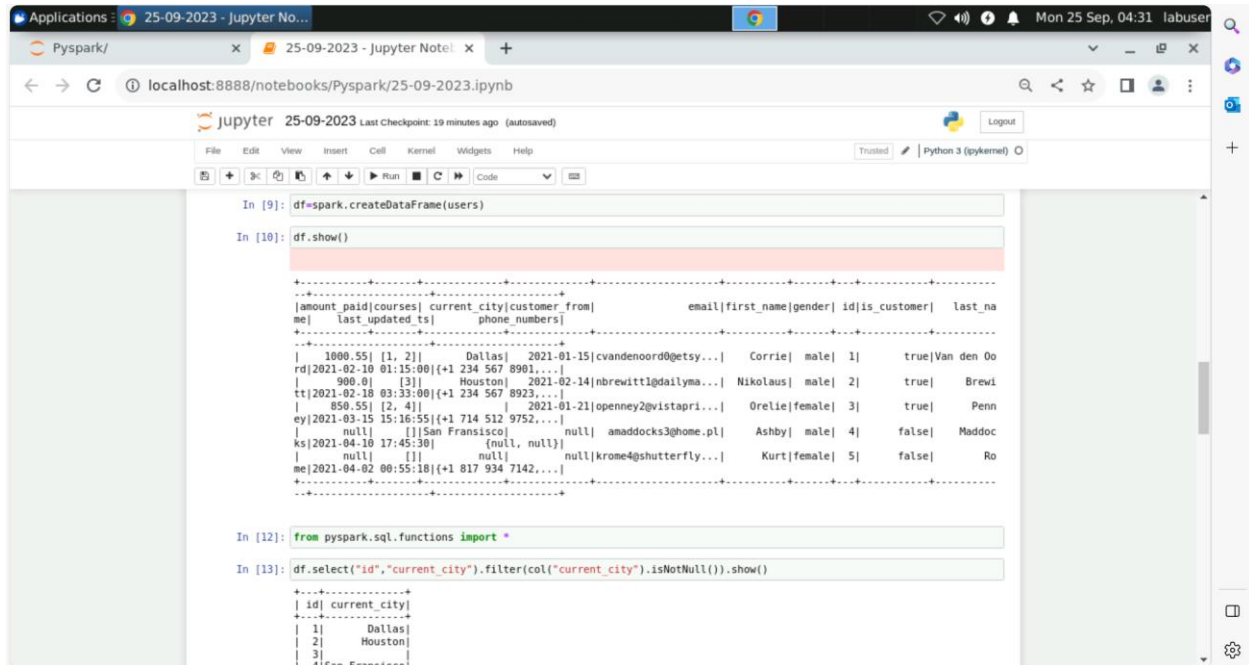


PYSPARK



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

```
In [9]: df=spark.createDataFrame(users)
```

