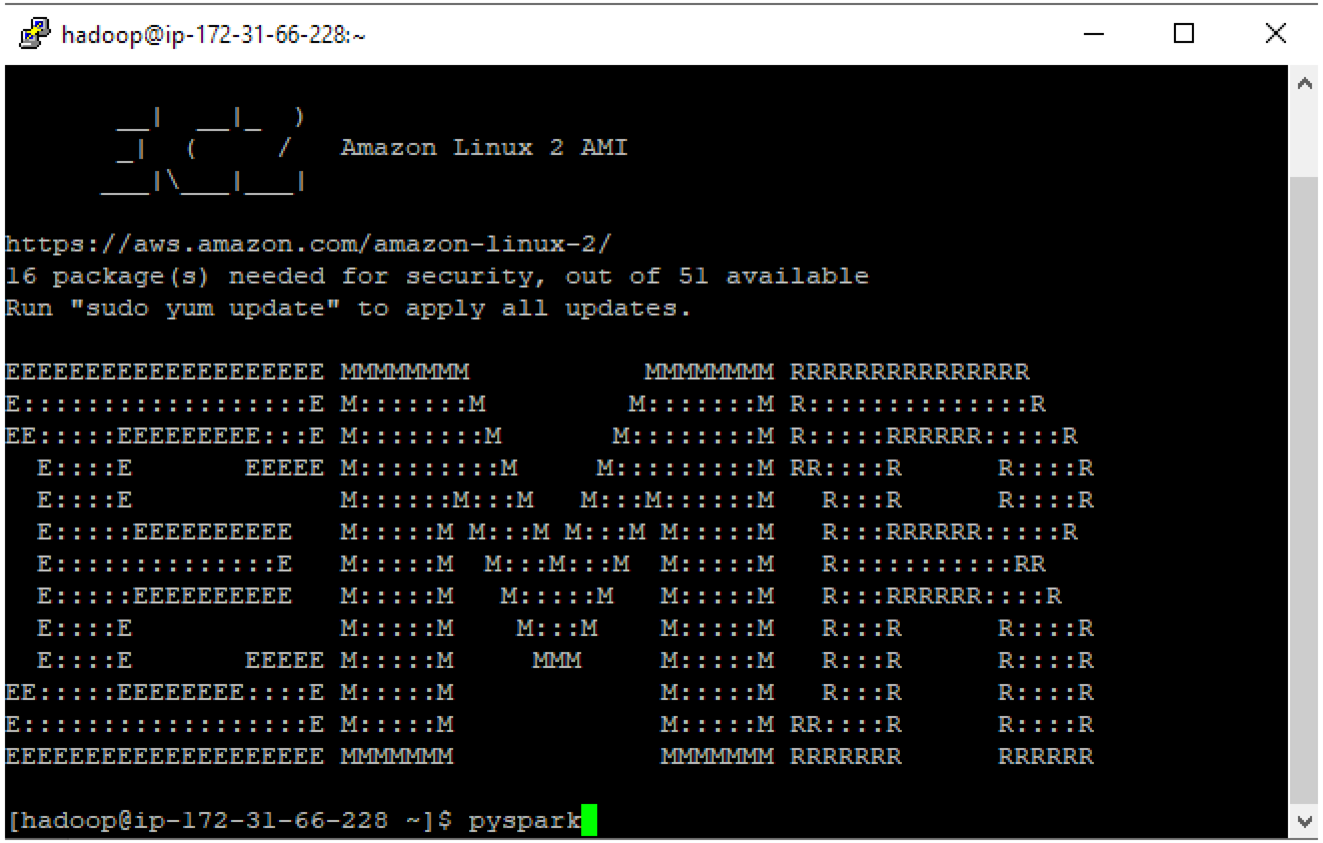
Spark Hands-On

Dr. Villanes

Once you see this screen.. Type pyspark



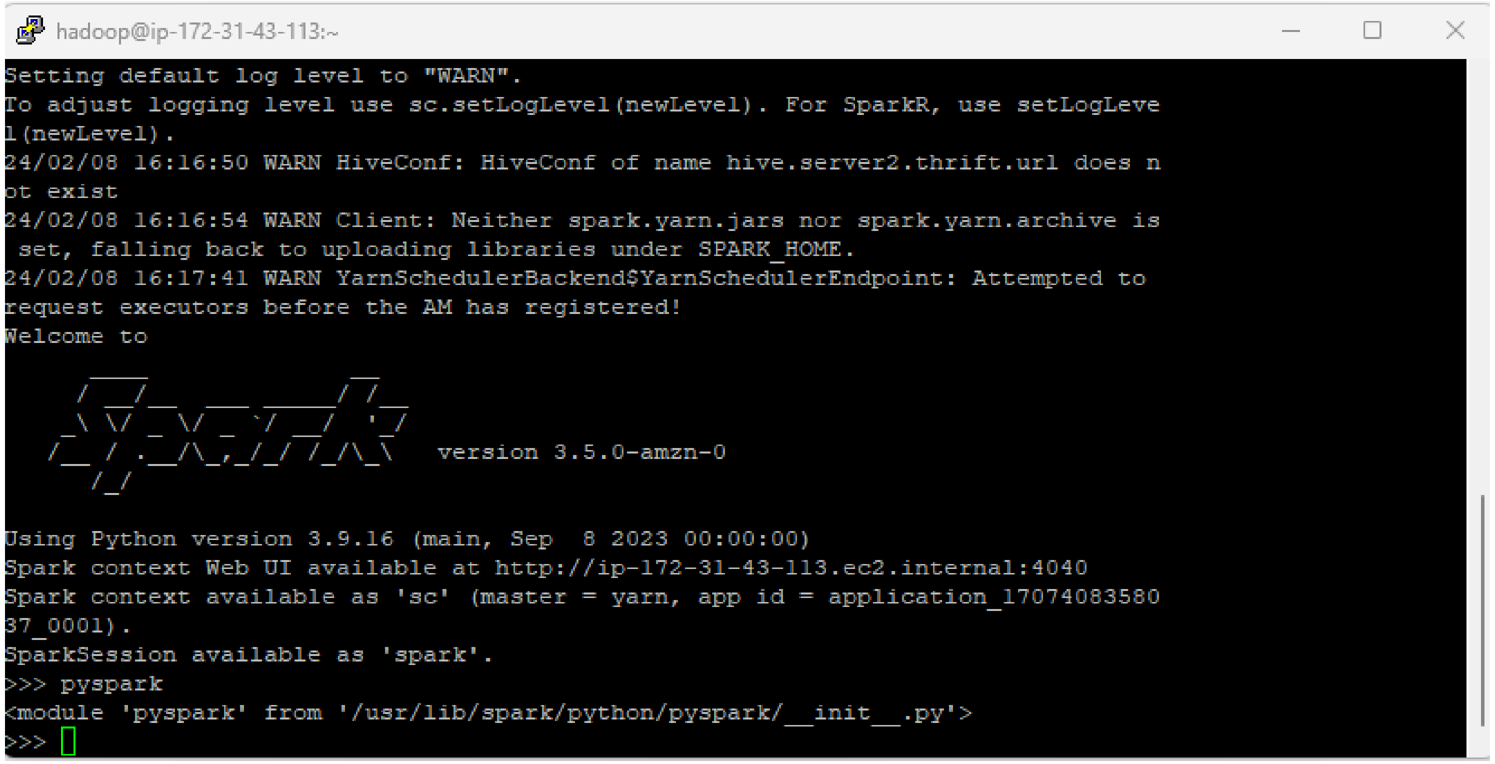
```
       __|  __|_  )
       _|  (     /   Amazon Linux 2 AMI
      ___|\___|___|

https://aws.amazon.com/amazon-linux-2/
16 package(s) needed for security, out of 51 available
Run "sudo yum update" to apply all updates.

EEEEEEEEEEEEEEEEEEEE MMMMMMMM           MMMMMMMM RRRRRRRRRRRRRRR
E::::::::::::::::::E M:::::::M         M:::::::M R::::::::::::::R
EE:::::EEEEEEEEE:::E M::::::::M       M::::::::M R:::::RRRRRR:::::R
  E::::E       EEEEE M:::::::::M     M:::::::::M RR::::R      R::::R
  E::::E             M::::::M:::M   M:::M::::::M   R:::R      R::::R
  E:::::EEEEEEEEEE   M:::::M M:::M M:::M M:::::M   R:::RRRRRR:::::R
  E::::::::::::::E   M:::::M  M:::M:::M  M:::::M   R:::::::::::RR
  E:::::EEEEEEEEEE   M:::::M   M:::::M   M:::::M   R:::RRRRRR::::R
  E::::E             M:::::M    M:::M    M:::::M   R:::R      R::::R
  E::::E       EEEEE M:::::M     MMM     M:::::M   R:::R      R::::R
EE:::::EEEEEEEE::::E M:::::M             M:::::M   R:::R      R::::R
