Welcome to Apache Nutch#

Table of contents

1 What is Apache Nutch?

Apache Nutch is a highly extensible and scalable open source web crawler software project. Stemming from Apache Lucene#, the project has diversified and now comprises two codebases, namely:

1. Nutch 1.x: A well matured, production ready crawler. 1.x enables fine grained configuration, relying on Apache Hadoop# data structures, which are great for batch processing.
2. Nutch 2.x: An emerging alternative taking direct inspiration from 1.x, but which differs in one key area; storage is abstracted away from any specific underlying data store by using Apache Gora# for handling object to persistent mappings. This means we can implement an extremely flexibile model/stack for storing everything (fetch time, status, content, parsed text, outlinks, inlinks, etc.) into a number of NoSQL storage solutions.

Being pluggable and modular of course has it's benefits, Nutch provides extensible interfaces such as Parse, Index and ScoringFilter's for custom implementations e.g. Apache Tika# for parsing. Additonally, pluggable indexing exists for Apache Solr#, Elastic Search, etc.

Nutch can run on a single machine, but gains a lot of its strength from running in a Hadoop cluster

You can download Nutch here.

For more information about Apache Nutch, please see the Nutch wiki.

Nutch is a project of the Apache Software Foundation and is part of the larger Apache community of developers and users.

2 Getting Started

To get started, begin here:

1. Learn about Nutch by reading the documentation.
2. Download Nutch from the release page.
3. Discuss Nutch on the mailing lists.

3 Download Nutch

Please head to the releases page to download a release of Apache Nutch.

4 Apache Nutch News

4.1 17 March 2014 - Apache Nutch v1.8 Released

The Apache Nutch PMC are pleased to announce the immediate release of Apache Nutch v1.8, we advise all current users and developers of the 1.X series to upgrade to this release.

Alhough this release includes library upgrades to Crawler Commons 0.3 and Apache Tika 1.4, it also provides over 30 bug fixes as well as 18 improvements. Please see the list of changes for a full breakdown, or see the release report. As usual in the 1.X series, this release is made available both as source and binary. Additionally developers can find Maven artifacts within Maven Central. The release is available here.

4.2 02 July 2013 - Apache Nutch v2.2.1 Released

The Apache Nutch PMC are pleased to announce the immediate release of Apache Nutch v2.2.1, we advise all current users and developers of the 2.X series to upgrade to this release ASAP. Although this release includes library upgrades to Apache Hadoop 1.2.0 and Apache Tika 1.3, it is predominantly a bug fix for NUTCH-1591 - Incorrect conversion of ByteBuffer to String. Please see the list of changes for a full breakdown, or see the release report. As usual in the 2.x series, this release is made available only as source, but is also available within Maven Central. The release is available here.

4.3 24th June 2013 - Apache Nutch v1.7 Released

The Apache Nutch PMC are extremely pleased to announce the immediate release of Apache Nutch v1.7. This release includes over 20 bug fixes, as many improvements; most noticeably featuring a new pluggable indexing architecture which currently supports Apache Solr and Elastic Search. Shadowing the recent Nutch 2.2 release, parsing of Robots.txt is now delegated to Crawler-Commons. Key library upgrades have been made to Apache Hadoop 1.2.0 and Apache Tika 1.3. Please see the list of changes or the release report made in this version for a full breakdown. As usual in the 1.x series, the release is made available as binary and source (zip + tar.gz) and is also available within Maven Central. The release is available here.

4.4 08 June 2013 - Apache Nutch v2.2 Released

The Apache Nutch PMC are extremely pleased to announce the immediate release of Apache Nutch v2.2. This release includes over 30 bug fixes and over 25 improvements representing the third release of increasingly popular 2.x Nutch series. This release features inclusion of Crawler-Commons which Nutch now utilizes for improved robots.txt parsing, library upgrades to Apache Hadoop 1.1.1, Apache Gora 0.3, Apache Tika 1.2 and Automaton 1.11-8. Please see the list of changes or the release report made in this version for a full breakdown. As usual in the 2.x series, this release is made available only as source, but is also available within Maven Central. The release is available here.

4.5 06 December 2012 - Apache Nutch v1.6 Released

The Apache Nutch PMC are extremely pleased to announce the release of Apache Nutch v1.6. This release includes over 20 bug fixes, the same in improvements, as well as new functionalities including a new HostNormalizer, the ability to dynamically set fetchInterval by MIME-type and functional enhancements to the Indexer API inluding the normalization of URL's and the deletion of robots noIndex documents. Other notable improvements include the upgrade of key dependencies to Tika 1.2 and Automaton 1.11-8. Please see the list of changes or the release report made in this version for a full breakdown. The release is available here.

