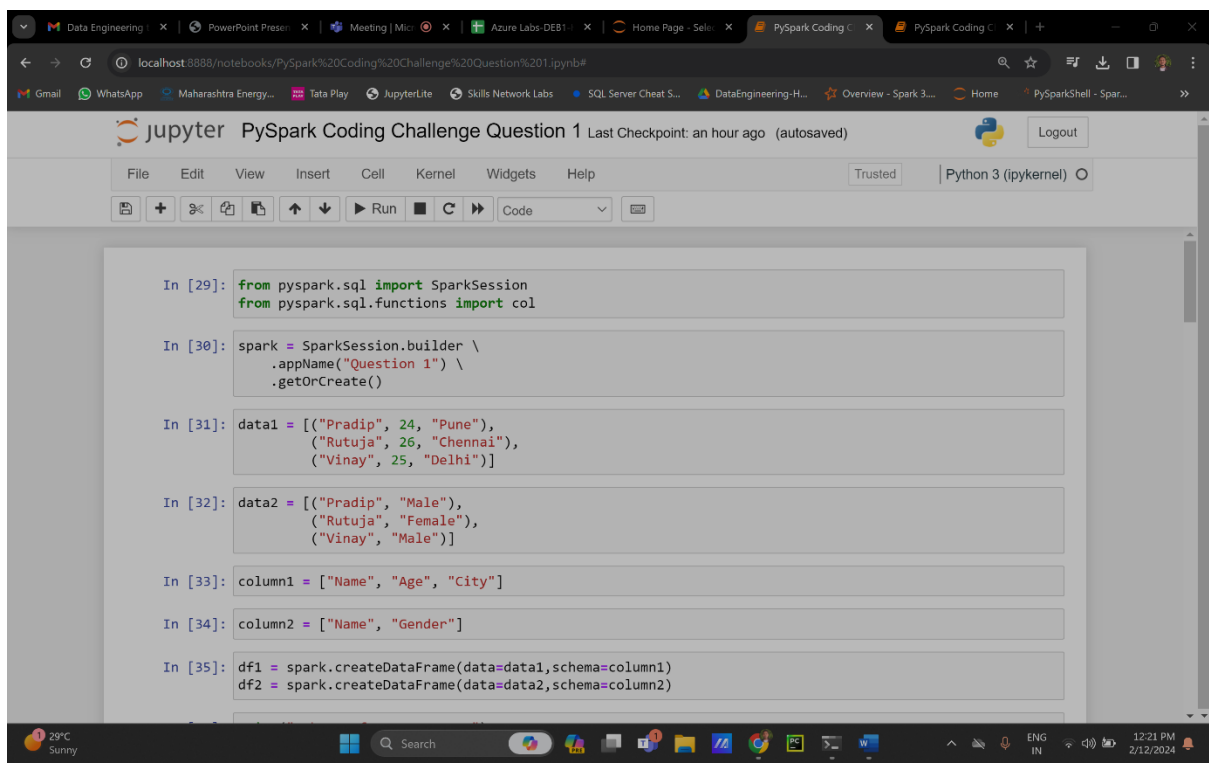


PySpark Coding Challenge Question 1

🚀 Execute Manipulating, Dropping, Sorting, Aggregations, Joining, GroupBy DataFrames



The screenshot shows a Jupyter Notebook titled "PySpark Coding Challenge Question 1" running on a local host. The notebook contains the following code cells:

```
In [29]: from pyspark.sql import SparkSession
        from pyspark.sql.functions import col

In [30]: spark = SparkSession.builder \
        .appName("Question 1") \
        .getOrCreate()

In [31]: data1 = [("Pradip", 24, "Pune"),
        ("Rutuja", 26, "Chennai"),
        ("Vinay", 25, "Delhi")]

In [32]: data2 = [("Pradip", "Male"),
        ("Rutuja", "Female"),
        ("Vinay", "Male")]

In [33]: column1 = ["Name", "Age", "City"]

In [34]: column2 = ["Name", "Gender"]

In [35]: df1 = spark.createDataFrame(data=data1, schema=column1)
        df2 = spark.createDataFrame(data=data2, schema=column2)
```

The notebook interface includes a menu bar (File, Edit, View, Insert, Cell, Kernel, Widgets, Help), a toolbar with icons for file operations and execution, and a status bar at the bottom showing the temperature (29°C Sunny) and the time (12:21 PM 2/12/2024).

