
Atelier : Comparaison entre Hadoop MapReduce et Spark (PySpark) sur un cas réel

Objectif de l'atelier

Cet atelier a pour but de **comparer les environnements Hadoop MapReduce et Spark (PySpark)** à travers un **cas d'usage réel déjà fonctionnel** sous Google Colab : un programme de **comptage de mots à partir d'un fichier de logs**.

L'objectif pédagogique est de :

- Comprendre les différences entre Hadoop MapReduce et Spark.
- Exécuter un même traitement sur deux environnements distribués.
- Explorer le traitement de données en local via Docker (Hadoop/Spark) et dans le cloud via Google Colab (PySpark).
- Monter en compétence sur les outils Big Data.


Déroulement de l'atelier

◆ **Partie 1 : Programme de référence (existant)**

- Un programme PySpark de **comptage de mots** est déjà fourni et **fonctionne sur Google Colab**.
- Un fichier de logs est également fourni directement dans Colab.
- Les participants commenceront par **analyser ce programme** et comprendre son fonctionnement.

◆ **Partie 2 : Exécution locale avec Docker**

- Mise en place d'un environnement local avec :
 - **Docker (via WSL pour Windows, sans Docker Desktop)**.
 - Conteneurs Hadoop et Spark standalone.
- Exécution du **même programme de comptage de mots**, mais cette fois :
 - Sur Hadoop avec un job MapReduce.
 - Sur un environnement Spark local avec PySpark.
- Comparaison des résultats, du temps de traitement et des logs d'exécution.

 Pour Windows, l'usage de **WSL (Ubuntu)** est recommandé afin d'éviter Docker Desktop qui consomme plus de ressources.

◆ **Partie 3 : Extension du programme – Analyse de logs**

- Reprise du fichier de logs utilisé dans le WordCount.
- Extension du traitement pour en extraire d'autres types d'informations :
 - Nombre de requêtes par adresse IP.
 - Répartition des codes de réponse HTTP.
 - Fréquence d'accès à certaines URLs.
- Exécution :
 - **Sur Google Colab avec PySpark.**
 - **En local via Docker avec Spark.**

✓ Objectif : utiliser la puissance des transformations RDD/DataFrame pour des cas concrets d'analyse de logs.

Livrables attendus

Chaque participant devra produire :

1. Une **analyse comparative** entre Hadoop MapReduce et Spark (temps, simplicité, performance).
 2. Une **exécution fonctionnelle du programme WordCount** :
 - Sous Google Colab (fourni).
 - En local sous Hadoop et Spark (à réaliser).
 3. Une **analyse enrichie des logs**, extraite via PySpark.
 4. Un mini rapport synthétique (readme.ms) ou une présentation.
-