Q1: Unable to invoke spark session and not able the execute the boiler plate code given

Ans:

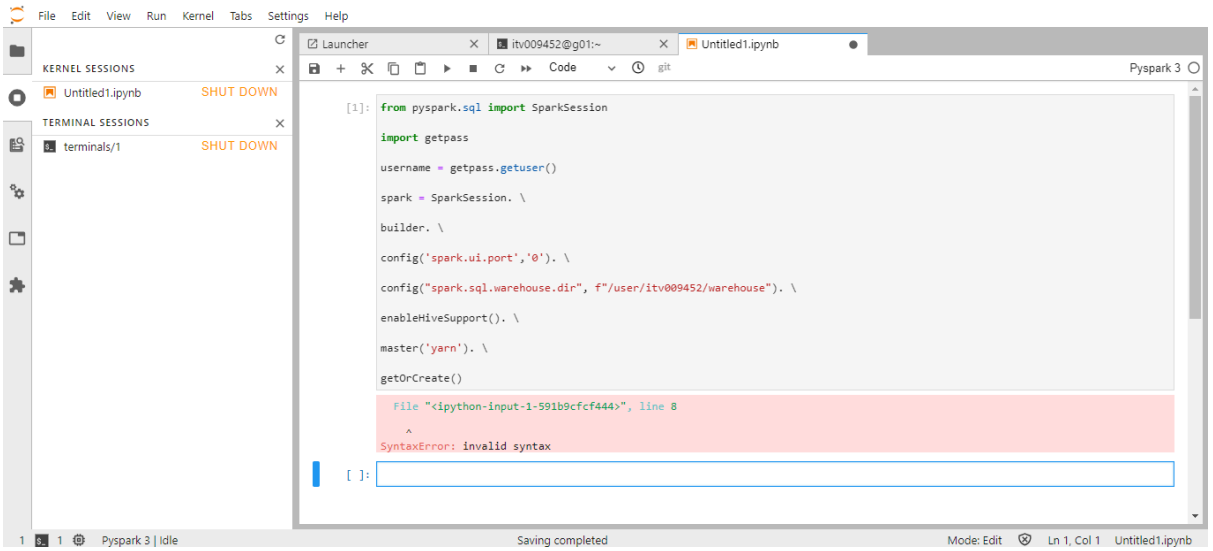
Please use the below boiler plate code.

Note: In below code please replace the {username} with your id and try. Like if your username is itv006753 then configuration becomes config("spark.sql.warehouse.dir", "/user/itv006753/warehouse").

```
from pyspark.sql import SparkSession

spark = SparkSession. \
builder. \
config('spark.ui.port','0'). \
config("spark.sql.warehouse.dir", "/user/{username}/warehouse"). \
enableHiveSupport(). \
master('yarn'). \
getOrCreate()
```

Q2. Boilerplate code giving error :



Ans:

Please remove spaces between the lines.

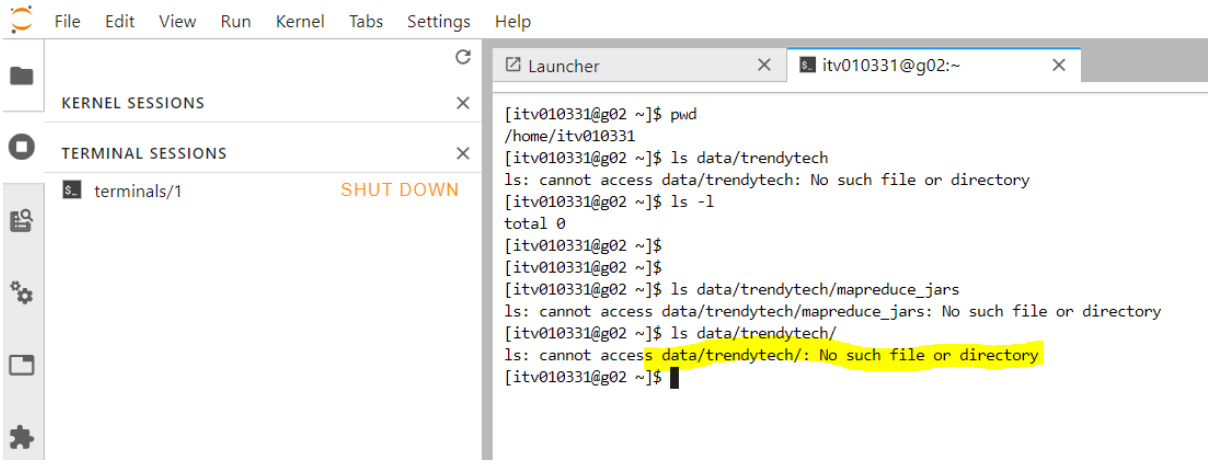
And use the below boiler plate code.

