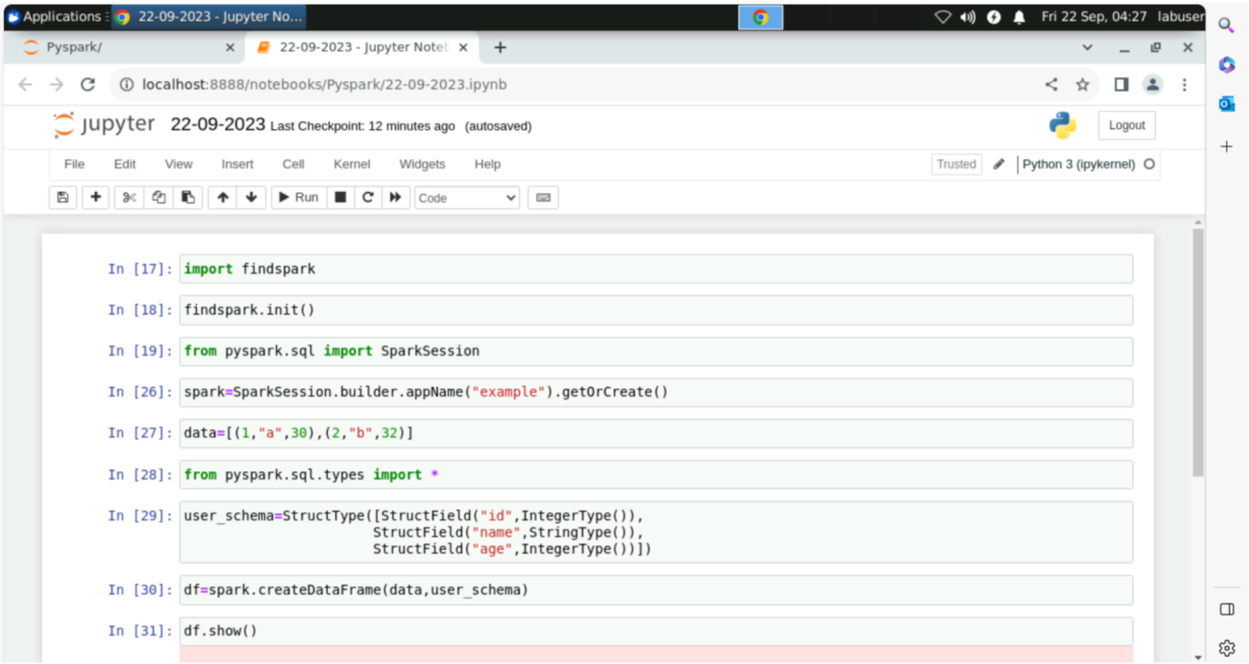
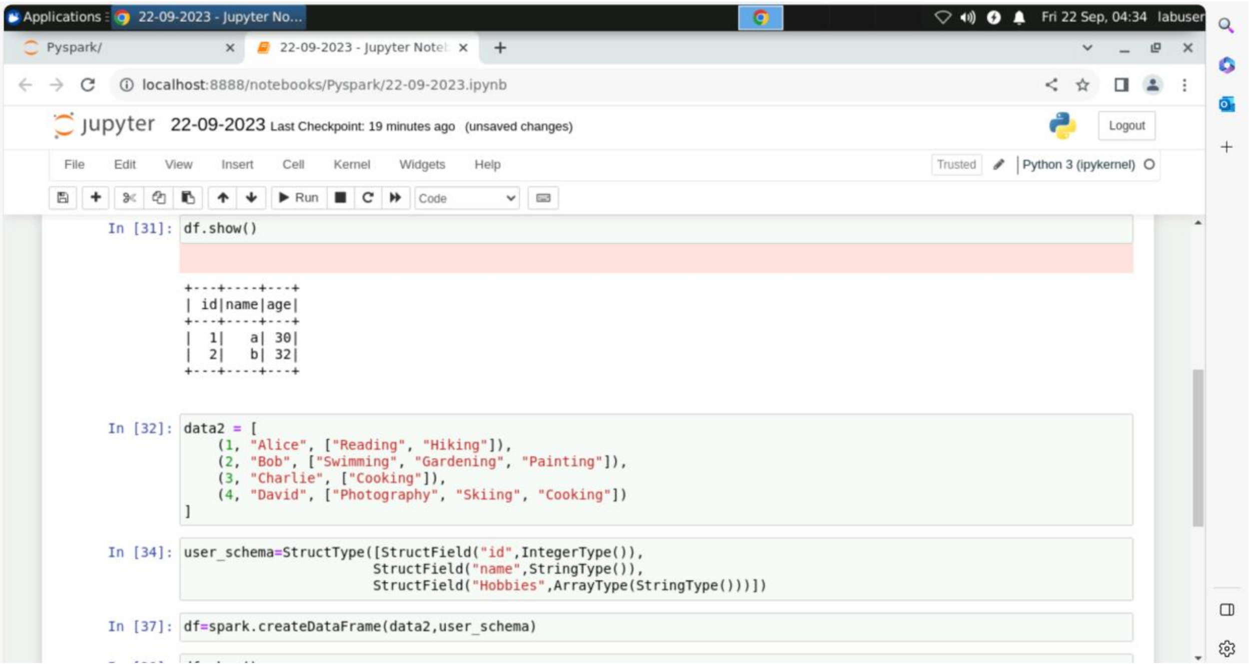
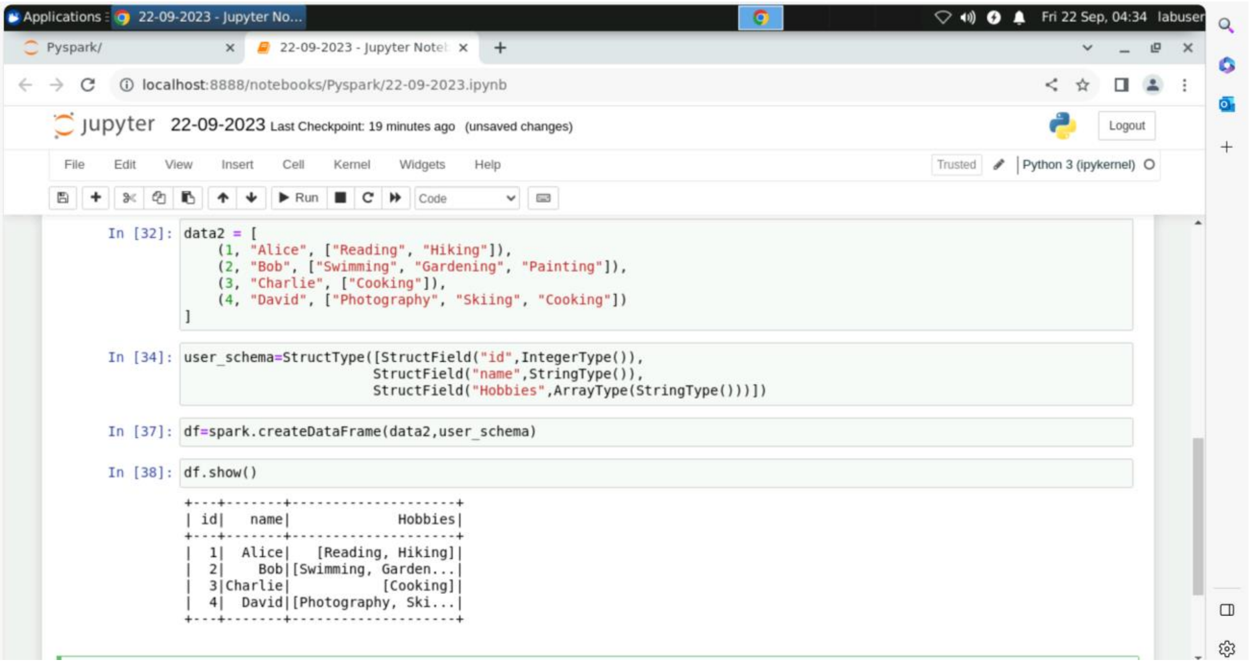
PYSPARK


Applications
22-09-2023 - Jupyter No...
Fri 22 Sep, 04:27 labuser
Pyspark/
22-09-2023 - Jupyter Note
localhost:8888/notebooks/Pyspark/22-09-2023.ipynb
jupyter 22-09-2023 Last Checkpoint: 12 minutes ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code

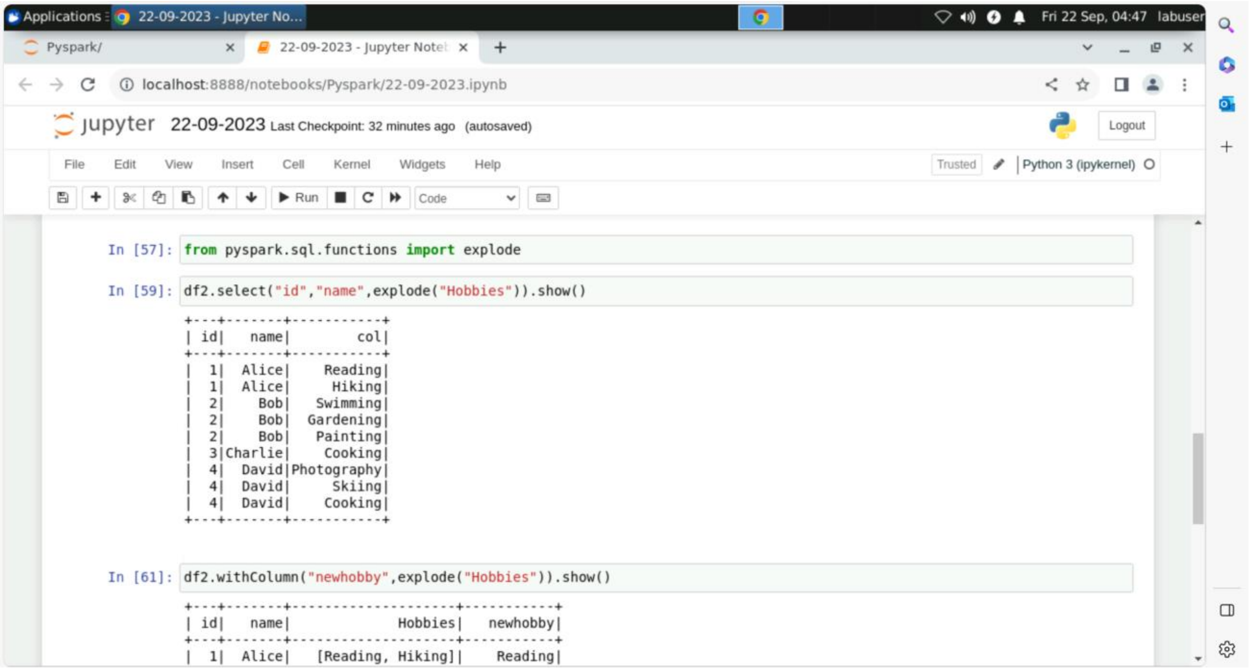
```
In [17]: import findspark

In [18]: findspark.init()

In [19]: from pyspark.sql import SparkSession

In [26]: spark=SparkSession.builder.appName("example").getOrCreate()

In [27]: data=[(1,"a",30),(2,"b",32)]

In [28]: from pyspark.sql.types import *

In [29]: user_schema=StructType([StructField("id",IntegerType()),
                                 StructField("name",StringType()),
                                 StructField("age",IntegerType())])

In [30]: df=spark.createDataFrame(data,user_schema)

In [31]: df.show()
```


Applications
22-09-2023 - Jupyter No...
Fri 22 Sep, 04:34 labuser
Pyspark/
22-09-2023 - Jupyter Note
localhost:8888/notebooks/Pyspark/22-09-2023.ipynb
jupyter 22-09-2023 Last Checkpoint: 19 minutes ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code

