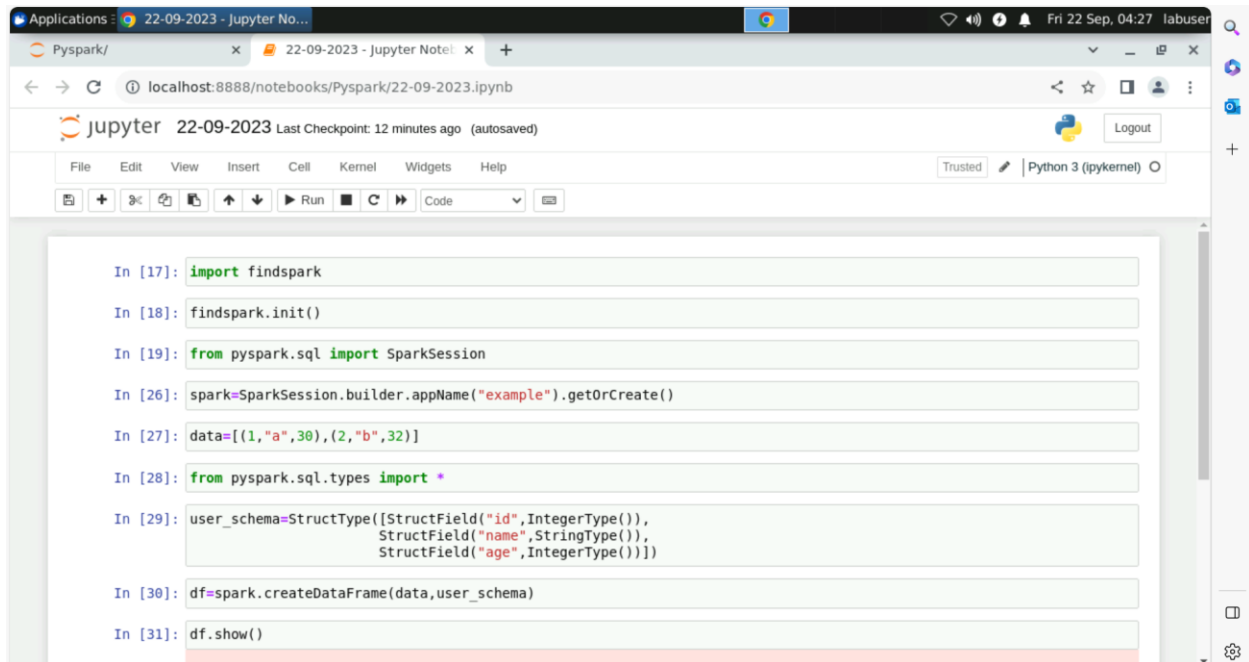


PYSPARK



```
In [17]: import findspark

In [18]: findspark.init()

In [19]: from pyspark.sql import SparkSession

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

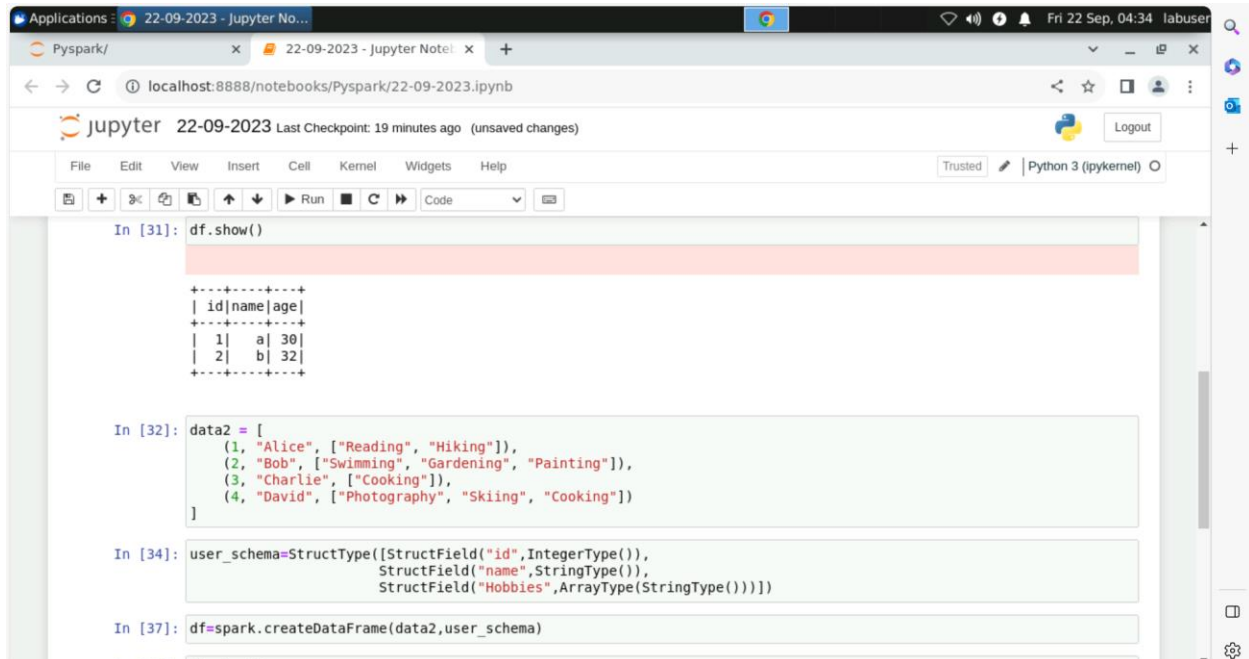
In [27]: data=[(1,"a",30),(2,"b",32)]

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()
```



```
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...
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)
File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)
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...
Pyspark/ 22-09-2023 - Jupyter Note...
localhost:8888/notebooks/Pyspark/22-09-2023.ipynb
jupyter 22-09-2023 Last Checkpoint: 32 minutes ago (autosaved)
File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)
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

Applications: 22-09-2023 - Jupyter No...
 Pyspark/ 22-09-2023 - Jupyter Note...
 localhost:8888/notebooks/Pyspark/22-09-2023.ipynb
 jupyter 22-09-2023 Last Checkpoint: 32 minutes ago (autosaved)
 File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

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, Gardening]	Swimming
