Week 2 Assignment

Guruprasad Velikadu Krishnamoorthy

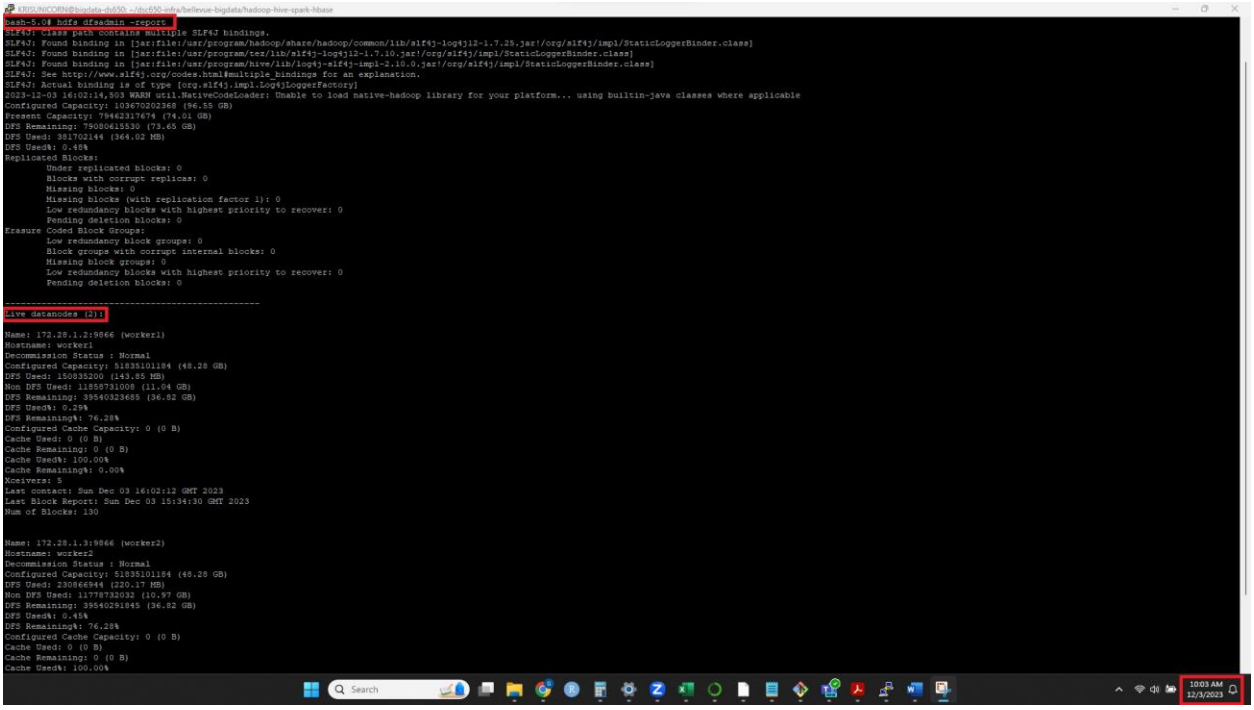
College of Science and Technology, Bellevue University

