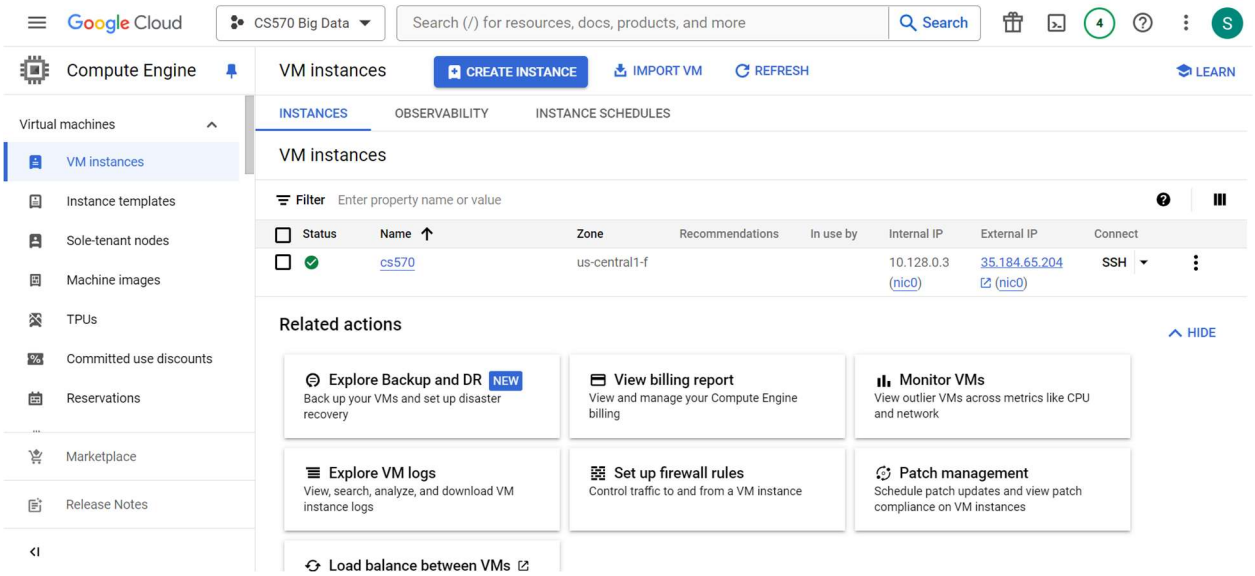
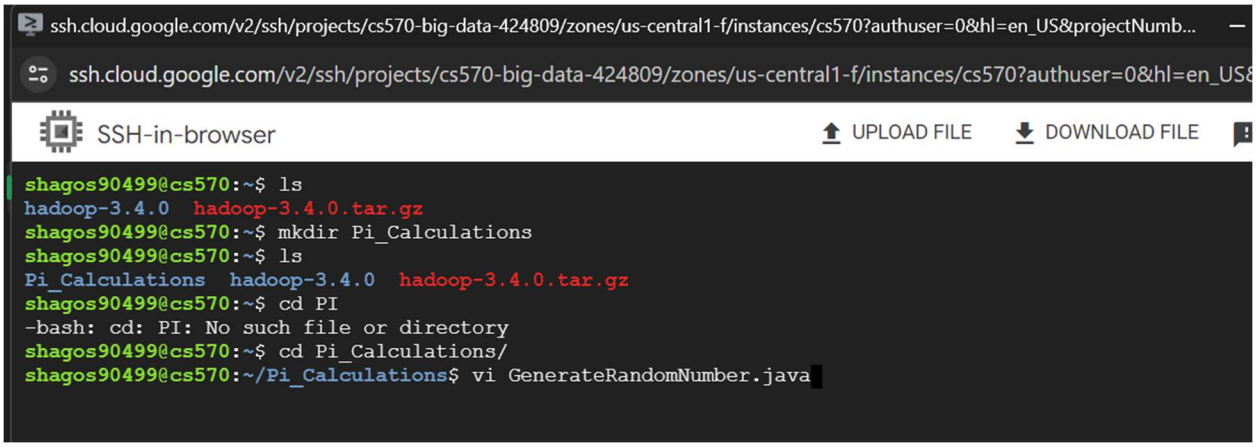
Week 3: Homework 1: Project: Creating MapReduce program to calculating Pi

