

Урок 1

1. Разверните у себя hadoop кластер внутри docker контейнера
2. Проверьте работоспособность кластера, посмотрев на статус ресурс менеджера, нейм ноды и дата ноды
3. Остановите кластер
4. Вы пришли в компанию, в которой планируют строить Data Lake и DWH с нуля. Текущих данных около 15 Тб. Ежегодный прирост данных составляет ~500 Гб. Какую технологию вы бы предложили использовать и почему?

Задание 1

Поднимаем контейнер с кластером в интерактивном режиме

```
docker run -it --name gbhdp \ -p 50090:50090 \ -p 50075:50075 \ -p 50070:50070 \ -p 8042:8042 \ -p 8088:8088 \ -p 8888:8888 \ -p 4040:4040 \ -p 4044:4044 \ --hostname localhost \ img-hdp-hadoop
```

```
/*****
SHUTDOWN_MSG: Shutting down NameNode at localhost/127.0.0.1
*****/
Starting namenodes on [localhost]
localhost: Warning: Permanently added 'localhost' (ECDSA) to the list of known hosts.
localhost: starting namenode, logging to /home/hduser/hadoop/logs/hadoop-hduser-namenode-localhost.out
localhost: Warning: Permanently added 'localhost' (ECDSA) to the list of known hosts.
localhost: starting datanode, logging to /home/hduser/hadoop/logs/hadoop-hduser-datanode-localhost.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: Warning: Permanently added '0.0.0.0' (ECDSA) to the list of known hosts.
0.0.0.0: starting secondarynamenode, logging to /home/hduser/hadoop/logs/hadoop-hduser-secondarynamenode-localhost.out
starting yarn daemons
starting resourcemanager, logging to /home/hduser/hadoop/logs/yarn--resourcemanager-localhost.out
localhost: Warning: Permanently added 'localhost' (ECDSA) to the list of known hosts.
localhost: starting nodemanager, logging to /home/hduser/hadoop/logs/yarn-hduser-nodemanager-localhost.out
hduser@localhost:~$
```

Задание 2

Ресурс менеджер

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Used Resources	Total Resources	Reserved Resources
0	0	0	0	0	<memory:0, vCores:0>	<memory:4096, vCores:8>	<memory:0, vCores:0>

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Nodes	Shutdown Nodes
0	0	0	0	0	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation	Maximum Cluster Application Priority
Capacity Scheduler	[<name=memory-mb default-unit=Mi type=COUNTABLE>, <name=vcores default-unit=type=COUNTABLE>]	<memory:1024, vCores:1>	<memory:2048, vCores:4>	0

Showing 0 to 0 of 0 entries

Нод менеджер

NodeManager information

Property	Value
Total Vmem allocated for Containers	8.40 GB
Vmem enforcement enabled	true
Total Pmem allocated for Container	4 GB
Pmem enforcement enabled	true
Total VCoers allocated for Containers	8
NodeHealthyStatus	true
LastNodeHealthTime	Mon May 02 06:44:30 GMT 2022
NodeHealthReport	
NodeManager started on	Mon May 02 06:28:29 GMT 2022
NodeManager Version	2.10.1 from 1827467c9a56f133025f28557b6c2c562d78e16 by centos source checksum 2da9946fa5679794b77621bdc0b1a on 2020-09-14T13:24Z
Hadoop Version	2.10.1 from 1827467c9a56f133025f28557b6c2c562d78e16 by centos source checksum 3114ede668f13824e7d0f68b0c3650 on 2020-09-14T13:17Z

Нейм нода

← → ↺

localhost:50070/dfshealth.html#tab-overview

80% ☆

🔒 ⬇️ ☰

Hadoop

Overview

Datanodes

Datanode Volume Failures

Snapshot

Startup Progress

Utilities

Overview

'localhost:9000' (active)

Started:

Mon May 02 09:28:14 +0300 2022

Version:

2.10.1, r1827467c9a56f133025f28557b1c2c562d78e816

Compiled:

Mon Sep 14 16:17:00 +0300 2020 by centos from branch-2.10.1

Cluster ID:

CID-026b11aa-cfab-40b8-97f4-5a8a2d538119

Block Pool ID:

BP-1748043258-127.0.0.1-1651472891094

Summary

Security is off.

Safemode is off.

7 files and directories, 0 blocks = 7 total filesystem object(s).

Heap Memory used 81.77 MB of 150 MB Heap Memory. Max Heap Memory is 889 MB.

Non Heap Memory used 45.68 MB of 46.72 MB Committed Non Heap Memory. Max Non Heap Memory is -unbounded>.

Configured Capacity:

86.37 GB

DFS Used:

28 KB (0%)

Non DFS Used:

46.69 GB

DFS Remaining:

35.26 GB (40.82%)

Block Pool Used:

28 KB (0%)

DataNodes usages% (Min/Median/Max/stdDev):

0.00% / 0.00% / 0.00% / 0.00%

Live Nodes

1 (Decommissioned: 0, In Maintenance: 0)

Dead Nodes

0 (Decommissioned: 0, In Maintenance: 0)

Decommissioning Nodes

0

Entering Maintenance Nodes

0

Total Datanode Volume Failures

0 (0 B)

Number of Under-Replicated Blocks

0

Number of Blocks Pending Deletion

0

Block Deletion Start Time

Mon May 02 09:28:14 +0300 2022

Last Checkpoint Time

Mon May 02 09:28:11 +0300 2022

Дата нода

← → ↺

localhost:50075/datanode.html

80% ☆

🔒 ⬇️ ☰

Hadoop

Overview

Utilities

DataNode on

localhost:50010

Cluster ID:

CID-026b11aa-cfab-40b8-97f4-5a8a2d538119

Version:

2.10.1

Block Pools

namenode Address

Block Pool ID

Actor State

Last Heartbeat

Last Block Report

Last Block Report Size (Max Size)

localhost:9000

BP-1748043258-127.0.0.1-1651472891094

RUNNING

1s

8 minutes

0 B (64 MB)

Volume Information

Directory

StorageType

Capacity Used

Capacity Left

Capacity Reserved

Reserved Space for Replicas

Blocks

/tmp/hadoop-huser/dfs/data/current

DISK

24 KB

35.26 GB

0 B

0 B

0

Hadoop, 2020.

Задание 3

Выключаем кластер командой **exit**

```
hduser@localhost:~$ exit
exit
Stopping namenodes on [localhost]
localhost: Warning: Permanently added 'localhost' (ECDSA) to the list of known hosts.
localhost: stopping namenode
localhost: Warning: Permanently added 'localhost' (ECDSA) to the list of known hosts.
localhost: stopping datanode
Stopping secondary namenodes [0.0.0.0]
0.0.0.0: Warning: Permanently added '0.0.0.0' (ECDSA) to the list of known hosts.
0.0.0.0: stopping secondarynamenode
stopping yarn daemons
stopping resourcemanager
localhost: Warning: Permanently added 'localhost' (ECDSA) to the list of known hosts.
localhost: stopping nodemanager
localhost: nodemanager did not stop gracefully after 5 seconds: killing with kill -9
no proxyserver to stop
Bye, bye!
```

Задание 4

В целом это зависит от самих данных. Если они реалиционные, то при таком приросте и изначальном объеме данных можно обойтись RDBMS. Если же данные малоструктурированы, то можно задуматься об использовании hadoop, хотя опять же при таком небольшом объеме и приросте данных это не очень целесообразно