

Instalacja hadoop

Środowisko

Przygotowaliśmy vagrantfile konfigurujący środowisko:

```
# -*- mode: ruby -*-
# vi: set ft=ruby :

# All Vagrant configuration is done below. The "2" in Vagrant.configure
# configures the configuration version (we support older styles for
# backwards compatibility). Please don't change it unless you know what
# you're doing.
Vagrant.configure("2") do |config|
  config.vm.network "private_network", ip:"192.168.56.10"

  config.vm.provider "virtualbox" do |v|
    v.memory = 4096
    v.cpus = 4
  end

  config.vm.network "forwarded_port", guest: 8088, host: 8088
  config.vm.network "forwarded_port", guest: 9870, host: 9870
  config.vm.box = "bento/ubuntu-20.04"
  config.vm.provision "shell", path: "init.sh"
end
```

Init.sh:

```
apt update
s
curl -fsSL https://get.docker.com -o get-docker.sh
sh get-docker.sh

sudo curl -L "https://github.com/docker/compose/releases/download/1.29.2/docker-
compose-$(uname -s)-$(uname -m)" -o /usr/local/bin/docker-compose
sudo chmod +x /usr/local/bin/docker-compose

sudo groupadd docker
sudo usermod -aG docker vagrant
newgrp docker

sudo apt install -y python3
sudo apt install -y python3-pip
pip3 install docker
```

Przygotowaliśmy skrypt tworzący obrazy dockerowe, który należy ręcznie uruchomić za pierwszym razem kiedy maszyna jest utworzona:

```
./compose-up.sh 3.3.0 3 /tmp/hadoop /tmp/hadoop_logs /tmp/hbase_logs  
/tmp/hive_logs /tmp/sqoop_logs mariadb /Users/Shared/workspace/docker-ws/maria-  
data
```

Uruchomienie hadoop

Hadoopa uruchamiamy zgodnie z instrukcją skryptem `./hadoop_start`

The screenshot shows a terminal window on the left and a web browser on the right. The terminal displays the output of the `./hadoop_start` script, showing the initialization of the Hadoop cluster. The web browser shows the Hadoop Overview page, indicating that the cluster is active and providing details about the master node and the data nodes.

Terminal Output:

```
2023-04-17 22:43:03,527 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.num.users = 10  
2023-04-17 22:43:03,528 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.windows.minutes = 1,5,25  
2023-04-17 22:43:03,545 INFO namenode.FSNamesystem: Retry cache on namenode is enabled  
2023-04-17 22:43:03,546 INFO namenode.FSNamesystem: Retry cache will use 0.03 of total heap and retry cache entry expiry time is 600000 millis  
2023-04-17 22:43:03,561 INFO util.GSet: Computing capacity for map NameNodeRetryCache  
2023-04-17 22:43:03,562 INFO util.GSet: VM type = 64-bit  
2023-04-17 22:43:03,565 INFO util.GSet: 0.029999999329447746% max memory 875 MB = 268.8 KB  
2023-04-17 22:43:03,565 INFO util.GSet: capacity = 2^15 = 32768 entries  
2023-04-17 22:43:03,671 INFO namenode.FSImage: Allocated new BlockPoolId: BP-945495881-10.1.2.3-1681738983643  
2023-04-17 22:43:03,753 INFO common.Storage: Storage directory /data/hadoop/dfs/name has been successfully formatted.  
2023-04-17 22:43:03,919 INFO namenode.FSImageFormatProtobuf: Saving image file /data/hadoop/dfs/name/current/fsimage.ckpt_000000000000000000 using no compression  
2023-04-17 22:43:04,405 INFO namenode.FSImageFormatProtobuf: Image file /data/hadoop/dfs/name/current/fsimage.ckpt_000000000000000000 of size 399 bytes saved in 0 seconds  
2023-04-17 22:43:04,486 INFO namenode.NNStorageRetentionManager: Going to retain 1 image from 1 images  
2023-04-17 22:43:04,516 INFO namenode.FSImageSaver: clean checkpoint: txid=0 when meet shutdown.  
2023-04-17 22:43:04,519 INFO namenode.NameNode: SHUTDOWN_MSG: /*****  
SHUTDOWN_MSG: Shutting down NameNode at master/10.1.2.3  
*****/  
Starting namenodes on [master]  
Last login: Sun Apr 16 00:27:58 KST 2023 on pts/0  
master: Warning: Permanently added 'master,10.1.2.3' (ECDSA) to the list of known hosts  
Starting datanodes  
Last login: Mon Apr 17 22:43:08 KST 2023 on pts/0  
slave1: Warning: Permanently added 'slave1,10.1.2.4' (ECDSA) to the list of known hosts  
slave3: Warning: Permanently added 'slave3,10.1.2.6' (ECDSA) to the list of known hosts  
master: Warning: Permanently added 'master,10.1.2.3' (ECDSA) to the list of known hosts  
slave2: Warning: Permanently added 'slave2,10.1.2.5' (ECDSA) to the list of known hosts  
Starting secondary namenodes [master]  
Last login: Mon Apr 17 22:43:11 KST 2023 on pts/0  
master: Warning: Permanently added 'master,10.1.2.3' (ECDSA) to the list of known hosts  
Starting resource manager  
Last login: Mon Apr 17 22:43:31 KST 2023 on pts/0
```