DSC650-T301: Big Data

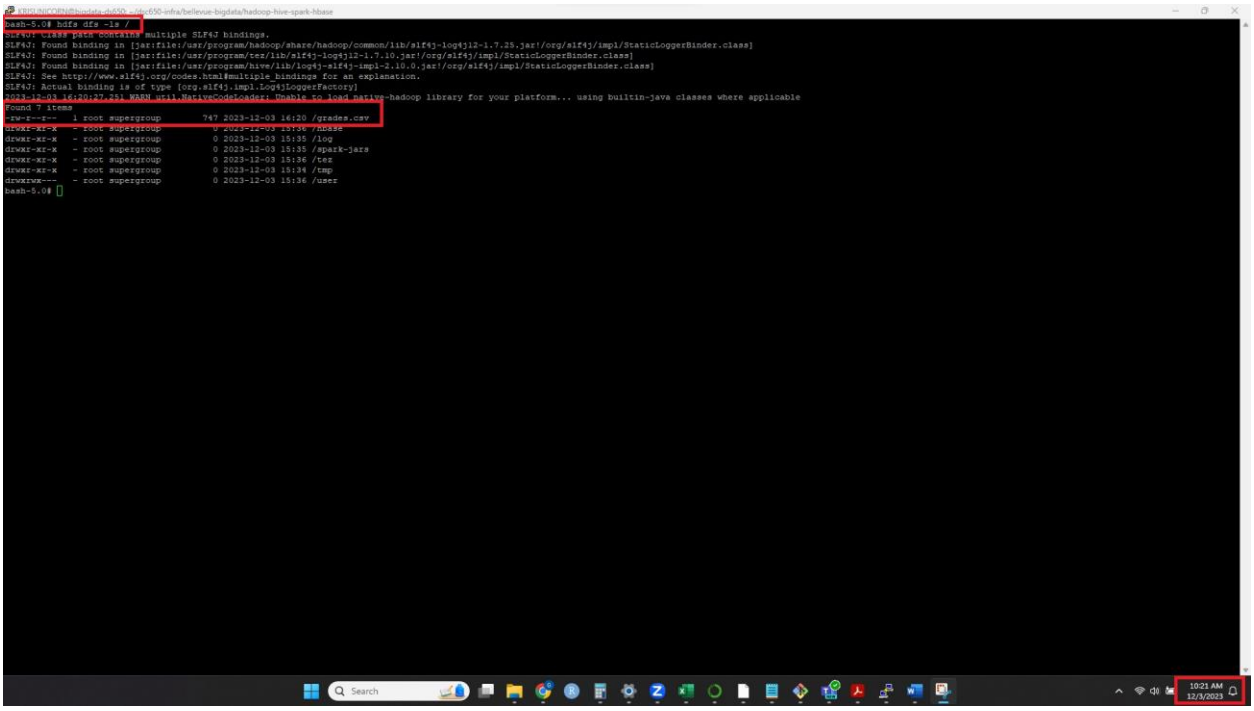
Professor. Nasheb Ismaily

December 8, 2023

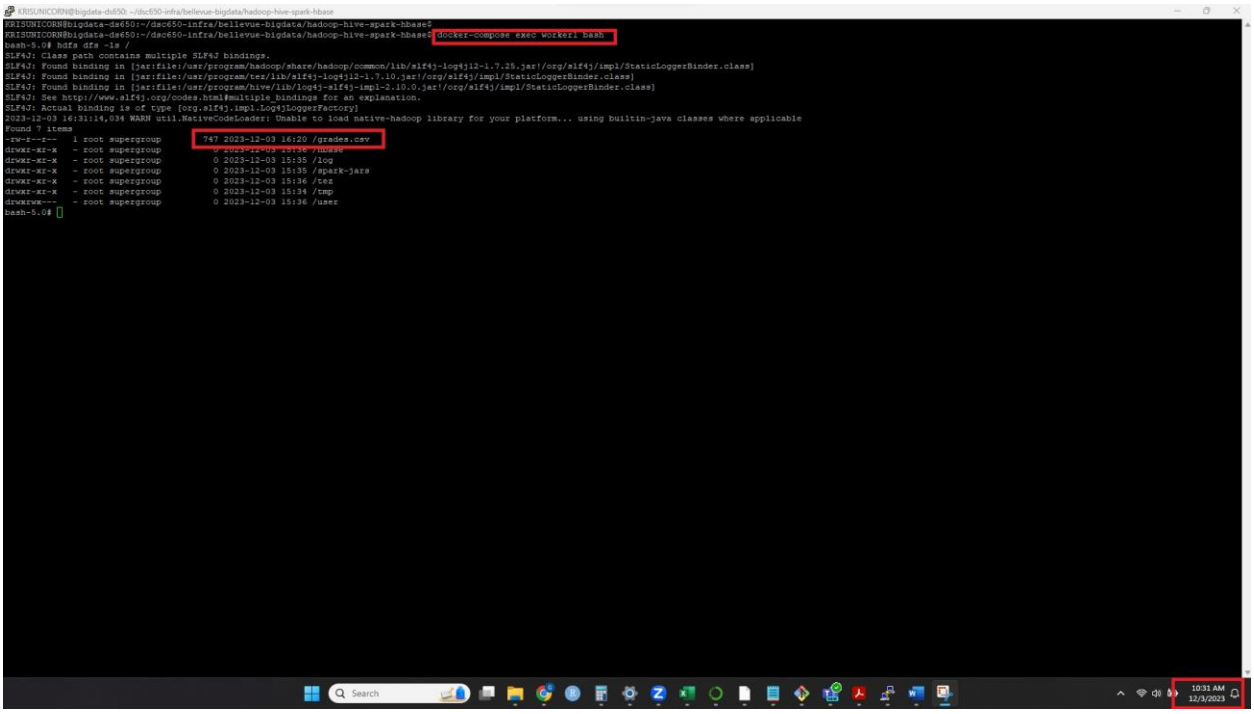
1. Screenshot of dfsadmin report command:



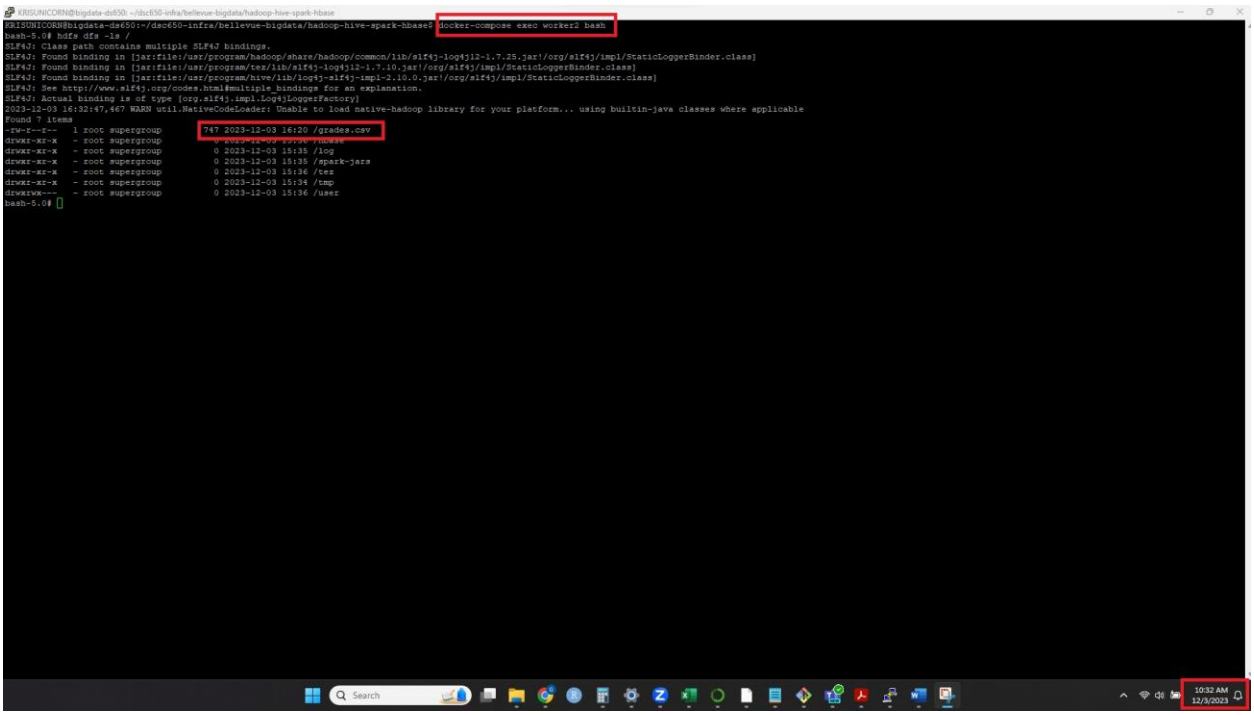
2.1. Screenshot showing the grades.csv file is copied in Hadoop master node / directory:



