1.7 History of Hadoop

Hadoop started out as a subproject of Nutch, which in turn was a subproject of Apache Lucene. Doug Cutting founded all three projects, and each project was a logical progression of the previous one.

Lucene is a full-featured text indexing and searching library. Given a text collection, a developer can easily add search capability to the documents using the Lucene engine. Desktop search, enterprise search, and many domain-specific search engines have been built using Lucene. Nutch is the most ambitious extension of Lucene. It tries to build a complete web search engine using Lucene as its core component. Nutch has parsers for HTML, a web crawler, a link-graph database, and other extra components necessary for a web search engine. Doug Cutting envisions Nutch to be an open democratic alternative to the proprietary technologies in commercial offerings such as Google.

Besides having added components like a crawler and a parser, a web search engine differs from a basic document search engine in terms of scale. Whereas Lucene is targeted at indexing millions of documents, Nutch should be able to handle billions of web pages without becoming exorbitantly expensive to operate. Nutch will have to run on a distributed cluster of commodity hardware. The challenge for the Nutch team is to address scalability issues in software. Nutch needs a layer to handle distributed processing, redundancy, automatic failover, and load balancing. These challenges are by no means trivial.

Around 2004, Google published two papers describing the Google File System (GFS) and the MapReduce framework. Google claimed to use these two technologies for scaling its own search system. Doug Cutting immediately saw the applicability of these technologies to Nutch, and his team implemented the new framework and ported Nutch to it. The new implementation immediately boosted Nutch's scalability. It started to handle several hundred million web pages and could run on clusters of dozens of nodes. Doug realized that a dedicated project to flesh out the two technologies was needed to get to web scale, and Hadoop was born. Yahoo! hired Doug in January

2006 to work with a dedicated team on improving Hadoop as an open source project. Two years later, Hadoop achieved the status of an Apache Top Level Project. Later, on February 19, 2008, Yahoo! announced that Hadoop running on a 10,000+ core Linux cluster was its production system for indexing the Web (http://developer.yahoo.net/blogs/hadoop/2008/02/yahoo-worlds-largest-production-hadoop.html). Hadoop had truly hit web scale!

What's up with the names?

When naming software projects, Doug Cutting seems to have been inspired by his family. *Lucene* is his wife's middle name, and her maternal grandmother's first name. His son, as a toddler, used *Nutch* as the all-purpose word for *meal* and later named a yellow stuffed elephant *Hadoop*. Doug said he "was looking for a name that wasn't already a web domain and wasn't trademarked, so I tried various words that were in my life but not used by anybody else. Kids are pretty good at making up words."