Processing Streaming Data Using Spark

Kodjo Klouvi | kodjo.klouvi@gmail.com

Spark Streaming

- API to process continuously inbounding events
- Can process streaming data (continuously)
- Micro batch data (at-once given a timeframe)

What do we need ?

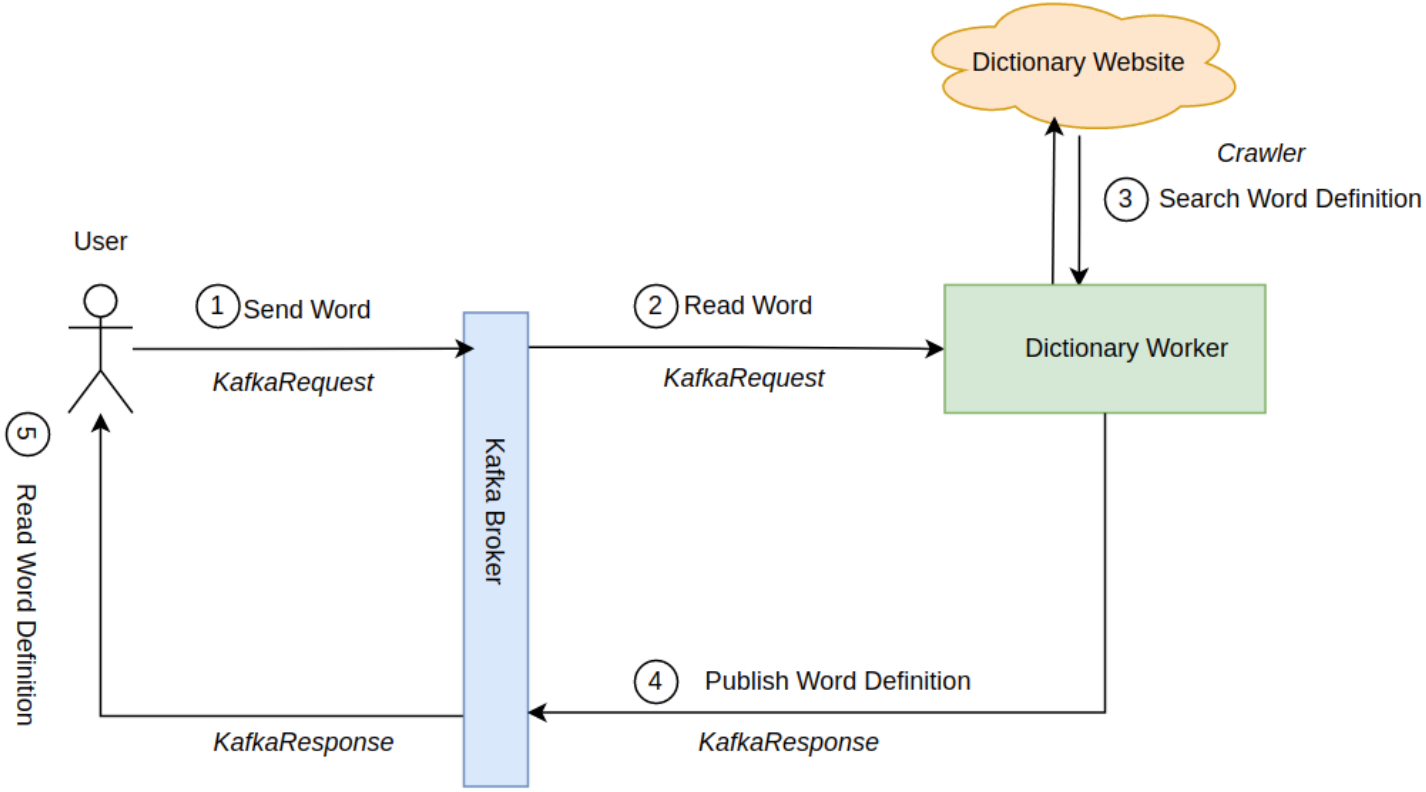
- Streaming sources
  - **Kafka**, zeromq, logs, elastic search, hdfs, etc.
- Spark application
  - Read data from the source
  - Transform the data using spark frameworks
  - Reroute the data as per the business requirements…

