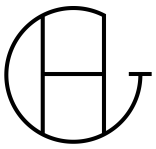

HoGent

Faculteit Bedrijf en Organisatie

Analysis and Design of an Efficient NoSQL Data Storage Schema for an Interlinked Recent Activity Feed

Florian Dejonckheere

Scriptie voorgedragen tot het bekomen van de graad van
professionele bachelor in de toegepaste informatica

Promotor:
Chantal Teerlinck

Instelling: Open Webslides

Academiejaar: 2017-2018

Tweede examenperiode

Faculty of Business and Information Management

Analysis and Design of an Efficient NoSQL Data Storage Schema for an Interlinked Recent Activity Feed

Florian Dejonckheere

Thesis submitted in partial fulfilment of the requirements for the degree of
professional bachelor of applied computer science

Promotor:
Chantal Teerlinck

Institution: Open Webslides

Academic year: 2017-2018

Second examination period

Preface

Samenvatting

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus.

Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Abstract

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus.

Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Contents

List of Figures

List of Tables

Introduction

0.1 Context

In this age of computers and smartphones, older and deprecated methods of teach are quickly being replaced by the digital equivalent. Course content has shifted from being printed in full-text on paper, to static slides on an overhead projector and to the digital screen present in every modern classroom. However, there is a lot more potential to gain from the modernization of course content. Education is still too often a one-way street, where students are obligated to process the course content without being able to provide much challenge or activity. This also does not allow for any dialogue to take place between students and teachers concerning feedback and improvement of the material itself.

Technological constraints in current iterations of educational software do not allow this co-creation discourse easily. Material being locked to specific versions of proprietary software is just one of the many problems teachers might encounter when trying to apply this concept in real life.

By utilizing interactive tools and applications, the teacher can engage the students more directly.

By building on modern, open standards, the Open Webslides project (Open Webslides, 2017a) aims to provide a platform that solves these problems. It creates a user-friendly environment where teachers can create courses based on open source technologies and standards, and it allows them to apply the co-creation narrative easily. This also enables users to share their material not just with their immediate environment but with a much broader educational audience.

0.2 Problem statement

The Open Webslides platform incorporates several ways to stimulate spontaneous co-creation between teachers and students. One of the most prominent elements is the *Recent Activity* feed. This reverse chronologically ordered list enumerates the most recent user interactions with the platform and with other users. Feed items range from simple actions such as a user having created or modified course content, to more complex social interactions like a discussion being held using annotations or a student's changes being incorporated in a teacher's courses.

The size of the activity feed data set is directly correlated to the size and activity of the userbase. It has the potential to grow explosively, especially in timespans critical to teachers and students such as examination periods. In order for the infrastructure to be able to handle the deluge of the queried information, designing a system that allows for efficient querying and easy scalability is of paramount importance. Further decoupling of this subsystem from the business-critical processes is also important to ensure round the clock availability of course content to the users.

This research thesis will provide a framework for the Open Webslides project to implement an efficient, scalable data storage system in the context of the Recent Activity feed already incorporated into the platform.

0.3 Research questions

0.4 Research goal and objectives

In the chapter 1 an overview will be presented of current research, applied solutions and other related work in the research domain based on a literature study.

The chapter 2 will clarify the analytical research approaches used in this thesis to attempt to formulate an answer to the research questions.

Finally, the **??** chapter will present a conclusion and an applied answer to the research questions.

1. State of the Art

Ever since the rise of the NoSQL databases in 2009 (Sadalage & Fowler, 2012) it has been a subject for vigorous academic and professional research. The contrast with relational databases, optimal use cases, performance and scalability are only some of the aspects that have been analyzed with great regularity. This chapter will summarize previous publications relevant to this thesis.

The book by Sadalage and Fowler (2012) provides an excellent entry into the world of NoSQL. It explains the motivation behind the use of NoSQL techniques, and how this differs from relational data storage. Furthermore it introduced the segregation of NoSQL data stores into four main categories: key-value, document, column and graph databases. Second, Sadalage and Fowler touch the concept of polyglot persistence. This discribes the application's opportunity to use multiple types of data stores to store heterogeneously structured data. This technique is relevant in particular to this thesis, as the describe data schema only relates to one of the database management systems integrated in the Open Webslides platform. The second part of the book provides a more practical approach to using polyglot persistence in an enterprise application. The authors have written down many pointers and guides in order to pick the right database for the use case.