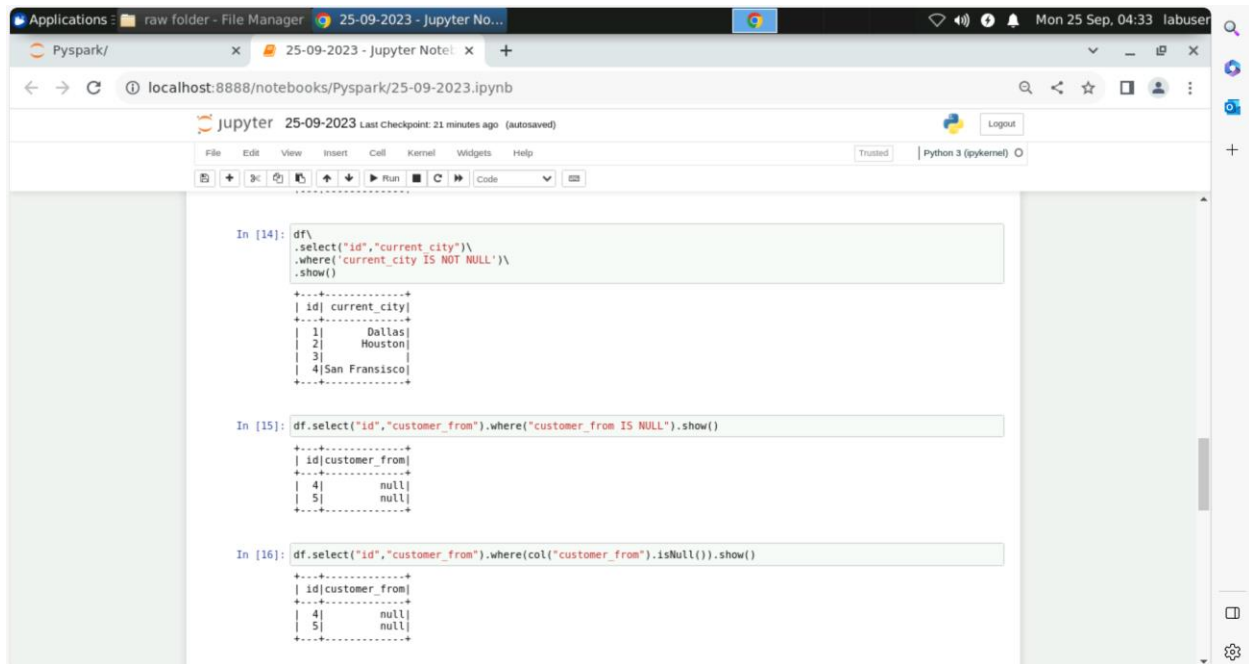
```
In [10]: df.show()
```

amount_paid	courses	current_city	customer_from	email	first_name	gender	id	is_customer	last_name
1000.55	[1, 2]	Dallas	2021-01-15	cvandenoord00getsy...	Corrie	male	1	true	Van den Oo
2021-02-10 01:15:00	[+1 234 567 8901,...]	Houston	2021-02-14	nbrewitt1@daily...	Nikolaus	male	2	true	Brewi
900.0	[3]	Houston	2021-02-14	nbrewitt1@daily...	Nikolaus	male	2	true	Brewi
850.55	[2, 4]	Houston	2021-01-21	openney2@vistapri...	Orelie	female	3	true	Penn
2021-03-15 15:16:55	[+1 714 512 9752,...]	San Francisco	amaddocks3@home.pl	Ashby	male	4	false	Maddoc	
2021-04-10 17:45:30	[null, null]	San Francisco	amaddocks3@home.pl	Ashby	male	4	false	Maddoc	
2021-04-02 00:55:18	[+1 817 934 7142,...]	San Francisco	amaddocks3@home.pl	Ashby	male	4	false	Maddoc	

```
In [12]: from pyspark.sql.functions import *
```

```
In [13]: df.select("id","current_city").filter(col("current_city").isNull()).show()
```

id	current_city
1	Dallas
2	Houston
3	San Francisco



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

```
In [14]: df\
        .select("id","current_city")\
        .where("current_city IS NOT NULL")\
        .show()
```

id	current_city
1	Dallas
2	Houston
3	San Francisco

```
In [15]: df.select("id","customer_from").where("customer_from IS NULL").show()
```

id	customer_from
4	null
5	null

```
In [16]: df.select("id","customer_from").where(col("customer_from").isNull()).show()
```

id	customer_from
4	null
5	null

Applications - raw folder - File Manager 25-09-2023 - Jupyter No... Mon 25 Sep, 04:33 labuser

Pyspark/ 25-09-2023 - Jupyter Note... +

localhost:8888/notebooks/Pyspark/25-09-2023.ipynb

jupyter 25-09-2023 Last Checkpoint: 21 minutes ago (autosaved) Logout

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

In [17]: `df.select("id","current_city","customer_from").orderBy("customer_from").show()`

```
++-----+
| id| current_city|customer_from|
++-----+
| 4|San Francisco| null|
| 5| null| null|
| 1| Dallas| 2021-01-15|
| 3| null| 2021-01-21|
| 2| Houston| 2021-02-14|
++-----+
```

In [18]: `df.select("id","current_city","customer_from").orderBy("current_city").show()`

```
++-----+
| id| current_city|customer_from|
++-----+
| 5| null| null|
| 3| null| 2021-01-21|
| 1| Dallas| 2021-01-15|
| 2| Houston| 2021-02-14|
| 4|San Francisco| null|
++-----+
```

