



**14\_Purchasing Activity by Product Type**

We have been given purchasing activity DF and we need to find out cumulative purchases of each product over time.

**Difficult Level :** EASY

**DataFrame:**

**# Define schema for the DataFrame**

**schema = StructType([**

**StructField("order\_id", IntegerType(), True),**

**StructField("product\_type", StringType(), True),**

**StructField("quantity", IntegerType(), True),**

**StructField("order\_date", StringType(), True),**

**])**

**# Define data**

**# Define data**

**data = [**

**(213824, 'printer', 20, "2022-06-27 "),**

**(212312, 'hair dryer', 5, "2022-06-28 "),**

**(132842, 'printer', 18, "2022-06-28 "),**

**(284730, 'standing lamp', 8, "2022-07-05 ")**

**]**

|  |
| --- |



**INPUT**

| **INPUT** | | | |
| --- | --- | --- | --- |
| **ORDER\_ID** | **PRODUCT\_TYPE** | **QUANTITY** | **ORDER\_DATE** |
| 213824 | printer | 20 | 2022-06-27 12:00:00 |
| 212312 | hair dryer | 5 | 2022-06-28 12:00:00 |
| 132842 | printer | 18 | 2022-06-28 12:00:00 |
| 284730 | standing lamp | 8 | 2022-07-05 12:00:00 |

**OUTPUT**

| **OUTPUT** | | |
| --- | --- | --- |
| **ORDER\_DATE** | **PRODUCT\_TYPE** | **CUM\_PURCHASED** |
| 2022-06-27 12:00:00 | printer | 20 |
| 2022-06-28 12:00:00 | hair dryer | 5 |
| 2022-06-28 12:00:00 | printer | 38 |
| 2022-07-05 12:00:00 | standing lamp | 8 |



**# Creating Spark Session**

**from pyspark.sql import SparkSession**

**from pyspark.sql.types import StructType,StructField,IntegerType,StringType**

**#creating spark session**

**spark = SparkSession. \**

**builder. \**

**config('spark.shuffle.useOldFetchProtocol', 'true'). \**

**config('spark.ui.port','0'). \**

**config("spark.sql.warehouse.dir", "/user/itv008042/warehouse"). \**

**enableHiveSupport(). \**

**master('yarn'). \**

**getOrCreate()**

**# Define schema for the DataFrame**

**schema = StructType([**

**StructField("order\_id", IntegerType(), True),**

**StructField("product\_type", StringType(), True),**

**StructField("quantity", IntegerType(), True),**

**StructField("order\_date", StringType(), True),**

**])**

**# Define data**

**# Define data**

**data = [**

**(213824, 'printer', 20, "2022-06-27 "),**

**(212312, 'hair dryer', 5, "2022-06-28 "),**

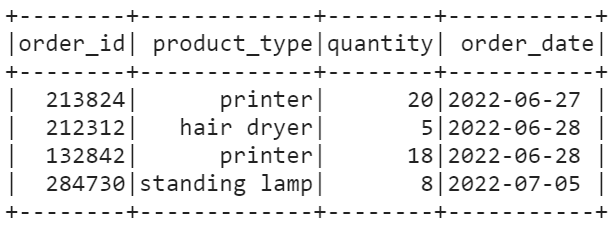
**(132842, 'printer', 18, "2022-06-28 "),**

**(284730, 'standing lamp', 8, "2022-07-05 ")**

**]**

**order\_df=spark.createDataFrame(data,schema)**

**order\_df.show()**

****

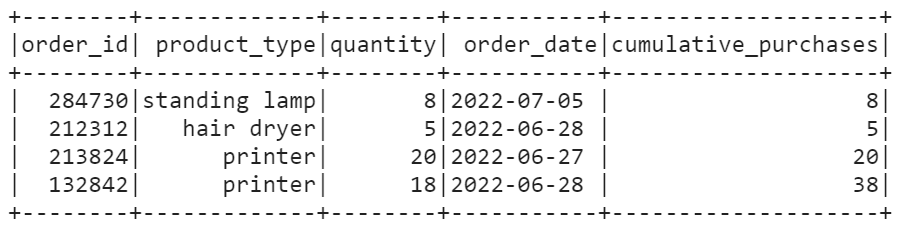
**# Define a Window specification based on the 'order\_date' column**

**window\_spec = Window.partitionBy("product\_type").orderBy("order\_date").rowsBetween(Window.unboundedPreceding, 0)**

**# Add a new column 'cumulative\_purchases' representing the cumulative sum**

**result\_df = order\_df.withColumn("cumulative\_purchases", F.sum("quantity").over(window\_spec))**

**result\_df.show()**

****