Grolinger, Higashino, Tiwari, and Capretz (2013) present a use-case based approach to comparing different NoSQL and NewSQL data stores. The survey incorporates a feature-based comparison over different aspects such as querying, scalability and security, and analyzes these concepts in the context of a select number of NoSQL data stores.

Hecht and Jablonski (2011) provides a feature-based comparison of different NoSQL database types and vendors. The researchers compare the data model, querying access, concurrency, partioning and replication. They use a duality-based approach, where a minus

indicates that the feature is not supported by the database system, and a plus if the feature is supported. The paper also presents the problem of a lack of unified querying interface for NoSQL databases. Furthermore, the importance of choosing the right NoSQL database type for the use case is emphasized, however Hecht and Jablonski do not present a specific case study.

The proceedings of the 2013 IEEE International Conference on Big data by Kaur and Rani (2013) describe the theoretical modeling and querying of SQL and NoSQL data stores. The paper then proceeds with a case study of a social networking site similar to Slashdot (Malda & Bates, 1997). Starting from an entity-relationship diagram (ERD), the researchers then proceed by modeling the entities in both a document and a graph database. Finally, a set of seven queries related to the use case is then drawn up and compared for the PostgreSQL, MongoDB and Neo4j data stores.

Zhao (2015) explores the use of NoSQL data stores to store huge amounts of observational data generated by astronomical research. It briefly discusses using filesystems and relational data stores, before comparing NoSQL alternatives. A concrete data model to store the astronomical data in a MongoDB data store is then presented, together with eight scenarios and queries that may be used in a production system. Furthermore, performance measurements of MongoDB are also analyzed. Data insertion, querying and deletion using the aforementioned data scheme and real observational data are used in this section.

The proceedings of the AGILE 2015 conference by Schmid, Galicz, and Reinhardt (2015) present an overview of selected SQL and NoSQL databases, focusing on the geo-functionalities of the systems. It uses performance tests between two document-based NoSQL data stores (MongoDB and CouchBase). The researchers conclude that geospatial calculations in NoSQL database systems are still only supported for basic queries. Relational databases still perform superior to NoSQL databases in small to larger data sets for queries with geo-functions. However the NoSQL response time only increases slightly relative to data set size.

The technical report by Barahmand, Ghandeharizadeh, and Li (2015) quantifies the scalability of MongoDB and HBase for processing simple operations using the social networking benchmark BG (Barahmand & Ghandeharizadeh, 2013). It considers both horizontal and vertical scalability of the data stores using the Social Action Rating (SoAR) introduced by the benchmarking tool. In order to perform these benchmarks, two logical data models for the database design are presented. The report concludes that while both data stores scale superlinearly, their speedup is limited by the resources of a few nodes out of many becoming fully utilized.

The main differences between this study and the previous studies are:

1. Many studies have been conducted to understand the motivation between the NoSQL principles and the shift from relational data stores. The division of NoSQL data store types into four categories and elaboration upon this is usually also a topic in these studies. This research paper builds upon that knowledge, providing only a brief introduction in the world of NoSQL and NewSQL concepts.

2. Some of the previously mentioned research papers also discuss a case study applied to a specific use case. This is mostly related to business critical systems that store and process large volumes of data. The use case described in this thesis is very specific in that it's a complementary subsystem that does not affect critical data. Therefor, certain comparative attributes such as security and availability are not considered in this research.

This thesis aims to provide a case study of data storage the Open Webslides (2017a) platform. Several use case based surveys and studies already exist, however they aim at replacing a relational database in an application with a NoSQL database without bringing polyglot persistence into account. Sadalage and Fowler (2012) is one notable exception in this aspect. In the case of Open Webslides, the NoSQL data store only complements the relational database and does not fulfill a critical function. Therefore, several constraints such as security and availability differ in interpretation from existing studies.

