Day 4, Travaux Pratiques, Partie 2

**exercice 1 : architecture**

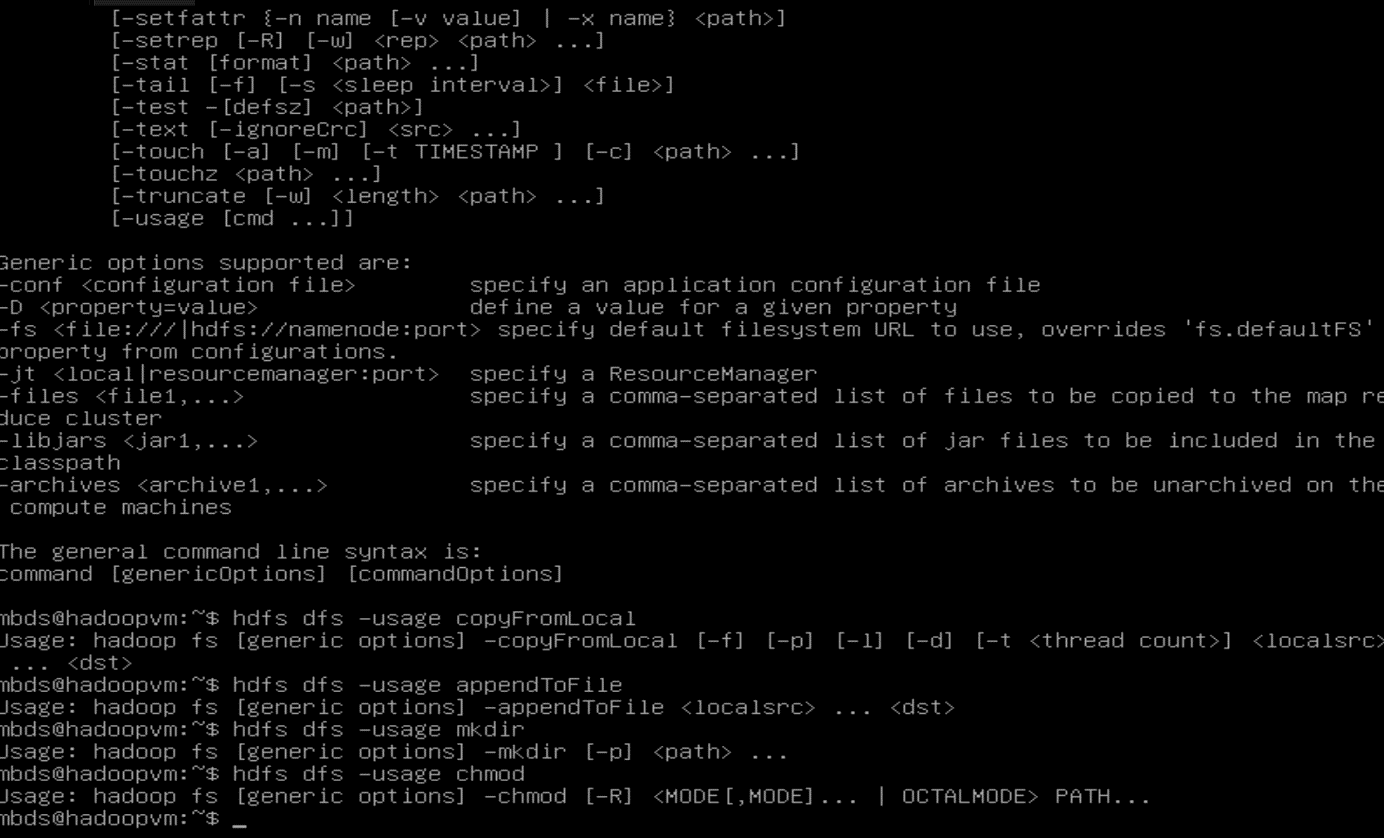
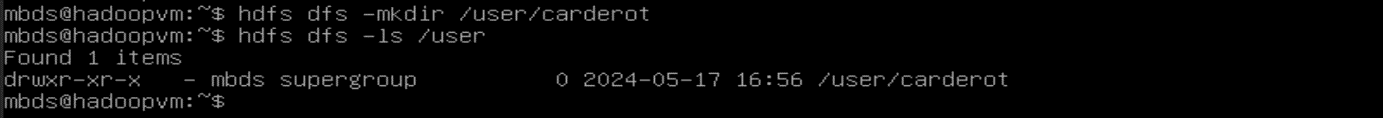
* décrire l'architecture de Hadoop

