Lucidworks

reddit

CASE STUDY

The Search for Better Search at Reddit

How the team at Reddit destroys global productivity with cat photos and Lucidworks Fusion

Reddit — "the front page of the internet" — is in the top ten most popular websites in the U.S. with 330+ million users, ~10+ million post submissions per month, almost 3 million comments, and ~60 million upvotes and downvotes across hundreds of thousands of communities each day. Reddit is organized around communities, or "subreddits," which cover every topic and passion imaginable. People use Reddit to gather news and information, share experiences, and find communities and conversations about shared interests.

"The Front Page of the Internet"

Reddit has a deluge of content that's constantly being submitted, voted on, and discussed; so building a great search experience is no easy feat. The team needed a scalable solution that could quickly locate relevant content for millions of searches per day. Over the past 12 years, Reddit implemented five different search solutions with varying success.

When the site first launched in 2005, it used a very basic search built with Postgres. This solution lasted three years, but traffic logs were showing that while search was just 2% of site traffic, it was using more than half of system resources. After that, the team tried using open source Apache Lucene with a Python RPC client. Then later, a solution using open source Apache Solr was deployed. In 2010, as Reddit reached an unprecedented number of pageviews per month, the team deployed another search - this time built with Memcached and Cassandra - to try and keep ahead of the Reddit's user base and growth trajectory. The next approach was to outsource search to an outside company. But it was a short-lived partnership due to the vendor's acquisition several months later. Then Reddit switched to a cloud-based search provicder, but it did not meet Reddit's growing requirements. But query response times got longer, latency was slowing, searches started timing out, and users grew increasingly impatient. At a virtual town hall in 2016, Reddit's CEO was asked, "Where do you see Reddit in 10 years?" The CEO responded, "Reddit search might work by then."

The Challenge

Reddit handles ~10M monthly posts, almost 3M daily comments and 58M daily votes across 138K active communities. Search is a key way Reddit's more than 330M monthly users find content. After trying five different search stacks in 12 years, they needed a solution that could scale and quickly locate relevant content for almost 47M daily searches.

Our Plan

Lucidworks and Reddit partnered to deliver a solution that was fast, scalable, responsive and relevant that laid the foundation for future improvements.

The Results

Launched in September 2017, the response to the new search app has been overwhelmingly positive. Search cluster reduced from 200 to 30 nodes. Uptime at 99.99% with 400 queries and 1,000 updates per second including live streaming updates.

The Solution

Reddit partnered with Lucidworks to build the site's new search with Fusion, an advanced platform for developing smart apps. Reddit wanted a solution that included Apache Solr, the open source search technology known for its reliability and scalability. Fusion includes Solr as part of its search stack, so the team was confident the new search solution would be able to scale to the magnitude needed for such a popular site.

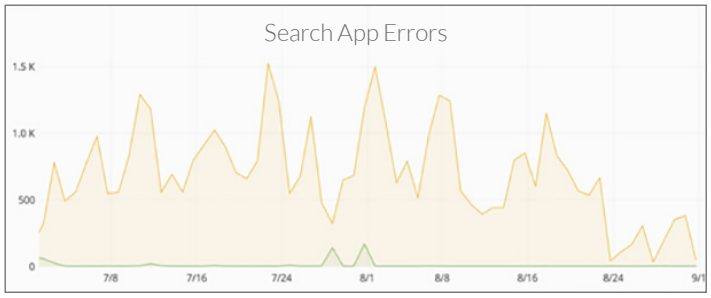
From 200 Nodes to 24

The new search app was separated into three parts for faster, more iterative development. The first part is a search microservice which runs two to three instances a day. Then the existing data pipeline architecture was used used to pull in nightly batch updates, building a canonical view of all content as well as the live updates streaming in through Kafka. And the third piece is the Fusion cluster itself, which is 24 nodes, indexing all the content of the front page of the internet. The previous search app had weighed in at 200 instances.

Faster Indexing, Faster Everything

Index time was significantly reduced as well with a full corpus index that would require two weeks reduced to just two days with ingestion rates of five thousand updates per second. To test scaling and performance of the Fusion cluster, all search traffic sent to the live site was also routed to Fusion. Then over a few weeks, more and more queries were gradually sent to and processed by the Fusion cluster with results sent back to the live site and viewed by more and more users. Error rates dropped significantly to practically zero.

Comparing errors from the legacy search app (yellow) and Fusion (green) implementations:



**Green spikes indicate when nodes were taken down for updates.*

Maintaining Relevance

In addition to scalability and performance, the new search app had to achieve the same level of relevance in search results. By tracking clicks on search results, the team was able to determine that the search results coming from Fusion had the same quality as the existing search, so users would not experience a drop in relevance. Easy wins were achieved with filtering out spam and other bad queries. Relevancy is also being tuned to aggregate results by Reddit communities (a.k.a. "Subreddits"), so a search for news about the Super Bowl won't be incorrectly routed to the community dedicated to sharing photos of majestic owls, /r/SuperbOwl.

Positive Reception

The new search app built with Fusion reached full 100% availability to all of Reddit's 330+M million users in September 2017, and so far the reaction has been overwhelmingly positive. The search cluster is indexing a quarter of a billion posts with 1000+ updates per second, serving 400 queries per second to both users and to third-party apps through Reddit's APIs.

Future of Search at Reddit

- Future plans for the Reddit search team include ingesting other document types, like comments and messages, to improve the accuracy and relevancy of search results, so users find exactly what they are looking for.
- Better understand each user's interests to present optimized results and a feed full of relevant content.
- Finally, the search interface is going to become more integral to Reddit's user experience, serving users richer, more relevant (and directional) content.

"Reddit relies heavily on content discovery, as our primary value proposition is giving our people a home for discovering, sharing, and discussing the things they're most passionate about. We expect Fusion's customization and machine learning functionality will significantly elevate our search capabilities and transform the way people discover content on the site."

— Nick Caldwell, Vice President of Engineering at Reddit

Get Started or Learn More

For more information or to start using Lucidworks Fusion, contact us today at lucidworks.com/contact or call 415-329-6515.