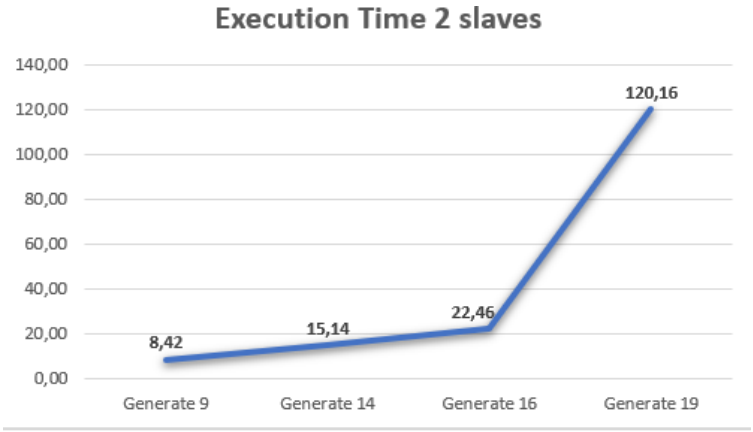
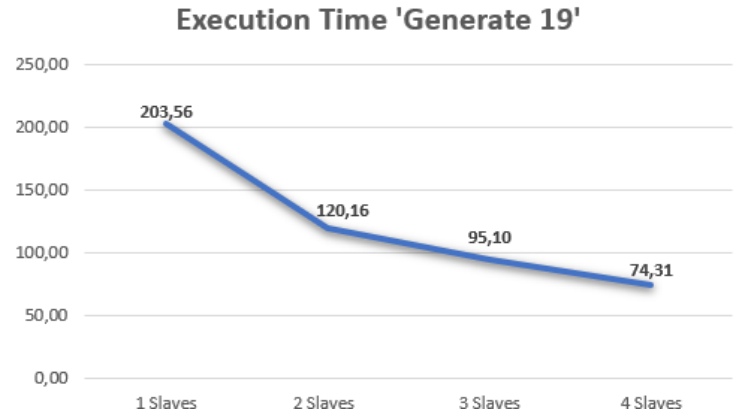
Projet Big Data
Spark Streaming scalability testing

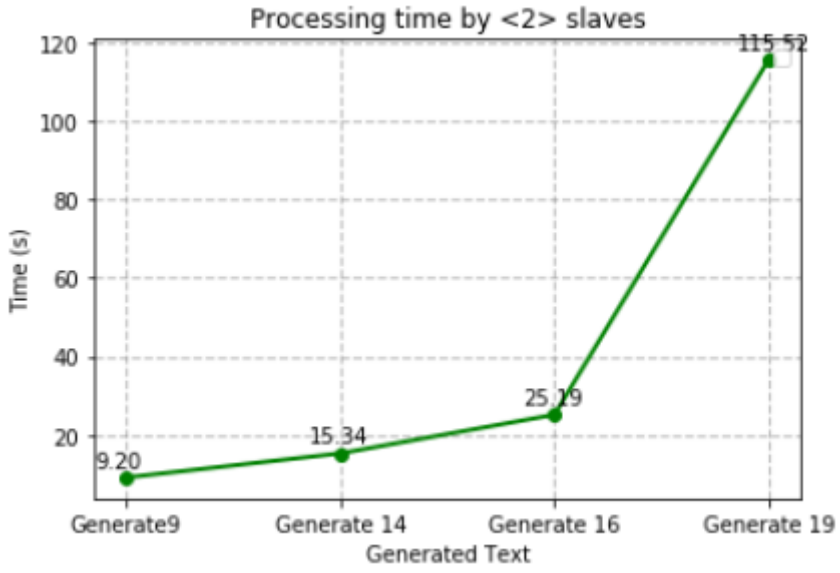
Groupe : GHARBI Aymen, HAMID Achraf, NEMRI Chouaieb

Docker Spark Word Count program performance evaluation :
Device : 16 RAM, CPU i7 10th Gen 8 Cores 1.6 GHz, SSD Hard Drive.