E::::::::::::::::::E M:::::M             M:::::M RR::::R      R::::R
EEEEEEEEEEEEEEEEEEEE MMMMMMM             MMMMMMM RRRRRRR      RRRRRR

[hadoop@ip-172-31-66-228 ~]$ pyspark
```

You should now see something like this…

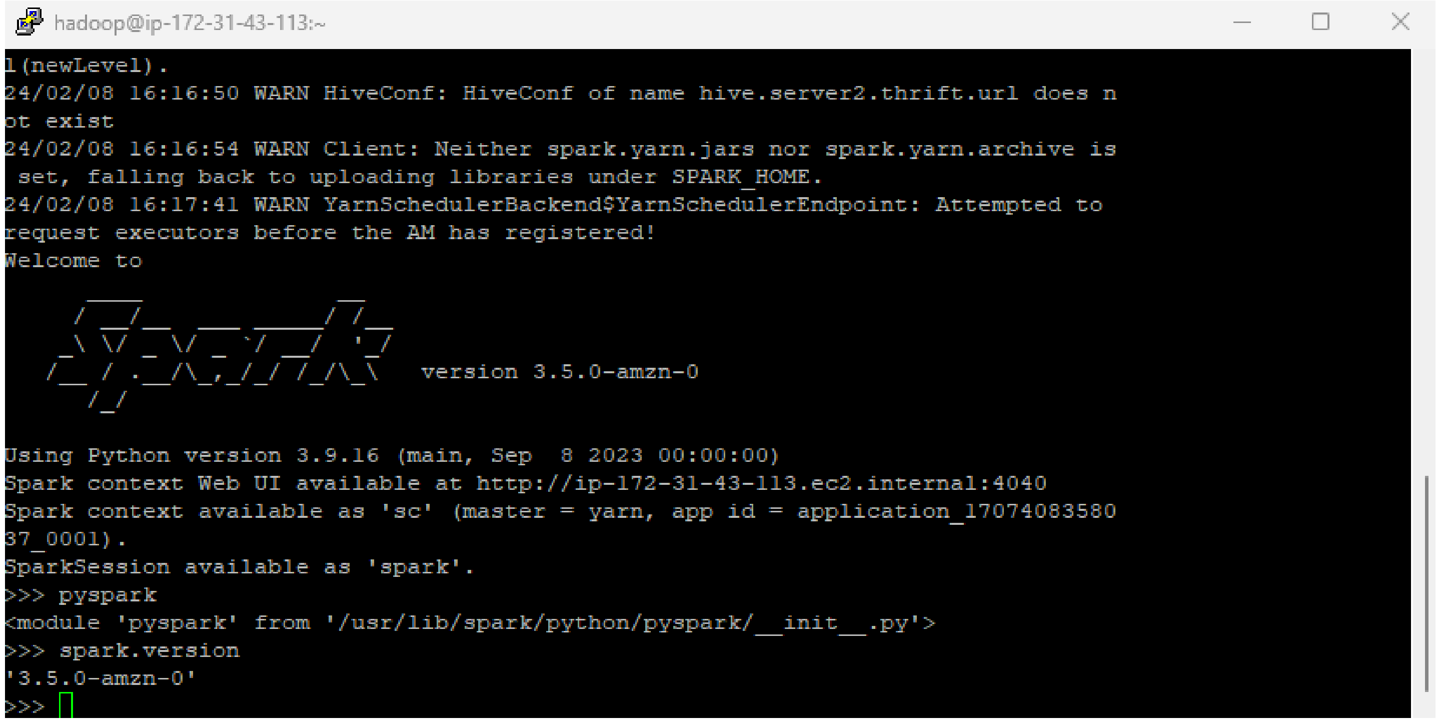


```
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLeve
l(newLevel).
24/02/08 16:16:50 WARN HiveConf: HiveConf of name hive.server2.thrift.url does n
ot exist
24/02/08 16:16:54 WARN Client: Neither spark.yarn.jars nor spark.yarn.archive is
 set, falling back to uploading libraries under SPARK_HOME.
24/02/08 16:17:41 WARN YarnSchedulerBackend$YarnSchedulerEndpoint: Attempted to
request executors before the AM has registered!
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 3.5.0-amzn-0
      /_/

Using Python version 3.9.16 (main, Sep  8 2023 00:00:00)
Spark context Web UI available at http://ip-172-31-43-113.ec2.internal:4040
Spark context available as 'sc' (master = yarn, app id = application_17074083580
37_0001).
SparkSession available as 'spark'.
>>> pyspark
<module 'pyspark' from '/usr/lib/spark/python/pyspark/__init__.py'>
>>> 
```

