Big Data Project: Benchmark testing procedure version 2
(Version 1.02)

Purpose:
This document describes the necessary steps that a person needs to follow to run the benchmarks

Configuration:
The Big data cluster consisting of 6 DataNodes and 2 NameNodes.
Hardware:
- 8 x Cisco C240 M3 rack server.
- 256GB total memory for each node (16GB DDR3-1600-MHz RDIMM/PC3-12800/dual rank/1.35v)
- DataNodes: 24 RAID0 with 1 HDD.
- NameNodes: 8 HDDs configured as RAID6 with 4 hot-spare HDDs (12 HDDs free)
- Each HDD is a 600GB 6Gb SAS 10K RPM SFF HDD

Software:
- CentOS: 6.5 (Final)
- UCSM version: 2.2(1d)
- BIOS C240 M3: 3.1.5.4f.0.1113201
- Cloudera Hadoop 5.0.1

Cloudera post install configuration optimizations:
All other values are at their specified defaults except for the following

Configuration item	Final tuning
dfs.namenode.handler.count	160
fs.trash.interval	1440
io.file.buffer.size	64kb
yarn.nodemanager.resource.memory-mb (all resource groups)	152GB
yarn.nodemanager.resource.cpu-vcores	TBC during some tests
yarn_scheduler_minimum_allocation_mb	1GB
yarn.scheduler.maximum-allocation-mb	96gb
mapreduce.map.memory.mb	1GB
mapreduce.reduce.memory.mb	1GB
mapreduce.map.java.opts	CD manager Defaults
mapreduce.reduce.java.opts	CD manager Defaults
mapred_compress_map_output	TRUE
mapred_map_output_compression_codec	org.apache.hadoop.io.compress.SnappyCodec
mapred_reduce_parallel_copies	32
mapreduce.task.io.sort.mb	512
mapreduce.map.sort.spill.percent	0.8
mapreduce.task.io.sort.factor	64
CDH version	5.0.1

Mainly changed

Additional Linux & clusters changes:

- CPU scaling set to performance to force CPU to always be running at highest frequency
  - Impala/Hbase systems need to be set to performance
- vm.swappiness set to vm.swapiness=0
- Transparent Huge Pages Defrag set to OFF.
- Network tuning
  - Set ring buffer to 4096

- Tuned memory settings in sysctl.conf
- Enabled jumbo frames on the LAN and the servers
- lowered IRQ rate and increased throughput

Changes in sysctl.conf:

```
#10 Gbps settings
TCP/IP memory tuning
net.ipv4.tcp_timestamps = 0
net.ipv4.tcp_sack = 0
net.ipv4.tcp_rmem = 4096 134217728 134217728
net.ipv4.tcp_wmem = 4096 134217728 134217728
Core memory tuning
net.core.rmem_max = 134217728
net.core.wmem_max = 134217728
net.core.rmem_default = 134217728
net.core.rmem_default = 134217728
net.core.optme_max = 134217728
net.core.netdev_max_backlog = 250000
# recommended default congestion control is htcp
net.ipv4.tcp_congestion_control=htcp
# recommended for hosts with jumbo frames enabled
net.ipv4.tcp_mtu_probing=1
vm.swappiness = 0
```

- onBoot.sh script added to /etc/rc.local to apply on every boot.

onBoot.sh script:
```
#Set all CPU to max freq
for i in {0..47};do echo performance>
/sys/devices/system/cpu/cpu$i/cpufreq/scaling_governor;done
#Disable Transparent Huge Page defrag
echo never> /sys/kernel/mm/transparent_hugepage/defrag
#Transmit to NIC card 4096 Bytes at a time
#setpci -v -d 1137:0043 e6.b=2e
#Jumbo frames
ifconfig eth0 mtu 9000
ifconfig eth0 txqueuelen 1000
```

Defined benchmarking test

Ten unique tests are identified (see table). Every test has a unique name i.e. test<x>. Every test is run 3 times. A test is run by launching the script benchmark-v2.sh.
Benchmark loops automatically through the 3 iterations of a benchmark run – kick it off once to run all 3 runs.

