

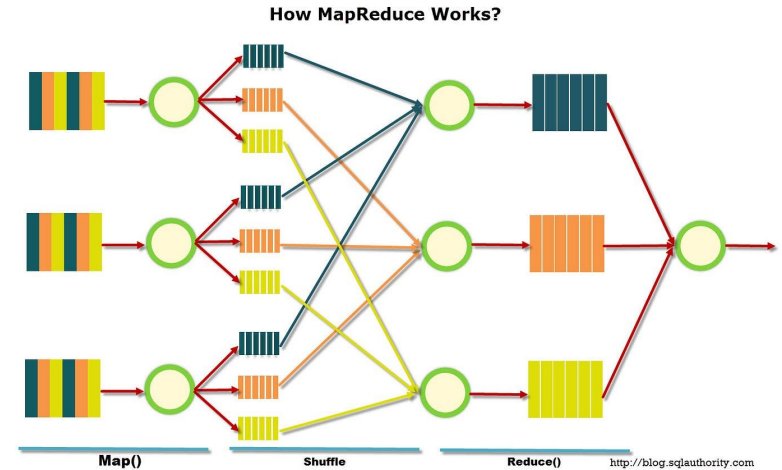


Calculating Pi using MapReduce & Pyspark

CS570: Big Data Processing & Analytics
Submitted by Imran Noor Saleh Student ID 19648

Introduction: Map Reduce

MapReduce is a programming model and an associated implementation for processing and generating big data sets with a parallel, distributed algorithm on a cluster.



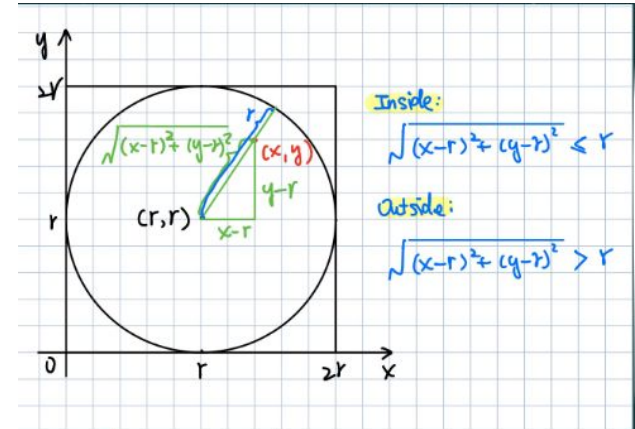


The number π is a mathematical constant that is the ratio of a circle's circumference to its diameter, approximately equal to 3.14159.



Overview: Pi Calculation using Map Reduce

- Throw N darts on the board . Each dart lands at a random position (x,y) on the board.
- 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: $4(S/N) = \pi$

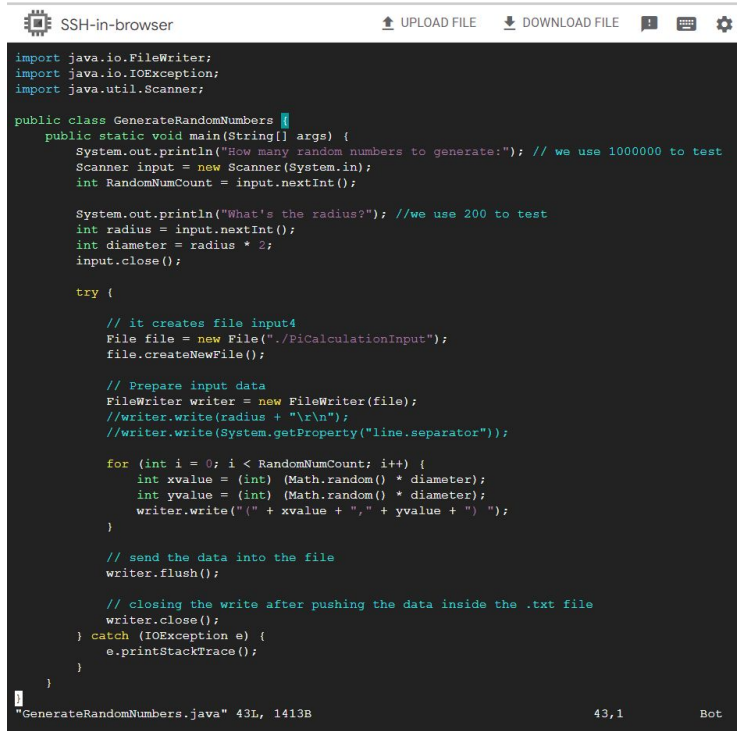


Design

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

```
isaleh@instance-1:~$ ls
Pi  hadoop-3.3.4  hadoop-3.3.4.tar.gz
isaleh@instance-1:~$ cd Pi
isaleh@instance-1:~/Pi$ vi GenerateRandomNumbers.java
```



```
SSH-in-browser  ⬆️ UPLOAD FILE  ⬇️ DOWNLOAD 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
            writer.flush();

            // closing the write after pushing the data inside the .txt file
            writer.close();
        } catch (IOException e) {
            e.printStackTrace();
        }
    }
}

"GenerateRandomNumbers.java" 43L, 1413B  43,1  Bot
```

