UNIVERSIDAD EAFIT
INGENIERÍA DE SISTEMAS
ST0263 TÓPICOS ESPECIALES EN TELEMÁTICA, 2022-1

UNIDAD 3 BIG DATA

DOCUMENTACIÓN y EVIDENCIAS DE LAB 6

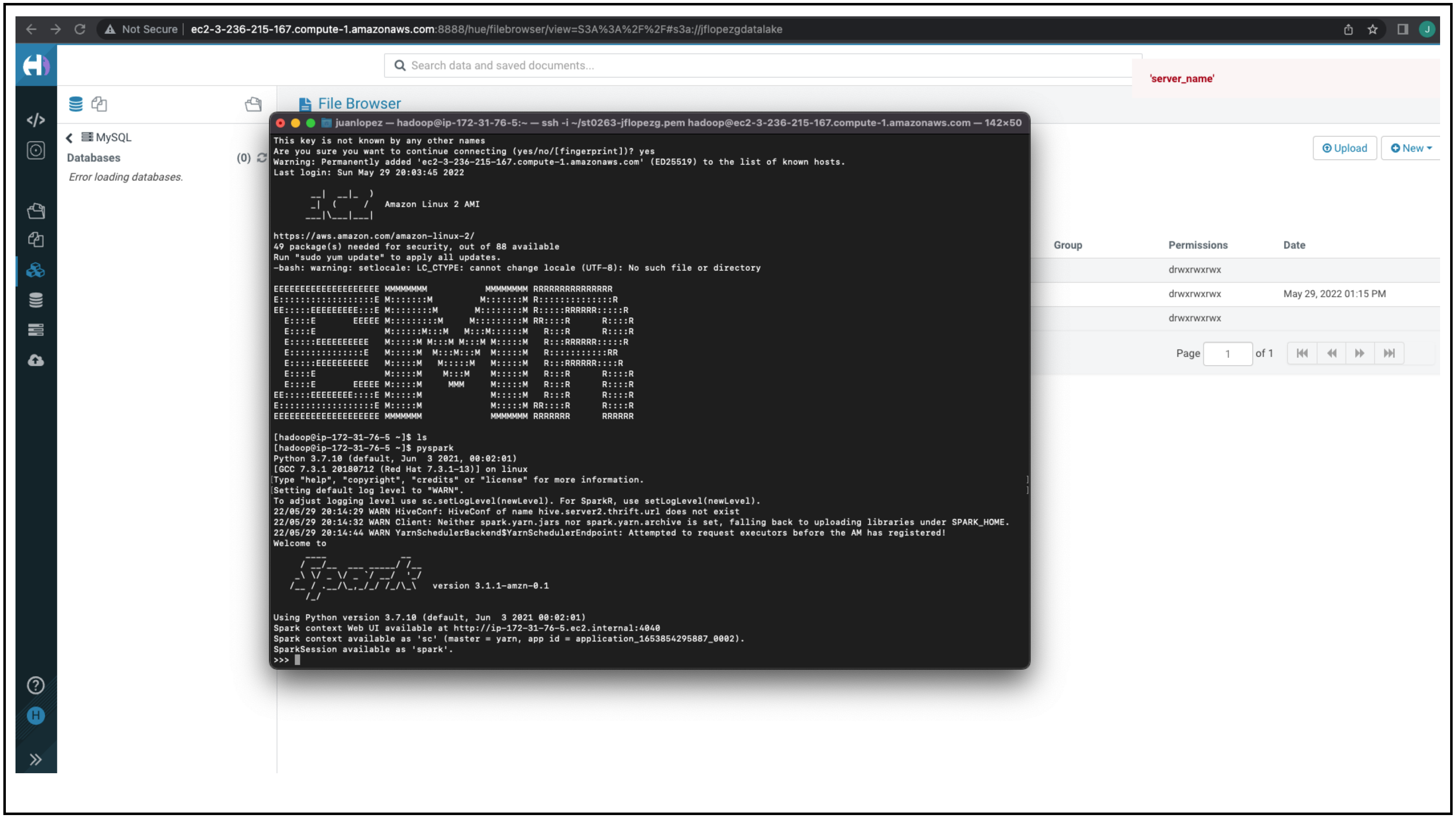
ESTE DOCUMENTO DEBE SER ENTREGADO POR BUZON DE ENTREGAS DE LA MATERIA
Además entregar por buzón la URL del repo github donde tiene todo de los LABs 5 y 6

NOMBRE DEL ALUMNO: Juan Felipe López Gutiérrez
EMAIL: jflopezg@eafit.edu.co
GITHUB DEL ALUMNO PARA LOS LABS https://github.com/JuanL-Code/st0263-lab6
Realizar las actividades relacionadas con el lab6 en: https://github.com/st0263eafit/st0263-2261/tree/main/bigdata

LAB6-EVIDENCIAS DE LA EJECUCIÓN DEL WORDCOUNG EN:

1. pyspark desde CLI interactivo en EMR
2. Pyspark desde CLI como archivo de entrada EMR
3. Pyspark desde jupyter notebooks en EMR
4. Ejecución del notebook jupyter: Data_processing_using_PySpark.ipynb en EMR con datos en S3

párrafos descriptivos, Screenshots, códigos fuente, extractos de código, scripts, urls, etc.



Not Secure | ec2-3-236-215-167.compute-1.amazonaws.com:8888/hue/filebrowser/view=S3A%3A%2F%2F#s3a://jflopezgdatalake
Search data and saved documents...
'server_name'
File Browser
MySQL
Upload
New
Page 1 of 1

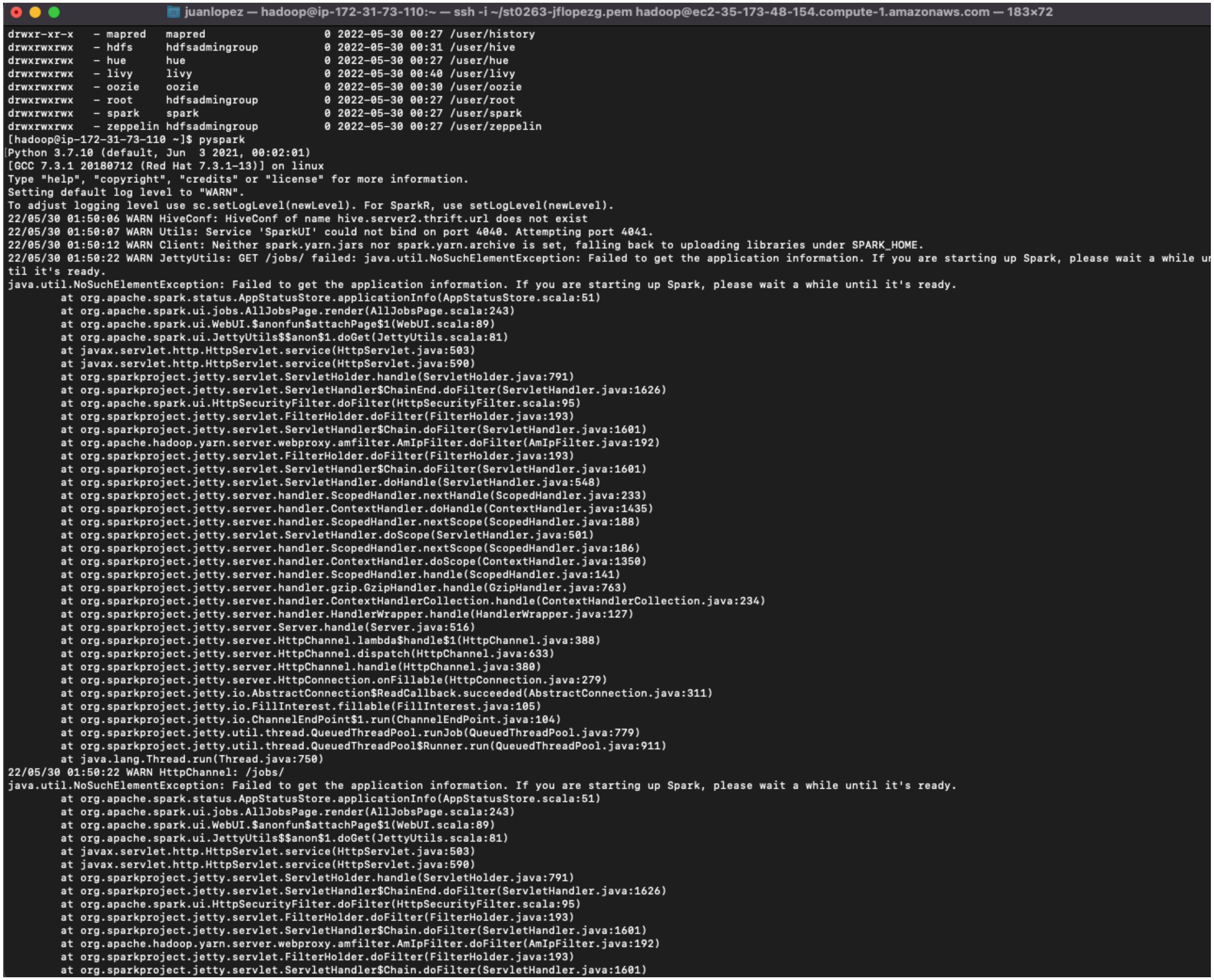
Databases (0)

Error loading databases.

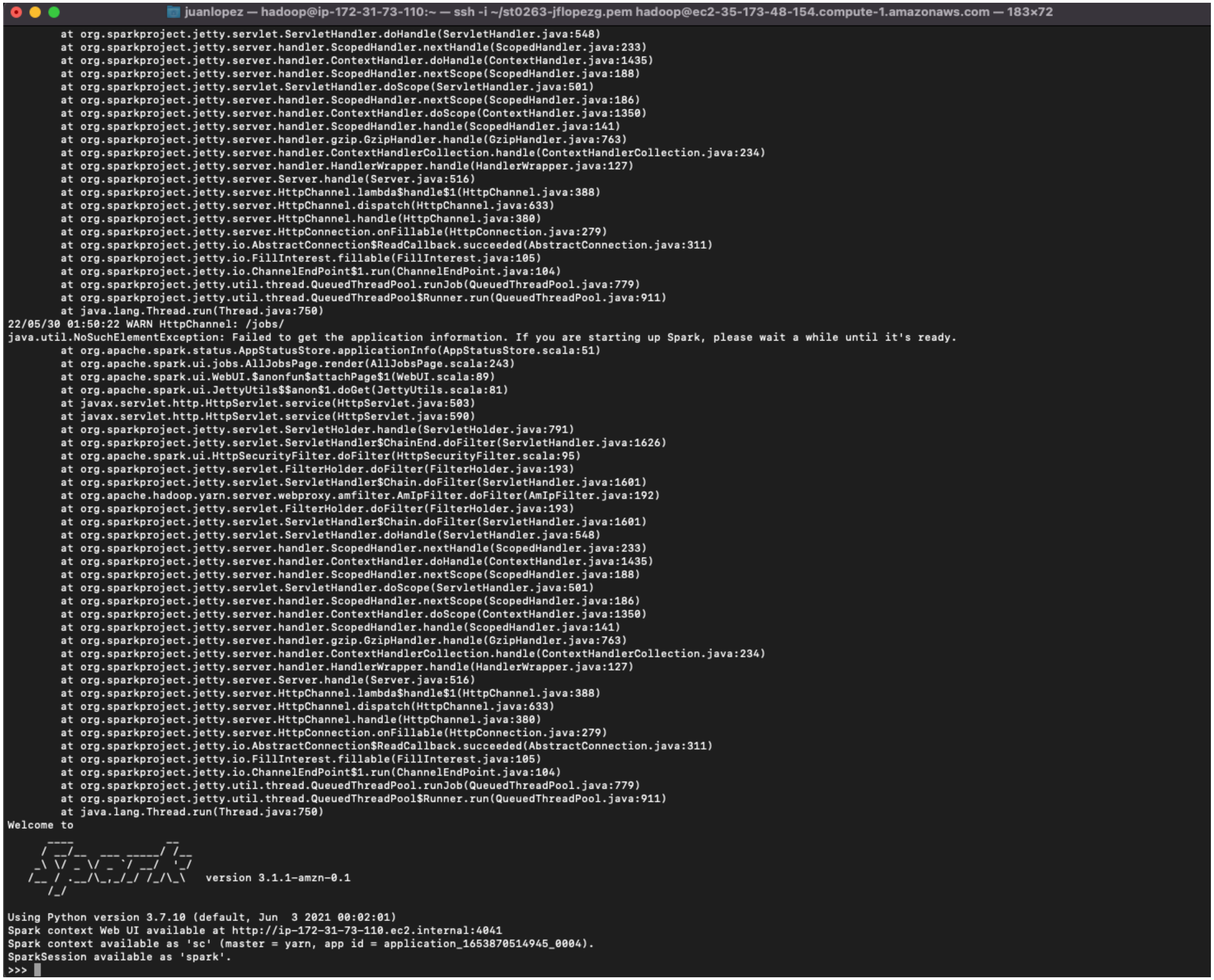
Group	Permissions	Date
	drwxrwxrwx	
	drwxrwxrwx	May 29, 2022 01:15 PM
	drwxrwxrwx	

juanlopez — hadoop@ip-172-31-76-5:~ — ssh -i ~/st0263-jflopezg.pem hadoop@ec2-3-236-215-167.compute-1.amazonaws.com — 142×50

```
This key is not known by any other names
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added 'ec2-3-236-215-167.compute-1.amazonaws.com' (ED25519) to the list of known hosts.
Last login: Sun May 29 20:03:45 2022

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

https://aws.amazon.com/amazon-linux-2/
49 package(s) needed for security, out of 88 available
Run "sudo yum update" to apply all updates.
-bash: warning: setlocale: LC_CTYPE: cannot change locale (UTF-8): No such file or directory

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-172-31-76-5 ~]$ ls
[hadoop@ip-172-31-76-5 ~]$ pyspark
Python 3.7.10 (default, Jun  3 2021, 00:02:01)
[GCC 7.3.1 20180712 (Red Hat 7.3.1-13)] on linux
Type "help", "copyright", "credits" or "license" for more information.
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
22/05/29 20:14:29 WARN HiveConf: HiveConf of name hive.server2.thrift.url does not exist
22/05/29 20:14:32 WARN Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploading libraries under SPARK_HOME.
22/05/29 20:14:44 WARN YarnSchedulerBackend$YarnSchedulerEndpoint: Attempted to request executors before the AM has registered!
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 3.1.1-amzn-0.1
      /_/

Using Python version 3.7.10 (default, Jun  3 2021 00:02:01)
Spark context Web UI available at http://ip-172-31-76-5.ec2.internal:4040
Spark context available as 'sc' (master = yarn, app id = application_1653854295887_0002).
SparkSession available as 'spark'.
>>> 
```

