Join Operations

Joining two streams

- Define two streams, and then using the join function on stream 1 and 2.
- Third stream where each batch interval, the RDD generated by the first stream is joined with the RDD generated by the second stream

```
stream1 = ...
stream2 = ...
joinedStream = stream1.join(stream2)
```

- can also use the `leftOuterJoin`, `rightOuterJoin`, and `fullOuterJoin` on the DStream

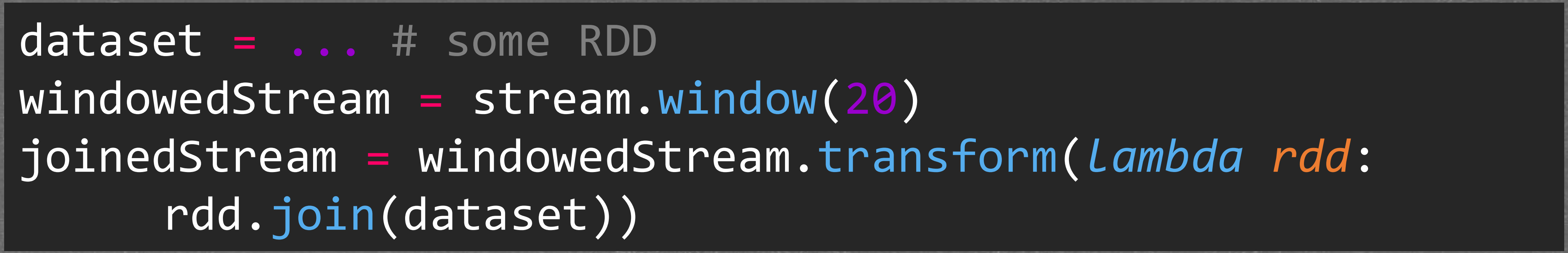
Joining two Streams (cont.)

- Often very useful to do joins over windows of the streams.

```
windowedStream1 = stream1.window(20)
windowedStream2 = stream2.window(60)
joinedStream = windowedStream1.join(windowedStream2)
```

Joining Streams with Dataframes

- Involves using the join operator in a function provided to the transform operator.
- The function will be evaluated every batch interval and will therefore use the current dataset that the dataset refers to.



```
dataset = ... # some RDD
windowedStream = stream.window(20)
joinedStream = windowedStream.transform(lambda rdd:
     rdd.join(dataset))
```