

DSC 650

Week 2 Assignment

Eyram Kueviakoe

March 22, 2024

## Deep Dive into HDFS

### Screenshot 1: *hdfs dfsadmin -report*

```
rsa-key-20240315@dsc650-kueviakoe: ~/dsc650-infra/bellevue-bigdata/hadoop-hive-spark-hbase
bash-5.0$ hdfs dfsadmin -report
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/program/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/program/tez/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/program/hive/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
2024-03-22 19:24:37,054 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Configured Capacity: 103670202368 (96.55 GB)
Present Capacity: 79552949328 (74.09 GB)
DFS Remaining: 79172745097 (73.74 GB)
DFS Used: 350204231 (362.59 MB)
DFS Used%: 0.48%
Replicated Blocks:
  Under replicated blocks: 0
  Blocks with corrupt replicas: 0
  Missing blocks: 0
  Missing blocks (with replication factor 1): 0
  Low redundancy blocks with highest priority to recover: 0
  Pending deletion blocks: 0
Erasure Coded Block Groups:
  Low redundancy block groups: 0
  Block groups with corrupt internal blocks: 0
  Missing block groups: 0
  Low redundancy blocks with highest priority to recover: 0
  Pending deletion blocks: 0
-----
Live datanodes (2):
Name: 172.28.1.2:9866 (worker1)
Hostname: worker1
Decommission Status : Normal
Configured Capacity: 51835101184 (48.28 GB)
DFS Used: 141872673 (135.30 MB)
Non DFS Used: 11821674975 (11.01 GB)
DFS Remaining: 39586343235 (36.87 GB)
DFS Used%: 0.27%
DFS Remaining%: 76.37%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceiver: 5
Last contact: Fri Mar 22 19:24:37 GMT 2024
Last Block Report: Fri Mar 22 19:17:59 GMT 2024
Num of Blocks: 125
```

```
rsa-key-20240315@dsc650-kueviakoe: ~/dsc650-infra/bellevue-bigdata/hadoop-hive-spark-hbase
Block groups with corrupt internal blocks: 0
Missing block groups: 0
Low redundancy blocks with highest priority to recover: 0
Pending deletion blocks: 0

-----
Live datanodes (2):
Name: 172.28.1.2:9866 (worker1)
Hostname: worker1
Decommission Status : Normal
Configured Capacity: 51835101184 (48.28 GB)
DFS Used: 141872673 (135.30 MB)
Non DFS Used: 11821674975 (11.01 GB)
DFS Remaining: 39586343235 (36.87 GB)
DFS Used%: 0.27%
DFS Remaining%: 76.37%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 5
Last contact: Fri Mar 22 19:24:37 GMT 2024
Last Block Report: Fri Mar 22 19:17:59 GMT 2024
Num of Blocks: 125

Name: 172.28.1.3:9866 (worker2)
Hostname: worker2
Decommission Status : Normal
Configured Capacity: 51835101184 (48.28 GB)
DFS Used: 238331558 (227.29 MB)
Non DFS Used: 11725216090 (10.92 GB)
DFS Remaining: 39586369094 (36.87 GB)
DFS Used%: 0.46%
DFS Remaining%: 76.37%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 5
Last contact: Fri Mar 22 19:24:38 GMT 2024
Last Block Report: Fri Mar 22 19:17:59 GMT 2024
Num of Blocks: 151

bash-5.0#
```

Screenshot 2: Screenshot proving the data has been loaded  
*hdfs dfs -ls /*

```
rsa-key-20240315@dsc650-kueviakoe: ~/dsc650-infra/bellevue-bigdata/hadoop-hive-spark-hbase
bash-5.0# hdfs dfs -ls /
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/program/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/program/tez/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/program/hive/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
2024-03-25 02:13:25,975 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 7 items
-rw-r--r-- 1 root supergroup          747 2024-03-25 02:12 /grades.csv
drwxr-xr-x  - root supergroup           0 2024-03-25 02:07 /hbase
drwxr-xr-x  - root supergroup           0 2024-03-25 02:05 /log
drwxr-xr-x  - root supergroup           0 2024-03-25 02:06 /spark-jars
drwxr-xr-x  - root supergroup           0 2024-03-25 02:06 /tez
drwxr-xr-x  - root supergroup           0 2024-03-25 02:09 /tmp
drwxrwx---  - root supergroup           0 2024-03-25 02:06 /user
bash-5.0#
```