2. Methodology

Etiam pede massa, dapibus vitae, rhoncus in, placerat posuere, odio. Vestibulum luctus commodo lacus. Morbi lacus dui, tempor sed, euismod eget, condimentum at, tortor. Phasellus aliquet odio ac lacus tempor faucibus. Praesent sed sem. Praesent iaculis. Cras rhoncus tellus sed justo ullamcorper sagittis. Donec quis orci. Sed ut tortor quis tellus euismod tincidunt. Suspendisse congue nisl eu elit. Aliquam tortor diam, tempus id, tristique eget, sodales vel, nulla. Praesent tellus mi, condimentum sed, viverra at, consectetuer quis, lectus. In auctor vehicula orci. Sed pede sapien, euismod in, suscipit in, pharetra placerat, metus. Vivamus commodo dui non odio. Donec et felis.

Etiam suscipit aliquam arcu. Aliquam sit amet est ac purus bibendum congue. Sed in eros. Morbi non orci. Pellentesque mattis lacinia elit. Fusce molestie velit in ligula. Nullam et orci vitae nibh vulputate auctor. Aliquam eget purus. Nulla auctor wisi sed ipsum. Morbi porttitor tellus ac enim. Fusce ornare. Proin ipsum enim, tincidunt in, ornare venenatis, molestie a, augue. Donec vel pede in lacus sagittis porta. Sed hendrerit ipsum quis nisl. Suspendisse quis massa ac nibh pretium cursus. Sed sodales. Nam eu neque quis pede dignissim ornare. Maecenas eu purus ac urna tincidunt congue.

Donec et nisl id sapien blandit mattis. Aenean dictum odio sit amet risus. Morbi purus. Nulla a est sit amet purus venenatis iaculis. Vivamus viverra purus vel magna. Donec in justo sed odio malesuada dapibus. Nunc ultrices aliquam nunc. Vivamus facilisis pellentesque velit. Nulla nunc velit, vulputate dapibus, vulputate id, mattis ac, justo. Nam mattis elit dapibus purus. Quisque enim risus, congue non, elementum ut, mattis quis, sem. Quisque elit.

Maecenas non massa. Vestibulum pharetra nulla at lorem. Duis quis quam id lacus dapibus interdum. Nulla lorem. Donec ut ante quis dolor bibendum condimentum. Etiam egestas

tortor vitae lacus. Praesent cursus. Mauris bibendum pede at elit. Morbi et felis a lectus interdum facilisis. Sed suscipit gravida turpis. Nulla at lectus. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Praesent nonummy luctus nibh. Proin turpis nunc, congue eu, egestas ut, fringilla at, tellus. In hac habitasse platea dictumst.

Vivamus eu tellus sed tellus consequat suscipit. Nam orci orci, malesuada id, gravida nec, ultricies vitae, erat. Donec risus turpis, luctus sit amet, interdum quis, porta sed, ipsum. Suspendisse condimentum, tortor at egestas posuere, neque metus tempor orci, et tincidunt urna nunc a purus. Sed facilisis blandit tellus. Nunc risus sem, suscipit nec, eleifend quis, cursus quis, libero. Curabitur et dolor. Sed vitae sem. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Maecenas ante. Duis ullamcorper enim. Donec tristique enim eu leo. Nullam molestie elit eu dolor. Nullam bibendum, turpis vitae tristique gravida, quam sapien tempor lectus, quis pretium tellus purus ac quam. Nulla facilisi.

3. Conclusion

Curabitur nunc magna, posuere eget, venenatis eu, vehicula ac, velit. Aenean ornare, massa a accumsan pulvinar, quam lorem laoreet purus, eu sodales magna risus molestie lorem. Nunc erat velit, hendrerit quis, malesuada ut, aliquam vitae, wisi. Sed posuere. Suspendisse ipsum arcu, scelerisque nec, aliquam eu, molestie tincidunt, justo. Phasellus iaculis. Sed posuere lorem non ipsum. Pellentesque dapibus. Suspendisse quam libero, laoreet a, tincidunt eget, consequat at, est. Nullam ut lectus non enim consequat facilisis. Mauris leo. Quisque pede ligula, auctor vel, pellentesque vel, posuere id, turpis. Cras ipsum sem, cursus et, facilisis ut, tempus euismod, quam. Suspendisse tristique dolor eu orci. Mauris mattis. Aenean semper. Vivamus tortor magna, facilisis id, varius mattis, hendrerit in, justo. Integer purus.