Note: In below code please replace the {username} with your id and try. Like if your username is itv006753 then configuration becomes config("spark.sql.warehouse.dir", "/user/itv006753/warehouse").

```
from pyspark.sql import SparkSession

spark = SparkSession. \
builder. \
config('spark.ui.port','0'). \
config("spark.sql.warehouse.dir", "/user/{username}/warehouse"). \
enableHiveSupport(). \
master('yarn'). \
getOrCreate()
```

Q3. Unable to find the mapReduce related jar files in the data/trendytech directory in my gateway node. Could you please help me to locate the mapReduce jar files? Please find attached the screenshot of the command which i have used to view the files under data/trendytech



Ans:

The jar files are in the location : /data/trendytech/mapreduce_jars

i.e. it starts with /data.... and not like data/...

Use the below command to see the list of jars:

ls /data/trendytech/mapreduce_jars

Q4. How to run the program jar file? I did the same process mentioned in the session but jar files were not executed.

Ans:

Use the following command for executing the jar files:

hadoop jar <path_of_jar> <input_path> <output_path>

Ex: If path of jar is “/data/trendytech/mapreduce_jars/mapreduce_prog_0_reducer.jar”, path of input file is “/user/itv006277/data/inputfile.txt” and you want to store the output to the directory “/user/itv006277/data/output” then command will be

hadoop jar /data/trendytech/mapreduce_jars/mapreduce_prog_0_reducer.jar /user/itv006277/data/inputfile.txt /user/itv006277/data/output

```
[itv006277@g01 ~]$ ls /data/trendytech/mapreduce_jars
mapreduce_prog_0_reducer.jar  mapreduce_prog_2_reducer.jar  mapreduce_prog_combiner.jar  mapreduce_prog_cpartitioner.jar  mapreduce_prog.jar
[itv006277@g01 ~]$ hadoop jar /data/trendytech/mapreduce_jars/mapreduce_prog_0_reducer.jar /user/itv006277/data/inputfile.txt /user/itv006277/data/output
2023-12-07 23:47:53,759 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at m02.itversity.com/172.16.1.104:8032
2023-12-07 23:47:53,994 INFO client.AHSProxy: Connecting to Application History server at m01.itversity.com/172.16.1.103:10200
2023-12-07 23:47:54,220 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/itv006277/.staging/job_1692342312988_8
2158
2023-12-07 23:48:23,445 INFO input.FileInputFormat: Total input files to process : 1
2023-12-07 23:48:23,721 INFO mapreduce.JobSubmitter: number of splits:1
2023-12-07 23:48:23,975 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1692342312988_82158
2023-12-07 23:48:23,975 INFO mapreduce.JobSubmitter: Executing with tokens: []
2023-12-07 23:48:24,233 INFO conf.Configuration: resource-types.xml not found
2023-12-07 23:48:24,233 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2023-12-07 23:48:24,275 INFO impl.YarnClientImpl: Submitted application application_1692342312988_82158
2023-12-07 23:48:24,294 INFO mapreduce.Job: The url to track the job: http://m02.itversity.com:19088/proxy/application_1692342312988_82158/
2023-12-07 23:48:24,295 INFO mapreduce.Job: Running job: job_1692342312988_82158
2023-12-07 23:48:29,381 INFO mapreduce.Job: Job job_1692342312988_82158 running in uber mode : false
2023-12-07 23:48:29,384 INFO mapreduce.Job:  map 0% reduce 0%
2023-12-07 23:48:33,558 INFO mapreduce.Job:  map 100% reduce 0%
2023-12-07 23:48:33,575 INFO mapreduce.Job: Job job_1692342312988_82158 completed successfully
2023-12-07 23:48:33,681 INFO mapreduce.Job: Counters: 33
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=266166
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=275
                HDFS: Number of bytes written=203
                HDFS: Number of read operations=7
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=1
```