Web Browser Overview:

Overview 'master:9000' (✓active)

Started:	Mon Apr 17 15:43:24 +0200 2023
Version:	3.3.0, raa96f1871bfdb858f9bac59cf2a81ec470da649af
Compiled:	Mon Jul 06 20:44:00 +0200 2020 by brahma from branch-3.3.0
Cluster ID:	CID-71c15936-4fbd-402b-af6b-af092c9eb277
Block Pool ID:	BP-945495881-10.1.2.3-1681738983643

Summary

Security is off.
Safemode is off.
1 files and directories, 0 blocks (0 replicated blocks, 0 erasure coded block groups) = 1 total filesystem object(s).
Heap Memory used 57.46 MB of 245.5 MB Heap Memory. Max Heap Memory is 875 MB.
Non Heap Memory used 45.96 MB of 47.75 MB Committed Non Heap Memory. Max Non Heap Memory is <unbounded>.

Configured Capacity:	0 B
Configured Remote Capacity:	0 B
DFS Used:	0 B (100%)
Non DFS Used:	0 B
DFS Remaining:	0 B (0%)

Uruchomienie map-reduce na klastrze

Przygotowaliśmy projekt maven w IntelliJ. Po wygenerowaniu jara możemy umieścić go i uruchomić na klastrze za pomocą własnego skryptu w języku python:

```
#!/usr/bin/env python3  
  
import docker  
import os  
import tarfile  
import sys  
  
client = docker.from_env()  
filePath = sys.argv[1]  
params = sys.argv[2:]
```

```
def copy_to(src, dst):
    name, dst = dst.split(':')
    container = client.containers.get(name)
    src = os.path.abspath(src)
    os.chdir(os.path.dirname(src))
    srcname = os.path.basename(src)
    tar = tarfile.open(src + '.tar', mode='w')
    try:
        tar.add(srcname)
    finally:
        tar.close()
    data = open(src + '.tar', 'rb').read()
    container.put_archive(os.path.dirname(dst), data)

copy_to(filePath, "master:/home/test.jar")

container = client.containers.get('master')
res = container.exec_run("yarn jar /home/test.jar " + " ".join(params))
print(res)
```