GCP instance:



1. Creating folder Pi_calculations and program to generate random numbers in (x,y) format.



The program to generate random numbers:

```java
import java.io.File;
import java.io.FileWriter;
import java.io.IOException;
import java.util.Scanner;

public class GenerateRandomNumbers {
    public static void main(String[] args) {
        System.out.println("How many random numbers to generate:"); // we use 1000000 to test
        Scanner input = new Scanner(System.in);
        int RandomNumCount = input.nextInt();

        System.out.println("What's the radius?"); //we use 200 to test
        int radius = input.nextInt();
        int diameter = radius * 2;
        input.close();

        try {

            // it creates file input4
            File file = new File("./PiCalculationInput");
            file.createNewFile();

            // Prepare input data
            FileWriter writer = new FileWriter(file);
            //writer.write(radius + "\r\n");
            //writer.write(System.getProperty("line.separator"));

            for (int i = 0; i < RandomNumCount; i++) {
                int xvalue = (int) (Math.random() * diameter);
                int yvalue = (int) (Math.random() * diameter);
                writer.write("(" + xvalue + "," + yvalue + ") ");
            }

            // send the data into the file
```

Generate an input file to the Pi MapReduce program and compiling the program:

```
shagos90499@cs570:~/Pi_Calculations$ vi GenerateRandomNumbers.java
shagos90499@cs570:~/Pi_Calculations$ javac GenerateRandomNumbers.java
shagos90499@cs570:~/Pi_Calculations$ java -cp . GenerateRandomNumbers
How many random numbers to generate:
1000000
What's the radius?
200
shagos90499@cs570:~/Pi_Calculations$ ls
GenerateRandomNumbers.class  GenerateRandomNumbers.java  PiCalculationInput
shagos90499@cs570:~/Pi_Calculations$
```