Q5. I am getting an error NameNode is in safe mode, how to resolve this.

: org.apache.hadoop.hdfs.server.namenode.SafeModeException: Cannot create directory /user/itv010248/.sparkStaging/application_1707552082651_0011. Name node is in safe mode.

Caused by: org.apache.hadoop.ipc.RemoteException(org.apache.hadoop.hdfs.server.namenode.SafeModeException): Cannot create directory /user/itv010248/.sparkStaging/application_1707552082651_0049. Name node is in safe mode.

Ans:

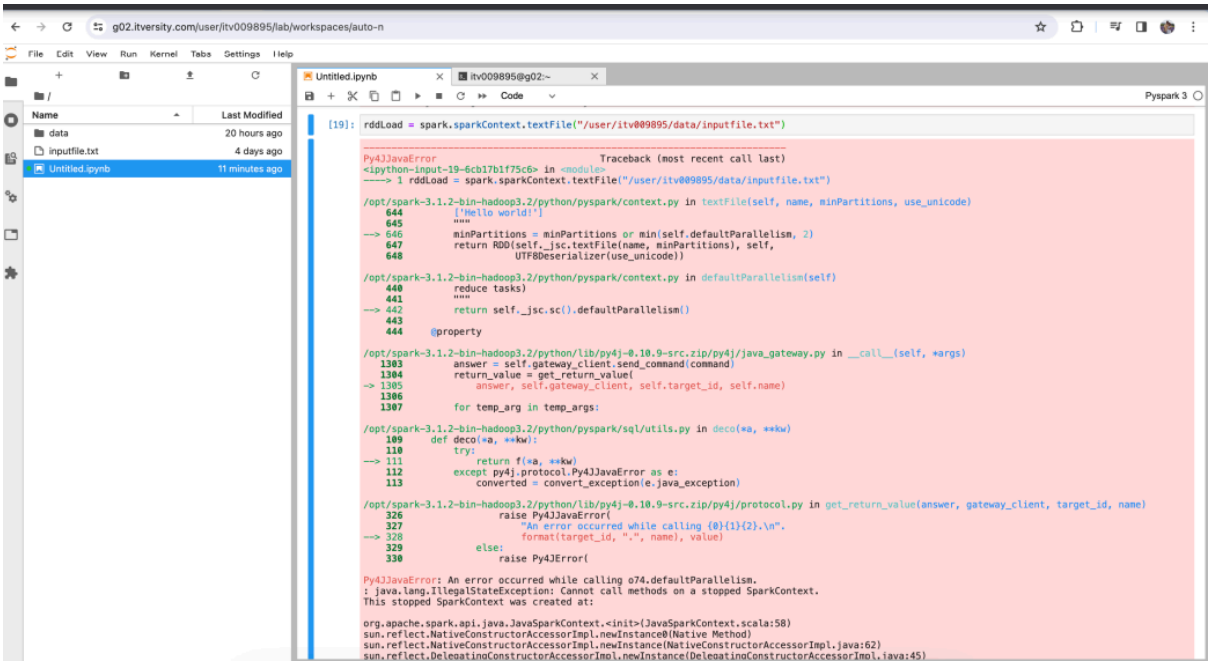
To resolve the "NameNode is in safe mode" error, you can take the following steps:

Use the command “**hdfs dfsadmin -safemode get**” command to check the current status of safe mode.

