CLOUDERA

Cloudera Professional Services

Sabic Cluster Health Check

Engagement Report

Zuling Kang, Senior Machine Learning Solutions Architect

Extract 23-DEC-2020

Revision History

Version	Author	Description	Date
0.1		Document creation	15/12/2020

TABLE OF CONTENTS

Executive Summary 6

Hardware and OS Environment Review 7

- Server rack layout 7
- OS and kernel configuration for QA 7
- OS and kernel configuration for PROD 9
- File system layout and storage configurations for QA 10
- File system layout and storage configurations for PROD 12
- Network configurations for QA 13
- Network configurations for PROD 15

Cluster Overview 18

- Nodes and components 18
- Resource allocation 18
- Cluster resource utilization 20
  - QA Cluster 20
  - PROD Cluster 21
- Log file locations 23
  - QA Cluster 23
  - PROD Cluster 24

QA Cluster Component Review 25

- Role distribution and HA configuration for QA 25
- HDFS review 26
  - Major configurations and deployment settings 26
  - Namenode GC, process pause and swap 28
  - Namenode resource usage 29
  - R/W performance 30
- YARN and container allocation 31
  - Major configurations and deployment settings 31
  - Overall load and health of YARN and MapReduce 33
- ZOOKEEPER review 34
  - Major configurations and deployment settings 34
  - Major Zookeeper health metrics 34
- HIVE review 37
  - Major configurations and deployment settings 37

Table and partition counts 37
IMPALA review 40
Major configurations and deployment settings 40
Workload and success rate 41
Locality of assignment 43
Memory usage of IMPALA-DAEMON 43
SPARK review 44
SOLR review 45
Major configurations and deployment settings 45
Major Solr health metrics 46
PROD Cluster Component Review 48
Role distribution and HA configuration for PROD 48
HDFS review 49
Major configurations and deployment settings 49
Namenode GC, process pause and swap 51
Namenode resource usage 53
R/W performance 54
YARN and container allocation 55
Major configurations and deployment settings 55
Overall load and health of YARN and MapReduce 57
ZOOKEEPER review 60
Major configurations and deployment settings 60
Major Zookeeper health metrics 61
HIVE review 62
Major configurations and deployment settings 62
Table and partition counts 63
IMPALA review 65
Major configurations and deployment settings 65
Workload and success rate 66
Locality of assignment 68
Memory usage of IMPALA-DAEMON 68
SPARK review 69
SOLR review (not in use?) 70
Major configurations and deployment settings 70
Major Solr health metrics 70
Data Ingestion Flow Review 73

QA cluster (?) 73
PROD cluster 73
Security Configuration Review for QA 74
Kerberos & LDAP 74
Kerberos enablement and /etc/krb5.conf 74
LDAP client configuration at OS and /etc/openldap/ldap.conf 74
LDAP integration in OS and /etc/sssd/sssd.conf 75
Sentry review 76
Encryption status 77
Security configuration review with major components 77
Security review with HDFS/YARN/ZOOKEEPER 77
Security review with HIVE/IMPALA 79
Security review with Solr 81
Security Configuration Review for PROD 82
Kerberos & LDAP 82
Kerberos enablement and /etc/krb5.conf 82
LDAP client configuration at OS and /etc/openldap/ldap.conf 82
LDAP integration in OS and /etc/sssd/sssd.conf 83
Sentry review 84
Encryption status 86
Security configuration review with major components 86
Security review with HDFS/YARN/ZOOKEEPER 86
Security review with HIVE/IMPALA 88
Security review with Solr 90
KTS and KMS review 90
Security Zone of KTS 90
Major configurations review 91

IMPORTANT NOTICE

Cloudera, Inc.

395 Page Mill Road
Palo Alto, CA 94306
info@cloudera.com
US: 1-888-789-1488
Intl: 1-650-362-0488
www.cloudera.com

Release Information
Version: 1.0 Date: 15/02/2019

1 Executive Summary

(To be added later.)

2 Hardware and OS Environment Review

2.1 Server rack layout

The bare metal servers (excluding VMs) are placed in the racks as follows:

	SADAF	SADAF - Rack1	SADAF - Rack2
QA Cluster	(Rack: default) CDSW Worker 1	(Rack: /rack1) Master-1 Worker-1/2/5/6/9 Edge-1/2	(Rack: /rack2) Master-2 Worker-3/4/7/8/10
Production Cluster		(Rack: /rack1) Master-1 Worker-1/2/3 Edge-1	(Rack: /rack2) Master-2/3 Worker-4/5 Edge-2

The VM servers are placed in the racks as follows:

	HADEED-VM	HQ-VM
QA Cluster	(Rack: default) CDSW Master KMS-1/2 Dataiku_Design_QA Dataiku_Automation_QA	(Rack: default) KTS-1/2 CDSW Worker 2 Dataiku_API_QA Datapred_QA
Production Cluster	(Rack: default) KTS-1/2 KMS-1/2 CDSW Worker Dataiku_API_Prod-2 DevOps_Prod-1 Datapred_Prod	(Rack: default) CDSW Master Dataiku_Automation_Prod Dataiku_API_Prod-1 DevOps_Prod-2 WebProtege_Prod

Comments:

- In summary, rack awareness is set properly in both QA and PROD clusters.
- The rack names for Worker 9 and 10 are incorrect with their physical rack placement. Please rename them from default to /rack1 and /rack2, respectively.
- If the high availability of HADEEM-VM and HQ-VM are guaranteed by the virtual machine manager, it'll be of no problem for the current KMS and KTS placement. That is, if both KTS or KMS happen to be running under the same rack, the virtual machine manager will transfer the KTS or KMS instances to elsewhere when the rack's power supply goes off.

2.2 OS and kernel configuration for QA

	ITEMS	***RESULTS***	***RECOMMENDATIONS***

1	Hostname	FIRST: /etc/hosts THEN: DNS Some concerns are discovered for DNS lookups (see the comments below)	Recommended to remove the DNS entry 172.17.0.1 for d02ycdsmas001.
2	NTP upstream servers	10.1.150.222 10.33.12.230 Two upstream servers with Chrony	
3	NTP status	Approximately in sync within ten minutes.	
4	OS kernels versions	No kernel versions that are known to be bad are running	
5	vm.swappiness	Host d02ydkuaut001 and d02ydkudes001 is set to 30	Recommended to turn to the related software vendor(s) as these 2 nodes are not used to deploy CDH components. However, if that vendor is neutral on vm.swappiness, we still recommend setting it to 1 as long as this node is used for data management or data mining.
6	Transparent huge pages	THP setting for d02ydkudes001 is not set properly	Recommended to turn to the related software vendor(s) as this node is not used to deploy CDH components. However, if that vendor is neutral on THP, we still recommend disabling it as long as this node is used for data management or data mining.
7	JDK versions	OpenJDK 11.0.3	
8	Number of opening files (nofiles)	hdfs - nofile 32768 mapred - nofile 32768 hbase - nofile 32768	Recommend to specify the nofile parameter for the following users (or all users): • yarn • solr • impala

Comments on DNS lookup:

- Some DNS lookups return multiple entries like the following (using nslookup command in d02ycdpwrk006)

```
[hadoop@d02ycdpwrk006 ~]$ nslookup d02ycdsmas001.sabic.com
Server:         10.32.43.25
Address:        10.32.43.25#53

Name:   d02ycdsmas001.sabic.com
Address: 10.32.180.14
Name:   d02ycdsmas001.sabic.com
Address: 172.17.0.1
Name:   d02ycdsmas001.sabic.com
Address: 100.66.0.2
```

It can be seen that the entry 10.32.180.14 is the host's real IP address. Meanwhile, as d02ycdsmas001 is the CDSW master, it has a Kubernetes container's IP address, i.e. 100.66.0.2. While 172.17.0.1 is the Docker's gateway address.

Generally speaking, the entry 172.17.0.1 could be safely removed.

2.3 OS and kernel configuration for PROD

	ITEMS	***RESULTS***	***RECOMMENDATIONS***
1	Hostname	FIRST: /etc/hosts THEN: DNS Some concerns are discovered for DNS lookups (see the comments below)	Recommended to remove the DNS entry 172.17.0.1 for d02ycdsmas001.
2	NTP upstream servers	10.1.150.222 10.33.12.230 Two upstream servers with Chrony	
3	NTP status	Approximately in sync within ten minutes.	
4	OS kernels versions	No kernel versions that are known to be bad are running	
5	vm.swappiness	Host d02pdkuaut001 is set to 30	Recommended to turn to the related software vendor(s) as this node is not used to deploy CDH components. However, if that vendor is neutral on vm.swappiness, we still recommend setting it to 1 as long as this node is used for data management or data mining.
6	Transparent huge pages	THP setting for d02pdkuaut001 is not set properly	Recommended to turn to the related software vendor(s) as this node is not used to deploy CDH

			components. However, if that vendor is neutral on THP, we still recommend disabling it as long as this node is used for data management or data mining.
7	JDK versions	OpenJDK 1.8.0_232 for d02pdkuaut001 while OpenJDK 11.0.3 for the others	Recommended to turn to the related software vendor(s) as d02pdkuaut001 is not used to deploy CDH components.
8	Number of opening files (nofiles)	hdfs - nofile 32768 mapred - nofile 32768 hbase - nofile 32768	Recommend to specify the nofile parameter for the following users (or all users): • yarn • solr • impala

Comments on DNS lookup:

- Some DNS lookups return multiple entries like the following (using nslookup command in d02ycdpwrk006)

```
[hadoop@d02pcdpedg001 ~]$ nslookup d02pcdsmas001.sabic.com
Server:         10.32.43.25
Address:        10.32.43.25#53

Name:   d02pcdsmas001.sabic.com
Address: 10.32.181.18
Name:   d02pcdsmas001.sabic.com
Address: 172.17.0.1
Name:   d02pcdsmas001.sabic.com
Address: 100.66.0.2
```

  It can be seen that the entry 10.32.181.18 is the host's real IP address. Meanwhile, as d02ycdsmas001 is the CDSW master, it has a Kubernetes container's IP address, i.e. 100.66.0.2. While 172.17.0.1 is the Docker's gateway address.

  Generally speaking, the entry 172.17.0.1 could be safely removed.

- It is reported in CM Host Inspector about the following error message:

```
DNS reverse lookup of IP 10.32.181.12 on host d02pcdpedg001.sabic.com failed.
Expected d02pcdpmas003.sabic.com but got jusdrhhdmp3r2.sabic.com.
```

  However, as double checked from d02pcdpedg001, when issuing 'nslookup 10.32.181.12', the returning entry is 'd02pcdpmas003.sabic.com'. We can thus ignore the error message from CM Host Inspector.

2.4 File system layout and storage configurations for QA

- Local disks are placed as JBOD.

- The mount configuration master and worker nodes are shown below. To speed up the execution of dump and fsck, the 5th and 6th columns of data disks can be considered to be configured as 0.

```
[hadoop@d02ycdpmas001 ~]$ cat /etc/fstab

#
# /etc/fstab
# Created by anaconda on Tue Mar  6 12:11:59 2018
#
# Accessible filesystems, by reference, are maintained under '/dev/disk'
# See man pages fstab(5), findfs(8), mount(8) and/or blkid(8) for more info
#
LABEL=root                                /       ext4    defaults        1 1
LABEL=boot                                /boot           ext4    defaults        1 2
UUID="EE85-F07C"                          /boot/efi       vfat    umask=0077,shortname=winnt 0 0
LABEL=home                                /home           ext4    defaults,nodev        1 2
LABEL=tmp                                 /tmp            ext4    defaults        1 2
LABEL=var                                 /var            ext4    defaults        1 2
LABEL=vartmp                              /var/tmp        ext4    defaults,nodev,noexec,nosuid        1 2
LABEL=swap                                swap            swap    defaults        0 0
tmpfs                   /dev/shm          tmpfs   defaults,nodev,nosuid,noexec        0 0
#
#
#UUID=d984516b-4da1-47a6-bfa8-e9367e0187ee       /data/db        ext4    defaults,noatime        1 2
UUID=d984516b-4da1-47a6-bfa8-e9367e0187ee       /data/raidl_db   ext4    defaults,noatime        1 2

#UUID=7baa825f-5334-493d-ae43-9635442577bc       /data/hdfs      ext4    defaults,noatime        1 2
UUID=7baa825f-5334-493d-ae43-9635442577bc       /data/raidl_hdfs ext4    defaults,noatime        1 2

UUID=ca220fee-277e-4190-86fd-0c0fb229ad58       /data/01        ext4    defaults,noatime        1 2
UUID=e821e423-e3c0-4a01-b58c-21b55d4fb95f       /data/02        ext4    defaults,noatime        1 2
UUID=ccd7e58f-4071-4823-8268-da614b914df3       /data/03        ext4    defaults,noatime        1 2
UUID=d6b46496-c826-4934-826d-4af74b27a9af       /data/04        ext4    defaults,noatime        1 2
UUID=3b177b24-68f9-4fc9-b1c2-c61c4d3d95ae       /data/05        ext4    defaults,noatime        1 2
UUID=1e577a38-d2a8-4891-a426-96cc9d789cce       /data/06        ext4    defaults,noatime        1 2
```

```
[hadoop@d02ycdpwrk002 ~]$ cat /etc/fstab

#
# /etc/fstab
# Created by anaconda on Thu Mar 15 10:06:45 2018
#
# Accessible filesystems, by reference, are maintained under '/dev/disk'
# See man pages fstab(5), findfs(8), mount(8) and/or blkid(8) for more info
#
LABEL=/                                 /               ext4    defaults        1 1
LABEL=/boot                             /boot           ext4    defaults        1 2
UUID="CA3E-5405"                        /boot/efi       vfat    umask=0077,shortname=winnt 0 0
LABEL=/home                             /home           ext4    defaults,nodev        1 2
LABEL=/tmp                              /tmp            ext4    defaults        1 2
LABEL=/var                              /var            ext4    defaults        1 2
#LABEL=/var/tmp                                 /var/tmp        ext4    defaults,nodev,noexec,nosuid        1 2
LABEL=/var/tmp                                /var/tmp        ext4    defaults,nodev        1 2
LABEL=swap                              swap            swap    defaults        0 0
tmpfs                   /dev/shm        tmpfs   defaults,nodev,nosuid,noexec        0 0
#
#
UUID=95e03f82-2344-461a-9996-976c0ba28aa0       /data/01        ext4    defaults,noatime        1 2
#UUID=b8d7f1a2-9ece-4d18-9f01-0019113b893f       /data/02        ext4    defaults,noatime        1 2
UUID=b0d5c225-6a05-4524-b8b9-83116e89b274       /data/03        ext4    defaults,noatime        1 2
UUID=cc2fcb2a-115d-42c3-bffe-06a5686a1f08       /data/04        ext4    defaults,noatime        1 2
UUID=fcd2b9d5-298b-423f-9721-edc8b8cd0eb8       /data/05        ext4    defaults,noatime        1 2
UUID=69388af3-3d9b-4444-b525-04df15a61dff       /data/06        ext4    defaults,noatime        1 2
UUID=394c5320-6eaf-4c64-a1d9-101085ef30c9       /data/07        ext4    defaults,noatime        1 2
UUID=87d59876-a8f1-4985-888e-83aaf6125f5f       /data/08        ext4    defaults,noatime        1 2
UUID=cad9f038-d555-4fe6-b0c5-15d6be72144a       /data/09        ext4    defaults,noatime        1 2
UUID=e634bd08-65b4-4a35-a1d8-cfbc2aa66bd9       /data/10        ext4    defaults,noatime        1 2
UUID=8d2019a5-cb6d-41b4-a732-7d8475399859       /data/11        ext4    defaults,noatime        1 2
UUID=1892745c-7f34-43f0-80c3-cb232046bcba       /data/12        ext4    defaults,noatime        1 2
UUID=7f034fbf-d935-4a4f-b9dc-ec36d53da4be       /data/13        ext4    defaults,noatime        1 2
UUID=4bea9cd7-d5c0-4389-bf58-e13b55ec4dab       /data/14        ext4    defaults,noatime        1 2
UUID=d638fd39-7b1d-4d58-8aed-eab936021711       /data/15        ext4    defaults,noatime        1 2
UUID=3a5c28ab-b1e6-4aac-a26c-139e6c763dc4       /data/16        ext4    defaults,noatime        1 2
UUID=77c9b5fd-660c-4c46-9452-2c4e0be602b9       /data/17        ext4    defaults,noatime        1 2
UUID=81b0c6e0-17fa-4d9e-9829-1db7b75ff5ff       /data/18        ext4    defaults,noatime        1 2
UUID=9df88d44-b4d1-41e8-90a0-a04ae9540fcf       /data/19        ext4    defaults,noatime        1 2
UUID=9c119662-b8df-4422-9b95-4ca86df15b54       /data/20        ext4    defaults,noatime        1 2
UUID=c2324475-16cb-4990-bec0-c6acccf244e5       /data/21        ext4    defaults,noatime        1 2
UUID=70ccd234-0aa1-4e10-8fe8-235292d27843       /data/22        ext4    defaults,noatime        1 2
UUID=722da432-0266-4cb7-8a7a-69afb5d79104       /data/02        ext4    defaults,noatime        1 2
```

2.5 File system layout and storage configurations for PROD

- Local disks are placed as JBOD.
- The mount configuration master and worker nodes are shown below. To speed up the execution of dump and fsck, the 5th and 6th columns of data disks can be considered to be configured as 0.

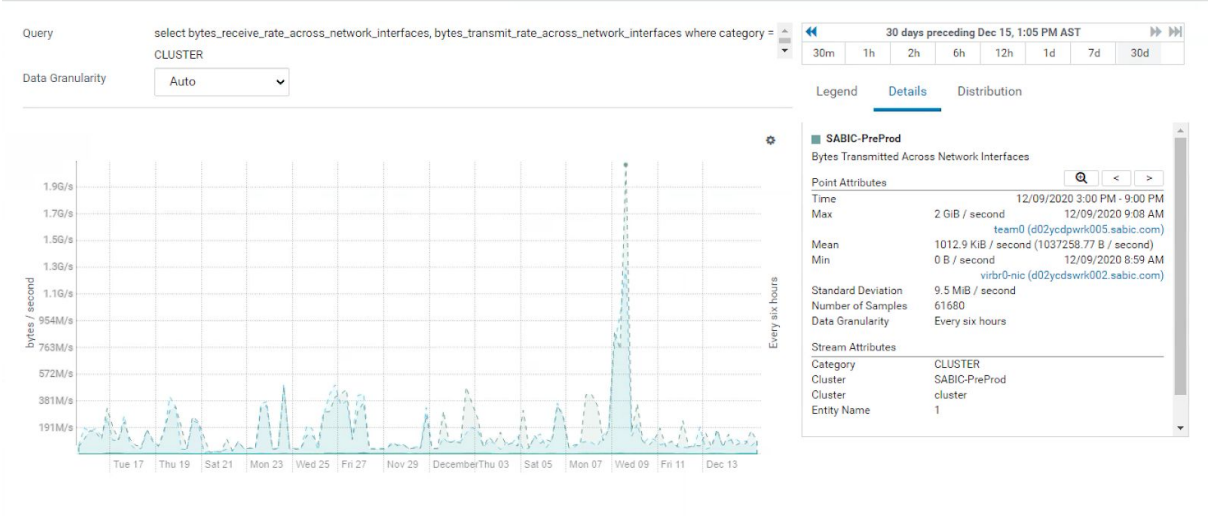
```
[hadoop@d02pcdpmas002 ~]$ cat /etc/fstab

#
# /etc/fstab
# Created by anaconda on Mon Apr  2 12:37:32 2018
#
# Accessible filesystems, by reference, are maintained under '/dev/disk'
# See man pages fstab(5), findfs(8), mount(8) and/or blkid(8) for more info
#
LABEL=root                              /                       ext4    defaults        1 1
LABEL=boot                              /boot                   ext4    defaults        1 2
UUID="1D00-399E"                        /boot/efi               vfat    umask=0077,shortname=winnt 0 0
LABEL=home                              /home                   ext4    defaults,nodev        1 2
LABEL=tmp                               /tmp                    ext4    defaults        1 2
LABEL=var                               /var                    ext4    defaults        1 2
#LABEL=vartmp                           /var/tmp                ext4    defaults,nodev,nosuid,noexec        1 2
LABEL=vartmp                           /var/tmp                ext4    defaults        1 2
LABEL=swap                              swap                    swap    defaults        0 0
tmpfs                   /dev/shm                tmpfs   defaults,nodev,nosuid,noexec        0 0
UUID=d1dfc990-4976-4f19-a814-263f5f79484e       /data/01        ext4    defaults,noatime        1 2
UUID=75c61583-39cd-4738-a04d-a3605c241f58       /data/02        ext4    defaults,noatime        1 2
UUID=81c92c34-df6d-4738-bc4c-7459716ba434       /data/03        ext4    defaults,noatime        1 2
UUID=1f2786b6-bfa9-46bc-a951-d380bf50c72d       /data/04        ext4    defaults,noatime        1 2
UUID=e751afca-1209-4937-afd4-cbd2a9b7b31c       /data/05        ext4    defaults,noatime        1 2
UUID=a057a93b-8fee-4e48-ae3c-a898b3e82fbd       /data/06        ext4    defaults,noatime        1 2
#UUID=13c31678-d5f8-4777-a80c-f864dd0674fc      /data/07        ext4    defaults,noatime        1 2
#UUID=4649b5a6-f0cc-4681-9c5e-259a526d6255      /data/08        ext4    defaults,noatime        1 2
#UUID=6073ca0a-ac70-4815-abff-35f6f3d7e43c      /data/09        ext4    defaults,noatime        1 2
#UUID=522e7d7a-c4fd-4874-af27-cff46f158670      /data/10        ext4    defaults,noatime        1 2
UUID=f87ff8ce-48e8-4be5-8e57-f27e9ddc32a3       /data/raid1_hdfs        ext4    defaults,noatime        1 2
UUID=6b56cd26-3965-4daa-a4c1-aa9d8cb667fe       /data/raid1_db  ext4    defaults,noatime        1 2
```

```
#
# /etc/fstab
# Created by anaconda on Tue Mar  6 10:10:16 2018
#
# Accessible filesystems, by reference, are maintained under '/dev/disk'
# See man pages fstab(5), findfs(8), mount(8) and/or blkid(8) for more info
LABEL=root              /                       ext4    defaults        1 1
LABEL=boot              /boot                   ext4    defaults        1 2
UUID="816C-2377"        /boot/efi               vfat    umask=0077,shortname=winnt 0 0
LABEL=home              /home                   ext4    defaults,nodev        1 2
LABEL=tmp               /tmp                    ext4    defaults        1 2
LABEL=var               /var                    ext4    defaults        1 2
LABEL=vartmp            /var/tmp                ext4    defaults,nodev,noexec,nosuid        1 2
LABEL=swap              swap                    swap    defaults        0 0
tmpfs                   /dev/shm                tmpfs   defaults,nodev,nosuid,noexec        0 0
################################################################################
#
#    HADOOP FILESYSTEMS
#
################################################################################

#/dev/sdd
UUID=4adc0287-d8fd-4343-9500-d19c4a94698d    /data/01        ext4    defaults,noatime        1 2
#/dev/sde
UUID=cbfa5b6d-46e7-424e-9d2a-9717e60c7172    /data/02        ext4    defaults,noatime        1 2
#/dev/sdf
UUID=5e4c2948-934c-46d1-a150-09aab80a1164    /data/03        ext4    defaults,noatime        1 2
#/dev/sdg
UUID=c3de0ac3-0724-42aa-9519-f1fa9b091c65    /data/04        ext4    defaults,noatime        1 2
#/dev/sdh
UUID=5b0ceac7-18ed-40ad-8e85-7a24dcdc452b    /data/05        ext4    defaults,noatime        1 2
#/dev/sdi
UUID=7d365351-4b1b-47d0-826d-65704dd9dc5b    /data/06        ext4    defaults,noatime        1 2
#/dev/sdj
UUID=6877a5f2-ce8d-412b-903f-af663a966969    /data/07        ext4    defaults,noatime        1 2
#/dev/sdk
UUID=75df98ca-b086-4269-a12d-2d57c44cc2ac    /data/08        ext4    defaults,noatime        1 2
#/dev/sdl
UUID=8548f885-2139-4f99-bf30-09d1a3090c49    /data/09        ext4    defaults,noatime        1 2
#/dev/sdm
UUID=b05bb9a9-5740-49c4-a7ce-19e17094643f    /data/10        ext4    defaults,noatime        1 2
```

2.6 Network configurations for QA

- Two 10-Gbps Ethernet interfaces are bound together in active-active mode, which creates an aggregate interface of 20 Gbps.

  In the past 30 days, the peak of network traffic for different nodes is around 200MB/s to 300MB/s, thus still has large space for further growth.



- The shell command ‘egrep -o "Slow.*?(took|cost)" /var/log/hadoop-hdfs/* | sort | uniq -c’ is executed on Worker 2 and 6 to check data transfer speed between datanodes. The following results (Worker 2) came out, and we can see that slow-block-transfer are not serious for the QA cluster.

```
      1 /var/log/hadoop-hdfs/hadoop-cmf-hdfs-DATANODE-d02ycdpwrk002.sabic.com.log.out.10:Slow BlockReceiver write packet to mirror took
      2 /var/log/hadoop-hdfs/hadoop-cmf-hdfs-DATANODE-d02ycdpwrk002.sabic.com.log.out.10:Slow flushOrSync took
     10 /var/log/hadoop-hdfs/hadoop-cmf-hdfs-DATANODE-d02ycdpwrk002.sabic.com.log.out.1:Slow BlockReceiver write data to disk cost
    171 /var/log/hadoop-hdfs/hadoop-cmf-hdfs-DATANODE-d02ycdpwrk002.sabic.com.log.out.1:Slow BlockReceiver write packet to mirror took
      9 /var/log/hadoop-hdfs/hadoop-cmf-hdfs-DATANODE-d02ycdpwrk002.sabic.com.log.out.1:Slow flushOrSync took
      9 /var/log/hadoop-hdfs/hadoop-cmf-hdfs-DATANODE-d02ycdpwrk002.sabic.com.log.out.2:Slow BlockReceiver write data to disk cost
     75 /var/log/hadoop-hdfs/hadoop-cmf-hdfs-DATANODE-d02ycdpwrk002.sabic.com.log.out.2:Slow BlockReceiver write packet to mirror took
     12 /var/log/hadoop-hdfs/hadoop-cmf-hdfs-DATANODE-d02ycdpwrk002.sabic.com.log.out.2:Slow flushOrSync took
     14 /var/log/hadoop-hdfs/hadoop-cmf-hdfs-DATANODE-d02ycdpwrk002.sabic.com.log.out.3:Slow BlockReceiver write data to disk cost
     46 /var/log/hadoop-hdfs/hadoop-cmf-hdfs-DATANODE-d02ycdpwrk002.sabic.com.log.out.3:Slow BlockReceiver write packet to mirror took
      5 /var/log/hadoop-hdfs/hadoop-cmf-hdfs-DATANODE-d02ycdpwrk002.sabic.com.log.out.3:Slow flushOrSync took
      9 /var/log/hadoop-hdfs/hadoop-cmf-hdfs-DATANODE-d02ycdpwrk002.sabic.com.log.out.4:Slow BlockReceiver write data to disk cost
     57 /var/log/hadoop-hdfs/hadoop-cmf-hdfs-DATANODE-d02ycdpwrk002.sabic.com.log.out.4:Slow BlockReceiver write packet to mirror took
      9 /var/log/hadoop-hdfs/hadoop-cmf-hdfs-DATANODE-d02ycdpwrk002.sabic.com.log.out.4:Slow flushOrSync took
      3 /var/log/hadoop-hdfs/hadoop-cmf-hdfs-DATANODE-d02ycdpwrk002.sabic.com.log.out.5:Slow BlockReceiver write packet to mirror took
     15 /var/log/hadoop-hdfs/hadoop-cmf-hdfs-DATANODE-d02ycdpwrk002.sabic.com.lo
```

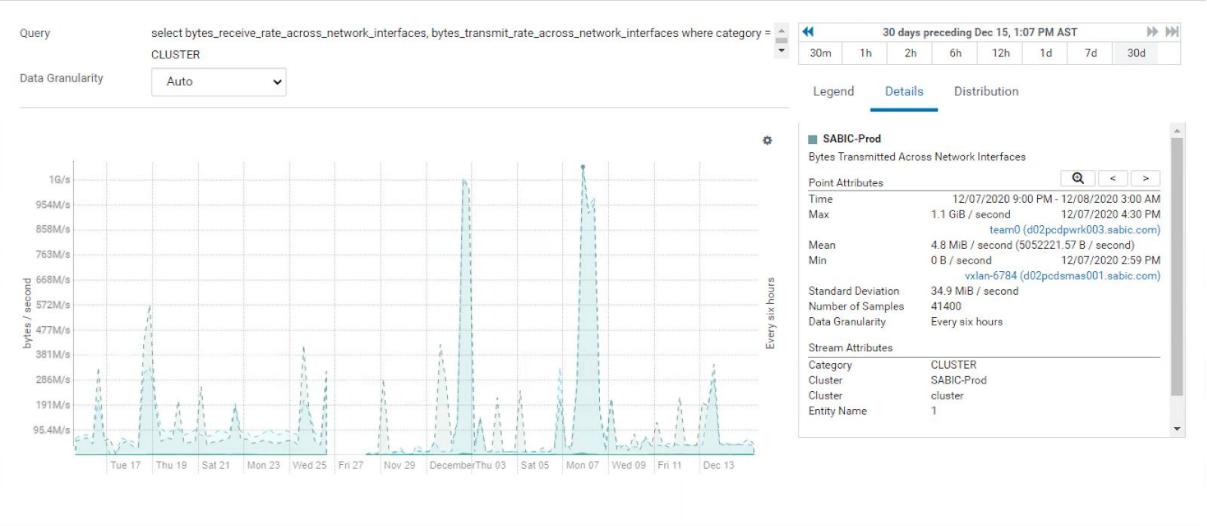
```
g.out.5:Slow flushOrSync took
    1
/var/log/hadoop-hdfs/hadoop-cmf-hdfs-DATANODE-d02ycdpwrk002.sabic.com.lo
g.out.6:Slow BlockReceiver write data to disk cost
   36
/var/log/hadoop-hdfs/hadoop-cmf-hdfs-DATANODE-d02ycdpwrk002.sabic.com.lo
g.out.6:Slow BlockReceiver write packet to mirror took
   24
/var/log/hadoop-hdfs/hadoop-cmf-hdfs-DATANODE-d02ycdpwrk002.sabic.com.lo
g.out.6:Slow flushOrSync took
    1
/var/log/hadoop-hdfs/hadoop-cmf-hdfs-DATANODE-d02ycdpwrk002.sabic.com.lo
g.out.7:Slow BlockReceiver write data to disk cost
   18
/var/log/hadoop-hdfs/hadoop-cmf-hdfs-DATANODE-d02ycdpwrk002.sabic.com.lo
g.out.7:Slow BlockReceiver write packet to mirror took
    6
/var/log/hadoop-hdfs/hadoop-cmf-hdfs-DATANODE-d02ycdpwrk002.sabic.com.lo
g.out.7:Slow flushOrSync took
    5
/var/log/hadoop-hdfs/hadoop-cmf-hdfs-DATANODE-d02ycdpwrk002.sabic.com.lo
g.out.8:Slow BlockReceiver write data to disk cost
   24
/var/log/hadoop-hdfs/hadoop-cmf-hdfs-DATANODE-d02ycdpwrk002.sabic.com.lo
g.out.8:Slow BlockReceiver write packet to mirror took
    9
/var/log/hadoop-hdfs/hadoop-cmf-hdfs-DATANODE-d02ycdpwrk002.sabic.com.lo
g.out.8:Slow flushOrSync took
    7
/var/log/hadoop-hdfs/hadoop-cmf-hdfs-DATANODE-d02ycdpwrk002.sabic.com.lo
g.out.9:Slow BlockReceiver write packet to mirror took
    7
/var/log/hadoop-hdfs/hadoop-cmf-hdfs-DATANODE-d02ycdpwrk002.sabic.com.lo
g.out.9:Slow flushOrSync took
    4
/var/log/hadoop-hdfs/hadoop-cmf-hdfs-DATANODE-d02ycdpwrk002.sabic.com.lo
g.out:Slow flushOrSync took
```

- Jumbo frames with 9000 mtu are NOT configured.
- 'RX errors' and 'TX errors' are not detected in Worker 2 and 6.

2.7 Network configurations for PROD

- Two 10-Gbps Ethernet interfaces are bound together in active-active mode, which creates an aggregate interface of 20 Gbps.

  In the past 30 days, the peak of network traffic for different nodes is around 200MB/s to 300MB/s. However, there are 2 peak time points at which time, the network traffic is arriving at around 1GB/s for Worker 3. The network bandwidth for each node is 20 Gbps (roughly 2000 - 2500MB/s), thus still enough to hold the highest peak time. However, we still recommend observing, to make sure there are not too many such peaks.



- The shell command 'egrep -o "Slow.*?(took|cost)" /var/log/hadoop-hdfs/* | sort | uniq -c' is executed on Worker 2 and 6 to check data transfer speed between datanodes. The following results (Worker 2) came out, and we can see that slow-block-transfer are not serious for the PROD cluster (even better than the QA cluster).

```
    1 /var/log/hadoop-hdfs/hadoop-cmf-hdfs-DATANODE-d02pcdpwrk002.sabic.com.log.out.10:Slow PacketResponder send ack to upstream took
    1 /var/log/hadoop-hdfs/hadoop-cmf-hdfs-DATANODE-d02pcdpwrk002.sabic.com.log.out.1:Slow BlockReceiver write packet to mirror took
    2 /var/log/hadoop-hdfs/hadoop-cmf-hdfs-DATANODE-d02pcdpwrk002.sabic.com.log.out.2:Slow BlockReceiver write data to disk cost
    1 /var/log/hadoop-hdfs/hadoop-cmf-hdfs-DATANODE-d02pcdpwrk002.sabic.com.log.out.4:Slow BlockReceiver write packet to mirror took
    1 /var/log/hadoop-hdfs/hadoop-cmf-hdfs-DATANODE-d02pcdpwrk002.sabic.com.log.out.5:Slow BlockReceiver write data to disk cost
    1 /var/log/hadoop-hdfs/hadoop-cmf-hdfs-DATANODE-d02pcdpwrk002.sabic.com.log.out.5:Slow BlockReceiver write packet to mirror took
    1 /var/log/hadoop-hdfs/hadoop-cmf-hdfs-DATANODE-d02pcdpwrk002.sabic.com.log.out.7:Slow BlockReceiver write data to disk cost
    2 /var/log/hadoop-hdfs/hadoop-cmf-hdfs-DATANODE-d02pcdpwrk002.sabic.com.log.out.7:Slow BlockReceiver write packet to mirror took
    2 /var/log/hadoop-hdfs/hadoop-cmf-hdfs-DATANODE-d02pcdpwrk002.sabic.com.log.out.8:Slow BlockReceiver write packet to mirror took
    1 /var/log/hadoop-hdfs/hadoop-cmf-hdfs-DATANODE-d02pcdpwrk002.sabic.com.log.out.9:Slow BlockReceiver write packet to mirror took
    1
```

/var/log/hadoop-hdfs/hadoop-cmf-hdfs-DATANODE-d02pcdpwrk002.sabic.com.log.out:Slow BlockReceiver write packet to mirror took

- Jumbo frames with 9000 mtu are NOT configured.
- ‘RX errors’ and ‘TX errors’ are not detected in Worker 2 and 3.