PySpark Coding Challenge Question 1 Last Checkpoint: an hour ago (autosaved)

```
In [35]: df1 = spark.createDataFrame(data=data1,schema=column1)
df2 = spark.createDataFrame(data=data2,schema=column2)

In [36]: print("Schema of DataFrame 1:")
df1.printSchema()
print("\nSchema of DataFrame 2:")
df2.printSchema()

Schema of DataFrame 1:
root
 |-- Name: string (nullable = true)
 |-- Age: long (nullable = true)
 |-- City: string (nullable = true)

Schema of DataFrame 2:
root
 |-- Name: string (nullable = true)
 |-- Gender: string (nullable = true)

In [37]: df1.show()
df2.show()
```

Name	Age	City
Pradip	24	Pune
Rutuja	26	Chennai
Vinay	25	Delhi

```
In [37]: df1.show()
df2.show()

+-----+-----+
| Name|Age|  City|
+-----+-----+
|Pradip| 24|   Pune|
|Rutuja| 26|Chennai|
| Vinay| 25|   Delhi|
+-----+-----+

+-----+-----+
| Name|Gender|
+-----+-----+
|Pradip|  Male|
|Rutuja|Female|
| Vinay|  Male|
+-----+-----+

In [38]: # manipulating dataframe
print("df1 after manipulating:")
df1.withColumnRenamed("City","Location").show()

df1 after manipulating:
+-----+-----+
| Name|Age|Location|
+-----+-----+
|Pradip| 24|    Pune|
|Rutuja| 26|Chennai|
| Vinay| 25|    Delhi|
+-----+-----+
```

PySpark Coding Challenge Question 1 Last Checkpoint: an hour ago (autosaved)

In [38]:

```
# manipulating dataframe
print("df1 after manipulating:")
df1.withColumnRenamed("City","Location").show()
```

df1 after manipulating:

Name	Age	Location
Pradip	24	Pune
Rutuja	26	Chennai
Vinay	25	Delhi

In [39]:

```
# sorting dataframe
print("df1 after sorting:")
df1_sorted = df1.orderBy("Age").show()
```

df1 after sorting:

Name	Age	City
Pradip	24	Pune
Vinay	25	Delhi
Rutuja	26	Chennai

PySpark Coding Challenge Question 1 Last Checkpoint: an hour ago (autosaved)

In [40]:

```
# aggregations
print("aggregated df1:")
df1_agg = df1.groupBy("Name").count().show()
```

aggregated df1:

Name	count
Pradip	1
Rutuja	1
Vinay	1

In [41]:

```
# Joining Dataframes
print("Joined Dataframe:")
joined_df = df1.join(df2, on="Name", how="inner").show()
```

Joined Dataframe:

Name	Age	City	Gender
Pradip	24	Pune	Male
Rutuja	26	Chennai	Female
Vinay	25	Delhi	Male

Browser tabs: Data Engineering, PowerPoint Present..., Meeting | Mic..., Azure Labs-DEB1, Home Page - Sele..., PySpark Coding C..., PySpark Coding C..., +

Address bar: localhost:8888/notebooks/PySpark%20Coding%20Challenge%20Question%201.ipynb#

Browser extensions: Gmail, WhatsApp, Maharashtra Energy..., Tata Play, JupyterLite, Skills Network Labs, SQL Server Cheat S..., DataEngineering-H..., Overview - Spark 3..., Home, PySparkShell - Spar...

Jupyter PySpark Coding Challenge Question 1

Last Checkpoint: an hour ago (autosaved) [Logout](#)

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

Run Code

```
In [45]: Data_list = ["Emp Name", "Age (Year)", "Location"]
new_df = df1.toDF(*Data_list)
new_df.show()

+-----+
|Emp Name|Age (Year)|Location|
+-----+
| Pradip |      24 |    Pune |
| Rutuja |      26 | Chennai |
| Vinay  |      25 |    Delhi |
+-----+
```

```
In [46]: # Dropping column from DataFrame
print("df2 after dropping column Gender:")
df2 = df2.drop("Gender").show()

df2 after dropping column Gender:
+-----+
| Name |
+-----+
| Pradip |
| Rutuja |
| Vinay  |
+-----+
```

23°C Sunny Search 12:22 PM 2/12/2024