

# Building your big data search stack with Apache Nutch 2.x

Lewis John McGibbney, ApacheCon NA 2014, Denver, CO



# Building your big data search stack with Apache Nutch 2.x

John McGibbney, ApacheCon NA 2014, Denver, CO



**Picture when we do the video shoot!**



**Personal Bio**

- 10+ years in the Big Data space
- 5+ years in the Nutch 2.x space
- 5+ years in the Hadoop space
- 5+ years in the Java space
- 5+ years in the Linux space
- 5+ years in the Open Source space
- 5+ years in the Apache space
- 5+ years in the ApacheCon space

**Apache Nutch 2.x**

- 5+ years in the Apache Nutch 2.x space
- 5+ years in the Apache Nutch 2.x space
- 5+ years in the Apache Nutch 2.x space
- 5+ years in the Apache Nutch 2.x space
- 5+ years in the Apache Nutch 2.x space

**Key Concepts**

- Controlled by configuration from which Hadoop Shadoos architecture enables plugin customization.
- Pluggable indexing architecture
- Like Hiera for storage abstraction.
- Maintains the concept of WebPage object containing fields we populate with data...

**Nutch Configuration**

- Nutch default.yml, Default options and values
- Multi-site.yml, Override, always takes precedence

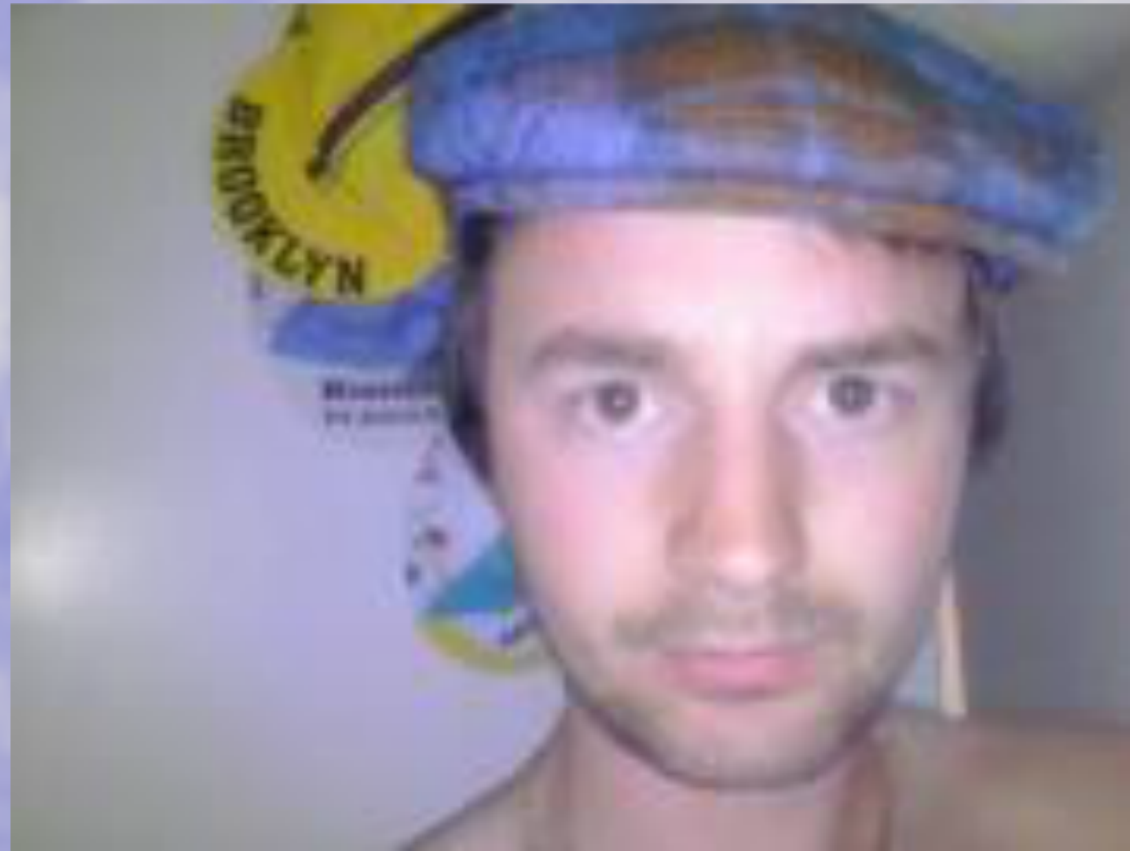
**Links look at some configuration options**

- <http://wiki.apache.org/nutch/nutch-configuration.html>
- <http://wiki.apache.org/nutch/nutch-configuration.html>
- <http://wiki.apache.org/nutch/nutch-configuration.html>

**Building the 2.x Project**

- <http://wiki.apache.org/nutch/nutch-configuration.html>
- <http://wiki.apache.org/nutch/nutch-configuration.html>
- <http://wiki.apache.org/nutch/nutch-configuration.html>
- <http://wiki.apache.org/nutch/nutch-configuration.html>
- <http://wiki.apache.org/nutch/nutch-configuration.html>

**Please allow me to introduce myself...**



Apache Member, V.P. Gora, PMC Nutch,  
PMC Any23, PMC OODT, TAC, PPMC  
Usergrid (incubating)

# Presentation Agenda

- Introduction to Apache Nutch (1.x + 2.x)
- Getting some code
- Key concepts
- Nutch configuration
- Setting it up!
- `./bin/nutch` (`./bin/crawl`)
- Sample local crawl
- Sample plugin
- Running on Hadoop

2.x

Get Ready 2.2.1  
<https://www.apache.org/hackathon>

Introduction  
<https://apache.org/2>

Running on Kubernetes Cluster

It's an article explaining that the configuration is changed (Docker) building the job file. There are 3 ways to do it: based on Docker, Kubernetes, along with the job file.