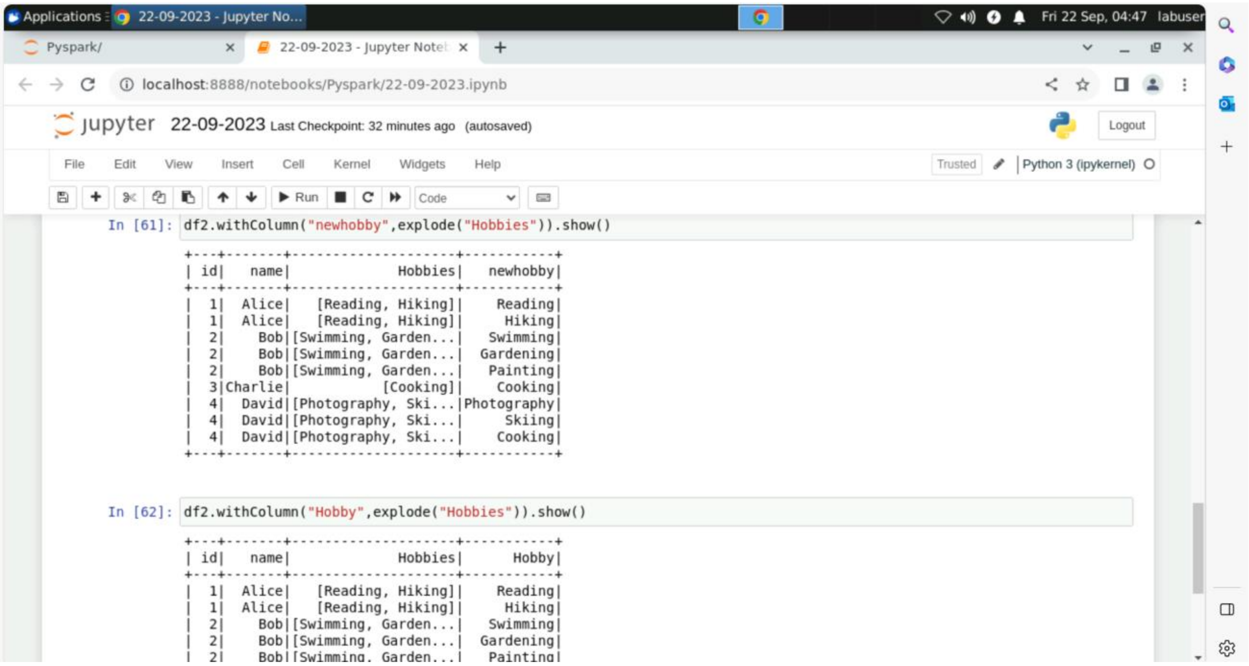
```
In [31]: df.show()

         +---+----+---+
         | id|name|age|
         +---+----+---+
         |  1|   a| 30|
         |  2|   b| 32|
         +---+----+---+

In [32]: data2 = [
             (1, "Alice", ["Reading", "Hiking"]),
             (2, "Bob", ["Swimming", "Gardening", "Painting"]),
             (3, "Charlie", ["Cooking"]),
             (4, "David", ["Photography", "Skiing", "Cooking"])
         ]

In [34]: user_schema=StructType([StructField("id",IntegerType()),
                                 StructField("name",StringType()),
                                 StructField("Hobbies",ArrayType(StringType()))])

In [37]: df=spark.createDataFrame(data2,user_schema)
```


Applications
22-09-2023 - Jupyter No...
Fri 22 Sep, 04:34 labuser
Pyspark/
22-09-2023 - Jupyter Noteb
localhost:8888/notebooks/Pyspark/22-09-2023.ipynb
jupyter 22-09-2023 Last Checkpoint: 19 minutes ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code

```
In [32]: data2 = [
             (1, "Alice", ["Reading", "Hiking"]),
             (2, "Bob", ["Swimming", "Gardening", "Painting"]),
             (3, "Charlie", ["Cooking"]),
             (4, "David", ["Photography", "Skiing", "Cooking"])
         ]

In [34]: user_schema=StructType([StructField("id",IntegerType()),
                                 StructField("name",StringType()),
                                 StructField("Hobbies",ArrayType(StringType()))])

In [37]: df=spark.createDataFrame(data2,user_schema)

In [38]: df.show()

         +---+-------+--------------------+
         | id|   name|             Hobbies|
         +---+-------+--------------------+
         |  1|  Alice|   [Reading, Hiking]|
         |  2|    Bob|[Swimming, Garden...|
         |  3|Charlie|           [Cooking]|
         |  4|  David|[Photography, Ski...|
         +---+-------+--------------------+
```


Applications
22-09-2023 - Jupyter No...
Fri 22 Sep, 04:47 labuser
Pyspark/
22-09-2023 - Jupyter Noteb
localhost:8888/notebooks/Pyspark/22-09-2023.ipynb
jupyter 22-09-2023 Last Checkpoint: 32 minutes ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code

```
In [57]: from pyspark.sql.functions import explode

In [59]: df2.select("id","name",explode("Hobbies")).show()

         +---+-------+-----------+
         | id|   name|        col|
         +---+-------+-----------+
         |  1|  Alice|    Reading|
         |  1|  Alice|     Hiking|
         |  2|    Bob|   Swimming|
         |  2|    Bob|  Gardening|
         |  2|    Bob|   Painting|
         |  3|Charlie|    Cooking|
         |  4|  David|Photography|
         |  4|  David|     Skiing|
         |  4|  David|    Cooking|
         +---+-------+-----------+

In [61]: df2.withColumn("newhobby",explode("Hobbies")).show()

         +---+-------+--------------------+-----------+
         | id|   name|             Hobbies|   newhobby|
         +---+-------+--------------------+-----------+
         |  1|  Alice|   [Reading, Hiking]|    Reading|
```

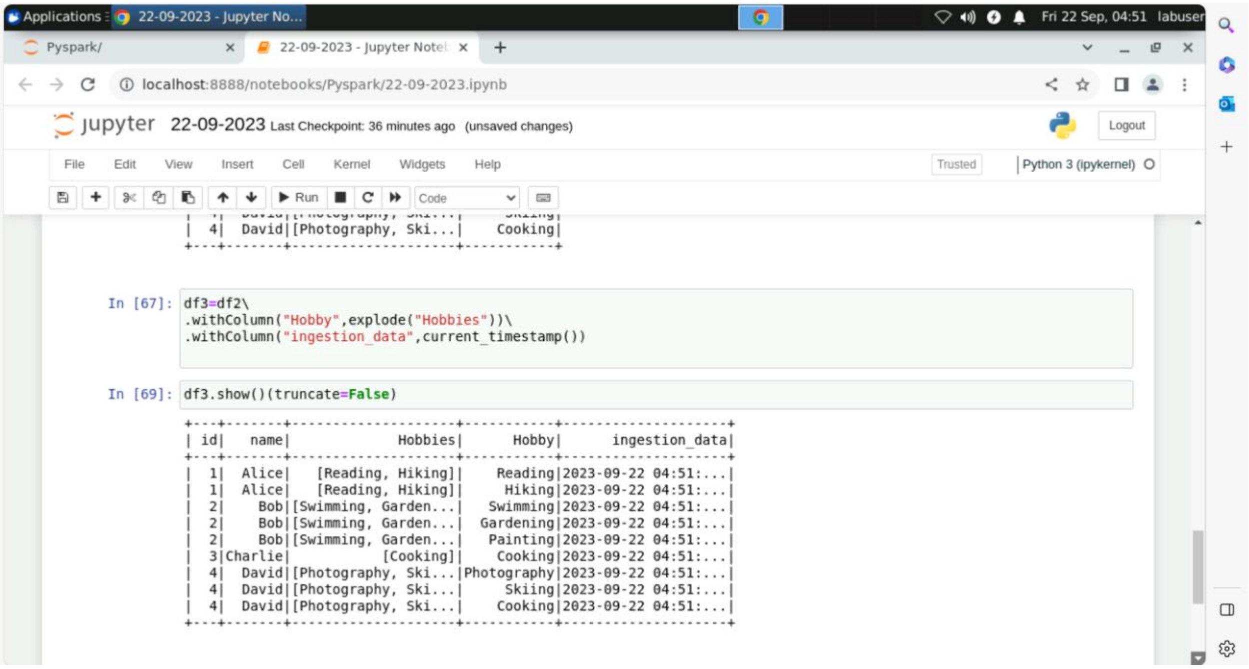


In [61]:
```
df2.withColumn("newhobby",explode("Hobbies")).show()
```

```
+---+-------+--------------------+-----------+
| id|   name|             Hobbies|   newhobby|
+---+-------+--------------------+-----------+
|  1|  Alice|   [Reading, Hiking]|    Reading|
|  1|  Alice|   [Reading, Hiking]|     Hiking|
|  2|    Bob|[Swimming, Garden...|   Swimming|
|  2|    Bob|[Swimming, Garden...|  Gardening|
|  2|    Bob|[Swimming, Garden...|   Painting|
|  3|Charlie|           [Cooking]|    Cooking|
|  4|  David|[Photography, Ski...|Photography|
|  4|  David|[Photography, Ski...|     Skiing|
|  4|  David|[Photography, Ski...|    Cooking|
+---+-------+--------------------+-----------+
```

In [62]:
```
df2.withColumn("Hobby",explode("Hobbies")).show()
```

```
+---+-------+--------------------+-----------+
| id|   name|             Hobbies|      Hobby|
+---+-------+--------------------+-----------+
|  1|  Alice|   [Reading, Hiking]|    Reading|
|  1|  Alice|   [Reading, Hiking]|     Hiking|
|  2|    Bob|[Swimming, Garden...|   Swimming|
|  2|    Bob|[Swimming, Garden...|  Gardening|
|  2|    Bob|[Swimming, Garden...|   Painting|
```



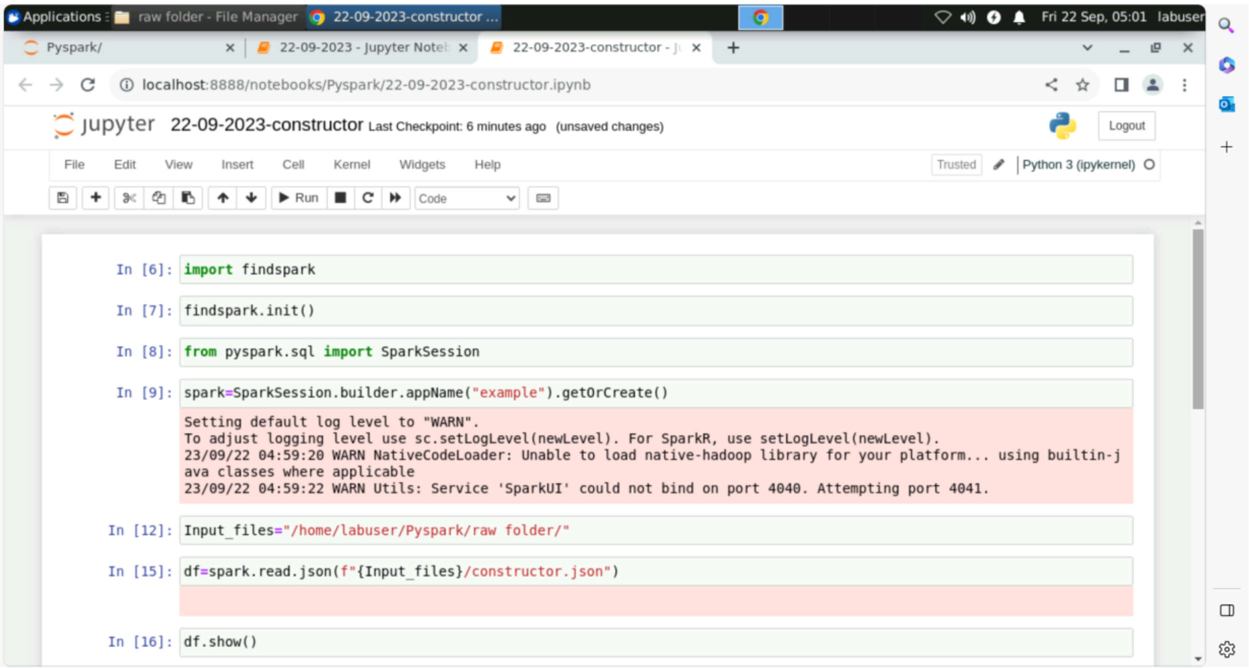
```
|  4|  David|[Photography, Ski...|     Skiing|
|  4|  David|[Photography, Ski...|    Cooking|
+---+-------+--------------------+-----------+
```

In [67]:
```
df3=df2\
.withColumn("Hobby",explode("Hobbies"))\
.withColumn("ingestion_data",current_timestamp())
```

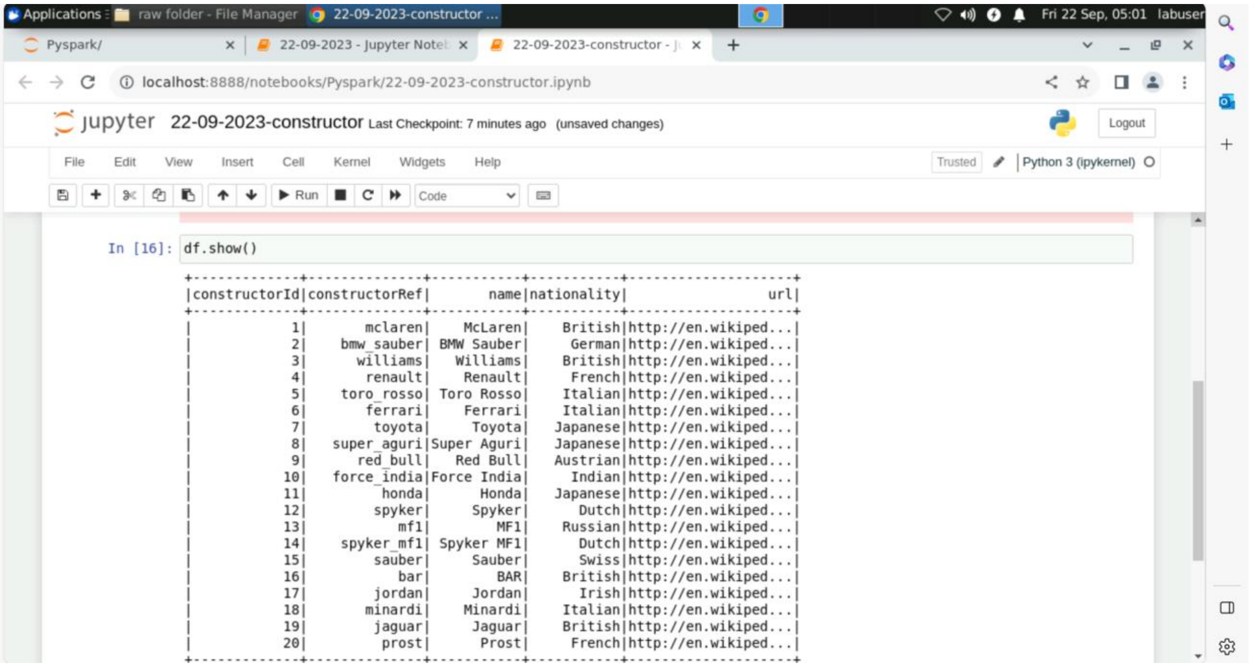
In [69]:
```
df3.show()(truncate=False)
```

```
+---+-------+--------------------+-----------+--------------------+
| id|   name|             Hobbies|      Hobby|      ingestion_data|
+---+-------+--------------------+-----------+--------------------+
|  1|  Alice|   [Reading, Hiking]|    Reading|2023-09-22 04:51:...|
|  1|  Alice|   [Reading, Hiking]|     Hiking|2023-09-22 04:51:...|
|  2|    Bob|[Swimming, Garden...|   Swimming|2023-09-22 04:51:...|
|  2|    Bob|[Swimming, Garden...|  Gardening|2023-09-22 04:51:...|
|  2|    Bob|[Swimming, Garden...|   Painting|2023-09-22 04:51:...|
|  3|Charlie|           [Cooking]|    Cooking|2023-09-22 04:51:...|
|  4|  David|[Photography, Ski...|Photography|2023-09-22 04:51:...|
|  4|  David|[Photography, Ski...|     Skiing|2023-09-22 04:51:...|
|  4|  David|[Photography, Ski...|    Cooking|2023-09-22 04:51:...|
+---+-------+--------------------+-----------+--------------------+
```