Screenshot:

*docker-compose exec worker1 bash*  
*hdfs dfs -ls /*

and

*docker-compose exec worker2 bash*  
*hdfs dfs -ls /*

```

rsa-key-20240315@dsc650-kueviakoe: ~/dsc650-infra/bellevue-bigdata/hadoop-hive-spark-hbase
bash-5.0# hdfs dfs -ls /
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/program/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/program/tez/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/program/hive/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
2024-03-25 02:39:45,072 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 6 items
-rw-r--r-- 1 root supergroup          747 2024-03-25 02:39 /grades.csv
drwxr-xr-x - root supergroup          0 2024-03-25 02:39 /hbase
drwxr-xr-x - root supergroup          0 2024-03-25 02:39 /log
drwxr-xr-x - root supergroup          0 2024-03-25 02:39 /spark-jars
drwxr-xr-x - root supergroup          0 2024-03-25 02:38 /tmp
drwxrwx--- - root supergroup          0 2024-03-25 02:39 /user
bash-5.0# exit
exit
rsa-key-20240315@dsc650-kueviakoe: ~/dsc650-infra/bellevue-bigdata/hadoop-hive-spark-hbase$ sudo docker-compose exec worker2 bash
bash-5.0# hdfs dfs -ls /
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/program/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/program/tez/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/program/hive/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
2024-03-25 02:40:06,629 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 6 items
-rw-r--r-- 1 root supergroup          747 2024-03-25 02:39 /grades.csv
drwxr-xr-x - root supergroup          0 2024-03-25 02:39 /hbase
drwxr-xr-x - root supergroup          0 2024-03-25 02:39 /log
drwxr-xr-x - root supergroup          0 2024-03-25 02:40 /spark-jars
drwxr-xr-x - root supergroup          0 2024-03-25 02:38 /tmp
drwxrwx--- - root supergroup          0 2024-03-25 02:40 /user
bash-5.0#

```

Screenshots of the three chosen HDFS command outputs.

Command 1

*hdfs dfs -mkdir assignment2*

and

*hdfs dfs -ls*

to show the folder was created

```

rsa-key-20240315@dsc650-kueviakoe: ~/dsc650-infra/bellevue-bigdata/hadoop-hive-spark-hbase
bash-5.0# hdfs dfs -mkdir assignment2
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/program/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/program/tez/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/program/hive/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
2024-03-22 20:28:25,865 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
bash-5.0# hdfs dfs -ls
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/program/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/program/tez/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/program/hive/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
2024-03-22 20:28:41,673 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 3 items
drwxr-xr-x - root supergroup          0 2024-03-22 19:19 .hiveJars
drwxr-xr-x - root supergroup          0 2024-03-22 20:28 assignment2
-rw-r--r-- 1 root supergroup          747 2024-03-22 19:35 grades.csv
bash-5.0#

```

## Command 2

*hdfs dfs -getfacl assignment2*

```
bash-5.0# hdfs dfs -getfacl assignment2
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/program/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/program/tez/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/program/hive/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
2024-03-22 20:30:22,904 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
# file: assignment2
# owner: root
# group: supergroup
user::rwx
group::r-x
other::r-x
bash-5.0#
```

## Command 3

*hdfs dfs -rm -r assignment2*

and

*hdfs dfs -ls*

to show the folder is deleted

```
bash-5.0# hdfs dfs -rm -r assignment2
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/program/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/program/tez/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/program/hive/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
2024-03-22 20:31:08,013 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Deleted assignment2
bash-5.0# hdfs dfs -ls
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/program/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/program/tez/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/program/hive/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
2024-03-22 20:31:15,772 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 2 items
drwxr-xr-x   - root supergroup          0 2024-03-22 19:19 .hiveJars
-rw-r--r--   1 root supergroup       747 2024-03-22 19:35 grades.csv
bash-5.0#
```

