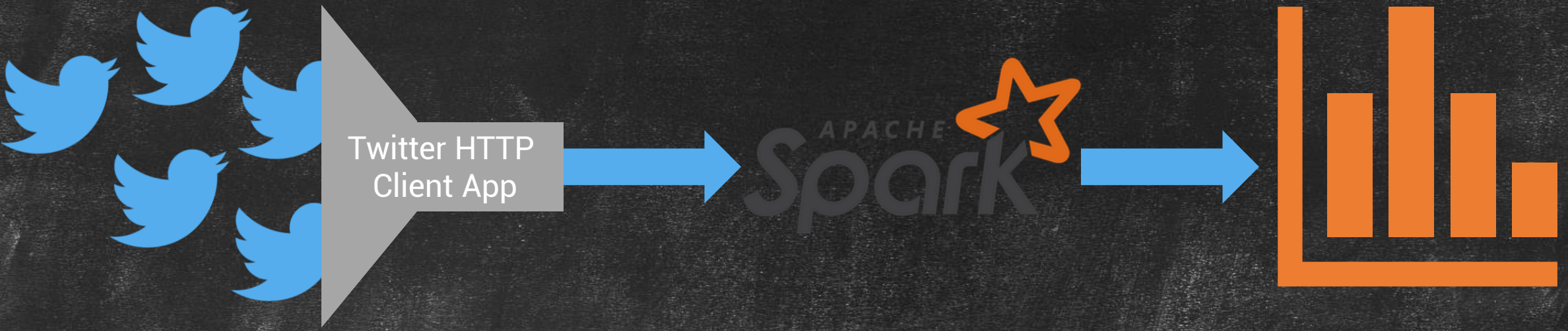
Tracking the Top 5 Most Popular Hashtags

Using Streaming with

Overview

- Spark streaming allows for tracking frequently-updated datasets
- Can use it to track most popular hashtags in 5 mins windows based on their counts in a Twitter stream, and by using the `StreamingContext` function.



To the Code!

What is `reduceByKeyAndWindow`?

- reduceByKeyAndWindow(*func*, *windowLength*, *slideInterval*, [*numTasks*])
- Aggregates datastream of (K,V) Pairs where values for each key are aggregated using the *func* reduce functions over a windowLength

What is `reduceByKeyAndWindow`?

- reduceByKeyAndWindow(*func*, *windowLength*, *slideInterval*, [*numTasks*])
- Aggregates datastream of (K,V) Pairs where values for each key are aggregated using the *func* reduce functions over a windowLength
- Improve by adding *inv* to input arguments
- reduceByKeyAndWindow(*func*, *inv*, *windowLength*, *slideInterval*, [*numTasks*])
- Reducing the new data that enters the sliding window, and “inverse reducing” the old data that leaves the window.

Thank you for Watching!

Like and Subscribe for More Content

Resources

- Spark Streaming Guide
- https://spark.apache.org/docs/latest/streaming-programming-guide.html
- https://youtu.be/AqcszswBRFQ
- https://www.toptal.com/apache/apache-spark-streaming-twitter
- https://github.com/databricks/spark-training/blob/master/website/realtime-processing-with-spark-streaming.md
- https://apps.twitter.com/app/14399887