Yewno

Data Engineering Assignment

Thanks for your interest in Yewno. At Yewno, we don't believe in arbitrary, onerous "how would you code a linked list" type interviews. At work every day, you'll be dealing with a range of challenges, some software development, some software engineering, all fun. The objective of this exercise is to see how you deal with challenges in a realistic setting, rather than in an artificial one hour interview.

The process goes like this:
1. You: Thoroughly read the exercise below, if you have any questions, email matt@yewno.com.
2. You: Complete the exercise within 3 days of receiving this document.
3. We: Contact you to setup a time to chat about your submission.

For your solution, we are interested in seeing a couple of things:
1. How do you approach a problem?
2. How do you manage your work?
3. What path do you take?

Introduction

Data and data processing is the foundation of Yewno. With our goal to ingest the world's knowledge, we are working to consume both public and private data sources in both batch and streaming methods. Both data pipelines are built around sets of algorithms that are ran against the datasets to build the Yewno inference engine.

One of the key roles within Yewno will be ingesting large volumes of data in disparate formats and processing them at scale. This role works closely with the data science team to turn their algorithms into performant, production-ready systems.

Task

Using Spark's Streaming capabilities, create a Spark process which will read tweets from Twitter's Streaming API. These tweets should be filtered for a particular topic of your choice such as programming languages or sports teams. With this stream of tweets calculate a 5 minute window outputting the top 25 hashtag count. Capture at least three windows worth of data. The result should be a list of tuples similar to:

```
(5, lakers)
(4, cowboys)
(4, rangers)
...
```

bonus points (optional)

- run sentiment analysis over the tweets to determine if they are positive or negative
- determine the most popular tweet based on the number of retweets during the window
- integration tests
- impress us!

When you are finished, send us a link to the code repository -Github or BitBucket are great. Please be sure to **save the outputs of your test** run so we can take a look. Remember, we care for as much about **how you think** about the problem as the code itself! Document the code as needed and be ready to discuss your project.

Above all, have fun and reach out if you have any questions. The task is designed to take approximately 4 hours to complete with the assumption that you may not have used one or more of the technologies required.

Links

Spark Streaming - http://spark.apache.org/docs/latest/streaming-programming-guide.html
Twitter Streaming - https://dev.twitter.com/streaming/overview
Spark on Docker - https://hub.docker.com/r/sequenceiq/spark/
Twitter on Spark
-http://ampcamp.berkeley.edu/3/exercises/realtime-processing-with-spark-streaming.html