3 Cluster Overview

3.1 Nodes and components

The QA cluster is composed of:

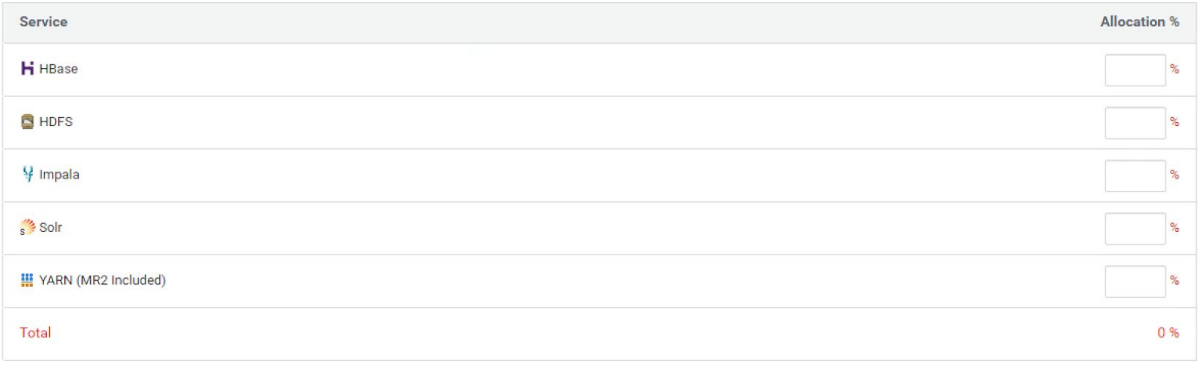
- Master nodes: 2
- Worker (Datanode) nodes: 10 (2 of them are not used)
- KMS: 2
- KTS: 2
- CDSW: 3
- Edge nodes: 2
- Others: 2

The PROD cluster is composed of:

- Master nodes: 3
- Worker (Datanode) nodes: 5
- KMS: 2
- KTS: 2
- CDSW: 1
- Edge nodes: 2
- Others: 1

3.2 Resource allocation

The static resource allocation for QA cluster is not configured in this cluster, as shown in the following diagram.



The QA cluster memory allocation among these components are shown in the following (from perspective of worker nodes).

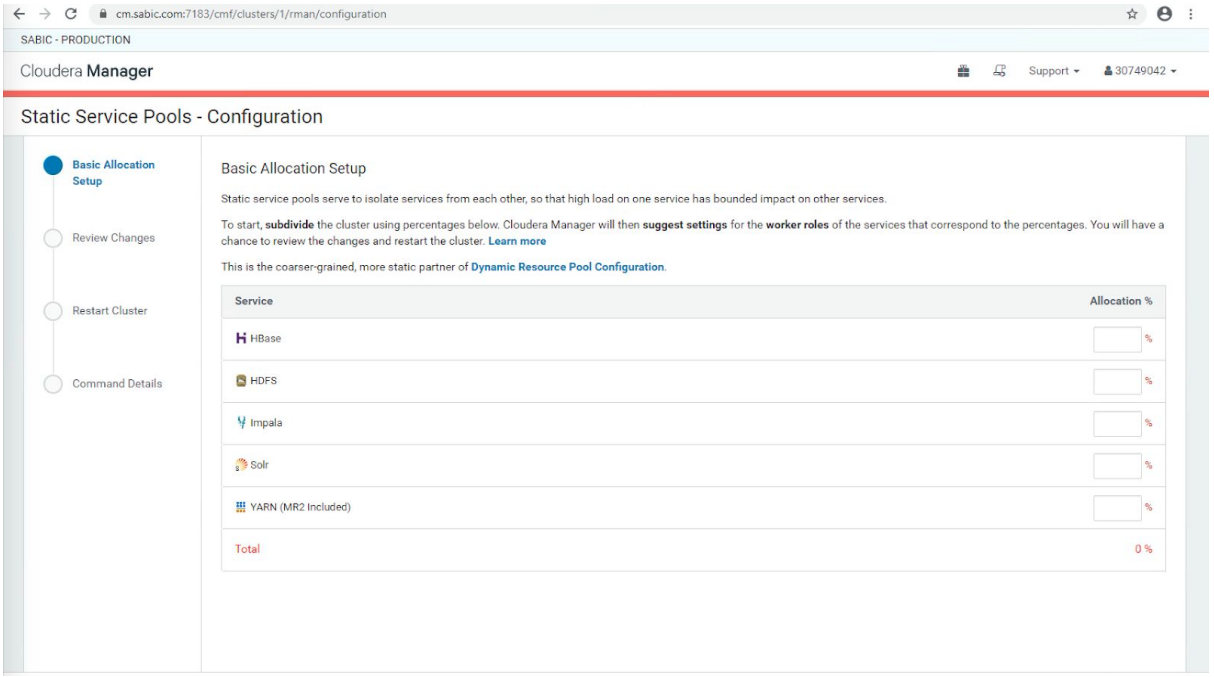
	Components	Allocated	Percentage
1	TOTAL	512GB	
2	HDFS	DATANODE HEAPSIZE: 4GB HDFS CACHE: 4GB	1.56%
3	YARN	NM HEAPSIZE: 2GB	63.31%

		CONTAINER: 322.17GB	
4	IMPALA	IMPALA_DAEMON (mem_limit): 128GB Embeded JVM HEAPSIZE: 32GB	31.25%
5	HBASE	1GB	0.20%
6	SOLR	JVM HEAPSIZE: 32GB JVM Direct Memory: 32GB	12.5%

Major concerns are:

- YARN NodeManager heapsize is recommended to be increased to 4GB.
- Although HBASE is not in heavy load, the RegionServer JVM heapsize of 1GB is still too little. Recommend to increase to 12GB (the same as the PROD cluster).
- Although the summary of the memory allocation is already going beyond the physical memory (557.17GB in total), it is still safe as the actual memory usage is still not high. Especially when it is not a production environment. Recommend to monitor the actual memory usage for each worker node.

The static resource allocation for the PROD cluster is not configured in this cluster, as shown in the following diagram.



The PROD cluster memory allocation among these components are (worker node):

	Components	Allocated	Percentage
1	TOTAL	512GB	
2	HDFS	**DATANODE HEAPSIZE: 1GB** HDFS CACHE: 4GB	0.98%
3	YARN	NM HEAPSIZE: 2GB	35.55%

		CONTAINER: 180GB	
4	IMPALA	IMPALA_DAEMON (mem_limit): 128GB Embeded JVM HEAPSIZE: 32GB	31.25%
5	HBASE	12GB	2.34%
6	SOLR	JVM HEAPSIZE: 16GB JVM Direct Memory: 5GB	4.10%

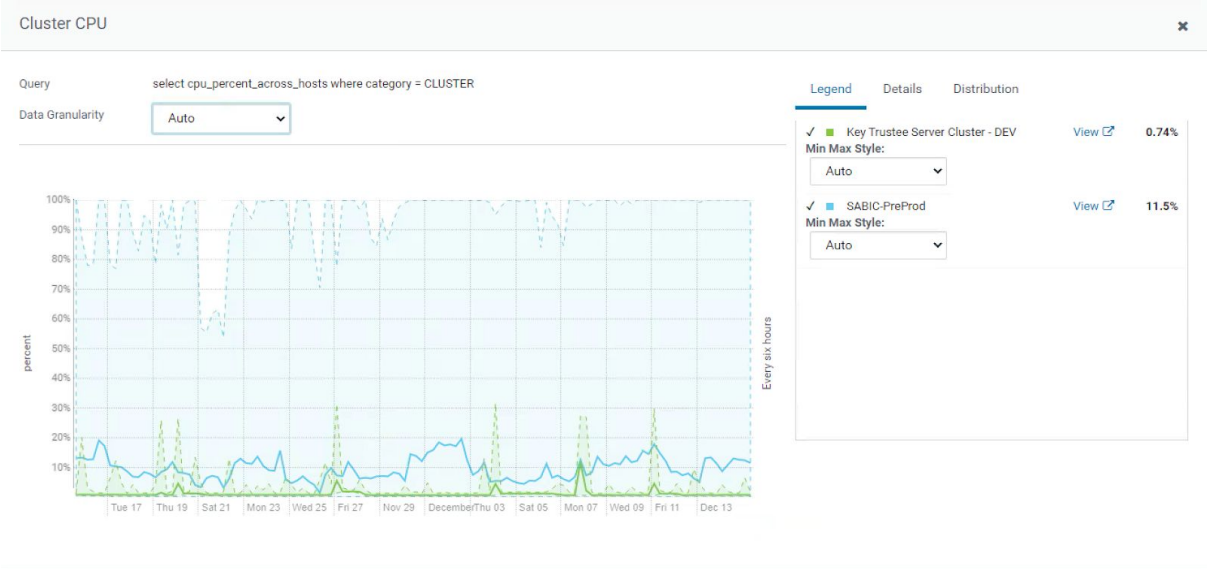
Major concerns are:

- HDFS Datanode JVM heapsize is highly recommended to be increased to 4GB or above.
- YARN NodeManager JVM heapsize is recommended to be increased to 4GB.
- As the summary of the memory allocation is only 380GB, it is recommended to increase the memory allocation for YARN and IMPALA to further improve the performance of them. Generally speaking, it is safe to increase the memory allocation from current 342GB to 462GB.

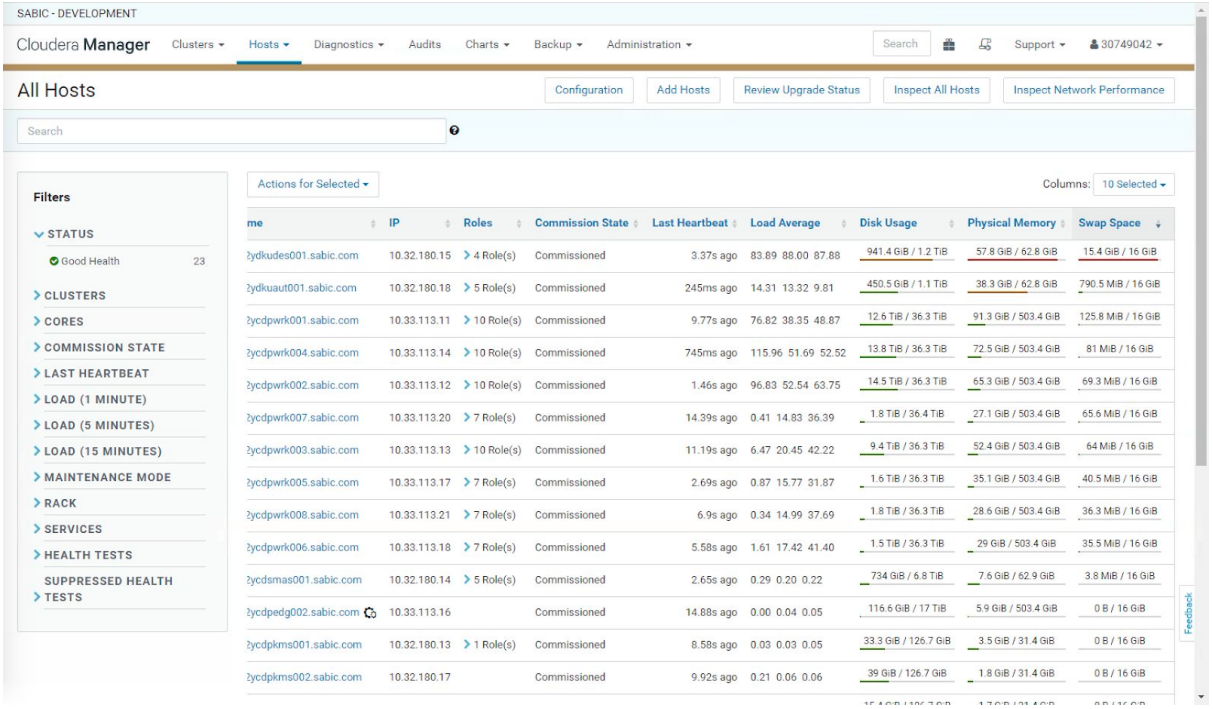
3.3 Cluster resource utilization

3.3.1 QA Cluster

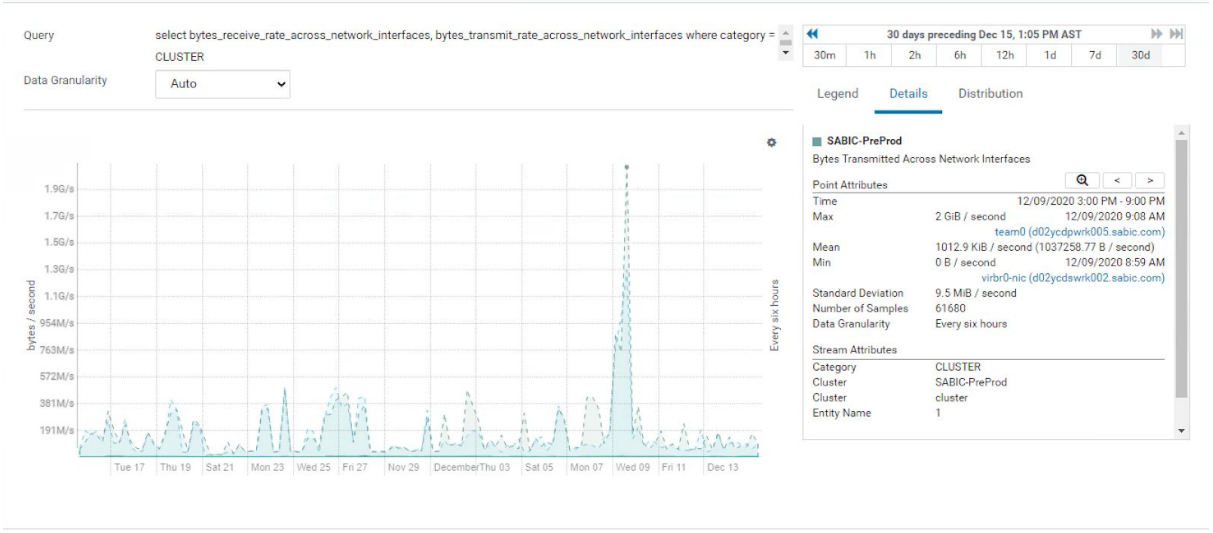
The cluster CPU utilization over the past 30 days is around 10% - 20%, so that we can consider that the CPU will not be the bottleneck for this cluster.



The following is the memory usage of each cluster node, which shows that the memory is also far below its capacity limits **except for d02ydkudes001. As is checked, d02ydkudes001 is running other components than CDH.**

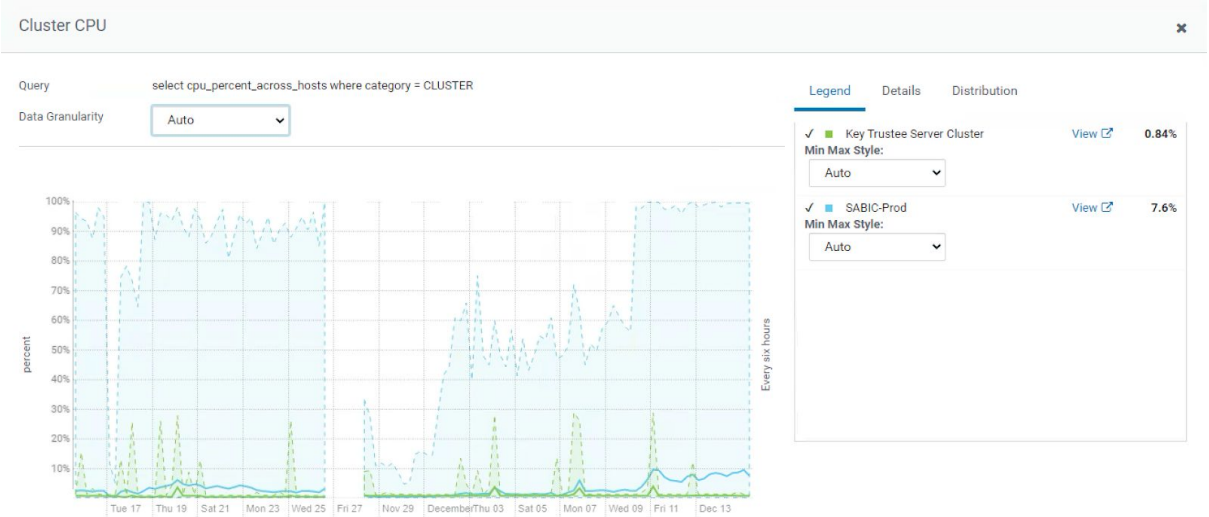


In the past 30 days, the peak of network traffic for different nodes is around 200MB/s to 300MB/s. The network bandwidth for each node is 20 Gbps (roughly 2000 - 2500MB/s), thus still has large space for further growth.

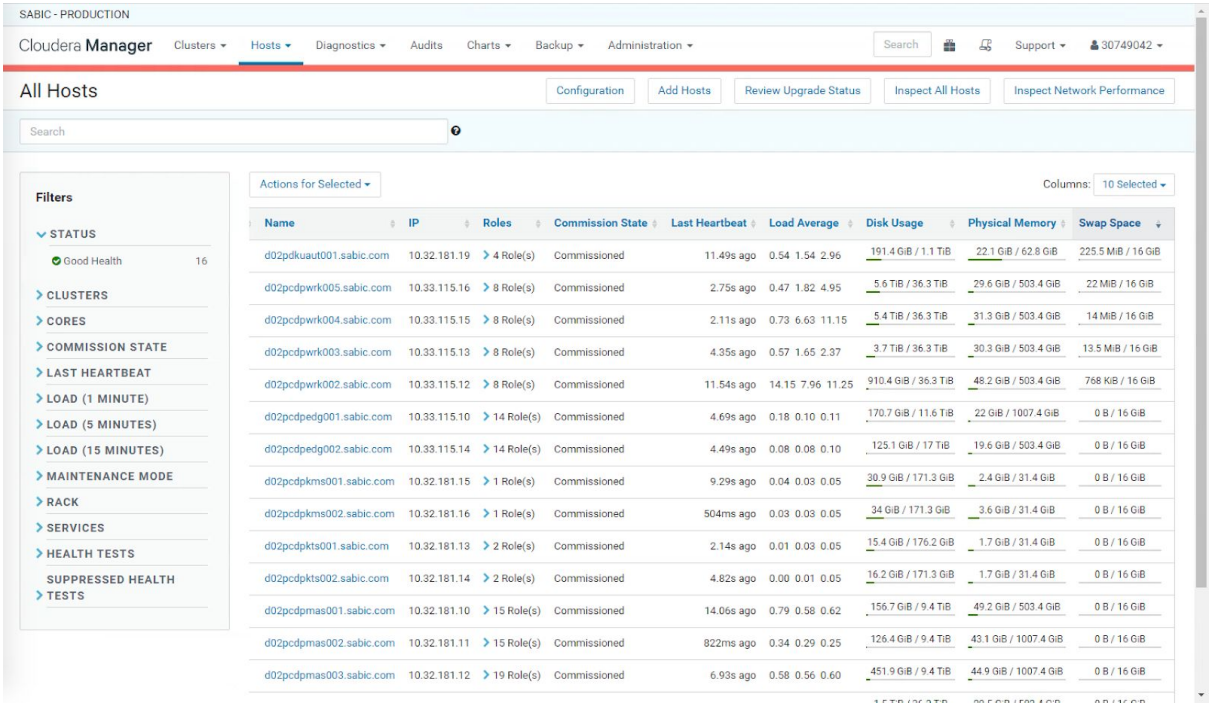


3.3.2 PROD Cluster

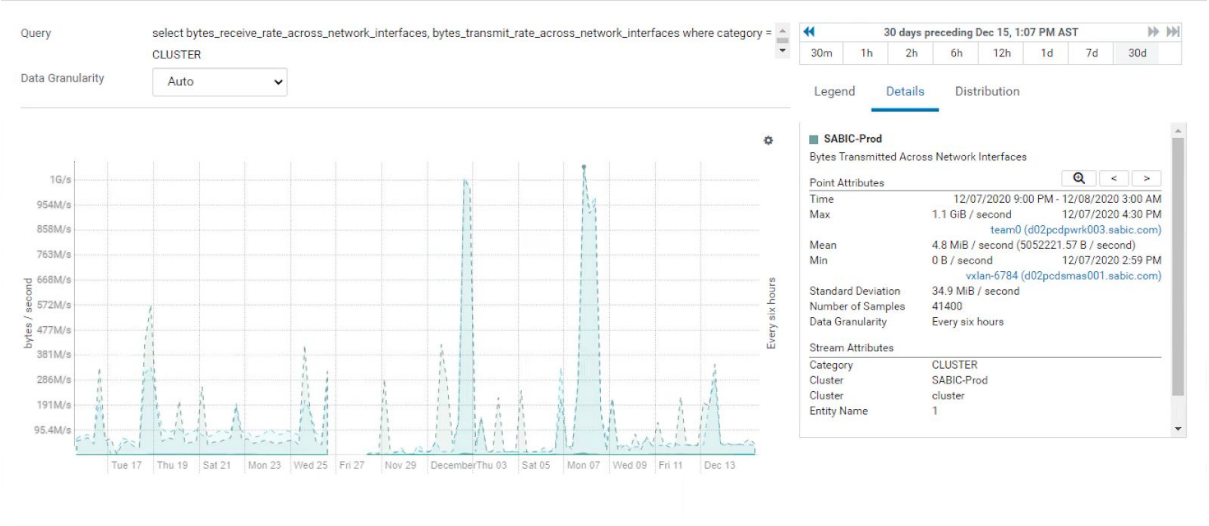
The cluster CPU utilization over the past 30 days is seldom going beyond 10%, so that we can consider that the CPU will not be the bottleneck for this cluster.



The following is the memory usage of each cluster node, which shows that the memory is also far below its capacity limits.



In the past 30 days, the peak of network traffic for different nodes is around 200MB/s to 300MB/s. However, there are 2 peak time points at which time, the network traffic is arriving at around 1GB/s for Worker 3. The network bandwidth for each node is 20 Gbps (roughly 2000 - 2500MB/s), thus still enough to hold the highest peak time. However, we still recommend observing, to make sure there are not too many such peaks.



3.4 Log file locations

3.4.1 QA Cluster

	CDH Components	Log Path
1	HDFS	/var/log/hadoop-hdfs /var/log/hadoop-hdfs/audit
2	YARN	RM and NM: /var/log/hadoop-yarn JHS: /var/log/hadoop-mapreduce
3	ZOOKEEPER	/var/log/zookeeper
4	HBASE	/var/log/hbase /var/log/hbase/audit
5	SPARK	/var/log/spark /var/log/spark/lineage
6	HIVE	/var/log/hive /var/log/audit /var/log/lineage /var/log/operation_logs /var/log/hcatalog
7	IMPALA	/var/log/impalad /var/log/impalad/audit /var/log/impalad/lineage /var/log/catalogd /var/log/statestore
8	SOLR	/var/log/solr /var/log/solr/audit
9	KAFKA	/var/log/kafka

All of the logs are placed in the /var file system. The overall capacity for this file system is 707GB. As is checked from d02ycdpwrk002 and d02ycdpwrk006, the current usage is only 2% and 11% (11GB and 70GB), thus far from being full.

3.4.2 PROD Cluster

	CDH Components	Log Path
1	HDFS	/var/log/hadoop-hdfs /var/log/hadoop-hdfs/audit
2	YARN	RM and NM: /var/log/hadoop-yarn JHS: /var/log/hadoop-mapreduce
3	ZOOKEEPER	/var/log/zookeeper
4	HBASE	/var/log/hbase /var/log/hbase/audit
5	SPARK	/var/log/spark /var/log/spark/lineage
6	HIVE	/var/log/hive /var/log/audit /var/log/lineage /var/log/operation_logs /var/log/hcatalog
7	IMPALA	/var/log/impalad /var/log/impalad/audit /var/log/impalad/lineage /var/log/catalogd /var/log/statestore
8	SOLR	/var/log/solr /var/log/solr/audit
9	KAFKA	/var/log/kafka

All of the logs are placed in the /var file system. The overall capacity for this file system is 707GB. As is checked from d02pcdpwrk002 and d02pcdpwrk003, the current usage is only 3% (19GB), thus far from being full.

4 QA Cluster Component Review

4.1 Role distribution and HA configuration for QA

	Master-1 (d02ycdpmas001)	Master-2 (d02ycdpmas002)	Edge-1	Worker 1-8	CDSW 1-3
CM-SERVER	X				
CMS	X				
CLOUDERA NAVIGATOR		X			
HDFS/NN	X	X			
HDFS/JN	X	X	X		
HDFS/DN				X	
HDFS/HTTPFS	X	X			
YARN/RM	X	X			
YARN/JHS	X				
YARN/NM				X	
ZK	X	X	X		
HBASE/MASTER	X	X			
HBASE/RS				X	
SPARK/JHS	X				
HIVE METASTORE	X	X			
HIVE SERVER2	X	X			
IMPALA STATESTORE	X				
IMPALA CATALOG	X				
IMPALA DAEMON				Worker 1-4	
SOLR				X	

SENTRY	X	X			
KAFKA			X		
FLUME			X		
OOZIE	X				
HUE			X		
CDSW					X

Issues:

- Highly recommend not to deploy HDFS HTTPFS at the master nodes, as they might lead to high network traffic and impact the performance of service masters like Namenode and/or HIVE METASTORE.

  It is recommended to put HDFS HTTPFS at Edge-1.

- Highly recommend not to build the single-node Kafka cluster. It is normally at least 3 nodes for a Kafka cluster, even in a pre-production environment. As the cluster will sometimes perform differently in a single-node cluster and a real-distributed cluster.

  A single-node Kafka cluster is also in lack of HA protection.

  It is recommended to deploy a 3-node Kafka cluster in Master-1, Master-2, and Edge-1.

Recommendations & concerns:

- IMPALA-D processes are deployed only on the first 4 worker nodes, while the workers of the other services (like HDFS, HBASE, etc) are all deployed across the worker nodes.

  Considering the balance amongst the worker nodes, please consider of deploying SOLR to only Work 5 - 8, unless the performance of only 4 SOLR nodes are not enough.

- Edge-2 is tentatively offline for maintenance now. Assuming it will be back to service after maintenance, we can put another HUE node in Edge-2, so as to build the high availability of HUE.
- Both YARN history server and SPARK history server can be deployed in the edge node, especially Edge-2.
- OOZIE is better to be deployed in the edge node than the master node, especially Edge-2.

4.2 HDFS review

4.2.1 Major configurations and deployment settings

	ITEMS	***CURRENT STATUS***
1	HDFS HA	Enabled with QJM
2	HDFS HA Automatic Failover	Enabled

3	HDFS HA Fencing	Built-in HDFS fencing mechanism VIA shell(true)
4	Default Blocksize	128 MB
5	Default Replica	3
6	HDFS Total Space	281.92 TB
7	HDFS Used	59.31 TB (21.04%)
8	HDFS Trash	Enabled by default fs.trash.interval: 1 day fs.trash.checkpoint.interval: 1 hour
9	Namenode Heapsize	4GB
10	Namenode Heap Usage	2.03 GB
11	Namenode GC	JAVA_GC_ARGS by default
12	Namenode RPC Handler dfs.namenode.handler.count	60
13	Datanode RPC Handler dfs.datanode.handler.counnt	3
14	Namenode Checkpoint dfs.namenode.checkpoint.period	1 hour
15	Namenode Leave Safemode dfs.namenode.safemode.threshold-pct	0.999
16	FSIMAGE	**Path (dfs.namenode.name.dir): /data/raid1_hdfs/dfs/nn** SIZE: 567MB
17	dfs.datanode.du.reserved	10GB
18	dfs.datanode.failed.volumes.tolerated	10 or 11 (22 disks in total)
19	dfs.client.read.shortcircuit	Enabled
20	dfs.datanode.balance.bandwidthPerSec	10MB

Issues:

- Although the disk of FSIMAGE is already protected by RAID 1, there is still a likelihood of logical error during writing FSIMAGE. Thus it is highly recommended to store FSIMAGE in more than 1 disk. As for now, we can simply add another disk path (either RAID-1 or JBOD) to dfs.namenode.name.dir.

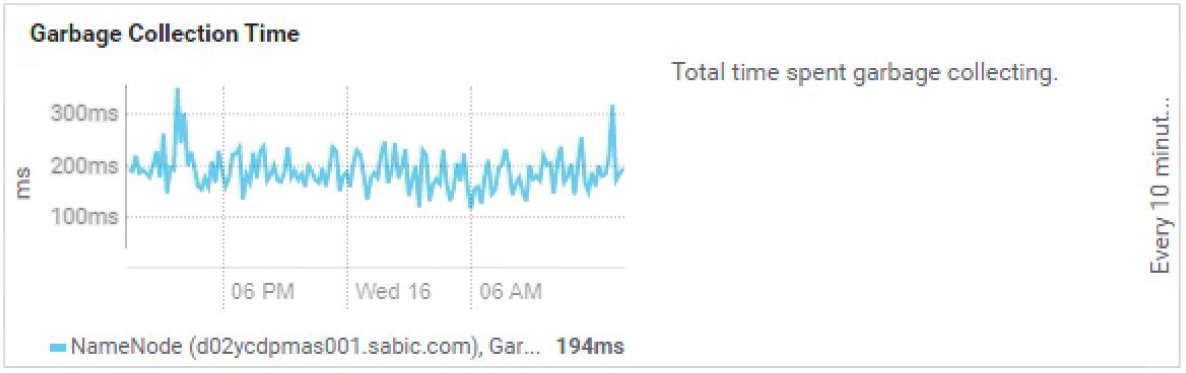
Concerns:

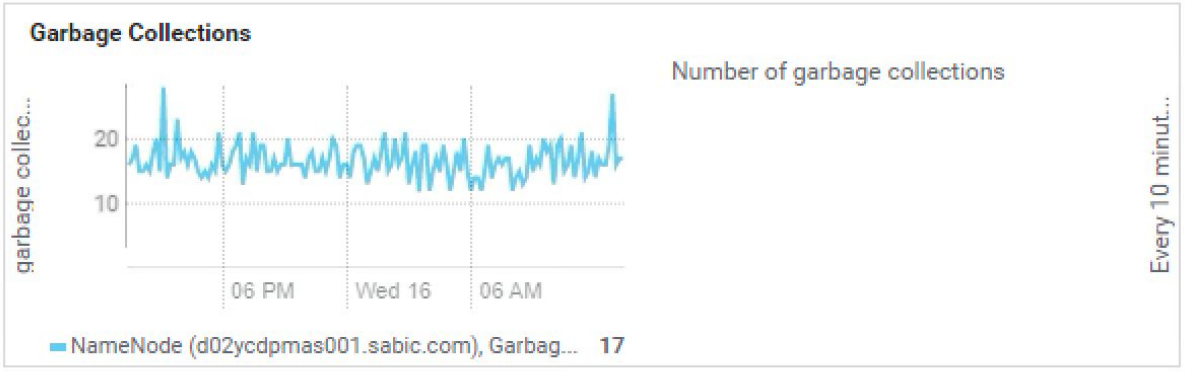
- The NAMENODE JVM heapsize of 4GB is a bit small, especially when the real usage is already more than 2GB. And as is seen in the next section, there are already GCs appearing in Namenode.

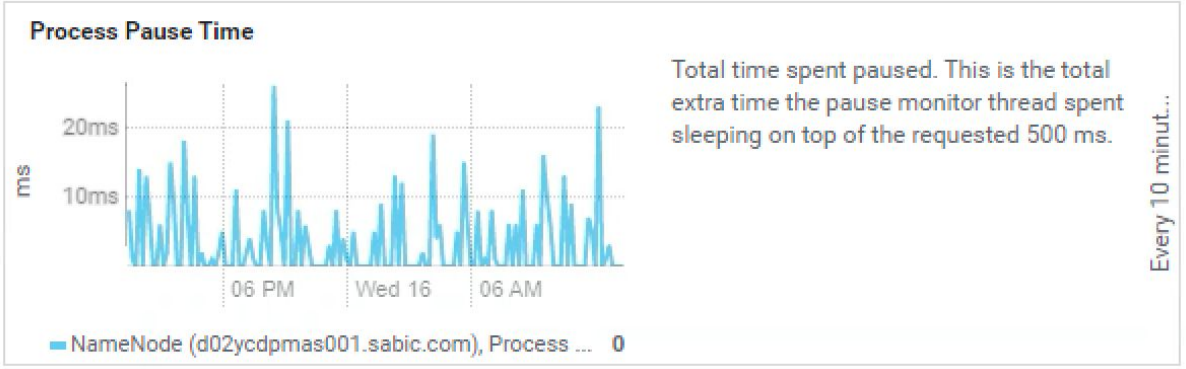
  Considering the further growth of data, the size of the cluster and the capacity of the whole disk volume, it can be set to 8GB.

