Name: Rohan Vinayak Chaudhari

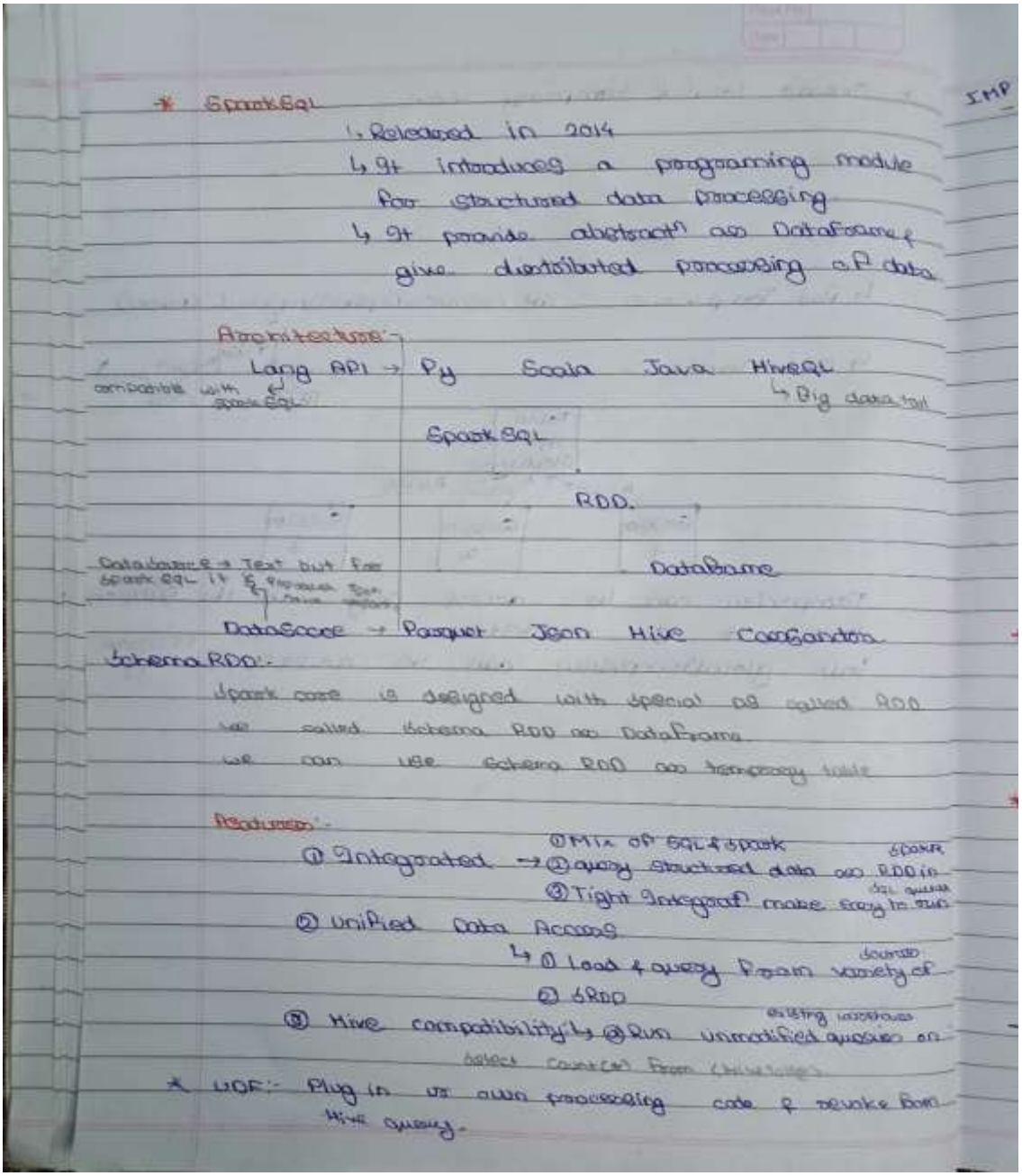
Batch: Data Engineering

Date:10/02/2024

Topic: Pyspark(Sparksql)

Solution:

1.SparkSQL:



IMP

- Spark RDD:
  - ↳ Fundamental data structure of Spark
  - ↳ Immutable distributed collection of object that can be stored in memory
  - ↳ Each dataset in RDD is divided into logical partitions, which maybe comp-uted on diff. nodes of the cluster.
  - ↳ Automatic retrit on failure
  - ↳ It can be operated in parallel
  - ↳ RDD is a read only partitioned collection of data.
  - ↳ 2 way to create RDD:
    - Parallelizing an existing collection
    - refereeing a database/dataset in external storage.