2.2. Screenshot showing the grades.csv file is present in worker1 node / directory:

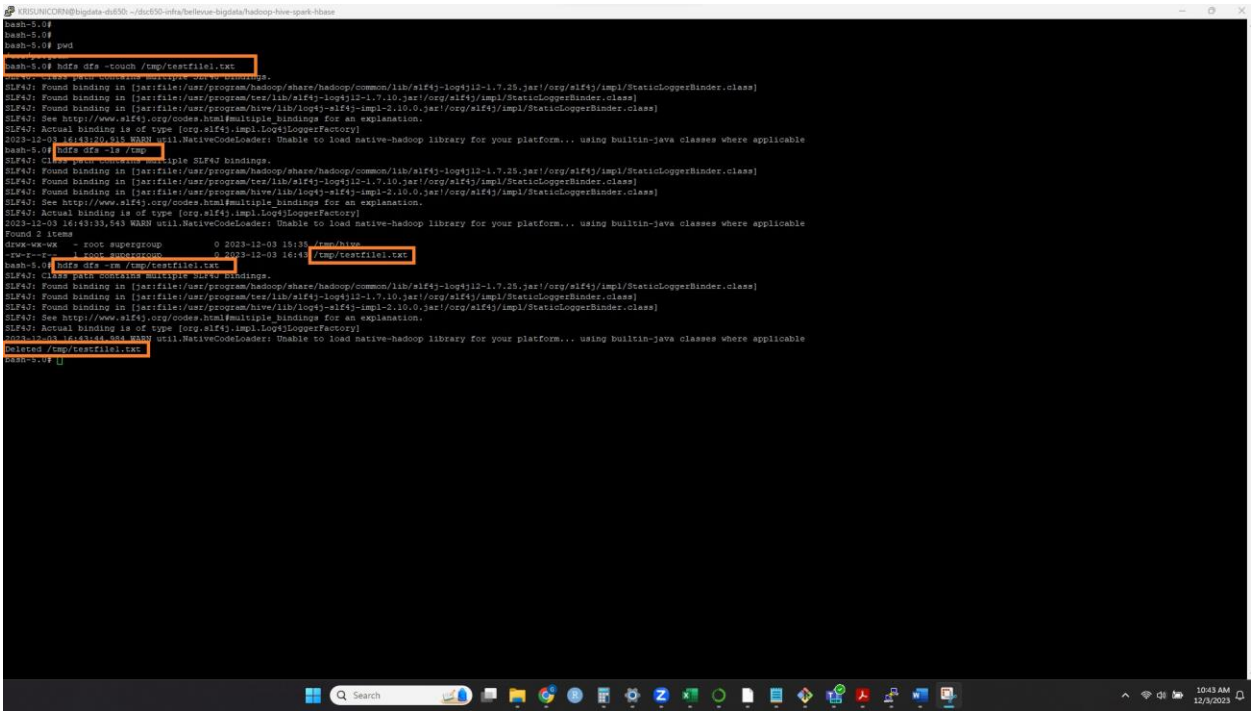


2.3. Screenshot showing the grades.csv file is present in worker2 node / directory:



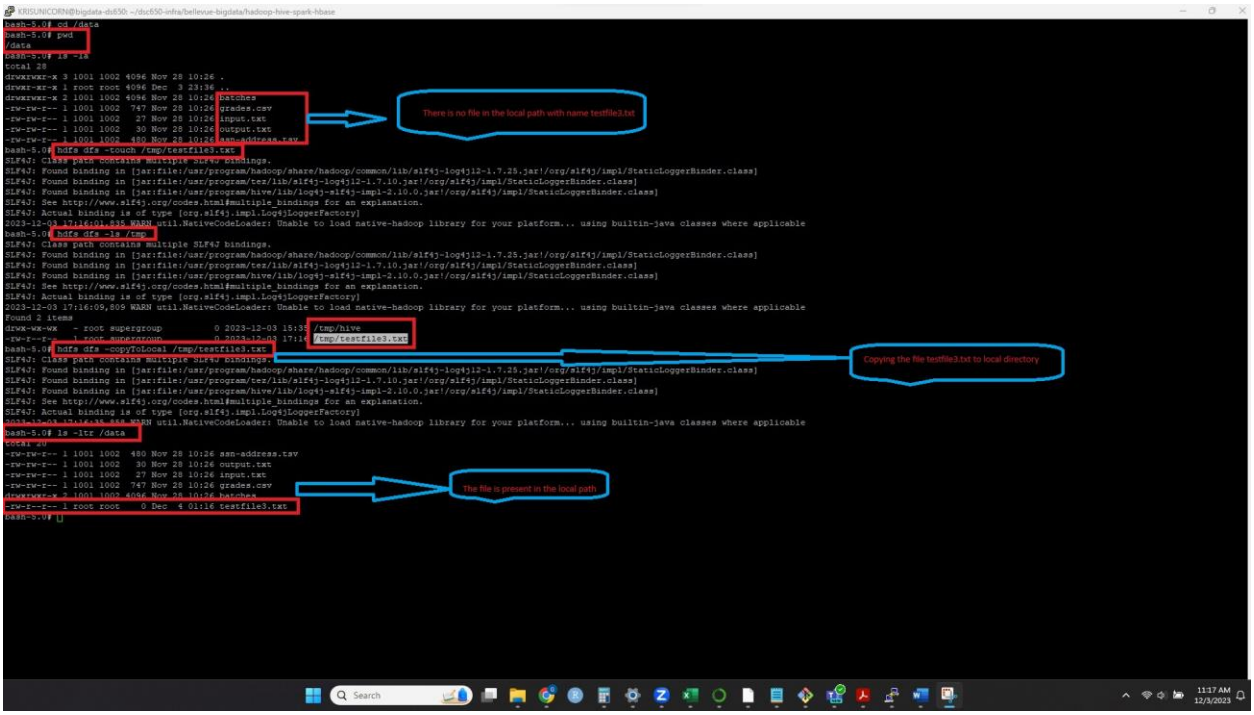
3.1. Screenshot of HDFS command 1 → touch & rm commands.

A new file named testfile1.txt was created in hdfs /tmp directory using touch command and removed using rm command.



