The NPD Benchmark: Reality Check for OBDA Systems

EDBT 2015
Bruxelles, Belgium
25/03/15

Speaker:

- **Davide Lanti**
  - **FUB — Free University of Bozen-Bolzano**
  - `davide.lanti@unibz.it`

Joint Work With:

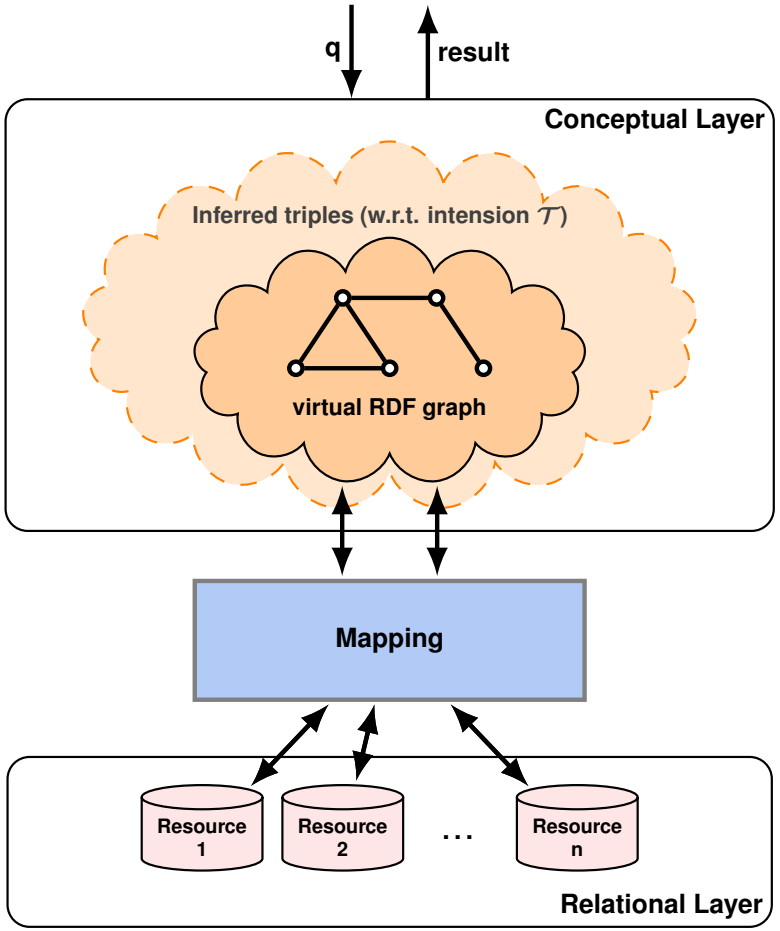
- **Martin Rezk, Guohui Xiao, and Diego Calvanese**
  - **FUB — Free University of Bozen-Bolzano**
  - `{mrezk,xiao,calvanese}@inf.unibz.it`

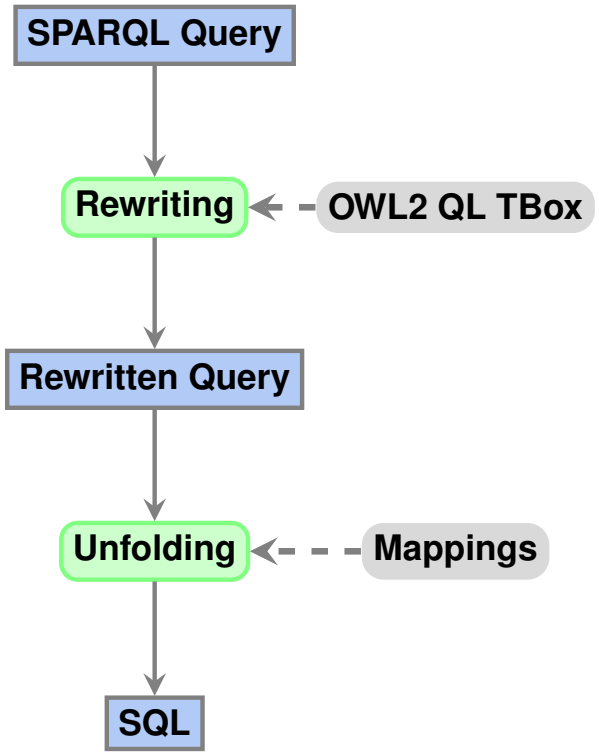
Outline

- **OBDA**
- **Benchmarking OBDA Systems**
- **The NPD Benchmark**
- **Empirical Evaluation of OBDA Systems**
- **Conclusions**

