Evaluating the Effects of Architectural Documentation: A Case Study of a Large Scale Open Source Project

Presented by Vagner Clementino

Department of Computer Science
Federal University of Minas Gerais (UFMG)
Software Archicteture - 2016

Outline

About the Paper

Context

Motivation

Proposed Work

Study Design

Results

Conclusions

Threats to Validity

About the Paper

- Who:
  - Kazman, Rick (Senior Member IEEE)
  - Goldenson, Dennis (Senior Member IEEE)
  - Monarch, Ira (Software Engineering Institute)
  - Nichols, William (Senior Member IEEE)
  - Valetto, Giuseppe (Member IEEE)
- When: 2015
- Where:
  - IEEE Transactions on Software Engineering (Volume:42 , Issue: 3)

Contribution in Open Source System (OSS)

- Sustaining large Open Source System (OSS) requires continuos recruitment new participants.
- The number of contributors can be used as metric of project sucess.

Objective of Architectural Documentation

- Architectural documentation is believed to serve three major purposes [1]:
  1. providing a means of introducing new project members to the system
  2. serving as a vehicle for communication among stakeholders
  3. being the basis for system analysis and construction

Architectural Documentation in OSS

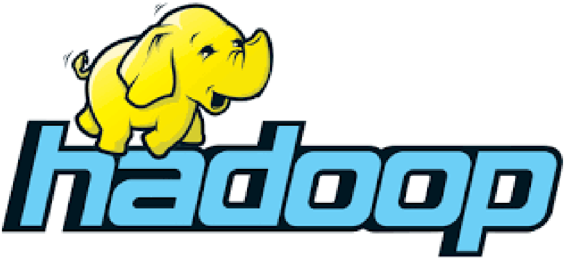
- A lack of architectural documentation might inhibit new participants since large amounts of project knowledge are unavailable to newcomers.
- In 5.4 percent of open source projects have any software architecture documentation [2]

Proposed Work

- This is a multitrait, multimethod analysis of the effects of introducing architectural documentation into a substantial open source project
- The objective is to investigate if and how architecture documentation adds value to a software project.

Open Source System Selection

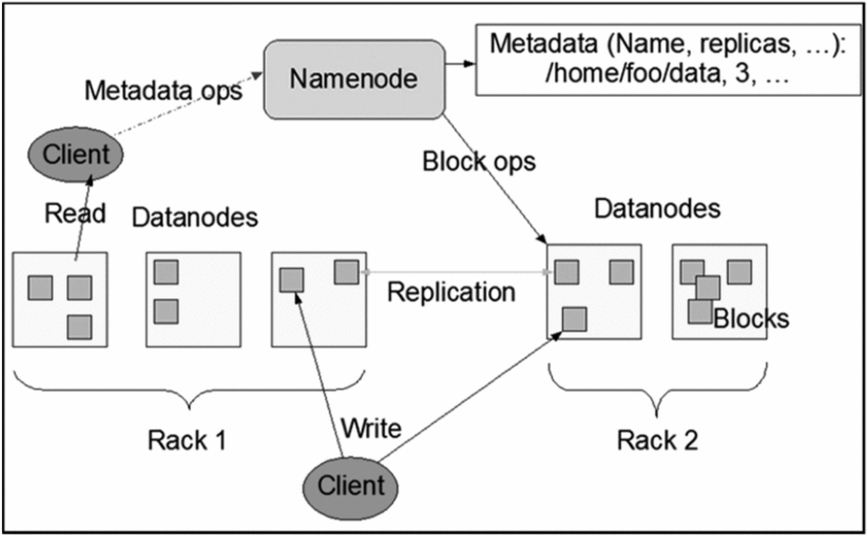
- The study used Hadoop Distributed File System (HDFS)[3].
- The Apache Hadoop project is widely used by large companies such as Yahoo!, eBay, Facebook, and others.



