PySpark : traitement des données

Lien : https://innov-systems.com/formation/pyspark-traitement-des-donnees

DURÉE
4 jours (28h)

RÉFÉRENCE
BSI20

CATÉGORIE
Big Data : développement, langages et NOSQL

OBJECTIFS DE LA FORMATION

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

- Comprendre le principe de fonctionnement de Spark
- Utiliser l'API PySpark pour interagir avec Spark en Python
- Utiliser les méthodes de Machine Learning avec la librairie MLlib de Spark
- Traiter les flux de données avec Spark Streaming
- Manipuler les données avec Spark SQL

POUR QUI ?

- Statisticiens
- Data analysts
- Data scientists

Innov Systems

Programme détaillé

1 / Introduction à Hadoop

- Big Data
- Architecture et composants de la plateforme Hadoop
- L'architecture d'HDFS
- Prise en main des concepts de base de l'architecture Hadoop (NameNode, DataNode, ResourceManager…)
- YARN et MapReduce : comprendre les principes de fonctionnement

2 / Introduction à Spark

- Présentation Spark, origine du projet
- Apports, principe de fonctionnement (RDD, DataFrames, Data Sets)
- Spark vs Mapreduce
- Comparaison avec l'environnement Apache Hadoop
- Les différents modules de Spark
- Comment interagir avec Spark ?
- PySpark : programmer avec Spark en Python

3 / Installation de Spark

- Sur une infrastructure distribuée
- En local
- En Cloud (présentation avec Amazon AWS et Microsoft Azure)

4 / Spark pour la manipulation des données - SparkSQL

- SparkSQL et DataFrames pour manipuler des données
- Charger des données depuis Hadoop, depuis des fichiers CSV, texte, JSON…
- Transformer des données (création de DataFrames, ajout de colonnes, filtres…)

Inov Systems

5 / L'utilisation de spark.ml pour le machine learning

- Apprentissage supervisé
- Forêts aléatoires avec Spark
- Mise en place d'un outil de recommandation
- Traitement de données textuelles
- Automatiser vos analyses avec des pipelines

6 / Apache spark streaming

- Introduction à Spark Streaming
- La notion de "DStream"
- Principales sources de données
- Utilisation de l'API
- Manipulation des données

7 / Dataframe et apache spark sql

- Apache Spark SQL et le SQL Context
- Création de DataFrames
- Transformer et requêter un Dataframe
- Dataframes et RDD
- Comparaison entre Spark SQL, Impala et Hive On Spark
- Chargement et stockage de données (avec Hive, JSON…)

8 / GraphX et graphframes

- Présentation de GraphX
- Principe de création des graphes
- API GraphX
- Présentation de GraphFrames
- GraphX vs GraphFrames

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

05 au 08 Jan. 2027	Présentiel - Casablanca
02 au 05 Mars 2027	Présentiel - Casablanca
27 au 30 Avr. 2027	Présentiel - Casablanca
22 au 25 Juin 2027	Présentiel - Casablanca
24 au 27 Août 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 — Tous droits réservés