Spark Version

>>> spark.version

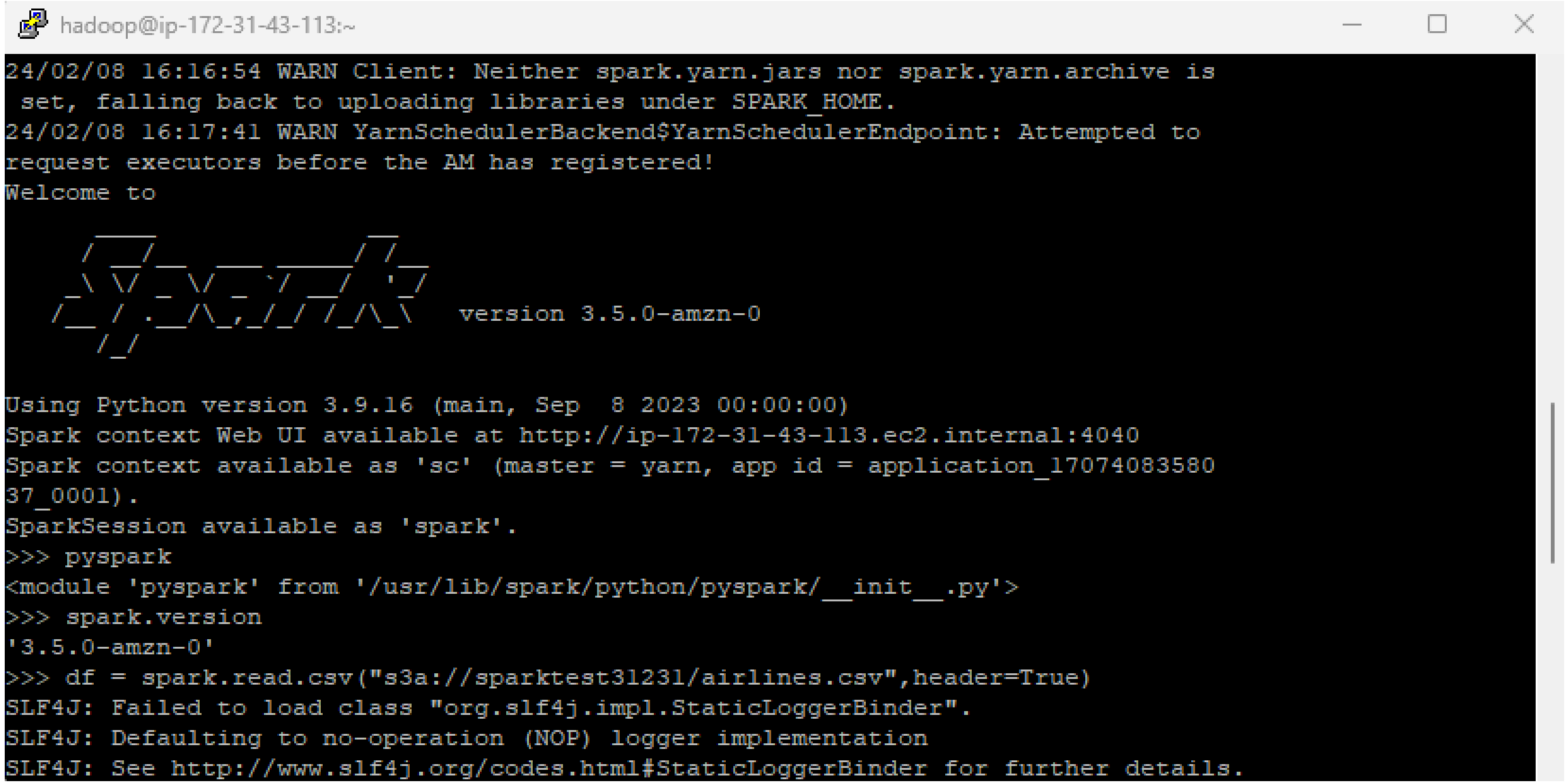


```
l(newLevel).
24/02/08 16:16:50 WARN HiveConf: HiveConf of name hive.server2.thrift.url does n
ot exist
24/02/08 16:16:54 WARN Client: Neither spark.yarn.jars nor spark.yarn.archive is
 set, falling back to uploading libraries under SPARK_HOME.
24/02/08 16:17:41 WARN YarnSchedulerBackend$YarnSchedulerEndpoint: Attempted to
request executors before the AM has registered!
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 3.5.0-amzn-0
      /_/

Using Python version 3.9.16 (main, Sep  8 2023 00:00:00)
Spark context Web UI available at http://ip-172-31-43-113.ec2.internal:4040
Spark context available as 'sc' (master = yarn, app id = application_17074083580
37_0001).
SparkSession available as 'spark'.
>>> pyspark
<module 'pyspark' from '/usr/lib/spark/python/pyspark/__init__.py'>
>>> spark.version
'3.5.0-amzn-0'
>>>
```