L’architecture d'Hadoop est un package de MapReduce engine (MapReduce/MR1 ou YARN/MR2), du HGFS (Hadoop Distributed File System) et du Hadoop Common

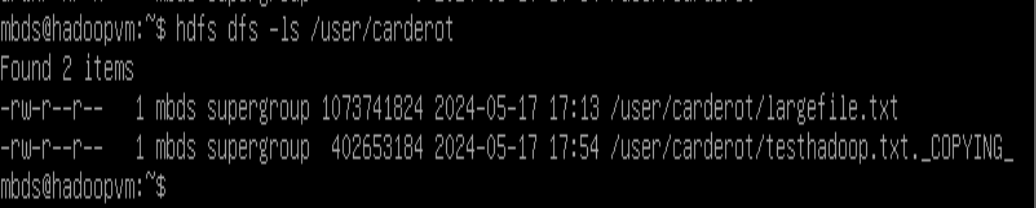
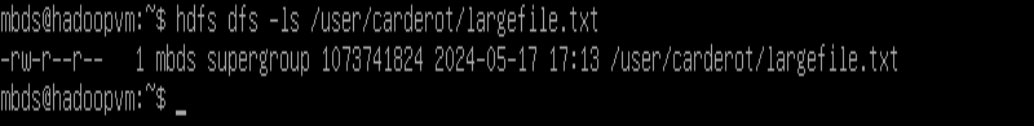
* MapReduce : un algorithme de la structure de données basé sur la framework YARN. Son rôle principal c’est de performer des processus en parallèle dans le cluster. Il permet donc d’effectuer un traitement de données en volume important de manière efficace. Il contient plusieurs fonctions Map et Reduce. Les Maps sont responsables pour la génération des rapport sur les données. Chaque fonction génère, parce, transforme et filtre les données avant de les passer vers la fonction suivante.
* YARN - Yet Another Resource Negotiator est le gestionnaire de ressources qui permet de programmer et de faire du monitoring des jobs dans Hadoop. Il sépare les tâches en petits jobs et les attribue à chaque node esclave dans le cluster. Il suit aussi l'importance de chaque job et définit leurs priorités.
* HDFS - est utilisé pour les permission de stockage.