2	Bob	[Swimming, Gardening]	Gardening
2	Bob	[Swimming, Gardening]	Painting
3	Charlie	[Cooking]	Cooking
4	David	[Photography, Skiing]	Photography
4	David	[Photography, Skiing]	Skiing
4	David	[Photography, Skiing]	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, Gardening]	Swimming
2	Bob	[Swimming, Gardening]	Gardening
2	Bob	[Swimming, Gardening]	Painting

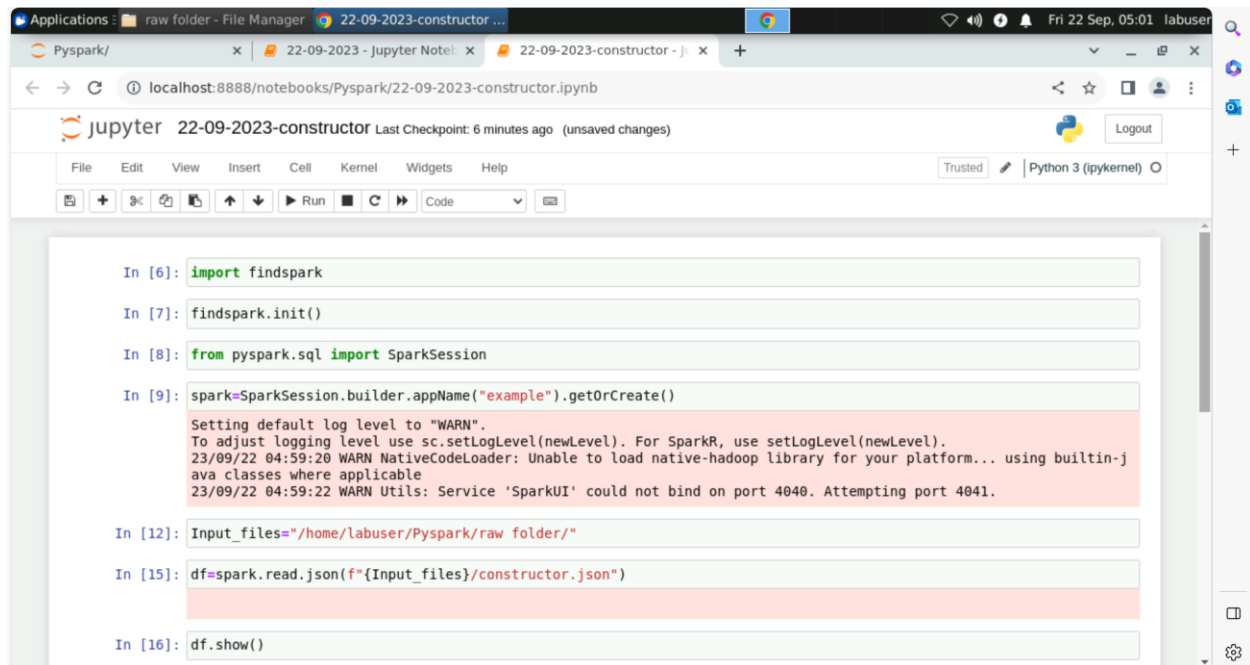
Applications: 22-09-2023 - Jupyter No...
 Pyspark/ 22-09-2023 - Jupyter Note...
 localhost:8888/notebooks/Pyspark/22-09-2023.ipynb
 jupyter 22-09-2023 Last Checkpoint: 36 minutes ago (unsaved changes)
 File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

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, Gardening]	Swimming	2023-09-22 04:51:...
2	Bob	[Swimming, Gardening]	Gardening	2023-09-22 04:51:...
2	Bob	[Swimming, Gardening]	Painting	2023-09-22 04:51:...
3	Charlie	[Cooking]	Cooking	2023-09-22 04:51:...
4	David	[Photography, Skiing]	Photography	2023-09-22 04:51:...
4	David	[Photography, Skiing]	Skiing	2023-09-22 04:51:...
4	David	[Photography, Skiing]	Cooking	2023-09-22 04:51:...

Constructor:



The screenshot shows a Jupyter Notebook interface with the following code cells:

