



Cassandra is already making life easier for Formspring's engineering team. With Cassandra, if a node has a temporary blip, things continue to hum along just fine, which is really amazing.

Company

Formspring

Data Size

3 clusters in production

Challenge

The need for a NoSQL database solution to support Formspring's explosive growth and allow the social network to provide reliable service and new features to its users worldwide, efficiently store and easily access large amounts of data, and seamlessly syndicate content to large groups of users.

Solution

The open source, elastically scalable and reliable Apache Cassandra™ platform, which delivers significantly faster performance than other solutions, while reducing data storage costs and IT's system management responsibilities.

Users of Formspring engage with and learn more about each other by asking and responding to questions. However, software engineer Kyle Ambroff is quick to emphasize that the social network is more than just a Q&A site. "Think of it as bringing cocktail conversation to the Web," he says. "You can ask a group of friends or colleagues – anonymously, if you choose – to offer their feelings on a topic, share advice, talk about what movies they like – whatever you want to know."

Launched in 2009, San Francisco-based Formspring, which recently released its first native iOS application for mobile users, has more than 26 million registered members worldwide who provide more than 10 million responses daily. Formspring receives over 30 million unique visitors per month. "We have about 3.8 billion responses in our system," says Ambroff. "That's a lot of content."

Formspring's explosive growth is what prompted the company to explore other database solutions and ultimately, move to the elastically scalable and reliable Apache Cassandra™ platform. "We were initially using a mix of MySQL, Amazon SimpleDB, Redis and Memcached," Ambroff recalls. (Formspring, which is hosted on Amazon's ec2 infrastructure, had nearly 64 MySQL databases as of September 2011.)

Within just a month of the site going online, Formspring had more than a million users and was already having scaling issues. "We were rapidly outgrowing our original roots as a PHP web application using MySQL," says Ambroff. "By the time I joined the company in April 2011, our engineers were already starting to denormalize the MySQL schemas and had to do a lot of sharding."

Some Formspring users – including musicians, comedians and other celebrities – had begun to collect so many followers that the social network's existing infrastructure was overwhelmed. "Say a user has 2 million followers and asks all of them a question. Just syndicating that question to every follower was essentially impossible using the site's original architecture," Ambroff explains.

Other solutions could not support Formspring's new features

In addition to scaling problems, Formspring's existing infrastructure could not support the addition of new features either at all, or without significantly degrading the site's performance. "We needed to address some limitations," says Ambroff. "For instance, the original way we recorded relationships in the site was with two tables in a MySQL database: one for followed relationships and one for blocked relationships. Essentially, User A follows User B or User A blocks User B. That works fine when you have a million users. But with 26 million users, and between 8 and 10 million users active at any time, you need to be able to do really complicated queries on tables with a billion rows. We couldn't."

Just so the site could continue to operate, Formspring's engineers started to cache the entire collection of relationships for each user. "That sort of worked for a while, although not for any of our high-profile users," says Ambroff. "There's a finite size to the value you can stuff in Memcached and we were exceeding that. So, we needed to rethink our approach."



“Cassandra has allowed us to build bigger features faster and more reliably, while using less money and without needing to expand our staff.”

—KYLE AMBROFF, Formspring

Formspring’s engineering team evaluated several solutions, including Membase and Riak, before choosing Cassandra. “We wanted to build a social graph service to record all the relationships for any user,” explains Ambroff. “This is a rather large data set, and we need to randomly access it at between five and 50 milliseconds. Cassandra was really the only solution that could do it. We built a web service in production that was the interface for the social graph service that had Cassandra sitting behind it, and started writing all of our production data to it. The transition was really painless.”

Formspring currently has three Cassandra clusters. The social graph service is the largest, with 12 nodes. A second cluster, with eight nodes, is for Formspring’s response service, which indexes all responses to users’ questions (users can view the collective results to a particular question on a separate webpage). The third cluster, which has just four nodes, holds an index of all the questions that have been submitted by Formspring users, along with the sent boxes for every member.

Ambroff says that as Formspring continues to grow and add new features, the number of Cassandra clusters will increase too. “We try to maintain a really well-defined, service-oriented architecture on the backend of our site, so anytime we add a new feature that needs to store data in Cassandra, we will definitely create a separate cluster just to gain even more reliability,” he says.

