Enterprises struggling with the inherent performance and scalability limitations of relational databases are increasingly turning to modern distributed information management technologies such as Cassandra. Mahalo.com was able to easily upgrade their original relational database management system and employ Cassandra as a key component of a speedy, cost-effective and highly scalable architecture.

In operation since 2007, Mahalo.com is dedicated to helping people easily find the information and resources they can trust. This mission resonates with users: in just a few years, Mahalo.com has grown to be a top 200 website with 12 million visitors per month. Hosted at two datacenters, Mahalo.com is deployed on the FreeBSD operating system and constructed on the Apache, Python, and Django stack. Python serves as the primary software development language. To control hardware costs, Mahalo.com runs on Intel-based commodity hardware, using HAProxy for cluster load balancing.

Cassandra at Mahalo.com

Until recently, a replicated MySQL cluster had served as Mahalo.com’s sole storage mechanism. However, a series of RDBMS performance and application functionality shortfalls contributed to the selection of Cassandra as an important addition to the Mahalo.com infrastructure, capable of supporting current and projected future activity.

First, from a performance perspective, crucial portions of the Mahalo.com application are heavily write-oriented. Its workload continually fluctuates, making it very difficult for administrators to predict operational peaks and valleys.

This volatility and variability mandated a new information storage and retrieval approach. Furthermore, the response time for queries that required scanning multiple MySQL tables began to exceed acceptable tolerances.

As Mahalo.com sought to implement new application functionality in support of user profiles, it became apparent that the limitations of a traditional RDBMS would add extra effort and complexity to this vital task. Given the performance and functionality drawbacks of their RDBMS, architects at Mahalo.com determined that a modern key/value store would be a more appropriate information management strategy to support their growing enterprise.

Originally, Apache HBase had been identified as a possible storage solution, but after an intensive four week evaluation Cassandra was determined to be the best option.

“While MySQL had been a big part of our technology stack, we realized that it wasn’t the ideal destination for all of our information. We also needed a database that would not get in the way of us adding new features to our site”, says Jacob Burch Chief Technical Officer, Mahalo.com

The migration to Cassandra took less than two months. Database administrators began by loading the previous month’s worth of data into Cassandra. They used this information as the baseline for going live, and were happy to observe that their Cassandra cluster had no problem digesting the large amounts of data that was fed in during the migration. Once the initial data load was completed, the remainder of the historical MySQL-based information was transferred into

With a minimal amount of effort, Mahalo.com used Cassandra to build a data infrastructure that will support future throughput and scalability requirements while helping to control costs.
“With the Cassandra conversion completed and running smoothly, we’re now free to focus on our primary mission, knowing that we’ll be able to deliver the excellent responsiveness and capabilities that our user community has come to expect”
—JACOB BURCH, Mahalo.com

Cassandra. Figure 1 illustrates the Mahalo.com architecture.

Cassandra has been in production for over a year, with 120 GB of data under management. Mahalo.com is averaging 46 transactions per second for read activity, and 42 transactions per second for write operations. As data volumes and application loads have grown, Mahalo.com was able to seamlessly scale from four nodes to eight nodes during this time.

Cassandra is now an essential underpinning of two key Mahalo.com capabilities, including:

- **Activity log.** Each time any Mahalo.com user takes any action, a record is created in the activity log. In addition to the large volumes of information generated by these transactions, each activity log change must also be propagated to many other feeds. Cassandra has been able to handle the enormous write volumes created by all of these events, as well as provide near real-time access to data.

- **Topics.** Each question and answer maintained in Mahalo.com may be tagged with up to five topics. Cassandra makes it possible to quickly interrogate the database and correlate the hundreds of thousands of topics that have been tagged by Mahalo.com users.

Now that the initial Cassandra implementation has been completed, Mahalo.com plans to continue migrating additional portions of their product and double the size of the Cassandra cluster in multiple data centers and convert other applications as well.

**Conclusion**

Faced with untenable RDBMS-driven performance and application functionality roadblocks, Mahalo.com chose Cassandra as a critical component of their future technology architecture. With a minimal amount of effort, Mahalo used Cassandra to build a data infrastructure that will support future throughput and scalability requirements while helping to control costs.

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 combines the performance of Cassandra with analytics powered by Apache Hadoop and enterprise search with Apache Solr, creating a smartly integrated, big data platform. With DataStax Enterprise, real-time, analytic, and search workloads never conflict, giving you maximum performance with the added benefit of only managing a single database. The company has over 140 customers, including leaders such as Netflix, Disney, Cisco, Rackspace and Constant Contact, and spans 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 visit www.datastax.com.

© 2012 DataStax. All rights reserved.