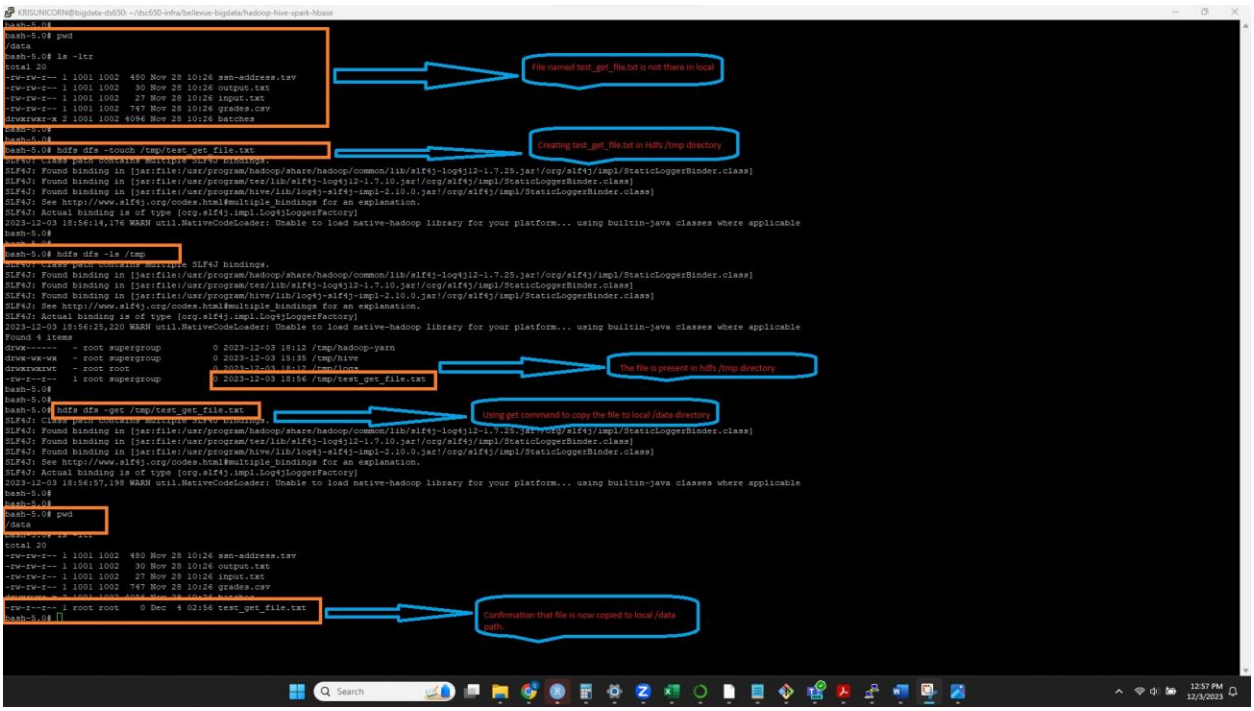
3.2. Screenshot of HDFS command 2 → copyToLocal command.

A new file testfile3.txt was created in hdfs /tmp directory and was copied to local /data directory using copyToLocal command

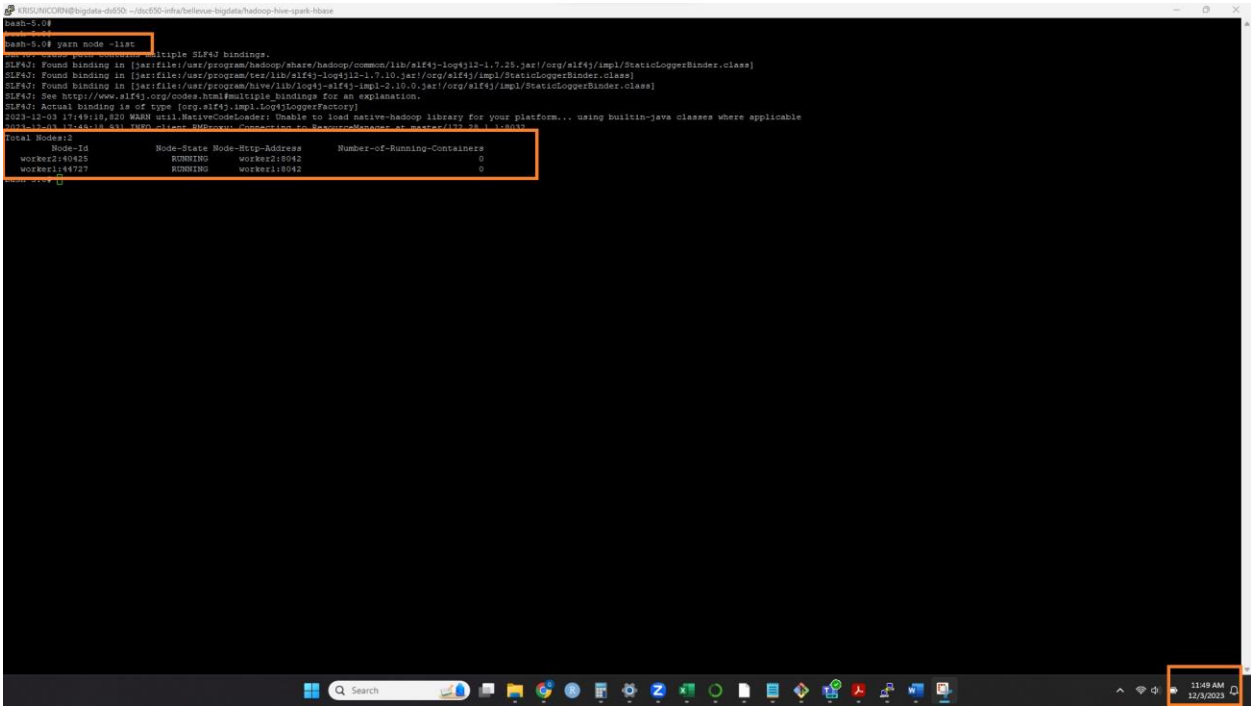


3.3. Screenshot of HDFS command 3 → get command.

A new file test_get_file.txt was created in hdfs /tmp directory and was copied to local /data directory using get command