Implementation Cntd.

```
isaleh@instance-1:~$ cd hadoop-3.3.4
isaleh@instance-1:~/hadoop-3.3.4$ cd etc/hadoop/
isaleh@instance-1:~/hadoop-3.3.4/etc/hadoop$ vi hadoop-env.sh
isaleh@instance-1:~/hadoop-3.3.4/etc/hadoop$ cd
```

```
isaleh@instance-1:~/hadoop-3.3.4$ bin/hadoop
Usage: hadoop [OPTIONS] SUBCOMMAND [SUBCOMMAND OPTIONS]
or   hadoop [OPTIONS] CLASSNAME [CLASSNAME OPTIONS]
     where CLASSNAME is a user-provided Java class

    OPTIONS is none or any of:

--config dir      Hadoop config directory
--debug           turn on shell script debug mode
--help           usage information
buildpaths        attempt to add class files from build tree
hostnames list[,of,host,names] hosts to use in slave mode
hosts filename    list of hosts to use in slave mode
loglevel level    set the log4j level for this command
workers          turn on worker mode

SUBCOMMAND is one of:

Admin Commands:

daemonlog         get/set the log level for each daemon

Client Commands:

archive           create a Hadoop archive
checknative       check native Hadoop and compression libraries availability
classpath         prints the class path needed to get the Hadoop jar and the required libraries
conftest         validate configuration XML files
credential        interact with credential providers
distch           distributed metadata changer
distcp           copy file or directories recursively
dutil            operations related to delegation tokens
envvars          display computed Hadoop environment variables
fs              run a generic filesystem user client
gridmix          submit a mix of synthetic job, modeling a profiled from production load
jar <jar>        run a jar file. NOTE: please use "yarn jar" to launch YARN applications, not
                 this command.
jnipath          prints the java.library.path
kdiag            Diagnose Kerberos Problems
kerbname         show auth_to_local principal conversion
key             manage keys via the KeyProvider
rumenfolder      scale a rumen input trace
```

Implementation Cntd.

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
Licensed under the Apache License, Version 2.0 (the "License");
you may not use this file except in compliance with the License.
You may obtain a copy of the License at

    http://www.apache.org/licenses/LICENSE-2.0

Unless required by applicable law or agreed to in writing, software
distributed under the License is distributed on an "AS IS" BASIS,
WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
See the License for the specific language governing permissions and
limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->

<configuration>
  <property>
    <name>fs.defaultFS</name>
    <value>hdfs://localhost:9000</value>
  </property>
</configuration>
~
~
~
```

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
Licensed under the Apache License, Version 2.0 (the "License");
you may not use this file except in compliance with the License.
You may obtain a copy of the License at

    http://www.apache.org/licenses/LICENSE-2.0

Unless required by applicable law or agreed to in writing, software
distributed under the License is distributed on an "AS IS" BASIS,
WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
See the License for the specific language governing permissions and
limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->

<configuration>
  <property>
    <name>dfs.replication</name>
    <value>1</value>
  </property>
</configuration>
~
```


