УДК 519.7: 004.896

Е.Л. Шевченко

Харьковский национальный университет радиоэлектроники, Харьков

ПРАКТИЧЕСКАЯ АПРОБАЦИЯ ФУНКЦИОНАЛЬНЫХ ВОЗМОЖНОСТЕЙ И ПРОИЗВОДИТЕЛЬНОСТИ ХРАНИЛИЩ ОНТОЛОГИЧЕСКИХ БАЗ ЗНАНИЙ

Приведен детальный сравнительный анализ наиболее перспективных RDF-хранилищ. Сравнению подлежали функциональные и нефункциональные характеристики, а также производительность фреймворков на тестовом наборе, по размерам и вариантам использования максимально приближенном к реальным задачам. В результате проведенного анализа даны рекомендации по выбору соответствующего инструментария.

***Ключевые слова**: онтологические системы, системы управления базами знаний, триплет, OWL, RDF, SPARQL.*

Введение

На смену существующей парадигме построения RDBMS, использующей реляционную модель представления данных и методы работы с ними, приходят системы, построенные по другому принципу. В их основе лежит применение семантических технологий, таких как использование онтологической модели представления знаний, методов логического вывода на них, методов получения выборок, соответствующих клиентским запросам. Это направление хорошо финансируется, результатом чего явилась разработка множества систем управления базами знаний, реализующих подходы SW (RDF-хранилищ). В настоящий момент RDBMS характеризуются большей скоростью обработки и компактностью хранения полезной информации. Онтологические же системы предоставляют более гибкий механизм описания модели хранимых знаний [1], её последующего расширения, хранение знаний в неявном виде и т.д. (распределенные хранилища). Использование в современных высокоэффективных производственных системах предъявляет к базам знаний жесткие требования по скорости и эффективности работы. В данной статье произведено сравнение RDF-хранилищ, сделана оценка возможности их использования в высоконагруженных информационных системах.

Анализ последних исследований и публикаций. *Выбор RDF-хранилищ, подлежащих тестированию.* В настоящий момент в индустрии SW можно перечислить множество фреймворков, потенциально претендующих на роль ядра системы управления базами знаний. Они различаются архитектурой хранения базы фактов, набором поддерживаемых интерфейсов взаимодействия с клиентом, мощностью используемого метода логического вывода неявных фактов при ответе на запросы пользователя. Различие подходов в реализации приводит к отличиям в функциональных возможностях и производительности систем. Также иногда один и тот же производитель предоставляет две и более версии своего фреймворка, отличных по функциональным возможностям и по лицензии, определяющей условия использования.

В данном исследовании при выборе фреймворков для анализа автора интересовали следующие факты, необходимые для успешности потенциальной *системы управления базами знаний*: наличие бесплатной для некоммерческого использования версии фреймворка, поддержание взаимодействия с API фреймворка как на уровне динамического линкования так и на уровне высокоуровневых сетевых протоколов взаимодействия, активное состояние разработки, наличие развернутой документации и поддержки разработчиков, активное сообщество пользователей. Под выделенный перечень критериев попали следующие продукты: 4store, Sesame/OWLIM (для исследователей), BigData, Jena TDB, Virtuoso.

Обоснование метода тестирования. К настоящему моменту разработано несколько тестовых наборов для оценки производительности RDF-хранилищ. Среди них общедоступными являются: Berlin SPARQL Benchmark (BSBM), Lehigh University Benchmark (LUBM), Ontology Benchmark (UOBM), DBpedia SPARQL Benchmark. Все они направлены на тестирование RDF-хранилищ, интерфейсная часть которых поддерживает язык запросов SPARQL. Тестовые наборы традиционно состоят из: **а)** тестовой онтологии (набор фактов заданных явно и в виде правил логического вывода), посвященной определенной предметной области; **б)** набора SPARQL запросов на выборку; **в)** четко оговоренного набора метрик, для контроля которых предназначен тестовый набор.

Описываемая идея тестирования систем управления базами знаний на основе реалистичной рабочей нагрузки, созданной путем выполнения запросов, соответствующих различным сценариям использования тестируемого хранилища на практике, заимствована и хорошо зарекомендовала себя при сравнении и анализе RDBMS. Это делает её обоснованной для решения подобной задачи при сравнении RDF-хранилищ.