In [19]: `df.select("id","current_city","customer_from").orderBy(df.customer_from.desc()).show()`

```
++-----+
| id| current_city|customer_from|
++-----+
| 2| Houston| 2021-02-14|
| 3| null| 2021-01-21|
| 1| Dallas| 2021-01-15|
| 4|San Francisco| null|
| 5| null| null|
++-----+
```

Applications - raw folder - File Manager 25-09-2023 - Jupyter No... Mon 25 Sep, 04:34 labuser

Pyspark/ 25-09-2023 - Jupyter Note... +

localhost:8888/notebooks/Pyspark/25-09-2023.ipynb

jupyter 25-09-2023 Last Checkpoint: 22 minutes ago (autosaved) Logout

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

In [20]: `df.select("id","current_city","customer_from").orderBy(desc_nulls_last("customer_from")).show()`

```
++-----+
| id| current_city|customer_from|
++-----+
| 2| Houston| 2021-02-14|
| 3| null| 2021-01-21|
| 1| Dallas| 2021-01-15|
| 4|San Francisco| null|
| 5| null| null|
++-----+
```

In [21]: `df.select("id","current_city","customer_from").orderBy(df["customer_from"].asc_nulls_last()).show()`

```
++-----+
| id| current_city|customer_from|
++-----+
| 1| Dallas| 2021-01-15|
| 3| Houston| 2021-01-21|
| 2| Houston| 2021-02-14|
| 4|San Francisco| null|
| 5| null| null|
++-----+
```

In []:

Applications - raw folder - File Manager 25-09-2023 - Jupyter No... Mon 25 Sep, 04:40 labuser

Pyspark/ 25-09-2023 - Jupyter Note... +

localhost:8888/notebooks/Pyspark/25-09-2023.ipynb

jupyter 25-09-2023 Last Checkpoint: 28 minutes ago (unsaved changes) Logout

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

In [23]: `zipdf=spark.read.option("header",True).option("inferSchema",True).csv("/home/labuser/Pyspark/raw folder/zipcode.csv")`

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

	id zipcode	type	city state population
1	704 STANDARD		null PR 30100
2	704	null PASEO COSTA DEL SUR	PR null
3	709	null	BDA SAN LUIS PR 3700
4	76166	UNIQUE	CINGULAR WIRELESS TX 84000
5	76177 STANDARD		null TX null
1	704 STANDARD		null PR 30100
1	704 STANDARD		null PR 30100

In [26]: `zipdf.dropDuplicates().show()`

	id zipcode	type	city state population
2	704	null PASEO COSTA DEL SUR	PR null
5	76177 STANDARD		null TX null
3	709	null	BDA SAN LUIS PR 3700
1	704 STANDARD		null PR 30100
4	76166	UNIQUE	CINGULAR WIRELESS TX 84000

In []:

Applications - raw folder - File Manager 25-09-2023 - Jupyter No... Mon 25 Sep, 04:45 labuser

Pyspark/ 25-09-2023 - Jupyter Note... +

localhost:8888/notebooks/Pyspark/25-09-2023.ipynb

jupyter 25-09-2023 Last Checkpoint: 33 minutes ago (unsaved changes) Logout

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

In [23]: `zipdf=spark.read.option("header",True).option("inferSchema",True).csv("/home/labuser/Pyspark/raw folder/zipcode.csv")`

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

	id zipcode	type	city state population
1	704 STANDARD		null PR 30100
2	704	null PASEO COSTA DEL SUR	PR null
3	709	null	BDA SAN LUIS PR 3700
4	76166	UNIQUE	CINGULAR WIRELESS TX 84000
5	76177 STANDARD		null TX null
1	704 STANDARD		null PR 30100
1	704 STANDARD		null PR 30100

In [28]: `zipdf1=zipdf.dropDuplicates()`

In [29]: `zipdf1.show()`

	id zipcode	type	city state population
2	704	null PASEO COSTA DEL SUR	PR null
5	76177 STANDARD		null TX null
3	709	null	BDA SAN LUIS PR 3700
1	704 STANDARD		null PR 30100
4	76166	UNIQUE	CINGULAR WIRELESS TX 84000

