Calculating Pi - PySpark implementation

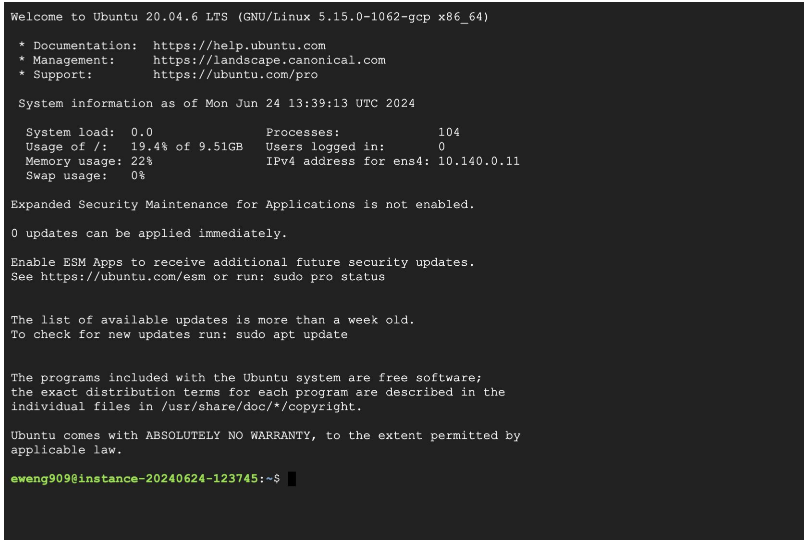
Emily Weng

Overview:

1. Use previous project: Project: Creating MapReduce program to calculating Pi
2. Add the implementation of Pi Calculation using PySpark

Follow the step for Pi on MapReduce

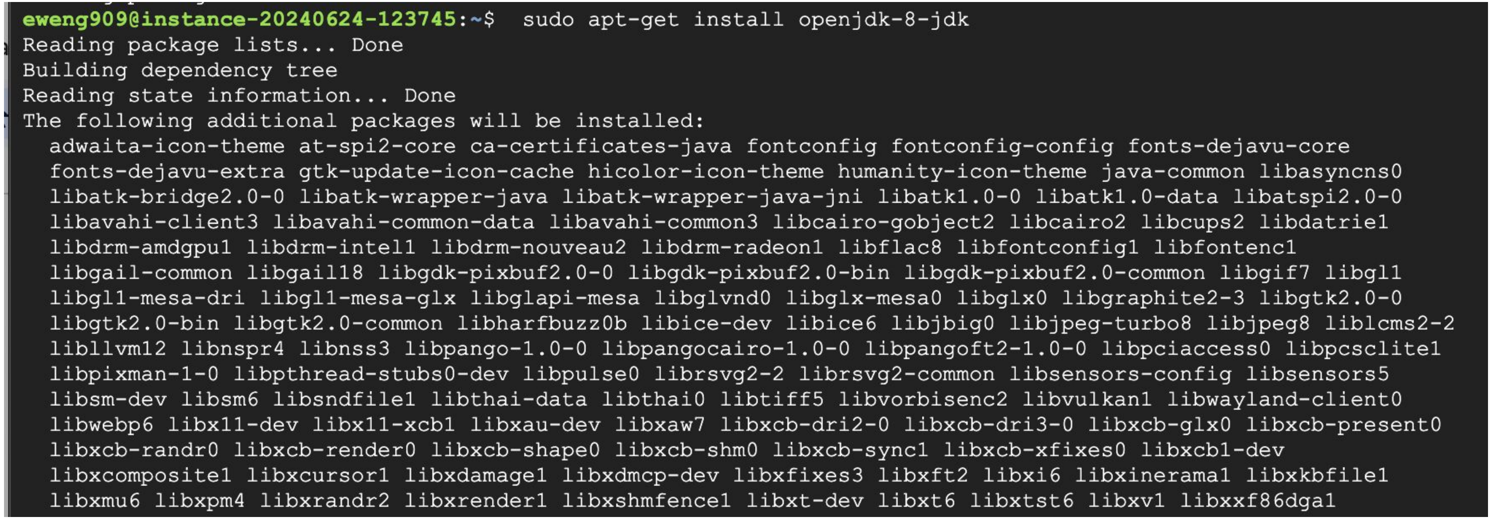
1. Make your instance and open SSH Browser
   a.



```
Welcome to Ubuntu 20.04.6 LTS (GNU/Linux 5.15.0-1062-gcp x86_64)

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

 System information as of Mon Jun 24 13:39:13 UTC 2024

  System load:  0.0               Processes:             104
  Usage of /:   19.4% of 9.51GB   Users logged in:       0
  Memory usage: 22%               IPv4 address for ens4: 10.140.0.11
  Swap usage:   0%

Expanded Security Maintenance for Applications is not enabled.

0 updates can be applied immediately.

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


The list of available updates is more than a week old.
To check for new updates run: sudo apt update


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

Ubuntu comes with ABSOLUTELY NO WARRANTY, to the extent permitted by
applicable law.

eweng909@instance-20240624-123745:~$
```