Implementation Cntd.

```
>
> bin/hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-3.3.4.jar grep input output
'dfs[a-z,1]'
2023-08-20 02:19:24,715 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2023-08-20 02:19:24,960 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 seconds
2023-08-20 02:19:24,961 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2023-08-20 02:19:25,237 INFO input.FileInputFormat: Total input files to process : 10
2023-08-20 02:19:25,275 INFO mapreduce.JobSubmitter: number of splits:10
2023-08-20 02:19:25,673 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local449076150_0001
2023-08-20 02:19:25,673 INFO mapreduce.JobSubmitter: Executing with tokens: []
2023-08-20 02:19:25,906 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2023-08-20 02:19:25,907 INFO mapreduce.Job: Running job: job_local449076150_0001
2023-08-20 02:19:25,918 INFO mapred.LocalJobRunner: OutputCommitter set in config null
2023-08-20 02:19:25,930 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2023-08-20 02:19:25,930 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup_temporary folders under output directory:false, ignore cleanup failures: false
2023-08-20 02:19:25,931 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
2023-08-20 02:19:26,010 INFO mapred.LocalJobRunner: Waiting for map tasks
2023-08-20 02:19:26,012 INFO mapred.LocalJobRunner: Starting task: attempt_local449076150_0001_m_000000_0
2023-08-20 02:19:26,050 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2023-08-20 02:19:26,054 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup_temporary folders under output directory:false, ignore cleanup failures: false
2023-08-20 02:19:26,092 INFO mapred.Task: Using ResourceCalculatorProcessTree : [ ]
2023-08-20 02:19:26,097 INFO mapred.MapTask: Processing split: file:/home/isaleh/hadoop-3.3.4/input/hadoop-policy.xml:0+11765
2023-08-20 02:19:26,232 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
2023-08-20 02:19:26,232 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
2023-08-20 02:19:26,232 INFO mapred.MapTask: soft limit at 83886080
2023-08-20 02:19:26,232 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
2023-08-20 02:19:26,232 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
2023-08-20 02:19:26,265 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
2023-08-20 02:19:26,307 INFO mapred.LocalJobRunner:
2023-08-20 02:19:26,308 INFO mapred.MapTask: Starting flush of map output
2023-08-20 02:19:26,335 INFO mapred.Task: Task:attempt_local449076150_0001_m_000000_0 is done.
And is in the process of committing
```

```
isaleh@instance-1:~/hadoop-3.3.4$ cat output/*
isaleh@instance-1:~/hadoop-3.3.4$ ls ./output
_SUCCESS part-r-00000
isaleh@instance-1:~/hadoop-3.3.4$ vi ./etc/hadoop/core-site.xml
isaleh@instance-1:~/hadoop-3.3.4$ vi ./etc/hadoop/core-site.xml
isaleh@instance-1:~/hadoop-3.3.4$ vi ./etc/hadoop/hdfs-site.xml
isaleh@instance-1:~/hadoop-3.3.4$ ssh localhost
The authenticity of host 'localhost (::1)' can't be established.
ECDSA key fingerprint is SHA256:V3AauVR5/6odrQv6gtNx1g5GkfzXsD1wDCfenZQ/1sA.
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added 'localhost' (ECDSA) to the list of known hosts.
isaleh@localhost: Permission denied (publickey).
isaleh@instance-1:~/hadoop-3.3.4$ ssh-keygen -t rsa -P '' -f ~/.ssh/id_rsa
Generating public/private rsa key pair.
Your identification has been saved in /home/isaleh/.ssh/id_rsa
Your public key has been saved in /home/isaleh/.ssh/id_rsa.pub
The key fingerprint is:
SHA256:KZGuSwaFuc4rB9J+IZBBQSSziIE5RQMxKwJCrySride isaleh@instance-1
The key's randomart image is:
+---[RSA 3072]-----+
|  %Bo                |
| XX+o .              |
| X+E . o             |
| == + . . .          |
| = * + o S           |
| o B + .             |
| + O                 |
| * .                 |
| .                   |
+-----[SHA256]-----+
isaleh@instance-1:~/hadoop-3.3.4$
```

Implementation Cntd.

```
isaleh@instance-1:~/hadoop-3.3.4$ ssh localhost
Linux instance-1 5.10.0-24-cloud-amd64 #1 SMP Debian 5.10.179-5 (2023-08-08) x86_64
```

The programs included with the Debian GNU/Linux system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/copyright.

Debian GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent
permitted by applicable law.