Constructor:



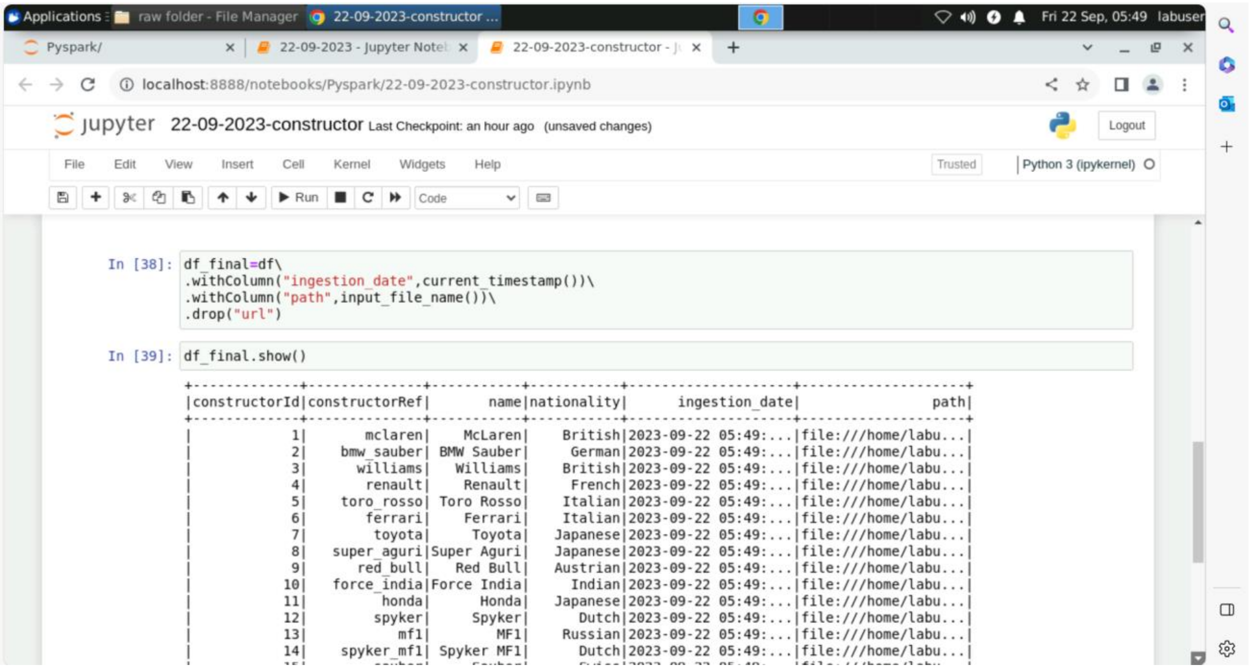
```
In [6]: import findspark

In [7]: findspark.init()

In [8]: from pyspark.sql import SparkSession

In [9]: spark=SparkSession.builder.appName("example").getOrCreate()

Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
23/09/22 04:59:20 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-j
ava classes where applicable
23/09/22 04:59:22 WARN Utils: Service 'SparkUI' could not bind on port 4040. Attempting port 4041.

In [12]: Input_files="/home/labuser/Pyspark/raw folder/"

In [15]: df=spark.read.json(f"{Input_files}/constructor.json")

In [16]: df.show()
```



```
In [16]: df.show()

+-------------+--------------+-----------+-----------+--------------------+
|constructorId|constructorRef|       name|nationality|                 url|
+-------------+--------------+-----------+-----------+--------------------+
|            1|       mclaren|    McLaren|    British|http://en.wikiped...|
|            2|    bmw_sauber| BMW Sauber|     German|http://en.wikiped...|
|            3|      williams|   Williams|    British|http://en.wikiped...|
|            4|       renault|    Renault|     French|http://en.wikiped...|
|            5|    toro_rosso| Toro Rosso|    Italian|http://en.wikiped...|
|            6|       ferrari|    Ferrari|    Italian|http://en.wikiped...|
|            7|        toyota|     Toyota|   Japanese|http://en.wikiped...|
|            8|   super_aguri|Super Aguri|   Japanese|http://en.wikiped...|
|            9|      red_bull|   Red Bull|   Austrian|http://en.wikiped...|
|           10|   force_india|Force India|     Indian|http://en.wikiped...|
|           11|         honda|      Honda|   Japanese|http://en.wikiped...|
|           12|        spyker|     Spyker|      Dutch|http://en.wikiped...|
|           13|           mf1|        MF1|    Russian|http://en.wikiped...|
|           14|    spyker_mf1| Spyker MF1|      Dutch|http://en.wikiped...|
|           15|        sauber|     Sauber|      Swiss|http://en.wikiped...|
|           16|           bar|        BAR|    British|http://en.wikiped...|
|           17|        jordan|     Jordan|      Irish|http://en.wikiped...|
|           18|       minardi|    Minardi|    Italian|http://en.wikiped...|
|           19|        jaguar|     Jaguar|    British|http://en.wikiped...|
|           20|         prost|      Prost|     French|http://en.wikiped...|
+-------------+--------------+-----------+-----------+--------------------+
```

from pyspark.sql.functions import *

from pyspark.sql.types import *


Applications
raw folder - File Manager
22-09-2023-constructor ...
Fri 22 Sep, 05:49 labuser
Pyspark/
22-09-2023 - Jupyter Noteb
22-09-2023-constructor - Ju
localhost:8888/notebooks/Pyspark/22-09-2023-constructor.ipynb
jupyter 22-09-2023-constructor Last Checkpoint: an hour ago (unsaved changes)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Python 3 (ipykernel)
Run
Code

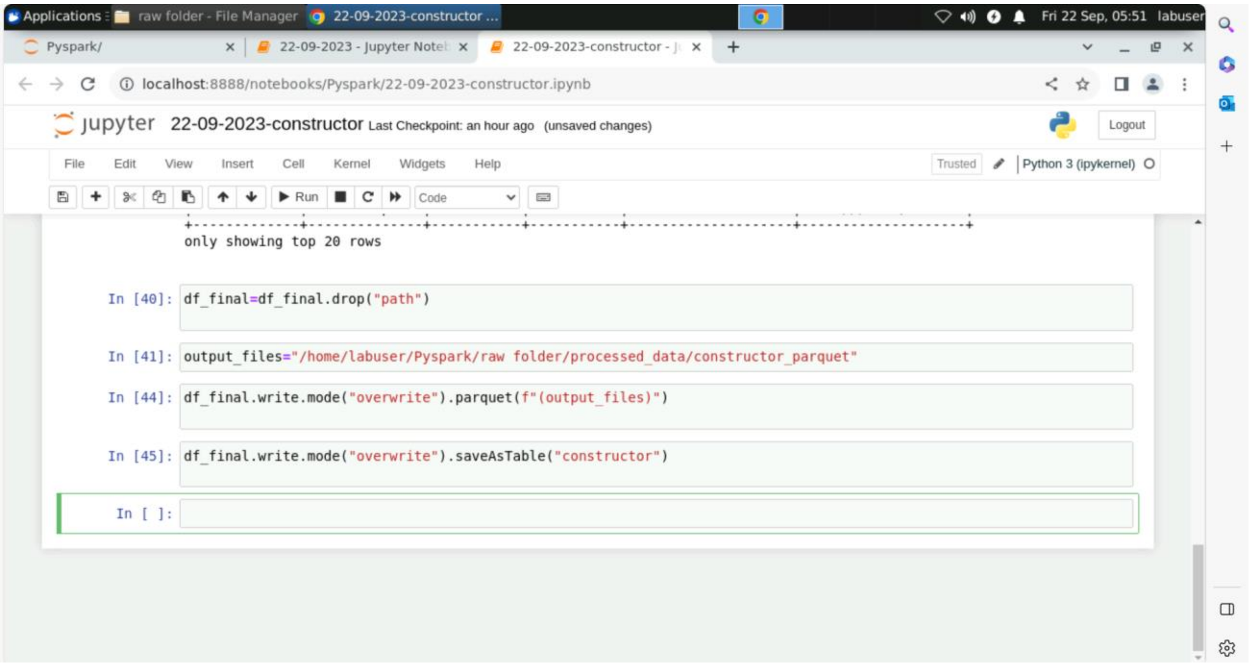
In [38]:

```
df_final=df\
.withColumn("ingestion_date",current_timestamp())\
.withColumn("path",input_file_name())\
.drop("url")
```

In [39]:

```
df_final.show()
```

```
+-------------+--------------+-----------+-----------+--------------------+--------------------+
|constructorId|constructorRef|       name|nationality|      ingestion_date|                path|
+-------------+--------------+-----------+-----------+--------------------+--------------------+
|            1|       mclaren|    McLaren|    British|2023-09-22 05:49:...|file:///home/labu...|
|            2|    bmw_sauber| BMW Sauber|     German|2023-09-22 05:49:...|file:///home/labu...|
|            3|      williams|   Williams|    British|2023-09-22 05:49:...|file:///home/labu...|
|            4|       renault|    Renault|     French|2023-09-22 05:49:...|file:///home/labu...|
|            5|    toro_rosso| Toro Rosso|    Italian|2023-09-22 05:49:...|file:///home/labu...|
|            6|       ferrari|    Ferrari|    Italian|2023-09-22 05:49:...|file:///home/labu...|
|            7|        toyota|     Toyota|   Japanese|2023-09-22 05:49:...|file:///home/labu...|
|            8|   super_aguri|Super Aguri|   Japanese|2023-09-22 05:49:...|file:///home/labu...|
|            9|      red_bull|   Red Bull|   Austrian|2023-09-22 05:49:...|file:///home/labu...|
|           10|   force_india|Force India|     Indian|2023-09-22 05:49:...|file:///home/labu...|
|           11|         honda|      Honda|   Japanese|2023-09-22 05:49:...|file:///home/labu...|
|           12|        spyker|     Spyker|      Dutch|2023-09-22 05:49:...|file:///home/labu...|
|           13|           mf1|        MF1|    Russian|2023-09-22 05:49:...|file:///home/labu...|
|           14|    spyker_mf1| Spyker MF1|      Dutch|2023-09-22 05:49:...|file:///home/labu...|
```


Applications
raw folder - File Manager
22-09-2023-constructor ...
Fri 22 Sep, 05:51 labuser
Pyspark/
22-09-2023 - Jupyter Noteb
22-09-2023-constructor - Ju
localhost:8888/notebooks/Pyspark/22-09-2023-constructor.ipynb
jupyter 22-09-2023-constructor 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 [40]:

```
df_final=df_final.drop("path")
```

In [41]:

```
output_files="/home/labuser/Pyspark/raw folder/processed_data/constructor_parquet"
```