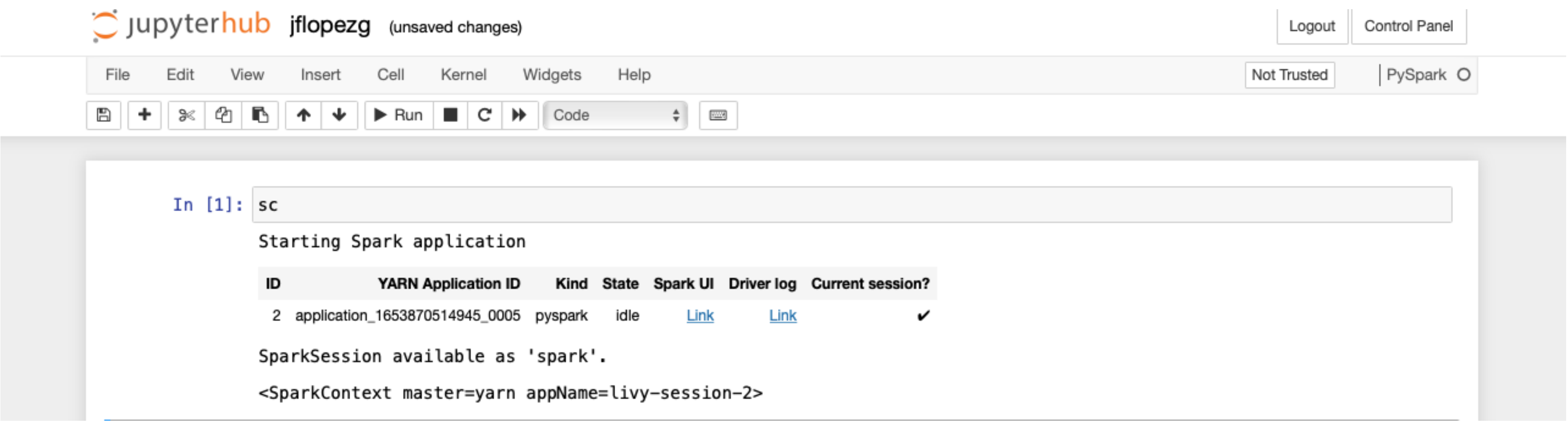

juanlopez — hadoop@ip-172-31-73-110:~ — ssh -i ~/st0263-jflopezg.pem hadoop@ec2-35-173-48-154.compute-1.amazonaws.com — 183×72

```
drwxr-xr-x   - mapred   mapred                   0 2022-05-30 00:27 /user/history
drwxrwxrwx   - hdfs     hdfsadmingroup           0 2022-05-30 00:31 /user/hive
drwxrwxrwx   - hue      hue                      0 2022-05-30 00:27 /user/hue
drwxrwxrwx   - livy     livy                     0 2022-05-30 00:40 /user/livy
drwxrwxrwx   - oozie    oozie                    0 2022-05-30 00:30 /user/oozie
drwxrwxrwx   - root     hdfsadmingroup           0 2022-05-30 00:27 /user/root
drwxrwxrwx   - spark    spark                    0 2022-05-30 00:27 /user/spark
drwxrwxrwx   - zeppelin hdfsadmingroup           0 2022-05-30 00:27 /user/zeppelin
[hadoop@ip-172-31-73-110 ~]$ pyspark
[Python 3.7.10 (default, Jun  3 2021, 00:02:01)
[GCC 7.3.1 20180712 (Red Hat 7.3.1-13)] on linux
Type "help", "copyright", "credits" or "license" for more information.
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
22/05/30 01:50:06 WARN HiveConf: HiveConf of name hive.server2.thrift.url does not exist
22/05/30 01:50:07 WARN Utils: Service 'SparkUI' could not bind on port 4040. Attempting port 4041.
22/05/30 01:50:12 WARN Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploading libraries under SPARK_HOME.
22/05/30 01:50:22 WARN JettyUtils: GET /jobs/ failed: java.util.NoSuchElementException: Failed to get the application information. If you are starting up Spark, please wait a while un
til it's ready.
java.util.NoSuchElementException: Failed to get the application information. If you are starting up Spark, please wait a while until it's ready.
        at org.apache.spark.status.AppStatusStore.applicationInfo(AppStatusStore.scala:51)
        at org.apache.spark.ui.jobs.AllJobsPage.render(AllJobsPage.scala:243)
        at org.apache.spark.ui.WebUI.$anonfun$attachPage$1(WebUI.scala:89)
        at org.apache.spark.ui.JettyUtils$$anon$1.doGet(JettyUtils.scala:81)
        at javax.servlet.http.HttpServlet.service(HttpServlet.java:503)
        at javax.servlet.http.HttpServlet.service(HttpServlet.java:590)
        at org.sparkproject.jetty.servlet.ServletHolder.handle(ServletHolder.java:791)
        at org.sparkproject.jetty.servlet.ServletHandler$ChainEnd.doFilter(ServletHandler.java:1626)
        at org.apache.spark.ui.HttpSecurityFilter.doFilter(HttpSecurityFilter.scala:95)
        at org.sparkproject.jetty.servlet.FilterHolder.doFilter(FilterHolder.java:193)
        at org.sparkproject.jetty.servlet.ServletHandler$Chain.doFilter(ServletHandler.java:1601)
        at org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter.doFilter(AmIpFilter.java:192)
        at org.sparkproject.jetty.servlet.FilterHolder.doFilter(FilterHolder.java:193)
        at org.sparkproject.jetty.servlet.ServletHandler$Chain.doFilter(ServletHandler.java:1601)
        at org.sparkproject.jetty.servlet.ServletHandler.doHandle(ServletHandler.java:548)
        at org.sparkproject.jetty.server.handler.ScopedHandler.nextHandle(ScopedHandler.java:233)
        at org.sparkproject.jetty.server.handler.ContextHandler.doHandle(ContextHandler.java:1435)
        at org.sparkproject.jetty.server.handler.ScopedHandler.nextScope(ScopedHandler.java:188)
        at org.sparkproject.jetty.servlet.ServletHandler.doScope(ServletHandler.java:501)
        at org.sparkproject.jetty.server.handler.ScopedHandler.nextScope(ScopedHandler.java:186)
        at org.sparkproject.jetty.server.handler.ContextHandler.doScope(ContextHandler.java:1350)
        at org.sparkproject.jetty.server.handler.ScopedHandler.handle(ScopedHandler.java:141)
        at org.sparkproject.jetty.server.handler.gzip.GzipHandler.handle(GzipHandler.java:763)
        at org.sparkproject.jetty.server.handler.ContextHandlerCollection.handle(ContextHandlerCollection.java:234)
        at org.sparkproject.jetty.server.handler.HandlerWrapper.handle(HandlerWrapper.java:127)
        at org.sparkproject.jetty.server.Server.handle(Server.java:516)
        at org.sparkproject.jetty.server.HttpChannel.lambda$handle$1(HttpChannel.java:388)
        at org.sparkproject.jetty.server.HttpChannel.dispatch(HttpChannel.java:633)
        at org.sparkproject.jetty.server.HttpChannel.handle(HttpChannel.java:380)
        at org.sparkproject.jetty.server.HttpConnection.onFillable(HttpConnection.java:279)
        at org.sparkproject.jetty.io.AbstractConnection$ReadCallback.succeeded(AbstractConnection.java:311)
        at org.sparkproject.jetty.io.FillInterest.fillable(FillInterest.java:105)
        at org.sparkproject.jetty.io.ChannelEndPoint$1.run(ChannelEndPoint.java:104)
        at org.sparkproject.jetty.util.thread.QueuedThreadPool.runJob(QueuedThreadPool.java:779)
        at org.sparkproject.jetty.util.thread.QueuedThreadPool$Runner.run(QueuedThreadPool.java:911)
        at java.lang.Thread.run(Thread.java:750)
22/05/30 01:50:22 WARN HttpChannel: /jobs/
java.util.NoSuchElementException: Failed to get the application information. If you are starting up Spark, please wait a while until it's ready.
        at org.apache.spark.status.AppStatusStore.applicationInfo(AppStatusStore.scala:51)
        at org.apache.spark.ui.jobs.AllJobsPage.render(AllJobsPage.scala:243)
        at org.apache.spark.ui.WebUI.$anonfun$attachPage$1(WebUI.scala:89)
        at org.apache.spark.ui.JettyUtils$$anon$1.doGet(JettyUtils.scala:81)
        at javax.servlet.http.HttpServlet.service(HttpServlet.java:503)
        at javax.servlet.http.HttpServlet.service(HttpServlet.java:590)
        at org.sparkproject.jetty.servlet.ServletHolder.handle(ServletHolder.java:791)
        at org.sparkproject.jetty.servlet.ServletHandler$ChainEnd.doFilter(ServletHandler.java:1626)
        at org.apache.spark.ui.HttpSecurityFilter.doFilter(HttpSecurityFilter.scala:95)
        at org.sparkproject.jetty.servlet.FilterHolder.doFilter(FilterHolder.java:193)
        at org.sparkproject.jetty.servlet.ServletHandler$Chain.doFilter(ServletHandler.java:1601)
        at org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter.doFilter(AmIpFilter.java:192)
        at org.sparkproject.jetty.servlet.FilterHolder.doFilter(FilterHolder.java:193)
        at org.sparkproject.jetty.servlet.ServletHandler$Chain.doFilter(ServletHandler.java:1601)
```



juanlopez — hadoop@ip-172-31-73-110:~ — ssh -i ~/st0263-jflopezg.pem hadoop@ec2-35-173-48-154.compute-1.amazonaws.com — 183×72

```
        at org.sparkproject.jetty.servlet.ServletHandler.doHandle(ServletHandler.java:548)
        at org.sparkproject.jetty.server.handler.ScopedHandler.nextHandle(ScopedHandler.java:233)
        at org.sparkproject.jetty.server.handler.ContextHandler.doHandle(ContextHandler.java:1435)
        at org.sparkproject.jetty.server.handler.ScopedHandler.nextScope(ScopedHandler.java:188)
        at org.sparkproject.jetty.servlet.ServletHandler.doScope(ServletHandler.java:501)
        at org.sparkproject.jetty.server.handler.ScopedHandler.nextScope(ScopedHandler.java:186)
        at org.sparkproject.jetty.server.handler.ContextHandler.doScope(ContextHandler.java:1350)
        at org.sparkproject.jetty.server.handler.ScopedHandler.handle(ScopedHandler.java:141)
        at org.sparkproject.jetty.server.handler.gzip.GzipHandler.handle(GzipHandler.java:763)
        at org.sparkproject.jetty.server.handler.ContextHandlerCollection.handle(ContextHandlerCollection.java:234)
        at org.sparkproject.jetty.server.handler.HandlerWrapper.handle(HandlerWrapper.java:127)
        at org.sparkproject.jetty.server.Server.handle(Server.java:516)
        at org.sparkproject.jetty.server.HttpChannel.lambda$handle$1(HttpChannel.java:388)
        at org.sparkproject.jetty.server.HttpChannel.dispatch(HttpChannel.java:633)
        at org.sparkproject.jetty.server.HttpChannel.handle(HttpChannel.java:380)
        at org.sparkproject.jetty.server.HttpConnection.onFillable(HttpConnection.java:279)
        at org.sparkproject.jetty.io.AbstractConnection$ReadCallback.succeeded(AbstractConnection.java:311)
        at org.sparkproject.jetty.io.FillInterest.fillable(FillInterest.java:105)
        at org.sparkproject.jetty.io.ChannelEndPoint$1.run(ChannelEndPoint.java:104)
        at org.sparkproject.jetty.util.thread.QueuedThreadPool.runJob(QueuedThreadPool.java:779)
        at org.sparkproject.jetty.util.thread.QueuedThreadPool$Runner.run(QueuedThreadPool.java:911)
        at java.lang.Thread.run(Thread.java:750)
22/05/30 01:50:22 WARN HttpChannel: /jobs/
java.util.NoSuchElementException: Failed to get the application information. If you are starting up Spark, please wait a while until it's ready.
        at org.apache.spark.status.AppStatusStore.applicationInfo(AppStatusStore.scala:51)
        at org.apache.spark.ui.jobs.AllJobsPage.render(AllJobsPage.scala:243)
        at org.apache.spark.ui.WebUI.$anonfun$attachPage$1(WebUI.scala:89)
        at org.apache.spark.ui.JettyUtils$$anon$1.doGet(JettyUtils.scala:81)
        at javax.servlet.http.HttpServlet.service(HttpServlet.java:503)
        at javax.servlet.http.HttpServlet.service(HttpServlet.java:590)
        at org.sparkproject.jetty.servlet.ServletHolder.handle(ServletHolder.java:791)
        at org.sparkproject.jetty.servlet.ServletHandler$ChainEnd.doFilter(ServletHandler.java:1626)
        at org.apache.spark.ui.HttpSecurityFilter.doFilter(HttpSecurityFilter.scala:95)
        at org.sparkproject.jetty.servlet.FilterHolder.doFilter(FilterHolder.java:193)
        at org.sparkproject.jetty.servlet.ServletHandler$Chain.doFilter(ServletHandler.java:1601)
        at org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter.doFilter(AmIpFilter.java:192)
        at org.sparkproject.jetty.servlet.FilterHolder.doFilter(FilterHolder.java:193)
        at org.sparkproject.jetty.servlet.ServletHandler$Chain.doFilter(ServletHandler.java:1601)
        at org.sparkproject.jetty.servlet.ServletHandler.doHandle(ServletHandler.java:548)
        at org.sparkproject.jetty.server.handler.ScopedHandler.nextHandle(ScopedHandler.java:233)
        at org.sparkproject.jetty.server.handler.ContextHandler.doHandle(ContextHandler.java:1435)
        at org.sparkproject.jetty.server.handler.ScopedHandler.nextScope(ScopedHandler.java:188)
        at org.sparkproject.jetty.servlet.ServletHandler.doScope(ServletHandler.java:501)
        at org.sparkproject.jetty.server.handler.ScopedHandler.nextScope(ScopedHandler.java:186)
        at org.sparkproject.jetty.server.handler.ContextHandler.doScope(ContextHandler.java:1350)
        at org.sparkproject.jetty.server.handler.ScopedHandler.handle(ScopedHandler.java:141)
        at org.sparkproject.jetty.server.handler.gzip.GzipHandler.handle(GzipHandler.java:763)
        at org.sparkproject.jetty.server.handler.ContextHandlerCollection.handle(ContextHandlerCollection.java:234)
        at org.sparkproject.jetty.server.handler.HandlerWrapper.handle(HandlerWrapper.java:127)
        at org.sparkproject.jetty.server.Server.handle(Server.java:516)
        at org.sparkproject.jetty.server.HttpChannel.lambda$handle$1(HttpChannel.java:388)
        at org.sparkproject.jetty.server.HttpChannel.dispatch(HttpChannel.java:633)
        at org.sparkproject.jetty.server.HttpChannel.handle(HttpChannel.java:380)
        at org.sparkproject.jetty.server.HttpConnection.onFillable(HttpConnection.java:279)
        at org.sparkproject.jetty.io.AbstractConnection$ReadCallback.succeeded(AbstractConnection.java:311)
        at org.sparkproject.jetty.io.FillInterest.fillable(FillInterest.java:105)
        at org.sparkproject.jetty.io.ChannelEndPoint$1.run(ChannelEndPoint.java:104)
        at org.sparkproject.jetty.util.thread.QueuedThreadPool.runJob(QueuedThreadPool.java:779)
        at org.sparkproject.jetty.util.thread.QueuedThreadPool$Runner.run(QueuedThreadPool.java:911)
        at java.lang.Thread.run(Thread.java:750)
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 3.1.1-amzn-0.1
      /_/

Using Python version 3.7.10 (default, Jun  3 2021 00:02:01)
Spark context Web UI available at http://ip-172-31-73-110.ec2.internal:4041
Spark context available as 'sc' (master = yarn, app id = application_1653870514945_0004).
SparkSession available as 'spark'.
>>> 
```



```
In [1]: sc
```

