* + 1. Twitter Firehose VS Twitter API

Users send 400 million tweets every day. Ranked as the 8th most popular site in the world by Alexa rank in January 2015, Twitter boasts 500 million registered users.

The only way to access 100% of those tweets in real-time is through the Twitter “Firehose”. The other option for accessing tweets is using one of Twitter’s direct API offerings.

What is an API?

An API, or Application Programming Interface, is the instruction set created for developers to interact with some type of technology. In this case, Twitter has data and lots of it! Twitter created an open API allowing external developers to develop technology which rely on Twitter’s data.

What is the advantage of offering an open API?

The major advantage of offering an open API is to promote external innovation, further strengthening the base technology, service or data. Offering data externally allows developers to create products, platforms, and interfaces without the need to expose the raw data. It’s common for technology companies to acquire other innovative technologies rather than building innovations internally. Twitter has capitalized on this model as evidenced by their recent acquisitions of 10 different technology companies in 2012 built around their open API.

Twitter Search API vs. Twitter Streaming API vs. Twitter Firehose

There are three different ways to access Twitter data .

* Twitter’s Search API
* Twitter’s Streaming API
* Twitter’s Firehose

Twitter’s Search API

It involves polling Twitter’s data through a search or username. Twitter’s Search API gives you access to a data set that already exists from tweets that have occurred. Through the Search API users request tweets that match some sort of “search” criteria. The criteria can be keywords, usernames, locations, named places, etc. A good way to think of the Twitter Search API is by thinking how an individual user would do a search directly at Twitter (navigating tosearch.twitter.comand entering in keywords).

How much data can you get with the Twitter Search API?

With the Twitter Search API, developers query (or poll) tweets that have occurred and are limited by Twitter’s rate limits. For an individual user, the maximum number of tweets you can receive is the last 3,200 tweets, regardless of the query criteria. With a specific keyword, you can typically only poll the last 5,000 tweets per keyword. You are further limited by the number of requests you can make in a certain time period. The Twitter request limits have changed over the years but are currently limited to 180 requests in a 15 minute period.

Twitter’s Streaming API

Unlike Twitter’s Search API where you are polling data from tweets that have already happened, Twitter’s Streaming API is a push of data as tweets happen in near real-time. With Twitter’s Streaming API, users register a set of criteria (keywords, usernames, locations, named places, etc.) and as tweets match the criteria, they are pushed directly to the user. Think of this as an agreement between the end user and Twitter – you agree with Twitter that whenever they receive tweets that match keywords relating to “hockey”, they will deliver the tweet directly to you as they happen. This is a push of data by Twitter, rather than a pull of data initiated by the end user.

The major drawback of the Streaming API is that Twitter’s Steaming API provides only a sample of tweets that are occurring. The actual percentage of total tweets users receive with Twitter’s Streaming API varies heavily based on the criteria users request and the current traffic. Studies have estimated that using Twitter’s Streaming API users can expect to receive anywhere from 1% of the tweets to over 40% of tweets in near real-time.The reason that you do not receive all of the tweets from the Twitter Streaming API is simply because Twitter doesn’t have the current infrastructure to support it, and they don’t want to; hence, the Twitter Firehose.

Twitter Firehose

The final way to access data is by having access to the full Twitter Firehose. The Twitter Firehose is in fact very similar to the Twitter’s Streaming API as it pushes data to end users in near real-time, but the Twitter Firehose guarantees delivery of 100% of the tweets that match your criteria.

The Twitter Firehose is handled by two data providers, GNIP and DataSift, which have tight relationships with Twitter. Similar to the streaming API, the firehose consists of an agreement between an end user and distributors of the Firehose (GNIP or Datasift) on what tweets the end user should receive in near real-time. As the data providers receive tweets they are pushed directly to the end user.

The two differences between Twitter’s Streaming API and Twitter’s Firehose access is that you are guaranteed delivery of 100% of the tweets and it’s not free. The Twitter Streaming API is free to use but gives you limited results (and limited licensing usage of the data). Access to the Twitter Firehose removes a lot of the usage restrictions imposed by Twitter but is fairly costly for access to all the tweets.

Why the Difference Matters

The Twitter Search API and Twitter Streaming API work well for a lot of individuals that just want to access Twitter data for light analytics or statistical analysis. Marketing companies and social media analytic companies use Twitter’s Search API to analyze trends in social media. However,these differences are significant when you are in a situation that requires you to monitor Twitter in real-time during a specific event or critical situation.

For example, professional sports teams provide security during games for spectators. It is critical that they be able to see what is happening in real-time at the venue.

Real-time, full access is also imperative for law enforcement. Whether it’s a specific situation that is evolving minute by minute or a high-profile event that is happening in their jurisdiction, the police need to know what is happening, when it is happening, and where it is happening to keep citizens safe. They can’t rely on just a sample of the information and have it delivered after the fact.Twitter Firehose Solution

BlueJay is a Twitter monitor for law enforcement, intelligence, and security professionals that provides users with full access to the Twitter Firehose. Users can monitor all publicly-available tweets against specific keywords, locations, and users in real-time.

The technology was used by law enforcement in Long Island during the Hofstra University presidential debates and is currently being used by police departments and security companies around the country.