# On Distributed Querying of Linked Data

**Martin Svoboda, Jakub Stárka, Irena Mlýnková**

**XML and Web Engineering Research Group**

**Charles University in Prague**

The Czech Republic

20 April 2012

DATESO

Žernov, Rovensko pod Troskami

# Outline

- Introduction
- Problem
- Challenges
- Framework
- Issues
- Conclusion

# Introduction

- Motivation
  - **Web of Documents**
  - **Web of Data**
- Linked Data
  - Principles
    - Identifiers (URIs)
    - Descriptions (HTTP, RDF)
    - Links

# Introduction

- **RDF** (Resource Description Framework)
  - Triples
    - **Subject Predicate Object**
  - Graph
    - Directed labeled multigraph
    - Vertices for subjects and objects
    - Edges represent particular triples

# Problem

- Querying framework
  - Context
    - **Distributed datasets**
    - **Transparent querying**
  - Issues
    - Physical storage
    - Index structures
    - Query processor

# Problem

- Architecture
  - **Local**
    - Efficient processing
    - Independent data
    - Storage requirements
  - **Distributed**
    - Runtime requests
    - Up-to-date data
    - Network throughput

# Challenges

- **Data distribution**
  - Datasets are distributed
  - Architecture compromise
- **Data dynamicity**
  - Data become obsolete
  - Dynamic structures

# Challenges

- **Data scalability**
  - Motivation
    - Web of Data size explosion
      - September 2011:
      - 295 datasets, 31 billion triples, 504 million links
  - Problems
    - Scalable storages and indices
    - Efficient query evaluation
    - Quality, provenance and trust