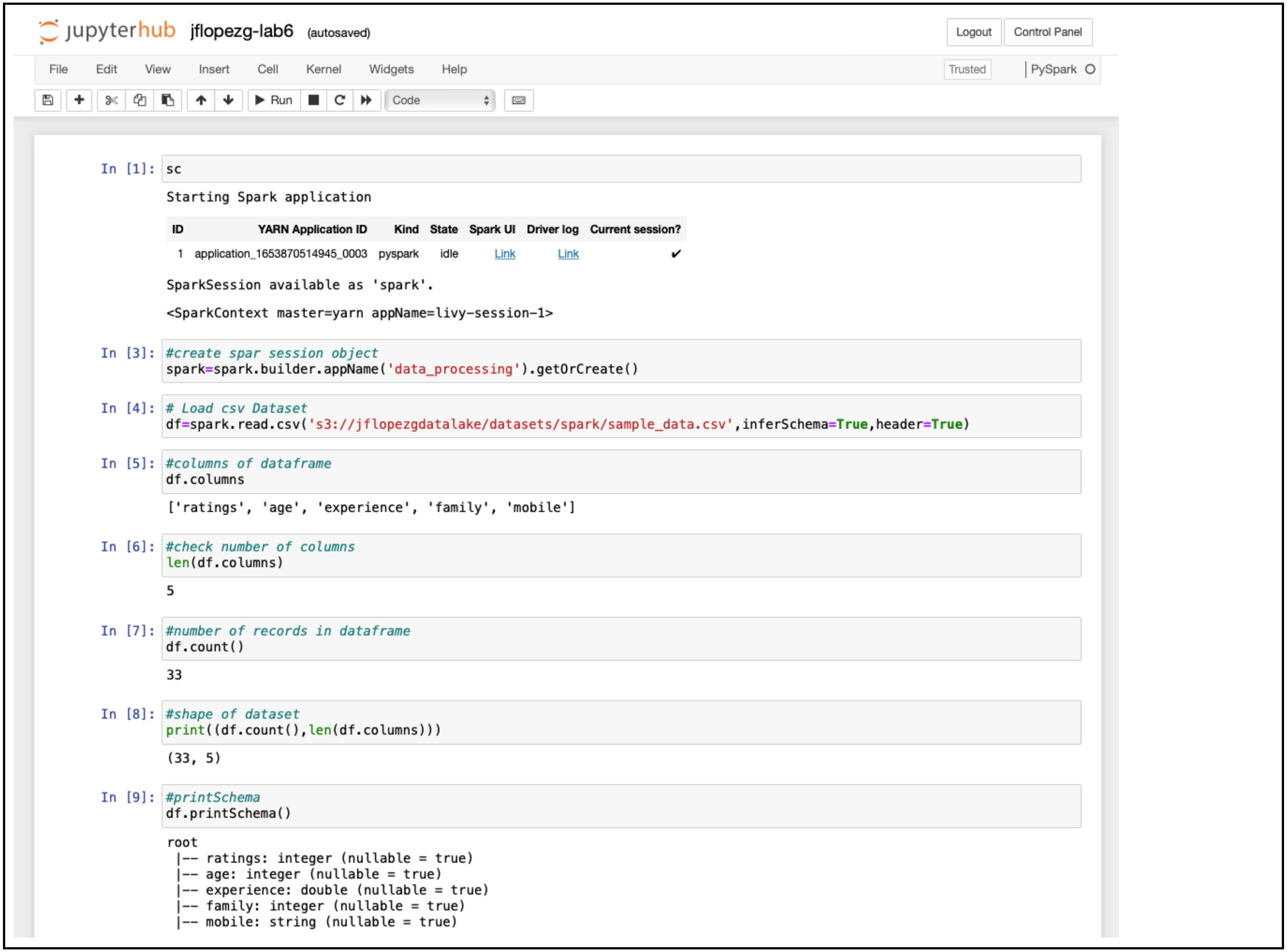
```
Starting Spark application
```

ID	YARN Application ID	Kind	State	Spark UI	Driver log	Current session?
2	application_1653870514945_0005	pyspark	idle	Link	Link	✔

```
SparkSession available as 'spark'.
<SparkContext master=yarn appName=livy-session-2>
```

LAB6- Ejecución del notebook jupyter: Data_processing_using_PySpark.ipynb en EMR con datos en S3

párrafos descriptivos, Screenshots, códigos fuente, extractos de código, scripts, urls, etc.



```
In [1]: sc
```

```
Starting Spark application
```

ID	YARN Application ID	Kind	State	Spark UI	Driver log	Current session?
1	application_1653870514945_0003	pyspark	idle	Link	Link	✔

```
SparkSession available as 'spark'.

<SparkContext master=yarn appName=livy-session-1>
```

```
In [3]: #create spar session object
        spark=spark.builder.appName('data_processing').getOrCreate()
```

```
In [4]: # Load csv Dataset
        df=spark.read.csv('s3://jflopezgdatalake/datasets/spark/sample_data.csv',inferSchema=True,header=True)
```

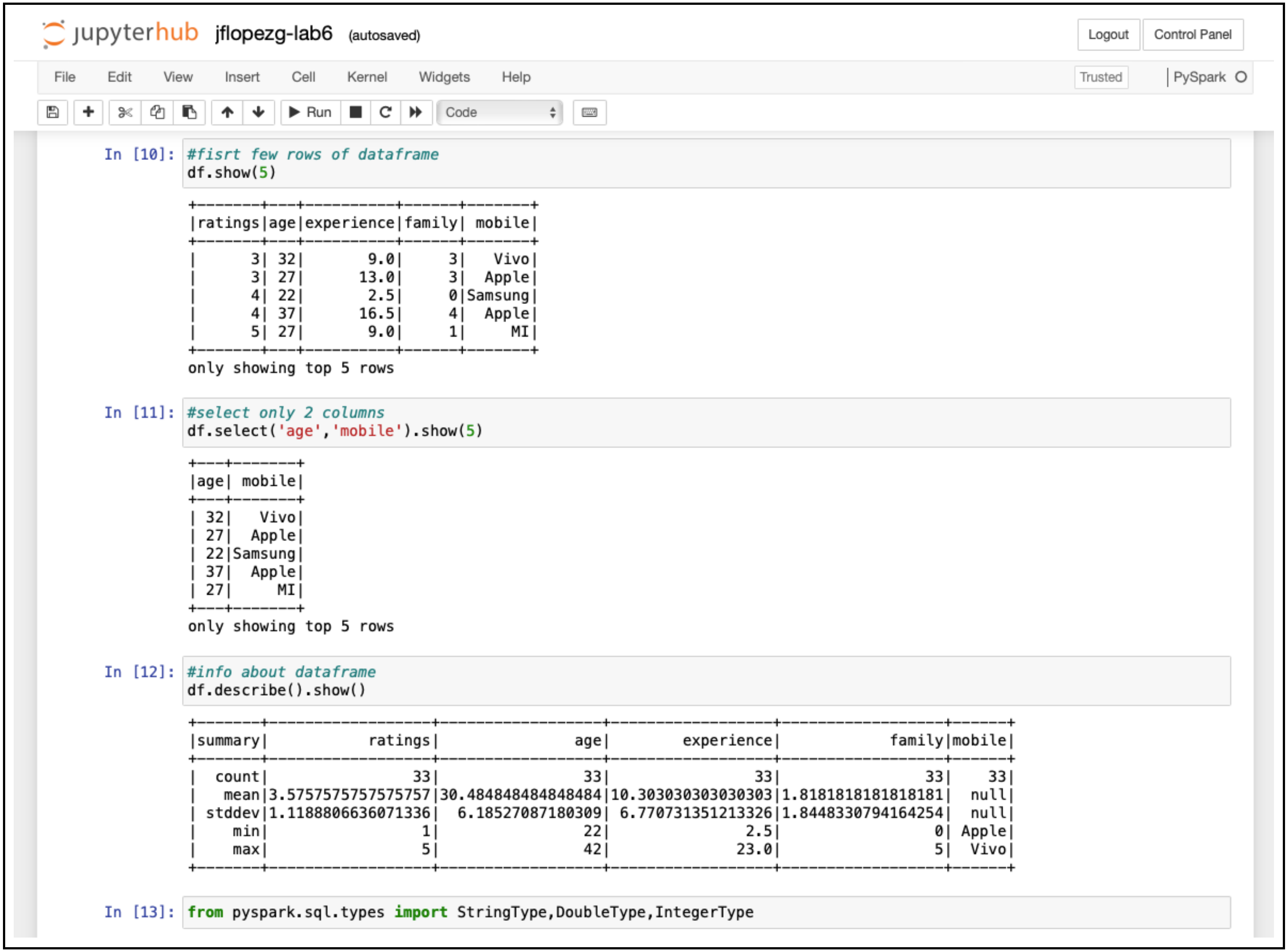
```
In [5]: #columns of dataframe
        df.columns

        ['ratings', 'age', 'experience', 'family', 'mobile']
```

```
In [6]: #check number of columns
        len(df.columns)

        5
```

```
In [7]: #number of records in dataframe
        df.count()

        33
```

```
In [8]: #shape of dataset
        print((df.count(),len(df.columns)))

        (33, 5)
```

```
In [9]: #printSchema
        df.printSchema()

        root
         |-- ratings: integer (nullable = true)
         |-- age: integer (nullable = true)
         |-- experience: double (nullable = true)
         |-- family: integer (nullable = true)
         |-- mobile: string (nullable = true)
```



In [10]:
```
#fisrt few rows of dataframe
df.show(5)
```

```
+-------+---+----------+------+-------+
|ratings|age|experience|family| mobile|
+-------+---+----------+------+-------+
|      3| 32|       9.0|     3|   Vivo|
|      3| 27|      13.0|     3|  Apple|
|      4| 22|       2.5|     0|Samsung|
|      4| 37|      16.5|     4|  Apple|
|      5| 27|       9.0|     1|     MI|
+-------+---+----------+------+-------+
only showing top 5 rows
```

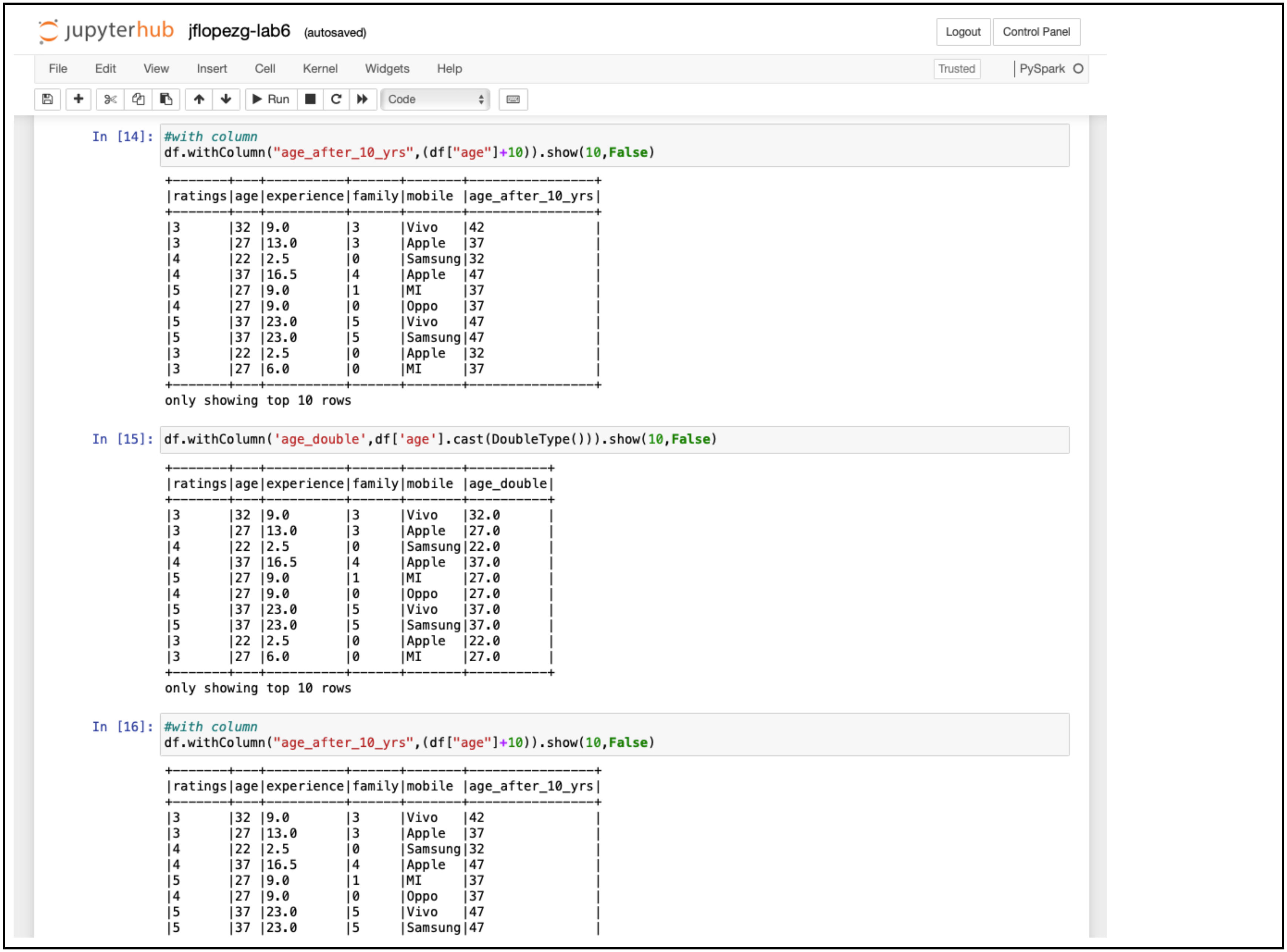
In [11]:
```
#select only 2 columns
df.select('age','mobile').show(5)
```

```
+---+-------+
|age| mobile|
+---+-------+
| 32|   Vivo|
| 27|  Apple|
| 22|Samsung|
| 37|  Apple|
| 27|     MI|
+---+-------+
only showing top 5 rows
```

