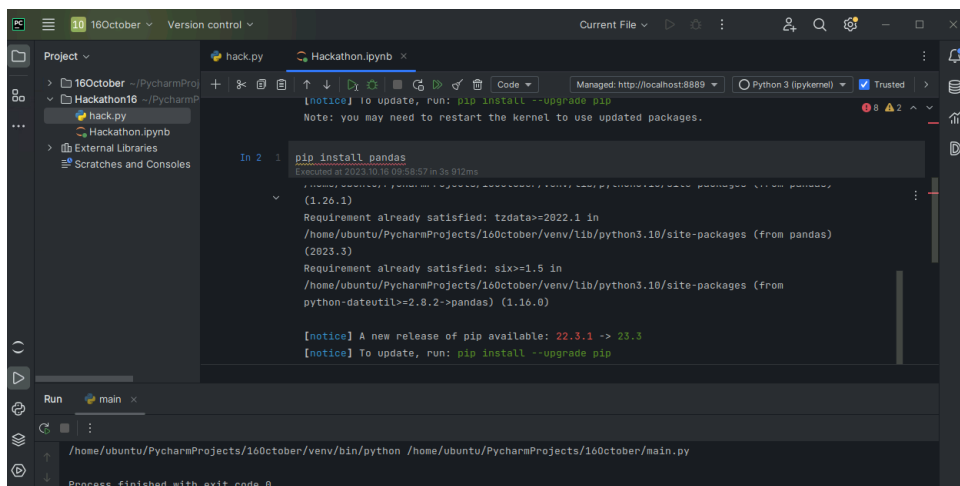
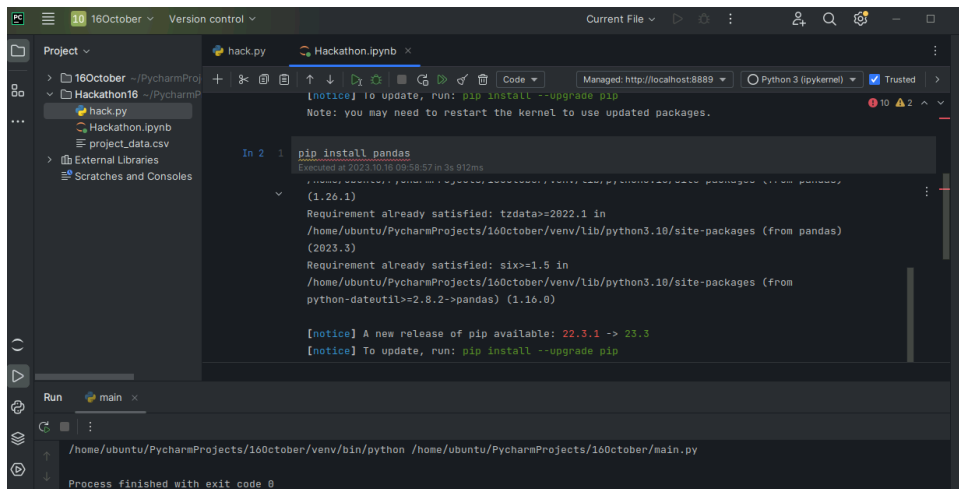


1. PIP install pyspark

2. pip install pandas



3. upload project_data.csv



The screenshot shows the PyCharm IDE interface. The left sidebar displays the project structure for '16October', including files like 'hack.py', 'Hackathon.ipynb', 'project_data.csv', and 'External Libraries'. The main editor window shows a Jupyter Notebook cell with the command `pip install pandas`. The output of the cell shows that pandas is already installed (version 1.26.1) and that the requirements are satisfied. A notice at the bottom indicates that a new release of pip is available (22.3.1 -> 23.3) and suggests running `pip install --upgrade pip` to update it. The bottom status bar shows 'Process finished with exit code 0'.