Screenshot

yarn node -list

```
rsa-key-20240315@dsc650-kueviakoe: ~/dsc650-infra/bellevue-bigdata/hadoop-hive-spark-hbase
bash-5.0# yarn node -list
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/program/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/program/tez/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/program/hive/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
2024-03-22 20:34:40,135 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
2024-03-22 20:34:40,308 INFO client.RMProxy: Connecting to ResourceManager at master/172.28.1.1:8032
Total Nodes:2
Node-Id      Node-State Node-Http-Address  Number-of-Running-Containers
worker1:43507 RUNNING     worker1:8042       0
worker2:38701 RUNNING     worker2:8042       0
bash-5.0#
```

Screenshot from the YARN UI showing the updated maximum memory

All Applications

localhost:8088/cluster

hadoop

Cluster

About

Nodes

Node Labels

Applications

NEW

NEW SAVING

SUBMITTED

ACCEPTED

RUNNING

FINISHED

FAILED

KILLED

Scheduler

Tools

All Applications

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Used Resources	Total Resources	Reserved Resources	Physical Memory
0	0	0	0	0	<memory:0 B, vCores:0>	<memory:8 GB, vCores:16>	<memory:0 B, vCores:0>	85

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Nodes
2	0	0	0	0	0

User Metrics for dr.who

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Containers Pending	Containers Reserved	Memory Used	Memory Pending	Memory Reserved	vCores
0	0	0	0	0	0	0	0 B	0 B	0 B	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation	Maximum Cluster Application P
Fair Scheduler	[memory-mb (unit=Mi), vcores]	<memory:512, vCores:1>	<memory:2048, vCores:4>	0

Show 20 entries

ID	User	Name	Application Type	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCores	Allocated Memory MB	Allocated GPUs	Reserved CPU VCores	Reserved Memory MB	Reserved GPUs	% of Queue
No data available in table																		

Showing 0 to 0 of 0 entries

Experimenting with MapReduce

```
rsa-key-20240315@dsc650-kueviako: ~/dsc650-infra/bellevue-bigdata/hadoop-hive-spark-hbase
Data-local map tasks=1
Rack-local map tasks=1
Total time spent by all maps in occupied slots (ms)=22494
Total time spent by all reduces in occupied slots (ms)=13716
Total time spent by all map tasks (ms)=11247
Total time spent by all reduce tasks (ms)=3429
Total vcore-milliseconds taken by all map tasks=11247
Total vcore-milliseconds taken by all reduce tasks=3429
Total megabyte-milliseconds taken by all map tasks=11516928
Total megabyte-milliseconds taken by all reduce tasks=7022592
Map-Reduce Framework
  Map input records=2
  Map output records=4
  Map output bytes=36
  Map output materialized bytes=56
  Input split bytes=286
  Combine input records=0
  Combine output records=0
  Reduce input groups=2
  Reduce shuffle bytes=56
  Reduce input records=4
  Reduce output records=0
  Spilled Records=8
  Shuffled Maps =2
  Failed Shuffles=0
  Merged Map outputs=2
  GC time elapsed (ms)=251
  CPU time spent (ms)=2030
  Physical memory (bytes) snapshot=744058880
  Virtual memory (bytes) snapshot=7062241280
  Total committed heap usage (bytes)=67040064
  Peak Map Physical memory (bytes)=285478912
  Peak Map Virtual memory (bytes)=2355625984
  Peak Reduce Physical memory (bytes)=180830208
  Peak Reduce Virtual memory (bytes)=2354847744
Shuffle Errors
  BAD_ID=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONG_REDUCE=0
File Input Format Counters
  Bytes Read=236
File Output Format Counters
  Bytes Written=97
Job Finished in 31.133 seconds
Estimated value of Pi is 3.800000000000000000000000
bash-5.0#
```

```
WRONG_REDUCE=0
File Input Format Counters
  Bytes Read=236
File Output Format Counters
  Bytes Written=97
Job Finished in 31.133 seconds
Estimated value of Pi is 3.800000000000000000000000
bash-5.0#
```

### Summary and significance of the result:

The estimated Pi value of 3.8 from the MapReduce job is higher than the actual Pi value. This means that the calculation wasn't very accurate.

There are many reasons that can explain the inaccuracy:

- We are not using enough samples or data points
- The random numbers are not properly generated.

To get an accurate result, we will need a larger sample. This will reduce the estimation error. Also the random number generation should be tuned to generate points that are evenly distributed.