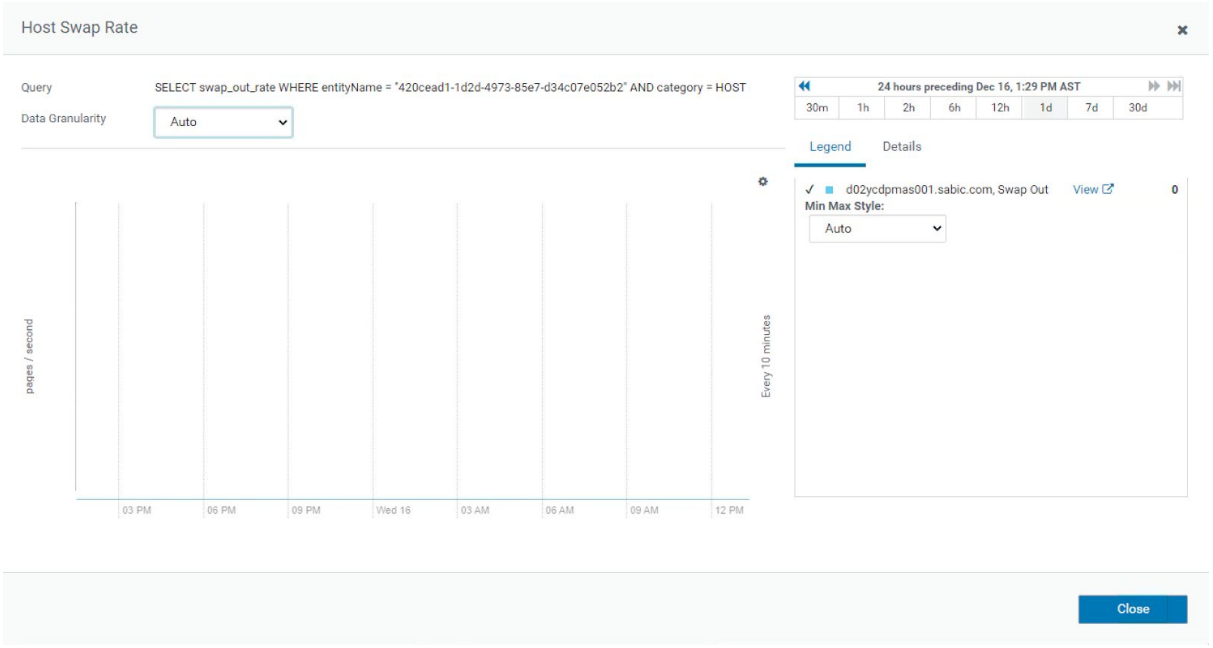
- Recommend to increase the Datanode RPC handler from 3 to 10.
- Recommend to decrease the tolerated failed volumes to a relative small number like 3 or 4.
- Recommend to strictly set the dfs.datanode.du.reserved to 15% - 20% of the local disk space.
- Recommend to increase dfs.datanode.balance.bandwidthPerSec to around 10% - 20% of the total server bandwidth. Note that the 10% - 20% value is estimated per the current network usage

4.2.2 Namenode GC, process pause and swap







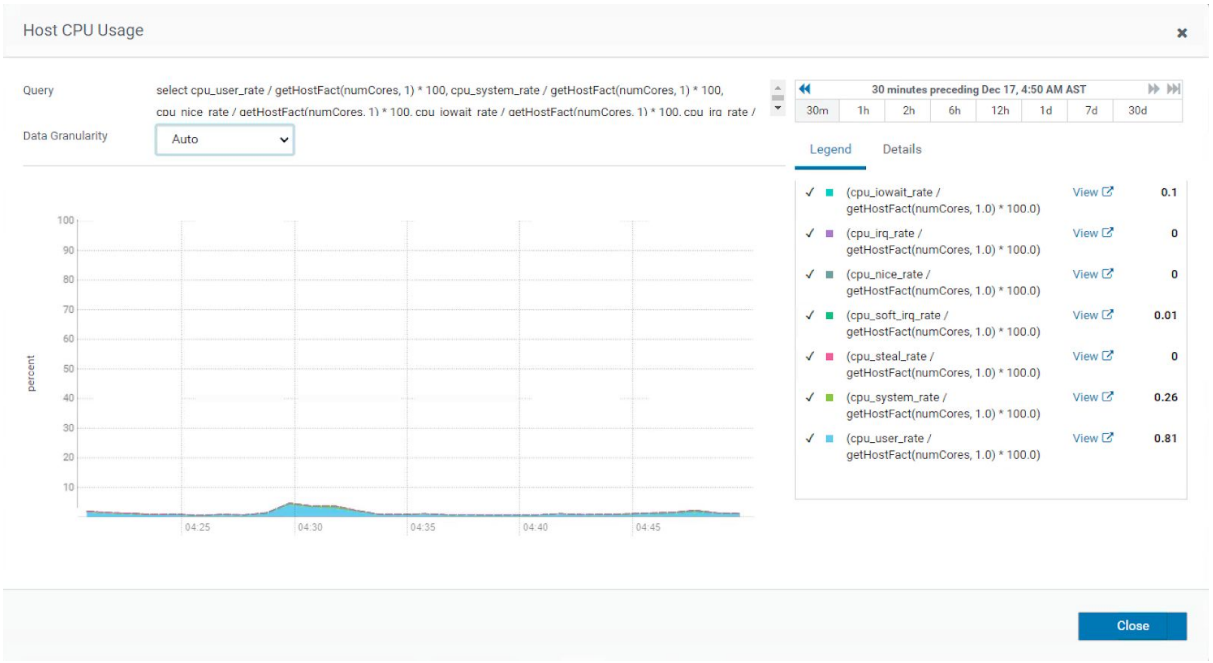


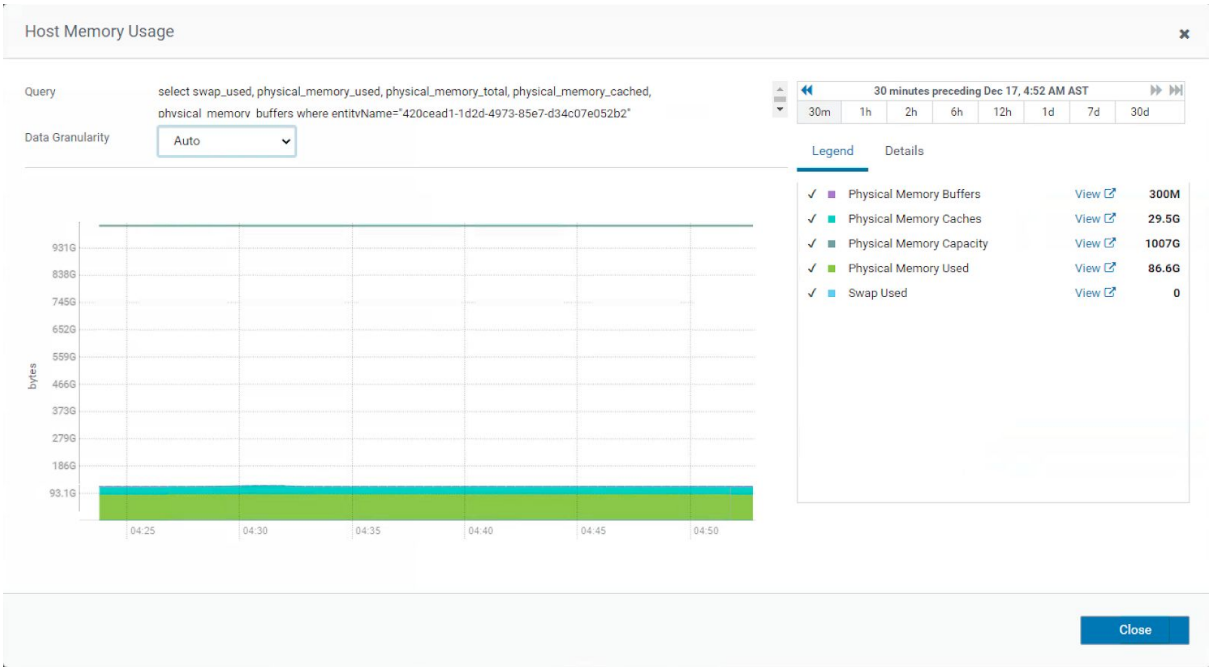
Concerns:

- Although the NAMENODE GC and process pause time is still within the 'normal' range. However, by considering the number of Datanode and the overall data volume, it is really a bit higher. Combining with the NAMENODE heapsize and the current usage, it can be tuned by configuring more heap space (8GB is recommended) for NAMENODE.

4.2.3 Namenode resource usage

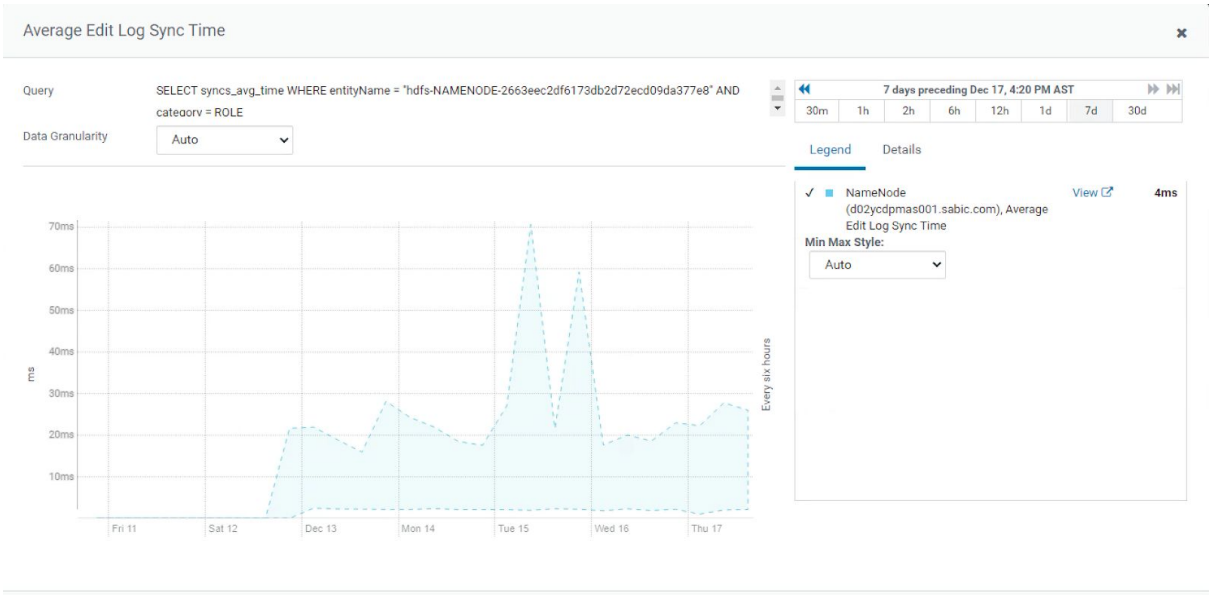
Below is the Namenode resource usage diagram. Per this diagram, the host resource usage of Namenode is fairly low and, of course, totally within the acceptance range.





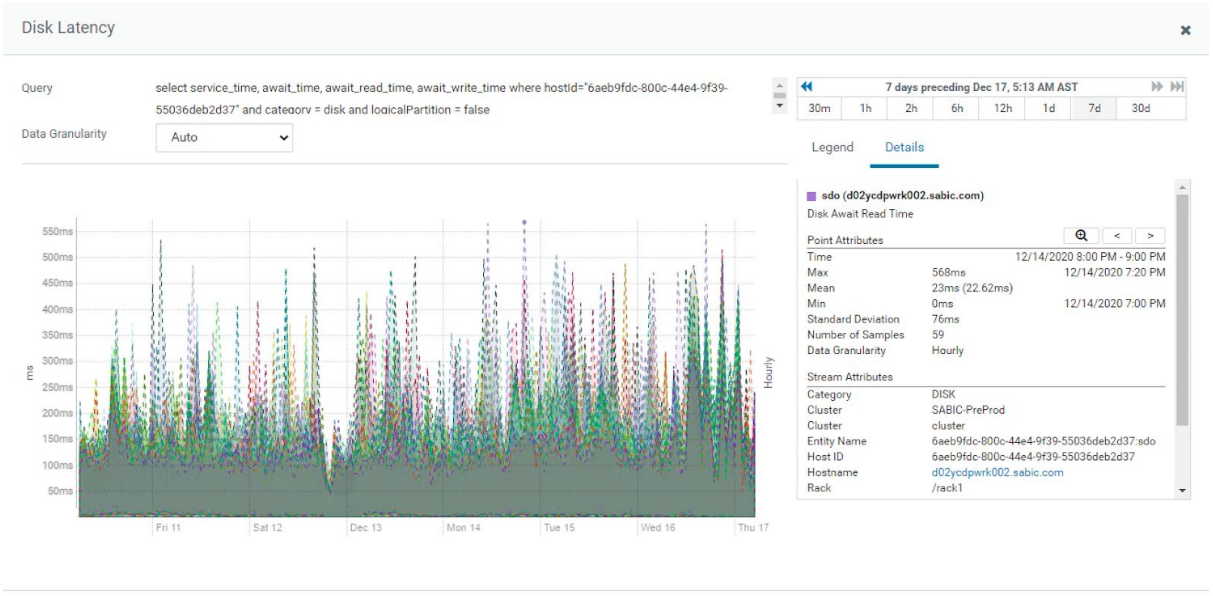
4.2.4 R/W performance

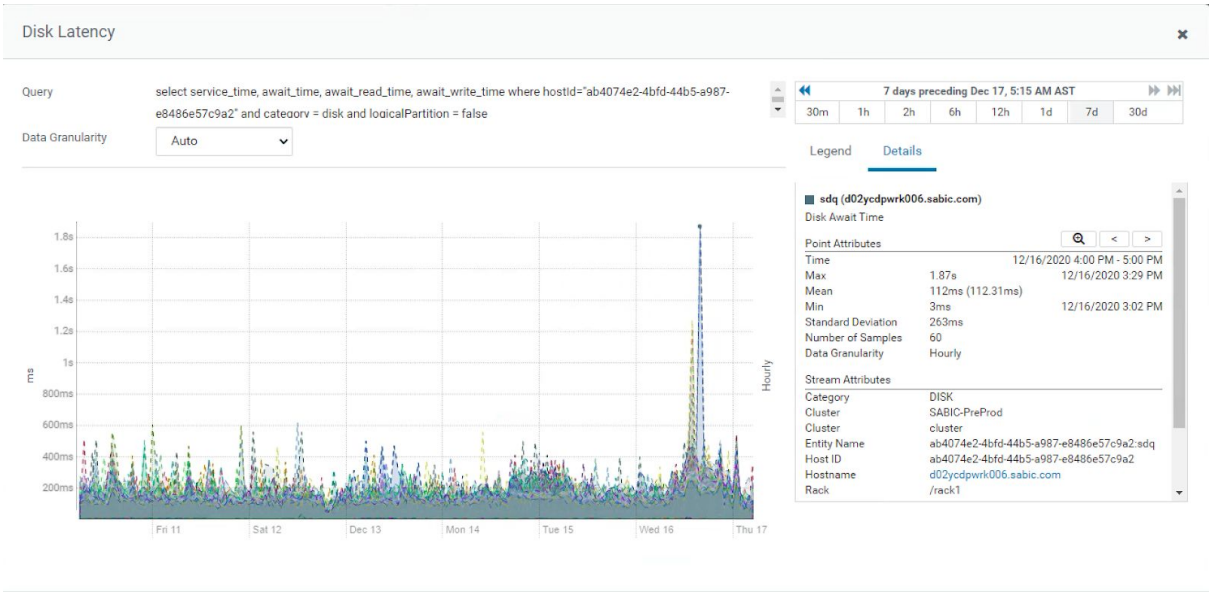
Below is the Editlog Sync time for Namenode. Its average time is below 10ms, and the maximum time is 70m in the past 7 days. Thus fairly good.



Below are the IO latency diagrams for Datanode 2 and 6. The concerns are:

- The disks' I/O latency is already at the high end. Especially for the Datanode 6, one of its disks' average disk latency goes beyond 100ms, and the maximum latency is even approaching 2s. Highly recommended to take care about the R/W performance of the upper-level services (like IMPALA, HIVE, HBASE, etc).





4.3 YARN and container allocation

4.3.1 Major configurations and deployment settings

	ITEMS	*CURRENT STATUS*
1	CPU shares per node (VCORES)	72 (with 36 physical cores for each node)
2	Memory shares per nodes (GB)	322.17GB
3	CPU allocation policy (VCORES)	Maximum: 72 Incremental: 1 Minimum: 1
4	Memory allocation policy (GB)	Maximum: 322.17 Incremental: 1 Minimum: 1

5	Resource allocation policy	DRF
6	Default CPU allocating shares for M/R (VCORES)	Map: 1 Reduce: 1
7	**Default memory allocating shares for M/R (GB)**	**Map: 12GB** **Reduce: 48GB**
8	Heap to Container Size Ratio mapreduce.job.heap.memory-mb.ratio	0.8
9	LOCAL-DIR yarn.nodemanager.local-dirs	Distributed at every local disks
10	Default compression for M/R tasks	Map output compression: SnappyCodec Job output compression: DISABLED
11	Speculative execution mapreduce.map.speculative mapreduce.reduce.speculative	DISABLED
12	mapreduce.task.io.sort.mb	256MB (Could be tuned according to default memory size for mapper and reducer)
13	mapreduce.task.io.sort.factor	64
14	mapreduce.map.sort.spill.percent	0.8
15	Slow starter of reducer mapreduce.job.reduce.slowstart.completedmaps	0.8

Issues:

- The default memory allocation for mappers (12GB) and reducer (48GB) is far too large, which will largely decrease the parallelism of platform's execution, thus impacting the overall performance. Recommend to set it to 4GB for both mappers and reducers.
    - Normally, seldom M/R containers require so much memory.
    - The actual memory for each container can be specified for each job within the range of yarn.scheduler.minimum-allocation-mb (currently 1GB) and yarn.scheduler.maximum-allocation-mb (currently 322.17GB).

Concerns:

- It is NOT recommended to set the container's upper CPU and memory limit to all of the CPU and memory resources allocated to the host. This creates a risk that a M/R job is able to require a large-size container (like 60 VCORES and 300GB of memory), and make all of the other jobs and container requests pending.
    - Recommend to set the upper limit for CPU VCORE as 8, and memory as 16GB.

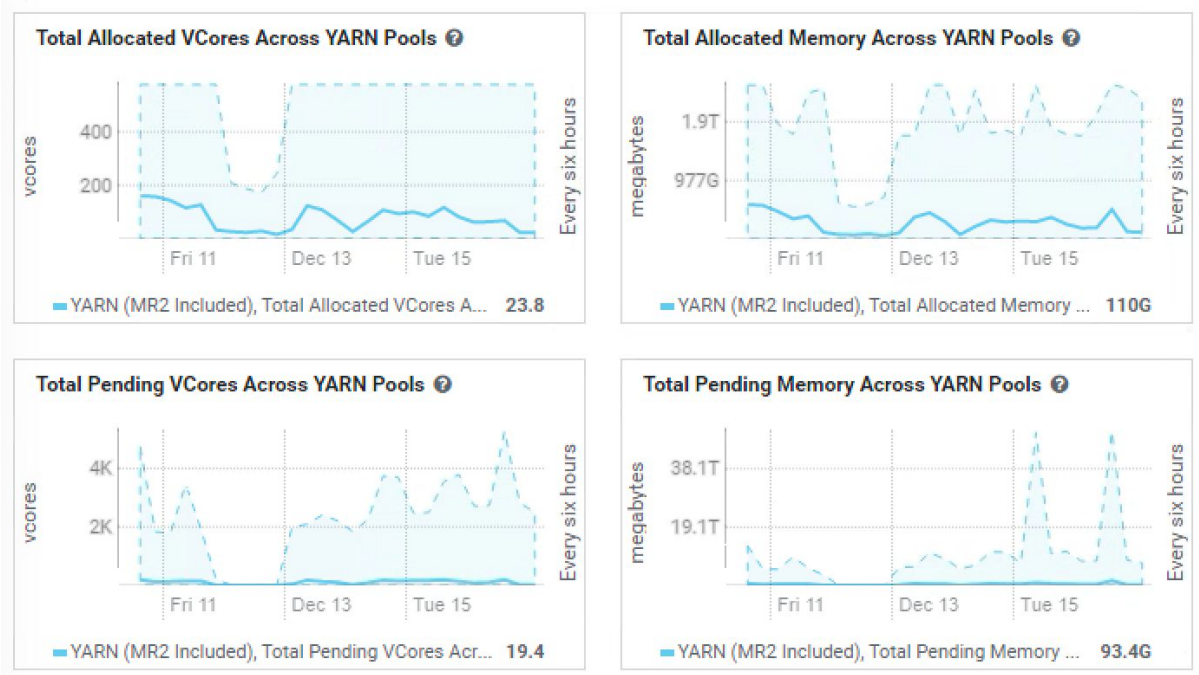
Observations:

- Make sure the VCORES (CPU share per node) will not become the bottleneck in resource allocation.

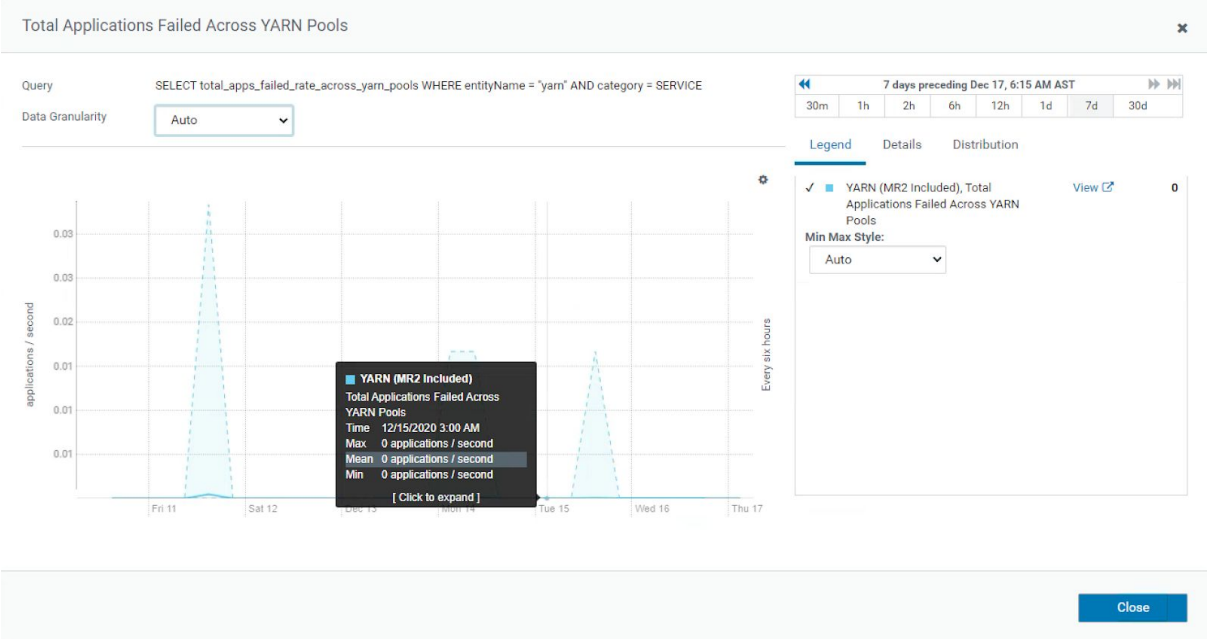
- If the average container size is 4GB, that'll be 80 containers for each node. So that we could slightly increase this value from 72 to 80.

- The sort-area memory (mapreduce.task.io.sort.mb) could be set roughly as 256MB for EACH 4GB.

4.3.2 Overall load and health of YARN and MapReduce



Above are the diagrams for VCORE and memory allocation/pending of YARN. We can see that sometime, the VCOREs of the cluster will be fully occupied, leading to the pending VCORE allocation. So it is recommended to increase the number VCORE definition or decrease the number of maximum allowed VCOREs per container (like the recommendation in Section 4.3.2).



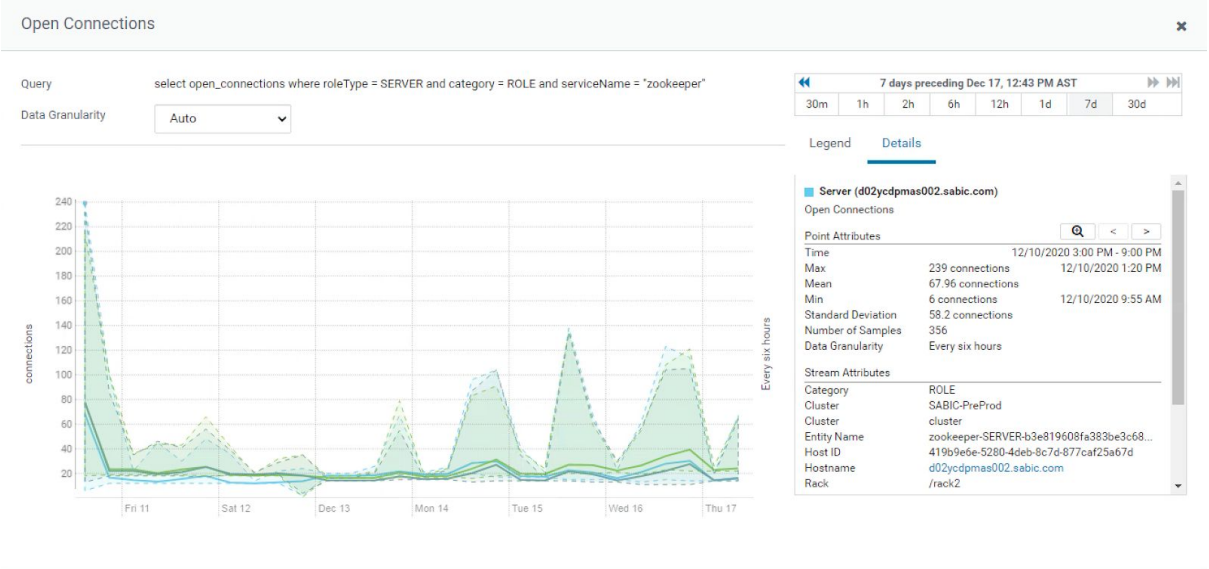
From the diagram above, we can also see that there are rarely failed applications within the past 7 days.

4.4 ZOOKEEPER review

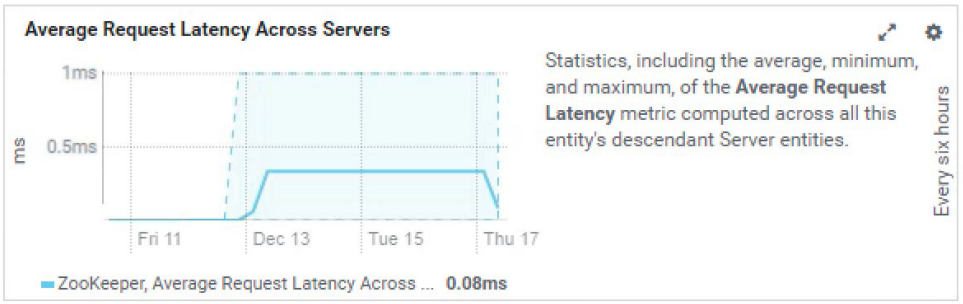
4.4.1 Major configurations and deployment settings

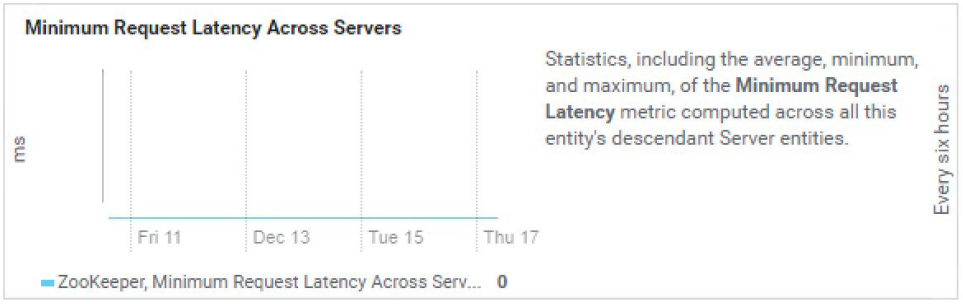
	ITEMS	CURRENT STATUS
1	Data directory	/data/01/zookeeper
2	Transaction log directory	/data/01/zookeeper
3	Maximum client connections	300 (See the next section for recommendation)
4	Maximum session timeout	60 sec
5	JVM Heapsize	1GB

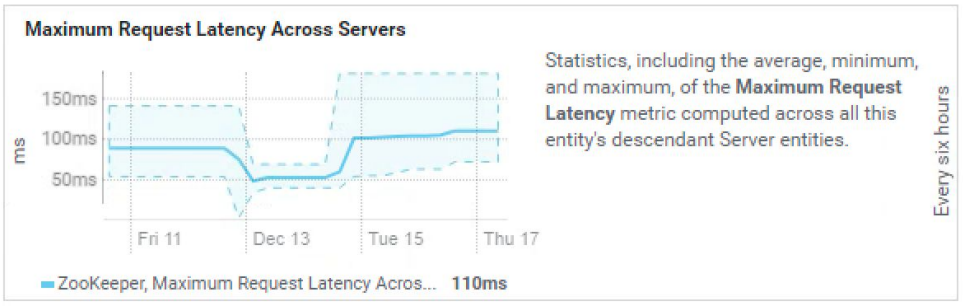
4.4.2 Major Zookeeper health metrics



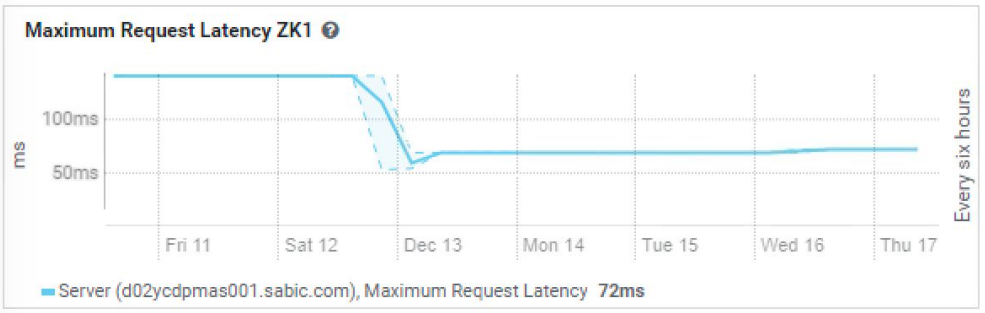
As seen from the figure above, in the past 7 days, the maximum number of connections has reached 239. However, the maximum client connection is now configured to 300. It is recommended to increase the maximum client connections to a larger value like 600 (or even 1000) so as to provide more spaces for further workload increment.

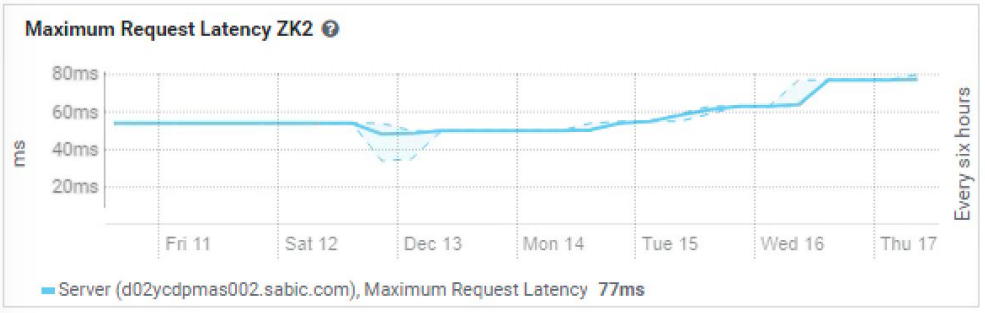


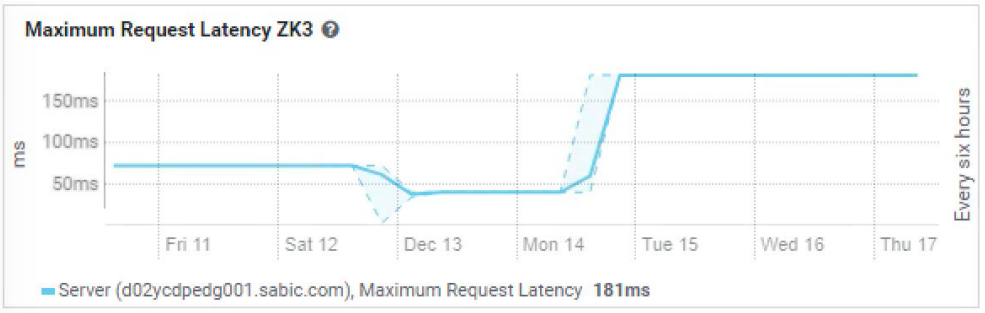




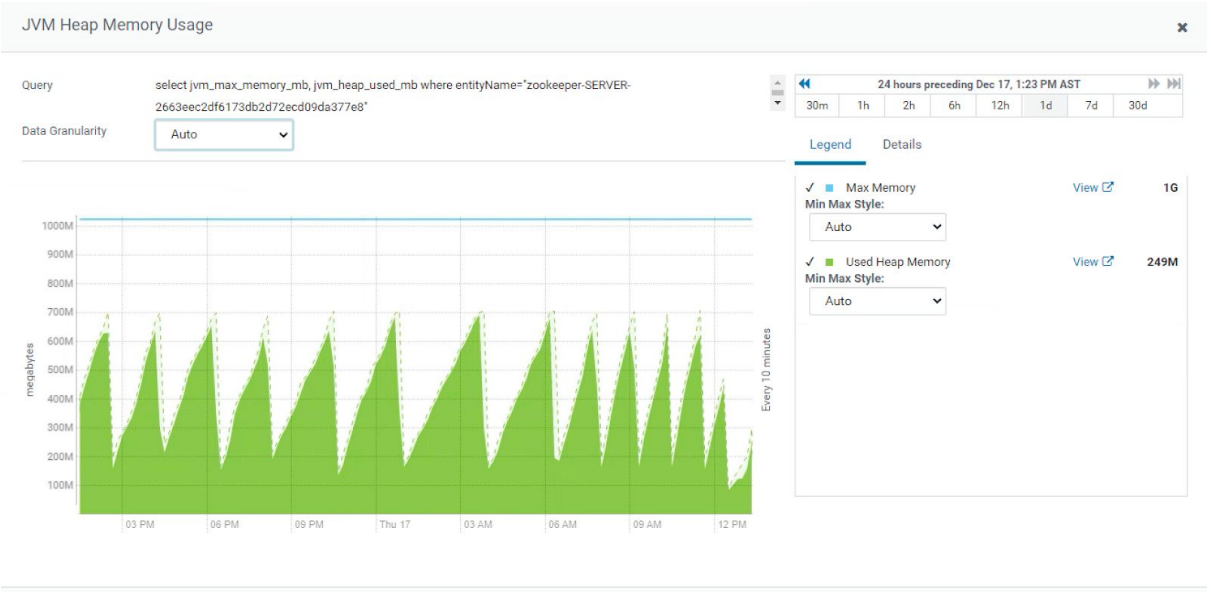
As shown above, although the average request latency is very good, the maximum request latency is a bit high. Further investigation shows that it is the same for all of 3 instances, as shown below.







By investigating the related metrics and parameters, JVM GC is the most suspicious factor. The figure below shows that there is an obvious GC process for every 2 to 3 hours. It is recommended that 1) increase the ZOOKEEPER JVM HEAPSIZE from 1GB to 4GB, and 2) continue to observe whether it will disappear or be different.



4.5 HIVE review

4.5.1 Major configurations and deployment settings

	ITEMS	*CURRENT STATUS*
1	HIVE metastore heapsize	12GB
2	HIVE server2 heapsize	12GB
3	Kerberos and Impersonation	Disabled
4	Stored notification in database	Enabled
5	HIVE execution engine	MapReduce
6	Enable Stats Optimization hive.compute.query.using.stats	Disabled
7	Enable MapJoin Optimization hive.auto.convert.join	Enabled
8	hive.mapjoin.smalltable.filesize	25MB by default
9	Hive Auto Convert Join Noconditional Size hive.auto.convert.join.noconditionaltask.size	20MB
10	Enable Cost-Based Optimizer for Hive hive.cbo.enable	Disabled
11	Enable Vectorization Optimization hive.vectorized.execution.enabled	Enabled
12	Enable Stored Notifications in Database	Enabled

Observations:

- If encountering OOM during HIVE map-join, consider to decrease hive.mapjoin.smalltable.filesize.
  - This parameter can be tuned on HIVE session level.
- See the section of Sentry review for details about Enable Stored Notifications in Database.

4.5.2 Table and partition counts

- The number of HIVE databases:

```
mysql> select count(*) from DBS;
+----------+
| count(*) |
+----------+
|       59 |
+----------+
```

```
1 row in set (0.00 sec)
```

It does not go over the recommended upper limits of 100.

- The number of table in each database:

```
mysql> select TBLS.DB_ID,NAME,count(*) TAB_NUM from TBLS
join DBS on TBLS.DB;
+--------+----------------------------------+---------+
| DB_ID  | NAME                             | TAB_NUM |
+--------+----------------------------------+---------+
| 495883 | petchem_mi_af                    |       1 |
| 483892 | supplier_management_raw          |       1 |
| 495904 | web_data                         |       1 |
| 501863 | sap                              |       1 |
| 143951 | log_analysis                     |       1 |
| 512043 | grid_db                          |       1 |
| 296774 | yanpet                           |       1 |
| 461649 | ibn_ss                           |       1 |
| 235583 | yokogawa                         |       2 |
| 504040 | pi                               |       2 |
| 444371 | smart_inspection                 |       2 |
| 474383 | pe_refined                       |       3 |
| 400972 | poly_viny_chlo                   |       3 |
| 434009 | temp                             |       3 |
| 412538 | petro_kemya                      |       4 |
| 461136 | ibn_sina                         |       5 |
| 483754 | proc_mro_refined                 |       6 |
| 372765 | data_science                     |       6 |
| 337513 | ibn_zahr                         |       7 |
| 209048 | petrokemya                       |       9 |
| 186026 | retail_db                        |      10 |
| 524190 | price_prediction_mmr             |      10 |
| 482202 | plant_efficiency_intermediate    |      13 |
| 471817 | plant_efficiency                 |      13 |
| 469728 | asset_healthcare                 |      14 |
|  94360 | lims                             |      18 |
|  93272 | meridium                         |      20 |
| 481430 | proc_mro_spd                     |      24 |
| 475004 | dpp                              |      27 |
| 468497 | mro_sample                       |      30 |
| 483751 | proc_sm_refined                  |      34 |
| 524196 | price_prediction_ltts            |      38 |
| 492434 | datapred                         |      42 |
| 483737 | proc_sm_inter                    |      47 |
| 483755 | proc_pp_refined                  |      48 |
| 483741 | proc_mro_inter                   |      73 |
| 483752 | proc_st_refined                  |      82 |
|      1 | default                          |     140 |
| 483739 | proc_st_inter                    |     141 |
| 101464 | osipi                            |     155 |
| 455261 | dataiku_dss_test                 |     238 |
```

```
| 483743 | proc_pp_inter                       |      241 |
| 483744 | manuf_pe_refined                    |      252 |
| 482209 | supplier_management_intermediate    |      289 |
| 462206 | mro_spd                             |      295 |
| 492442 | manuf_ahi_inter                     |      519 |
| 483736 | manuf_pe_inter                      |     3922 |
+--------+-------------------------------------+----------+
47 rows in set (0.00 sec)
```

The number of tables for database manuf_pe_inter goes beyond the recommended upper limit of 1000. Attention is recommended to be paid for its metadata access performance.