Hands-On

- https://github.com/osekoo/hands-on-spark-streaming
- Practice 1 : Pull the previous Kafka lab https://github.com/osekoo/hands-on-kafka
  - Add code to publish the word's definition into kafka
- Practice 2: Pull spark source code from https://github.com/osekoo/hands-on-spark-streaming
  - Implement the additional work

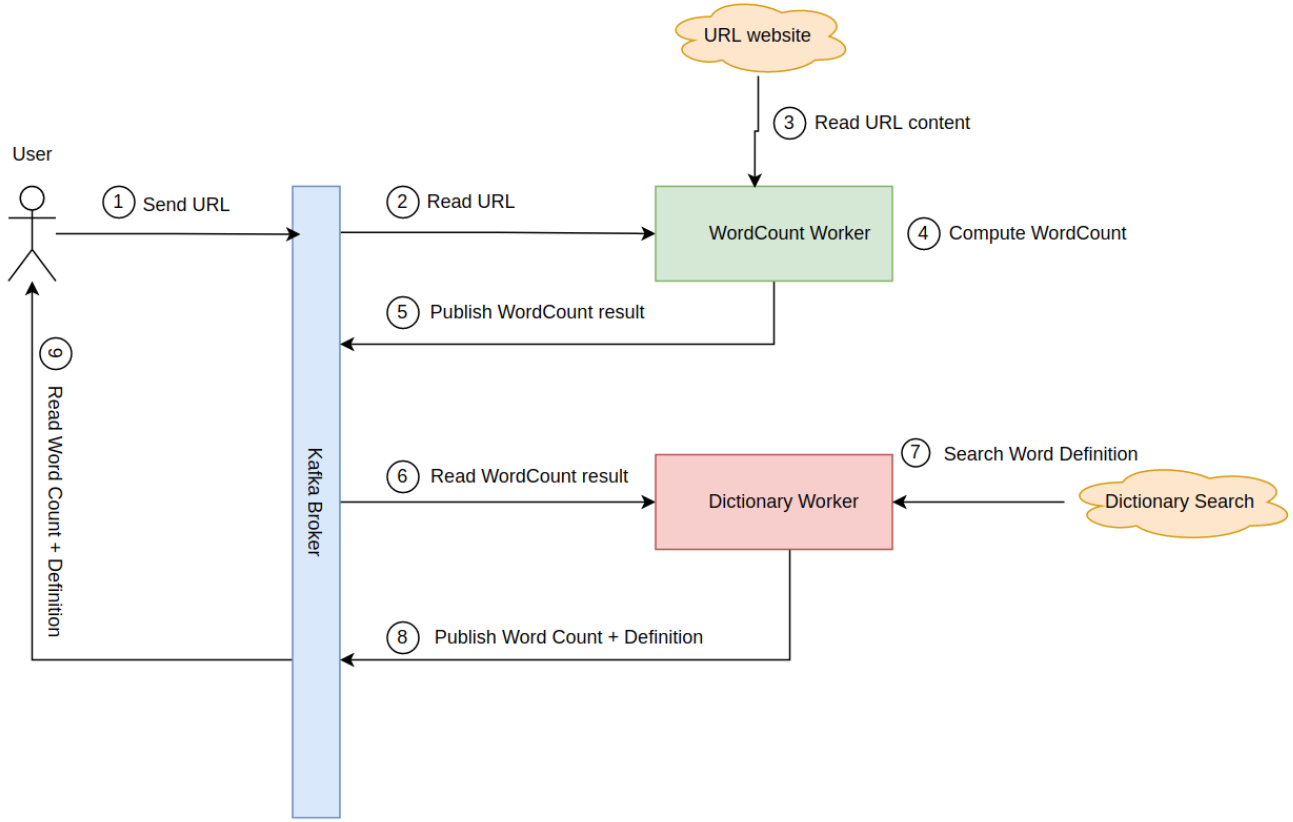
Practice 1: Online dictionary

- Build a Kafka based word definition search dictionary
- User specifies his nickname and the dictionary language (FR, EN)
- He/she interacts with the application by sending a word
- Our application search for the word definition by scrapping online dictionaries



Practice (Spark)

- Homework
- Searching word definition for website
- Extend language support (ES, CN, GE, etc.)
- Further details are available on Github



Troubleshooting

- Raise any question during the session or via email kodjo.klouvi@gmail.com