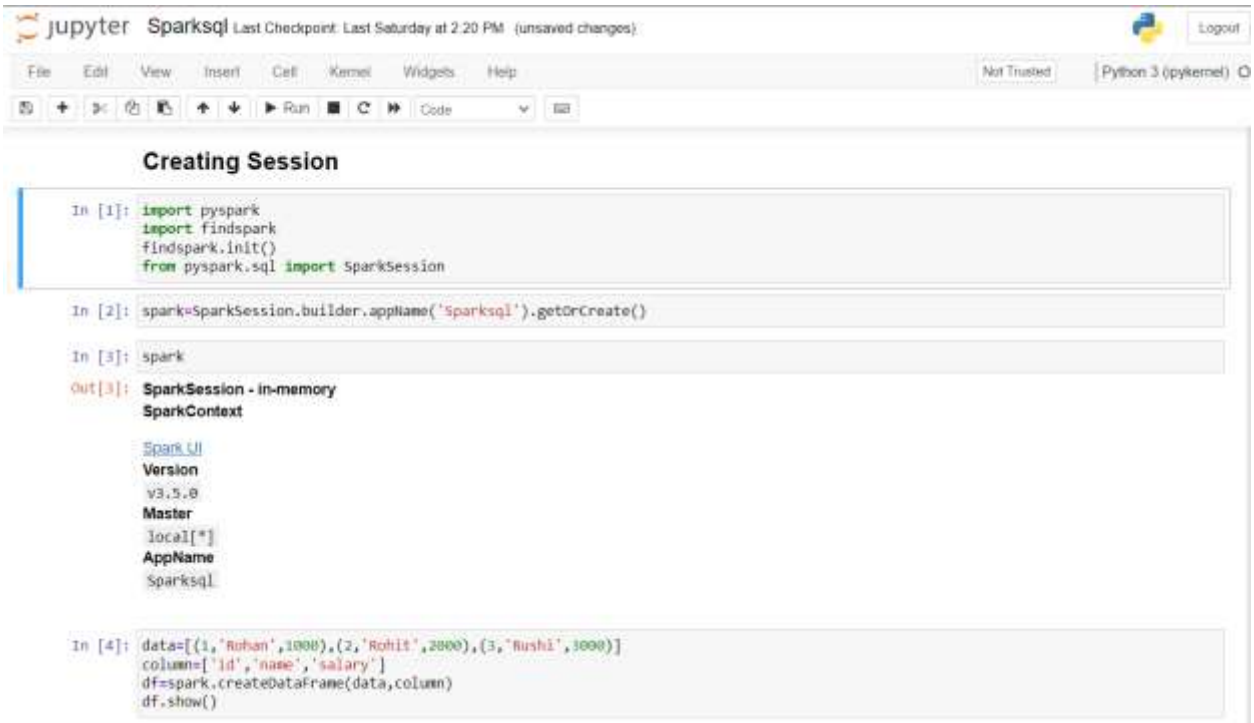
* Dataset
  - ↳ Distributed collection of data, organised into named columns equivalent to relational columns.
  - ↳ Static typing & runtime safety

* Dataframe:
  - ↳ Data is organised into named column, like a table
  - ↳ Features: supports Data formats, Ability to process data in huge size available on a single node to large clus

- Plan optimization & execution

DataFrame → Analysis → Logical → Physical → Code → R
optimization planning Gene

DataFrame & SQL share same optimization pipeline


jupyter Sparksql Last Checkpoint: Last Saturday at 2:20 PM (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (pykernel)
Run
Code
Creating Session
In [1]:
import pyspark
import findspark
findspark.init()
from pyspark.sql import SparkSession
In [2]:
spark=SparkSession.builder.appName('Sparksql').getOrCreate()
In [3]:
spark
Out[3]:
SparkSession - in-memory
SparkContext
Spark UI
Version
v3.5.0
Master
local[*]
AppName
Sparksql
In [4]:
data=[(1,'Rohan',1000),(2,'Rohit',2000),(3,'Rushi',3000)]
column=['id','name','salary']
df=spark.createDataFrame(data,column)
df.show()

jupyter Sparksql Last Checkpoint: Last Saturday at 2:20 PM (unsaved changes)

Logout

File Edit View Insert Cell Kernel Widgets Help

Not Trusted | Python 3 (ipykernel)

Run Code

In [5]:
```
help(df.createOrReplaceTempView)
```

```
Help on method createOrReplaceTempView in module pyspark.sql.dataframe:

createOrReplaceTempView(name: str) -> None method of pyspark.sql.dataframe.DataFrame instance
    Creates or replaces a local temporary view with this :class:`DataFrame`.

    The lifetime of this temporary table is tied to the :class:`SparkSession`
    that was used to create this :class:`DataFrame`.

    .. versionadded:: 2.0.0

    .. versionchanged:: 3.4.0
        Supports Spark Connect.

    Parameters
    ----------
    name : str
        Name of the view.

    Examples
    --------
    Create a local temporary view named 'people'.

    >>> df = spark.createDataFrame([(2, "Alice"), (5, "Bob")], schema=["age", "name"])
    >>> df.createOrReplaceTempView("people")

    Replace the local temporary view.

    >>> df2 = df.filter(df.age > 3)
    >>> df2.createOrReplaceTempView("people")
    >>> df3 = spark.sql("SELECT * FROM people")
    >>> sorted(df3.collect()) == sorted(df2.collect())
    True
```

jupyter Sparksql Last Checkpoint: Last Saturday at 2:20 PM (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Not Trusted | Python 3 (ipykernel)