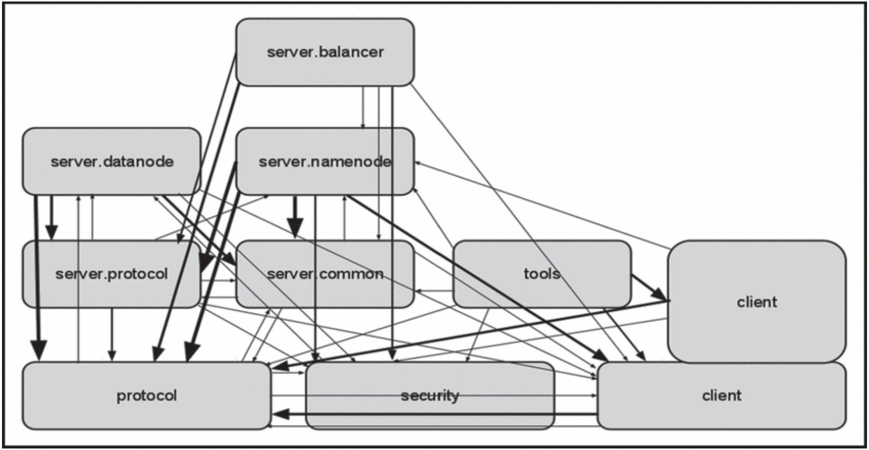
Documenting the HDFS Architecture

- The architecture documentation captured the main abstractions employed in HDFS
- It is also to connect those abstractions to the code (files) that developers work on every day.

HDFS Run-Time Concepts



Documented Module Relationships in HDFS



Research Question

RQ 1.1 Was the architecture document read and if so, how much and how did it change?

RQ 1.2 Was the architecture document referred to by the project Contributors and Committers?

Research Question

RQ 2.1 Was the introduction of the architecture document associated with a change in submission activity?

RQ 2.2 Was the introduction of the architecture document associated with a change in the quality of submissions?

Research Question

RQ 3.1 Was the introduction of the architecture document associated with faster promotion from Commenter to Contributor to Committer?

RQ 3.2 Was the introduction of the architecture document associated with changes in project communication patterns?

Research Question

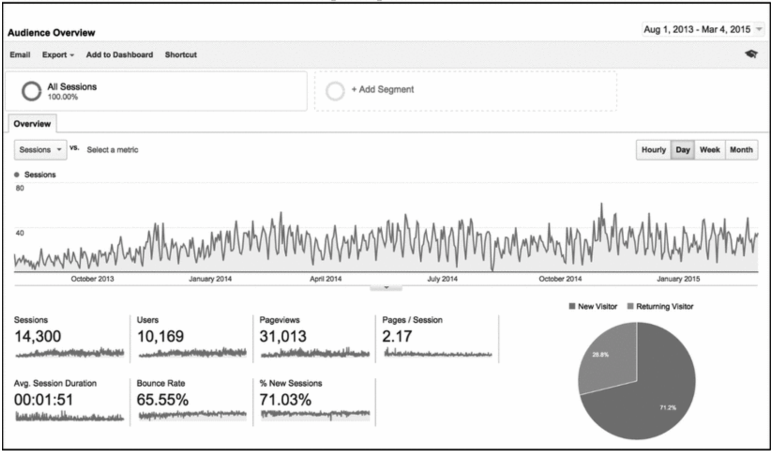
RQ 4.1 Were there any measurable differences in the use of architectural concepts in discussion of issues before and after the architecture document was introduced?

RQ 5.1 How did the Contributors and Committers use the key concepts outlined in the architecture document?

Methodology

Research Question	**Methodology**
RQ 1.1	Tracking how often the architecture documentation is downloaded and how often it is mentioned in discussion groups
RQ 2.1 & 2.2	Tracking whether any changes have occurred in the interactions and activities of the HDFS developer (Contributors and Committers)
RQ 3.1 & 3.2	Tracking project community health measures, such as the growth of the committer group, and the time lag between someone’s appearance as a Contributor and their acceptance as a Committer
RQ 4.1	Tracking whether the introduction of the architecture documentation changed how the project community discussed the system
RQ 2.1 & 2.2	Tracking product performance indicators, such as project capacity—how often Contributors made submissions to the system—and submission quality—how often submissions were rejected by the Committers
RQ 1.2 & 5.1	Surveying the HDFS Contributor and Committer community on their opinions of the value of the architecture documentation that we created

RQ 1.1 Was the architecture document read and if so, how much and how did it change?