4.6 05 October 2012 - Apache Nutch v2.1 Released

The Apache Nutch PMC are very pleased to announce the release of Apache Nutch v2.1. This release continues to provide Nutch users with a simplified Nutch distribution building on the 2.x development drive which is growing in popularity amongst the community. As well as addressing ~20 bugs this release also offers improved properties for better Solr configuration, upgrades to various Gora dependencies and the introduction of the option to build indexes in elastic search. Please see the list of changes made in this version for a full breakdown. The release is available here.

4.7 10 August 2012 - Happy 10th Birthday Apache Nutch!!

It's official, Apache Nutch is now a decade old! The project has come a long long way since inception, through acceptance into the Apache Incubator way back in Janurary 2005, to the Top Level Project it became on 21st April 2010. Happy birthday Nutch and thanks to all contributors past and present! See Doug Cutting's tweet.

4.8 10 July 2012 - Apache Nutch v1.5.1 Released

The Apache Nutch PMC are very pleased to announce the release of Apache Nutch v1.5.1. This release is a maintainence release of the popular 1.5.X mainstream version of Nutch which has been widely adopted within the community. Please see the list of changes made in this version for a full breakdown. The release is available here.

4.9 07 July 2012 - Apache Nutch v2.0 Released

The Apache Nutch PMC are very pleased to announce the release of Apache Nutch v2.0. This release offers users an edition focused on large scale crawling which builds on storage abstraction (via Apache Gora#) for big data stores such as Apache Accumulo#, Apache Avro#, Apache Cassandra#, Apache HBase#, HDFS#, an in memory data store and various high profile SQL stores. After some two years of development Nutch v2.0 also offers all of the mainstream Nutch functionality and it builds on Apache Solr# adding web-specifics, such as a crawler, a link-graph database and parsing support handled by Apache Tika# for HTML and an array other document formats. Nutch v2.0 shadows the latest stable mainstream release (v1.5.X) based on Apache Hadoop# and covers many use cases from small crawls on a single machine to large scale deployments on Hadoop clusters. Please see the list of changes made in this version for a full breakdown. The release is available here.

4.10 07 June 2012 - Apache Nutch 1.5 Released

The 1.5 release of Nutch is now available. This release includes several improvements including upgrades of several major components including Tika 1.1 and Hadoop 1.0.0, improvements to LinkRank and WebGraph elements as well as a number of new plugins covering blacklisting, filering and parsing to name a few. Please see the list of changes made in this version for a full breakdown of the 50 odd improvements the release boasts. The release is available here.

4.11 26 November 2011 - Apache Nutch 1.4 Released

The 1.4 release of Nutch is now available. This release includes several improvements including allowing Parsers to declare support for multiple MIME types, configurable Fetcher Queue depth, Fetcher speed improvements, tigther Tika integration, and support for HTTP auth in Solr indexing. Please see the list of changes made in this version. The release is available here.

4.12 23 September 2011 - Apache Nutch focuses on 1.x series for main development

After some discussion and a vote about the issue, the Nutch development community decided to focus their efforts on maintaining and releasing the 1.x series of Nutch, and to branch the now former Nutch trunk based on Gora, allowing others to try and improve it, while the mainline development goes on.

4.13 7 June 2011 - Apache Nutch 1.3 Released

The 1.3 release of Nutch is now available. This release includes several improvements (improved RSS parsing support, tighter integration with Apache Tika, external parsing support, improved language identification and an order of magnitude smaller source release tarball -- only about 2MB!). Please see the list of changes made in this version. The release is available here.

4.14 24 September 2010 - Apache Nutch 1.2 Released

The 1.2 release of Nutch is now available. This release includes several improvements (addition of parse-html as a selectable parser again, configurable per-field indexing), new features (including adding timing information to all Tool classes, and implementation of parser timeouts), and bug fixes (fixing an NPE in distributed search, fixing of XML formatting issues per Document fields). Please see the list of changes made in this version. The release is available here.

4.15 06 June 2010 - Apache Nutch 1.1 Released

The 1.1 release of Nutch is now available. This release includes several major upgrades of existing libraries (Hadoop, Solr, Tika, etc.) on which Nutch depends. Various bug fixes, and speedups (e.g., to Fetcher2) have also been included. See list of changes made in this version. The release is available here.

4.16 21 April 2010 - Apache Nutch graduates to TLP

Passed by unanimous approval of the Apache Board, Nutch graduated to TLP status. We are in the process of updating the website, and moving things around, so if you notice anything out of place, please let us know.