Random numbers generated:

```
7,297) (265,396) (236,33) (159,310) (323,287) (56,67) (102,8) (303,198) (335,34) (16,27) (260,1) (348,287) (324,282) (89,392) (331,2
86) (134,345) (347,334) (108,328) (330,253) (170,383) (246,345) (226,283) (293,181) (133,119) (263,139) (106,392) (135,281) (308,275
) (285,54) (272,16) (329,149) (204,310) (348,235) (273,244) (127,319) (267,75) (362,87) (46,99) (111,71) (317,178) (318,95) (162,82)
 (382,152) (39,351) (310,340) (320,6) (184,378) (149,399) (236,113) (366,141) (68,190) (171,35) (325,339) (229,26) (197,313) (38,376
) (144,245) (29,180) (92,189) (55,397) (235,153) (146,44) (22,357) (250,132) (234,49) (330,8) (338,223) (228,104) (222,11) (213,246)
 (358,204) (369,183) (163,330) (122,18) (94,104) (87,351) (376,24) (343,352) (233,80) (185,377) (99,179) (341,368) (270,267) (19,269
) (90,200) (286,373) (388,27) (313,320) (117,24) (179,169) (3,287) (144,11) (296,169) (352,302) (353,124) (316,375) (120,73) (13,116
) (253,65) (273,113) (336,2) (211,121) (208,272) (143,235) (114,366) (112,337) (184,257) (78,135) (76,341) (342,341) (42,121) (284,7
5) (211,224) (94,65) (139,152) (82,373) (257,22) (356,70) (91,158) (35,82) (174,100) (156,150) (203,9) (50,51) (83,199) (382,0) (171
,134) (86,157) (305,32) (268,181) (150,44) (212,191) (257,382) (113,130) (198,256) (10,399) (267,103) (321,49) (187,262) (285,155) (
41,159) (120,11) (399,123) (75,161) (48,289) (256,127) (283,147) (117,67) (169,34) (28,112) (208,12) (369,287) (63,325) (341,281) (2
82,3) (142,114) (195,358) (156,126) (57,323) (237,319) (194,160) (218,305) (217,144) (250,116) (73,167) (24,388) (250,166) (233,90)
(374,22) (122,388) (203,137) (73,377) (188,116) (318,231) (330,111) (387,281) (100,276) (49,196) (123,298) (80,16) (14,119) (127,261
) (335,78) (181,169) (298,91) (300,398) (99,36) (89,168) (22,290) (37,363) (138,186) (50,342) (117,308) (383,1) (364,49) (198,267) (
221,184) (92,30) (285,349) (185,188) (64,304) (83,120) (304,180) (320,399) (335,139) (22,23) (115,30) (56,374) (85,262) (92,379) (10
8,308) (139,341) (228,124) (328,354) (170,341) (51,84) (375,243) (6,129) (395,111) (79,266) (362,281) (39,397) (297,293) (223,52) (3
32,256) (285,296) (33,258) (286,136) (94,178) (124,123) (34,204) (357,352) (252,336) (379,61) (349,76) (225,359) (13,153) (380,267)
(295,54) (346,384) (360,12) (14,157) (329,249) (289,332) (118,171) (372,18) (130,193) (287,21) (180,190) (242,203) (201,378) (25,259
) (33,371) (244,32) (386,201) (84,140) (45,31) (66,253) (13,394) (374,41) (261,325) (394,181) (51,95) (373,330) (150,228) (9,383) (1
98,18) (119,248) (278,391) (210,219) (288,99) (152,376) (132,89) (70,98) (328,72) (359,164) (65,170) (71,380) (360,348) (133,113) (2
05,221) (223,165) (256,248) (381,308) (81,267) (13,102) (42,364) (3,269) (22,213) (95,275) (253,319) (365,39) (272,350) (61,275) (32
4,65) (204,72) (282,325) (3,333) (321,213) (283,26) (323,111) (19,19) (199,137) (20,242) (74,171) (351,371) (110,327) (29,203) (143,
134) (130,184) (145,354) (186,356) (160,214) (399,89) (17,302) (155,197) (399,55) (398,183) (95,50) (204,392) (396,252) (188,223) (7
1,44) (220,37) (78,231) (375,347) (91,143) (176,243) (297,316) (237,340) (13,395) (154,129) (92,307) (7,5) (236,107) (157,219) (136,
157) (398,57) (362,67) (44,324) (331,89) (151,296) (127,101) (8,314) (383,349) (75,290) (271,209) (120,245) (337,114) (155,305) (381
,48) (377,57) (84,378) (14,328) (372,43) (367,304) (235,118) (295,320) (48,262) (188,335) (378,351) (247,131) (8,143) (344,387) (92,
19) (86,28) (119,228) (42,198) (157,86) (11,62) (258,365) (220,22) (319,387) (251,227) (112,142) (136,35) (58,222) (220,235) (149,23
1) (27,219) (111,225) (154,265) (357,185) (95,126) (200,315) (309,99) (254,367) (117,388) (125,244) (98,214) (167,168) (119,381) (25
8,151) (39,43) (287,386) (246,132) (5,249) (292,362) (96,114) (39,241) (77,46) (68,30) (76,226) (263,124) (130,19) (273,233) (87,397
) (95,163) (220,357) (281,257) (136,344) (216,246) (220,288) (196,141) (114,37) (20,87) (265,392) (91,42) (26,154) (72,22) (164,36)
(195,256) (303,138) (245,266) (231,120) (102,333) (293,120) (52,314) (99,272) (394,215) (351,280) (214,286) (150,85) (129,54) (207,1
00) (177,98) (326,125) (195,240) (31,72) (14,52) (373,119) (192,68) (288,145) (32,125) (58,231) (223,2) (185,41) (319,161) (20,258)
(271,166) (218,343) (142,43) (12,29) (338,195) (9,82) (48,239) (164,61) (222,94) (345,320) (362,10) (246,349) (308,245) (20,63) (339
,267) (359,129) (56,392) (171,192) (279,175) (223,313) (150,348) (124,367) (233,276) (302,23) (393,69) (38,363) (235,241) (123,326)
(331,24) (162,207) (276,220) (40,390) (64,103) (193,383) (355,98) (392,177) (40,210) (348,269) (232,227) (111,283) (292,166) (248,25
```

