retengr

Apache Spark FOAD

Durée : 3 jours

Date de mise à jour : 21/03/2026

Méthode pédagogique

Composée à 70% de pratique, cette formation à distance utilise des exercices illustrés, présentés et accomplis avec le formateur, et des exercices à réaliser en autonomie.

Une journée se décompose de la façon suivante :

Matin : 2h de théorie en visioconférence, 10 à 15 minutes de présentation d'exercices en visio, 1h de TP en autonomie avec possibilité de solliciter le formateur (partage d'écran à distance pour une assistance efficace).

Après-midi : 2h de théorie en visioconférence, 10 à 15 minutes de présentation d'exercices en visio, 1h de TP en autonomie avec possibilité de solliciter le formateur (partage d'écran à distance pour une assistance efficace).

Une évaluation quotidienne de l'acquisition des connaissances de la veille est effectuée.

Une synthèse est proposée en fin de formation.

Une évaluation à chaud sera proposée au stagiaire à la fin du cours.

Un support de cours (version électronique) sera remis à chaque participant comprenant les slides sur la théorie, les exercices. L'émargement par demi-journée de présence se fera de façon numérique.

Enfin, une attestation de formation sera envoyée si le stagiaire a bien assisté à la totalité de la session.

En ce qui concerne le matériel informatique du stagiaire, il est seulement préconisé un ordinateur et une connexion internet. Nous nous chargeons du reste.

Chaque participant se verra attribuer une Machine Virtuelle qui sera exécutée dans le Cloud d'Amazon. Il disposera alors de la puissance et des outils nécessaires pour le bon déroulement de la formation. Aucune installation de la part du participant n'est requise avant la formation.

40 rue des Filatiers, 31000 Toulouse - Organisme de formation : 73 31 08023 31 - SIRET : 812 673 903 00017

Présentation

Apache Spark est un moteur de traitements distribués sur des gros volumes de données.

Souvent mis en opposition au modèle mapreduce implémenté dans Hadoop, il en est en fait une extension qui peut en diviser les temps d'exécution jusqu'à un facteur de 100 en maximisant le travail «in-memory».

Spark exploite les principes de programmation fonctionnelle afin d'optimiser l'empreinte mémoire nécessaire à son exécution. Conçu pour mettre en œuvre des traitements distribués, Spark peut s'appuyer sur plusieurs types de clusters, dont YARN le négociateur de ressources intégré à Hadoop.

Objectifs

- Concevoir une application avec Spark
- Comprendre le principe de distribution des traitements
- Maîtriser les concepts fondamentaux des et des Resilient Distributed Dataset
- Utiliser les dataframes via Spark SQL
- Utiliser SparkUI afin d'analyser les jobs et tâches de Spark
- Positionner SparkML dans un contexte de data science
- Traiter des données en continu avec Spark Streaming

Audience

Architectes, Chefs de projet, Data Scientists, Développeurs

40 rue des Filatiers, 31000 Toulouse - Organisme de formation : 73 31 08023 31 - SIRET : 812 673 903 00017

Le formateur

Le formateur est un expert du domaine qui intervient sur le sujet depuis plusieurs années en formation mais aussi en conseil.

Doté d'une grande qualité d'écoute, sa pédagogie et sa compétence technique vous permettront d'acquérir les compétences sur SPARK.

Il saura alterner entre théorie, pratique, et retours d'expérience.

Pré-requis

Connaissance d'un langage de programmation

Programme

Présentation de Spark [3.5h]

- Spark : un besoin de distribuer vos traitements
- Architecture de Spark runtime : driver, executor, master
- Positionner Spark vs Hadoop
- Les langages du framework : Java | Scala | Python | R

RDD : Resilient Distributed Dataset [3.5h]

- RDD : Le composant fondateur du fonctionnement de Spark
- Les partitions : la base de la distribution
- Transformations, actions et directed acyclic Graph
- Manipuler un RDD : Une API riche
- Le cas particulier des Pairs RDD

40 rue des Filatiers, 31000 Toulouse - Organisme de formation : 73 31 08023 31 - SIRET : 812 673 903 00017

SparkSQL, Dataframes et Datasets [3.5h]

- Un modèle de programmation haut niveau
- Initialisation d'un dataframe
- Manipulation : sélection, tri et fonctions d'agrégation.
- Dataset : une surcouche typée des dataframes
- Comprendre le plan d'exécution d'une requête
- Bonnes et mauvaises pratiques avec SparkSQL

Mise en cluster : Les infrastructures de déploiement [3.5h]

- Les composants d'une exécution Spark : Jobs, stages et tasks
- Un principe important : Data locality
- Distribution des données dans le cadre d'un cluster : les partitions
- Redistribution des données : le shuffle
- Bonnes pratiques et performance

Machine Learning [3.5h]

- Comprendre les principes fondamentaux du Machine Learning
- Apprentissage et création d'un modèle avec SparkML

Spark streaming [3.5h]

- Collecte et traitement des données en continu
- Stream processing avec Spark
- Comprendre le principe du micro-batching

40 rue des Filatiers, 31000 Toulouse - Organisme de formation : 73 31 08023 31 - SIRET : 812 673 903 00017

Modalités et délais d'accès à la formation

Les inscriptions sont possibles jusqu'à 48 heures ouvrées avant le début de la formation, en interentreprises, dans la limite des places disponibles. Pour les formations organisées en intra entreprise, la liste des participants peut être modifiée jusqu'à 24h ouvrés avant le début de la formation.

Accessibilité

RETENGR facilite l'accessibilité de ses formations.

Cette formation est accessible aux personnes en situation de handicap.

Si vous avez un besoin d'accès spécifique, contactez Céline BOURREIL (celine.bourreil@retengr.com) qui étudiera avec Handifiel's (notre référent handicap) votre demande et vous proposera les meilleures solutions

Vous allez nous adorer si comme nous vous pensez que...

Une formation doit être au service de la performance du collaborateur et de l'entreprise

Ceci nécessite une quête constante d'excellence de la part de l'organisme formateur avec une adaptation systématique aux enjeux de l'entreprise, la mise à jour régulière des supports de cours et une veille technologique indispensables pour toujours être à la pointe du domaine.

L'expertise technique est aussi importante que les qualités pédagogiques

Nos formateurs sont tous des experts de leur domaine. Mais qu'ont-ils de plus que les autres ? Nous les sélectionnons en plus pour leurs qualités de pédagogue et leurs méthodes d'enseignements. Nous plaçons les qualités pédagogiques au même niveau que l'expertise afin que nos stagiaires tirent le meilleur de leurs formations.

L’excellence naît de l’excellence

Beaucoup de nos clients se classent parmi les leaders de leurs industries respectives ou parmi les start-ups les plus prometteuses. Nous savons que former les collaborateurs de telles entreprises nécessite de prêter attention à chaque détail en prodiguant un accompagnement à la hauteur de l’ambition de nos stagiaires. C’est pourquoi nous savons faire des leaders d’aujourd’hui les champions de demain !

retengr

Faire du leader d'aujourd'hui, le champion de demain