Vivamus adipiscing. Curabitur imperdiet tempus turpis. Vivamus sapien dolor, congue venenatis, euismod eget, porta rhoncus, magna. Proin condimentum pretium enim. Fusce fringilla, libero et venenatis facilisis, eros enim cursus arcu, vitae facilisis odio augue vitae orci. Aliquam varius nibh ut odio. Sed condimentum condimentum nunc. Pellentesque eget massa. Pellentesque quis mauris. Donec ut ligula ac pede pulvinar lobortis. Pellentesque euismod. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent elit. Ut laoreet ornare est. Phasellus gravida vulputate nulla. Donec sit amet arcu ut sem tempor malesuada. Praesent hendrerit augue in urna. Proin enim ante, ornare vel, consequat ut, blandit in, justo. Donec felis elit, dignissim sed, sagittis ut, ullamcorper a, nulla. Aenean pharetra vulputate odio.

Quisque enim. Proin velit neque, tristique eu, eleifend eget, vestibulum nec, lacus. Vivamus odio. Duis odio urna, vehicula in, elementum aliquam, aliquet laoreet, tellus. Sed velit. Sed vel mi ac elit aliquet interdum. Etiam sapien neque, convallis et, aliquet vel, auctor non, arcu. Aliquam suscipit aliquam lectus. Proin tincidunt magna sed wisi. Integer blandit

lacus ut lorem. Sed luctus justo sed enim.

Morbi malesuada hendrerit dui. Nunc mauris leo, dapibus sit amet, vestibulum et, commodo id, est. Pellentesque purus. Pellentesque tristique, nunc ac pulvinar adipiscing, justo eros consequat lectus, sit amet posuere lectus neque vel augue. Cras consectetuer libero ac eros. Ut eget massa. Fusce sit amet enim eleifend sem dictum auctor. In eget risus luctus wisi convallis pulvinar. Vivamus sapien risus, tempor in, viverra in, aliquet pellentesque, eros. Aliquam euismod libero a sem.

Nunc velit augue, scelerisque dignissim, lobortis et, aliquam in, risus. In eu eros. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Curabitur vulputate elit viverra augue. Mauris fringilla, tortor sit amet malesuada mollis, sapien mi dapibus odio, ac imperdiet ligula enim eget nisl. Quisque vitae pede a pede aliquet suscipit. Phasellus tellus pede, viverra vestibulum, gravida id, laoreet in, justo. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Integer commodo luctus lectus. Mauris justo. Duis varius eros. Sed quam. Cras lacus eros, rutrum eget, varius quis, convallis iaculis, velit. Mauris imperdiet, metus at tristique venenatis, purus neque pellentesque mauris, a ultrices elit lacus nec tortor. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent malesuada. Nam lacus lectus, auctor sit amet, malesuada vel, elementum eget, metus. Duis neque pede, facilisis eget, egestas elementum, nonummy id, neque.

A. Research proposal

The subject of this thesis is based upon a research proposal that was graded by the promotor in advance. This proposal has been included as an attachment.

A.1 Introduction

The Open Webslides (2017b) project provides a user-friendly platform to collaborate on webslides - slides made with modern web technologies such as HTML, CSS and JavaScript. One of the core features this application provides is *co-creation*. The co-creation aspect manifests itself in several forms within the application; annotations on slides and a change suggesting system resembling GitHub's pull request feature are the main mechanisms. Because of the inherent social nature of co-creation, a basic notifications feed was also implemented. This feed is tailored to the user, and reflects the most recent changes, additions and comments relevant to the slide decks the user is interested in.

However, at the moment the functionality implemented in the system contains only the bare necessities. The module will be expanded in the future, and doing so requires a structural and conceptual rethinking of how the notifications are generated, stored and queried. The size of the dataset is also expected to grow linearly with user activity, therefore scalability is a requirement as well.

This paper has two research questions.

1. What frameworks and software packages currently exist in the industry to store structured non-relational graph or document data?
2. How is the social graph provided by the Open Webslides' notification feed structured

and how is this data consumed?