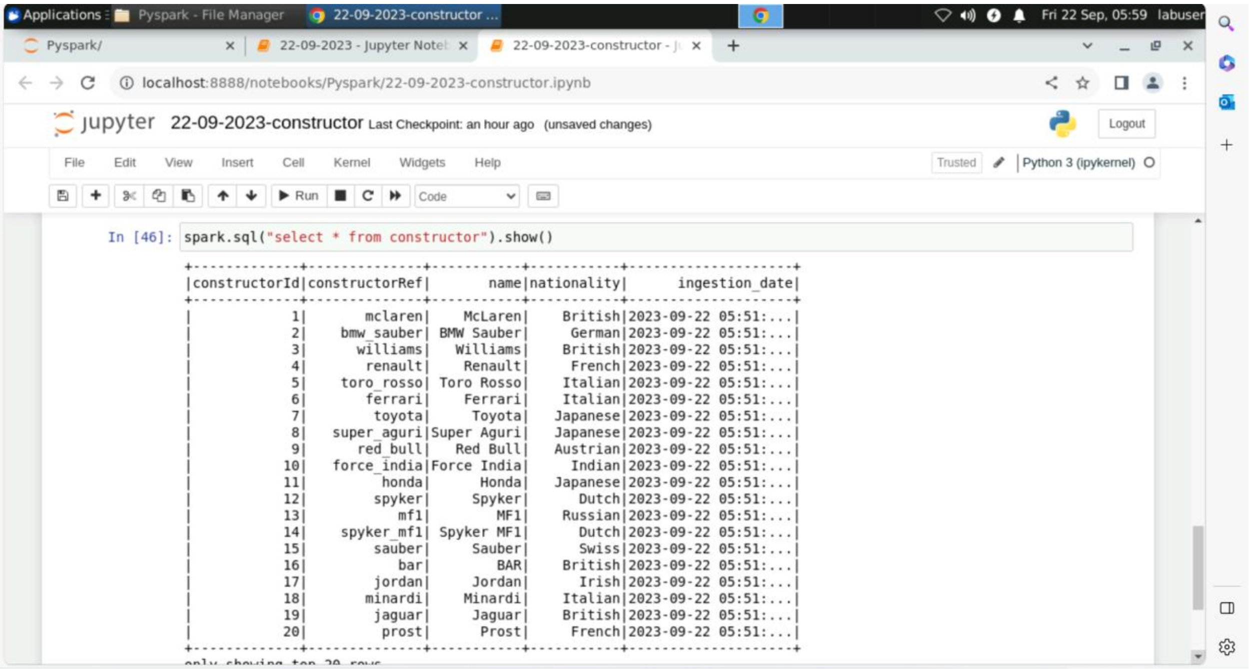
In [44]:

```
df_final.write.mode("overwrite").parquet(f"(output_files)")
```

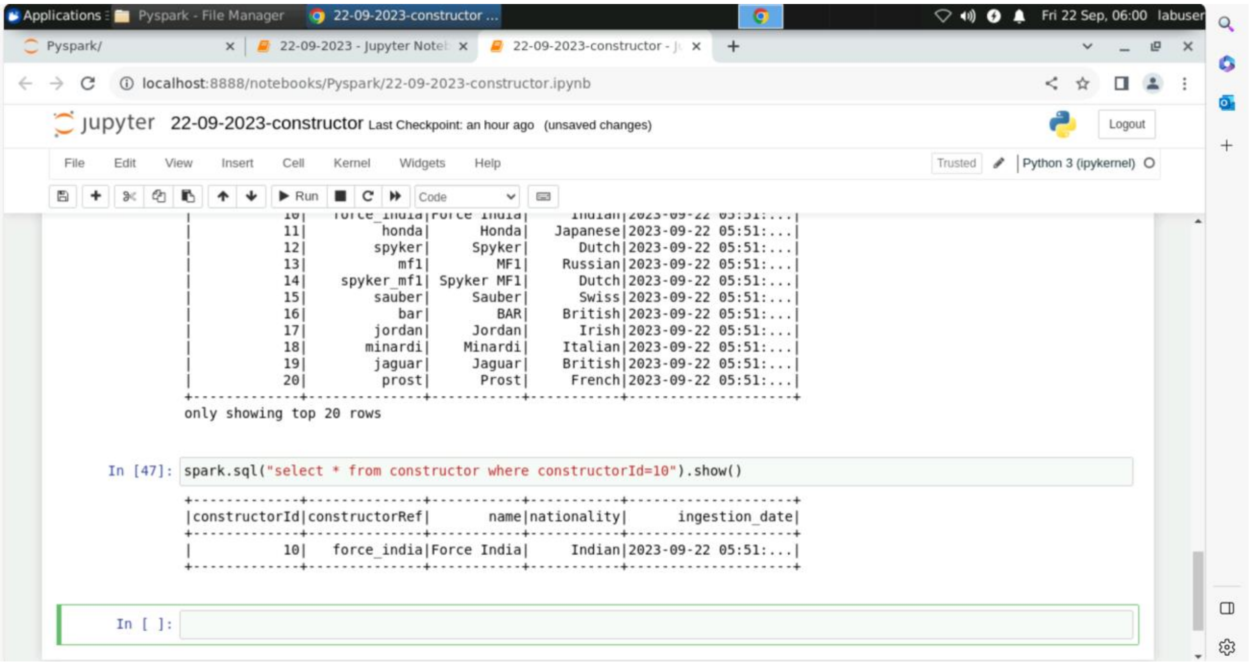
In [45]:

```
df_final.write.mode("overwrite").saveAsTable("constructor")
```

In [ ]:

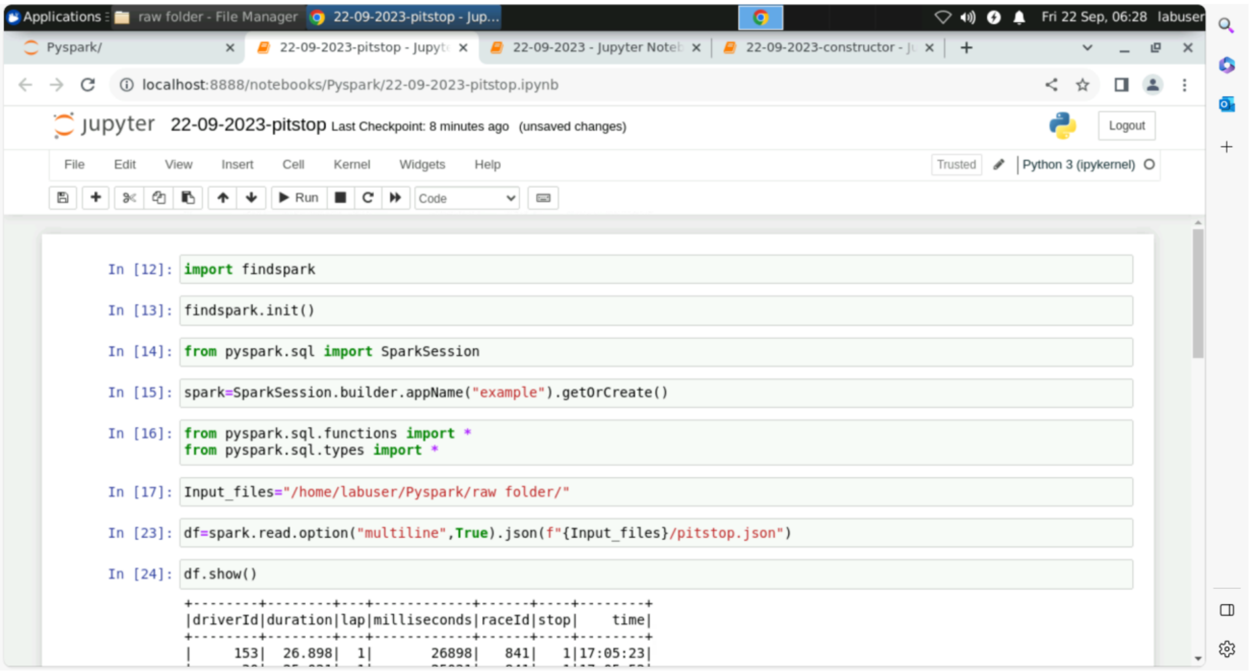

Applications
Pyspark - File Manager
22-09-2023-constructor ...
Fri 22 Sep, 05:59 labuser
Pyspark/
22-09-2023 - Jupyter Noteb
22-09-2023-constructor - J
localhost:8888/notebooks/Pyspark/22-09-2023-constructor.ipynb
jupyter 22-09-2023-constructor Last Checkpoint: an hour ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code

```
In [46]: spark.sql("select * from constructor").show()

         +-------------+-------------+-----------+-----------+--------------------+
         |constructorId|constructorRef|       name|nationality|      ingestion_date|
         +-------------+-------------+-----------+-----------+--------------------+
         |            1|      mclaren|    McLaren|    British|2023-09-22 05:51:...|
         |            2|   bmw_sauber| BMW Sauber|     German|2023-09-22 05:51:...|
         |            3|     williams|   Williams|    British|2023-09-22 05:51:...|
         |            4|      renault|    Renault|     French|2023-09-22 05:51:...|
         |            5|   toro_rosso| Toro Rosso|    Italian|2023-09-22 05:51:...|
         |            6|      ferrari|    Ferrari|    Italian|2023-09-22 05:51:...|
         |            7|       toyota|     Toyota|   Japanese|2023-09-22 05:51:...|
         |            8|  super_aguri|Super Aguri|   Japanese|2023-09-22 05:51:...|
         |            9|     red_bull|   Red Bull|   Austrian|2023-09-22 05:51:...|
         |           10|  force_india|Force India|     Indian|2023-09-22 05:51:...|
         |           11|        honda|      Honda|   Japanese|2023-09-22 05:51:...|
         |           12|       spyker|     Spyker|      Dutch|2023-09-22 05:51:...|
         |           13|          mf1|        MF1|    Russian|2023-09-22 05:51:...|
         |           14|   spyker_mf1|  Spyker MF1|     Dutch|2023-09-22 05:51:...|
         |           15|       sauber|     Sauber|      Swiss|2023-09-22 05:51:...|
         |           16|          bar|        BAR|    British|2023-09-22 05:51:...|
         |           17|       jordan|     Jordan|      Irish|2023-09-22 05:51:...|
         |           18|      minardi|    Minardi|    Italian|2023-09-22 05:51:...|
         |           19|       jaguar|     Jaguar|    British|2023-09-22 05:51:...|
         |           20|        prost|      Prost|     French|2023-09-22 05:51:...|
         +-------------+-------------+-----------+-----------+--------------------+
         only showing top 20 rows
```

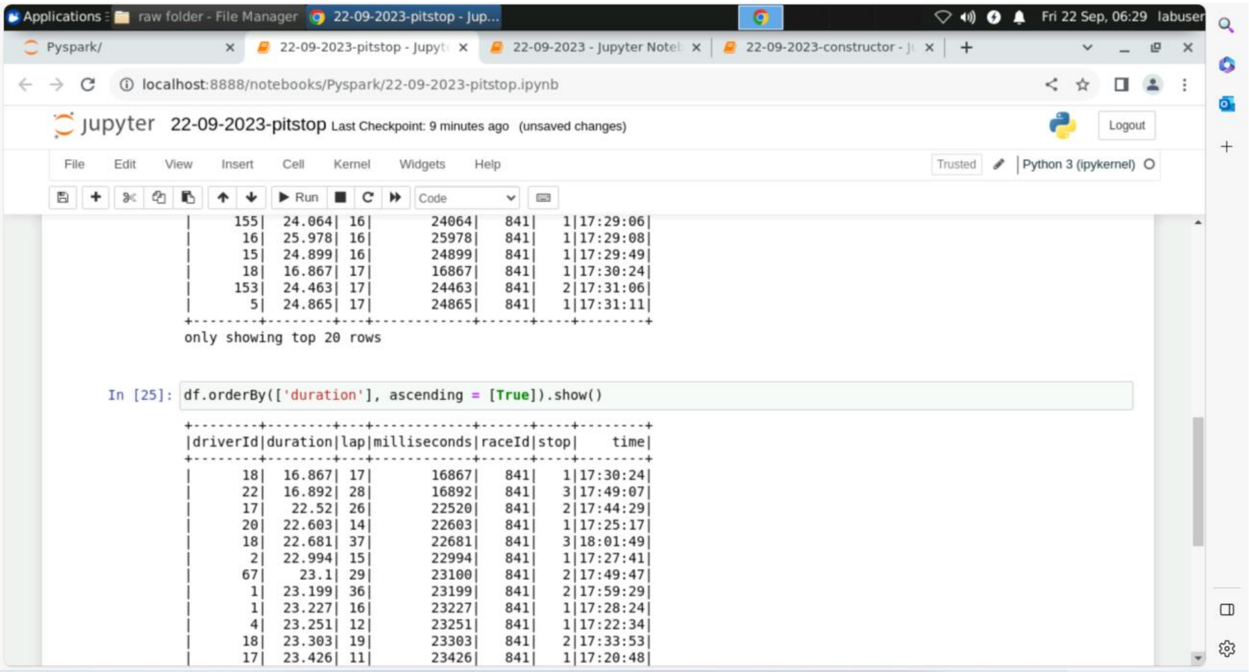

Applications
Pyspark - File Manager
22-09-2023-constructor ...
Fri 22 Sep, 06:00 labuser
Pyspark/
22-09-2023 - Jupyter Noteb
22-09-2023-constructor - J
localhost:8888/notebooks/Pyspark/22-09-2023-constructor.ipynb
jupyter 22-09-2023-constructor Last Checkpoint: an hour ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code

```
         |           10|  force_india|Force India|     Indian|2023-09-22 05:51:...|
         |           11|        honda|      Honda|   Japanese|2023-09-22 05:51:...|
         |           12|       spyker|     Spyker|      Dutch|2023-09-22 05:51:...|
         |           13|          mf1|        MF1|    Russian|2023-09-22 05:51:...|
         |           14|   spyker_mf1|  Spyker MF1|     Dutch|2023-09-22 05:51:...|
         |           15|       sauber|     Sauber|      Swiss|2023-09-22 05:51:...|
         |           16|          bar|        BAR|    British|2023-09-22 05:51:...|
         |           17|       jordan|     Jordan|      Irish|2023-09-22 05:51:...|
         |           18|      minardi|    Minardi|    Italian|2023-09-22 05:51:...|
         |           19|       jaguar|     Jaguar|    British|2023-09-22 05:51:...|
         |           20|        prost|      Prost|     French|2023-09-22 05:51:...|
         +-------------+-------------+-----------+-----------+--------------------+
         only showing top 20 rows

In [47]: spark.sql("select * from constructor where constructorId=10").show()

         +-------------+-------------+-----------+-----------+--------------------+
         |constructorId|constructorRef|       name|nationality|      ingestion_date|
         +-------------+-------------+-----------+-----------+--------------------+
         |           10|  force_india|Force India|     Indian|2023-09-22 05:51:...|
         +-------------+-------------+-----------+-----------+--------------------+

In [ ]:
```

