



**UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH**

Barcelona School of Informatics (FIB)
Master in Information Technology

Improving Acquia Search and the Apache Solr Search Integration Drupal module

by

Nick VEENHOF

Supervisor: B. Chris BROOKINS

Co-Supervisor: Dr. Ir. Peter WOLANIN

Tutor/Professor: Prof. Dr. Ir. Carles FARRÉ TOST

Academic Year 2011–2012

Preface

This report is a reflection of 6 months work within the Acquia company and involved a high learning curve regarding the Apache Solr software and Drupal (Web Framework/CMS). This report will explain you the specifics of each of these. Also it will explain the work that has been done for Acquia Search (Solr as a Service).

Nick Veenhof, February 2012

License Statement

This work is licensed under the NonCommercialAcknowledgementWithoutDerivedWork license from Creative Commons. This license, which is the most restrictive from Creative Commons, doesn't allow derived works, and authorizes, in all cases, the reproduction, distribution and public communication of the work as long as its author is mentioned and as no commercial use is done.

Nick Veenhof, February 2012

Improving Acquia search and the ApacheSolr Module

by

Nick VEENHOF

Master Thesis in order to acquire a Master in Information Technology

Academic Year 2011–2012

Supervisor: B. Chris BROOKINS

Co-Supervisor: Dr. Ir. Peter WOLANIN

Tutor/Professor: Prof. Dr. Ir. Carles FARRÉ TOST

Barcelona School of Informatics (FIB)

Master in Information Technology

BarcelonaTech

Abstract

This report is a reflection of 6 months work within the Acquia company and involved a high learning curve regarding the Apache Solr software and Drupal (Web Framework/CMS). This report will explain you the specifics of each of these. Also it will explain the work that has been done for Acquia Search (Solr as a Service).

Keywords

Drupal, Apache Solr, Lucene, Acquia, Acquia Search, Acquia Network, Search Technology

Contents

1	Introduction	1
1.1	Community	1
2	Objectives	2
3	Description and Terms	4
3.1	Acquia	4
3.2	Drupal	4
3.3	Apache Solr	4
3.4	Personal History	5
4	Analysis	6
4.1	Apache Solr	6
4.2	Apache Solr 1.4	6
4.3	Apache Solr 3.x	6
4.4	Apachesolr module for Drupal 6 version 6.x-2.x	6
4.5	Apachesolr module for Drupal 7 version 7.x-1.x-beta5	6
4.6	Facetapi module for Drupal 7 version 7.x-1.x	6
4.7	Acquia Search for Drupal 6 and 7	6
5	Requirements	7
5.1	Apachesolr module for Drupal 6 version 6.x-3.x	7
5.2	Apachesolr module for Drupal 7 version 7.x-1.0	7
5.3	Functional requirements	7
5.3.1	Search Environments	7
5.3.2	Search pages	7

5.3.3	Query Object	7
5.3.4	Apaches Solr Document	7
5.3.5	Entity layer	7
5.4	Non-Functional Requirements	7
5.4.1	User interface	7
5.4.2	Usability	7
5.4.3	Performance	7
5.4.4	Security	7
5.4.5	Legal	7
6	Implementation	8
6.1	Apachesolr module for Drupal 6 version 6.x-3.x	9
6.2	Apachesolr module for Drupal 7 version 7.x-1.0	9
6.3	Functional requirements	9
6.3.1	Search Environments	9
6.3.2	Search pages	9
6.3.3	Query Object	9
6.3.4	Apaches Solr Document	9
6.3.5	Entity layer	9
6.4	Non-Functional Requirements	9
6.4.1	User interface	9
6.4.2	Usability	9
6.4.3	Performance	9
6.4.4	Security	9
6.4.5	Legal	9
7	Related Work	10
7.1	Search Appliances	10
7.1.1	Elastic Cloud	10
7.1.2	Sphinx	10
7.1.3	Some Other	10
7.2	Drupal Search Solutions	10
7.2.1	Search API	10

