Projet Big Data - Partie 2

Groupe 35 : Abderrazzak Mahraye - Noemane Chahid - Sophie Marimbordes

Nous avons traité la partie infrastructure du projet en utilisant l'option Docker.

Le but du projet est de compter le nombre de mots d'un gros fichier texte (512 Mo) en utilisant la technique MapReduce. Le framerwork Spark est utilisé pour distribuer la tache sur 1 à 4 slaves qui sont des containers Docker. L'opération est coordonnée par un master, également un container Docker. A chaque container est associé un cœur de la machine ce qui permet de paralléliser l'exécution de la tâche.

Nous observons effectivement une diminution du temps d'exécution de l'application WordCount lorsque le nombre de slaves augmente (cf. page 2).

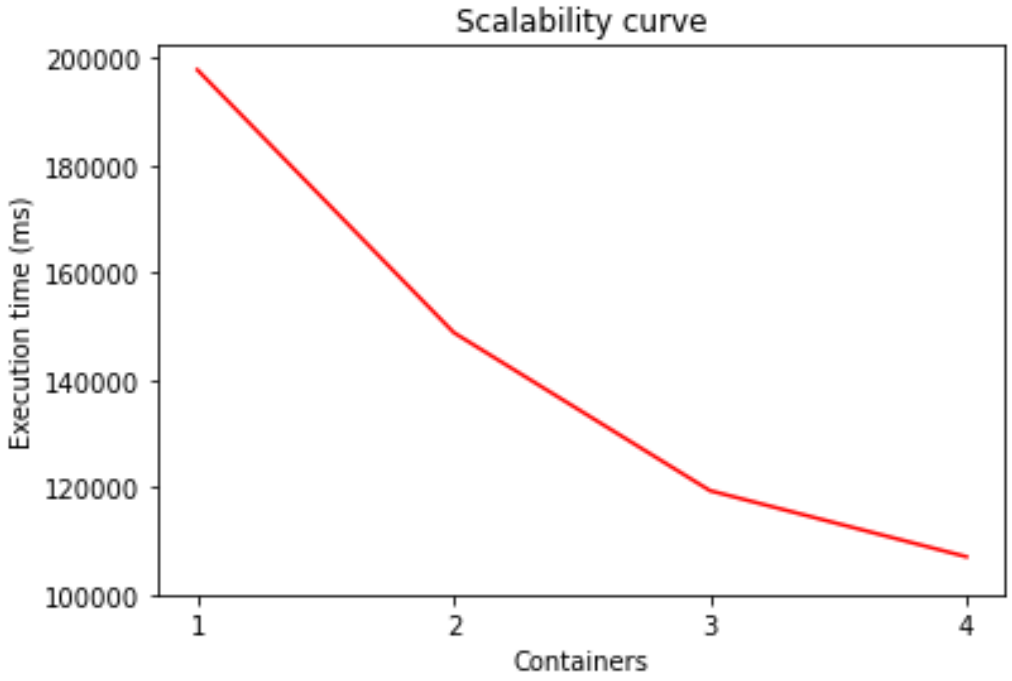
1 seul container ➔ 197 855 ms soit 3.3 min

2 containers ➔ 148 856 ms soit 2.5 min

3 containers ➔ 119 362 ms soit 2 min

4 containers ➔ 107 045 ms soit 1.8 min

Courbe de scalabilité :



En exécutant le programme Count.java (exécution séquentielle), nous avons un temps d'exécution plus petit (11 390 ms soit environ 11 s) par rapport au MapReduce.

Si nous avions utilisé véritablement une architecture répartie, nous aurions obtenu des temps d'exécution plus petits avec l'utilisation des slaves. Dans notre cas, les containers (slaves) sont simulés sur notre PC via Docker, nous n'utilisons pas de vraies machines. Chaque slave ne possède pas des performances équivalentes à un node AWS.