- The number of partitions for each table (top 50):

```
mysql> select TBLS.TBL_ID,TBL_NAME,count(*) as part_num from PARTITIONS
join TBLS on PARTITIONS.TBL_ID=TBLS.TBL_ID group by TBL_ID,TBL_NAME order
by part_num desc limit 50;
+--------+-----------------------------------------------------+----------+
| TBL_ID | TBL_NAME                                            | part_num |
+--------+-----------------------------------------------------+----------+
| 278720 | petro_minute                                        |  1367911 |
| 474304 | pk_olf1_caustic_wash_tower_date_partitioned         |     2067 |
| 474422 | pk_olf1_deethenizer_date_partitioned                |     2067 |
| 474424 | pk_olf2_furnace10_date_partitioned                  |     2067 |
| 310755 | pk_olefin3_nrt_interpolated                         |     1716 |
| 431550 | pk_compacted                                        |     1630 |
| 323009 | petro                                               |     1061 |
| 474287 | pk_olf2_furnace10_tag_partitioned                   |      447 |
| 290899 | pi_int_3                                            |      294 |
| 472302 | olf1_caustic_wash_tower_tag_part                    |      230 |
| 474213 | pk_olf1_caustic_wash_tower_tag_partitioned          |      230 |
| 474266 | pk_olf1_deethenizer_tag_partitioned                 |      155 |
| 475884 | pk_olf2_furnace10_monthly_partitioned               |      108 |
| 476045 | pk_olf1_caustic_wash_tower_monthly_partitioned      |      108 |
| 499608 | quench_105                                          |      108 |
| 475408 | pk_olf1_deethenizer_monthly_partitioned_1           |      108 |
| 506181 | olf2_c3_16448b_100                                  |      100 |
| 506133 | olf2_c3_16431_100                                   |      100 |
| 506183 | olf2_c3_16448b_200                                  |      100 |
| 506135 | olf2_c3_16431_200                                   |      100 |
| 506185 | olf2_c3_16448b_300                                  |      100 |
| 506137 | olf2_c3_16431_300                                   |      100 |
| 506139 | olf2_c3_16431_400                                   |      100 |
| 506141 | olf2_c3_16431_500                                   |      100 |
| 506107 | olf2_c3_comp_default                                |       92 |
| 504402 | olf2_furn_common                                    |       86 |
| 506142 | olf2_c3_16431_583                                   |       83 |
| 505738 | olf2_butadiene_558                                  |       78 |
| 479737 | pk_olf1_monthly_partitioned                         |       70 |
| 480093 | pk_olf2_monthly_partitioned                         |       70 |
| 546113 | pk_olf2_de_methanizer_stripper_monthly_partitioned  |       70 |
| 485089 | pk_olf1_monthly_partitioned_full                    |       69 |
| 485128 | pk_olf2_monthly_partitioned_full                    |       69 |
| 488323 | comp_30_2                                           |       69 |
| 475224 | pk_olf1_deethenizer_monthly_partitioned             |       68 |
| 506214 | olf2_c3_16702                                       |       63 |
| 506257 | olf2_seperator_180                                  |       60 |
| 506259 | olf2_seperator_240                                  |       60 |
| 506266 | olf2_seperator_300                                  |       60 |
| 505703 | olf2_butadiene_60                                   |       60 |
| 505705 | olf2_butadiene_120                                  |       60 |
| 506237 | olf1_cgc_60                                         |       60 |
```

```
| 505717 | olf2_butadiene_180                                    |       60 |
| 506239 | olf1_cgc_120                                          |       60 |
| 505719 | olf2_butadiene_240                                    |       60 |
| 505721 | olf2_butadiene_300                                    |       60 |
| 506243 | olf1_cgc_240                                          |       60 |
| 505733 | olf2_butadiene_360                                    |       60 |
| 505735 | olf2_butadiene_420                                    |       60 |
| 506253 | olf2_seperator_60                                     |       60 |
+--------+-------------------------------------------------------+----------+
50 rows in set (0.91 sec)
```

The number of partitions for table petro_minute goes beyond the recommended upper limit of 10,000. Attention is recommended to be paid for its metadata access performance.

- The summary number of partitions for each database:

```
mysql> select DBS.DB_ID,NAME,count(*) as part_num from
PARTITIONS,TBLS,DBS where PARTITIONS.TBL_ID=TBLS.TBL_ID and
TBLS.DB_ID=DBS.DB_ID group by DB_ID,NAME order by part_num
desc limit 100;
+--------+------------------+----------+
| DB_ID  | NAME             | part_num |
+--------+------------------+----------+
| 209048 | petrokemya       |  1370698 |
| 471817 | plant_efficiency |     7635 |
| 483736 | manuf_pe_inter   |     5669 |
| 434009 | temp             |     1630 |
|  94360 | lims             |      338 |
| 474383 | pe_refined       |      230 |
| 483741 | proc_mro_inter   |       21 |
| 455261 | dataiku_dss_test |       18 |
| 101464 | osipi            |        9 |
+--------+------------------+----------+
9 rows in set (0.92 sec)
```

The summary number of partitions for database petrokemya goes beyond the recommended upper limit of 100,000. Attention is recommended to be paid for its metadata access performance.

4.6 IMPALA review

4.6.1 Major configurations and deployment settings

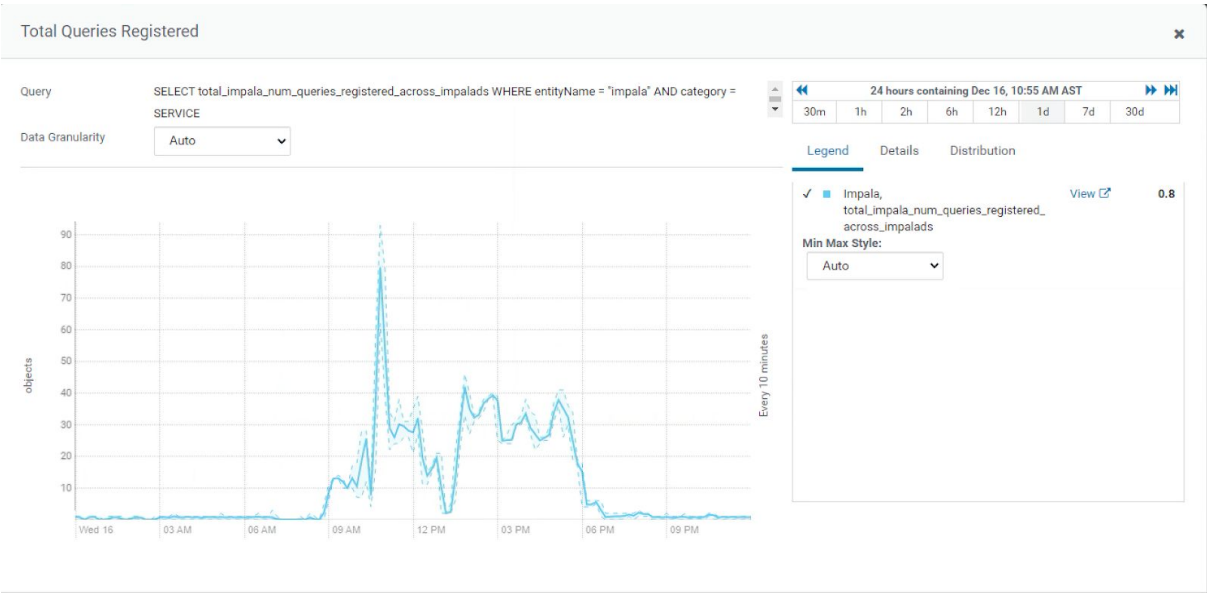
	ITEMS	*CURRENT STATUS*
1	IMPALA_DAEMON (mem_limit)	128GB
2	Embeded JVM HEAPSIZE	32GB
3	Impala Daemon Scratch Directories scratch_dirs	Every disk of the IMPALA-D nodes
4	Local UDF Library Dir local_library_dir	/var/lib/impala/udfs

5	Result Cache Maximum Size max_result_cache_size	100000

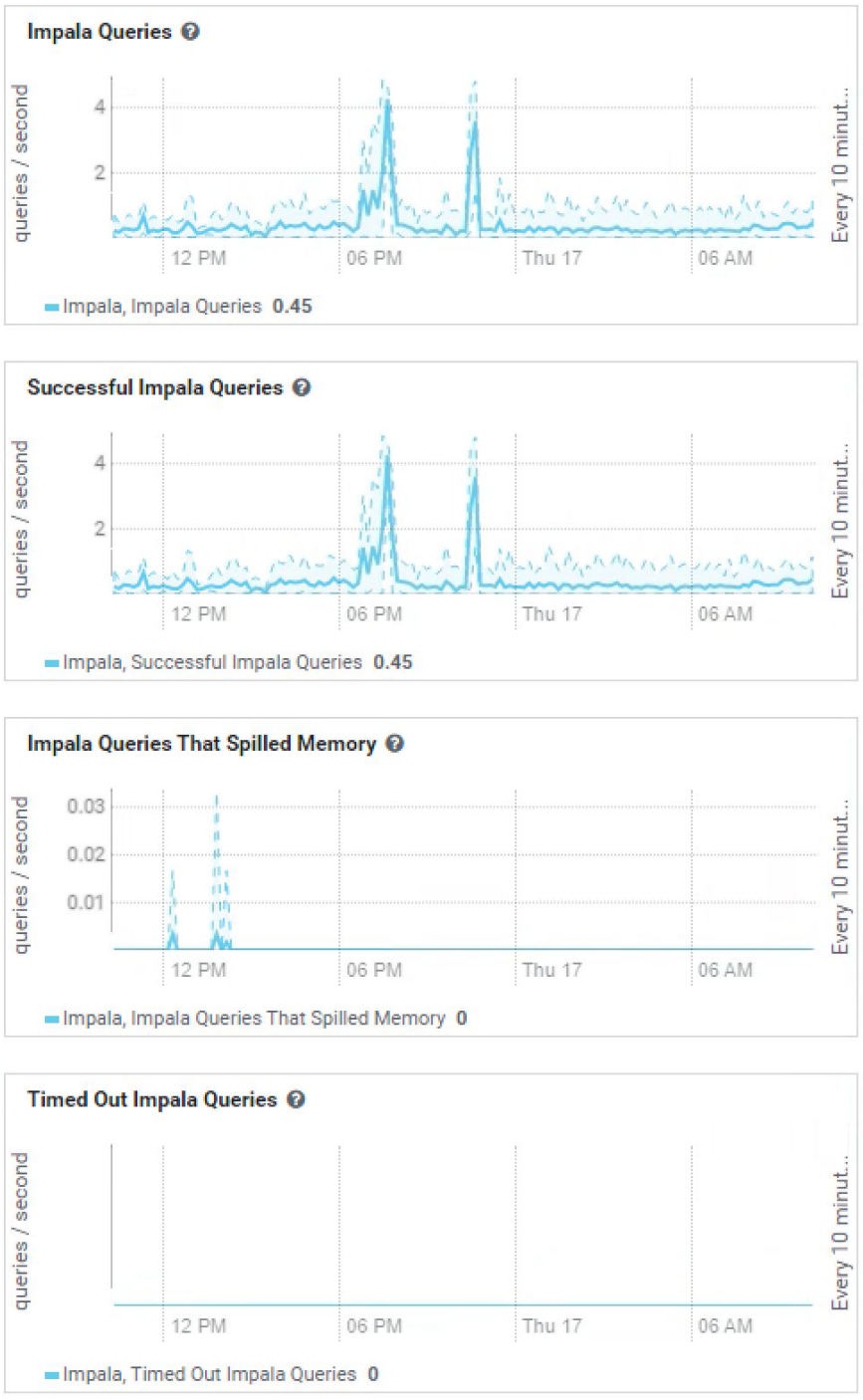
Comments:

- As a friendly reminder, make sure to back up (or migrate) /var/lib/impala/udfs in case of losing your IMPALA user-defined functions.

4.6.2 Workload and success rate

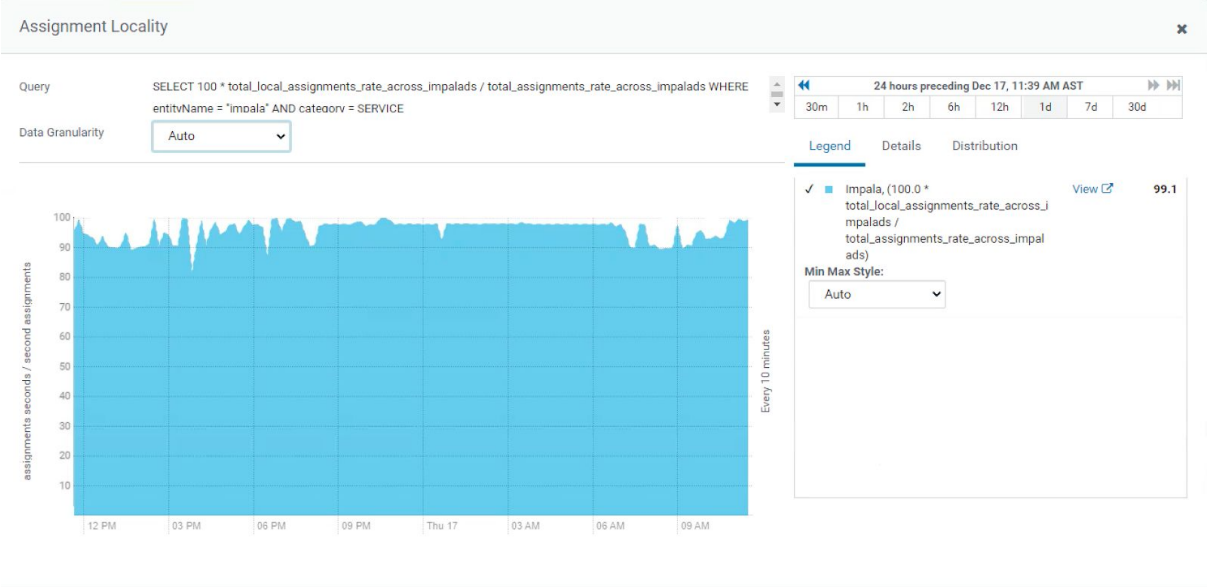


The figure above shows the total number of queries running in (or stopped but tracked by) IMPALA during a whole day (Wed 16 Dec). It can be seens that during the daytime, the concurrency is generally 30 - 40, and 90 during the peak time.



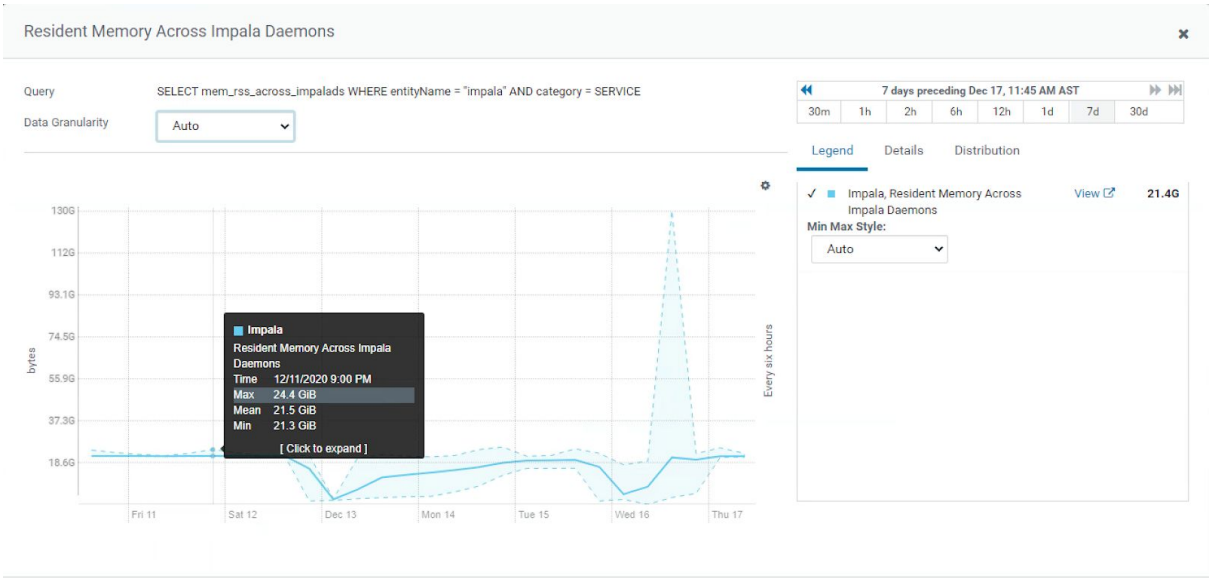
From the figures above, we can see that the IMPALA query incoming rate is almost less than 0.5 query/sec, and the peak rate is only around 4 queries/sec. So the workload is not heavy, and no failures are detected.

4.6.3 Locality of assignment



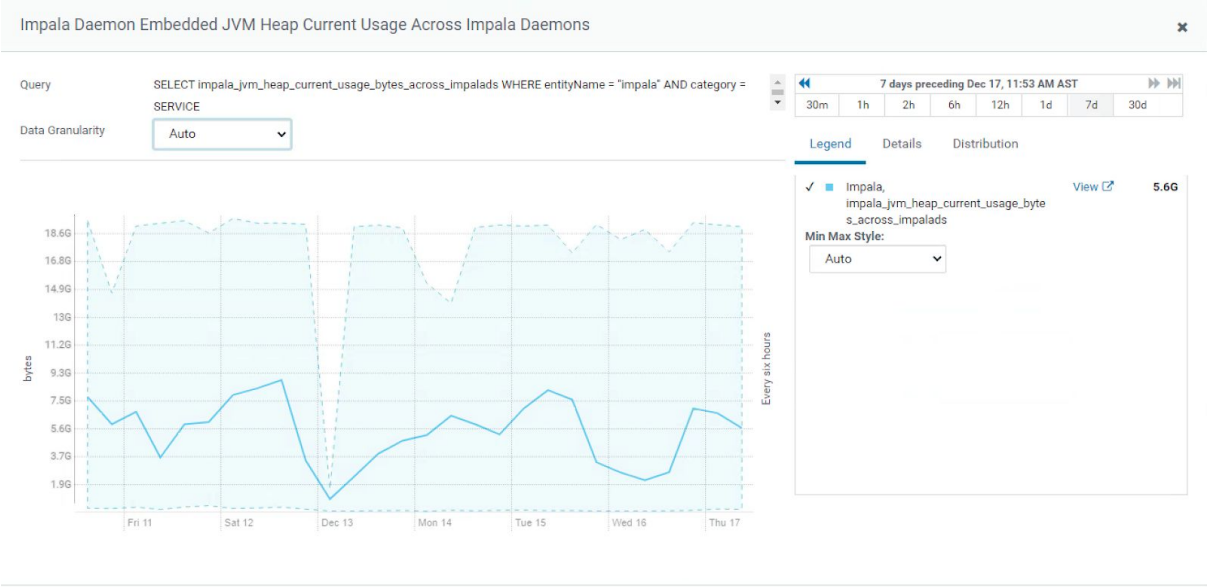
Although the IMPALA-DAEMON is only deployed in Worker 1~4, however, the assignment locality is also maintained at a high level. Thus no problem in this deployment.

4.6.4 Memory usage of IMPALA-DAEMON

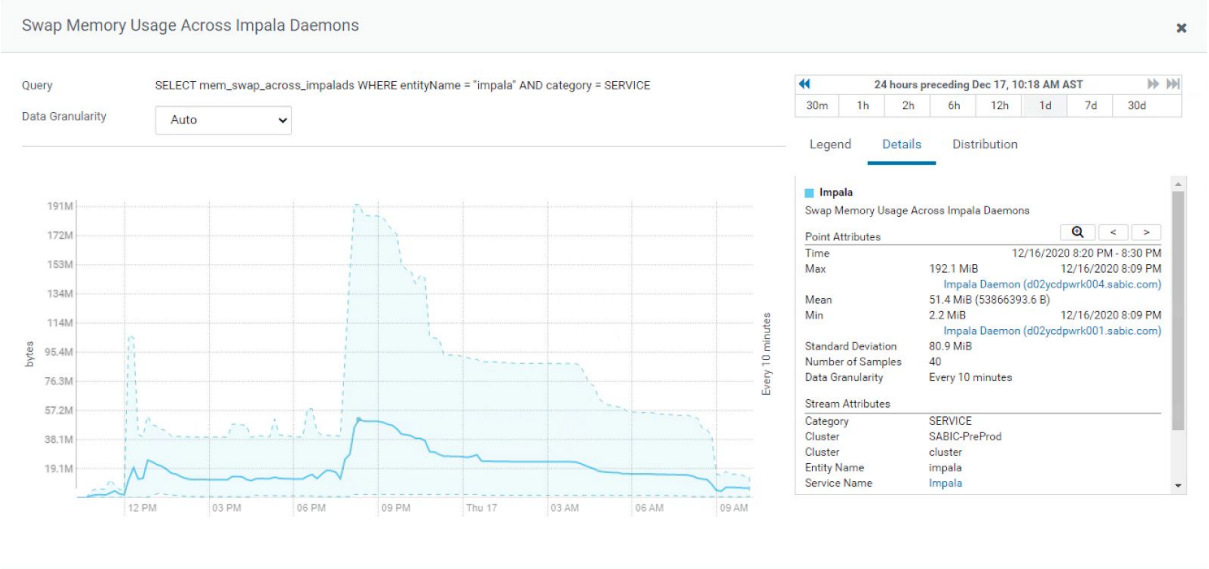


In the past 7 days, the memory usage of IMPALA-DAEMON is normally maintained at around 21GB to 25 GB (configured to be 128GB). Only one purge to almost the configured limits. No changes required.

NOTE that the resident memory here means the physical memory of the host, instead of the virtual memory.



In the past 7 days, the embedded JVM's heap memory usage is less than 10GB in average (configured to be 32GB in maximum), and the highest IMPALA-DAEMON instance amongst 4 instances is still less than 20GB. No change required.



Some nodes, like Worker 4, DO start using swap memory, however still less than 200MB in total. Recommend to observe.

4.7 SPARK review

	ITEMS	***CURRENT STATUS***
1	Enable Dynamic Allocation spark.dynamicAllocation.enabled	Enabled
2	Initial Executor Count spark.dynamicAllocation.initialExecutors	N/A

3	Minimum Executor Count spark.dynamicAllocation.minExecutors	0
4	Maximum Executor Count spark.dynamicAllocation.maxExecutors	N/A
5	ARROW_PRE_0_15_IPC_FORMAT	1

Comments:

- It is great that Spark dynamic allocation is enabled with default parameters, and also great to proactively use Apache Arrow as Python is used as the major language for the Spark environment.
- Since ARROW_PRE_0_15_IPC_FORMAT is set, it is largely because Apache Arrow 0.15 (or higher) is installed as well. The default Arrow version for Spark 2.4.x is 0.8.
    - To further ease application development, we can consider to further set the following Spark parameters directly in of CM:

      spark.sql.execution.arrow.pyspark.enabled=true

      spark.sql.execution.arrow.pyspark.fallback.enabled=true

4.8 SOLR review

4.8.1 Major configurations and deployment settings

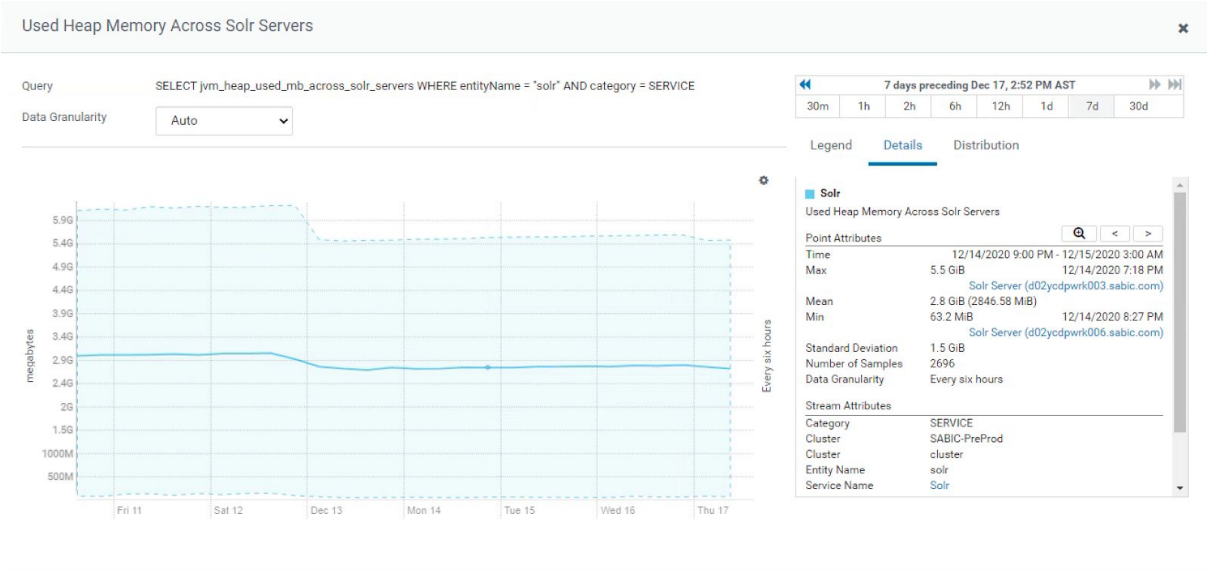
	ITEMS	*CURRENT STATUS*
1	Java Heap Size of Solr Server	32GB
2	Java Direct Memory Size of Solr Server	32GB
3	Solr Data Directory	/var/lib/solr
4	HDFS Block Cache solr.hdfs.blockcache.enabled	Enabled
5	HDFS Block Cache Off-Heap Memory solr.hdfs.blockcache.direct.memory.allocation	Enabled
6	HDFS Block Cache Blocks per Slab solr.hdfs.blockcache.blocksperban	16384
7	**HDFS Block Cache Number of Slabs solr.hdfs.blockcache.slab.count**	**32**

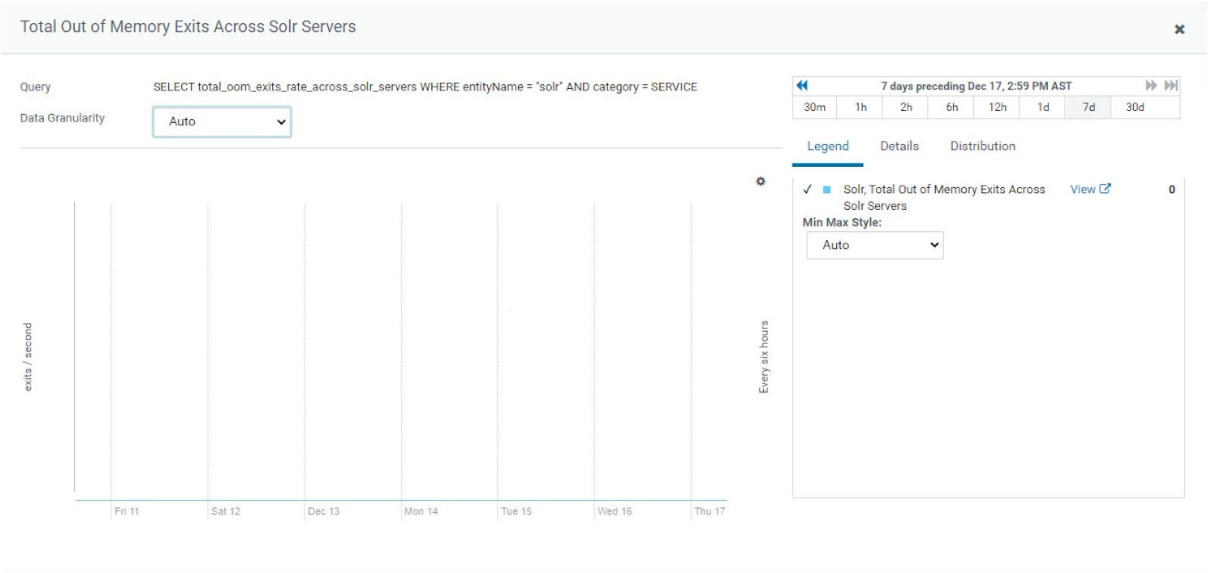
Issues:

- HDFS block cache number of slabs is set to far too small, which leads to inefficient direct memory usage, and thus very low cache hit ratio (see the next sub-section for details). It is highly recommended to set to 179 according to the practise experience formula below:

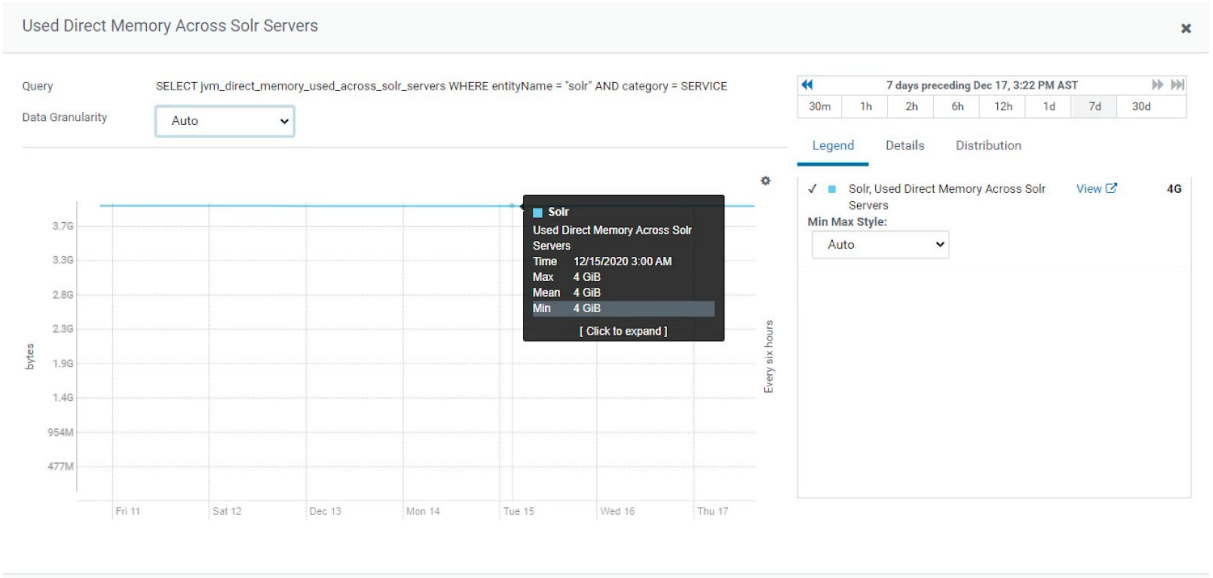
Slab count = direct memory size * 0.7 / 128M
SEE: https://blog.cloudera.com/apache-solr-memory-tuning-for-production/

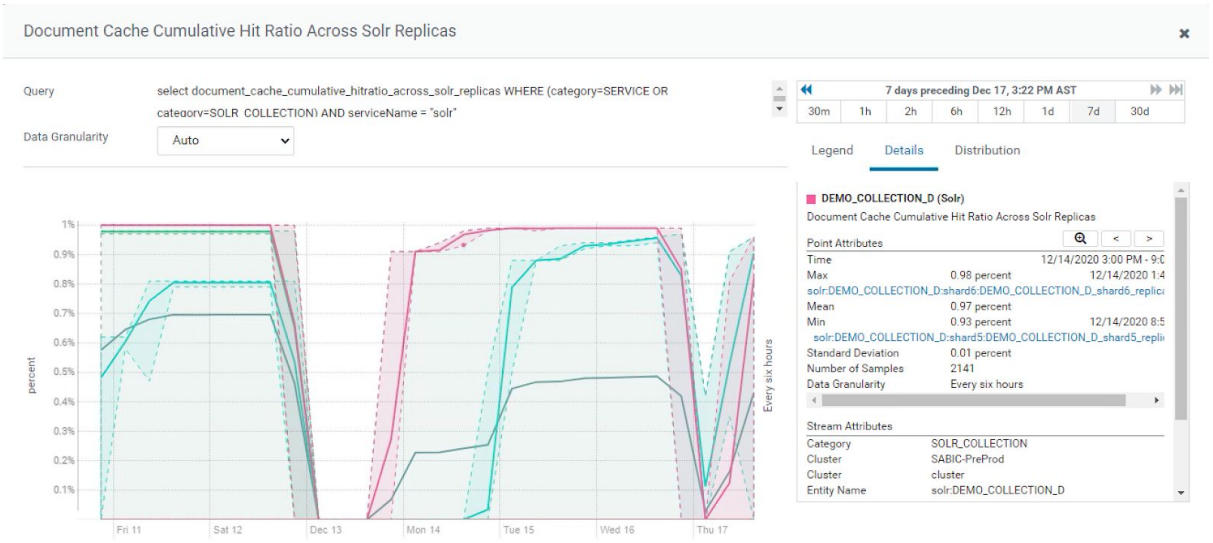
4.8.2 Major Solr health metrics





As is shown in the figures above, the SOLR JVM heapsize is fairly enough. The configured size is 32GB, while the real maximum usage is around 6GB. And there is no OOM occurred with SOLR JVM.





The cache area in SOLR direct memory is totally full. Its capacity is now configured as 4GB (= 8KB*16384*32), which directly results in a very low cache hit ratio of less than 1%. Recommended to be configured as 22GB (= 8KB*16384*179).

5 PROD Cluster Component Review

5.1 Role distribution and HA configuration for PROD

	Master-1 (d~mas001)	Master-2 (d~mas002)	Master-3 (d~mas003)	Edge 1-2	Worker 1-5	CDSW-1
CM-SERVER			X			
CMS			X			
CLOUDERA NAVIGATOR			X			
HDFS/NN	X	X				
HDFS/JN	X	X	X			
HDFS/DN					X	
HDFS/HTTPFS				X		
YARN/RM	X	X				
YARN/JHS			X			
YARN/NM					X	
ZK	X	X	X			
HBASE/MASTER	X	X	(Could make it the 3rd one)			
HBASE/RS					X	
SPARK			X			
HIVE METASTORE	X	X				
HIVE SERVER2	X	X				
IMPALA STATESTORE			X			

IMPALA CATALOG			X			
IMPALA DAEMON					X	
SOLR					X	
SENTRY	X	X				
KAFKA				X		
FLUME				X		
OOZIE			X			
HUE				X		
CDSW						X

Issues:

- Highly recommend not to build the single-node Kafka cluster. It is normally at least 3 nodes for a Kafka cluster, even in a pre-production environment. As the cluster will sometimes perform differently in a single-node cluster and a real-distributed cluster.

