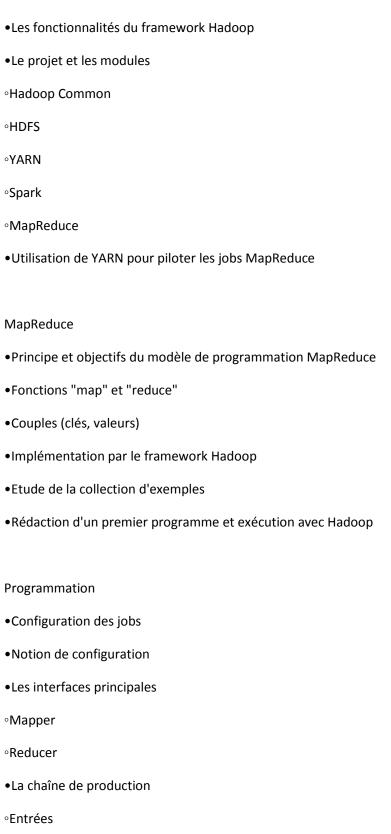
Hadoop Développement

Introduction



olnput 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
·\$3
°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