7.2.2	Drupal Lucene API	10
7.2.3	Google Search Appliance	10
8	Conclusions	11
8.1	Overview of work	11
8.2	Reflection on Apache Solr	11
8.3	Reflection on Drupal 6 and Drupal 7 in regards to search integration	11
8.4	Future Work	11
8.4.1	Apache Solr Search Integration	11
8.4.2	Acquia Search	11
9	Acknowledgements	12

Chapter 1

Introduction

1.1 Community

Before you, the reader, will dive into deep specifics of Apache Solr and Drupal there are a couple things you should know in advance. Firstly this work is the result of many hours hard work and not only from me but from a complete community. These communities have changed the way I look at software. In programming classes in university a student is taught a different way of designing software, namely that the control is fully his. I had numerous courses going from basic Java to Advanced Web Technologies to IBM rational rose project management. And while I personally have learned a lot from these courses it has always been my understanding that what is taught is just another way of creating solutions for existing problems.

Working in a community is, similarly, another way of creating solutions for a set of existing problems but involves a different way of making decisions and looking at software. It is great if your code can be shared and used across thousands of people and be corrected by those same group of people. While code will never be perfect, different people have used the same codebase to solve existing problems and they have been saving time and resources. I would like to thank Acquia for being very clear in regards of the benefits of sharing knowledge because it is also my personal belief that this is important for everyone to grow and to contribute to do good.

Chapter 2

Objectives

The student will be asked to be on-site at the headquarters in Boston for a couple of weeks in order to meet the team and to get to know the company in order to gather all the information necessary to reach the objectives set further in this document. He will follow and join meetings to obtain a good insight in the requirements of the project and learn how to work under a Agile/Scrum based development methodology.

Being responsible for improving the Drupal Apache Solr search integration [1] project and the Acquia Search service is the common theme of the whole internship. This means adding additional features, keeping high quality and create upgrades/updates. The objective will be to exploit as much as possible from the latest Apache Solr 3.x branch while merging and keeping the software compatible with Apache Solr 1.4.

Communication will be a crucial part in order to succeed. The project has a worldwide scope, reaching out to more companies than just Acquia. This means he will have to be able to consult and make decisions after talking with a lot of end-users and other stakeholders. English will be the language of choice. This can happen by means of chat (IRC), on the Drupal community website [14] , giving presentations in conferences or taking interviews. Finally the ability to work remotely, over a large distance and in a team, is an important skill to acquire.

In the beginning of the internship this was itemized in a list.

- Bring Facet Api [2] for Drupal 7 to a stable Release Candidate.
- Create a multisite module to search between 2 or more Drupal sites.

-
- Update the Acquia Search service to the latest stable Apache Solr version. Upgrade the custom java code that was written to be able to authenticate customers.
 - Backport to a new Drupal 6 branch all the new features that have been programmed into the Drupal 7 version of the Apache Solr Search Integration Module. This includes the backporting of the multisite module.
 - Achieve mastery of the agile/scrum process, the open source software engineering methods, and the team communication processes used by Acquia.
 - Empower the community to use the Apache Solr Search Integration project by means of Presentations, Blog posts and other interactions with community members.

Chapter 3

Description and Terms

3.1 Acquia

Acquia is a commercial open source software company providing products, services, and technical support for the open source Drupal social publishing system and was founded by Dries Buytaert, the original creator and project lead of the Drupal project. With over two million downloads since inception, Drupal is used by web developers worldwide to build sophisticated community websites. Diverse organizations use Drupal as their core social publishing system for external facing websites and internal collaboration applications.

Acquia Search is a plug-and-play service within the Acquia Network [16], built on Apache Solr [4] and available for any Drupal 6 or Drupal 7 site. Acquia Search offers site visitors faceted search navigation and content recommendations to help them find valuable information faster. It is a fully redundant, high performance cloud service, with no software to install or servers to manage.

3.2 Drupal

Information about Drupal + History etc etc..

3.3 Apache Solr

Information about Apache Solr + History etc etc..

3.4 Personal History

My story with Drupal starts in the beginning of 2007. During my Bachelor I was asked, together with a group of people, to make a community site in Drupal to see what it was capable of. Now, almost 5 years later I still don't fully know what it is capable of since it keeps evolving and growing.

