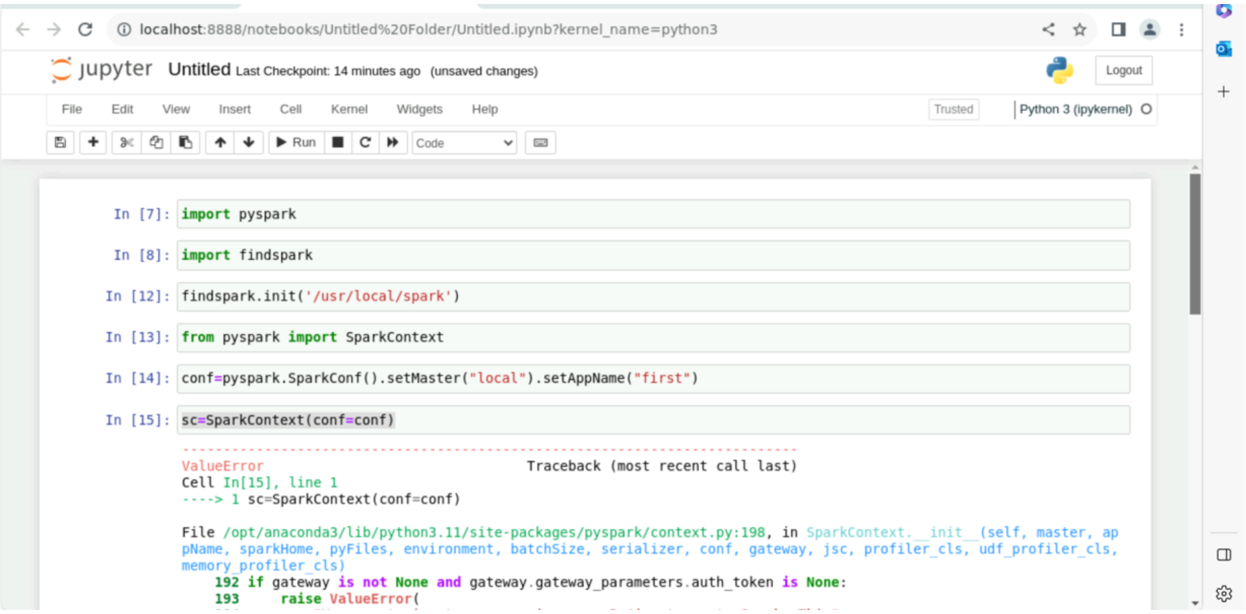
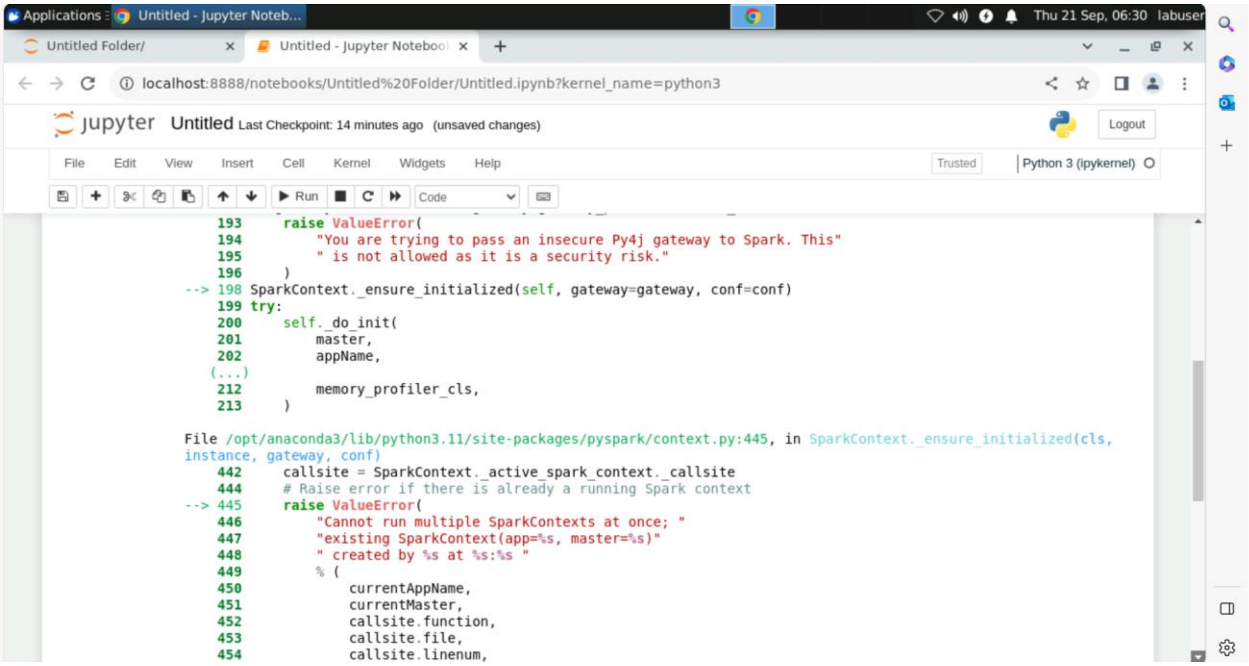
Pyspark



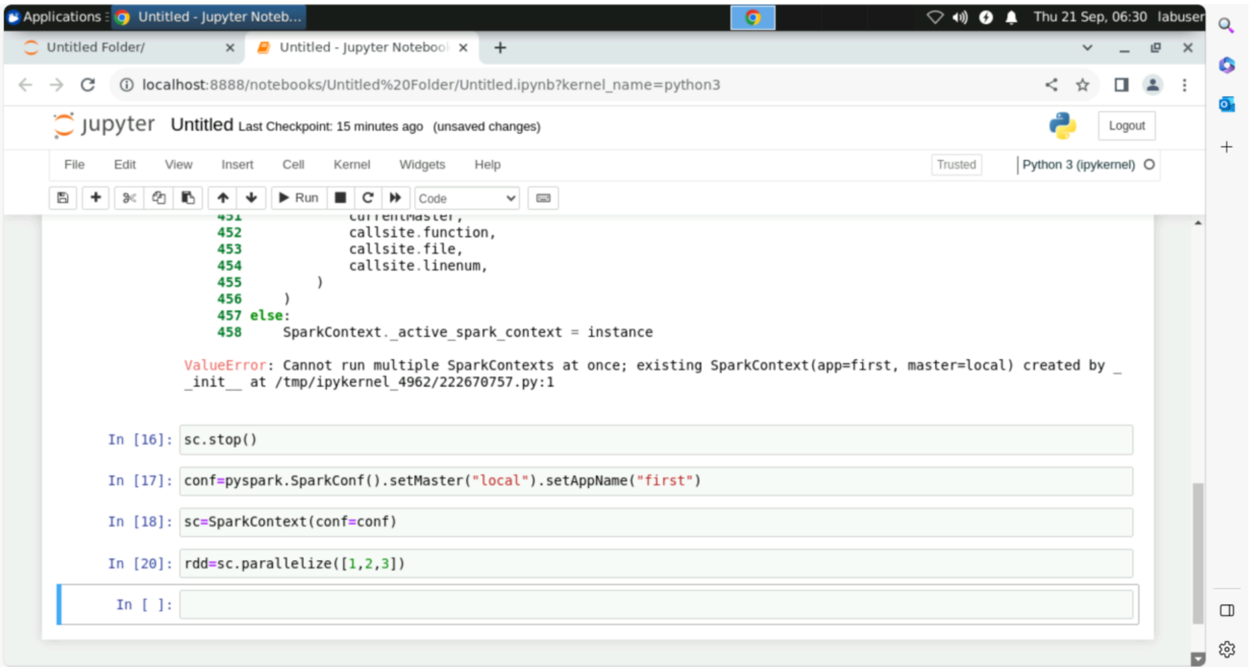
localhost:8888/notebooks/Untitled%20Folder/Untitled.ipynb?kernel_name=python3
jupyter Untitled Last Checkpoint: 14 minutes ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code

```
In [7]: import pyspark

In [8]: import findspark

In [12]: findspark.init('/usr/local/spark')

In [13]: from pyspark import SparkContext

In [14]: conf=pyspark.SparkConf().setMaster("local").setAppName("first")

In [15]: sc=SparkContext(conf=conf)

---------------------------------------------------------------------------
ValueError                                Traceback (most recent call last)
Cell In[15], line 1
----> 1 sc=SparkContext(conf=conf)

File /opt/anaconda3/lib/python3.11/site-packages/pyspark/context.py:198, in SparkContext.__init__(self, master, ap
pName, sparkHome, pyFiles, environment, batchSize, serializer, conf, gateway, jsc, profiler_cls, udf_profiler_cls,
memory_profiler_cls)
    192 if gateway is not None and gateway.gateway_parameters.auth_token is None:
    193     raise ValueError(
```



