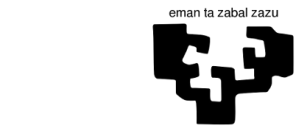
OpenLink Virtuoso as the universal database engine for the semantic web applications

Master Thesis

June 27, 2012

Łukasz Andrzej Grądzki

Supervisor:
Arantza Illarramendi, Ph.D.



Acknowledgements

This part is optional. The following are usually mentioned in the Acknowledgments:

- Supervisor and committee
- Grant support
- Helpful fellow students, lab mates etc.
- Family support

II

Preface

This thesis is submitted in complete fulfilment of the requirements for the author in the Masters's Degree in Advanced Computer Systems on The University of the Basque Country. It contains work done from May to September 2012. The supervisor of the project was Arantza Illarramendi, Ph.D. from The Facuty of Computer Science of San Sebastián. The document of this thesis has been made solely by the author who focused mainly on the analysis of the OpenLink Virtuoso database engine.

IV

Contents

List of Figures

List of Tables

Abstract

You can put an abstract of what the Thesis is about here.

Chapter 1

Introduction

1.1 Motivation

Why do we make the research? Requirements of the Semantics Web Research. Current solutions. The main research papers that are going to be used as the basis for the study. There is no definite guide for the software (feel the gap between the detailed documentation and superficial general-purpose manuals).

1.2 The OpenLink Virtuoso System

The general info about the Virtuoso System. It's properties and applications - general description.

1.3 Objectives

The main objective of the thesis is to analyse and describe the properties of the Virtuoso System that are crucial for the research over the semantic webs.

Chapter 2

Problem analysis

2.1 Semantic Web

2.1.1 Definitions

Semantic Web is a project focused on definition and publication of the standards regarding content descriptions on the Internet. The main objective of these standards is to provide the content in a form convenient for the effective information processing by the software and hardware. The semantic Web standards include OWL (*Web Ontology Language*), RDF (*Resource Description Framework*, Sec. 2.2.2) and RDFS (*RDF Schema*). The meanings of the informational resources are defined using ontologies (Sec. 2.1.2) - this representation are discussed in details in the next section.

Semantic Web was an idea of Tim Berners-Lee who is the chef of W3C, the creator of the WWW standard and the first Web browser. In principle, semantic Web should be based on the existing communication protocols that set up the contemporary Internet. The main difference is that the published data must be also *comprehensible* for the machines. To achieve this objective, the resources are presented in a form that allows to identify their context and the relations between them.

2.1.2 Ontologies

In general, ontology is a formal representation of a knowledge domain, that is composed of sets of concepts and relations between them. This system creates a conceptual schema that provides a description of some domain. In addition, the conceptual schema can be used as a basis to draw conclusions about the properties of the terms described by a given ontology.

According to Yu Juan and Dang Yanzhong[1], ontologies are classified into two groups, according to their expressiveness:

- lightweight ontologies
- heavylight ontologies

The lightweight ontologies, also referred as terminologies, are simple taxonomic structures of concepts with simplified relationships between them. On the other hand, the heavyweight ontologies are composed of concepts, relations and rules that represent ontological commitment explicitly. Sometimes is difficult to clearly assign an ontology to one of these types.

2.1.3 Field of study

2.1.4 Research trends

2.1.5 Data Storage

2.2 Data Storage Formats

2.2.1 Relational DBs

2.2.2 Resource Description Framework

2.3 Overview of existing systems

...

2.3.1 Commercial Platforms

Platform 1

...

2.3.2 OpenSource Platforms

Platform 11

...