If the NameNode is in safe mode, you can take it out of safe mode using the command “**hdfs dfsadmin -safemode leave**”

This command will transition the NameNode to the active state, allowing you to create directories and perform other write operations.

Q6. Trying to load the file using the spark.sparkContext.textFile("/user/itv009895/data/inputfile.txt")
However the context is already created when I check using the spark.sparkContext command



Ans:

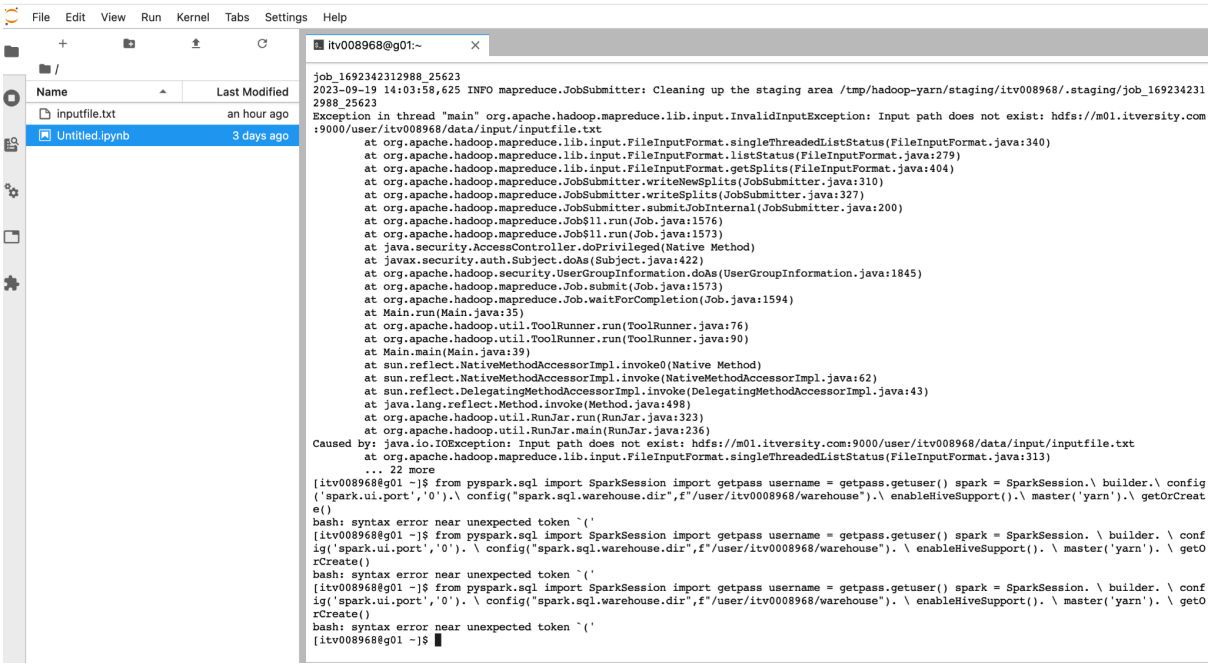
Follow the steps below.

1. Run the command spark.stop()

2. Now restart the kernel.

3. Using Boilerplate code, create the spark session and try to run all the codes.

Q7: Query regarding boiler plate code in Week-3



Ans:

The boiler plate code is used for session creation in a notebook. But if you want to execute the pyspark in terminal then simply use the command : pyspark

Please refer to the attached screenshots for more clarification.

```
[itv006277@g01 ~]$ pyspark
Multiple versions of Spark are installed but SPARK_MAJOR_VERSION is not set
Spark2 will be picked by default
Python 2.7.5 (default, Jun 20 2023, 11:36:40)
[GCC 4.8.5 20150623 (Red Hat 4.8.5-44)] on linux2
Type "help", "copyright", "credits" or "license" for more information.
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/opt/spark-2.4.7-bin-hadoop2.7/jars/slf4j-log4j12-1.7.16.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/opt/hadoop-3.3.0/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
Setting default log level to "WARN".
```

Q8:I am getting the below error in Week 3 Assignment, how can I resolve this?