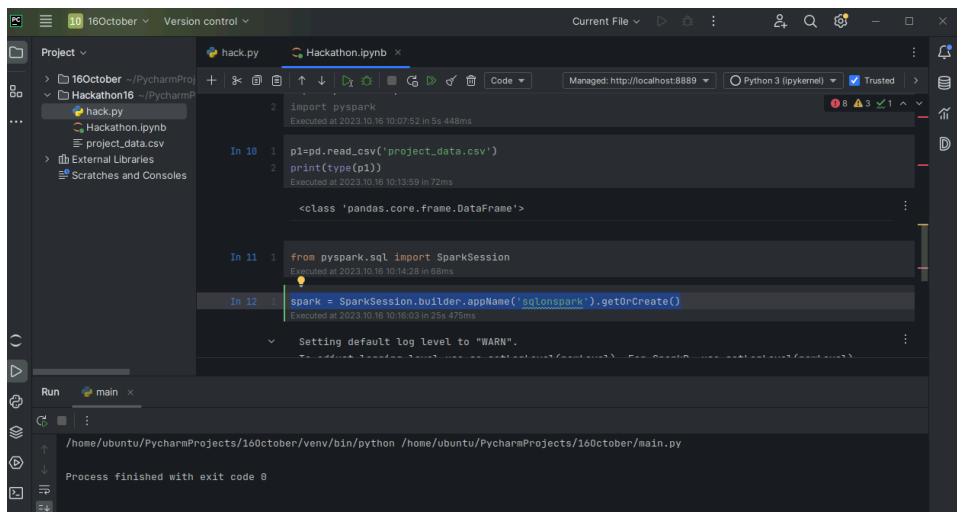
```
p1=pd.read_csv('project_data.csv')
```

```
print(type(p1))
```

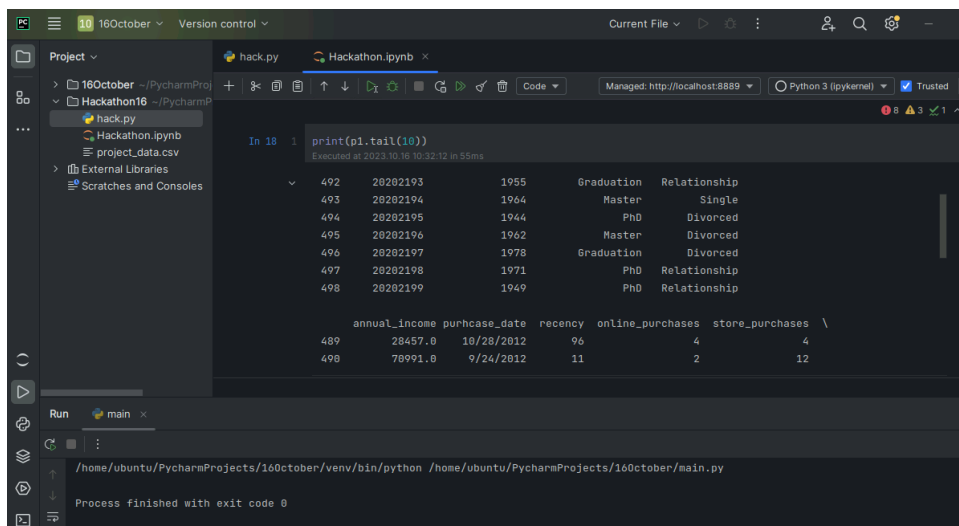
```
print(p1.tail(10))
```

```
from pyspark.sql import SparkSession
```