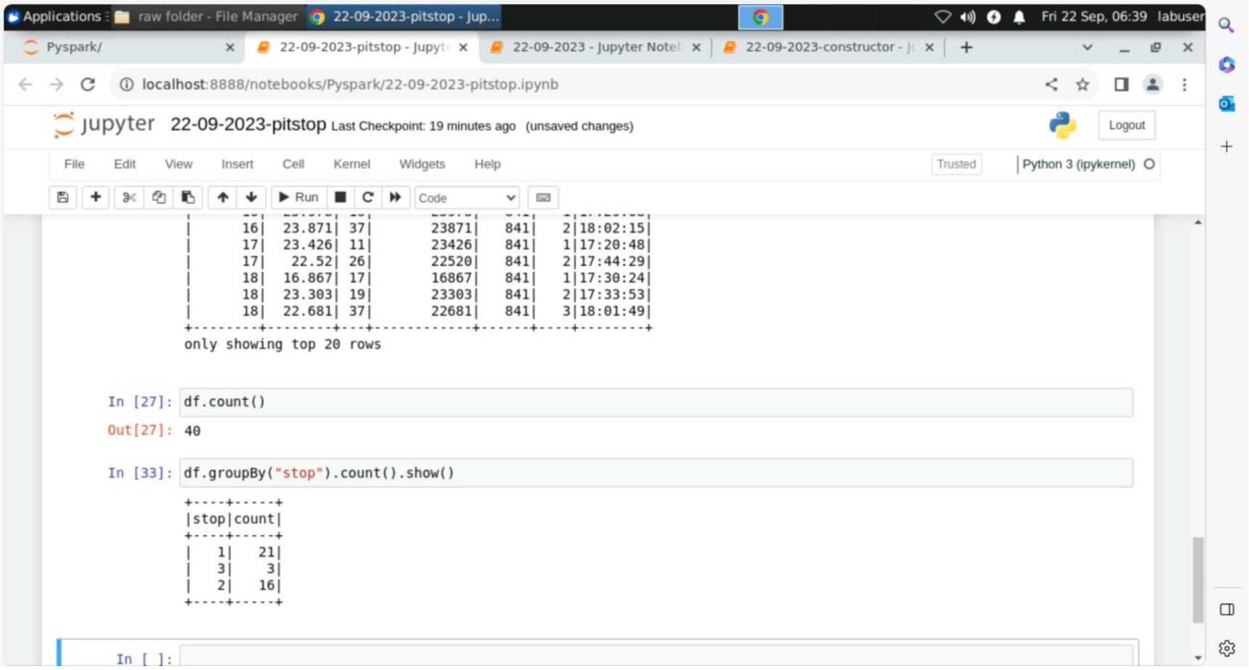
df_final.write.mode("overwrite").option("path","/home/labuser/Pyspark/raw folder/processed_data/constructor_table").saveAsTable("constructor")


Applications
raw folder - File Manager
22-09-2023-pitstop - Jup...
Fri 22 Sep, 06:28 labuser
Pyspark/
22-09-2023-pitstop - Jupyt
22-09-2023 - Jupyter Noteb
22-09-2023-constructor - J
localhost:8888/notebooks/Pyspark/22-09-2023-pitstop.ipynb
jupyter
22-09-2023-pitstop Last Checkpoint: 8 minutes ago (unsaved changes)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Python 3 (ipykernel)
Run
Code

```
In [12]: import findspark

In [13]: findspark.init()

In [14]: from pyspark.sql import SparkSession

In [15]: spark=SparkSession.builder.appName("example").getOrCreate()

In [16]: from pyspark.sql.functions import *
         from pyspark.sql.types import *

In [17]: Input_files="/home/labuser/Pyspark/raw folder/"

In [23]: df=spark.read.option("multiline",True).json(f"{Input_files}/pitstop.json")

In [24]: df.show()

         +--------+--------+---+------------+------+----+--------+
         |driverId|duration|lap|milliseconds|raceId|stop|    time|
         +--------+--------+---+------------+------+----+--------+
         |     153|  26.898|  1|       26898|   841|   1|17:05:23|
```


Applications
raw folder - File Manager
22-09-2023-pitstop - Jup...
Fri 22 Sep, 06:29 labuser
Pyspark/
22-09-2023-pitstop - Jupyt
22-09-2023 - Jupyter Noteb
22-09-2023-constructor - J
localhost:8888/notebooks/Pyspark/22-09-2023-pitstop.ipynb
jupyter
22-09-2023-pitstop Last Checkpoint: 9 minutes ago (unsaved changes)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Python 3 (ipykernel)
Run
Code

```
         |     155|  24.064| 16|       24064|   841|   1|17:29:06|
         |      16|  25.978| 16|       25978|   841|   1|17:29:08|
         |      15|  24.899| 16|       24899|   841|   1|17:29:49|
         |      18|  16.867| 17|       16867|   841|   1|17:30:24|
         |     153|  24.463| 17|       24463|   841|   2|17:31:06|
         |       5|  24.865| 17|       24865|   841|   1|17:31:11|
         +--------+--------+---+------------+------+----+--------+
         only showing top 20 rows


In [25]: df.orderBy(['duration'], ascending = [True]).show()

         +--------+--------+---+------------+------+----+--------+
         |driverId|duration|lap|milliseconds|raceId|stop|    time|
         +--------+--------+---+------------+------+----+--------+
         |      18|  16.867| 17|       16867|   841|   1|17:30:24|
         |      22|  16.892| 28|       16892|   841|   3|17:49:07|
         |      17|   22.52| 26|       22520|   841|   2|17:44:29|
         |      20|  22.603| 14|       22603|   841|   1|17:25:17|
         |      18|  22.681| 37|       22681|   841|   3|18:01:49|
         |       2|  22.994| 15|       22994|   841|   1|17:27:41|
         |      67|    23.1| 29|       23100|   841|   2|17:49:47|
         |       1|  23.199| 36|       23199|   841|   2|17:59:29|
         |       1|  23.227| 16|       23227|   841|   1|17:28:24|
         |       4|  23.251| 12|       23251|   841|   1|17:22:34|
         |      18|  23.303| 19|       23303|   841|   2|17:33:53|
         |      17|  23.426| 11|       23426|   841|   1|17:20:48|
```


Applications
raw folder - File Manager
22-09-2023-pitstop - Jup...
Fri 22 Sep, 06:39 labuser
Pyspark/
22-09-2023-pitstop - Jupyt
22-09-2023 - Jupyter Noteb
22-09-2023-constructor - J
localhost:8888/notebooks/Pyspark/22-09-2023-pitstop.ipynb
jupyter 22-09-2023-pitstop Last Checkpoint: 19 minutes ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code

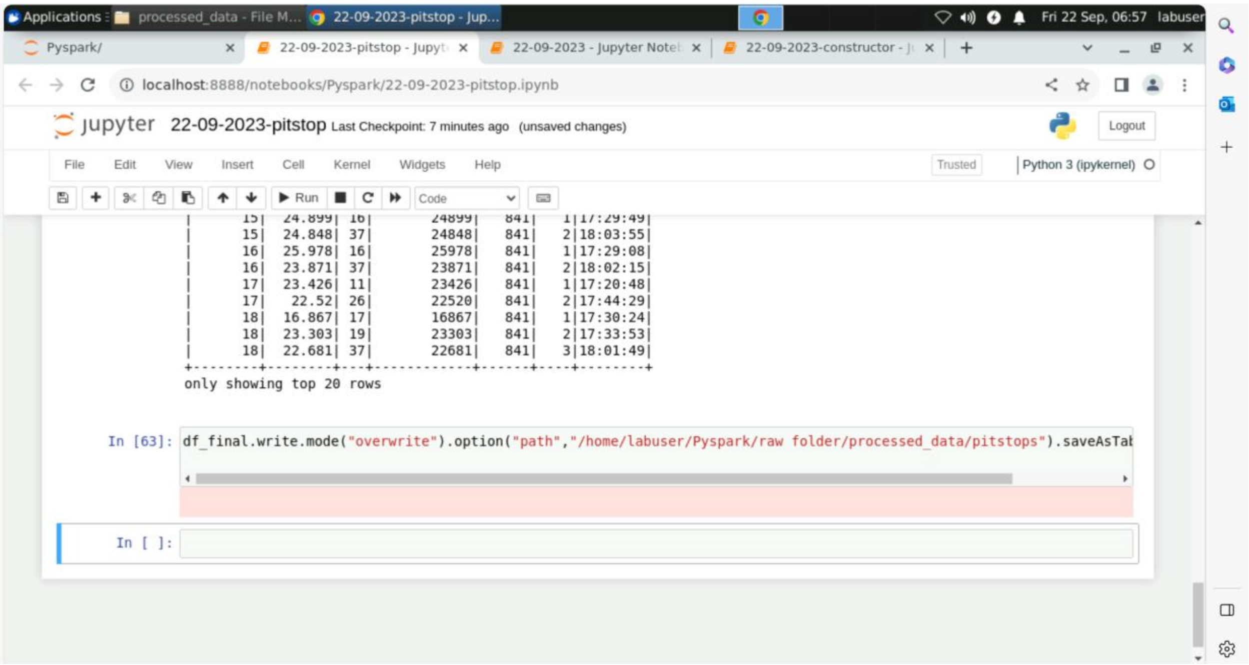
```
|       16| 23.871| 37|       23871|   841|   2|18:02:15|
|       17| 23.426| 11|       23426|   841|   1|17:20:48|
|       17|  22.52| 26|       22520|   841|   2|17:44:29|
|       18| 16.867| 17|       16867|   841|   1|17:30:24|
|       18| 23.303| 19|       23303|   841|   2|17:33:53|
|       18| 22.681| 37|       22681|   841|   3|18:01:49|
+---------+-------+---+------------+------+----+--------+
only showing top 20 rows
```

```
In [27]: df.count()
Out[27]: 40
```

```
In [33]: df.groupBy("stop").count().show()
```

```
+----+-----+
|stop|count|
+----+-----+
|   1|   21|
|   3|    3|
|   2|   16|
+----+-----+
```

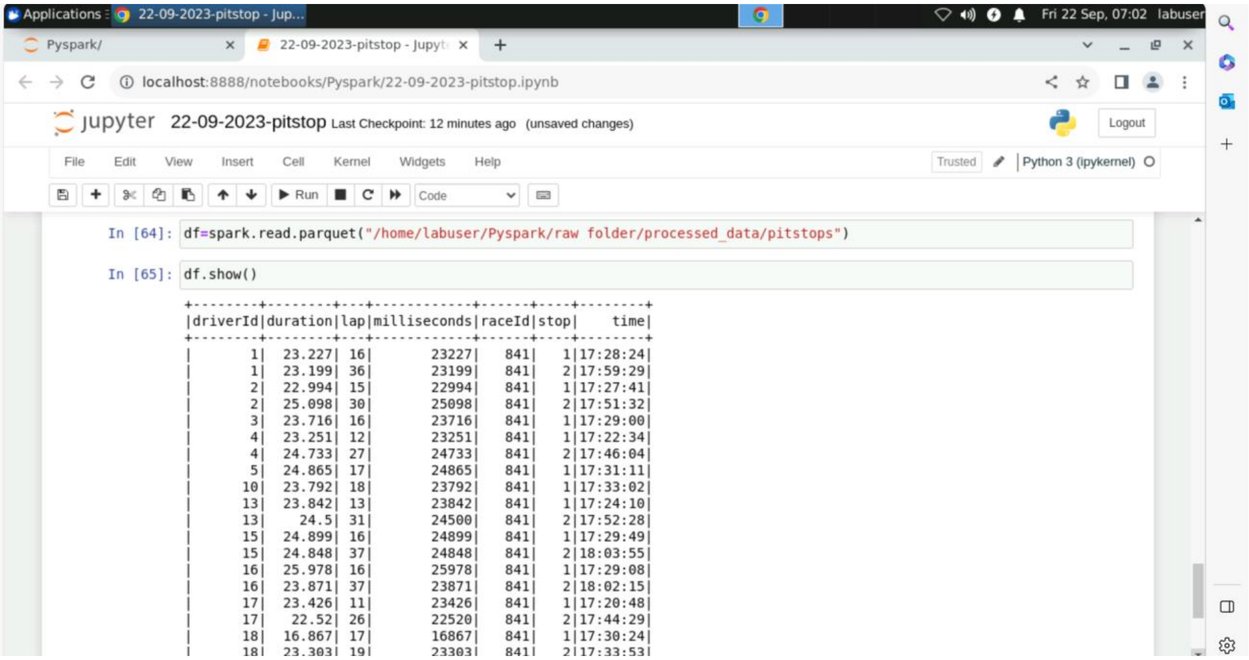
```
In [ ]:
```


Applications
processed_data - File M...
22-09-2023-pitstop - Jup...
Fri 22 Sep, 06:57 labuser
Pyspark/
22-09-2023-pitstop - Jupyt
22-09-2023 - Jupyter Noteb
22-09-2023-constructor - J
localhost:8888/notebooks/Pyspark/22-09-2023-pitstop.ipynb
jupyter 22-09-2023-pitstop Last Checkpoint: 7 minutes ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code

```
|       15| 24.899| 16|       24899|   841|   1|17:29:49|
|       15| 24.848| 37|       24848|   841|   2|18:03:55|
|       16| 25.978| 16|       25978|   841|   1|17:29:08|
|       16| 23.871| 37|       23871|   841|   2|18:02:15|
|       17| 23.426| 11|       23426|   841|   1|17:20:48|
|       17|  22.52| 26|       22520|   841|   2|17:44:29|
|       18| 16.867| 17|       16867|   841|   1|17:30:24|
|       18| 23.303| 19|       23303|   841|   2|17:33:53|
|       18| 22.681| 37|       22681|   841|   3|18:01:49|
+---------+-------+---+------------+------+----+--------+
only showing top 20 rows
```

```
In [63]: df_final.write.mode("overwrite").option("path","/home/labuser/Pyspark/raw folder/processed_data/pitstops").saveAsTat
```

```
In [ ]:
```