4.17 14 August 2009 - Lucene at US ApacheCon

ApacheCon US is once again in the Bay Area and Lucene is coming along for the ride! The Lucene community has planned two full days of talks, plus a meetup and the usual bevy of training. With a well-balanced mix of first time and veteran ApacheCon speakers, the Lucene track at ApacheCon US promises to have something for everyone. Be sure not to miss:

Training:

- Lucene Boot Camp - A two day training session, Nov. 2nd & 3rd
- Solr Day - A one day training session, Nov. 2nd

Thursday, Nov. 5th

- Introduction to the Lucene Ecosystem - Grant Ingersoll @ 9:00
- Lucene Basics and New Features - Michael Busch @ 10:00
- Apache Solr: Out of the Box - Chris Hostetter @ 14:00
- Introduction to Nutch - Andrzej Bialecki @ 15:00
- Lucene and Solr Performance Tuning - Mark Miller @ 16:30

Friday, Nov. 6th

- Implementing an Information Retrieval Framework for an Organizational Repository - Sithu D Sudarsan @ 9:00
- Apache Mahout - Going from raw data to Information - Isabel Drost @ 10:00
- MIME Magic with Apache Tika - Jukka Zitting @ 11:30
- Building Intelligent Search Applications with the Lucene Ecosystem - Ted Dunning @ 14:00
- Realtime Search - Jason Rutherglen @ 15:00

4.18 23 March 2009 - Apache Nutch 1.0 Released

The 1.0 release of Nutch is now available. This release includes several major feature improvements such as new indexing framework, new scoring framework, Apache Solr integration just to mention a few. See list of changes made in this version. The release is available here.

4.19 09 February 2009 - Lucene at ApacheCon Europe 2009 in Amsterdam

Lucene will be extremely well represented at ApacheCon EU 2009 in Amsterdam, Netherlands this March 23-27, 2009:

- Lucene Boot Camp - A two day training session, March 23 & 24th
- Solr Boot Camp - A one day training session, March 24th
- Introducing Apache Mahout - Grant Ingersoll. March 25th @ 10:30
- Lucene/Solr Case Studies - Erik Hatcher. March 25th @ 11:30
- Advanced Indexing Techniques with Apache Lucene - Michael Busch. March 25th @ 14:00
- Apache Solr - A Case Study - Uri Boness. March 26th @ 17:30
- Best of breed - httpd, forrest, solr and droids - Thorsten Scherler. March 27th @ 17:30
- Apache Droids - an intelligent standalone robot framework - Thorsten Scherler. March 26th @ 15:00

4.20 2 April 2007: Nutch 0.9 Released

The 0.9 release of Nutch is now available. This is the second release of Nutch based entirely on the underlying Hadoop platform. This release includes several critical bug fixes, as well as key speedups described in more detail at Sami Siren's blog. See list of changes made in this version. The release is available here.

4.21 24 September 2006: Nutch 0.8.1 Released

The 0.8.1 release of Nutch is now available. This is a maintenance release to 0.8 branch fixing many serous bugs found in version 0.8. See list of changes made in this version. The release is available here.

4.22 25 July 2006: Nutch 0.8 Released

The 0.8 release of Nutch is now available. This is the first release of Nutch based on hadoop architecure. See CHANGES.txt for list of changes made in this version. The release is available here.

4.23 31 March 2006: Nutch 0.7.2 Released

The 0.7.2 release of Nutch is now available. This is a bug fix release for 0.7 branch. See CHANGES.txt for details. The release is available here.

4.24 1 October 2005: Nutch 0.7.1 Released

The 0.7.1 release of Nutch is now available. This is a bug fix release. See CHANGES.txt for details. The release is available here.

4.25 17 August 2005: Nutch 0.7 Released

This is the first Nutch release as an Apache Lucene sub-project. See CHANGES.txt for details. The release is available here.

4.26 June 2005: Nutch graduates from Incubator

Nutch has now graduated from the Apache incubator, and is now a Subproject of Lucene.

4.27 January 2005: Nutch Joins Apache Incubator

Nutch is a two-year-old open source project, previously hosted at Sourceforge and backed by its own non-profit organization. The non-profit was founded in order to assign copyright, so that we could retain the right to change the license. We have now determined that the Apache license is the appropriate license for Nutch and no longer require the overhead of an independent non-profit organization. Nutch's board of directors and its developers were both polled and supported the move to the Apache foundation.

4.28 September 2004: Creative Commons launches Nutch-based Search

Creative Commons unveiled a beta version of its search engine, which scours the web for text, images, audio, and video free to re-use on certain terms a search refinement offered by no other company or organization.

See the Creative Commons Press Release for more details.

4.29 September 2004: Oregon State University switches to Nutch

Oregon State University is converting its searching infrastructure from Googletm to the open source project Nutch. The effort to replace the Googletm will realize significant cost savings for Oregon State University, while promoting both the Nutch Search Engine and transparency in search engine use and management.

For more details see the announcement by OSU's Open Source Lab.

4.30 Apache Nutch Trademark Attributions

Apache Nutch, Nutch, Apache, the Apache feather logo, and the Apache Nutch project logo are trademarks of The Apache Software Foundation.