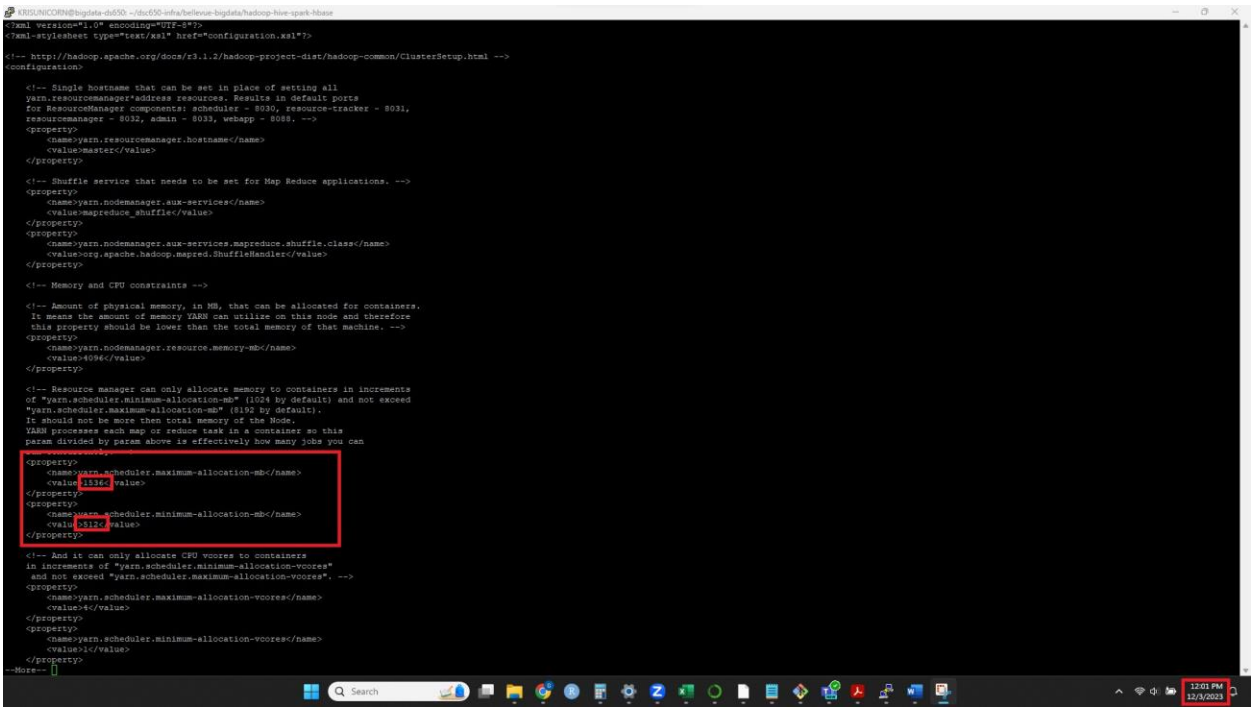


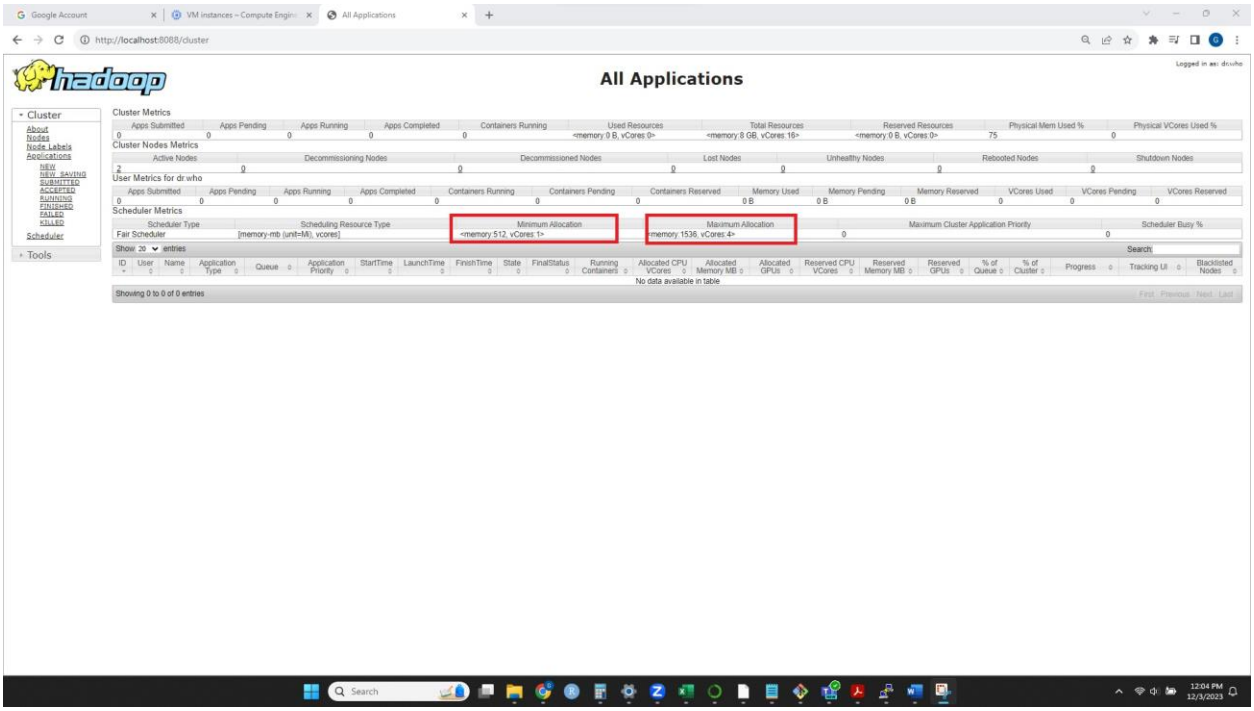
4. **Screenshot of yarn list command:**



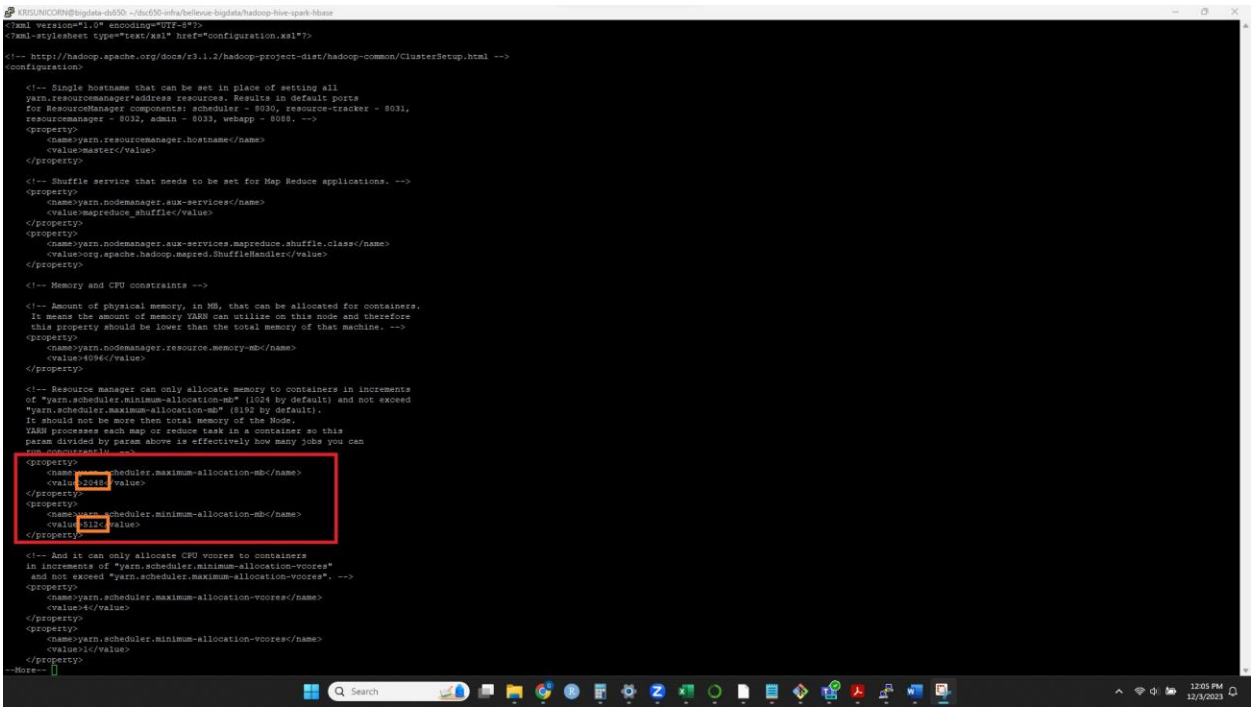
5. Screenshots of Yarn yarn.scheduler.maximum-allocation-mb :