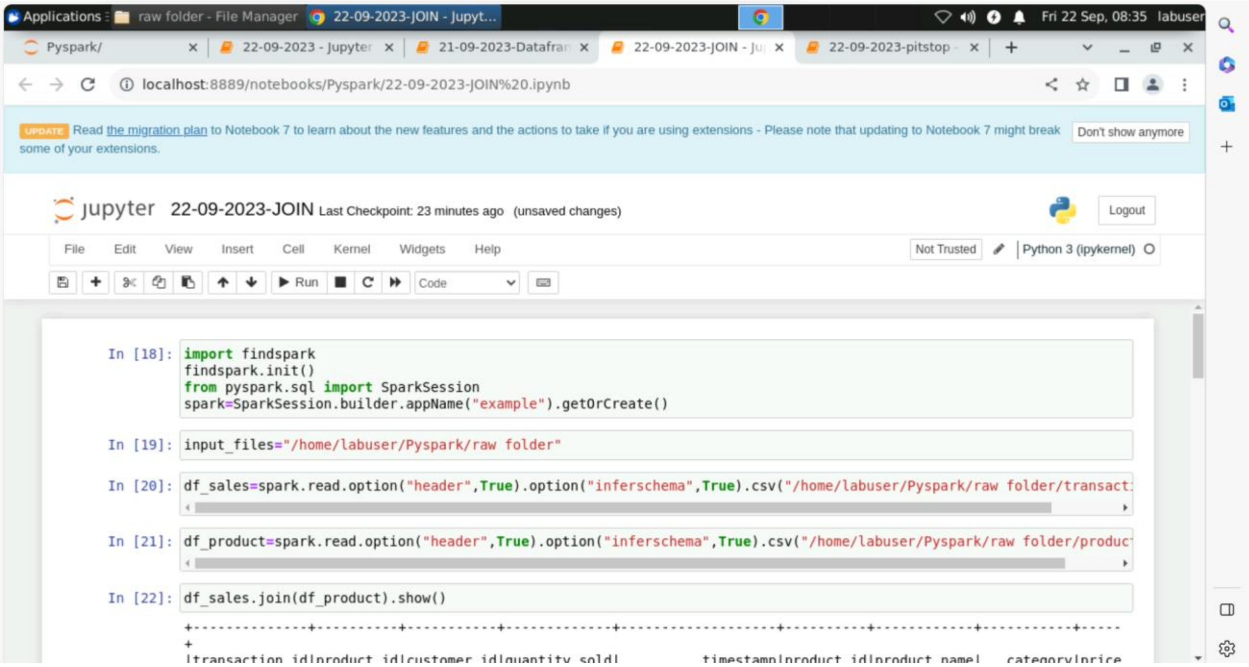
Applications
22-09-2023-pitstop - Jup...
Fri 22 Sep, 07:02 labuser
Pyspark/
22-09-2023-pitstop - Jupyte
localhost:8888/notebooks/Pyspark/22-09-2023-pitstop.ipynb
jupyter 22-09-2023-pitstop Last Checkpoint: 12 minutes ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code

```
In [64]: df=spark.read.parquet("/home/labuser/Pyspark/raw folder/processed_data/pitstops")
```

```
In [65]: df.show()
```

```
+--------+--------+---+------------+------+----+--------+
|driverId|duration|lap|milliseconds|raceId|stop|    time|
+--------+--------+---+------------+------+----+--------+
|       1|  23.227| 16|       23227|   841|   1|17:28:24|
|       1|  23.199| 36|       23199|   841|   2|17:59:29|
|       2|  22.994| 15|       22994|   841|   1|17:27:41|
|       2|  25.098| 30|       25098|   841|   2|17:51:32|
|       3|  23.716| 16|       23716|   841|   1|17:29:00|
|       4|  23.251| 12|       23251|   841|   1|17:22:34|
|       4|  24.733| 27|       24733|   841|   2|17:46:04|
|       5|  24.865| 17|       24865|   841|   1|17:31:11|
|      10|  23.792| 18|       23792|   841|   1|17:33:02|
|      13|  23.842| 13|       23842|   841|   1|17:24:10|
|      13|    24.5| 31|       24500|   841|   2|17:52:28|
|      15|  24.899| 16|       24899|   841|   1|17:29:49|
|      15|  24.848| 37|       24848|   841|   2|18:03:55|
|      16|  25.978| 16|       25978|   841|   1|17:29:08|
|      16|  23.871| 37|       23871|   841|   2|18:02:15|
|      17|  23.426| 11|       23426|   841|   1|17:20:48|
|      17|   22.52| 26|       22520|   841|   2|17:44:29|
|      18|  16.867| 17|       16867|   841|   1|17:30:24|
|      18|  23.303| 19|       23303|   841|   2|17:33:53|
```

JOIN:



Applications
raw folder - File Manager
22-09-2023-JOIN - Jupyt...
Fri 22 Sep, 08:35 labuser
Pyspark/
22-09-2023 - Jupyter
21-09-2023-Datafran
22-09-2023-JOIN - Ju
22-09-2023-pitstop -
localhost:8889/notebooks/Pyspark/22-09-2023-JOIN%20.ipynb
UPDATE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions.
Don't show anymore
jupyter 22-09-2023-JOIN Last Checkpoint: 23 minutes ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code

```
In [18]: import findspark
         findspark.init()
         from pyspark.sql import SparkSession
         spark=SparkSession.builder.appName("example").getOrCreate()
```

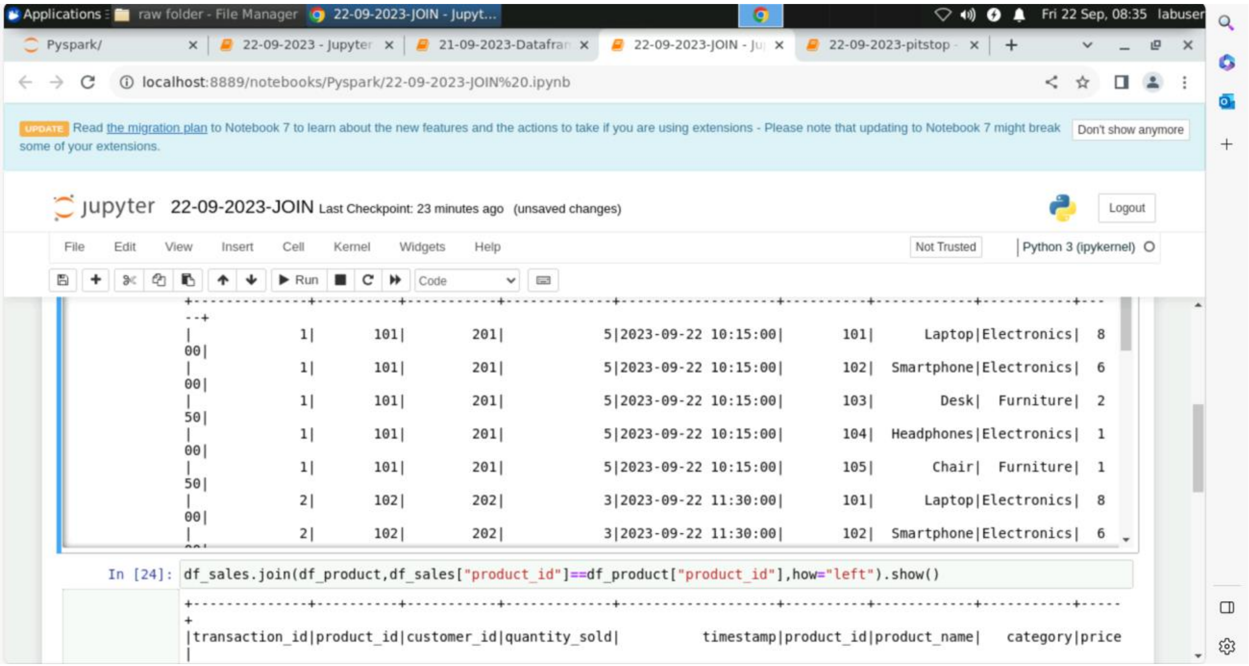
```
In [19]: input_files="/home/labuser/Pyspark/raw folder"
```

```
In [20]: df_sales=spark.read.option("header",True).option("inferschema",True).csv("/home/labuser/Pyspark/raw folder/transact
```

```
In [21]: df_product=spark.read.option("header",True).option("inferschema",True).csv("/home/labuser/Pyspark/raw folder/produc
```

```
In [22]: df_sales.join(df_product).show()
```

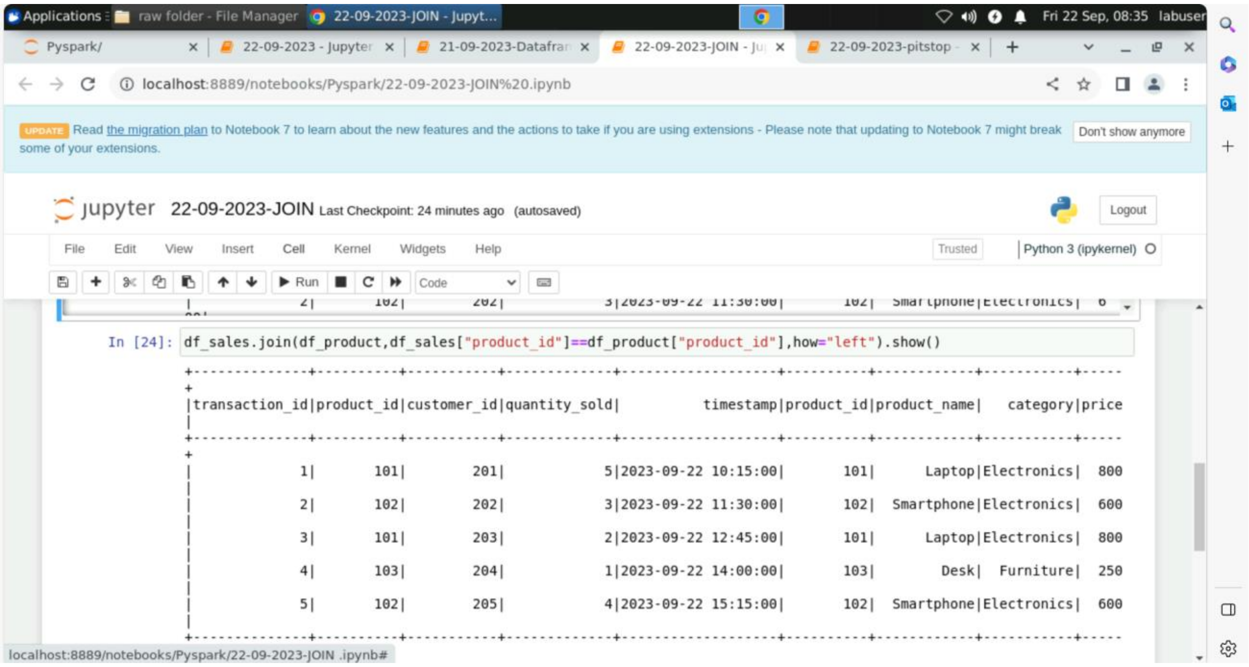
```
+--------------+----------+-----------+-------------+-------------------+----------+------------+-----------+-----
+
|transaction_id|product_id|customer_id|quantity_sold|          timestamp|product_id|product_name|   category|price
```