Creating Spark Dataframe

>>> df = spark.read.csv("s3a://bucket_name/airlines.csv",header=True)

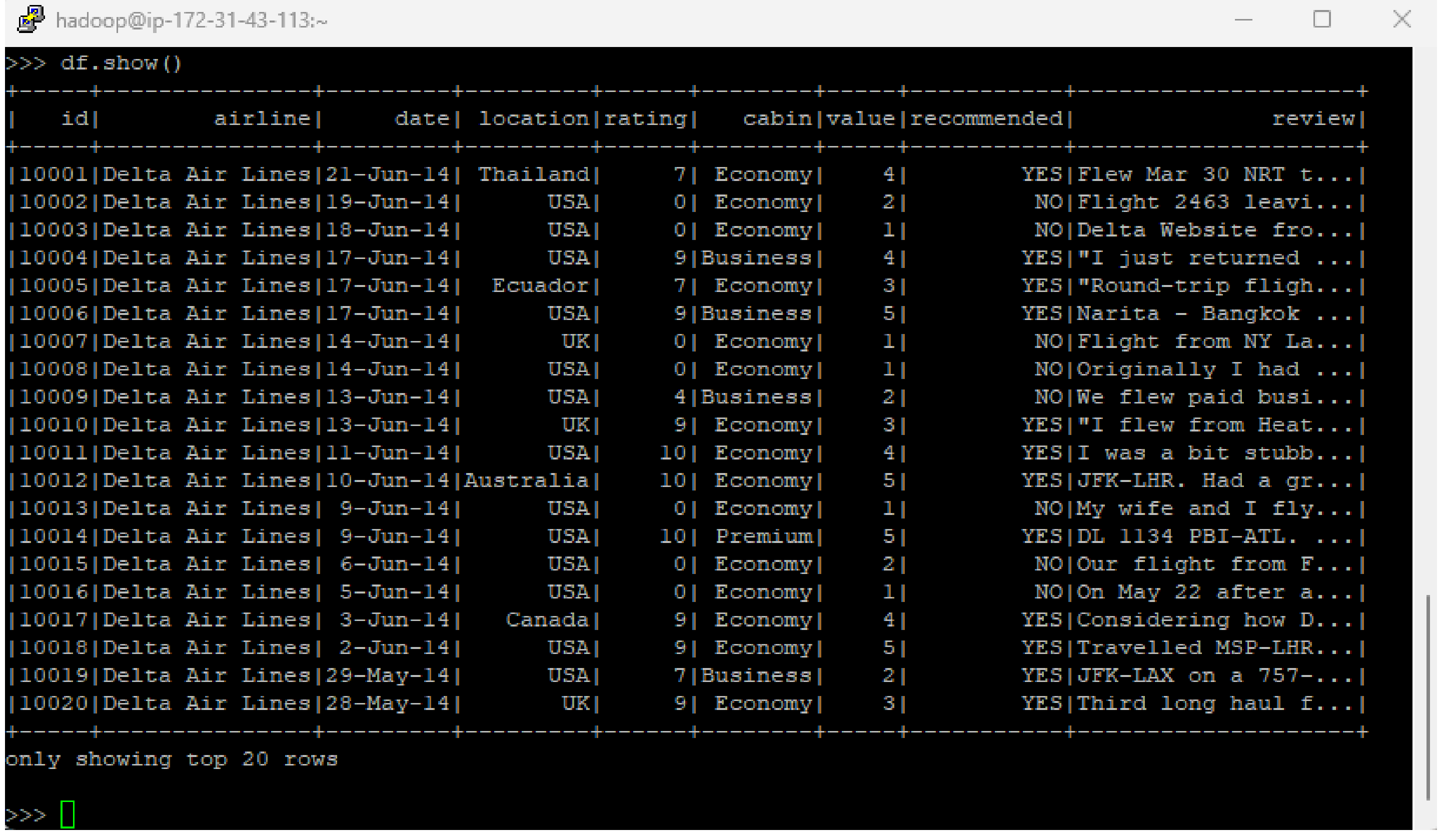


```
24/02/08 16:16:54 WARN Client: Neither spark.yarn.jars nor spark.yarn.archive is
 set, falling back to uploading libraries under SPARK_HOME.
24/02/08 16:17:41 WARN YarnSchedulerBackend$YarnSchedulerEndpoint: Attempted to
request executors before the AM has registered!
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 3.5.0-amzn-0
      /_/

Using Python version 3.9.16 (main, Sep  8 2023 00:00:00)
Spark context Web UI available at http://ip-172-31-43-113.ec2.internal:4040
Spark context available as 'sc' (master = yarn, app id = application_17074083580
37_0001).
SparkSession available as 'spark'.
>>> pyspark
<module 'pyspark' from '/usr/lib/spark/python/pyspark/__init__.py'>
>>> spark.version
'3.5.0-amzn-0'
>>> df = spark.read.csv("s3a://sparktest31231/airlines.csv",header=True)
SLF4J: Failed to load class "org.slf4j.impl.StaticLoggerBinder".
SLF4J: Defaulting to no-operation (NOP) logger implementation
SLF4J: See http://www.slf4j.org/codes.html#StaticLoggerBinder for further details.
```

View the Dataframe

>>> df.show()

