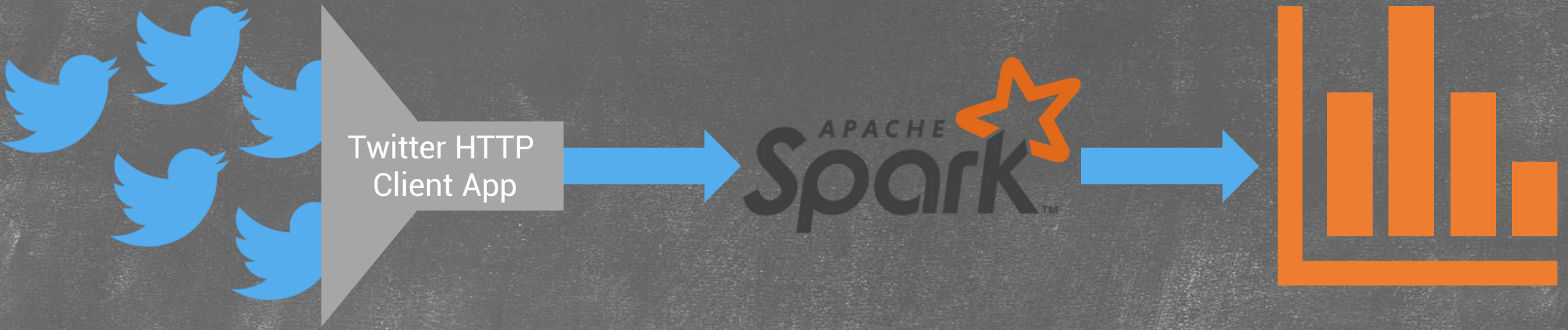
Overview of Apache Spark Streaming

What is Apache Spark Streaming?

- 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.



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

User Case of Apache Spark Streaming

- 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.