonctionnement.

7) 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 : 24 juin, 16 sept., 16 déc.