The results are stored individually for each sequence & benchmark run, in a separate directory.
The individual directories with the results are located on server 1 (192.168.201.20) in the directory */benchmark*.
Every test has a unique configuration. The configuration <u>must not</u> be changed during the 3 runs of each test.

Testname	CPU	DataNode/ node manager	1Gb /10Gb	HT ON / OFF	Cluster Tuning Vcores (default group)	Nbr. of HDD	Line rate	DCM power logging
Test 1	E5-2695v2	6/5	10Gb	ON	38	24	4 links - 5Gb/s	GroupBD
test2	E5-2660v2	6/5	10Gb	ON	32	24	4 links - 5Gb/s	GroupBD
test3	E5-2643v2	6/5	10Gb	ON	19	24	4 links - 5Gb/s	GroupBD
test4	E5-2667v2	6/5	10Gb	ON	26	24	4 links - 5Gb/s	GroupBD
test5	E5-2690v2	6/5	10Gb	ON	32	24	4 links - 5Gb/s	GroupBD

Testname	CPU	DataNode/ node manager	1Gb / 10Gb	HT ON / OFF	Cluster Tuning Vcores (default group)	Nbr. of HDD	Line rate	DCM power logging
test6	E5-2690v2	6/5	10Gb	ON	32	24	FE removed	GroupBD
test7	E5-2690v2	6/5	10Gb	OFF	16	24	FE removed	GroupBD

test8	E5-2690v2	6/5	1Gb	ON	32	24	FE removed	GroupBD
test9	E5-2690v2	5/5	10Gb	ON	32	24	FE removed	GroupBD-8
test10	E5-2690v2	4/4	10Gb	ON	32	24	FE removed	GroupBD-8&7
test11	E5-2690v2	6/6	10Gb	ON	32	24	FE removed	GroupBD
test12	E5-2690v2	6/6	10Gb	ON	32	12	FE removed	GroupBD

(table 1)

The configuration will consist of:
- Specific CPU
- Number of DataNodes powered on & number of node managers running.
    - 6/5 = 6 data nodes and 5 node managers running (no node manager on DataNode 3)
- Hyper-threading switched ON or OFF
- Server Network card set to 1Gb or 10Gb
- Cluster tuning to reflect the number of cores. (see table 1)
    - With hyper threading enabled the virtual cores are set to 1.6 times total hyper threaded cores
        - I.e.: 12C => 12 * 2 * 1.6 = 38
- Number of HDDs used on the DataNodes.
- Line rate
    - 4 uplinks per 8 servers is typically recommended (5Gbit/sec per server) (test 1 – test 5)
    - Full line rate (10Gbit/sec per server) – FEX removed. Servers directly connected to FI

The above table shows the sequence in which all the tests must be run. <u>The person performing the benchmark tests shall not deviate from the sequence as shown in table 1.</u>

Testing prerequisites:
There are two required Jar files required to execute this benchmark
- Jar 1 - hadoop-mapreduce-client-jobclient.jar
- Jar 2 - hadoop-mapreduce-examples.jar

Copy these files from /opt/cloudera/parcels/<distribution>/lib/hadoop-mapreduce to the working directory where the tests will be run from (where benchmark-test.sh is located)

Unpack and copy the inputtext.zip file to hdfs as follows:
- Unzip inputtext.xzip

Note the location where files are extracted.
Edit the benchmark-v2.sh script line as follows:
- hadoop fs put <pattern> /user/root/wordcount/in/
- Once script has run once, put a # in front of the command to avoid the copy action happening for every script run

The script will start 3 Hadoop reference benchmarks:
1. TeraGen, TeraSort, TeraValidate (10 billion rows)
2. TeraGen, TeraSort, TeraValidate (20 billion rows)
3. Pig Wordcount.