Install Java JDK

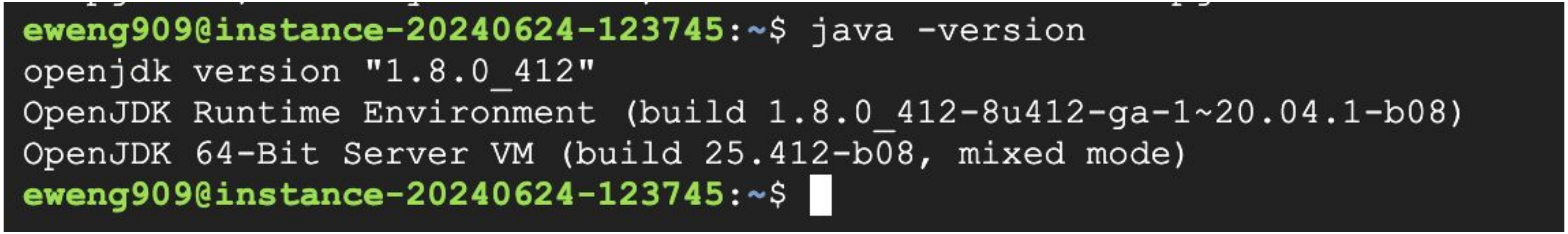
1. sudo apt-get install openjdk-8-jdk



```
eweng909@instance-20240624-123745:~$  sudo apt-get install openjdk-8-jdk
Reading package lists... Done
Building dependency tree
Reading state information... Done
The following additional packages will be installed:
  adwaita-icon-theme at-spi2-core ca-certificates-java fontconfig fontconfig-config fonts-dejavu-core
  fonts-dejavu-extra gtk-update-icon-cache hicolor-icon-theme humanity-icon-theme java-common libasyncns0
  libatk-bridge2.0-0 libatk-wrapper-java libatk-wrapper-java-jni libatk1.0-0 libatk1.0-data libatspi2.0-0
  libavahi-client3 libavahi-common-data libavahi-common3 libcairo-gobject2 libcairo2 libcups2 libdatrie1
  libdrm-amdgpu1 libdrm-intel1 libdrm-nouveau2 libdrm-radeon1 libflac8 libfontconfig1 libfontenc1
  libgail-common libgail18 libgdk-pixbuf2.0-0 libgdk-pixbuf2.0-bin libgdk-pixbuf2.0-common libgif7 libgl1
  libgl1-mesa-dri libgl1-mesa-glx libglapi-mesa libglvnd0 libglx-mesa0 libglx0 libgraphite2-3 libgtk2.0-0
  libgtk2.0-bin libgtk2.0-common libharfbuzz0b libice-dev libice6 libjbig0 libjpeg-turbo8 libjpeg8 liblcms2-2
  libllvm12 libnspr4 libnss3 libpango-1.0-0 libpangocairo-1.0-0 libpangoft2-1.0-0 libpciaccess0 libpcsclite1
  libpixman-1-0 libpthread-stubs0-dev libpulse0 librsvg2-2 librsvg2-common libsensors-config libsensors5
  libsm-dev libsm6 libsndfile1 libthai-data libthai0 libtiff5 libvorbisenc2 libvulkan1 libwayland-client0
  libwebp6 libx11-dev libx11-xcb1 libxau-dev libxaw7 libxcb-dri2-0 libxcb-dri3-0 libxcb-glx0 libxcb-present0
  libxcb-randr0 libxcb-render0 libxcb-shape0 libxcb-shm0 libxcb-sync1 libxcb-xfixes0 libxcb1-dev
  libxcomposite1 libxcursor1 libxdamage1 libxdmcp-dev libxfixes3 libxft2 libxi6 libxinerama1 libxkbfile1
  libxmu6 libxpm4 libxrandr2 libxrender1 libxshmfence1 libxt-dev libxt6 libxtst6 libxv1 libxxf86dga1
```

Check version

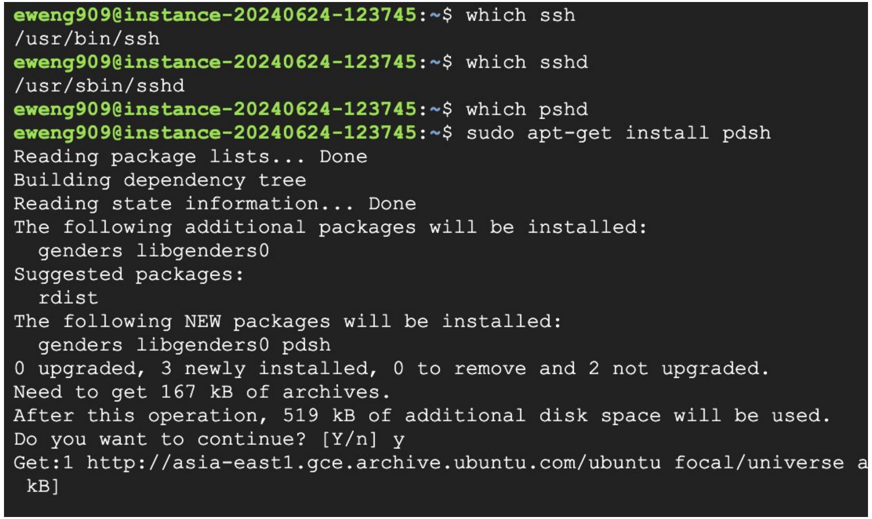
1. $ java -version



```
eweng909@instance-20240624-123745:~$ java -version
openjdk version "1.8.0_412"
OpenJDK Runtime Environment (build 1.8.0_412-8u412-ga-1~20.04.1-b08)
OpenJDK 64-Bit Server VM (build 25.412-b08, mixed mode)
eweng909@instance-20240624-123745:~$
```

Install ssh, sshd, phsd

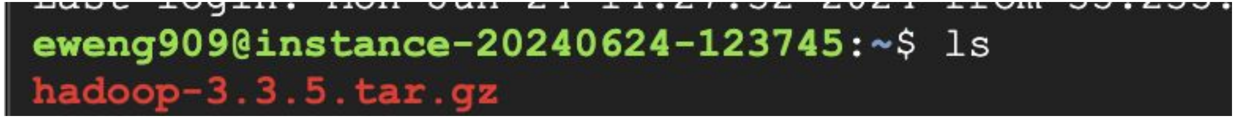
1. Check if ssh/sshd/pdsh exists already, if not, install them
   a. which ssh
   b. which sshd
   c. which pshd



```
eweng909@instance-20240624-123745:~$ which ssh
/usr/bin/ssh
eweng909@instance-20240624-123745:~$ which sshd
/usr/sbin/sshd
eweng909@instance-20240624-123745:~$ which pshd
eweng909@instance-20240624-123745:~$ sudo apt-get install pdsh
Reading package lists... Done
Building dependency tree
Reading state information... Done
The following additional packages will be installed:
  genders libgenders0
Suggested packages:
  rdist
The following NEW packages will be installed:
  genders libgenders0 pdsh
0 upgraded, 3 newly installed, 0 to remove and 2 not upgraded.
Need to get 167 kB of archives.
After this operation, 519 kB of additional disk space will be used.
Do you want to continue? [Y/n] y
Get:1 http://asia-east1.gce.archive.ubuntu.com/ubuntu focal/universe a
 kB]
```