  It is recommended to deploy a 3-node Kafka cluster either in Master 1 - 3, or in Edge 1-2 and Master-3.

Concerns & recommendations:

- For the distribution balance of the master nodes, HBASE MASTER can be considered to be deployed in all of Master 1~3.
- OOZIE is better to be deployed in the edge node than the master node.

5.2 HDFS review

5.2.1 Major configurations and deployment settings

	ITEMS	***CURRENT STATUS***
1	HDFS HA	Enabled with QJM
2	HDFS HA Automatic Failover	Enabled
3	HDFS HA Fencing	Built-in HDFS fencing mechanism VIA shell(true)
4	Default Blocksize	128 MB
5	Default Replica	3
6	HDFS Total Space	176.2 TB
7	HDFS Used	16.91 TB (9.6%)
8	HDFS Trash	Enabled by default

		fs.trash.interval: 1 day fs.trash.checkpoint.interval: 1 hour
9	Namenode Heapsize	4GB
10	Namenode Heap Usage	Ranging from 1.3 - 2.5GB
11	Namenode GC	JAVA_GC_ARGS by default
12	**Namenode RPC Handler dfs.namenode.handler.count**	**32**
13	Datanode RPC Handler dfs.datanode.handler.counnt	3
14	Namenode Checkpoint dfs.namenode.checkpoint.period	1 hour
15	Namenode Leave Safemode dfs.namenode.safemode.threshold-pct	0.999
16	**FSIMAGE**	**Path (dfs.namenode.name.dir): /data/raid1_hdfs/dfs/nn** SIZE: 337MB
17	dfs.datanode.du.reserved	10GB
18	dfs.datanode.failed.volumes.tolerated	11 (22 disks in total)
19	dfs.client.read.shortcircuit	Enabled
20	dfs.datanode.balance.bandwidthPerSec	10MB

Issues:

- Although the disk of FSIMAGE is already protected by RAID 1, there is still a likelihood of logical error during writing FSIMAGE. Thus it is highly recommended to store FSIMAGE in more than 1 disk. As for now, we can simply add another disk path (either RAID-1 or JBOD) to dfs.namenode.name.dir.

Concerns:

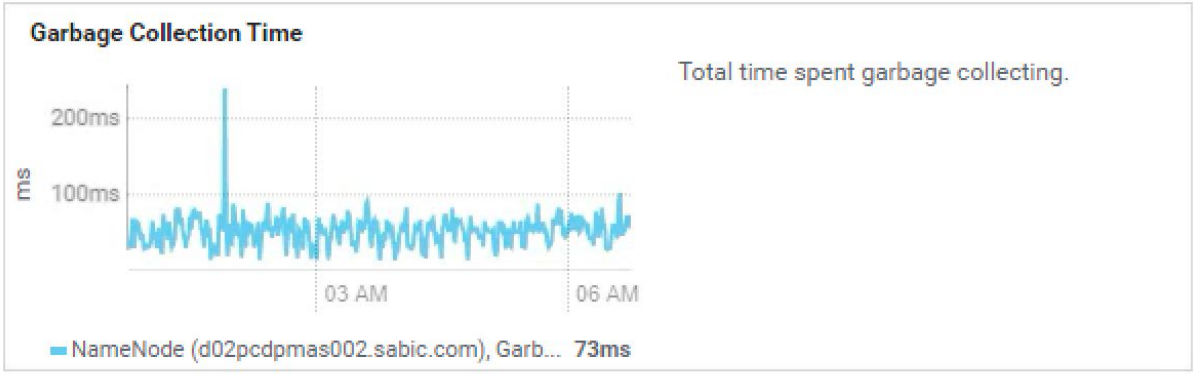
- The NAMENODE JVM heapsize of 4GB is a bit small, especially when the real usage is already more than 2GB during its peak time. And as is seen in the next section, there are already GCs and process pauses appearing in Namenode.

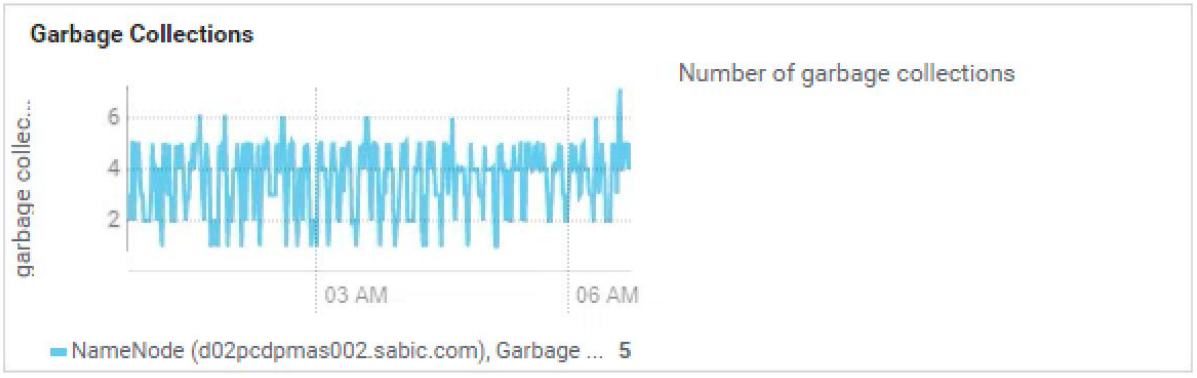
  Considering the further growth of data, the size of the cluster and the capacity of the whole disk volume, it can be set to 8GB.

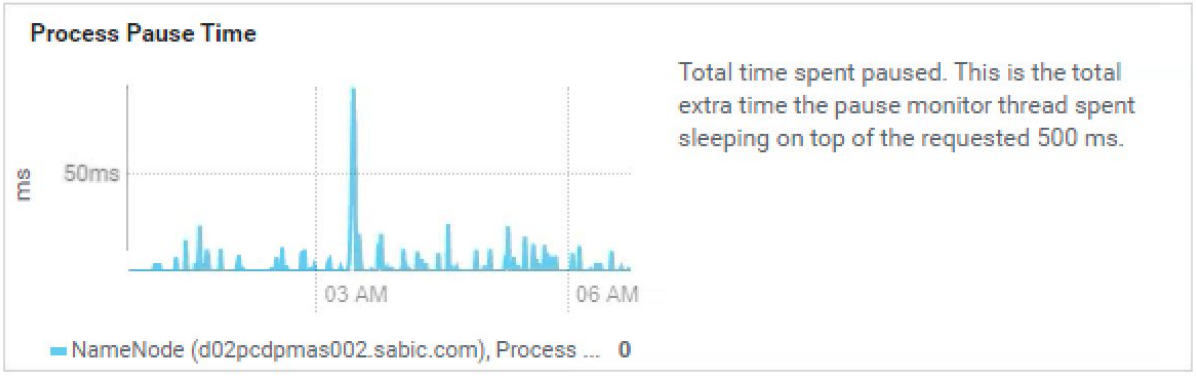
- Recommend to increase the Namenode RPC handler from 32 to 300. See RPC queue status for detail in the next sub-section.

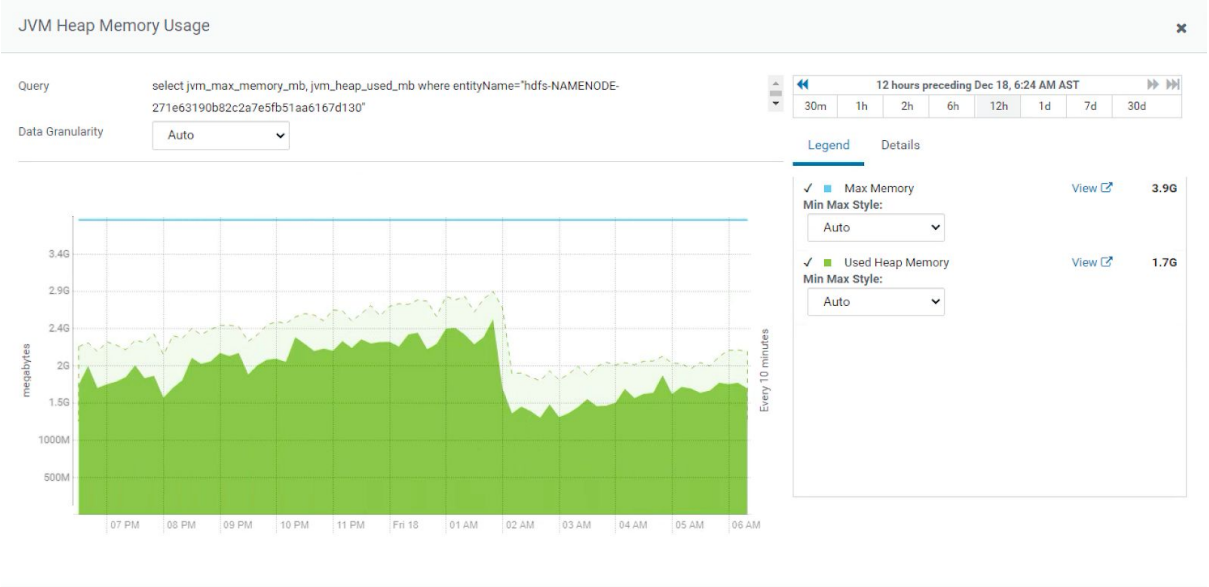
- Recommend to increase the Datanode RPC handler from 3 to 10.
- Recommend to decrease the tolerated failed volumes to a relative small number like 3 or 4.
- Recommend to strictly set the dfs.datanode.du.reserved to 15% - 20% of the local disk space.
- Recommend to increase dfs.datanode.balance.bandwidthPerSec to around 10% - 20% of the total server bandwidth. Note that the 10% - 20% value is estimated per the current network usage=

5.2.2 Namenode GC, process pause and swap



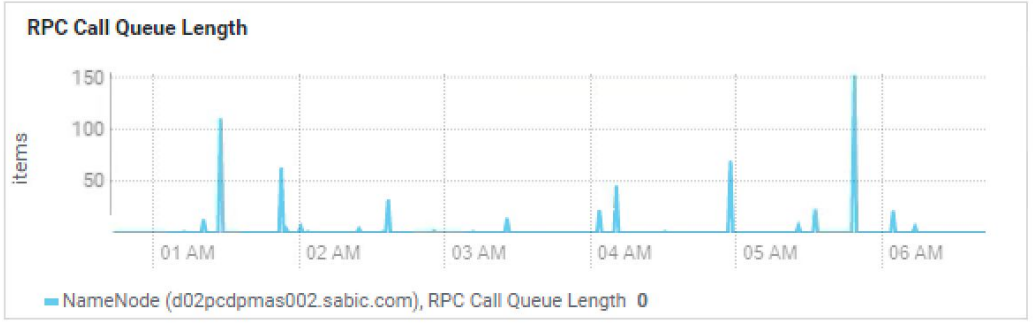


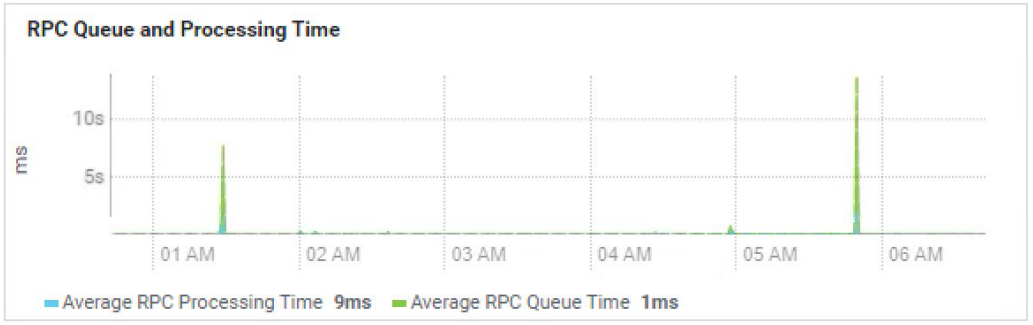




Concerns:

- Although the NAMENODE GC and process pause time is still within the 'normal' range. However, by considering the number of Datanode and the overall data volume, it is really a bit higher. Combining with the NAMENODE heapsize and the current usage, it can be tuned by configuring more heap space (8GB is recommended) for NAMENODE.



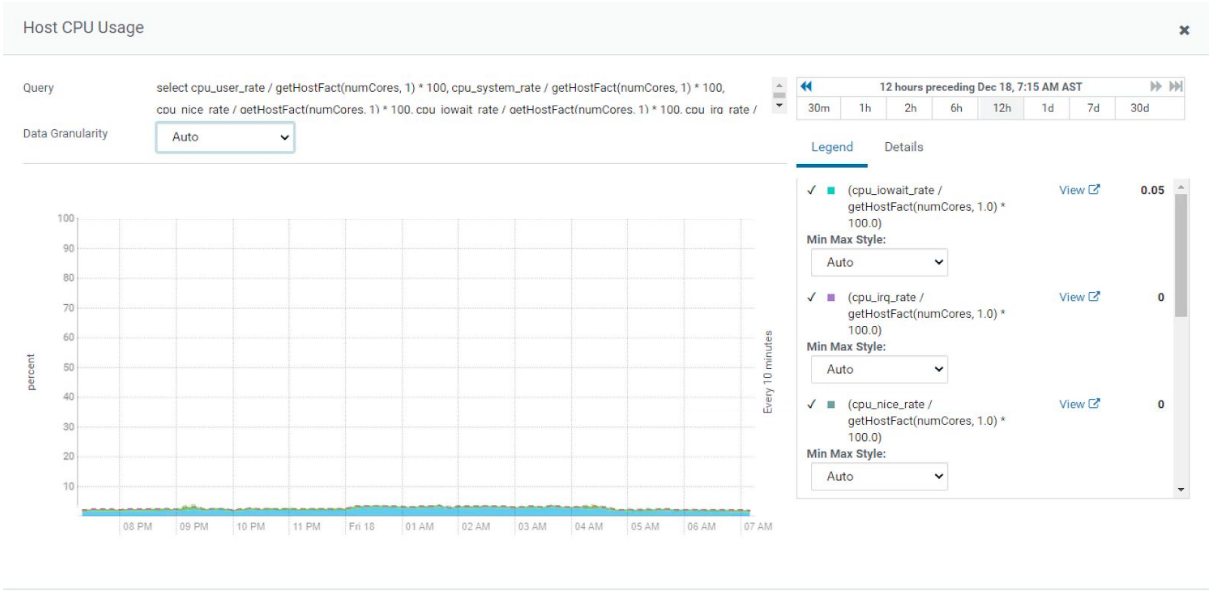


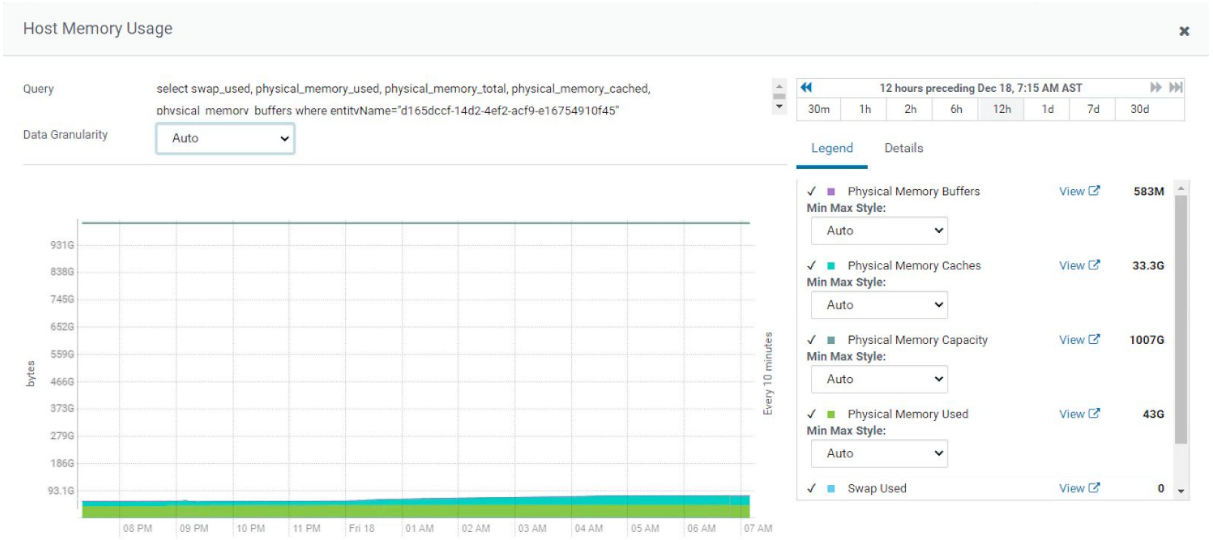
Issues:

- As seen from the above figures, both of the RPC queue length (150 in max) and queue time (12 sec in max) are high. Combining the Namenode RPC handler count setting (currently 32), it is highly recommended to tune the Namenode RPC handler count to 300 (greater than 150+32).

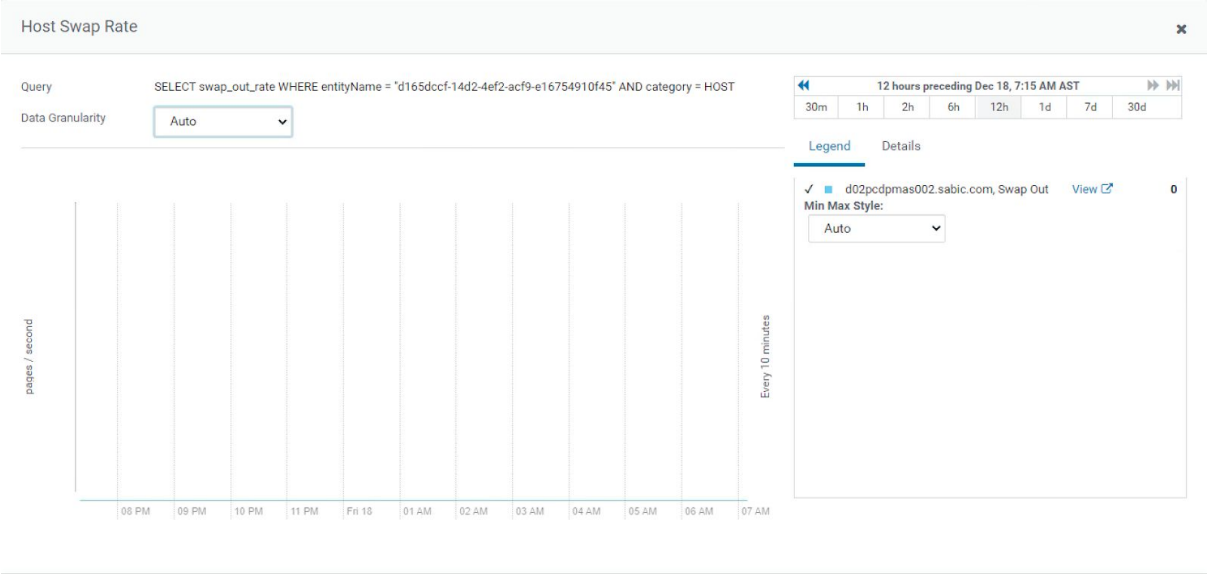
5.2.3 Namenode resource usage

Below is the Namenode resource usage diagram. Per this diagram, the host CPU and memory resource usage of Namenode is fairly low and, of course, totally within the acceptance range.



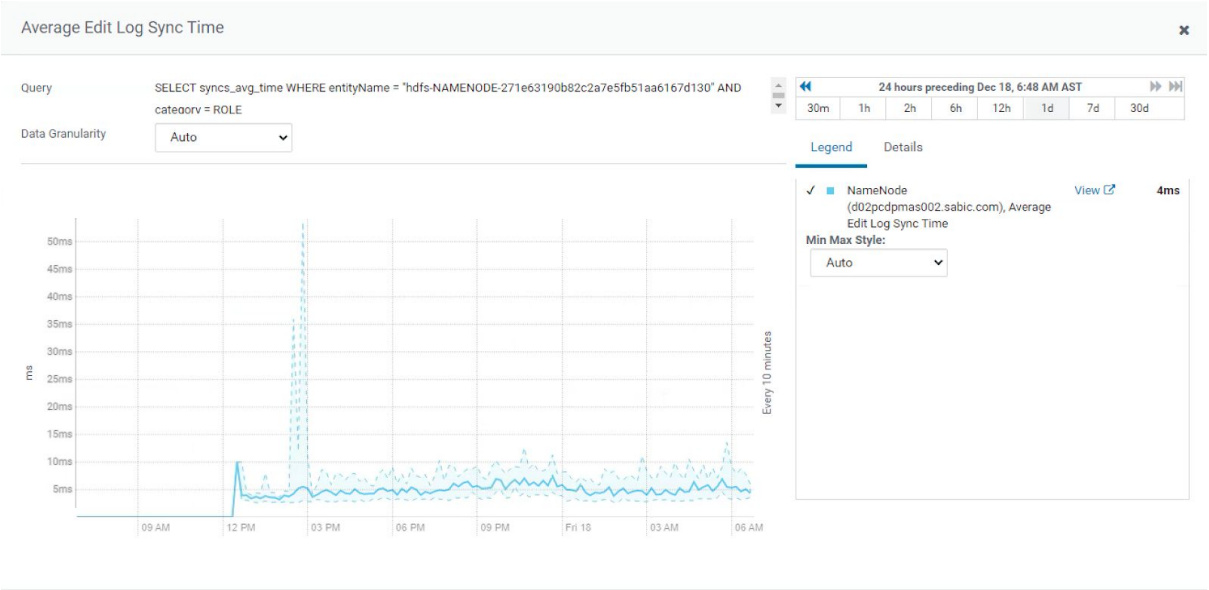


No host swap usage is detected.



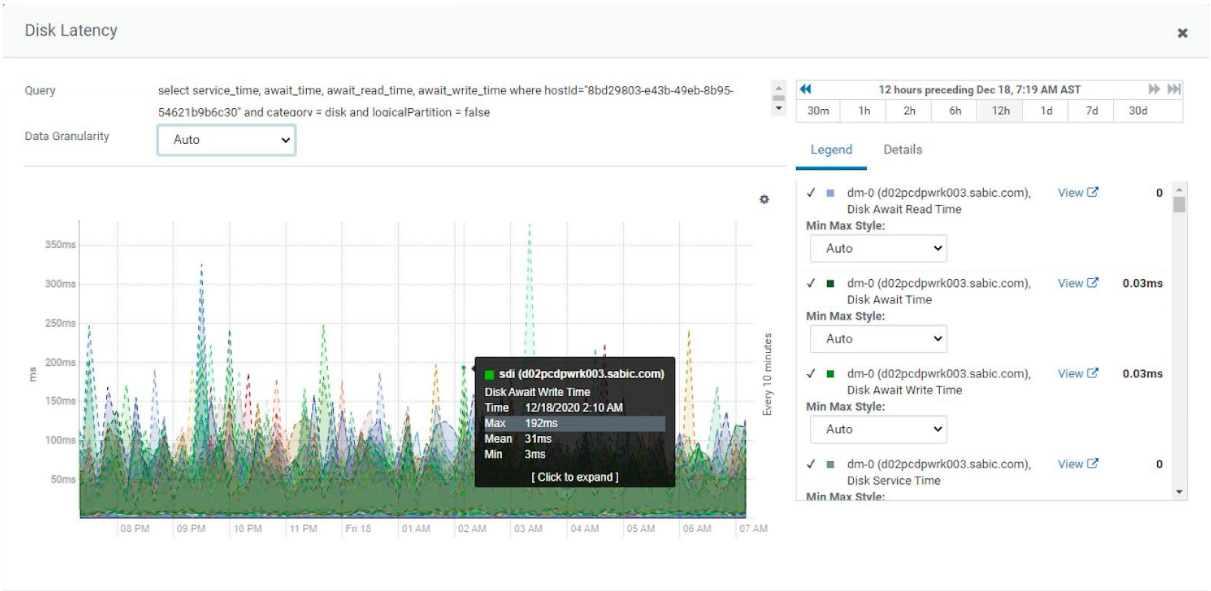
5.2.4 R/W performance

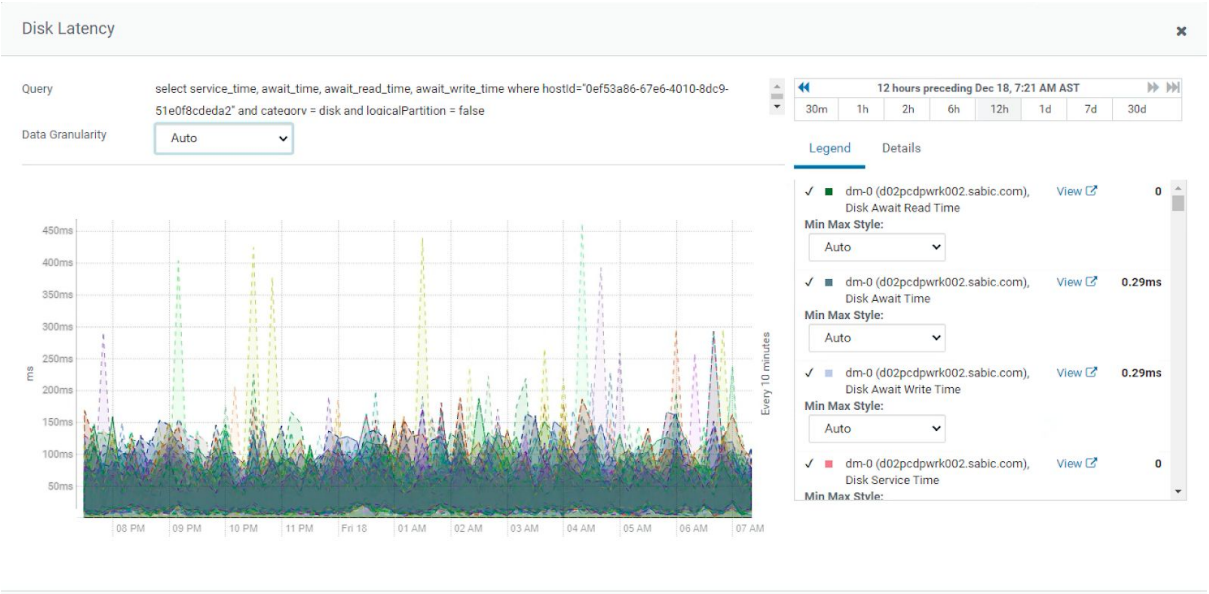
Below is the Editlog Sync time for Namenode. Its average time is around 5ms, and the maximum time is around 55ms in the past 1 day. Thus fairly good.



Below are the IO latency diagrams for Datanode 2 and 3. The concerns are:

- The disks' I/O latency is already at the high end. The peak latency time is ranging from 350ms to 450ms. Highly recommended to take care about the R/W performance of the upper-level services (like IMPALA, HIVE, HBASE, etc).





5.3 YARN and container allocation

5.3.1 Major configurations and deployment settings

	ITEMS	*CURRENT STATUS*
1	CPU shares per node (VCORES)	72 (with 36 physical cores for each node)
2	Memory shares per nodes (GB)	180GB
3	CPU allocation policy (VCORES)	Maximum: 72 Incremental: 1 Minimum: 1
4	Memory allocation policy (GB)	Maximum: 250GB Incremental: 0.5GB Minimum: 1GB

5	Resource allocation policy	DRF
6	Default CPU allocating shares for M/R (VCORES)	Map: 1 Reduce: 1
7	Default memory allocating shares for M/R (GB)	Map: 0 (Using 1GB by CM) Reduce: 0 (Using 1GB by CM)
8	Default JVM heapsize for M/R tasks (GB)	0.8
9	LOCAL-DIR yarn.nodemanager.local-dirs	Distributed at every local disks
10	Default compression for M/R tasks	Map output compression: SnappyCodec Job output compression: DISABLED
11	Speculative execution mapreduce.map.speculative mapreduce.reduce.speculative	DISABLED
12	mapreduce.task.io.sort.mb	256MB (Could be tuned according to default memory size for mapper and reducer)
13	mapreduce.task.io.sort.factor	64
14	mapreduce.map.sort.spill.percent	0.8
15	Slow starter of reducer mapreduce.job.reduce.slowstart.completedmaps	0.8

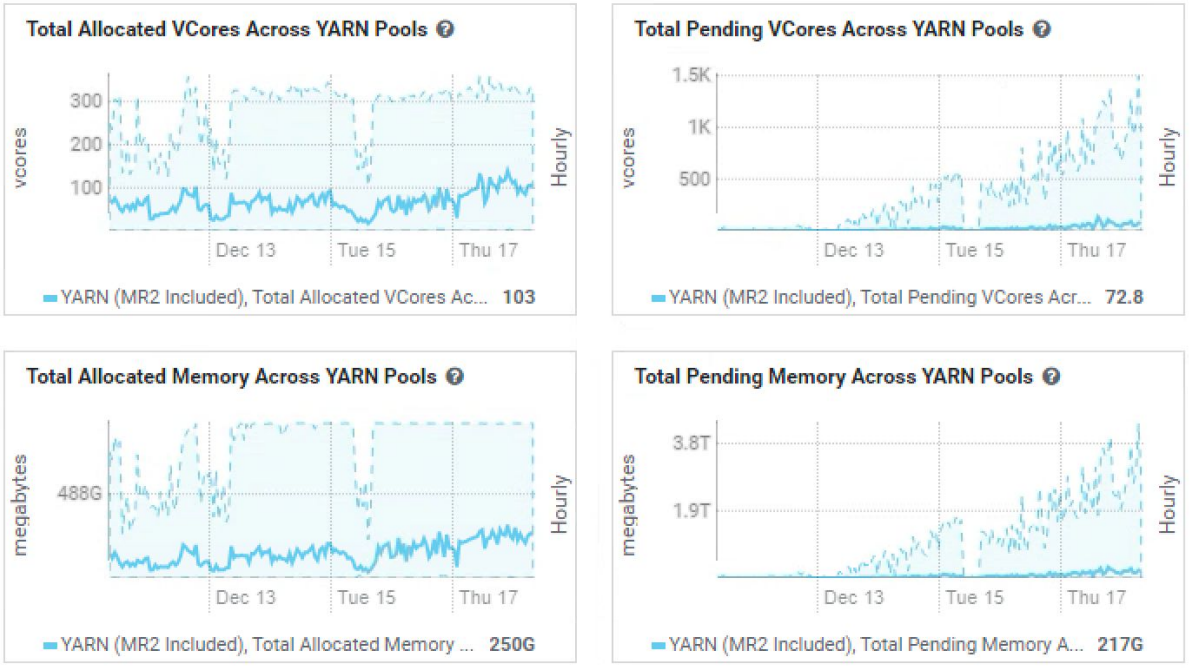
Concerns:

- Default memory allocating shares for M/R are now both set to 0. According to CM, it will be default to 1GB when setting 0. Recommended to be set to 2GB or 4GB.
- CPU VCORE and memory could be configured more to increase performance and resource utilization. Deeper investigation in the next subsection further shows that the current resource allocation configuration is not enough.
- It is NOT recommended to set the container's upper CPU and memory limit to all of the CPU and memory resources allocated to the host. This creates a risk that a M/R job is able to require a large-size container (like 60 VCORES and 300GB of memory), and make all of the other jobs and container requests pending.
    - Recommend to set the upper limit for CPU VCORE as 8, and memory as 16GB.

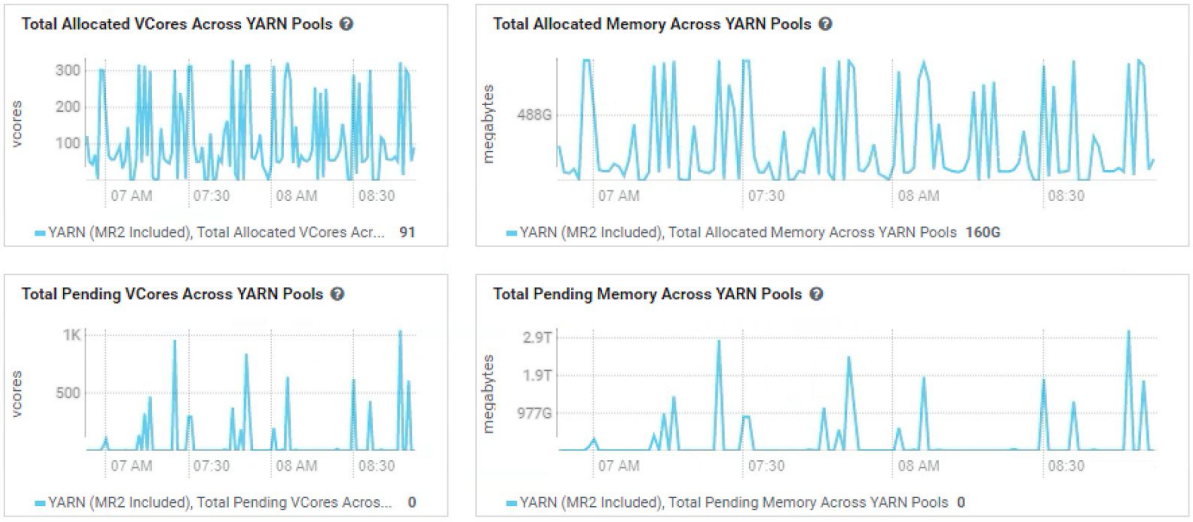
Observations:

- The sort-area memory (mapreduce.task.io.sort.mb) could be set roughly as 256MB for EACH 4GB.

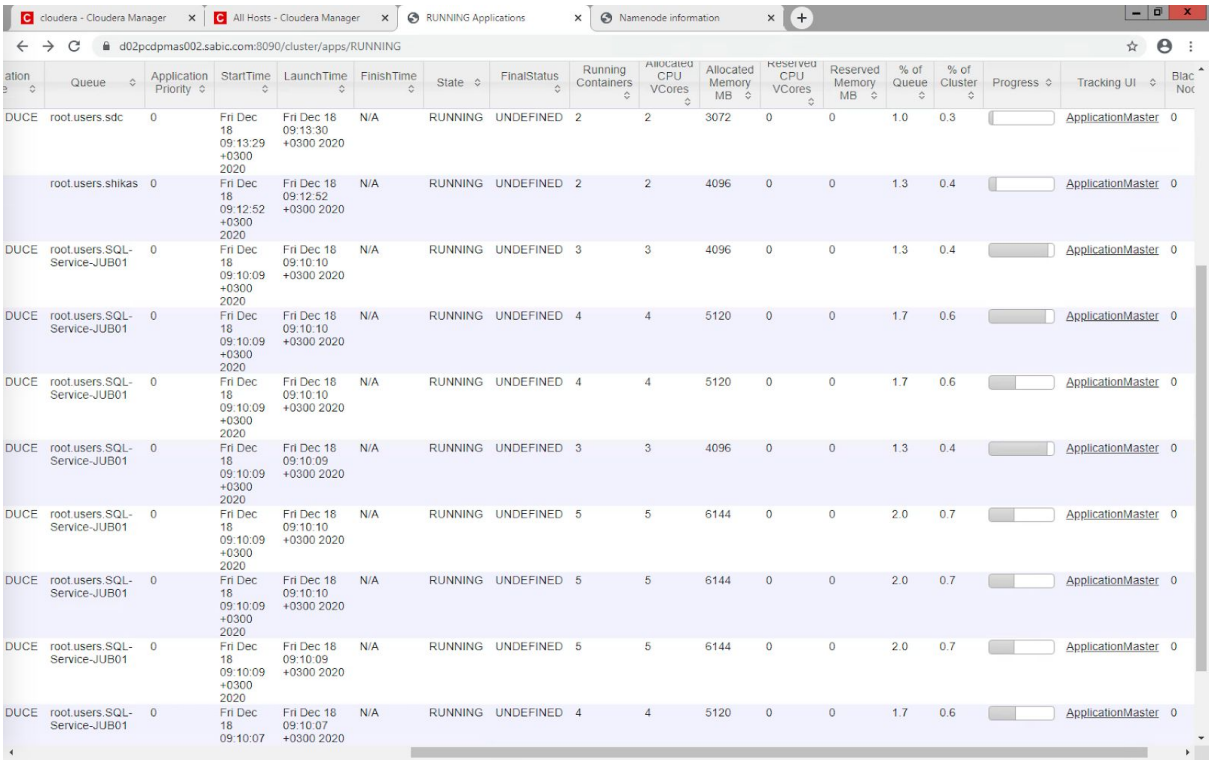
5.3.2 Overall load and health of YARN and MapReduce



The figures above show the YARN VCORE and memory allocated and pending in the past 7 days. And it can be seen that there are a lot of pendings in VCOREs and memory. Further investing the more detailed allocations and pending information (figures below) shows that the VCORE and memory pending appear at the same time.