TeraGen benchmark (1 ..2)
Parameters:
- hadoop jar hadoop-mapreduce-examples.jar teragen -Dmapred.map.tasks=119 <teragensize> /user/root/terasort<1..2>-out

TeraSort benchmark
Parameters:
- hadoop jar hadoop-mapreduce-examples.jar terasort /user/root/terasort<1..2>-out/ /user/root/terasort<1..2>-sorted/

TeraValdate benchmark
Parameters:
- hadoop jar hadoop-mapreduce-examples.jar teravalidate /user/root/terasort<1..2>-sorted/ /user/root/terasort<1..2>-validated/

Pig Wordcount

Parameters:

- Datasets of 11.5GB of text files

  hadoop fs -rm -r -skipTrash wordcount/out-ba
  hadoop fs -expunge
  pig wordcount.pig

How to run the benchmark script:

1. Copy to management node with HDFS gateway and yarn gateway roles enabled
2. Create a directory called /benchmark
3. Copy benchmark-v2.sh and the Jar files described above to this location
4. Execute the script via ./benchmark-v2.sh
   a. You will be prompted for 3 entries –
      i. Enter your surname
      ii. Enter the test name
   b. The script will then run and execute three benchmarks
   c. Record performance results as described in this document
      i. Ensure to use yarn applications list within the Cloudera UI to record results
      ii. Ensure to record results from DCM for power consumption
      iii. Ensure to record relevant results from Ganglia
      iv. Record results in test spreadsheet
   d. Once benchmark script is completed, execute again for subsequent test runs and be sure to enter a unique test sequence number
   e. Rerun to complete a total of 3 test sequences per test
   f. Once completed, tar all files and store the entire test directory contents on the Internal sharepoint for analysis
   g. Modify system configuration as required and rerun tests as required

Once the reference benchmarks are terminated the collected information will be stored.
A unique directory, using the above provided info, will be created for every test.

After finishing a test, 3 sub-directories will be created with the structure: <Surname>-<Testname>-<Sequence>
   i.e.: smith-test5-1 AND smith-test5-2 AND smith-test5-3.

At every run the data is refreshed, it means that every new test will wipe the old data and use fresh data.

Collecting the benchmark information

1) Ganglia:
An agent (service must be started on every DataNode and NameNode). During the tests the agents will send the info to the Ganglia server. The information can be found in at http://192.168.201.222/ganglia. After every test the information must be copied to the SAN.

How to collect the data:

- Go to the ganglia webpage at http://192.168.201.222/ganglia and make certain all the data has been recorded. (due to the sampling nature it can take a few minutes before all the graphs have been updated) make certain to click the "get fresh data" button.
- Log into the ganglia server (ssh 192.168.201.222)
- Go to the /var/lib/ganglia directory
- Create an archive of the RRD files. Via "tar –zcf <testname>.tar.gz ./rrds"
   o i.e. bin case of test1: "tar –zcf test1.tar.gz ./rrds"
- Copy test1.tar.gz on the SAN
- (Make certain you don't overwrite the previous archives in case you leave the archives in place)

- The *.rrd files can be manipulated later with the rrdtools , to extract xml info or turned into graphs etc.

2) Cloudera:
After termination of every test the script will collect all information and copy it to sub-directory location <Surname>-<Testname>-<Sequence>.

Additionally, following info via the Cloudera manager must be retrieved:

1. Open a connection to Cloudera manager and log in with Admin rights
2. Click on the yarn instances from the left hand list of services –

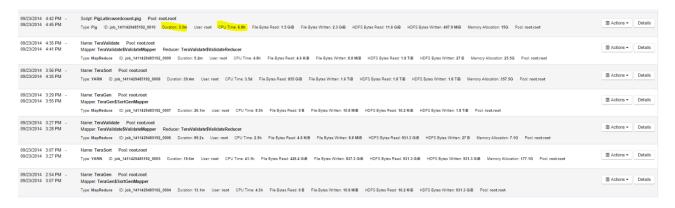3. Click on Applications from the yarn top level menu