Download Hadoop 3.3.5

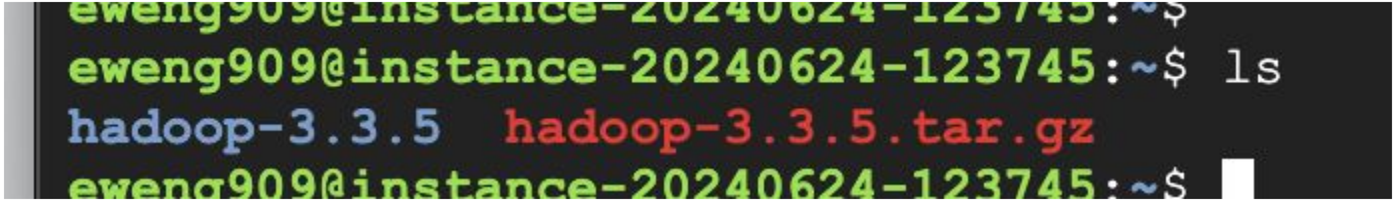
1. wget https://dlcdn.apache.org/hadoop/common/hadoop-3.3.5/hadoop-3.3.5.tar.gz



```
Last login: Mon Jun 24 14:27:52 2024 from 35.235.
eweng909@instance-20240624-123745:~$ ls
hadoop-3.3.5.tar.gz
```

Unzip the tar file

1. Unzip the tar file
   a. $ tar xzf hadoop-3.3.5.tar.gz



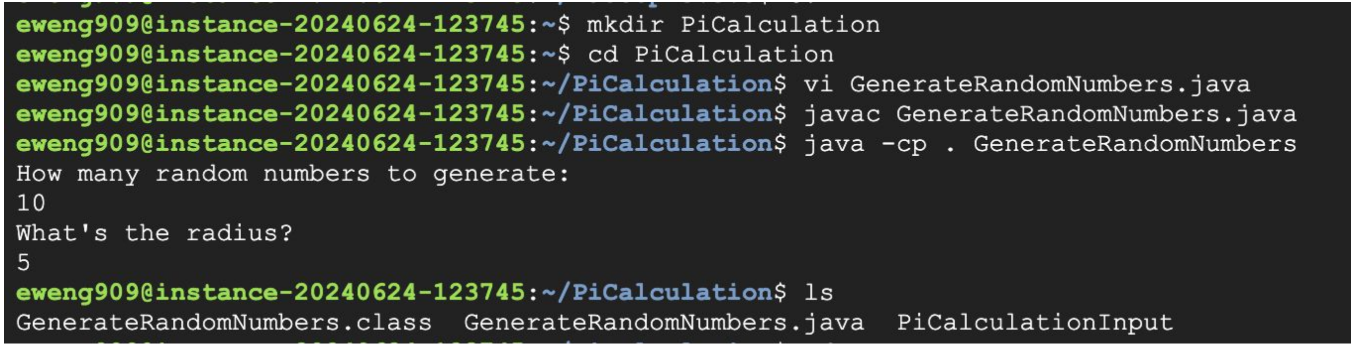
```
eweng909@instance-20240624-123745:~$
eweng909@instance-20240624-123745:~$ ls
hadoop-3.3.5  hadoop-3.3.5.tar.gz
eweng909@instance-20240624-123745:~$
```

Set up the rest of Hadoop Environment:

1. Modify bashrc file and set java and Hadoop environment
2. Configure HDFS

Prepare input data

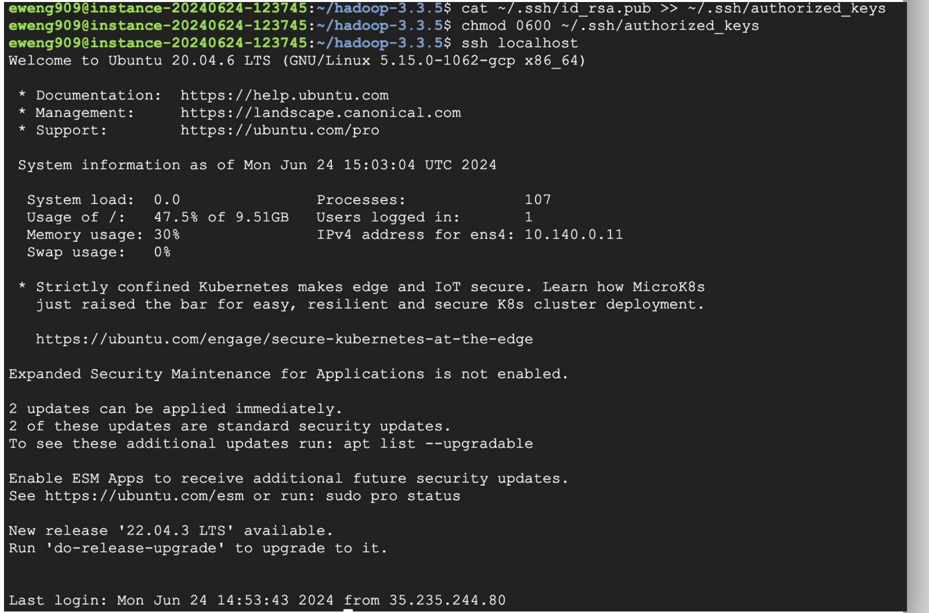
1. $ mkdir PiCalculation
2. $ cd PiCalculation
3. $ vi GenerateRandomNumbers.java
4. $ javac GenerateRandomNumbers.java
5. $ java -cp . GenerateRandomNumbers



```
eweng909@instance-20240624-123745:~$ mkdir PiCalculation
eweng909@instance-20240624-123745:~$ cd PiCalculation
eweng909@instance-20240624-123745:~/PiCalculation$ vi GenerateRandomNumbers.java
eweng909@instance-20240624-123745:~/PiCalculation$ javac GenerateRandomNumbers.java
eweng909@instance-20240624-123745:~/PiCalculation$ java -cp . GenerateRandomNumbers
How many random numbers to generate:
10
What's the radius?
5
eweng909@instance-20240624-123745:~/PiCalculation$ ls
GenerateRandomNumbers.class  GenerateRandomNumbers.java  PiCalculationInput
```

