Распределенная обработка «больших» данных в семантических NoSQL-приложениях

Докладчик:
к.т.н, доц., с.н.с.
Щербак Сергей Сергеевич

Рассматриваемые вопросы

- Big Data
- Linked Data
- NoSQL
- Endpoints
- Sparql
- Facebook Open Graph

Buzzword -Big Data

"Большие данные: новый рубеж для инноваций, конкуренции и производительности" — McKinsey Institute

"Большие данные объединяют техники и технологии, которые извлекают смысл из данных на экстремальном пределе практичности" — Консалтинговая компания Forrester

На пути к йотта байтам…

- мега — 10^6 — 1 000 000 (10^6) или 1 048 576 (2^20) байт
- гига — 10^9 — 1 000 000 000 (10^9) или 2^30 = 1 073 741 824 байт
- тера — 10^12 — 1 099 511 627 776 байт или 1024 гигабайт
- пета — 10^15 или 2^50 байт
- экса — 10^18 или 2^60 байт
- зетта — 10^21 или 2^70 байт
- йотта — 10^24 или 2^80 байт

Big Data - проблема или решение???

- основное их предназначение в том, чтобы помочь ***понять*** прошлое ***предсказать*** будущее

Большую часть данных, которая будет произведена в период с 2012 по 2020 годы, сгенерируют не люди, а машины в ходе взаимодействия друг с другом и другими сетями данных.

три «V»


Variety
Velocity
Volume
Structured & Unstructured
Batch
Structured
Big Data
Streaming Data
Zettabytes
Terabytes

Методики анализа больших данных

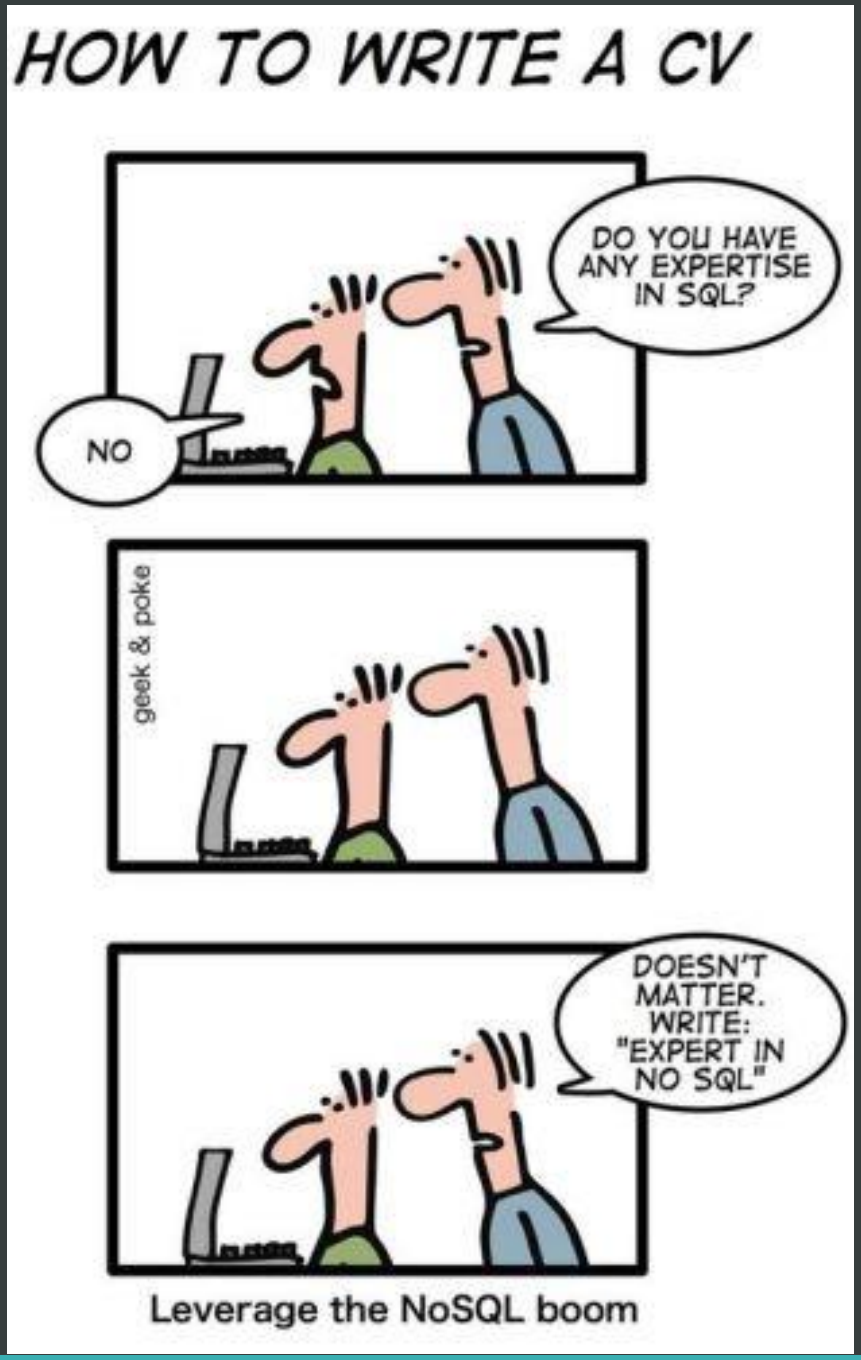
- A/B testing
- Association rule learning
- Classification
- Cluster analysis
- Crowdsourcing.
- Data fusion and data integration
- Data mining
- Ensemble learning
- Genetic algorithms