Audience Overview

Aug 1, 2013 - Mar 4, 2015

Sessions 14,300

Users 10,169

Pageviews 31,013

Pages / Session 2.17

Avg. Session Duration 00:01:51

Bounce Rate 65.55%

% New Sessions 71.03%

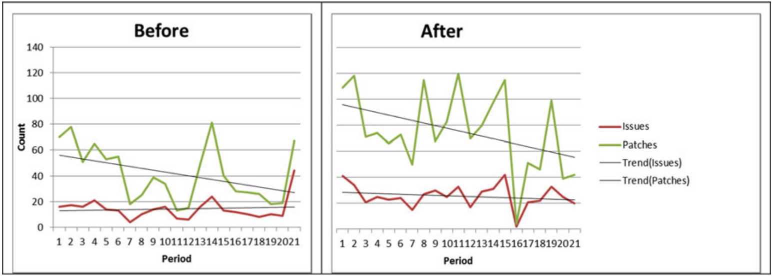
RQ 1.1 Was the architecture document read and if so, how much and how did it change?

- The architecture documentation website was visited over 14,000 times in 19 months
- Almost 30 percent of these visits were from return visitors
- The access rate increased steadily during the study period.

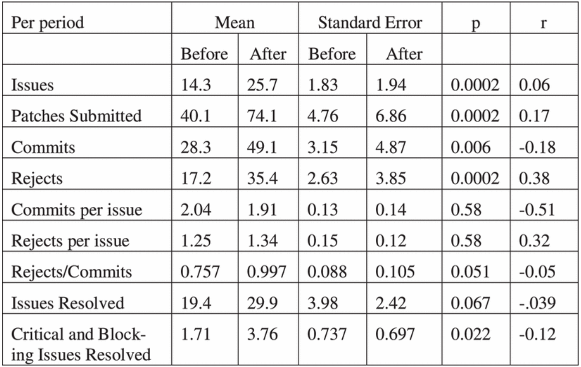
RQ 1.2 Was the architecture document referred to by the project Contributors and Committers?

- Committers and Contributors reported having made relatively little reference to the architectural document itself
- The HDFS Committers appeared to be beginning to advise others about the existence of the documentation available on the Internet.

RQ 2.1 Was the introduction of the architecture document associated with a change in submission activity?



RQ 2.2 Was the introduction of the architecture document associated with a change in the quality of submissions?



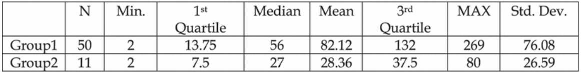
Per period	Mean		Standard Error		p	r
	Before	After	Before	After		
Issues	14.3	25.7	1.83	1.94	0.0002	0.06
Patches Submitted	40.1	74.1	4.76	6.86	0.0002	0.17
Commits	28.3	49.1	3.15	4.87	0.006	-0.18
Rejects	17.2	35.4	2.63	3.85	0.0002	0.38
Commits per issue	2.04	1.91	0.13	0.14	0.58	-0.51
Rejects per issue	1.25	1.34	0.15	0.12	0.58	0.32
Rejects/Commits	0.757	0.997	0.088	0.105	0.051	-0.05
Issues Resolved	19.4	29.9	3.98	2.42	0.067	-.039
Critical and Blocking Issues Resolved	1.71	3.76	0.737	0.697	0.022	-0.12

Discussion RQ 2.1 & 2.2

- The numbers of issues addressed with at least one patch, patches submitted, committed, and rejected all increased by statistically and practi- cally significant amounts
- Although the observed pre/post differences in patches sub- mitted, committed, and rejected was significant, there is a doubt of a causal relationship.

RQ 3.1 Was the introduction of the architecture document associated with faster promotion from Commenter to Contributor to Committer?

$$\text{Commenter} \rightarrow \text{Contributor} \rightarrow \text{Committer} \tag{1}$$



	N	Min.	1st Quartile	Median	Mean	3rd Quartile	MAX	Std. Dev.
Group1	50	2	13.75	56	82.12	132	269	76.08
Group2	11	2	7.5	27	28.36	37.5	80	26.59

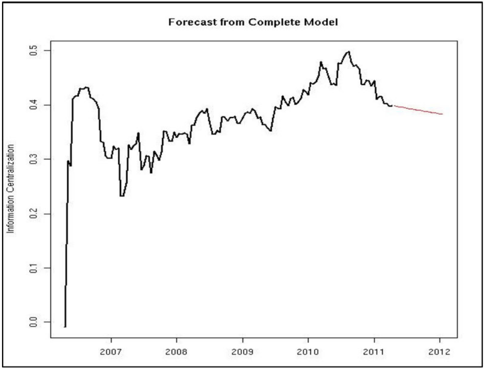
RQ 3.1 Was the introduction of the architecture document associated with faster promotion from Commenter to Contributor to Committer?