Applications : raw folder - File Manager 25-09-2023 - Jupyter No... Mon 25 Sep, 04:50 labuser

Pyspark/ 25-09-2023 - Jupyter Note... localhost:8888/notebooks/Pyspark/25-09-2023.ipynb

jupyter 25-09-2023 Last Checkpoint: 37 minutes ago (unsaved changes) Logout

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

```
In [29]: zipdf1.show()
```

	id	zipcode	type	city	state	population
2	704	null	PASEO COSTA DEL SUR	PR	null	
5	76177	STANDARD	null	TX	null	
3	709	null	BDA SAN LUIS	PR	3700	
1	704	STANDARD	null	PR	30100	
4	76166	UNIQUE	CINGULAR WIRELESS	TX	84000	

```
In [31]: zipdf1.dropna().show()
```

	id	zipcode	type	city	state	population
4	76166	UNIQUE	CINGULAR WIRELESS	TX	84000	

```
In [32]: zipdf1.dropna("all").show()
```

	id	zipcode	type	city	state	population
2	704	null	PASEO COSTA DEL SUR	PR	null	
5	76177	STANDARD	null	TX	null	
3	709	null	BDA SAN LUIS	PR	3700	
1	704	STANDARD	null	PR	30100	
4	76166	UNIQUE	CINGULAR WIRELESS	TX	84000	

```
In [ ]:
```

Applications : raw folder - File Manager 25-09-2023 - Jupyter No... Mon 25 Sep, 04:56 labuser

Pyspark/ 25-09-2023 - Jupyter Note... localhost:8888/notebooks/Pyspark/25-09-2023.ipynb

jupyter 25-09-2023 Last Checkpoint: 44 minutes ago (unsaved changes) Logout

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

```
In [32]: zipdf1.dropna("all").show()
```

	id	zipcode	type	city	state	population
2	704	null	PASEO COSTA DEL SUR	PR	null	
5	76177	STANDARD	null	TX	null	
3	709	null	BDA SAN LUIS	PR	3700	
1	704	STANDARD	null	PR	30100	
4	76166	UNIQUE	CINGULAR WIRELESS	TX	84000	

```
In [34]: zipdf1.dropna("all", subset="type").show()
```

	id	zipcode	type	city	state	population
5	76177	STANDARD	null	TX	null	
1	704	STANDARD	null	PR	30100	
4	76166	UNIQUE	CINGULAR WIRELESS	TX	84000	

```
In [35]: zipdf1.na.drop(subset="population").show()
```

	id	zipcode	type	city	state	population
3	709	null	BDA SAN LUIS	PR	3700	
1	704	STANDARD	null	PR	30100	
4	76166	UNIQUE	CINGULAR WIRELESS	TX	84000	

Applications - raw folder - File Manager 25-09-2023 - Jupyter No... Mon 25 Sep, 05:25 labuser

Pyspark/ 25-09-2023 - Jupyter Note... +

localhost:8888/notebooks/Pyspark/25-09-2023.ipynb

jupyter 25-09-2023 Last Checkpoint: an hour ago (autosaved) Logout

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

In [37]: zipdf1.show()

	id zipcode	type	city state population
2	704	null	PASEO COSTA DEL SUR PR null
5	76177	STANDARD	null TX null
3	709	null	BDA SAN LUIS PR 3700
1	704	STANDARD	null PR 30100
4	76166	UNIQUE	CINGULAR WIRELESS TX 84000

In [39]: zipdf1.na.fill(30000).show()

	id zipcode	type	city state population
2	704	null	PASEO COSTA DEL SUR PR 30000
5	76177	STANDARD	null TX 30000
3	709	null	BDA SAN LUIS PR 3700
1	704	STANDARD	null PR 30100
4	76166	UNIQUE	CINGULAR WIRELESS TX 84000

In [40]: zipdf1.fillna("Mumbai",subset="city").show()

	id zipcode	type	city state population
1	704	STANDARD	Mumbai PR 30100
2	704	null	PASEO COSTA DEL SUR PR null
3	709	null	BDA SAN LUIS PR 3700
4	76166	UNIQUE	CINGULAR WIRELESS TX 84000
5	76177	STANDARD	Mumbai TX null
1	704	STANDARD	Mumbai PR 30100