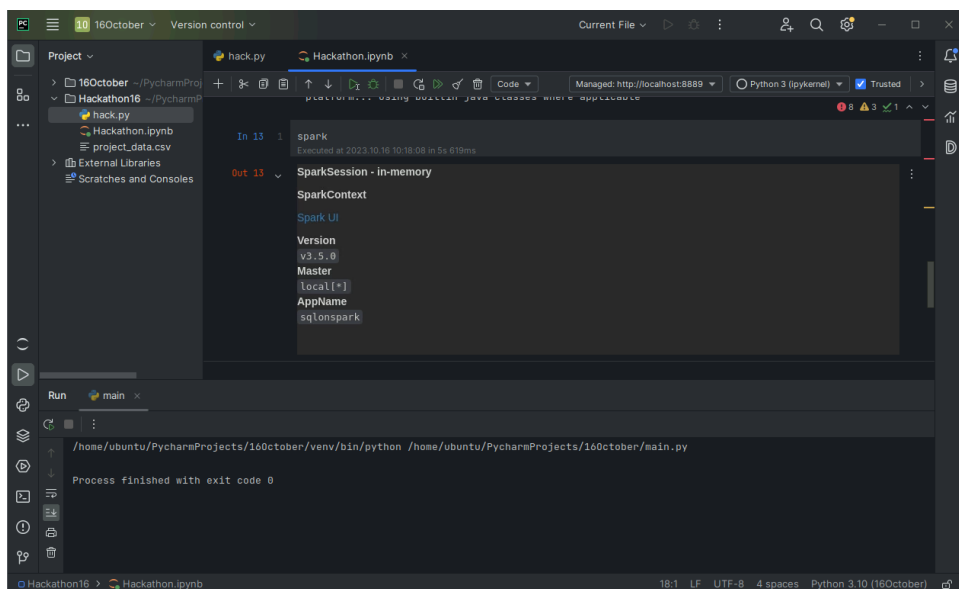
```
spark = SparkSession.builder.appName('sqlonspark').getOrCreate()
```



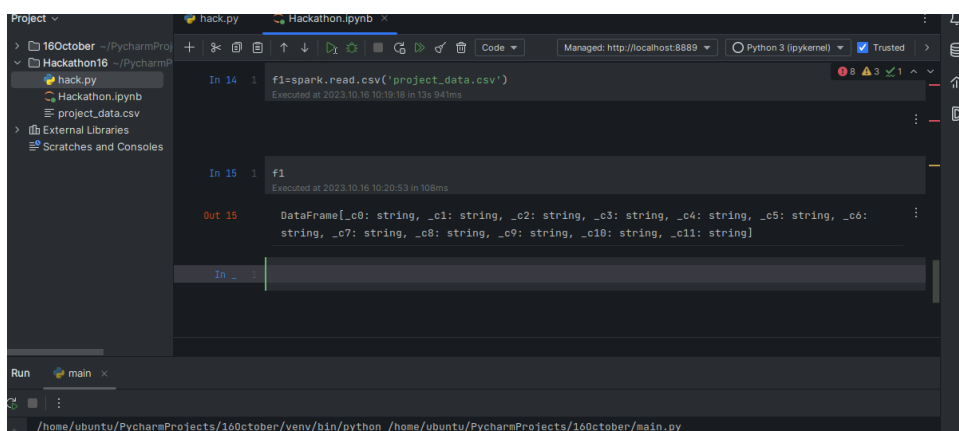
The screenshot shows the PyCharm IDE interface with a Jupyter Notebook. The left sidebar shows the project structure. The main editor window displays several code cells. The first cell contains `import pyspark`. The second cell contains `p1=pd.read_csv('project_data.csv')` and `print(type(p1))`, with the output showing `<class 'pandas.core.frame.DataFrame'>`. The third cell contains `from pyspark.sql import SparkSession`. The fourth cell contains `spark = SparkSession.builder.appName('sqlonspark').getOrCreate()`. The bottom status bar shows 'Process finished with exit code 0'.



