In [10]:
```
import findspark
findspark.init()
from pyspark.sql import SparkSession
spark=SparkSession.builder.appName("example").getOrCreate()
```

In [11]:
```
import datetime
users = [
    {
        "id": 1,
        "first_name": "Corrie",
        "last_name": "Van den Oord",
        "email": "cvandenoord0@etsy.com",
        "gender": "male",
        "current_city": "Dallas",
        "phone_numbers": Row(mobile="+1 234 567 8901", home="+1 234 567 8911"),
        "courses": [1, 2],
        "is_customer": True,
        "amount_paid": 1000.55,
        "customer_from": datetime.date(2021, 1, 15),
        "last_updated_ts": datetime.datetime(2021, 2, 10, 1, 15, 0)
    },
    {
        "id": 2,
        "first_name": "Nikolaus",
        "last_name": "Brewitt",
        "email": "nbrewitt1@dailymail.co.uk",
        "gender": "male"
```

In [12]:
```
from pyspark.sql import Row
```

In [13]:
```
df=spark.createDataFrame(users)
```

In [15]:
```
df.show()
```

```
+-----------+-------+-------------+-------------+--------------------+----------+------+---+-----------+----------
--+-------------------+--------------------+
|amount_paid|courses| current_city|customer_from|               email|first_name|gender| id|is_customer|   last_na
me|    last_updated_ts|       phone_numbers|
+-----------+-------+-------------+-------------+--------------------+----------+------+---+-----------+----------
--+-------------------+--------------------+
|    1000.55| [1, 2]|       Dallas|   2021-01-15|cvandenoord0@etsy...|    Corrie|  male|  1|       true|Van den Oo
rd|2021-02-10 01:15:00|{+1 234 567 8901,...|
|      900.0|    [3]|      Houston|   2021-02-14|nbrewitt1@dailyma...|  Nikolaus|  male|  2|       true|     Brewi
tt|2021-02-18 03:33:00|{+1 234 567 8923,...|
|     850.55| [2, 4]|             |   2021-01-21|openney2@vistapri...|    Orelie|female|  3|       true|      Penn
ey|2021-03-15 15:16:55|{+1 714 512 9752,...|
|       null|     []|San Fransisco|         null|  amaddocks3@home.pl|     Ashby|  male|  4|      false|    Maddoc
ks|2021-04-10 17:45:30|        {null, null}|
|       null|     []|         null|         null|krome4@shutterfly...|      Kurt|female|  5|      false|        Ro
me|2021-04-02 00:55:18|{+1 817 934 7142,...|
+-----------+-------+-------------+-------------+--------------------+----------+------+---+-----------+----------
--+-------------------+--------------------+
```

In [16]:
```
df\
.select("id","current_city")\
.where('current_city IS NOT NULL')\
.show()
```

```
+---+-------------+
| id| current_city|
+---+-------------+
|  1|       Dallas|
|  2|      Houston|
|  3|             |
|  4|San Fransisco|
+---+-------------+
```

In [18]:
```
from pyspark.sql.types import *
from pyspark.sql.functions import *
```

In [19]:
```
df.select("id","customer_from").where(col("customer_from").isNull()).show()
```

```
+---+-------------+
| id|customer_from|
+---+-------------+
|  4|         null|
|  5|         null|
+---+-------------+
```

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

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

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

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

In [22]:
```
df.select('id','current_city','customer_from').orderBy(df.customer_from.desc()).show()
```

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

In [25]:
```
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|             |   2021-01-21|
|  1|       Dallas|   2021-01-15|
|  4|San Fransisco|         null|
|  5|         null|         null|
+---+-------------+-------------+
```

In [27]:
```
df1 = spark.read.option("header",True).csv("zipcode.csv")
```

In [28]:
```
df1.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 [30]:
```
df1.drop_duplicates().show()
```

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

In [33]:
```
df1.drop_duplicates(["zipcode"]).show()
```

```
+---+-------+--------+-----------------+-----+----------+
| id|zipcode|    type|             city|state|population|
+---+-------+--------+-----------------+-----+----------+
|  1|    704|STANDARD|             null|   PR|     30100|
|  3|    709|    null|     BDA SAN LUIS|   PR|      3700|
|  4|  76166|  UNIQUE|CINGULAR WIRELESS|   TX|     84000|
|  5|  76177|STANDARD|             null|   TX|      null|
+---+-------+--------+-----------------+-----+----------+
```

In [34]:
```
df1.dropna().show()
```

```
+---+-------+------+-----------------+-----+----------+
| id|zipcode|  type|             city|state|population|
+---+-------+------+-----------------+-----+----------+
|  4|  76166|UNIQUE|CINGULAR WIRELESS|   TX|     84000|
+---+-------+------+-----------------+-----+----------+
```

In [35]:
```
df1.dropna("all").show()
```