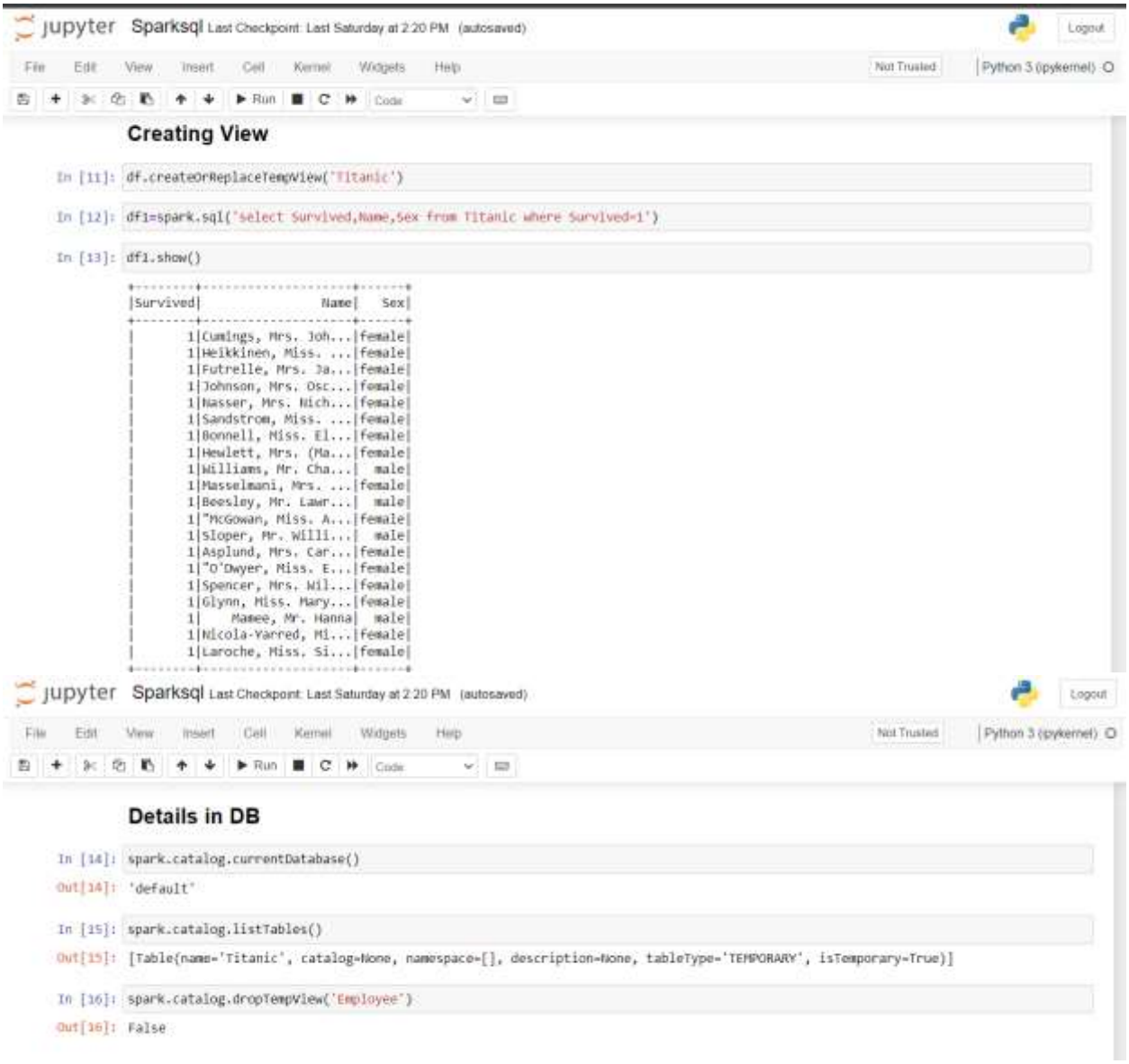
Run Code

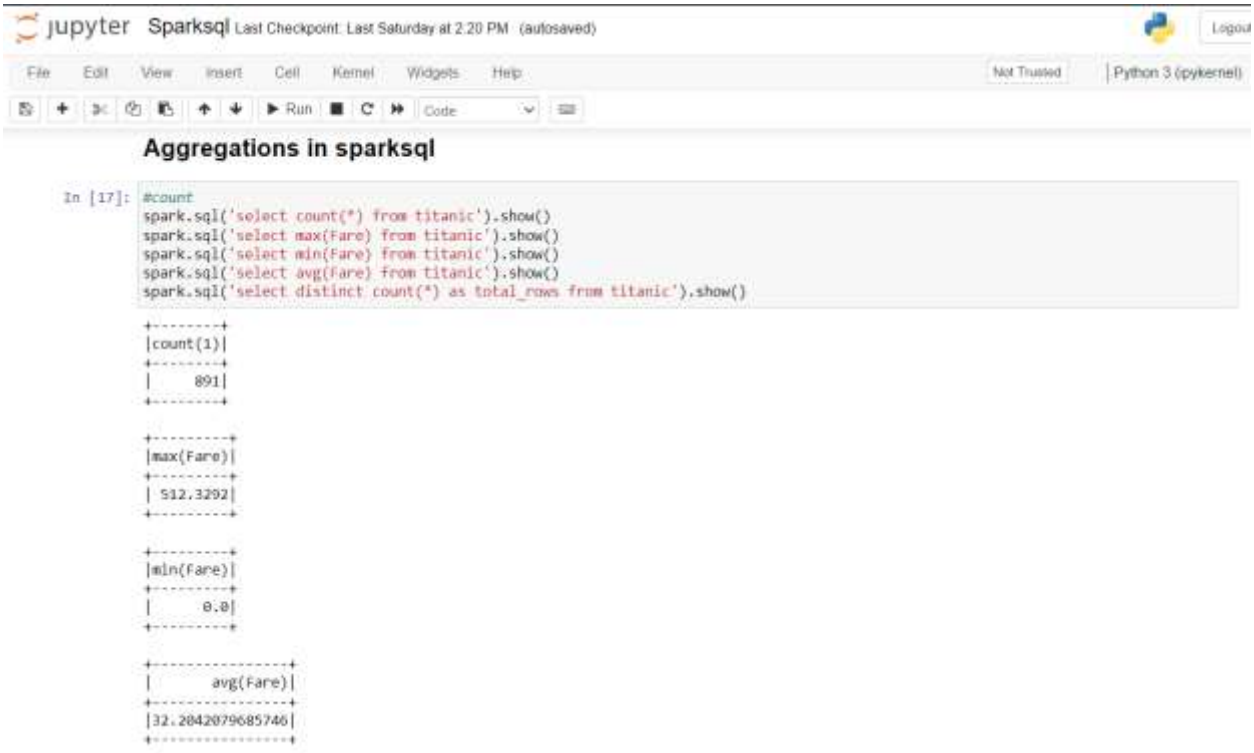
Reading CSV

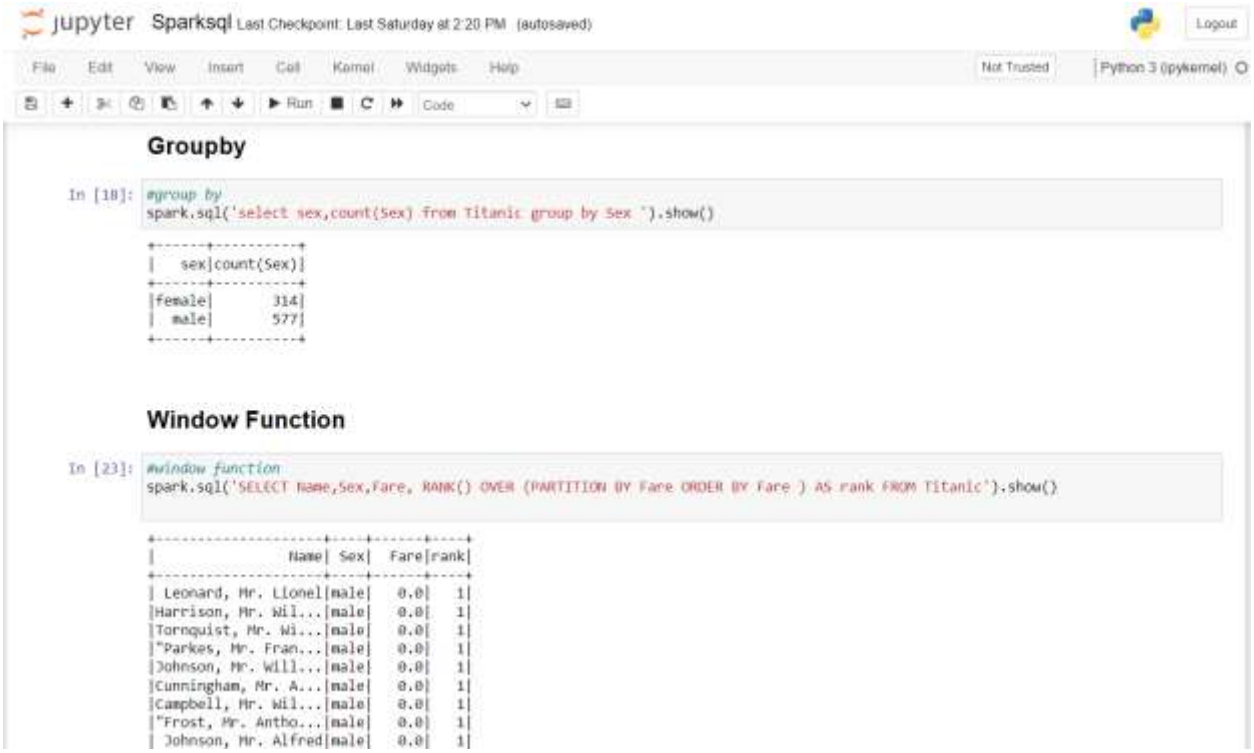
In [9]:
```
df=spark.read.csv('D:\Hexaware\Data_Engineering\Python\output_file.csv',header=True,inferSchema=True)
```