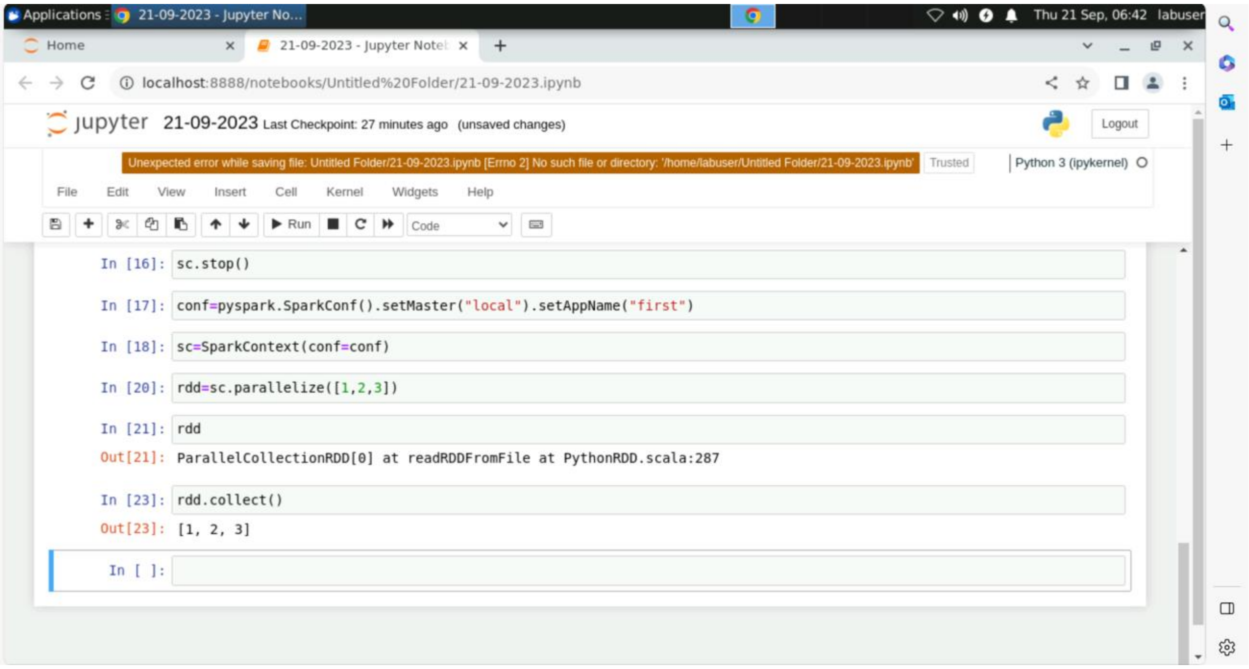
Applications
Untitled - Jupyter Noteb...
Thu 21 Sep, 06:30 labuser
Untitled Folder/
Untitled - Jupyter Noteboo
localhost:8888/notebooks/Untitled%20Folder/Untitled.ipynb?kernel_name=python3
jupyter Untitled Last Checkpoint: 14 minutes ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code

```
    193     raise ValueError(
    194         "You are trying to pass an insecure Py4j gateway to Spark. This"
    195         " is not allowed as it is a security risk."
    196     )
--> 198 SparkContext._ensure_initialized(self, gateway=gateway, conf=conf)
    199 try:
    200     self._do_init(
    201         master,
    202         appName,
   (...)
    212         memory_profiler_cls,
    213     )

File /opt/anaconda3/lib/python3.11/site-packages/pyspark/context.py:445, in SparkContext._ensure_initialized(cls,
instance, gateway, conf)
    442     callsite = SparkContext._active_spark_context._callsite
    444     # Raise error if there is already a running Spark context
--> 445     raise ValueError(
    446         "Cannot run multiple SparkContexts at once; "
    447         "existing SparkContext(app=%s, master=%s)"
    448         " created by %s at %s:%s "
    449         % (
    450             currentAppName,
    451             currentMaster,
    452             callsite.function,
    453             callsite.file,
    454             callsite.linenum,
```


Applications
Untitled - Jupyter Noteb...
Thu 21 Sep, 06:30 labuser
Untitled Folder/
Untitled - Jupyter Noteboo
localhost:8888/notebooks/Untitled%20Folder/Untitled.ipynb?kernel_name=python3
jupyter Untitled Last Checkpoint: 15 minutes ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code

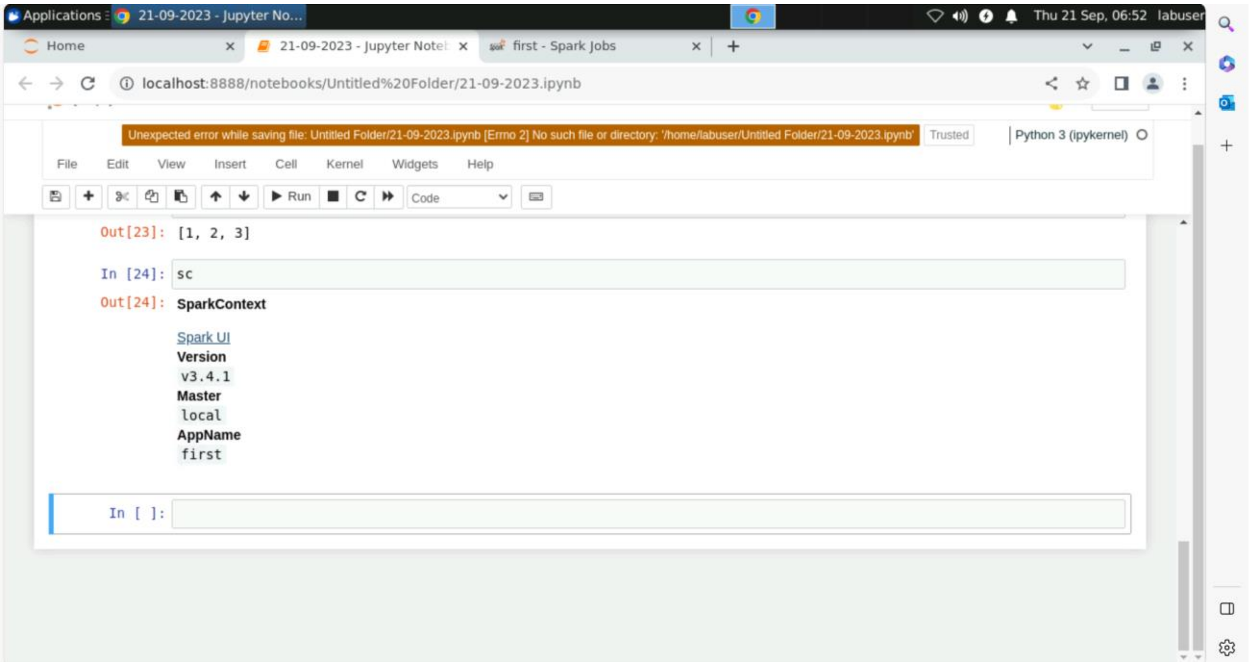
```
    451            currentMaster,
    452            callsite.function,
    453            callsite.file,
    454            callsite.linenum,
    455        )
    456    )
    457 else:
    458    SparkContext._active_spark_context = instance

ValueError: Cannot run multiple SparkContexts at once; existing SparkContext(app=first, master=local) created by _
_init__ at /tmp/ipykernel_4962/222670757.py:1
```

```
In [16]: sc.stop()

In [17]: conf=pyspark.SparkConf().setMaster("local").setAppName("first")

In [18]: sc=SparkContext(conf=conf)

In [20]: rdd=sc.parallelize([1,2,3])

In [ ]:
```


Applications
21-09-2023 - Jupyter No...
Thu 21 Sep, 06:42 labuser
Home
21-09-2023 - Jupyter Noteb
localhost:8888/notebooks/Untitled%20Folder/21-09-2023.ipynb
jupyter 21-09-2023 Last Checkpoint: 27 minutes ago (unsaved changes)
Logout
Trusted
Python 3 (ipykernel)
File Edit View Insert Cell Kernel Widgets Help
Run
Code

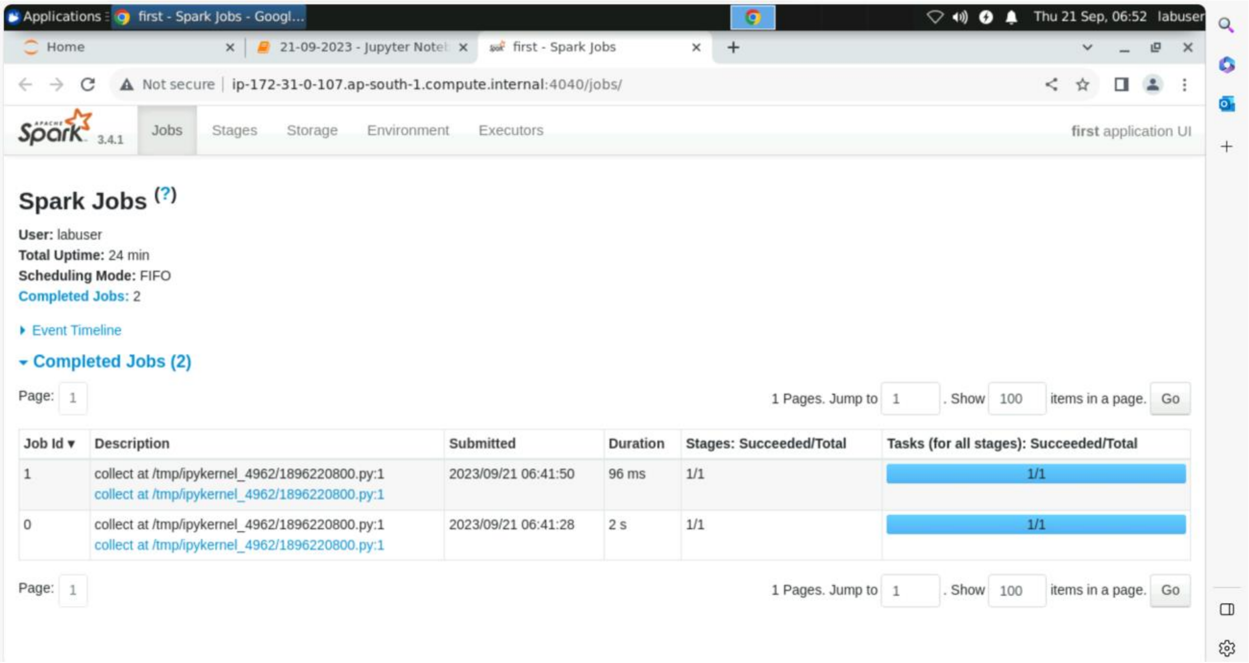
Unexpected error while saving file: Untitled Folder/21-09-2023.ipynb [Errno 2] No such file or directory: '/home/labuser/Untitled Folder/21-09-2023.ipynb'

```
In [16]: sc.stop()

In [17]: conf=pyspark.SparkConf().setMaster("local").setAppName("first")

In [18]: sc=SparkContext(conf=conf)

In [20]: rdd=sc.parallelize([1,2,3])

In [21]: rdd
Out[21]: ParallelCollectionRDD[0] at readRDDFromFile at PythonRDD.scala:287

In [23]: rdd.collect()
Out[23]: [1, 2, 3]

In [ ]:
```


Applications
21-09-2023 - Jupyter No...
Thu 21 Sep, 06:52 labuser
Home
21-09-2023 - Jupyter Noteb
first - Spark Jobs
localhost:8888/notebooks/Untitled%20Folder/21-09-2023.ipynb
Trusted
Python 3 (ipykernel)
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Run
Code

Unexpected error while saving file: Untitled Folder/21-09-2023.ipynb [Errno 2] No such file or directory: '/home/labuser/Untitled Folder/21-09-2023.ipynb'

```
Out[23]: [1, 2, 3]

In [24]: sc

Out[24]: SparkContext

         Spark UI
         Version
         v3.4.1
         Master
         local
         AppName
         first

In [ ]:
```


Applications
first - Spark Jobs - Googl...
Thu 21 Sep, 06:52 labuser
Home
21-09-2023 - Jupyter Noteb
first - Spark Jobs
Not secure
ip-172-31-0-107.ap-south-1.compute.internal:4040/jobs/
Spark 3.4.1
Jobs
Stages
Storage
Environment
Executors
first application UI

Spark Jobs (?)

