Traitement de données Big Data en temps réel avec Spark et Storm

Cours Pratique de 3 jours - 21h
Réf : DSS - Prix 2025 : 2 330 HT

Le Big Data, connu pour sa capacité à traiter des données massives, intègre avec l'arrivée d'outils comme Spark et Storm la composante temps réel. Vous découvrirez dans ce cours les atouts de ces outils, leur système de calcul distribué temps réel ainsi que la notion de diffusion en continue (Streaming Big Data).

OBJECTIFS PÉDAGOGIQUES

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

Comprendre les fondamentaux du développement d'applications Big Data en temps réel

Evaluer les outils Spark et Storm

Appliquer les systèmes de calcul distribué temps réel de Storm et de Spark

Traiter des grosses quantités de données en temps réel

LE PROGRAMME

dernière mise à jour : 05/2024

1) Introduction à l'architecture temps réel

- Traitements temps réel.
- Les architectures LAMBDA.
- Les architectures KAPPA.
- Les architectures SMACK.

Travaux pratiques : Etude de l'implémentation d'une architecture KAPPA pour Spark et Strom.

2) L'architecture Kafka

- L'aperçu de Kafka Producers, Brokers, Consumers.
- Les fichiers journaux de Kafka.
- Les schémas Avro. Utilisation de ZooKeeper.

Travaux pratiques : Étude de la configuration de Kafka dans l'architecture KAPPA.

3) L'architecture Apache Storm

- Définition de l'environnement de développement.
- Création de projets basés sur Storm.
- Définition des composants Storm (Spout et Bolt).
- Définition des flux Storm.
- Modèle de données (clé, valeur).
- Rôles des Nimbus et ZooKeeper.

Etude de cas : Étude de l'implémentation de l'architecture KAPPA pour Storm.

4) Le traitement des messages Storm

- Programmation de services avec Clojure, Java, Python.
- Cycle de vie d'un message.
- L'API Storm pour la définition de la fiabilité.

PARTICIPANTS

Concepteurs, développeurs, architectes.

PRÉREQUIS

Bonnes connaissances en développement logiciel. Une compréhension des problèmes liés au Big Data est un plus.

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

Pour toute question ou besoin relatif à l'accessibilité, vous pouvez joindre notre équipe PSH par e-mail à l'adresse psh-accueil@orsys.fr.

- Stratégie de mise en place de la fiabilité pour une application utilisant le Big Data.
Travaux pratiques : Mise en œuvre d'un projet de traitement des réseaux sociaux en temps réel dans l'architecture KAPPA.

5) L'architecture d'Apache Spark
- Les différentes versions de Spark (Scala, Python, R et Java).
- Comparaison avec l'environnement Storm.
- Les différents modules de Spark.
- Les différents types d'architecture : Standalone, Apache Mesos ou Hadoop YARN.
Travaux pratiques : Etude de l'implémentation de l'architecture SMACK pour Spark.

6) Le temps réel avec Spark Streaming
- Présentation des RDD.
- Créer, manipuler et réutiliser des RDD.
- Accumulateurs et variables broadcastées.
- Principe de fonctionnement.
- Les différents types de sources.
- Comparaison avec Apache Storm.
Travaux pratiques : Mise en œuvre d'un projet de traitement des réseaux sociaux en temps réel.

7) Les autres acteurs du marché
- Comparaison de tous les outils de streaming de l'écosystème (Storm, Spark Streaming, Flink, Samza).
- Focus sur l'architecture Samza.
Travaux pratiques : Étude de l'implémentation de l'architecture KAPPA avec Samza.

LES DATES

CLASSE À DISTANCE
2025 : 17 déc.