Set up paraphrase less SSH

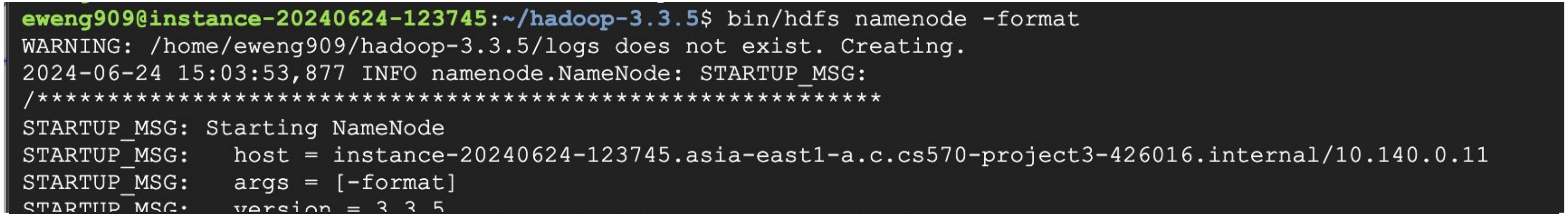
1. ssh-keygen -t rsa -P " -f ~/.ssh/id_rsa
2. cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
3. chmod 0600 ~/.ssh/authorized_keys
4. ssh localhost



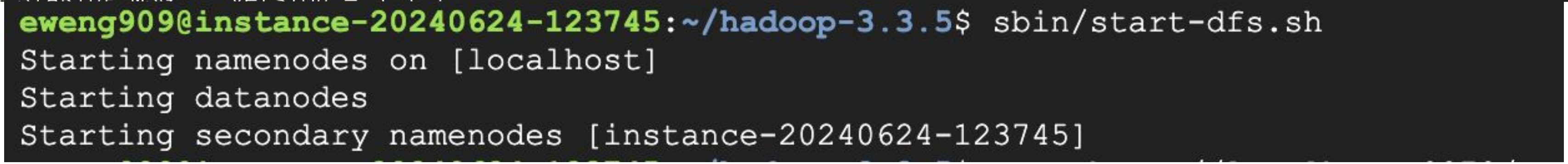
```
eweng909@instance-20240624-123745:~/hadoop-3.3.5$ cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
eweng909@instance-20240624-123745:~/hadoop-3.3.5$ chmod 0600 ~/.ssh/authorized_keys
eweng909@instance-20240624-123745:~/hadoop-3.3.5$ ssh localhost
Welcome to Ubuntu 20.04.6 LTS (GNU/Linux 5.15.0-1062-gcp x86_64)

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

 System information as of Mon Jun 24 15:03:04 UTC 2024

  System load:  0.0                Processes:             107
  Usage of /:   47.5% of 9.51GB    Users logged in:       1
  Memory usage: 30%                IPv4 address for ens4: 10.140.0.11
  Swap usage:   0%

 * Strictly confined Kubernetes makes edge and IoT secure. Learn how MicroK8s
   just raised the bar for easy, resilient and secure K8s cluster deployment.

   https://ubuntu.com/engage/secure-kubernetes-at-the-edge

Expanded Security Maintenance for Applications is not enabled.

2 updates can be applied immediately.
2 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 24 14:53:43 2024 from 35.235.244.80
```

Make the HDFS directories required to execute MapReduce jobs

1. cd hadoop-3.3.5
2. bin/hdfs namenode -format
3. sbin/start-dfs.sh



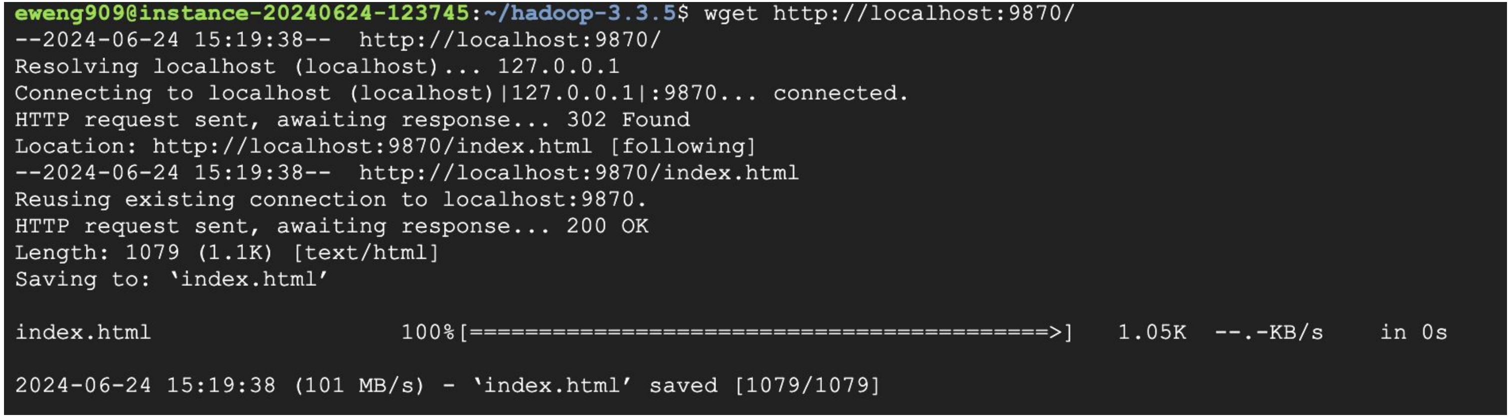
```
eweng909@instance-20240624-123745:~/hadoop-3.3.5$ bin/hdfs namenode -format
WARNING: /home/eweng909/hadoop-3.3.5/logs does not exist. Creating.
2024-06-24 15:03:53,877 INFO namenode.NameNode: STARTUP_MSG:
/************************************************************
STARTUP_MSG: Starting NameNode
STARTUP_MSG:   host = instance-20240624-123745.asia-east1-a.c.cs570-project3-426016.internal/10.140.0.11
STARTUP_MSG:   args = [-format]
STARTUP_MSG:   version = 3.3.5
```



```
eweng909@instance-20240624-123745:~/hadoop-3.3.5$ sbin/start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [instance-20240624-123745]
```

