

BUILDING A GRAPH BASED RDF STORE TO MANIPULATE RDF DATA EFFICIENTLY

Ravindra Sampath Ranwala

138227T



Department Computer Science and Engineering

University of Moratuwa

Sri Lanka

March 2015

BUILDING A GRAPH BASED RDF STORE TO MANIPULATE RDF DATA EFFICIENTLY

Ravindra Sampath Ranwala

138227T



University of Moratuwa, Sri Lanka.
Electronic Theses & Dissertations
www.lib.mrt.ac.lk

Thesis/Dissertation submitted in partial fulfillment of the requirements for the degree Master
of Science/Master of Engineering in Computer Science and Engineering

Department of Computer Science and Engineering

University of Moratuwa

Sri Lanka

March 2015

DECLARATION

I declare that this is my own work and this thesis/dissertation 2 does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my thesis/dissertation, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature:

Date:

The above candidate has carried out research for the Masters/MPhil/PhD thesis/Dissertation under my supervision.



University of Moratuwa, Sri Lanka.
Electronic Theses & Dissertations
www.lib.mrt.ac.lk

Signature of the supervisor:

Date:

ABSTRACT

Due to the expansion of semantic web technologies, Resource Description frameworks (RDFs) and triple stores became more prevalent. Since there is a huge amount of RDF data available, managing them in a proper and efficient manner is challenging. Many Triple stores were implemented to support the queries related to semantic web. The queries submitted in this context is called as SPARQL queries which are read dominant. These SPARQL queries needs to be answered quickly and efficiently. RDF data is stored in <subject, predicate, object> form and which is called as a triple. A typical triple store contains billions of triples in the above form.

Much work has been devoted to handle RDF data efficiently. But state of the art systems still cannot handle web scale RDF data effectively. Most existing systems store and index data in particular ways. For an example some systems uses relational tables, bitmap matrix to optimize SPARQL query processing on RDF data. This relational approach suffers from high Join cost and large intermediate results. Some have used prolog inference engine to handle RDF data. This also have some limitations given a huge amount of RDF data.

A modern approach is to model the RDF data in its native Graph form. This approach requires new algorithms to build the graph and graph exploration techniques to answer SPARQL queries. This yields no join cost and very small intermediary results. Also this approach yields less query execution time for complex SPARQL queries.

The objective of this research is to build a graph based triple store for Apache Cassandra. It uses Apache Jena Graph Processing framework to build and explore the RDF graph. Towards the end, it conducts a performance benchmark of this RDE store with some other RDF store implementations using DBPedia dataset and sample queries and proves that this graph based approach outperforms other RDF store implementations.

ACKNOWLEDGEMENT

I would like to express my special appreciation and thanks to my supervisor Dr. Amal Shehan Perera, you have been a tremendous mentor to me. I would like to thank you for guiding me through your experience to make this research more worthwhile. Also I would like to appreciate the extended support and guidance given by Dr. Srinath Perera through his experience and knowledge, which was really helpful for me to complete this successfully.

I would also like to take this opportunity to thank Dr. Malaka Walpola for guiding us towards this work throughout Research Seminar lecture sessions over a year and for the extended support and kindness granted to us. At last but not the least, I would prefer thank to all the academic staff members for helping, guiding, encouraging us and disseminating knowledge throughout the programme.



University of Moratuwa, Sri Lanka.
Electronic Theses & Dissertations
www.lib.mrt.ac.lk

TABLE OF CONTENTS



University of Moratuwa, Sri Lanka.
Electronic Theses & Dissertations
www.lib.mrt.ac.lk

Table of Contents

DECLARATION	ii
ABSTRACT	iii
ACKNOWLEDGEMENT	iv
TABLE OF CONTENTS	v
LIST OF FIGURES	viii
LIST OF TABLES	ix
LIST OF ABBREVIATIONS	x
1 INTRODUCTION.....	1
1.1 Semantic web, RDF and Triples.....	2
1.2 Challenges faced by existing RDF management systems.....	3
1.3 New approaches in managing RDF data efficiently	4
1.4 Existing Graph Processing Frameworks.....	6
1.5 The Problem/Opportunity.....	7
1.6 Objectives	7
2 LITERATURE REVIEW	9
2.1. RDF Store (What, Why and How).....	10
2.2 Apache Cassandra.....	12
2.3 Different approaches to build a triple store	16
2.3.1 Relational Approach	16
2.3.2 RDF data centric storage	17
2.3.3 RDF graph based approach	18
2.3.4 Hybrid Approaches	25
2.4 Benchmarking RDF stores for performance evaluation	27
3 METHODOLOGY	29
3.1. Existing Solutions	30
3.2 Proposed Solution	31
3.3 Solution Architecture.....	32
3.4 Solution Implementation	34
3.4.1 Populating data into Cassandra Cluster	36
3.4.2. Building the RDF Graph	37
3.4.3. Querying the RDF Graph	38
3.4.4: Dropping the RDF Store	39

