PI CALCULATION USING MAPREDUCE AND PYSPARK

CS570 Big Data Processing Project

By Arsiema Yohannes

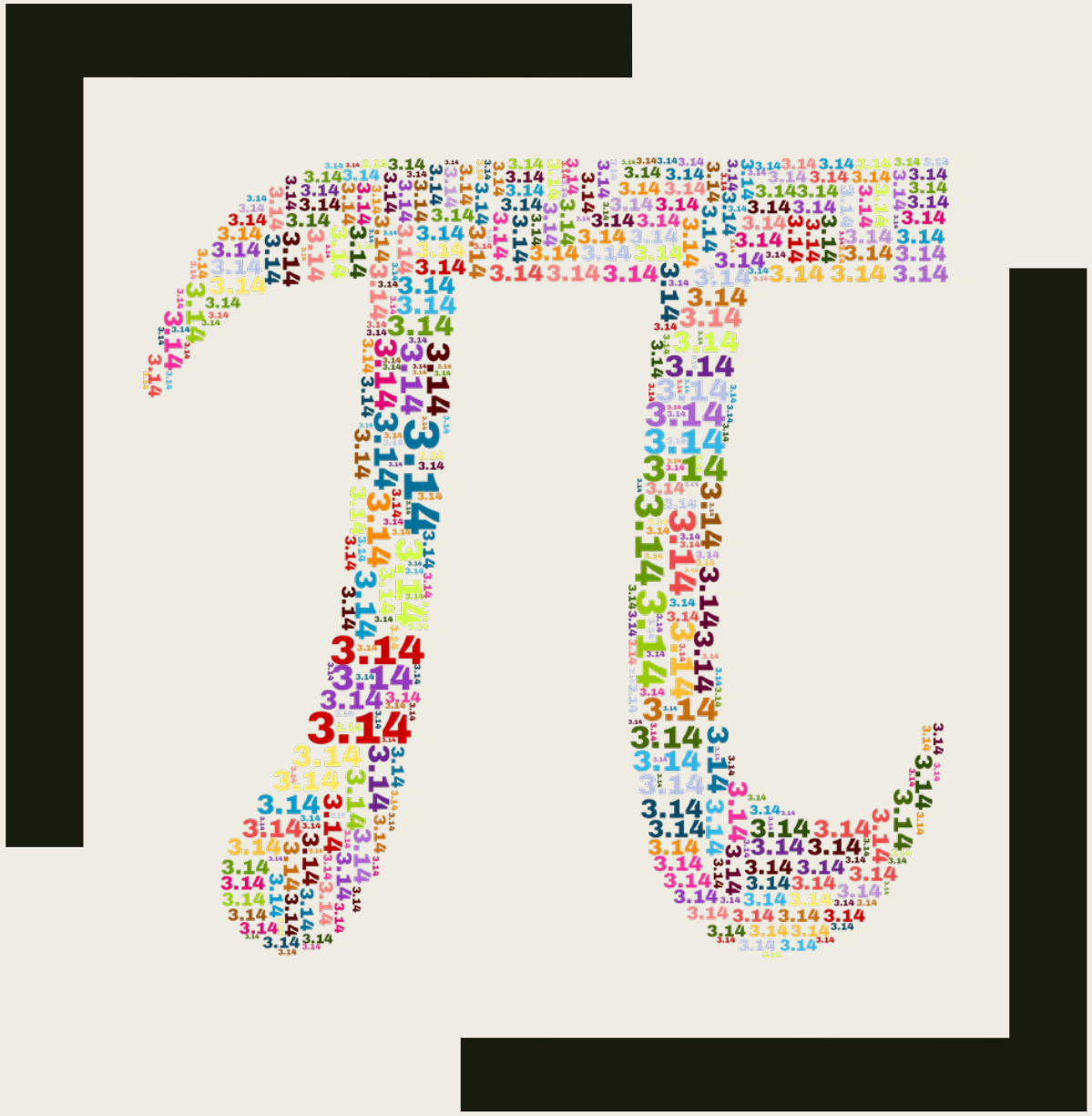
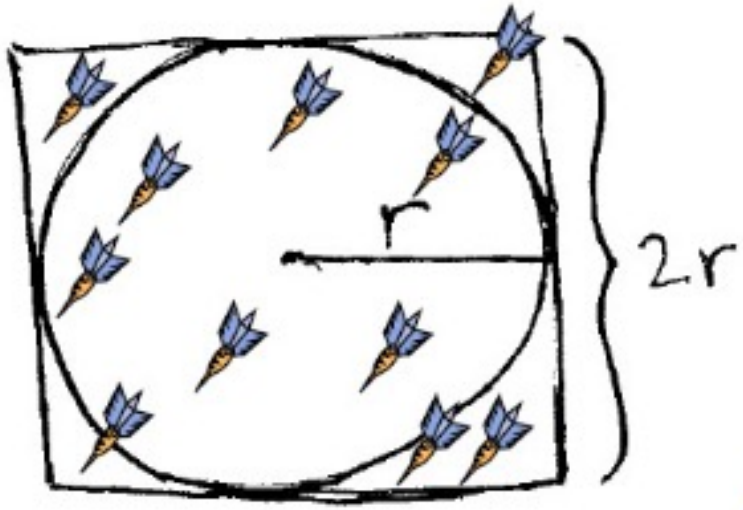