- It is a positive answer to RQ 3.1 on the role of architectural documentation in facilitating promotion within the OSS community of HDFS.

RQ 3.2 Was the introduction of the architecture document associated with changes in project communication patterns?

- The study builds a social network derived from the communications data set
- It was collected structural metrics of that social network change over the course of the project.

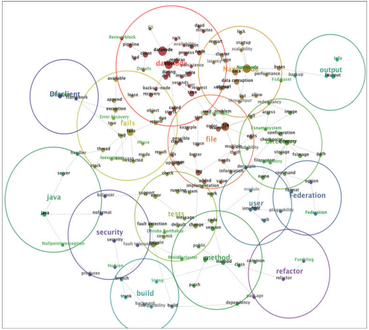
RQ 3.2 Was the introduction of the architecture document associated with changes in project communication patterns?



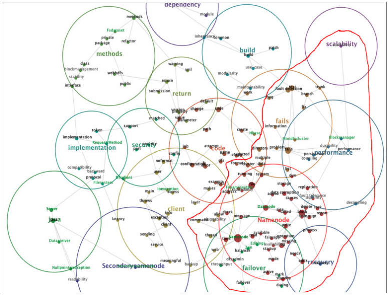
RQ 3.2 Was the introduction of the architecture document associated with changes in project communication patterns?

- The statistical tests on the communication data set suggest also a benefit to the HDFS community as a whole, in that the know-how necessary to participants to be effective Contributors may have become more diffused.

RQ 4.1 Were there any measurable differences in the use of architectural concepts in discussion of issues before and after the architecture document was introduced?



RQ 4.1 Were there any measurable differences in the use of architectural concepts in discussion of issues before and after the architecture document was introduced?



RQ 4.1 Were there any measurable differences in the use of architectural concepts in discussion of issues before and after the architecture document was introduced?

- The analysis revealed concepts and relationships correspond-ing to those described in our architectural documentation.
- Text analysis of issue discussions shows that the discussions themselves can be a source for sharing architectural concepts

RQ 5.1 How did the Contributors and Committers use the key concepts outlined in the architecture document?

- Contributors and Committers we surveyed clearly recognized the importance of the concepts that were covered in the architectural document and used them in their own work on the HDFS codebase.
- They sometimes tended to focus more heavily on implementation details than architectural considerations *per se*.

Conclusions

- The HDFS Committers appear to maintain intellectual control over their code base.
- The HDFS community of Committers and the most active Contributors is actively interested in architectural concepts.

Conclusions

- The architecture documentation to have an effect on the project, but principally on less active Contributors and newcomers.
- The project’s social network became less centralized and the speed of promotion from Commenter to Contributor was quicker after the introduction of the documentation.

Threats to Validity

- The findings must be intended as observed correlations, whose causal linkage to architectural documentation still remains to be explored and validated
- Limitations on generalization of Study Cases

Threats to Validity

- The HDFS community itself may have matured during the course of this study, creating a greater inherent interest in architectural documentation, irrespective of any activity on our part.
- The new users might have produced poorer quality code because of inexperience

Questions?

References I

[1] P. Clements, D. Garlan, L. Bass, J. Stafford, R. Nord, J. Ivers, and R. Little, *Documenting software architectures: views and beyond*. Pearson Education, 2002.

[2] W. Ding, P. Liang, A. Tang, H. v. Vliet, and M. Shahin, “How do open source communities document software architecture: An exploratory survey,” in *Engineering of Complex Computer Systems (ICECCS), 2014 19th International Conference on*, Aug 2014, pp. 136–145.

References II

[3] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, “The hadoop distributed file system,” in *Mass Storage Systems and Technologies (MSST), 2010 IEEE 26th Symposium on*. IEEE, 2010, pp. 1–10.