Рассмотрим более детально структуру тестового набора на примере наиболее задокументированного BSBM. BSBM посвящен классической для RDBMS предметной области - электронной коммерции (товары, поставщики, покупатели). SPARQL запросы в BSBM разбиты на три потенциальных практических сценария: **а)** простые запросы выборки покупателя, ищущего определенные товары; **б)** запросы на выборку и обновление набора фактов и правил (расширение предыдущего тестового сценария); **в)** запросы бизнес-аналитики, симулирующие как различные заинтересованные стороны могут изучать ситуацию на рынке по имеющемуся набору фактов.

Среди общепринятых тестовых метрик в BSBM выделяют:

Метрики для одного запроса

– Average Query Execution Time (aQET) – среднее время исполнения одного запроса определенного типа (запрос выполняется множество раз с различными параметрами);

– Queries per Second (QpS) – среднее кол-во запросов определенного типа, которые удалось выполнить за секунду;

– Min/Max Query Execution Time (minQET, maxQET) – наименьшее и наибольшее время выполнения запроса определенного типа;

Метрики для набора запросов

– *Queries Mixes per Hour (QMpH)* – кол-во запросов с различными параметрами, которые возможно исполнить за час;

– *Overall Runtime (oaRT)* – время выполнения всего набора тестов;

– *Composite Query Execution Time (cQET)* – среднее время выполнения смешанного по типу тестового набора запросов, вычисленного на основе множества повторений этого набора с различными параметрами;

– *Average Query Execution Time over all Queries (aQEToA)* – сумма времен запуска N смешанных тестовых наборов, деленных на кол-во запросов в каждом наборе.

Метрики для оценки "конкурентоспособности" системы (сочетания цена/качество)

– *$/QMpH, где $* – средняя цена установки и эксплуатации системы в течение 5-ти лет.

В данной работе для тестирования был выбран BSBM v3.1 от 26.08.2011 г. Как LUBM и UOBM этот тестовый набор является синтетическим. Но в отличие от своих более ранних аналогов, он наиболее сбалансирован, на искусственных данных позволяет смоделировать реальную нагрузку и близкие к реальным сценарии использования, в которых несколько клиентов одновременно выполняют осмысленные, мотивированные запросы к системе. Также тест содержит запросы, контролирующие возможности SPARQL 1.1, такие как группировка и агрегация.

Описание эксперимента

При разработке любой информационной системы основными определяющими факторами выбора той или иной системы хранения данных являются: необходимое стороннее ПО, перечень предоставляемых функциональных возможностей, скорость обработки клиентских запросов, компактность хранения информации. Рассмотрим выбранные инструментальные средства подробнее.

Качественный анализ

4store – это RDF-хранилище, разработанное Garlik Inc. Оно реализовано на ANSIC99 и доступно по GNU General Public License (GPL). Доступно для Unix-подобных систем и запускается как сервер на одной машине или в режиме кластера на 64-битных машинах. Клиентские библиотеки для связи с 4store через API доступны на многих языках программирования: PHP, Ruby, Python, Java. Механизмы кастомизации сервера не предусмотрены.

Jena TDB – является компонентом Jena Semantic Web framework и доступно как OpenSource ПО по лицензии BSD. Оно может быть установлено как на 32, так и на 64-битные системы с поддержкой Java. Обращение возможно через основанное на java API, через утилиты командной строки. Joseki - это веб-сервер, поддерживающий язык запросов SPARQL для обращения к RDF-данным в Jena.

BigOWLIM – это коммерческое RDF-хранилище семейства OWLIM от компании Ontotext, разработанное для больших объемов данных (биллионы триплетов). Бесплатное использование сервиса допускается в исследовательских проектах и для проведения тестирования.

Sesame – это OpenSource фреймворк (Aduna BSD лицензия) для хранения, логического вывода и обработки пользовательских запросов к RDF-данным. Структура Sesame включает RDF-парсер и генератор (Sesame Rio), слой хранения и логического вывода (storage and inference layer – SAIL API), который абстрагирует от деталей хранения, API для обработки RDF и HTTP-сервер для доступа к репозиторию по HTTP. Sesame может быть установлен на любой системе с поддержкой Java, и доступен через java API сторонним приложениям. Имеет множество расширений и плагинов от сторонних разработчиков. Хранение RDF может быть настроено в двух вариантах: собственная реализация или отображение на внешнюю по отношению к Sesame реляционную БД.