In [12]:
```
#info about dataframe
df.describe().show()
```

```
+-------+------------------+------------------+------------------+------------------+------+
|summary|           ratings|               age|        experience|            family|mobile|
+-------+------------------+------------------+------------------+------------------+------+
|  count|                33|                33|                33|                33|    33|
|   mean|3.5757575757575757|30.484848484848484|10.303030303030303|1.8181818181818181|  null|
| stddev|1.1188806636071336|  6.18527087180309| 6.770731351213326|1.8448330794164254|  null|
|    min|                 1|                22|               2.5|                 0| Apple|
|    max|                 5|                42|              23.0|                 5|  Vivo|
+-------+------------------+------------------+------------------+------------------+------+
```

In [13]:
```
from pyspark.sql.types import StringType,DoubleType,IntegerType
```



In [14]:
```
#with column
df.withColumn("age_after_10_yrs",(df["age"]+10)).show(10,False)
```

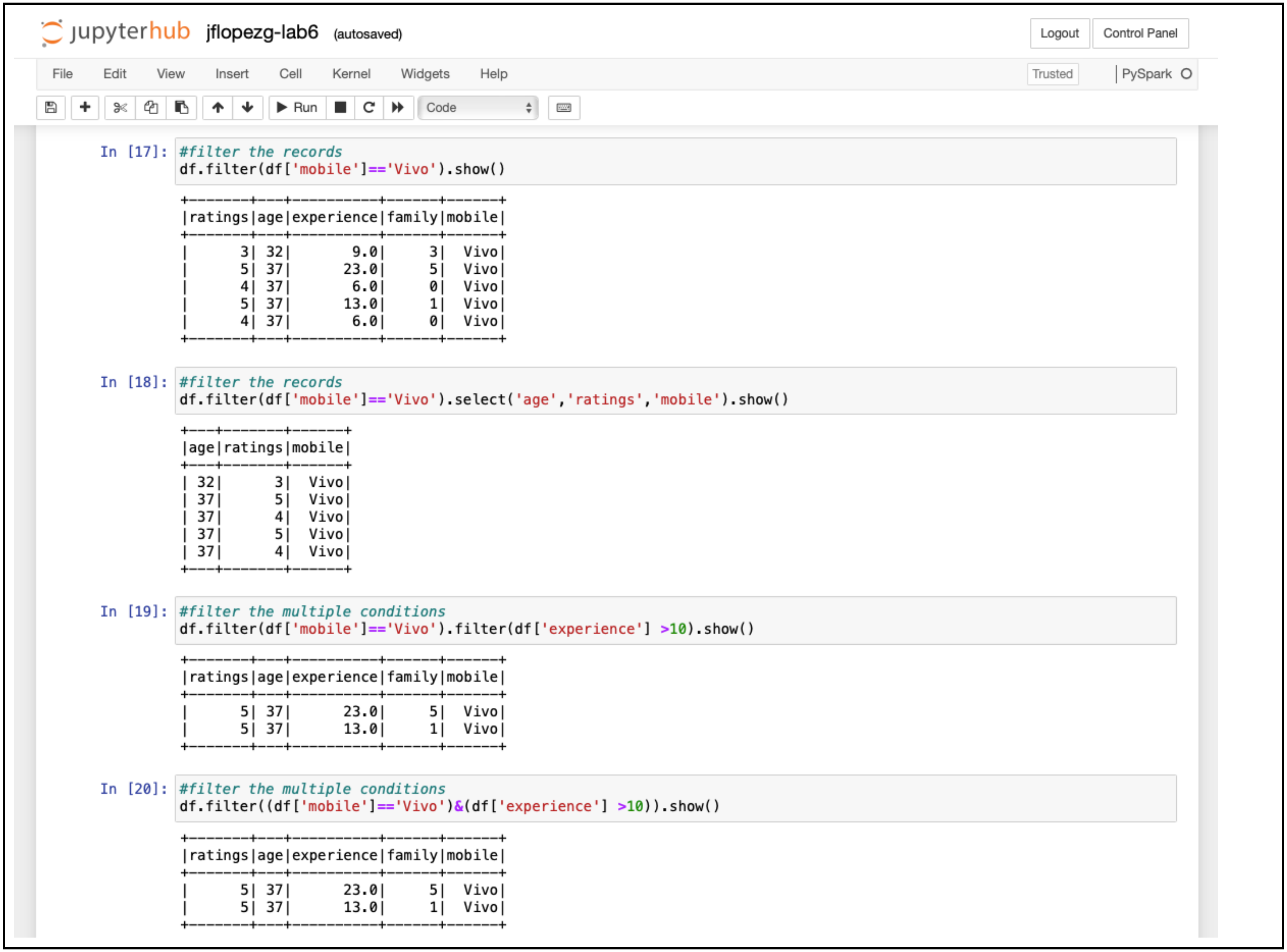
```
+-------+---+----------+------+-------+----------------+
|ratings|age|experience|family|mobile |age_after_10_yrs|
+-------+---+----------+------+-------+----------------+
|3      |32 |9.0       |3     |Vivo   |42              |
|3      |27 |13.0      |3     |Apple  |37              |
|4      |22 |2.5       |0     |Samsung|32              |
|4      |37 |16.5      |4     |Apple  |47              |
|5      |27 |9.0       |1     |MI     |37              |
|4      |27 |9.0       |0     |Oppo   |37              |
|5      |37 |23.0      |5     |Vivo   |47              |
|5      |37 |23.0      |5     |Samsung|47              |
|3      |22 |2.5       |0     |Apple  |32              |
|3      |27 |6.0       |0     |MI     |37              |
+-------+---+----------+------+-------+----------------+
only showing top 10 rows
```

In [15]:
```
df.withColumn('age_double',df['age'].cast(DoubleType())).show(10,False)
```

```
+-------+---+----------+------+-------+----------+
|ratings|age|experience|family|mobile |age_double|
+-------+---+----------+------+-------+----------+
|3      |32 |9.0       |3     |Vivo   |32.0      |
|3      |27 |13.0      |3     |Apple  |27.0      |
|4      |22 |2.5       |0     |Samsung|22.0      |
|4      |37 |16.5      |4     |Apple  |37.0      |
|5      |27 |9.0       |1     |MI     |27.0      |
|4      |27 |9.0       |0     |Oppo   |27.0      |
|5      |37 |23.0      |5     |Vivo   |37.0      |
|5      |37 |23.0      |5     |Samsung|37.0      |
|3      |22 |2.5       |0     |Apple  |22.0      |
|3      |27 |6.0       |0     |MI     |27.0      |
+-------+---+----------+------+-------+----------+
only showing top 10 rows
```

In [16]:
```
#with column
df.withColumn("age_after_10_yrs",(df["age"]+10)).show(10,False)
```

```
+-------+---+----------+------+-------+----------------+
|ratings|age|experience|family|mobile |age_after_10_yrs|
+-------+---+----------+------+-------+----------------+
|3      |32 |9.0       |3     |Vivo   |42              |
|3      |27 |13.0      |3     |Apple  |37              |
|4      |22 |2.5       |0     |Samsung|32              |
|4      |37 |16.5      |4     |Apple  |47              |
|5      |27 |9.0       |1     |MI     |37              |
|4      |27 |9.0       |0     |Oppo   |37              |
|5      |37 |23.0      |5     |Vivo   |47              |
|5      |37 |23.0      |5     |Samsung|47              |
```



In [17]:
```
#filter the records
df.filter(df['mobile']=='Vivo').show()
```

```
+-------+---+----------+------+------+
|ratings|age|experience|family|mobile|
+-------+---+----------+------+------+
|      3| 32|       9.0|     3|  Vivo|
|      5| 37|      23.0|     5|  Vivo|
|      4| 37|       6.0|     0|  Vivo|
|      5| 37|      13.0|     1|  Vivo|
|      4| 37|       6.0|     0|  Vivo|
+-------+---+----------+------+------+
```

In [18]:
```
#filter the records
df.filter(df['mobile']=='Vivo').select('age','ratings','mobile').show()
```

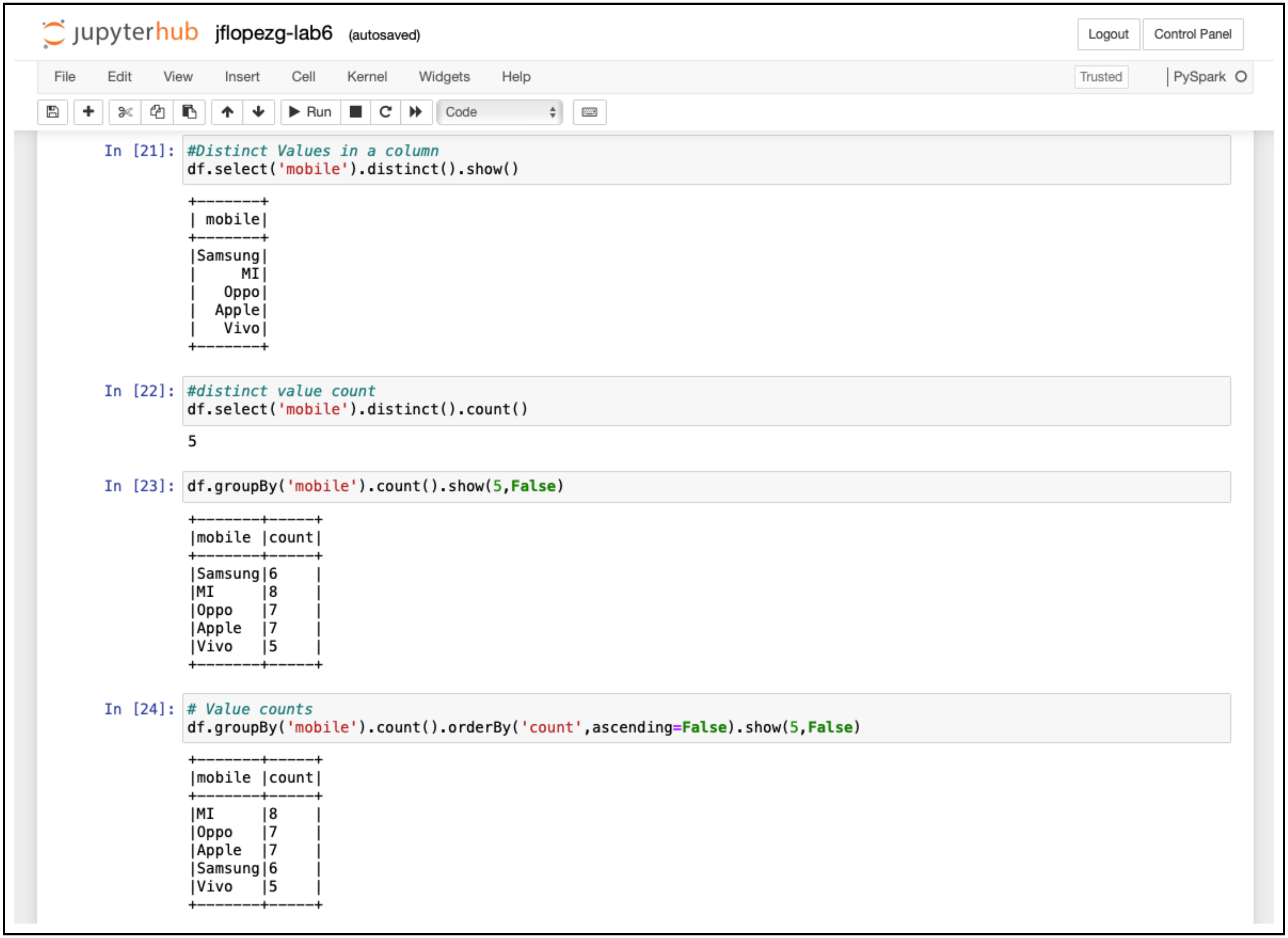
```
+---+-------+------+
|age|ratings|mobile|
+---+-------+------+
| 32|      3|  Vivo|
| 37|      5|  Vivo|
| 37|      4|  Vivo|
| 37|      5|  Vivo|
| 37|      4|  Vivo|
+---+-------+------+
```

In [19]:
```
#filter the multiple conditions
df.filter(df['mobile']=='Vivo').filter(df['experience'] >10).show()
```

```
+-------+---+----------+------+------+
|ratings|age|experience|family|mobile|
+-------+---+----------+------+------+
|      5| 37|      23.0|     5|  Vivo|
|      5| 37|      13.0|     1|  Vivo|
+-------+---+----------+------+------+
```

In [20]:
```
#filter the multiple conditions
df.filter((df['mobile']=='Vivo')&(df['experience'] >10)).show()
```

```
+-------+---+----------+------+------+
|ratings|age|experience|family|mobile|
+-------+---+----------+------+------+
|      5| 37|      23.0|     5|  Vivo|
|      5| 37|      13.0|     1|  Vivo|
+-------+---+----------+------+------+
```



In [21]:
```
#Distinct Values in a column
df.select('mobile').distinct().show()
```

```
+-------+
| mobile|
+-------+
|Samsung|
|     MI|
|   Oppo|
|  Apple|
|   Vivo|
+-------+
```

In [22]:
```
#distinct value count
df.select('mobile').distinct().count()
```

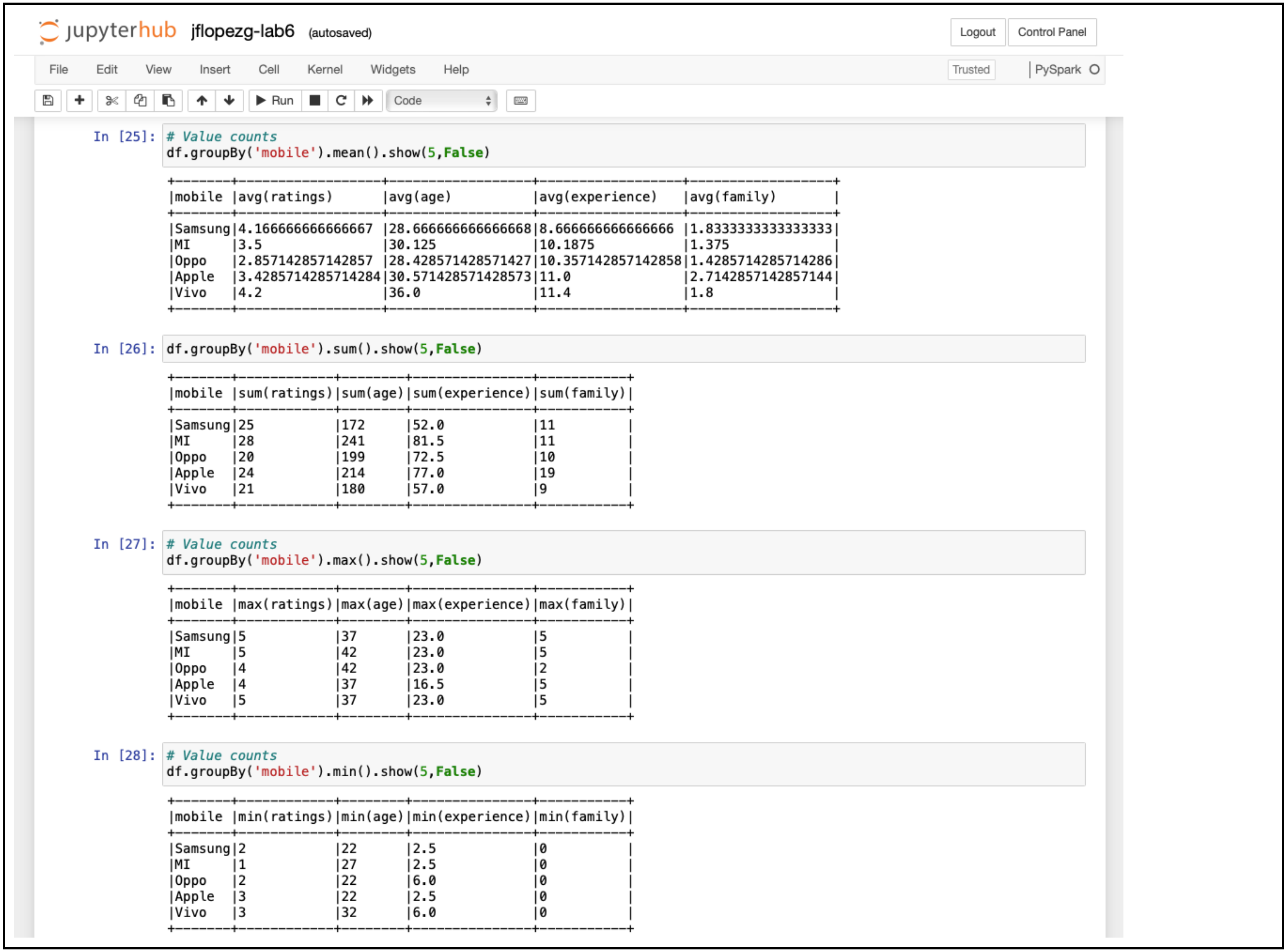
```
5
```

In [23]:
```
df.groupBy('mobile').count().show(5,False)
```

```
+-------+-----+
|mobile |count|
+-------+-----+
|Samsung|6    |
|MI     |8    |
|Oppo   |7    |
|Apple  |7    |
|Vivo   |5    |
+-------+-----+
```

In [24]:
```
# Value counts
df.groupBy('mobile').count().orderBy('count',ascending=False).show(5,False)
```

```
+-------+-----+
|mobile |count|
+-------+-----+
|MI     |8    |
|Oppo   |7    |
|Apple  |7    |
|Samsung|6    |
|Vivo   |5    |
+-------+-----+
```



In [25]:
```
# Value counts
df.groupBy('mobile').mean().show(5,False)
```

```
+-------+------------------+------------------+------------------+------------------+
|mobile |avg(ratings)      |avg(age)          |avg(experience)   |avg(family)       |
+-------+------------------+------------------+------------------+------------------+
|Samsung|4.166666666666667 |28.666666666666668|8.666666666666666 |1.8333333333333333|
|MI     |3.5               |30.125            |10.1875           |1.375             |
|Oppo   |2.857142857142857 |28.428571428571427|10.357142857142858|1.4285714285714286|
|Apple  |3.4285714285714284|30.571428571428573|11.0              |2.7142857142857144|
|Vivo   |4.2               |36.0              |11.4              |1.8               |
+-------+------------------+------------------+------------------+------------------+
```

In [26]:
```
df.groupBy('mobile').sum().show(5,False)
```

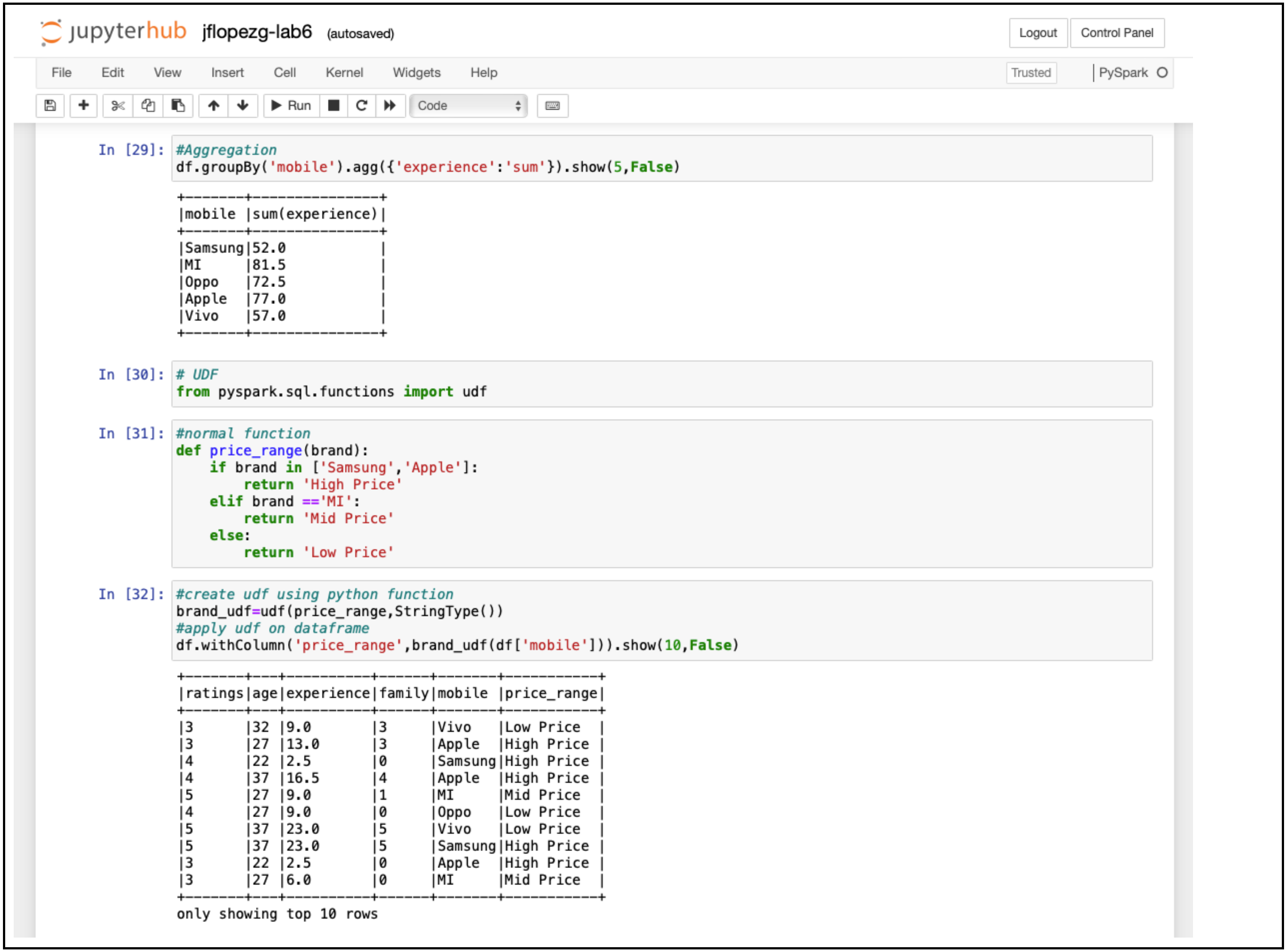
```
+-------+------------+--------+---------------+-----------+
|mobile |sum(ratings)|sum(age)|sum(experience)|sum(family)|
+-------+------------+--------+---------------+-----------+
|Samsung|25          |172     |52.0           |11         |
|MI     |28          |241     |81.5           |11         |
|Oppo   |20          |199     |72.5           |10         |
|Apple  |24          |214     |77.0           |19         |
|Vivo   |21          |180     |57.0           |9          |
+-------+------------+--------+---------------+-----------+
```

In [27]:
```
# Value counts
df.groupBy('mobile').max().show(5,False)
```

```
+-------+------------+--------+---------------+-----------+
|mobile |max(ratings)|max(age)|max(experience)|max(family)|
+-------+------------+--------+---------------+-----------+
|Samsung|5           |37      |23.0           |5          |
|MI     |5           |42      |23.0           |5          |
|Oppo   |4           |42      |23.0           |2          |
|Apple  |4           |37      |16.5           |5          |
|Vivo   |5           |37      |23.0           |5          |
+-------+------------+--------+---------------+-----------+
```

In [28]:
```
# Value counts
df.groupBy('mobile').min().show(5,False)
```

```
+-------+------------+--------+---------------+-----------+
|mobile |min(ratings)|min(age)|min(experience)|min(family)|
+-------+------------+--------+---------------+-----------+
|Samsung|2           |22      |2.5            |0          |
|MI     |1           |27      |2.5            |0          |
|Oppo   |2           |22      |6.0            |0          |
|Apple  |3           |22      |2.5            |0          |
|Vivo   |3           |32      |6.0            |0          |
+-------+------------+--------+---------------+-----------+
```



In [29]:
```
#Aggregation
df.groupBy('mobile').agg({'experience':'sum'}).show(5,False)
```

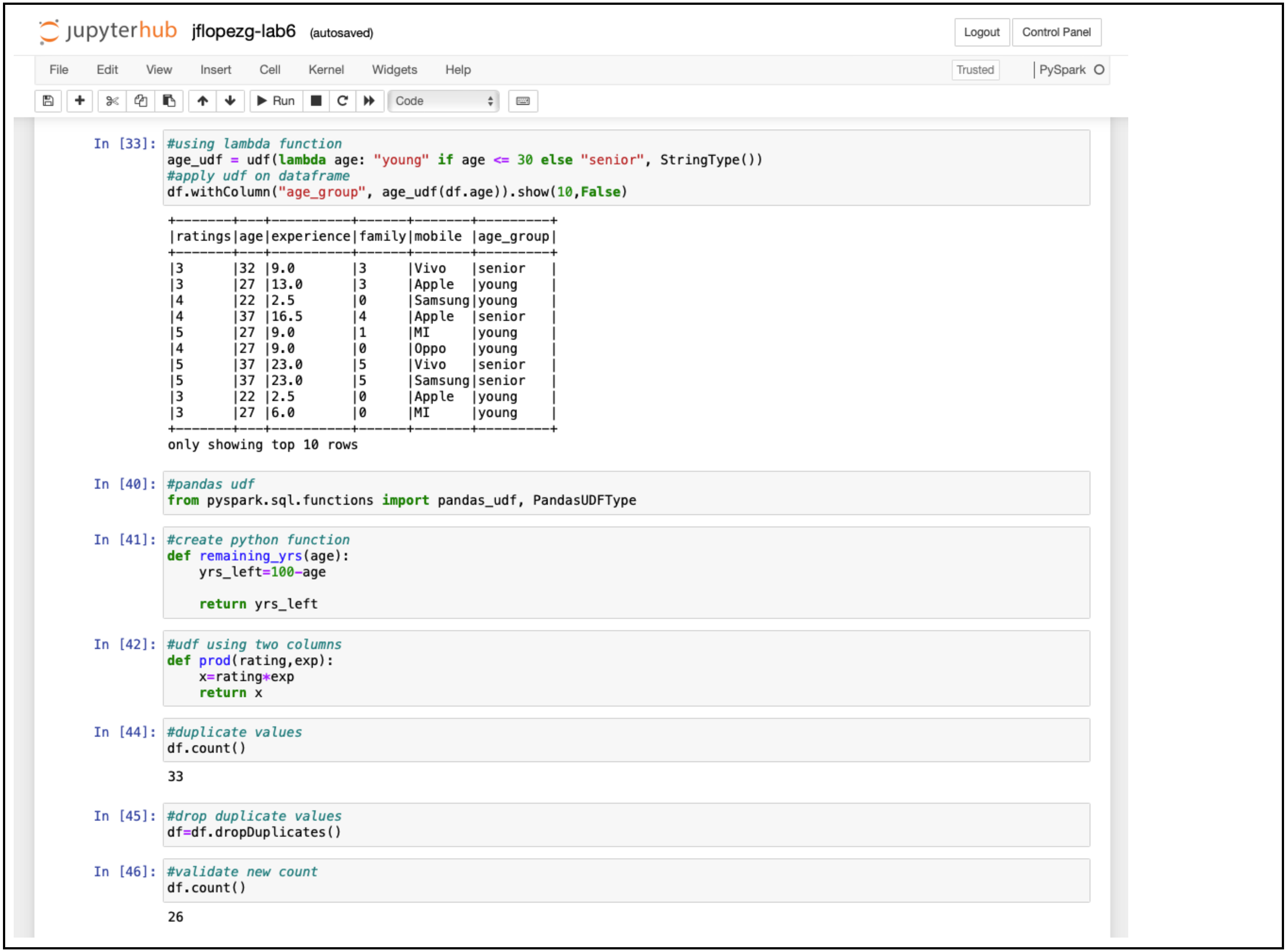
```
+-------+---------------+
|mobile |sum(experience)|
+-------+---------------+
|Samsung|52.0           |
|MI     |81.5           |
|Oppo   |72.5           |
|Apple  |77.0           |
|Vivo   |57.0           |
+-------+---------------+
```

In [30]:
```
# UDF
from pyspark.sql.functions import udf
```

In [31]:
```
#normal function
def price_range(brand):
    if brand in ['Samsung','Apple']:
        return 'High Price'
    elif brand =='MI':
        return 'Mid Price'
    else:
        return 'Low Price'
```

In [32]:
```
#create udf using python function
brand_udf=udf(price_range,StringType())
#apply udf on dataframe
df.withColumn('price_range',brand_udf(df['mobile'])).show(10,False)
```

```
+-------+---+----------+------+-------+-----------+
|ratings|age|experience|family|mobile |price_range|
+-------+---+----------+------+-------+-----------+
|3      |32 |9.0       |3     |Vivo   |Low Price  |
|3      |27 |13.0      |3     |Apple  |High Price |
|4      |22 |2.5       |0     |Samsung|High Price |
|4      |37 |16.5      |4     |Apple  |High Price |
|5      |27 |9.0       |1     |MI     |Mid Price  |
|4      |27 |9.0       |0     |Oppo   |Low Price  |
|5      |37 |23.0      |5     |Vivo   |Low Price  |
|5      |37 |23.0      |5     |Samsung|High Price |
|3      |22 |2.5       |0     |Apple  |High Price |
|3      |27 |6.0       |0     |MI     |Mid Price  |
+-------+---+----------+------+-------+-----------+
only showing top 10 rows
```



In [33]:
```
#using lambda function
age_udf = udf(lambda age: "young" if age <= 30 else "senior", StringType())
#apply udf on dataframe
df.withColumn("age_group", age_udf(df.age)).show(10,False)
```

```
+-------+---+----------+------+-------+---------+
|ratings|age|experience|family|mobile |age_group|
+-------+---+----------+------+-------+---------+
|3      |32 |9.0       |3     |Vivo   |senior   |
|3      |27 |13.0      |3     |Apple  |young    |
|4      |22 |2.5       |0     |Samsung|young    |
|4      |37 |16.5      |4     |Apple  |senior   |
|5      |27 |9.0       |1     |MI     |young    |
|4      |27 |9.0       |0     |Oppo   |young    |
|5      |37 |23.0      |5     |Vivo   |senior   |
|5      |37 |23.0      |5     |Samsung|senior   |
|3      |22 |2.5       |0     |Apple  |young    |
|3      |27 |6.0       |0     |MI     |young    |
+-------+---+----------+------+-------+---------+
only showing top 10 rows
```

In [40]:
```
#pandas udf
from pyspark.sql.functions import pandas_udf, PandasUDFType
```

In [41]:
```
#create python function
def remaining_yrs(age):
    yrs_left=100-age

    return yrs_left
```

In [42]:
```
#udf using two columns
def prod(rating,exp):
    x=rating*exp
    return x
```

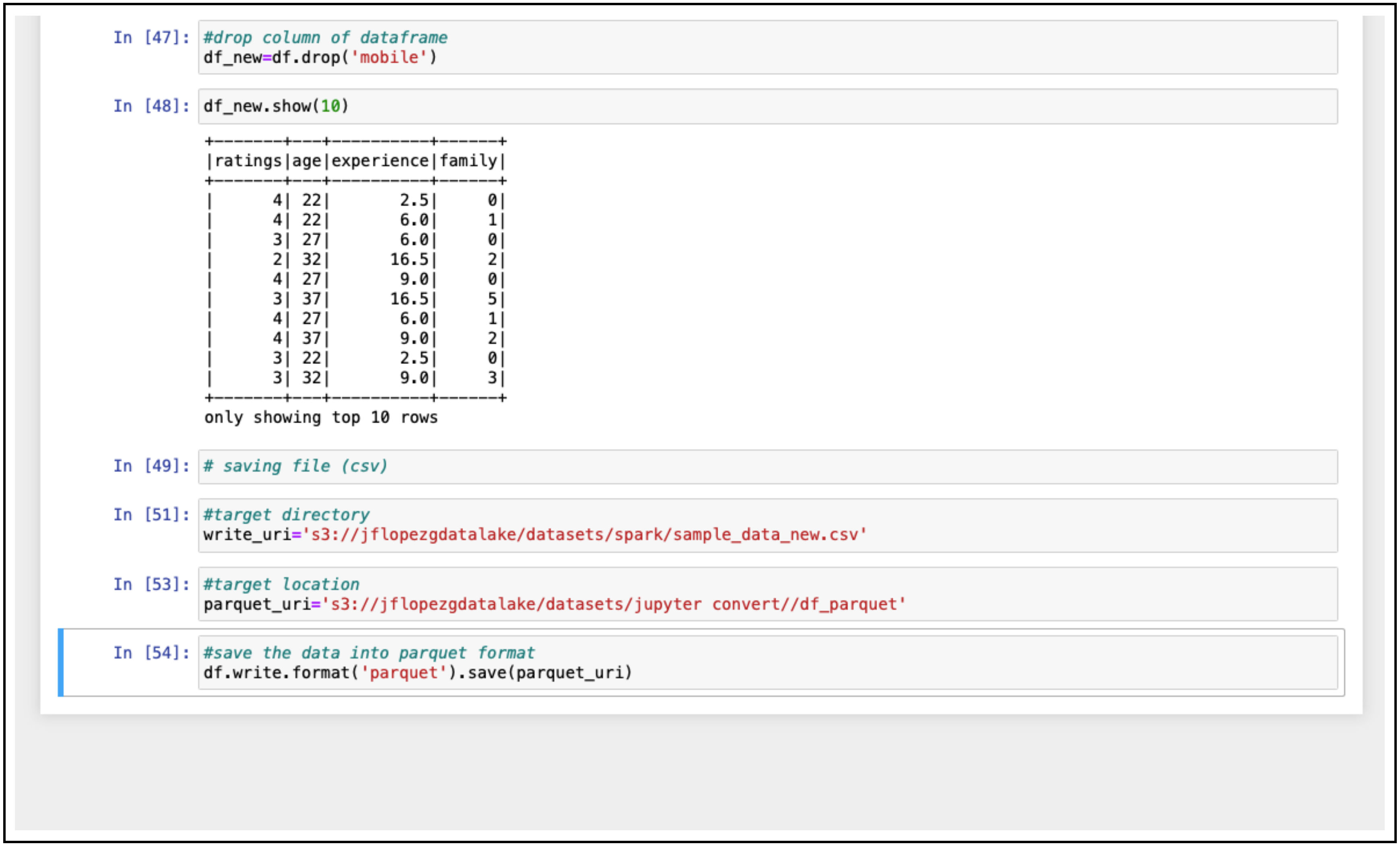
In [44]:
```
#duplicate values
df.count()
```

```
33
```

In [45]:
```
#drop duplicate values
df=df.dropDuplicates()
```

In [46]:
```
#validate new count
df.count()
```

```
26
```



In [47]:
```
#drop column of dataframe
df_new=df.drop('mobile')
```

In [48]:
```
df_new.show(10)
```

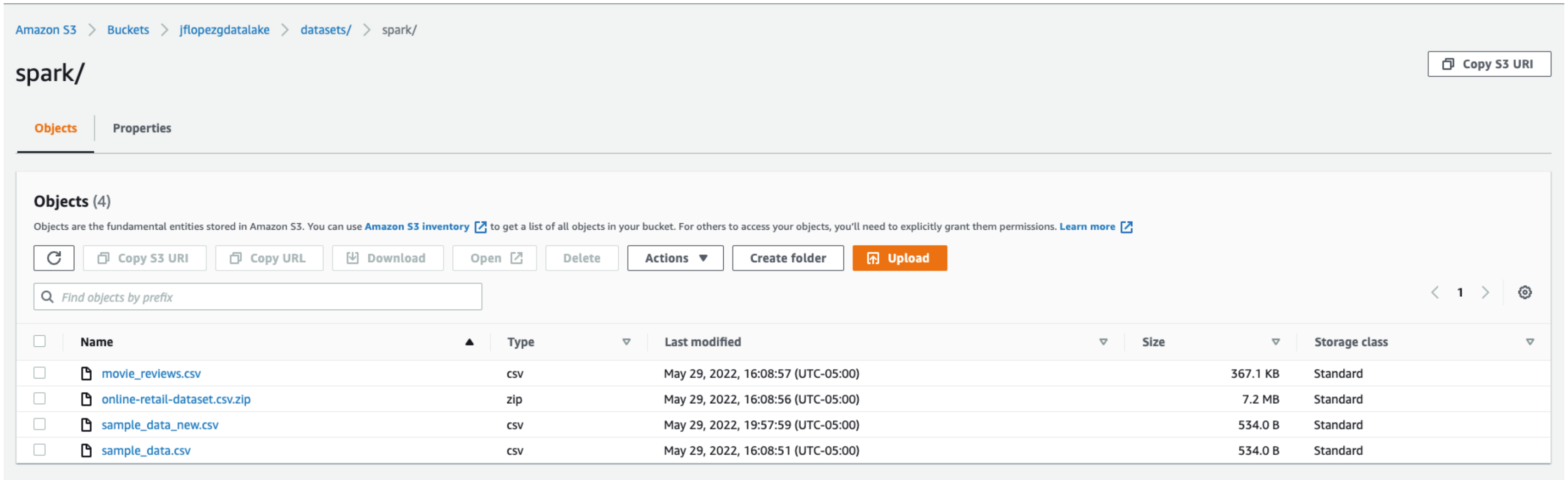
```
+-------+---+----------+------+
|ratings|age|experience|family|
+-------+---+----------+------+
|      4| 22|       2.5|     0|
|      4| 22|       6.0|     1|
|      3| 27|       6.0|     0|
|      2| 32|      16.5|     2|
|      4| 27|       9.0|     0|
|      3| 37|      16.5|     5|
|      4| 27|       6.0|     1|
|      4| 37|       9.0|     2|
|      3| 22|       2.5|     0|
|      3| 32|       9.0|     3|
+-------+---+----------+------+
only showing top 10 rows
```

In [49]:
```
# saving file (csv)
```

In [51]:
```
#target directory
write_uri='s3://jflopezgdatalake/datasets/spark/sample_data_new.csv'
```

In [53]:
```
#target location
parquet_uri='s3://jflopezgdatalake/datasets/jupyter convert//df_parquet'
```

In [54]:
```
#save the data into parquet format
df.write.format('parquet').save(parquet_uri)
```



spark/

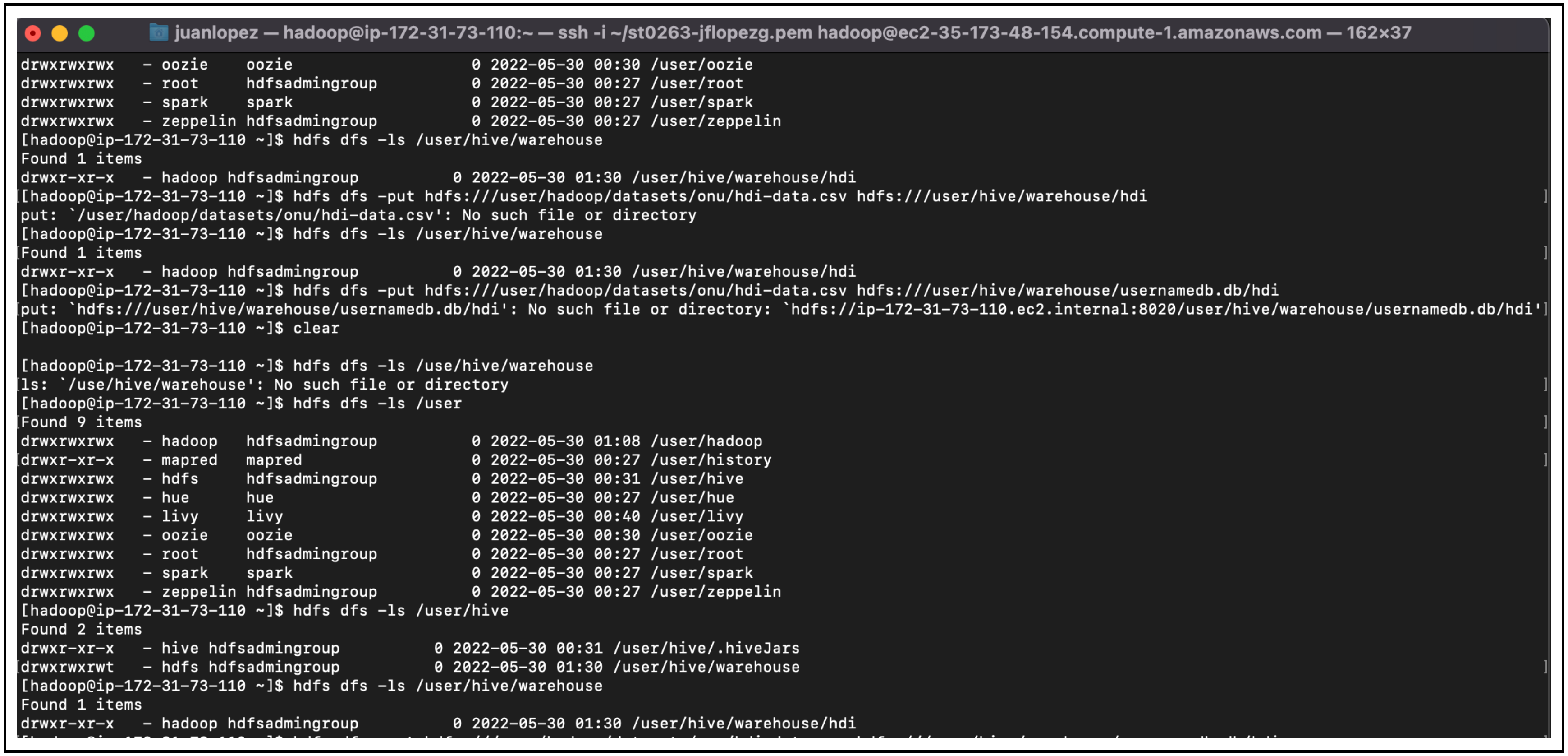
Objects (4)

Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

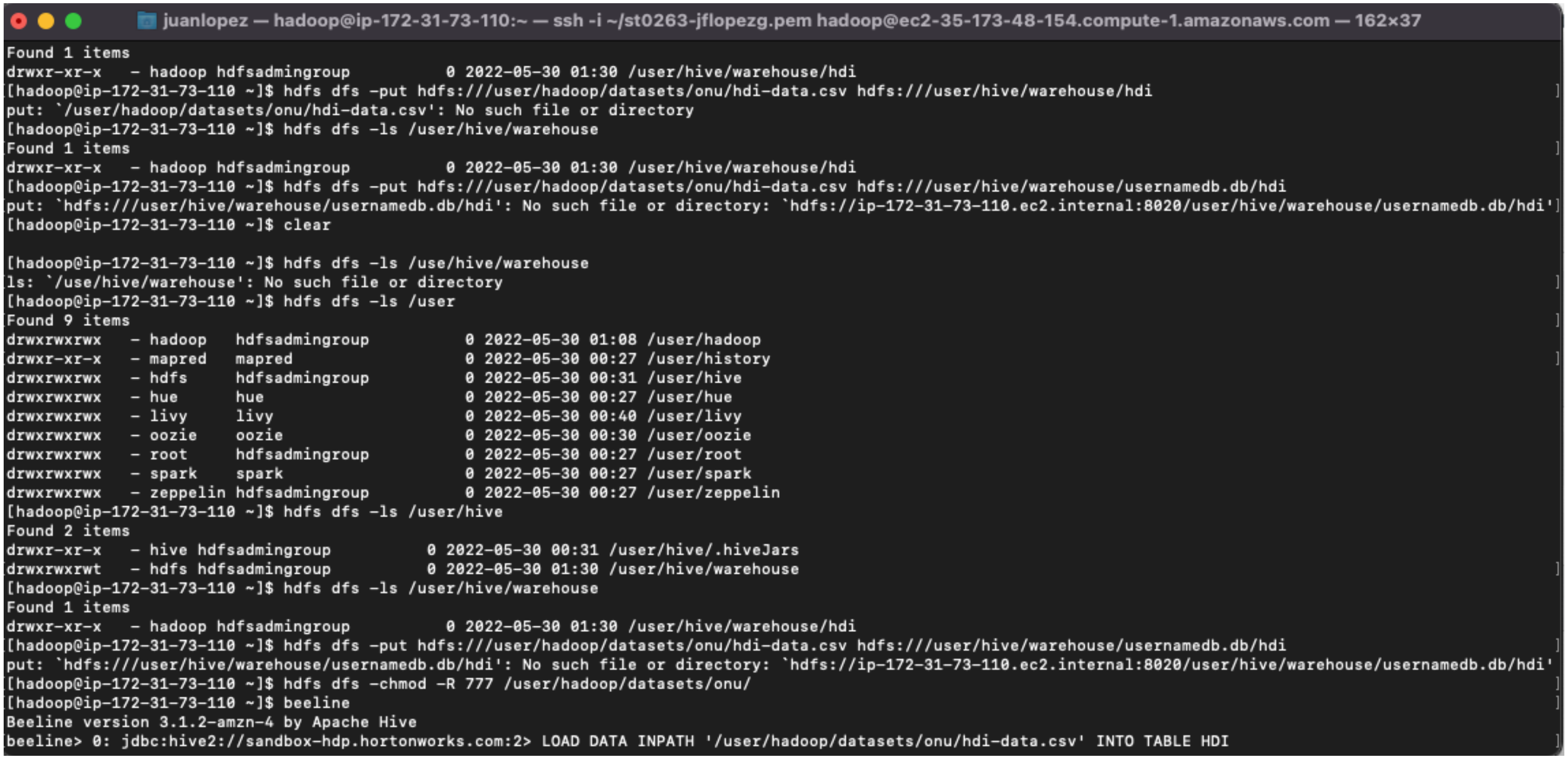
□	Name ▲	Type	Last modified	Size	Storage class
□	movie_reviews.csv	csv	May 29, 2022, 16:08:57 (UTC-05:00)	367.1 KB	Standard
□	online-retail-dataset.csv.zip	zip	May 29, 2022, 16:08:56 (UTC-05:00)	7.2 MB	Standard
□	sample_data_new.csv	csv	May 29, 2022, 19:57:59 (UTC-05:00)	534.0 B	Standard
□	sample_data.csv	csv	May 29, 2022, 16:08:51 (UTC-05:00)	534.0 B	Standard

LAB6-Evidencias de la gestión de tablas (creación y consultas) en Hive para los datos de la ONU en EMR

párrafos descriptivos, Screenshots, códigos fuente, extractos de código, scripts, urls, etc.


juanlopez — hadoop@ip-172-31-73-110:~ — ssh -i ~/st0263-jflopezg.pem hadoop@ec2-35-173-48-154.compute-1.amazonaws.com — 162×37

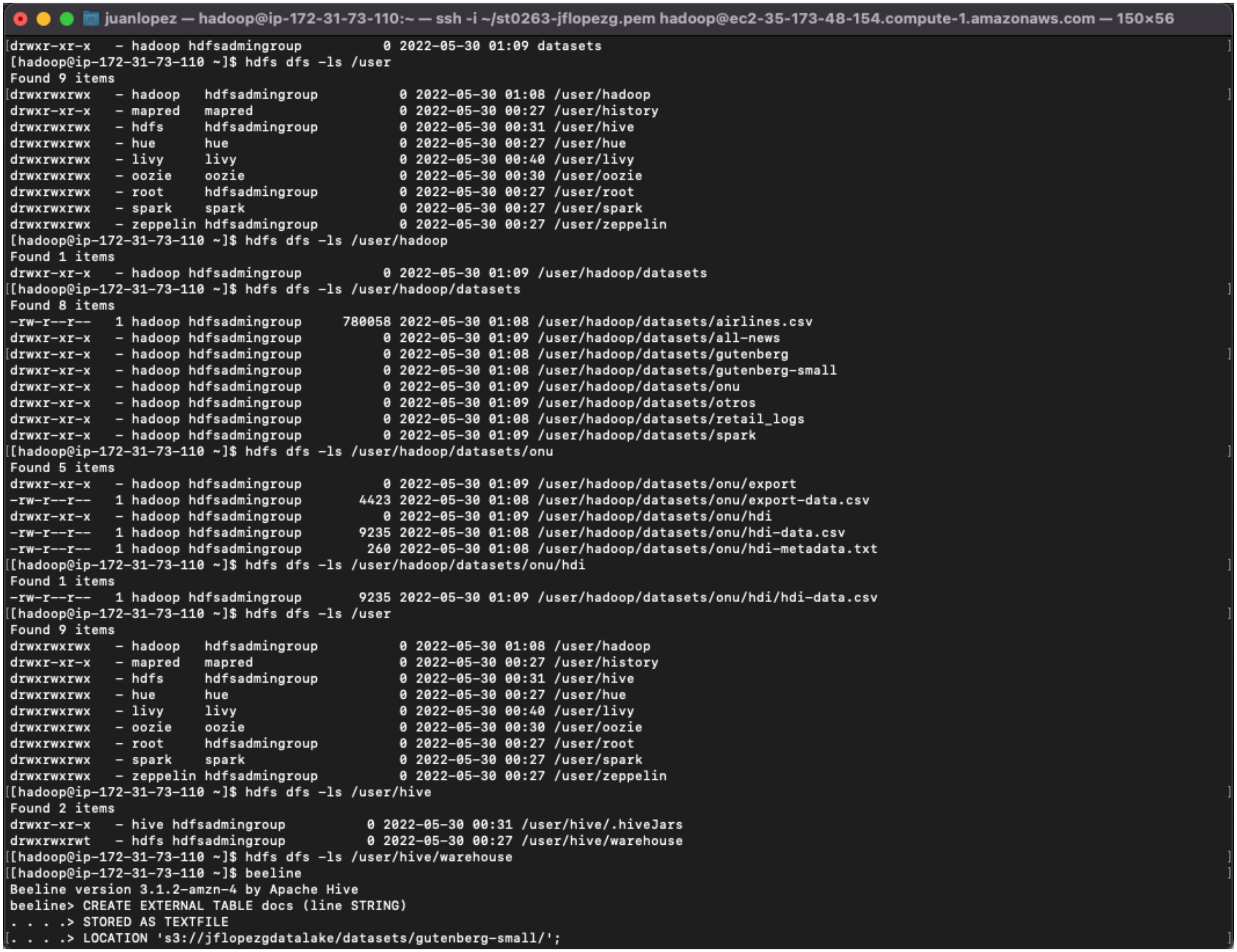
```
drwxrwxrwx   - oozie    oozie                   0 2022-05-30 00:30 /user/oozie
drwxrwxrwx   - root     hdfsadmingroup          0 2022-05-30 00:27 /user/root
drwxrwxrwx   - spark    spark                   0 2022-05-30 00:27 /user/spark
drwxrwxrwx   - zeppelin hdfsadmingroup          0 2022-05-30 00:27 /user/zeppelin
[hadoop@ip-172-31-73-110 ~]$ hdfs dfs -ls /user/hive/warehouse
Found 1 items
drwxr-xr-x   - hadoop hdfsadmingroup          0 2022-05-30 01:30 /user/hive/warehouse/hdi
[hadoop@ip-172-31-73-110 ~]$ hdfs dfs -put hdfs:///user/hadoop/datasets/onu/hdi-data.csv hdfs:///user/hive/warehouse/hdi
put: `/user/hadoop/datasets/onu/hdi-data.csv': No such file or directory
[hadoop@ip-172-31-73-110 ~]$ hdfs dfs -ls /user/hive/warehouse
Found 1 items
drwxr-xr-x   - hadoop hdfsadmingroup          0 2022-05-30 01:30 /user/hive/warehouse/hdi
[hadoop@ip-172-31-73-110 ~]$ hdfs dfs -put hdfs:///user/hadoop/datasets/onu/hdi-data.csv hdfs:///user/hive/warehouse/usernamedb.db/hdi
put: `hdfs:///user/hive/warehouse/usernamedb.db/hdi': No such file or directory: `hdfs://ip-172-31-73-110.ec2.internal:8020/user/hive/warehouse/usernamedb.db/hdi'
[hadoop@ip-172-31-73-110 ~]$ clear

[hadoop@ip-172-31-73-110 ~]$ hdfs dfs -ls /use/hive/warehouse
ls: `/use/hive/warehouse': No such file or directory
[hadoop@ip-172-31-73-110 ~]$ hdfs dfs -ls /user
Found 9 items
drwxrwxrwx   - hadoop   hdfsadmingroup          0 2022-05-30 01:08 /user/hadoop
drwxr-xr-x   - mapred   mapred                  0 2022-05-30 00:27 /user/history
drwxrwxrwx   - hdfs     hdfsadmingroup          0 2022-05-30 00:31 /user/hive
drwxrwxrwx   - hue      hue                     0 2022-05-30 00:27 /user/hue
drwxrwxrwx   - livy     livy                    0 2022-05-30 00:40 /user/livy
drwxrwxrwx   - oozie    oozie                   0 2022-05-30 00:30 /user/oozie
drwxrwxrwx   - root     hdfsadmingroup          0 2022-05-30 00:27 /user/root
drwxrwxrwx   - spark    spark                   0 2022-05-30 00:27 /user/spark
drwxrwxrwx   - zeppelin hdfsadmingroup          0 2022-05-30 00:27 /user/zeppelin
[hadoop@ip-172-31-73-110 ~]$ hdfs dfs -ls /user/hive
Found 2 items
drwxr-xr-x   - hive hdfsadmingroup          0 2022-05-30 00:31 /user/hive/.hiveJars
drwxrwxrwt   - hdfs hdfsadmingroup          0 2022-05-30 01:30 /user/hive/warehouse
[hadoop@ip-172-31-73-110 ~]$ hdfs dfs -ls /user/hive/warehouse
Found 1 items
drwxr-xr-x   - hadoop hdfsadmingroup          0 2022-05-30 01:30 /user/hive/warehouse/hdi
```