3.4.5: Techniques used to render RDF/XML results on the webpage	40
3.4.6: Solution Extensibility and Flexibility	43
4 USE CASE SCENARIOS	44
4.1. Build the RDF graph.....	45
4.2 Executing SPARQL query.....	47
4.3 Rendering Results	48
5 EVALUATION AND RESULT.....	49
5.1 RDF Store Benchmarking	50
5.1.1 Dataset Generation	51
5.1.1.1 RDF/XML	51
5.1.1.2 Turtle	53
5.1.1.3 N-Triples	54
5.1.2 Tested RDF Stores	55
5.1.3 Query Generation	56
5.1.4. Benchmark Configuration	60
5.1.5. Benchmark Metrics	60
6 FUTURE WORK.....	63
7 CONCLUSION.....	66
References	68



University of Moratuwa, Sri Lanka.
 Electronic Theses & Dissertations
www.lib.mrt.ac.lk

LIST OF FIGURES

Figure Index	Name	Page
Figure 1.1	Set of triples	02
Figure 2.1	RDF Storage Example	12
Figure 2.2	Peer to Peer architecture of Apache Cassandra	13
Figure 2.3	Keyspace in Cassandra	15
Figure 2.4	An example RDF graph	18
Figure 2.5	Distributed query processing framework	19
Figure 2.6	The query graph	20
Figure 2.7:	TripleRush index graph that is created for the triple vertex	22
Figure 2.8	Query execution on the relevant part of the index	23
Figure 2.9	A Motivated Example	26
Figure 2.10	Newly Proposed Idea	26
Figure 3.2.1	Usecase diagram for RDF Graph building	31
Figure: 3.3.1	High level class diagram of the Graph based RDF store	32
Figure 3.4.1	Jena Framework Architecture	35
Figure: 3.4.2	Populating data into Cassandra Cluster	36
Figure: 3.4.3	Building the RDF Graph	37
Figure 3.4.4	Querying the RDF Graph	38
Figure 3.4.5	Drop the RDF Store	39
Figure 3.4.6	SPARQL query result in RDF/XML form	40
Figure 3.4.7	XSLT used for transformation	41
Figure 3.4.8	Sample jsp/struts code used to render results	42
Figure 4.1	Web Client Building RDF Graph	46
Figure 4.2	Execute Query against the RDF graph	47
Figure 4.3	Rendering Results of the SPARQL query	48
Figure 5.1.1.1	Multiple resources as RDF/XML	52
Figure 5.1.1.2	Multiple resources as Turtle	53
Figure 5.1.1.3	Multiple Resources as N-Triples	54
Figure 5.1.3.1	Query 1	56
Figure 5.1.3.2	Query 2	56
Figure 5.1.3.3	Query 3	57
Figure 5.1.3.4	Query 4	58
Figure 5.1.3.5	Query 5	59
Figure 5.1.5.1	SPARQL Query Execution time	61
Figure 6.1	Distributed Implementation of the RDF store	64

LIST OF TABLES

Table Index	Name	Page
Table 1	Base tables and bound variables	17
Table 2	Benchmark Configuration	60
Table 3	Performance Benchmark results	60



University of Moratuwa, Sri Lanka.
Electronic Theses & Dissertations
www.lib.mrt.ac.lk

LIST OF ABBREVIATIONS

Abbreviation	Description
CQL	Cassandra Query Language
DBMS	Database Management Systems
RDBMS	Relational Database Management Systems
RDF	Resource Description Framework
SPARQL	RDF Query language



University of Moratuwa, Sri Lanka.
Electronic Theses & Dissertations
www.lib.mrt.ac.lk