

Question 1:

Install Apache-Spark in your laptop/desktop and upload the screenshots. The expected screenshot should contain a terminal (on mac and linux systems)/command prompt with the welcome page.

Screenshot of creating a Spark Installation directory and un-tarring the installation file.

```
~/miniforge3/envs/ENV_SPARK/SparkInstallation/spark-3.2.1-bin-hadoop2.7 -- zsh
himanideshpande@Himanis-MacBook-Air SparkInstallation % ls -l
total 557864
-rw-r--r--@ 1 himanideshpande staff 272637746 Jan 30 18:11 spark-3.2.1-bin-hadoop2.7.tgz
himanideshpande@Himanis-MacBook-Air SparkInstallation % tar -xvf spark-3.2.1-bin-hadoop2.7.tgz
himanideshpande@Himanis-MacBook-Air SparkInstallation % ls -l
total 557864
drwxr-xr-x@ 17 himanideshpande staff 544 Jan 28 15:52 spark-3.2.1-bin-hadoop2.7
-rw-r--r--@ 1 himanideshpande staff 272637746 Jan 30 18:11 spark-3.2.1-bin-hadoop2.7.tgz
himanideshpande@Himanis-MacBook-Air SparkInstallation % cd spark-3.2.1-bin-hadoop2.7
himanideshpande@Himanis-MacBook-Air spark-3.2.1-bin-hadoop2.7 % ls -l
total 192
-rw-r--r--@ 1 himanideshpande staff 22878 Jan 28 15:52 LICENSE
-rw-r--r--@ 1 himanideshpande staff 57677 Jan 28 15:52 NOTICE
drwxr-xr-x@ 3 himanideshpande staff 96 Jan 28 15:52 R
-rw-r--r--@ 1 himanideshpande staff 4512 Jan 28 15:52 README.md
-rw-r--r--@ 1 himanideshpande staff 167 Jan 28 15:52 RELEASE
drwxr-xr-x@ 29 himanideshpande staff 928 Jan 28 15:52 bin
drwxr-xr-x@ 8 himanideshpande staff 256 Jan 28 15:52 conf
drwxr-xr-x@ 5 himanideshpande staff 160 Jan 28 15:52 data
drwxr-xr-x@ 4 himanideshpande staff 128 Jan 28 15:52 examples
drwxr-xr-x@ 269 himanideshpande staff 8608 Jan 28 15:52 jars
drwxr-xr-x@ 4 himanideshpande staff 128 Jan 28 15:52 kubernetes
drwxr-xr-x@ 68 himanideshpande staff 1920 Jan 28 15:52 licenses
drwxr-xr-x@ 15 himanideshpande staff 576 Jan 28 15:52 python
drwxr-xr-x@ 29 himanideshpande staff 928 Jan 28 15:52 sbin
drwxr-xr-x@ 3 himanideshpande staff 96 Jan 28 15:52 yarn
himanideshpande@Himanis-MacBook-Air spark-3.2.1-bin-hadoop2.7 %
```

