SPARQLytics: Multidimensional Analytics for RDF

Michael Rudolf

Database Technology Group, Technische Universität Dresden

March 8, 2017

Agenda

Dresden Database Systems Group

Motivation

RDF and SPARQL

Multidimensional Analytics for RDF

Motivation

Focus of Interest

Focus moved from single entity (OLTP)

- Bookkeeping
- Where is what?

To aggregations over sets of entities of the same kind (OLAP)

- Reporting
- What are the sales figures?

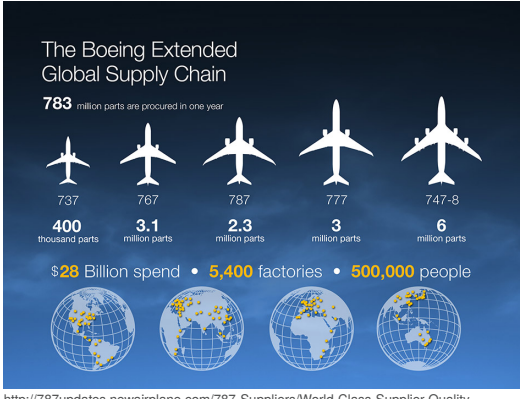
To connections between entities

- Who likes what and why?
- What do the friends of your customers buy?

Business Use Cases

Supply Chain Management

- Transportation & logistics: routing, tendering, tracking, auditing, payment

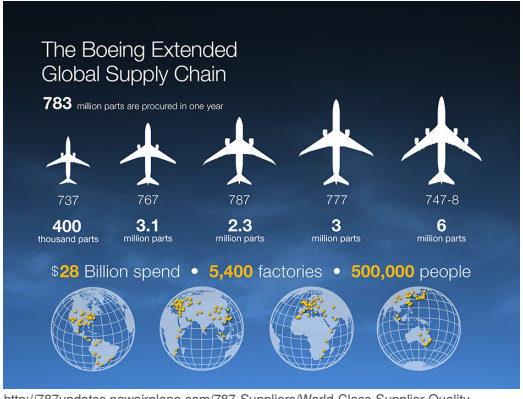


http://787updates.newairplane.com/787-Suppliers/World-Class-Supplier-Quality

Business Use Cases

Supply Chain Management

- Transportation & logistics: routing, tendering, tracking, auditing, payment



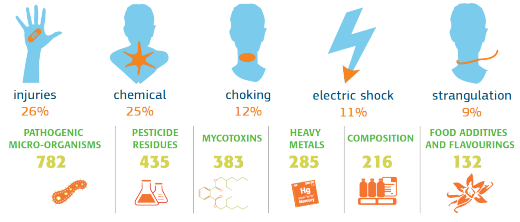
http://787updates.newairplane.com/787-Suppliers/World-Class-Supplier-Quality

Track & Trace

- Pinpoint product recalls
- Mandated by law for certain industries (e.g. pharmaceuticals, food, waste)

EU Commission's Rapid Alert System

	non-food (RAPEX)	food & feed (RASFF)
2013	2364	3137
2014	2435	3157



RDF and SPARQL

Resource Description Framework (RDF) [WLC14]

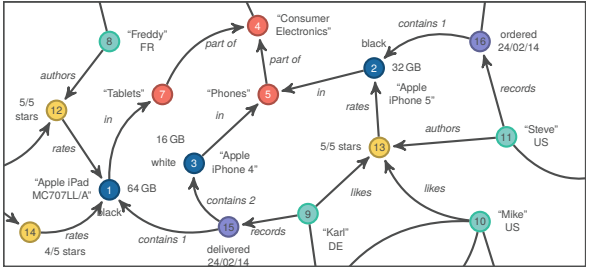
Dresden Database Systems Group

- **Subjects** name an entity
- **Predicates** describe the relationship
- **Objects** can be literals or name

- no built-in schema
- can re-use vocabularies and ontologies
- suitable for inferencing facts

```
@prefix amazon:   <http://www.amazon.com/#> .
@prefix customer: <http://www.amazon.com/customer#> .
@prefix product:  <http://www.amazon.com/product#> .
@prefix category: <http://www.amazon.com/category#> .

product:1  amazon:capacity "64 GB" .
product:1  amazon:color    "black" .
product:1  amazon:in       category:7 .
category:7 amazon:name     "Tablets" .
category:7 amazon:partOf   category:6 .
category:6 amazon:name     "Computers & Accessories" .
user:8     amazon:country  "FR" .
user:8     amazon:rates    product:1 .
```



