Introduction to Streaming

using

APACHE Spark™

What is Streaming?

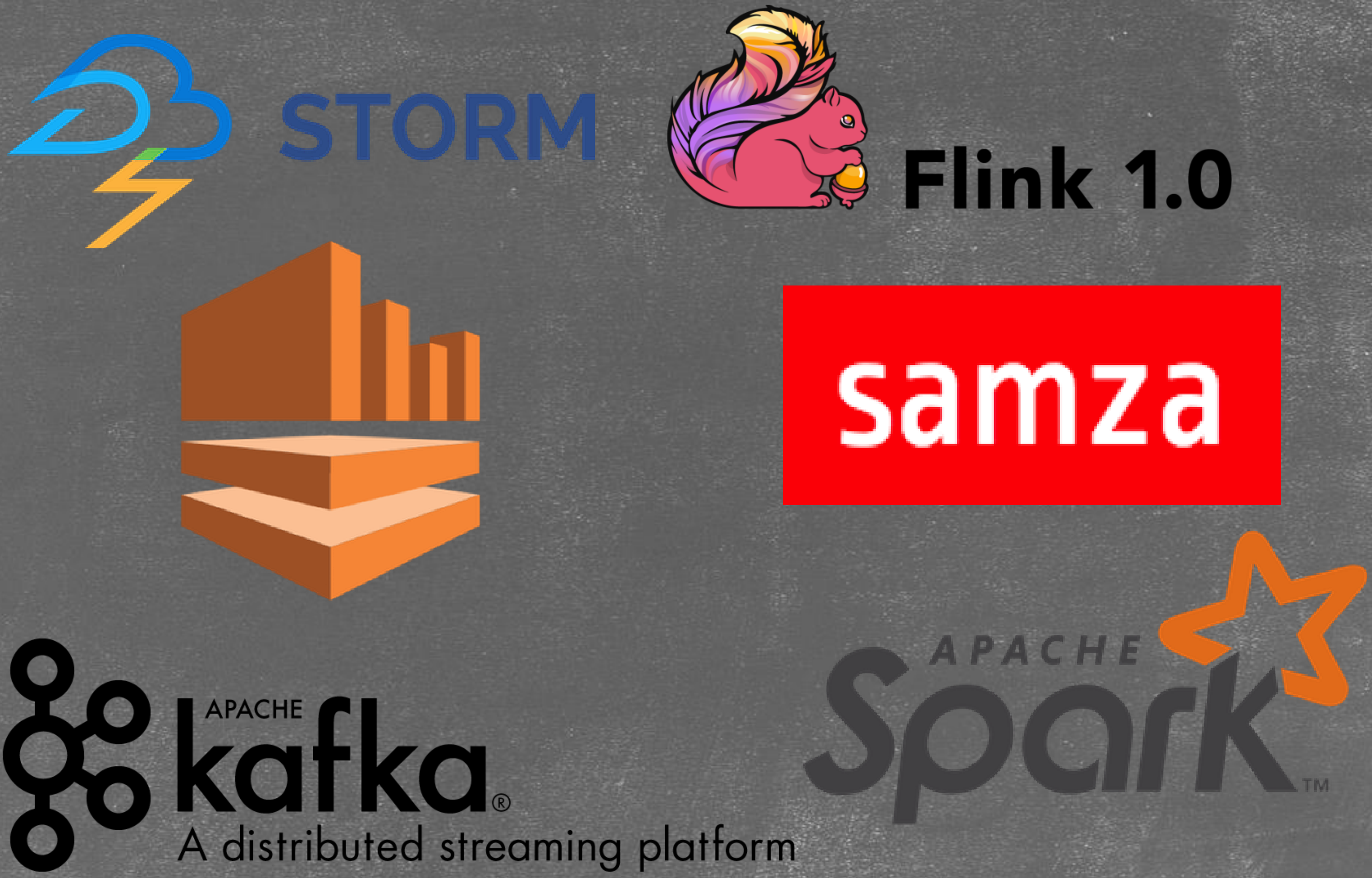
- Data is generated continuously from one or many sources
- Sources typically send in data simultaneously
- Data comes in small packages (kilobyte scale) in succession

Why use Streaming?

