Ontology-based Data Access and Integration

Guohui Xiao

KRDB Research Centre for Knowledge and Data,
Faculty of Computer Science, Free University of Bozen-Bolzano, Italy

Freie Universität Bozen
Libera Università di Bolzano
Free University of Bozen-Bolzano

Sep 24 2017
Workshop on Novel Techniques for Integrating Big Data (BigNovelTI 2017)

What is big data?

Big data in reality

Challenges in "Big Data" Era

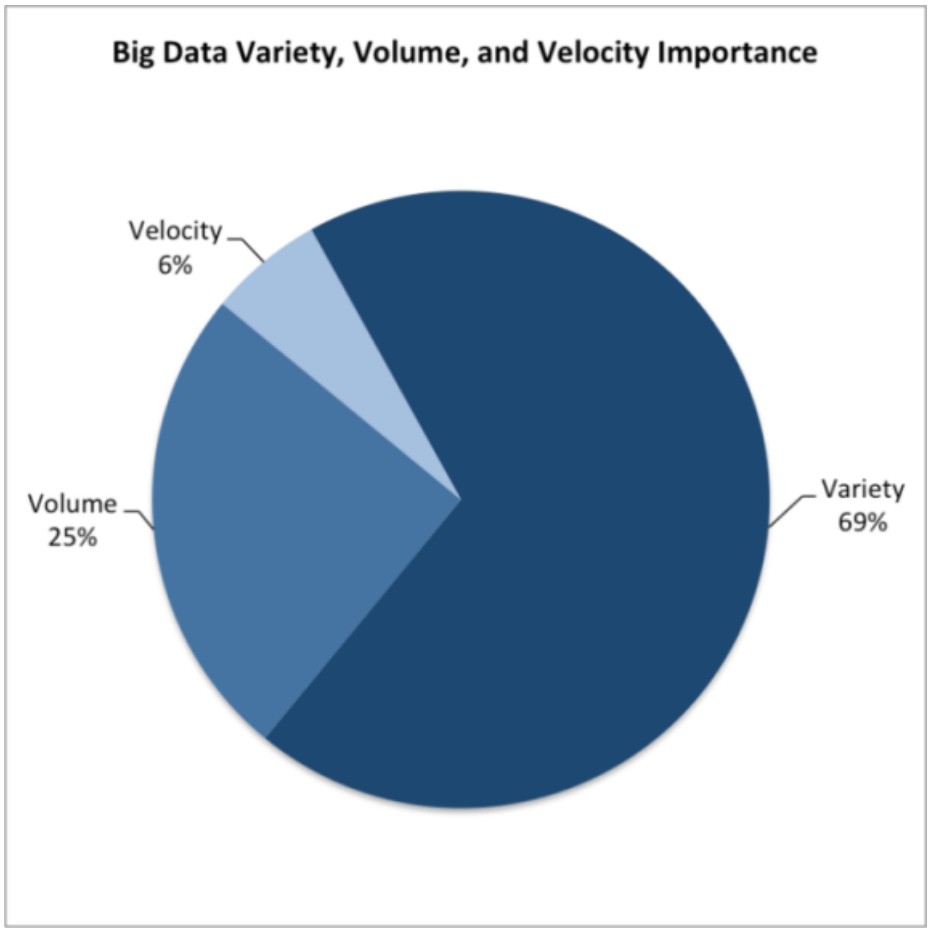
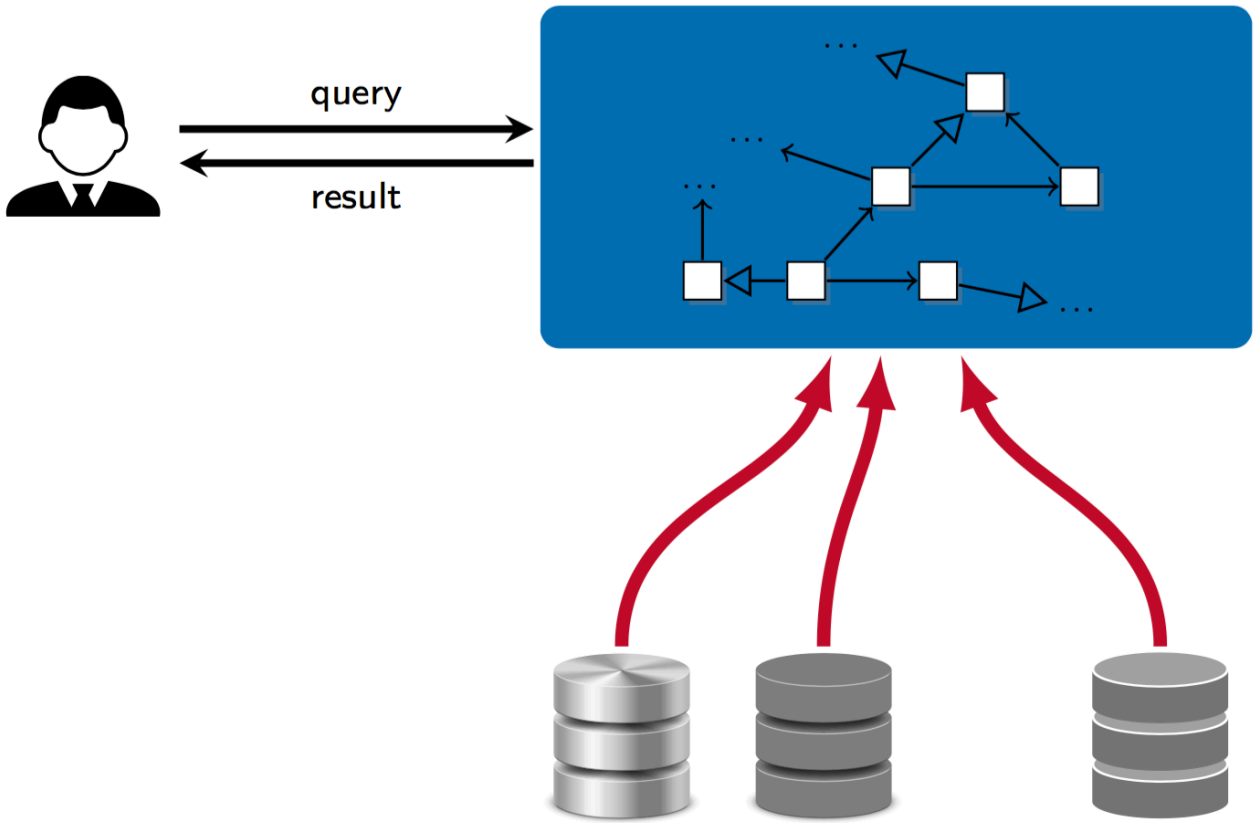


