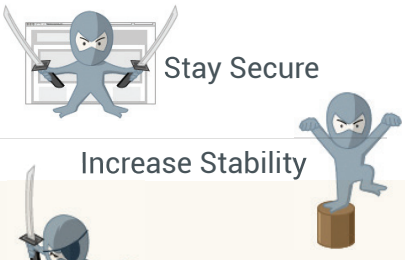
DATASTAX



SourceNinja

SourceNinja is benefitting from powerful, scalable enterprise search capabilities built into DataStax Enterprise and supplied by Apache Solr™. It doesn't need to trade high performance to have more flexibility in querying data.

SourceNinja aims to be the leading resource for all information about open source software and updates for anyone who works with the model – from individual developers to large enterprises. The Sunnyvale, California-based company, which was founded in 2011, monitors and tracks all open source software versions across all platforms to identify third-party libraries that are out of date and prevent problems in applications that use open source.

"We want to help open source developers save time and money and avoid potential legal problems by notifying them when updates are released or problems occur," explains Matt Stump, SourceNinja's chief technology officer. "Issues that SourceNinja can help to prevent include license changes, security vulnerabilities, and performance problems."

Anyone working with open source software knows that staying apprised of the latest updates can be an arduous task. "Tracking packages takes a lot of time," says Stump. "The results of doing it on your own aren't always the best, and by the time you're done, you have to start all over again." He adds that even large enterprises struggle to keep pace with open source updates. "That's why we started SourceNinja."

Big data challenges emerged in the alpha phase

SourceNinja, propelled by a round of seed funding from an angel incubator in late 2011, was in early 2012 making plans to emerge from the alpha phase and open up its data list to a wider audience. At the time, the company had been using the DataStax Enterprise big data management platform, which is built on Apache Cassandra software, for about four months. SourceNinja had decided to implement a NoSQL solution after discovering that its PostgreSQL ("Postgres") object-relational database management system simply could not handle big data challenges, even in the alpha phase.

"We're tracking every single piece of open source out there – hundreds of millions of pieces of information – from products to vendors to updates," says Stump. "Postgres was under strain with just a few hundred users querying the data, so we knew we had a problem."

SourceNinja first evaluated HBase, but ran into issues immediately, according to Stump. "First, there are single points of failure with HBase," he says. "You have to name nodes. And if a node goes down, or you lose data on those nodes, the whole cluster is lost. We didn't like that. Also, we found that HBase is simply not fast enough to serve web requests."

Stump then decided to explore the use of Cassandra, a solution he was a ready familiar with but had not evaluated since before DataStax existed. "I was really impressed by the progress DataStax had made in just a short time in making Cassandra more robust and easy for enterprises to use with offerings like DataStax Enterprise," he says.

Lower costs, less maintenance with embedded Hadoop

The DataStax Enterprise platform provides integrated Apache Hadoop, which was particularly appealing to SourceNinja's engineering team. "We do a lot of background processing via Hadoop," Stump explains. "We pull data from many different data sources and do transforms, calculations and merges and then populate the Cassandra database with that data. Previously, we used Backgroundworkers, but that approach didn't work well for us. It's expensive and takes a lot of time to maintain. We're a startup, so we're really resource-conscious."

SourceNinja

Company
SourceNinja

Data Size
1 3-node cluster

Business Challenge
SourceNinja, a fast-growing startup, needed to find a way to handle its big data challenges more efficiently and cost-effectively as it prepared to open its data list to a wider audience.

Technical Challenge
SourceNinja wanted a highly reliable and easily scalable NoSQL distributed database solution that could deliver both high performance and greater flexibility in querying – but also reduce its development team's workload.

Solution
DataStax Enterprise, a fully integrated big data platform for managing real-time, analytic, and enterprise search data all in the same database, which is powered by Apache Cassandra™ software and features continuously available Apache Hadoop™ and powerful search support with Apache Solr™.

"Having DataStax Enterprise embedded with Solr means a lot of work is taken away and transitioning to NoSQL is much simpler. I get all the performance, scalability, reliability, and redundancy of NoSQL, but also keep a really powerful query language."

