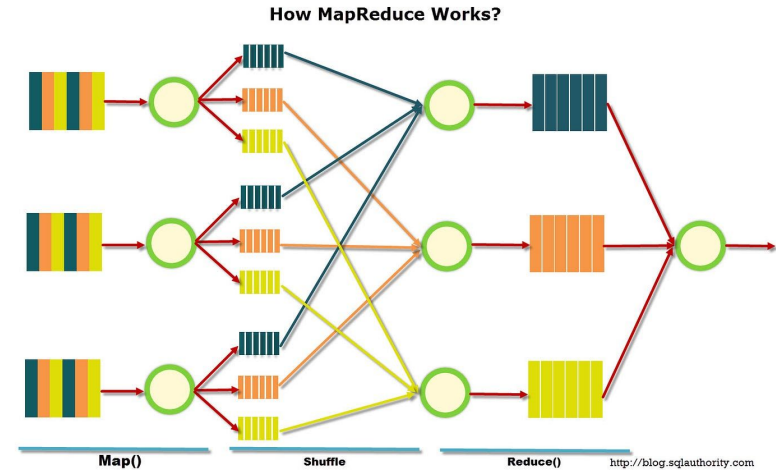# Calculating Pi using MapReduce
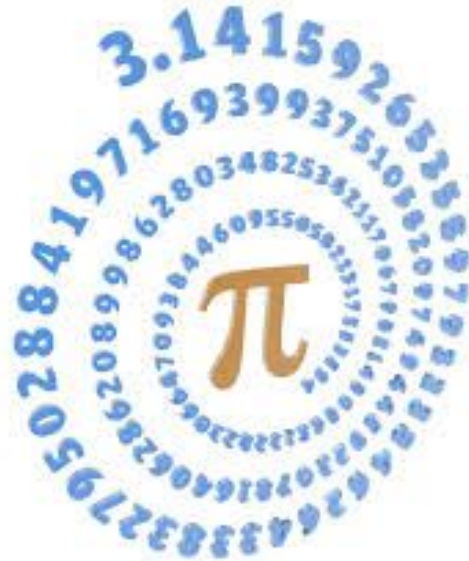
CS570: Big Data Processing & Analytics
Submitted by Imran Noor Saleh   Student ID 19648

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.
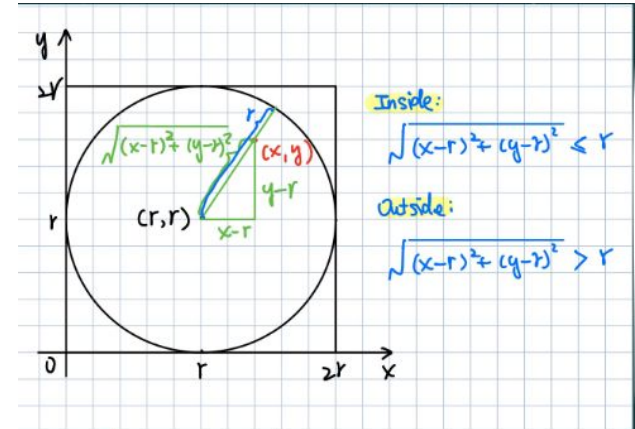

How MapReduce Works?