SPARQL Protocol and RDF Query Language [HS13]

- Built around pattern matching, produces pattern variable bindings
- Grouping and aggregation, CRUD operations
- No multidimensional concepts ➔ complex and error-prone queries

```
PREFIX amazon: <http://www.amazon.com/#>
SELECT (AVG(?capacity) AS ?avgCap) (?name AS ?categoryName)
WHERE {
  ?product  amazon:in       ?category .
  ?category amazon:name     ?name .
  ?category amazon:partOf+  category:6 .
  ?product  amazon:capacity ?capacity
}
GROUP BY ?categoryName
```

Dresden Database
Systems Group

Multidimensional Analytics for RDF

TECHNISCHE UNIVERSITÄT DRESDEN

Multidimensional Data Model [KR13]

(Base) Facts

- Describe events and measurements
- Mostly numeric and continuous

Dimensions

- Provide context for facts
- If numeric, then often discrete
- Can embody structure

Measures

- Are computed from grouped facts
- Are “arranged” in (hyper-)cubes

Multidimensional Data Model [KR13]

(Base) Facts

- Describe events and measurements
- Mostly numeric and continuous

Dimensions

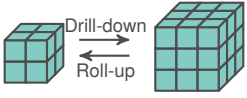
- Provide context for facts
- If numeric, then often discrete
- Can embody structure

Measures

- Are computed from grouped facts
- Are “arranged” in (hyper-)cubes

Slice

Dice



Multidimensional Data Model [KR13]

Dresden Database Systems Group

(Base) Facts

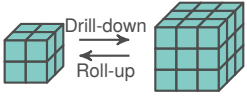
- Describe events and measurements
- Mostly numeric and continuous

Dimensions

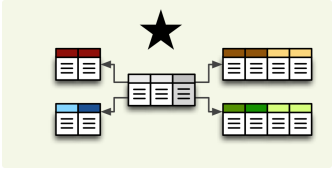
- Provide context for facts
- If numeric, then often discrete
- Can embody structure

Measures

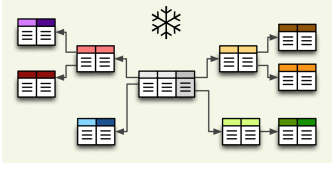
- Are computed from grouped facts
- Are “arranged” in (hyper-)cubes



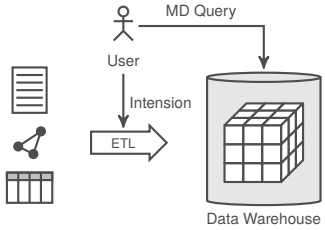
Star schema



Snowflake schema



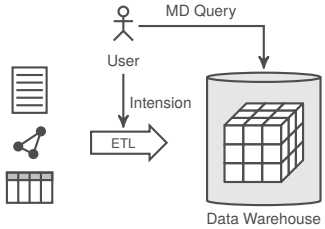
From Intensional to Extensional Analytics

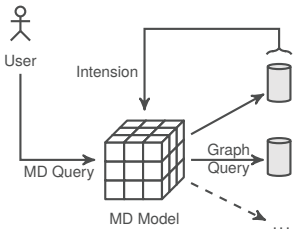


Data Transformation

- Intension fixed by domain expert or metadata
- Import data using ETL process

From Intensional to Extensional Analytics





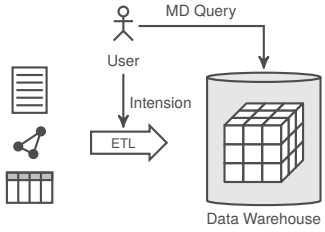
Data Transformation

- Intension fixed by domain expert or metadata
- Import data using ETL process

Query Generation

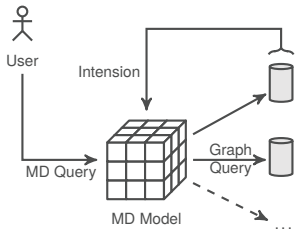
- Intension fixed by metadata
- Generate SPARQL queries from model