Answering these questions is paramount for the final section of the paper, which describes a data storage mechanism that performs well given the functional requirements of the project's data flow.

A.2 Use case

This thesis is a study of NoSQL data storage techniques applied to the Open Webslides (2017b) project. The online, interactive platform that the project provides includes a list of notifications in reverse chronological order, tailored to the user. This feature is called the *social feed*. It enumerates the most recent social activity on the platform. For example, if a user updates a slide deck, the user's friends would be able to find a change notification in their respective social feeds.

From the perspective of the application code that generates the social feed, the user, slide deck and notification should be considered separate entities. The notification itself has relations to the other entities: the author (the *subject*), the slide deck (the *object*), along with the *predicate* property that provides information on what operation was performed (for example creating or updating a slide deck).

This data model is also characterized by the *write-once read-many* nature of the information; once the notification has been generated, it does not need to be modified again. It will also only be queried in a very specific way: the application server will always attempt to retrieve the most recent notifications starting from the user entity. This principle is an important aspect to take into consideration in the choice of data storage mechanism.

A.3 State-of-the-art

In current literature, studies such as Moniruzzaman and Hossain (2013, 4), Nayak, Poriya, and Poojary (2013) and Dayne Hammes and Mitchell (2014) have already analyzed the disparity between traditional relational database systems and NoSQL stores. However, the conceptual and technical difference between these data storage models will not be scrutinized any further, since this paper presents a data storage solution applied to the social notification feed of the Open Webslides (2017b) project.

There are already many existing free and commercial products for the storage of NoSQL data, such as Redis (Sanfilippo, 2009), CouchDB (Apache Software Foundation, 2005a), MongoDB (Inc., 2009) and Neo4j (Technology, 2007). Finding the right database model for this use case (section A.2) is one of the hurdles this paper intends to handle. Zhao (2015) describes the development of a messaging system for astrophysical transient event notifications. Part of this paper is a qualitative comparison between document-based NoSQL storage solutions fit for this particular use case. We expect this paper to provide

a solid base of reasoning in order to find a scalable and efficient solution for resolving similar computational challenges.

The goal of this paper is to provide a performant, scalable and maintainable data storage schema for the Open Webslides (2017b) platform, regarding the linked social graph that powers the *Social Feed* functionality present in the platform.

A.4 Methodology

First, a range of industry-standard NoSQL database management systems such as MongoDB (Inc., 2009), HBase (Apache Software Foundation, 2005b) and Neo4j (Technology, 2007) will be qualitatively analyzed. Three of the five types of NoSQL database types (Nayak et al., 2013) will be included in the study: column-oriented, document based and graph databases. Criteria for comparison include how the database management system concretely stores its data on disk, the query format and specific programming language bindings. Another important aspect is the distributed nature of many NoSQL databases. Using Brewer's conjecture (Gilbert & Lynch, 2002, 2) – often called the CAP theorem – the existing types of data storage systems will be examined and summarized. There is also a practical factor present in the research; this includes the license of the project, its active maintainability and future prospects. Common types of NoSQL databases include key-value store, column-oriented, document store and graph databases (Nayak et al., 2013). This paper will provide a short introduction to these types, before proceeding with the type that fits our use case the most.

Second, the data model specific to the Open Webslides project will be examined. We will start from the data model that is already implemented in the current iteration of the platform. At the time of writing, the existing base implementation of the social notification feed only contains two types of notifications. This paper will try to extrapolate this concept into a more generalized, abstract system in which developers can easily plug additional notification types. The physical properties of the data model will also be taken into account: the data will be written to the data storage only once, but read many times. It is also highly interlinked information, as a notification will always relate to one or more users as a subject, and a target object as well – most likely a slide deck or collection of slide decks. These links need to be maintained, and efficiently reconstructed when queried.

Finally, a sample dataset will be constructed using the aforementioned detailed analysis. Empirical testing will be conducted against multiple database management systems, and the results will be summarized and interpreted. Various information flows will be tested; however, the most important process remains efficiently querying the stored data.

Using the comparative study of storage engines, data model analysis and the empirical results an implementation plan will be constructed. This plan will serve as a recommendation for future development.

