Ahmed GUIZANI
- 5A- SWE

III. Architecture Map Reduce

1-Expliquez les propriété du challenge de « Big Data » concernant les données et la problématique du ce Projet

le principal enjeu et défi de ce sujet réside dans la difficulté à traiter correctement les données, non seulement à cause de leur voluminosité mais aussi à cause de leur mauvaise qualité et donc c'est compliqué de faire un prétraitement et une normalisation appropriés.
Les moyens nécessaires pour implémenter les différents processus et algorithmes sont compliquées.

-----> nous devons traiter de grands flux de données en direct et fournir des résultats en temps quasi réel

▪ Nécessite de grands clusters pour gérer les charges de travail
▪ Nécessite des latences de quelques secondes

----->MapReduce a grandement simplifié l'analyse des données volumineuses sur des clusters volumineux et peu fiables.
Mais dès qu'il est devenu populaire, les utilisateurs en ont voulu plus :
Tâches itératives, par exemple, algorithmes d'apprentissage automatique Analyses interactives

--->Les requêtes itératives et interactives ont besoin d'une chose qui manque à MapReduce (primitives efficaces pour le partage de données)

Dans MapReduce, le seul moyen de partager des données entre les étapes de traitement est un stockage stable (disque)
La réplication ralentit également le système, mais elle est nécessaire pour la tolérance aux pannes.

2-Présentez et expliquez l'avantage d'utiliser Spark par rapport à aux techniques classiques dans ce contexte.

Nous avons du big data avec des traitements exigeants en termes de ressources et de calculs. Le principal avantage de l'utilisation de Spark est:
-sa vitesse car il peut exécuter des programmes 100 fois plus rapidement que Hadoop MapReduce en mémoire et 10 fois plus rapidement sur disque.
-Le moteur d'exécution DAG avancé prend en charge le flux de données acyclique et le calcul en mémoire, Il est également facile à utiliser et permet le développement d'applications en Java,

Scala, Python et R.

-Son modèle de programmation est plus simple que celui de Hadoop. Avec plus de 80 opérateurs de haut niveau, le logiciel facilite le développement d'applications en parallèle.

-Quantité de bibliothèques d'algorithmes MLib pour l'apprentissage automatique.

-Ces bibliothèques peuvent être facilement combinées dans la même application.

- son immense communauté.

-Un grand nombre d'entreprises utilisent Apache Spark pour traiter de grands ensembles de données.

-En tant que plate-forme open source, Apache Spark est développé par un grand nombre de développeurs de plus de 200 entreprises. Depuis 2009, plus de 1 000 développeurs ont contribué au projet.

-Cela fournit plusieurs didacticiels Spark, des réponses et des corrections potentielles en cas de problème ou autre.

-Avec sa vitesse de traitement des données, sa capacité à standardiser de nombreux types de bases de données et à exécuter différentes applications d'analyse, il peut aider à standardiser toutes vos applications Spark Big Data.

-----> Principales caractéristiques du streaming Spark

▪Scalable aux grands clusters

▪Latences de seconde échelle

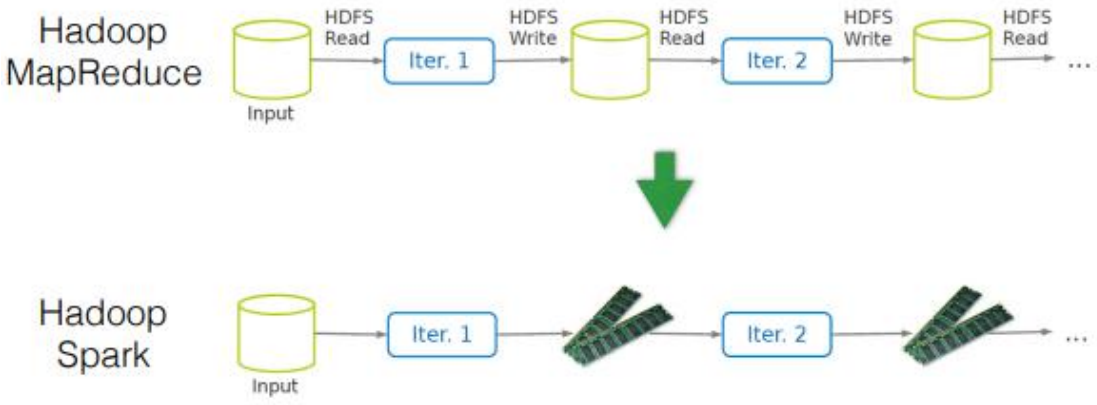
▪Modèle de programmation simple

▪ Intégré au traitement par lots et interactif

▪Tolérance aux pannes efficace dans les calculs avec état

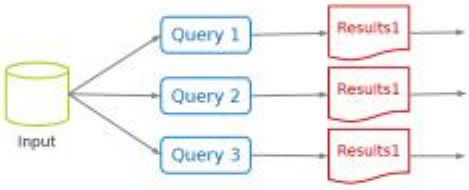
---->Solution

Traitement et partage des données en mémoire



Sharing

Hadoop MapReduce



Hadoop Spark