```
In [6]: import findspark

In [7]: findspark.init()

In [8]: from pyspark.sql import SparkSession

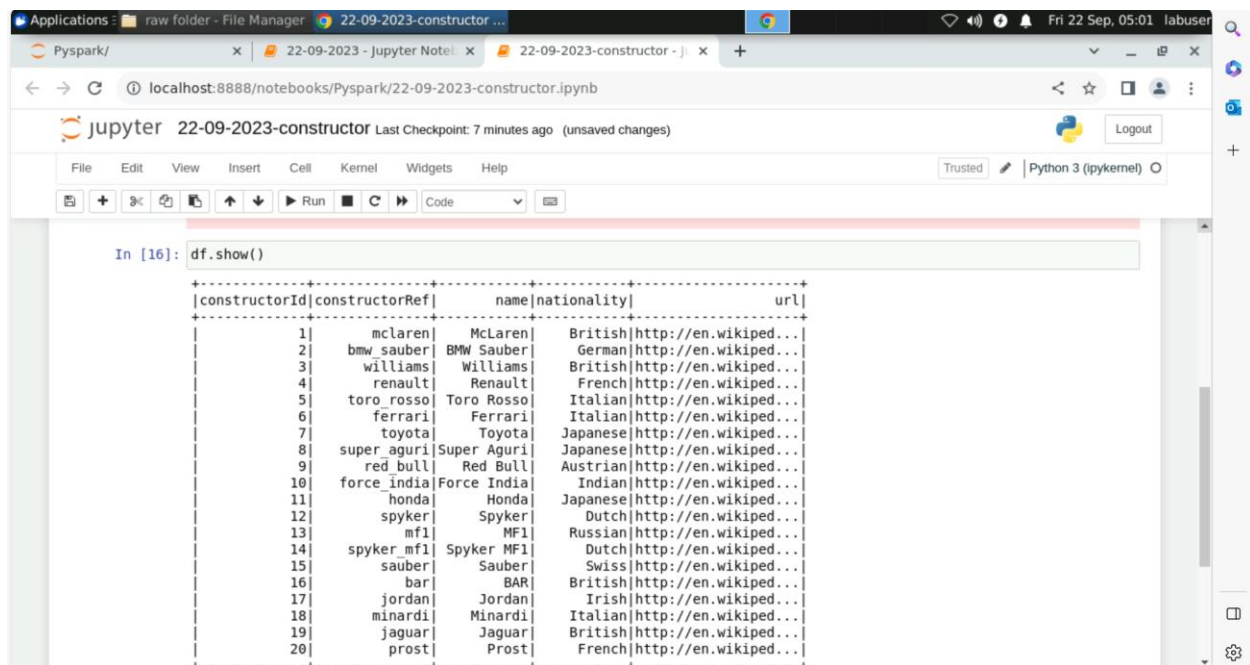
In [9]: spark=SparkSession.builder.appName("example").getOrCreate()