Last login: Sun Aug 20 01:25:39 2023 from 35.235.243.64

```
isaleh@instance-1:~$
```

```
isaleh@instance-1:~$ cd hadoop-3.3.4
isaleh@instance-1:~/hadoop-3.3.4$ bin/hdfs namenode -format
WARNING: /home/isaleh/hadoop-3.3.4/logs does not exist. Creating.
2023-08-20 02:29:10,027 INFO namenode.NameNode: STARTUP_MSG:
/*****
STARTUP_MSG: Starting NameNode
STARTUP_MSG:   host = instance-1/10.182.0.2
STARTUP_MSG:   args = [-format]
STARTUP_MSG:   version = 3.3.4
STARTUP_MSG:   classpath = /home/isaleh/hadoop-3.3.4/etc/hadoop:/home/isaleh/hadoop-3.3.4/share/hadoop/common/lib/
e/hadoop/common/lib/slf4j-reload4j-1.7.36.jar:/home/isaleh/hadoop-3.3.4/share/hadoop/common/lib/jul-to-slf4j-1.7.36.j
ar:/home/isaleh/hadoop-3.3.4/share/hadoop/common/lib/jackson-xc-1.9.13.jar:/home/isaleh/hadoop-3.3.4/share/hadoop/co
mmon/lib/jar395-3.0.2.jar:/home/isaleh/hadoop-3.3.4/share/hadoop/common/lib/commons-net-3.6.jar:/home/isaleh/hadoop-3.3.4/share/hadoop/common/lib/jcip-annotations-1.0-1.jar:/home/isaleh/hadoop-3.3.4/share/hadoop/common/lib/zookeeper-jute-3.5.6.jar:/home/isaleh/hadoop-3.3.4/share/hadoop/common/lib/jetty-webapp-9.4.43.v20210629.jar:/home/isaleh/hadoop-3.3.4/share/hadoop/common/lib/b/s/zookeeper-3.5.6.jar:/home/isaleh/hadoop-3.3.4/share/hadoop/common/lib/jetty-util-ajax-9.4.43.v20210629.jar:/home/isaleh/hadoop-3.3.4/share/hadoop/common/lib/kerb-identity-1.0.1.jar:/home/isaleh/hadoop-3.3.4/share/hadoop/common/lib/accenseore-smart-2.4.7.jar:/home/isaleh/hadoop-3.3.4/share/hadoop/common/lib/jackson-core-api-1.9.13.jar:/home/isaleh/hadoop-3.3.4/share/hadoop/common/lib/javaservlet-api-3.1.0.jar:/home/isaleh/hadoop-3.3.4/share/hadoop/common/lib/commons-beanutils-1.9.4.jar:/home/isaleh/hadoop-3.3.4/share/hadoop/common/lib/asm-5.0.4.jar:/home/isaleh/hadoop-3.3.4/share/hadoop/common/lib/jackson-core-2.12.7.jar:/home/isaleh/hadoop-3.3.4/share/hadoop/common/lib/curator-client-4.2.0.jar:/home/isaleh/hadoop-3.3.4/share/hadoop/common/lib/httpclient-4.5.13.jar:/home/isaleh/hadoop-3.3.4/share/hadoop/common/lib/listenablefuture-9999.0-empty-to-avoid-conflict-with-guava.jar:/home/isaleh/hadoop-3.3.4/share/hadoop/common/lib/gson-2.8.9.jar:/home/isaleh/hadoop-3.3.4/share/hadoop/common/lib/commons-logging-1.1.3.jar:/home/isaleh/hadoop-3.3.4/share/hadoop/common/lib/jersey-servlet-1.19.jar:/home/isaleh/hadoop-3.3.4/share/hadoop/common/lib/erby-xdr-1.0.1.jar:/home/isaleh/hadoop-3.3.4/share/hadoop/common/lib/jaxb-api-2.2.11.jar:/home/isaleh/hadoop-3.3.4/share/hadoop/common/lib/commons-configuration2-2.11.1.jar:/home/isaleh/hadoop-3.3.4/share/hadoop/common/lib/jetty-security-9.4.43.v20210629.jar:/home/isaleh/hadoop-3.3.4/share/hadoop/common/lib/jersey-core-1.19.jar:/home/isaleh/hadoop-3.3.4/share/hadoop/common/lib/httpcore-4.4.13.jar:/home/isaleh/hadoop-3.3.4/share/hadoop/common/lib/jch-0.1.55.jar:/home/isaleh/hadoop-3.3.4/share/hadoop/common/lib/commons-math3-3.1.1.jar:/home/isaleh/hadoop-3.3.4/share/hadoop/common/lib/audience-annotations-0.5.0.jar:/home/isaleh/hadoop-3.3.4/share/hadoop/common/lib/kerb-serve-1.0.1.jar:/home/isaleh/hadoop-3.3.4/share/hadoop/common/lib/jetty-util-9.4.43.v20210629.jar:/home/isaleh/hadoop-3.3.4/share/hadoop/common/lib/snappy-java-1.1.8.2.jar:/home/isaleh/hadoop-3.3.4/share/hadoop-3.3.4/share/hadoop/common/lib/jetty-http-9.4.43.v20210629.jar:/home/isaleh/hadoop-3.3.4/share/hadoop-3.3.4/share/
```



Implementation Cntd.

```
isaleh@instance-1:~/hadoop-3.3.4$ sbin/start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [instance-1]
instance-1: Warning: Permanently added 'instance-1,10.182.0.2' (ECDSA) to the list of known ho
sts.
isaleh@instance-1:~/hadoop-3.3.4$
```

Implementation Cntd.

```
isaleh@instance-1:~/hadoop-3.3.4$ wget http://localhost:9870/
--2023-08-20 02:30:54-- http://localhost:9870/
Resolving localhost (localhost)... ::1, 127.0.0.1
Connecting to localhost (localhost)|::1|:9870... failed: Connection refused.
Connecting to localhost (localhost)|127.0.0.1|:9870... connected.
HTTP request sent, awaiting response... 302 Found
Location: http://localhost:9870/index.html [following]
--2023-08-20 02:30:54-- http://localhost:9870/index.html
Reusing existing connection to localhost:9870.
HTTP request sent, awaiting response... 200 OK
Length: 1079 (1.1K) [text/html]
Saving to: 'index.html'

index.html          100%[=====>]    1.05K  --.-KB/s    in 0s

2023-08-20 02:30:54 (130 MB/s) - 'index.html' saved [1079/1079]
```

