Résumé des Commandes - TP3 : Programmation avec MapReduce

Démarrage du cluster

- docker start hadoop-master hadoop-slave1 hadoop-slave2
- docker exec -it hadoop-master bash
- ./start-hadoop.sh
- jps

WordCount en Java (MapReduce)

- mvn clean package
- cp target/hadoop-hdfs-WordCount.jar ~/documents/BigData/hadoop_project/
- hadoop jar /shared_volume/hadoop-hdfs-WordCount.jar /user/root/web_input/alice.txt /user/root/output_wordcount

Vérification des résultats

- hdfs dfs -ls /user/root/output_wordcount
- hdfs dfs -cat /user/root/output_wordcount/part-r-00000 | head -20
- hdfs dfs -cat /user/root/output_wordcount/part-r-00000 | sort -t$'\t' -k2 -nr | head -10
- hdfs dfs -cat /user/root/output_wordcount/part-r-00000 | wc -l

WordCount en Python (Hadoop Streaming)

- cat alice.txt | python mapper.py
- cat alice.txt | python mapper.py | sort -k1,1 | python reducer.py
- find / -name 'hadoop-streaming*.jar'
- hadoop jar /usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-3.2.0.jar \
- -files /shared_volume/hadoop_python_TP3_E2/mapper.py,/shared_volume/hadoop_python_TP3_E2/reducer.py \
- -mapper "python3 mapper.py" \
- -reducer "python3 reducer.py" \
- -input /user/root/web_input/alice.txt \
- -output /user/root/output_python

Vérification des résultats Python

- hdfs dfs -ls /user/root/output_python
- hdfs dfs -cat /user/root/output_python/part-00000 | head -20