Untitled2.ipynb | itv009051@g01:~

Code | git

[1]:
```
from pyspark.sql import SparkSession
import getpass
username=getpass.getuser()
spark=SparkSession.\
builder.\
config('spark.ui.port','0').\
config("spark.sql.warehouse.dir",f"/user/itv000173/warehouse").\
enableHiveSupport().\
master('yarn').\
getOrCreate()
```

[3]:
```
spark
```

[3]: **SparkSession - hive**

SparkContext

Spark UI

Version	v3.1.2
Master	yarn
AppName	pyspark-shell

[4]:
```
rdd1 = spark.sparkContext.textFile("/user/itv009051/data/input/linkedin_views.csv")
```

Untitled2.ipynb | itv009051@g01:~

Code | git

[17]:
```
rdd1.collect()
```

[17]:
```
['1,Manasa,Sumit',
 '2,Deepa,Sumit',
 '3,Sumit,Manasa',
 '4,Deepa,Manasa',
 '5,Manasa,Deepa',
 '6,Shilpy,Manasa',
 '']
```

[18]:
```
rdd2 = rdd1.flatMap(lambda x: x.split(",")[2])
```

[19]:
```
rdd2.collect()
```

```
---------------------------------------------------------------------------
Py4JJavaError                             Traceback (most recent call last)
<ipython-input-19-83517eaf6d43> in <module>
----> 1 rdd2.collect()

/opt/spark-3.1.2-bin-hadoop3.2/python/pyspark/rdd.py in collect(self)
    947         """
    948         with SCCallSiteSync(self.context) as css:
--> 949             sock_info = self.ctx._jvm.PythonRDD.collectAndServe(self._jrdd.rdd())
    950         return list(_load_from_socket(sock_info, self._jrdd_deserializer))
    951

/opt/spark-3.1.2-bin-hadoop3.2/python/lib/py4j-0.10.9-src.zip/py4j/java_gateway.py in __call__(self, *args)
   1303         answer = self.gateway_client.send_command(command)
   1304         return_value = get_return_value(
-> 1305             answer, self.gateway_client, self.target_id, self.name)
   1306
   1307         for temp_arg in temp_args:

/opt/spark-3.1.2-bin-hadoop3.2/python/pyspark/sql/utils.py in deco(*a, **kw)
    109     def deco(*a, **kw):
    110         try:
```

```
Untitled2.ipynb                    itv009051@g01:~
Code                git                                                                    Pyspark

    1307         for temp_arg in temp_args:

/opt/spark-3.1.2-bin-hadoop3.2/python/pyspark/sql/utils.py in deco(*a, **kw)
    109     def deco(*a, **kw):
    110         try:
--> 111             return f(*a, **kw)
    112         except py4j.protocol.Py4JJavaError as e:
    113             converted = convert_exception(e.java_exception)

/opt/spark-3.1.2-bin-hadoop3.2/python/lib/py4j-0.10.9-src.zip/py4j/protocol.py in get_return_value(answer, gateway_client, target_id, name)
    326                 raise Py4JJavaError(
    327                     "An error occurred while calling {0}{1}{2}.\n".
--> 328                     format(target_id, ".", name), value)
    329             else:
    330                 raise Py4JError(

Py4JJavaError: An error occurred while calling z:org.apache.spark.api.python.PythonRDD.collectAndServe.
: org.apache.spark.SparkException: Job aborted due to stage failure: Task 1 in stage 10.0 failed 4 times, most recent failure: Lost task 1.3
in stage 10.0 (TID 33) (w02.itversity.com executor 2): org.apache.spark.api.python.PythonException: Traceback (most recent call last):
  File "/opt/spark-3.1.2-bin-hadoop3.2/python/pyspark/worker.py", line 604, in main
    process()
  File "/opt/spark-3.1.2-bin-hadoop3.2/python/pyspark/worker.py", line 596, in process
    serializer.dump_stream(out_iter, outfile)
  File "/opt/spark-3.1.2-bin-hadoop3.2/python/pyspark/serializers.py", line 259, in dump_stream
    vs = list(itertools.islice(iterator, batch))
  File "/opt/spark-3.1.2-bin-hadoop3.2/python/pyspark/util.py", line 73, in wrapper
    return f(*args, **kwargs)
  File "<ipython-input-18-c31632bbb291>", line 1, in <lambda>
IndexError: list index out of range

        at org.apache.spark.api.python.BasePythonRunner$ReaderIterator.handlePythonException(PythonRunner.scala:517)
        at org.apache.spark.api.python.PythonRunner$$anon$3.read(PythonRunner.scala:652)
        at org.apache.spark.api.python.PythonRunner$$anon$3.read(PythonRunner.scala:635)
        at org.apache.spark.api.python.BasePythonRunner$ReaderIterator.hasNext(PythonRunner.scala:470)
        at org.apache.spark.InterruptibleIterator.hasNext(InterruptibleIterator.scala:37)
        at scala.collection.Iterator.foreach(Iterator.scala:941)
        at scala.collection.Iterator.foreach$(Iterator.scala:941)
        at org.apache.spark.InterruptibleIterator.foreach(InterruptibleIterator.scala:28)
        at scala.collection.generic.Growable.$plus$plus$eq(Growable.scala:62)
```

