XML and NOSQL DBMS: Migration and Benchmarking

Author:
Prakash Thapa

Prof. Dr. Marc H. Scholl
Prof. Dr. Daniel Keim

Christian Gruen

Konstanz University

November 4, 2014

Abstract

XML and NoSQL database are two growing field in second generation database system, They share some similarities as well as they have some significant difference. This thesis focus on the comparative analysis of these two database system based on the Use cases and existing solution, we will discuss the data processing, query pattern and Information Retrieval(IR)

....

Zusammenfassung(German Abstract)

XML und NoSQL

Acknowledgments

The completion of this master thesis would not have been possible without the support of many people.

Contents

1 Introduction

1.1 Motivation

Few years of time XML was *de facto* data exchange format which enabled people to do previously not that easy thing that time like exchange of content of Microsoft's office documents exchange through HTTP connections. But in recent years a bold transformation has been a foot in the world of Data exchange. The more light weight, less bandwidth consumer JSON(JavaScript Object Notation)[1] has been emerge not just as an alternative to the XML but as rather as potential full Blown successor[2]. Even though these two format has their own pros and cons, the rise of JSON as key in data exchange format, new database technologies so called NoSQL are also emerges and getting success in their own way. The rate of new research papers in these system are increasing in recent years.

1.2 Contribution

The main contribution of this thesis is that it provide the necessary techniques and algorithms migration of data from XML database to NoSQL databases. More specifically, It will focus on Document store databases MongoDB, Couchbase and Re-thinkdb. To complete this task it is necessary to understand general architecture and data model of each of these database as well and the Information Retrieval(IR). At the second part, conversion of Queries in XML data to individual NoSQL database is also ......

.....

1.3 Overview

This thesis is divided into three main sections. The first section define the Techniques and necessary algorithms to convert XML to JSON Data format. The second section will Systems and scope of work. On third section, we see the performance and comparative analysis of each of these systems. The work is structured as follows: Chapter 2: Chapter 3: Chapter 4: Chapter 5:

2 Preliminaries

2.1 Semi-structured data

2.1.1 XML and JSON

2.2 XML Database

2.2.1 XML Query Language

2.3 NoSQL database

2.4 Key/Value storage

2.5 document oriented database

2.5.1 Querying NoSQL database
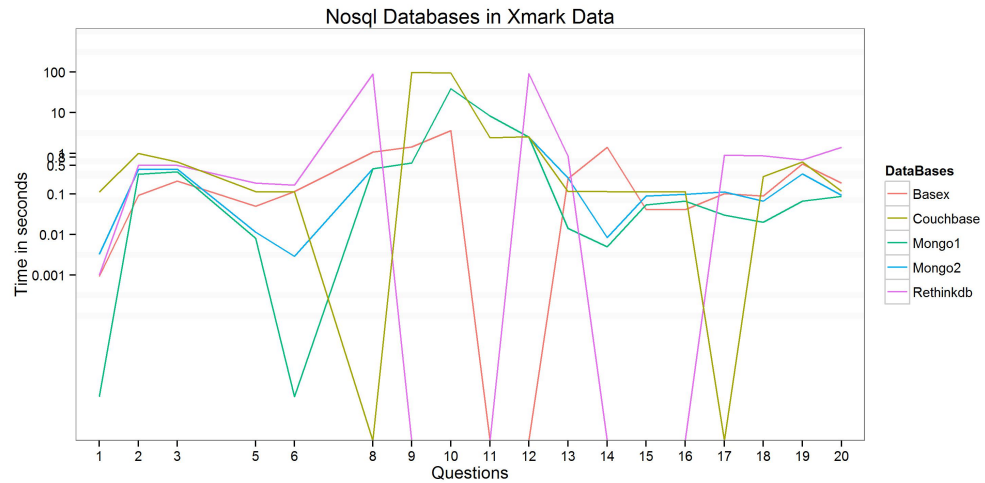
3 Related work

Figure 1: An overview of some important indexing structures developed over years

4 System/Environment

4.1 BaseX

Code 1: A simple KML example representing a Point

```xml
<?xml version="1.0" encoding="UTF-8" ?>
<kml xmlns="http://www.opengis.net/kml/2.2">
<document>
<placemark>
  <name>New York City</name>
  <description>New York City</description>
  <point>
    <coordinates>-74.006393,40.714172,0</coordinates>
  </point>
</placemark>
</document>
</kml>
```

Code 2: JSON Data

```json
{"prakash":"Thapa"}
```

4.2   MongoDB

4.3   Couchbase

4.4   Rethinkdb

4.5   Summary

5 Performance/Experiments

5.1 XMark

5.1.1 Dataset

5.1.2 Queries

5.2 XMark data into NoSQL Database

5.2.1 MongoDB

5.2.2 Couchbase

5.2.3 Rethinkdb

5.3 Benchmarking

5.4 Summary

6 Discussion

7 Conclusion

8 Future Work

storing in the memory

References

List of Figures

List of Tables

Listings