Applications
raw folder - File Manager
22-09-2023-JOIN - Jupyt...
Fri 22 Sep, 08:35 labuser
Pyspark/
22-09-2023 - Jupyter
21-09-2023-Datafran
22-09-2023-JOIN - Ju
22-09-2023-pitstop -
localhost:8889/notebooks/Pyspark/22-09-2023-JOIN%20.ipynb
UPDATE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions.
Don't show anymore
jupyter 22-09-2023-JOIN Last Checkpoint: 23 minutes ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code

```
+-------------+----------+-----------+-------------+-------------------+----------+------------+-----------+---
--+
|            1|       101|        201|            5|2023-09-22 10:15:00|       101|      Laptop|Electronics|  8
00|
|            1|       101|        201|            5|2023-09-22 10:15:00|       102|  Smartphone|Electronics|  6
00|
|            1|       101|        201|            5|2023-09-22 10:15:00|       103|        Desk|  Furniture|  2
50|
|            1|       101|        201|            5|2023-09-22 10:15:00|       104|  Headphones|Electronics|  1
00|
|            1|       101|        201|            5|2023-09-22 10:15:00|       105|       Chair|  Furniture|  1
50|
|            2|       102|        202|            3|2023-09-22 11:30:00|       101|      Laptop|Electronics|  8
00|
|            2|       102|        202|            3|2023-09-22 11:30:00|       102|  Smartphone|Electronics|  6
```

```
In [24]: df_sales.join(df_product,df_sales["product_id"]==df_product["product_id"],how="left").show()
```

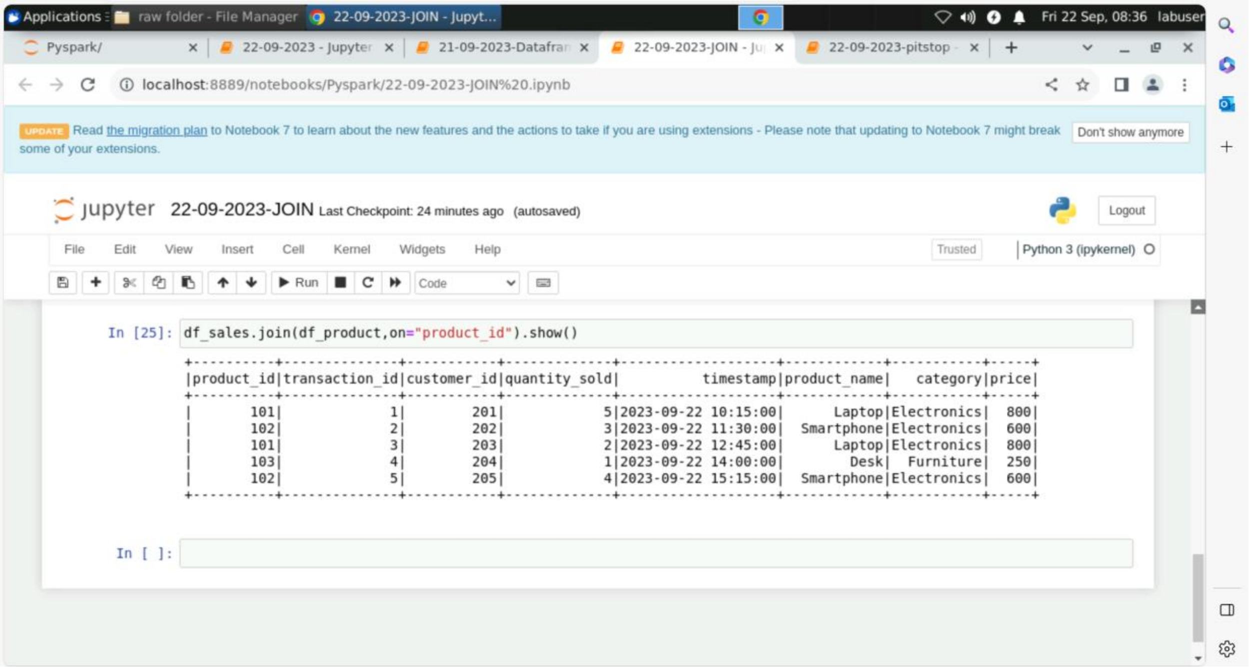
```
+-------------+----------+-----------+-------------+-------------------+----------+------------+-----------+-----
+
|transaction_id|product_id|customer_id|quantity_sold|          timestamp|product_id|product_name|   category|price
|
```


Applications
raw folder - File Manager
22-09-2023-JOIN - Jupyt...
Fri 22 Sep, 08:35 labuser
Pyspark/
22-09-2023 - Jupyter
21-09-2023-Datafran
22-09-2023-JOIN - Ju
22-09-2023-pitstop -
localhost:8889/notebooks/Pyspark/22-09-2023-JOIN%20.ipynb
UPDATE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions.
Don't show anymore
jupyter 22-09-2023-JOIN Last Checkpoint: 24 minutes ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code
localhost:8889/notebooks/Pyspark/22-09-2023-JOIN .ipynb#

```
|            2|       102|        202|            3|2023-09-22 11:30:00|       102|  Smartphone|Electronics|  6
```

```
In [24]: df_sales.join(df_product,df_sales["product_id"]==df_product["product_id"],how="left").show()
```

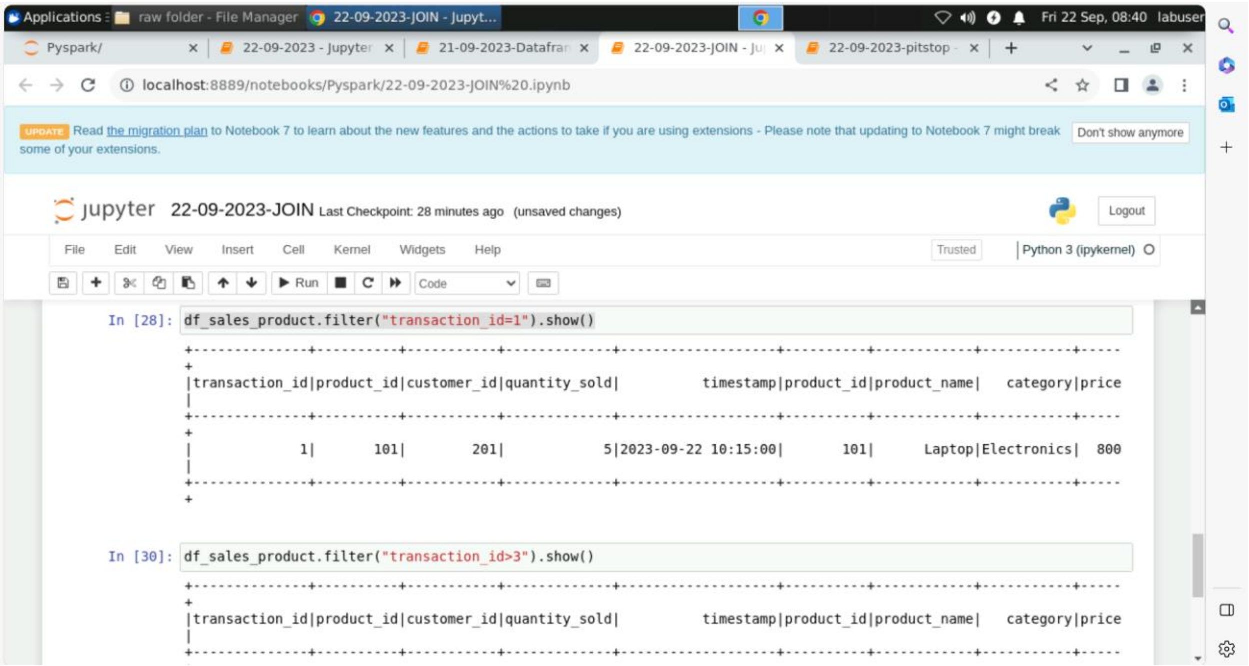
```
+-------------+----------+-----------+-------------+-------------------+----------+------------+-----------+-----
+
|transaction_id|product_id|customer_id|quantity_sold|          timestamp|product_id|product_name|   category|price
|
+-------------+----------+-----------+-------------+-------------------+----------+------------+-----------+-----
+
|            1|       101|        201|            5|2023-09-22 10:15:00|       101|      Laptop|Electronics|  800
|
|            2|       102|        202|            3|2023-09-22 11:30:00|       102|  Smartphone|Electronics|  600
|
|            3|       101|        203|            2|2023-09-22 12:45:00|       101|      Laptop|Electronics|  800
|
|            4|       103|        204|            1|2023-09-22 14:00:00|       103|        Desk|  Furniture|  250
|
|            5|       102|        205|            4|2023-09-22 15:15:00|       102|  Smartphone|Electronics|  600
|
+-------------+----------+-----------+-------------+-------------------+----------+------------+-----------+-----
```



UPDATE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions.

22-09-2023-JOIN Last Checkpoint: 24 minutes ago (autosaved)

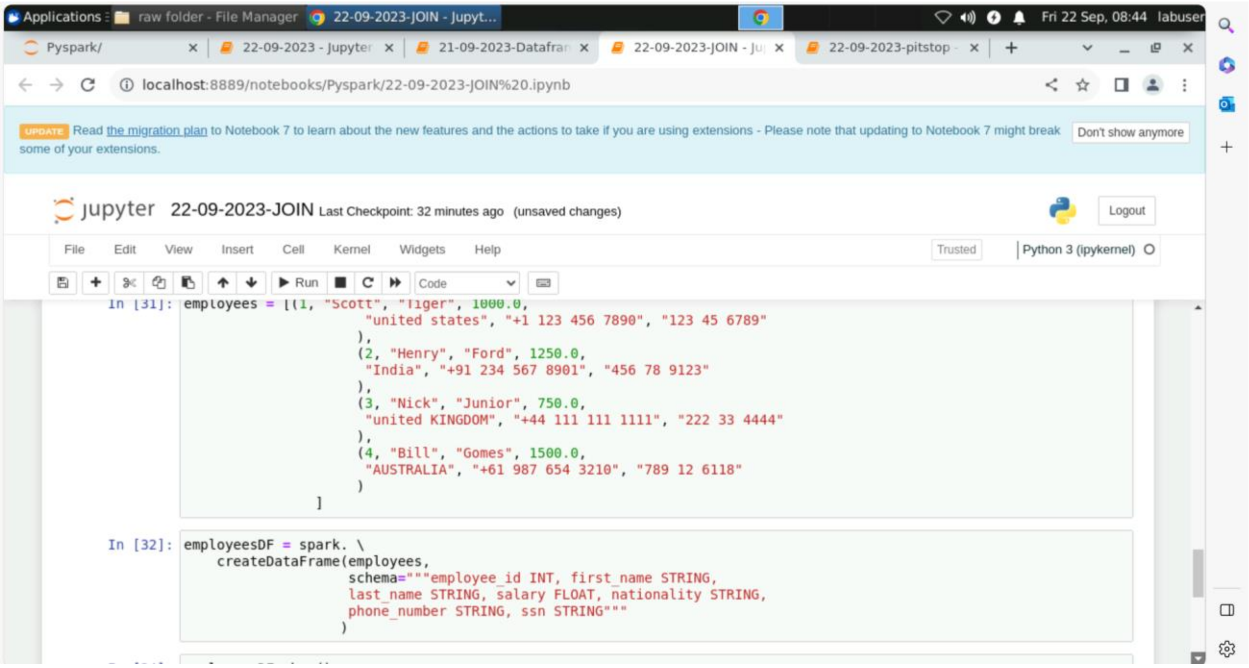
```
In [25]: df_sales.join(df_product,on="product_id").show()

+----------+--------------+-----------+-------------+-------------------+------------+-----------+-----+
|product_id|transaction_id|customer_id|quantity_sold|          timestamp|product_name|   category|price|
+----------+--------------+-----------+-------------+-------------------+------------+-----------+-----+
|       101|             1|        201|            5|2023-09-22 10:15:00|      Laptop|Electronics|  800|
|       102|             2|        202|            3|2023-09-22 11:30:00|  Smartphone|Electronics|  600|
|       101|             3|        203|            2|2023-09-22 12:45:00|      Laptop|Electronics|  800|
|       103|             4|        204|            1|2023-09-22 14:00:00|        Desk|  Furniture|  250|
|       102|             5|        205|            4|2023-09-22 15:15:00|  Smartphone|Electronics|  600|
+----------+--------------+-----------+-------------+-------------------+------------+-----------+-----+

In [ ]:
```



UPDATE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions.