Map()   Shuffle   Reduce()   http://blog.sqlauthority.com

Introduction: Pi

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) = π

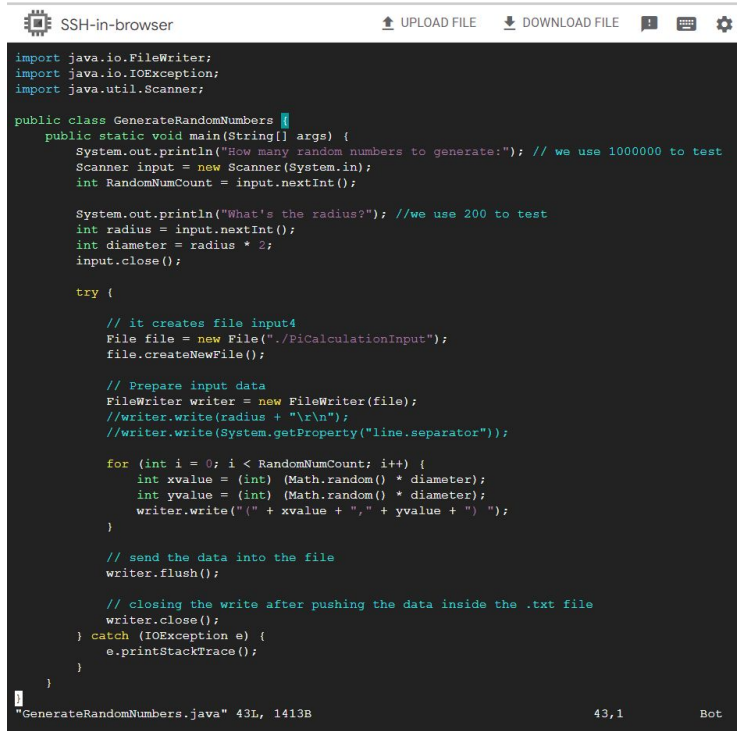# Design

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

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
dtutil        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
<?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
<?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.

Implementation Cntd.

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-exam
ples-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.proper
ties
2023-08-20 08:04:04,590 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 se
cond(s).
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_local88791
5342_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 versi
on is 2
2023-08-20 08:04:05,635 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _tem
porary folders under output directory:false, ignore cleanup failures: false
2023-08-20 08:04:05,636 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapre
duce.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 versi
on is 2
2023-08-20 08:04:05,794 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _tem
porary 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/isal
eh/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.ma
pred.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
```

```
                          bytes written 29
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

```java
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();
        }
    }
}
```

```java
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-s
cheduler.xml._COPYING_ could only be written to 0 of the 1 minReplication nodes. There are 0 d
atanode(s) running and 0 node(s) are excluded in this operation.
        at org.apache.hadoop.hdfs.server.blockmanagement.BlockManager.chooseTarget4NewBlock(Bl
ockManager.java:2315)
        at org.apache.hadoop.hdfs.server.namenode.FSDirWriteFileOp.chooseTargetForNewBlock(FSD
irWriteFileOp.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.addB
lock(ClientNamenodeProtocolServerSideTranslatorPB.java:593)
        at org.apache.hadoop.hdfs.protocol.proto.ClientNamenodeProtocolProtos$ClientNamenodePr
otocol$2.callBlockingMethod(ClientNamenodeProtocolProtos.java)
        at org.apache.hadoop.ipc.ProtobufRpcEngine2$Server$ProtoBufRpcInvoker.call(ProtobufRpc
Engine2.java:604)
        at org.apache.hadoop.ipc.ProtobufRpcEngine2$Server$ProtoBufRpcInvoker.call(ProtobufRpc
Engine2.java:572)
        at org.apache.hadoop.ipc.ProtobufRpcEngine2$Server$ProtoBufRpcInvoker.call(ProtobufRpc
Engine2.java:556)
        at org.apache.hadoop.ipc.RPC$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.getRpcResponse(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(Clien
tNamenodeProtocolTranslatorPB.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-exam
ples-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.proper
ties
2023-08-20 08:37:44,692 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 se
cond(s).
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_local50006
8553_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 versi
on is 2
2023-08-20 08:37:45,672 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _tem
porary folders under output directory:false, ignore cleanup failures: false
2023-08-20 08:37:45,676 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapre
duce.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 versi
on is 2
2023-08-20 08:37:45,792 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _tem
porary 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 initial
ized!
2023-08-20 08:37:45,857 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=722259136, ma
xSingleShuffleLimit=180564784, mergeThreshold=476691040, ioSortFactor=10, memToMemMergeOutputs
Threshold=10
2023-08-20 08:37:45,878 INFO reduce.EventFetcher: attempt_local500068553_0001_r_000000_0 Threa
d 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-o
utputs 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
```