SPARK fondamentaux

Référence et durée

SPARKFD
1 jour

Objectif

Comprendre le principe des traitements distribués
Comment traiter les sources de données avec SPARK
Utiliser SPARK SQL

Public

Datascientistes

Prérequis

Connaissance du framework Hadoop
Connaissance du langage SQL

Programme

- Apache SPARK – Les principes
  - Qu'est ce que Apache Spark ?
  - Framework Spark
  - Pour quels usages ?
  - Architecture distribuée
  - Session Spark (Spark context)
  - Lancement d'une application Spark
  - Web UI relatives à Spark
  - Notebook et spark
- SPARK et le RDD
  - Créer des RDD
  - Opérations principales avec les RDD
  - Redistribution RDD (shuffle)
  - Persistance d'un RDD
- SPARK et ses langages
  - SCALA
  - Python avec PYSPARK
  - Java
  - Lequel choisir ?
- DataFrame et SPARK SQL
  - Apache Spark SQL et le SQL Context
  - Création des Dataframes
  - Transformer et requêter un Dataframe
  - Persister un Dataframe
  - Dataframes et RDD
  - Comparaison entre Spark SQL, Hive on Spark
- Spark et R
  - Lien entre SPARK et R
  - Présentation de SparkLyr