User: labuser
Total Uptime: 24 min
Scheduling Mode: FIFO
Completed Jobs: 2

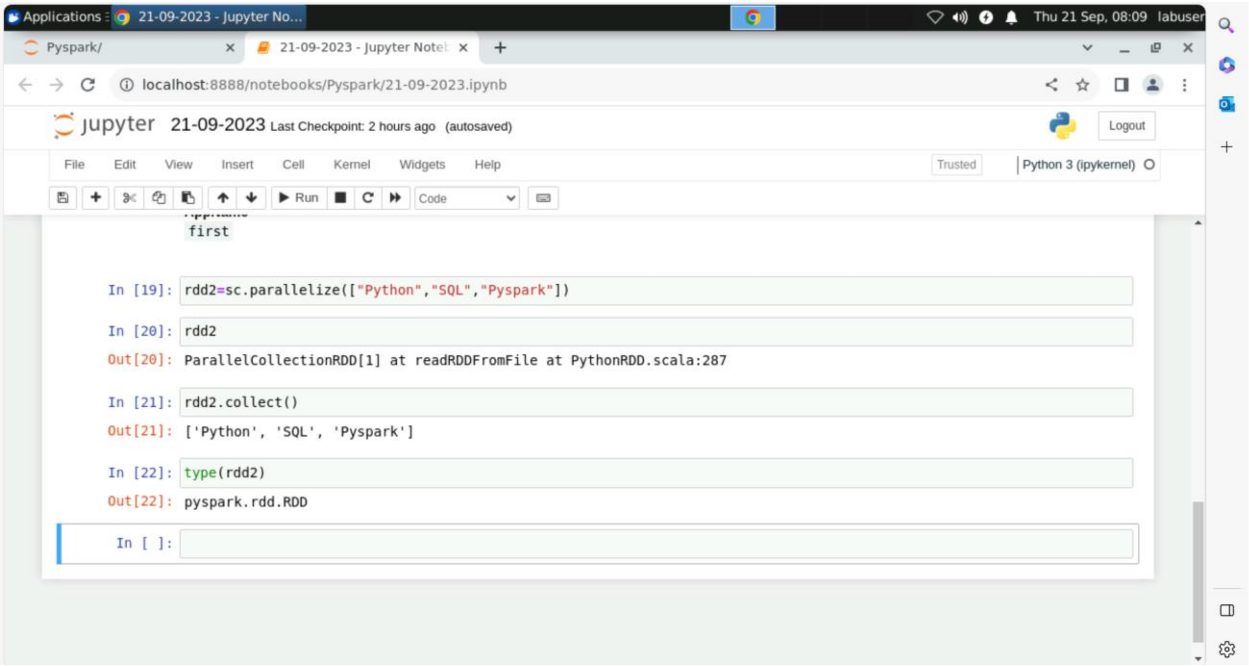
▸ Event Timeline

▾ Completed Jobs (2)

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

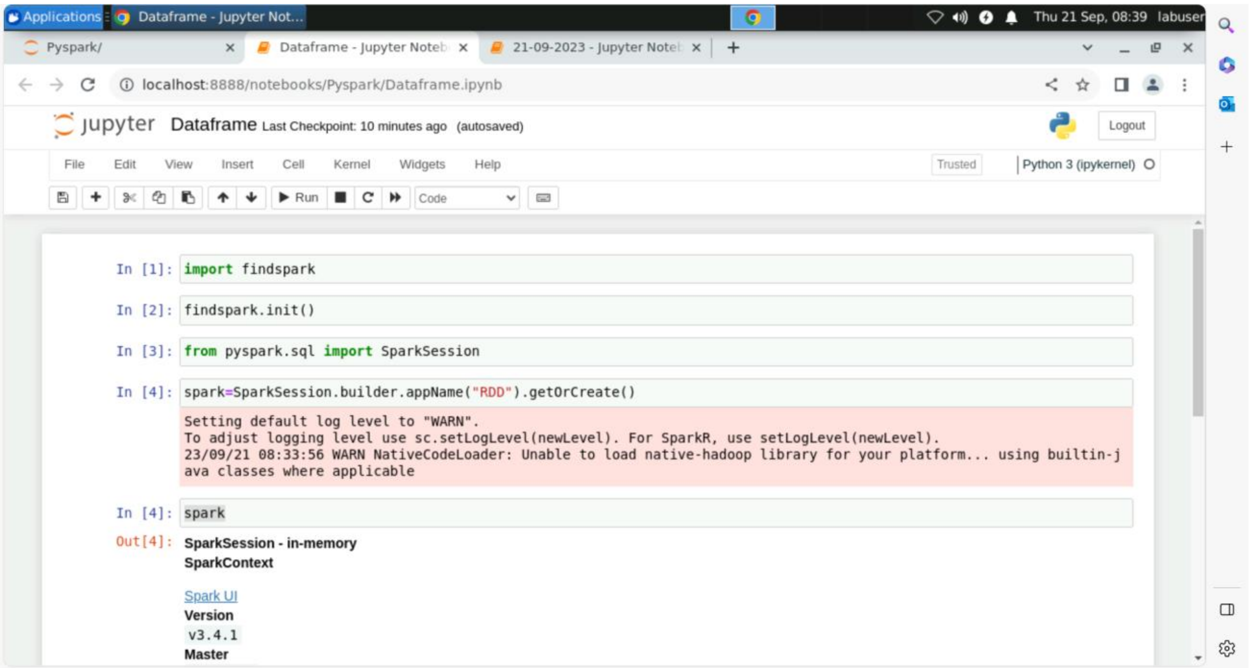
Job Id ▾	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
1	collect at /tmp/ipykernel_4962/1896220800.py:1 collect at /tmp/ipykernel_4962/1896220800.py:1	2023/09/21 06:41:50	96 ms	1/1	1/1
0	collect at /tmp/ipykernel_4962/1896220800.py:1 collect at /tmp/ipykernel_4962/1896220800.py:1	2023/09/21 06:41:28	2 s	1/1	1/1

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

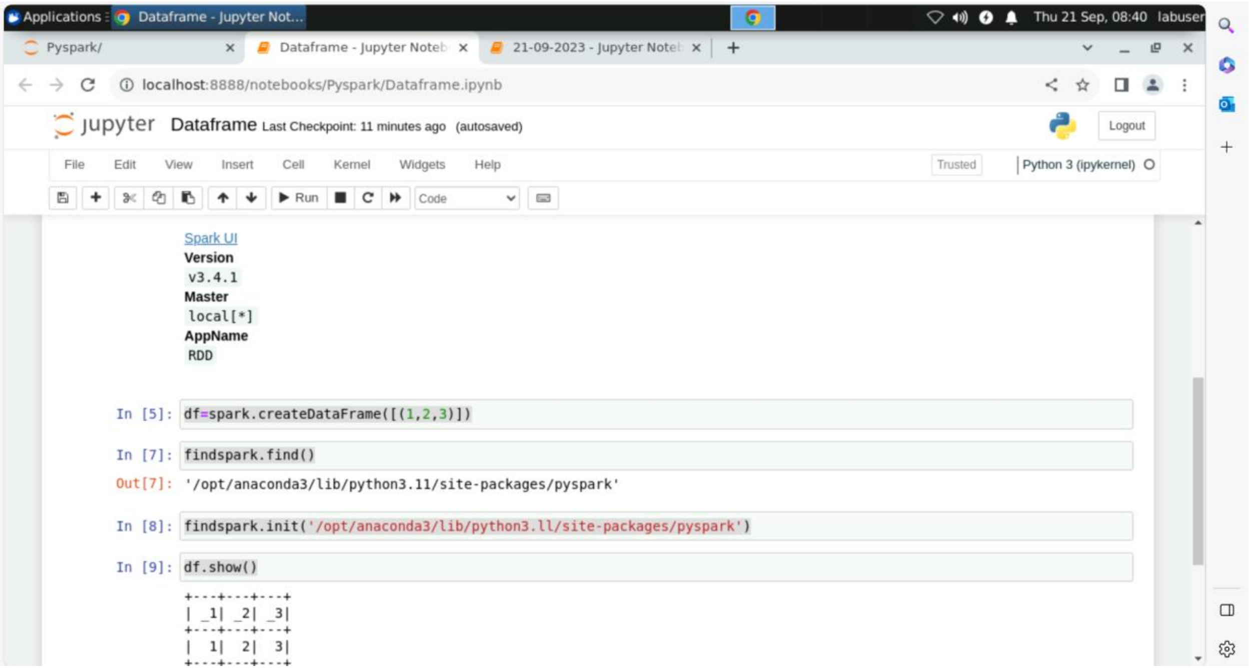

localhost:8888/notebooks/Pyspark/21-09-2023.ipynb
jupyter 21-09-2023 Last Checkpoint: 2 hours ago (autosaved)
Thu 21 Sep, 08:09

```
first

In [19]: rdd2=sc.parallelize(["Python","SQL","Pyspark"])

In [20]: rdd2
Out[20]: ParallelCollectionRDD[1] at readRDDFromFile at PythonRDD.scala:287

In [21]: rdd2.collect()
Out[21]: ['Python', 'SQL', 'Pyspark']

In [22]: type(rdd2)
Out[22]: pyspark.rdd.RDD

In [ ]:
```