Table of Content

Introduction

- Objective: Calculate Pi using Hadoop
- Note: Hadoop is not ideal for computationally intensive tasks, but this project aims to demonstrate its capabilities
- Based on a practice question to gain hands-on experience with Hadoop technology

- Throw N darts on the board. Each dart lands at a random position (x,y) on the board.
  - Note if each dart landed inside the circle or not
    - Check if $x^2+y^2<r$
  - Take the total number of darts that landed in the circle as S



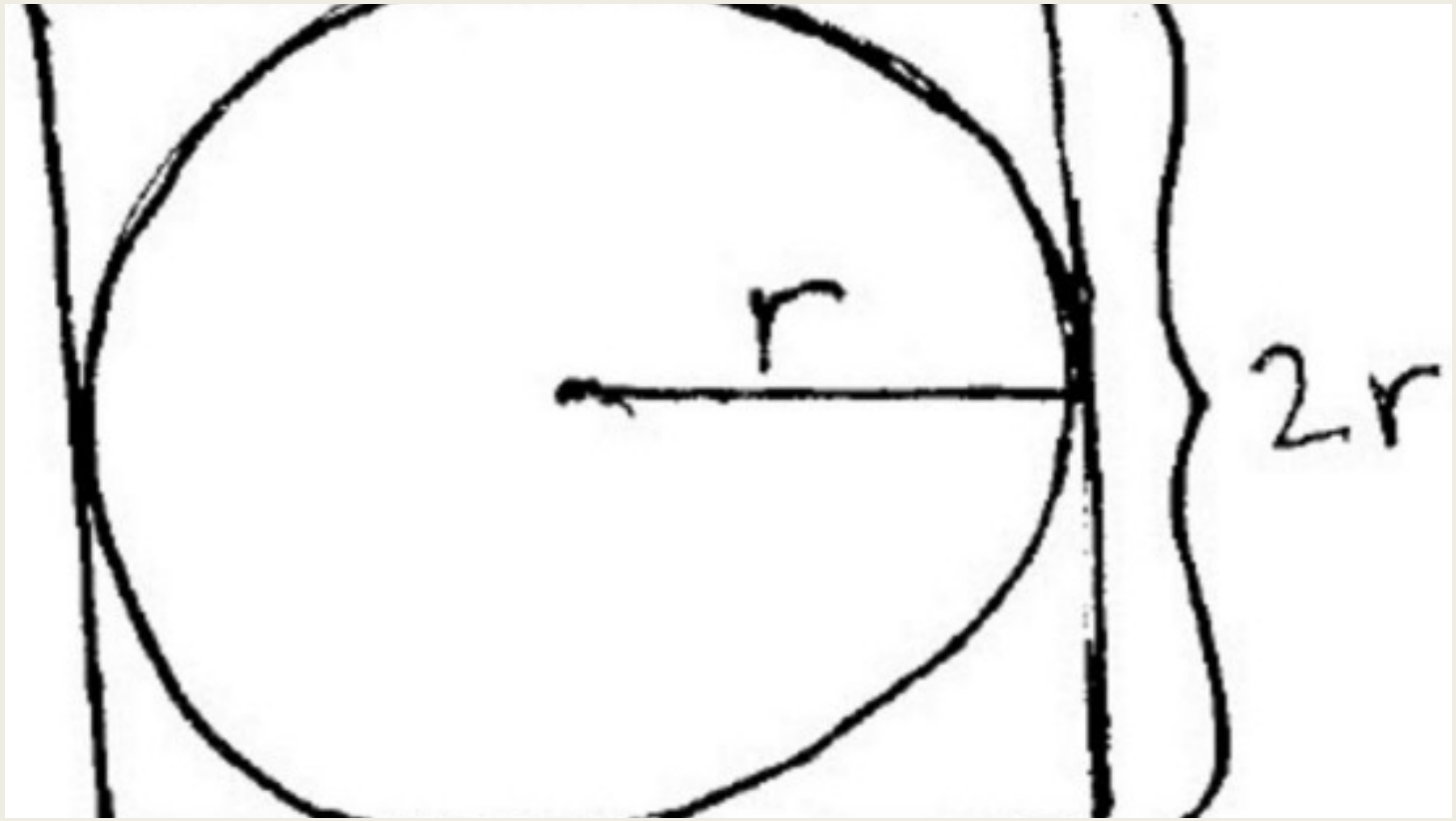
$$4\left(\frac{S}{N}\right)=\pi$$

Formula:

4 * S / N = 4 * (pi * r * r) / (4 * r * r) = pi

Pi Calculation

- As the illustrated on the right, the value of pi can be calculated by counting the number of random darts that falls in the circle and outside the circle.



