Untitled5

June 5, 2025

[1]:
```
from pyspark.sql import SparkSession
from pyspark.sql.functions import col, avg, countDistinct, desc, row_number,␣
 ↪explode, split
from pyspark.sql.window import Window
```

[2]:
```
spark = SparkSession.builder\
                    .master("local")\
                    .appName("demo")\
                    .getOrCreate()
```

[3]:
```
df = spark.read.option("header", True).csv("gs://datapyspark/employee_data.
 ↪csv", inferSchema=True)
```

[4]:
```
df.show()
```

```
+-----+-------------------+---------+---------+--------------------+----------+
|empId|            empName|empGender|empSalary|          empCountry|  hireDate|
+-----+-------------------+---------+---------+--------------------+----------+
|    1|   Richard Williams|   Female|    47329|              Malawi|2018-05-05|
|    2|         Devin Khan|     Male|   103880|          Azerbaijan|2015-12-23|
|    3|  Mr. Michael Ayala|   Female|   108197|              Rwanda|2019-06-08|
|    4|    Deborah Burnett|     Male|    80102|            Colombia|2022-10-10|
|    5|     Heather Greene|   Female|    96146|             Somalia|2018-11-22|
|    6|     Patricia Weiss|   Female|    98525|        South Africa|2021-06-21|
|    7|  Nicholas Woodward|     Male|   107971|              Jordan|2020-11-15|
|    8|      Amy Rodriguez|   Female|    62957|             Vietnam|2022-09-23|
|    9|     Anthony Holt V|     Male|    90547|    Marshall Islands|2016-07-02|
|   10|          Paul Wood|     Male|    92935|             Bahrain|2020-02-13|
|   11|     Jeremy Johnson|   Female|    77341|United Arab Emirates|2017-01-05|
|   12|         John Brady|   Female|   110007|            Slovenia|2022-10-02|
|   13|    Andrew Peterson|   Female|    80136|            Colombia|2016-08-08|
|   14|Dr. Bailey Anderson|     Male|    75087|Netherlands Antilles|2016-06-30|
|   15|       Paul Johnson|     Male|    85158|             Belgium|2021-03-08|
|   16|      Daniel Rivera|     Male|    87814|             Ukraine|2020-11-07|
|   17|        Evelyn Reed|   Female|   115953|           Gibraltar|2024-06-08|
|   18|       David Newton|   Female|    80115|              Uganda|2023-12-24|
|   19|         Raymond Lee|   Female|    87908|             Bermuda|2024-08-20|
|   20|        Kathryn Yang|     Male|   108813|    Marshall Islands|2023-03-30|
+-----+--------------------+---------+---------+--------------------+----------+
only showing top 20 rows
```

[5]:

```
# 1. Top 3 most occurring words in empName
words_df = df.select(explode(split(col("empName"), " ")).alias("word"))
top_words = words_df.groupBy("word").count().orderBy(desc("count")).limit(3)
top_words.show()
```

```
[Stage 5:===================>                                     (1 + 1) / 3]

+--------+-----+
|    word|count|
+--------+-----+
|Jennifer|    3|
|  Daniel|    3|
|    Paul|    3|
+--------+-----+
```

[6]:

```
# 2. Remove duplicate records
df_no_duplicates = df.dropDuplicates()
df_no_duplicates.show()
```

```
+-----+----------------+---------+---------+--------------------+----------+
|empId|         empName|empGender|empSalary|          empCountry|  hireDate|
+-----+----------------+---------+---------+--------------------+----------+
|   44|Adrienne Sanchez|   Female|    86290|       Liechtenstein|2017-03-05|
|   43|     Jaime Perez|     Male|    67462|       New Caledonia|2024-07-22|
|   27|     Paul Nguyen|   Female|    60067|         Netherlands|2018-12-14|
|    4| Deborah Burnett|     Male|    80102|            Colombia|2022-10-10|
|   25| Denise Gonzales|   Female|    58877|           Indonesia|2018-01-10|
|    2|      Devin Khan|     Male|   103880|          Azerbaijan|2015-12-23|
|   28| Allison Roberts|   Female|   118978|               Sudan|2018-09-25|
|   39|   John Williams|     Male|    61439|               Haiti|2019-06-09|
|   26|     Andrea Rose|     Male|    93428|              Cyprus|2016-10-05|
|   33|Amber Wright DVM|   Female|   116155|Bosnia and Herzeg…|2019-10-20|
|   10|       Paul Wood|     Male|    92935|             Bahrain|2020-02-13|
|   31|     Ellen Baker|   Female|    37040|               Spain|2020-07-21|
|   42|  Tiffany Deleon|     Male|    48674|  Dominican Republic|2018-09-19|
|   29|Jennifer Salinas|     Male|    65388|French Southern T…|2019-07-18|
|    9|   Anthony Holt V|     Male|    90547|    Marshall Islands|2016-07-02|
|   21|   David Compton|   Female|    71507|Saint Kitts and N…|2019-10-03|
|   49|  Terry Mitchell|   Female|    67579|              Guinea|2019-02-06|
|   32|    Kayla Nguyen|     Male|    81046|      American Samoa|2022-09-03|
|    22|    Daniel Boone|   Female|    95277|     Solomon Islands|2016-05-13|
|    36|       Jill Long|     Male|    72665|             Bermuda|2018-11-22|
+------+----------------+---------+---------+--------------------+----------+
only showing top 20 rows
```