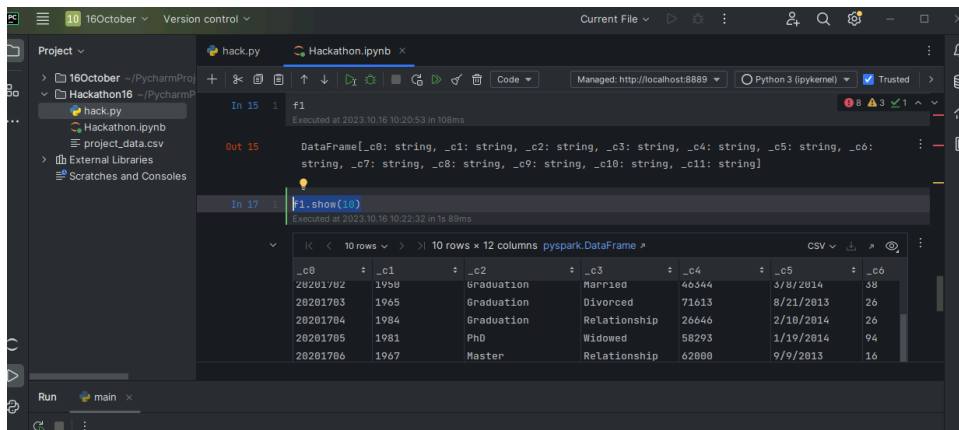
Spark



F1



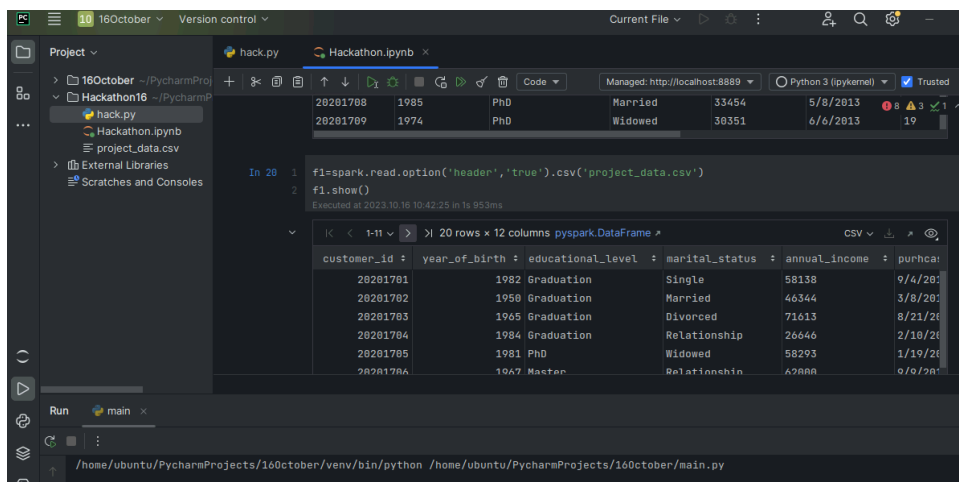
f1.show(10)



To create data frame.

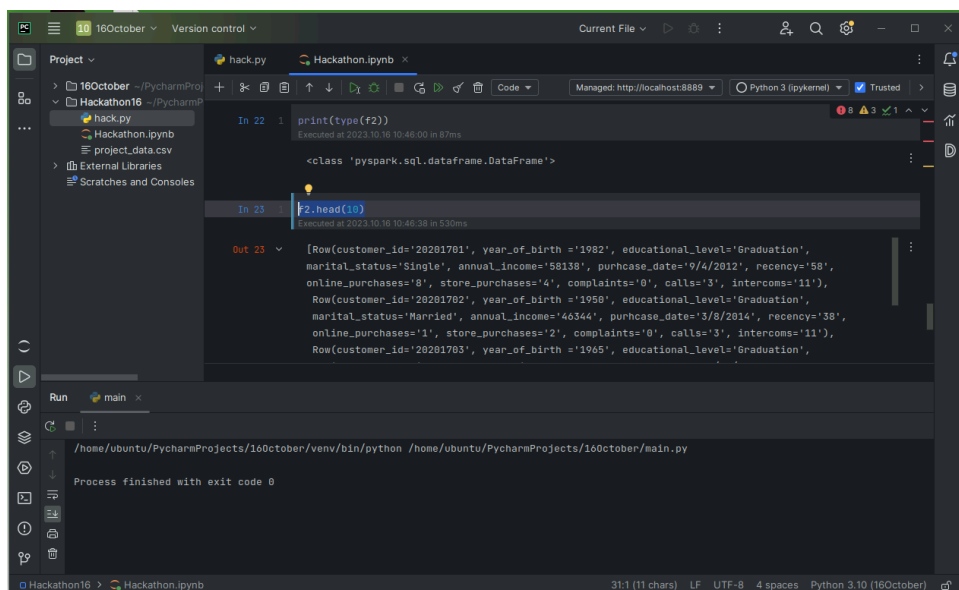
```
f2=spark.read.option('header','true').csv('project_data.csv')
```

```
F2.show()
```



```
print(type(f2))
```