Pi Calculation

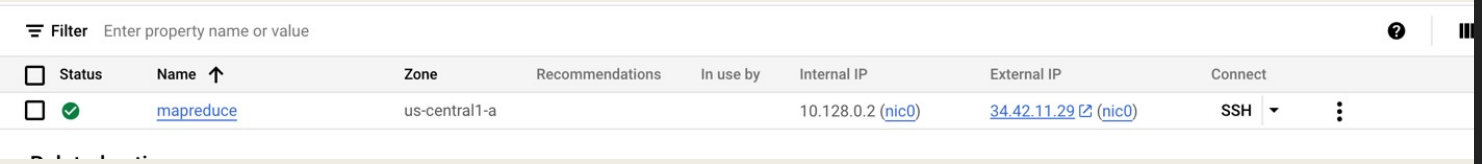
- How to check if one dart (x,y) lands inside the circle with radius r or not?

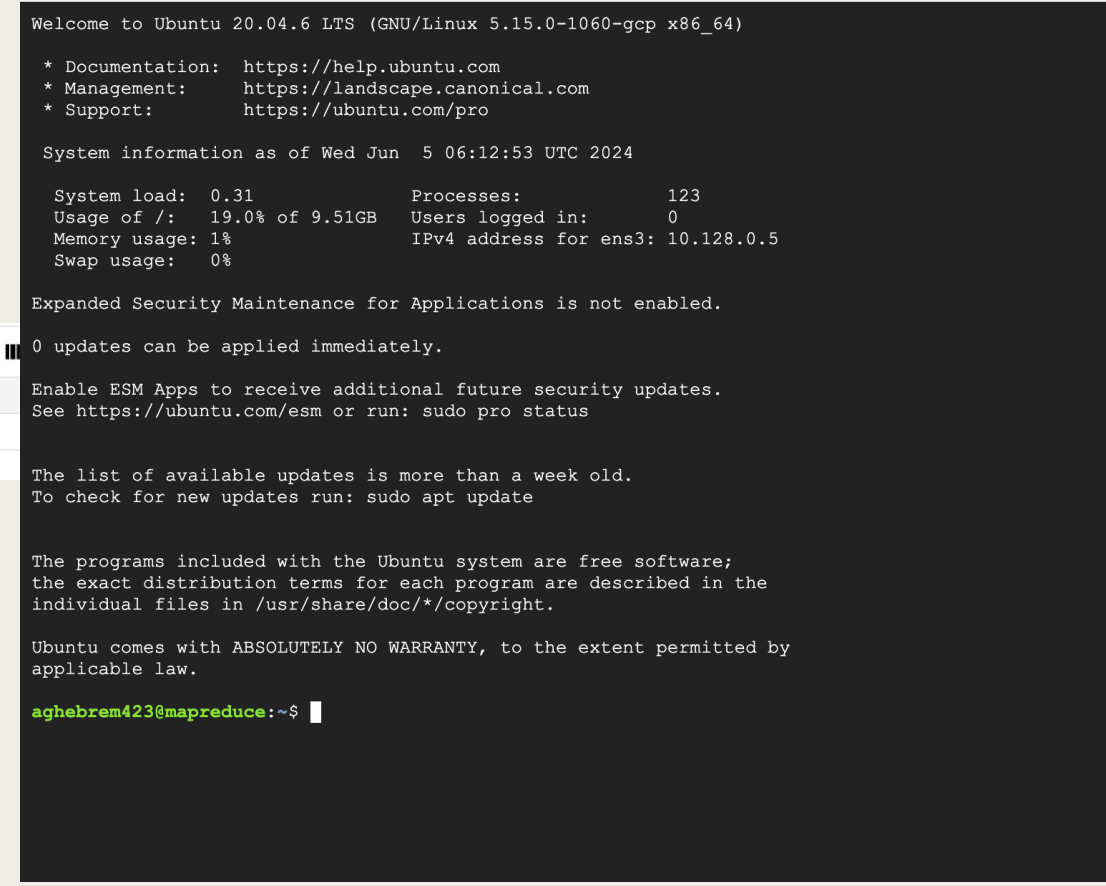
(x - center_x)2 + (y - center_y)2 compare r2

- Inside: if compare is <
- Outside: if compare is >
- On the circle: if compare is =

DESIGN

Job: Pi										
Map Task								Reduce Task		
map()				combine()				reduce()		
Input (Given)		Output (Program)		Input (Given)		Output (Program)		Input (Given)		Output (Program)
Key	Value (radius=2)	Key	Value (radius=2)	Key	Values	Key	Value	Key	Values	
file1	(0, 1)	Outside	1	Inside	[1]	Inside	1	Inside	[1, 3, 1]	Inside 5
	(1, 3)	Inside	1	Outside	[1, 1]	Outside	2	Outside	[2, 1, 4]	Outside 7
	(4, 3)	Outside	1							
file2	(2, 3)	Inside	1	Inside	[1, 1, 1]	Inside	3			
	(1, 3)	Inside	1	Outside	[1]	Outside	1			
	(1, 4)	Outside	1							
	(3, 2)	Inside	1							
file3	(3, 0)	Outside	1	Inside	[1]	Inside	1			
	(3, 3)	Inside	1	Outside	[1, 1, 1, 1]	Outside	4			
	(3, 4)	Outside	1							
	(0, 0)	Outside	1							
	(4, 4)	Outside	1							