4. Select the correct time range to see previously run applications

Set the range to cover at least the entire test run. A list of run applications will then be presented

5. From the list of application, many metrics are available. The most important are CPU time and Time to complete the job as described below. Please note also each benchmark has its own job. Note that the terasort benchmark has 3 phases. Combine all three terasort jobs into a single aggregated result when recording results.
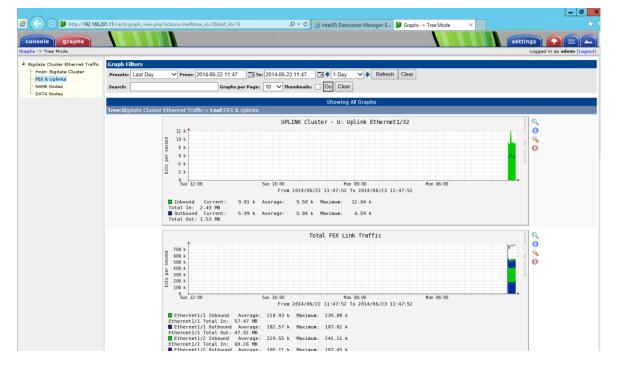

3) *Power management of the cluster:*
   1. Connect to the openVPN.
   2. Connect to DCM-ED via Http://192.168.105.200:5013/DcmConsole (it is not needed to go via server 192.168.201.12)
   3. Log into DCM-ED
   4. Select: Group (left tab)
      o Depending of the test a different "Group" may need to be selected (see table 1)
         ▪ GroupBD
         ▪ GroupBD-8
         ▪ GroupBD-8&7

   5. Select : Temperature/Power

6. Select: Export data
   o The start time (begin point of the collection) and the end time (end point of the collection) should be wide enough
     ▪ The start time: minimum 3 hours before the test started
     ▪ The end time: until the time of collection

7. Specify the correct data range and time range
8. **Attention** do use **Save as** .
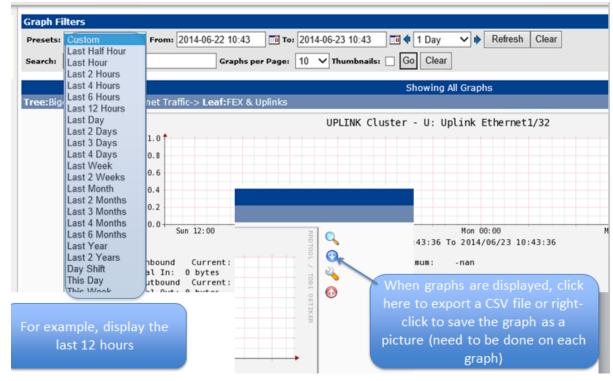9. Save the CSV file with the name **power -<test-name>.CSV**

4) *Network traffic via Cacti*
   1. Connect to Cacti server via Http://192.168.201.11/cacti
   2. Log into Cacti
   3. Select: Graph TAB
   4. Select Flex & Uplink (left tree)

5. Specify date
6. Export file.
   ➔ After graphs are displayed (display last 4 hours)
7. Save the CSV file with the name: **network -<test-name>.CSV**

Presets: Custom
Last Half Hour
Last Hour
Last 2 Hours
Last 4 Hours
Last 6 Hours
Last 12 Hours
Last Day
Last 2 Days
Last 3 Days
Last 4 Days
Last Week
Last 2 Weeks
Last Month
Last 2 Months
Last 3 Months
Last 4 Months
Last 6 Months
Last Year
Last 2 Years
Day Shift
This Day
This Week

For example, display the last 12 hours

When graphs are displayed, click here to export a CSV file or right-click to save the graph as a picture (need to be done on each graph)

Test 1

Description: Testing of Ivy Bridge CPU <u>E5-2695 v2 with Cloudera optimization to12C</u>