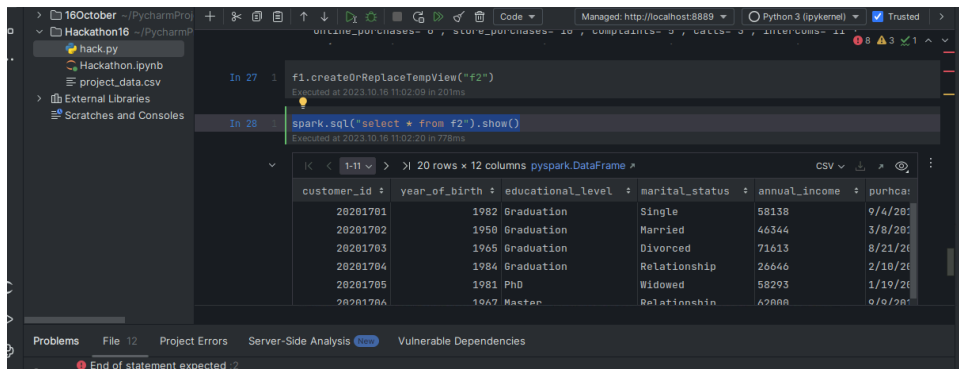
```
f2.head(10)
```



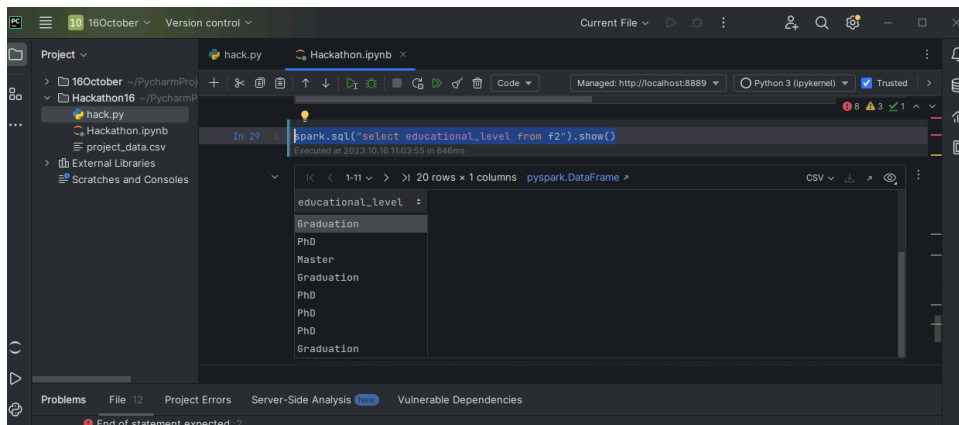
Spark sql operation

```
f1.createOrReplaceTempView("f2")
```

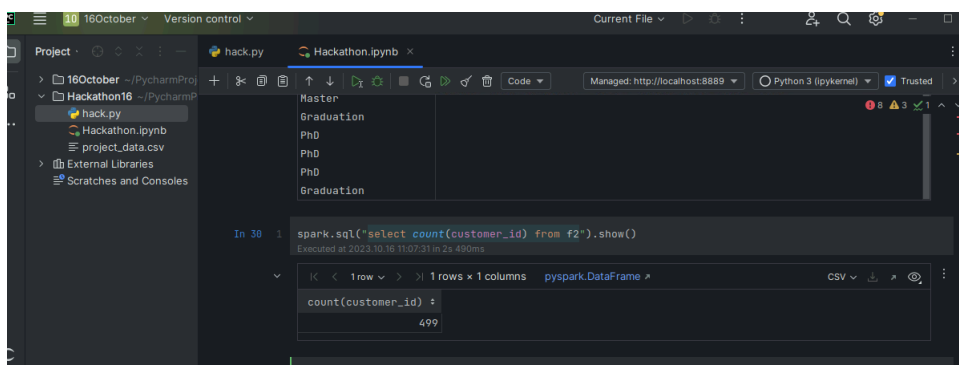
```
spark.sql("select * from f2").show()
```