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-
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()
```



The screenshot shows the output of the `df.show()` command, displaying a table of constructor data:

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

from pyspark.sql.functions import *

from pyspark.sql.types import *

Applications - raw folder - File Manager 22-09-2023-constructor ...

Pyspark/ 22-09-2023 - Jupyter Noteb 22-09-2023-constructor - J- x

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)

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

Applications - raw folder - File Manager 22-09-2023-constructor ...

Pyspark/ 22-09-2023 - Jupyter Noteb 22-09-2023-constructor - J- x

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)

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 ...
Pyspark/ x 22-09-2023 - Jupyter Note x 22-09-2023-constructor - J x +
localhost:8888/notebooks/Pyspark/22-09-2023-constructor.ipynb
jupyter 22-09-2023-constructor Last Checkpoint: an hour ago (unsaved changes)
Python 3 (ipykernel)
File Edit View Insert Cell Kernel Widgets Help
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|mfl|MF1|Russian|2023-09-22 05:51:...|
|14|spyker_mfl|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 ...
Pyspark/ x 22-09-2023 - Jupyter Note x 22-09-2023-constructor - J x +
localhost:8888/notebooks/Pyspark/22-09-2023-constructor.ipynb
jupyter 22-09-2023-constructor Last Checkpoint: an hour ago (unsaved changes)
Python 3 (ipykernel)
File Edit View Insert Cell Kernel Widgets Help
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|mfl|MF1|Russian|2023-09-22 05:51:...|
14|spyker_mfl|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 - Jupy... 22-09-2023 - Jupyter Note... 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)

```
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 - Jupy... 22-09-2023 - Jupyter Note... 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)

```
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 - Jup... 22-09-2023 - Jupyter Note... 22-09-2023-creator - 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)

only showing top 20 rows

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

```
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 - Jup... 22-09-2023 - Jupyter Note... 22-09-2023-creator - 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)

only showing top 20 rows

15	24.899	10	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

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

In []:

Applications: 22-09-2023-pitstop - Jup... Fri 22 Sep, 07:02 labuser

Pyspark/ 22-09-2023-pitstop - Jupyt... 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)

```
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 - Jupyt... 21-09-2023-Datafram... 22-09-2023-JOIN - Ju... 22-09-2023-pitstop - x... localhost:8889/notebooks/Pyspark/22-09-2023-JOIN%20.ipynb

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)

```
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/transaction_id.csv")

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

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

transaction_id	product_id	customer_id	quantity	timestamp	product_name	category
1	1	1	1	2017-09-01 10:00:00	Product 1	Category 1
2	2	2	2	2017-09-01 11:00:00	Product 2	Category 2
3	3	3	3	2017-09-01 12:00:00	Product 3	Category 3
4	4	4	4	2017-09-01 13:00:00	Product 4	Category 4
5	5	5	5	2017-09-01 14:00:00	Product 5	Category 5
6	6	6	6	2017-09-01 15:00:00	Product 6	Category 6
7	7	7	7	2017-09-01 16:00:00	Product 7	Category 7
8	8	8	8	2017-09-01 17:00:00	Product 8	Category 8
9	9	9	9	2017-09-01 18:00:00	Product 9	Category 9
10	10	10	10	2017-09-01 19:00:00	Product 10	Category 10
11	11	11	11	2017-09-01 20:00:00	Product 11	Category 11
12	12	12	12	2017-09-01 21:00:00	Product 12	Category 12
13	13	13	13	2017-09-01 22:00:00	Product 13	Category 13
14	14	14	14	2017-09-01 23:00:00	Product 14	Category 14
15	15	15	15	2017-09-02 00:00:00	Product 15	Category 15
16	16	16	16	2017-09-02 01:00:00	Product 16	Category 16
17	17	17	17	2017-09-02 02:00:00	Product 17	Category 17
18	18	18	18	2017-09-02 03:00:00	Product 18	Category 18
19	19	19	19	2017-09-02 04:00:00	Product 19	Category 19
20	20	20	20	2017-09-02 05:00:00	Product 20	Category 20
21	21	21	21	2017-09-02 06:00:00	Product 21	Category 21
22	22	22	22	2017-09-02 07:00:00	Product 22	Category 22
23	23	23	23	2017-09-02 08:00:00	Product 23	Category 23
24	24	24	24	2017-09-02 09:00:00	Product 24	Category 24
25	25	25	25	2017-09-02 10:00:00	Product 25	Category 25
26	26	26	26	2017-09-02 11:00:00	Product 26	Category 26
27	27	27	27	2017-09-02 12:00:00	Product 27	Category 27
28	28	28	28	2017-09-02 13:00:00	Product 28	Category 28
29	29	29	29	2017-09-02 14:00:00	Product 29	Category 29
30	30	30	30	2017-09-02 15:00:00	Product 30	Category 30
31	31	31	31	2017-09-02 16:00:00	Product 31	Category 31
32	32	32	32	2017-09-02 17:00:00	Product 32	Category 32
33	33	33	33	2017-09-02 18:00:00	Product 33	Category 33
34	34	34	34	2017-09-02 19:00:00	Product 34	Category 34
35	35	35	35	2017-09-02 20:00:00	Product 35	Category 35
36	36	36	36	2017-09-02 21:00:00	Product 36	Category 36
37	37	37	37	2017-09-02 22:00:00	Product 37	Category 37
38	38	38	38	2017-09-02 23:00:00	Product 38	Category 38
39	39	39	39	2017-09-03 00:00:00	Product 39	Category 39
40	40	40	40	2017-09-03 01:00:00	Product 40	Category 40
41	41	41	41	2017-09-03 02:00:00	Product 41	Category 41
42	42	42	42	2017-09-03 03:00:00	Product 42	Category 42
43	43	43	43	2017-09-03 04:00:00	Product 43	Category 43
44	44	44	44	2017-09-03 05:00:00	Product 44	Category 44
45	45	45	45	2017-09-03 06:00:00	Product 45	Category 45
46	46	46	46	2017-09-03 07:00:00	Product 46	Category 46
47	47	47	47	2017-09-03 08:00:00	Product 47	Category 47
48	48	48	48	2017-09-03 09:00:00	Product 48	Category 48
49	49	49	49	2017-09-03 10:00:00	Product 49	Category 49
50	50	50	50	2017-09-03 11:00:00	Product 50	Category 50
51	51	51	51	2017-09-03 12:00:00	Product 51	Category 51
52	52	52	52	2017-09-03 13:00:00	Product 52	Category 52
53	53	53	53	2017-09-03 14:00:00	Product 53	Category 53
54	54	54	54	2017-09-03 15:00:00	Product 54	Category 54
55	55	55	55	2017-09-03 16:00:00	Product 55	Category 55
56	56	56	56	2017-09-03 17:00:00	Product 56	Category 56
57	57	57	57	2017-09-03 18:00:00	Product 57	Category 57
58	58	58	58	2017-09-03 19:00:00	Product 58	Category 58
59	59	59	59	2017-09-03 20:00:00	Product 59	Category 59
60	60	60	60	2017-09-03 21:00:00	Product 60	Category 60
61	61	61	61	2017-09-03 22:00:00	Product 61	Category 61
62	62	62	62	2017-09-03 23:00:00	Product 62	Category 62
63	63	63	63	2017-09-04 00:00:00	Product 63	Category 63
64	64	64	64	2017-09-04 01:00:00	Product 64	Category 64
65	65	65	65	2017-09-04 02:00:00	Product 65	Category 65
66	66	66	66	2017-09-04 03:00:00	Product 66	Category 66
67	67	67	67	2017-09-04 04:00:00	Product 67	Category 67
68	68	68	68	2017-09-04 05:00:00	Product 68	Category 68
69	69	69	69	2017-09-04 06:00:00	Product 69	Category 69
70	70	70	70	2017-09-04 07:00:00	Product 70	Category 70
71	71	71	71	2017-09-04 08:00:00	Product 71	Category 71
72	72	72	72	2017-09-04 09:00:00	Product 72	Category 72
73	73	73	73	2017-09-04 10:00:00	Product 73	Category 73
74	74	74	74	2017-09-04 11:00:00	Product 74	Category 74
75	75	75	75	2017-09-04 12:00:00	Product 75	Category 75
76	76	76	76	2017-09-04 13:00:00	Product 76	Category 76
77	77	77	77	2017-09-04 14:00:00	Product 77	Category 77
78	78	78	78	2017-09-04 15:00:00	Product 78	Category 78
79	79	79	79	2017-09-04 16:00:00	Product 79	Category 79
80	80	80	80	2017-09-04 17:00:00	Product 80	Category 80
81	81	81	81	2017-09-04 18:00:00	Product 81	Category 81
82	82	82	82	2017-09-04 19:00:00	Product 82	Category 82
83	83	83	83	2017-09-04 20:00:00	Product 83	Category 83
84	84	84	84	2017-09-04 21:00:00	Product 84	Category 84
85	85	85	85	2017-09-04 22:00:00	Product 85	Category 85
86	86	86	86	2017-09-04 23:00:00	Product 86	Category 86
87	87	87	87	2017-09-05 00:00:00	Product 87	Category 87
88	88	88	88	2017-09-05 01:00:00	Product 88	Category 88
89	89	89	89	2017-09-05 02:00:00	Product 89	Category 89
90	90	90	90	2017-09-05 03:00:00	Product 90	Category 90
91	91	91	91	2017-09-05 04:00:00	Product 91	Category 91
92	92	92	92	2017-09-05 05:00:00	Product 92	Category 92
93	93	93	93	2017-09-05 06:00:00	Product 93	Category 93
94	94	94	94	2017-09-05 07:00:00	Product 94	Category 94
95	95	95	95	2017-09-05 08:00:00	Product 95	Category 95
96	96	96	96	2017-09-05 09:00:00	Product 96	Category 96
97	97	97	97	2017-09-05 10:00:00	Product 97	Category 97
98	98	98	98	2017-09-05 11:00:00	Product 98	Category 98
99	99	99	99	2017-09-05 12:00:00	Product 99	Category 99
100	100	100	100	2017-09-05 13:00:00	Product 100	Category 100

