Reddit Case Study
February 2011

IndexTank Eases Growing Pains

After a bout with infamously disappointing search, social news site Reddit found IndexTank, which easily handles the company's huge index, delivers real-time results, and has helped re-polish their reputation.

At social news site Reddit, people submit links to Internet content or create "self-posts," in which they write their own original text. Other users in the community can comment on the listings and give each one an up or down vote. The posts with the most positive votes rise to the top, meaning that the most popular items from the last 24 hours are always on the front page.

Behind the web site, Reddit is four software engineers working in a room in San Francisco's South of Market neighborhood. The walls are plastered with postcards from around the world, because the company offers a free month of Reddit gold, a set of extended account features, to anyone "who takes the time to send us a postcard."

The Challenge: Scalable Search

"We'd never had a really good search implementation," says David King of Reddit. "We were continually outgrowing whatever we had in place."

The company's list of tried-and-failed solutions includes Sphinx, TSearch, Xapian and PyLucene. Of PyLucene, Jeremy Edberg of Reddit says, "We didn't really scale out of that one, but the results were bad and it crashed all the time." Next, they switched to Solr, which worked well for a time. "That was a great short-term solution," says King, "but like the others before it, we hit a wall where the quality of results dropped off very quickly."

This cycle of outgrowing and replacing seemed like it was all part of being a successful startup until early 2010. That's when users decided they'd had enough, and rebelled by posting angry messages about how bad the search feature was. Thousands of frustrated users voted those posts up all the way to the front page, where they appeared over and over again.

The Back Story

Founded by two friends in June 2005 and bought by Condé Nast Digital in October 2006, Reddit started small and grew quickly. The company's user community of *redditors* has grown exponentially from the get-go. As the number of users increased, so did the total number of posts and votes, creating not only a huge pile of data, but also the need for frequent updates to the huge pile of data to refresh each item's score. Because remember, a news item's total number of "up" votes determines its priority ranking on the site, rewarding the most popular with a front page slot, and relegating the rejects to the bottom of the pile.

Every time someone votes for or against a post, its score updates. With the site receiving an average of about 3,000 votes per minute, and traffic doubling every six months, that number is high and getting higher.

In order to deliver results in order of popularity, the search solution needs each item's score in real-time,

or close to it. Historically, it's been these increasingly frequent updates that have broken each successive search solution.

"All of the scoring updates to Solr fragmented its index, which slowed it way down and made the results unusable," explains King. On top of Solr's fragmentation and performance problems, Reddit's two-phase commit process (when anything changes, it commits first, and then is made available in the index) caused the solution to go offline for 5-10 minutes each time it updated. "So, actually, it was both unusable and crashing," Edberg adds. "We tried to fix it by adding more search servers, but we scaled out of that too."

During this time is when users started writing their own stories (complete with long comment threads) on the site about how terrible the search was. A new one showed up on the front page a couple of times a month, voted there by the community. King says, "It was consistently embarrassing, knowing that millions of people were seeing our own users call us incompetent."

The Solution: IndexTank

As Reddit shopped around for a new solution, they found that many of the options simply wouldn't scale to fit their needs. Meanwhile, their problem was becoming infamous. "Lots of companies were coming to us, offering to help," says Edberg, "but the problem was, we couldn't get through the salespeople to talk to anyone who knew what they were talking about. We knew exactly what our problem was; we just couldn't find a solution that we knew would work."

The team also talked to representatives from Google Search. "Their salespeople called me every week for two months," King says. "When we finally worked out what the API would look like, we got the pricing chart, which is based on the number of user searches. We qualified for the scary category at the bottom of the page, which doesn't have numbers. It just says 'contact sales.'" The final bill was going to be well over $200,000 per year. That's an order of magnitude higher than what they were spending on search.

That's when Reddit learned about IndexTank. "Diego, the CEO, actually came by the office to talk to us," says King. "Since he's one of the principal authors of IndexTank, and also a Reddit user, he cut right to the chase and we started talking about merge trees and how their tool handles updates."

Confident that IndexTank would at least improve on what they had, Reddit agreed to do an opt-in beta on the site. Because Reddit is open-source, two IndexTank engineers quickly wrote and sent the patch. "We didn't even have to write the code," says King. "That was huge. And then right away, IndexTank was crazy fast."

Business Benefits

During the six-month beta, Reddit added a question to both its old and new search pages that asked people if they're satisfied with the results. Users can answer yes or no. On the old Solr implementation, about 50 percent said yes. IndexTank scored 70 percent yes right out of the gate. These days, that number is at 80 percent—and growing.

IndexTank's pricing structure fits Reddit, too. It's far less than a dedicated resource, and feature-comparable to the hundreds-of-thousands Google Commerce Search that doesn't provide the real-time indexing that Reddit needs. Plus, King says, "We're not just getting the product itself and hosting like we were before. We're getting access to a team of experts."

IndexTank has also cut down on administration time, from between five and eight hours a week to about one hour a month. Plus, it saves worry, embarrassment, and other unquantifiable things. “It may not sound like a big deal, but there are only four of us here. To have an item like search off our to-do list is a massive relief. I don’t know how to turn that into dollars, but it would be a lot,” King says.

“If we had unlimited programming resources, we could have built something great, but it was much easier to outsource, saving us the time, hassle and headache,” adds Edberg,

As engineers themselves, the Reddit team did wrestle with the idea of outsourcing their search instead of building it themselves. “The fact that search works so much better than if we’d built it, that makes up for any downside.”

IndexTank has also become a partner to Reddit as their community and needs continue to grow. The IndexTank team’s ability to provide search expertise and 1:1 support to the Reddit team remains a strong selling point.“The team is very competent, and that was a huge decision point for us,” says King. “We actually have access to the whole team on instant message.” Beyond access, IndexTank has taken an active role in understanding Reddit’s business and what users are trying to do. King says, “They suggest new things, and if we try something new, they help us test it. I appreciate that they’re excited about Reddit, because they have a lot of great ideas.”

Not only that, but since the introduction of IndexTank, many users who taunted the team with angry posts in the past have apologized. “It’s like a running joke now,” says Edberg. “Someone will write ‘search sucks’ in a comment, and someone else will say, ‘No, you’re living in the past.’”

Pullquotes
“We were continually outgrowing whatever we had in place.”
--David King, Reddit

“If we had unlimited programming resources, we could have built something great, but it was much easier to outsource, saving us the time, hassle and headache”
--Jeremy Edberg, Reddit

At a Glance

Company: Reddit
Location: San Francisco, CA
Industry: Web: Social media
Business Challenge: After consistently out-scaling search solutions, social news site Reddit had to find a tool that could handle their fast-paced growth.

Solution: IndexTank Real-Time Search

Results: With IndexTank, Reddit:

- Delivers the right search results to 30 percent more users
- Saves 20-30 hours a month
- No longer suffers from ridicule by their own users

For more information, visit:
IndexTank
www.indextank.com

Reddit
www.reddit.com