Implementation and Demo

CS 321-005 Team 3 Project Deliverable 4

Baker, Guo, Oh, Syed

Compiled on November 17, 2019 at 10:19pm

Contents

Changelog

> Document the changes if any to your requirements, architecture, and/or design. Briefly discuss when those changes were identified and how your team accommodated those changes.

Team Big Data's project has essentially remained unchanged over the course of the semester. Given the choice of tooling, our initial set of requirements, architecture, and design have largely not needed to change.

One point of change, however, was an implementation detail. Instead of receiving streaming data from Twitter we fetched data from Twitter in bulk with a Python script to skirt API limitations.

Technology Used

> Briefly discuss the technologies that you have used (language, platform, patterns, etc.).

Team Big Data (TBD) uses three main Commercial Off The Shelf (COTS) software components to facilitate its data flow (or pipeline) architecture:

1. NiFi
2. Elasticsearch
3. Kibana

TBD also uses Tweepy to retrieve data from Twitter, and VADER Sentiment to perform sentiment analysis.

NiFi

Apache NiFi is a dataflow system based on flow-based programming which is used to automate the flow of data between systems. NiFi processors can perform a combination of data routing, transformation or mediation between systems.

NiFi was selected as it is open source software, has a great deal of documentation, and provides an API that can be used to write new processors to add functionality.

Elasticsearch

Elasticsearch is a distributed, scalable, and high-performance search engine. Elasticsearch offers simple REST based APIs, a simple HTTP interface, and uses schema-free JSON documents. It also provides support for various languages including JAVA, PHP, Node.js etc.

Elasticsearch was selected as a data store as it is open source software, is high-performing and its tight integration with the visualization and analytics tool, Kibana.

Kibana

Kibana an analytics and visualization tool designed to work with Elasticsearch. Kibana can perform advanced data analysis and visualize the data in a variety of charts, tables, and maps.

Kibana was selected as pairing Kibana and Elasticsearch yields a tightly integrated solution with best-in-class performance, maintainability, and ease of use.

Tweepy

Tweepy is an open source Python library built to ease fetching and processing data from Twitter. Tweepy wraps Twitter's public API so that data from Twitter can be retrieved from within Python scripts.

VADER Sentiment

VADER Sentiment is a tool which performs sentiment analysis of text. It is tuned so that it performs best on text sourced from social media. TBD uses VADER Sentiment to perform sentiment analysis on the text of tweets of people that TBD follows.

Contributions

> Describe how the work was distributed for development purposes and mention who implemented each piece. If multiple people have worked on one piece, document the % of contribution of each member with respect to that piece. Share your version control project repository with the GTA. If you are enrolled in section 001 or 004, please share it with Bhargavi. If you are enrolled in section 002 or 005, please share it with Roberto. Their email addresses are listed on Piazza.

The progression of TBD's project can be split in two: the first portion of the project involved setting up the data pipeline and ingesting data into our data store; the second portion of the project involved creating analytics and visualizations on that ingested data.

With respect to the first portion of the project, Connor handled almost all of it. He is the primary account holder of the remote servers that TBD is using, and as such, is responsible for the content on and status of the account. Additionally, given his prior experience with NiFi, Elasticsearch, and Kibana he was the de-facto person to construct the pipeline. He also handled all of the typesetting of the deliverables to ensure consistent branding.

The second portion of the project was split evenly between Ghousia, Shin, and Ziyan. Together, they created all the analytics and visualizations on the data. Connor provided minor editorial notes on the verbiage used to describe visualizations, their placement, and their presentation, but did not create any of the visualizations – he served in a position like an advisor, answering questions if they arose.

Software Reuse

> If you have reused code (methods, classes, components, etc.), provide details of the elements that have been reused and include attributions to the original author(s) [if you followed tutorials, stack overflow posts, etc., include links to those as well]. Also discuss if reuse was beneficial (why/why not).

Team Big Data's project is made possible by use of COTS software and public APIs. Without the use of existing software, a project of this size would simply not have been possible. TBD's work builds upon decades of research and ingenuity – attempting to rediscover and implement such a vast amount of knowledge over the span of nine weeks would have been irresponsible.

NiFi

https://nifi.apache.org/ and https://github.com/apache/nifi

NiFi is an open source tool which is maintained by the Apache Software Foundation.

Elasticsearch and Kibana

https://www.elastic.co/products/elasticsearch and https://github.com/elastic/elasticsearch

https://www.elastic.co/products/kibana and https://github.com/elastic/kibana

Elasticsearch and Kibana are both open source tools which are maintained by Elastic NV.

Twitter

https://developer.twitter.com/en.html and https://developer.twitter.com/en/docs/api-reference-index

Twitter provides developer accounts free of charge. These developer accounts have access to well-documented APIs.

Tweepy

http://www.tweepy.org/ and https://github.com/tweepy/tweepy

Tweepy is an open source Python library built to ease fetching and processing data from Twitter. Tweepy wraps Twitter's public API so that data from Twitter can be retrieved from within Python scripts. TBD uses Tweepy to fetch tweets from only users we follow.

VADER Sentiment

https://github.com/cjhutto/vaderSentiment

VADER Sentiment is an open source Python library with eight contributors. We use it to perform sentiment analysis on the tweets of the people we follow.