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

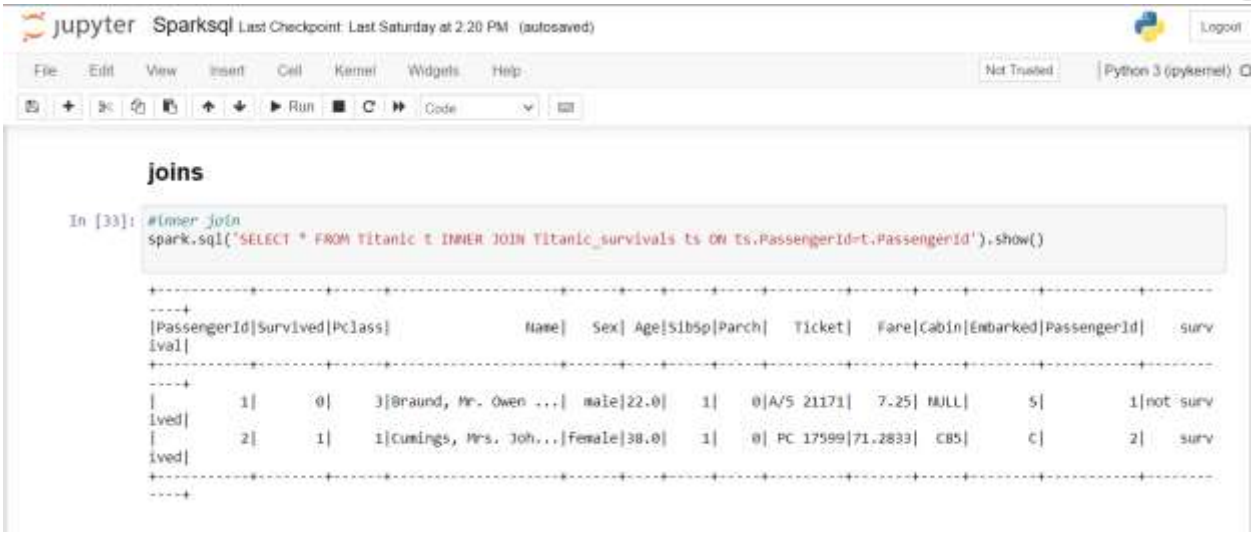
```
+-----------+--------+------+--------------------+------+----+-----+-----+----------------+-------+-----+--------+
|PassengerId|Survived|Pclass|                Name|   Sex| Age|SibSp|Parch|          Ticket|   Fare|Cabin|Embarked|
+-----------+--------+------+--------------------+------+----+-----+-----+----------------+-------+-----+--------+
|          1|       0|     3|Braund, Mr. Owen ...|  male|22.0|    1|    0|       A/5 21171|   7.25| NULL|       S|
|          2|       1|     1|Cumings, Mrs. Joh...|female|38.0|    1|    0|        PC 17599|71.2833|  C85|       C|
|          3|       1|     3|Heikkinen, Miss. ...|female|26.0|    0|    0|STON/O2. 3101282|  7.925| NULL|       S|
|          4|       1|     1|Futrelle, Mrs. Ja...|female|35.0|    1|    0|          113803|   53.1| C123|       S|
|          5|       0|     3|Allen, Mr. Willia...|  male|35.0|    0|    0|          373450|   8.05| NULL|       S|
|          6|       0|     3|    Moran, Mr. James|  male|NULL|    0|    0|          330877| 8.4583| NULL|       Q|
|          7|       0|     1|McCarthy, Mr. Tim...|  male|54.0|    0|    0|           17463|51.8625|  E46|       S|
|          8|       0|     3|Palsson, Master. ...|  male| 2.0|    3|    1|          349909| 21.075| NULL|       S|
|          9|       1|     3|Johnson, Mrs. Osc...|female|27.0|    0|    2|          347742|11.1333| NULL|       S|
|         10|       1|     2|Nasser, Mrs. Nich...|female|14.0|    1|    0|          237736|30.0708| NULL|       C|
|         11|       1|     3|Sandstrom, Miss. ...|female| 4.0|    1|    1|         PP 9549|   16.7|   G6|       S|
|         12|       1|     1|Bonnell, Miss. El...|female|58.0|    0|    0|          113783|  26.55| C103|       S|
|         13|       0|     3|Saundercock, Mr. ...|  male|20.0|    0|    0|       A/5. 2151|   8.05| NULL|       S|
|         14|       0|     3|Andersson, Mr. An...|  male|39.0|    1|    5|          347082| 31.275| NULL|       S|
|         15|       0|     3|Vestrom, Miss. Hu...|female|14.0|    0|    0|          350406| 7.8542| NULL|       S|
|         16|       1|     2|Hewlett, Mrs. (Ma...|female|55.0|    0|    0|          248706|   16.0| NULL|       S|
|         17|       0|     3|Rice, Master. Eugene|  male| 2.0|    4|    1|          382652| 29.125| NULL|       Q|
|         18|       1|     2|Williams, Mr. Cha...|  male|NULL|    0|    0|          244373|   13.0| NULL|       S|
|         19|       0|     3|Vander Planke, Mr...|female|31.0|    1|    0|          345763|   18.0| NULL|       S|
|         20|       1|     3|Masselmani, Mrs. ...|female|NULL|    0|    0|            2649|  7.225| NULL|       C|
+-----------+--------+------+--------------------+------+----+-----+-----+----------------+-------+-----+--------+
only showing top 20 rows
```