Continue:

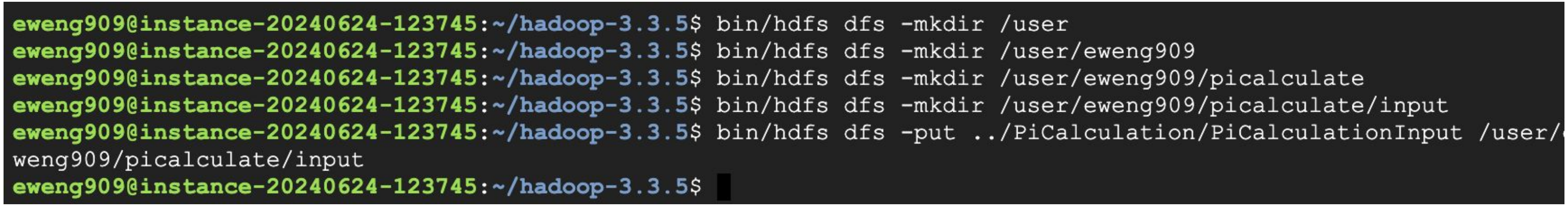
1. Update hdfs-site.xml and core-site.xml file
2. wget http://localhost:9870/



```
eweng909@instance-20240624-123745:~/hadoop-3.3.5$ wget http://localhost:9870/
--2024-06-24 15:19:38--  http://localhost:9870/
Resolving localhost (localhost)... 127.0.0.1
Connecting to localhost (localhost)|127.0.0.1|:9870... connected.
HTTP request sent, awaiting response... 302 Found
Location: http://localhost:9870/index.html [following]
--2024-06-24 15:19:38--  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

2024-06-24 15:19:38 (101 MB/s) - ‘index.html’ saved [1079/1079]
```

Continue:

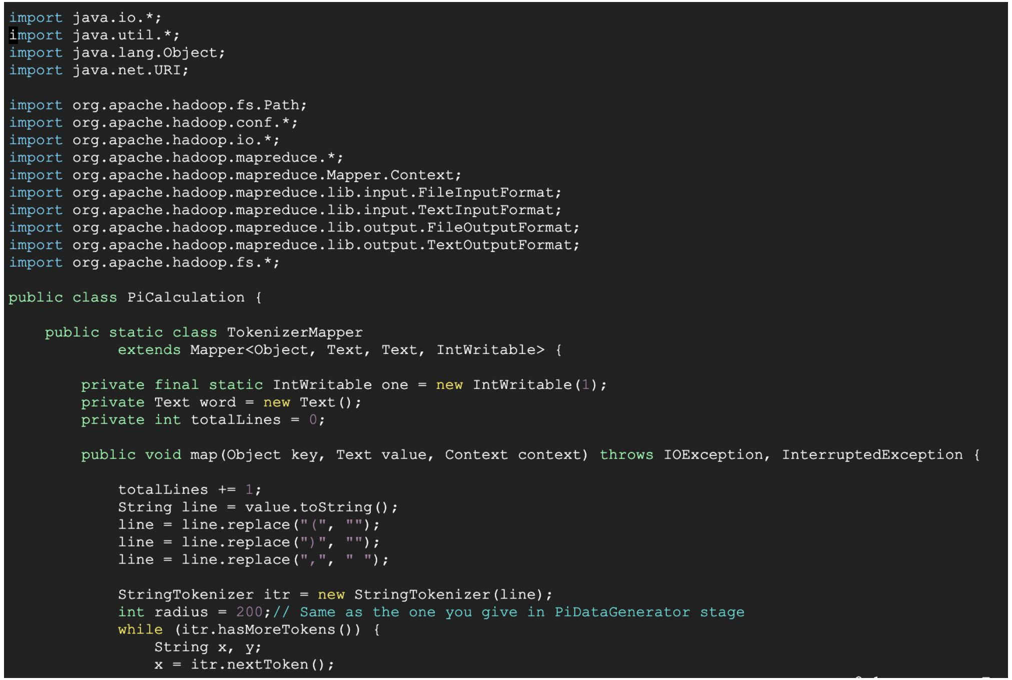
1. bin/hdfs dfs -mkdir /user
2. bin/hdfs dfs -mkdir /user/eweng909
3. bin/hdfs dfs -mkdir /user/eweng909/picalculate
4. bin/hdfs dfs -mkdir /user/eweng909/picalculate/input
5. bin/hdfs dfs -put ../PiCalculation/PiCalculationInput /user/eweng909/picalculate/input



```
eweng909@instance-20240624-123745:~/hadoop-3.3.5$ bin/hdfs dfs -mkdir /user
eweng909@instance-20240624-123745:~/hadoop-3.3.5$ bin/hdfs dfs -mkdir /user/eweng909
eweng909@instance-20240624-123745:~/hadoop-3.3.5$ bin/hdfs dfs -mkdir /user/eweng909/picalculate
eweng909@instance-20240624-123745:~/hadoop-3.3.5$ bin/hdfs dfs -mkdir /user/eweng909/picalculate/input
eweng909@instance-20240624-123745:~/hadoop-3.3.5$ bin/hdfs dfs -put ../PiCalculation/PiCalculationInput /user/
weng909/picalculate/input
eweng909@instance-20240624-123745:~/hadoop-3.3.5$ █
```

