Hadoop Développement

Introduction

- Les fonctionnalités du framework Hadoop
- Le projet et les modules
  - Hadoop Common
  - HDFS
  - YARN
  - Spark
  - MapReduce
- Utilisation de YARN pour piloter les jobs MapReduce

MapReduce

- Principe et objectifs du modèle de programmation MapReduce
- Fonctions "map" et "reduce"
- Couples (clés, valeurs)
- Implémentation par le framework Hadoop
- Etude de la collection d'exemples
- Rédaction d'un premier programme et exécution avec Hadoop

Programmation

- Configuration des jobs
- Notion de configuration
- Les interfaces principales
  - Mapper
  - Reducer
- La chaîne de production
  - Entrées

◦Input splits

◦Mapper

◦Combiner

◦Shuffle / sort

◦Reducer

◦Sortie

◦Partitioner

◦OutputCollector

◦Codecs

◦Compresseurs

•Format des entrées et sorties d'un job MapReduce

◦InputFormat

◦OutputFormat

•Type personnalisé : création d'un Writable spécifique

•Utilisation

•Contraintes

Outils complémentaires

•Mise en oeuvre du cache distribué

•Paramétrage d'un job

◦ToolRunner

◦Transmission de propriétés

•Accès à des systèmes externes

◦S3

◦HDFS

◦HAR

•Répartition du job sur la ferme au travers de YARN

Streaming

•Définition du streaming MapReduce

•Création d'un job MapReduce dans Python

•Répartition sur la ferme

•Avantages et inconvénients

•Liaisons avec des systèmes externes

•Introduction au pont Hadoop

•Suivi d'un job en streaming

Pig

•Pattern et best practices MapReduce

•Introduction à Pig

•Caractéristiques du langage : latin

•Installation / lancement

•Ecriture d'un script Pig

•Les fonctions de bases

•Ajouts de fonctions personnalisées

•Les UDF

•Mise en oeuvre

Hive

•Simplification du requêtage

•Syntaxe de base

•Création de tables

•Ecriture de requêtes

•Comparaison Pig / Hive

Sécurité en environnement Hadoop

- Mécanisme de gestion de l'authentification
- Configuration des ACL