Implementation: Map Reduce

```
isaleh@instance-1:~/hadoop-3.3.4$ bin/hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-3.3.4.jar grep input output 'dfs[a-z.]+'
2023-08-20 08:04:04,449 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2023-08-20 08:04:04,590 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 seconds.
2023-08-20 08:04:04,590 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2023-08-20 08:04:05,118 INFO input.FileInputFormat: Total input files to process : 10
2023-08-20 08:04:05,147 INFO mapreduce.JobSubmitter: number of splits:10
2023-08-20 08:04:05,405 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local887915342_0001
2023-08-20 08:04:05,405 INFO mapreduce.JobSubmitter: Executing with tokens: {}
2023-08-20 08:04:05,617 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2023-08-20 08:04:05,618 INFO mapreduce.Job: Running job: job_local887915342_0001
2023-08-20 08:04:05,625 INFO mapred.LocalJobRunner: OutputCommitter set in config null
2023-08-20 08:04:05,635 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2023-08-20 08:04:05,635 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup temporary folders under output directory:false, ignore cleanup failures: false
2023-08-20 08:04:05,636 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
2023-08-20 08:04:05,748 INFO mapred.LocalJobRunner: Waiting for map tasks
2023-08-20 08:04:05,749 INFO mapred.LocalJobRunner: Starting task: attempt_local887915342_0001_m_000000_0
2023-08-20 08:04:05,792 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2023-08-20 08:04:05,794 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup temporary folders under output directory:false, ignore cleanup failures: false
2023-08-20 08:04:05,826 INFO mapred.Task: Using ResourceCalculatorProcessTree : [ ]
2023-08-20 08:04:05,832 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/user/isaleh/input/hadoop-policy.xml:0+11765
2023-08-20 08:04:05,916 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
2023-08-20 08:04:05,916 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
2023-08-20 08:04:05,916 INFO mapred.MapTask: soft limit at 83886080
2023-08-20 08:04:05,917 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
2023-08-20 08:04:05,917 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
2023-08-20 08:04:05,923 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
2023-08-20 08:04:06,195 INFO mapred.LocalJobRunner:
2023-08-20 08:04:06,205 INFO mapred.MapTask: Starting flush of map output
2023-08-20 08:04:06,205 INFO mapred.MapTask: Spilling map output
```

```
isaleh@instance-1:~/hadoop-3.3.4$ bin/hdfs dfs -get output output1
isaleh@instance-1:~/hadoop-3.3.4$ cat output1/*
1      dfsadmin
1      dfs.replication
isaleh@instance-1:~/hadoop-3.3.4$
```

Implementation: Input

```
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
            writer.flush();

            // closing the write after pushing the data inside the .txt file
            writer.close();
        } catch (IOException e) {
            e.printStackTrace();
        }
    }
}
```

```
job.setJarByClass(PiCalculation.class);
job.setMapperClass(TokenizerMapper.class);
job.setCombinerClass(IntSumReducer.class);
job.setReducerClass(IntSumReducer.class);
job.setOutputKeyClass(Text.class);
job.setOutputValueClass(IntWritable.class);
FileInputFormat.addInputPath(job, new Path(args[0]));
FileOutputFormat.setOutputPath(job, new Path(args[1]));
// System.exit(job.waitForCompletion(true) ? 0 : 1);
job.waitForCompletion(true);
String filePath = args[1] + "/" + "part-r-00000";
Path path = new Path(filePath);
FileSystem fs = FileSystem.get(path.toUri(), conf);

BufferedReader br = new BufferedReader(new InputStreamReader(fs.open(path)));

String z, inside = null, outside = null;

String line1, line2;

line1 = br.readLine();
System.out.println(line1);
line2 = br.readLine();
System.out.println(line2);

line1 = line1.replace("inside", "").trim();
line2 = line2.replace("outside", "").trim();

System.out.println("Inside:" + line1 + ", Outside:" + line2);

if (line1 != null && line2 != null) {
    double invalue = Double.valueOf(line1);
    double outvalue = Double.valueOf(line2);
    double pi = 4 * (invalue / (invalue + outvalue));
    System.out.println("PI:" + pi);
}

fs.close();
}
```


