Zachary Erwin

TC 362

February 24th, 2015

Professor Siarto

Many Eons ago, back in the dark ages, or what most youngsters would now consider the early to mid 2000's, there was a news aggregation website that had quite a bit of traffic. That website is Digg or rather digg.com. I confess, that I do not know if Digg receives more traffic than Reddit, which seems to be the preferred method of information gather amongst my generation. With all of that said Digg's technology stack is something that is what concerns us today.

In terms of the technology specifications of digg.com there were not documents that were publically available about what kind of power any of their servers had. With that said Digg does have something interesting going on in terms of the kind of open source software it uses. "**Digg.com** credits two particular features of its LAMP (Linux, Apache, MySQL and PHP) server cluster for helping the news aggregation site maintain speedy performance in the face of high growth.[1]" Now LAMP may not sound that impressive, but consider that fact that most of that most of the software being used in that cluster is technically open source, so they do not really have to worry about paying that much for it.

---

[1] http://www.computerworld.com/article/2544850/enterprise-applications/how-digg-com-uses-the-lamp-stack-to-scale-upward.html

MySQL is used as their database software while, PHP is what scripting language of choice for Digg, so Javascipt seems to be nowhere in sight. What is also interesting about the kind of technology that Digg uses is that it employs special software for distributed computing[2]. "The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures. [3]" The ability to scale software is important because Digg would want to have their software scale so they could grow or shrink as technology and the company demanded.

In fact "Today, Digg.com boasts 100 servers scattered in multiple data centers that host a total of 30GB of data, but the site started off in late 2004 as a single Linux server running Apache 1.3, PHP 4, and MySQL 4.0 using the default **MyISAM** storage engine,[4]". Having software that can scale would be a great benefit to a company that would want to expand beyond that number of servers.

The final aspect of that has to addressed is what kind of servers Digg.com is using and how many of them there are. "Digg's current architecture includes about

---

[2] http://www.developerfusion.com/news/91331/a-look-behind-the-numbers-on-digg-and-stack-overflow/
[3] http://hadoop.apache.org/#What+Is+Apache+Hadoop%3F
[4] http://www.computerworld.com/article/2544850/enterprise-applications/how-digg-com-uses-the-lamp-stack-to-scale-upward.html

20 database servers, 30 Web servers, and a few search servers running **Lucene**; the balance operate as backup servers. All but one of the database servers run some version of MySQL 5.[5]" The Lucene that is mentioned in that quote is a "free open source information retrieval software library.[6]" Digg has the backbone to handle the kind of traffic it gets from millions of users. This information is out of date by a few years, so the infrastructure the have installed with LAMP has been scaled to a larger size allowing Digg to start scaling to a larger amount of servers. Thank you for your time and have a great day!

[5] http://www.computerworld.com/article/2544850/enterprise-applications/how-digg-com-uses-the-lamp-stack-to-scale-upward.html
[6] http://en.wikipedia.org/wiki/Lucene