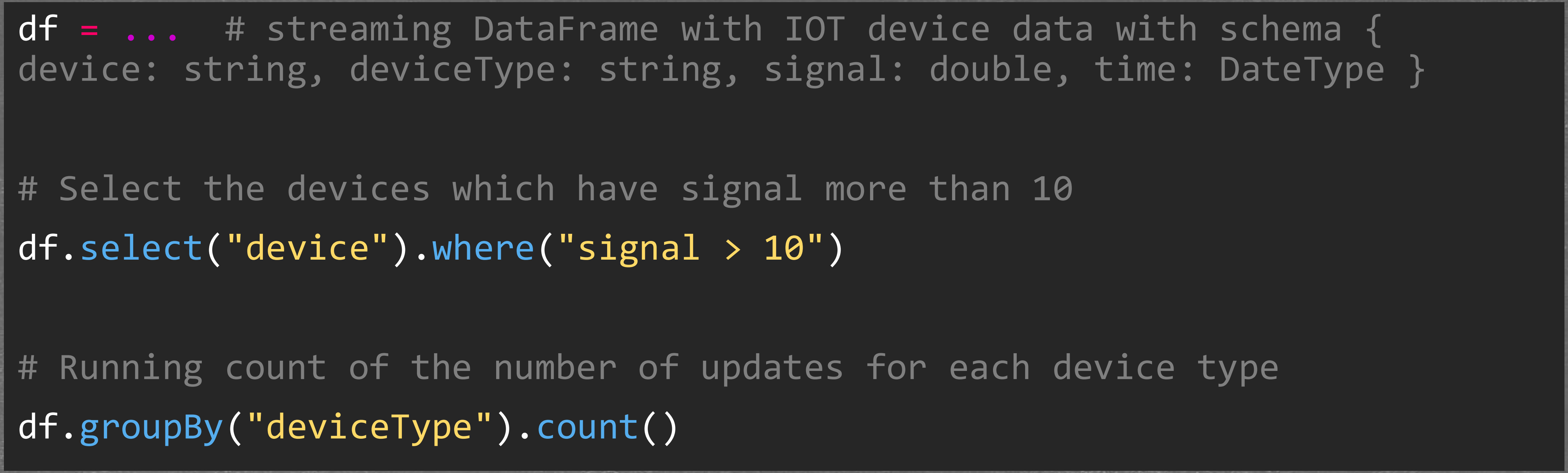
Operations on Streaming Dataframes/DataSets

Overview

- Spark streaming allows for tracking frequently-updated datasets
- can apply all kinds of operations:
  - untyped, SQL-like operations (e.g. `select`, `where`, `groupBy`),
  - typed RDD-like operations (e.g. `map`, `filter`, `flatMap`)



```
df = ...  # streaming DataFrame with IOT device data with schema {
device: string, deviceType: string, signal: double, time: DateType }


# Select the devices which have signal more than 10

df.select("device").where("signal > 10")


# Running count of the number of updates for each device type

df.groupBy("deviceType").count()
```

Advantages of Apache Spark Streaming

- Spark offers high-speed batch processing and micro-batch processing for streaming.
- Useful for mixed workloads compared to tools like Flink
- Can use many different data sources