Spark Python, développer des applications pour le big data

Cours Pratique de 3 jours - 21h
Réf : QNC - Prix 2024 : 1 870€ HT

Spark est un framework de calcul distribué permettant de manipuler des données volumineuses. Conçu au départ pour accélérer les traitements d'Hadoop, il est devenu un système autonome. Il peut se programmer avec quatre langages, dont Python, devenu prédominant. Ce cours vous fait découvrir Spark Python.

OBJECTIFS PÉDAGOGIQUES

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

Découvrir les concepts fondamentaux de Spark

Utiliser le concept des RDD de Spark

Exploiter des données avec Spark SQL

Effectuer de l'analyse en temps réel avec Spark Streaming

Utiliser Spark avec les notebooks Jupyter, manipuler les données avec Pyspark comme avec Pandas

Aborder le machine learning avec Spark

MÉTHODES PÉDAGOGIQUES

Chaque sujet est illustré par des démonstrations se déroulant sur un cluster dans le cloud. Les participants réalisent des exercices après la présentation des concepts.

EXERCICE

De nombreux exercices sont réalisés pour illustrer les sujets.

LE PROGRAMME

dernière mise à jour : 02/2023

1) Présentation d'Apache Spark

- Historique du framework.
- Les quatre principaux composants : Spark SQL, Spark Streaming, MLlib et GraphX.
- Les outils et les librairies Python pour Spark : PySpark, notebooks Jupyter, Koalas.
- Les concepts de programmation de Spark.
- Exécuter Spark dans un environnement distribué.

Travaux pratiques : Mise en place de l'environnement Python pour Spark. Mise en œuvre de scripts manipulant des concepts de Spark.

2) Utiliser Spark avec Python : les resilient distributed datasets (RDD)

- Configurer son environnement Python.
- Se connecter à Spark avec Python : les contextes et les sessions.
- Présentation des RDD. Créer, manipuler et réutiliser des RDD.
- Les principales fonctions/transformations, mise en œuvre d'algorithmes de type map/reduce.
- Accumulateurs et variables broadcastées.
- Utiliser des partitions.
- Utiliser les notebooks et soumettre des jobs Python.

Travaux pratiques : Manipulation de contextes et de sessions. Création et réutilisation de RDD. Soumission de travaux.

3) Manipuler des données structurées

- Présentation de Spark SQL et des DataFrames et datasets.

PARTICIPANTS

Toute personne connaissant Python souhaitant découvrir le framework Spark de la fondation Apache.

PRÉREQUIS

Bonne pratique du langage Python.

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.

- Les différents types/formats de sources de données.
- Interopérabilité avec les RDD.
- Utiliser la librairie PySpark Pandas.

Travaux pratiques tutorés : Exécution de requêtes avec Spark SQL. Mise en œuvre de DataFrames et datasets. Manipulation de DataFrame.

4) Machine learning avec Spark

- Introduction au machine learning.
- Les différentes classes d'algorithmes.
- Présentation de MLlib.
- Implémentation des différents algorithmes dans MLlib.

Travaux pratiques : Mise en œuvre d'apprentissages supervisés au travers d'une classification.

5) Analyser en temps réel avec Spark Streaming

- Comprendre l'architecture du streaming.
- Présentation des Discretized Streams (DStreams).
- Les différents types de sources.
- Manipulation de l'API (agrégations, watermarking...).
- Machine learning en temps réel.

Travaux pratiques : Création de statistiques en temps réel à partir d'une source de données et prédictions à l'aide du machine learning.

6) Théorie des graphes

- Introduction à la théorie des graphes (nœuds, arêtes, graphes orientés, chemins, principaux algorithmes).
- Utilisation de l'API.
- Présentation des librairies GraphX et GraphFrame.

Travaux pratiques : Mise en œuvre d'un algorithme de recherche du plus court chemin ou page rank et visualisation du graphe.

LES DATES

CLASSE À DISTANCE
2024 : 03 juil., 02 oct., 09 déc.

PARIS
2024 : 26 juin, 25 sept., 02 déc.