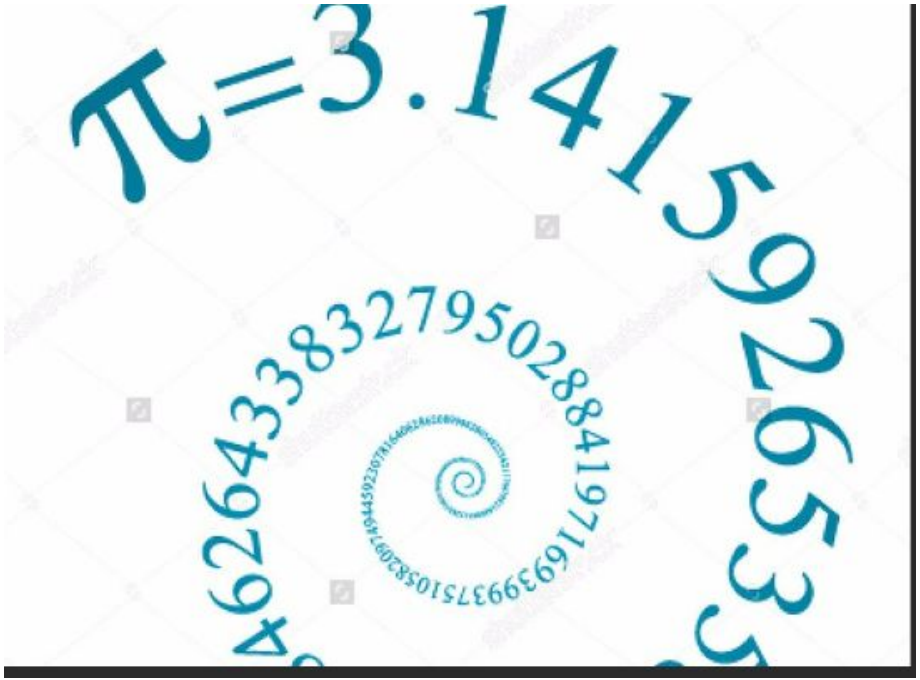
MapReduce-Pi

FATEMA NAGORI 19635

TABLE OF CONTENT:

Introduction



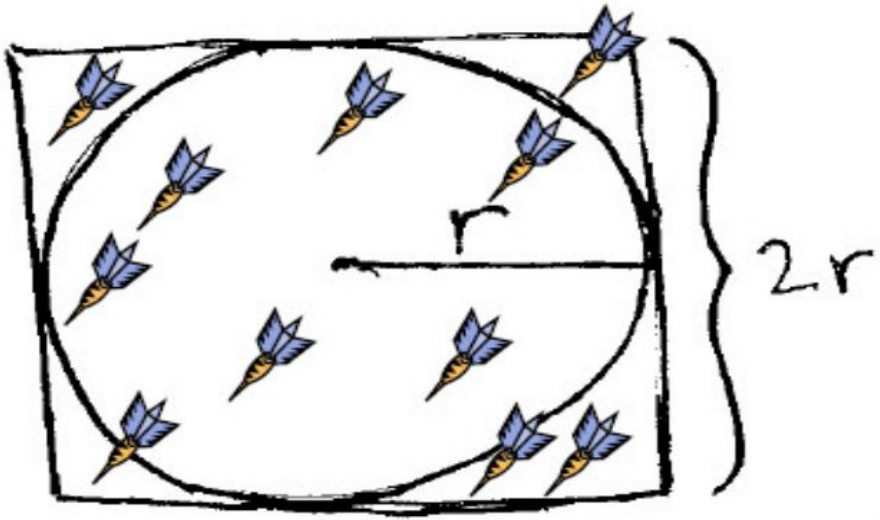
In this Project we find the value of **Pi** by generating random numbers, we also demonstrate manual solution to find the value of **Pi** using **Map Reduce** method and implement the same program at **Google Cloud Platform.**

Design Approach

a. **Pi- Using Random Numbers**

Generate an input file to the Pi MapReduce program

- 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

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

Explanation

If radius is 5, then based on the input values in Map, we can calculate pi = 4 * (S / N)

= 4 * (Insde / (Inside + Outside))

= 4 * (5 / (5 + 7))

= 4 * (5 / 12)