Figure: "Variety, Not Volume, Is Driving Big Data Initiatives" MIT Sloan Management Review (2016年3月28日)

http://sloanreview.mit.edu/article/variety-not-volume-is-driving-big-data-initiatives/

unibz

Ontology-based Data Access/Integration, OBDA/I)



- A common data model for data integration
- Virtual approach vs. ETL
- Two levels of data integration:
  - At the data source level (using DB integration techniques)
  - At the ontology level (e.g. SPARQL Federation)
- OBDA Systems: Ontop, Mastro, Morph, D2RQ, Ultrawrap

unibz

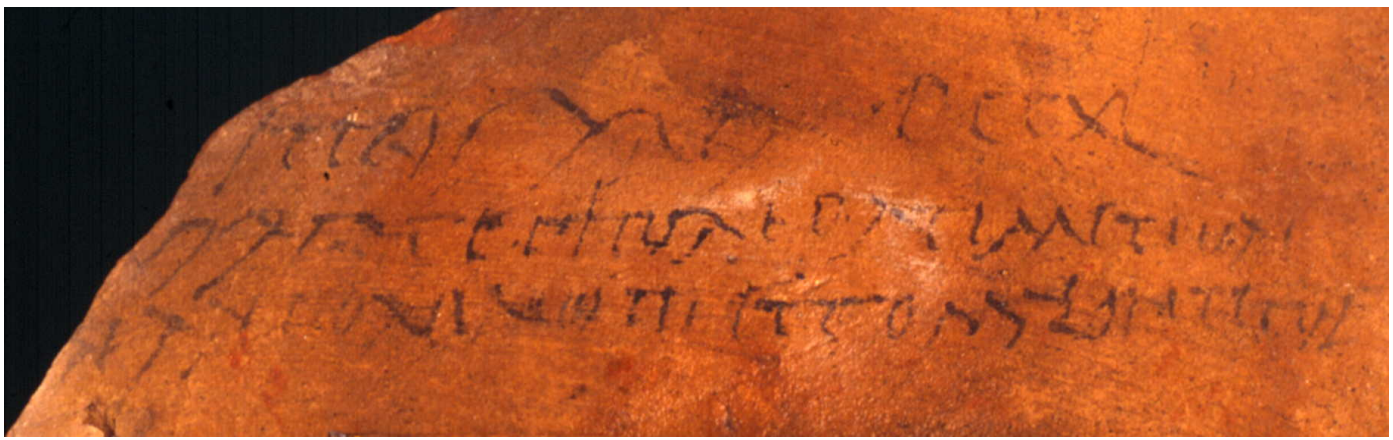
Ontop

- State-of-the-art OBDA system
- Compliant with the RDFS, OWL 2 QL, R2RML, and SPARQL standards.
- Supports all major relational DBs
  - Oracle, DB2, MS SQL Server, Postgres, MySQL, Teiid, Exareme, etc.
- Open-source and released under Apache 2 license
- Development of Ontop:
  - development started in 2009
  - already well established:
    - +300 members in the mailing list
    - 10000 downloads since May 2015
  - main development carried out in the context of the EU project Optique

unibz

Use Case: EPNet project

- Ontology-based data integration for humanities and archaeologists
- ERC advanced grant EPNet "Production and distribution of food during the Roman Empire: Economics and Political Dynamics".
- Linking three datasets (using ETL):
  1. the EPNet relational repository
  2. the Epigraphic Database Heidelberg
  3. the Pleiades dataset



unibz

Use Case: EMSec project

- German BMBF project EMSec: real-time services for maritime security
- SPARQL federation to access different kinds of data sources:
  - SPARQL endpoints of Ontop over *in situ* data
  - open SPARQL endpoints: Geonames, DBPedia

unibz

Challenges in OBDI

- How to map heterogeneous data sources in different data formats?
- How to link entities from different data sources?
- How to perform reasoning?
- How to deal with data quality issues?
- How to efficiently evaluate queries?
- How to explain query answers?
- How to handle streaming data in real time?

unibz