The image shows two terminal windows. The left window displays Hadoop logs from a master node, including namenode startup, datanode registration, and secondary namenode startup. The right window shows a Python script execution that runs a Hadoop job. The script uses the `copy_to` function to upload a JAR file to the master node and then runs the job using `yarn jar`.

```
2023-04-17 22:43:03,671 INFO namenode.FSImage: Allocated new BlockPoolId: BP-945495881-10.1.2.3-1681738983643
2023-04-17 22:43:03,753 INFO common.Storage: Storage directory /data/hadoop/dfs/name has been successfully formatted.
2023-04-17 22:43:03,919 INFO namenode.FSImageFormatProtobuf: Saving image file /data/hadoop/dfs/name/current/fsimage.ckpt_00000000000000000000 using no compression
2023-04-17 22:43:04,405 INFO namenode.FSImageFormatProtobuf: Image file /data/hadoop/dfs/name/current/fsimage.ckpt_00000000000000000000 of size 399 bytes saved in 0 seconds.
2023-04-17 22:43:04,486 INFO namenode.NNStorageRetentionManager: Going to retain 1 images with txid >= 0
2023-04-17 22:43:04,516 INFO namenode.FSImage: FSImageSaver clean checkpoint: txid=0 when meet shutdown.
2023-04-17 22:43:04,519 INFO namenode.NameNode: SHUTDOWN_MSG:
/*****
Shutting down NameNode at master/10.1.2.3
*****/
Starting namenodes on [master]
Last login: Sun Apr 16 00:27:58 KST 2023 on pts/0
master: Warning: Permanently added 'master,10.1.2.3' (ECDSA) to the list of known hosts

Starting datanodes
Last login: Mon Apr 17 22:43:08 KST 2023 on pts/0
slave1: Warning: Permanently added 'slave1,10.1.2.4' (ECDSA) to the list of known hosts
slave3: Warning: Permanently added 'slave3,10.1.2.6' (ECDSA) to the list of known hosts
master: Warning: Permanently added 'master,10.1.2.3' (ECDSA) to the list of known hosts
slave2: Warning: Permanently added 'slave2,10.1.2.5' (ECDSA) to the list of known hosts

Starting secondary namenodes [master]
Last login: Mon Apr 17 22:43:11 KST 2023 on pts/0
master: Warning: Permanently added 'master,10.1.2.3' (ECDSA) to the list of known hosts

Starting resourcemanager
Last login: Mon Apr 17 22:43:31 KST 2023 on pts/0
Starting nodemanagers
Last login: Mon Apr 17 22:43:48 KST 2023 on pts/0
slave2: Warning: Permanently added 'slave2,10.1.2.5' (ECDSA) to the list of known hosts
slave1: Warning: Permanently added 'slave1,10.1.2.4' (ECDSA) to the list of known hosts
slave3: Warning: Permanently added 'slave3,10.1.2.6' (ECDSA) to the list of known hosts

master: Warning: Permanently added 'master,10.1.2.3' (ECDSA) to the list of known hosts
WARNING: Use of this script to start the MR JobHistory daemon is deprecated.
WARNING: Attempting to execute replacement "mapred --daemon start" instead.
```

```
filePath = sys.argv[1]
IndexError: list index out of range
vagrant@vagrant:~$ python3 ./utils/run_jar.py ./examples/test.jar pi 2 5
ExecResult(exit_code=0, output=b'Number of Maps = 2\nSamples per Map = 5\nWrote input for Map #0\nWrote input for Map #1\nStarting Job\n2023-04-17 22:52:32,720 INFO client.DefaultHadoopFailoverProxyProvider: Connecting to ResourceManager at master/10.1.2.3:8050\n2023-04-17 22:52:33,774 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_1681739055355_0001\n2023-04-17 22:52:34,207 INFO input.FileInputFormat: Total input files to process: 2\n2023-04-17 22:52:34,479 INFO mapreduce.JobSubmitter: number of splits:2\n2023-04-17 22:52:35,019 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1681739055355_0001\n2023-04-17 22:52:35,021 INFO mapreduce.JobSubmitter: Executing with tokens: []\n2023-04-17 22:52:35,538 INFO conf.Configuration: resource-types.xml not found\n2023-04-17 22:52:35,539 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'\n2023-04-17 22:52:36,338 INFO impl.YarnClientImpl: Submitted application application_1681739055355_0001\n2023-04-17 22:52:36,556 INFO mapreduce.Job: The url to track the job: http://master:8088/xy/application_1681739055355_0001\n2023-04-17 22:52:36,559 INFO mapreduce.Job: Running job: job_1681739055355_0001\n2023-04-17 22:52:56,822 INFO mapreduce.Job: Job job_1681739055355_0001 running in uber mode : false\n2023-04-17 22:52:56,826 INFO mapreduce.Job: map 0% reduce 0%\n2023-04-17 22:53:21,801 INFO mapreduce.Job: map 100% reduce 0%\n2023-04-17 22:53:38,411 INFO mapreduce.Job: map 100% reduce 100%\n2023-04-17 22:53:40,515 INFO mapreduce.Job: Job job_1681739055355_0001 completed successfully\n2023-04-17 22:53:40,810 INFO mapreduce.Job: Counters: 54\n\nFile System Counters\nFile: Number of bytes read=50\nFile: Number of bytes written=796245\nFile: Number of read operations=0\nFile: Number of large read operations=0\nFile: Number of write operations=0\nHDFS: Number of bytes read=520\nHDFS: Number of bytes written=215\nHDFS: Number of read operations=13\nHDFS: Number of large read operations=0\nHDFS: Number of write operations=3\n\nLaunched map tasks=2\nLaunched reduce tasks=1\nRack-local map tasks=2\nTotal time spent by all maps in occupied slots (ms)=41784\nTotal time spent by all reduces in occupied slots (ms)=13672\nTotal time spent by all map tasks (ms)=41784\nTotal time spent by all reduce tasks (ms)=13672\nTotal vcore-milliseconds taken by all map tasks=13672\nTotal megabyte-milliseconds taken by all map tasks=42786816\nTotal megabyte-milliseconds taken by all reduce tasks=14000120\n\nMap-Reduce Framework\nMap input records=2\nMap output records=36\nMap output materialized bytes=56\nInput split bytes=28\nCombine input records=0\nCombine output records=0\nReduce input groups=2\nReduce input records=56\nReduce output records=4\nReduce output records=0\nSpilled Records=8\nShuffled Map s=2\nFailed Shuffles=0\nMerged Map outputs=2\nGC time elapsed (ms)=705\nCPU time spent (ms)=6030\nPhysical memory (bytes) snapshot=711092288\nVirtual memory (bytes) snapshot=7772073984\nTotal committed heap usage (bytes)=614989824\nPeak Map Physical memory (bytes)=271388672\nPeak Map Virtual memory (bytes)=2588893184\nPeak Reduce Physical memory (bytes)=169152512\nPeak Reduce Virtual memory (bytes)=2596298752\nShuffle Errors\nWRONG_LENGTH=0\nWRONG_MAP=0\nWRONG_REDUCE=0\nFile Input Format Counter Counters\nBytes Written=97\nJob Finish Pi is 3.6000000000000000\n")
```