From Intensional to Extensional Analytics



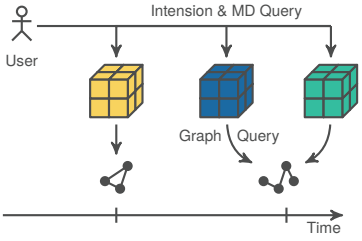
Data Transformation

- Intension fixed by domain expert or metadata
- Import data using ETL process



Query Generation

- Intension fixed by metadata
- Generate SPARQL queries from model

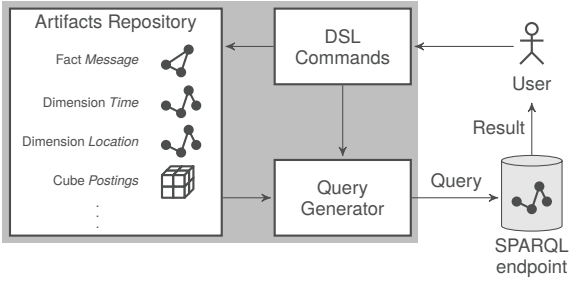


Extensional

- Intension not fixed up-front
- Generate graph queries from user-specified intension

SPARQLytics for the Data Enthusiast

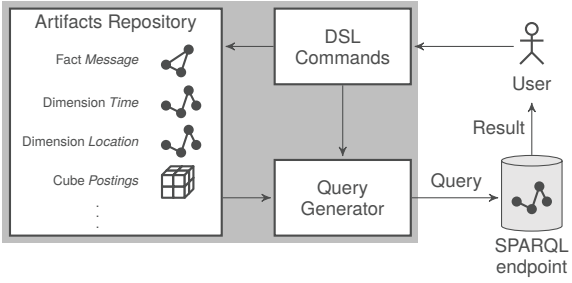
SPARQLytics Workflow



SPARQLytics for the Data Enthusiast

Dresden Database Systems Group

SPARQLytics Workflow



1. Create artifacts in repository

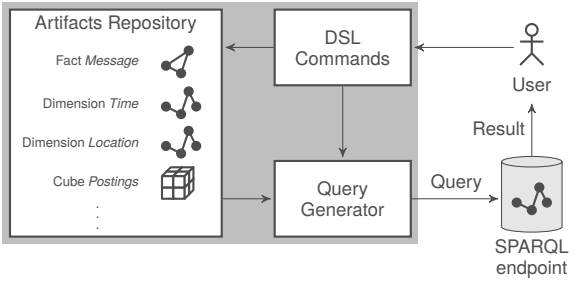
Example

```
USING REPOSITORY "myrepo";
SELECT FACTS {
 ?person rdf:type snvoc:Person ;
         snvoc:birthday ?birthday .
 FILTER (YEAR(NOW()) - YEAR(?birthday) >= 18)
};
DEFINE DIMENSION "Location" FROM (
  ?person  snvoc:isLocatedIn ?city .
  ?city    snvoc:isPartOf ?country .
  ?country snvoc:isPartOf ?continent
) WITH (
  LEVEL "City" AS ?city,
  LEVEL "Country" AS ?country,
  LEVEL "Continent" AS ?continent
);
DEFINE MEASURE "Avg. No. Languages"
 AS COUNT(DISTINCT ?language) WHERE (
  ?person snvoc:speaks ?language
 ) WITH "AVG";
CREATE CUBE "QB" FROM "Location", ...
WITH "Avg. No. Languages", ...;
```

SPARQLytics for the Data Enthusiast

Dresden Database Systems Group

SPARQLytics Workflow



1. Create artifacts in repository
2. Start session re-using artifacts

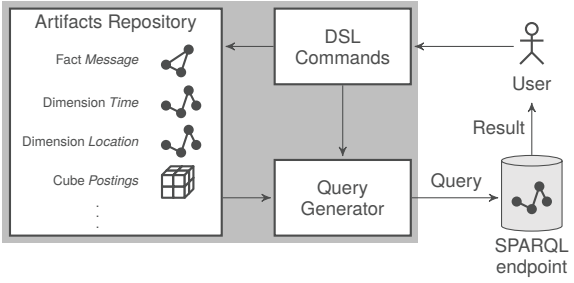
Example

```
USING CUBE "QB" OVER <http://localhost:3030/ds/sparql>;
SLICE("Location", "Country", dbpedia:Italy);
COMPUTE ("Avg. No. Languages");
```