Implementation: Copy in Distributed File

```
isaleh@instance-1:~/hadoop-3.3.4$ bin/hdfs dfs -put etc/hadoop/*.xml input
2023-08-20 08:37:31,056 WARN hdfs.DataStreamer: DataStreamer Exception
org.apache.hadoop.ipc.RemoteException(java.io.IOException): File /user/isaleh/input/capacity-scheduler.xml, COPYING could only be written to 0 of the 1 minReplication nodes. There are 0 datanode(s) running and 0 node(s) are excluded in this operation.
    at org.apache.hadoop.hdfs.server.blockmanagement.BlockManager.chooseTarget4NewBlock(BlockManager.java:2315)
    at org.apache.hadoop.hdfs.server.namenode.FSDirWriteFileOp.chooseTargetForNewBlock(FSDirWriteFileOp.java:294)
    at org.apache.hadoop.hdfs.server.namenode.FSNamesystem.getAdditionalBlock(FSNamesystem.java:2960)
    at org.apache.hadoop.hdfs.server.namenode.NameNodeRpcServer.addBlock(NameNodeRpcServer.java:904)
    at org.apache.hadoop.hdfs.protocolPB.ClientNameNodeProtocolServerSideTranslatorPB.addBlock(ClientNameNodeProtocolServerSideTranslatorPB.java:593)
    at org.apache.hadoop.hdfs.protocol.proto.ClientNameNodeProtocolProtos$ClientNameNodeProtocol$2.callBlockingMethod(ClientNameNodeProtocolProtos.java)
    at org.apache.hadoop.ipc.ProtobufRpcEngine2$Server$ProtobufRpcInvoker.call(ProtobufRpcEngine2.java:604)
    at org.apache.hadoop.ipc.ProtobufRpcEngine2$Server$ProtobufRpcInvoker.call(ProtobufRpcEngine2.java:572)
    at org.apache.hadoop.ipc.ProtobufRpcEngine2$Server$ProtobufRpcInvoker.call(ProtobufRpcEngine2.java:550)
    at org.apache.hadoop.ipc.RPC$Server.call(RPC.java:1093)
    at org.apache.hadoop.ipc.Server$RpcCall.run(Server.java:1043)
    at org.apache.hadoop.ipc.Server$RpcCall.run(Server.java:971)
    at java.base/java.security.AccessController.doPrivileged(Native Method)
    at java.base/javax.security.auth.Subject.doAs(Subject.java:423)
    at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1878)
)
    at org.apache.hadoop.ipc.Server$Handler.run(Server.java:2976)
    at org.apache.hadoop.ipc.Client.getResponse(Client.java:1612)
    at org.apache.hadoop.ipc.Client.call(Client.java:1558)
    at org.apache.hadoop.ipc.Client.call(Client.java:1455)
    at org.apache.hadoop.ipc.ProtobufRpcEngine2$Invoker.invoke(ProtobufRpcEngine2.java:242)
)
    at org.apache.hadoop.ipc.ProtobufRpcEngine2$Invoker.invoke(ProtobufRpcEngine2.java:129)
)
    at com.sun.proxy.$Proxy12.addBlock(Unknown Source)
    at org.apache.hadoop.hdfs.protocolPB.ClientNameNodeProtocolTranslatorPB.addBlock(ClientNameNodeProtocolTranslatorPB.java:530)
    at java.base/jdk.internal.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
```

```
isaleh@instance-1:~/hadoop-3.3.4$ bin/hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-3.3.4.jar grep input output 'dfs[a-z.]+'
2023-08-20 08:37:44,511 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2023-08-20 08:37:44,692 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 seconds
2023-08-20 08:37:44,692 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2023-08-20 08:37:45,089 INFO input.FileInputFormat: Total input files to process : 0
2023-08-20 08:37:45,100 INFO mapreduce.JobSubmitter: number of splits:0
2023-08-20 08:37:45,362 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local500068553_0001
2023-08-20 08:37:45,362 INFO mapreduce.JobSubmitter: Executing with tokens: []
2023-08-20 08:37:45,649 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2023-08-20 08:37:45,650 INFO mapreduce.Job: Running job: job_local500068553_0001
2023-08-20 08:37:45,657 INFO mapred.LocalJobRunner: OutputCommitter set in config null
2023-08-20 08:37:45,671 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2023-08-20 08:37:45,672 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup temporary folders under output directory:false, ignore cleanup failures: false
2023-08-20 08:37:45,676 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
2023-08-20 08:37:45,736 INFO mapred.LocalJobRunner: Waiting for map tasks
2023-08-20 08:37:45,736 INFO mapred.LocalJobRunner: map task executor complete.
2023-08-20 08:37:45,743 INFO mapred.LocalJobRunner: Waiting for reduce tasks
2023-08-20 08:37:45,745 INFO mapred.LocalJobRunner: Starting task: attempt_local500068553_0001_r_000000_0
2023-08-20 08:37:45,788 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2023-08-20 08:37:45,792 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup temporary folders under output directory:false, ignore cleanup failures: false
2023-08-20 08:37:45,831 INFO mapred.Task: Using ResourceCalculatorProcessTree : [ ]
2023-08-20 08:37:45,835 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task.reduce.Shuffle@2b0d4db3
2023-08-20 08:37:45,838 WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
2023-08-20 08:37:45,857 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=722259136, maxSingleShuffleLimit=180564784, mergeThreshold=476691040, ioSortFactor=10, memToMemMergeOutputsThreshold=10
2023-08-20 08:37:45,878 INFO reduce.EventFetcher: attempt_local500068553_0001_r_000000_0 Thread started: EventFetcher for fetching Map Completion Events
2023-08-20 08:37:45,884 INFO mapred.LocalJobRunner:
2023-08-20 08:37:45,887 INFO reduce.MergeManagerImpl: finalMerge called with 0 in-memory map-outputs and 0 on-disk map-outputs
```



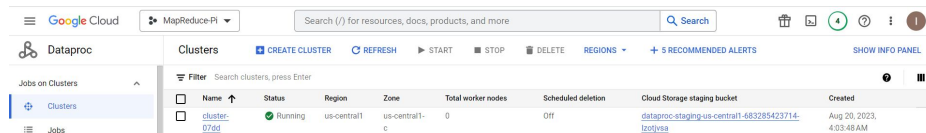
Implementation: Output

```
isaleh@instance-1:~/hadoop-3.3.4$ bin/hdfs dfs -get output output1
isaleh@instance-1:~/hadoop-3.3.4$ cat output1/*
cat: output1/output: Is a directory
1      dfsadmin
1      dfs.replication
isaleh@instance-1:~/hadoop-3.3.4$
```

```
isaleh@instance-1:~/hadoop-3.3.4/output$ ls
_SUCCESS part-r-00000
```