- A lot of applications use continuously-updated data
- Examples:
    - Sensors in vehicles, industrial equipment, and machinery send data to streaming for performance measurement.
    - A website tracking geo-location data from customer’s phones, which is gathered by streaming, so the website can make recommendations of which restaurants to visit
    - Solar power company monitoring panel performance through streaming
    - Online gaming company collecting streaming data about player-game interactions

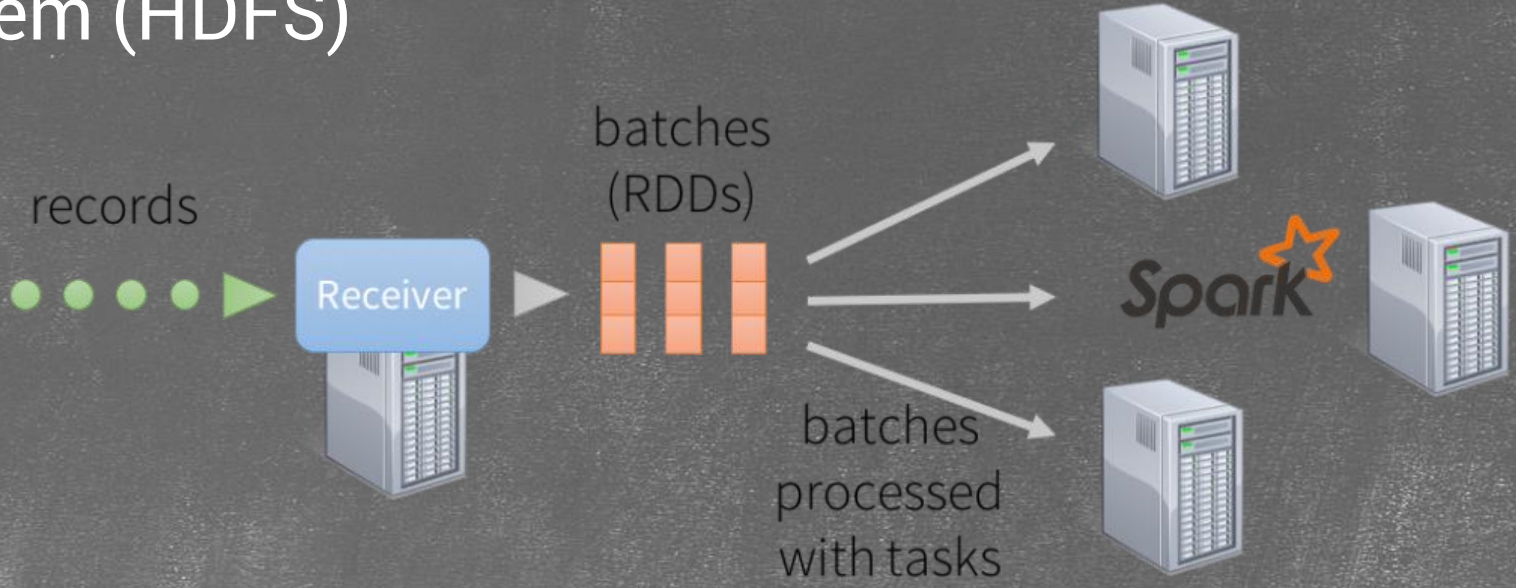
Popular Streaming Tools

- Storm
- Flink
- Kinesis