Applications | raw folder - File Manager | 22-09-2023-JOIN - Jupyter... | Fri 22 Sep, 08:35 | labuser

Pyspark/ | 22-09-2023 - Jupyter | 21-09-2023-Datafram | 22-09-2023-JOIN - Ju | 22-09-2023-pitstop - x | + | - | x

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

```
--+
|00|      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 - Jupyter... | Fri 22 Sep, 08:35 | labuser

Pyspark/ | 22-09-2023 - Jupyter | 21-09-2023-Datafram | 22-09-2023-JOIN - Ju | 22-09-2023-pitstop - x | + | - | x

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

```
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
```

localhost:8889/notebooks/Pyspark/22-09-2023-JOIN .ipynb#

Applications - raw folder - File Manager 22-09-2023-JOIN - Jupyter... Fri 22 Sep, 08:36 labuser

Pyspark/ x 22-09-2023 - Jupyter x 21-09-2023-Datafram x 22-09-2023-JOIN - Ju x 22-09-2023-pitstop - x +

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)

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 []:

Applications - raw folder - File Manager 22-09-2023-JOIN - Jupyter... Fri 22 Sep, 08:40 labuser

Pyspark/ x 22-09-2023 - Jupyter x 21-09-2023-Datafram x 22-09-2023-JOIN - Ju x 22-09-2023-pitstop - x +

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: 28 minutes ago (unsaved changes) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

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
----------------	------------	-------------	---------------	-----------	------------	--------------	----------	-------

Applications : raw folder - File Manager 22-09-2023-JOIN - Jupyter... Fri 22 Sep, 08:44 labuser

Pyspark/ x 22-09-2023 - Jupyter x 21-09-2023-Datafram x 22-09-2023-JOIN - Ju x 22-09-2023-pitstop x +

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: 32 minutes ago (unsaved changes) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

```
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""")
```

Applications : raw folder - File Manager 22-09-2023-JOIN - Jupyter... Fri 22 Sep, 08:45 labuser

Pyspark/ x 22-09-2023 - Jupyter x 21-09-2023-Datafram x 22-09-2023-JOIN - Ju x 22-09-2023-pitstop x +

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: 33 minutes ago (unsaved changes) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

```
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 [ ]:
```

```
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(employeesDF
```

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

+-----+-----+-----+-----+-----+-----+-----+
|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.catalog.clearCache()