Settings		
CPU model	E5-2695 (Ivy Bridge)	
Cluster tuning	yarn.nodemanager.resource.cpu-vcores = 38	
Hyper-threading On / OFF	Set to ON (default setting)	
Network (1Gb or 10Gbit)		10Gb (default setting)
Number of DataNodes / node managers		6 DataNodes - 5 node managers
Number of NameNodes switched ON		2 (default setting)
Number of HDDs used by a DataNode		24 (default setting)
Line rate		4 cables to FEX - 5Gb/S

Test1: Start script benchmark-v2.sh		
	Surname	<>.
	Testname	Test1

Collect Ganglia information	Link: Collect_Ganglia_information
Collect Cloudera report data	Link: Collect_Cloudera_information
Collect Power measurement	Link: Collect_Power_information (GroupBD)
Collect Network measurement	Link: Collect_Network_information

Before moving to any other test, it shall be ensured that the collected data make sense!

Copy the three directories <surname>-test1-1, <surname>-test1-2, <surname>-test1-3 to the SAN.

(See also table 1)

Test 2

Description: Testing of Ivy Bridge CPU <u>E5-2660 v2 with Cloudera optimization to 10C</u>

Settings		
CPU model	E5-2660 v2 (Ivy Bridge)	**Before running the test, the CPU E5-2695v2 needs to be replaced by E5-2660 v2 on all nodes. Clean application shutdown needed!**
Cluster tuning	yarn.nodemanager.resource.cpu-vcores = 32	
Hyper-threading On / OFF	Set to ON (default setting)	
Network (1Gb or 10Gbit)	10Gb (default setting)	
Number of DataNodes / node managers	6 DataNodes - 5 node managers	
Number of NameNodes switched ON	2 (default setting)	
Number of HDDs used by a DataNode	24 (default setting)	
Line rate	4 cables to FEX - 5Gb/S	

Test2: Start script benchmark-v2.sh		
	Surname	<>.
	Testname	Test2

Collect Ganglia information	Link: Collect_Ganglia_information
Collect Cloudera report data	Link: Collect_Cloudera_information
Collect Power measurement	Link: Collect_Power_information (GroupBD)
Collect Network measurement	Link: Collect_Network_information

Before moving to any other test, it shall be ensured that the collected data make sense!

Copy the three directories <surname>-test2-1, <surname>-test2-2, <surname>-test2-3 to the SAN.

(See also table 1)

Test 3

Description: Testing of Ivy Bridge CPU <u>E5-2643 v2 with Cloudera optimization to 6C</u>

Settings			
CPU model	E5-2643 v2 (Ivy Bridge)	**Before running the test the CPU E5-2660 v2 needs to be replaced by E5-2643 v2 on all nodes. Clean application shutdown needed!**	
Cluster tuning	yarn.nodemanager.resource.cpu-vcores = 19		
Hyper-threading On / OFF	Set to ON (default setting)		
Network (1Gb or 10Gbit)	10Gb (default setting)		
Number of DataNodes / node managers	6 DataNodes - 5 node managers		
Number of NameNodes switched ON	2 (default setting)		
Number of HDDs used by a DataNode	24 (default setting)		
Line rate	4 cables to FEX - 5Gb/S		

Test3: Start script benchmark-v2.sh		
	Surname	<>.
	Testname	Test3

Collect Ganglia information	Link: Collect_Ganglia_information
Collect Cloudera report data	Link: Collect_Cloudera_information
Collect Power measurement	Link: Collect_Power_information (GroupBD)
Collect Network measurement	Link: Collect_Network_information

Before moving to any other test, it shall be ensured that the collected data make sense!

Copy the three directories <surname>-test3-1, <surname>-test3-2, <surname>-test3-3 to the SAN.