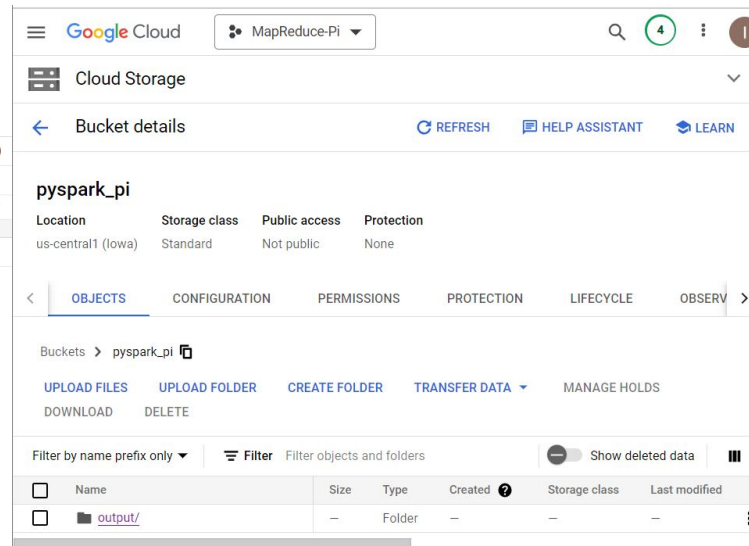
```
inside 106
outside 44
```


Implementation: Pyspark



The screenshot shows the Google Cloud Dataproc Clusters page. The left sidebar has a 'Clusters' link highlighted. The main content area shows a table of clusters. One cluster is listed with the name 'cluster-07ed', status 'Running', region 'us-central1', zone 'us-central1-c', 0 total worker nodes, and a scheduled deletion of 'Off'. The Cloud Storage staging bucket is 'dataproc-staging-us-central1-483285423714-lzotlva'. The cluster was created on Aug 20, 2023, at 4:03:48 AM.

Name	Status	Region	Zone	Total worker nodes	Scheduled deletion	Cloud Storage staging bucket	Created
cluster-07ed	Running	us-central1	us-central1-c	0	Off	dataproc-staging-us-central1-483285423714-lzotlva	Aug 20, 2023, 4:03:48 AM



The screenshot shows the Google Cloud Cloud Storage Bucket details page for 'pyspark_pi'. The bucket is located in 'us-central1 (Iowa)', has a 'Standard' storage class, is 'Not public', and has 'None' protection. The 'OBJECTS' tab is selected, showing a list of objects. There is one object named 'output/' which is a folder. The bucket also has options for 'UPLOAD FILES', 'UPLOAD FOLDER', 'CREATE FOLDER', 'TRANSFER DATA', 'MANAGE HOLDS', 'DOWNLOAD', and 'DELETE'.

Name	Size	Type	Created	Storage class	Last modified
output/	-	Folder	-	-	-



Implementation: Pyspark

```
Welcome to Cloud Shell! Type "help" to get started.
Your Cloud Platform project in this session is set to chrome-unity-396501.
Use "gcloud config set project [PROJECT_ID]" to change to a different project.
isaleh@cloudshell:~ (chrome-unity-396501)$ gcloud auth login
```

```
You are already authenticated with gcloud when running
inside the Cloud Shell and so do not need to run this
command. Do you wish to proceed anyway?
```

```
Do you want to continue (Y/n)? y
```

```
Go to the following link in your browser:
```

```
https://accounts.google.com/o/oauth2/auth?response_type=code&client_id=32555940559.apps.googleusercontent.com&redirect_uri=https%3A%2F%2Fsdk.cloud.google.com%2Fauthcode.html&scope=openid+https%3A%2F%2Fwww.googleapis.com%2Fauth%2Fuserinfo.email+https%3A%2F%2Fwww.googleapis.com%2Fauth%2Fcloud-platform+https%3A%2F%2Fwww.googleapis.com%2Fauth%2Fappengine.admin+https%3A%2F%2Fwww.googleapis.com%2Fauth%2Fsqlservice.login+https%3A%2F%2Fwww.googleapis.com%2Fauth%2Fcompute+https%3A%2F%2Fwww.googleapis.com%2Fauth%2Faccounts.reauth&state=DBCJMPMYJuuMOG7UZrvamYBnKacEf2&prompt=consent&access_type=offline&code_challenge=ppaOQDzoHdo1T6kMkQLSTrwoRx3HNwsA6rJVZ_MrRQ&code_challenge_method=S256
```

```
Enter authorization code: 4/0Adeu5BUA5J2-EMOA4S16b5uyG0mSmXJj616AfLD0pjL7sw0KVSUNUN7wGvAuGccrEvMeJOW
```

```
You are now logged in as [isaleh@student.sfbu.edu].
Your current project is [chrome-unity-396501]. You can change this setting by running:
$ gcloud config set project PROJECT_ID
isaleh@cloudshell:~ (chrome-unity-396501)$
```