A.5 Expected results and conclusions

The NoSQL ecosystem, unlike relational databases, is headed towards specialization, so different solutions are headed in different directions (Maroo, 2013). In this paper, we expect to find one type of NoSQL database that is a better fit for the Open Webslides use case, in clear contrast with the other types of storage engines. Due to the inherently highly interlinked nature of the stored data, we suspect a graph-based database management system to provide most advantages, and generally the most performant experience.

This expectation is amplified by the availability and good community support of Ruby bindings to the most popular graph database management systems.

Since the platform being discussed only caters to a small to medium user base, we do not expect the need to scale horizontally beyond one instance. However, the vertical scalability is still a topic for discussion, and we expect to determine the computational order of magnitude in order to efficiently query the given dataset during this study.

Finally, the implementation plan should describe a concrete roadmap, stretching over a development period with a baseline expectation of one to three months. Roll-out of this mechanism should also be included in this plan.

We also expect that this thesis will provide a good reference to a further stable, scalable and extendable implementation of the *social feed* feature in the Open Webslides (2017b) project as outlined in section A.2.

Bibliography

Apache Software Foundation. (2005a). CouchDB. Retrieved from https://couchdb.apache.org/

Apache Software Foundation. (2005b). Neo4j. Retrieved from https://hbase.apache.org/

Barahmand, S. & Ghandeharizadeh. (2013). *BG: A Benchmark to Evaluate Interactive Social Networking Actions*. University of Southern California.

Barahmand, S., Ghandeharizadeh, S., & Li, J. (2015). *On Scalability of Two NoSQL Data Stores for Processing Interactive Social Networking Actions*. University of Southern California.

Dayne Hammes, H. M. & Mitchell, H. (2014). *Comparison of NoSQL and SQL Databases in the Cloud*. Southern Association for Information Systems.

Gilbert, S. & Lynch, N. (2002). Brewer's conjecture and the feasibility of consistent, available, partition-tolerant web services. *ACM SIGACT News*, *33*, 51–59.

Grolinger, K., Higashino, W. A., Tiwari, A., & Capretz, M. A. (2013). Data Management in Cloud Environments: NoSQL and NewSQL Data Stores. *Journal of Cloud Computing: Advances, Systems and Applications*, *2*(1), 22. doi:10.1186/2192-113X-2-22

Hecht, R. & Jablonski, S. (2011). NoSQL Evaluation: A Use Case Oriented Survey. In *2011 International Conference on Cloud and Service Computing* (pp. 336–341).

Inc., M. (2009). MongoDB. Retrieved from https://www.mongodb.com/

Kaur, K. & Rani, R. (2013). Modeling and Querying Data in NoSQL Databases. In *2013 IEEE International Conference on Big Data* (pp. 1–7).

Malda, R. & Bates, J. (1997). Slashdot. Retrieved from https://www.mongodb.com/

Maroo, T. (2013). *Handling with Dynamic, Large Data Sets - NoSQL a Buzzword or Savior?* JECRC Foundation.

Moniruzzaman, A. B. M. & Hossain, S. A. (2013). NoSQL Database: New Era of Databases for Big data Analytics - Classification, Characteristics and Comparison. *International Journal of Database Theory and Application*, *6*.

Nayak, A., Poriya, A., & Poojary, D. (2013, March). Type of NoSQL Databases and its Comparison with Relational Databases. *International Journal of Applied Information Systems (IJAIS)*, *5*(4).

Open Webslides. (2017a, March). Open Webslides. Retrieved from http://openwebslides.github.io/

Open Webslides. (2017b). Open Webslides. Retrieved from http://openwebslides.github.io/

Sadalage, P. J. & Fowler, M. (2012). *NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence*. Pearson Education.

Sanfilippo, S. (2009). Redis. Retrieved from https://redis.io/

Schmid, S., Galicz, E., & Reinhardt, W. (2015). Performance Investigation of Selected SQL and NoSQL Databases. In *AGILE 2015*.

Technology, N. (2007). Neo4j. Retrieved from https://neo4j.com

Zhao, Y. (2015). *Event Based Transient Notification Architecture and NoSQL Solution for Astronomical Data Management* (Doctoral dissertation, Massey University).