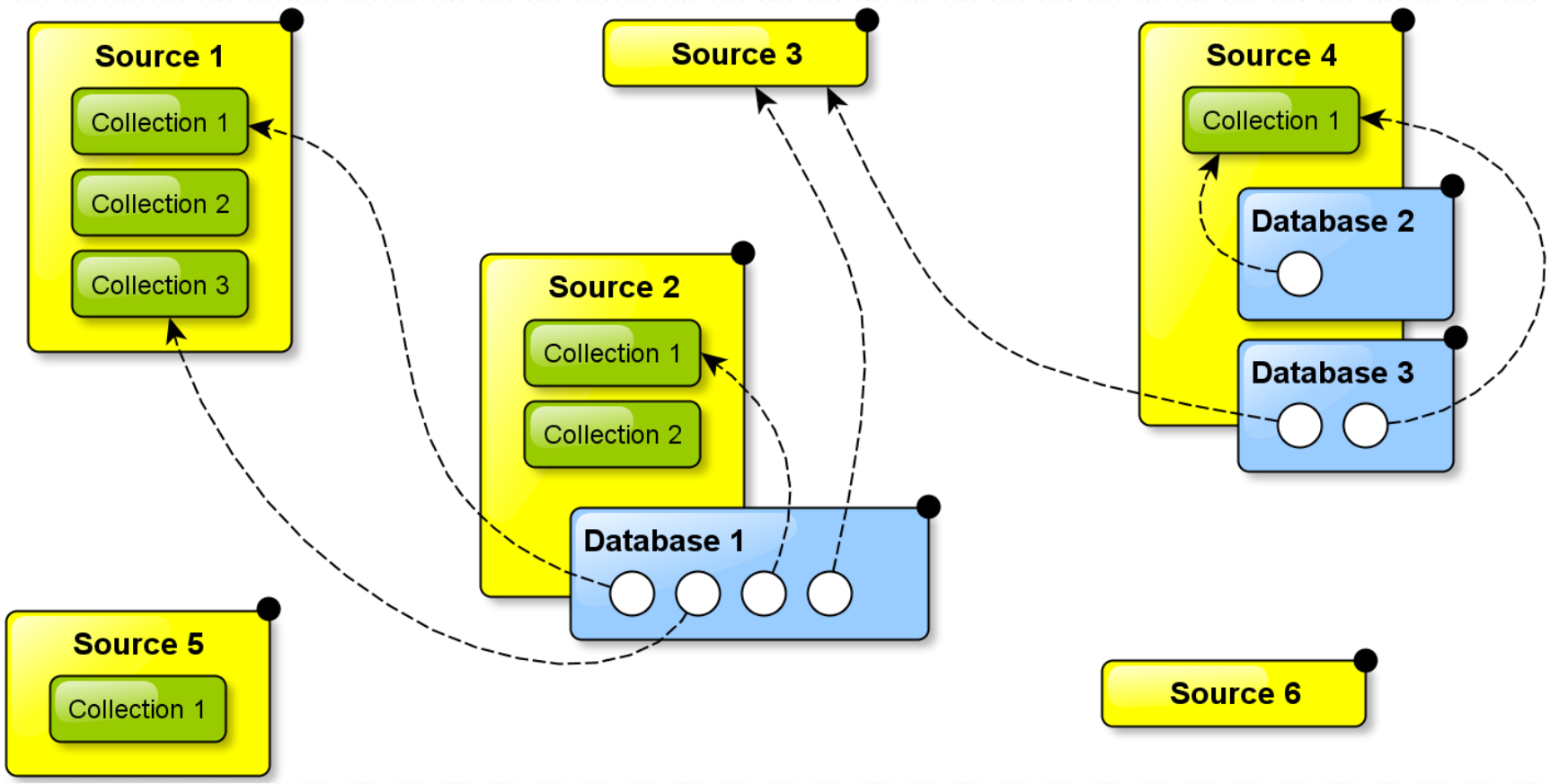
# Ideas

- **String compression**
  - Repeating string values
    - URIs and literals
  - Unique integer identifiers
    - Efficient processing
    - Space requirements
  - Translation maps
    - Both directions
    - Based on B-trees

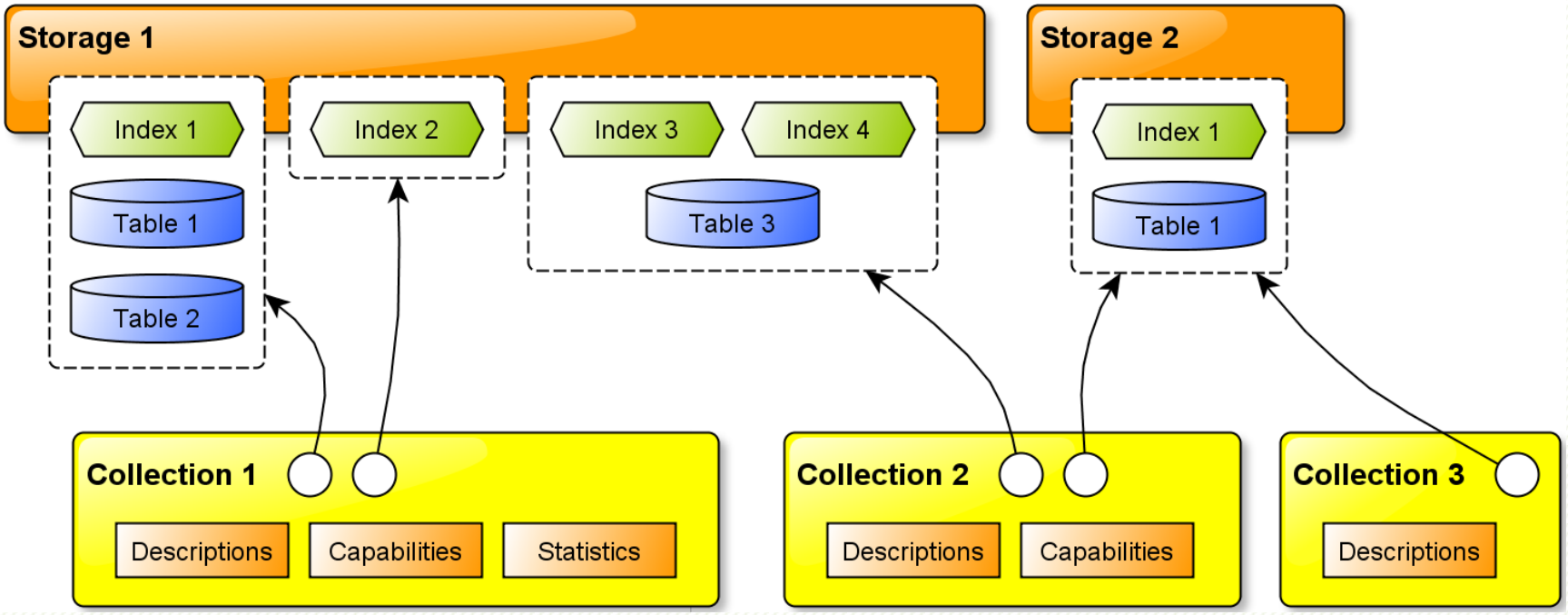
# Ideas

- **Data pruning**
  - Idea
    - Query optimization
    - Relevant data
  - Methods
    - Filtering selections
    - Join ordering
  - Problem
    - Partial knowledge