- Kaho - Project (Reference?) - Worked 1 year at Krimson (Reference, maybe recommendation?) where my interest in Solr and Search started and created my first real module (apach-esolr_ubercart for Drupal 6)

- Worked at Ateneatech (Officially using UPC contracts) - Worked at AtSistemas as one of the reference engineers for a huge Solr and Drupal powered website. (<http://www.elsevier.es>) During my studies at UPC I kept following the Drupal development and made numerous discussions with people and teachers on how software engineering should look at these projects. In the course of Advanced Web Technologies I presented Drupal a couple of times from different angles. (References)

There was only 1 logical step possible and that was doing my Master thesis with Acquia. During my Erasmus period in Portugal I attended a Drupal Camp (And even presented some Solr technologies, see reference) and I've met Robert Douglas, one of the creators of the Apache Solr Integration Project for Drupal and approached him with the question if I would be able to do my internship with Acquia. After a long process with the UPC and with Acquia everything was set and the pieces of the puzzle fell in place.

Chapter 4

Analysis

4.1 Apache Solr

4.2 Apache Solr 1.4

4.3 Apache Solr 3.x

4.4 Apachesolr module for Drupal 6 version 6.x-2.x

4.5 Apachesolr module for Drupal 7 version 7.x-1.x-beta5

4.6 Facetapi module for Drupal 7 version 7.x-1.x

4.7 Acquia Search for Drupal 6 and 7

Chapter 5

Requirements

5.1 Apachesolr module for Drupal 6 version 6.x-3.x

5.2 Apachesolr module for Drupal 7 version 7.x-1.0

5.3 Functional requirements

5.3.1 Search Environments

5.3.2 Search pages

5.3.3 Query Object

5.3.4 Apaches Solr Document

5.3.5 Entity layer

5.4 Non-Functional Requirements

5.4.1 User interface

5.4.2 Usability

5.4.3 Performance

5.4.4 Security

5.4.5 Legal

Chapter 6

Implementation

I started my internship September 22nd in Boston. While being on-site and I learned how Acquia manages processes and works together with the community in order to reach business goals. Also watching them work with a lot of remote employees was already a very valuable lesson. More about this experience can be read in the article [9] I wrote about this. In addition to this I have helped with the creation of some new modules such as Facet Slider [12] and Apachesolr sort UI [13].

As for addressing the public on this subject, I have recently given a presentation ‘Drupal Search’ [6] explaining to more than 60 attendees what was done with the module and where it was heading to in Belgium. Lots of open communication [10] has happened within the community in the Apache Solr Issue Queue. [11] In total I have given 4 presentations with a combined total of more then 200 attendants. Since my involvement with the project there are about 2500 websites using the Drupal 7 version of the module and about 10000 users in total using the module for Drupal 6 and Drupal 7 combined.

I’ve contributed several blog posts about this topic and this internship

- Presentation about Drupal Search for the DUG group in November 2011. [10]
- Simple guide to install Apache Solr 3.x for Drupal 7 on a unix machine [7]
- Adding a custom plugin to the Apache Solr Project [8]
- A Story of an intern at Acquia [9]

Objectives 1, 2 and 3 are in progress and are nearing its completion status. Objective 4, updating the Acquia Search service to the latest stable Apache Solr version, made great progress

and is currently being tested by a client of Acquia that required this change. Every day there is a daily call with Peter Wolanin to keep the daily objectives clear and to clear out any issues that could block progress.

