SPARK ET SON UTILISATION

David RIGAUX, Mehdi DALAA, Maxime LUNDQUIST et Mandry MBUNDU

CPI2 2016-2017 TIPE

EISTI

SOMMAIRE

1

QU'EST CE QUE SPARK ?

Quelques notions de bases à propos de Spark

2

LES FONCTIONNALITÉS DE SPARK

Comment est ce que Spark fonctionne ?

3

L'ÉCOSYSTÈME DE SPARK

Librairies additionnelles permettant de travailler dans le domaine des analyses big data et du machine learning

EISTI

QU'EST CE QUE SPARK

LE DÉVELOPPEMENT

Développé par AMPLab, de l'Université UC Berkeley et passé sous forme de projet Apache en 2010

HADOOP

Permet à des applications sur clusters Hadoop d'être exécutées jusqu'à 100 fois plus vite en mémoire et 10 fois plus vite sur disque

APACHE SPARK

Framework de traitements Big Data open source réalisé pour effectuer des analyses sophistiquées et conçu pour la rapidité et la facilité d'utilisation

LE FRAMEWORK

Spark propose un framework complet et unifié pour subvenir aux besoins de traitements Big Data pour divers jeux de données, divers par leur nature aussi bien que par le type de source

APPLICATIONS

Écrire rapidement des applications en Java, Scala ou Python et inclut plus de 80 opérateurs haut-niveau. Possible d'utiliser de façon interactive pour requêter les données depuis un shell. Spark supporte les requêtes SQL, le streaming de données et des fonctionnalités de machine learning et de traitements orientés graphe

EISTI

LES FONCTIONNALITÉS DE SPARK

LANGAGE DE SPARK

Spark est écrit en Scala et s'exécute sur la Machine Virtuelle Java (JVM)

CINQ LANGAGES

Spark propose une interface de programmation de haut-niveau pour une meilleure productivité (API en Java, Scala et Python) et un shell interactif pour Scala et Python. Spark supporte également le Clojure et le R

AMÉLIORE MAPREDUCE

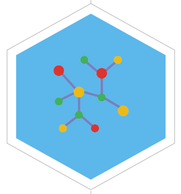
Apporte des améliorations à MapReduce grâce à des étapes de shuffle moins coûteuses. Mais Spark n'en reste pas là et propose d'autres fonctions que Map et Reduce

EVALUATION PARESSEUSE

Spark supporte les évaluations paresseuses ("lazy evaluation") des requêtes, ce qui aide à l'optimisation des étapes de traitement.

FLEXIBLE

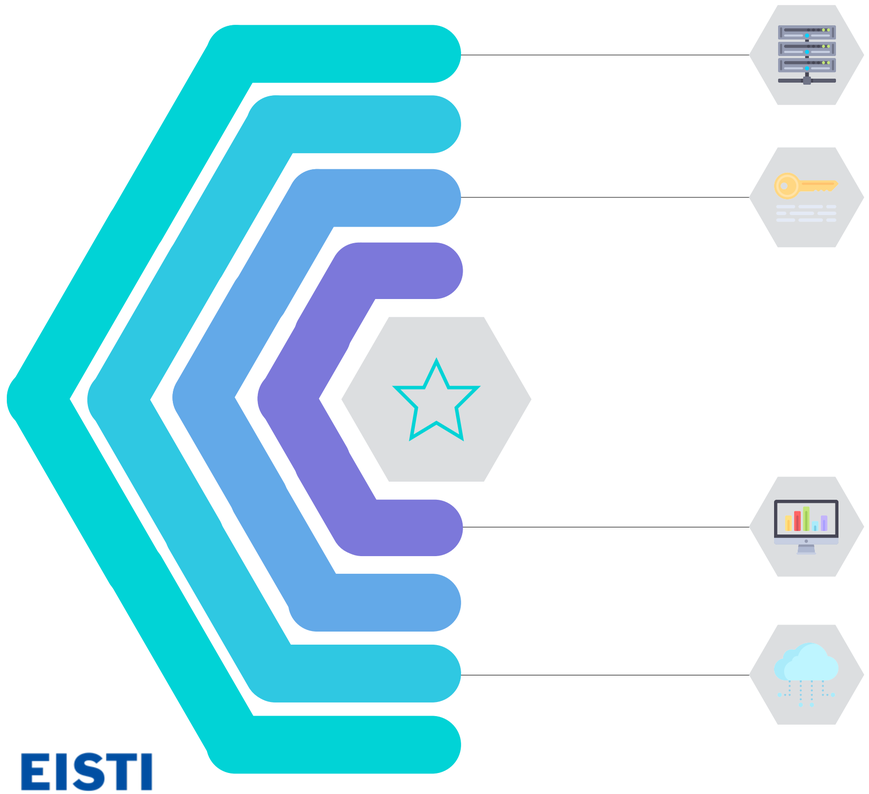
Spark maintient les résultats intermédiaires en mémoire plutôt que sur disque. Mais le moteur d'exécution est conçu pour travailler aussi bien sur les deux.



OPTIMISE LES GRAPHES

L'optimisation de graphes d'opérateurs arbitraires

L'ÉCOSYSTÈME DE SPARK



SPARK STREAMING

Peut être utilisé pour traitement temps-réel des données en flux.

SPARK SQL

permet d'exposer les jeux de données Spark via API JDBC et d'exécuter des requêtes de type SQL en utilisant les outils BI et de visualisation traditionnels.

SPARK MLLIB

Librairie de machine learning qui contient tous les algorithmes et utilitaires d'apprentissage classiques, comme la classification, la régression, le clustering, le filtrage collaboratif, la réduction de dimensions, en plus des primitives d'optimisation sous-jacentes.

SPARK GRAPHX

La nouvelle API (en version alpha) pour les traitements de graphes et de parallélisation de graphes.

SPARK ET SON UTILISATION

EISTI