Applications - raw folder - File Manager 25-09-2023 - Jupyter No... Mon 25 Sep, 05:25 labuser

Pyspark/ 25-09-2023 - Jupyter Note... +

localhost:8888/notebooks/Pyspark/25-09-2023.ipynb

jupyter 25-09-2023 Last Checkpoint: an hour ago (autosaved) Logout

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

In [41]: zipdf1.fillna({"type":"VIP","City":"Mumbai","population":300000}).show()

	id zipcode	type	city state population
2	704	VIP	PASEO COSTA DEL SUR PR 300000
5	76177	STANDARD	Mumbai TX 300000
3	709	VIP	BDA SAN LUIS PR 3700
1	704	STANDARD	Mumbai PR 30100
4	76166	UNIQUE	CINGULAR WIRELESS TX 84000

In [42]: zipdf1.fillna("Mumbai",subset="city").show()

	id zipcode	type	city state population
2	704	null	PASEO COSTA DEL SUR PR null
5	76177	STANDARD	Mumbai TX null
3	709	null	BDA SAN LUIS PR 3700
1	704	STANDARD	Mumbai PR 30100
4	76166	UNIQUE	CINGULAR WIRELESS TX 84000

In []:

Applications - raw folder - File Manager 25-09-2023 - Jupyter No... Mon 25 Sep, 05:25 labuser

Pyspark/ 25-09-2023 - Jupyter Note... +

localhost:8888/notebooks/Pyspark/25-09-2023.ipynb

jupyter 25-09-2023 Last Checkpoint: an hour ago (autosaved) Logout

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

In [41]:

```
zipdf1.fillna({"type": "VIP", "City": "Mumbai", "population": 300000}).show()
```

	id	zipcode	type	city	state	population
2	704	VIP	PASEO COSTA DEL SUR	PR	300000	
5	76177	STANDARD	Mumbai	TX	300000	
3	709	VIP	BDA SAN LUIS	PR	3700	
1	704	STANDARD	Mumbai	PR	30100	
4	76166	UNIQUE	CINGULAR WIRELESS	TX	84000	

In [42]:

```
zipdf1.fillna("Mumbai", subset="city").show()
```

	id	zipcode	type	city	state	population
2	704	null	PASEO COSTA DEL SUR	PR	null	
5	76177	STANDARD	Mumbai	TX	null	
3	709	null	BDA SAN LUIS	PR	3700	
1	704	STANDARD	Mumbai	PR	30100	
4	76166	UNIQUE	CINGULAR WIRELESS	TX	84000	

In []:

Applications - raw folder - File Manager PartitionBy-25-09-2023 ... Mon 25 Sep, 05:49 labuser

Pyspark/ 25-09-2023 - Jupyter Note... PartitionBy-25-09-2023 - Ju... +

localhost:8888/notebooks/Pyspark/PartitionBy-25-09-2023.ipynb

jupyter PartitionBy-25-09-2023 Last Checkpoint: 5 minutes ago (unsaved changes) Logout

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

In [5]:

```
from pyspark.sql import Row
```

In [6]:

```
import findspark
findspark.init()
from pyspark.sql import SparkSession
spark = SparkSession.builder.appName("example").getOrCreate()
```

In [7]:

```
names_df = spark.read.option("header", True).option("inferSchema", True).csv("/home/labuser/Pyspark/raw_folder/babynames.csv")
```

In [8]:

```
names_df.show()
```

Year	First Name	County	Sex	Count
2007	ZOEY	KINGS	F	11
2007	ZOEY	SUFFOLK	F	6
2007	ZOEY	MONROE	F	6
2007	ZOEY	ERIE	F	9
2007	ZOE	ULSTER	F	5
2007	ZOE	WESTCHESTER	F	24
2007	ZOE	BRONX	F	13
2007	ZOE	NEW YORK	F	55
2007	ZOE	NASSAU	F	15
2007	ZOE	ERIE	F	6
2007	ZOE	SUFFOLK	F	14
2007	ZOE	KINGS	F	34
2007	ZOE	MONROE	F	9
2007	ZOE	QUEENS	F	26
2007	ZOE	ALBANY	F	5
2007	ZISSY	ROCKLAND	F	5
2007	ZISSY	KINGS	F	27
2007	ZION	KINGS	M	15
2007	ZION	BRONX	M	14

Applications: raw folder - File Manager PartitionBy-25-09-2023 ...