juanlopez — hadoop@ip-172-31-73-110:~ — ssh -i ~/st0263-jflopezg.pem hadoop@ec2-35-173-48-154.compute-1.amazonaws.com — 162×37

```
Found 1 items
drwxr-xr-x   - hadoop hdfsadmingroup          0 2022-05-30 01:30 /user/hive/warehouse/hdi
[hadoop@ip-172-31-73-110 ~]$ hdfs dfs -put hdfs:///user/hadoop/datasets/onu/hdi-data.csv hdfs:///user/hive/warehouse/hdi
put: `/user/hadoop/datasets/onu/hdi-data.csv': No such file or directory
[hadoop@ip-172-31-73-110 ~]$ hdfs dfs -ls /user/hive/warehouse
Found 1 items
drwxr-xr-x   - hadoop hdfsadmingroup          0 2022-05-30 01:30 /user/hive/warehouse/hdi
[hadoop@ip-172-31-73-110 ~]$ hdfs dfs -put hdfs:///user/hadoop/datasets/onu/hdi-data.csv hdfs:///user/hive/warehouse/usernamedb.db/hdi
put: `hdfs:///user/hive/warehouse/usernamedb.db/hdi': No such file or directory: `hdfs://ip-172-31-73-110.ec2.internal:8020/user/hive/warehouse/usernamedb.db/hdi'
[hadoop@ip-172-31-73-110 ~]$ clear

[hadoop@ip-172-31-73-110 ~]$ hdfs dfs -ls /use/hive/warehouse
ls: `/use/hive/warehouse': No such file or directory
[hadoop@ip-172-31-73-110 ~]$ hdfs dfs -ls /user
Found 9 items
drwxrwxrwx   - hadoop   hdfsadmingroup          0 2022-05-30 01:08 /user/hadoop
drwxr-xr-x   - mapred   mapred                  0 2022-05-30 00:27 /user/history
drwxrwxrwx   - hdfs     hdfsadmingroup          0 2022-05-30 00:31 /user/hive
drwxrwxrwx   - hue      hue                     0 2022-05-30 00:27 /user/hue
drwxrwxrwx   - livy     livy                    0 2022-05-30 00:40 /user/livy
drwxrwxrwx   - oozie    oozie                   0 2022-05-30 00:30 /user/oozie
drwxrwxrwx   - root     hdfsadmingroup          0 2022-05-30 00:27 /user/root
drwxrwxrwx   - spark    spark                   0 2022-05-30 00:27 /user/spark
drwxrwxrwx   - zeppelin hdfsadmingroup          0 2022-05-30 00:27 /user/zeppelin
[hadoop@ip-172-31-73-110 ~]$ hdfs dfs -ls /user/hive
Found 2 items
drwxr-xr-x   - hive hdfsadmingroup          0 2022-05-30 00:31 /user/hive/.hiveJars
drwxrwxrwt   - hdfs hdfsadmingroup          0 2022-05-30 01:30 /user/hive/warehouse
[hadoop@ip-172-31-73-110 ~]$ hdfs dfs -ls /user/hive/warehouse
Found 1 items
drwxr-xr-x   - hadoop hdfsadmingroup          0 2022-05-30 01:30 /user/hive/warehouse/hdi
[hadoop@ip-172-31-73-110 ~]$ hdfs dfs -put hdfs:///user/hadoop/datasets/onu/hdi-data.csv hdfs:///user/hive/warehouse/usernamedb.db/hdi
put: `hdfs:///user/hive/warehouse/usernamedb.db/hdi': No such file or directory: `hdfs://ip-172-31-73-110.ec2.internal:8020/user/hive/warehouse/usernamedb.db/hdi'
[hadoop@ip-172-31-73-110 ~]$ hdfs dfs -chmod -R 777 /user/hadoop/datasets/onu/
[hadoop@ip-172-31-73-110 ~]$ beeline
Beeline version 3.1.2-amzn-4 by Apache Hive
beeline> 0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> LOAD DATA INPATH '/user/hadoop/datasets/onu/hdi-data.csv' INTO TABLE HDI
```

LAB6-Evidencias del Wordcount en Hive con datos en S3 en ERM

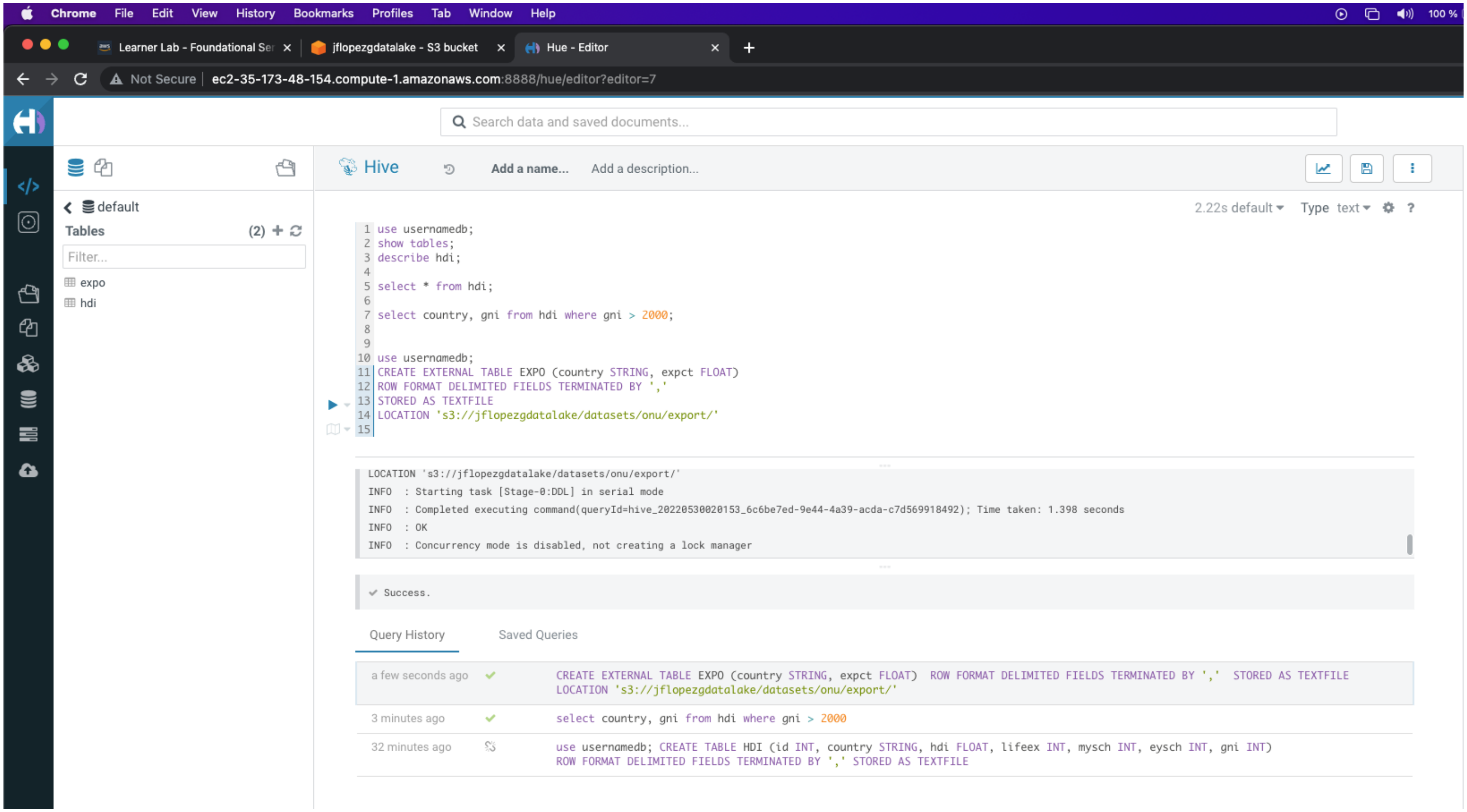
párrafos descriptivos, Screenshots, códigos fuente, extractos de código, scripts, urls, etc.



juanlopez — hadoop@ip-172-31-73-110:~ — ssh -i ~/st0263-jflopezg.pem hadoop@ec2-35-173-48-154.compute-1.amazonaws.com — 150×56

```
[drwxr-xr-x   - hadoop hdfsadmingroup          0 2022-05-30 01:09 datasets
 [hadoop@ip-172-31-73-110 ~]$ hdfs dfs -ls /user
 Found 9 items
[drwxrwxrwx   - hadoop   hdfsadmingroup          0 2022-05-30 01:08 /user/hadoop
 drwxr-xr-x   - mapred   mapred                  0 2022-05-30 00:27 /user/history
 drwxrwxrwx   - hdfs     hdfsadmingroup          0 2022-05-30 00:31 /user/hive
 drwxrwxrwx   - hue      hue                     0 2022-05-30 00:27 /user/hue
 drwxrwxrwx   - livy     livy                    0 2022-05-30 00:40 /user/livy
 drwxrwxrwx   - oozie    oozie                   0 2022-05-30 00:30 /user/oozie
 drwxrwxrwx   - root     hdfsadmingroup          0 2022-05-30 00:27 /user/root
 drwxrwxrwx   - spark    spark                   0 2022-05-30 00:27 /user/spark
 drwxrwxrwx   - zeppelin hdfsadmingroup          0 2022-05-30 00:27 /user/zeppelin
 [hadoop@ip-172-31-73-110 ~]$ hdfs dfs -ls /user/hadoop
 Found 1 items
 drwxr-xr-x   - hadoop hdfsadmingroup          0 2022-05-30 01:09 /user/hadoop/datasets
[[hadoop@ip-172-31-73-110 ~]$ hdfs dfs -ls /user/hadoop/datasets
 Found 8 items
 -rw-r--r--   1 hadoop hdfsadmingroup     780058 2022-05-30 01:08 /user/hadoop/datasets/airlines.csv
 drwxr-xr-x   - hadoop hdfsadmingroup          0 2022-05-30 01:09 /user/hadoop/datasets/all-news
[drwxr-xr-x   - hadoop hdfsadmingroup          0 2022-05-30 01:08 /user/hadoop/datasets/gutenberg
 drwxr-xr-x   - hadoop hdfsadmingroup          0 2022-05-30 01:08 /user/hadoop/datasets/gutenberg-small
 drwxr-xr-x   - hadoop hdfsadmingroup          0 2022-05-30 01:09 /user/hadoop/datasets/onu
 drwxr-xr-x   - hadoop hdfsadmingroup          0 2022-05-30 01:09 /user/hadoop/datasets/otros
 drwxr-xr-x   - hadoop hdfsadmingroup          0 2022-05-30 01:08 /user/hadoop/datasets/retail_logs
 drwxr-xr-x   - hadoop hdfsadmingroup          0 2022-05-30 01:09 /user/hadoop/datasets/spark
[[hadoop@ip-172-31-73-110 ~]$ hdfs dfs -ls /user/hadoop/datasets/onu
 Found 5 items
 drwxr-xr-x   - hadoop hdfsadmingroup          0 2022-05-30 01:09 /user/hadoop/datasets/onu/export
 -rw-r--r--   1 hadoop hdfsadmingroup       4423 2022-05-30 01:08 /user/hadoop/datasets/onu/export-data.csv
 drwxr-xr-x   - hadoop hdfsadmingroup          0 2022-05-30 01:09 /user/hadoop/datasets/onu/hdi
 -rw-r--r--   1 hadoop hdfsadmingroup       9235 2022-05-30 01:08 /user/hadoop/datasets/onu/hdi-data.csv
 -rw-r--r--   1 hadoop hdfsadmingroup        260 2022-05-30 01:08 /user/hadoop/datasets/onu/hdi-metadata.txt
[[hadoop@ip-172-31-73-110 ~]$ hdfs dfs -ls /user/hadoop/datasets/onu/hdi
 Found 1 items
 -rw-r--r--   1 hadoop hdfsadmingroup       9235 2022-05-30 01:09 /user/hadoop/datasets/onu/hdi/hdi-data.csv
[[hadoop@ip-172-31-73-110 ~]$ hdfs dfs -ls /user
 Found 9 items
 drwxrwxrwx   - hadoop   hdfsadmingroup          0 2022-05-30 01:08 /user/hadoop
 drwxr-xr-x   - mapred   mapred                  0 2022-05-30 00:27 /user/history
 drwxrwxrwx   - hdfs     hdfsadmingroup          0 2022-05-30 00:31 /user/hive
 drwxrwxrwx   - hue      hue                     0 2022-05-30 00:27 /user/hue
 drwxrwxrwx   - livy     livy                    0 2022-05-30 00:40 /user/livy
 drwxrwxrwx   - oozie    oozie                   0 2022-05-30 00:30 /user/oozie
 drwxrwxrwx   - root     hdfsadmingroup          0 2022-05-30 00:27 /user/root
 drwxrwxrwx   - spark    spark                   0 2022-05-30 00:27 /user/spark
 drwxrwxrwx   - zeppelin hdfsadmingroup          0 2022-05-30 00:27 /user/zeppelin
[[hadoop@ip-172-31-73-110 ~]$ hdfs dfs -ls /user/hive
 Found 2 items
 drwxr-xr-x   - hive hdfsadmingroup          0 2022-05-30 00:31 /user/hive/.hiveJars
 drwxrwxrwt   - hdfs hdfsadmingroup          0 2022-05-30 00:27 /user/hive/warehouse
[[hadoop@ip-172-31-73-110 ~]$ hdfs dfs -ls /user/hive/warehouse
[[hadoop@ip-172-31-73-110 ~]$ beeline
 Beeline version 3.1.2-amzn-4 by Apache Hive
 beeline> CREATE EXTERNAL TABLE docs (line STRING)
 . . . .> STORED AS TEXTFILE
[. . . .> LOCATION 's3://jflopezgdatalake/datasets/gutenberg-small/';
```

LAB6-Evidencias de la consulta de tablas Hive, desde SparkSQL en un jupyter notebook en EMR con datos en S3 párrafos descriptivos, Screenshots, códigos fuente, extractos de código, scripts, urls, etc.


Chrome File Edit View History Bookmarks Profiles Tab Window Help
100 %
Learner Lab - Foundational Se
jflopezgdatalake - S3 bucket
Hue - Editor
Not Secure | ec2-35-173-48-154.compute-1.amazonaws.com:8888/hue/editor?editor=7
Search data and saved documents...
default
Tables (2)
Filter...
expo
hdi
Hive
Add a name...
Add a description...
2.22s default
Type text

```
use usernamedb;
show tables;
describe hdi;

select * from hdi;

select country, gni from hdi where gni > 2000;


use usernamedb;
CREATE EXTERNAL TABLE EXPO (country STRING, expct FLOAT)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
STORED AS TEXTFILE
LOCATION 's3://jflopezgdatalake/datasets/onu/export/'
```

```
LOCATION 's3://jflopezgdatalake/datasets/onu/export/'
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20220530020153_6c6be7ed-9e44-4a39-acda-c7d569918492); Time taken: 1.398 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
```

✓ Success.

```
a few seconds ago  ✓   CREATE EXTERNAL TABLE EXPO (country STRING, expct FLOAT)  ROW FORMAT DELIMITED FIELDS TERMINATED BY ','  STORED AS TEXTFILE
                       LOCATION 's3://jflopezgdatalake/datasets/onu/export/'
3 minutes ago      ✓   select country, gni from hdi where gni > 2000
32 minutes ago         use usernamedb; CREATE TABLE HDI (id INT, country STRING, hdi FLOAT, lifeex INT, mysch INT, eysch INT, gni INT)
                       ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE
```

LAB6-evidencias de algo no considerado en las anteriores evidencias y requeridas en el LAB

párrafos descriptivos, Screenshots, códigos fuente, extractos de código, scripts, urls, etc.