В связке Sesame/BigOWLIM последний используется как SAIL.

BigData – это кластерное RDF-хранилище. Доступно по лицензии GNU GPL для UNIX систем. Java-клиентом как и в случае OWLIM является Sesame. Дальнейшее масштабирование достигается путем внедрения плагинов динамически в режиме runtime в большинство сервисов. Плагины самостоя-

тельно регистрируются через централизированный сервис и начиают управлять данными.

OpenLink Virtuoso Universal Server – это гибридное хранилище для многих моделей представления данных: реляционные, RDF, XML, текст. Поэтому может также использоваться как средство отображения из RDF в другие форматы или как средство интеграции гетерогенных источников. Основной функционал предоставляется как OpenSource, также существует несколько вариантов коммерческих лицензий для сборок с расширенной функциональностью, например, обработка геопространственных данных и т.д. Virtuoso используется ка хранилище для многих важных Linked Data источников, например, DBpedia. Устанавливается как серверная служба на все основные платформы.

Для упрощения проведения анализа составлены краткие характеристики выбранных инструментальных средств, которые затем были разделены на несколько сравнительных таблиц, позволяющих делать выбор конкретного инструмента в зависимости от особенностей решаемой задачи. В табл. 1 выделены основные, по мнению автора, нефункциональные характеристики анализируемых продуктов. Жирным шрифтом отмечены параметры фреймворков, которые превосходят конкурентов по тому или иному критерию сравнения.

Таблица 1

Сравнение нефункциональных особенностей RDF-хранилищ

Инструмент. средство	OS	Способ хранения информации	Лицензия	Допускает клиентов	Последний релиз
4Store	Linux Mac	Native Store	GPL	**shell, SPARQL http, REST, PHP, Ruby, Python, Java, Django API**	v1.1.4 2011 г.
Joseki/Jena TDB	Indep	Native store	BSD	shell, SPARQL http, REST, Java	v2.7 2011 г.
Sesame/ BigOWLIM	Indep	Native store	commerc., free of charge for research	SPARQL HTTP, REST, java	OWLIM v4.3 Sesame v2.6 2011 г.
BigData	Linux	Hybrid (Native, RDBMS) Store, Custom rules	GPL	SPARQL HTTP, REST, java	**v1.0.6 2012 г.**
Virtuoso	**Win, Linux, Mac, Solaris**	Hybrid (Native, RDBMS) store	GPL	SPARQL HTTP, REST, java	v6.1.4 2011 г.

В табл. 2 выделены основные, по мнению автора, функциональные характеристики RDF-хранилищ. Данные про максимальный объем загружаемой онтологии получены из [2] и обозначают кол-во триплетов, которые возможно загрузить в систему "явно", размер загруженной онтологии с учетом кол-ва неявно выведенных триплетов не анализировался.

Здесь и далее в сравнительных таблицах используется введенный автором параметр "выигрыш". Он отражает кол-во сравниваемых характеристик, по которым инструментальное средство показало преимущество по сравнению со всеми своими конкурентами. Выигравшие характеристики выделены жирным шрифтом.

Таблица 2

Сравнение функциональных возможностей RDF-хранилищ

Инструментальное средство	Макс. объем загружаемой онтологии (явный)	Языки запросов	Поддержка логического вывода	Поддержка изменения данных	Поддержка транзакций	Поддержка репликаций	Поддержка резервного копир./ восстановл.
4store (0 выигр.)	15 млрд.	SPARQL	нет	Files import	нет	да сегменты	БД / File system
Joseki/Jena TDB (2 выигр.)	1,7 млрд.	SPARQL	**RDFS, OWL, DAML, Rules**	**Files import, SPARQL INSERT**	нет	нет	нет
Sesame/ BigOWLIM (1 выигр.)	12 млрд.	SPARQL SeRQL	RDFS, Custom Rules	Files, API	**да**	нет	нет
BigData (1 выигр.)	12,7 млрд.	SPARQL SeRQL	RDFS, OWL (limited)	Files import	**да**	да БД	БД / File system
Virtuoso (5 выигр.)	**15,4** млрд.	SPARQL SQL	RDFS, OWL (limited)	**Files import, SPARQL INSERT**	**да**	**да, Snapshot, Transactional**	**да**

Проведенное сравнение позволяет выделить Virtuoso как наиболее полнофункциональное RDF-хранилище, которое может быть использовано на большем количестве современных ОС.