(See also table 1)

Test 4

Description: Testing of Ivy Bridge CPU <u>E5-2667 v2 with Cloudera optimization to8C</u>

Settings			
CPU model	E5-2667 v2 (Ivy Bridge)	**Before running the test the CPU E5-2643 v2 needs to be replaced by E5-2667 v2 on all nodes. Clean application shutdown needed!**	
Cluster tuning	yarn.nodemanager.resource.cpu-vcores = 26		
Hyper-threading On / OFF	Set to ON (default setting)		
Network (1Gb or 10Gbit)	10Gb (default setting)		
Number of DataNodes / node managers	6 DataNodes - 5 node managers		
Number of NameNodes switched ON	2 (default setting)		
Number of HDDs used by a DataNode	24 (default setting)		
Line rate	4 cables to FEX - 5Gb/S		

Test4: Start script benchmark-v2.sh		
	Surname	<>.
	Testname	Test4

Collect Ganglia information	Link: Collect_Ganglia_information
Collect Cloudera report data	Link: Collect_Cloudera_information
Collect Power measurement	Link: Collect_Power_information (GroupBD)
Collect Network measurement	Link: Collect_Network_information

Before moving to any other test, it shall be ensured that the collected data make sense!
Copy the three directories <surname>-test4-1, <surname>-test4-2, <surname>-test4-3 to the SAN.

(See also table 1)

Test 5

Description: Testing of Ivy Bridge CPU <u>E5-2690 v2 with Cloudera optimization to 10C</u>

Settings			
CPU model	E5-2690 v2 (Ivy Bridge)	**Before running the test, the CPU E5-2667 v2 needs to be replaced by E5-2690 v2 on all nodes. Clean application shutdown needed!**	
Cluster tuning	yarn.nodemanager.resource.cpu-vcores = 32		
Hyper-threading On / OFF	Set to ON (default setting)		
Network (1Gb or 10Gbit)	10Gb (default setting)		
Number of DataNodes / node managers	6 DataNodes - 5 node managers		
Number of NameNodes switched ON	2 (default setting)		
Number of HDDs used by a DataNode	24 (default setting)		
Line rate	4 cables to FEX - 5Gb/S		

Test5: Start script benchmark-v2.sh		
	Surname	<>.
	Testname	Test5

Collect Ganglia information	Link: Collect_Ganglia_information
Collect Cloudera report data	Link: Collect_Cloudera_information
Collect Power measurement	Link: Collect_Power_information (GroupBD)
Collect Network measurement	Link: Collect_Network_information

Before moving to any other test, it shall be ensured that the collected data make sense!
Copy the three directories <surname>-test5-1, <surname>-test5-2, <surname>-test5-3 to the SAN.

(See also table 1)

Test 6

Description: Moving the cluster to full Line rate (10Gbps) by reconfiguring the UCS and connecting 8 cables between FE1 and FI1

Settings		
CPU model	E5-2690v2	
Cluster tuning	yarn.nodemanager.resource.cpu-vcores = 32	
Hyper-threading On / OFF	Set to ON (default setting)	
Network (1Gb or 10Gbit)	10Gb (default setting)	
Number of DataNodes / node managers	6 DataNodes - 5 node managers	
Number of NameNodes switched ON	2 (default setting)	
Number of HDDs used by a DataNode	24 (default setting)	
Line rate	**FEX removed – Server directly connected to FI - Full line rate 10Gb/S**	

Test6: Start script benchmark-v2.sh		
	Surname	<>.
	Testname	Test6

Collect Ganglia information	Link: Collect_Ganglia_information
Collect Cloudera report data	Link: Collect_Cloudera_information
Collect Power measurement	Link: Collect_Power_information (GroupBD)
Collect Network measurement	Link: Collect_Network_information

Before moving to any other test, it shall be ensured that the collected data make sense!
Copy the three directories <surname>-test6-1, <surname>-test6-2, <surname>-test6-3 to the SAN

(See also table 1)