Аналитический инструментарий

- 1010data
- Apache Chukwa
- Apache Hadoop
- Apache Hive
- Apache Pig!
- Jaspersoft
- MapReduce

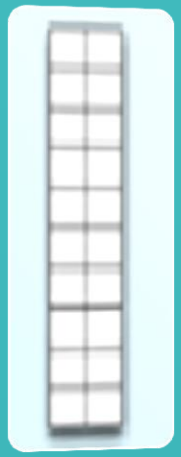
Технологии

NoSQL-решения



NoSQL WORLD

Key-value

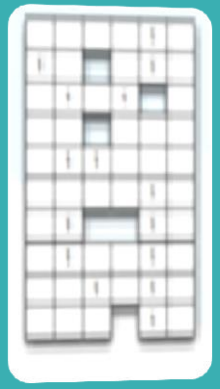


- Dynamo
- Voldemo
- Redis
- Riak

Column (tabular)

- HBase
- Hypertable
- Cassandra

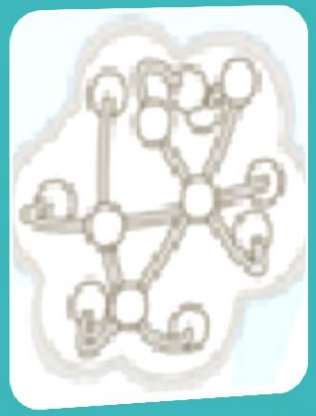
Graph



- Neo4J
- Virtuoso
- AllegroGraph
- CloudGRAPH

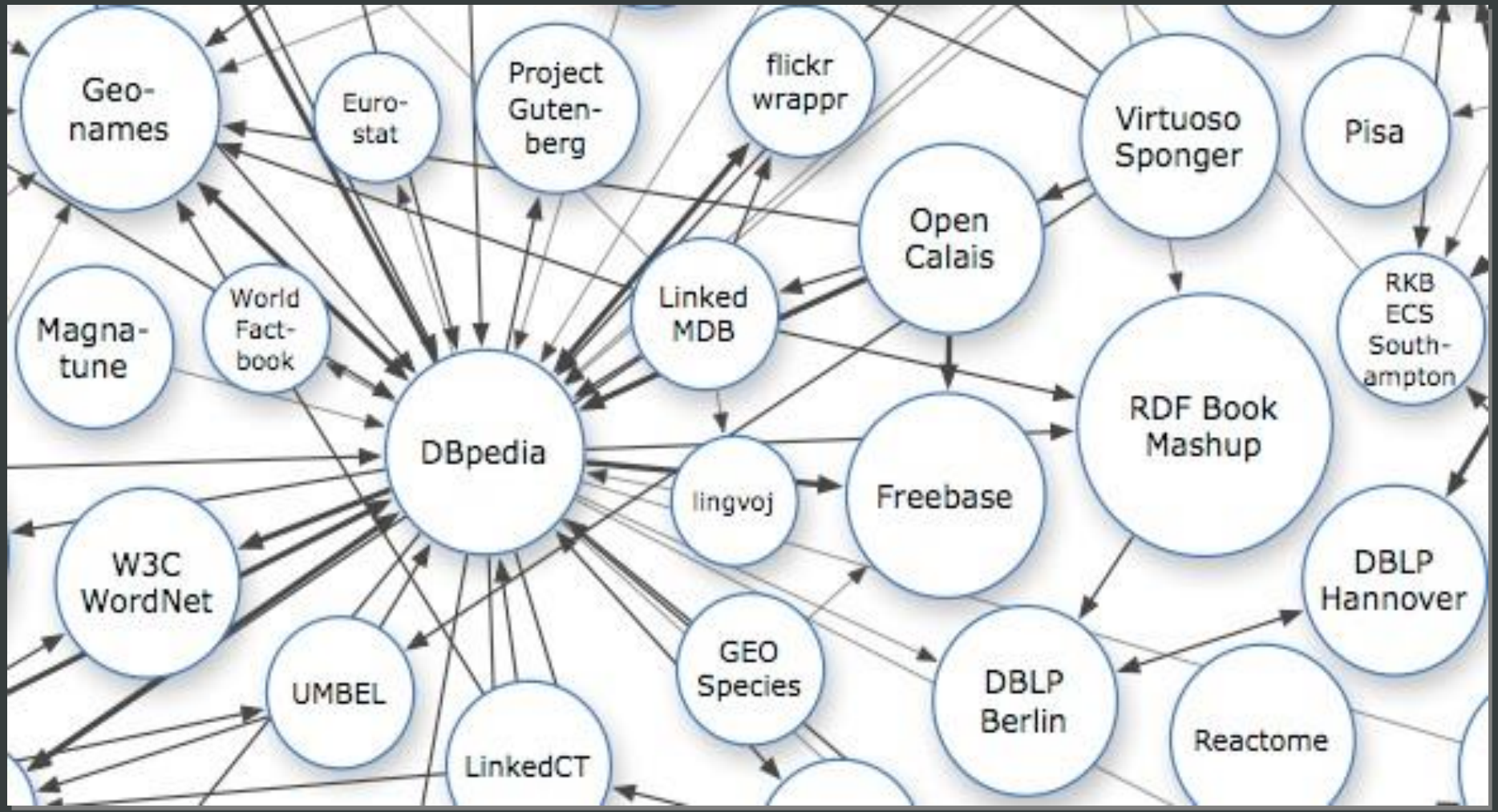