Formspring preserving resources, saving costs with Cassandra

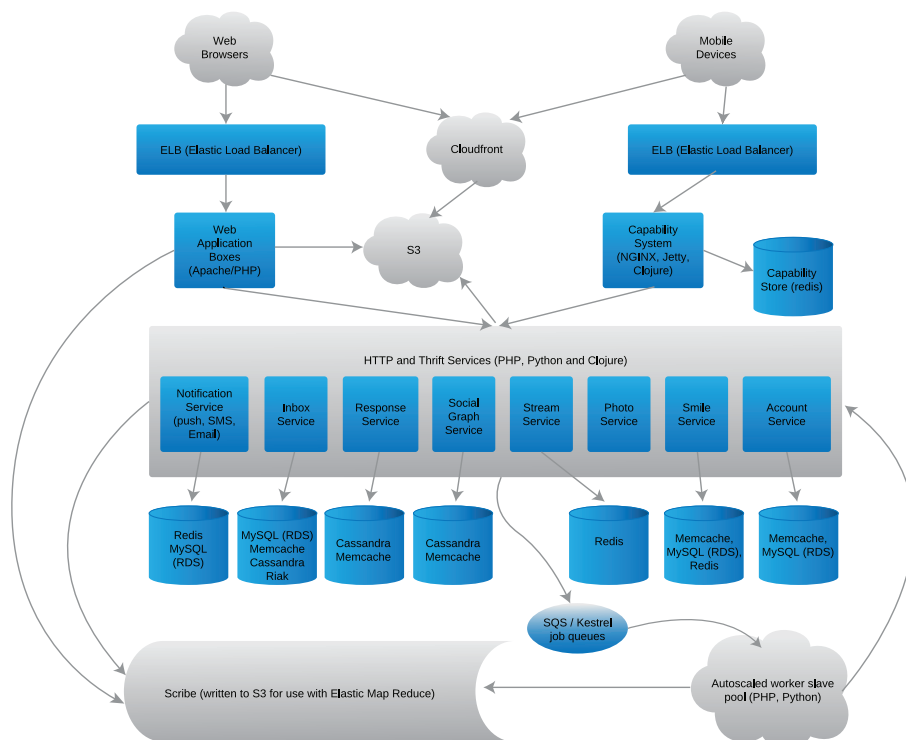
From a management perspective, Cassandra is already making life easier for Formspring’s engineering team. “For one, we have to do a lot less handholding,” Ambroff explains. “When we were sharding our MySQL or Redis servers, if one of those servers had a problem, like a latency spike or a net split, some percentage of our users didn’t have access to the site. With Cassandra, if a node has a temporary blip, things continue to hum along just fine, which is really amazing.”

He adds, “I can safely say that if we had tried to implement some of Formspring’s new features without using Cassandra, we would’ve had to double the size of our operations staff just because we’d be adding more single points of failure.”

Another anticipated benefit of using Cassandra is cost savings related to storage, according to Ambroff. “We have many features that were implemented using Redis; in fact, we have [one of] the largest Redis deployments. We’re running about a

terabyte of data, and it all needs to be in memory, which is a huge cost for us,” he explains. “On ec2, we have about 81 instances running Redis for our event stream feature, which allows you to see a list of all questions and responses coming in from your group of friends. We’re looking to migrate that data to Cassandra, so we can cut the number of ec2 instances down to a dozen or so. The cost savings will be significant.”

Ambroff says the fact that Cassandra is an open source product, and has such an engaged community around it, has been invaluable to his team’s success so far. “Cassandra has allowed us to build bigger features faster and more reliably, while using less money and without needing to expand our staff. That’s particularly important, as we are trying to grow the website with a lean engineering team. When we’ve encountered issues during production, we’ve been able to just dive into the code and quickly figure out what’s going on. We couldn’t do that if we were using something proprietary. It’s reassuring to know that you have that power.”



About DataStax DataStax is the developer of DataStax Enterprise, a distributed, scalable, and highly available database platform that delivers optimal performance either on premise or in the cloud for modern enterprise applications that manage both real-time and analytic workloads. 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 Burlingame, CA with offices in Austin, TX and Stamford, CT. For more information, visit www.datastax.com



DATASTAX

270 East Lane #1
Burlingame, CA 94010