6.1 Apachesolr module for Drupal 6 version 6.x-3.x

6.2 Apachesolr module for Drupal 7 version 7.x-1.0

6.3 Functional requirements

6.3.1 Search Environments

6.3.2 Search pages

6.3.3 Query Object

6.3.4 Apaches Solr Document

6.3.5 Entity layer

6.4 Non-Functional Requirements

6.4.1 User interface

6.4.2 Usability

6.4.3 Performance

6.4.4 Security

6.4.5 Legal

Chapter 7

Related Work

7.1 Search Appliances

7.1.1 Elastic Cloud

7.1.2 Sphinx

7.1.3 Some Other

7.2 Drupal Search Solutions

7.2.1 Search API

7.2.2 Drupal Lucene API

7.2.3 Google Search Appliance

Chapter 8

Conclusions

8.1 Overview of work

8.2 Reflection on Apache Solr

8.3 Reflection on Drupal 6 and Drupal 7 in regards to search integration

8.4 Future Work

8.4.1 Apache Solr Search Integration

8.4.2 Acquia Search

Chapter 9

Acknowledgements

Bibliography

- [1] Apache Solr 3 Enterprise Search Server RAW Book & eBook | Packt Publishing Technical & IT Book and eBook Store
<http://www.packtpub.com/apache-solr-3-enterprise-search-server/book>
- [2] <http://people.apache.org/~hossman/apachecon2009us/apache-solr-out-of-the-box.pdf> people.apache.org/~hossman/apachecon2009us/apache-solr-out-of-the-box.pdf
- [3] [solr-user] Limit number of docs that can be indexed (security) - Search by Lucid Imagination
http://www.lucidimagination.com/search/document/23d645855e8417a1/limit_number_of_docs_that
- [4] <https://acquia.com/sites/default/files/blog/DCChicago2011SolrChopsv3.pdf>
- [5] FieldCollapsing Solr Wiki
<http://wiki.apache.org/solr/FieldCollapsing>
- [6] Issues for Apache Solr Search Integration | drupal.org
<http://drupal.org/project/issues/apachesolr?text=&status=Open&priorities=All&categories=All&ver>
- [7] CoreAdmin Solr Wiki
<http://wiki.apache.org/solr/CoreAdmin>
- [8] Double ellipses on search snippets. | drupal.org
<http://drupal.org/node/1264786>
- [9] Dynamic queries | drupal.org
<http://drupal.org/node/310075>
- [10] [#SOLR-232] let Solr set request headers (for logging) - ASF JIRA
<https://issues.apache.org/jira/browse/SOLR-232>

-
- [11] [#SOLR-2452] rewrite solr build system - ASF JIRA
<https://issues.apache.org/jira/browse/SOLR-2452>
- [12] Compiling with Ant – genoviz
[http://sourceforge.net/apps/trac/genoviz/wiki/Compiling with Ant](http://sourceforge.net/apps/trac/genoviz/wiki/Compiling%20with%20Ant)
- [13] Update war file for report reader - Attias
http://attias.myftp.org/attias/index.php/Update_war_file_for_report_reader
- [14] solr at master from distilledmedia/munin-plugins GitHub
<https://github.com/distilledmedia/munin-plugins/tree/master/solr>
- [15] Date-boosting Solr / Drupal search results | Metal Toad Media
<http://www.metaltoad.com/blog/date-boosting-solr-drupal-search-results>
- [16] Major Solr 4 Highlights | Javalobby
<http://java.dzone.com/videos/major-solr-4-highlights>
- [17] Solr Black Belt Preconference
<http://www.slideshare.net/erikhatcher/solrblackbeltpreconference>
- [18] Some tips for Solr tuning - Vizrt forum
<http://forum.vizrt.com/showthread.php?t=5177>
- [19] Getting started faster with LucidWorks for Solr
<http://www.slideshare.net/LucidImagination/improving-findability>
- [20] Spatial Indexes: Solr — Derick Rethans
<http://derickrethans.nl/spatial-indexes-solr.html>
- [21] Using Apache Access Logs with JMeter
<http://minaret.biz/tips/jmeter.html>
- [22] Jmeter used to playback Apache access logs to generate live-like server load | artur.ejsmont.org
<http://artur.ejsmont.org/blog/content/jmeter-used-to-playback-apache-access-logs-to-generate-live-like-server-load>
- [23] Drupal Patching, Committing, and Squashing with Git | RandyFay.com
<http://randyfay.com/node/97>