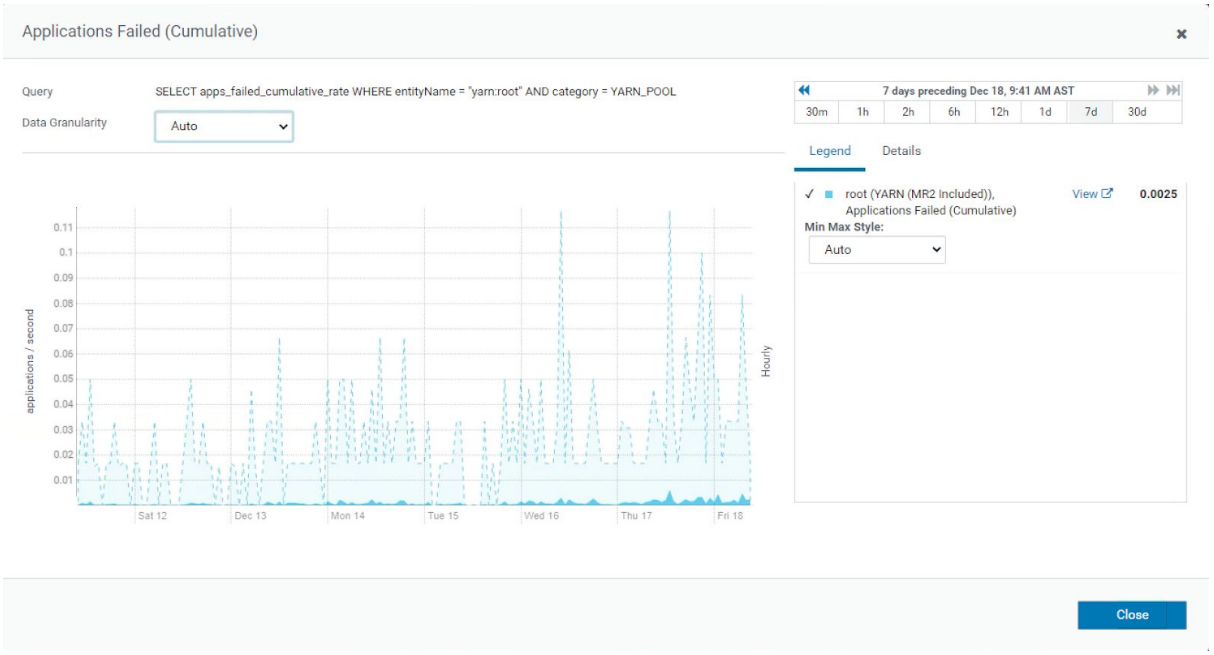
Meanwhile, by looking deeply into the YARN WEB CONSOLE below, we also see that there are no large-sized containers allocated. Thus we can conclude that the VCORE and memory resources are not configured enough.



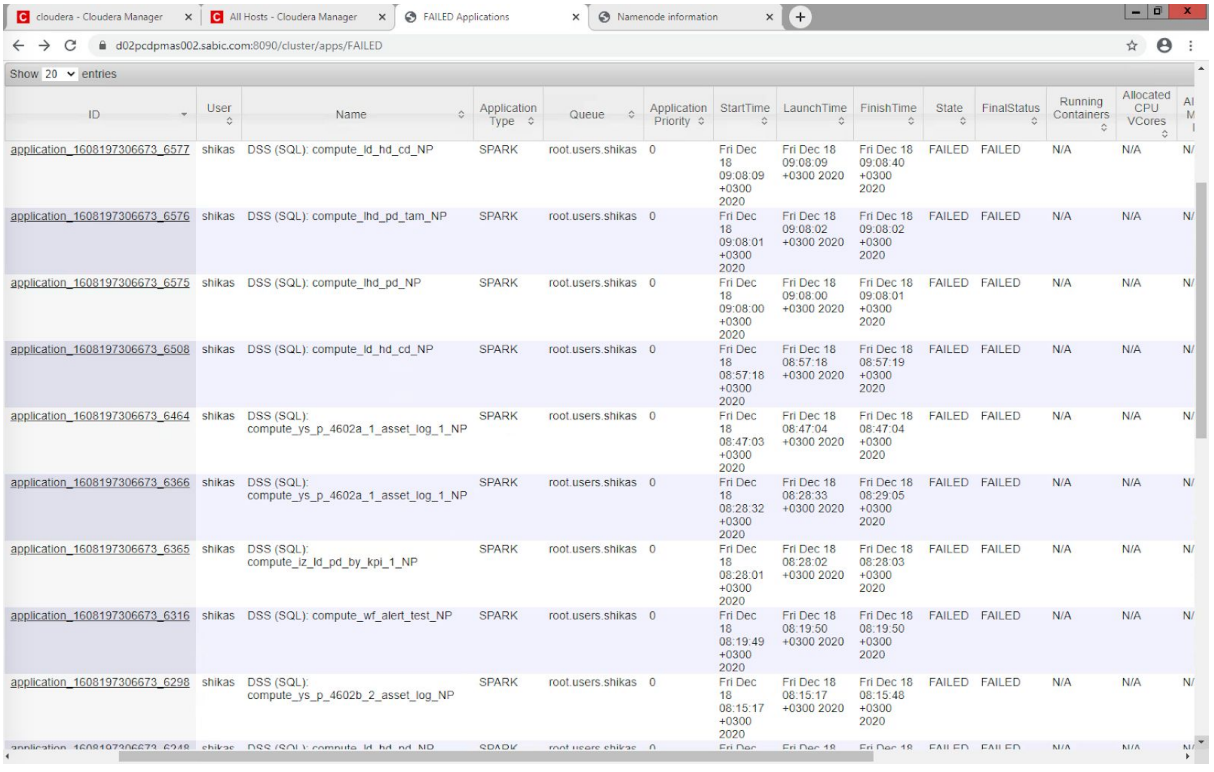
ation	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCores	Allocated Memory MB	Reserved CPU VCores	Reserved Memory MB	% of Queue	% of Cluster	Progress	Tracking UI	Blac Noc
DUCE	root.users.sdc	0	Fri Dec 18 09:13:29 +0300 2020	Fri Dec 18 09:13:30 +0300 2020	N/A	RUNNING	UNDEFINED	2	2	3072	0	0	1.0	0.3		ApplicationMaster	0
	root.users.shikas	0	Fri Dec 18 09:12:52 +0300 2020	Fri Dec 18 09:12:52 +0300 2020	N/A	RUNNING	UNDEFINED	2	2	4096	0	0	1.3	0.4		ApplicationMaster	0
DUCE	root.users.SQL-Service-JUB01	0	Fri Dec 18 09:10:09 +0300 2020	Fri Dec 18 09:10:10 +0300 2020	N/A	RUNNING	UNDEFINED	3	3	4096	0	0	1.3	0.4		ApplicationMaster	0
DUCE	root.users.SQL-Service-JUB01	0	Fri Dec 18 09:10:09 +0300 2020	Fri Dec 18 09:10:10 +0300 2020	N/A	RUNNING	UNDEFINED	4	4	5120	0	0	1.7	0.6		ApplicationMaster	0
DUCE	root.users.SQL-Service-JUB01	0	Fri Dec 18 09:10:09 +0300 2020	Fri Dec 18 09:10:10 +0300 2020	N/A	RUNNING	UNDEFINED	4	4	5120	0	0	1.7	0.6		ApplicationMaster	0
DUCE	root.users.SQL-Service-JUB01	0	Fri Dec 18 09:10:09 +0300 2020	Fri Dec 18 09:10:09 +0300 2020	N/A	RUNNING	UNDEFINED	3	3	4096	0	0	1.3	0.4		ApplicationMaster	0
DUCE	root.users.SQL-Service-JUB01	0	Fri Dec 18 09:10:09 +0300 2020	Fri Dec 18 09:10:10 +0300 2020	N/A	RUNNING	UNDEFINED	5	5	6144	0	0	2.0	0.7		ApplicationMaster	0
DUCE	root.users.SQL-Service-JUB01	0	Fri Dec 18 09:10:09 +0300 2020	Fri Dec 18 09:10:10 +0300 2020	N/A	RUNNING	UNDEFINED	5	5	6144	0	0	2.0	0.7		ApplicationMaster	0
DUCE	root.users.SQL-Service-JUB01	0	Fri Dec 18 09:10:09 +0300 2020	Fri Dec 18 09:10:09 +0300 2020	N/A	RUNNING	UNDEFINED	5	5	6144	0	0	2.0	0.7		ApplicationMaster	0
DUCE	root.users.SQL-Service-JUB01	0	Fri Dec 18 09:10:07	Fri Dec 18 09:10:07 +0300 2020	N/A	RUNNING	UNDEFINED	4	4	5120	0	0	1.7	0.6		ApplicationMaster	0

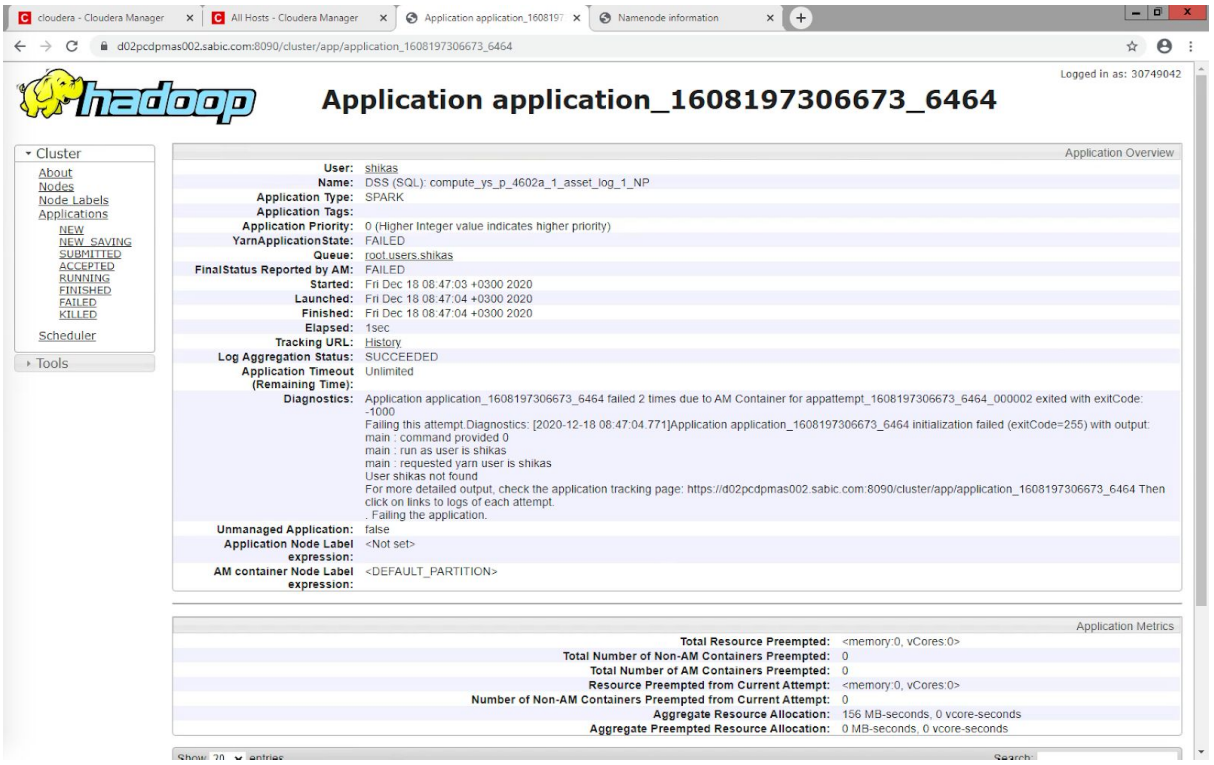
Thus it is recommended to adjust the YARN resource configuration as the following table. Also recommend to closely observe the change of CPU utilization, physical memory, and disk latency of all workers after adjustment.

	ITEMS	***CURRENT STATUS***
1	CPU shares per node (VCORES)	160 (with 36 physical cores for each node)
2	Memory shares per nodes (GB)	320GB
3	CPU allocation policy (VCORES)	Maximum: 8 Incremental: 1 Minimum: 1
4	Memory allocation policy (GB)	Maximum: 16GB Incremental: 0.5GB Minimum: 1GB



Failed applications are detectly constantly, as shown above.





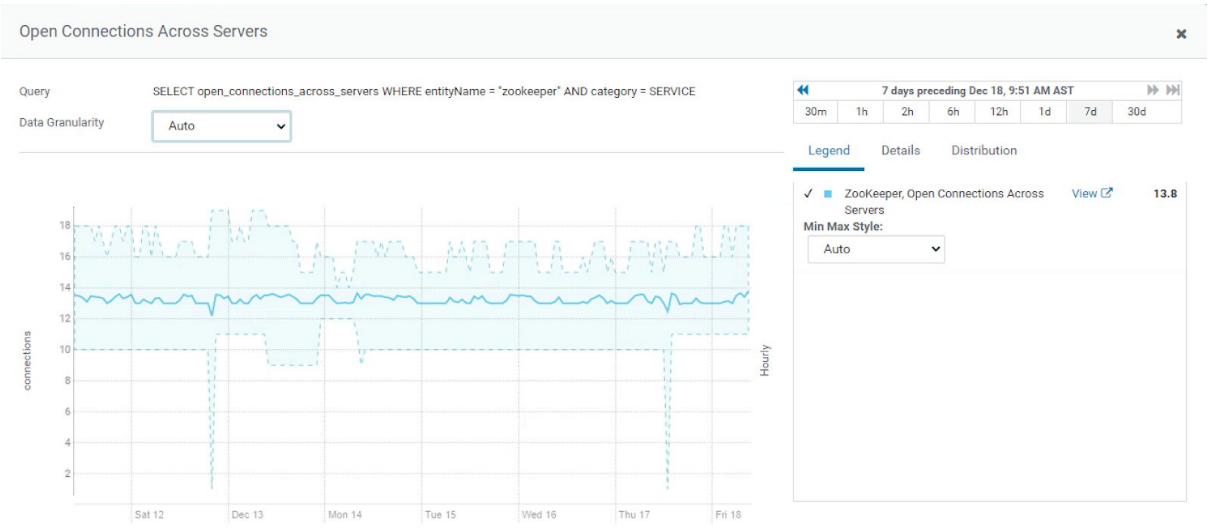
Further investigation shows that all the failed applications are the SPARK application submitted by the SAME user *shikas* with the SAME failure reason that the user does NOT exist.

5.4 ZOOKEEPER review

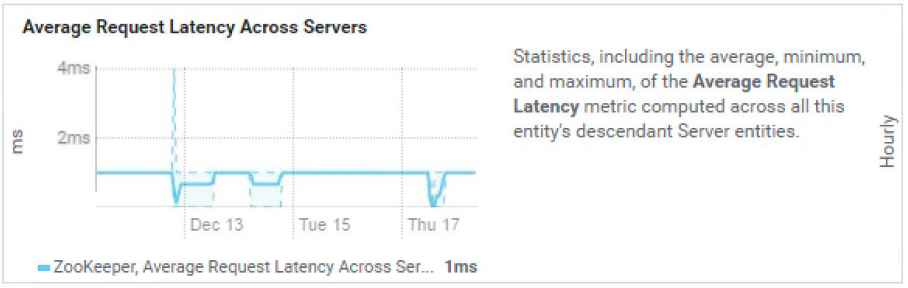
5.4.1 Major configurations and deployment settings

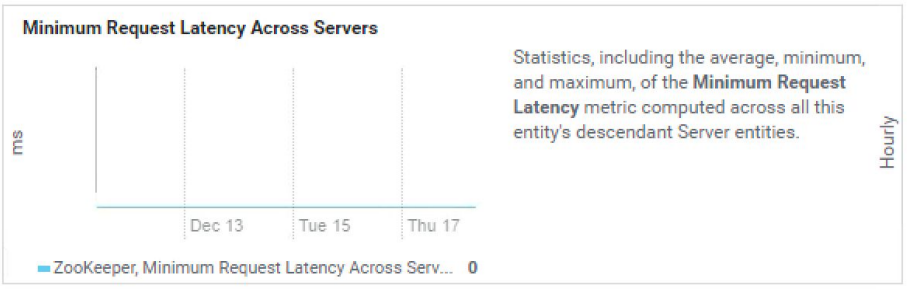
	ITEMS	*CURRENT STATUS*
1	Data directory	/data/01/zookeeper
2	Transaction log directory	/data/01/zookeeper
3	Maximum client connections	60 (Configure as 300 in QA cluster)
4	Maximum session timeout	60 sec
5	JVM Heapsize	2GB

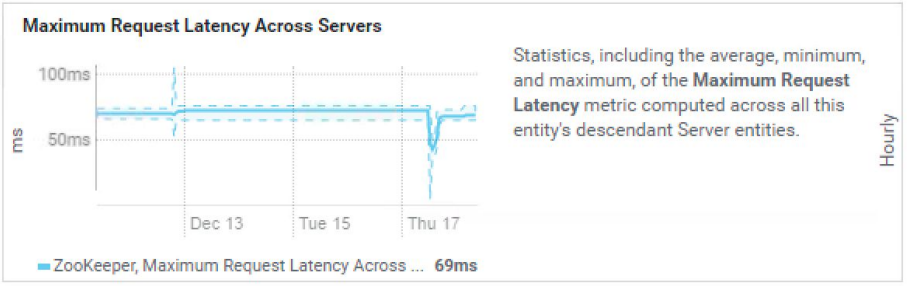
5.4.2 Major Zookeeper health metrics

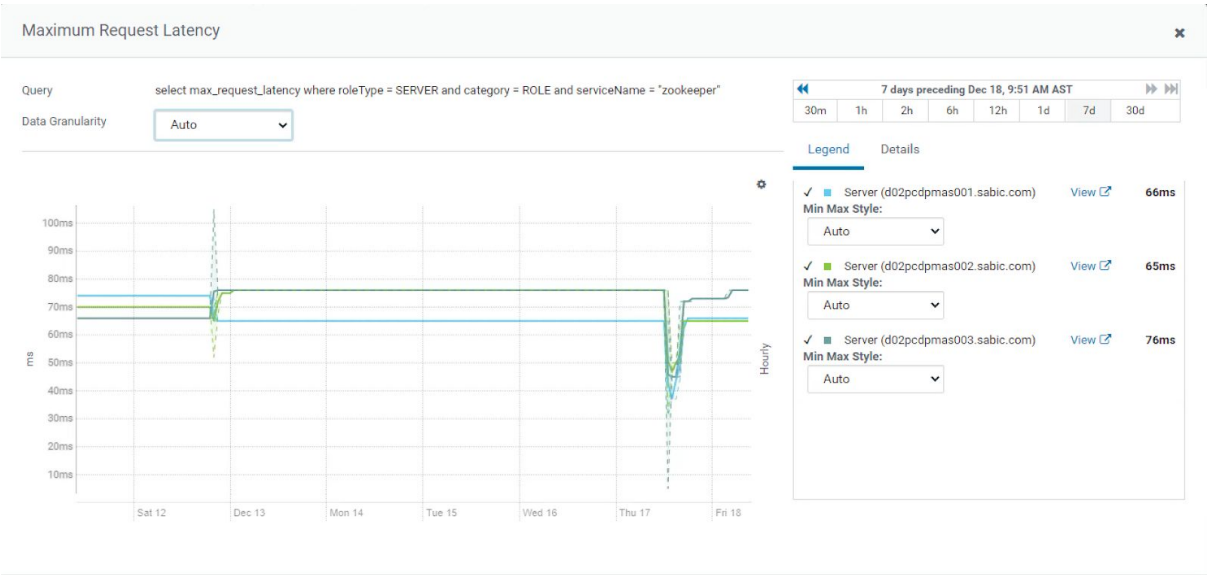


In the past 7 days, the largest number of open connections are around 20. Thus the configuration of 300 for this is enough for now. However, it is recommended to observe the connection number variation.

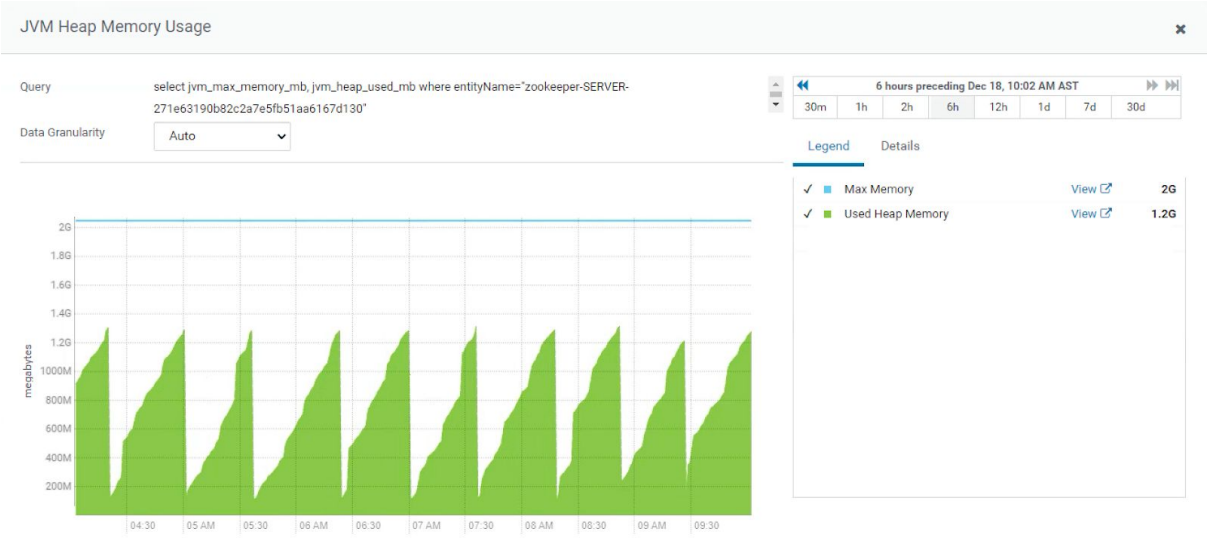








As shown above, although the average request latency is very good, the maximum request latency is a bit high but still good. And it is the same for all of the three zookeeper instances. Recommend for further observation. No action is required as long as it won't constantly go beyond 100ms.



The heap memory is using a bit faster but still good, as shown above. Increasing the heapsize from 2GB to 4GB could make it better.

5.5 HIVE review

5.5.1 Major configurations and deployment settings

	ITEMS	*CURRENT STATUS*
1	HIVE metastore heapsize	16GB
2	HIVE server2 heapsize	16GB
3	Kerberos and Impersonation	Disabled

4	Stored notification in database	Enabled
5	HIVE execution engine	MapReduce
6	Enable Stats Optimization hive.compute.query.using.stats	Disabled
7	Enable MapJoin Optimization hive.auto.convert.join	Enabled
8	hive.mapjoin.smalltable.filesize	25MB by default
9	Hive Auto Convert Join Noconditional Size hive.auto.convert.join.noconditionaltask.size	20MB
10	Enable Cost-Based Optimizer for Hive hive.cbo.enable	Disabled
11	Enable Vectorization Optimization hive.vectorized.execution.enabled	Enabled
12	Enable Stored Notifications in Database	Enabled

Observations:

- If encountering OOM during HIVE map-join, consider to decrease hive.mapjoin.smalltable.filesize.
    - This parameter can be tuned on HIVE session level.
- See the section of Sentry review for details about *Enable Stored Notifications in Database*.

5.5.2 Table and partition counts

- The number of HIVE databases:

```
mysql> select count(*) from DBS;
+----------+
| count(*) |
+----------+
|       29 |
+----------+
1 row in set (0.00 sec)
```

It does not go over the recommended upper limits of 100.

- The number of table in each database:

```
mysql> select TBLS.DB_ID,NAME,count(*) TAB_NUM from TBLS
join DBS on TBLS.DB_ID=DBS.DB_ID group by DB_ID,NAME order
by TAB_NUM;
+--------+----------------------------------+---------+
| DB_ID  | NAME                             | TAB_NUM |
+--------+----------------------------------+---------+
```

```
| 537116 | web_data                          |        1 |
| 108194 | meridium                          |        2 |
| 531494 | proc_st_refined                   |        2 |
| 367185 | sadaf                             |        2 |
| 108201 | lims                              |        3 |
| 423854 | petrokemya                        |        4 |
| 537578 | proc_mro_refined                  |        4 |
| 368711 | arrazi                            |        5 |
| 537577 | proc_mro_spd                      |        5 |
| 541136 | asset_healthcare                  |        5 |
| 395633 | ibn_zahr                          |        7 |
| 538156 | plant_efficiency_intermediate     |        7 |
| 531505 | proc_sm_refined                   |        7 |
| 466207 | dataiku_dss_test                  |        8 |
| 531501 | proc_pp_refined                   |       10 |
| 367211 | yanpet                            |       10 |
| 108202 | sap                               |       11 |
| 531506 | datapred                          |       11 |
| 367212 | yansab                            |       15 |
| 531508 | mro_spd                           |       29 |
|      1 | default                           |       35 |
| 534086 | manuf_pe_refined                  |       42 |
| 542373 | manuf_ahi_inter                   |       78 |
| 108195 | osipi                             |       99 |
| 541737 | manuf_pe_inter                    |     3203 |
+--------+-----------------------------------+----------+
25 rows in set (0.00 sec)
```

The number of tables for database manuf_pe_inter goes beyond the recommended upper limit of 1000. Attention is recommended to be paid for its metadata access performance.

6 The number of partitions for each table (top 30):

```
mysql> select TBLS.TBL_ID,TBL_NAME,count(*) as part_num
from PARTITIONS join TBLS on PARTITIONS.TBL_ID=TBLS.TBL_ID
group by TBL_ID,TBL_NAME order by part_num desc limit 30;
+--------+-----------------------------------+-----------+
| TBL_ID | TBL_NAME                          | part_num  |
+--------+-----------------------------------+-----------+
| 424088 | pk_olefin3_nrt_interpolated       |      1749 |
| 541978 | quench_105                        |       108 |
| 541423 | olf2_c3_16448b_200                |       100 |
| 541926 | olf2_c3_16448b_100                |       100 |
| 541927 | olf2_c3_16448b_300                |       100 |
| 540916 | olf2_c3_16431_300                 |       100 |
| 541916 | olf2_c3_16431_100                 |       100 |
| 541917 | olf2_c3_16431_400                 |       100 |
| 541918 | olf2_c3_16431_200                 |       100 |
| 541920 | olf2_c3_16431_500                 |       100 |
| 541434 | olf2_c3_comp_default              |        92 |
| 541939 | olf2_furn_common                  |        86 |
```

```
| 541921 | olf2_c3_16431_583                 |        83 |
| 541907 | olf2_butadiene_558                |        78 |
| 541976 | pk_olf2_monthly_partitioned_full  |        69 |
| 541479 | pk_olf1_monthly_partitioned_full  |        69 |
| 540874 | comp_30_2                         |        69 |
| 541934 | olf2_c3_16702                     |        63 |
| 541973 | olf2_seperator_180                |        60 |
| 541475 | olf2_seperator_240                |        60 |
| 541478 | olf2_seperator_60                 |        60 |
| 540904 | olf2_butadiene_60                 |        60 |
| 541880 | olf1_cgc_60                       |        60 |
| 541881 | olf1_cgc_240                      |        60 |
| 541399 | olf2_butadiene_300                |        60 |
| 541902 | olf2_butadiene_360                |        60 |
| 541401 | olf2_butadiene_180                |        60 |
| 541903 | olf2_butadiene_240                |        60 |
| 541404 | olf2_butadiene_480                |        60 |
| 541904 | olf2_butadiene_120                |        60 |
+--------+-----------------------------------+-----------+
30 rows in set (0.01 sec)
```

No table's partition number goes beyond the recommended upper limit of 10,000.

- The summary number of partitions for each database:

```
mysql> select DBS.DB_ID,NAME,count(*) as part_num from
PARTITIONS,TBLS,DBS where PARTITIONS.TBL_ID=TBLS.TBL_ID and
TBLS.DB_ID=DBS.DB_ID group by DB_ID,NAME order by part_num
desc limit 100;
+--------+----------------+----------+
| DB_ID  | NAME           | part_num |
+--------+----------------+----------+
| 541737 | manuf_pe_inter |     5669 |
| 423854 | petrokemya     |     1751 |
+--------+----------------+----------+
2 rows in set (0.00 sec)
```

No database's partition summary number goes beyond the recommended upper limit of 100,000.

6.1 IMPALA review

6.1.1 Major configurations and deployment settings

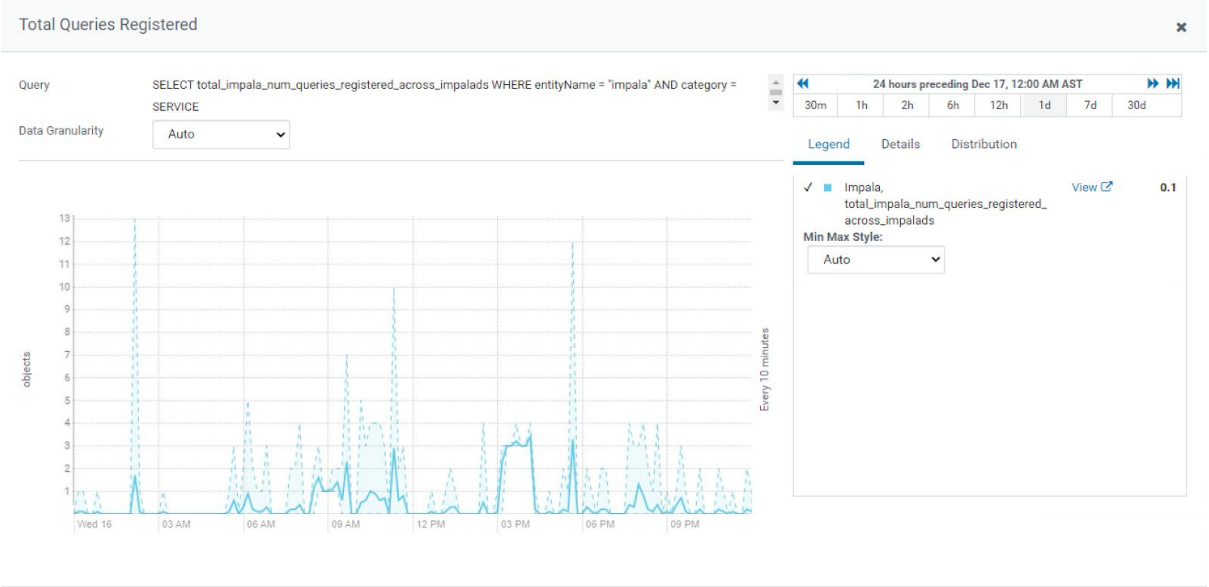
	ITEMS	*CURRENT STATUS*
1	IMPALA_DAEMON (mem_limit)	128GB
2	Embeded JVM HEAPSIZE	32GB
3	Impala Daemon Scratch Directories scratch_dirs	Every disk of the IMPALA-D nodes
4	Local UDF Library Dir	/var/lib/impala/udfs

	local_library_dir	
5	Result Cache Maximum Size max_result_cache_size	100000

Comments:

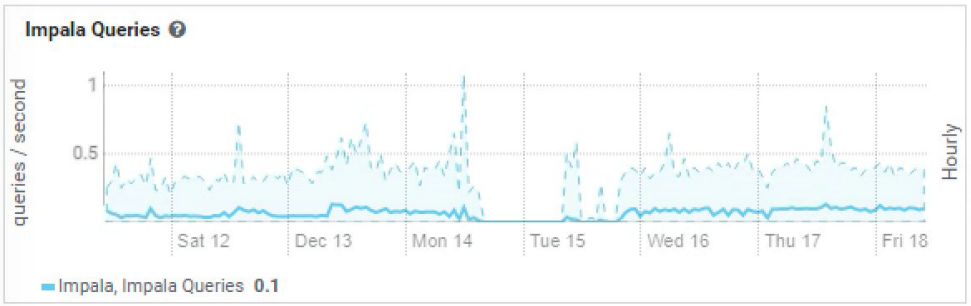
- As a friendly reminder, make sure to back up (or migrate) /var/lib/impala/udfs in case of losing your IMPALA user-defined functions.

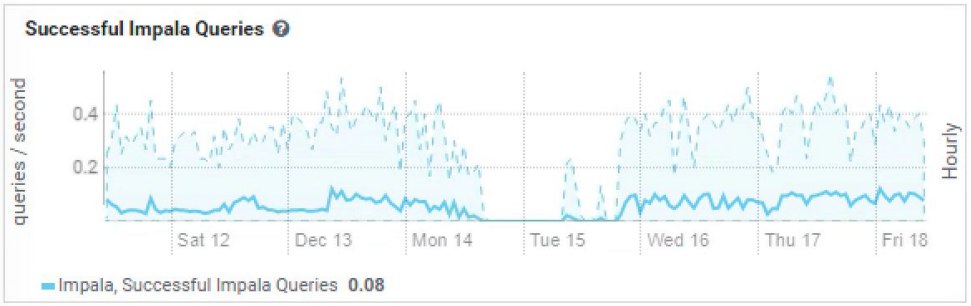
6.1.2 Workload and success rate

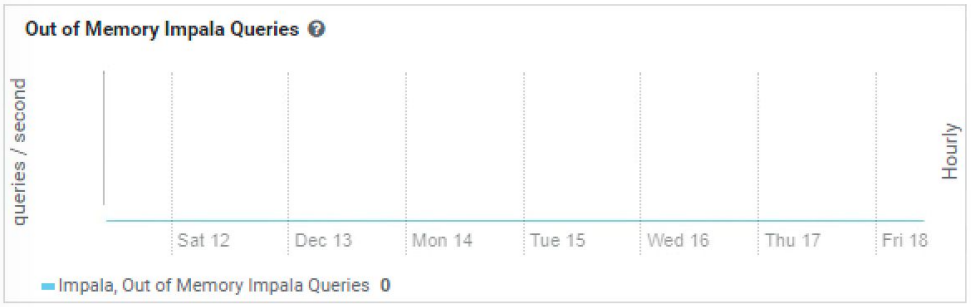


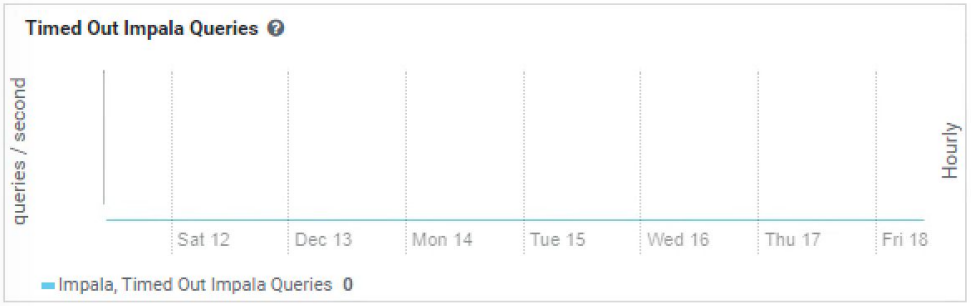
The figure above shows the total number of queries running in (or stopped but tracked by) IMPALA during a whole day (Wed 16 Dec). Unlike that of the QA cluster, it can be seen that:

- The overall concurrency is far less than that of the QA cluster, only 2 - 3 in general, and 19 in max.
- It is more evenly distributed than that of the QA cluster.