Pyspark/ 25-09-2023 - Jupyter Noteb PartitionBy-25-09-2023 - Jupyter Noteb

localhost:8888/notebooks/Pyspark/PartitionBy-25-09-2023.ipynb

jupyter PartitionBy-25-09-2023 Last Checkpoint: 8 minutes ago (autosaved) Logout

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

```
2007| ZISSY| KINGS| F| 27|
2007| ZION| KINGS| M| 15|
2007| ZION| BRONX| M| 14|
2007| ZEV| ROCKLAND| M| 6|
.....
only showing top 20 rows
```

```
In [10]: names_df.groupby("Year").count().show()

[Stage 6:>] (0 + 1) / 1]
+-----+
|Year|count|
+-----+
|2007| 6367|
|2013| 6158|
|2014| 8362|
|2012| 6164|
|2009| 6312|
|2010| 6192|
|2011| 6216|
|2008| 6481|
+-----+
```

In []:

Applications: babyparquet - File Mana... PartitionBy-25-09-2023 ...

Pyspark/ 25-09-2023 - Jupyter Noteb PartitionBy-25-09-2023 - Jupyter Noteb

localhost:8888/notebooks/Pyspark/PartitionBy-25-09-2023.ipynb

jupyter PartitionBy-25-09-2023 Last Checkpoint: 20 minutes ago (autosaved) Logout