Outline

What is OBDA


q
result
Conceptual Layer
Inferred triples (w.r.t. intension $\mathcal{T}$)
virtual RDF graph
Mapping
Resource 1
Resource 2
...
Resource n
Relational Layer


SPARQL Query
Rewriting
OWL2 QL TBox
Rewritten Query
Unfolding
Mappings
SQL

What is OBDA Useful For

- It provides a high-level conceptual view of the data
- It makes it easy to query information otherwise difficult to retrieve
- It allows for easy integration of legacy data sources

Example

At SIEMENS Energy, the interaction between the database users and the IT experts can take weeks before the right query is formulated

Industrial Use-Cases of OBDA

- **Italian Ministry for Economy and Commerce**
- **SELEX SI (Finmeccanica)**
- **Monte dei Paschi di Siena (Bank)**
- **Telecom Italia**
- **Statoil**
- **Siemens Energy Services**

Remark

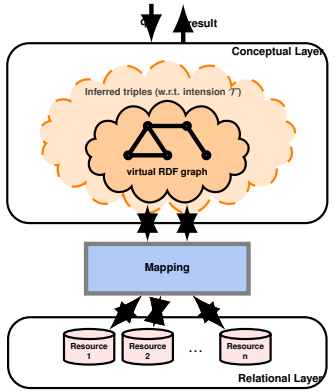
OBDA is aimed at organizations and enterprises dealing with huge amounts of structured data

Outline

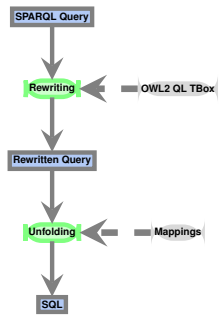
Need for OBDA Benchmarks

- **Performance Issues in OBDA**
  - **Query translation is worst-case exponential**
  - **Semantic and structural optimizations allow to exploit DB indexes**
  - $\rightsquigarrow$ **Some queries cannot be executed efficiently, unless optimized**
- **Main Challenge for OBDA**
  - **Develop optimization techniques (ongoing work in Bolzano)**
  - **Assess the effectiveness of optimizations**
  - $\rightsquigarrow$ **Need for benchmarks**

Benchmarking OBDA Systems: Metrics



- **Metrics to assess the complexity of the input**
  - **Ontology:** Size, language expressiveness, complexity
  - **Mappings:** Size, redundancy, complexity of queries over sources
  - **Database:** Size, schema structure
  - **User Queries:** Size, complexity
- **Metrics for a fine-grained analysis on each execution phase**
  - **Start-up:** Time
  - **Rewriting:** Size, complexity, time to compute
  - **Unfolding:** Size, complexity, time to compute
  - **SQL execution:** Time



Remark

Scalability analysis driven by these measures is valuable but difficult to achieve in real-world settings

Shortcomings of Existing Benchmarks

name	Ontology	Queries	Data Instance	Mappings	Standards
adolena	✓	✓	✓	✓	✗
lubm	✓	✓	✓	✗	✓
dbpedia	✗	✓	✓	✗	✓
bsbm	✗	✓	✓	✗	✓
fishmark	✓	✓	✓	✓	✗
wisconsin	✗	✗	✓	✗	✓
TPC	✗	✗	✓	✗	✓

Legenda:

- ✓ : adequate
- ✓ : partially adequate
- ✗ : inadequate

Outline

The NPD Benchmark: Starting Point

Based on Norwegian Petroleum Directorate FactPages[1] dataset, mappings, queries, and ontology

- **Pro:**
  - **Rich, real-world ontology: developed by Univ. Oslo**
  - **Complex mappings: developed by Univ. Oslo**
  - **Real-world queries: developed in collaboration with Univ. Oslo and users**
- **Cons:**
  - **Inconsistencies in the ontology**
  - **Database constraints violations**
  - **Incomplete mappings, problems with datatypes, and non-standard language**
  - **Unsupported constructs in queries**
  - **Small data instance**

[1] `http://sws.ifi.uio.no/project/npd-v2/`

The NPD Benchmark: Our Contributions

- **Fixed the ontology, mappings and database**
- **Ported the mappings into the W3C standard (R2RML)**
- **Adapted the query set to OBDA**
- **Developed a synthetic data generator**
- **Developed an automatized testing platform**

