Spark Traitement des données

Introduction à Scala pour Apache Spark

- Présentation de Scala
- Pourquoi Scala avec Spark ?
- Scala dans les autres framework
- Introduction à Scala REPL
- Les opérations basiques sur Scala
- Les types de variables dans Scala
- Les structures de contrôles dans Scala
  - Les boucles
  - Les fonctions
  - Les procédures
- Les collections dans Scala (Array, ArrayBuffer, Map, Tuples, Lists...)

Introduction au Big Data et Apache Spark

- Introduction au Big Data
- Les challenges du Big Data
- Batch vs le temps réel dans le Big Data Analytics
- Analyse en Batch Hadoop
- Vue d'ensemble de l'ecosystème
- Les options de l'analyse en temps réel
- Streaming Data Spark
- In-memory Data Spark
- Présentation de Spark
- Ecosystème Spark
- Les modes de Spark

•Installation de Spark

•Vue d'ensemble de Spark en cluster

•Spark Standalone Cluster

•Spark Web UI

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 du 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

Introduction aux RDDs

•Transformations dans le RDD

•Actions dans le RDD

•Chargement de données dans RDD

•Enregistrement des données à travers RDD

•Paire clé-valeur "RDD MapReduce" et les paires "RDD Operations"

•Intégration HDFS avec Spark et Hadoop

•Intégration YARN avec Spark et Hadoop

•Gestion des fichiers de séquences et les partitionner

Spark Streaming et MLlib

•Architecture de Spark Streaming

•Premier programme avec Spark Streaming

•Les transformations dans Spark Streaming

•La "fault tolerance" dans Spark Streaming

•Checkpointing

•Niveaux de parallélismes

•Machine Learning avec Spark

•Types de données

•Algorithmes et statistiques

•Classification et régression

•Clustering

•Filtrage collaboratif

GraphX, SparkSQL et amélioration des performances dans Spark

•Analyse de l'architecture de Hive et Spark SQL

•SQLContext dans Spark SQL

•Travailler avec les DataFrames

•Implémentation d'un exemple pour Spark SQL

•Intégration de Hive et Spark SQL

•Support pour JSON et les formats des "Parquet File"

•Implémentation de la Data Visualization avec Spark

•Chargement de données

•Les requêtes Hive à travers Spark

•Les techniques de tests dans Scala

•Les astuces d'amélioration de performance dans Spark

- Les variables partagées
- Diffusion des variables
- Partage de variables
- Accumulateurs