Хотя нельзя не заметить, что Jena в настоящий момент наиболее полно реализует теоретически разработанные стандарты-концепции SW (логический вывод в OWL, поддержка последних версий языков запросов).

Анализ производительности

Для анализа производительности RDF-хранилищ приведем результаты их тестирования на основе BSBM. Для тестирования использовалась следующая аппаратная конфигурация: процессор Intel i7 950, 3.07GHz (4 ядра), ОЗУ 24GB, жесткий диск 2 x 1.8TB SATA2.

Используемое ПО: Ubuntu 10.04 64-bit, Kernel 2.6.32-24-generic, файловая система ext4, Java v. 1.6, OpenJDK.

Стратегия тестирования включает оценку времени загрузки модели (табл. 3) в хранилище и время выполнения запросов в автономном (табл. 4) и конкурентном (таблица 5) режимах [3].

Однозначно лидирует BigOWLIM.

Таблица 3

Сравнение производительности RDF-хранилищ
при начальной загрузке и индексировании данных

	4store	**Joseki/Jena TDB**	**Sesame/BigOWLIM (2 выигрыша)**	**BigData**	**Virtuoso**
BSBM 100M	26:42	1:14:48	**17:22**	1:03:47	1:49:26
BSBM 200M	1:12:04	2:45:13	**38:36**	3:24:25	3:59:38

Таблица 4

Сравнение производительности RDF-хранилищ
при выполнении пользовательских запросов к базе знаний BSBM200

	4store	**Joseki/Jena TDB**	**Sesame/ BigOWLIM**	**BigData**	**Virtuoso**
Кол-во выигрышей	**3**	**0**	**2**	**0**	**7**
Q1	145,3	62,2	45,0	64,4	**163,0**
Q2	55,7	44,1	**111,6**	35,3	73,8
Q3	122,9	66,1	48,1	14,4	**195,4**
Q4	62,9	47,3	37,8	40,0	**94,8**
Q5	5,0	1,2	1,8	1,7	**9,3**
Q6	не выполнялся	не выполнялся	не выполнялся	не выполнялся	не выполнялся
Q7	**49,3**	15,0	11,0	21,9	15,4
Q8	**57,1**	15,9	12,6	14,1	22,5
Q9	117,3	97,1	67,9	44,6	**160,1**
Q10	52,8	26,4	22,4	28,8	**69,9**
Q11	33,3	23,4	18,7	26,7	**39,5**
Q12	**40,4**	28,3	29,1	31,1	39,8

Числовые показатели в табл. 4 – усредненный результат для 500 запусков запросов одного типа, единицы измерения – кол-во выполненных запросов за секунду (QpS).

Эксперимент показывает, что ни один из фреймворков не имеет приоритета перед остальными по всем типам запросов.

Хотя можно отметить закономерность, что для сложных выборок преимущество делят между собой **4store** и **Virtuoso**.

Упорядоченность по производительности, приведенная в табл. 5, сохраняется и для набора BSBM 100M.

Таблица 5

Сравнение производительности
RDF-хранилищ при различной нагрузке

RDF-хранилище	**Кол-во клиентов**			
	1	**4**	**8**	**64**
4store	4593	*	*	*
Jena TDB	1443	2206	1474	*
BigOwlim	1795	3713	4041	3622
BigData	1795	3040	3167	2689
Virtuoso	**4669**	**13265**	**18264**	**16564**

Это дает основания охарактеризовать **Virtuoso,** как наиболее быстрый сервер баз знаний, широко поддерживающий логический вывод неявных фактов и все основные функциональные возможности RDBMS. * в ячейке означает технические проблемы в хранилище, остальные показатели указаны в QMpH.

В дополнение приведем интересные факты по оценке необходимых объемов дискового пространства и оперативной памяти для обработки RDF.

По оценкам разработчиков Virtuoso [4], в случае, если при обработке запросов пользователя задействовано 4 индекса (без необходимости специфицировать граф, к которому производится запрос) каждый триплет представляется 39 байтами. При этом (при учете еще и полнотекстового индекса по всем литералам) файл базы знаний на миллиард (1115M) триплетов занимает, примерно, 120 Гб на жестком диске.

