GoogleSummerOfCode SecurityLayer

Title :		GSoC 2016 Proposal		
Issue :		NUTCH-1756 - Security layer for NutchServer		
<ac:structured-macro ac:name="unmigrated-wiki-markup" ac:schema-version="1" ac:macro-id="01696137-3ba3-4889-add0-5ea401d1053f"><ac:plain-text-body><![CDATA[	**Student :**	Furkan KAMACI - furkankamaci [at] gmail.com		]]></ac:plain-text-body></ac:structured-macro>
Mentor :		Lewis John McGibbney		

Abstract

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:
Nutch 1.x and Nutch 2.x. This proposal aims to develop a security layer for Nutch 2.x.

Additional Info

http://www.furkankamaci.com/

Furkan KAMACI

Istanbul Technical University

Graduate School of Science, Engineering and Technology

Computer Engineering

Istanbul, Turkey

1. Introduction

Apache Nutch comprises two codebases, namely:

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.

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 flexible model/stack for storing everything (fetch time, status, content, parsed text, outlinks, inlinks, etc.) into a number of NoSQL storage solutions.

2. Definition of the Problem

Nutch 2.x has a REST API and web application but it doesn't have a security layer on it. A security layer should be implemented which covers security functionality (authentication, authorization), different authentication mechanisms , documentation and refactoring existing code. This project will therefore propose the design, development and implementation of the security agenda as described above. This work will be specifically applicable to the Nutch 2.X codebase.

3. Proposed Method

A. Background

There has been implemented an API which lets to interact with Nutch via REST API. Administration and configuration tasks can be done via this API.

This proposal offers a comprehensive security layer under NUTCH-1756. Existing code should be re-factored, security layer should be added and a documentation should be done programmatically (i.e. Miredot, Swagger).

B. Suggested Steps for Proposed Method

1) Authentication/Authorization Implementation

2) API Documentation Generator Implementation

4. Schedule & Timeline

Suggested schedule and timeline is as follows:

1) _Analyzing the Problem (1 Week - 30 May 2016)_
a) Problem will be analyzed with more detail.

2) _Authentication Implementation (5 weeks - 4 July 2016)_

a) HTTP basic authentication

b) HTTP digest authentication

c) SSL support

d) Kerberos authentication

3) _Authorization Implementation (3 weeks - 25 July 2016)_
a) Authorization will be implemented

4) _API documentation (1 week - 1 August 2016)_
a) API documentation implementation

5) _Test (1 week - 8 August 2016)_
a) Implementation tests will be written and run.

6) _Documentation (1 week - 15 August 2016)_
a) Documentation will be prepared.

7) _Tidy up (1 week - 23 August 2016)_
a) Tidy up will be done.

References

[1] https://issues.apache.org/jira/browse/NUTCH-1756

[2] https://wiki.apache.org/nutch/NutchRESTAPI

[3] https://issues.apache.org/jira/browse/GORA-386

[4] https://issues.apache.org/jira/browse/NUTCH-2243

[5] https://issues.apache.org/jira/browse/NUTCH-2022

[6] https://github.com/apache/nutch

[7] https://en.wikipedia.org/wiki/Apache_Nutch

[8] https://github.com/apache/gora

[9] http://en.wikipedia.org/wiki/Apache_Gora

[10] http://gora.apache.org/current/tutorial.html#introduction

Reports

Weekly Reports

Midterm Report

Final Report

Development Repo

kamaci/nutch

List of Merged Commits

apache/nutch/commits/2.x?author=kamaci