Resource

`https://github.com/ontop/npd-benchmark`

The Ontology and The Mappings

- The ontology

#classes	#obj_prop	#tbox axioms	max_depth	avg_siblings
343	380	1451	10	4.83

The Ontology and The Mappings

- The ontology

#classes	#obj_prop	#tbox axioms	max_depth	avg_siblings
343	380	1451	10	4.83

- The mappings

#mappings	#mapped terms	avg rules	avg joins	max term maps
1190	464	2.6	1.7	116

Queries I

query	#join	#BGPs	#opts	Agg	Filt.	Mod.
Q1	4	5	0	N	Y	N
Q2	5	6	0	N	Y	N
Q3	3	4	0	N	Y	Y
Q4	5	6	0	N	Y	Y
Q5	5	6	0	N	Y	Y
Q6	6	7	0	N	Y	Y
Q7	7	8	0	N	Y	N
Q8	3	4	0	N	Y	N
Q9	3	4	0	N	Y	Y
Q10	2	3	0	N	Y	Y
Q11	7	8	0	N	Y	Y
Q12	8	10	0	N	Y	Y
Q13	2	3	2	N	Y	N
Q14	2	5	2	N	Y	N
Q15	4	5	0	Y	Y	N

Queries II

query	#join	#BGPs	#opts	Agg	Filt.	Mod.
Q16	3	3	0	Y	Y	N
Q17	8	8	0	Y	N	Y
Q18	4	5	0	Y	N	N
Q19	8	8	0	Y	N	N
Q20	3	3	0	Y	N	N
Q21	3	3	0	Y	N	N
Q22	1	2	0	N	N	Y
Q23	2	3	0	N	N	Y
Q24	2	3	0	N	N	Y
Q25	1	1	0	N	N	Y
Q26	2	1	0	N	N	Y
Q27	1	2	0	N	N	Y
Q28	2	3	0	N	Y	Y
Q29	4	5	0	N	Y	Y
Q30	6	7	0	N	Y	Y

An Example of Query (Fragment of q28)

```
SELECT DISTINCT ?wellbore ?well
WHERE  ?wellbore :wellboreForDiscovery ?discovery;
                 :belongsToWell ?well.
```

An Example of Query (Fragment of q28)

```
SELECT DISTINCT ?wellbore ?well
WHERE  ?wellbore :wellboreForDiscovery ?discovery;
                 :belongsToWell ?well.
```

Ontology

```
DevelopmentWellbore SubClassOf wellboreForDiscovery some Discovery
ExplorationWellbore   SubClassOf wellboreForDiscovery some Discovery
```

An Example of Query (Fragment of q28)

```
SELECT DISTINCT ?wellbore ?well
WHERE  ?wellbore :wellboreForDiscovery ?discovery;
                 :belongsToWell ?well.
```

Ontology

```
DevelopmentWellbore SubClassOf wellboreForDiscovery some Discovery
ExplorationWellbore   SubClassOf wellboreForDiscovery some Discovery
```

$\rightsquigarrow$ All individuals in DevelopmentWellbore and ExplorationWellbore must be retrieved and joined with the wellbores belonging to some well

The Synthetic Data Generator

- **Goal: Efficiently generate large datasets with real-world characteristics**
- **Idea: Collect relevant statistics and constraints from an initial dataset and the ontology**
  - **Physical correlation**
  - **Column-based duplicates ratio**
  - **Multiplicities related to ontology relations**
  - **Constraints in the ontology (disjointness)**
- **Implementation:**
  - **Physical correlation for foreign keys**
  - **Column-based duplicates ratio**
  - **Multiplicities related to ontology relations addressed only partially**
  - **Disjointness satisfied by mapping design**

Outline

Full-fledged OBDA Systems

- **Ontop**[2]
  - **Core component of the Optique** [3] **platform**
  - **Code released under Apache Licence version 2.0**
- **Mastro**[4]
  - **Successfully used in several projects carried out in collaboration with public administrations and large companies**
  - **Closed system**
- **Ultrawrap**[5]
  - **Adopted in several projects in the areas of healthcare, law, and transportation**
  - **Proprietary system**

[2] http://ontop.inf.unibz.it/
[3] http://www.optique-project.eu/
[4] http://www.dis.uniroma1.it/~mastro/download/api.php
[5] http://capsenta.com/

OBDA-like Systems

Systems that do not support reasoning

- **Virtuoso-views**[6]
  - **Popular multi-model (relationa, graph, etc.) data server**
  - **Proprietary and open source**
- **Morph**[7]
  - **RDB2RDF engine**
  - **Open source**

Remark

We were not granted the rights to test Ultrawrap. None of these systems, except for Ontop, could load the R2RML mappings for NPD.