SPARQLytics for the Data Enthusiast

Dresden Database Systems Group

SPARQLytics Workflow



Example

```
USING CUBE "QB" OVER <http://localhost:3030/ds/sparql>;
SLICE("Location", "Country", dbpedia:Italy);
COMPUTE ("Avg. No. Languages");

RESET FILTER("Location", "Country");
ROLLUP("Location", 1);
COMPUTE ("Avg. No. Languages");
...
```

1. Create artifacts in repository
2. Start session re-using artifacts
3. Iteratively explore data,
   optionally create additional artifacts

Summary

Big Graph Data

- Not just social networks, also business scenarios
- Not enough data scientists, enable data enthusiasts

RDF and SPARQL

- Linked Open Data a rich source of information
- SPARQL does not expose multidimensional concepts

SPARQLytics

- Re-use core SPARQL elements for defining multidimensional model
- Generate complex SPARQL queries from analytical session
- Stateful approach integrates well with data enthusiasts workflow

Additional Material & References

References I

Charu C. Aggarwal and Haixun Wang.
A Survey of Clustering Algorithms for Graph Data.
In Charu C. Aggarwal and Haixun Wang, editors, *Managing and Mining Graph Data*, volume 40 of *Advances in Database Systems*, chapter 9, pages 275–301. Springer US, 2010.

Seyed-Mehdi-Reza Beheshti, Boualem Benatallah, Hamid Reza Motahari-Nezhad, and Mohammad Allahbakhsh.
A framework and a language for on-line analytical processing on graphs.
In *Proceedings of the 13th International Conference on Web Information Systems Engineering (WISE)*, volume 7651 of *Lecture Notes in Computer Science*, pages 213–227. Springer, 2012.

Peter Boncz.
LDBC: Benchmarks for Graph and RDF Data Management.
In *Proc. IDEAS*, pages 1–2. ACM, 2013.

Fabio Crestani.
Application of spreading activation techniques in information retrieval.
Artificial Intelligence Review, 11(6):453–482, December 1997.

Chen Chen, Xifeng Yan, Feida Zhu, Jiawei Han, and Philip S. Yu.
Graph OLAP: Towards Online Analytical Processing on Graphs.
In *Proceedings of the 8th International Conference on Data Mining*, pages 103–112. IEEE, December 2008.

Hartmut Ehrig, Gregor Engels, Hans-Jörg Kreowski, and Grzegorz Rozenberg, editors.
Handbook of Graph Grammars and Computing by Graph Transformation: Applications, Languages and Tools, volume 2.
World Scientific, 1997.

References II


Steven Harris and Andy Seaborne.
SPARQL 1.1 query language.
W3C recommendation, W3C, March 2013.


Dirk Koschützki, Katharina Anna Lehmann, Leon Peeters, Stefan Richter, Dagmar Tenfelde-Podehl, and Oliver Zlotowski.
Centrality Indices, volume 3418 of *Lecture Notes in Computer Science*, chapter 3, pages 16–61.
Springer, 2005.


Sven Kosub.
Local Density, volume 3418 of *Lecture Notes in Computer Science*, chapter 6, pages 112–142.
Springer, 2005.


Ralph Kimball and Margy Ross.
The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling.
Wiley, 3rd edition, 2013.


Kristen LeFevre and Evimaria Terzi.
Grass: Graph structure summarization.
In *Proc. SDM*, pages 454–465. SIAM, 2010.


Saket Navlakha, Rajeev Rastogi, and Nisheeth Shrivastava.
Graph summarization with bounded error.
In *Proc. SIGMOD*, pages 419–432. ACM, 2008.

References III


Satu Elisa Schaeffer.
Graph clustering.
Computer Science Review, 1(1):27–64, August 2007.


Yuanyuan Tian and Jignesh M. Patel.
TALE: A Tool for Approximate Large Graph Matching.
In *2008 IEEE 24th International Conference on Data Engineering*, pages 963–972. IEEE, April 2008.


David Wood, Markus Lanthaler, and Richard Cyganiak.
RDF 1.1 concepts and abstract syntax.
W3C recommendation, W3C, February 2014.
http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/.


Peixiang Zhao, Xiaolei Li, Dong Xin, and Jiawei Han.
Graph Cube: On Warehousing and OLAP Multidimensional Networks.
In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 853–864. ACM, 2011.


Ning Zhang, Yuanyuan Tian, and Jignesh M. Patel.
Discovery-Driven Graph Summarization.
In *Proceedings of the 26th International Conference on Data Engineering*, pages 880–891. IEEE, 2010.