Для комфортной работы сервера при одновременной обработке 500M триплетов рекомендуется около 16Гб ОЗУ. В случае, если триплеты короткие и повторяющиеся (как в данных тестового набора LUBM), 16Гб может быть достаточно для обработки одного миллиарда триплетов. Все также зависит от сложности SPARQL запроса. Запросы, использующие аналитику, последовательности соединений, агрегации по всему набору увеличивают потребность в оперативной памяти.

Выводы

Исходя из всего вышесказанного, современные rdf-хрнилища, построенные на онтологическом принципе, вполне могут обеспечить потребности небольших производственных баз знаний, ограниченных объемом в 15 миллиардов триплетов. При этом скорость выборки информации из такого хранилища сохраняется на приемлемом уровне при исполнении простых запросов выборки, эквивалентных сценариям BSBM. По результатам эксперимента, в настоящее время наиболее эффективным онтологическим хранилищем можно считать Openlink Virtuoso, обеспечивающий максимальную производительность при наиболее полной реализации стандартов SW (логический вывод на основе OWL и SPARQL v1.1) и всех базовых возможностей RDBMS.

В дальнейшем представляет интерес тестирование определившихся лидеров линейки RDF-хранилищ на естественном, а не синтетическом наборе, например, DBpedia. Для этого необходимо формирование, на основе онтологии DBpedia реалистичного набора сценариев использования (запросов) с возможностью проверки полноты и точности ответов, формируемых RDF-хранилищем. Выполнение такого сравнения позволит получить более точные результаты на задаче, максимально приближенной к реальности.

Список литературы

1. *Шевченко А.Ю. Модели распределенной онтологической базы знаний для интеллектуальных информационных систем [Текст] / А.Ю. Шевченко, Е.Л. Шевченко // Системи обробки інформації. – Х.: ХУПС, 2010. – Вип. 6. – С. 25 – 29.*

2. *Large Triple Stores / W3C WIKI. Режим доступа: www / URL: http://www.w3.org/wiki/LargeTripleStores. – 21.03.2012. – Загл. с экрана.*

3. *BSBM V3 Results (February 2011) / Prof.'s Dr. Christian Bizer Home page, Institut für Wirtschaftsinformatik. – Режим доступа: www / URL: http://www4.wiwiss.fu-berlin.de/bizer/BerlinSPARQLBenchmark/results/V6/index.html. 21.03.2012. – Загл. с экрана.*

4. *Virtuoso 6 FAQ / официальный сайт компании Openlink Software. – Режим доступа: www / URL: http://virtuoso.openlinksw.com/dataspace/dav/wiki/Main/VOSVirtuoso6FAQ. 21.03.2012. – Загл. с экрана.*

Поступила в редколлегию 30.10.2012

Рецензент: д-р техн. наук, проф. Е.И. Кучеренко, Харьковский национальный университет радиоэлектроники, Харьков.

ПРАКТИЧНА АПРОБАЦІЯ ФУНКЦІОНАЛЬНИХ МОЖЛИВОСТЕЙ ТА ПРОДУКТИВНОСТІ СХОВИЩ ОНТОЛОГІЧНИХ БАЗ ЗНАНЬ

О.Л. Шевченко

Наведено детальний порівняльний аналіз найбільш перспективних RDF-сховищ. Порівнювалися функціональні та нефункціональні характеристики, а також продуктивність фреймворків на тестовому наборі, за розмірами та варіантами використання максимально наближеному до реальних задач. У результаті наведеного аналізу надані рекомендації з вибору відповідного інструментарію.

***Ключові слова**: онтологічні системи, системи управління базами знань, триплет OWL, RDF, SPARQL.*

BANCHMARKING FUNCTIONAL AND PERFORMANCE CAPABILITIES OF THE SEMANTIC STORAGES

E.L. Shevchenko

Detailed comparison of the most perspective RDF-stores is provided. Nonfunctional and functional characteristics as well as tool's performance with test triple set that is similar to real tasks by size and test cases have been analyzed. Recommendations to choose the particular tool are given as a result of comparison.

***Keywords**: ontological systems, control the system by the bases of knowledge's, triplet OWL, RDF, SPARQL.*