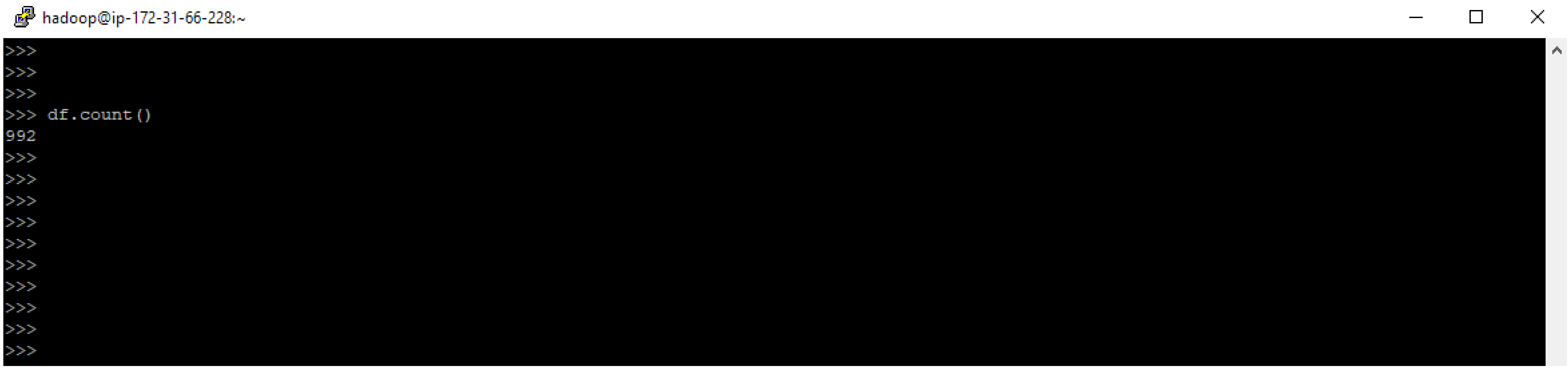


```
>>> df.show()
+-----+---------------+---------+---------+------+--------+-----+-----------+--------------------+
|   id|        airline|     date| location|rating|   cabin|value|recommended|              review|
+-----+---------------+---------+---------+------+--------+-----+-----------+--------------------+
|10001|Delta Air Lines|21-Jun-14| Thailand|     7| Economy|    4|        YES|Flew Mar 30 NRT t...|
|10002|Delta Air Lines|19-Jun-14|      USA|     0| Economy|    2|         NO|Flight 2463 leavi...|
|10003|Delta Air Lines|18-Jun-14|      USA|     0| Economy|    1|         NO|Delta Website fro...|
|10004|Delta Air Lines|17-Jun-14|      USA|     9|Business|    4|        YES|"I just returned ...|
|10005|Delta Air Lines|17-Jun-14|  Ecuador|     7| Economy|    3|        YES|"Round-trip fligh...|
|10006|Delta Air Lines|17-Jun-14|      USA|     9|Business|    5|        YES|Narita - Bangkok ...|
|10007|Delta Air Lines|14-Jun-14|       UK|     0| Economy|    1|         NO|Flight from NY La...|
|10008|Delta Air Lines|14-Jun-14|      USA|     0| Economy|    1|         NO|Originally I had ...|
|10009|Delta Air Lines|13-Jun-14|      USA|     4|Business|    2|         NO|We flew paid busi...|
|10010|Delta Air Lines|13-Jun-14|       UK|     9| Economy|    3|        YES|"I flew from Heat...|
|10011|Delta Air Lines|11-Jun-14|      USA|    10| Economy|    4|        YES|I was a bit stubb...|
|10012|Delta Air Lines|10-Jun-14|Australia|    10| Economy|    5|        YES|JFK-LHR. Had a gr...|
|10013|Delta Air Lines| 9-Jun-14|      USA|     0| Economy|    1|         NO|My wife and I fly...|
|10014|Delta Air Lines| 9-Jun-14|      USA|    10| Premium|    5|        YES|DL 1134 PBI-ATL. ...|
|10015|Delta Air Lines| 6-Jun-14|      USA|     0| Economy|    2|         NO|Our flight from F...|
|10016|Delta Air Lines| 5-Jun-14|      USA|     0| Economy|    1|         NO|On May 22 after a...|
|10017|Delta Air Lines| 3-Jun-14|   Canada|     9| Economy|    4|        YES|Considering how D...|
|10018|Delta Air Lines| 2-Jun-14|      USA|     9| Economy|    5|        YES|Travelled MSP-LHR...|
|10019|Delta Air Lines|29-May-14|      USA|     7|Business|    2|        YES|JFK-LAX on a 757-...|
|10020|Delta Air Lines|28-May-14|       UK|     9| Economy|    3|        YES|Third long haul f...|
+-----+---------------+---------+---------+------+--------+-----+-----------+--------------------+
only showing top 20 rows

>>> 
```

How many records are in the Dataframe?