Implementation

- GCP
- Hadoop
- Java

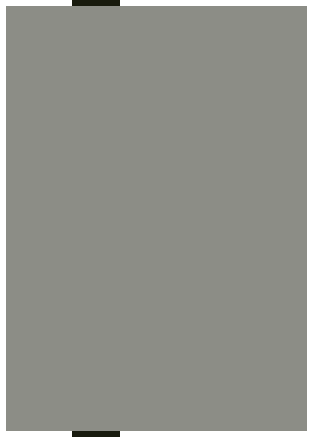
Implementation

- Create 3 java files.
  - *GenerateDots.java : Java Program to generate random dot pairs with command line arguments taken in as radius and number of pairs. Output format: x y radius*
  - *CalculatePiMR.java: Map(), Reduce() and main() for MapReduce*
  - *CalculatePi.java: Java Program to calculate pi value with MapReduce result taken in by reading the file.*

```java
import java.io.IOException;
import java.util.Random;

public class GenerateDots {
    public static void main(String[] args) throws Exception {
        //args[0]=>radius args[1]=>pairs of (x,y) to create
        //convert arguments to integer
        double radius = Double.parseDouble(args[0]);
        int num = Integer.parseInt(args[1]);
        for (int i=0; i< num; i++){
            double x = Math.random()*2*radius;
            double y = Math.random()*2*radius;

            System.out.println( Double.toString(x) + ' ' + Double.toString(y) + ' '
        }
    }
}
```



```java
import java.io.*;
public class CalculatePi {
       public static void main(String[] args) throws Exception{
              String file = "../hadoop-3.3.4/"+args[0]+"/part-r-00000";
              BufferedReader bufferedReader = new BufferedReader(new FileReader(file));

              String curLine="", line1="", line2="";
              while ((curLine = bufferedReader.readLine()) != null){
                     line1 = curLine;
                     if((curLine = bufferedReader.readLine()) != null){
                            line2 = curLine;
                     }
              }
              System.out.println(line1);
              System.out.println(line2);

              //System.out.println(line1.length() + " " + line2.length());
              String in = line1.substring(line1.length()-(line1.length()-6-1));
              String out = line2.substring(line2.length()-(line2.length()-7-1));

              double inside = Double.parseDouble(in);
              //System.out.println(inside);
              double outside = Double.parseDouble(out);
              //System.out.println(outside);
              double pi = 4 * ( inside / ( inside + outside ) );
              System.out.println("PI value is: " + pi );

              bufferedReader.close();
       }
```

```java
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.conf.*;
import org.apache.hadoop.io.*;
import org.apache.hadoop.mapreduce.*;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.input.TextInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
import org.apache.hadoop.mapreduce.lib.output.TextOutputFormat;

public class CalculatePiMR {
   public static class Map extends Mapper<LongWritable, Text, Text,IntWritable>
   {
        private final static IntWritable one = new IntWritable(1);
        private Text word = new Text();

        public void map(LongWritable key, Text value, Context context) throws IOException, InterruptedException
        {
            String line = value.toString();
            StringTokenizer tokenizer = new StringTokenizer(line);

            while(tokenizer.hasMoreTokens()){
                String xStr="0", yStr="0", rStr="5";
                xStr = tokenizer.nextToken();
                if(tokenizer.hasMoreTokens()){
                        yStr = tokenizer.nextToken();
                }
                if(tokenizer.hasMoreTokens()){
                        rStr = tokenizer.nextToken();
                }

                Double x = (Double)(Double.parseDouble(xStr));
                Double y = (Double)(Double.parseDouble(yStr));
```

```
aghebrem423@mapreduce:~$ cd PiProject
aghebrem423@mapreduce:~/PiProject$ ls
CalculatePi.java  CalculatePiMR.java  GenerateDots.java
aghebrem423@mapreduce:~/PiProject$ javac GenerateDots.java
aghebrem423@mapreduce:~/PiProject$ ls
CalculatePi.java  CalculatePiMR.java  GenerateDots.class  GenerateDots.java
aghebrem423@mapreduce:~/PiProject$ java GenerateDots 5 1000 > ./Input/dots.txt
-bash: ./Input/dots.txt: No such file or directory
aghebrem423@mapreduce:~/PiProject$ mkdir input
aghebrem423@mapreduce:~/PiProject$ mkdir testing
aghebrem423@mapreduce:~/PiProject$ java GenerateDots 5 1000 > ./Input/dots.txt
-bash: ./Input/dots.txt: No such file or directory
aghebrem423@mapreduce:~/PiProject$ java GenerateDots 5 1000 > ./input/dots.txt
aghebrem423@mapreduce:~/PiProject$ cat ./input/dots.txt
6.243485654899436 8.720902745249195 5.0
3.961982875273465 2.0603535616270072 5.0
6.944032255197917 9.14937060292062 5.0
4.1670804916734525 6.510901466477144 5.0
5.56317705185764 2.2999783442050923 5.0
7.122244620095847 8.687458006223917 5.0
1.1944942807520376 9.651704128046518 5.0
6.1491771457558775 8.096778241386295 5.0
3.363587131426502 7.771983363274392 5.0
7.0078683195231415 0.7037010652815912 5.0
6.275338317935631 1.6120607397503195 5.0
9.383518226601439 5.9490581946391305 5.0
0.9179566187996502 6.1694695295507 5.0
6.770568132279963 2.0707783180633967 5.0
5.17915852317535 0.484010352744515 5.0
5.12890668090272 9.265843260143402 5.0
9.834893227951563 1.9667988356913069 5.0
9.173636747779542 6.295208617976121 5.0
0.03860477128069251 9.657662015334724 5.0
2.708621902706162 3.7607545463136995 5.0
7.769098153793496 2.1643582014605665 5.0
6.025358311620774 6.472783803503457 5.0
0.5876580490628824 9.757807231891173 5.0
9.062544440165128 8.078537995191184 5.0
```