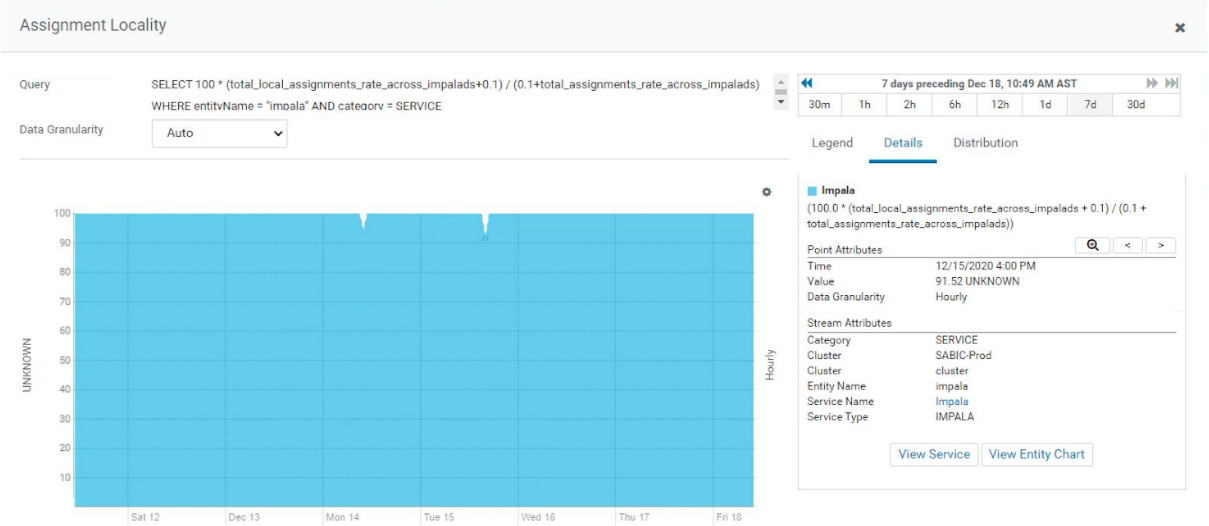






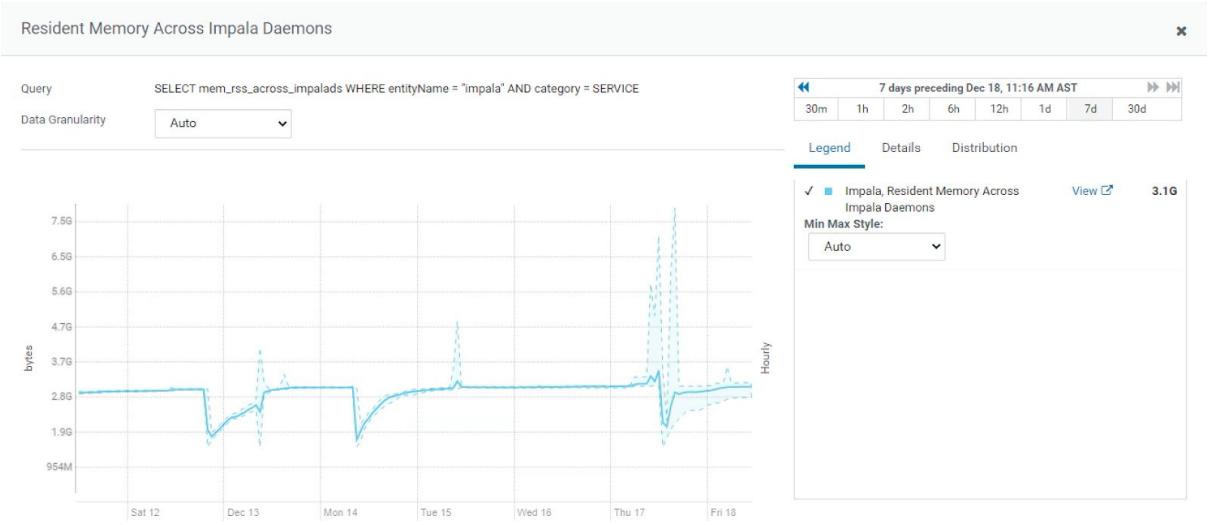
From the figures above, we can see that the IMPALA query incoming rate is around 0.1 query/sec, and the peak rate is only around 0.4 queries/sec. So the workload is not heavy (even less heavier than that of the QA cluster), and no failures are detected.

6.1.3 Locality of assignment



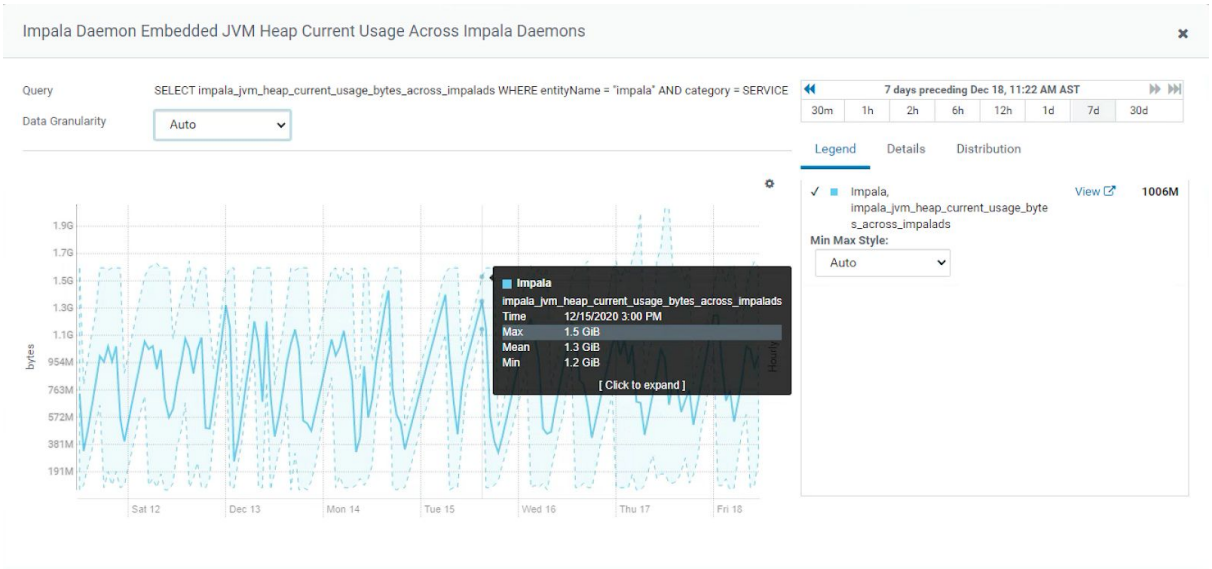
The assignments are almost 100% locally executed.

6.1.4 Memory usage of IMPALA-DAEMON



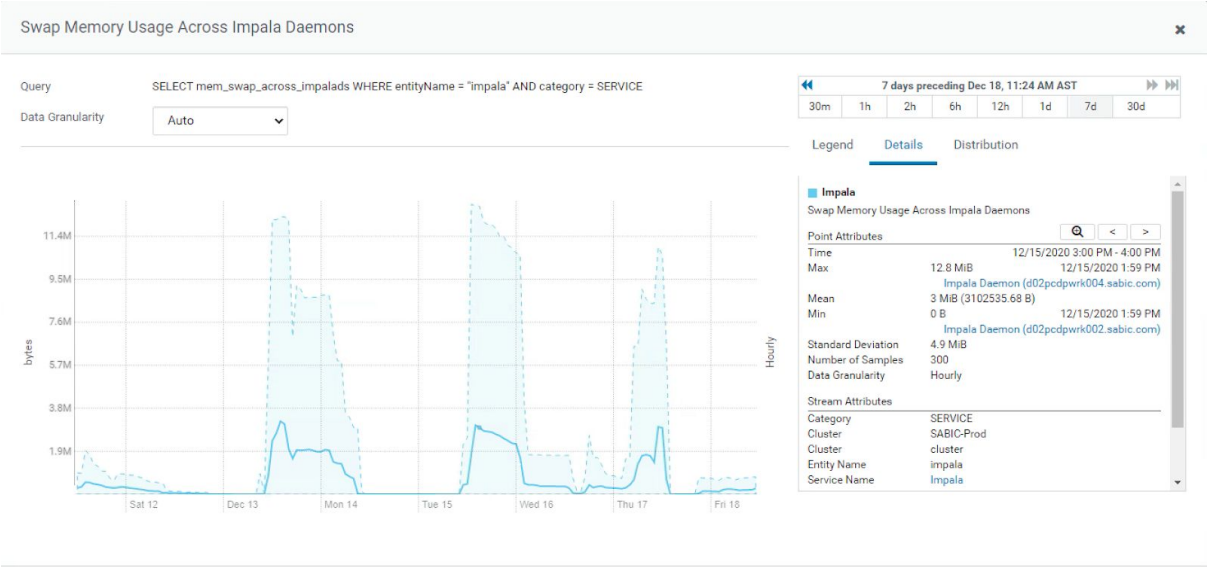
In the past 7 days, the memory usage of IMPALA-DAEMON is normally maintained at around 2 to 3.5 GB (configured to be 128GB), and 7.8 GB by maximum. No increment required.

NOTE that the resident memory here means the physical memory of the host, instead of the virtual memory.



In the past 7 days, the embedded JVM's heap memory usage is less than 1.5 GB in average (configured to be 32GB in maximum), and the highest IMPALA-DAEMON instance amongst 5 instances is still less than 2 GB. No increment required.

Considering the gap between the real usage and the configured capacity of IMPALA DAEMON memory and embedded JVM heap space, it is reasonable to reduce them in half, i.e. 64GB and 16GB respectively, when YARN memory is not enough.



In the past 7 days, no outstanding swap space usage has been detected, with no more than 4 MB in average, and 12.8 MB in max.

6.2 SPARK review

	ITEMS	***CURRENT STATUS***
1	Enable Dynamic Allocation spark.dynamicAllocation.enabled	Enabled
2	Initial Executor Count	N/A

	spark.dynamicAllocation.initialExecutors	
3	Minimum Executor Count spark.dynamicAllocation.minExecutors	0
4	Maximum Executor Count spark.dynamicAllocation.maxExecutors	N/A
5	ARROW_PRE_0_15_IPC_FORMAT	UNSET

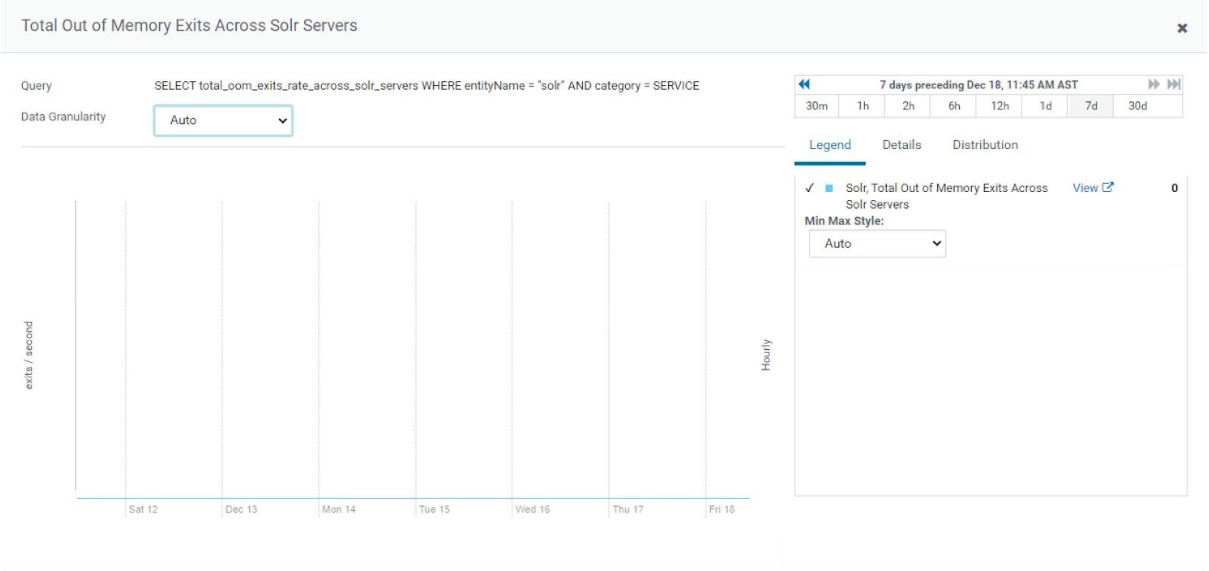
6.3 SOLR review (not in use?)

6.3.1 Major configurations and deployment settings

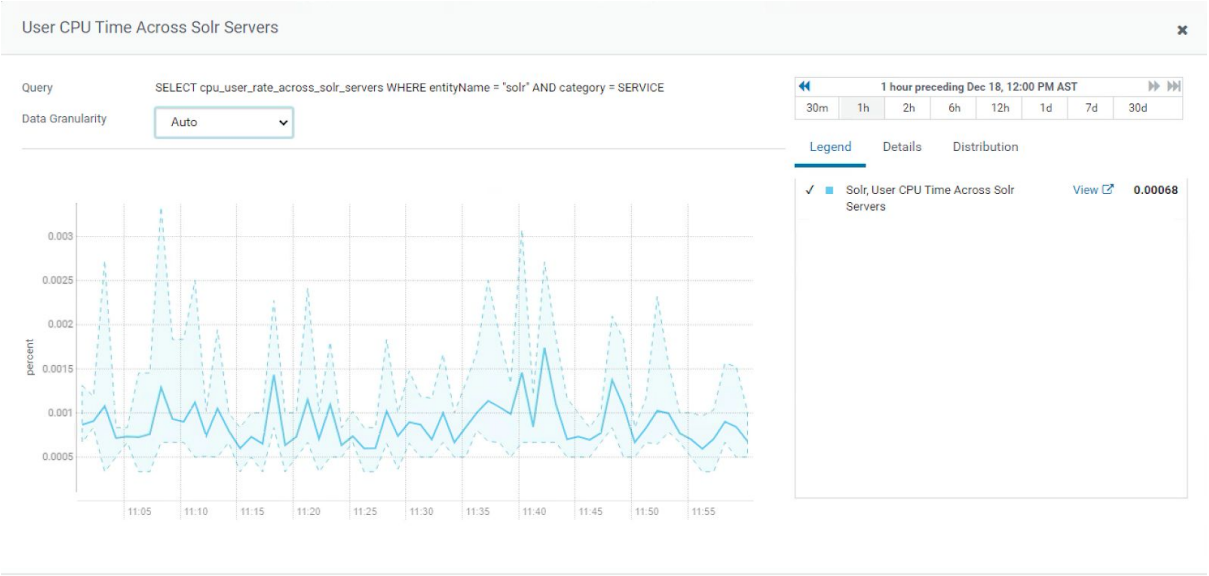
	ITEMS	*CURRENT STATUS*
1	Java Heap Size of Solr Server	16GB
2	Java Direct Memory Size of Solr Server	5GB
3	Solr Data Directory	/var/lib/solr
4	HDFS Block Cache solr.hdfs.blockcache.enabled	Enabled
5	HDFS Block Cache Off-Heap Memory solr.hdfs.blockcache.direct.memory.allocation	Enabled
6	HDFS Block Cache Blocks per Slab solr.hdfs.blockcache.blocksperban	16384
7	HDFS Block Cache Number of Slabs solr.hdfs.blockcache.slab.count	32

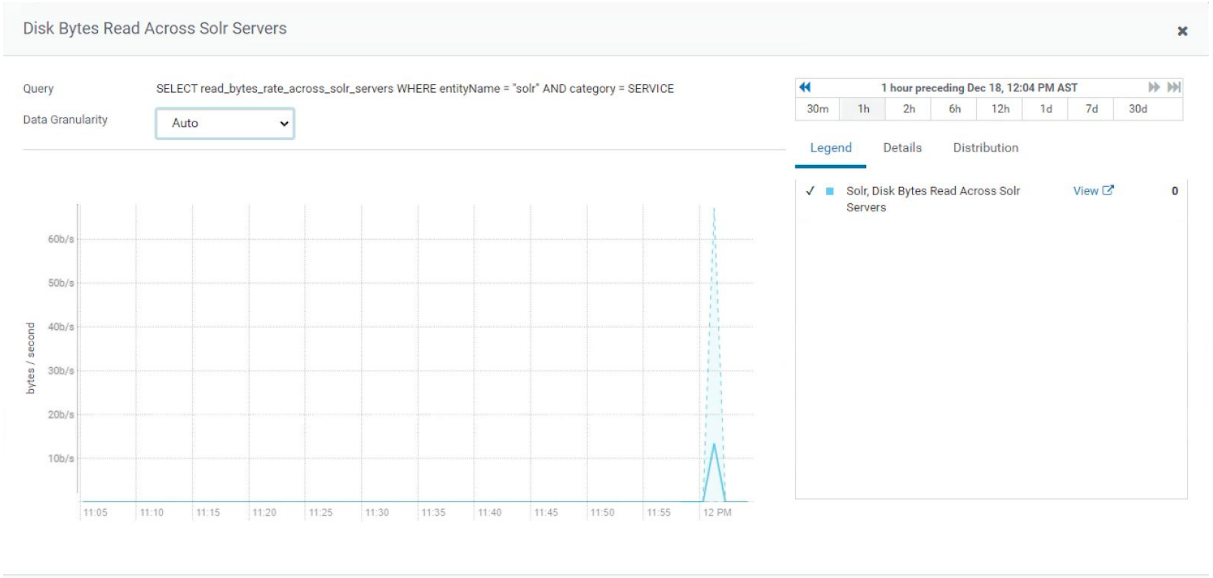
6.3.2 Major Solr health metrics

There is no OOM occurred with SOLR JVM, thus don’t have to worry about the JVM heap size configuration.



For this Solr cluster, a lot of metrics cannot be collected, like 'Used Direct Memory Across Solr Servers'. However, as seen from other visible metrics, the Solr service seems not in use now, as its CPU utilization is 0.001% (almost none), and disk read rate is 0 for most of time, and occasionally 10 bytes / sec.





7 Data Ingestion Flow Review

7.1 QA cluster (?)

7.2 PROD cluster

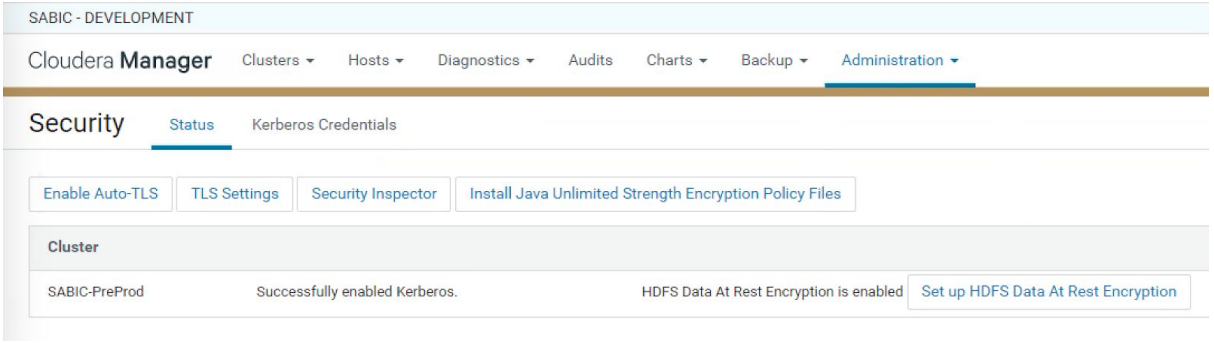
(To be added)

8 Security Configuration Review for QA

8.1 Kerberos & LDAP

8.1.1 Kerberos enablement and /etc/krb5.conf

By checking CM → Administration → Security → Status, it can be seen that Kerberos is enabled in the QA cluster.



Further check with the client-side Kerberos configuration file /etc/krb5.conf, as shown below.

```
[hadoop@d02ycdpmas001 ~]$ cat /etc/krb5.conf
[libdefaults]
default_realm = SABICCORP.SABIC.COM
dns_lookup_kdc = false
dns_lookup_realm = false
ticket_lifetime = 86400
renew_lifetime = 604800
forwardable = true
default_tgs_enctypes = aes256-cts aes128-cts
default_tkt_enctypes = aes256-cts aes128-cts
permitted_enctypes = aes256-cts aes128-cts
udp_preference_limit = 1
kdc_timeout = 3000
[realms]
SABICCORP.SABIC.COM = {
kdc = SS-JHQ-MEA-DC20.SABICCORP.SABIC.COM
admin_server = SS-JHQ-MEA-DC20.SABICCORP.SABIC.COM
}
[domain_realm]
.sabic.com = SABICCORP.SABIC.COM
```

As only one KDC address is configured in /etc/krb5.conf, make sure the KDC (or Active Directory) itself has configured with the HA mechanism. That is said, when the master KDC is halted, its IP address can be switched to the standby node automatically.

8.1.2 LDAP client configuration at OS and /etc/openldap/ldap.conf

```
[hadoop@d02ycdpmas001 ~]$ cat /etc/openldap/ldap.conf
```

```
#
# LDAP Defaults
#

# See ldap.conf(5) for details
# This file should be world readable but not world writable.

#BASE   dc=example,dc=com
#URI    ldap://ldap.example.com ldap://ldap-master.example.com:666

#SIZELIMIT      12
#TIMELIMIT      15
#DEREF          never

TLS_CACERTDIR   /etc/openldap/certs

# Turning this off breaks GSSAPI used with krb5 when rdns = false
SASL_NOCANON    on
```

8.1.3 LDAP integration in OS and /etc/sssd/sssd.conf

```
[sssd]
domains = SABICCORP.SABIC.COM
config_file_version = 2
services = nss, pam

[domain/SABICCORP.SABIC.COM]
ad_domain = SABICCORP.SABIC.COM
krb5_realm = SABICCORP.SABIC.COM
realmd_tags = manages-system joined-with-samba
cache_credentials = True
id_provider = ad
krb5_store_password_if_offline = True
default_shell = /bin/bash
ldap_id_mapping = True
use_fully_qualified_names = False
fallback_homedir = /home/%u@%d
access_provider = ad
ad_access_filter
=(&(memberOf=CN=SABIC_Big-Data_CDH_Admins-Qlty,OU=Groups-Qlty,OU=BigData-Qlty,
OU=Apps,DC=SABICCORP,DC=SABIC,DC=com))
enumerate = false
ignore_group_members = true
ldap_schema =ad
ldap_force_upper_case_realm = true
case_sensitive = false
```

As neither ldap_uri nor ldap_backup_uri is specified in sssd.conf, the service discovery mechanism is used to find LDAP servers. And by checking user information with the 'id' command, it can output successfully with the required information.

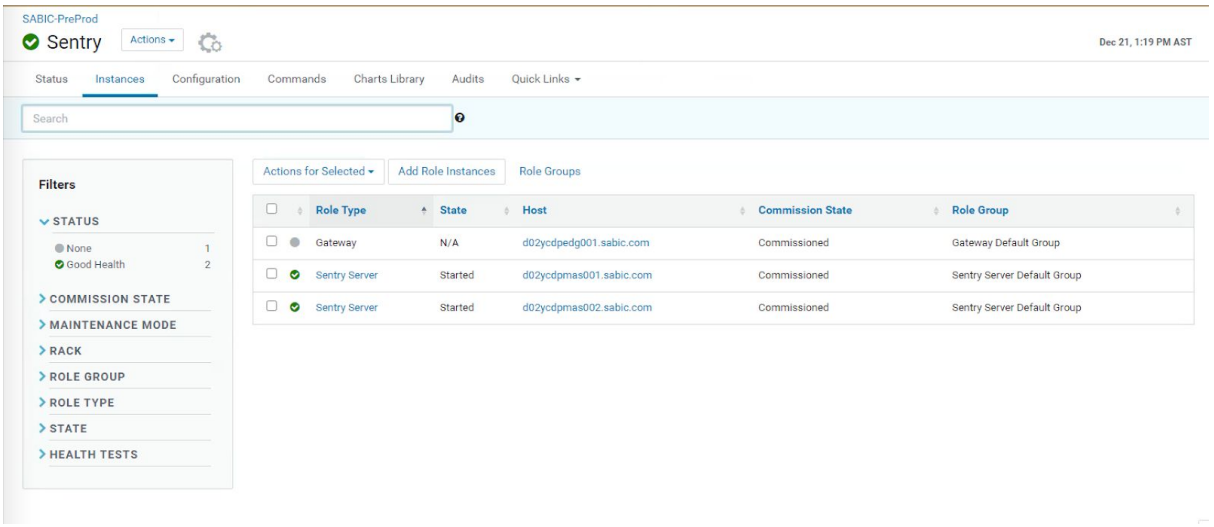
```
[root@d02ycdpmas001 sssd]# id 30749042
uid=12915377(30749042) gid=266800513(domain users) groups=266800513(domain users),1625560062(_sabic azure aip user license),1281
4492(sabic_big-data_cm_admins-qlty),1625544686(count_citrix_groups),266809684(_sabic mea sabicju all users),12889086(sa02-vmpc00
096_rdpadmins),12814208(sabic_big-data_sntry_admins),12814205(sabic_big-data_hue_admins),12814498(sabic_big-data_hue_users-qlty)
,12827936(sabic_big-data_cdsw_admins-qlty),12814500(sabic_big-data_nav_admins-qlty),1625535286(eur-car_maintenance_lxlmechanical
_condition-monitoring_rx),12814499(sabic_big-data_key_admins-qlty),12814497(sabic_big-data_hue_admins-qlty),12814206(sabic_big-d
ata_key_admins),12814207(sabic_big-data_nav_admins),12811043(mea_pam_digital_transformation_admins),12835611(sabic_big-data_cdsw
_admins),266818035(_sabic enhanced ocs users),266876594(eur-car_mtt_mtt02_r),12802072(eur-car_mtt_wfd_public_r),12814496(sabic_b
ig-data_cm_users-qlty),12814502(sabic_big-data_sntry_admins-qlty),12814203(sabic_big-data_cm_users),1625575270(sfpit006v-wass_r)
,12814202(sabic_ big-data_cm_admins),12814204(sabic_big-data_hue_users)
```

Further by checking the LDAP servers (via DNS service discovery) with the 'nslookup -type=srv _ldap._tcp.SABICCORP.SABIC.COM' command, we can see that it uses the LDAP server's unencrypted port (389), instead of the encrypted port (636).

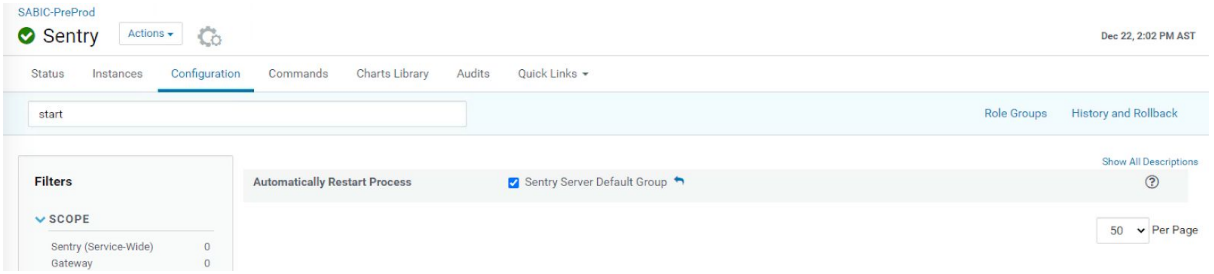
```
[root@d02ycdpmas001 sssd]# nslookup -type=srv _ldap._tcp.SABICCORP.SABIC.COM
;; Truncated, retrying in TCP mode.
Server:         10.32.43.25
Address:        10.32.43.25#53

_ldap._tcp.sabiccorp.sabic.com  service = 0 100 389 ss-jhq-mea-dc20.sabiccorp.sabic.com.
_ldap._tcp.sabiccorp.sabic.com  service = 0 100 389 ss-jhq-mea-dc21.sabiccorp.sabic.com.
_ldap._tcp.sabiccorp.sabic.com  service = 0 100 389 ss-jhq-mea-dc22.sabiccorp.sabic.com.
_ldap._tcp.sabiccorp.sabic.com  service = 0 100 389 ss-jhq-mea-dc23.sabiccorp.sabic.com.
_ldap._tcp.sabiccorp.sabic.com  service = 0 100 389 ss-jhq-mea-dc24.sabiccorp.sabic.com.
_ldap._tcp.sabiccorp.sabic.com  service = 0 100 389 ss-jhq-mea-dc25.sabiccorp.sabic.com.
_ldap._tcp.sabiccorp.sabic.com  service = 0 100 389 ss-kor-roa-dc01.sabiccorp.sabic.com.
_ldap._tcp.sabiccorp.sabic.com  service = 0 100 389 ss-mtv-amr-dc01.sabiccorp.sabic.com.
_ldap._tcp.sabiccorp.sabic.com  service = 0 100 389 ss-rhq-mea-dc01.sabiccorp.sabic.com.
_ldap._tcp.sabiccorp.sabic.com  service = 0 100 389 ss-rhq-mea-dc02.sabiccorp.sabic.com.
_ldap._tcp.sabiccorp.sabic.com  service = 0 100 389 ss-rhq-mea-dc03.sabiccorp.sabic.com.
_ldap._tcp.sabiccorp.sabic.com  service = 0 100 389 ss-rhq-mea-dc04.sabiccorp.sabic.com.
_ldap._tcp.sabiccorp.sabic.com  service = 0 100 389 ss-rhq-mea-dc05.sabiccorp.sabic.com.
_ldap._tcp.sabiccorp.sabic.com  service = 0 100 389 ss-rhq-mea-dc20.sabiccorp.sabic.com.
_ldap._tcp.sabiccorp.sabic.com  service = 0 100 389 ss-rhq-mea-dc21.sabiccorp.sabic.com.
_ldap._tcp.sabiccorp.sabic.com  service = 0 100 389 ss-rhq-mea-dc22.sabiccorp.sabic.com.
_ldap._tcp.sabiccorp.sabic.com  service = 0 100 389 ss-rhq-mea-dc26.sabiccorp.sabic.com.
_ldap._tcp.sabiccorp.sabic.com  service = 0 100 389 ss-rhq-mea-dc27.sabiccorp.sabic.com.
_ldap._tcp.sabiccorp.sabic.com  service = 0 100 389 ss-rhq-mea-dc28.sabiccorp.sabic.com.
_ldap._tcp.sabiccorp.sabic.com  service = 0 100 389 ss-sin-roa-dc01.sabiccorp.sabic.com.
_ldap._tcp.sabiccorp.sabic.com  service = 0 100 389 ss-sin-roa-dc02.sabiccorp.sabic.com.
_ldap._tcp.sabiccorp.sabic.com  service = 0 100 389 ss-wtn-eur-dc01.sabiccorp.sabic.com.
_ldap._tcp.sabiccorp.sabic.com  service = 0 100 389 ss-hou-amr-dc02.sabiccorp.sabic.com.
_ldap._tcp.sabiccorp.sabic.com  service = 0 100 389 ss-boz-dc01.sabiccorp.sabic.com.
_ldap._tcp.sabiccorp.sabic.com  service = 0 100 389 ss-jhd-dc04.sabiccorp.sabic.com.
_ldap._tcp.sabiccorp.sabic.com  service = 0 100 389 ss-jhq-dc10.sabiccorp.sabic.com.
```

8.2 Sentry review

- Sentry JVM HEAPSIZE: 10GB.
- As is shown in the diagram below, two Sentry instances are deployed.



- Meanwhile, the option 'Automatically Restart Process' is also enabled. However, the dual Sentry instances and 'Automatically Restart Process' are all serving the purpose of Sentry high availability. Thus we can sometimes only have to enable one of them, when the performance consideration is required (See Section 8.4.2 for detailed discussion).



- Kerberos is enabled.

8.3 Encryption status

- Wire encryption is enabled.
- HDFS at-rest encryption is enabled.

8.4 Security configuration review with major components

8.4.1 Security review with HDFS/YARN/ZOOKEEPER

	ITEMS	*CURRENT STATUS*
1	Hadoop Secure Authentication for HDFS hadoop.security.authentication	Enabled
2	Enable Kerberos Authentication for HTTP Web-Consoles	Enabled
3	Hadoop User Group Mapping Implementation hadoop.security.group.mapping	org.apache.hadoop.security.ShellBasedUnixGroupsMapping

4	Hadoop Secure Authorization hadoop.security.authorization	Enabled
5	Enable Access Control Lists dfs.namenode.acls.enabled	Enabled
6	Enable Sentry Synchronization	Enabled
7	Hadoop RPC Protection hadoop.rpc.protection	ALL of authentication, integrity and privacy
8	Hadoop User Group Mapping LDAP TLS/SSL Enabled hadoop.security.group.mapping.ldap.use.ssl	Disabled
9	Enable Data Transfer Encryption dfs.encrypt.data.transfer	Enabled
10	Hadoop TLS/SSL Enabled hadoop.ssl.enabled	Enabled (Enable TLS/SSL encryption for HDFS, MapReduce, and YARN web UIs, as well as encrypted shuffle for MapReduce and YARN.)
11	Enable Log and Query Redaction redaction_policy_enabled	Enabled
12	Enable TLS/SSL for HttpFS	Enabled
13	Enable Kerberos Authentication for YARN HTTP Web-Consoles	Enabled
14	YARN shuffle and web UIs encryptions	Enabled
15	Enable Zookeeper Kerberos Authentication enableSecurity	Enabled
16	Enable Zookeeper Server to Server SASL Authentication quorum.auth.enableSasl	Enabled
17	TLS/SSL for ZooKeeper JMX	Configured but disabled

Comments:

- As is seen from the security configuration of HDFS, YARN, and ZOOKEEPER, everything is well protected (including WEB UI authentication and encryption, shuffle encryption, HTTPFS encryption, log redaction, etc) except for the LDAP-based user group mapping and Zookeeper JMX encryption.

- As for Zookeeper JMX, it is only used for collecting ZOOKEEPER JVM monitoring data, and won't be of large security risk even if not encrypted.
- As for the LDAP-based user group mapping, CDH is now configured with shell-based unix group mapping, i.e. CDH asks OS for group information and OS further uses SSSD to acquire group information from LDAP. Currently, SSSD is using plain text to exchange information with LDAP.

  Theoretically speaking, if the internal network of CDH (including the network from CDH to LDAP) cannot be trusted, it will have the risk of getting the false group name (like the hdfs group) by the man-in-the-middle attack and ARP spoofing. The recommendations are:

  - If the internal network can be trusted, we can safely keep the current SSSD group mapping configuration. Moreover, that also means the data encryption between Datanodes (Item 9) is NOT required and can be disabled.
  - Otherwise if the internal network cannot be trusted, it is highly recommended to use LDAPS instead of LDAP in the SSSD configuration.