Things to come ..

- StartUp - getting with standard containers
- WebServer based Match Gull for 2x on top of REST API, Google Server of Code
- More plugins
- <https://issues.apache.org/jira/browse/BUCKET>
- Ask for features on the dev@ list

Questions?

A HUGE thank you for caring

Enjoy the rest of ApacheCon Hackathon on some Match-related projects!

[user@dev@match.apache.org](mailto:user@dev@match.apache.org)  
[levians@apache.org](mailto:levians@apache.org)  
[hector@apache.org](mailto:hector@apache.org)

Key Concepts

- Control by configuration from source (Hadoop) clusters architecture and external plug-in
- Plug-in for adding activities
- Managing the context of WebPage about containing WebPage plugins and data.

Match Configuration

- Set up Match on the WebPage cluster with custom configuration file.
- Let's look at some configuration options

<https://issues.apache.org/jira/browse/BUCKET>

DEV@BUCKET

Building on 2.x Project

My 2.x dependency management

Building BUCKET plugin for use with Apache 2.x

Let's look at some configuration options

<https://issues.apache.org/jira/browse/BUCKET>

Simple / Matchbook

This is the easiest everything we need to use a simple case. It's a full stack of other features

Simple / Matchbook

Building dependent Core Data

<https://issues.apache.org/jira/browse/BUCKET>

Example used

Example of application using 2.x and 2.x

<https://issues.apache.org/jira/browse/BUCKET>

Let's look at some configuration options

<https://issues.apache.org/jira/browse/BUCKET>

Let's try...

Sample plugin: Apache Any23

Loads of plugins are available to be used into WebPage, including: REST API, Foursquare, RSS, etc.

More about and relevant implementation for Any23

<https://issues.apache.org/jira/browse/BUCKET>

# Get Nutch 2.2.1

<http://nutch.apache.org/downloads.html>

## Introduction

<http://s.apache.org/eST>

# Key Concepts

- Controlled by configuration from which Hadoop shadows.
- Implements an extensible plugin architecture enabling customization.
- Pluggable indexing architecture
- Use Gora for storage abstraction.
- Maintains the concept of WebPage object containing fields we populate with data...

# Nutch Configuration

- nutch-default.xml: Default options and values
- nutch-site.xml: Override, always takes precedence

**Lets look at some configuration options**

<http://wiki.apache.org/nutch/NutchPropertiesCompleteList>

gora.properties

gora-\$backend\_mapping.xml



## Building the 2.x Project

*Ivy for dependency management*

*Editing \$NUTCH\_HOME/ivy/ivy.xml*

*ant -projecthelp - provides all ant targets*

*default target is 'ant runtime'*

*\$NUTCH\_HOME/runtime/local*

*run on one node*

*\$NUTCH\_HOME/runtime/deploy*

*run on Hadoop cluster via .job file*

**Scripts: ./bin/nutch**

*This script contains everything we need to run a sample crawl. It is also used to 'chain' together arguments for continuous crawls.*

**Scripts: ./bin/crawl**

*Replaces deprecated Crawl class.*

***crawl <seedDir> <crawlID> <solrURL> <numberOfRounds>***

# Example crawl

- Sequence of operations same in 1.X and 2.x
  - Inject: populates WebTable/CrawlDB from seed list
  - Generate: Selects URLs to fetch and assigns a batchId for this grouping.
  - Fetch: Fetches URL's with the batchId
  - Parse: Parses URL's with the batchId
  - UpdateDB: Updates the WebTable/CrawlDB with new URL's (outlinks, inlinks) and URL status
  - InvertLinks: Builds a WebGraph (Apache Giraph)
  - Index: Send docs to [Solr | ES | SolrCloud | MongoDBc| CSV writer]

**Lets try...**

# Sample plugin: Apache Any23

Loads of implicit markup/structure embedded into WebPage's. Microdata, Rel-Tag's, Microformat's, RDFa, JSON-LD, etc.

Write parser and indexing implementation for Nutch.



## Running on Hadoop Cluster

It's as simple as ensuring that the configuration is completed BEFORE building the .job file.

Once .job is built, it can be found in \$NUTCH\_HOME/deploy along with the bin scripts

A Nutch job, any job, is submitted to the Hadoop Jobtracker. It knows where the cluster is and what config is to be used. The bin/nutch script does little more than submitting the job to the tracker with job specific parameters.

## Things to come...

- SiteMap parsing with crawler-commons
- Wicket-based Nutch GUI for 2.x on top of REST API. Google Summer of Code.
- More plugins
- <https://issues.apache.org/jira/browse/NUTCH>
- Ask for features on the dev@ list

**Questions?**



**A HUGE thank you for coming**

**Enjoy the rest of ApacheCon**

**Hackathon on some Nutch-  
related queries?**

**user@ dev@**

**nutch.apache.org**

**lewismc@apache.org**

**@hectorMcSpector**