BEST

Développer des applications pour Apache Spark 2.X avec Python ou Scala

Formation officielle Hortonworks "DEV 343 – HDP Developer: Spark 2.x Developer"

DESCRIPTION

Spark est un framework qui permet d'écrire simplement des applications distribuées complexes qui permettent de prendre des meilleures décisions plus rapidement et des actions en temps réel.
Cette formation s'adresse aux développeurs qui souhaitent créer et déployer des applications Big Data complètes et uniques en combinant batchs, le streaming et analyses interactives sur l'ensemble des données.

La formation couvre une introduction technique sur l'architecture et le fonctionnement de Spark 2.X, les éléments de base de Spark (e.g. RDDs et calcul distribué), ainsi que les abstractions plus haut niveau qui fournissent une interface plus simple et plus complète (e.g. Spark SQL, les Dataframes, les Datasets). Cette formation traitera également des problèmes de performances et stratégies d'optimisation ainsi que de l'utilisation de Spark streaming pour traiter les données en temps réel.

OBJECTIFS PÉDAGOGIQUES

Appréhender le fonctionnement et l'architecture de Spark
Développer des applications avec Apache Spark
Optimiser une application Spark
Utiliser Spark SQL, les dataframes et les datasets
Faire de l'analyse en temps réel avec Spark streaming

PUBLIC CIBLE

Data Analyste
Développeur d'applications in-memory ou avec des contraintes temps réel
Ingénieur d'études
Architecte technique
Chef de projet technique

PRÉ-REQUIS

- Connaissances de base en programmation ou en scripting (Python/Scala)
- Expérience basique en ligne de commande
- Connaissances de base sur Hadoop
- Connaissances en SQL et conception d'application temps réel utiles mais non obligatoire

MÉTHODE PÉDAGOGIQUE

Formation avec apports théoriques, échanges sur les contextes des participants et retours d'expérience pratique du formateur, complétés de travaux pratiques et de mises en situation.
Cette formation prépare à la certification éditeur Hortonworks.

PROFILS DES INTERVENANTS

Toutes nos formations sont animées par des consultants-formateurs

Stage pratique en présentiel
HADOOP HORTONWORKS

Code :
HWSPK

Durée :
4 jours (28 heures)

Certification :
300 € HT

Exposés :
40%

Cas pratiques :
50%

Échanges d'expérience :
10%

Sessions à venir :

17 - 20 déc. 2018
Paris / 2 600 eur

18 - 21 fév. 2019
Paris / 3 450 eur

8 - 11 avr. 2019
Paris / 3 450 eur

11 - 14 juin 2019
Paris / 3 450 eur

9 - 12 sept. 2019
Paris / 3 450 eur

Tarif & dates intra :
Sur demande

expérimentés et reconnus par leurs pairs.

MODALITÉS D'ÉVALUATION ET FORMALISATION À L'ISSUE DE LA FORMATION

L'évaluation des acquis se fait tout au long de la session au travers des ateliers et des mises en pratique. Une évaluation à chaud sur la satisfaction des stagiaires est réalisée systématiquement en fin de session et une attestation de formation est délivrée aux participants mentionnant les objectifs de la formation, la nature, le programme et la durée de l'action de formation ainsi que la formalisation des acquis.

POUR ALLER PLUS LOIN :

- Livre blanc "Hadoop, feuille de route"
- Toutes nos formations Hortonworks
- Formation officielle Hortonworks "Administrer la plateforme Hadoop 2.X Hortonworks : fondamentaux" ("HDP Hadoop Administration 1 foundations") (HWADM)
- Formation officielle Hortonworks "Analyse de données pour Hadoop 2.X Hortonworks avec Pig, Hive et Spark" (HDP Developer: Apache Pig and Hive) (HWAPH)
- Parcours de formation "Intelligence artificielle par la pratique : des fondamentaux à l'industrialisation"
- Formation "Fondamentaux de la Data Science" (DSFDX)
- Formation "Data Science : niveau avancé" (DSNVA)
- Formation "Cadrage et pilotage d'un projet de Data Science" (DSGDP)
- Formation "Industrialisation d'un projet de Data Science" (DSIND)
- Formation "Architecture des données : stockage et accès" (DSARC)

