Apache Spark : développer des applications pour le Big Data

Lien :
https://innov-systems.com/formation/apache-spark-developper-des-applications-pour-le-big-data

DURÉE
4 jours (28h)

RÉFÉRENCE
BSI19

CATÉGORIE
Big Data : développement, langages et NOSQL

Innov Systems

OBJECTIFS DE LA FORMATION

À l'issue de cette formation, vous serez capable de :

- Comprendre le fonctionnement de Spark
- Intégrer Spark dans un environnement Hadoop
- Manipuler des données structurées avec Spark sql
- Faire de la programmation parallèle avec Spark sur un cluster

POUR QUI ?

- Développeurs
- Architectes

Programme détaillé

1 / Introduction au Big Data

- Introduction au Big Data
- Les challenges du Big Data
- Batch vs le temps réel dans le Big Data Analytics
- Vue d'ensemble de l'écosystème
- Les options de l'analyse en temps réel
- Quelles solutions apportent Spark au Big Data ?

2 / Introduction à Spark

- Historique du Framework
- Origine du projet
- Streaming Data - Spark
- In-memory Data - Spark
- Présentation de Spark
- Principes de base du fonctionnement de Spark
- Environnement et outils de Spark
- Comparaison avec l'environnement Apache Hadoop
- Les différents modules de Spark
- Langages supportés
- Ecosystème Spark
- Les modes de Spark
- Installation de Spark
- Vue d'ensemble de Spark en cluster
- Spark Standalone cluster
- Spark Web UI

Innov Systems

3 / Les opérations communes sur Spark

- Utilisation de Spark Shell
- Création d'un contexte Spark
- Chargement d'un fichier en Shell
- Réalisation d'opérations basiques sur un fichier avec Spark Shell
- Présentation de l'environnement de développement SBT
- Créer un projet Spark avec SBT
- Exécuter un projet Spark avec SBT
- Le mode local
- Le mode Spark
- Le caching sur Spark
- Persistance distribuée

4 / Règles de développement

- Mise en pratique en Java, Scala et Python
- Notion de contexte Spark
- Présentation des RDD (Resilient Distributed Dataset)
- Différentes méthodes de création des RDD : depuis un fichier texte, un stockage externe
- Manipuler et réutiliser des RDD
- Accumulateurs et variables broadcastées
- Utiliser des partitions
- Fonctions, gestion de la persistance
- RDD et MapReduce

5 / Analyser des données avec des requêtes sur DataFrames

- Requêter des DataFrames avec des expressions sur les colonnes nommées
- Les requêtes de groupement et d’aggrégation
- Les jointures

Inov Systems

6 / Les RDD – structure fondamentale de Spark

- Introduction aux RDD
- Les sources de données de RDD
- Créer et sauvegarder des RDD
- Les opérations sur les RDD

7 / Transformer les données avec des RDD

- Écrire et passer des fonctions de transformation
- Fonctionnement des transformations de Spark
- Conversion entre RDD et DataFrames

8 / Agrégation de données avec les rdd de paires

- Les RDD clé-valeur
- Map-Reduce : principe et usage dans Spark
- Autres opérations sur les RDD de paires

9 / Traitements

- Lecture/écriture de données : texte, JSon, Parquet, HDFS, fichiers séquentiels
- Filtrage de données, enrichissement
- Calculs distribués de base
- Travail sur les RDDs
- Lazy execution
- Impact du shuffle sur les performances
- RDD de base, key-pair RDDs

10 / Manipuler des données structurées avec Spark sql

- Spark et SQL
- L'API Dataset et DataFrames

Inov Systems

- Optimisation des requêtes
- Mise en oeuvre des Dataframes et DataSet
- Les différents types de sources de données
- Interopérabilité avec les RDD
- Performance de Spark SQL
- JDBC/ODBC server et Spark SQL CLI
- Requêter des fichiers et des vues
- L'API catalogue de Spark
- Comptabilité Hive

11 / Spark en cluster

- Différents cluster managers : Spark en autonome, avec Mesos, avec Yarn, avec Amazon EC2
- Architecture : SparkContext, Cluster Manager, Executor sur chaque noeud
- Définitions : Driver program, Cluster manager, deploy mode, Executor, Task, Job
- Mise en oeuvre avec Spark et Amazon EC2
- Soumission de jobs, supervision depuis l'interface web
- Configurer un cluster en mode Standalone
- Packager une application avec ses dépendances
- Déployer des applications avec Spark-submit
- Dimensionner un cluster

12 / Partitionnement et programmation parallèle

- Localiser les données du HDFS
- Partitionner les RDD
- Programmer et exécuter les opérations parallèles
- Mettre en cache le partitionnement des données
- Gérer la persistance des données

13 / Concevoir une application avec Spark

Inov Systems

- Présentation et configuration des propriétés de Spark
- Prototypage d’opérations avec Spark Shell
- Compilation et génération d’une application

14 / Traiter les données en temps réel avec Spark streaming

- Notion de Streaming
- Objectifs , principe de fonctionnement : stream processing
- Présentation des Discretized Streams (DStreams)
- Source de données : HDFS, Flume, Kafka, ...
- Manipulation de l'API
- Comparaison avec Apache Storm

15 / Manipuler des graphes avec GraphX

- Présentation de GraphX
- Les différentes opérations
- Créer des graphes
- Vertex and Edge RDD
- Présentation de différents algorithmes

16 / Machine learning et implémentation d’algorithmes

- Introduction au Machine Learning
- Fonctionnalités : Machine Learning avec Spark, gestion de la persistance
- Les fonctionnalités d'extraction et les statistiques basiques
- Les différentes classes d'algorithmes
- Présentation de SparkML et MLlib
- Implémentations des différents algorithmes dans MLlib
- Support de RDD
- Mise en oeuvre avec les DataFrames
- Algorithmes itératifs et résolution de problèmes

Inov Systems

- Opérations sur les données graphiques
- Les terminologies communes au Machine Learning
- Applications du Machine Learning
- Spark ML API
- Transformateurs et estimateurs
- Les pipelines
- Travailler avec un pipeline
- DAG Pipelines
- La vérification pendant l'exécution
- Passage de paramètres
- Sélection de modèles via une validation croisée
- Les types de données
- Mettre en place le Clustering en utilisant K-Means
- Gaussian Mixture
- Power Iteration Clustering (PIC)
- Latent Dirichlet Allocation (LDA)
- Le filtrage collaboratif
- Classification
- Régression
- Exemple de régression
- Linéaire

17 / Optimisation de Spark

- Gestion des variables partagées
- Données broadcastées
- Accumulateurs
- Méthodes et outils d’optimisation des performances

Innov Systems

Approche pédagogique

- Support Ecrit et Projection
- Exposés Interactifs, Podcasts et Vidéos
- Brainstorming et Jeux de Rôle
- Cas Pratiques et Labs inclus pour leur impact opérationnel
- Test de Validation des Acquis des Connaissances

Prochaines dates programmées

26 au 29 Jan. 2027	Présentiel - Casablanca
23 au 26 Mars 2027	Présentiel - Casablanca
18 au 21 Mai 2027	Présentiel - Casablanca
13 au 16 Juil. 2027	Présentiel - Casablanca
14 au 17 Sep. 2027	Présentiel - Casablanca

Autres dates possibles sur demande. Contactez-nous pour organiser une session intra-entreprise.

Réservation & Renseignements

Téléphone : +212 522 247 210
Email : contact@innov-systems.com
Web : https//www.innov-systems.com

Innov Systems

Innov Systems — Tous droits réservés