Build PiCalculation java file

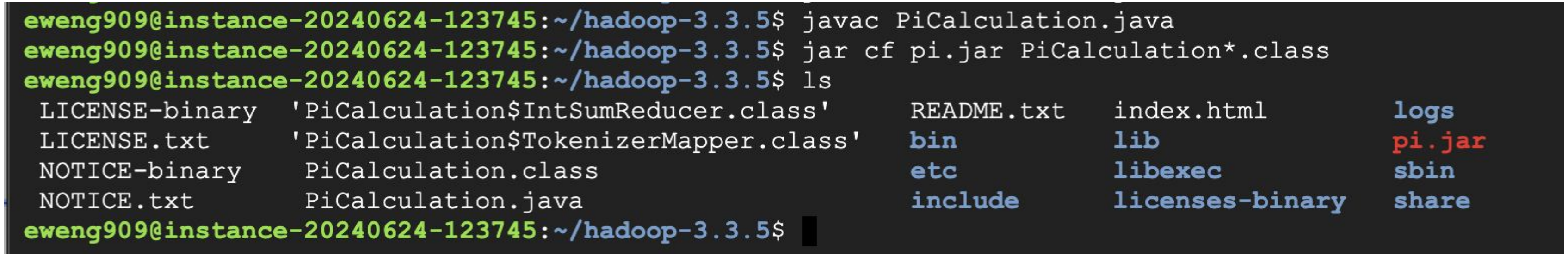
1. cd hadoop-3.3.5
2. vi PiCalculation.java



```java
import java.io.*;
import java.util.*;
import java.lang.Object;
import java.net.URI;

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.Mapper.Context;
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;
import org.apache.hadoop.fs.*;

public class PiCalculation {

    public static class TokenizerMapper
            extends Mapper<Object, Text, Text, IntWritable> {

        private final static IntWritable one = new IntWritable(1);
        private Text word = new Text();
        private int totalLines = 0;

        public void map(Object key, Text value, Context context) throws IOException, InterruptedException {

            totalLines += 1;
            String line = value.toString();
            line = line.replace("(", "");
            line = line.replace(")", "");
            line = line.replace(",", " ");

            StringTokenizer itr = new StringTokenizer(line);
            int radius = 200;// Same as the one you give in PiDataGenerator stage
            while (itr.hasMoreTokens()) {
                String x, y;
                x = itr.nextToken();
```

Compile PiCalculation.java and create a jar

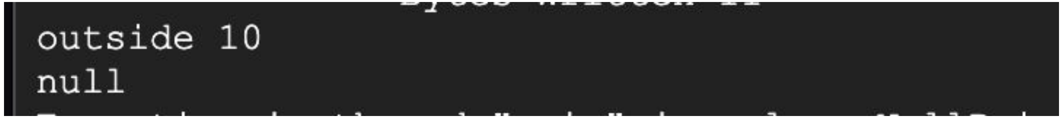
1. javac PiCalculation.java
2. jar cf wc.jar PiCalculation*class



```
eweng909@instance-20240624-123745:~/hadoop-3.3.5$ javac PiCalculation.java
eweng909@instance-20240624-123745:~/hadoop-3.3.5$ jar cf pi.jar PiCalculation*.class
eweng909@instance-20240624-123745:~/hadoop-3.3.5$ ls
 LICENSE-binary   'PiCalculation$IntSumReducer.class'     README.txt   index.html        logs
 LICENSE.txt      'PiCalculation$TokenizerMapper.class'   bin          lib               pi.jar
 NOTICE-binary     PiCalculation.class                    etc          libexec           sbin
 NOTICE.txt        PiCalculation.java                     include      licenses-binary   share
eweng909@instance-20240624-123745:~/hadoop-3.3.5$ █
```

Results

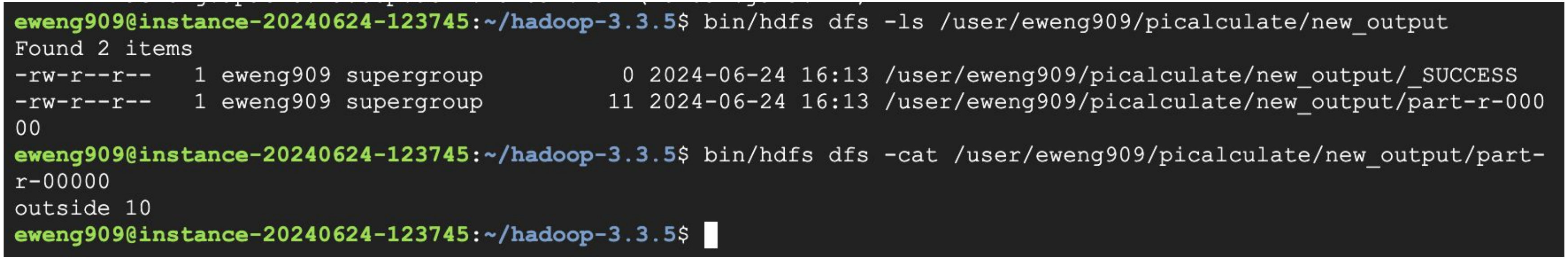
hadoop jar pi.jar PiCalculation /user/eweng909/picalculate/input /user/eweng909/picalculate/output



```
outside 10
null
```

Output

1. bin/hdfs dfs -ls /user/eweng909/picalculate/new_output
2. bin/hdfs dfs -cat /user/eweng909/picalculate/new_output/part-r-00000