Dataframe:


localhost:8888/notebooks/Pyspark/Dataframe.ipynb
jupyter Dataframe Last Checkpoint: 10 minutes ago (autosaved)
Thu 21 Sep, 08:39

```
In [1]: import findspark

In [2]: findspark.init()

In [3]: from pyspark.sql import SparkSession

In [4]: spark=SparkSession.builder.appName("RDD").getOrCreate()

        Setting default log level to "WARN".
        To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
        23/09/21 08:33:56 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-j
        ava classes where applicable

In [4]: spark
Out[4]: SparkSession - in-memory
        SparkContext

        Spark UI
        Version
         v3.4.1
        Master
```


Applications
Dataframe - Jupyter Not...
Thu 21 Sep, 08:40 labuser
Pyspark/
Dataframe - Jupyter Noteb
21-09-2023 - Jupyter Noteb
localhost:8888/notebooks/Pyspark/Dataframe.ipynb
jupyter
Dataframe Last Checkpoint: 11 minutes ago (autosaved)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Python 3 (ipykernel)
Run
Code

Spark UI

Version

`v3.4.1`

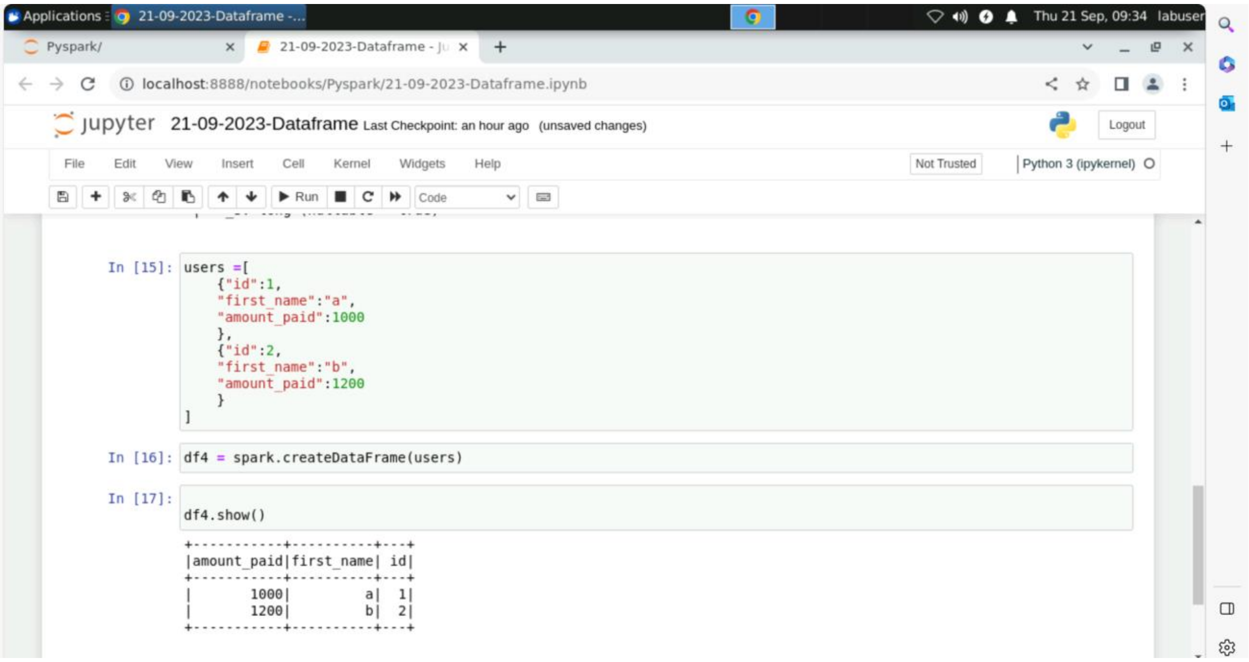
Master

`local[*]`

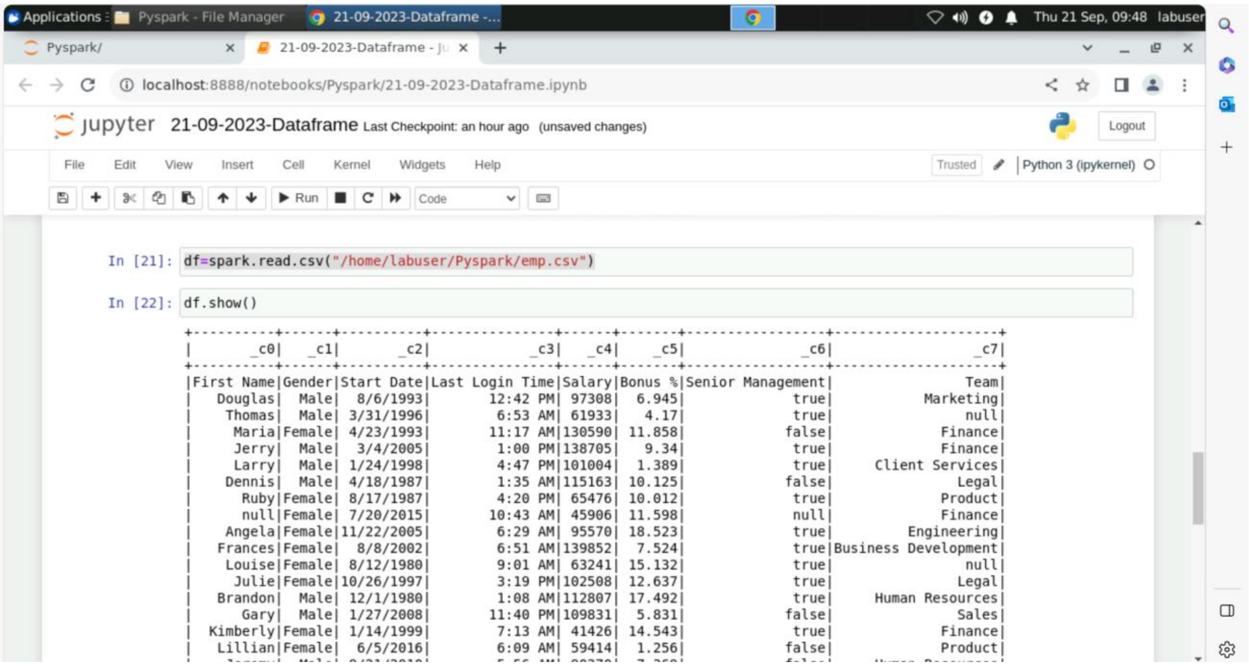
AppName

`RDD`

```
In [5]: df=spark.createDataFrame([(1,2,3)])

In [7]: findspark.find()

Out[7]: '/opt/anaconda3/lib/python3.11/site-packages/pyspark'

In [8]: findspark.init('/opt/anaconda3/lib/python3.ll/site-packages/pyspark')

In [9]: df.show()

        +---+---+---+
        | _1| _2| _3|
        +---+---+---+
        |  1|  2|  3|
        +---+---+---+
```


Applications
21-09-2023-Dataframe -...
Thu 21 Sep, 09:34 labuser
Pyspark/
21-09-2023-Dataframe - Ju
localhost:8888/notebooks/Pyspark/21-09-2023-Dataframe.ipynb
jupyter
21-09-2023-Dataframe Last Checkpoint: an hour ago (unsaved changes)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Not Trusted
Python 3 (ipykernel)
Run
Code

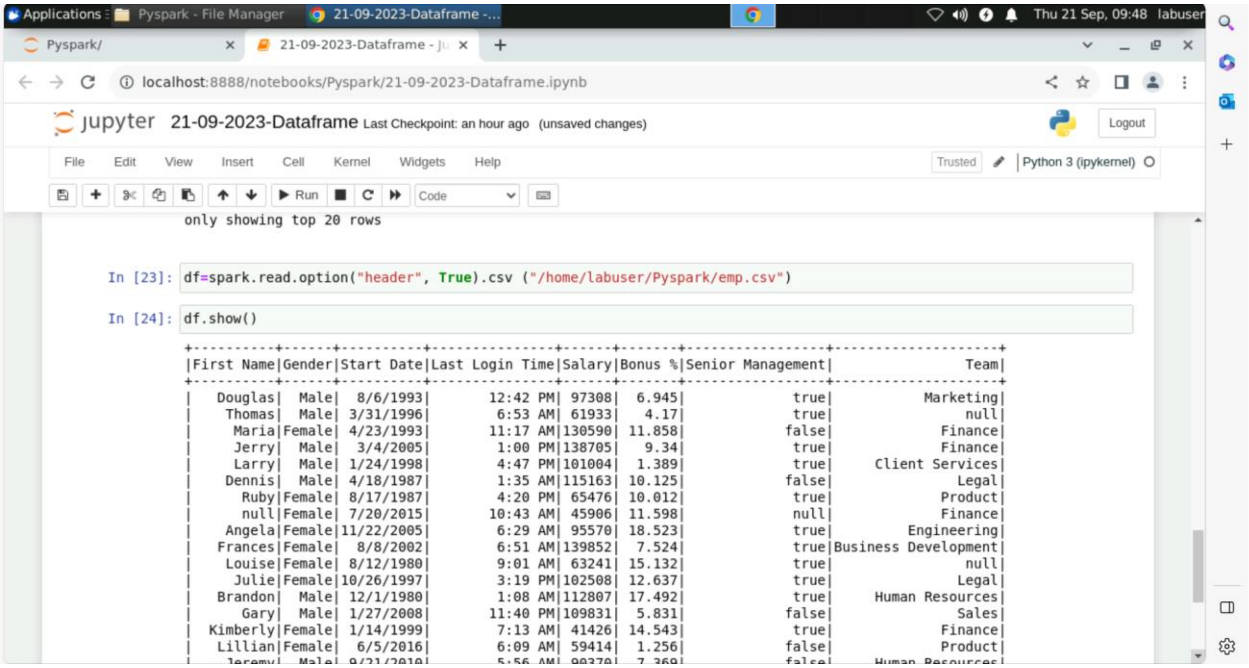
```
In [15]: users =[
             {"id":1,
             "first_name":"a",
             "amount_paid":1000
             },
             {"id":2,
             "first_name":"b",
             "amount_paid":1200
             }
         ]

In [16]: df4 = spark.createDataFrame(users)

In [17]:
         df4.show()

         +-----------+----------+---+
         |amount_paid|first_name| id|
         +-----------+----------+---+
         |       1000|         a|  1|
         |       1200|         b|  2|
         +-----------+----------+---+
```