Now lets go back to Hadoop and see if we can connect to the localhost:

```
shagos90499@cs570:~$ pwd
/home/shagos90499
shagos90499@cs570:~$ ls
Pi_Calculations  hadoop-3.4.0  hadoop-3.4.0.tar.gz
shagos90499@cs570:~$ cd hadoop-3.4.0/
shagos90499@cs570:~/hadoop-3.4.0$ ssh localhost
Welcome to Ubuntu 20.04.6 LTS (GNU/Linux 5.15.0-1060-gcp x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/pro

 System information as of Mon Jun  3 09:47:33 UTC 2024

  System load:  0.0               Processes:             109
  Usage of /:   53.7% of 9.51GB   Users logged in:       1
  Memory usage: 8%                IPv4 address for ens4: 10.128.0.3
  Swap usage:   0%


Expanded Security Maintenance for Applications is not enabled.

15 updates can be applied immediately.
7 of these updates are standard security updates.
To see these additional updates run: apt list --upgradable

Enable ESM Apps to receive additional future security updates.
See https://ubuntu.com/esm or run: sudo pro status

New release '22.04.3 LTS' available.
Run 'do-release-upgrade' to upgrade to it.


Last login: Mon Jun  3 09:20:39 2024 from 127.0.0.1
shagos90499@cs570:~$
```

Next, we will create the HDFS directories required to execute MapReduce jobs:

```
shagos90499@cs570:~/hadoop-3.4.0$ bin/hdfs dfs -mkdir /user
shagos90499@cs570:~/hadoop-3.4.0$ bin/hdfs dfs -mkdir /user/shagos90499
shagos90499@cs570:~/hadoop-3.4.0$ bin/hdfs dfs -mkdir /user/shagos90499/picalculate
shagos90499@cs570:~/hadoop-3.4.0$ bin/hdfs dfs -mkdir /user/shagos90499/picalculate/input
shagos90499@cs570:~/hadoop-3.4.0$ bin/hdfs dfs -put ../Pi_Calculations/PiCalculationInput /user/shagos90499/picalculate/input
shagos90499@cs570:~/hadoop-3.4.0$
```

Step 2: create a MapReduce program to calculate the numbers of inside darts and outside darts.

```
shagos90499@cs570:~/hadoop-3.4.0$ vi PiCalculation.java
shagos90499@cs570:~/hadoop-3.4.0$ bin/hadoop com.sun.tools.javac.Main PiCalculation.java
shagos90499@cs570:~/hadoop-3.4.0$ jar cf wc.jar PiCalculation*class
shagos90499@cs570:~/hadoop-3.4.0$ ls
 LICENSE-binary  'PiCalculation$IntSumReducer.class'     README.txt   index.html   licenses-binary   share
 LICENSE.txt     'PiCalculation$TokenizerMapper.class'   bin          input        logs              wc.jar
 NOTICE-binary    PiCalculation.class                    etc          lib          output
 NOTICE.txt       PiCalculation.java                     include      libexec      sbin
shagos90499@cs570:~/hadoop-3.4.0$
```

Step 3: Use the file generated in Step 1.2 as the input to execute the MapReduce program created in Step 2:

```
shagos90499@cs570:~/hadoop-3.4.0$ bin/hadoop jar wc.jar PiCalculation /user/shagos90499/picalculate/input /user/90499/picalculate/output5
2024-06-03 10:32:44,508 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2024-06-03 10:32:44,643 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2024-06-03 10:32:44,643 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2024-06-03 10:32:44,959 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execut
e your application with ToolRunner to remedy this.
2024-06-03 10:32:45,255 INFO input.FileInputFormat: Total input files to process : 1
2024-06-03 10:32:45,293 INFO mapreduce.JobSubmitter: number of splits:1
2024-06-03 10:32:45,498 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1485709741_0001
2024-06-03 10:32:45,498 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-06-03 10:32:45,804 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2024-06-03 10:32:45,807 INFO mapreduce.Job: Running job: job_local1485709741_0001
2024-06-03 10:32:45,812 INFO mapred.LocalJobRunner: OutputCommitter set in config null
2024-06-03 10:32:45,844 INFO output.PathOutputCommitterFactory: No output committer factory defined, defaulting to FileOutputCommitterFactory
2024-06-03 10:32:45,849 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2024-06-03 10:32:45,850 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cle
anup failures: false
2024-06-03 10:32:45,852 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
```

Step 4: Calculate Pi in the driver program based on the numbers of inside darts and outside darts