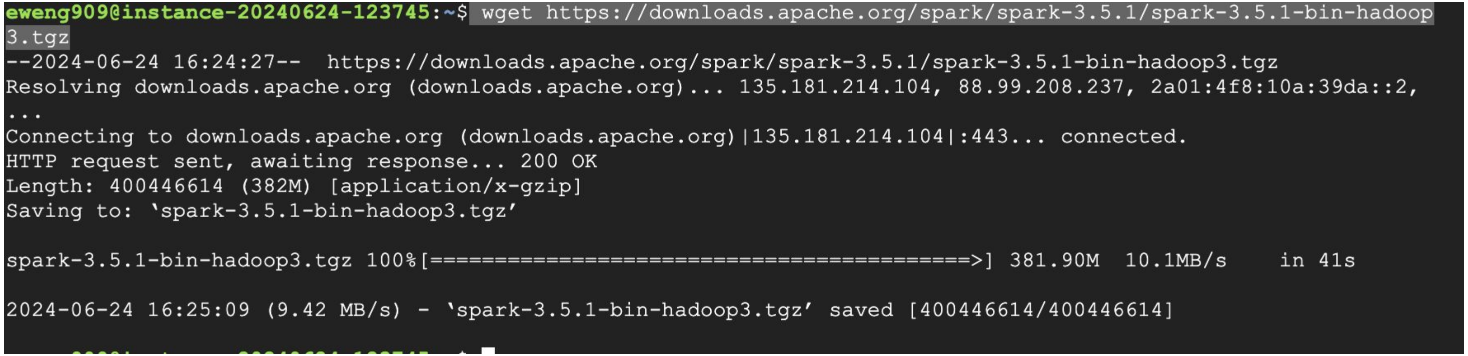


```
eweng909@instance-20240624-123745:~/hadoop-3.3.5$ bin/hdfs dfs -ls /user/eweng909/picalculate/new_output
Found 2 items
-rw-r--r--   1 eweng909 supergroup          0 2024-06-24 16:13 /user/eweng909/picalculate/new_output/_SUCCESS
-rw-r--r--   1 eweng909 supergroup         11 2024-06-24 16:13 /user/eweng909/picalculate/new_output/part-r-000
00
eweng909@instance-20240624-123745:~/hadoop-3.3.5$ bin/hdfs dfs -cat /user/eweng909/picalculate/new_output/part-
r-00000
outside 10
eweng909@instance-20240624-123745:~/hadoop-3.3.5$
```

With PySpark

Using pyspark

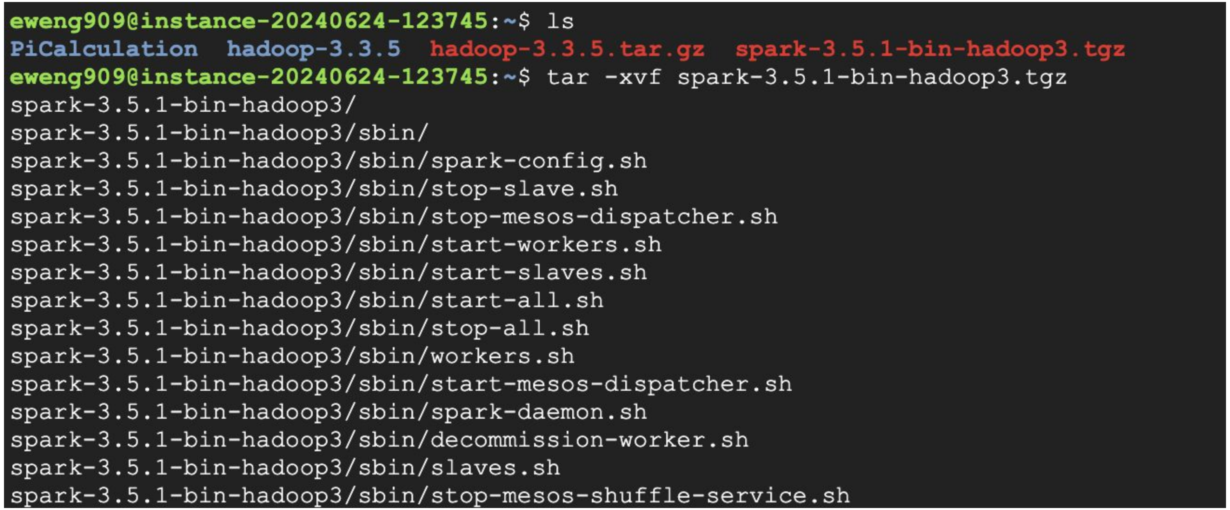
1. wget https://downloads.apache.org/spark/spark-3.5.1/spark-3.5.1-bin-hadoop3.tgz



```
eweng909@instance-20240624-123745:~$ wget https://downloads.apache.org/spark/spark-3.5.1/spark-3.5.1-bin-hadoop
3.tgz
--2024-06-24 16:24:27--  https://downloads.apache.org/spark/spark-3.5.1/spark-3.5.1-bin-hadoop3.tgz
Resolving downloads.apache.org (downloads.apache.org)... 135.181.214.104, 88.99.208.237, 2a01:4f8:10a:39da::2,
...
Connecting to downloads.apache.org (downloads.apache.org)|135.181.214.104|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 400446614 (382M) [application/x-gzip]
Saving to: ‘spark-3.5.1-bin-hadoop3.tgz’

spark-3.5.1-bin-hadoop3.tgz 100%[==========================================>] 381.90M  10.1MB/s    in 41s

2024-06-24 16:25:09 (9.42 MB/s) - ‘spark-3.5.1-bin-hadoop3.tgz’ saved [400446614/400446614]

eweng909@instance-20240624-123745:~$
```

