Home

Welcome to the Apache Nutch Wiki

Please contribute your knowledge about Nutch here!

Or browse the open issues, open a new Jira ticket, or check the Nutch source code on git.

Table of Contents

What is Apache Nutch?

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

Nutch 1.x (**ACTIVE**): A well matured, production ready crawler. 1.x enables fine grained configuration, relying on Apache Hadoop data structures, which are great for batch processing.

Nutch 2.x (**INACTIVE**): 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. No more releases or bug fixes are anticipated for this codebase.

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. Additionally, 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.

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

Nutch Version Administration

- DownloadingNutch
- Current CommandLineOptions: Command line options for 1.X and 2.X
- JavaDocs – for the most recent Nutch-1.X release
- JavaDocs – of Nutch-1.X nightly builds
- JavaDocs – or the most recent Nutch-2.X release.

Tutorials

Nutch 1.X tutorial(s)

- NutchTutorial - How to configure Nutch to crawl in local mode and post to Apache Solr for search/index.
- QuickStartparseChecker - Quick start tutorial on how to use the ParseChecker tool to quickly scrape a website.
- https://wiki.apache.org/nutch/Nutch_1.X_RESTAPI - An overview of the entire Nutch 1.X REST API.
- Running Nutch on Tez - Covers using Apache Tez as the YARN execution engine

Other Tutorial(s)

- Focused Crawling with Nutch using Cosine Similarity, Naive Bayes or the Anthelion mechanisms.
- Hadoop Tutorial Nutch being based Hadoop, it helps to have a better understanding of Hadoop.
- Running Nutch in (pseudo) distributed mode - How to setup and run Nutch in Hadoop pseudo-distributed mode.
- RunNutchInEclipse - How to configure, build, crawl and debug Nutch within Eclipse
- Intranet Document Search - Index and search Microsoft Office, PDF etc. documents in a file system hierarchy with a Solr backend.
- Recrawling with Nutch - How to re-crawl with Nutch.
- Ajax-Solr Tutorial: Nutch - Quick and easy guide to getting a nice UI on top of your Nutch crawl data.
- AJAX/JavaScript Enabled Parsing with Apache Nutch and Selenium
- SetupProxyForNutch - using Tinyproxy on Ubuntu
- SetupNutchAndTor - Crawling .onion hidden services using Nutch behind Polipo HTTP Proxy
- CloudSearch - Step by step instructions on using Nutch with Cloudsearch, including pseudo distributed mode
- Webcast : running Apache Nutch on Elastic MapReduce

Configuration

- OverviewDeploymentConfigs ⚠️ :This full page requires a complete update to reflect recent Nutch releases: ⚠️
- NutchConfigurationFiles: An overview from Nutch developers.
- NutchPropertiesCompleteList: A fine grained account of all Nutch property configuration.
- HttpAuthenticationSchemes - How to enable Nutch to authenticate itself using NTLM, Basic or Digest authentication schemes.
- NonDefaultIntranetCrawlingOptions - Desirable options to add to your Nutch intranet crawling configuration.
- OptimizingCrawls - How to optimise your crawling/fetching speed with Nutch.
- ErrorMessages – What they mean and suggestions for getting rid of them. ⚠️ :This requires extensive updating to reflect recent Nutch releases. In addition the legacy indexing and searching material should be archived. ⚠️
- IndexStructure ⚠️ :This page needs a slight update to provide more information on plugins and the data they send to Solr for indexing: ⚠️
- IndexWriters: How to configure the index writers for indexing step.
- Exchanges: How to configure the exchanges for indexing step.
- Logging: Details of logging using slf4j and log4j2
- Metrics: A narrative on Nutch application metrics. It details which metrics are captured for which Nutch Job's within which Tasks.

General Information

- Nutch Website
- Features ⚠️ :TODO:This needs to be completely overhauled to reflect recent Nutch features. ⚠️
- Current Nutch Gotchas
- PublicServers running Nutch
- Presentations on Nutch
- Press Articles
- Evaluations of Search Quality
- Commercial Support & developers for hire
- Mailing Lists
- AcademicArticles that deal with Nutch
- FAQ
- HardwareRequirements
- NutchResources
- NutchScoring - The whats and wheres of Scoring implementations in Apache Nutch
- NutchFileFormats - Provides information on the Nutch file formats

Nutch Development

- Becoming a Nutch Developer - Start developing and contributing to Nutch.
- PluginCentral – How to write your own plugins and use other people's.
- InternalDocumentation – How Nutch works.
- Nutch Version Control
- UsingGit - a guide to leveraging Git and Nutch. Nutch's source code is no longer managed in Subversion, it's managed in Git.
- HowToContribute
- Committer's_Rules – Committers should follow these guidelines when deciding, which branch to use for committing the patches and when to commit.
- Release_HOWTO
- Nutch website repository (see README there how to edit and deploy changes to the website)
- Image_Search_Design
- StrategicGoals
- Getting_Started
- NutchMeetUps - Records of previous Nutch community meetup, hackathons, barcamps etc.
- Using Nutch as a Maven dependency
- GoogleSummerOfCode - An area dedicated to GSoC projects and student/mentor development/documentation sandbox.

- AdvancedAjaxInteraction - Discussion centered on enabling Nutch to not only fetch, but also interact with JavaScript
- WhiteListRobots - User guide for the new host robots.txt whitelist capability

Archive and Old Nutch Versions

- Archive and Legacy

How to edit this Wiki

This Wiki is a collaborative site, anyone can contribute and share. To help avoid spam the Nutch wiki is only editable by known accounts. If you would like to help out with the Nutch wiki, add a new page, or work on an existing one, please first create a wiki account by clicking on "Sign Up" or "Log in" if you already have an account.