Spark ML

Introduction à 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'écosystème

•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 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

Spark Machine Learning

•Introduction au Machine Learning

•Les Terminologies communes au Machine Learning

•Applications du Machine Learning

•Machine Learning dans Spark

•Spark ML API

•DataFrames

•Transformateurs et estimateurs

•Les pipelines

•Travailler avec un pipeline

•DAG Pipelines

•La vérification pendant l'exécution

•Passage de paramètres

•General Machine Learning Pipeline

•Sélection de modèles via une validation croisée

•Les types supportés, les algorithmes et les utilitaires

•Les types de données

•Les fonctionnalités d'extraction et les statistiques basiques

•Clustering

•K-Means

•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
- Mettre en place une classification en utilisation la régression
- Linéaire
- Mettre en place un système de recommandations utilisant le filtrage collaboratif