[7]:

```python
# 3. Word count using PySpark (on empName)
word_count = words_df.groupBy("word").count()
word_count.show()
```

```
[Stage 10:>                                                         (0 + 1) / 1]

+--------+-----+
|    word|count|
+--------+-----+
| Compton|    1|
|  Andrea|    2|
| Sanchez|    1|
|   Brady|    1|
|   James|    1|
|   Jaime|    1|
| Gilmore|    1|
|  Denise|    1|
|  Nguyen|    2|
|Jennifer|    3|
|     DVM|    1|
|    Rose|    1|
|Williams|    2|
|  Rivera|    1|
|   Amber|    1|
| Deborah|    1|
|Peterson|    1|
|  Miller|    1|
|  Wright|    1|
|  Howell|    1|
+--------+-----+
only showing top 20 rows
```

[8]:

```python
# 4. Group by empCountry and calculate average salary
avg_salary = df.groupBy("empCountry").agg(avg("empSalary").alias("avgSalary"))
avg_salary.show()
```

```
+--------------------+---------+
|          empCountry|avgSalary|
+--------------------+---------+
|Heard Island and …|  39311.0|
|French Southern T…|  65388.0|
|              Turkey|  66545.0|
|              Malawi|  47329.0|
|              Rwanda| 108197.0|
|              Jordan| 107971.0|
|               Sudan| 118978.0|
|   Equatorial Guinea|  63383.0|
|             Belgium|  85158.0|
|             Ecuador|  62624.0|
|       New Caledonia|  82432.0|
|      American Samoa|  81046.0|
|             Somalia|  96146.0|
|    Marshall Islands|  99680.0|
|Netherlands Antilles|  75087.0|
|               Spain|  37040.0|
|  Russian Federation|  82395.0|
|       Liechtenstein|  86290.0|
|            Thailand|  46582.0|
|             Ukraine|  87814.0|
+--------------------+---------+
only showing top 20 rows
```

```
[9]: # 5. Handle missing/null values
df_cleaned = df.na.fill({"empName": "Unknown", "empSalary": 0})
df_cleaned.show()
```

```
+-----+-------------------+---------+---------+--------------------+----------+
|empId|            empName|empGender|empSalary|          empCountry|  hireDate|
+-----+-------------------+---------+---------+--------------------+----------+
|    1|   Richard Williams|   Female|    47329|              Malawi|2018-05-05|
|    2|         Devin Khan|     Male|   103880|          Azerbaijan|2015-12-23|
|    3|  Mr. Michael Ayala|   Female|   108197|              Rwanda|2019-06-08|
|    4|    Deborah Burnett|     Male|    80102|            Colombia|2022-10-10|
|    5|     Heather Greene|   Female|    96146|             Somalia|2018-11-22|
|    6|     Patricia Weiss|   Female|    98525|        South Africa|2021-06-21|
|    7|  Nicholas Woodward|     Male|   107971|              Jordan|2020-11-15|
|    8|      Amy Rodriguez|   Female|    62957|             Vietnam|2022-09-23|
|    9|     Anthony Holt V|     Male|    90547|    Marshall Islands|2016-07-02|
|   10|          Paul Wood|     Male|    92935|             Bahrain|2020-02-13|
|   11|     Jeremy Johnson|   Female|    77341|United Arab Emirates|2017-01-05|
|   12|         John Brady|   Female|   110007|            Slovenia|2022-10-02|
|   13|    Andrew Peterson|   Female|    80136|            Colombia|2016-08-08|
|   14|Dr. Bailey Anderson|     Male|    75087|Netherlands Antilles|2016-06-30|
|   15|       Paul Johnson|     Male|    85158|             Belgium|2021-03-08|
|   16|      Daniel Rivera|     Male|    87814|             Ukraine|2020-11-07|
|   17|        Evelyn Reed|   Female|   115953|           Gibraltar|2024-06-08|
|   18|       David Newton|   Female|    80115|              Uganda|2023-12-24|
|   19|        Raymond Lee|   Female|    87908|             Bermuda|2024-08-20|
|   20|       Kathryn Yang|     Male|   108813|    Marshall Islands|2023-03-30|
+-----+-------------------+---------+---------+--------------------+----------+
only showing top 20 rows
```

