Real-time Twitter Sentiment Analytics System

Team 5

Ruifeng Cui 001029411

Ashish Roy 001044804

Anxi Liu 001029165

Summary

We aim to design and build a real-time sentiment analytics system which can achieve the sentiment analysis for the streaming tweets with input specific keywords by users. This system consists of data ingestion module, data processing and storge module, data visualization module, and this system will be deployed on the cloud.

Use Case

Real users input a keyword which they want to know. This system will ingest the real-time tweets which is relative to this keyword on the Twitter. Then, system will process batches of data in the real–time and perform the sentiment analysis by using machine learning algorithms. Finally, this system will display the sentiment analysis result according to this keyword for the user.

If users want to know what people think and attitudes towards a certain university or a certain team on Twitter in real-time, they can just type the name of that university or that team. Then, in seconds, this system will do the processing and computing. Finally, system will return the analysis results in the charts for users. Users can intuitively see the attitude of Twitter users about this team or school.

Methodology

data ingestion : Twitter API + Kafka

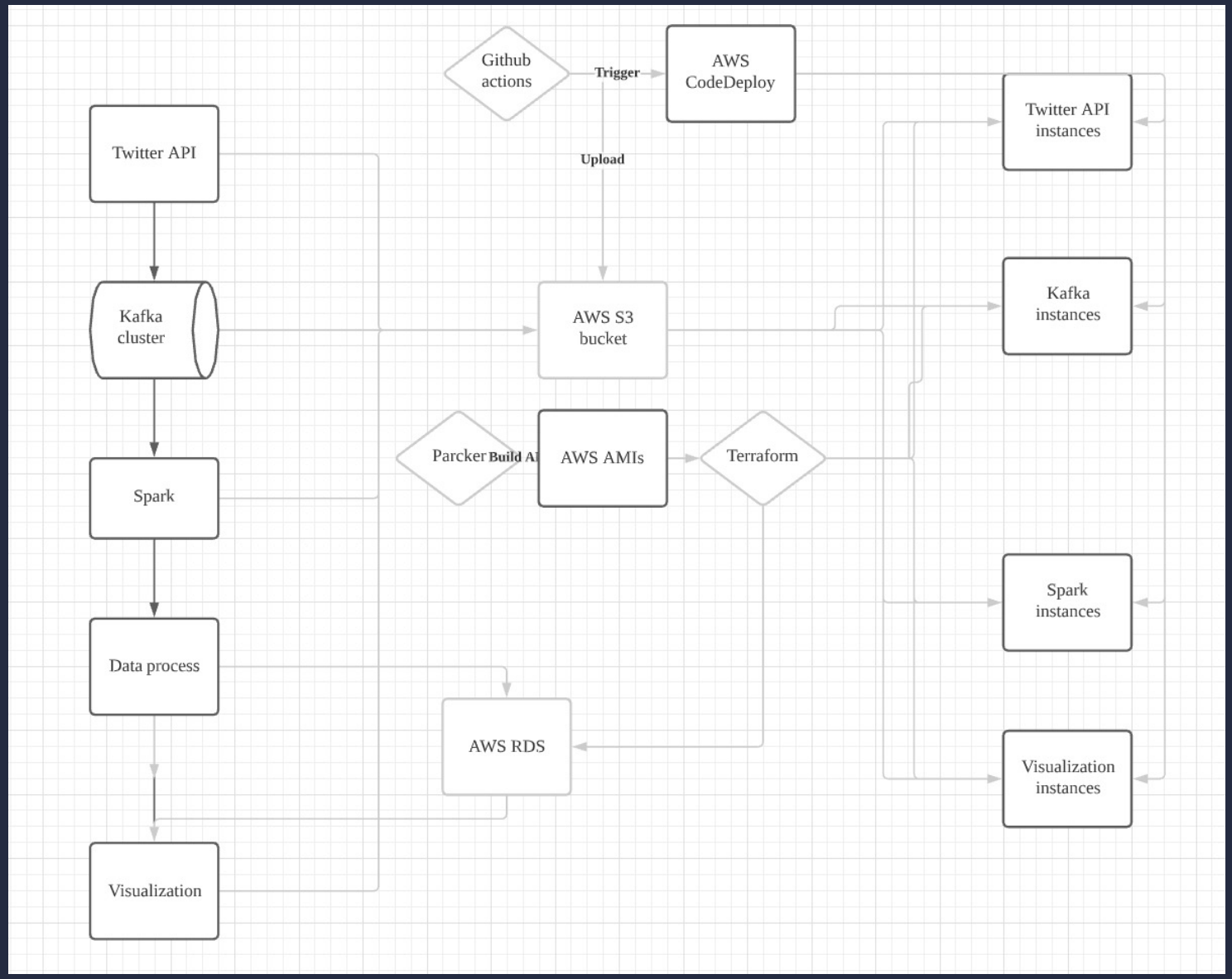
data processing: Spark Streaming + Spark MLlib

data storage: HDFS + MySQL (Maybe a little bit change when we implement)

data visualization: web or visualization tools

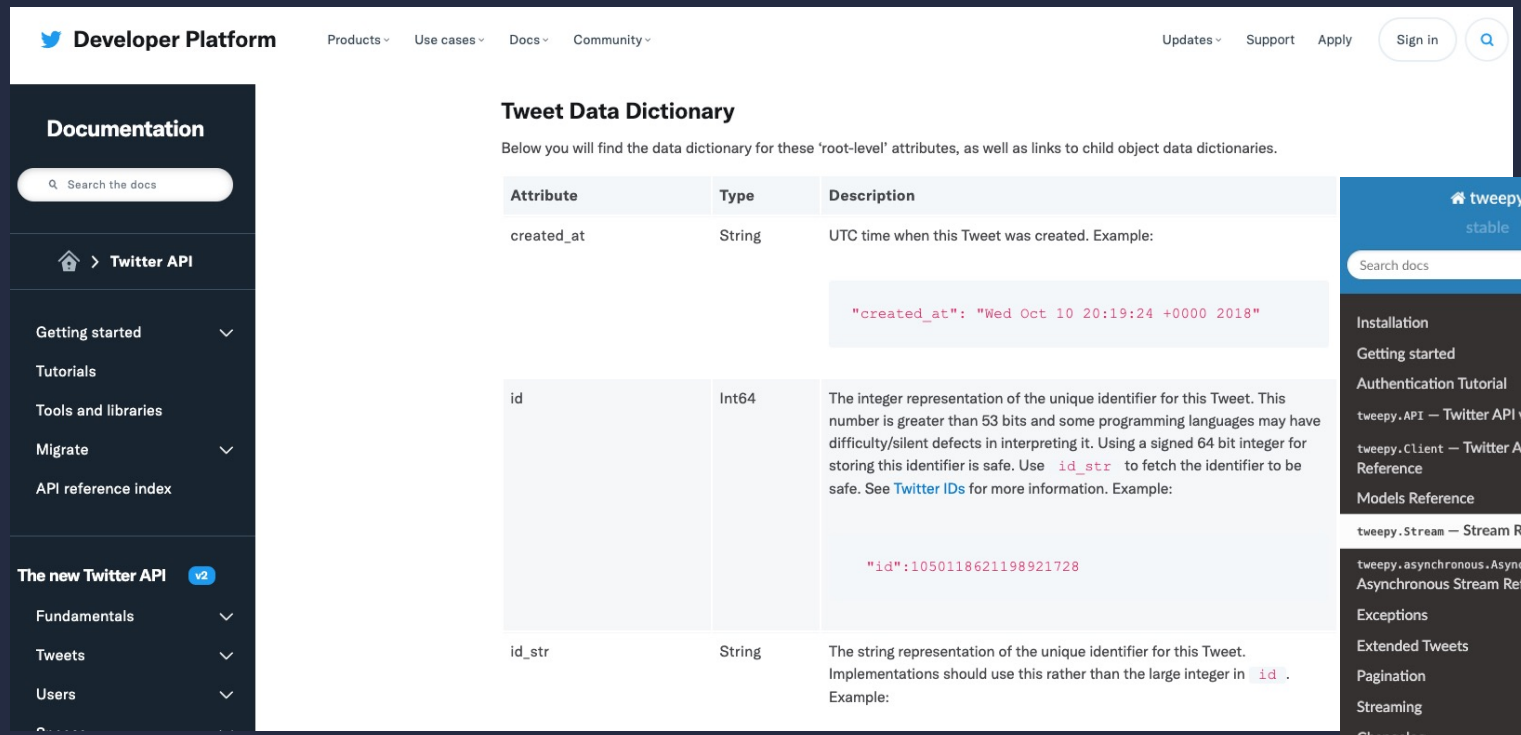
deploying and resource management：AWS

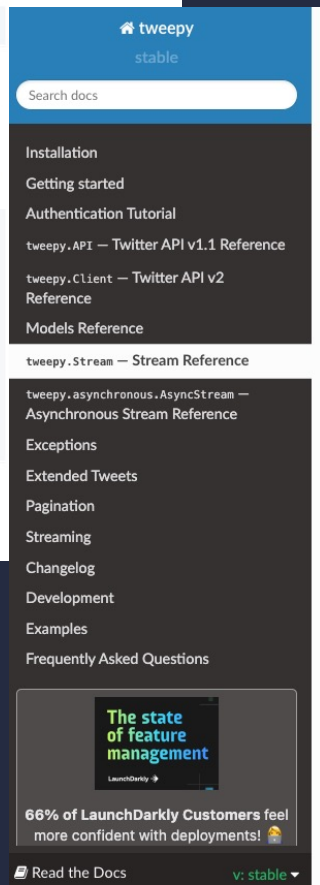
Architecture and Workflow

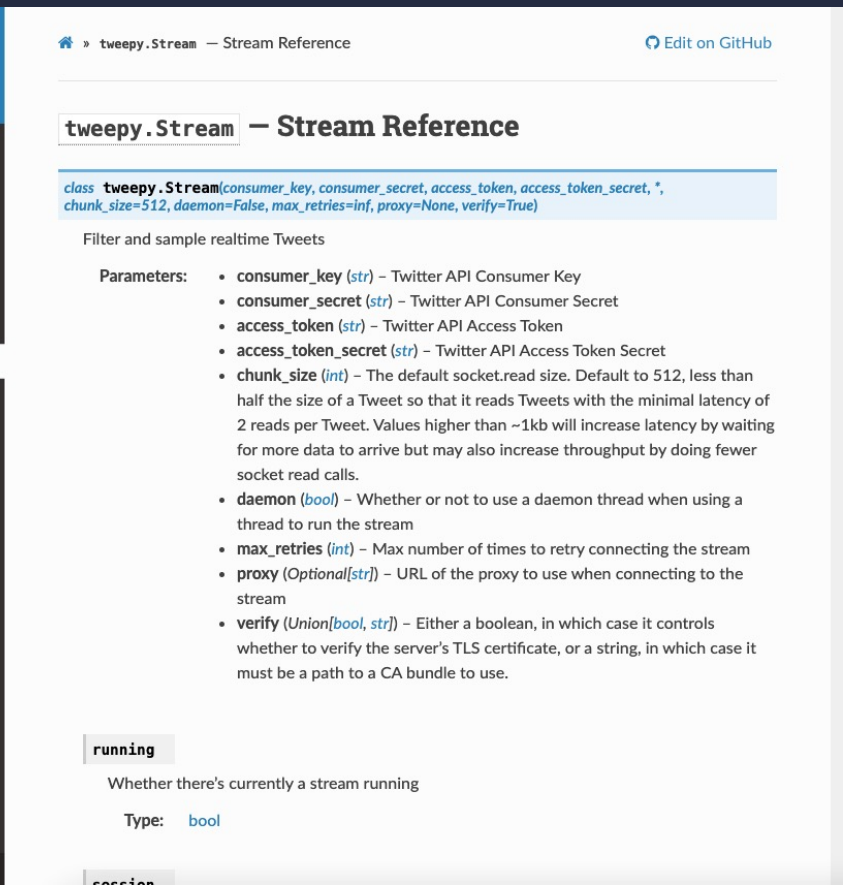


Data sources

Streaming data from Twitter API







Sprint

		Anxi	Ruifeng	Ashish
Sprint1 (Nov 1 - Nov 7)	Learn what we need to do	Learning		
Sprint2 (Nov 8 - Nov 14)	Implement	Twitter API / Kafka	Could deploy / CICD	Data process / ML
Sprint3 (Nov 15 - Nov 21)				
Sprint4 (Nov22 - Nov 28)	Integration	Integration		
Sprint5 (Nov29 - Nov 5)	Final review / bugs fix / presentation	Review and test		

Programming in Scala and Code Repository

Programming in Scala：As we can see, the essential big data frameworks which we are going to use are Kafka and Spark, and they are written by Scala primarily. Therefore, we are going to develop our system in Scala especially in the data ingestion and data processing module.

Other language: We may also use Python, shell script, query language and other language if we need to use for the data visualization.

Code repository： https://github.com/CSYE7200-21FALL-TEAM6

Acceptance criteria

1. This system can ingest the relative streaming tweets from Twitter according to the input keywords by user.
2. This system can process about 150 tweets for every query and perform the sentiment analysis for these real-time tweets in the seconds.
3. This system will be expected to achieve 70% sentiment analysis accuracy.
4. This system will show the sentiment analysis results in the visualization chart as the feedback to the user.
5. This system will be deployed on AWS.

Goals

This project will help us to know how to design and build the big data system. We will be better acquainted with big data frameworks, tools and cloud computing. Most importantly, we will achieve good, practical, working knowledge of Scala

Thank You !