semantic

Document store



- CouchDB
- MongoDB

Множество источников данных, доступных в Internet



Datasets

- DBpedia
- FOAF
- CKAN
- GeoNames
- Freebase

Интерфейсы доступа к ресурсам Data Sets

ODBC

- **JDBC**

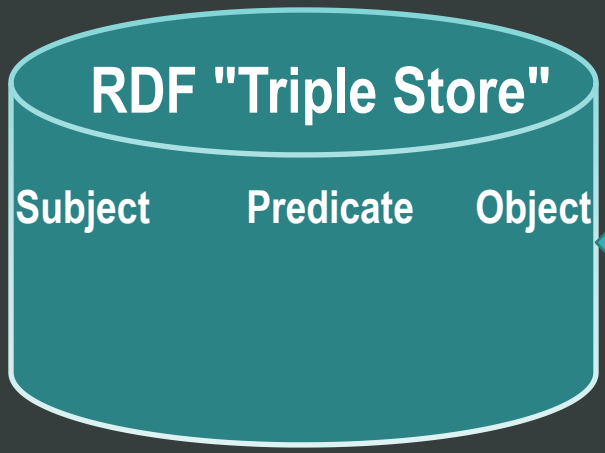
ADO.NET

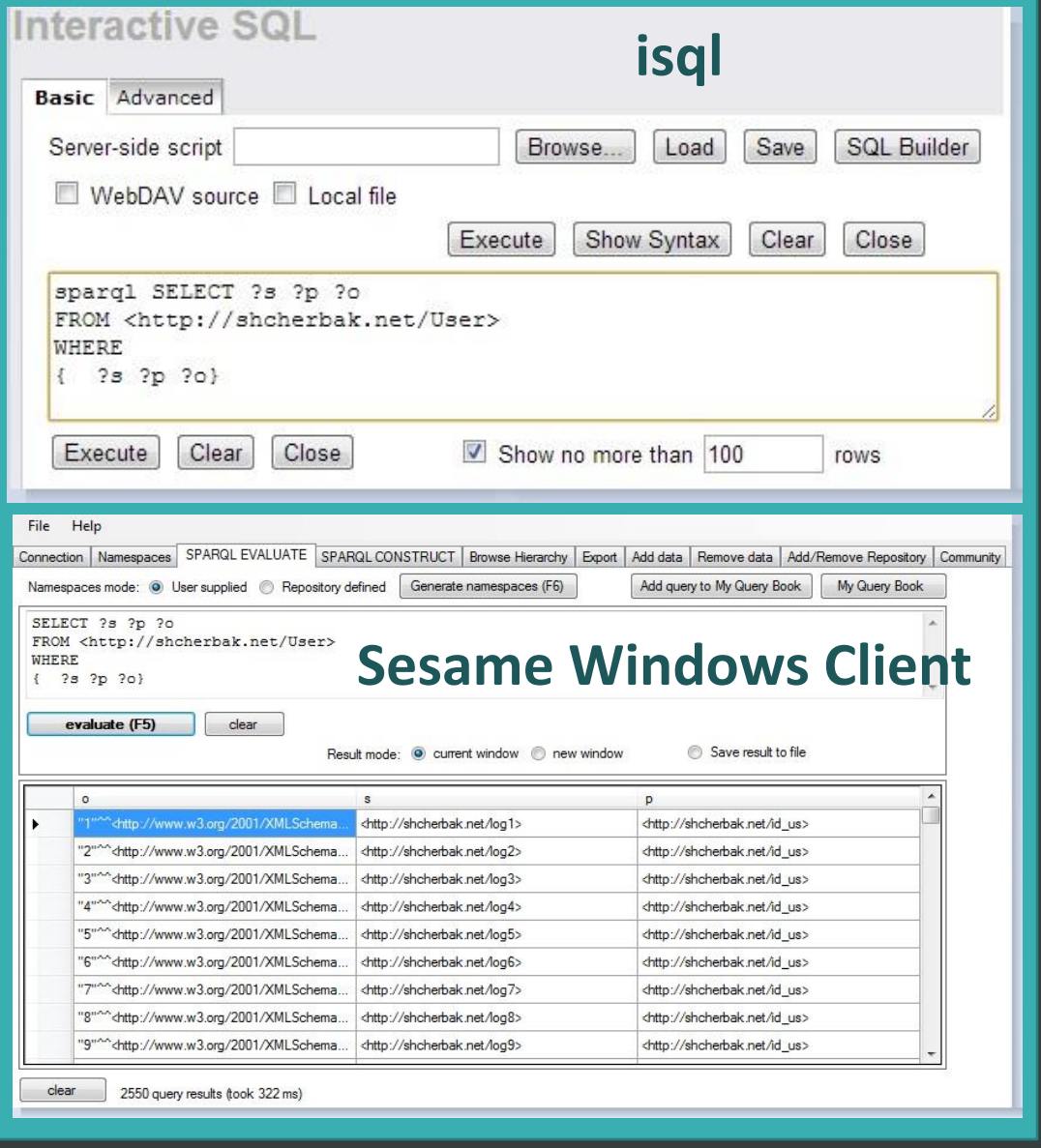
- **OLE DB**

ADO.NET Entity Framework

Обеспечение унифицированного программного API для доступа к ресурсам Graph Triple-, Quad Store

Примеры клиентов для доступа к ресурсам Dataset

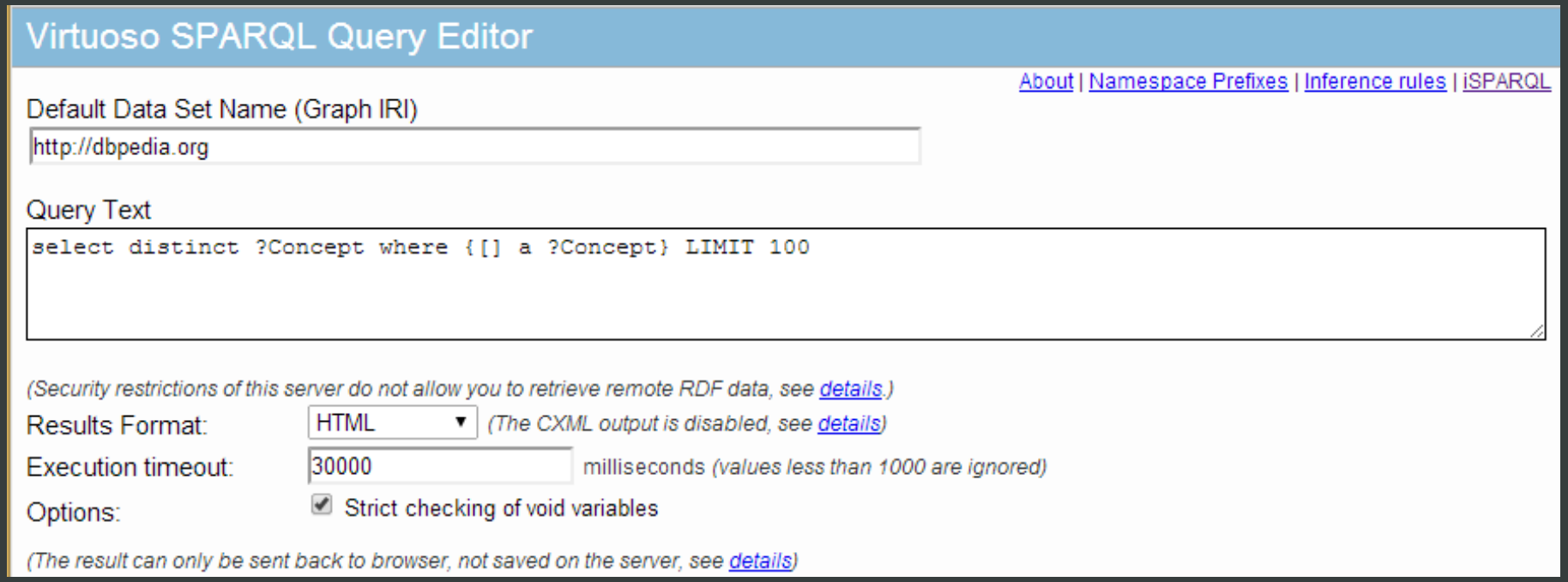




см. http://simile.mit.edu/reports/stores/

SPARQL-точка доступа DBPedia

http://dbpedia.org/sparql/



SPARQL запрос для получения перечня названий первых десяти фильмов с указанием lang-тега


```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX dbpedia-owl: <http://dbpedia.org/ontology/>
SELECT ?film
WHERE
{
  ?a rdf:type   dbpedia-owl:Film;
     rdfs:label ?film.
        FILTER( lang(?film) = "ru" )
}
limit 10
```