Spark 2.4.3 Spark Master at spark://master:7077

URL: spark://master:7077
Alive Workers: 4
Cores in use: 4 Total, 0 Used
Memory in use: 27.0 GB Total, 0.0 B Used
Applications: 0 Running, 1 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

Workers (4)

Worker Id	Address	State	Cores	Memory
worker-20210117230241-172.18.0.3-41545	172.18.0.3:41545	ALIVE	1 (0 Used)	6.7 GB (0.0 B Used)
worker-20210117230241-172.18.0.4-39893	172.18.0.4:39893	ALIVE	1 (0 Used)	6.7 GB (0.0 B Used)
worker-20210117230241-172.18.0.5-39385	172.18.0.5:39385	ALIVE	1 (0 Used)	6.7 GB (0.0 B Used)
worker-20210117230241-172.18.0.6-37259	172.18.0.6:37259	ALIVE	1 (0 Used)	6.7 GB (0.0 B Used)

Running Applications (0)

Application ID	Name	Cores	Memory per Executor	Submitted Time	User	State	Duration

Completed Applications (1)

Application ID	Name	Cores	Memory per Executor	Submitted Time	User	State	Duration
app-20210117150313-0000	WordCount	4	1024.0 MB	2021/01/17 15:03:13	root	FINISHED	1.8 min

Spark 2.4.3 **Spark Master at spark://master:7077**

URL: spark://master:7077
Alive Workers: 3
Cores in use: 3 Total, 0 Used
Memory in use: 20.2 GB Total, 0.0 B Used
Applications: 0 Running, 1 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

Workers (3)

Worker Id	Address	State	Cores	Memory
worker-20210117225341-172.18.0.3-44705	172.18.0.3:44705	ALIVE	1 (0 Used)	6.7 GB (0.0 B Used)
worker-20210117225341-172.18.0.4-44507	172.18.0.4:44507	ALIVE	1 (0 Used)	6.7 GB (0.0 B Used)
worker-20210117225341-172.18.0.5-42147	172.18.0.5:42147	ALIVE	1 (0 Used)	6.7 GB (0.0 B Used)

Running Applications (0)

Application ID	Name	Cores	Memory per Executor	Submitted Time	User	State	Duration

Completed Applications (1)

Application ID	Name	Cores	Memory per Executor	Submitted Time	User	State	Duration
app-20210117145502-0000	WordCount	3	1024.0 MB	2021/01/17 14:55:02	root	FINISHED	2.0 min

Spark 2.4.3 **Spark Master at spark://master:7077**

URL: spark://master:7077
Alive Workers: 2
Cores in use: 2 Total, 0 Used
Memory in use: 13.5 GB Total, 0.0 B Used
Applications: 0 Running, 1 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

Workers (2)

Worker Id	Address	State	Cores	Memory
worker-20210117222430-172.18.0.3-43973	172.18.0.3:43973	ALIVE	1 (0 Used)	6.7 GB (0.0 B Used)
worker-20210117222430-172.18.0.4-38509	172.18.0.4:38509	ALIVE	1 (0 Used)	6.7 GB (0.0 B Used)

Running Applications (0)

Application ID	Name	Cores	Memory per Executor	Submitted Time	User	State	Duration

Completed Applications (1)

Application ID	Name	Cores	Memory per Executor	Submitted Time	User	State	Duration
app-20210117142955-0000	WordCount	2	1024.0 MB	2021/01/17 14:29:55	root	FINISHED	2.5 min

Spark 2.4.3 **Spark Master at spark://master:7077**

URL: spark://master:7077
Alive Workers: 1
Cores in use: 1 Total, 0 Used
Memory in use: 6.7 GB Total, 0.0 B Used
Applications: 0 Running, 1 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

Workers (1)

Worker Id	Address	State	Cores	Memory
worker-20210117231003-172.18.0.3-45517	172.18.0.3:45517	ALIVE	1 (0 Used)	6.7 GB (0.0 B Used)

Running Applications (0)