[10]:

```
# 6. Count distinct values in empCountry
distinct_countries = df.select("empCountry").distinct().count()
print(f"Distinct countries: {distinct_countries}")
```

```
Distinct countries: 45
```

[11]:

```
# 7. Filter records where salary > 50000
high_salary = df.filter(col("empSalary") > 50000)
high_salary.show()
```

```
+-----+-------------------+---------+---------+--------------------+----------+
|empId|            empName|empGender|empSalary|          empCountry|  hireDate|
+-----+-------------------+---------+---------+--------------------+----------+
|    2|         Devin Khan|     Male|   103880|          Azerbaijan|2015-12-23|
|    3|  Mr. Michael Ayala|   Female|   108197|              Rwanda|2019-06-08|
|    4|    Deborah Burnett|     Male|    80102|            Colombia|2022-10-10|
|    5|     Heather Greene|   Female|    96146|             Somalia|2018-11-22|
|    6|     Patricia Weiss|   Female|    98525|        South Africa|2021-06-21|
|    7|  Nicholas Woodward|     Male|   107971|              Jordan|2020-11-15|
|    8|      Amy Rodriguez|   Female|    62957|             Vietnam|2022-09-23|
|    9|     Anthony Holt V|     Male|    90547|    Marshall Islands|2016-07-02|
|   10|          Paul Wood|     Male|    92935|             Bahrain|2020-02-13|
|   11|     Jeremy Johnson|   Female|    77341|United Arab Emirates|2017-01-05|
|   12|         John Brady|   Female|   110007|            Slovenia|2022-10-02|
|   13|    Andrew Peterson|   Female|    80136|            Colombia|2016-08-08|
|   14|Dr. Bailey Anderson|     Male|    75087|Netherlands Antilles|2016-06-30|
|   15|       Paul Johnson|     Male|    85158|             Belgium|2021-03-08|
|   16|      Daniel Rivera|     Male|    87814|             Ukraine|2020-11-07|
|   17|        Evelyn Reed|   Female|   115953|           Gibraltar|2024-06-08|
|   18|       David Newton|   Female|    80115|              Uganda|2023-12-24|
|   19|        Raymond Lee|   Female|    87908|             Bermuda|2024-08-20|
|   20|       Kathryn Yang|     Male|   108813|    Marshall Islands|2023-03-30|
|   21|      David Compton|   Female|    71507|Saint Kitts and N…|2019-10-03|
+-----+-------------------+---------+---------+--------------------+----------+
only showing top 20 rows
```

[12]:

```
# 9. Find second highest salary
window_spec = Window.orderBy(col("empSalary").desc())
second_highest_salary = df.withColumn("rank", row_number().over(window_spec)).
 ↪filter(col("rank") == 2)
second_highest_salary.show()
```

```
25/06/05 10:57:04 WARN org.apache.spark.sql.execution.window.WindowExec: No
Partition Defined for Window operation! Moving all data to a single partition,
this can cause serious performance degradation.
25/06/05 10:57:04 WARN org.apache.spark.sql.execution.window.WindowExec: No
Partition Defined for Window operation! Moving all data to a single partition,
this can cause serious performance degradation.
25/06/05 10:57:05 WARN org.apache.spark.sql.execution.window.WindowExec: No
Partition Defined for Window operation! Moving all data to a single partition,
this can cause serious performance degradation.
25/06/05 10:57:05 WARN org.apache.spark.sql.execution.window.WindowExec: No
Partition Defined for Window operation! Moving all data to a single partition,
this can cause serious performance degradation.
[Stage 26:>                                                         (0 + 1) / 1]

+-----+----------------+---------+---------+--------------------+----------+----
+
|empId|         empName|empGender|empSalary|          empCountry|
hireDate|rank|
+-----+----------------+---------+---------+--------------------+----------+----
+
|   33|Amber Wright DVM|   Female|   116155|Bosnia and Herzeg…|2019-10-20|
2|
+-----+----------------+---------+---------+--------------------+----------+----
+
```