- Samza
- Kafka
- ***Apache Spar***

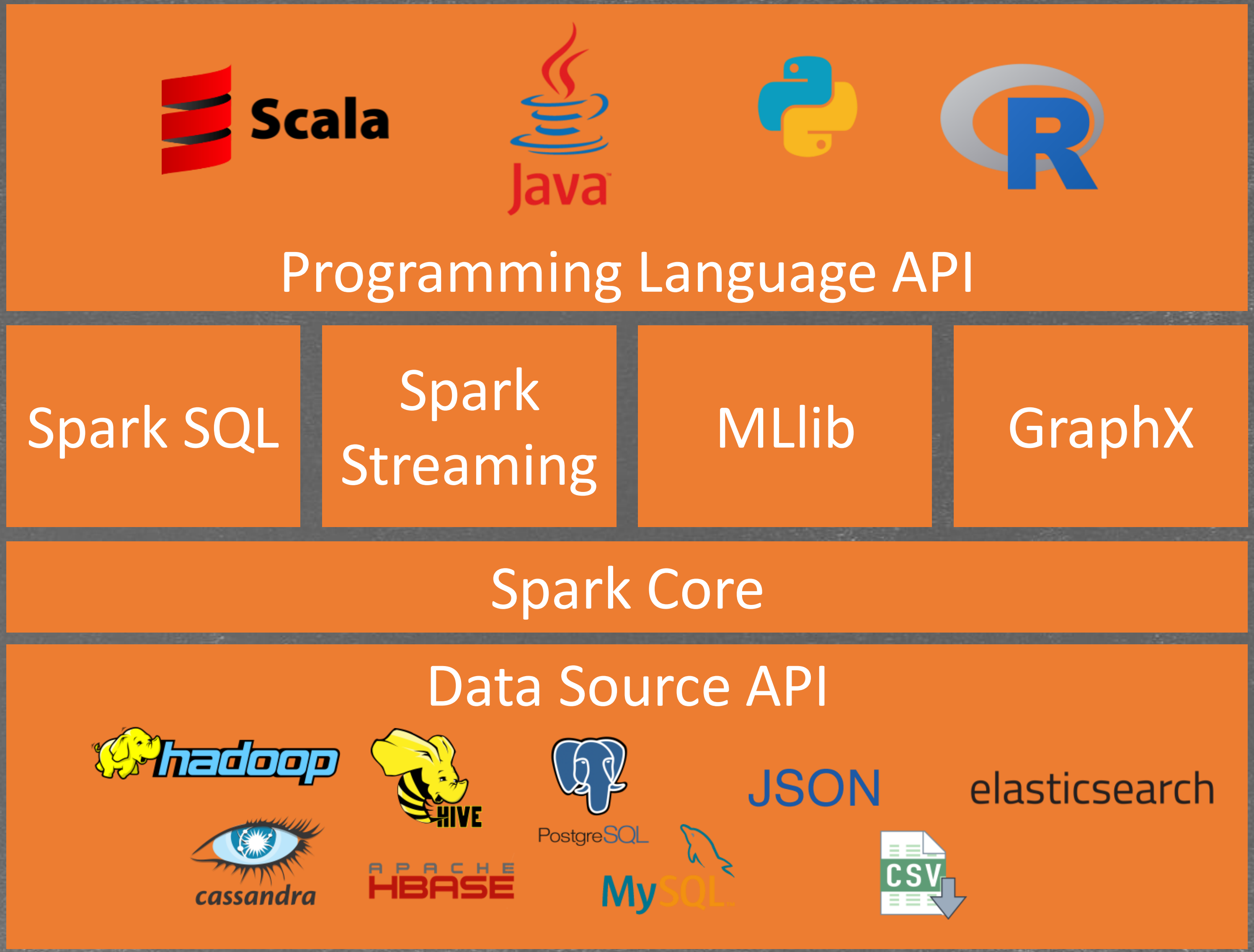


What is Apache Spark Streaming?

- Spark is general purpose and is widely used
- Spark connects with a lot of the previously mentioned streaming tools
- Fault tolerant thanks to projects like Hadoop Distributed File System (HDFS)



Apache Spark Structure



Programming Language API

Spark SQL

Spark Streaming

MLlib

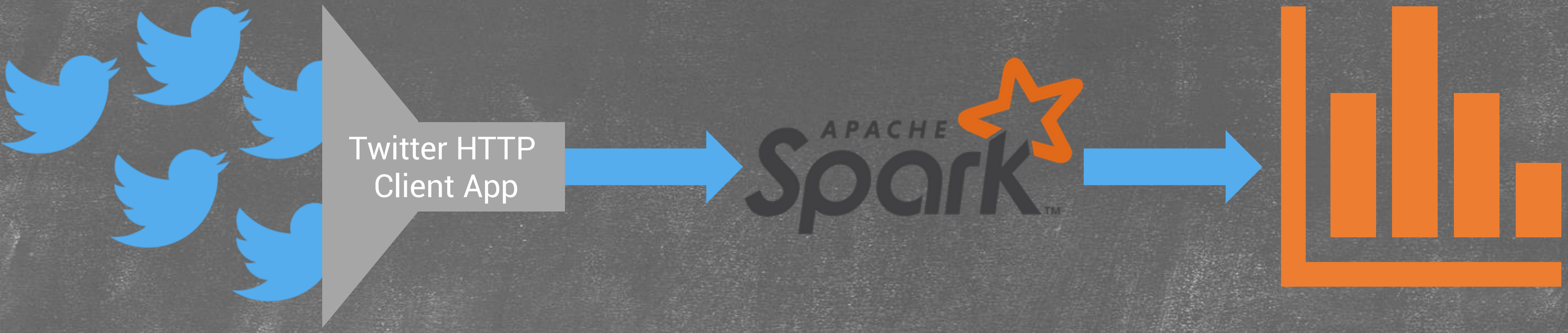
GraphX

Spark Core

Data Source API

In-depth Example Application (finally)

- 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.



Next Video: The Setup