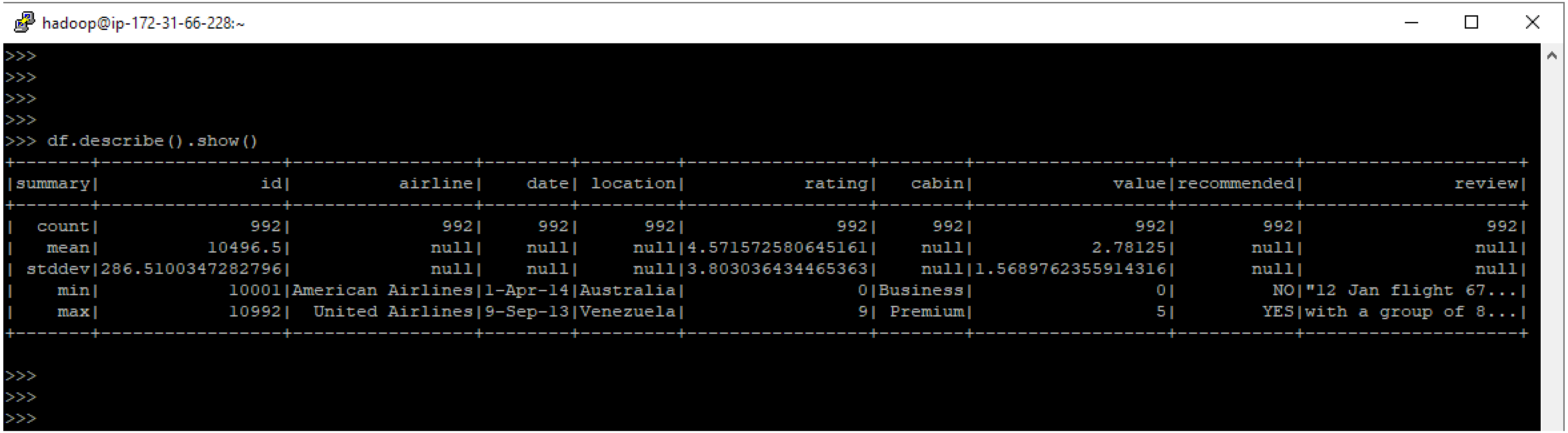
>>> df.count()



```
>>>
>>>
>>>
>>> df.count()
992
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
```

Display descriptive stats using the describe function

>>> df.describe().show()

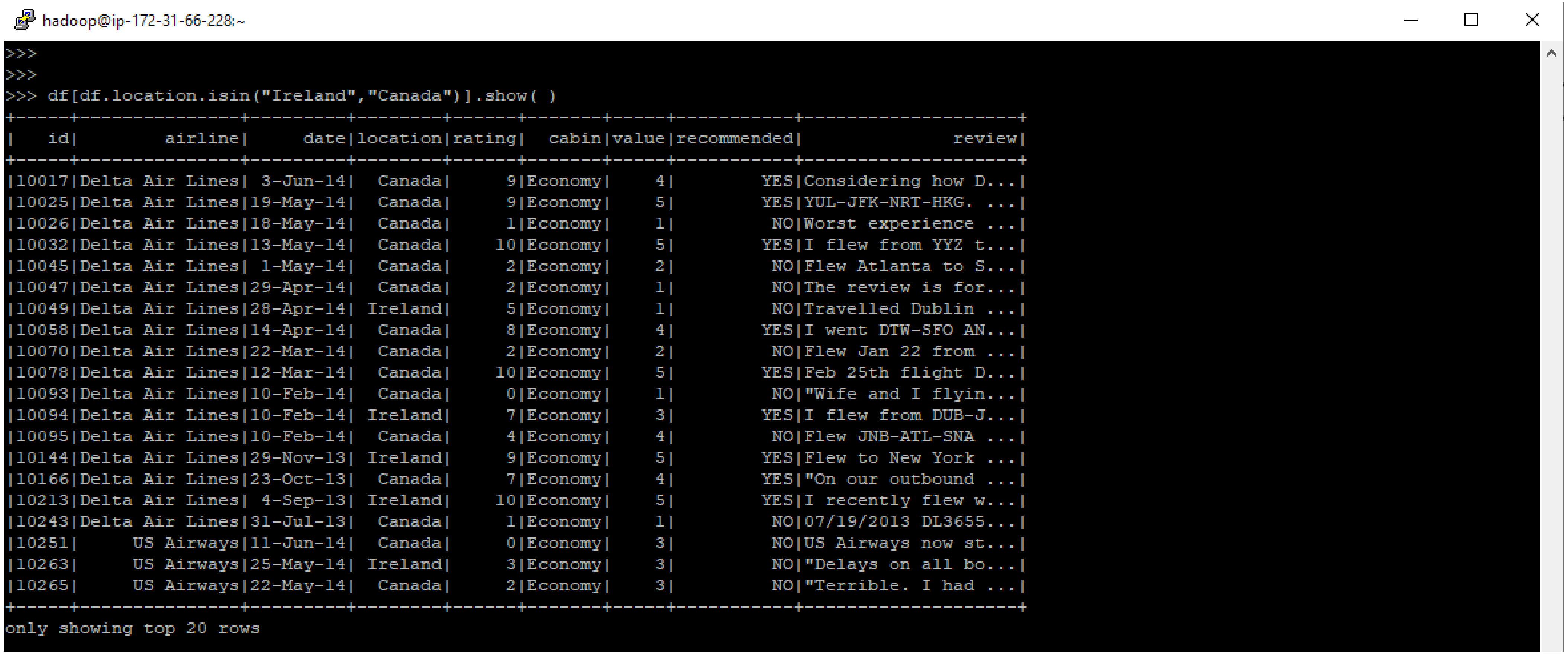


```
>>>
>>>
>>>
>>>
>>> df.describe().show()
+-------+-----------------+-----------------+--------+---------+-----------------+--------+------------------+-----------+--------------------+
|summary|               id|          airline|    date| location|           rating|   cabin|             value|recommended|              review|
+-------+-----------------+-----------------+--------+---------+-----------------+--------+------------------+-----------+--------------------+
|  count|              992|              992|     992|      992|              992|     992|               992|        992|                 992|
|   mean|          10496.5|             null|    null|     null|4.571572580645161|    null|           2.78125|       null|                null|
| stddev|286.5100347282796|             null|    null|     null|3.803036434465363|    null|1.5689762355914316|       null|                null|
|    min|            10001|American Airlines|1-Apr-14|Australia|                0|Business|                 0|         NO|"12 Jan flight 67...|
|    max|            10992|  United Airlines|9-Sep-13|Venezuela|                9| Premium|                 5|        YES|with a group of 8...|
+-------+-----------------+-----------------+--------+---------+-----------------+--------+------------------+-----------+--------------------+

>>>
>>>
>>>
```

Select only those observations where location is Ireland or Canada

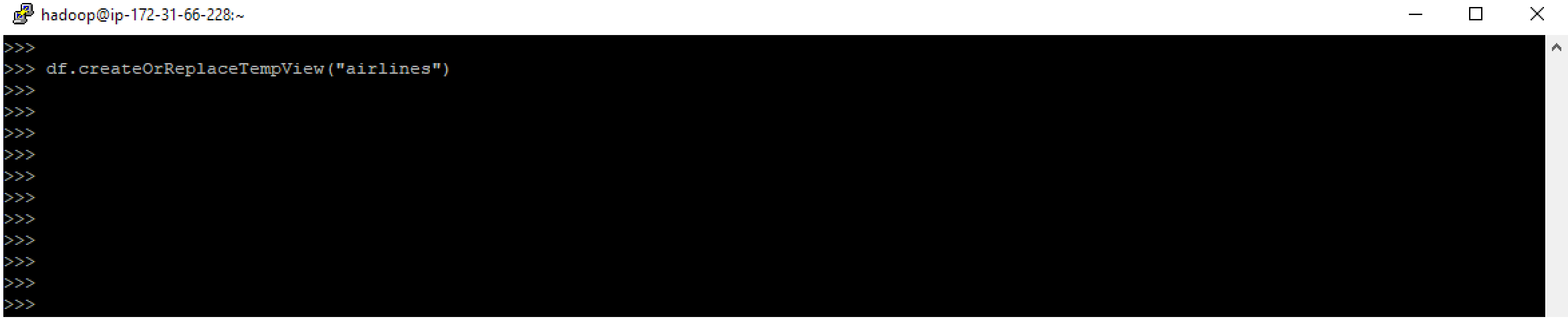
>>> df[df.location.isin("Ireland","Canada")].show( )



```
>>>
>>>
>>> df[df.location.isin("Ireland","Canada")].show( )
+-----+---------------+---------+--------+------+-------+-----+-----------+--------------------+
|   id|        airline|     date|location|rating|  cabin|value|recommended|              review|
+-----+---------------+---------+--------+------+-------+-----+-----------+--------------------+
|10017|Delta Air Lines| 3-Jun-14|  Canada|     9|Economy|    4|        YES|Considering how D...|
|10025|Delta Air Lines|19-May-14|  Canada|     9|Economy|    5|        YES|YUL-JFK-NRT-HKG. ...|
|10026|Delta Air Lines|18-May-14|  Canada|     1|Economy|    1|         NO|Worst experience ...|
|10032|Delta Air Lines|13-May-14|  Canada|    10|Economy|    5|        YES|I flew from YYZ t...|
|10045|Delta Air Lines| 1-May-14|  Canada|     2|Economy|    2|         NO|Flew Atlanta to S...|
|10047|Delta Air Lines|29-Apr-14|  Canada|     2|Economy|    1|         NO|The review is for...|
|10049|Delta Air Lines|28-Apr-14| Ireland|     5|Economy|    1|         NO|Travelled Dublin ...|
|10058|Delta Air Lines|14-Apr-14|  Canada|     8|Economy|    4|        YES|I went DTW-SFO AN...|
|10070|Delta Air Lines|22-Mar-14|  Canada|     2|Economy|    2|         NO|Flew Jan 22 from ...|
|10078|Delta Air Lines|12-Mar-14|  Canada|    10|Economy|    5|        YES|Feb 25th flight D...|
|10093|Delta Air Lines|10-Feb-14|  Canada|     0|Economy|    1|         NO|"Wife and I flyin...|
|10094|Delta Air Lines|10-Feb-14| Ireland|     7|Economy|    3|        YES|I flew from DUB-J...|
|10095|Delta Air Lines|10-Feb-14|  Canada|     4|Economy|    4|         NO|Flew JNB-ATL-SNA ...|
|10144|Delta Air Lines|29-Nov-13| Ireland|     9|Economy|    5|        YES|Flew to New York ...|
|10166|Delta Air Lines|23-Oct-13|  Canada|     7|Economy|    4|        YES|"On our outbound ...|
|10213|Delta Air Lines| 4-Sep-13| Ireland|    10|Economy|    5|        YES|I recently flew w...|
|10243|Delta Air Lines|31-Jul-13|  Canada|     1|Economy|    1|         NO|07/19/2013 DL3655...|
|10251|     US Airways|11-Jun-14|  Canada|     0|Economy|    3|         NO|US Airways now st...|
|10263|     US Airways|25-May-14| Ireland|     3|Economy|    3|         NO|"Delays on all bo...|
|10265|     US Airways|22-May-14|  Canada|     2|Economy|    3|         NO|"Terrible. I had ...|
+-----+---------------+---------+--------+------+-------+-----+-----------+--------------------+
only showing top 20 rows
```

Create a Spark temporary table (so that we can query the table using SparkSQL)