Ans:

Please remove the empty line in linkedin_views.csv file which is at the bottom of file. Or create a new linkedin_views.csv file and make sure there are no empty lines at the end.

Q9: Need answer for the correct workflow asked in week 3-ques12 quiz question.

QUESTION 12 OF 22

Which of the following is the right workflow?

Choose only ONE best answer.

A	Mapper -> Combiner -> Partitioner -> Sort -> Shuffle -> Reducer
B	Mapper -> Partitioner -> Combiner -> Sort -> Shuffle -> Reducer
C	Mapper -> Combiner -> Partitioner -> Shuffle -> Sort -> Reducer
D	Mapper -> Partitioner -> Combiner -> Shuffle -> Reducer -> Sort

Ans:

The typical flow of data and operations within a MapReduce job is as follows:
Mapper -> Combiner -> Partitioner -> Shuffle -> Sort -> Reducer

Mapper:
>> In the MapReduce process, data is divided into smaller chunks, and each chunk is processed by a separate Mapper task.
>> The Mapper's main function is to apply a user-defined transformation to the input data and emit intermediate key-value pairs.

Combiner:
>> The Combiner is an optional step that occurs on the Mapper side.
>> It's a mini-reducer that performs a local aggregation of the intermediate key-value pairs produced by the Mapper.
>> The purpose of the Combiner is to reduce the amount of data transferred during the shuffle phase by aggregating values with the same key within the same Mapper before sending them to the Reducers.

Partitioner:
>> After the Mappers have emitted intermediate key-value pairs, they are grouped by key and sent to Reducers for further processing.
>> The Practitioner's role is to determine which Reducer will receive which key-value pairs based on the key's hash value or some other logic.

Shuffle:
>> The Shuffle phase is a core part of MapReduce and involves the movement of data from the Mappers to the appropriate Reducers based on the keys.
>> This involves network communication and can be a resource-intensive step. The goal is to ensure that all key-value pairs with the same key end up at the same Reducer.

Sort:
Within each Reducer, the key-value pairs are sorted based on their keys.
Sorting the data allows Reducers to efficiently process and aggregate values for the same key.

Reducer:
>> Each Reducer receives a subset of the intermediate key-value pairs that share the same key.
>> The Reducer's main function is to apply a user-defined reduction operation to the values associated with each key, producing the final output of the MapReduce job.
The reduced results from all Reducers are the final results of the MapReduce job.

Q10: Does the combiner and partitioner happen in map nodes or reducer nodes?

Ans: Combiner is always a part of the Mapper phase whereas partition is an intermediate phase that takes place after the map phase and before the Reduce phase.