- Compile and run java program to generate dots with radius=5, number = 1000

```
aghebrem423@mapreduce:~/PiProject$ cd
aghebrem423@mapreduce:~$ cd hadoop-3.4.0
aghebrem423@mapreduce:~/hadoop-3.4.0$ bin/hdfs dfs -mkdir /user
aghebrem423@mapreduce:~/hadoop-3.4.0$ bin/hdfs dfs -mkdir /user/aghebrem423
aghebrem423@mapreduce:~/hadoop-3.4.0$ bin/hdfs dfs -mkdir /user/aghebrem423/PiProject/input
mkdir: `hdfs://localhost:9000/user/aghebrem423/PiProject': No such file or directory
aghebrem423@mapreduce:~/hadoop-3.4.0$ bin/hdfs dfs -mkdir /user/aghebrem423/PiProject
aghebrem423@mapreduce:~/hadoop-3.4.0$ bin/hdfs dfs -mkdir /user/aghebrem423/PiProject/input
aghebrem423@mapreduce:~/hadoop-3.4.0$
```

- **Copy file from local to Hadoop and check.**

```
aghebrem423@mapreduce:~/hadoop-3.4.0$ bin/hdfs dfs -mkdir /user/aghebrem423/PiProject/input
aghebrem423@mapreduce:~/hadoop-3.4.0$ bin/hdfs dfs -put ../PiProject/input/* PiProject/input
aghebrem423@mapreduce:~/hadoop-3.4.0$ bin/hdfs dfs -ls PiProject/input
Found 1 items
-rw-r--r--   1 aghebrem423 supergroup      40527 2024-06-05 07:24 PiProject/input/dots.txt
aghebrem423@mapreduce:~/hadoop-3.4.0$
```

- Compile MapReduce program in Hadoop with *.class files created

```
aghebrem423@mapreduce:~/hadoop-3.4.0$ bin/hadoop jar ~/hadoop-3.4.0/share/hadoop/mapreduce/hadoop-mapreduce-client-core-3.4.0.jar com.sun.tools.javac.Main ~/PiProject/CalculatePiMR.jav
a

Note: /home/aghebrem423/PiProject/CalculatePiMR.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
aghebrem423@mapreduce:~/hadoop-3.4.0$
aghebrem423@mapreduce:~/hadoop-3.4.0$
```

- Create .jar file with *.class files

```
aghebrem423@mapreduce:~/hadoop-3.4.0$ cd
aghebrem423@mapreduce:~$ cd PiProject
aghebrem423@mapreduce:~/PiProject$ ls
 CalculatePi.java  'CalculatePiMR$Map.class'  'CalculatePiMR$Reduce.class'   CalculatePiMR.class   CalculatePiMR.java   GenerateDots.class   GenerateDots.java   input   testing
aghebrem423@mapreduce:~/PiProject$ jar cf pi.jar CalculatePiMR*.class
aghebrem423@mapreduce:~/PiProject$
```

- Run MapReduce Program with input file and save result in Output

```
aghebrem423@mapreduce:~/hadoop-3.4.0$ bin/hadoop jar ~/PiProject/pi.jar CalculatePiMR /user/aghebrem423/PiProject/input /user/aghebrem423/PiProject/Output
2024-06-05 07:39:24,034 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2024-06-05 07:39:24,101 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2024-06-05 07:39:24,101 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2024-06-05 07:39:24,184 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner
to remedy this.
2024-06-05 07:39:24,298 INFO input.FileInputFormat: Total input files to process : 1
2024-06-05 07:39:24,312 INFO mapreduce.JobSubmitter: number of splits:1
2024-06-05 07:39:24,436 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local2016730749_0001
2024-06-05 07:39:24,436 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-06-05 07:39:24,540 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2024-06-05 07:39:24,541 INFO mapreduce.Job: Running job: job_local2016730749_0001
2024-06-05 07:39:24,542 INFO mapred.LocalJobRunner: OutputCommitter set in config null
2024-06-05 07:39:24,547 INFO output.PathOutputCommitterFactory: No output committer factory defined, defaulting to FileOutputCommitterFactory
2024-06-05 07:39:24,548 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2024-06-05 07:39:24,548 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2024-06-05 07:39:24,550 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
2024-06-05 07:39:24,579 INFO mapred.LocalJobRunner: Waiting for map tasks
2024-06-05 07:39:24,580 INFO mapred.LocalJobRunner: Starting task: attempt_local2016730749_0001_m_000000_0
2024-06-05 07:39:24,594 INFO output.PathOutputCommitterFactory: No output committer factory defined, defaulting to FileOutputCommitterFactory
2024-06-05 07:39:24,594 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2024-06-05 07:39:24,594 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
```