Programme pédagogique détaillé par journée

Jour 1

INTRODUCTION AU LANGAGE SCALA

- Travailler avec les variables, les différents types de données et les structures de contrôles
- L'interpréteur Scala
- Les collections et leurs méthodes (e.g. map)
- Travailler avec les fonctions, les méthodes et les « Function Literals »
- Définir les concepts suivants relativement au passage à l'échelle : Class, Object, Case Class

INTRODUCTION ET MOTIVATIONS POUR APACHE SPARK

- L'écosystème Spark
- Spark vs. Hadoop
- Obtenir et installer Spark
- La console Spark, et SparkContext

MISE EN PRATIQUE :

- Mettre en place l'environnement de lab
- Démarrer l'interpréteur Scala
- Premiers pas avec Apache Spark
- Premiers pas avec la console Spark

Jour 2

INTRODUCTION DES RDDS

- Les concepts de RDD, de cycle de vie, et de l'évaluation paresseuse.
- Travailler avec des RDDs : création et transformations (map, filter, etc.)
- Partitionnement et transformation des RDDs
- Transformations avancées (flatMap, explode, et split)

INTRODUCTION DES DATAFRAMES ET DATASETS

- Le concept de SparkSession
- Création et inférence de schéma
- Identification des formats supportés (dont JSON, CSV, Parquet, Text ...)
- Travailler avec l'API DataFrame
- Travailler avec l'API DataSet
- Transformations via des requêtes SQL (Spark SQL)

COMPARAISON ENTRE LES DATASETS, DATAFRAMES ET RDDS

MISE EN PRATIQUE :

- Les bases des RDD
- Opérations sur de multiples RDDs
- Les formats de données
- Les bases de Spark SQL
- Transformation de DataFrames

- L'API typée des DataSets
- Fractionner les données

Jour 3

OPTIMISATIONS :

- Shuffling, dépendances larges et étroites, et leur impact sur la performance
- L'optimiseur de requêtes Catalyst
- L'optimiseur Spark Tungsten (format binaire, gestion du cache…)
- Le caching Spark (concept, type de cache, recommandations)
- Minimiser le shuffling pour améliorer la performance
- Utilisation de la diffusion de variables et de l'accumulateur

RECOMMANDATIONS GLOBALES DE PERFORMANCES :

- L'interface Spark UI
- Les transformations efficaces
- Stockage de données
- Monitoring

MISE EN PRATIQUE :

- Comprendre le Shuffling
- Explorer l'optimiseur de requête Catalyst
- Explorer l'optimiseur Tungsten
- Travailler avec la mise en cache, le shuffling et la diffusion de variables
- Recommandations générales sur le broadcast

Jour 4

LES APPLICATIONS SPARK :

- Configurer et créer une SparkSession
- Construire et lancer des applications
- Cycle de vie des applications (Driver, Executors, et Tasks)
- Les modes d'executions (Standalone, YARN, Mesos)
- Logging et Debugging

INTRODUCTION AU TEMPS RÉEL

- Spark Streaming (Spark 1.0+)
  - DStreams, Receivers, Batching
  - Transformations Stateless
  - Transformations Windowed
  - Transformations Stateful
- Structured Streaming (Spark 2+)
  - Applications en continue
  - Le paradigme de Table, et de Result Tables
  - Les étapes du structured streaming
  - Les sources et puits
  - Introduction de Kafka

- Consommer des données Kafka
- Le Structured Streaming au format "kafka"

MISE EN PRATIQUE :

- Déclencher des jobs Spark
- Capacités additionnelles de Spark
- Spark Streaming
- Spark Structured Streaming
- Spark Structured Streaming avec Kafka