Matt Stump
Chief Technology Officer, SourceNinja

As part of the transition to Cassandra, Stump and his development team of four have been moving all of SourceNinja's big data to Hadoop. "We've been really happy with it," he says. "There's less code to write, fewer errors, and better tools – and the jobs take less time."

Because SourceNinja's data sources are open source software, Stump says variation is an issue – but one that Cassandra easily handles. "Cassandra is column-based, so we don't have to declare the schema up front and can allow for that variation," he explains. "Being able to choose how to denormalize the data is also very nice."

SourceNinja is also benefitting from the powerful and scalable enterprise search capabilities that are built into DataStax Enterprise and supplied by Solr (via Lucene and Apache Solr). "SQL is great for searching for absolute values, but if you want to do a wild card search – a "like" expression, for example – that can get really expensive," explains Stump. "There are database-specific add-ons for doing full-text indexing or trigram searching, but again, these approaches can be very costly, and you end up being wedded to that one database."

Letting go of an overly complex system

Having Solr built into DataStax Enterprise means SourceNinja does not need to trade high performance for more flexibility in querying data. "I can just put all my data into Cassandra, make sure it's indexed, and then query it however I want," says Stump. "Having DataStax Enterprise embedded with Solr means a lot of work is taken away and transitioning to NoSQL is much simpler. I get all the performance, scalability, reliability, and redundancy of NoSQL, but also keep a really powerful query language."

With Cassandra, and the DataStax Enterprise platform, SourceNinja has been able to lower both operational and development costs. "I now have a system that works with Hadoop, so I can transition everything off the Backgroundworker queue. I have the flexibility of a columnar database, and really powerful search capabilities built right into the stack with Solr. The code is also simplified, which has in turn simplified my operations and deployment, and actually sped up my system overall."

Extracting even more value from big data

Stump says SourceNinja is now exploring ways to make even better use of the powerful search capabilities provided by Cassandra and Solr through the DataStax Enterprise platform. "Our customers are now able to really drill down into the data to get more interesting information. For example, a user might want to ask, 'I'm looking for a piece of open source that lets me parse X amount in Java and doesn't have any open security exploits.' That type of query becomes much more doable just because of the way data is stored in Cassandra."

In the future, SourceNinja's engineering team plans to use Hadoop even more strategically, such as for running processes like machine learning and data mining. "We couldn't have done these things on the old stack," says Stump. "Our ability to deliver interesting, exciting, novel, and really game-changing services is now significantly augmented by the fact we're using Cassandra with DataStax Enterprise."

The expert service that DataStax provides is also helping SourceNinja to have a positive experience implementing Cassandra – and to get the most from robust capabilities like Hadoop and Solr, according to Stump. "I really enjoy working with the DataStax team," he says. "They're very responsive, take my bug reports seriously, and turn around fixes quickly."

> "I was really impressed by the progress DataStax had made in just a short time in making Cassandra more robust and easy for enterprises to use with offerings like DataStax Enterprise."

Matt Stump
Chief Technology Officer, SourceNinja

About DataStax

DataStax powers the apps that transform businesses. DataStax powers over 250 Big Data apps for start ups and 20 of the Fortune 100 with its flexible and massively scalable big data platform built on Cassandra, through multi-data centers.

DataStax Enterprise delivers enterpriseready Cassandra, then goes one step further by integrating the best of breed Big Data technologies — Apache Hadoop for analytics, and Apache Solr for search across multiple datacenters and the cloud.

Top companies such as Adobe, HealthCare Anytime, eBay, and Netflix rely on DataStax to transform their businesses. Based in San Mateo, Calif., DataStax is backed by industry-leading investors: Lightspeed Venture Partners, Crosslink Capital and Meritech Capital Partners. For more information, visit http://www.datastax.com/ and follow@DataStax.

DATASTAX
777 Mariners Island Blvd #510
San Mateo, CA 94404
650-389-6000