```
+---+-------+--------+-------------------+-----+----------+
| id|zipcode|    type|               city|state|population|
```

In [39]:
```
df1.na.drop(subset="type").show()
```

```
+---+-------+--------+-----------------+-----+----------+
| id|zipcode|    type|             city|state|population|
+---+-------+--------+-----------------+-----+----------+
|  1|    704|STANDARD|             null|   PR|     30100|
|  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 [40]:
```
df1.na.drop(subset="population").show()
```

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

In [41]:
```
df1.na.drop(thresh=5).show()
```

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

In [42]:
```
df1.na.fill(30000).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 [46]:
```
df1.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|
|  1|    704|STANDARD|            Mumbai|   PR|     30100|
+---+-------+--------+------------------+-----+----------+
```

In [50]:
```
df1.fillna({"type":"VIP","city":"Mumbai","population":30000}).show()
```

```
+---+-------+--------+------------------+-----+----------+
| id|zipcode|    type|              city|state|population|
+---+-------+--------+------------------+-----+----------+
|  1|    704|STANDARD|            Mumbai|   PR|     30100|
|  2|    704|     VIP|PASEO COSTA DEL SUR|   PR|     30000|
|  3|    709|     VIP|       BDA SAN LUIS|   PR|      3700|
```

In [51]:
```
df1.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 [53]:
```
df1=df1.drop_duplicates()
```

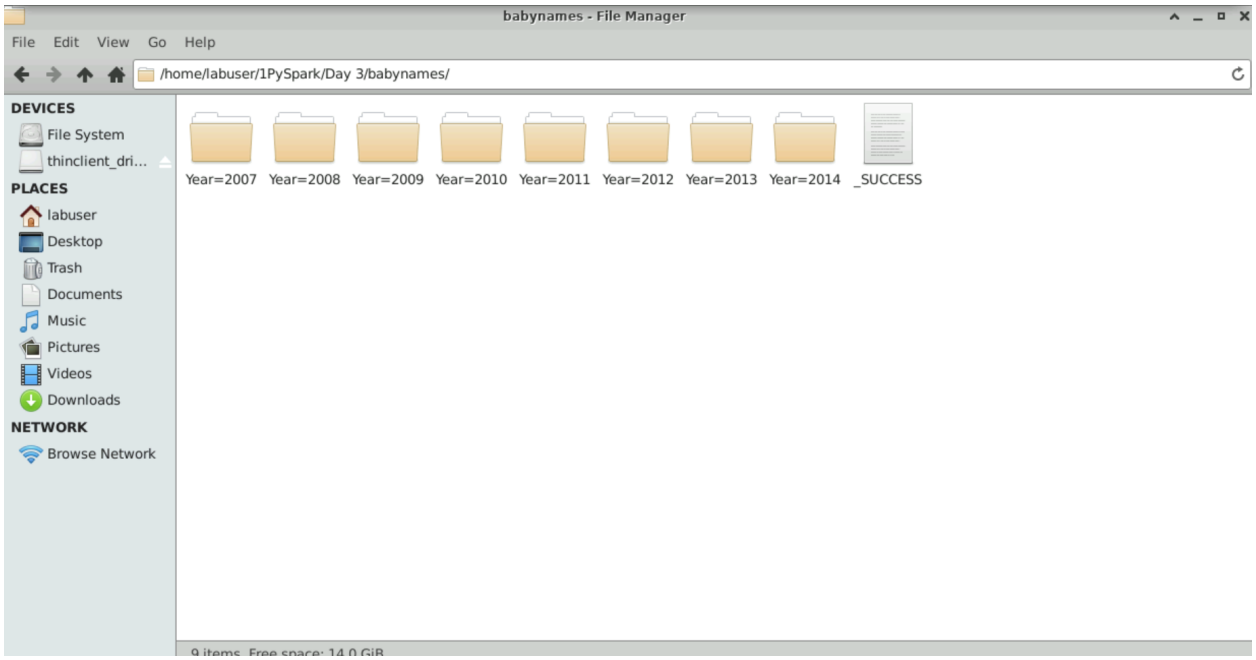
In [55]:
```
df.select("id","current_city","customer_from").replace("","LA").show()
```

```
+---+-------------+-------------+
| id| current_city|customer_from|
+---+-------------+-------------+
|  1|       Dallas|   2021-01-15|
```

In [9]:
```
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 [11]:
```
df.write\
.mode("overwrite")\
.partitionBy("Year")\
.option("path","/home/labuser/1PySpark/Day 3/babynames")\
.saveAsTable("babynames_year")
```


babynames - File Manager
File Edit View Go Help
/home/labuser/1PySpark/Day 3/babynames/
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: 14.0 GiB