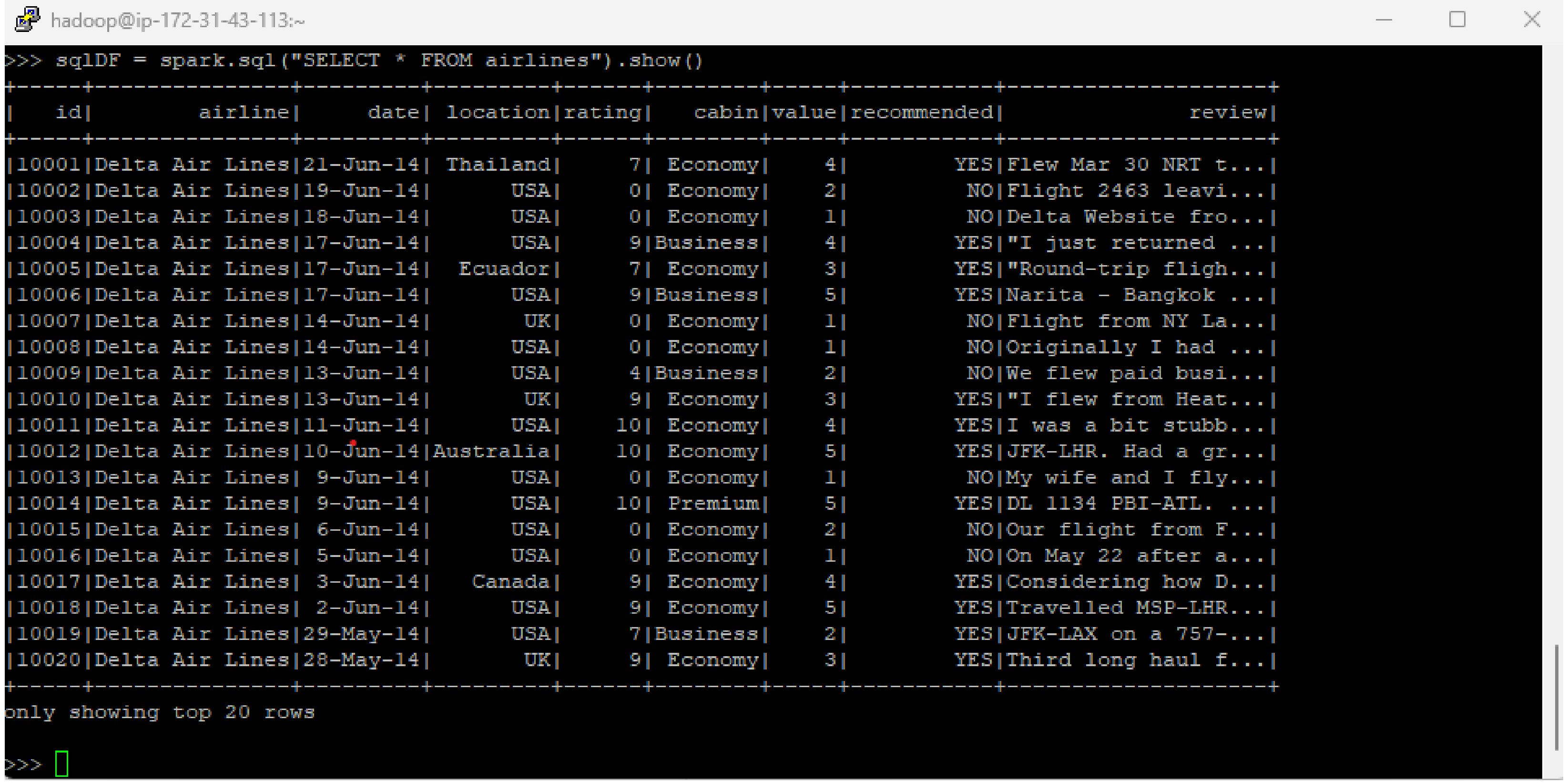
>>> df.createOrReplaceTempView("airlines")



```
>>>
>>> df.createOrReplaceTempView("airlines")
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
```

Use SparkSQL to query the temporary table that you just loaded in the previous step

>>> sqlDF = spark.sql("SELECT * FROM airlines").show()



```
>>> sqlDF = spark.sql("SELECT * FROM airlines").show()
+-----+---------------+---------+---------+------+--------+-----+-----------+--------------------+
|   id|        airline|     date| location|rating|   cabin|value|recommended|              review|
+-----+---------------+---------+---------+------+--------+-----+-----------+--------------------+
|10001|Delta Air Lines|21-Jun-14| Thailand|     7| Economy|    4|        YES|Flew Mar 30 NRT t...|
|10002|Delta Air Lines|19-Jun-14|      USA|     0| Economy|    2|         NO|Flight 2463 leavi...|
|10003|Delta Air Lines|18-Jun-14|      USA|     0| Economy|    1|         NO|Delta Website fro...|
|10004|Delta Air Lines|17-Jun-14|      USA|     9|Business|    4|        YES|"I just returned ...|
|10005|Delta Air Lines|17-Jun-14|  Ecuador|     7| Economy|    3|        YES|"Round-trip fligh...|
|10006|Delta Air Lines|17-Jun-14|      USA|     9|Business|    5|        YES|Narita - Bangkok ...|
|10007|Delta Air Lines|14-Jun-14|       UK|     0| Economy|    1|         NO|Flight from NY La...|
|10008|Delta Air Lines|14-Jun-14|      USA|     0| Economy|    1|         NO|Originally I had ...|
|10009|Delta Air Lines|13-Jun-14|      USA|     4|Business|    2|         NO|We flew paid busi...|
|10010|Delta Air Lines|13-Jun-14|       UK|     9| Economy|    3|        YES|"I flew from Heat...|
|10011|Delta Air Lines|11-Jun-14|      USA|    10| Economy|    4|        YES|I was a bit stubb...|
|10012|Delta Air Lines|10-Jun-14|Australia|    10| Economy|    5|        YES|JFK-LHR. Had a gr...|
|10013|Delta Air Lines| 9-Jun-14|      USA|     0| Economy|    1|         NO|My wife and I fly...|
|10014|Delta Air Lines| 9-Jun-14|      USA|    10| Premium|    5|        YES|DL 1134 PBI-ATL. ...|
|10015|Delta Air Lines| 6-Jun-14|      USA|     0| Economy|    2|         NO|Our flight from F...|
|10016|Delta Air Lines| 5-Jun-14|      USA|     0| Economy|    1|         NO|On May 22 after a...|
|10017|Delta Air Lines| 3-Jun-14|   Canada|     9| Economy|    4|        YES|Considering how D...|
|10018|Delta Air Lines| 2-Jun-14|      USA|     9| Economy|    5|        YES|Travelled MSP-LHR...|
|10019|Delta Air Lines|29-May-14|      USA|     7|Business|    2|        YES|JFK-LAX on a 757-...|
|10020|Delta Air Lines|28-May-14|       UK|     9| Economy|    3|        YES|Third long haul f...|
+-----+---------------+---------+---------+------+--------+-----+-----------+--------------------+
only showing top 20 rows

>>> 
```