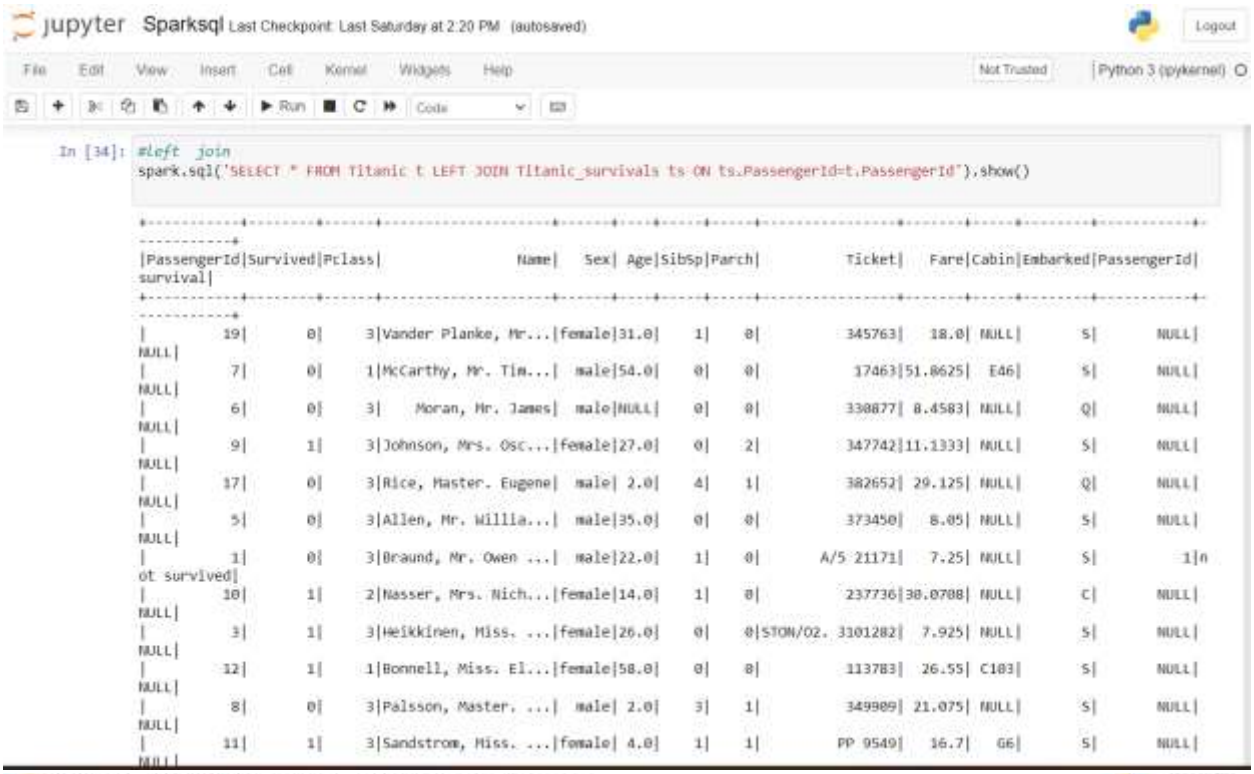

jupyter Sparksql Last Checkpoint: Last Saturday at 2:20 PM (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code
Creating View
In [11]: df.createOrReplaceTempView('Titanic')
In [12]: df1=spark.sql('Select Survived,Name,Sex from Titanic where Survived=1')
In [13]: df1.show()
+--------+--------------------+------+
|Survived| Name| Sex|
+--------+--------------------+------+
1	Cumings, Mrs. Joh...	female
1	Heikkinen, Miss. ...	female
1	Futrelle, Mrs. Ja...	female
1	Johnson, Mrs. Osc...	female
1	Nasser, Mrs. Nich...	female
1	Sandstrom, Miss. ...	female
1	Bonnell, Miss. El...	female
1	Hewlett, Mrs. (Ma...	female
1	Williams, Mr. Cha...	male
1	Masselmani, Mrs. ...	female
1	Beesley, Mr. Lawr...	male
1	"McGowan, Miss. A...	female
1	Sloper, Mr. Willi...	male
1	Asplund, Mrs. Car...	female
1	"O'Dwyer, Miss. E...	female
1	Spencer, Mrs. Wil...	female
1	Glynn, Miss. Mary...	female
1	Mamee, Mr. Hanna	male
1	Nicola-Yarred, Mi...	female
1	Laroche, Miss. Si...	female
+--------+--------------------+------+
jupyter Sparksql Last Checkpoint: Last Saturday at 2:20 PM (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code
Details in DB
In [14]: spark.catalog.currentDatabase()
Out[14]: 'default'
In [15]: spark.catalog.listTables()
Out[15]: [Table(name='Titanic', catalog=None, namespace=[], description=None, tableType='TEMPORARY', isTemporary=True)]
In [16]: spark.catalog.dropTempView('Employee')
Out[16]: False


jupyter Sparksql Last Checkpoint: Last Saturday at 2:20 PM (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (pykernel)
Run
Code
Aggregations in sparksql
In [17]:
#count
spark.sql('select count(*) from titanic').show()
spark.sql('select max(Fare) from titanic').show()
spark.sql('select min(Fare) from titanic').show()
spark.sql('select avg(Fare) from titanic').show()
