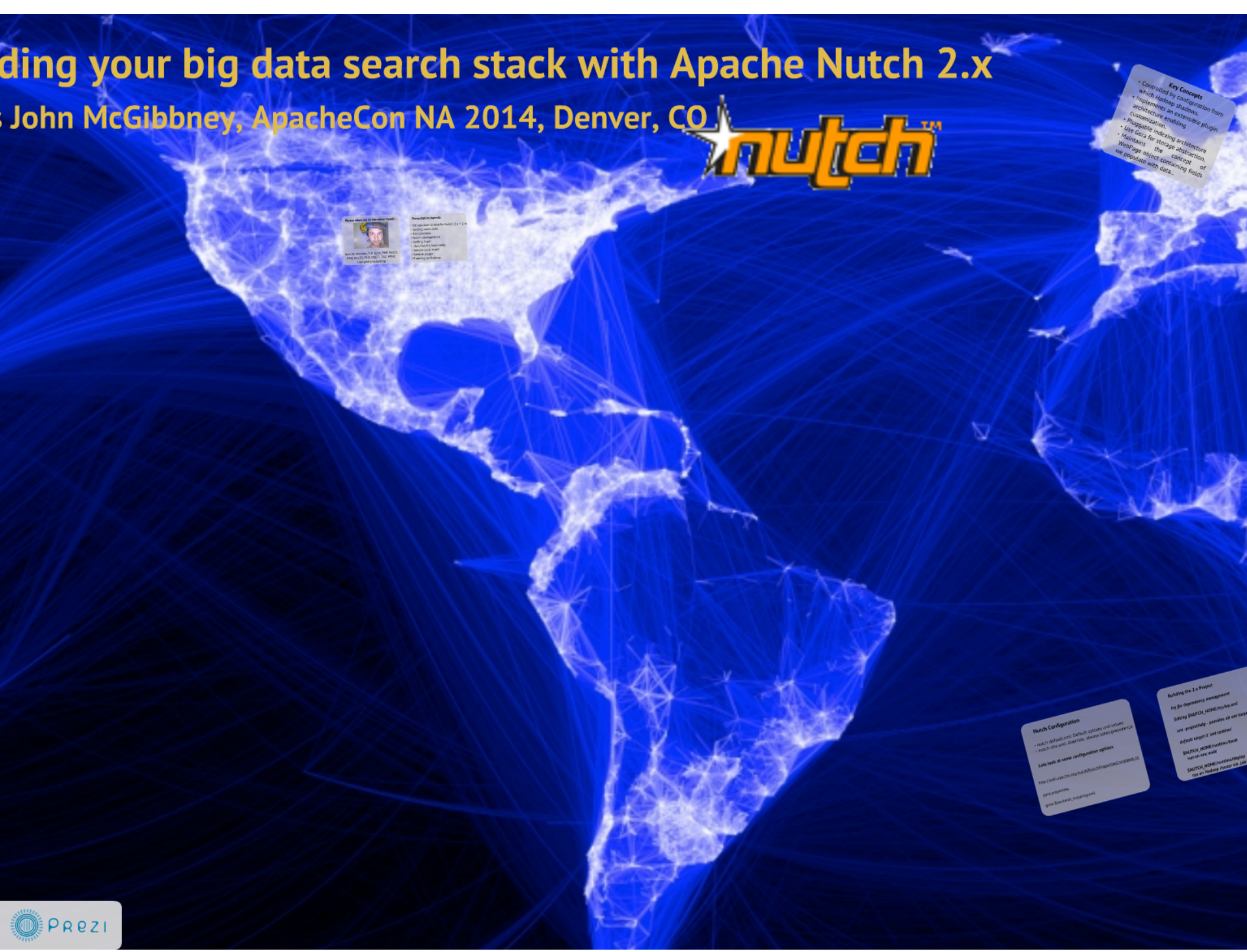
Building your big data search stack with Apache Nutch 2.x

Lewis John McGibbney, ApacheCon NA 2014, Denver, CO

ding your big data search stack with Apache Nutch 2.x

s John McGibbney, ApacheCon NA 2014, Denver, CO



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


Get Nutch 2.2.1
http://nutch.staging.apache.org/bin/nutch.html
Introduction
http://s.apache.org/aST
Running on Hadoop Cluster
Things to come...
• SiteMap parsing with crawler-commons
• Widget-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
2.x
Key Concepts
Hutch Configurations
Building the 2.x Project
Example crawl
Lets try...
Sample plugin: Apache Any23

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 it '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 litte 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