Output:

```
shagos90499@cs570:~/hadoop-3.4.0$ bin/hdfs dfs -ls /user/90499/picalculate/output5
Found 2 items
-rw-r--r--   1 shagos90499 supergroup          0 2024-06-03 10:32 /user/90499/picalculate/output5/_SUCCESS
-rw-r--r--   1 shagos90499 supergroup         29 2024-06-03 10:32 /user/90499/picalculate/output5/part-r-00000
shagos90499@cs570:~/hadoop-3.4.0$ bin/hdfs dfs -cat /user/90499/picalculate/output5/part-r-00000
inside  785377
outside 214623
shagos90499@cs570:~/hadoop-3.4.0$
```

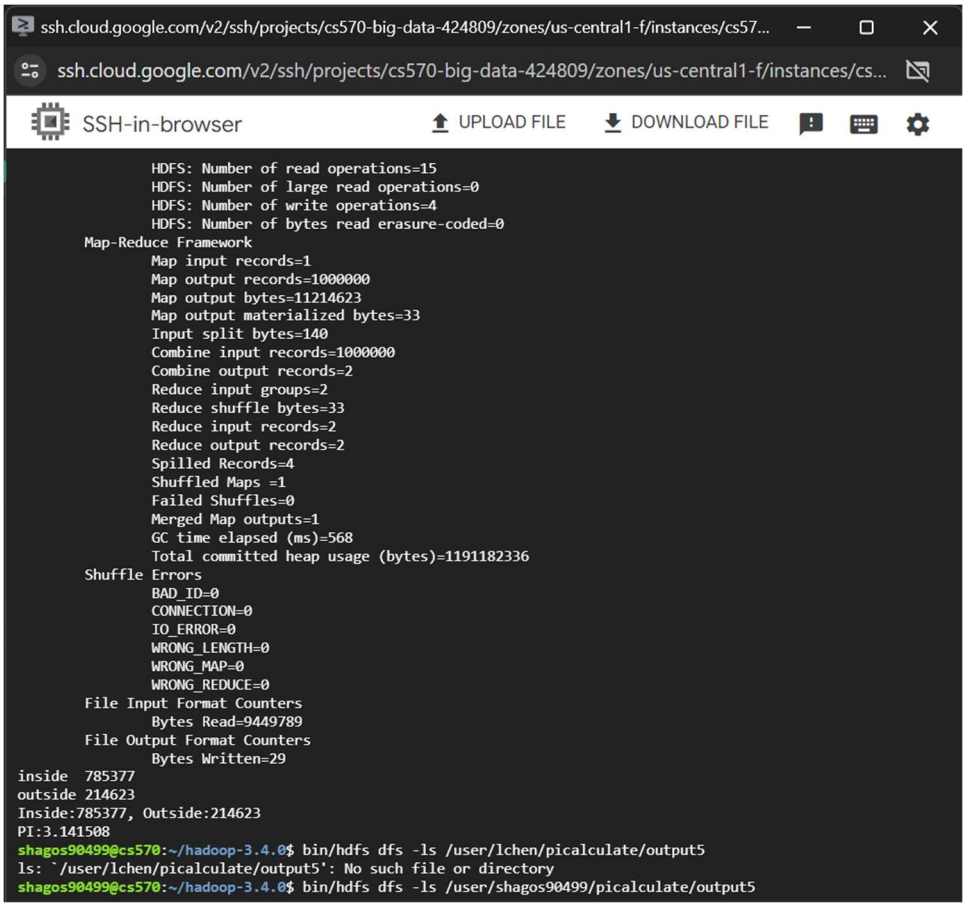
Test results:

```
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=9449789
        File Output Format Counters
                Bytes Written=29
inside  785377
outside 214623
Inside:785377, Outside:214623
PI:3.141508
shagos90499@cs570:~/hadoop-3.4.0$ bin/hdfs
```

Stop machine:

```
shagos90499@cs570:~/hadoop-3.4.0$ sbin/stop-dfs.sh
Stopping namenodes on [localhost]
Stopping datanodes
Stopping secondary namenodes [cs570]
shagos90499@cs570:~/hadoop-3.4.0$
```

Detailed result:



Shut down VM: