

УДК 004.658.2

Обзор постреляционных СУБД. AllegroGraph

*Супрун Д.Е., студент,
Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана
кафедра «Системы обработки информации и управления»*

*Научный руководитель: Ревунков Г.И., к.т.н., доцент
Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана
chernen@bmstu.ru*

1. AllegroGraph. Описание

AllegroGraph - это современная, высокопроизводительная, устойчивая графическая база данных. AllegroGraph совмещает эффективное использование памяти и хранения на основе жестких дисков, позволяя выполнять миллиарды quads (квадрациклов), сохраняя при этом высокую производительность. AllegroGraph поддерживает SPARQL, RDFS ++ и Prolog reasoning многочисленные клиентские приложения. AllegroGraph разрабатывает фирма FranzInc, которая широко известна благодаря таким продуктам, как TopBraidComposer (редактор онтологий), RacerPro (OWL DL reasoning) и др. Поддерживает 2 основных типа вывода:

1. RDFS++

Имеются в виду следующие предикаты: `rdf:type` (сокращённая запись для свойства), `rdfs:subClassOf` (отношение наследования между классами), `rdfs:range` (возможные значения), `rdfs:domain`, `rdfs:subpropertyof`, `owl:sameAs` (`owl` – язык веб-онтологий), `owl:inverseOf`, `owl:TransitiveProperty`, `owl:hasValue`.

2. RacerPro

AllegroGraph интегрирован с еще одним блоком мышления FranzInc - RacerPro, который имеет выразительность OWL DL (*Description Logic*). OWL DL рассчитано на пользователей, которым необходима максимальная степень выразительных возможностей языка без потери вычислительной полноты, без потери ни одного из семантических воплощений - содержательных толкований выводов, полученных формально-логическим

путем и разрешимости, что означает, что вычисления, будут закончены за конечное время. Уровень OWL DL ориентирован на существующие сегодня системы описания знаний и системы логического программирования.

Итак, AllegroGraph является RDF-хранилищем с поддержкой двух reasoning с полярными характеристиками.

FranzInc также разрабатывает проект AGWebView, который представляет собой Web-интерфейс для AllegroGraph.

2. AllegroGraph. Архитектура

AllegroGraph 4.12 обеспечивает архитектуру протокола REST (существенно расширенный набор Sesame HTTP клиент). Поддерживаются адаптеры для различных языков: Sesame Java, Sesame Jena, Python с использованием Sesame и Lisp (фреймворки). Есть Open Source адаптеры для проектов C #, Ruby, Clojure, Scala и Perl. Архитектура AllegroGraph представлена на рисунке 1.

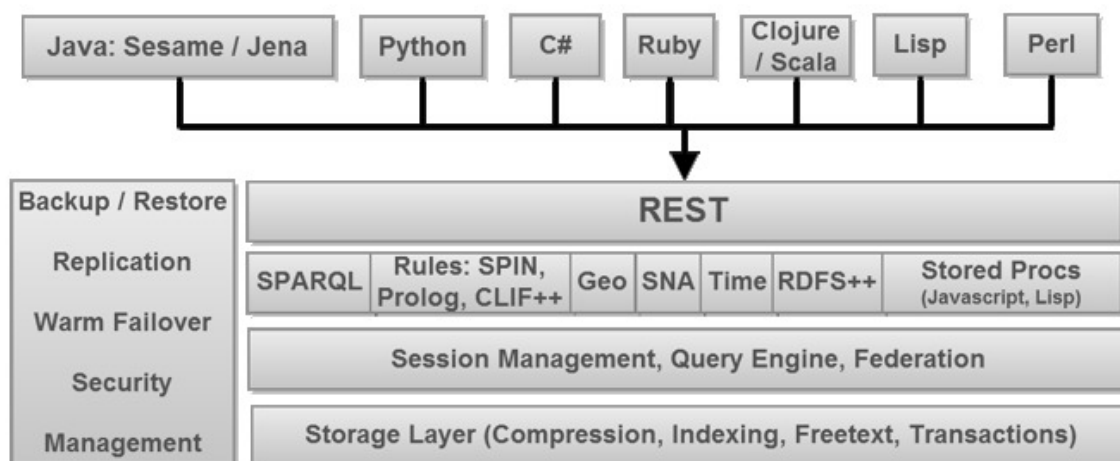


Рис. 1. Архитектура AllegroGraph

Поддерживаемые фреймворки: Jena, Sesame, Clojure - это язык программирования общего назначения с поддержкой разработки в интерактивном режиме, поощряющий функциональное программирование, и упрощающий поддержку многопоточности.

3. AllegroGraph. Технические особенности

3.1. Высокоэффективное хранение

AllegroGraph предназначен для максимальной скорости загрузки и скорости запроса. Загрузка квадранциков, через оптимизированный RDF / XML и N-Quads анализаторов, является бестселлером в своем классе, особенно с большими файлами. С 2004 года (версии 1.0) AllegroGraph стал первым продуктом, способным загружать и индексировать миллиард триплетов с использованием стандартных x86 64-разрядного оборудования. NoSQL Graph Database продолжал внедрять инновации (2008 Semantic conference) 10 млрд. квадранциков загружены на EC2 Amazon. Способность AllegroGraph автоматического управления доступом ко всем аппаратным ресурсам для максимизации загрузки, индексации и запросов повысила планку производительности для хранения RDF (среда описания ресурса). В таблице 1 представлены примеры выполнения AllegroGraph загрузки и индексации.

Таблица 1

AllegroGraph. Результаты тестов по загрузке и индексации

Испытание на нагрузку	Триплеты	Время	Нагрузка Тариф (Т / сек)
LUBM (8000) *	1.106 млрд.	36мин, 49 сек	500679
LUBM (160000) *	22.12 млрд.	12 часов, 18 м, 16s	499188
AllegroGraph Предварительная версия **	310,269 млрд.	78 ч., 9, 23s	1102737
AllegroGraph Предварительный выпуск ***	1,009 триллиона	338 ч., 5 м	829556

* 32 ядро Intel E5520, 2,0ГГц, с 1 RAM ТБ, RedHat v6.1.

** 64 ядро Intel X7560, 2,27 ГГц, 2 ТБ оперативной памяти, 22ТБ диска, Redhat v6.1.

LUBM-подобные данные.

*** 240 ядро Intel X5650, 2,66 ГГц, 1.28ТБ ОЗУ, 88ТБ диска, Redhat v6.1. LUBM-подобные данные.

3.2. Системные требования

AllegroGraph v4 Сервер БД работает на всех Linux x86-64 бит. Для запуска AllegroGraph v4 на других операционных систем (например, Windows, Mac) предлагается создать виртуальную машину Linux или EC2 (см. таблица 2). ОС AllegroGraph может быть любой: либо 32-битные или 64-бит.

Таблица 2

AllegroGraph v4.x. Виртуальная машина

AllegroGraph v4.x	Виртуальная машина Appliance (Требуется 64-разрядное оборудование)
Linux (x86-64), Glibc 2.4 и выше	Apple Mac OSX 10.6 и более новые версии
Amazon EC2 (Linux x86-64)	32 или 64 бит Microsoft Windows

Виртуальная машина Appliance позволит запустить версию AllegroGraphLinux в операционной системе Windows или Mac. Производительность будет медленнее, чем при запуске изначально, поэтому рекомендуется изначально установить AllegroGraph для оценки эффективности.

AllegroGraph 4 работает на Linux x86-64 бит. Для запуска AllegroGraph 4 в других операционных системах (например Windows, Mac) предлагается создать виртуальную машину Linux для экспериментов. Клиенты к серверу AllegroGraph может быть либо 32-битные или 64-битный.

Для 64-битной Mac, Windows и Solaris и всех 32-битных система предлагается использовать AllegroGraph 3.3.

4. AllegroGraph. Программирование

AllegroGraph является основой баз данных и приложений для создания семантических веб-приложений. Он может хранить данные и метаданные, как триплеты; запрашивать эти триплеты через различные API, запросы, как SPARQL (стандарт W3C языка запросов) и Prolog; применить RDFS + reasoning (мышление) с помощью встроенного в него блок мышления. AllegroGraph включает поддержку Federation (Федерации), Social Network Analysis (анализ социальных сетей), Geospatial capabilities (геопространственные возможности) и Temporal reasoning (временное мышление).

4.1. Семантическая Сеть

AllegroGraph представляет собой базу данных. Рассмотрим пример на рисунке 2. Здесь представлены некоторые сведения о Jans и его домашних животных, *Robbie is petOfJans* в данном случае явные отношения, (как и большая часть данных в Интернете), *petOf* обратная связь с *hasPet* и собака *subClassOf* млекопитающее - неявные отношения.

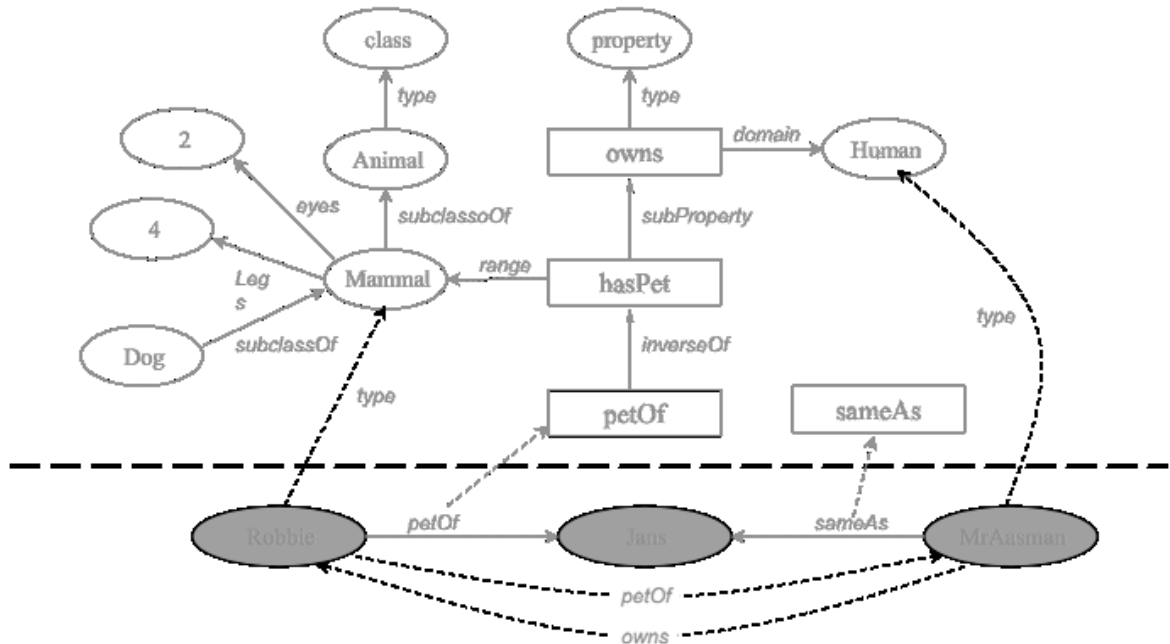


Рис. 2. БД «Домашние животные»

Для хранения такой информации существует несколько способов W3C стандартизированной в Resource Description Framework (RDF). Эти утверждения называются *триплетом*. Если у нас есть много триплетов из разных контекстов, мы можем добавить дополнительный слот для каждого утверждения, который будет называться *по имени графа*. Вот утверждения, соответствующие триплетам:

subject	predicate	object	graph
jans	Type	Human	jans's home page
robbie	petOf	jans	jans's home page
petOf	inverseOf	hasPet	english grammar
Dog	subClassOf	Mammal	science

В семантических сетях веб-страницы содержат достаточное количество описаний данных для того, чтобы машина смогла управлять ими так же легко, как это сейчас делают люди. AllegroGraph представляет собой высокопроизводительную базу данных для

хранения информации и её запрашивания. На сегодняшний день AllegroGraph становится мощной графоориентированной базой данных.

4.2. Блок-схема базы данных AllegroGraph

Рассмотрим более подробно структуру AllegroGraph, представленную на рис.3:

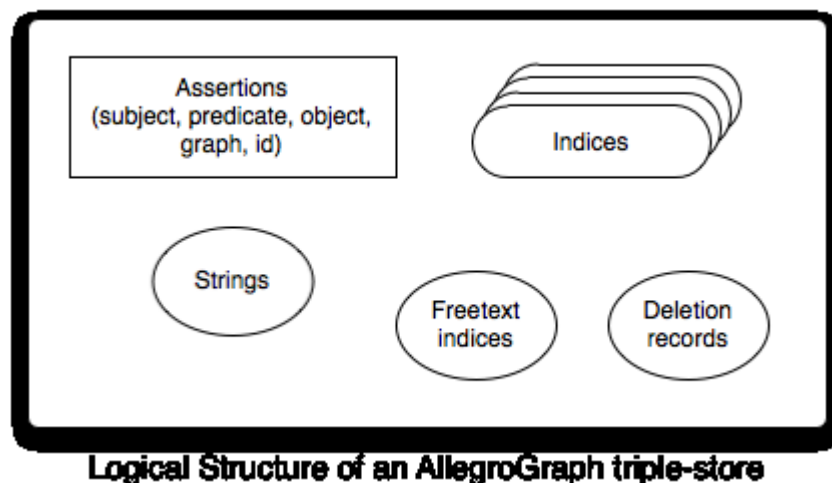


Рис. 3. Структура AllegroGraph

В RDF-Iend, assertion (assertion– используемый код, с помощью которого программа проверяет правильность своего выполнения), говорящее, что `subject predicate object (in the context of graph)`.

Блок Assertions в AllegroGraph состоит из пяти полей:

- Subject (заголовок) (s);
- Predicate (предикат) (p);
- Object (объект) (o);
- Graph (граф) (g);
- ID (id триплекса) (i).

Все s, p, o и g являются строками произвольного размера. Было бы очень неэффективно хранить все повторяющиеся строки, поэтому существует специальный номер (так называемый Unique Part Identifier или UPI). Таким образом, String Dictionary управляет этими строками, а UPI предотвращает дублирование.

Для ускорения запросов, AllegroGraph создает Indices (индексы), которые содержат утверждения, а также дополнительную информацию.

AllegroGraph может также выполнять поиск FREETEXT в утверждениях (для поиска документов, содержащих в столбце произвольные сочетания указанных слов), используя свои показатели FULLTEXT indexes.

AllegroGraph отслеживает удаленные триплеты.

4.2.1. Формат данных N-triple

Триплеты в целом могут вступать в AllegroGraph как строки, также из RDF / XML (см. пример) или в развернутом простом N-triple формате (см. пример). Имеется возможность легко импортировать данные из СУБД, CSV или любого другого формата пользовательского данных.

```
<http://www.franz.com/simple#Animal><http://www.w3.org/1999/02/22-rdf-syntax-ns#type><http://www.w3.org/2002/07/owl#Class> .
<http://www.franz.com/simple#Mammal><http://www.w3.org/2000/01/rdf-schema#subClassOf><http://www.franz.com/simple#Animal> .
<http://www.franz.com/simple#Mammal><http://www.franz.com/simple#eyes> "two".
```

4.2.2. RDF / XML формат данных

```
<?xml version="1.0"?><RDF
xmlns="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xml:base="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
xmlns:owl="http://www.w3.org/2002/07/owl#"><Description
rdf:about="http://www.franz.com/simple#Animal"><rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Class"/></Description><Description
rdf:about="http://www.franz.com/simple#Mammal"><rdfs:subClassOf rdf:resource="http://www.franz.com/simple#Animal"/><ns1:eyes>two</ns1:eyes></Description></RDF>
```

5. AllegroGraph и Сеть

Sesame 2.0 HTTP interface

AllegroGraph в Sesame 2.0 HTTP-сервер работает на любом сервере Java или с любым Lisp изображением, загруженным в AllegroGraph.

Протокол HTTP Sesame был расширен для того, чтобы появились дополнительные функции AllegroGraph, такие как индексирование и кодирование типов данных (эти расширения описаны в протоколе HTTP документации).

<http://sntbul.bmstu.ru/doc/699509.html>

Specialized Data types

AllegroGraph поддерживает несколько специализированных типов данных для эффективного хранения, манипуляции и поиска социальных сетей, геопространственной и временной информации.

Social Network Analysis

Просматривая взаимодействия - соединения в графе, мы можем рассматривать множество различных ситуаций с использованием инструментов Social Network Analysis(SNA). SNA позволяет нам отвечать на вопросы типа:

- Как тесно связаны какие-либо два человека?
- Каковы основные группы или кластеры в данных?
- Насколько важен этот человек (или компания) в потоке информации?
- Какова вероятность того, что эти два человека знают друг друга?

На рисунке 5 показана область с различными математическими методами и алгоритмами. SNA инструментарий AllegroGraph включает массив методов поиска, инструменты для измерения значений, building blocks для специальных целей.

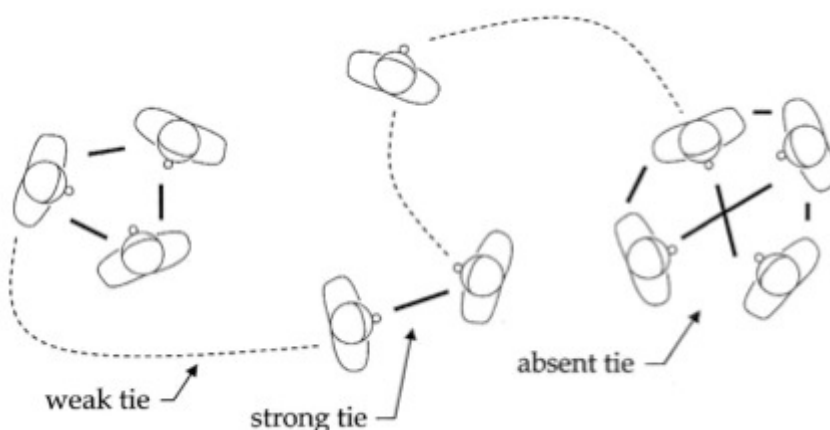


Рис. 5. SocialNetworkAnalysis

GeospatialPrimitives

AllegroGraph обеспечивает новый механизм для эффективного хранения и поиска геопространственных данных. Поддержка предоставляется как для декартовых систем координат, так и для сферических (например, поверхность земли или небесной сферы).

Координаты в двух измерениях кодируются в одной UPI. После того, как данные были закодированы таким образом, AllegroGraph может очень быстро выполнять следующие виды запросов как в декартовых, так и в сферических координатах:

- bounding-box (ограничительная рамка): возвращает курсор, который выполняет итерацию по всем триплетам на данном прямоугольном участке.

- `center/radius` (центр / радиус): возвращает курсор, который выполняет итерацию по всем триплетам в круговой области.

Геопространственные приложения AllegroGraph также имеют поддержку для быстрого определения многоугольников:

- лежит ли точка внутри или вне заданного многоугольника;
- перекрываются ли два многоугольника;
- получение всех триплетов, которые лежат внутри данной многоугольника.

TemporalPrimitives

AllegroGraph поддерживает эффективное хранение и поиск временных данных, (включая `DateTimes`) моменты времени и интервалы времени:

- `datetimes in ISO8601 format: "2008-02-01T00:00:00-08:00"`
- `time points: ex:point1, ex:h-hour, ex:when-the-meeting-began, etc`
- `time intervals: ex:delay-interval (say, from point ex:point1 to ex:h-hour)`

После того, как данные были закодированы, приложения могут выполнять запросы, связанные широким спектром временных ограничений на данные, в том числе отношений между:

- `points and date times`
- `intervals and datetimes`
- `two points`
- `two intervals`
- `points and intervals`

Free text Indexing

AllegroGraph можете построить FREETEXT индексы для строчных объектов, связанных с набором предикатов, которые будут указаны. Учитывая индекс FREETEXT, можно искать текст с помощью:

- `boolean expressions ("market" и "housing")`
- `wild cards ("science*" или "math*")`
- `phrases ("SemanticWebsearch")`

Конечно, FREETEXT индексации замедляет скорость вставки триплетов. Эксперименты показывают, что уменьшение где-то между 5 и 25% зависит от количества участвующих предикатов и виды строковых данных в данном приложении.

Программирование с AllegroGraph

AllegroGraph работает с несколькими языками программирования и средами.

Java

Клиентский интерфейс Java реализует большинство Sesame и Jena интерфейсов для доступа к хранилищам удаленных RDF.

HTTP

На данный момент есть возможность для веб-разработчиков и программистов взаимодействовать с AllegroGraph 4.12.1, полностью используя протокол RESTful HTTP (GET, PUT, POST) для добавления и удаления триплетов, для запроса отдельных триплетов, создания SPARQL, PrologSesam 2.0 HTTP-интерфейсом в некоторых расширениях.

Python

API для Python предлагает удобный и эффективный доступ к серверу AllegroGraph на основе приложения Python. Этот API предоставляет методы для создания, запросов и сохранения данных RDF, для управления сохраненными триплетами.

Lisp

Программисты Lisp могут открыть и использовать triple-store изнутри Lisp. Lispers могут создавать приложения в тот же образ, что и сервер AllegroGraph, управлением или использовать remote-triple-store для доступа к данным в режиме клиент / сервер.

Список литературы

1. Мартин Фаулер, Прамодкумар Дж. Садаладж NoSQL: новая методология разработки нереляционных баз данных = NoSQLDistilled. — М.: «Вильямс», 2013. — 192 с.
2. Леонид Черняк Смутное время СУБД // Открытые системы. — 2012.
3. Robinson, I. and Webber, J. and Eifrem, E. Graph Databases. — O'Reilly Media, Incorporated, 2013. — 178 с.
4. Franz Inc AllegroGraph 4.13 // Semantic Web Technologies. <http://www.franz.com/agraph/allegrograph/> (дата обращения: 21.01.2014г.).