

Challenges in installation and Running Program

The benefits of Hadoop distributed processing are obvious and do not need to be stated yet again.

For big data processing this tool is both a huge asset and a competitive necessity. But what does deserve much focus is how hard it is to install, implement and optimize an environment successfully.

The first step to avoid repeating that fate is to understand what kinds of challenges are commonly experienced.

Here are 3 challenges that I encountered in installing and running local Hadoop program,

1. Path variables getting reset after VMware Restart.

Ubuntu is installed on VMware. I observed that after machine restart path variables are getting lost in setup files. This issue is due to VMware.

2. Limited virtual memory allocation.

Hadoop on local machine assigns limited virtual memory to each container. This makes difficult to process files having size more than 250MB on local machine.

```
2020-12-14 01:44:25.971]Container killed on request. Exit code is 143
2020-12-14 01:44:25.976]Container exited with a non-zero exit code 143.

2020-12-14 01:44:36.523 INFO mapreduce.Job: Task Id : attempt_1607923678749_0002_m_000000_2, Status : FAILED
2020-12-14 01:44:35.016]Container [pid=24984,containerID=container_1607923678749_0002_01_000004] is running 322177536B beyond the 'VIRTUAL' memory limit.
Used; 2.4 GB of 2.1 GB virtual memory used. Killing container.
Dump of the process-tree for container_1607923678749_0002_01_000004 :
|- PID PPID PGRPID SESSID CMD_NAME USER MODE TIME(MILLIS) SYSTEM_TIME(MILLIS) VMEM_USAGE(BYTES) RSSMEM_USAGE(PAGES) FULL_CMD_LINE
|- 24993 24984 24984 24984 (java) 804 87 2566881280 115478 /usr/lib/jvm/java-8-openjdk-amd64/bin/java -Djava.net.preferIPv4Stack=true -Dhadoop.metrics
log.dir=/usr/local/hadoop/logs/userlogs/application_1607923678749_0002/container_1607923678749_0002_01_000004/tmp -Dlog4j.configuration=
org.apache.hadoop.mapred.YarnChild 127.0.1.1 36003 attempt_1607923678749_0002_m_000000_2 4
|- 24984 24982 24984 24984 (bash) 0 0 10153984 673 /bin/bash -c /usr/lib/jvm/java-8-openjdk-amd64/bin/java -Djava.net.preferIPv4Stack=true -Dhadoop
r=/usr/local/hadoop/htemp/nm-local-dir/usercache/hadoop/appcache/application_1607923678749_0002/container_1607923678749_0002_01_000004/tmp -Dlog4j.configu
iner.log.dir=/usr/local/hadoop/logs/userlogs/application_1607923678749_0002/container_1607923678749_0002_01_000004 -Dyarn.app.container.log.filesize=0 -Dh
syslog org.apache.hadoop.mapred.YarnChild 127.0.1.1 36003 attempt_1607923678749_0002_m_000000_2 4 1>/usr/local/hadoop/logs/userlogs/application_16079236787
out 2>/usr/local/hadoop/logs/userlogs/application_1607923678749_0002/container_1607923678749_0002_01_000004/stderr

2020-12-14 01:44:35.029]Container killed on request. Exit code is 143
2020-12-14 01:44:35.037]Container exited with a non-zero exit code 143.

2020-12-14 01:44:46.589 INFO mapreduce.Job: map 100% reduce 100%
2020-12-14 01:44:47.608 INFO mapreduce.Job: Job job_1607923678749_0002 failed with state FAILED due to: Task failed task_1607923678749_0002_m_000000
Job failed as tasks failed. failedAttempt=1 failedReducers=0 killedAttempt=0 killedReducers=0
```

3. Restart issue with Name node and data node.

Datanode and namenode process stops running after machine restart. I tried re-configuring it with clearing htemp folder and formatting namenode. However, this issue still persists, and final solution is to re-install Hadoop on machine