

Processing Time Windows



Objective

Practice window functions on structured streaming

Handle records by processing time = time they arrive at Spark



Takeaways

Add processing time to existing records

```
val enhancedDF = purchasesDF  
    .select(col("time"), current_timestamp(). as("processingTime"))
```

Do your window processing

- Spark doesn't care what the time-based column means
- the only criterion is that it's a time or date column

```
val groupByDay = enhancedDF  
    .groupBy(window(col("processingTime"), "1 day", "1 hour").as("time"))  
    .agg(sum("quantity").as("quantity"))
```

Window duration and sliding interval must be a multiple of the batch interval

Output mode will influence results

Spark rocks