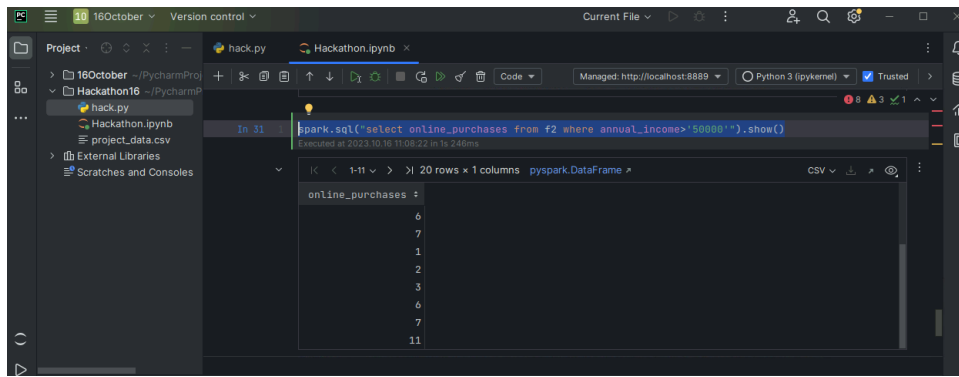
```
spark.sql("select educational_level from f2").show()
```



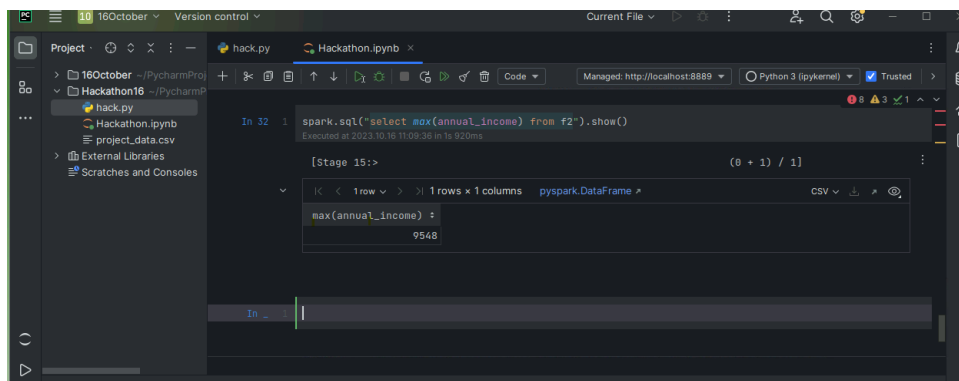
```
spark.sql("select count(customer_id) from f2").show()
```



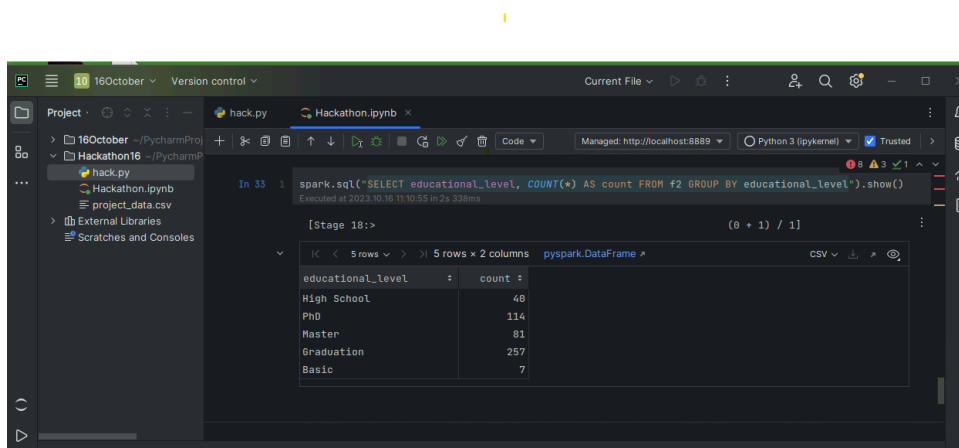
```
spark.sql("select online_purchases from f2 where annual_income>'50000']").show()
```