Test

- Get output and save to local, the show output.

```
        Bytes Written=23
aghebrem423@mapreduce:~/hadoop-3.4.0$ bin/hdfs dfs -get PiProject/Output Output
aghebrem423@mapreduce:~/hadoop-3.4.0$ cat Output/*
Inside  798
Outside 202
aghebrem423@mapreduce:~/hadoop-3.4.0$
```

- Using the output (local output folder as command line arguments) from MapReduce Program to compile and run java program to get pi value.

```
aghebrem423@mapreduce:~/PiProject$ javac CalculatePi.java
aghebrem423@mapreduce:~/PiProject$ java CalculatePi Output
Inside  798
Outside 202
PI value is: 3.192
aghebrem423@mapreduce:~/PiProject$
```

Enhancement

- Since, the PI number is close but a bit off we will try to decrease radius or increae number to see the changes they will make.
- **Enhancement: Decrease radius**

```
use --help for a list of possible options
aghebrem423@mapreduce:~/PiProject$ javac GenerateDots.java
aghebrem423@mapreduce:~/PiProject$ java GenerateDots 1 1000 > ./input/test1.txt
aghebrem423@mapreduce:~/PiProject$ ls ./input
dots.txt  test1.txt
aghebrem423@mapreduce:~/PiProject$ cat ./input/test1.txt
1.0349458824197373 0.6693962954049777 1.0
1.218808202659513 1.29143430090437 1.0
0.41060585984958387 1.8823914072892558
1.986189260669244 1.4661892713349345 1
1.0140297684695885 1.1410316263244047
0.4957752051101747 1.7959662232938904
0.8594215299498511 1.2377469855685717
0.58960770449765 1.93835792311259 1.0
```

```
aghebrem423@mapreduce:~/hadoop-3.4.0$ cd
aghebrem423@mapreduce:~$ cd PiProject
aghebrem423@mapreduce:~/PiProject$ java CalculatePi Test1
Inside  794
Outside 206
PI value is: 3.176
aghebrem423@mapreduce:~/PiProject$
```

- **Enhancement : Increase number.**

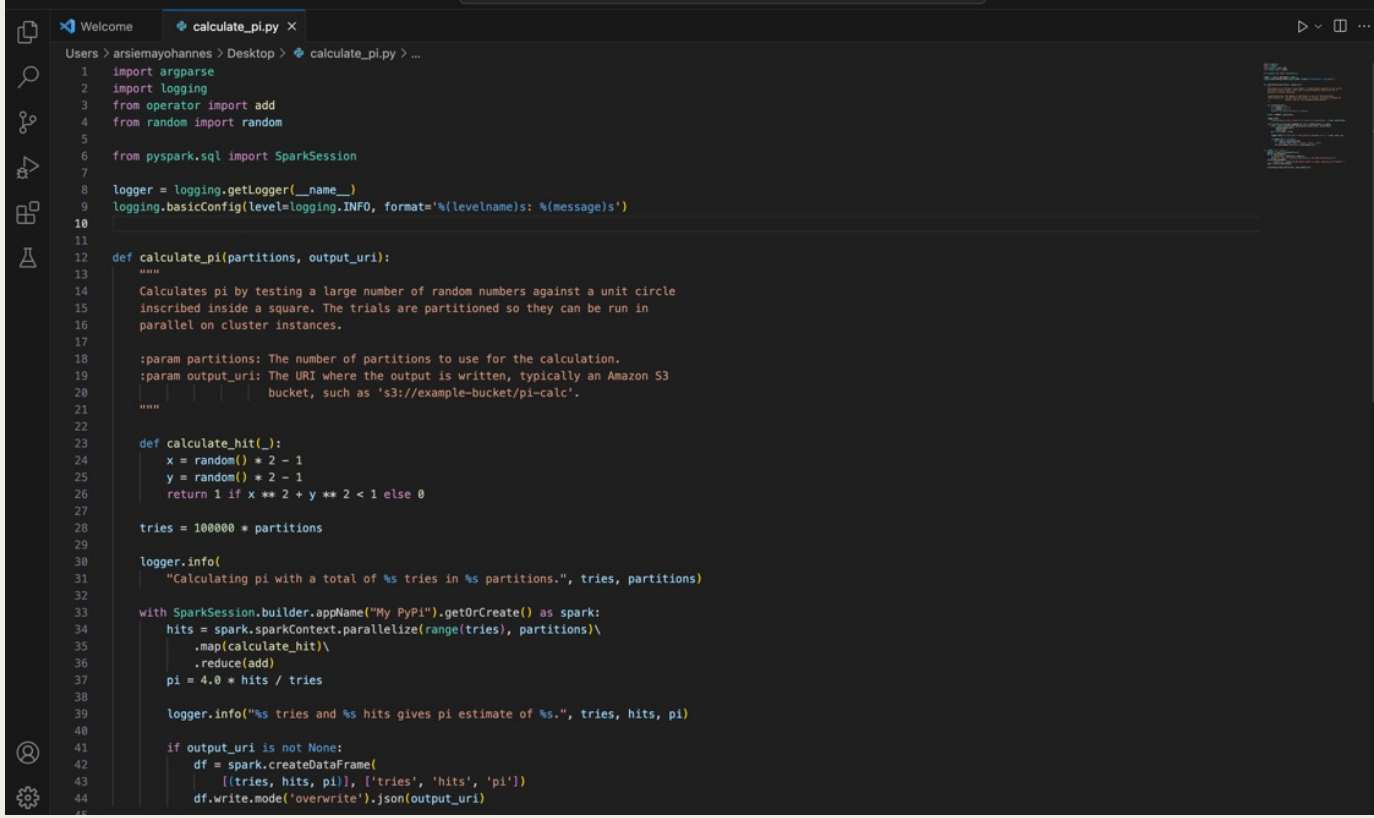
```
aghebrem423@mapreduce:~/PiProject$ java GenerateDots 5 1000000 > ./input/test2.txt
aghebrem423@mapreduce:~/PiProject$ ls ./input/test2.txt
./input/test2.txt
aghebrem423@mapreduce:~/PiProject$ ls ./input
dots.txt  test1.txt  test2.txt
aghebrem423@mapreduce:~/PiProject$ cat ./input/test2.txt
2.4409513178371336 1.9695968916104478 5.0
6.039596943158905 1.946277459843908 5.0
7.34317341682304 9.64860808775004 5.0
2.6616950654632565 2.589232923294439 5.0
3.495537161083142 8.291024720380582 5.0
6.371800950987319 0.4486674244486122 5.0
9.300473331723488 7.773773117188401 5.0
8.291425720800357 0.9219277488584798 5.0
5.642389490486829 0.0012242655171057493 5.0
```

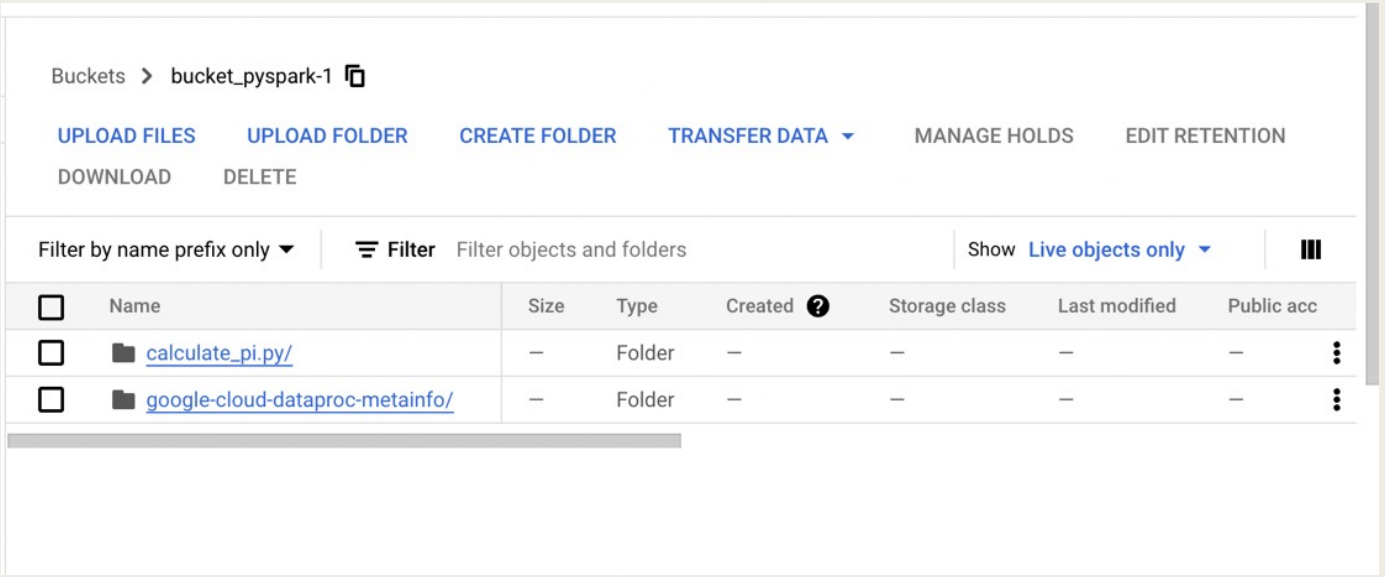
```
aghebrem423@mapreduce:~/hadoop-3.4.0$ bin/hdfs dfs -get PiProject/Test2 Test2
aghebrem423@mapreduce:~/hadoop-3.4.0$ cat Test2/*
Inside  785467
Outside 214533
aghebrem423@mapreduce:~/hadoop-3.4.0$ cd
aghebrem423@mapreduce:~$ cd PiProject
aghebrem423@mapreduce:~/PiProject$ java CalculatePi Test2
Inside  785467
Outside 214533
PI value is: 3.141868
aghebrem423@mapreduce:~/PiProject$
```

- Increasing number made the PI number very close.