SPARQL запрос для подключения к удаленному источнику DBpedia


```
sparql
SELECT ?s WHERE
   {
   SERVICE <http://dbpedia.org/sparql>
      { graph ?g
         {
            ?a rdf:type   dbpedia-owl:Film;
               rdfs:label ?film.
      FILTER( lang(?film) = "ru" )
         }
      }
   }
```

Пример страницы, посвященной фильму «Неудержимые 2»

About: Неудержимые 2

An Entity of Type : movie, from Named Graph : http://dbpedia.org, within Data Space : dbpedia.org

DBpedia

«Неудержимые 2» — сиквел вышедшего в 2010 году боевика «Неудержимые», снятый режиссёром Саймоном Уэстом по сценарию Сильвестра Сталлоне, Ричарда Уэнка, Кена Кауфмана и Дэвида Агосто. Прокатом в США занялась компания Lionsgate, в России — Universal Pictures. Сюжет фильма разворачивается вокруг отряда наёмников «Неудержимые», которые в ходе выполнения очередного задания сталкиваются с убийством одного из своих товарищей.

Property	Value
dbpedia-owl:Work/runtime	103.0
dbpedia-owl:abstract	The Expendables 2 ist ein US-amerikanischer Ac Richard Wenk, Ken Kaufman, David Agosto und Fortsetzung zu The Expendables aus dem Jahr 2 Statham, Jet Li, Dolph Lundgren, Chuck Norris, F Willis und Arnold Schwarzenegger über ein All-St Deutschland der 30. August 2012.
	The Expendables 2 is a 2012 American ensembl Stallone and based on a story by Ken Kaufman, the 2010 action film The Expendables, and stars Crews, Randy Couture, Liam Hemsworth, Jean-C the mercenary group known as "the Expendables revenge against rival mercenary Jean Vilain, who Principal photography took place over 14 weeks (locations included Bulgaria, Hong Kong and New environmental damage caused during filming in B America the following day. The Expendables 2 gr

Create account Log in

Article Talk

Read View source Search

WIKIPEDIA
The Free Encyclopedia

From Wikipedia, the free encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikimedia Shop

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools

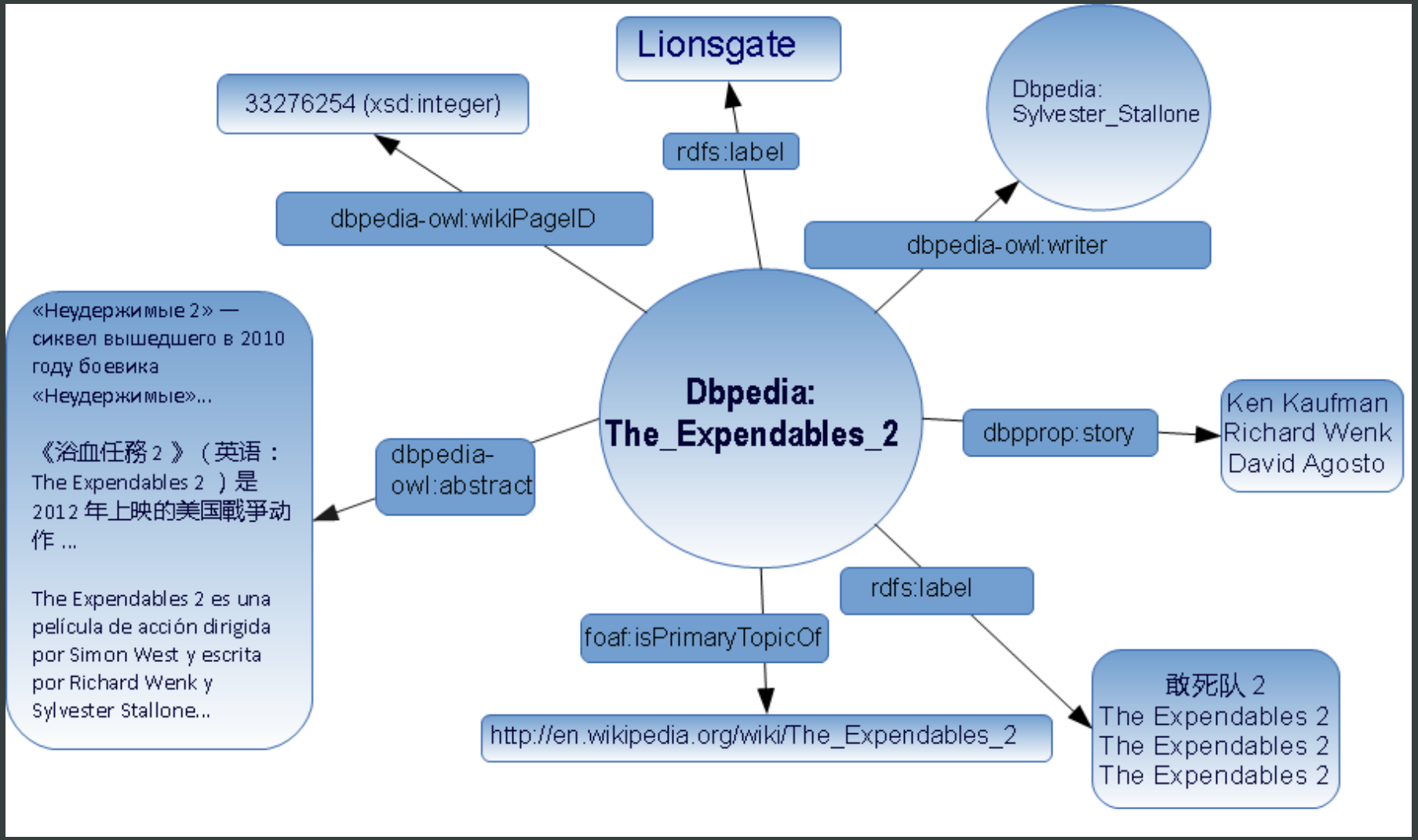
Print/export

Languages

The Expendables 2 is a 2012 American ensemble action film directed by Simon West, written by Richard Wenk and Sylvester Stallone and based on a story by Ken Kaufman, David Agosto and Wenk. Brian Tyler returned to score the film. It is a sequel to the 2010 action film *The Expendables*, and stars Sylvester Stallone, Jason

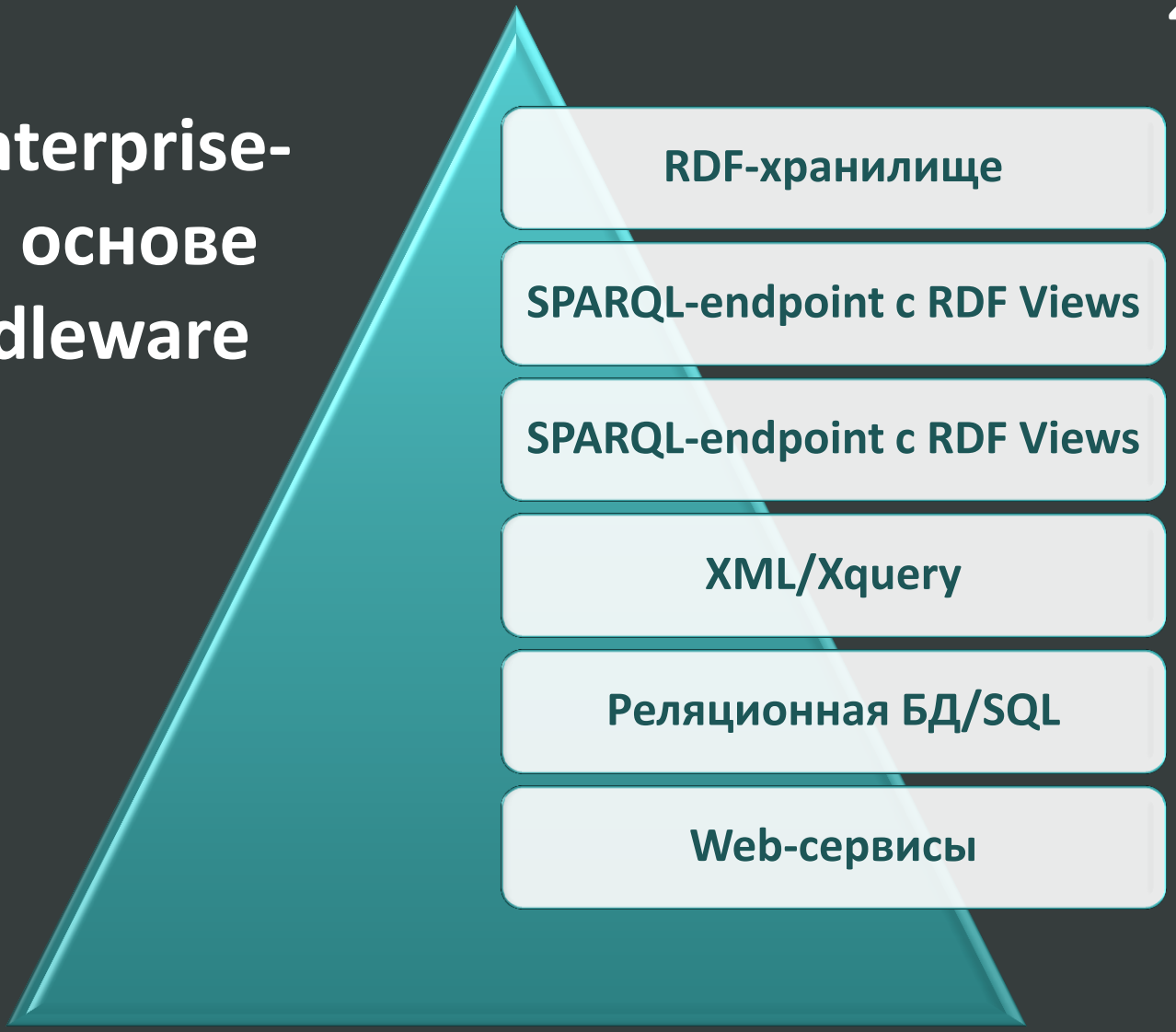
The Expendables 2

Графическое представление данных

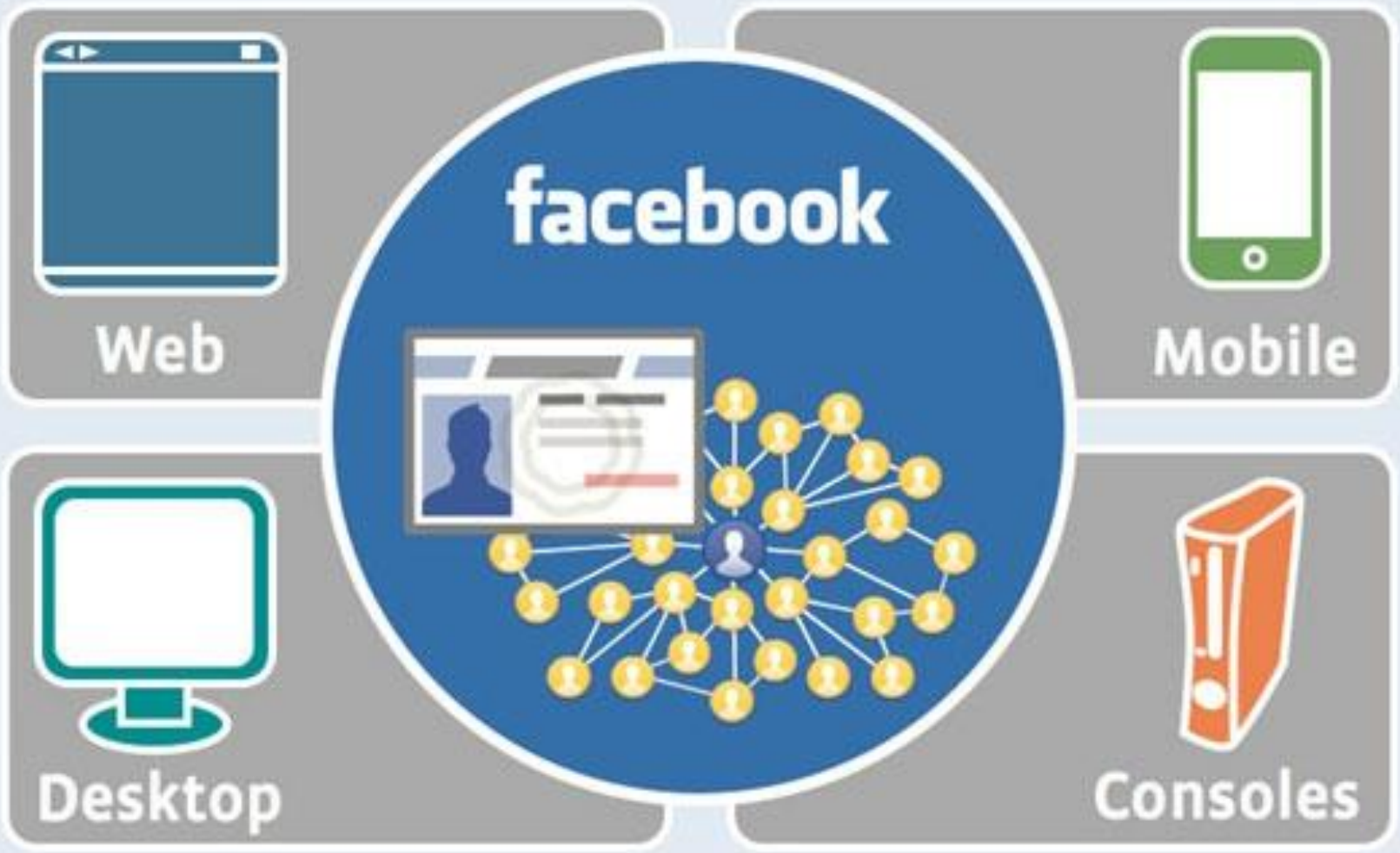


Хранилище триплетов Virtuoso

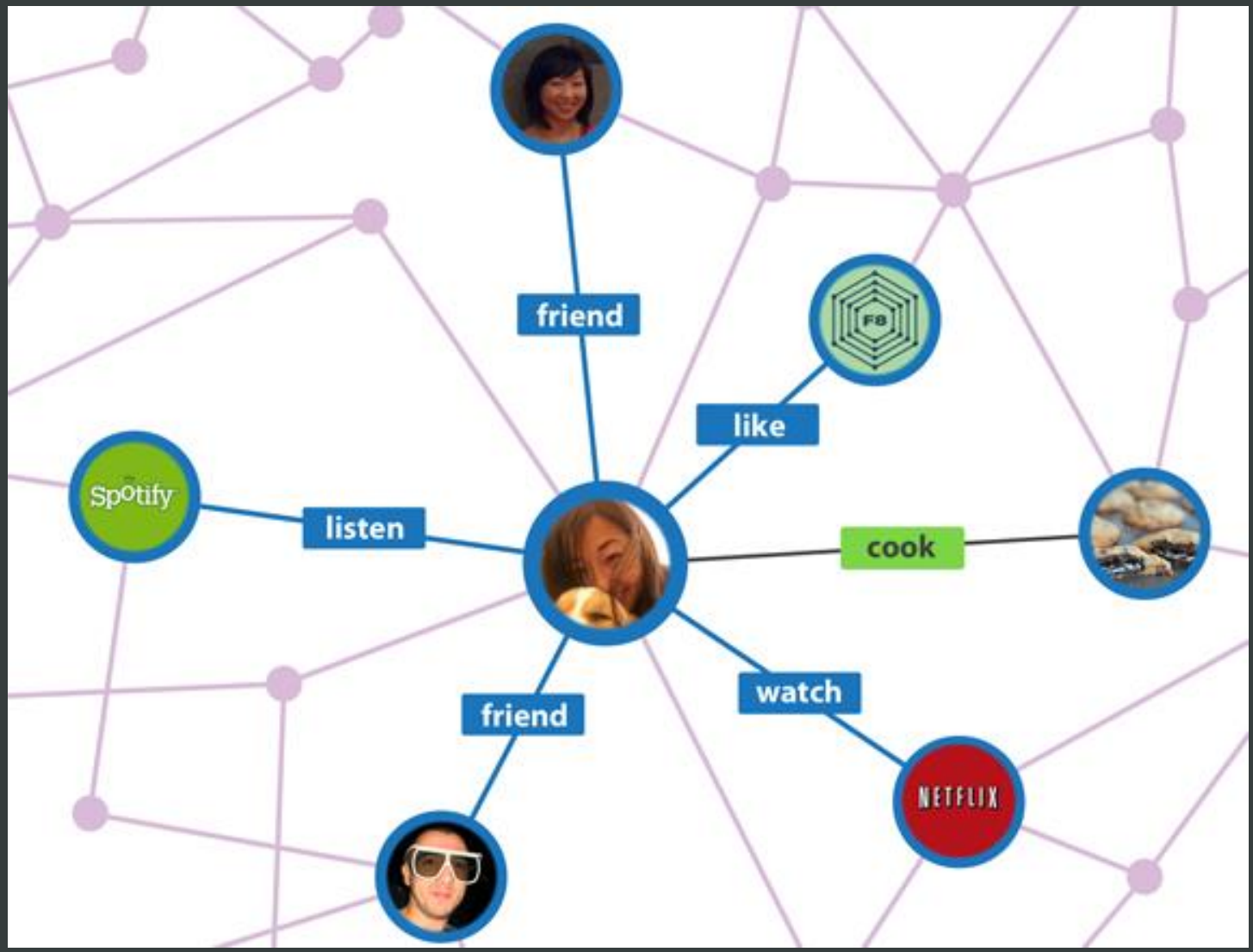
Virtuoso - Enterprise-решение на основе RDBMS Middleware компании OpenLink.



Open Graph



Пример Open Graph



Выводы

…

Wolfram Alpha:

Computational Knowledge Engine

IBM Watson

IBM Big Data Platform

Google Knowledge Graph

И многое другое из мира Big Data формирует будущее пользовательских приложений и сервисов.

Уже сегодня

Вопросы???

Спасибо за внимание

http://shcherbak.net