



— Phirio —

Spark : développer des applications pour le Big Data

CB037

Durée: 3 jours

2 550 €

26 au 28 janvier
1er au 3 juin

21 au 23 septembre
7 au 9 décembre

Public :

Chefs de projet, Data Scientists, Développeurs, Architectes...

Objectifs :

A l'issue de la formation, le stagiaire sera capable de maîtriser le framework Spark pour traiter des données hétérogènes et optimiser les calculs.

Connaissances préalables nécessaires :

avoir des connaissances de Java ou Python et des notions de calculs statistiques

Objectifs pédagogiques :

Maîtriser les concepts fondamentaux de Spark
Savoir intégrer Spark dans un environnement Hadoop
Développer des applications d'analyse en temps réel avec Spark Structured Streaming
Faire de la programmation parallèle avec Spark sur un cluster
Manipuler des données avec Spark SQL
Avoir une première approche du Machine Learning

Programme :

Maîtriser les concepts fondamentaux de Spark

Présentation Spark, origine du projet, apports, principe de fonctionnement. Langages supportés.
Modes de fonctionnement : batch/Streaming.
Bibliothèques : Machine Learning, IA
Mise en oeuvre sur une architecture distribuée. Architecture : clusterManager, driver, worker, ...
Architecture : SparkContext, SparkSession, Cluster Manager, Executor sur chaque noeud. Définitions : Driver program, Cluster manager, deploy mode, Executor, Task, Job

Savoir intégrer Spark dans un environnement Hadoop

Intégration de Spark avec HDFS, HBase,
Création et exploitation d'un cluster Spark/YARN. Intégration de données sqoop, kafka, flume vers une architecture Hadoop et traitements par Spark.
Intégration de données AWS S3.
Différents cluster managers : Spark interne, avec Mesos, avec Yarn, avec Amazon EC2

Atelier : Mise en oeuvre avec Spark sur Hadoop HDFS et Yarn.
Soumission de jobs, supervision depuis l'interface web



— Phirio —

Développer des applications d'analyse en temps réel avec Spark Structured Streaming

Objectifs , principe de fonctionnement: stream processing. Source de données : HDFS, Flume, Kafka, ...
Notion de StreamingContext, DStreams, démonstrations.

Atelier : traitement de flux DStreams en Scala. Watermarking. Gestion des micro-batches.

Intégration de Spark Structured Streaming avec Kafka

Atelier : mise en oeuvre d'une chaîne de gestion de données en flux tendu : IoT, Kafka, Spark Structured Streaming, Spark. Analyse des données au fil de l'eau.

Faire de la programmation parallèle avec Spark sur un cluster

Utilisation du shell Spark avec Scala ou Python. Modes de fonctionnement. Interprété, compilé.
Utilisation des outils de construction. Gestion des versions de bibliothèques.

Atelier : Mise en pratique en Java, Scala et Python. Notion de contexte Spark. Extension aux sessions Spark.

Manipuler des données avec Spark SQL

Spark et SQL

Traitement de données structurées. L'API Dataset et DataFrames

Jointures. Filtrage de données, enrichissement. Calculs distribués de base. Introduction aux traitements de données avec map/reduce.

Lecture/écriture de données : Texte, JSON, Parquet, HDFS, fichiers séquentiels.

Optimisation des requêtes. Mise en oeuvre des Dataframes et DataSet. Compatibilité Hive

Atelier : écriture d'un ETL entre HDFS et HBase

Atelier : extraction, modification de données dans une base distribuée.
Collections de données distribuées. Exemples.

Support Cassandra

Description rapide de l'architecture Cassandra. Mise en oeuvre depuis Spark. Exécution de travaux Spark s'appuyant sur une grappe Cassandra.

Spark GraphX

Fourniture d'algorithmes, d'opérateurs simples pour des calculs statistiques sur les graphes

Atelier : exemples d'opérations sur les graphes.



— Phirio —

Avoir une première approche du Machine Learning

Machine Learning avec Spark, algorithmes standards supervisés et non-supervisés (RandomForest, LogisticRegression, KMeans, ...)
Gestion de la persistance, statistiques.
Mise en oeuvre avec les DataFrames.

Atelier : mise en oeuvre d'une régression logistique sur Spark