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

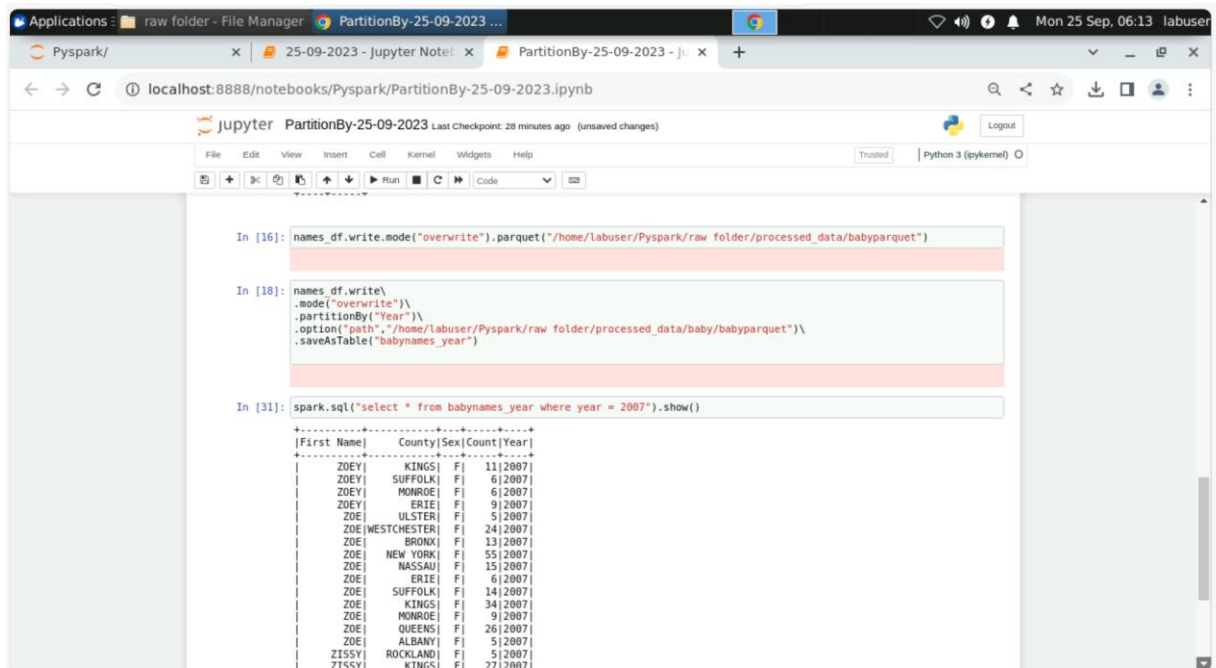
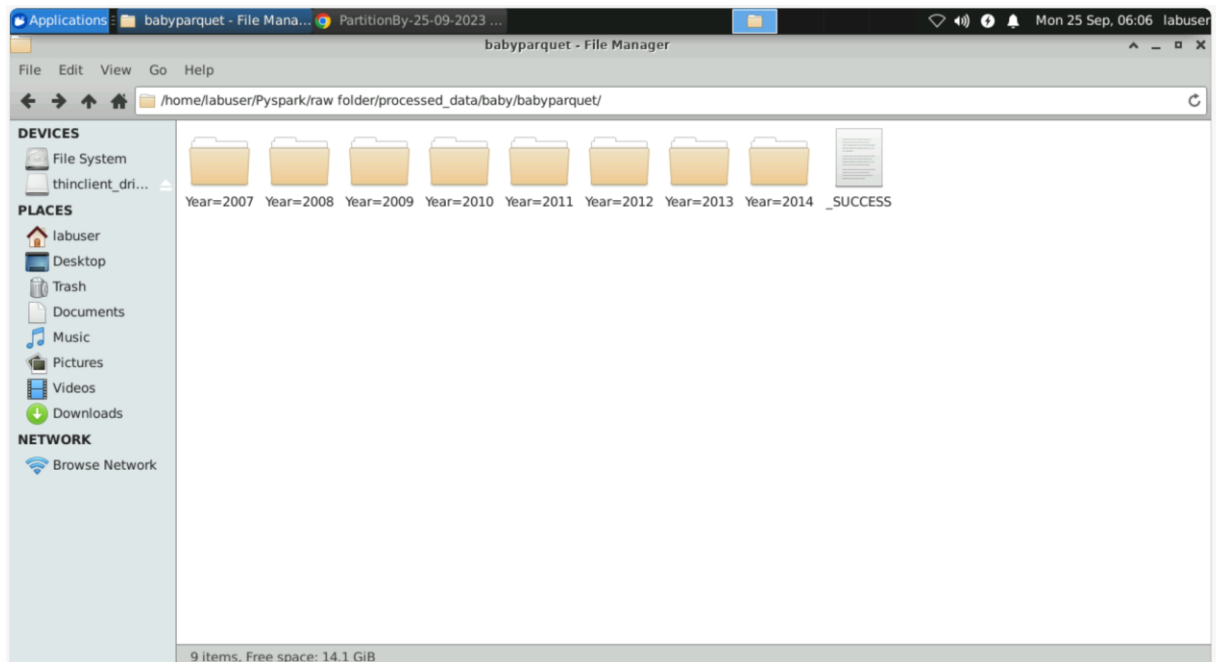
```
In [15]: names_df.groupby("Year").count().sort(col("Year").desc()).show()

+-----+
|Year|count|
+-----+
|2014| 8362|
|2013| 6158|
|2012| 6164|
|2011| 6216|
|2010| 6192|
|2009| 6312|
|2008| 6481|
|2007| 6367|
+-----+
```

```
In [16]: names_df.write.mode("overwrite").parquet("/home/labuser/Pyspark/raw_folder/processed_data/babyparquet")
```

```
In [18]: names_df.write\
    .mode("overwrite")\
    .partitionBy("Year")\
    .option("path", "/home/labuser/Pyspark/raw_folder/processed_data/baby/babyparquet")\
    .saveAsTable("babynames_year")
```

In []:



Applications: partitionbyyear&gender... PartitionBy-25-09-2023 ...

