Vision

There is a lot of data being generated today in by people who use social media, with their posts, tweets, and tags reflecting a good chunk of what is happening in society. This data can document events in real-time, demonstrate current trends, and reflect people's perceptions on certain things. With this vast repository of information, the Real Time Drug Safety Monitoring Project seeks to do the following: use information publicly available from social media to track and monitor the effects and symptoms taking certain drugs can cause.

What needs to be done for our project is straightforward. We intend to create an application that will gather tweets from Twitter regarding drug use and symptoms, analyze the tweets to get information on how drugs are affecting the users, and archive this information in a database for future reference and research. The overall goal of the application is that people would be able to tell from social media what effects drugs are having on people and be able to warn the public if there is an issue with certain drugs.

Scope

The data we will use will be strictly "tweets" from Twitter. While there are other social media platforms that exists (like Facebook), our focus is limited to using the Twitter API.

The deliverables will be the following:

1. An Apache Spark Streaming program that will be committed to a data ingestion task, gathering tweets from Twitter related to drug and drug usage
2. An Apache CTAKES/Analytics program that will analyze the tweets for useful information in a tweet, namely the drug being used and its effects on a user.
3. A NoSQL database that will store the tweets and the data of tweets analysis

Things to Explore

We will explore the frameworks, programming languages, and technologies that the client has proposed to us. They are the following: Scala, Apache BigTop, Apache Spark, cTAKES, SBT, OpenStack, GIT revisioning, and NoSQL.

Apache BigTop

BigTop allows for testing of leading open source big data components, specifically with Apache Spark. It packages RPMs and DEBs, provides smoke testing with over 50 test files, and provides virtualization. We will be working closely with BigTop for our project to make sure integration with cTakes and Spark goes well.

Apache cTAKES (Clinical Text Analysis and Knowledge Extraction System)

cTAKES is a processing system for information extraction from electronic medical record clinical free-text. This is our main engine for processing medical information from social media. We currently have one member of our team solely responsible for cTAKES ,but we will be allocating more people to be working on cTAKES if necessary.

Apache Spark

Spark is the engine for large scale data processing. It will be used as the main tool for collecting tweets from Twitter.

Scala (Scalable Language)

Scala is a programming language written under the Java virtual machine platform that has support for functional programming. It was created to make up for the shortcomings of Java and is extremely scalable. Everybody in the group will be responsible in learning Scala and contributing code to our project.

Openstack

Openstack is an open source software for creating private and public clouds. It controls large pools of compute, storage, and networking resources throughout a datacenter. We are not sure how it will fit in our overall design, but we will need to speak to Pei Chen to finalize Openstack implementation.

GIT Revisioning

GIT provides revision control and branching to allow for concurrent work to be done. We will focus on getting everybody in our group up to speed in using github so that we can make our commits to the repository to reflect the progression of our project.

NoSQL(Not Only SQL)

NoSQL provides a mechanism for storage and retrieval of data that is modeled in means other than tabular relations used in relational databases. It allows for simplicity and horizontal scaling (analogy being *having more nodes to the database* as opposed to *having more resources per node*). Within NoSQL, we have several databases that meet the requirement for our project. We require a column model for our drug collection database, so Accumulo, Cassandra, Druid, HBase, and Vertica stands out as our main candidates. In the end, we will be using Cassandra because of existing code and framework in the tutorials that our mentor Jay has created for us.

Our Roles

King Mui

I will be responsible for the learning NoSQL. This will include the research for how to use a NoSQL database, choosing a NOSQL database I would be comfortable working with, and setting up the server to ensure that it works with storing the data ingested from Twitter and the analyzed data that is gotten from Apache cTAKES.

Raymond Li

My role will be to learn Scala as in our project it will be the core language most of the open software programs will use. I will need to do research and learn how to use Scala with the IntelliJ IDEA. From there I will learn how to use Scala for both the Apache Sparks and Apache BigTop applications we will develop. At the moment, there is a lot of learning to do for myself.

Re-Hao Ni

Since Scala will be a huge component to this project, it was decided to ensure that there would need to be two people who would be proficient in Scala to divide up the future tasks involved with the data ingestion application with APache Sparks and Apache BigTop. CUrrently am also in the process of learning Scala on IntelliJ IDEA and setting up my laptop to has a Unix environment.

Xinzhu Xu

I will learn how to use cTAKES to extract data from twitter, also learn Scala to deal with the question. I think this task is closely related with database, so I still need to know some basic knowledge about machine learning and database.