
Dataset: `web_traffic_data`

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

web_data = [
    Row(UserID=1, Page="Home", Timestamp="2024-04-10 10:00:00", Duration=35,
Device="Mobile", Country="India"),
    Row(UserID=2, Page="Products", Timestamp="2024-04-10 10:02:00", Duration=120,
Device="Desktop", Country="USA"),
    Row(UserID=3, Page="Cart", Timestamp="2024-04-10 10:05:00", Duration=45,
Device="Tablet", Country="UK"),
    Row(UserID=1, Page="Checkout", Timestamp="2024-04-10 10:08:00", Duration=60,
Device="Mobile", Country="India"),
    Row(UserID=4, Page="Home", Timestamp="2024-04-10 10:10:00", Duration=15,
Device="Mobile", Country="Canada"),
    Row(UserID=2, Page="Contact", Timestamp="2024-04-10 10:15:00", Duration=25,
Device="Desktop", Country="USA"),
    Row(UserID=5, Page="Products", Timestamp="2024-04-10 10:20:00", Duration=90,
Device="Desktop", Country="India"),
]

df_web = spark.createDataFrame(web_data)
df_web.show(truncate=False)
```

PySpark Exercises – Set 5 (Web Traffic Data)

Data Exploration & Preparation

1. Display the schema of `web_traffic_data`.
2. Convert the `Timestamp` column to a proper `timestamp` type.
3. Add a new column `SessionMinute` by extracting the minute from the `Timestamp`.

Filtering and Conditions

4. Filter users who used a "Mobile" device and visited the "Checkout" page.
5. Show all entries with a `Duration` greater than 60 seconds.
6. Find all users from India who visited the "Products" page.

Aggregation and Grouping

7. Get the average duration per device type.
8. Count the number of sessions per country.
9. Find the most visited page overall.

Window Functions

10. Rank each user’s pages by timestamp (oldest to newest).
11. Find the total duration of all sessions per user using `groupBy`.

Spark SQL Tasks

12. Create a temporary view called `traffic_view`.
13. Write a SQL query to get the top 2 longest sessions by duration.
14. Get the number of unique users per page using SQL.

Export & Save

15. Save the final DataFrame to CSV.
16. Save partitioned by `Country` in Parquet format.