= 1.66

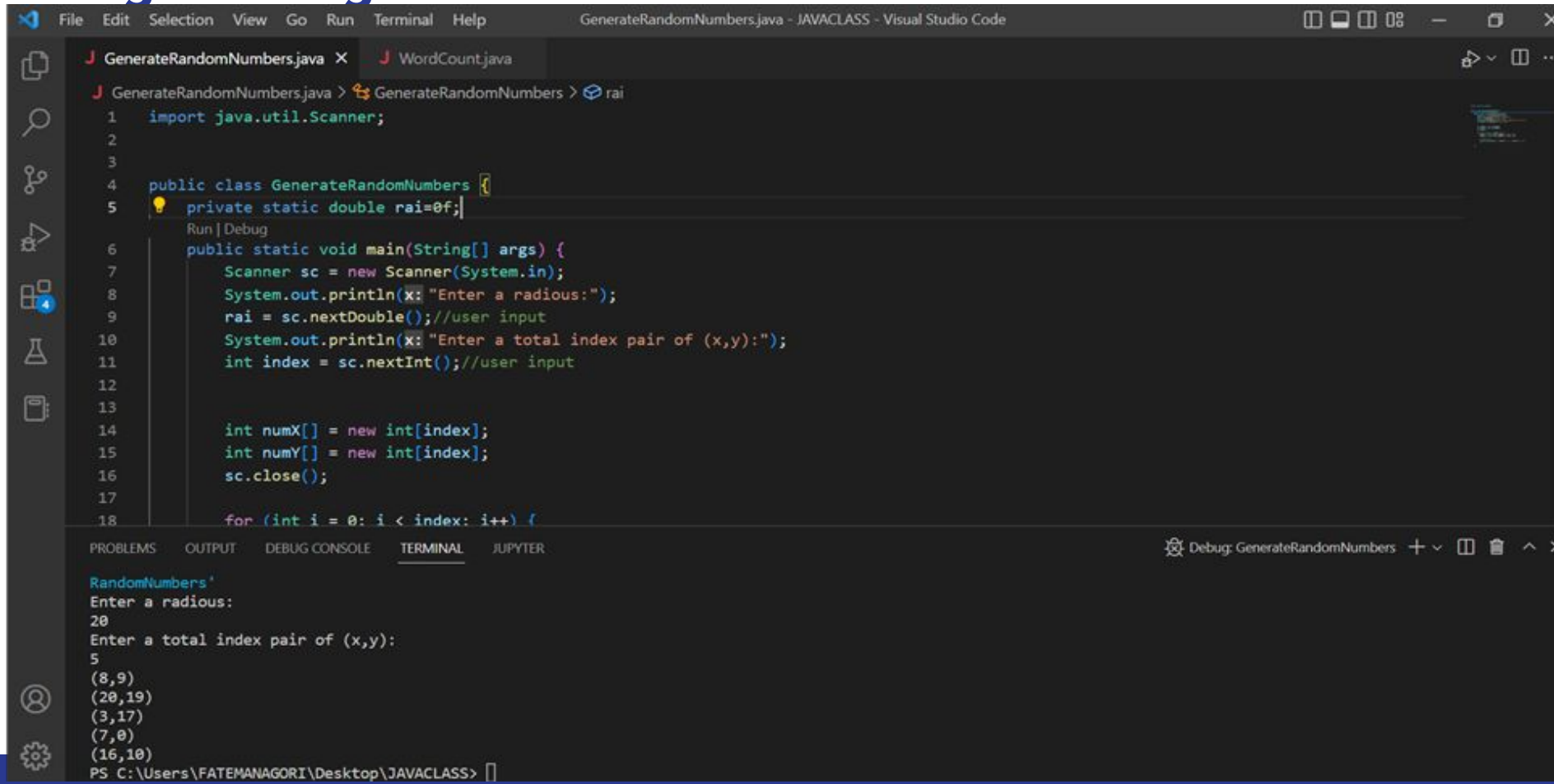
- Note:
  - To get more accurate pi value you can
    - increase the value of the radius, and
    - create much much more input values (e.g., a million values)

FORMULA:
inside/outside
$(x - center\_x)^2 + (x - center\_x)^2$ compare r^2

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

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							

Program to generate random numbers:



Implementation: GCP preparation

```
Last login: Mon Oct 10 22:39:28 2022 from 127.0.0.1
fnagori@instance-vm:~$ ls
MapReduce  hadoop-3.3.4  hadoop-3.3.4.tar.gz  pi
fnagori@instance-vm:~$ ls MapReduce
input
fnagori@instance-vm:~$ cd MapReduce
fnagori@instance-vm:~/MapReduce$ vi pi.java
fnagori@instance-vm:~/MapReduce$ ls
input  pi.java
fnagori@instance-vm:~/MapReduce$ cd input
fnagori@instance-vm:~/MapReduce/input$ vi file01
fnagori@instance-vm:~/MapReduce/input$
```

Connect to localhost

```
fnagori@instance-vm:~/MapReduce/input$ cd
fnagori@instance-vm:~$ vi ~/.bashrc
fnagori@instance-vm:~$ ssh-keygen -t rsa -P '' -f ~/.ssh/id_rsa
Generating public/private rsa key pair.
/home/fnagori/.ssh/id_rsa already exists.
Overwrite (y/n)? y
Your identification has been saved in /home/fnagori/.ssh/id_rsa.
Your public key has been saved in /home/fnagori/.ssh/id_rsa.pub.
The key fingerprint is:
SHA256:U3lnuE0GbmPwmfMUHMe1nMZqUdVwUXt+E8CvviKoq48 fnagori@instance-vm
The key's randomart image is:
+---[RSA 2048]----+
          . oo+*%|
           = **+*|
          o %.BO.|
         . + %+oo|
        S   .o+.o|
         .  ..  o|
        .   .    |
     .  . . . .  |
    Eo+o   . ...  |
+----[SHA256]-----+
fnagori@instance-vm:~$ cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
fnagori@instance-vm:~$ chmod 0600 ~/.ssh/authorized_keys
fnagori@instance-vm:~$ ssh localhost
Welcome to Ubuntu 18.04.6 LTS (GNU/Linux 5.4.0-1089-gcp x86_64)

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

https://ssh.cloud.google.com/v2/ssh/projects/new-cs570/zones/us-central1-a/instances/instance-vm?authuser=1&hl=en_US&projectNumber=1049843925247&useAdminProxy=true&troubleshoot4005Enable...

https://ssh.cloud.google.com/v2/ssh/projects/new-cs570/zones/us-central1-a/instances/instance-vm?authuser=1&hl=en_US&projectNumber=1049843925247&useAdminProxy=true&trouble...

SSH-in-browser

```
fnagori@instance-vm:~/hadoop-3.3.4$ bin/hadoop jar wc.jar PiCalculation /user/fnagori/mapreduce/input /user/fnagori/mapreduce/output
2022-10-11 02:28:06,896 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2022-10-11 02:28:07,004 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2022-10-11 02:28:07,004 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2022-10-11 02:28:07,377 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your appl
ication with ToolRunner to remedy this.
2022-10-11 02:28:07,646 INFO input.FileInputFormat: Total input files to process : 1
2022-10-11 02:28:07,691 INFO mapreduce.JobSubmitter: number of splits:1
2022-10-11 02:28:08,025 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1561569321_0001
2022-10-11 02:28:08,026 INFO mapreduce.JobSubmitter: Executing with tokens: []
2022-10-11 02:28:08,275 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2022-10-11 02:28:08,276 INFO mapreduce.Job: Running job: job_local1561569321_0001
2022-10-11 02:28:08,286 INFO mapred.LocalJobRunner: OutputCommitter set in config null
2022-10-11 02:28:08,311 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2022-10-11 02:28:08,311 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failur
es: false
2022-10-11 02:28:08,313 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
2022-10-11 02:28:08,400 INFO mapred.LocalJobRunner: Waiting for map tasks
2022-10-11 02:28:08,401 INFO mapred.LocalJobRunner: Starting task: attempt_local1561569321_0001_m_000000_0
2022-10-11 02:28:08,442 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2022-10-11 02:28:08,443 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failur
es: false
2022-10-11 02:28:08,489 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
2022-10-11 02:28:08,502 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/user/fnagori/mapreduce/input/file01:0+708
2022-10-11 02:28:08,719 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
2022-10-11 02:28:08,719 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
2022-10-11 02:28:08,719 INFO mapred.MapTask: soft limit at 83886080
2022-10-11 02:28:08,719 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
2022-10-11 02:28:08,719 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
2022-10-11 02:28:08,728 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
2022-10-11 02:28:08,942 INFO mapred.LocalJobRunner:
2022-10-11 02:28:08,946 INFO mapred.MapTask: Starting flush of map output
2022-10-11 02:28:08,947 INFO mapred.MapTask: Spilling map output
2022-10-11 02:28:08,947 INFO mapred.MapTask: bufstart = 0; bufend = 1130; bufvoid = 104857600
2022-10-11 02:28:08,947 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26214000(104856000); length = 397/6553600
2022-10-11 02:28:08,980 INFO mapred.MapTask: Finished spill 0
2022-10-11 02:28:08,997 INFO mapred.Task: Task:attempt_local1561569321_0001_m_000000_0 is done. And is in the process of committing
```

Type here to search

70°F

7:33 PM
10/10/2022

```
fnagori@instance-vm:~$ cd hadoop-3.3.4
fnagori@instance-vm:~/hadoop-3.3.4$ ls
 LICENSE-binary                      PiCalculation.java   input             output1
 LICENSE.txt                         README.txt           lib               sbin
 NOTICE-binary                       bin                  libexec           share
 NOTICE.txt                          etc                  licenses-binary   wc.jar
'PiCalculation$IntSumReducer.class'  include              logs
'PiCalculation$TokenizerMapper.class' index.html          output
 PiCalculation.class                 index.html.1         output.old
fnagori@instance-vm:~/hadoop-3.3.4$ bin/hadoop com.sun.tools.javac.Main PiCalculation.java
fnagori@instance-vm:~/hadoop-3.3.4$ jar cf wc.jar PiCalculation*.class
fnagori@instance-vm:~/hadoop-3.3.4$ ls
 LICENSE-binary                      PiCalculation.java   input             output1
 LICENSE.txt                         README.txt           lib               sbin
 NOTICE-binary                       bin                  libexec           share
 NOTICE.txt                          etc                  licenses-binary   wc.jar
'PiCalculation$IntSumReducer.class'  include              logs
'PiCalculation$TokenizerMapper.class' index.html          output
 PiCalculation.class                 index.html.1         output.old
```

output

SSH-in-browser

```
                Reduce input groups=2
                Reduce shuffle bytes=33
                Reduce input records=2
                Reduce output records=2
                Spilled Records=4
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=58
                Total committed heap usage (bytes)=246489088
        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=708
        File Output Format Counters
                Bytes Written=21
inside  70
outside 30
Inside:70, Outside:30
PI:2.8
fnagori@instance-vm:~/hadoop-3.3.4$ bin/hdfs dfs -get mapreduce/output output
get: `output/output/_SUCCESS': File exists
get: `output/output/part-r-00000': File exists
fnagori@instance-vm:~/hadoop-3.3.4$ cat output/*
cat: output/output: Is a directory
outside 50
fnagori@instance-vm:~/hadoop-3.3.4$ rm -rf output
fnagori@instance-vm:~/hadoop-3.3.4$ bin/hdfs dfs -get mapreduce/output output
fnagori@instance-vm:~/hadoop-3.3.4$ cat output/*
inside  70
outside 30
fnagori@instance-vm:~/hadoop-3.3.4$
```

Enhancement Ideas:

Millions of Darts!

- If you want to get an accurate estimate of Pi, you need a large number of random samples.
- Notice that each dart can be thrown at any time and it's position can be evaluated independently
- With one person throwing all the darts, it will take a long time to finish
- If we had N people throwing a dart each, this would be much faster!

Conclusion:

In this Project we successfully calculated the value of Pi with the random numbers generated. The value of Pi we got was 2.8 and the inside darts where 70 and outside 30 but we can get more accurate value of Pi by generating large numbers of random samples

REFERENCES:

- https://hc.labnet.sfbu.edu/~henry/npu/classes/mapreduce/pi/slide/exercise_pi.html
- https://hc.labnet.sfbu.edu/~henry/npu/classes/mapreduce/course/catherine_fang/W5%20-%20Pi%20Computation%20with%20MapReduce.pd