Test 7

Description: Testing the impact of Hyper-Threading

Settings			
CPU model	E5-2690v2		
Cluster tuning	yarn.nodemanager.resource.cpu-vcores = 16		
Hyper-threading On / OFF		**OFF**	**Set to the BIOS parameter: HyperThreading to OFF.** **Clean application shutdown needed!** **Apply the profile "HT-off" via the UCSM to all the servers** **Reboot of all the servers will be enforced by UCSM**
Network (1Gb or 10Gbit)			10Gb (default setting)
Number of DataNodes / node managers			6 DataNodes - 5 node managers
Number of NameNodes switched ON			2 (default setting)
Number of HDDs used by a DataNode			24 (default setting)
Line rate			Full line rate (10Gb/s)

Test7: Start script benchmark-v2.sh		
	Surname	<>.
	Testname	Test7

Collect Ganglia information	Link: Collect_Ganglia_information
Collect Cloudera report data	Link: Collect_Cloudera_information
Collect Power measurement	Link: Collect_Power_information (GroupBD)
Collect Network measurement	Link: Collect_Network_information

Before moving to any other test, it shall be ensured that the collected data make sense!
Copy the three directories <surname>-test7-1, <surname>-test7-2, <surname>-test7-3 to the SAN.

(See also table 1)

Test 8

Description: Testing the impact of changing the server network interfaces from 10Gb to 1Gb.

Settings			
CPU model	E5-2690v2		
Cluster tuning	yarn.nodemanager.resource.cpu-vcores = 32		
Hyper-threading On / OFF	**ON**	**Step 1: Set to the BIOS parameter: HyperThreading to ON.** **Clean application shutdown needed!** **Apply the profile "Bios_BigDataf" via the UCSM to all the servers** **Reboot of all the servers will be enforced by UCSM**	
Network (1Gb or 10Gbit)		**1Gb**	**Step 2: After reboot, reconfigure the network QoS policy to 1Gb through the UCSM Adaptor Performance Profile – QoS Policy = "1GB"**
Number of DataNodes / node managers	6 DataNodes - 5 node managers		
Number of NameNodes switched ON	2 (default setting)		
Number of HDDs used by a DataNode	24 (default setting)		
Line rate	Full line rate (10Gb/s)		

Test8: Start script benchmark-v2.sh		
	Surname	<>.
	Testname	Test8

Collect Ganglia information	Link: Collect_Ganglia_information
Collect Cloudera report data	Link: Collect_Cloudera_information
Collect Power measurement	Link: Collect_Power_information (GroupBD)
Collect Network measurement	Link: Collect_Network_information

Before moving to any other test, it shall be ensured that the collected data make sense!
Copy the three directories <surname>-test8-1, <surname>-test8-2, <surname>-test8-3 to the SAN.

(See also table 1)

Test 9

Description: Testing the impact of the cluster scaling. Reduce the number of DataNodes from 6 to 5.

Settings			
CPU model	E5-2690v2		
Cluster tuning	yarn.nodemanager.resource.cpu-vcores = 32		
Hyper-threading On / OFF		Set to ON (default setting)	
Network (1Gb or 10Gbit)		**10Gb**	**Step 1: Reconfigure the network to 10GB through the UCSM Reconfigure the network QoS policy to <not set> through the UCSM Adaptor Performance Profile – QoS Policy = <not set>**
Number of DataNodes / node managers		**5/5**	**Step 2: Decommission through Cloudera Manager 1 DataNodes. Bring the total number of DataNodes to 5.** **Have on each DataNode, node manager running (5 – 5)!**
Number of NameNodes switched ON		2 (default setting)	
Number of HDDs used by a DataNode		24 (default setting)	
Line rate		Full line rate (10Gb/s)	

Test9: Start script benchmark-v2.sh		
	Surname	<>.
	Testname	Test9

Collect Ganglia information	Link: Collect_Ganglia_information
Collect Cloudera report data	Link: Collect_Cloudera_information
Collect Power measurement	Link: Collect_Power_information (GroupBD-8)
Collect Network measurement	Link: Collect_Network_information

Before moving to any other test, it shall be ensured that the collected data make sense!
Copy the three directories <surname>-test9-1, <surname>-test9-2, <surname>-test9-3 to the SAN.