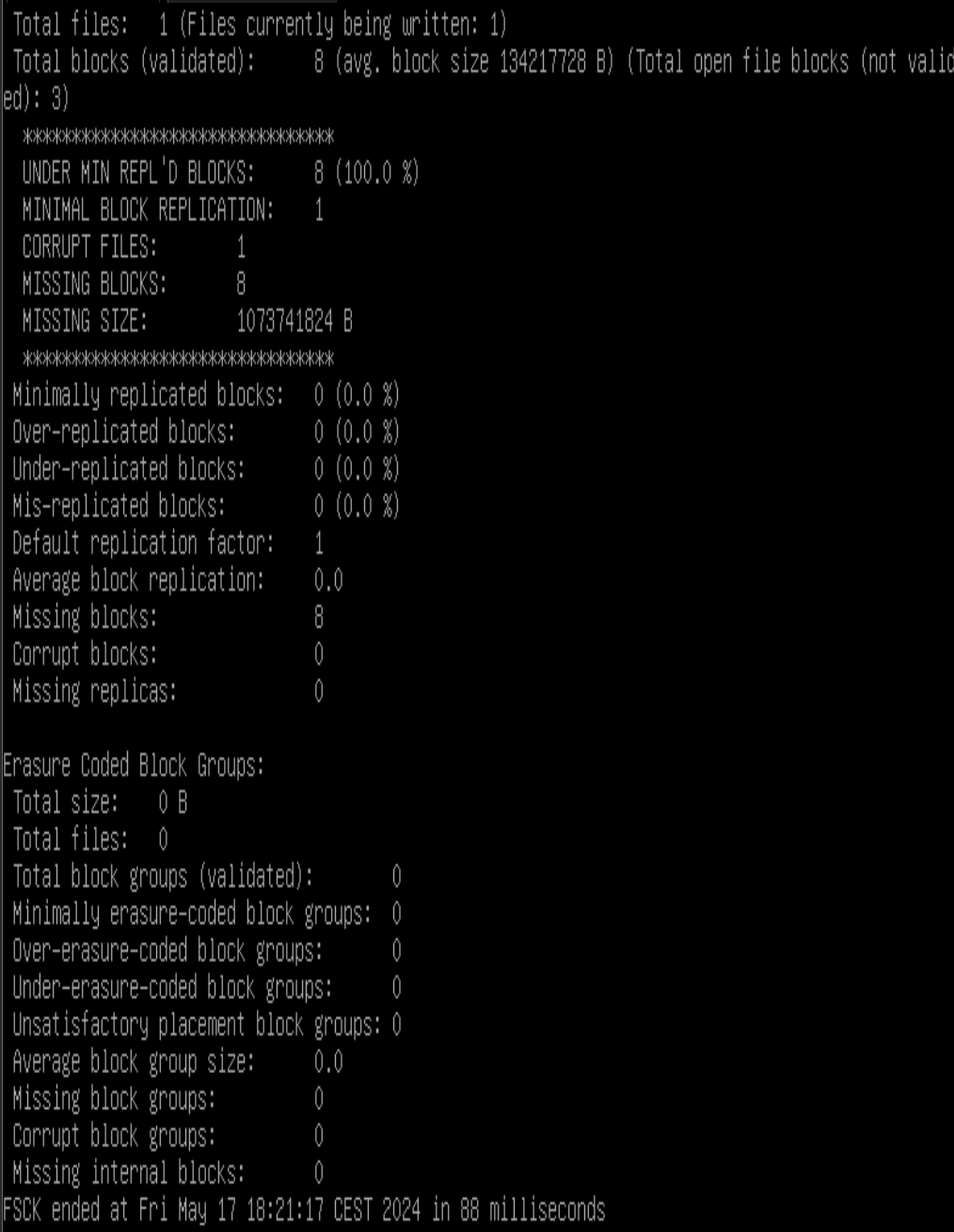
Le cluster Hadooop contient un node master et de multiples nodes esclaves. Le node master permet de gérer les nodes esclaves et inclut Job Tracker, Task Tracker, NameNode, DataNode et un node esclave posséde DataNode et TaskTracker.