22-09-2023-JOIN Last Checkpoint: 28 minutes ago (unsaved changes)

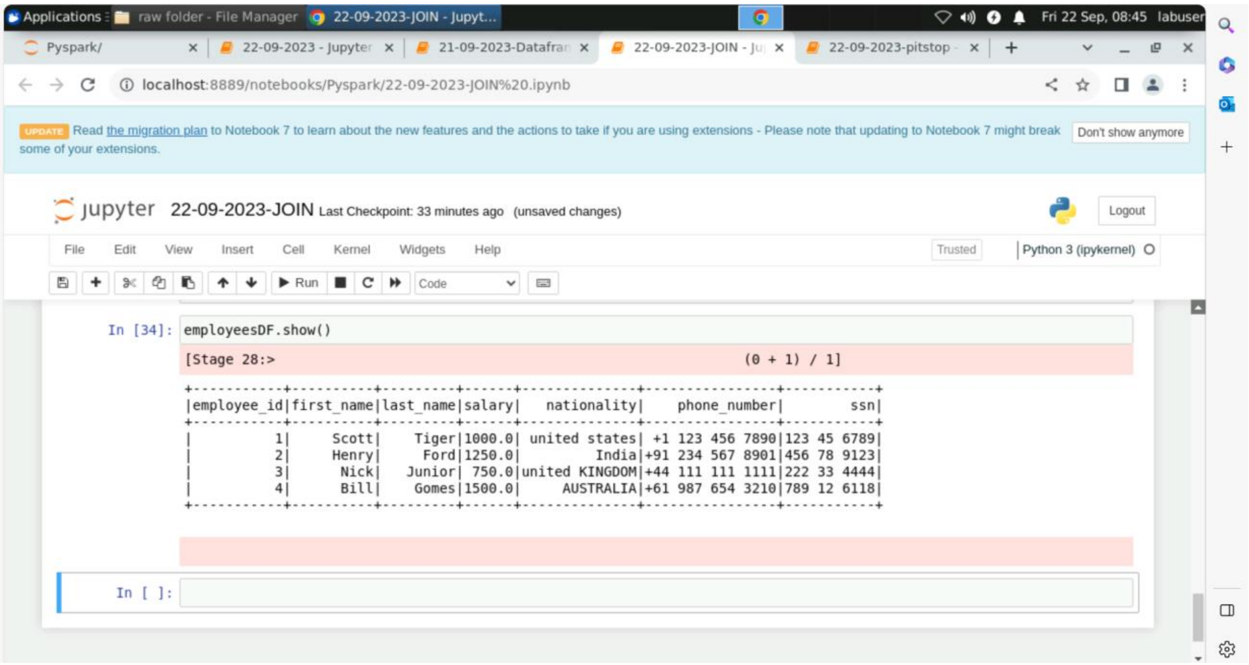
```
In [28]: df_sales_product.filter("transaction_id=1").show()

+--------------+----------+-----------+-------------+-------------------+----------+------------+-----------+-----
+
|transaction_id|product_id|customer_id|quantity_sold|          timestamp|product_id|product_name|   category|price
|
+--------------+----------+-----------+-------------+-------------------+----------+------------+-----------+-----
+
|             1|       101|        201|            5|2023-09-22 10:15:00|       101|      Laptop|Electronics|  800
|
+--------------+----------+-----------+-------------+-------------------+----------+------------+-----------+-----
+

In [30]: df_sales_product.filter("transaction_id>3").show()

+--------------+----------+-----------+-------------+-------------------+----------+------------+-----------+-----
+
|transaction_id|product_id|customer_id|quantity_sold|          timestamp|product_id|product_name|   category|price
|
+--------------+----------+-----------+-------------+-------------------+----------+------------+-----------+-----
```



Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions.

22-09-2023-JOIN Last Checkpoint: 32 minutes ago (unsaved changes)

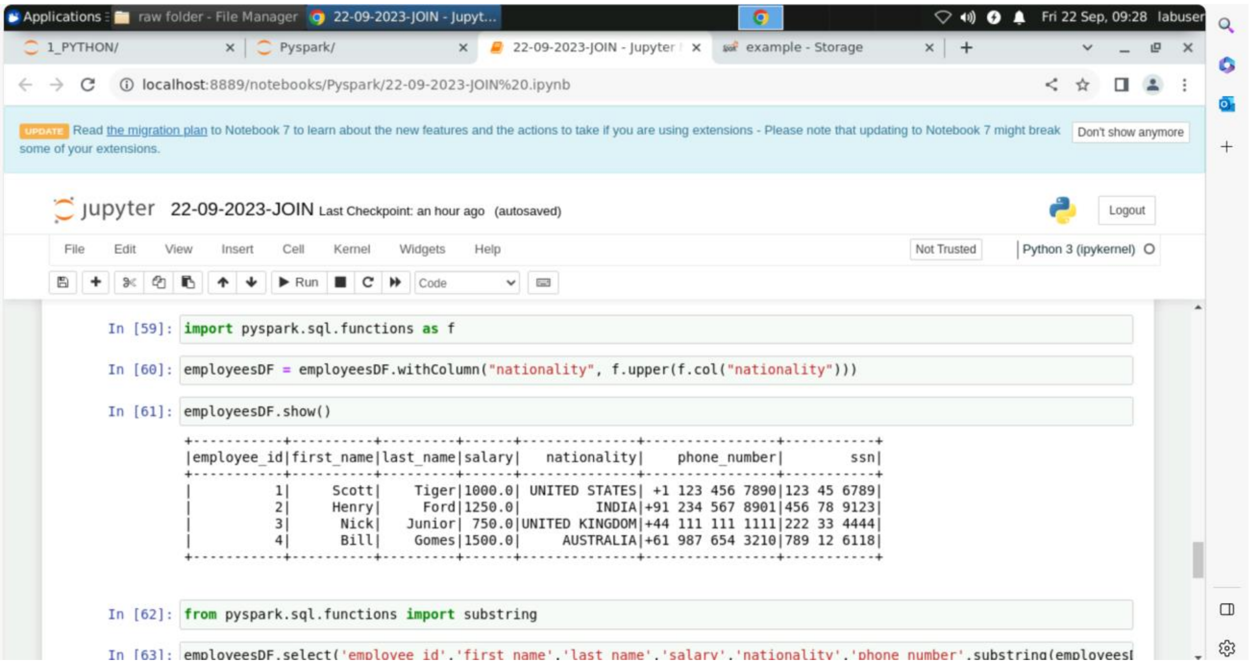
```
In [31]: employees = [(1, "Scott", "Tiger", 1000.0,
                       "united states", "+1 123 456 7890", "123 45 6789"
                      ),
                      (2, "Henry", "Ford", 1250.0,
                       "India", "+91 234 567 8901", "456 78 9123"
                      ),
                      (3, "Nick", "Junior", 750.0,
                       "united KINGDOM", "+44 111 111 1111", "222 33 4444"
                      ),
                      (4, "Bill", "Gomes", 1500.0,
                       "AUSTRALIA", "+61 987 654 3210", "789 12 6118"
                      )
                 ]

In [32]: employeesDF = spark. \
             createDataFrame(employees,
                             schema="""employee_id INT, first_name STRING,
                             last_name STRING, salary FLOAT, nationality STRING,
                             phone_number STRING, ssn STRING"""
                            )
```



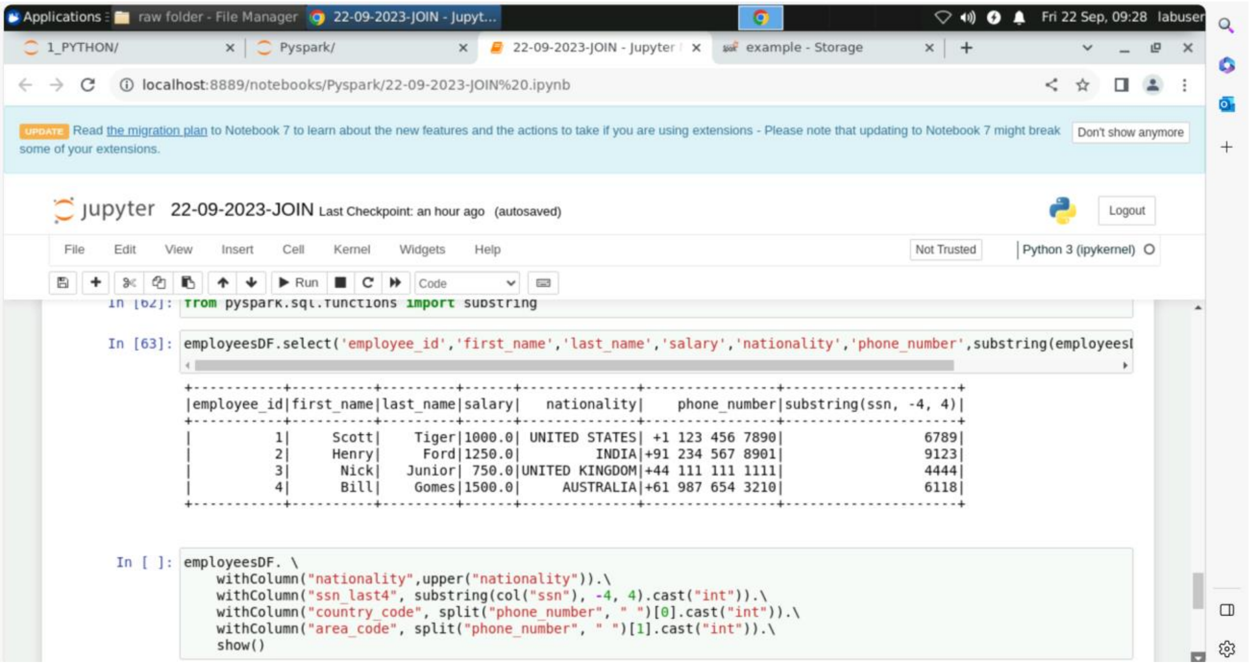
Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions.

22-09-2023-JOIN Last Checkpoint: 33 minutes ago (unsaved changes)

```
In [34]: employeesDF.show()

[Stage 28:>                                                         (0 + 1) / 1]

+-----------+----------+---------+------+--------------+----------------+-----------+
|employee_id|first_name|last_name|salary|   nationality|    phone_number|        ssn|
+-----------+----------+---------+------+--------------+----------------+-----------+
|          1|     Scott|    Tiger|1000.0| united states| +1 123 456 7890|123 45 6789|
|          2|     Henry|     Ford|1250.0|         India|+91 234 567 8901|456 78 9123|
|          3|      Nick|   Junior| 750.0|united KINGDOM|+44 111 111 1111|222 33 4444|
|          4|      Bill|    Gomes|1500.0|     AUSTRALIA|+61 987 654 3210|789 12 6118|
+-----------+----------+---------+------+--------------+----------------+-----------+

In [ ]:
```


Applications
raw folder - File Manager
22-09-2023-JOIN - Jupyt...
Fri 22 Sep, 09:28 labuser
1_PYTHON/
Pyspark/
22-09-2023-JOIN - Jupyter
example - Storage
localhost:8889/notebooks/Pyspark/22-09-2023-JOIN%20.ipynb
UPDATE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions.
Don't show anymore
jupyter 22-09-2023-JOIN Last Checkpoint: an hour ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code

```
In [59]: import pyspark.sql.functions as f

In [60]: employeesDF = employeesDF.withColumn("nationality", f.upper(f.col("nationality")))

In [61]: employeesDF.show()

+-----------+----------+---------+------+--------------+---------------+-----------+
|employee_id|first_name|last_name|salary|   nationality|   phone_number|        ssn|
+-----------+----------+---------+------+--------------+---------------+-----------+
|          1|     Scott|    Tiger|1000.0| UNITED STATES| +1 123 456 7890|123 45 6789|
|          2|     Henry|     Ford|1250.0|         INDIA|+91 234 567 8901|456 78 9123|
|          3|      Nick|   Junior| 750.0|UNITED KINGDOM|+44 111 111 1111|222 33 4444|
|          4|      Bill|    Gomes|1500.0|     AUSTRALIA|+61 987 654 3210|789 12 6118|
+-----------+----------+---------+------+--------------+---------------+-----------+

In [62]: from pyspark.sql.functions import substring

In [63]: employeesDF.select('employee_id','first_name','last_name','salary','nationality','phone_number',substring(employeesI
```


Applications
raw folder - File Manager
22-09-2023-JOIN - Jupyt...
Fri 22 Sep, 09:28 labuser
1_PYTHON/
Pyspark/
22-09-2023-JOIN - Jupyter
example - Storage
localhost:8889/notebooks/Pyspark/22-09-2023-JOIN%20.ipynb
UPDATE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions.
Don't show anymore
jupyter 22-09-2023-JOIN Last Checkpoint: an hour ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code

```
In [62]: from pyspark.sql.functions import substring

In [63]: employeesDF.select('employee_id','first_name','last_name','salary','nationality','phone_number',substring(employeesI

+-----------+----------+---------+------+--------------+---------------+---------------------+
|employee_id|first_name|last_name|salary|   nationality|   phone_number|substring(ssn, -4, 4)|
+-----------+----------+---------+------+--------------+---------------+---------------------+
|          1|     Scott|    Tiger|1000.0| UNITED STATES| +1 123 456 7890|                 6789|
|          2|     Henry|     Ford|1250.0|         INDIA|+91 234 567 8901|                 9123|
|          3|      Nick|   Junior| 750.0|UNITED KINGDOM|+44 111 111 1111|                 4444|
|          4|      Bill|    Gomes|1500.0|     AUSTRALIA|+61 987 654 3210|                 6118|
+-----------+----------+---------+------+--------------+---------------+---------------------+

In [ ]: employeesDF. \
    withColumn("nationality",upper("nationality")).\
    withColumn("ssn_last4", substring(col("ssn"), -4, 4).cast("int")).\
    withColumn("country_code", split("phone_number", " ")[0].cast("int")).\
    withColumn("area_code", split("phone_number", " ")[1].cast("int")).\
    show()
```

In databricks:

df.count()

spark.catelog.clearCache()