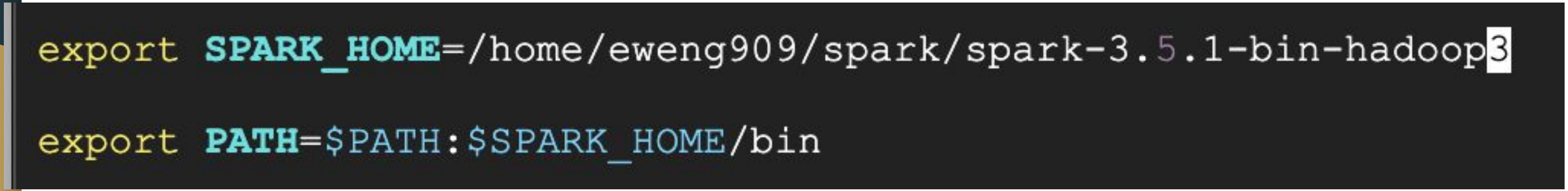
Unzip file

1. tar -xvf spark-3.5.1-bin-hadoop3.tgz



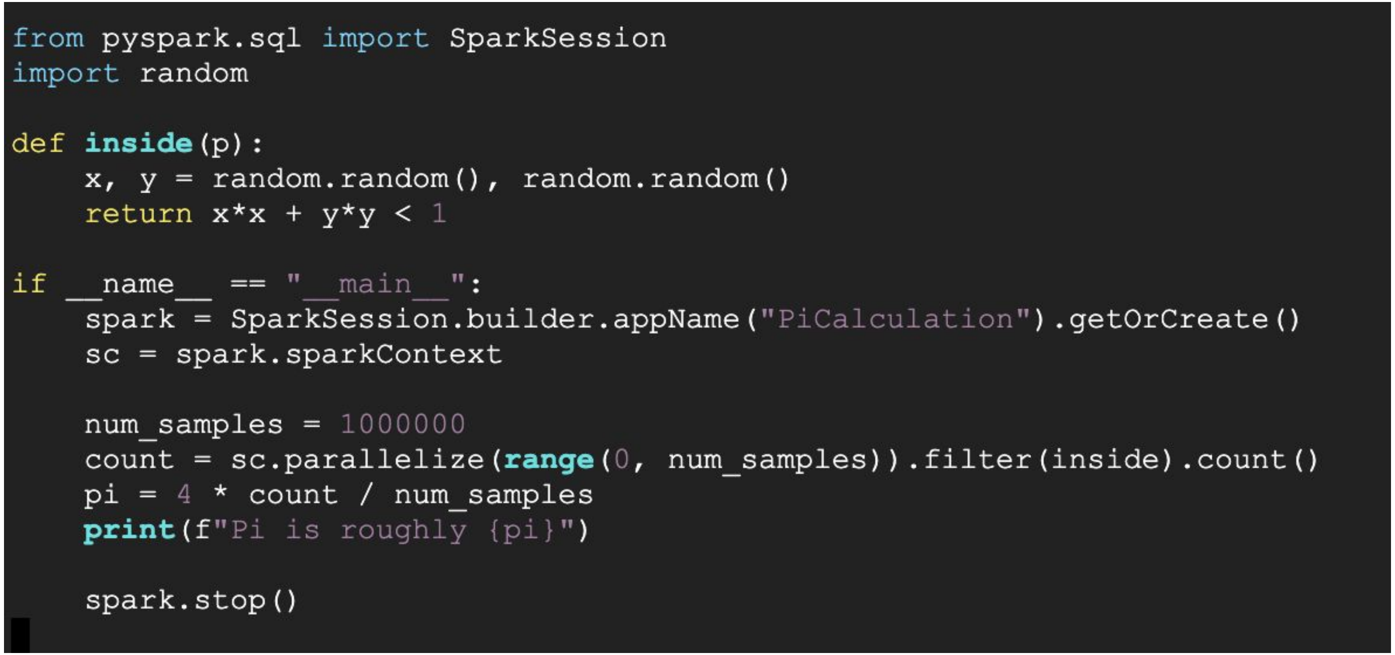
```
eweng909@instance-20240624-123745:~$ ls
PiCalculation  hadoop-3.3.5  hadoop-3.3.5.tar.gz  spark-3.5.1-bin-hadoop3.tgz
eweng909@instance-20240624-123745:~$ tar -xvf spark-3.5.1-bin-hadoop3.tgz
spark-3.5.1-bin-hadoop3/
spark-3.5.1-bin-hadoop3/sbin/
spark-3.5.1-bin-hadoop3/sbin/spark-config.sh
spark-3.5.1-bin-hadoop3/sbin/stop-slave.sh
spark-3.5.1-bin-hadoop3/sbin/stop-mesos-dispatcher.sh
spark-3.5.1-bin-hadoop3/sbin/start-workers.sh
spark-3.5.1-bin-hadoop3/sbin/start-slaves.sh
spark-3.5.1-bin-hadoop3/sbin/start-all.sh
spark-3.5.1-bin-hadoop3/sbin/stop-all.sh
spark-3.5.1-bin-hadoop3/sbin/workers.sh
spark-3.5.1-bin-hadoop3/sbin/start-mesos-dispatcher.sh
spark-3.5.1-bin-hadoop3/sbin/spark-daemon.sh
spark-3.5.1-bin-hadoop3/sbin/decommission-worker.sh
spark-3.5.1-bin-hadoop3/sbin/slaves.sh
spark-3.5.1-bin-hadoop3/sbin/stop-mesos-shuffle-service.sh
```

Add directory path into your bash file



```
export SPARK_HOME=/home/eweng909/spark/spark-3.5.1-bin-hadoop3

export PATH=$PATH:$SPARK_HOME/bin
```

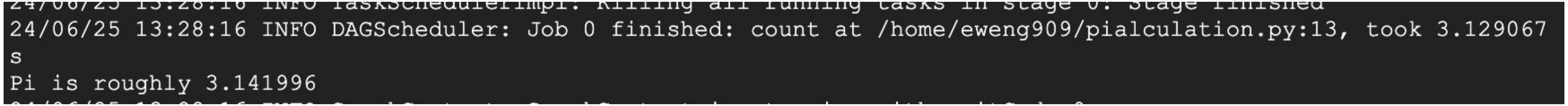
Create python file



```python
from pyspark.sql import SparkSession
import random

def inside(p):
    x, y = random.random(), random.random()
    return x*x + y*y < 1

if __name__ == "__main__":
    spark = SparkSession.builder.appName("PiCalculation").getOrCreate()
    sc = spark.sparkContext

    num_samples = 1000000
    count = sc.parallelize(range(0, num_samples)).filter(inside).count()
    pi = 4 * count / num_samples
    print(f"Pi is roughly {pi}")

    spark.stop()
```

Run the command for it to work

1. spark-submit picalculation.py
2. It should run the results.



```
24/06/25 13:28:16 INFO TaskSchedulerImpl: Killing all running tasks in stage 0: Stage finished
24/06/25 13:28:16 INFO DAGScheduler: Job 0 finished: count at /home/eweng909/pialculation.py:13, took 3.129067
s
Pi is roughly 3.141996
```