Calculating Pi - PySpark implementation

- Create the python file: Calculate_pi.py.
- Then Upload it to your bucket.





```
aghebrem423@cloudshell:~ (fluent-music-424804-d6)$ gcloud dataproc jobs submit pyspark gs://bucket_pyspark-1/calculate_pi.py     --cluster=cluster-2d8c     --region=us-central1     --
   --partitions=4     --output_uri=gs://bucket_pyspark-1/calculate_pi.py
Job [76cede9818c04d18bcadc4e45f627690] submitted.
Waiting for job output...
INFO: Calculating pi with a total of 400000 tries in 4 partitions.
24/06/19 23:51:54 INFO SparkEnv: Registering MapOutputTracker
24/06/19 23:51:54 INFO SparkEnv: Registering BlockManagerMaster
24/06/19 23:51:54 INFO SparkEnv: Registering BlockManagerMasterHeartbeat
24/06/19 23:51:54 INFO SparkEnv: Registering OutputCommitCoordinator
24/06/19 23:51:55 INFO DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at cluster-2d8c-m.us-central1-f.c.fluent-music-424804-d6.internal./10.128.0.18:8032
24/06/19 23:51:55 INFO AHSProxy: Connecting to Application History server at cluster-2d8c-m.us-central1-f.c.fluent-music-424804-d6.internal./10.128.0.18:10200
24/06/19 23:51:56 INFO Configuration: resource-types.xml not found
24/06/19 23:51:56 INFO ResourceUtils: Unable to find 'resource-types.xml'.
24/06/19 23:51:57 INFO YarnClientImpl: Submitted application application_1718839758942_0004
24/06/19 23:51:58 INFO DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at cluster-2d8c-m.us-central1-f.c.fluent-music-424804-d6.internal./10.128.0.18:8030
24/06/19 23:52:00 INFO GhfsStorageStatistics: Detected potential high latency for operation op_get_file_status. latencyMs=320; previousMaxLatencyMs=0; operationCount=1; context=gs://dat
aproc-temp-us-central1-660908043237-ctirguzj/888f724f-c8c1-4557-8d5a-83c80bd2a867/spark-job-history
24/06/19 23:52:00 INFO GoogleCloudStorageImpl: Ignoring exception of type GoogleJsonResponseException; verified object already exists with desired state.
24/06/19 23:52:00 INFO GhfsStorageStatistics: Detected potential high latency for operation op_mkdirs. latencyMs=191; previousMaxLatencyMs=0; operationCount=1; context=gs://dataproc-tem
p-us-central1-660908043237-ctirguzj/888f724f-c8c1-4557-8d5a-83c80bd2a867/spark-job-history
INFO: 400000 tries and 314664 hits gives pi estimate of 3.14664.
INFO: NumExpr defaulting to 4 threads.
24/06/19 23:52:14 INFO GhfsStorageStatistics: Detected potential high latency for operation op_delete. latencyMs=176; previousMaxLatencyMs=0; operationCount=1; context=gs://bucket_pyspa
rk-1/calculate_pi.py
24/06/19 23:52:14 INFO PathOutputCommitterFactory: No output committer factory defined, defaulting to FileOutputCommitterFactory
24/06/19 23:52:21 INFO GoogleCloudStorageFileSystem: Successfully repaired 'gs://bucket_pyspark-1/calculate_pi.py/' directory.
24/06/19 23:52:21 INFO GhfsStorageStatistics: Detected potential high latency for operation op_delete. latencyMs=338; previousMaxLatencyMs=176; operationCount=2; context=gs://bucket_pys
park-1/calculate_pi.py/_temporary
24/06/19 23:52:21 INFO GhfsStorageStatistics: Detected potential high latency for operation stream_write_close_operations. latencyMs=187; previousMaxLatencyMs=0; operationCount=1; conte
xt=gs://bucket_pyspark-1/calculate_pi.py/_SUCCESS
24/06/19 23:52:22 INFO GhfsStorageStatistics: Detected potential high latency for operation op_rename. latencyMs=338; previousMaxLatencyMs=0; operationCount=1; context=rename(gs://datap
roc-temp-us-central1-660908043237-ctirguzj/888f724f-c8c1-4557-8d5a-83c80bd2a867/spark-job-history/application_1718839758942_0004.inprogress -> gs://dataproc-temp-us-central1-66090804323
```

- Run the program.

gcloud dataproc jobs submit pyspark gs://bucket_pyspark-1/calculate_pi.py --cluster=cluster-2d8c --region=us-central1 -- --partitions=4 -- output_uri=gs://bucket_pyspark-1/calculate_pi.py

gsutil cat gs://bucket_pyspark-1/calculate_pi.py/part-00003-55ea2cec-fc96-4dca-8b58-75aa26eab974-c000.json

```
aghebrem423@cloudshell:~ (fluent-music-424804-d6)$ gsutil cat gs://bucket_pyspark-1/calculate_pi.py/part-00000-55ea2cec-fc96-4dca-8b58-75aa26eab974-c000.json
aghebrem423@cloudshell:~ (fluent-music-424804-d6)$ gsutil cat gs://bucket_pyspark-1/calculate_pi.py/part-00003-55ea2cec-fc96-4dca-8b58-75aa26eab974-c000.json
{"tries":400000,"hits":314664,"pi":3.14664}
aghebrem423@cloudshell:~ (fluent-music-424804-d6)$
```

Conclusion

- The Pi algorithm is concise and clear that,
  - *N should be large*
  - *Points should be chosen uniformly at random*
- Hence, we should increase the the number of random values to get accurate result of Pi.

Reference

- Hadoop mapreduce to calculate Pi
- MapReduce Tutorial
- Pi Computation With MapReduce
- Hadoop: Setting up a Single Node Cluster.