Applications
Pyspark - File Manager
21-09-2023-Dataframe -...
Thu 21 Sep, 09:48 labuser
Pyspark/
21-09-2023-Dataframe - Ju
localhost:8888/notebooks/Pyspark/21-09-2023-Dataframe.ipynb
jupyter 21-09-2023-Dataframe Last Checkpoint: an hour ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code

```
In [21]: df=spark.read.csv("/home/labuser/Pyspark/emp.csv")
```

```
In [22]: df.show()
```

```
+----------+------+----------+---------------+------+-------+-----------------+--------------------+
|       _c0|   _c1|       _c2|            _c3|   _c4|    _c5|              _c6|                 _c7|
+----------+------+----------+---------------+------+-------+-----------------+--------------------+
|First Name|Gender|Start Date|Last Login Time|Salary|Bonus %|Senior Management|                Team|
|   Douglas|  Male|  8/6/1993|       12:42 PM| 97308|  6.945|             true|           Marketing|
|    Thomas|  Male| 3/31/1996|        6:53 AM| 61933|   4.17|             true|                null|
|     Maria|Female| 4/23/1993|       11:17 AM|130590| 11.858|            false|             Finance|
|     Jerry|  Male|  3/4/2005|        1:00 PM|138705|   9.34|             true|             Finance|
|     Larry|  Male| 1/24/1998|        4:47 PM|101004|  1.389|             true|     Client Services|
|    Dennis|  Male| 4/18/1987|        1:35 AM|115163| 10.125|            false|               Legal|
|      Ruby|Female| 8/17/1987|        4:20 PM| 65476| 10.012|             true|             Product|
|      null|Female| 7/20/2015|       10:43 AM| 45906| 11.598|             null|             Finance|
|    Angela|Female|11/22/2005|        6:29 AM| 95570| 18.523|             true|         Engineering|
|   Frances|Female|  8/8/2002|        6:51 AM|139852|  7.524|             true|Business Development|
|    Louise|Female| 8/12/1980|        9:01 AM| 63241| 15.132|             true|                null|
|     Julie|Female|10/26/1997|        3:19 PM|102508| 12.637|             true|               Legal|
|   Brandon|  Male| 12/1/1980|        1:08 AM|112807| 17.492|             true|     Human Resources|
|      Gary|  Male| 1/27/2008|       11:40 PM|109831|  5.831|            false|               Sales|
|  Kimberly|Female| 1/14/1999|        7:13 AM| 41426| 14.543|             true|             Finance|
|   Lillian|Female|  6/5/2016|        6:09 AM| 59414|  1.256|            false|             Product|
|    Jeremy|  Male| 9/21/2010|        5:56 AM| 90370|  7.369|            false|     Human Resources|
```

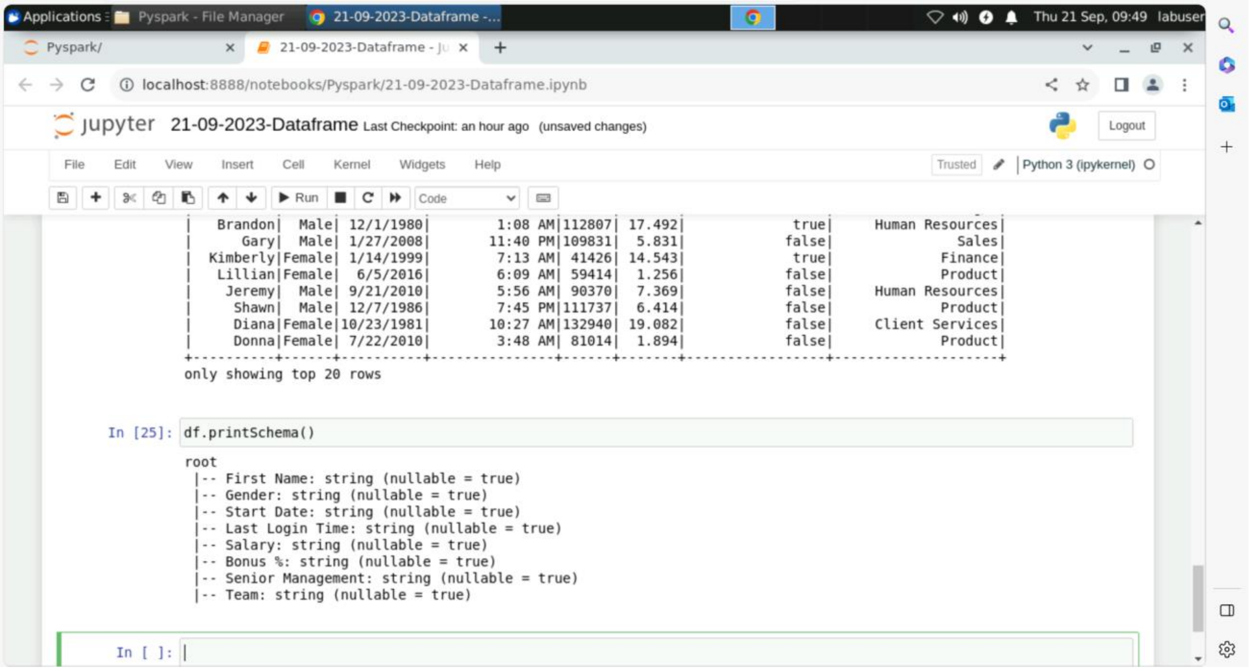

Applications
Pyspark - File Manager
21-09-2023-Dataframe -...
Thu 21 Sep, 09:48 labuser
Pyspark/
21-09-2023-Dataframe - Ju
localhost:8888/notebooks/Pyspark/21-09-2023-Dataframe.ipynb
jupyter 21-09-2023-Dataframe Last Checkpoint: an hour ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code

```
only showing top 20 rows
```

```
In [23]: df=spark.read.option("header", True).csv ("/home/labuser/Pyspark/emp.csv")
```

```
In [24]: df.show()
```

```
+----------+------+----------+---------------+------+-------+-----------------+--------------------+
|First Name|Gender|Start Date|Last Login Time|Salary|Bonus %|Senior Management|                Team|
+----------+------+----------+---------------+------+-------+-----------------+--------------------+
|   Douglas|  Male|  8/6/1993|       12:42 PM| 97308|  6.945|             true|           Marketing|
|    Thomas|  Male| 3/31/1996|        6:53 AM| 61933|   4.17|             true|                null|
|     Maria|Female| 4/23/1993|       11:17 AM|130590| 11.858|            false|             Finance|
|     Jerry|  Male|  3/4/2005|        1:00 PM|138705|   9.34|             true|             Finance|
|     Larry|  Male| 1/24/1998|        4:47 PM|101004|  1.389|             true|     Client Services|
|    Dennis|  Male| 4/18/1987|        1:35 AM|115163| 10.125|            false|               Legal|
|      Ruby|Female| 8/17/1987|        4:20 PM| 65476| 10.012|             true|             Product|
|      null|Female| 7/20/2015|       10:43 AM| 45906| 11.598|             null|             Finance|
|    Angela|Female|11/22/2005|        6:29 AM| 95570| 18.523|             true|         Engineering|
|   Frances|Female|  8/8/2002|        6:51 AM|139852|  7.524|             true|Business Development|
|    Louise|Female| 8/12/1980|        9:01 AM| 63241| 15.132|             true|                null|
|     Julie|Female|10/26/1997|        3:19 PM|102508| 12.637|             true|               Legal|
|   Brandon|  Male| 12/1/1980|        1:08 AM|112807| 17.492|             true|     Human Resources|
|      Gary|  Male| 1/27/2008|       11:40 PM|109831|  5.831|            false|               Sales|
|  Kimberly|Female| 1/14/1999|        7:13 AM| 41426| 14.543|             true|             Finance|
|   Lillian|Female|  6/5/2016|        6:09 AM| 59414|  1.256|            false|             Product|
|    Jeremy|  Male| 9/21/2010|        5:56 AM| 90370|  7.369|            false|     Human Resources|
```