spark.sql('select distinct count(*) as total_rows from titanic').show()
+--------+
|count(1)|
+--------+
| 891|
+--------+
+---------+
|max(Fare)|
+---------+
| 512.3292|
+---------+
+---------+
|min(Fare)|
+---------+
| 0.0|
+---------+
+-----------------+
| avg(Fare)|
+-----------------+
|32.2042079685746|
+-----------------+


jupyter Sparksql Last Checkpoint: Last Saturday at 2:20 PM (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (pykernel)
Run
Code
Groupby
In [18]:
#group by
spark.sql('select sex,count(Sex) from Titanic group by Sex ').show()
+------+----------+
| sex|count(Sex)|
+------+----------+
|female| 314|
| male| 577|
+------+----------+
Window Function
In [23]:
#window function
spark.sql('SELECT Name,Sex,Fare, RANK() OVER (PARTITION BY Fare ORDER BY Fare) AS rank FROM Titanic').show()
+--------------------+----+------+----+
| Name| Sex| Fare|rank|
+--------------------+----+------+----+
Leonard, Mr. Lionel	male	0.0	1
Harrison, Mr. Wil...	male	0.0	1
Tornquist, Mr. Wi...	male	0.0	1
"Parkes, Mr. Fran...	male	0.0	1
Johnson, Mr. Will...	male	0.0	1
Cunningham, Mr. A...	male	0.0	1
Campbell, Mr. Wil...	male	0.0	1
"Frost, Mr. Antho...	male	0.0	1
Johnson, Mr. Alfred	male	0.0	1


jupyter Sparksql Last Checkpoint: Last Saturday at 2:20 PM (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (pykernel)
Run
Code
Creating DF for Joins
In [25]:
data=[(1,'not survived'),(2,'survived')]
column=['PassengerId','survival']
df2=spark.createDataFrame(data,column)
df2.show()
+-----------+------------+
|PassengerId| survival|
+-----------+------------+
| 1|not survived|
| 2| survived|
+-----------+------------+
In [26]:
df2.createOrReplaceTempView('Titanic_survivals')
In [28]:
spark.sql('select * from Titanic_survivals').show()
+-----------+------------+
|PassengerId| survival|
+-----------+------------+
| 1|not survived|
| 2| survived|
+-----------+------------+