[6] http://virtuoso.openlinksw.com/

[7] http://mayor2.dia.fi.upm.es/oeg-upm/index.php/es/technologies/315-morph-rdb

Test Configuration I

- We used the NPD benchmark to test the Ontop OBDA system
- For queries without aggregates, we used Ontop v1.15 (next release)
- Queries with aggregates are not supported by Ontop v1.15, but they are supported experimentally by a pre-release version of the next main-release of Ontop (v2.0)
- System configuration
  - RDBMS: MySQL and PostgreSQL
  - CPU: HP Proliant server with 24 Intel Xeon X5690 CPUs (144 cores @3.47GHz)
  - Memory: 106GB of RAM and a 1TB 15K RPM HD
  - OS: Ubuntu 12.04 LTS
- Queries were executed sequentially

Test Configuration II

- We divided the input query set into two sets
  - Classically Optimizable Queries (9 Queries)
    - Queries that can be translated into a small (< 3) union of efficient SPJ queries without taking into account the data
    - 5 warm-up runs, plus 10 runs with different constants (Query Mix)
    - Calculation of the median running time and Query-mixes Per Hour (QmPh)
  - Non-classically Optimizable Queries (21 Queries)
    - Queries that cannot be translated into small unions of efficient SPJ queries without taking into account the data
    - One warm-up run and two test runs
    - Calculation of unfolding, rewriting, and response times
    - Calculation of unfolding and rewriting sizes (num. of queries)

Test Configuration III

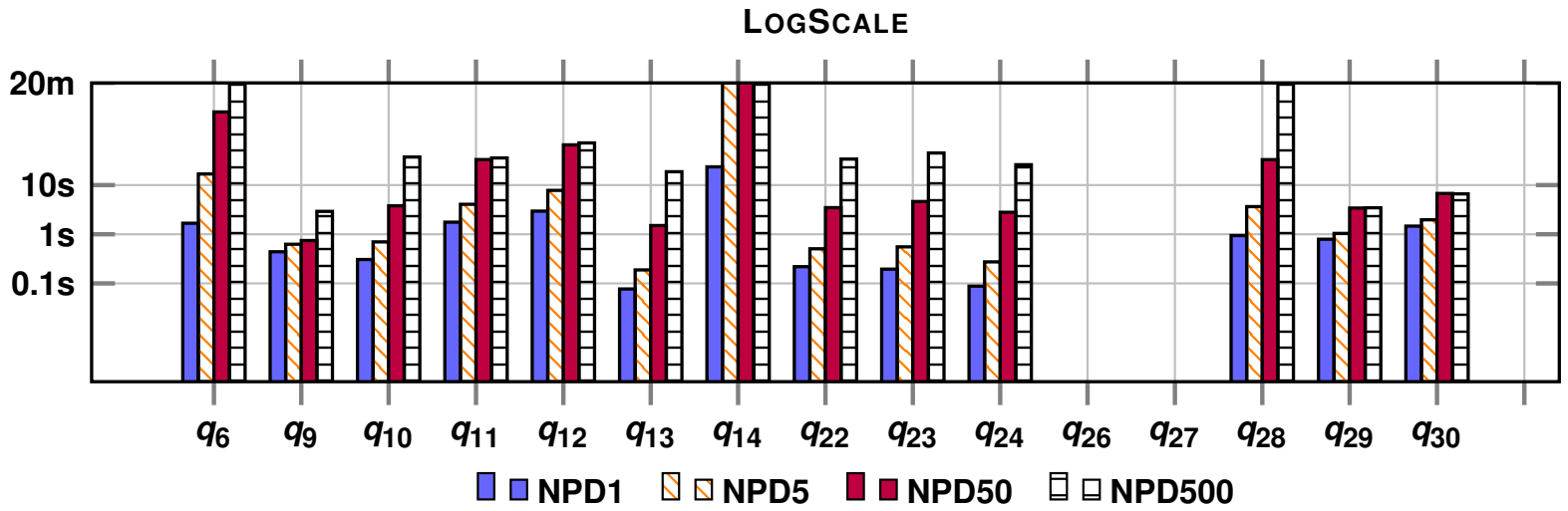
- **Run Parameters:**
  - **Reasoning Level:**
    - **Partial reasoning w.r.t. existentials axioms (≈RDFS)**
    - **Full reasoning w.r.t. existentials axioms (OWL 2 QL)**
  - **Underlying RDBMS:**
    - **MySQL**
    - **PostgreSQL**