Mon 25 Sep, 06:16 labuser

Pyspark/ 25-09-2023 - Jupyter Noteb... PartitionBy-25-09-2023 - ju x

localhost:8888/notebooks/Pyspark/PartitionBy-25-09-2023.ipynb

jupyter PartitionBy-25-09-2023 Last Checkpoint: 32 minutes ago (autosaved) Logout

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

Run Code

ZOEY	KINGS	F	11	2007
ZOEY	SUFFOLK	F	6	2007
ZOEY	MONROE	F	6	2007
ZOEY	ERIE	F	9	2007
ZOE	ULSTER	F	5	2007
ZOE	WESTCHESTER	F	24	2007
ZOE	BRONX	F	13	2007
ZOE	NEW YORK	F	55	2007
ZOE	NASSAU	F	15	2007
ZOE	ERIE	F	6	2007
ZOE	SUFFOLK	F	14	2007
ZOE	KINGS	F	34	2007
ZOE	MONROE	F	9	2007
ZOE	QUEENS	F	26	2007
ZOE	ALBANY	F	5	2007
ZISSY	ROCKLAND	F	5	2007
ZISSY	KINGS	F	27	2007
ZION	KINGS	M	15	2007
ZION	BRONX	M	14	2007
ZEV	ROCKLAND	M	6	2007

only showing top 20 rows

```
In [32]: names_df.write\
        .mode("overwrite")\
        .partitionBy("Year","Sex")\
        .option("path","/home/labuser/Pyspark/raw folder/processed_data/baby/partitionbyyear&gender")\
        .saveAsTable("babynames_yearr_gender")
```

In []:

Applications: partitionbyyear&gender... PartitionBy-25-09-2023 ...

Mon 25 Sep, 06:17 labuser

partitionbyyear&gender - File Manager

File Edit View Go Help

/home/labuser/Pyspark/raw folder/processed_data/baby/partitionbyyear&gender/

DEVICES

- File System
- thinclient_dri...

PLACES

- labuser
- Desktop
- Trash
- Documents
- Music
- Pictures
- Videos
- Downloads

NETWORK

- Browse Network

Year=2007 Year=2008 Year=2009 Year=2010 Year=2011 Year=2012 Year=2013 Year=2014 _SUCCESS

9 items, Free space: 14.1 GiB

Applications: partitionbyyear&gender... PartitionBy-25-09-2023 ... Mon 25 Sep, 06:24 labuser

Pyspark/ 25-09-2023 - Jupyter Note... PartitionBy-25-09-2023 - Ju x +

localhost:8888/notebooks/Pyspark/PartitionBy-25-09-2023.ipynb

jupyter PartitionBy-25-09-2023 Last Checkpoint: 39 minutes ago (unsaved changes) Logout

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

Run Code

ZISSY	ROCKLAND	F	5	2007
ZISSY	KINGS	F	27	2007
ZION	KINGS	M	15	2007
ZION	BRONX	M	14	2007
ZEV	ROCKLAND	M	6	2007

only showing top 20 rows

```
In [32]: names_df.write\
        .mode("overwrite")\
        .partitionBy("Year", "Sex")\
        .option("path", "/home/labuser/Pyspark/raw folder/processed_data/baby/partitionbyyear&gender")\
        .saveAsTable("babynames_yearr_gender")
```

```
In [34]: names_df.write\
        .mode("overwrite")\
        .option("maxRecordsPerFile", 4000)\
        .partitionBy("Year")\
        .option("path", "/home/labuser/Pyspark/raw folder/processed_data/baby/partitionbyyear_max4k")\
        .saveAsTable("babynames_year_max4k")
```

In []:

Applications: baby - File Manager PartitionBy-25-09-2023 ... Mon 25 Sep, 06:25 labuser

baby - File Manager

File Edit View Go Help

/home/labuser/Pyspark/raw folder/processed_data/baby/

DEVICES

- File System
- thinclient_dri...

PLACES

- labuser
- Desktop
- Trash
- Documents
- Music
- Pictures
- Videos
- Downloads

NETWORK

- Browse Network

babyparquet partitionbyyear&ge partitionbyyear_max4k

partitionbyyear_max4k - Properties

General Emblems Permissions

Name: partitionbyyear_max4k

Kind: folder

Location: ...e/labuser/Pyspark/raw folder/processed_data/baby

Modified: Today

Accessed: Today

Size: 45 items, totalling 240.8 KiB (246571 bytes)

Usage: 14.1 GiB of 29.0 GiB free (51% used)

"partitionbyyear_max4k": folder