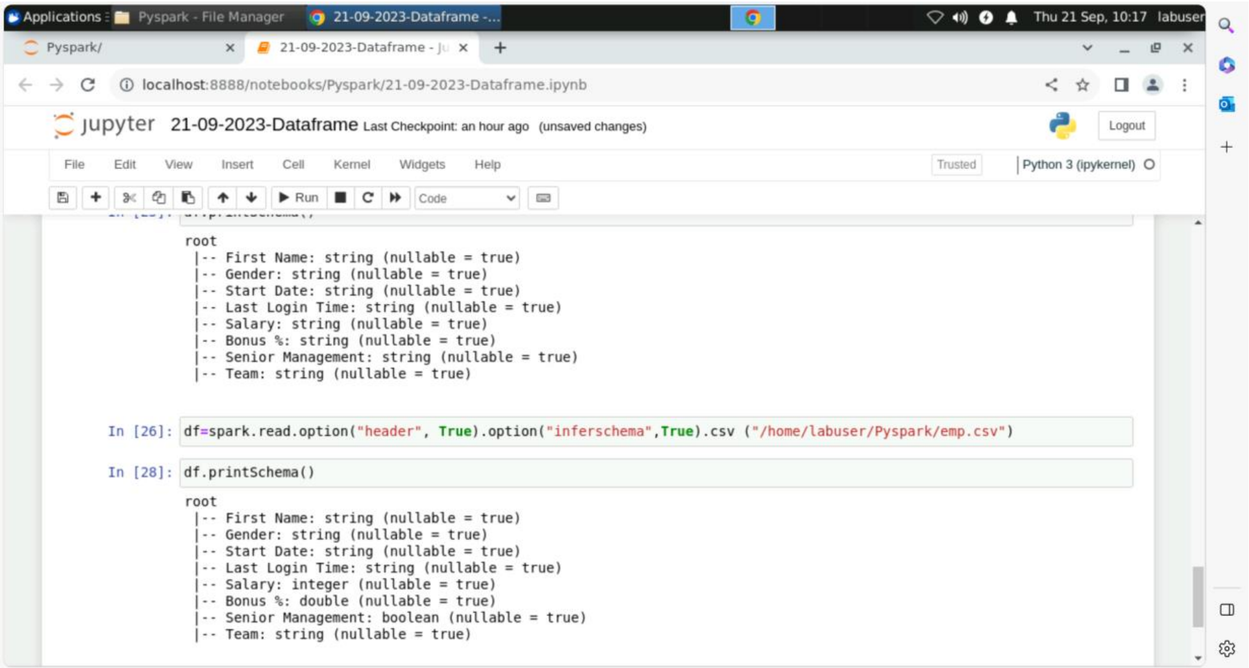

Applications
Pyspark - File Manager
21-09-2023-Dataframe -...
Thu 21 Sep, 09:49 labuser
Pyspark/
21-09-2023-Dataframe - Ju
localhost:8888/notebooks/Pyspark/21-09-2023-Dataframe.ipynb
jupyter 21-09-2023-Dataframe Last Checkpoint: an hour ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code

```
|   Brandon|  Male| 12/1/1980|       1:08 AM|112807| 17.492|            true|     Human Resources|
|      Gary|  Male| 1/27/2008|      11:40 PM|109831|  5.831|           false|               Sales|
|  Kimberly|Female| 1/14/1999|       7:13 AM| 41426| 14.543|            true|             Finance|
|   Lillian|Female|  6/5/2016|       6:09 AM| 59414|  1.256|           false|             Product|
|    Jeremy|  Male| 9/21/2010|       5:56 AM| 90370|  7.369|           false|     Human Resources|
|     Shawn|  Male| 12/7/1986|       7:45 PM|111737|  6.414|           false|             Product|
|     Diana|Female|10/23/1981|      10:27 AM|132940| 19.082|           false|     Client Services|
|     Donna|Female| 7/22/2010|       3:48 AM| 81014|  1.894|           false|             Product|
+----------+------+----------+--------------+------+-------+----------------+--------------------+
only showing top 20 rows
```

```
In [25]: df.printSchema()
```

```
root
 |-- First Name: string (nullable = true)
 |-- Gender: string (nullable = true)
 |-- Start Date: string (nullable = true)
 |-- Last Login Time: string (nullable = true)
 |-- Salary: string (nullable = true)
 |-- Bonus %: string (nullable = true)
 |-- Senior Management: string (nullable = true)
 |-- Team: string (nullable = true)
```

```
In [ ]:
```

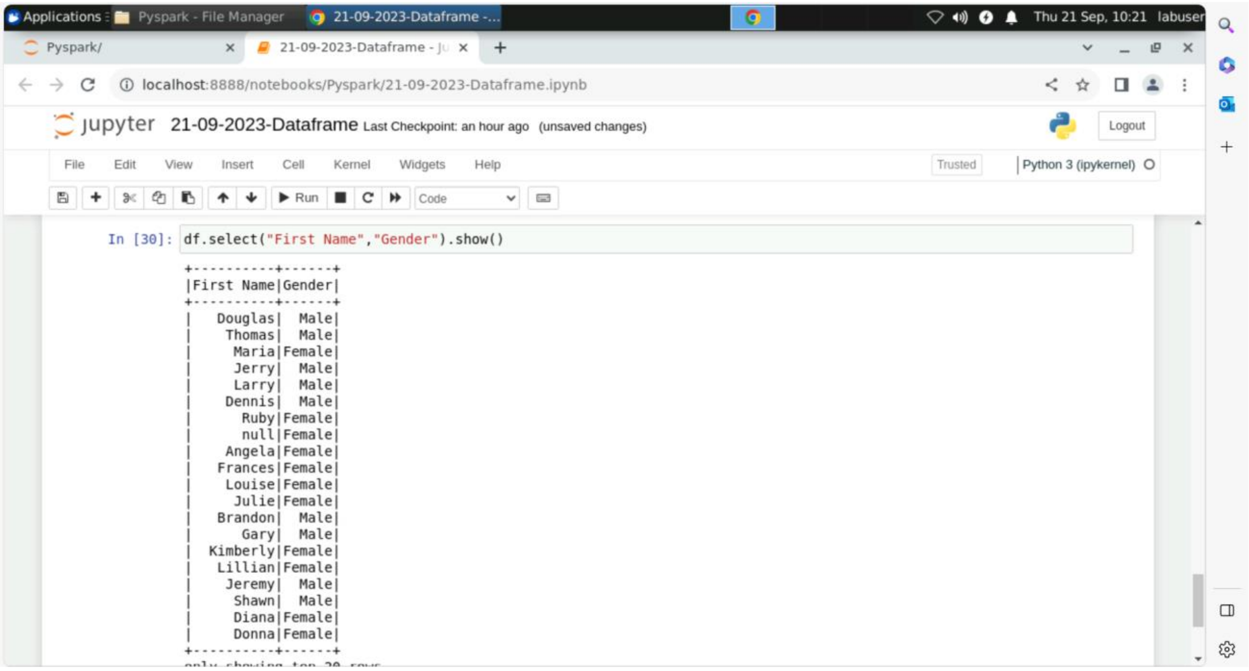

Applications
Pyspark - File Manager
21-09-2023-Dataframe -...
Thu 21 Sep, 10:17 labuser
Pyspark/
21-09-2023-Dataframe - Ju
localhost:8888/notebooks/Pyspark/21-09-2023-Dataframe.ipynb
jupyter 21-09-2023-Dataframe Last Checkpoint: an hour ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code

```
root
 |-- First Name: string (nullable = true)
 |-- Gender: string (nullable = true)
 |-- Start Date: string (nullable = true)
 |-- Last Login Time: string (nullable = true)
 |-- Salary: string (nullable = true)
 |-- Bonus %: string (nullable = true)
 |-- Senior Management: string (nullable = true)
 |-- Team: string (nullable = true)
```

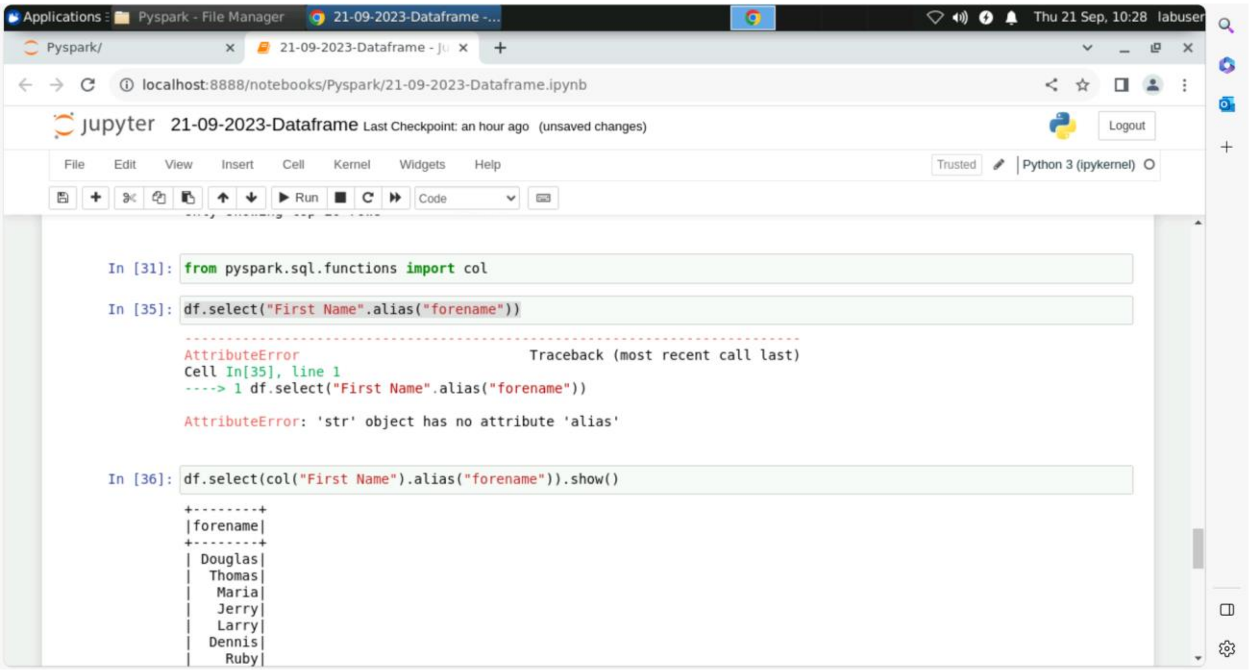
```
In [26]: df=spark.read.option("header", True).option("inferschema",True).csv ("/home/labuser/Pyspark/emp.csv")
```

```
In [28]: df.printSchema()
```

