GSoC Midterm Report

Security Layer for NutchServer (NUTCH-1756) Midterm Report

Furkan KAMACI[1]

[1] *Istanbul Technical University*

Abstract—This reports explains the progress up to midterm evaluations of GSoC 2016.

Index Terms— Apache Nutch, GSoC

I. INTRODUCTION

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: 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.

Nutch 2.x has a REST API but it doesn't have a security layer on it. It should be implemented a security layer which covers security functionality (authentication, authorization), different authentication mechanisms , documentation and refactoring existing code.

There has been implemented an API which lets to interact with Nutch via REST API. Administration, configuration and database tasks can be done via this API and NUTCH-1756 offers a comprehensive security layer for it.

During GSoC period, I've blogged at my personal website (http://furkankamaci.com/) and created a fork from Apache Nutch's master branch and worked on it: https://github.com/kamaci/nutch

This report explains what is done up to now and what will be the next steps. Currently, project progress is ahead from the proposed timeline. Here is the tag of this report:

Project Name	Security Layer for NutchServer (NUTCH-1756)
Project URL	https://issues.apache.org/jira/browse/NUTCH-1756
Report #	10
Report Compiled By	Furkan KAMACI
Report Date	27.06.2016

II. PROPOSED 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.

III. COMMUNITY BONDING

At community bonding period, Apache Nutch documentation is read and Apache Nutch source code is checked to be more familiar with project. Examples are run with Apache Nutch. Followed mail list and answered to asked questions. Organized/created Jira issues related to GSoC task. Attended ApacheCon and Apache: Big Data in Vancouver. Exchanged opinions about Nutch with community.

Jira issues worked on:

https://issues.apache.org/jira/browse/NUTCH-2089
https://issues.apache.org/jira/browse/NUTCH-2122
https://issues.apache.org/jira/browse/NUTCH-2222
https://issues.apache.org/jira/browse/NUTCH-2263
https://issues.apache.org/jira/browse/NUTCH-2264
https://issues.apache.org/jira/browse/NUTCH-2265

Merged Pull requests:

https://issues.apache.org/jira/browse/NUTCH-2122
https://issues.apache.org/jira/browse/NUTCH-2263

I've also blogged at my personal website about my acceptance of GSoC (http://furkankamaci.com/gsoc-2016-acceptance-for-apache-nutch/)

IV. CODING PERIOD

At coding period, I've organized source code and started to implement authentication mechanisms to NutchServer. Proposal documentation can be check from here: https://wiki.apache.org/nutch/GoogleSummerOfCode/SecurityLayer and weekly reports from here: https://wiki.apache.org/nutch/GoogleSummerOfCode/SecurityLayer/WeeklyReports

Jira issues worked on:

https://issues.apache.org/jira/browse/NUTCH-2266
https://issues.apache.org/jira/browse/NUTCH-2243
https://issues.apache.org/jira/browse/NUTCH-2284
https://issues.apache.org/jira/browse/NUTCH-2285
https://issues.apache.org/jira/browse/NUTCH-2288
https://issues.apache.org/jira/browse/NUTCH-2289

Whole code can be checked from code base: https://github.com/kamaci/nutch

V. ALIGNMENT WITH TIMELINE

Project progress is aligned with timeline. I've create pull requests for all tasks which are related to coding period.

VI. CONCLUSION & NEXT STEPS

I'm planning to finish SSL support and Kerberos authentication for NutchServer after midterm I've switched Kerberos authentication implementation and API Documentation tasks at timeline to go fast at implementations.

I'll implement authorization mechanism for security after security implementations. I'll finalize my task with tests and documentation. My previous reports can be found here:
https://wiki.apache.org/nutch/GoogleSummerOfCode/SecurityLayer/WeeklyReports

VII. REFERENCES

[1] https://issues.apache.org/jira/browse/NUTCH-1756
[2] https://wiki.apache.org/nutch/NutchRESTAPI
[3] https://issues.apache.org/jira/browse/NUTCH-2243
[4] https://issues.apache.org/jira/browse/NUTCH-2022
[5] https://github.com/apache/nutch
[6] https://en.wikipedia.org/wiki/Apache_Nutch
[7] https://issues.apache.org/jira/browse/GORA-386
[8] https://github.com/apache/gora
[9] http://en.wikipedia.org/wiki/Apache_Gora
[10] http://furkankamaci.com/
[11] https://wiki.apache.org/nutch/GoogleSummerOfCode
[12] https://wiki.apache.org/nutch/GoogleSummerOfCode/SecurityLayer/WeeklyReports
[13] https://issues.apache.org/jira/browse/NUTCH-2089
[14] https://issues.apache.org/jira/browse/NUTCH-2122
[15] https://issues.apache.org/jira/browse/NUTCH-2222
[16] https://issues.apache.org/jira/browse/NUTCH-2263
[17] https://issues.apache.org/jira/browse/NUTCH-2264
[18] https://issues.apache.org/jira/browse/NUTCH-2265
[19] https://issues.apache.org/jira/browse/NUTCH-2266
[20] https://issues.apache.org/jira/browse/NUTCH-2243
[21] https://issues.apache.org/jira/browse/NUTCH-2284
[22] https://issues.apache.org/jira/browse/NUTCH-2285
[23] https://issues.apache.org/jira/browse/NUTCH-2288
[24] https://issues.apache.org/jira/browse/NUTCH-2289