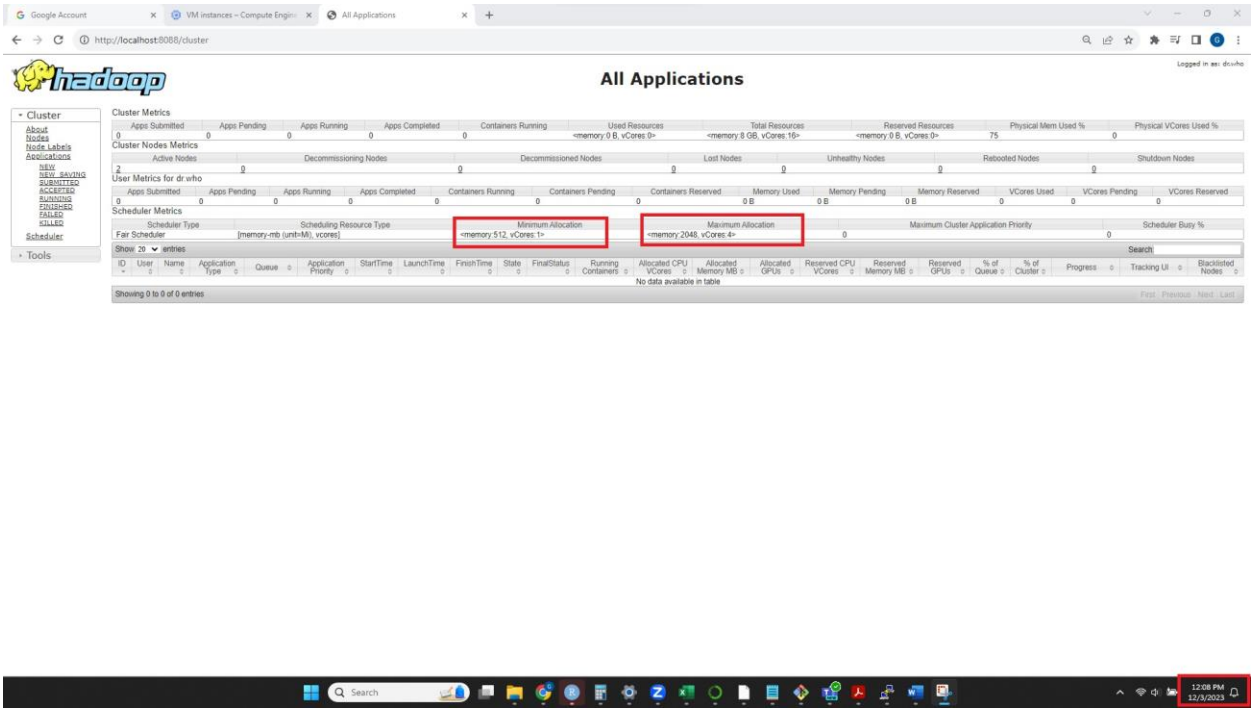
5.1. Before making the change:



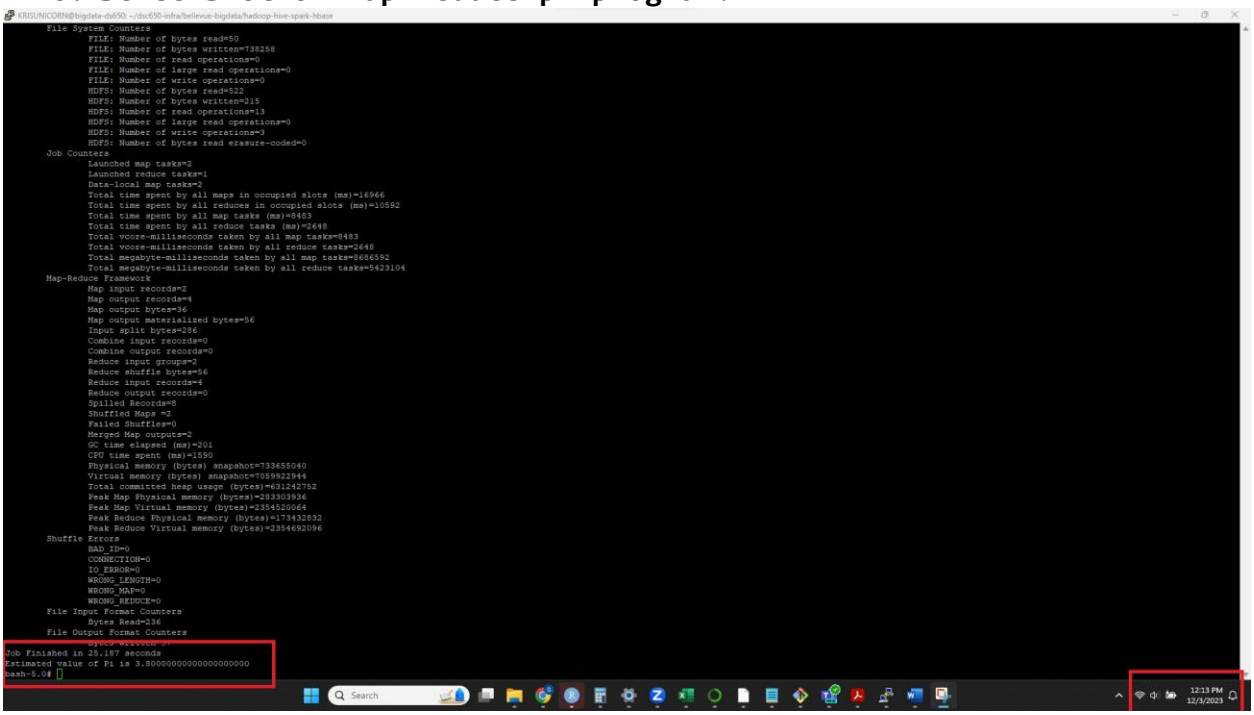


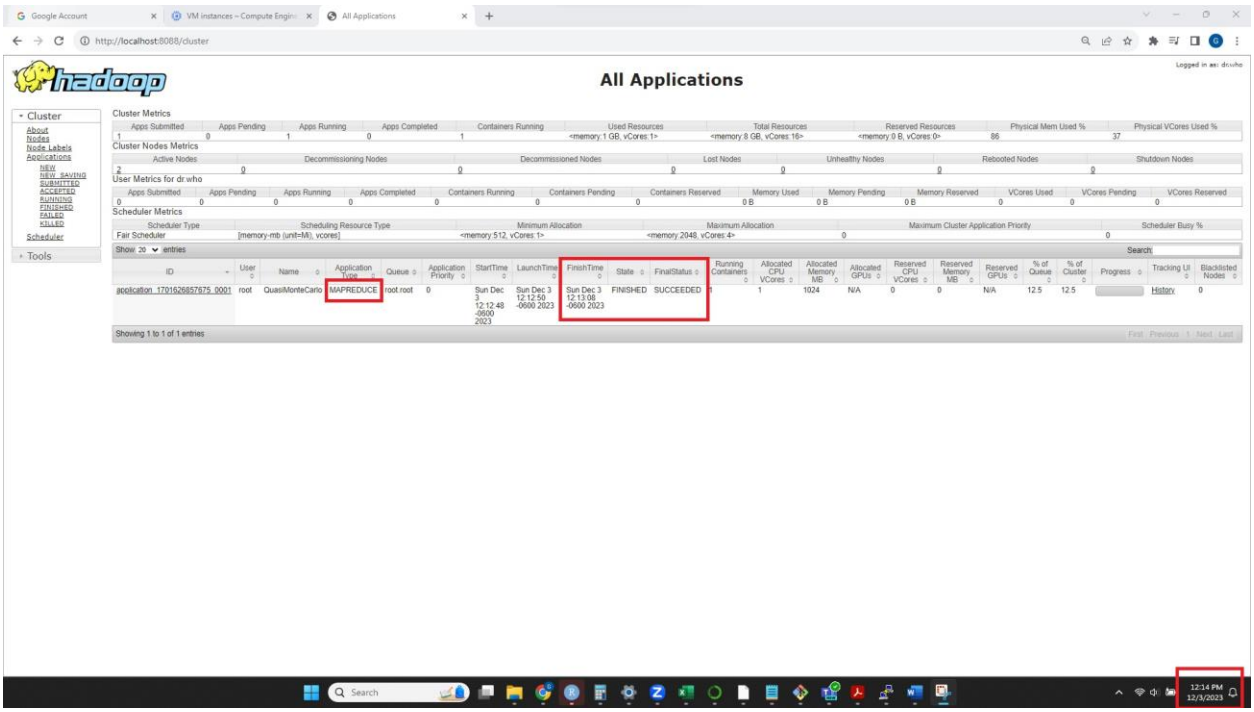
5.2. After making the changes:


KRISUNICORN@bigdata-ds650: ~/dsc650-infra/bellevue-bigdata/hadoop-hive-spark-hbase
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!-- http://hadoop.apache.org/docs/r3.1.2/hadoop-project-dist/hadoop-common/ClusterSetup.html -->
<configuration>
<!-- Single hostname that can be set in place of setting all
yarn.resourcemanager*address resources. Results in default ports
for ResourceManager components: scheduler - 8030, resource-tracker - 8031,
resourcemanager - 8032, admin - 8033, webapp - 8088. -->
<property>
<name>yarn.resourcemanager.hostname</name>
<value>master</value>
</property>
<!-- Shuffle service that needs to be set for Map Reduce applications. -->
<property>
<name>yarn.nodemanager.aux-services</name>
<value>mapreduce_shuffle</value>
</property>
<property>
<name>yarn.nodemanager.aux-services.mapreduce.shuffle.class</name>
<value>org.apache.hadoop.mapred.ShuffleHandler</value>
</property>
<!-- Memory and CPU constraints -->
<!-- Amount of physical memory, in MB, that can be allocated for containers.
It means the amount of memory YARN can utilize on this node and therefore
this property should be lower than the total memory of that machine. -->
<property>
<name>yarn.nodemanager.resource.memory-mb</name>
<value>4096</value>
</property>
<!-- Resource manager can only allocate memory to containers in increments
of "yarn.scheduler.minimum-allocation-mb" (1024 by default) and not exceed
"yarn.scheduler.maximum-allocation-mb" (8192 by default).
It should not be more then total memory of the Node.
YARN processes each map or reduce task in a container so this
param divided by param above is effectively how many jobs you can
run concurrently on a
<property>
<name>yarn.scheduler.maximum-allocation-mb</name>
<value>2048</value>
</property>
<property>
<name>yarn.scheduler.minimum-allocation-mb</name>
<value>512</value>
</property>
<!-- And it can only allocate CPU vcores to containers
in increments of "yarn.scheduler.minimum-allocation-vcores"
and not exceed "yarn.scheduler.maximum-allocation-vcores". -->
<property>
<name>yarn.scheduler.maximum-allocation-vcores</name>
<value>4</value>
</property>
<property>
<name>yarn.scheduler.minimum-allocation-vcores</name>
<value>1</value>
</property>
--More--
Search
12:05 PM
12/3/2023


Google Account
VM instances – Compute Engine
All Applications
http://localhost:8088/cluster
hadoop
All Applications
Logged in as: dr.who
Cluster
About
Nodes
Node Labels
Applications
NEW
NEW_SAVING
SUBMITTED
ACCEPTED
RUNNING
FINISHED
FAILED
KILLED
Scheduler
Tools
Cluster Metrics
Apps Submitted | Apps Pending | Apps Running | Apps Completed | Containers Running | Used Resources | Total Resources | Reserved Resources | Physical Mem Used % | Physical VCores Used %
0 | 0 | 0 | 0 | 0 | <memory:0 B, vCores:0> | <memory:8 GB, vCores:16> | <memory:0 B, vCores:0> | 75 | 0
Cluster Nodes Metrics
Active Nodes | Decommissioning Nodes | Decommissioned Nodes | Lost Nodes | Unhealthy Nodes | Rebooted Nodes | Shutdown Nodes
2 | 0 | 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 Used | VCores Pending | VCores Reserved
0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 B | 0 B | 0 B | 0 | 0 | 0
Scheduler Metrics
Scheduler Type | Scheduling Resource Type | Minimum Allocation | Maximum Allocation | Maximum Cluster Application Priority | Scheduler Busy %
Fair Scheduler | [memory-mb (unit=Mi), vcores] | <memory:512, vCores:1> | <memory:2048, vCores:4> | 0 | 0
Show 20 entries
Search
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 | % of Cluster | Progress | Tracking UI | Blacklisted Nodes
No data available in table
Showing 0 to 0 of 0 entries
First Previous Next Last
Search
12:08 PM
12/3/2023