- [24] svn get last commit message « Alec's Web Log
<http://www.alecjacobson.com/weblog/?p=2042>
- [25] GitX - See It
<http://gitx.frim.nl/seeit.html>
- [26] Add SSH key to Server
<http://oreilly.com/pub/h/66>
- [27] Maintaining patch series with Stacked GIT: a walk-through | drupal.org
<http://drupal.org/node/337933>
- [28] Git Best Practices: Upgrading the Patch Process | Lullabot
<http://www.lullabot.com/articles/git-best-practices-upgrading-patch-process>
- [29] Issues for Facet API | drupal.org
<http://drupal.org/project/issues/facetapi?categories=All>
- [30] Issues for Apache Solr Search Integration | drupal.org
<http://drupal.org/project/issues/apachesolr>
- [31] block_admin_display_form | Drupal API
http://api.drupal.org/api/drupal/modules/block/block.admin.inc/function/block_admin_display_form
- [32] Allow for vocab level facets | drupal.org
<http://drupal.org/node/1163880>
- [33] help | dgo.to
<http://dgo.to/>
- [34] All-day events missing or wrong in ical feed | drupal.org
<http://drupal.org/node/1284170>
- [35] Issues for Drupal core | drupal.org
<http://drupal.org/project/issues/drupal?status=1&categories=bug&version=7.x>
- [36] [http://bxl2011.drupaldays.org/sites/default/files/Search API Presentation.pdf](http://bxl2011.drupaldays.org/sites/default/files/Search_API_Presentation.pdf)
- [37] Interface text | drupal.org
<http://drupal.org/node/604342>

- [38] Backpack: Debugging Drupal
<https://ratatosk.backpackit.com/pub/1836982-debugging-drupal>
- [39] Using apachebench (ab) with Drupal 7 to load test site with authenticated users | Midwestern Mac, LLC
<http://www.midwesternmac.com/blogs/jeff-geerling/using-apachebench-ab-drupal-7>
- [40] Apache Bench (ab) | drupal.org
<http://drupal.org/node/659974>
- [41] Supercolliding a PHP array
<http://nikic.github.com/2011/12/28/Supercolliding-a-PHP-array.html>
- [42] Awesome Testing Party Cheat Sheet
<http://dmitrize.com/Awesome-Testing-Party-Cheat-Sheet.html>
- [43] VAT | Dries Buytaert
<http://buytaert.net/album/blog/vat>
- [44] Miscellaneous Simpletest Tips | drupal.org
<http://drupal.org/node/30011>
- [45] Apache Solr search integration module | drupal.org
<http://drupal.org/project/apachesolr>
- [46] Facet Api Module | drupal.org
<http://drupal.org/project/facetapi>
- [47] Acquia Search product information
<http://acquia.com/productsservices/acquia-search>
- [48] Apache Solr project page
<http://lucene.apache.org/solr/>
- [49] Drupal User Group presentation about Search in Drupal 7
<http://drupal.be/evenement/dug-over-search-in-drupal-7>
- [50] Drupal User Group presentation slides about Search in Drupal 7
<http://nickveenhof.be/blog/drupal-search-and-solr-dug-november-2011>

- [51] Simple Guide to install Apache Solr 3.x Drupal 7
<http://nickveenhof.be/blog/simple-guide-install-apache-solr-3x-drupal-7>
- [52] Adding custom plugins to Apache Solr
<http://nickveenhof.be/blog/adding-custom-plugin-solr>
- [53] Story of the first few weeks as an intern at Acquia
<http://nickveenhof.be/blog/story-intern-acquia>
- [54] Drupal.org user profile of Nick_vh (Nick Veenhof)
<http://drupal.org/user/122682/track>
- [55] Issue queue of Apache Solr search integration module
<http://drupal.org/project/issues/apachesolr>
- [56] Facet Api Slider module project page
http://drupal.org/project/facetapi_slider
- [57] Apache Solr Sort UI project page
http://drupal.org/project/apachesolr_sort
- [58] Drupal 7 search presentation in Toulouse
<http://toulouse2011.drupalcamp.fr/en>
- [59] Acquia Network product information
<http://www.acquia.com/products-services/acquia-network>

List of Figures

List of Tables