[13]:

```
# 10. Join two DataFrames and select specific columns
df2 = df.select("empId", "empCountry").withColumnRenamed("empCountry",␣
 ↪"country")
joined_df = df.join(df2, "empId").select("empId", "empName", "empSalary",␣
 ↪"country")
joined_df.show()
```

```
+-----+-------------------+---------+--------------------+
|empId|            empName|empSalary|             country|
+-----+-------------------+---------+--------------------+
|    1|   Richard Williams|    47329|              Malawi|
|    2|         Devin Khan|   103880|          Azerbaijan|
|    3|  Mr. Michael Ayala|   108197|              Rwanda|
|    4|    Deborah Burnett|    80102|            Colombia|
|    5|     Heather Greene|    96146|             Somalia|
|    6|     Patricia Weiss|    98525|        South Africa|
|    7|  Nicholas Woodward|   107971|              Jordan|
|    8|      Amy Rodriguez|    62957|             Vietnam|
|    9|     Anthony Holt V|    90547|    Marshall Islands|
|   10|          Paul Wood|    92935|             Bahrain|
|   11|     Jeremy Johnson|    77341|United Arab Emirates|
|   12|         John Brady|   110007|            Slovenia|
|   13|    Andrew Peterson|    80136|            Colombia|
|   14|Dr. Bailey Anderson|    75087|Netherlands Antilles|
|   15|       Paul Johnson|    85158|             Belgium|
|   16|      Daniel Rivera|    87814|             Ukraine|
|   17|        Evelyn Reed|   115953|           Gibraltar|
|   18|       David Newton|    80115|              Uganda|
|   19|        Raymond Lee|    87908|             Bermuda|
|   20|       Kathryn Yang|   108813|    Marshall Islands|
+-----+-------------------+---------+--------------------+
only showing top 20 rows
```

[16]:

```
# 8. Read JSON file and convert to DataFrame
json_df = spark.read.option("multiline", True).json("gs://datapyspark/
 ↪sample_json.json")
json_df.show()
```

```
25/06/05 11:00:04 WARN org.apache.hadoop.util.concurrent.ExecutorHelper: Thread
(Thread[GetFileInfo #0,5,main]) interrupted:
java.lang.InterruptedException
        at
com.google.common.util.concurrent.AbstractFuture.get(AbstractFuture.java:510)
        at com.google.common.util.concurrent.FluentFuture$TrustedFuture.get(Flue
ntFuture.java:88)
        at org.apache.hadoop.util.concurrent.ExecutorHelper.logThrowableFromAfte
rExecute(ExecutorHelper.java:48)
        at org.apache.hadoop.util.concurrent.HadoopThreadPoolExecutor.afterExecu
te(HadoopThreadPoolExecutor.java:90)
        at
java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1157)
        at
java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:624)
        at java.lang.Thread.run(Thread.java:750)
25/06/05 11:00:04 WARN org.apache.hadoop.util.concurrent.ExecutorHelper: Thread
(Thread[GetFileInfo #1,5,main]) interrupted:
java.lang.InterruptedException
        at
com.google.common.util.concurrent.AbstractFuture.get(AbstractFuture.java:510)
        at com.google.common.util.concurrent.FluentFuture$TrustedFuture.get(Flue
ntFuture.java:88)
        at org.apache.hadoop.util.concurrent.ExecutorHelper.logThrowableFromAfte
rExecute(ExecutorHelper.java:48)
        at org.apache.hadoop.util.concurrent.HadoopThreadPoolExecutor.afterExecu
te(HadoopThreadPoolExecutor.java:90)
        at
java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1157)
        at
```

```
java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:624)
        at java.lang.Thread.run(Thread.java:750)

+--------------------+--------+--------------------+-------+
|            location|    name|            sessions|user_id|
+--------------------+--------+--------------------+-------+
|     {New York, USA}|   Alice|[{[{30, /home}, {…|   U001|
|{San Francisco, USA}|     Bob|[{[{10, /home}, {…|   U002|
|     {Madrid, Spain}|  Carlos|                  []|   U003|
|   {Berlin, Germany}|   Diana|[{[{20, /home}, {…|   U004|
|   {Toronto, Canada}|   Ethan|[{[{10, /login}],…|   U005|
|        {Dubai, UAE}|  Fatima|                  []|   U006|
|        {London, UK}|  George|[{[{60, /home}], …|   U007|
|      {Tokyo, Japan}| Hiroshi|[{[{15, /home}, {…|   U008|
|     {Paris, France}|Isabelle|                  []|   U009|
|{Melbourne, Austr…|    Jack|[{[{5, /login}, {…|   U010|
+--------------------+--------+--------------------+-------+
```

[ ]: