



One of the best features of Cassandra is its simple schema, Cassandra does only a few things, but it does them extremely well.

The screenshot shows the SocialFlow website interface. At the top, there's a navigation bar with links like 'How SocialFlow Works', 'About Us', 'Blog', 'Sign Up', 'Follow', and 'Like'. The main content area features a headline: "Stop guessing at how to get maximum engagement on Twitter". Below this is a diagram illustrating the concept: "Relevant Message + Right Audience + Right Time = Maximum Engagement". A sub-headline reads: "SocialFlow sends your messages when they are most likely to connect with your audience." Below the diagram, there's a "Sign up now" button and a "Learn more" link. The page also includes sections for "What's New" and "Featured Video". The "What's New" section highlights a case study: "[Case Study] Congrats to Human Rights Watch (HRWH) on Surpassing 250,000 Followers on Twitter!". The "Featured Video" section shows a video player with a thumbnail of a man speaking.

Company

SocialFlow

Data Size

12-node cluster, 4-node cluster in development, and some single-cluster development boxes

Challenge

The need for a fault-tolerant, distributed database with incredible write ability to handle the real-time data stream from Twitter's "firehose," which channels about 250 million tweets per day.

Solution

The elastically scalable and reliable Apache Cassandra platform, which allows SocialFlow to support an extremely high volume of writes, publish the right tweet at the right time for its clients, and gives its IT team more time to work on new services.

SocialFlow is one of the few companies to enjoy full, unlimited access to Twitter's "firehose"—a real-time stream of every tweet. The fast-growing startup is the developer of the first and only social media optimization technology that uses real-time data from the streaming updates produced from the Twitter firehose, Bitly, Facebook, and other sources to help publishers, retailers and brands maximize potential engagement of Twitter followers. Through its SocialFlow AttentionScore™ algorithm, it can identify the precise time to publish a particular message so it's likely to be seen by the greatest number of interested followers.

Given that the Twitter firehose channels about 250 million tweets a day—and that most of SocialFlow's clients have millions of followers and receive thousands of clicks on just about everything they post—it's not surprising that SocialFlow faces some big data challenges. Drew Robb, director of data systems at SocialFlow, says there is simply no way for all the data SocialFlow takes in from the Twitter firehose to fit on one machine. "The stream equates to about 500 GB per day, uncompressed, and about 3,000 static updates per second," he says. "We need a big distributed system just to do anything with that data."

Not long after joining SocialFlow in early 2011, Robb began the search for a NoSQL database with "incredible write ability." He says, "We knew our write load was going to be huge. So when evaluating solutions, I looked hard at the write path. If a solution couldn't support an extremely high volume of writes, it wouldn't work. But it seemed Apache Cassandra would do what we needed—and do it right."

"The Bieber Stress Test"

Since implementing the elastically scalable Apache Cassandra database, Robb says he has been "pleasantly surprised" by its performance. He especially likes the solution's low latency—and the fact that it doesn't process non-sequential I/Os. Robb says, "When I first heard about big distributed systems, I thought, 'Well, that might work, but I'm sure it will be a game of wait-and-see.' But with Cassandra, everything does work. The write path is always fast and always writeable." One of the first things Robb did with Apache Cassandra is put it through what he calls "The Bieber Stress Test." He explains, "I fed the firehose into it and made a table of mentions. Then, I created a streaming thing that would give me all Justin Bieber mentions, and right away it was populating. I was seeing the data come out immediately, and we experienced no problems with the cluster." And why did Robb choose Justin Bieber for this critical performance test? "The kid gets about a 100 mentions per second on Twitter," he says. "I know if we don't see that people are mentioning him, something is really wrong!" Robb also has been impressed with Cassandra's linear scalability—and how quick and easy it is to add nodes. "The first time I did it, I doubled the node size," he says. "It only took about an hour, and it all just worked. I instantly saw the system load decrease by almost a factor of two."

Management made easier

Robb says that since implementing Cassandra, his database management responsibilities are much less burdensome. "Without a solution like Cassandra,



“We knew our write load was going to be huge. So when evaluating solutions, I looked hard at the write path. If a solution couldn’t support an extremely high volume of writes, it wouldn’t work. But it seemed Apache Cassandra would do what we needed—and do it right.”

—DREW ROBB, SocialFlow

we just wouldn’t be able to handle the insert volume,” he says. “We are using 12 machines now just to process the firehose integration full wave. It would be a nightmare to configure that. But with Cassandra, I can kill the Java process on any machine at any time and it won’t affect the quality of the data. There is just beautiful redundancy. It makes it much easier for managing and keeping uptime.” Apache Cassandra’s automatic replication to multiple nodes for fault-tolerance also has helped Robb to sleep at night—literally. He explains, “We use the Amazon EC2 web service for pretty much everything. Recently, Amazon forced us to reboot all of our machines. One machine had a hardware issue, so it needed a more serious reboot that I couldn’t control, and would be down sometime during the night. Because of Cassandra’s replication, I didn’t have to worry about the machine not rebooting correctly. So, I just went home and got some sleep.”

Expanding Cassandra’s role

New York City-based SocialFlow has been growing rapidly over the past year—tripling the size of its workforce, securing new office

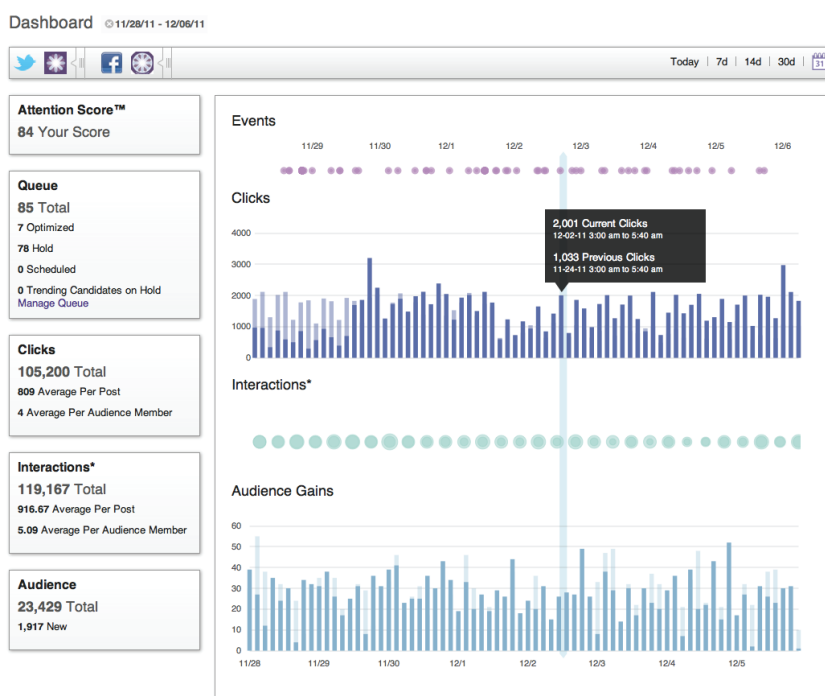
space, and receiving a funding boost. Robb says the company is now looking to Cassandra to help it grow and further differentiate its services for clients. “We are definitely using Cassandra in other ways that aren’t just active write/load firehose use cases,” says Robb. “One thing we’re trying to do is deploy some wide-scale machine learning, and we might use Cassandra for that in various ways. We’re also trying to dig deeper into the data from the firehose, examining things like, ‘Who is going to be more likely to respond to this type of content?’ and ‘What types of events would these followers likely find interesting?’”

Currently, SocialFlow stores all of its data in Amazon Simple Storage Service (S3). According to Robb, the company keeps about a month’s worth of rolling data and maintains a number of “streaming digests.” Those digests, built by Robb, allow the SocialFlow team to look back and see what was happening on Twitter at any point in time. “That’s all stored in Apache Cassandra and calculated on the fly,” he says.

The beauty of simplicity

Another key factor in Robb’s decision to implement Apache Cassandra at SocialFlow is that the platform is used by many leading companies—including Twitter. Robb also knew he could count on support from the Cassandra community. “There is no way I would use a closed-source database. If something really bad would happen, I couldn’t dig into the source and figure out the problem,” he says. “And if I can’t figure something out, I know I can turn to a really healthy community of active committers for help. They’re resolving all sorts of issues all the time and are always pushing out new releases.”

Something else that Robb likes about the Apache Cassandra database: its simplicity. “This may sound like a negative, but I think one of the best features of Cassandra is its simple schema,” says Robb. “Cassandra does only a few things, but it does them extremely well. So it forces you to think about modeling your data in a way that fits to that—and that causes you to design things a lot better.”



DataStax offers products and services based on the popular open-source database, Apache Cassandra™ that solve today’s most challenging big data problems. DataStax Enterprise (DSE) combines the performance of Cassandra with analytics powered by Apache Hadoop™, creating a smartly integrated, data-centric platform. With DSE, real-time and analytic workloads never conflict, giving you maximum performance with the added benefit of only managing a single database.

The company has over 100 customers, including leaders such as Netflix, Cisco, Rackspace and Constant Contact, and spanning verticals including web, financial services, telecommunications, logistics and government. DataStax is backed by industry leading investors, including Lightspeed Venture Partners and Crosslink Capital and is based in San Mateo, CA. For more information, visit www.datastax.com.



DATASTAX

270 East Lane #1
Burlingame, CA 94010