6. Screenshot of map-reduce pi program:





Significance and summary of the result:

The hadoop-mapreduce-examples-3.2.3.jar program is used to calculate the pi value using quasi-monte Carlo method. It takes 2 arguments, the first one is the number of maps and the second one is the number of samples per map. The

numerical approximation of arbitrary examples method is used in this program. The Mapper class generates points in the Unit square and then counts points inside or outside of the inscribed circle of the square. The Reducer combines the points inside or outside results from the mappers. Once the results are calculated it is sent to the output file.

Formula Calculation:

In this method, the value of pi is calculated using the formula $4I$.So the value can be calculated once the value of I is calculated.
The fraction numInside/numTotal is approximately equivalent to (Area of the circle)/(Area of the square) which is the same as I.

numInside/numTotal = (Area of the circle)/(Area of the square) = I
The Area of the unit square is 1.
The Area of the inscribed circle is pi/4

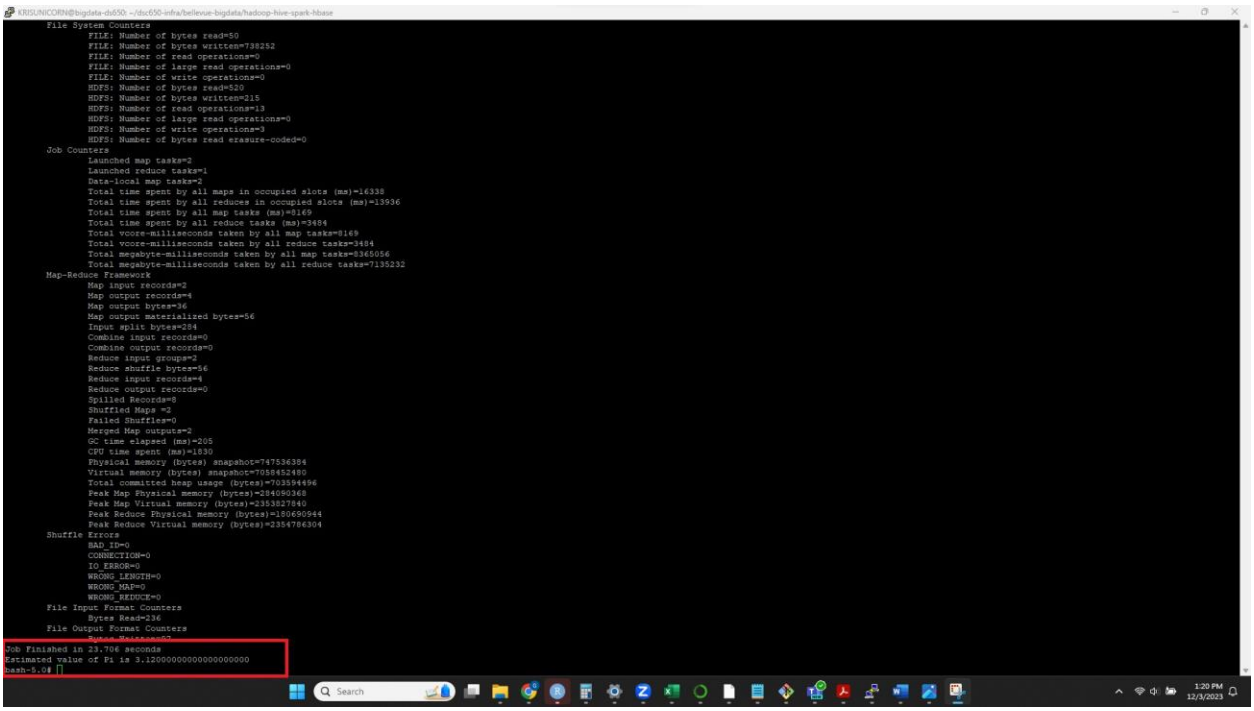
So, the equation can be broken down as follows:

numInside/numTotal = (pi/4) = I

so, pi = 4(numInside/numTotal)

Observations:

In the first screenshot, we used 2 maps and 10 samples per map and the value was calculated as 3.8 and the program completed in 25.1 seconds. In the second test run, 2 maps with 100 samples per map were used (the screenshot below) and the value pi was calculated as 3.12 in 23.7 seconds.


KRISUNICORN@bigdata-ds650: ~/dsc650-infra/bellevue-bigdata/hadoop-hive-spark-hbase
File System Counters
FILE: Number of bytes read=50
FILE: Number of bytes written=738252
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=520
HDFS: Number of bytes written=215
HDFS: Number of read operations=13
HDFS: Number of large read operations=0
HDFS: Number of write operations=3
HDFS: Number of bytes read erasure-coded=0
Job Counters
Launched map tasks=2
Launched reduce tasks=1
Data-local map tasks=2
Total time spent by all maps in occupied slots (ms)=16338
Total time spent by all reduces in occupied slots (ms)=13936
Total time spent by all map tasks (ms)=8169
Total time spent by all reduce tasks (ms)=3484
Total vcore-milliseconds taken by all map tasks=8169
Total vcore-milliseconds taken by all reduce tasks=3484
Total megabyte-milliseconds taken by all map tasks=8365056
Total megabyte-milliseconds taken by all reduce tasks=7135232
Map-Reduce Framework
Map input records=2
Map output records=4
Map output bytes=36
Map output materialized bytes=56
Input split bytes=284
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)=205
CPU time spent (ms)=1830
Physical memory (bytes) snapshot=747536384
Virtual memory (bytes) snapshot=7058452480
Total committed heap usage (bytes)=703594496
Peak Map Physical memory (bytes)=284090368
Peak Map Virtual memory (bytes)=2353827840
Peak Reduce Physical memory (bytes)=180690944
Peak Reduce Virtual memory (bytes)=2356776304
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
Job Finished in 23.706 seconds
Estimated value of Pi is 3.12000000000000000000
bash-5.0#
Search
1:20 PM
12/3/2023