Screenshot of Spark welcome page

```
~/miniforge3/envs/ENV_SPARK/SparkInstallation/spark-3.2.1-bin-hadoop2.7 -- java - python3
[ENV_SPARK] himanideshpande@Himanis-MacBook-Air spark-3.2.1-bin-hadoop2.7 % ./bin/pyspark
Python 3.8.12 | packaged by conda-forge | (default, Oct 12 2021, 21:25:58)
[Clang 11.1.0 ] on darwin
Type "help", "copyright", "credits" or "license" for more information.
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.spark.unsafe.Platform (file:/Users/himanideshpande/miniforge3/envs/ENV_SPARK/SparkInstallation/spark-3.2.1-bin-hadoop2.7/jars/spark-unsafe_2.12-3.2.1.jar)
to constructor java.nio.DirectByteBuffer(long,int)
WARNING: Please consider reporting this to the maintainers of org.apache.spark.unsafe.Platform
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
22/01/30 18:15:58 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Welcome to

  ____      _
 / ___|  __| | | |
 \___ \  | | | | | |
  ___) | | | | | | |
 |____|_|_|_|_|_|_|_|

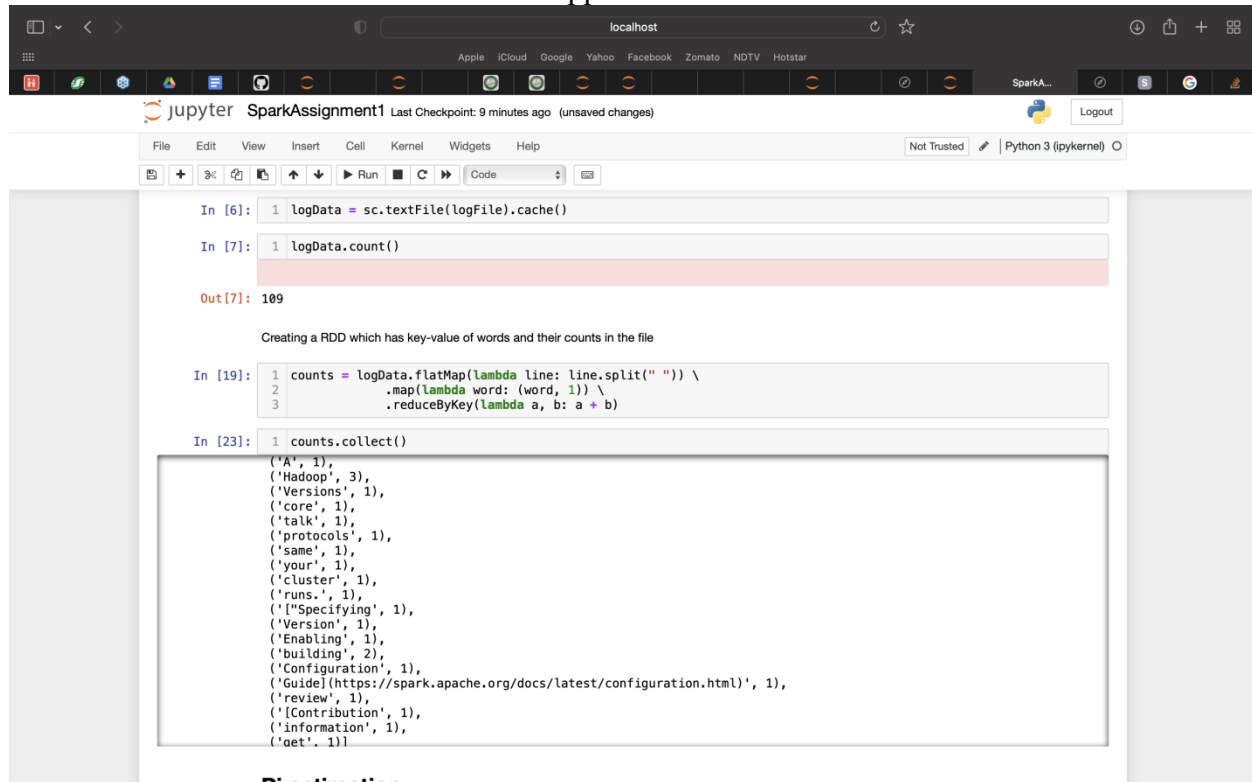
 version 3.2.1

Using Python version 3.8.12 (default, Oct 12 2021 21:25:58)
Spark context Web UI available at http://192.168.0.205:4040
Spark context available as 'sc' (master = local[*], app id = local-1643984551494).
SparkSession available as 'spark'.
>>>
```

Question 2:

Run 2 example spark codes and upload the resulting screenshots.

The first code I ran will count the number of words in a README.md file of Spark installation. I have also created an RDD of this same README.md file which stores the key as the word and the value as the number of times the word appears in this file.



```
In [6]: 1 logData = sc.textFile(logFile).cache()

In [7]: 1 logData.count()

Out[7]: 109

Creating a RDD which has key-value of words and their counts in the file

In [19]: 1 counts = logData.flatMap(lambda line: line.split(" ")) \
2         .map(lambda word: (word, 1)) \
3         .reduceByKey(lambda a, b: a + b)

In [23]: 1 counts.collect()

[('A', 1),
 ('Hadoop', 3),
 ('Versions', 1),
 ('core', 1),
 ('talk', 1),
 ('protocols', 1),
 ('same', 1),
 ('your', 1),
 ('cluster', 1),
 ('runs.', 1),
 ('["Specifying', 1),
 ('Version', 1),
 ('Enabling', 1),
 ('building', 2),
 ('Configuration', 1),
 ('Guide](https://spark.apache.org/docs/latest/configuration.html)', 1),
 ('review', 1),
 ('[Contribution', 1),
 ('information', 1),
 ('oet'. 1)]
```

The 2nd code is the Pi Estimation where the code estimates the value of π picking random points in the unit square ((0,0),(1,1)) and get an estimate of fraction $\pi/4$ of those which fall in the unit circle.

```
{
  'some', 1,
  ('your', 1),
  ('cluster', 1),
  ('runs.', 1),
  ('Specifying', 1),
  ('Version', 1),
  ('Enabling', 1),
  ('building', 2),
  ('Configuration', 1),
  ('Guide')(https://spark.apache.org/docs/latest/configuration.html)', 1),
  ('review', 1),
  ('Contribution', 1),
  ('information', 1),
  ('get', 1)]
```