Here we present results for Non-classically Optimizable queries and over PostgreSQL

Non-classically Optimizable Queries Rewriting and Unfolding

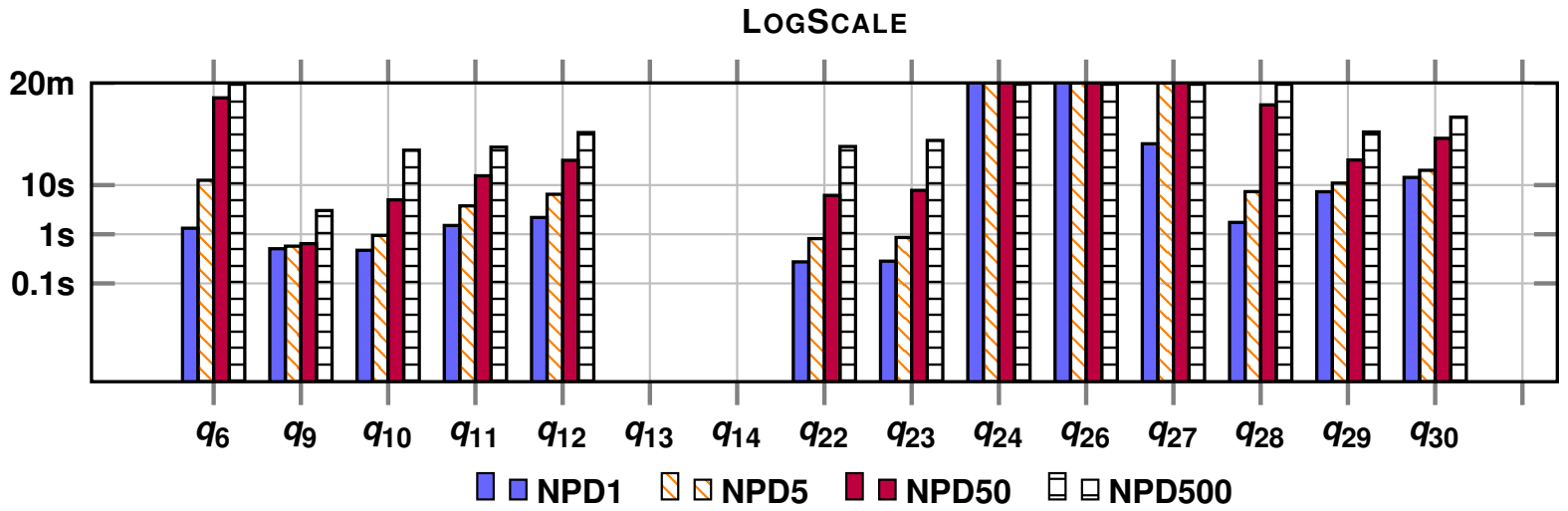
query	Ext. Reasoning OFF			Ext. Reasoning ON		
	#rw	#un	#chars(un)	#rw	#un	#chars(un)
q6	1	48	63556	1	48	63940
q9	1	150	96028	1	160	102178
q10	1	24	17464	1	48	38620
q11	1	24	37828	1	24	37636
q12	2	48	75436	2	48	75820
q13	1	4	1445	—	—	—
q14	1	2	2472	—	—	—
q22	1	2	2532	3	6	7428
q23	1	1	982	3	4	7849
q24	1	1	982	7	12	10842
q26	1	0	0	3	3	7602
q28	1	12	31677	3	24	61811
q29	1	6	18908	21	408	1253007
q30	2	12	37794	42	816	2505992

VIG: Ontop/PostgreSql (Hard Queries - No Ex. Reasoning)