`spark.sql("select max(annual_income) from f2").show()`



`spark.sql("SELECT educational_level, COUNT(*) AS count FROM f2 GROUP BY educational_level").show()`



Part 2 with other csv.file

Image_1000.csv

```
Project
  16October ~/PycharmProj
  Hackathon16 ~/PycharmProj
    spark-warehouse
      hack.py
      Hackathon.ipynb
      mdb_1000.csv
      project_data.csv
    External Libraries
    Scratches and Consoles

Run
hack

/home/ubuntu/PycharmProjects/16October/venv/bin/python /home/ubuntu/PycharmProjects/Hackathon16/hack.py
Streaming ingestion is started....
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
23/10/16 11:53:12 WARN NativeCodeLoader: Unable to load native-heapo library for your platform... using builtin-java classes where applicable
23/10/16 11:53:14 WARN Utils: Service 'SparkUI' could not bind on port 4040. Attempting port 4041.
23/10/16 11:53:17 WARN SparkSession: Using an existing Spark session; only runtime SQL configurations will take effect.
23/10/16 11:53:17 WARN TextSocketSourceProvider: The socket source should not be used for production applications! It does not support recovery
True
root
|-- value: string (nullable = true)

Process finished with exit code 0
```

Stackroute :: Subscription Details | Stackroute | NatWest - Google Docs | i-007eeb9304s782f0

wai.vlabs.stackroute.in/guacamole/#/client/s50wMDdZW5M2Q0YTM3ODJmMABjAGS1dmVsaW5r?hostname=172.31.7.231&pro...

ACID properties of T... spreadsheet DummyAPI - User... Office.com cheatSheet Digital Ocean HTML Code for Reg... Ready to check - N...

Menu

sqlonspark - Spark Jobs - Google Chrome

Document shared with you x sqlonspark - Spark Jobs x +

127.0.0.1:4040/jobs/

Spark 3.5.0 Jobs Stages Storage Environment Executors SQL / DataFrame sqlonspark application UI

Spark Jobs (?)

User: ubuntu
Total Uptime: 1.6 h
Scheduling Mode: FIFO
Completed Jobs: 18

Event Timeline

Completed Jobs (18)

Page: 1 1 Pages. Jump to 1 . Show 100 items in a page. Go

Job Id	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
17	showString at NativeMethodAccessorImpl.java:0 showString at NativeMethodAccessorImpl.java:0	2023/10/16 11:10:55	0.2 s	1/1 (1 skipped)	1/1 (1 skipped)
16	showString at NativeMethodAccessorImpl.java:0 showString at NativeMethodAccessorImpl.java:0	2023/10/16 11:10:54	0.7 s	1/1	1/1
15	showString at NativeMethodAccessorImpl.java:0 showString at NativeMethodAccessorImpl.java:0	2023/10/16 11:09:35	0.4 s	1/1 (1 skipped)	1/1 (1 skipped)
14	showString at NativeMethodAccessorImpl.java:0 showString at NativeMethodAccessorImpl.java:0	2023/10/16 11:09:34	0.9 s	1/1	1/1
13	showString at NativeMethodAccessorImpl.java:0 showString at NativeMethodAccessorImpl.java:0	2023/10/16 11:08:21	0.6 s	1/1	1/1
12	showString at NativeMethodAccessorImpl.java:0 showString at NativeMethodAccessorImpl.java:0	2023/10/16 11:07:30	0.3 s	1/1 (1 skipped)	1/1 (1 skipped)
11	showString at NativeMethodAccessorImpl.java:0	2023/10/16 11:07:30	0.5 s	1/1	1/1

ubuntu@ip-172-31-7-... [16October - hack.py ... sqlonspark - Spark Jo... [Hackathon16]

Search ENG IN 17:26 16-10-2023