Same application performance on another machine :
Device : 16 RAM, CPU i7 7th Gen 8 Cores 2.7 GHz, SSD Hard Drive.



- DataNode after ‘Generate 19’ on 2 slaves :

In operation

Node	Last contact	Admin State	Capacity	Used	Non DFS Used	Remaining	Blocks	Block pool used	Failed Volumes	Version
slave2:50010 (172.18.0.4:50010)	1	In Service	462.78 GB	322.55 MB	123.9 GB	338.56 GB	20	322.55 MB (0.07%)	0	2.7.1
slave1:50010 (172.18.0.3:50010)	1	In Service	462.78 GB	193.55 MB	124.03 GB	338.56 GB	12	193.55 MB (0.04%)	0	2.7.1

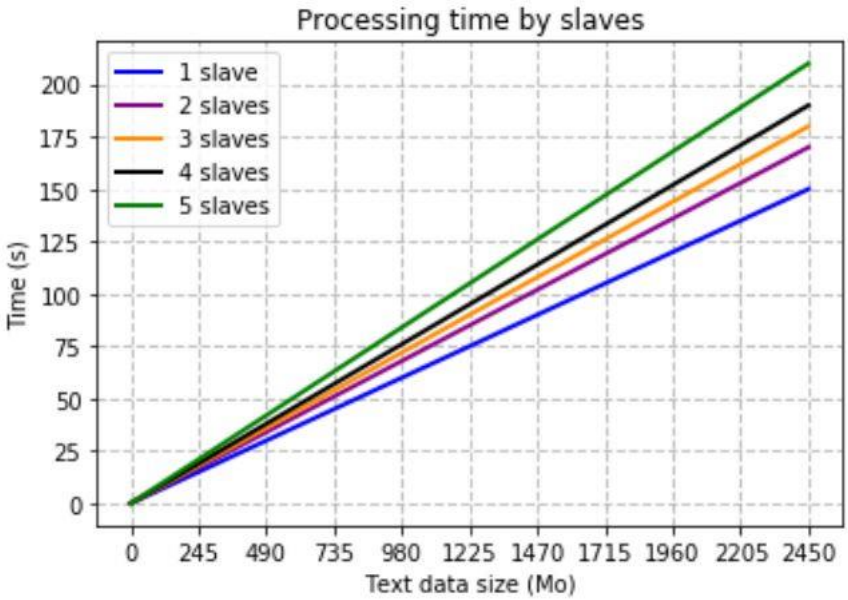
- DataNode after ‘Generate 14’ on 2 Slaves :

Node	Last contact	Admin State	Capacity	Used	Non DFS Used	Remaining	Blocks	Block pool used	Failed Volumes	Version
slave2:50010 (172.18.0.4:50010)	0	In Service	462.78 GB	16.24 MB	123.26 GB	339.5 GB	1	16.24 MB (0%)	0	2.7.1
slave1:50010 (172.18.0.3:50010)	0	In Service	462.78 GB	16.21 MB	123.26 GB	339.5 GB	1	16.21 MB (0%)	0	2.7.1

Spark streaming scalability testing:

We set up a Spark Streaming Word Count Program that connects into a python socket using TCP, the TCP server on our local machine sends batch of our generated text data which size is 245.7 Mo, the master consumes the data once and performs the word count job on the slaves using as much threads as nodes (master + workers), we made the tests for 1 to 5 slaves and the results obtained are shown on the plot below :

Device : 16 RAM, CPU i7 7th Gen 8 Cores 2.7 GHz, SSD Hard Drive.



As discussed, we remark that 1 slave performs better than multiple slaves, and that is because we are just simulating a distributed environment in the same machine, so we lose in processing time in order to perform parallelism. We will never get better results than sequential mode unless we use a real cluster.