Pi estimation

In [11]: 1 import random

In [24]: 1 NUM_SAMPLES = 1000000

In [25]: 1 def inside(p):
2 x, y = random.random(), random.random()
3 return x*x + y*y < 1
4
5 count = sc.parallelize(range(0, NUM_SAMPLES)) \
6 .filter(inside).count()
7 print("Pi is roughly %f" % (4.0 * count / NUM_SAMPLES))

Pi is roughly 3.140600

1 References:
2 <https://spark.apache.org/examples.html>

Using the Module for reference, I ran the code which creates a Spark Session and a dataframe using RDD API. Then we group by Names of people in the dataframe and take an aggregate of their age and show their average age.

```
-rw-r--r-- 1 himanideshpande staff 22070 Jan 20 15:52 LICENSE
-rw-r--r-- 1 himanideshpande staff 57677 Jan 20 15:52 NOTICE
drwxr-xr-x 3 himanideshpande staff 96 Jan 20 15:52 R
-rw-r--r-- 1 himanideshpande staff 4612 Jan 20 15:52 README.md
drwxr-xr-x 4 himanideshpande staff 160 Jan 20 15:52 RELEASE
drwxr-xr-x 29 himanideshpande staff 928 Jan 20 15:52 bin
drwxr-xr-x 8 himanideshpande staff 256 Jan 20 15:52 conf
drwxr-xr-x 5 himanideshpande staff 160 Jan 20 15:52 data
drwxr-xr-x 4 himanideshpande staff 128 Jan 20 15:52 examples
drwxr-xr-x 269 himanideshpande staff 8480 Jan 20 15:52 jars
drwxr-xr-x 4 himanideshpande staff 128 Jan 20 15:52 kubernetes
drwxr-xr-x 60 himanideshpande staff 1920 Jan 20 15:52 licenses
drwxr-xr-x 18 himanideshpande staff 876 Jan 20 15:52 python
drwxr-xr-x 29 himanideshpande staff 928 Jan 20 15:52 sbin
drwxr-xr-x 3 himanideshpande staff 96 Jan 20 15:52 yarn
himanideshpande@Himanis-MacBook-Air spark-3.2.1-bin-hadoop2.7 % pyspark
zsh: command not found: pyspark
himanideshpande@Himanis-MacBook-Air spark-3.2.1-bin-hadoop2.7 % ./bin/pyspark
Python 3.8.9 (default, Aug 3 2021, 19:21:54)
[Clang 13.0.0 (clang-1300.0.29.3)] on darwin
Type "help", "copyright", "credits" or "license()" for more information.
22/01/30 19:56:20 WARN Utils: Your hostname, Himanis-MacBook-Air.local resolves to a loopback address: 127.0.0.1; using 192.168.0.205 instead (on interface en0)
22/01/30 19:56:20 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.spark.unsafe.Platform (file:/Users/himanideshpande/miniforge3/envs/ENV_SPARK/SparkInstallation/spark-3.2.1-bin-hadoop2.7/jars/spark-unsafe-2.12-3.2.1.jar)
to constructor java.nio.DirectByteBuffer(long,int)
WARNING: Please consider reporting this to the maintainers of org.apache.spark.unsafe.Platform
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
22/01/30 19:56:22 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Welcome to

  ____      _
 / ___|  __| | | |
 \___ \  | | | | | |
  ___) | | | | | |
 |____|_|_|_|_|_|_|

version 3.2.1

Using Python version 3.8.9 (default, Aug 3 2021 19:21:54)
Spark context Web UI available at http://192.168.0.205:4040
Spark context available as 'sc' (master = local[*], app id = local-1643690683180).
SparkSession available as 'spark'.
>>> from pyspark.sql import SparkSession
>>> from pyspark.sql.functions import avg
>>> spark = SparkSession
... .builder
... .appName("AuthorsAges")
... .getOrCreate()
>>> data_df = spark.createDataFrame([("Brooke", 20), ("Denny", 31), ("Jules", 30), ("TD", 35), ("Brooke", 25)], ["name", "age"])
>>> avg_df = data_df.groupBy("name").agg(avg("age"))
>>> avg_df.show()
+-----+
| name | avg(age) |
+-----+
| Brooke | 22.5 |
| Denny | 31.0 |
| Jules | 30.0 |
| TD | 35.0 |
+-----+
```

References:

<https://spark.apache.org/examples.html>

Modules