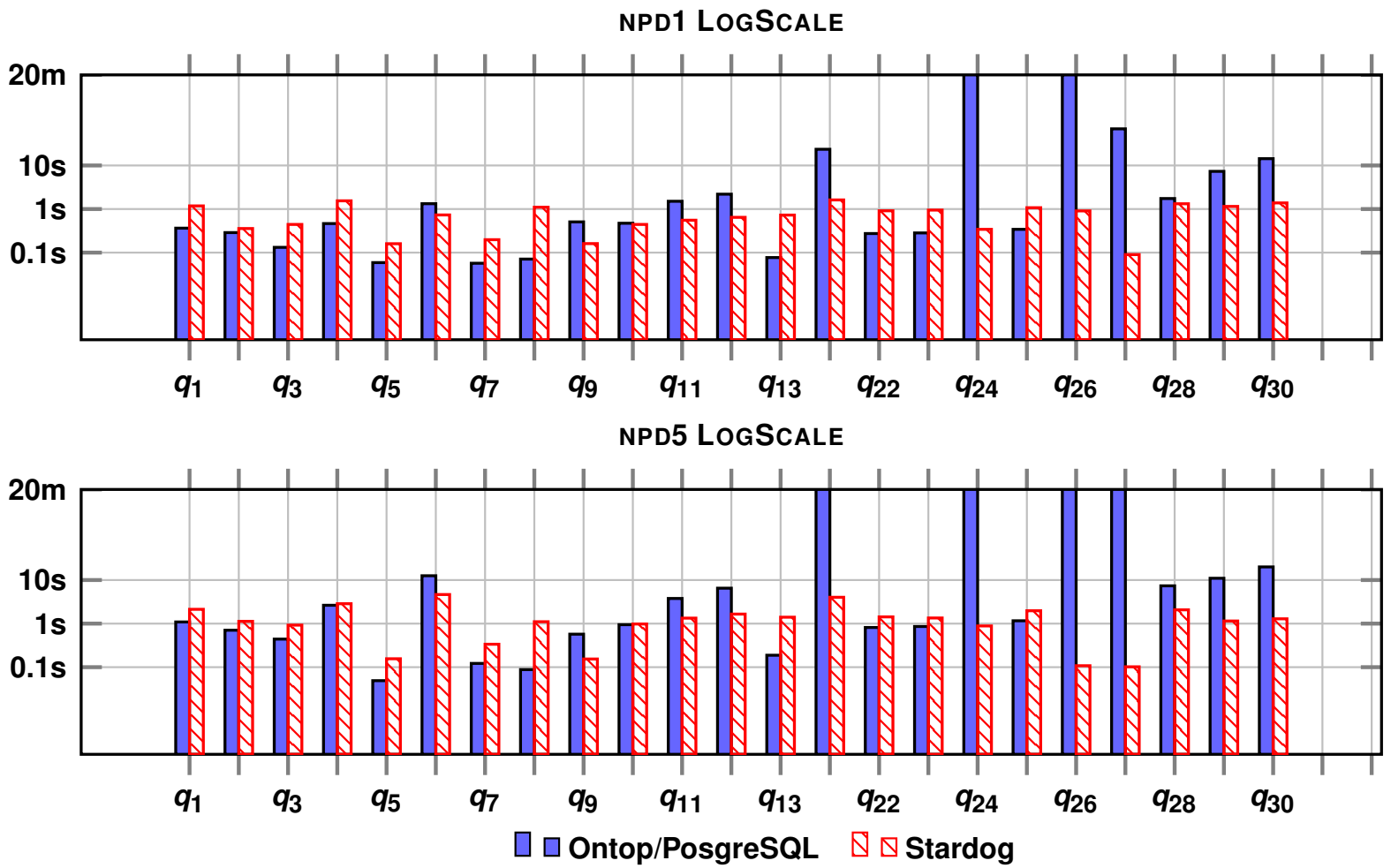
NPD500 ≈ 1.4 billion triples

VIG: Ontop/PostgreSql (Hard Queries - Ex. Reasoning)



NPD500 ≈ 1.4 billion triples

What about Triple Stores...



Outline

Conclusions

- **Existing benchmarks are not suitable for OBDA**
- **We developed a benchmark for OBDA systems**
  - **It is based on a real world scenario**
  - **Complex ontology, mappings and queries**

  $\rightsquigarrow$ **Challenging**
- **We used the benchmark to test the Ontop system, and we observed that**
  - **Optimizations currently applied by Ontop can make OBDA feasible, but more can be achieved and should be achieved**

Future Direction

- **Better data generation**
  - **“Learn” instead of “collect”**
  - **How to conciliate statistics at the virtual level with statistics at the data level**
- **Scalability analysis at the level of the mappings and the ontology**
- **Improve the metrics used in the benchmark**

Thank you!

The NPD Benchmark: Reality Check for OBDA Systems

EDBT 2015
Bruxelles, Belgium
25/03/15

Speaker:

- **Davide Lanti**
  - **FUB — Free University of Bozen-Bolzano**
  - `davide.lanti@unibz.it`

Joint Work With:

- **Martin Rezk, Guohui Xiao, and Diego Calvanese**
  - **FUB — Free University of Bozen-Bolzano**
  - `{mrezk,xiao,calvanese}@inf.unibz.it`