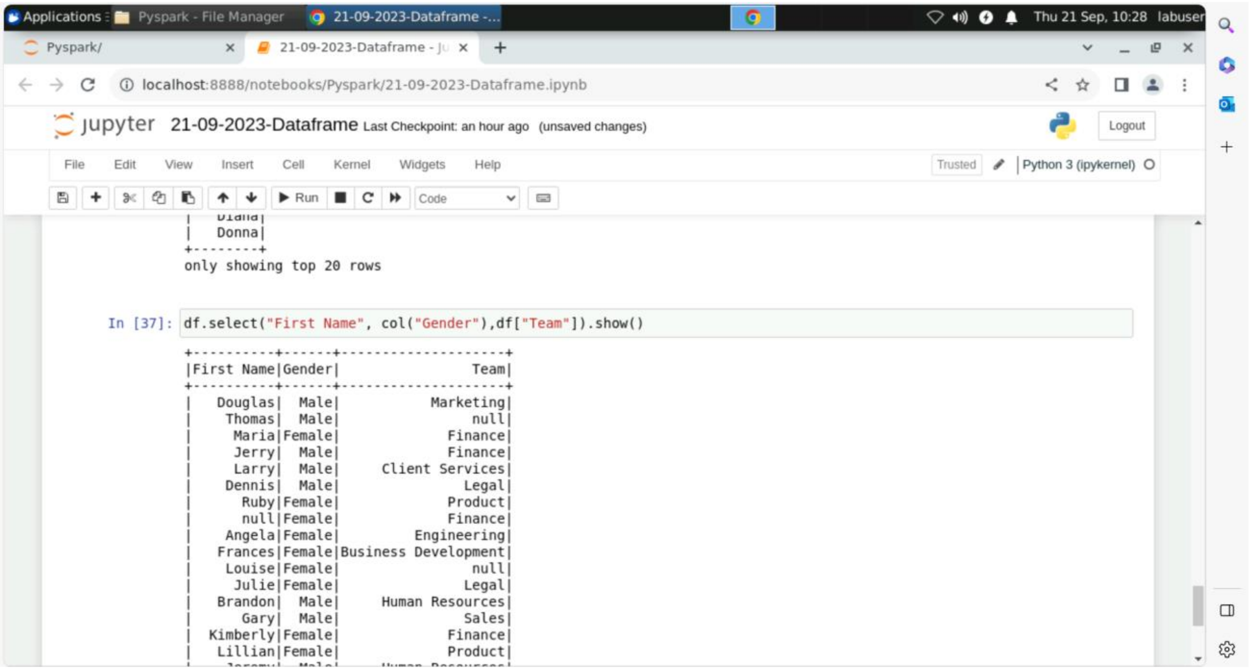
```
root
 |-- First Name: string (nullable = true)
 |-- Gender: string (nullable = true)
 |-- Start Date: string (nullable = true)
 |-- Last Login Time: string (nullable = true)
 |-- Salary: integer (nullable = true)
 |-- Bonus %: double (nullable = true)
 |-- Senior Management: boolean (nullable = true)
 |-- Team: string (nullable = true)
```


Applications
Pyspark - File Manager
21-09-2023-Dataframe -...
Thu 21 Sep, 10:21 labuser
Pyspark/
21-09-2023-Dataframe - Ju
localhost:8888/notebooks/Pyspark/21-09-2023-Dataframe.ipynb
jupyter 21-09-2023-Dataframe Last Checkpoint: an hour ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code

```
In [30]: df.select("First Name","Gender").show()

         +----------+------+
         |First Name|Gender|
         +----------+------+
         |   Douglas|  Male|
         |    Thomas|  Male|
         |     Maria|Female|
         |     Jerry|  Male|
         |     Larry|  Male|
         |    Dennis|  Male|
         |      Ruby|Female|
         |      null|Female|
         |    Angela|Female|
         |   Frances|Female|
         |    Louise|Female|
         |     Julie|Female|
         |   Brandon|  Male|
         |      Gary|  Male|
         |  Kimberly|Female|
         |   Lillian|Female|
         |    Jeremy|  Male|
         |     Shawn|  Male|
         |     Diana|Female|
         |     Donna|Female|
         +----------+------+
         only showing top 20 rows
```


Applications
Pyspark - File Manager
21-09-2023-Dataframe -...
Thu 21 Sep, 10:28 labuser
Pyspark/
21-09-2023-Dataframe - Ju
localhost:8888/notebooks/Pyspark/21-09-2023-Dataframe.ipynb
jupyter 21-09-2023-Dataframe Last Checkpoint: an hour ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code

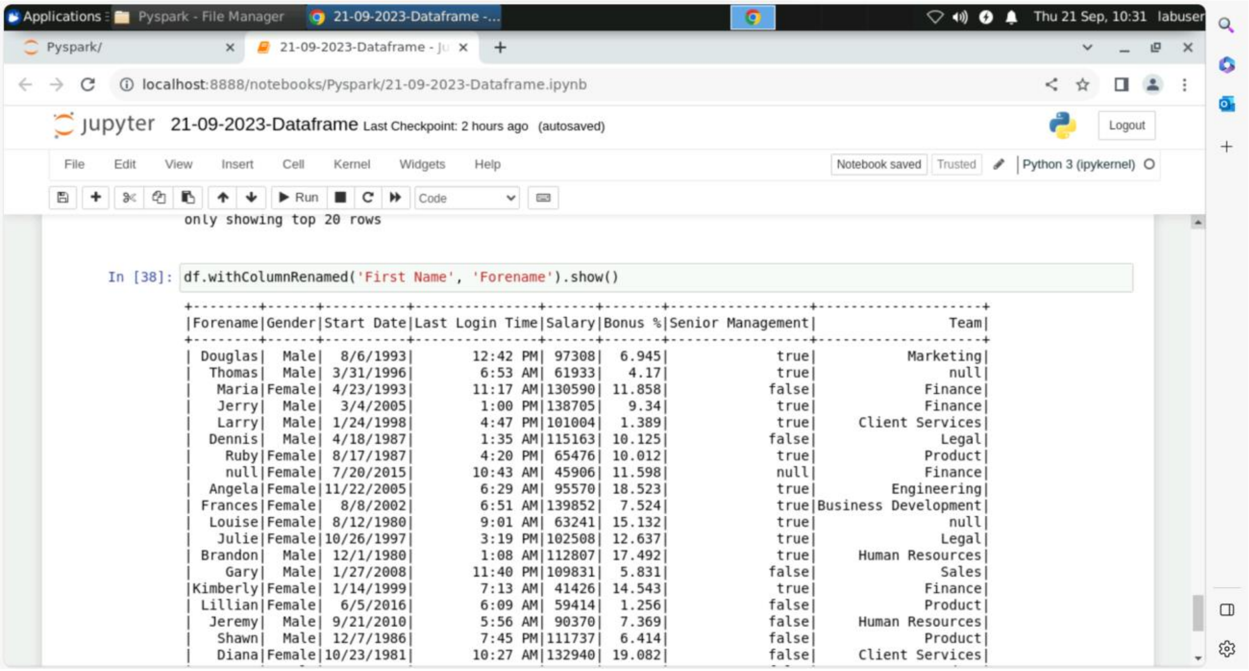
```
         only showing top 20 rows

In [31]: from pyspark.sql.functions import col

In [35]: df.select("First Name".alias("forename"))

         ---------------------------------------------------------------------------
         AttributeError                            Traceback (most recent call last)
         Cell In[35], line 1
         ----> 1 df.select("First Name".alias("forename"))

         AttributeError: 'str' object has no attribute 'alias'


In [36]: df.select(col("First Name").alias("forename")).show()

         +--------+
         |forename|
         +--------+
         | Douglas|
         |  Thomas|
         |   Maria|
         |   Jerry|
         |   Larry|
         |  Dennis|
         |    Ruby|
```


Applications
Pyspark - File Manager
21-09-2023-Dataframe -...
Thu 21 Sep, 10:28 labuser
Pyspark/
21-09-2023-Dataframe - Ju
localhost:8888/notebooks/Pyspark/21-09-2023-Dataframe.ipynb
jupyter 21-09-2023-Dataframe Last Checkpoint: an hour ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code

```
|   Diana|
|   Donna|
+--------+
only showing top 20 rows
```

In [37]:
```
df.select("First Name", col("Gender"),df["Team"]).show()
```

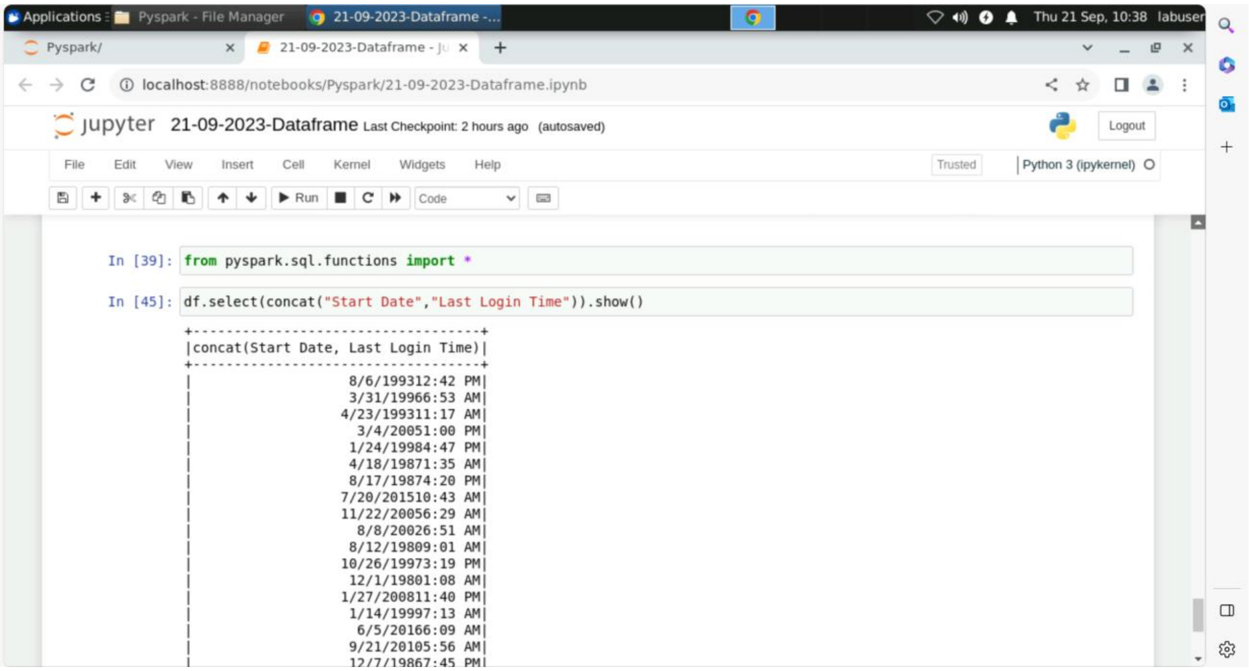
```
+----------+------+--------------------+
|First Name|Gender|                Team|
+----------+------+--------------------+
|   Douglas|  Male|           Marketing|
|    Thomas|  Male|                null|
|     Maria|Female|             Finance|
|     Jerry|  Male|             Finance|
|     Larry|  Male|     Client Services|
|    Dennis|  Male|               Legal|
|      Ruby|Female|             Product|
|      null|Female|             Finance|
|    Angela|Female|         Engineering|
|   Frances|Female|Business Development|
|    Louise|Female|                null|
|     Julie|Female|               Legal|
|   Brandon|  Male|     Human Resources|
|      Gary|  Male|               Sales|
|  Kimberly|Female|             Finance|
|   Lillian|Female|             Product|
```