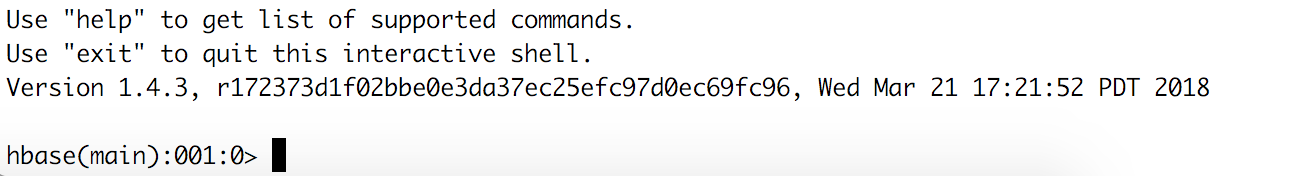
* NameNode : un service master unique qui exist dans le cluster HDFS.
* DataNode : contient de multiples blocs de données.
* JobTracker : accepte les MapReduce jobs, traite des données avec NameNode, réceptionne des métadonnées.
* TaskTracker : un node esclave de JobTracker
* décrire l'architecture de Hive
* Hive Clients : multilingue, prend en charge des applications et inclut Thrift Client, JDBC client, ODBC client.
* Hive Services : Beeline (un shell, permet d'exécuter des commandes), Hive Server 2 (permet d'exécuter des commandes vers Hive), Hive Driver (réceptionne des instructions HiveQL envoyées par utilisateur via le shell), Hive Compiler (parce la requête)
* Processing and Resource Management
* Distributed Storage
* décrire l'architecture de Hbase
* HMaster - une implémentation de Master Server dans HBase. Le processus dans lequel les régions sont attribuées à un serveur régional, ainsi que les opérations DDL. Il suit l’état des instances du Region Server dans le cluster.
* Region Server - les tables HBase sont séparées de manière horizontale, par une clé de ligne en Regions. Regions sont les éléments de base d’un cluster HBase qui contient la distribution des tables et des familles de colonnes. Il est exécuté sur un DataNode du HDFS.
* Zookeeper - est un coordinateur dans la HBase, il s' occupe des informations de configuration, de nommage, de synchronisation, d'erreurs etc.

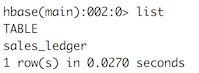
Exercice 2  
3)  
Exercice 3

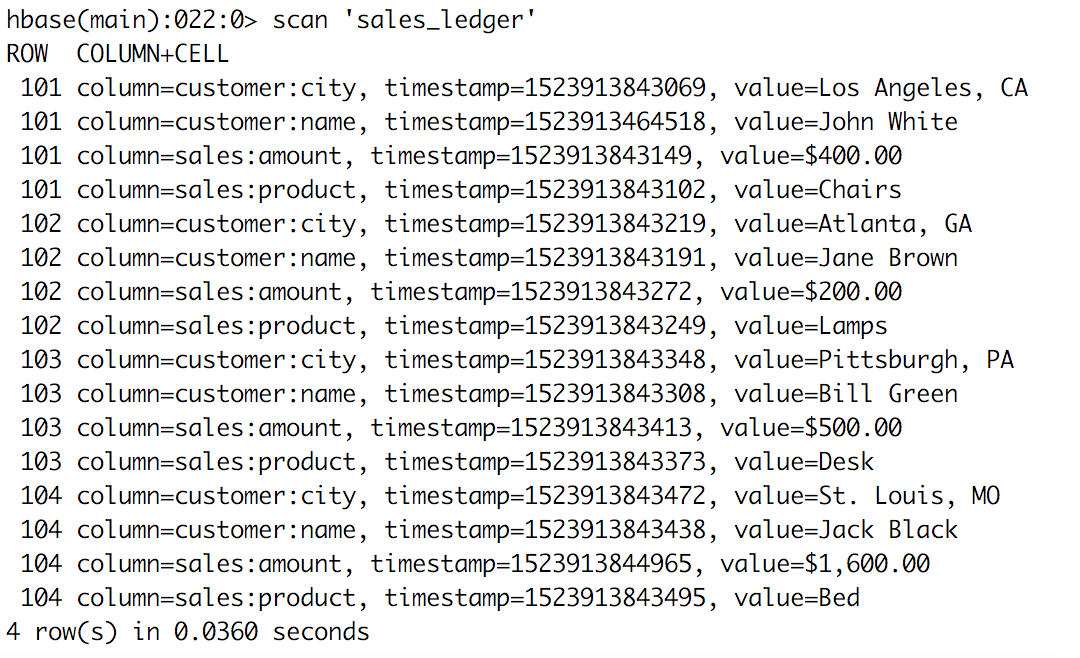
1)

2)

3)

Exercice 4   
  
1)   


  
2)

3)  


4) 