Implementation: Pyspark

```
isaleh@cloudshell:~ (chrome-unity-396501)$ gcloud dataproc jobs submit pyspark pi.py --cluster
=cluster-07dd --region=us-central1 -- --partitions 4 --output_uri gs://pyspark_pi/output
Job [b0e0a6e1532f4a74b49cd25689217e10] submitted.
Waiting for job output...
INFO: Calculating pi with a total of 400000 tries in 4 partitions.
23/08/20 11:15:51 INFO SparkEnv: Registering MapOutputTracker
23/08/20 11:15:51 INFO SparkEnv: Registering BlockManagerMaster
23/08/20 11:15:51 INFO SparkEnv: Registering BlockManagerMasterHeartbeat
23/08/20 11:15:51 INFO SparkEnv: Registering OutputCommitCoordinator
23/08/20 11:15:52 INFO DefaultNoHARMAFailoverProxyProvider: Connecting to ResourceManager at cl
uster-07dd-m.us-central1-c.c.chrome-unity-396501.internal./10.128.0.2:8032
23/08/20 11:15:52 INFO AHSProxy: Connecting to Application History server at cluster-07dd-m.us
-central1-c.c.chrome-unity-396501.internal./10.128.0.2:10200
23/08/20 11:15:53 INFO Configuration: resource-types.xml not found
23/08/20 11:15:53 INFO ResourceUtils: Unable to find 'resource-types.xml'.
23/08/20 11:15:55 INFO YarnClientImpl: Submitted application application_1692529499003_0001
23/08/20 11:15:56 INFO DefaultNoHARMAFailoverProxyProvider: Connecting to ResourceManager at cl
uster-07dd-m.us-central1-c.c.chrome-unity-396501.internal./10.128.0.2:8030
23/08/20 11:15:58 INFO GoogleCloudStorageImpl: Ignoring exception of type GoogleJsonResponseEx
ception: verified object already exists with desired state.
INFO: 400000 tries and 314188 hits gives pi estimate of 3.14188.
INFO: NumExpr defaulting to 4 threads.
23/08/20 11:16:26 INFO GoogleCloudStorageFileSystem: Successfully repaired 'gs://pyspark_pi/ou
tput/' directory.
INFO: Closing down clientserver connection
Job [b0e0a6e1532f4a74b49cd25689217e10] finished successfully.
done: true
driverControlFilesUri: gs://dataproc-staging-us-central1-683285423714-lzotjvsa/google-cloud-da
taproc-metainfo/caabff92-d2f2-4561-bc7a-4cd5f829fee1/jobs/b0e0a6e1532f4a74b49cd25689217e10/
driverOutputResourceUri: gs://dataproc-staging-us-central1-683285423714-lzotjvsa/google-cloud-
dataproc-metainfo/caabff92-d2f2-4561-bc7a-4cd5f829fee1/jobs/b0e0a6e1532f4a74b49cd25689217e10/d
riveroutput
jobUuid: af149d2d-174c-37be-97d9-fd586a5d0b92
placement:
```

Implementation: Pyspark Output

```
isaleh@cloudshell:~ (chrome-unity-396501)$ gsutil ls gs://pyspark_pi gs://pyspark_pi/output/
gs://pyspark_pi/:
gs://pyspark_pi/output/
gs://pyspark_pi/output/
gs://pyspark_pi/output/_SUCCESS
gs://pyspark_pi/output/part-00000-c2964914-2262-4f94-8cc2-5d325c4e0ab7-c000.json
gs://pyspark_pi/output/part-00003-c2964914-2262-4f94-8cc2-5d325c4e0ab7-c000.json
isaleh@cloudshell:~ (chrome-unity-396501)$ gsutil cp gs://pyspark_pi/output/*.
CommandException: Wrong number of arguments for "cp" command.
isaleh@cloudshell:~ (chrome-unity-396501)$ gsutil cp gs://pyspark_pi/output/* .
Copying gs://pyspark_pi/output/_SUCCESS...
Copying gs://pyspark_pi/output/part-00000-c2964914-2262-4f94-8cc2-5d325c4e0ab7-c000.json...
Copying gs://pyspark_pi/output/part-00003-c2964914-2262-4f94-8cc2-5d325c4e0ab7-c000.json...
/ [3 files][ 44.0 B/ 44.0 B]
Operation completed over 3 objects/44.0 B.
isaleh@cloudshell:~ (chrome-unity-396501)$ gs://pyspark_pi/output/part-00003-c2964914-2262-4f9
isaleh@cloudshell:~ (chrome-unity-396501)$ cat part-00000-c2964914-2262-4f94-8cc2-5d325c4e0ab7
-c000.json >> part-00003-c2964914-2262-4f94-8cc2-5d325c4e0ab7-c000.json
isaleh@cloudshell:~ (chrome-unity-396501)$ cat part-00003-c2964914-2262-4f94-8cc2-5d325c4e0ab7
-c000.json
{"tries":400000,"hits":314188,"pi":3.14188}
isaleh@cloudshell:~ (chrome-unity-396501)$
```