(See also table 1)

Test 10

Description: Testing the impact of the cluster scaling. Reduce the number of DataNodes from 5 to 4.

Settings			
CPU model	E5-2690v2		
Cluster tuning	yarn.nodemanager.resource.cpu-vcores = 32		
Hyper-threading On / OFF	Set to ON (default setting)		
Network (1Gb or 10Gbit)	10Gb (default setting)		
Number of DataNodes / node managers	4/4	**Step 2: Decommission through Cloudera Manager 1 DataNodes. Bring the total number of DataNodes to 4.** **Have on each DataNode, node manager running (4 – 4)!**	
Number of NameNodes switched ON	2 (default setting)		
Number of HDDs used by a DataNode	24 (default setting)		
Line rate	Full line rate (10Gb/s)		

Test10: Start script benchmark-v2.sh		
	Surname	<>.
	Testname	Test10

Collect Ganglia information	Link: Collect_Ganglia_information
Collect Cloudera report data	Link: Collect_Cloudera_information
Collect Power measurement	Link: Collect_Power_information (GroupBD-8&7)
Collect Network measurement	Link: Collect_Network_information

Before moving to any other test, it shall be ensured that the collected data make sense!
Copy the three directories <surname>-test10-1, <surname>-test10-2, <surname>-test10-3 to the SAN.

(See also table 1)

Test 11

Description: Testing the impact of the cluster scaling. Number of DataNodes set to maximum, 6.

Settings			
CPU model	E5-2690v2		
Cluster tuning	yarn.nodemanager.resource.cpu-vcores = 32		
Hyper-threading On / OFF	Set to ON (default setting)		
Network (1Gb or 10Gbit)	10Gb (default setting)		
Number of DataNodes / node managers	**6/6**	**Step 2: Re-commission through Cloudera Manager the all DataNodes. Bring the total number of DataNodes to 6.** **Have on each DataNode, node manager running (6 –6)!**	
Number of NameNodes switched ON	2 (default setting)		
Number of HDDs used by a DataNode	24 (default setting)		
Line rate	Full line rate (10Gb/s)		

Test11: Start script benchmark-v2.sh		
	Surname	<>.
	Testname	Test11

Collect Ganglia information	Link: Collect_Ganglia_information
Collect Cloudera report data	Link: Collect_Cloudera_information
Collect Power measurement	Link: Collect_Power_information (GroupBD)
Collect Network measurement	Link: Collect_Network_information

Before moving to any other test, it shall be ensured that the collected data make sense!
Copy the three directories <surname>-test11-1, <surname>-test11-2, <surname>-test11-3 to the SAN

(See also table 1)

Test 12

Description: Testing the impact of the number of HDDs per DataNodes. The number of HDDs per DataNode will be reduced from 24 to 12.

Settings		
CPU model	E5-2690v2	
Cluster tuning	yarn.nodemanager.resource.cpu-vcores = 32	
Hyper-threading On / OFF	Set to ON (default setting)	
Network (1Gb or 10Gbit)	10Gb (default setting)	
Number of DataNodes / node managers	6 DataNodes - 6 node managers	
Number of NameNodes switched ON	2 (default setting)	
Number of HDDs used by a DataNode	12	Step 2: Decommission 12 HDDs from every DataNode
Line rate	Full line rate (10Gb/s)	

Test12: Start script benchmark-v2.sh		
	Surname	<>.
	Testname	Test12

Collect Ganglia information	Link: Collect_Ganglia_information
Collect Cloudera report data	Link: Collect_Cloudera_information
Collect Power measurement	Link: Collect_Power_information (GroupBD)
Collect Network measurement	Link: Collect_Network_information

Before moving to any other test, it shall be ensured that the collected data make sense!
Copy the three directories <surname>-test12-1, <surname>-test12-2, <surname>-test12-3 to the SAN

(See also table 1)