jupyter Sparksql Last Checkpoint: Last Saturday at 2:20 PM (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (pykernel)
Run
Code
joins
In [33]:
#inner join
spark.sql('SELECT * FROM Titanic t INNER JOIN Titanic_survivals ts ON ts.PassengerId=t.PassengerId').show()
+-----------+--------+------+--------------------+------+----+-----+-----+---------+-------+-----+--------+-----------+--------
----+
|PassengerId|Survived|Pclass| Name| Sex| Age|SibSp|Parch| Ticket| Fare|Cabin|Embarked|PassengerId| surv
ival|
+-----------+--------+------+--------------------+------+----+-----+-----+---------+-------+-----+--------+-----------+--------
----+
| 1| 0| 3|Braund, Mr. Owen ...| male|22.0| 1| 0|A/5 21171| 7.25| NULL| S| 1|not surv
ived|
| 2| 1| 1|Cumings, Mrs. Joh...|female|38.0| 1| 0| PC 17599|71.2833| C85| C| 2| surv
ived|
+-----------+--------+------+--------------------+------+----+-----+-----+---------+-------+-----+--------+-----------+--------
----+


jupyter Sparksql Last Checkpoint: Last Saturday at 2:20 PM (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code
In [34]:
#Left join
spark.sql('SELECT * FROM Titanic t LEFT JOIN Titanic_survivals ts ON ts.PassengerId=t.PassengerId').show()
+-----------+--------+------+--------------------+------+----+-----+-----+----------------+-------+-----+--------+-----------+--------
-----------+
|PassengerId|Survived|Pclass| Name| Sex| Age|SibSp|Parch| Ticket| Fare|Cabin|Embarked|PassengerId|
survival|
+-----------+--------+------+--------------------+------+----+-----+-----+----------------+-------+-----+--------+-----------+--------
-----------+
| 19| 0| 3|Vander Planke, Mr...|female|31.0| 1| 0| 345763| 18.0| NULL| S| NULL|
NULL|
| 7| 0| 1|McCarthy, Mr. Tim...| male|54.0| 0| 0| 17463|51.8625| E46| S| NULL|
NULL|
| 6| 0| 3| Moran, Mr. James| male|NULL| 0| 0| 330877| 8.4583| NULL| Q| NULL|
NULL|
| 9| 1| 3|Johnson, Mrs. Osc...|female|27.0| 0| 2| 347742|11.1333| NULL| S| NULL|
NULL|
| 17| 0| 3|Rice, Master. Eugene| male| 2.0| 4| 1| 382652| 29.125| NULL| Q| NULL|
NULL|
| 5| 0| 3|Allen, Mr. Willia...| male|35.0| 0| 0| 373450| 8.05| NULL| S| NULL|
NULL|
| 1| 0| 3|Braund, Mr. Owen ...| male|22.0| 1| 0| A/5 21171| 7.25| NULL| S| 1|n
ot survived|
| 10| 1| 2|Nasser, Mrs. Nich...|female|14.0| 1| 0| 237736|30.0708| NULL| C| NULL|
NULL|
| 3| 1| 3|Heikkinen, Miss. ...|female|26.0| 0| 0|STON/O2. 3101282| 7.925| NULL| S| NULL|
NULL|
| 12| 1| 1|Bonnell, Miss. El...|female|58.0| 0| 0| 113783| 26.55| C103| S| NULL|
NULL|
| 8| 0| 3|Palsson, Master. ...| male| 2.0| 3| 1| 349909| 21.075| NULL| S| NULL|
NULL|
| 11| 1| 3|Sandstrom, Miss. ...|female| 4.0| 1| 1| PP 9549| 16.7| G6| S| NULL|
NULL|
+-----------+--------+------+--------------------+------+----+-----+-----+----------------+-------+-----+--------+-----------+--------
-----------+
only showing top 20 rows
In [35]:
#Right join
spark.sql('SELECT * FROM Titanic t RIGHT JOIN Titanic_survivals ts ON ts.PassengerId=t.PassengerId').show()
+-----------+--------+------+--------------------+------+----+-----+-----+---------+-------+-----+--------+-----------+--------
----+
|PassengerId|Survived|Pclass| Name| Sex| Age|SibSp|Parch| Ticket| Fare|Cabin|Embarked|PassengerId| surv
ival|
+-----------+--------+------+--------------------+------+----+-----+-----+---------+-------+-----+--------+-----------+--------
----+
| 1| 0| 3|Braund, Mr. Owen ...| male|22.0| 1| 0|A/5 21171| 7.25| NULL| S| 1|not surv
ived|
| 2| 1| 1|Cumings, Mrs. Joh...|female|38.0| 1| 0| PC 17599|71.2833| C85| C| 2| surv
ived|
+-----------+--------+------+--------------------+------+----+-----+-----+---------+-------+-----+--------+-----------+--------
----+