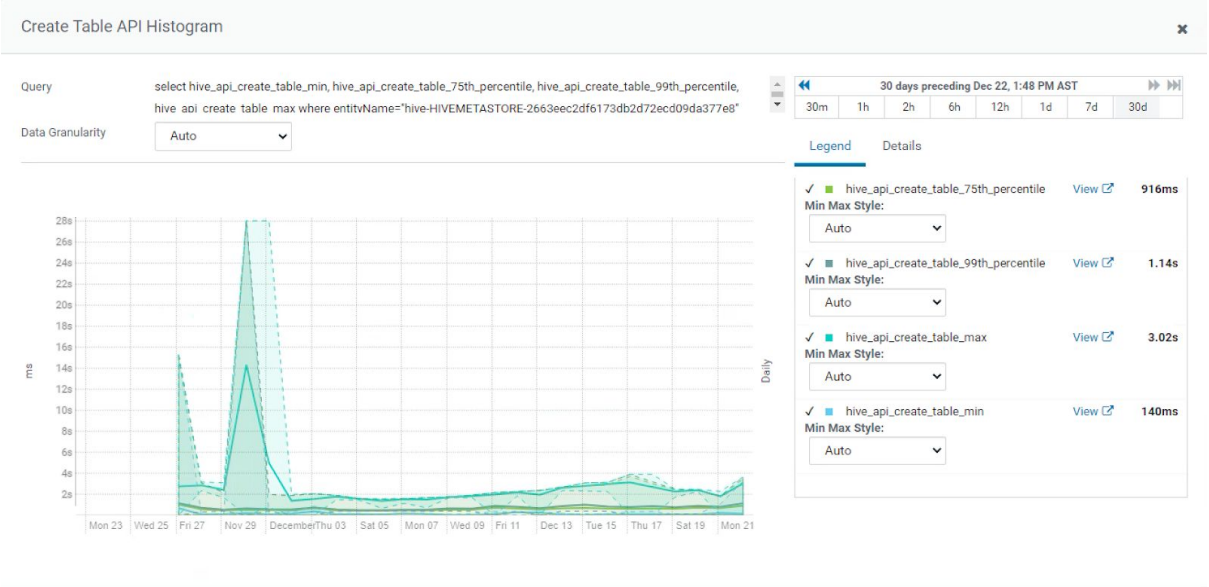
8.4.2 Security review with HIVE/IMPALA

	ITEMS	*CURRENT STATUS*
1	Kerberos authentication for HiveServer2	Enabled
2	LDAP authentication for HiveServer2	Disabled
3	Enable TLS/SSL for HiveServer2 hive.server2.use.SSL	Enabled
4	Enable TLS/SSL for HiveServer2 WebUI hive.server2.webui.use.ssl	Enabled
5	Enable Stored Notifications in Database	Enabled
6	Kerberos and Impersonation	Disabled
7	Kerberos authentication for IMPALA	Enabled
8	Enable TLS/SSL for Impala client_services_ssl_enabled (Encryption between client and Impala daemon)	Enabled
9	Impala TLS/SSL CA Certificate ssl_client_ca_certificate (Encryption between Impala daemons)	SET and enabled (Used to confirm the authenticity of SSL/TLS servers that the Impala daemons might connect to)
10	WebUI encryption for IMPALA	SET and enabled for all of CATALOG, STATESTORE, and IMPALA-DAEMON

11	Enable LDAP Authentication enable_ldap_auth	Disabled
12	Enable LDAP TLS ldap_tls	Disabled
13	Disk Spill Encryption disk_spill_encryption	Disabled
14	LDAP Server CA Certificate ldap_ca_certificate	NONE (Used to confirm the authenticity of the LDAP server certificate)
15	Log redaction for HIVE and IMPALA	Enabled (Configured in HDFS configuration page, see: https://docs.cloudera.com/documentation/enterprise/latest/topics/sg_redaction.html#concept_i2b_zt2_5y)

Comments:

- Both HIVE and IMPALA have Kerberos enabled, but LDAP disabled. That means, users of HIVE and IMPALA can only use Kerberos to log into HIVE and IMPALA.
- Communications between client and HIVE-SERVER2 and IMPALA-DAEMON are all encrypted and the certificates are all properly configured.
- Communications between IMPALA-DAEMONS are also encrypted and the certificates are all properly configured. However, it is required only when the internal network cannot be trusted.
    - Recommended to consider that together with the encryption between Datanodes, and the SSSD user group mapping encryption.
- WEB-UI encryptions are all enabled for HIVE-SERVER2, IMPALA CATALOG SERVER, IMPALA STATESTORE, and IMPALA-DAEMON, and the certificates are all properly configured.
- Enabling stored notifications in database (Item 5) will make the HIVE create-table API call wait until the information of the newly created table is totally synchronized with both HIVE METASTORE and both Sentry instances, which is sometimes very slow. In fact, the CM metrics chart 'Create Table API Histogram' has already shown that (28 seconds for a single create-table statement in max). The workaround approaches are either of the following:
    - Uncheck the 'Enable Stored Notification in Database' option, which actually makes the table creation information distributed in the asynchronous way, or in Eventual-Consistency, so that the client won't have to wait until the table creation information is totally synchronized. The side-effect of this approach is that the newly created table might be seen after seconds or tens of seconds.
    - Instead of using dual Sentry instances to implement Sentry HA, we can also use CM to implement the Sentry HA by checking 'Automatically Restart Process'.



8.4.3 Security review with Solr

	ITEMS	*CURRENT STATUS*
1	Kerberos authentication for Solr	Enabled
2	Enable LDAP Authentication	Disabled
3	Enable TLS/SSL for Solr (Encryption between client and Solr server)	Enabled
4	Solr TLS/SSL Client Trust Store File (Used when Solr is the client in a TLS/SSL connection)	SET

Comments:

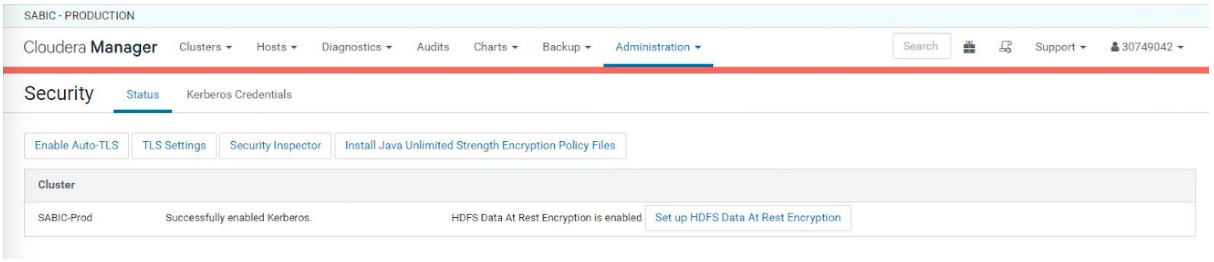
- Solr has Kerberos enabled, but LDAP disabled.
- Encryption between client and Solr server is enabled, and the certificate is properly configured.
- The certificate used when Solr server is the client is also properly configured.

9 Security Configuration Review for PROD

9.1 Kerberos & LDAP

9.1.1 Kerberos enablement and /etc/krb5.conf

By checking CM → Administration → Security → Status, it can be seen that Kerberos is enabled in the QA cluster.



Further check with the client-side Kerberos configuration file /etc/krb5.conf, as shown below.

```
[hadoop@d02pcdpmas002 ~]$ cat /etc/krb5.conf
[libdefaults]
default_realm = SABICCORP.SABIC.COM
dns_lookup_kdc = false
dns_lookup_realm = false
ticket_lifetime = 86400
renew_lifetime = 604800
forwardable = true
default_tgs_enctypes = aes256-cts aes128-cts
default_tkt_enctypes = aes256-cts aes128-cts
permitted_enctypes = aes256-cts aes128-cts
udp_preference_limit = 1
kdc_timeout = 10000000

[realms]
SABICCORP.SABIC.COM = {
kdc = SS-JHQ-MEA-DC20.SABICCORP.SABIC.COM
admin_server = SS-JHQ-MEA-DC20.SABICCORP.SABIC.COM
}
[domain_realm]
.sabic.com = SABICCORP.SABIC.COM
```

As only one KDC address is configured in /etc/krb5.conf, make sure the KDC (or Active Directory) itself has configured with the HA mechanism. That is said, when the master KDC is halted, its IP address can be switched to the standby node automatically.

9.1.2 LDAP client configuration at OS and /etc/openldap/ldap.conf

```
[hadoop@d02pcdpmas002 ~]$ cat /etc/openldap/ldap.conf
#
# LDAP Defaults
```

```
#

# See ldap.conf(5) for details
# This file should be world readable but not world writable.

#BASE   dc=example,dc=com
#URI    ldap://ldap.example.com ldap://ldap-master.example.com:666

#SIZELIMIT      12
#TIMELIMIT      15
#DEREF          never

TLS_CACERTDIR   /etc/openldap/certs

# Turning this off breaks GSSAPI used with krb5 when rdns = false
SASL_NOCANON    on
```

9.1.3 LDAP integration in OS and /etc/sssd/sssd.conf

```
[sssd]
domains = SABICCORP.SABIC.COM
config_file_version = 2
services = nss, pam

[domain/SABICCORP.SABIC.COM]
ad_domain = SABICCORP.SABIC.COM
krb5_realm = SABICCORP.SABIC.COM
realmd_tags = manages-system joined-with-samba
cache_credentials = True
id_provider = ad
krb5_store_password_if_offline = True
default_shell = /bin/bash
ldap_id_mapping = True
use_fully_qualified_names = False
fallback_homedir = /home/%u@%d
access_provider = ad
ad_access_filter
=(&(memberOf=CN=SABIC_Big-Data_CDH_Admins,OU=Groups,OU=Big-Data,OU=Apps,DC=
SABICCORP,DC=SABIC,DC=com))
enumerate = false
ignore_group_members = true
ldap_schema =ad
ldap_force_upper_case_realm = true
case_sensitive = false
```

As neither ldap_uri nor ldap_backup_uri is specified in sssd.conf, the service discovery mechanism is used to find LDAP servers. And by checking user information with the 'id' command (id shikas and id 30749042), it can output successfully with the required information.

```
[root@d02pcdpmas002 etc]# id shikas
uid=12884664(shikas) gid=266800513(domain users) groups=266800513(domain users),1325610586(s-1-5-21-111690494-1484066017-1836196843-10586@SABICHQ.
SABIC.COM),266945732(_sabic shared services staff - jubail),12888248(sabic_big-data_dataiku_users-qlty),12874026(sabic_big-data_cdsw_users-qlty),1
2814495(sabic_big-data_cdh_users-qlty),1625560062(_sabic azure aip user license),12888246(sabic_big-data_dataiku_users),12884348(mea_pam_cdp(pamus
er)_admins),1625445416(a_cm_acceptable_use_policy),12835613(sabic_big-data_cdsw_users),12888990(sabic_cdp_bd_ahi_tech_dev),266818035(_sabic enhanc
ed ocs users),12889052(sa02-vmpc00061_rdpadmins),266907996(internet_jub_ss_users),12866472(pi_sys_iz_pp3_sec),12842526(pi_sys_ys_hdpe_sec),2669329
68(ra_jub_ss),266939999(_sabic global except europe),266947567(_sabic (hq)),266945731(_sabic shared services staff),266918826(internet_mea_class_b
),266908077(internet_ksa_users),266810206(internet_sabic_global_class_b),1625535286(eur-car_maintenance_lxlmechanical_condition-monitoring_rx),128
02072(eur-car_mtt_wfd_public_r),266876594(eur-car_mtt_mtt02_r),1625575270(sfpit006v-wass_r),266932974(ra_ksa_users)
```

```
[hadoop@d02pcdpmas002 ~]$ id 30749042
uid=12915377(30749042) gid=266800513(domain users) groups=266800513(domain users),1625560062(_sabic azure aip user license),12814492(sabic_b
ig-data_cm_admins-qlty),1625544686(count_citrix_groups),266809684(_sabic mea sabicju all users),12889086(sa02-vmpc00096_rdpadmins),12814208(
sabic_big-data_sntry_admins),12814205(sabic_big-data_hue_admins),12814498(sabic_big-data_hue_users-qlty),12827936(sabic_big-data_cdsw_admins
-qlty),12814500(sabic_big-data_nav_admins-qlty),1625535286(eur-car_maintenance_lxlmechanical_condition-monitoring_rx),12814499(sabic_big-dat
a_key_admins-qlty),12814497(sabic_big-data_hue_admins-qlty),12814206(sabic_big-data_key_admins),12814207(sabic_big-data_nav_admins),12802072
(eur-car_mtt_wfd_public_r),266876594(eur-car_mtt_mtt02_r),1625575270(sfpit006v-wass_r),12811043(mea_pam_digital_transformation_admins),12835
611(sabic_big-data_cdsw_admins),266818035(_sabic enhanced ocs users),12814496(sabic_big-data_cm_users-qlty),12814502(sabic_big-data_sntry_ad
mins-qlty),12814203(sabic_big-data_cm_users),12814202(sabic_ big-data_cm_admins),12814204(sabic_big-data_hue_users)
```

Further by checking the LDAP servers (via DNS service discovery) with the 'nslookup -type=srv _ldap._tcp.SABICCORP.SABIC.COM' command, we can see that it uses the LDAP server's unencrypted port (389), instead of the encrypted port (636).

```
[hadoop@d02pcdpmas002 ~]$ nslookup -type=srv _ldap._tcp.SABICCORP.SABIC.COM
;; Truncated, retrying in TCP mode.
Server:         10.32.43.25
Address:        10.32.43.25#53

_ldap._tcp.sabiccorp.sabic.com  service = 0 100 389 ss-boz-dc01.sabiccorp.sabic.com.
_ldap._tcp.sabiccorp.sabic.com  service = 0 100 389 ss-jhd-dc04.sabiccorp.sabic.com.
_ldap._tcp.sabiccorp.sabic.com  service = 0 100 389 ss-jhq-dc10.sabiccorp.sabic.com.
_ldap._tcp.sabiccorp.sabic.com  service = 0 100 389 ss-rhq-dc10.sabiccorp.sabic.com.
_ldap._tcp.sabiccorp.sabic.com  service = 0 100 389 ss-slk-dc01.sabiccorp.sabic.com.
_ldap._tcp.sabiccorp.sabic.com  service = 0 100 389 ss-nan-gc-dc01.sabiccorp.sabic.com.
_ldap._tcp.sabiccorp.sabic.com  service = 0 100 389 ss-sha-gc-dc01.sabiccorp.sabic.com.
_ldap._tcp.sabiccorp.sabic.com  service = 0 100 389 ss-sha-gc-dc02.sabiccorp.sabic.com.
_ldap._tcp.sabiccorp.sabic.com  service = 0 100 389 ss-blr-roa-dc01.sabiccorp.sabic.com.
_ldap._tcp.sabiccorp.sabic.com  service = 0 100 389 ss-boz-eur-dc01.sabiccorp.sabic.com.
_ldap._tcp.sabiccorp.sabic.com  service = 0 100 389 ss-bur-amr-dc01.sabiccorp.sabic.com.
_ldap._tcp.sabiccorp.sabic.com  service = 0 100 389 ss-car-eur-dc01.sabiccorp.sabic.com.
_ldap._tcp.sabiccorp.sabic.com  service = 0 100 389 ss-ehv-eur-dc02.sabiccorp.sabic.com.
_ldap._tcp.sabiccorp.sabic.com  service = 0 100 389 ss-gln-eur-dc01.sabiccorp.sabic.com.
_ldap._tcp.sabiccorp.sabic.com  service = 0 100 389 ss-hou-amr-dc01.sabiccorp.sabic.com.
_ldap._tcp.sabiccorp.sabic.com  service = 0 100 389 ss-hou-amr-dc02.sabiccorp.sabic.com.
_ldap._tcp.sabiccorp.sabic.com  service = 0 100 389 ss-jhd-mea-dc01.sabiccorp.sabic.com.
_ldap._tcp.sabiccorp.sabic.com  service = 0 100 389 ss-jhd-mea-dc02.sabiccorp.sabic.com.
_ldap._tcp.sabiccorp.sabic.com  service = 0 100 389 ss-jhd-mea-dc03.sabiccorp.sabic.com.
_ldap._tcp.sabiccorp.sabic.com  service = 0 100 389 ss-jhd-mea-dc20.sabiccorp.sabic.com.
_ldap._tcp.sabiccorp.sabic.com  service = 0 100 389 ss-jhd-mea-dc21.sabiccorp.sabic.com.
_ldap._tcp.sabiccorp.sabic.com  service = 0 100 389 ss-jhd-mea-dc22.sabiccorp.sabic.com.
_ldap._tcp.sabiccorp.sabic.com  service = 0 100 389 ss-jhd-mea-dc26.sabiccorp.sabic.com.
_ldap._tcp.sabiccorp.sabic.com  service = 0 100 389 ss-jhd-mea-dc27.sabiccorp.sabic.com.
_ldap._tcp.sabiccorp.sabic.com  service = 0 100 389 ss-jhd-mea-dc28.sabiccorp.sabic.com.
_ldap._tcp.sabiccorp.sabic.com  service = 0 100 389 ss-jhq-mea-dc01.sabiccorp.sabic.com.
_ldap._tcp.sabiccorp.sabic.com  service = 0 100 389 ss-jhq-mea-dc02.sabiccorp.sabic.com.
_ldap._tcp.sabiccorp.sabic.com  service = 0 100 389 ss-jhq-mea-dc03.sabiccorp.sabic.com.
_ldap._tcp.sabiccorp.sabic.com  service = 0 100 389 ss-jhq-mea-dc20.sabiccorp.sabic.com.
_ldap._tcp.sabiccorp.sabic.com  service = 0 100 389 ss-jhq-mea-dc21.sabiccorp.sabic.com.
_ldap._tcp.sabiccorp.sabic.com  service = 0 100 389 ss-jhq-mea-dc22.sabiccorp.sabic.com.
```

9.2 Sentry review

- Sentry JVM HEAPSIZE: 1GB.
    - Cloudera recommends that for each Sentry host, you have 2.25 GB memory per million objects in the Hive database. Hive objects include servers, databases, tables, partitions, columns, URIs, and views. See:

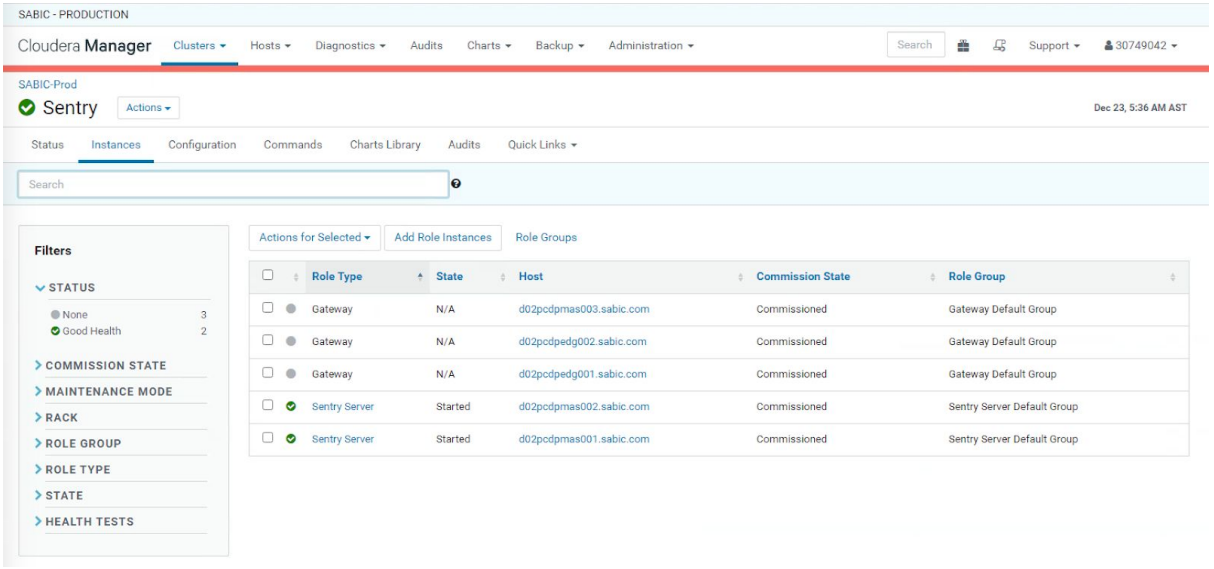
      https://docs.cloudera.com/documentation/enterprise/latest/topics/sg_sentry_before_you_install.html

    - According to our estimation, the total objects might be in the scale of 200,000 to 300,000. So that the current setting of 1GB should be just enough according to the figure below.
    - However, as we all know that such kind of 'just enough' setting is actually very risky in the production environment. If someone created a HIVE table with

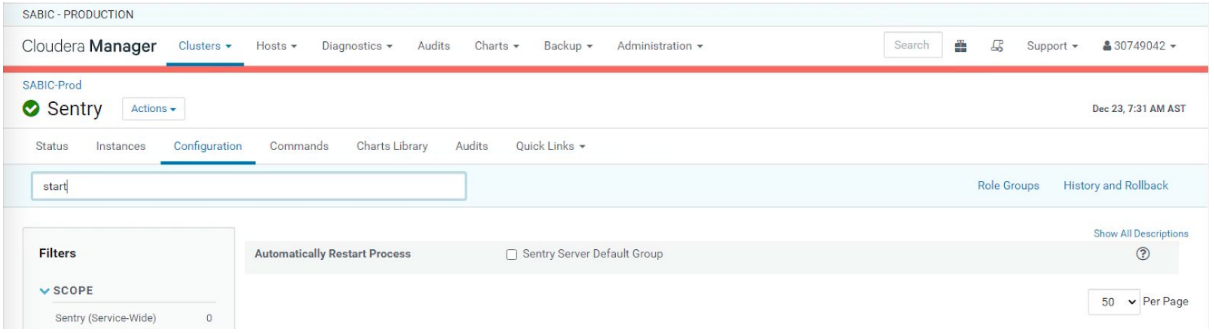
one-million partitions by mistake (e.g. mistaking the primary key as the partitioning key), the total object number would soon exceed one million. And thus fully occupied the heap space of Sentry JVM.

- Based on the consideration above, it is recommended to set the Sentry JVM heap size to around 2GB to 3GB for safety.
- NOTE that this parameter is set to 10GB in the QA cluster.

- As is shown in the diagram below, two Sentry instances are deployed.



- Meanwhile, the option 'Automatically Restart Process' is disabled. See Section 9.4.2 for the detailed discussion about Sentry instances number and this option.



- Kerberos is enabled.

9.3 Encryption status

- Wire encryption is enabled.
- HDFS at-rest encryption is enabled.

9.4 Security configuration review with major components

9.4.1 Security review with HDFS/YARN/ZOOKEEPER

	ITEMS	*CURRENT STATUS*
1	Hadoop Secure Authentication for HDFS hadoop.security.authentication	Enabled
2	Enable Kerberos Authentication for HTTP Web-Consoles	Enabled
3	Hadoop User Group Mapping Implementation hadoop.security.group.mapping	org.apache.hadoop.security.ShellBasedUnixGroupsMapping
4	Hadoop Secure Authorization hadoop.security.authorization	Enabled
5	Enable Access Control Lists dfs.namenode.acls.enabled	Enabled
6	Enable Sentry Synchronization	Enabled
7	Hadoop RPC Protection hadoop.rpc.protection	ALL of authentication, integrity and privacy
8	Hadoop User Group Mapping LDAP TLS/SSL Enabled hadoop.security.group.mapping.ldap.use.ssl	Disabled
9	Enable Data Transfer Encryption dfs.encrypt.data.transfer	Enabled
10	Hadoop TLS/SSL Enabled	Enabled

	hadoop.ssl.enabled	(Enable TLS/SSL encryption for HDFS, MapReduce, and YARN web UIs, as well as encrypted shuffle for MapReduce and YARN.)
11	Enable Log and Query Redaction redaction_policy_enabled	Enabled
12	Enable TLS/SSL for HttpFS	Enabled
13	Enable Kerberos Authentication for YARN HTTP Web-Consoles	Enabled
14	YARN shuffle and web UIs encryptions ssl.server.keystore.location	SET and enabled
15	Enable Zookeeper Kerberos Authentication enableSecurity	Enabled
16	Enable Zookeeper Server to Server SASL Authentication quorum.auth.enableSasl	Enabled
17	TLS/SSL for ZooKeeper JMX	Disabled

Comments:

- As is seen from the security configuration of HDFS, YARN, and ZOOKEEPER, everything is well protected (including WEB UI authentication and encryption, shuffle encryption, HTTPFS encryption, log redaction, etc) except for the LDAP-based user group mapping and Zookeeper JMX encryption.
- As for Zookeeper JMX, it is only used for collecting ZOOKEEPER JVM monitoring data, and won't be of large security risk even if not encrypted.
- As for the LDAP-based user group mapping, CDH is now configured with shell-based unix group mapping, i.e. CDH asks OS for group information and OS further uses SSSD to acquire group information from LDAP. Currently, SSSD is using plain text to exchange information with LDAP.

  Theoretically speaking, if the internal network of CDH (including the network from CDH to LDAP) cannot be trusted, it will have the risk of getting the false group name (like the hdfs group) by the man-in-the-middle attack and ARP spoofing. The recommendations are:

  - If the internal network can be trusted, we can safely keep the current SSSD group mapping configuration. Moreover, that also means the data encryption between Datanodes (Item 9) is NOT required and can be disabled.
  - Otherwise if the internal network cannot be trusted, it is highly recommended to use LDAPS instead of LDAP in the SSSD configuration.

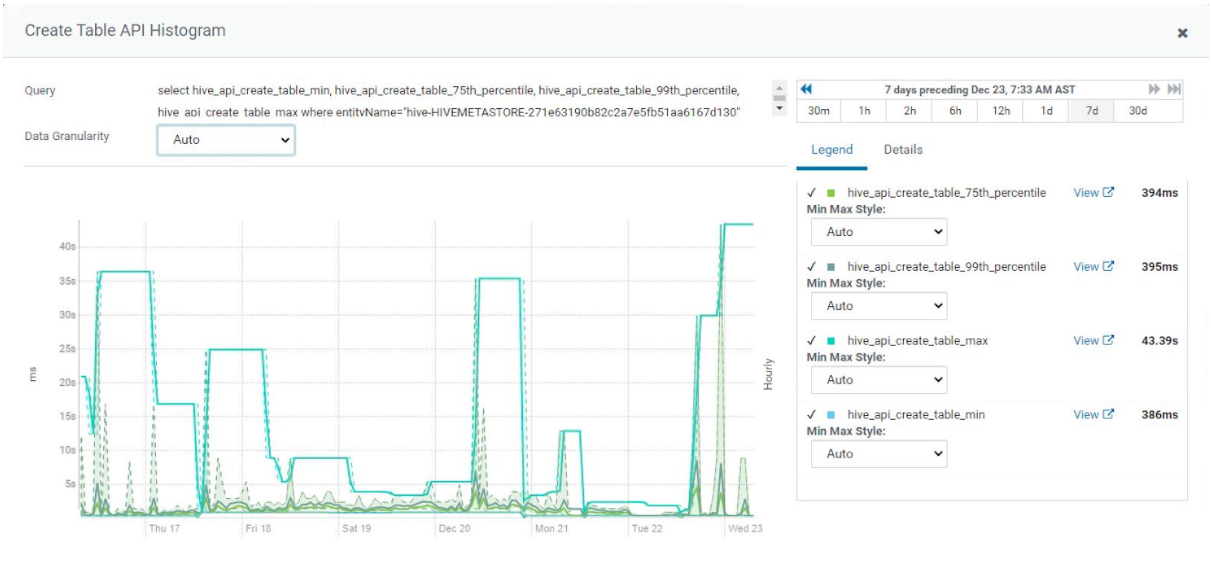
9.4.2 Security review with HIVE/IMPALA

	ITEMS	*CURRENT STATUS*
1	Kerberos authentication for HiveServer2	Enabled
2	LDAP authentication for HiveServer2	Disabled
3	Enable TLS/SSL for HiveServer2 hive.server2.use.SSL	Enabled
4	Enable TLS/SSL for HiveServer2 WebUI hive.server2.webui.use.ssl	Enabled
5	Enable Stored Notifications in Database	Enabled
6	Kerberos and Impersonation	Disabled
7	Kerberos authentication for IMPALA	Enabled
8	Enable TLS/SSL for Impala client_services_ssl_enabled (Encryption between client and Impala daemon)	Enabled
9	Impala TLS/SSL CA Certificate ssl_client_ca_certificate (Encryption between Impala daemons)	SET and enabled (Used to confirm the authenticity of SSL/TLS servers that the Impala daemons might connect to)
10	WebUI encryption for IMPALA	SET and enabled for all of CATALOG, STATESTORE, and IMPALA-DAEMON
11	Enable LDAP Authentication enable_ldap_auth	Disabled
12	Enable LDAP TLS ldap_tls	Enabled
13	Disk Spill Encryption disk_spill_encryption	Disabled
14	LDAP Server CA Certificate ldap_ca_certificate	NONE (Used to confirm the authenticity of the LDAP server certificate)
15	Log redaction for HIVE and IMPALA	Enabled (Configured in HDFS configuration page, see: https://docs.cloudera.com/documentation

		/enterprise/latest/topics/sg_redaction.html#concept_i2b_zt2_5y)

Comments:

- Both HIVE and IMPALA have Kerberos enabled, but LDAP disabled. That means, users of HIVE and IMPALA can only use Kerberos to log into HIVE and IMPALA.
- Communications between client and HIVE-SERVER2 and IMPALA-DAEMON are all encrypted and the certificates are all properly configured.
- Communications between IMPALA-DAEMONS are also encrypted and the certificates are all properly configured. However, it is required only when the internal network cannot be trusted.
    - Recommended to consider that together with the encryption between Datanodes, and the SSSD user group mapping encryption.
- WEB-UI encryptions are all enabled for HIVE-SERVER2, IMPALA CATALOG SERVER, IMPALA STATESTORE, and IMPALA-DAEMON, and the certificates are all properly configured.
- Enabling stored notifications in database (Item 5) will make the HIVE create-table API call wait until the information of the newly created table is totally synchronized with both HIVE METASTORE and both Sentry instances, which is sometimes very slow. In fact, the CM metrics chart 'Create Table API Histogram' has already shown that (more than 40 seconds for a single create-table statement in max, and even worse than the QA cluster). The workaround approaches are either of the following:
    - Uncheck the 'Enable Stored Notification in Database' option, which actually makes the table creation information distributed in the asynchronous way, or in Eventual-Consistency, so that the client won't have to wait until the table creation information is totally synchronized. The side-effect of this approach is that the newly created table might be seen after seconds or tens of seconds.
    - Instead of using dual Sentry instances to implement Sentry HA, we can also use CM to implement the Sentry HA by checking 'Automatically Restart Process'.



9.4.3 Security review with Solr

	ITEMS	*CURRENT STATUS*
1	Kerberos authentication for Solr	Enabled
2	Enable LDAP Authentication	Disabled
3	Enable TLS/SSL for Solr (Encryption between client and Solr server)	Enabled
4	Solr TLS/SSL Client Trust Store File (Used when Solr is the client in a TLS/SSL connection)	SET

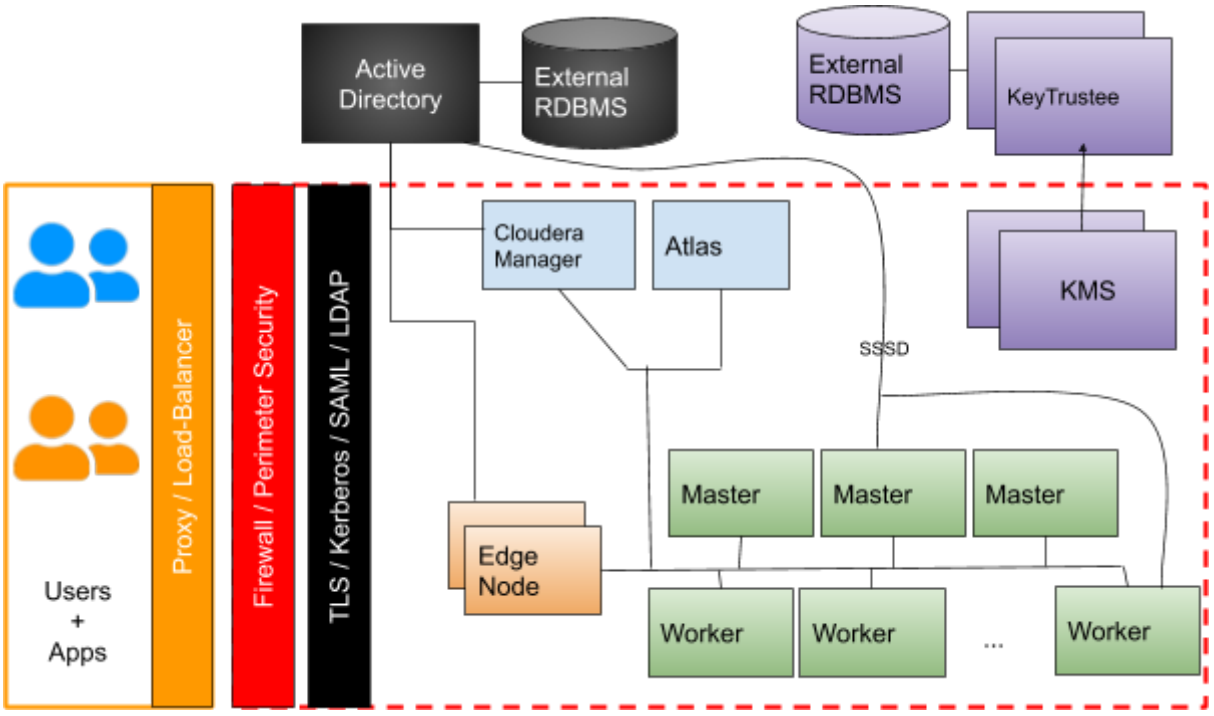
Comments:

- Solr has Kerberos enabled, but LDAP disabled.
- Encryption between client and Solr server is enabled, and the certificate is properly configured.
- The certificate used when Solr server is the client is also properly configured.

9.5 KTS and KMS review

9.5.1 Security Zone of KTS

As shown in CDH_servers_v3.xlsx, both KTS and KMS servers are now located in the DC zone. However, as the KTS servers are the locations where the real keys are stored, it is highly recommended that they are located in a separated zone and protected by the firewall. Cloudera's security reference architecture in the diagram below also shows that.



9.5.2 Major configurations review

	ITEMS	*CURRENT STATUS*
1	Database Storage Directory db_root	/data/01/keytrustee/db
2	Raid 1 or 1+0 for db_root	?? (/dev/sdb) Assuming the virtual disk is already deployed with RAID
3	Unix access attributes	drwx------
4	Backup policy	??
5	Encryption	Enabled for both active and passive servers
6	Automatically Restart Process (KTS)	Disabled
7	KMS Heap Size kms_heap_size	1GB
8	Kerberos for KMS	Enabled
9	Enable TLS/SSL for Key Management Server Proxy hadoop.kms.ssl.enabled	Enabled
10	KMS Max Threads hadoop.http.max.threads	250
11	KMS Accept Count	500

	hadoop.http.accept.queue.size	
12	Automatically Restart Process (KMS)	Disabled

Comments:

- The database of KTS stores all the keys of the encrypted zones in HDFS, and damaging the database or its storing disks will lead to the loss of all encrypted data. Thus RAID 1 or 1+0, and database backup are essential to the data safety of HDFS.
- Unlike KTS that has to be in the active-passive mode, KMS can be scaled out. However, KMS JVM heap size of 1GB is a bit too small, and consider to make it 2GB to 4GB.