Applications
Pyspark - File Manager
21-09-2023-Dataframe -...
Thu 21 Sep, 10:31 labuser
Pyspark/
21-09-2023-Dataframe - Ju
localhost:8888/notebooks/Pyspark/21-09-2023-Dataframe.ipynb
jupyter 21-09-2023-Dataframe Last Checkpoint: 2 hours ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Notebook saved
Trusted
Python 3 (ipykernel)
Run
Code

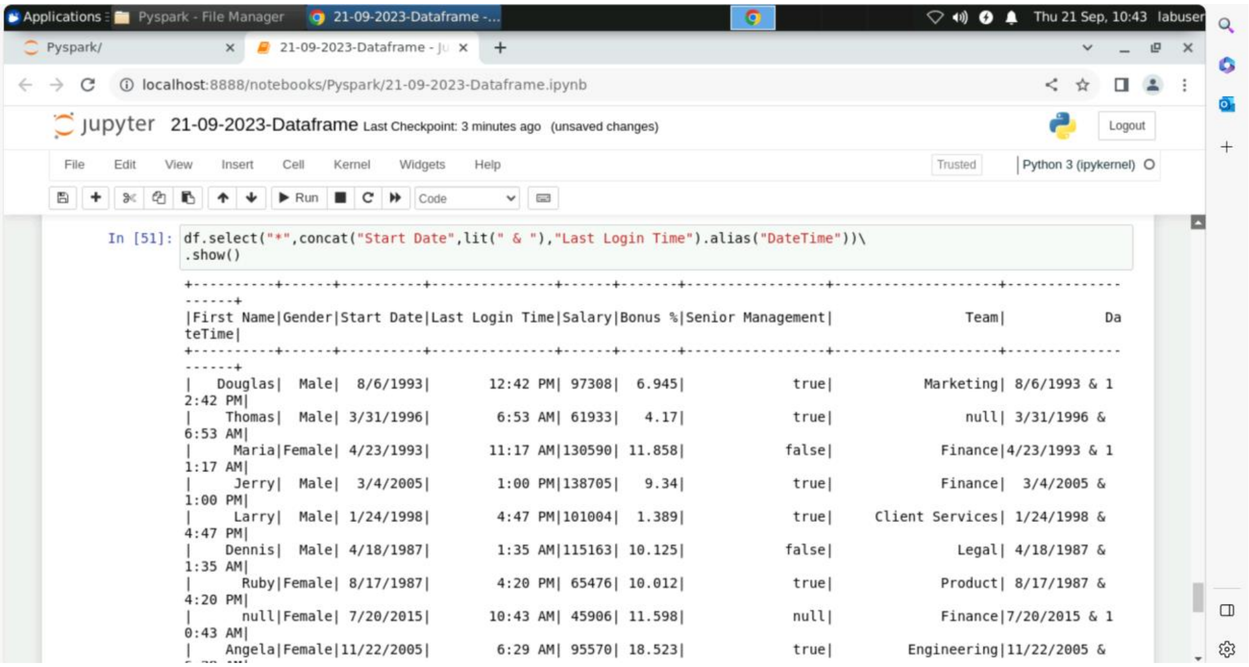
```
only showing top 20 rows
```

In [38]:
```
df.withColumnRenamed('First Name', 'Forename').show()
```

```
+--------+------+----------+---------------+------+-------+-----------------+--------------------+
|Forename|Gender|Start Date|Last Login Time|Salary|Bonus %|Senior Management|                Team|
+--------+------+----------+---------------+------+-------+-----------------+--------------------+
| Douglas|  Male|  8/6/1993|       12:42 PM| 97308|  6.945|             true|           Marketing|
|  Thomas|  Male| 3/31/1996|        6:53 AM| 61933|   4.17|             true|                null|
|   Maria|Female| 4/23/1993|       11:17 AM|130590| 11.858|            false|             Finance|
|   Jerry|  Male|  3/4/2005|        1:00 PM|138705|   9.34|             true|             Finance|
|   Larry|  Male| 1/24/1998|        4:47 PM|101004|  1.389|             true|     Client Services|
|  Dennis|  Male| 4/18/1987|        1:35 AM|115163| 10.125|            false|               Legal|
|    Ruby|Female| 8/17/1987|        4:20 PM| 65476| 10.012|             true|             Product|
|    null|Female| 7/20/2015|       10:43 AM| 45906| 11.598|             null|             Finance|
|  Angela|Female|11/22/2005|        6:29 AM| 95570| 18.523|             true|         Engineering|
| Frances|Female|  8/8/2002|        6:51 AM|139852|  7.524|             true|Business Development|
|  Louise|Female| 8/12/1980|        9:01 AM| 63241| 15.132|             true|                null|
|   Julie|Female|10/26/1997|        3:19 PM|102508| 12.637|             true|               Legal|
| Brandon|  Male| 12/1/1980|        1:08 AM|112807| 17.492|             true|     Human Resources|
|    Gary|  Male| 1/27/2008|       11:40 PM|109831|  5.831|            false|               Sales|
|Kimberly|Female| 1/14/1999|        7:13 AM| 41426| 14.543|             true|             Finance|
| Lillian|Female|  6/5/2016|        6:09 AM| 59414|  1.256|            false|             Product|
|  Jeremy|  Male| 9/21/2010|        5:56 AM| 90370|  7.369|            false|     Human Resources|
|   Shawn|  Male| 12/7/1986|        7:45 PM|111737|  6.414|            false|             Product|
|   Diana|Female|10/23/1981|       10:27 AM|132940| 19.082|            false|     Client Services|
```


Applications
Pyspark - File Manager
21-09-2023-Dataframe -...
Thu 21 Sep, 10:38 labuser
Pyspark/
21-09-2023-Dataframe - Ju
localhost:8888/notebooks/Pyspark/21-09-2023-Dataframe.ipynb
jupyter 21-09-2023-Dataframe Last Checkpoint: 2 hours ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code

```
In [39]: from pyspark.sql.functions import *

In [45]: df.select(concat("Start Date","Last Login Time")).show()

         +--------------------------------+
         |concat(Start Date, Last Login Time)|
         +--------------------------------+
         |                 8/6/199312:42 PM|
         |                 3/31/19966:53 AM|
         |                4/23/199311:17 AM|
         |                  3/4/20051:00 PM|
         |                 1/24/19984:47 PM|
         |                 4/18/19871:35 AM|
         |                 8/17/19874:20 PM|
         |                7/20/201510:43 AM|
         |                11/22/20056:29 AM|
         |                  8/8/20026:51 AM|
         |                 8/12/19809:01 AM|
         |                10/26/19973:19 PM|
         |                 12/1/19801:08 AM|
         |                1/27/200811:40 PM|
         |                 1/14/19997:13 AM|
         |                  6/5/20166:09 AM|
         |                 9/21/20105:56 AM|
         |                 12/7/19867:45 PM|
```


Applications
Pyspark - File Manager
21-09-2023-Dataframe -...
Thu 21 Sep, 10:43 labuser
Pyspark/
21-09-2023-Dataframe - Ju
localhost:8888/notebooks/Pyspark/21-09-2023-Dataframe.ipynb
jupyter 21-09-2023-Dataframe Last Checkpoint: 3 minutes ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code

```
In [51]: df.select("*",concat("Start Date",lit(" & "),"Last Login Time").alias("DateTime"))\
         .show()

         +----------+------+----------+---------------+------+-------+-----------------+--------------------+------------
         ------+
         |First Name|Gender|Start Date|Last Login Time|Salary|Bonus %|Senior Management|                Team|            Da
         teTime|
         +----------+------+----------+---------------+------+-------+-----------------+--------------------+------------
         ------+
         |   Douglas|  Male|  8/6/1993|       12:42 PM| 97308|  6.945|             true|           Marketing| 8/6/1993 & 1
         2:42 PM|
         |    Thomas|  Male| 3/31/1996|        6:53 AM| 61933|   4.17|             true|                null| 3/31/1996 &
         6:53 AM|
         |     Maria|Female| 4/23/1993|       11:17 AM|130590| 11.858|            false|             Finance|4/23/1993 & 1
         1:17 AM|
         |     Jerry|  Male|  3/4/2005|        1:00 PM|138705|   9.34|             true|             Finance|  3/4/2005 &
         1:00 PM|
         |     Larry|  Male| 1/24/1998|        4:47 PM|101004|  1.389|             true|     Client Services| 1/24/1998 &
         4:47 PM|
         |    Dennis|  Male| 4/18/1987|        1:35 AM|115163| 10.125|            false|               Legal| 4/18/1987 &
         1:35 AM|
         |      Ruby|Female| 8/17/1987|        4:20 PM| 65476| 10.012|             true|             Product| 8/17/1987 &
         4:20 PM|
         |      null|Female| 7/20/2015|       10:43 AM| 45906| 11.598|             null|             Finance|7/20/2015 & 1
         0:43 AM|
         |    Angela|Female|11/22/2005|        6:29 AM| 95570| 18.523|             true|         Engineering|11/22/2005 &
```

dffinal = df1.drop("Start Date","Last Login Time")

dffinal.write.csv("/home/labuser/Pandas/finalemp")