2.4 General properties of the Virtuoso System

2.4.1 History of development

2.4.2 Properties of Virtuoso

2.4.3 Main applications

2.4.4 Aspects important for the Semantic Webs

2.5 Used Tools

2.5.1 Installation of Virtuoso

The analysis was based on the newest stable version (ver. 6.1.5) of the OpenLink Virtuoso compiled and installed from the source code obtained from the SourceForge project page (ref. [2]). The software was installed on the following machine:

- Operating System: Linux Ubuntu 12.04 LTS 32-bit version
- Processor: Intel®Core™2 Duo CPU T7500 @ 2.20GHz × 2
- RAM: 2 x 2 GB SODIMM DDR2 Synchronous 667 MHz

In addition, the Virtuoso was configured with the following parameters (settings in the `virtuoso.ini` file):

- `NumberOfBuffers = 170000` (what corresponds to about 1410 MB of RAM)
- `MaxDirtyBuffers = 130000`
- `MaxCheckpointRemap = 42500` (25% of `NumberOfBuffers`)

2.5.2 The LUBM Benchmark

The study required a sample RDF data to be stored and processed by Virtuoso. For this purpose, the LUBM Benchmark (*The Lehigh University Benchmark*, ref. [3]) was used. According to its creators, LUBM was developed to facilitate the evaluation of Semantic Web repositories in a standard and systematic way. The tool evaluates the performance of those repositories with respect to extensional queries over a large data set that commits to a single realistic ontology.

The sample data used in the benchmark is an ontology describing a university structure. It can be dynamically generated in a customizable and repeatable way. In addition, it includes a set of test queries and provides several performance metrics.

To start with, the newest benchmark code (version 1.7) has been downloaded from the LUBM page: [3]. Apart from this, it is necessary to apply a special patch that makes the tool work with Linux paths. Then, the testing ontology data was generated for three universities with the following command:

```
java -cp classes edu.lehigh.swat.bench.uba.Generator -univ 3 \
-index 0 -seed 1234 \
-onto http://www.lehigh.edu/~zhp2/2004/0401/univ-bench.owl
```

According to the above `-onto` parameter, the appropriate ontology was downloaded [4]. Apart from this, the `VirtBulkRDFLoaderScript.vsql` script was prepared and loaded via `isql` to enable the bulk loading of all generated ontologies.

Chapter 3

System analysis

3.1 Internal Structure

3.1.1 SPARQL end-point

3.1.2 Advantages

3.1.3 Special queries and commands

3.2 Key Features

3.2.1 Reasoning

Forward/backward chaining. In essence, Virtuoso's support for subclasses and subproperties is backward chaining, i.e. it does not materialize all implied triples but rather looks for the basic facts implying these triples at query evaluation time.

3.2.2 Import Data Mechanisms

3.2.3 Export Data Mechanisms

3.3 Comparison with similar systems

3.3.1 Virtuoso vs Oracle NoSQL

Chapter 4

Conclusions

The section includes the following: Overall analysis and integration of the research and conclusions of the thesis in light of current research in the field Conclusions regarding goals or hypotheses of the thesis that were presented in the Introduction, and the overall significance and contribution of the thesis research Comments on strengths and limitations of the thesis research Discussion of any potential applications of the research findings An analysis of possible future research directions in the field drawing on the work of the thesis ...

Appendix I

First appendix

Contents

This is the first appendix.

I.a First section

This is the first section of the first appendix.

References

[1] Y. Juan and D. Yanzhong, "A framework of ontology management system," in *E-Business and E-Government (ICEE), 2010 International Conference on*, pp. 1719 –1722, may 2010.

[2] OpenLink, "The SourceForge project page of the OpenLink Virtuoso (Open-Source Edition)." http://sourceforge.net/projects/virtuoso/files/latest/download, 2012. [Online; accessed June 25, 2012].

[3] OpenLink, "The Lehigh University Benchmark." http://swat.cse.lehigh.edu/projects/lubm/, 2012. [Online; accessed June 26, 2012].

[4] OpenLink, "The Lehigh University Benchmark." http://www.lehigh.edu/ zhp2/2004/0401/univ-bench.owl, 2012. [Online; accessed June 26, 2012].

Publications

This is a list of publications.

- **My first article**
  M. Y. Name
  Journal **year**, *volume*, pages.