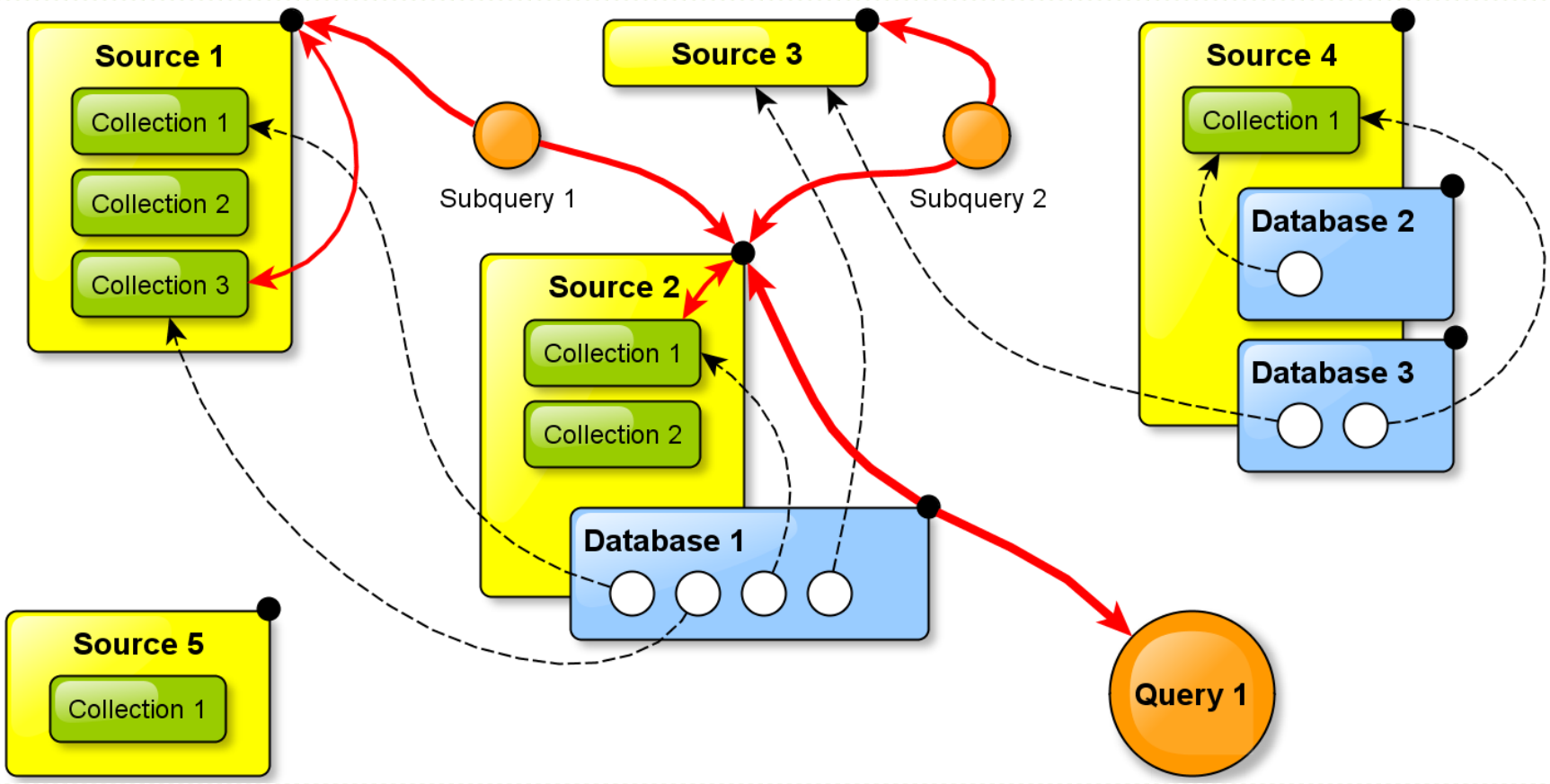
# Framework


Source 1
Collection 1
Collection 2
Collection 3
Source 3
Source 4
Collection 1
Database 2
Database 3
Source 2
Collection 1
Collection 2
Database 1
Source 5
Collection 1
Source 6

# Framework


Storage 1
Index 1
Index 2
Index 3
Index 4
Table 1
Table 2
Table 3
Storage 2
Index 1
Table 1
Collection 1
Descriptions
Capabilities
Statistics
Collection 2
Descriptions
Capabilities
Collection 3
Descriptions

# Framework


Source 1
Collection 1
Collection 2
Collection 3
Source 3
Subquery 1
Subquery 2
Source 4
Collection 1
Database 2
Database 3
Source 2
Collection 1
Collection 2
Database 1
Source 5
Collection 1
Query 1

# Framework

- Model
  - **Sources**
    - Distributed sources of data
  - **Collections**
    - Identified set of triples
    - Set of tables and indices
    - Descriptions, capabilities and statistics
  - **Databases**
    - Set of collections

# Issues

- Database
  - **Metadata**
    - Data descriptions
    - Querying capabilities
    - Auxiliary statistics
  - **Indices**
    - Database structure

# Issues

- Queries
  - **Processor**
    - Query decomposition
    - Source selection
    - Plan optimizations
    - Distributed evaluation

# Conclusion

- Problem
  - Transparent querying over distributed data
- Challenges
  - Data distribution, dynamicity and scaling
- Model
  - Sources, collections and databases
- Issues
  - Descriptions, indices and query processing

# Thank you for your attention…

XML and Web Engineering Research Group

**Charles University in Prague**