RESEARCH @ WISE

Large Scale Data Storage, Management and Processing

Prof. dr. Jan Hidders

March 26, 2018

VRIJE
UNIVERSITEIT
BRUSSEL

1. Introduction
2. Key-value stores and Document stores
3. Graph databases and RDF stores
4. Graph processing algorithms
5. Large-Scale Data-Processing Platforms

RELEVANT RESEARCH DOMAINS

- Databases
  - Key-value stores, Document stores
  - Graph stores, RDF stores
- Ontologies, Semantic Web and Data Integration
  - Ontology Description Languages
  - Data Mapping and Data Integration
- Graph Processing Algorithms
  - Community detection
  - Random Graph Generation
- Large-Scale Data-Processing Platforms
  - Declarative Data Processing Language Design
  - Query Optimization for Large Scale Data Processing

NOSQL DATABASES

MOTIVATION

- New types of DBMSs are emerging for different reasons:
    - disks are getting bigger, memory banks and SSDs are getting cheaper
    - distributed processing has become easier and more available, and so scaling out has become more important
    - processors are changing: more cores and more caches
- And so NoSQL databases emerged
- But DBMS technology is now converging:
    - RDBMSs are starting to incorporate NoSQL features
    - NoSQL database are starting to incorporate RDBMS features
- And so many things are happening in research at the moment.

NOSQL DATABASES

RESEARCH TOPICS

Developing an ORM-based DBMS

Develop a light-weight NoSQL DBMS whose native data model is ORM. We will investigate and develop update and query languages for the data model, as well as constraint languages. One specific topic is the efficient maintenance of database constraints as specified in ORM.

GRAPH DATABASES AND RDF STORES

MOTIVATION

- Special case of NoSQL databases
- Graph analysis is crucial in many new contexts:
  - social media, life sciences, telecommunication, crime fighting, ...
- And graphs actually give a very intuitive way of representing data
- Typical graph processing queries are hard to do in general DBMSs
  - e.g., often requires recursion in relational DBMS
- RDF Stores are essentially also graph databases
  - with similar research challenges for querying, indexing, optimization, ...

GRAPH DATABASES AND RDF STORES

RESEARCH TOPICS (1/2)

Typing for regular path queries

Investigate existing schema languages for graph databases, including ORM schemas, and develop typing mechanisms for typing (subsets of) graph query languages such as regular path queries.

Implementing extended regular path queries

Extended forms of path queries have been proposed in the literature. It is investigated how hard they are to implement in existing graph databases and graph processing frameworks.

GRAPH DATABASES AND RDF STORES

RESEARCH TOPICS (2/2)

Benchmarking graph datastores

Exploring existing benchmarks, including those for XML databases, and compare existing graph stores on them.

Indexing for graph databases

New types of indexes are explored, for typical graph queries. This includes indexes for queries over large graphs, as well as queries over large sets of graphs.

GRAPH PROCESSING ALGORITHMS

MOTIVATION

- Distributed processing brings new problems and opportunities for processing graphs
- Has lead to the development of new systems to program graph analytical computations, e.g., Giraph and GraphLab, and new algorithms
- We investigate algorithms for:
  - *community detection* and
  - *graph generation*

GRAPH PROCESSING ALGORITHMS

RESEARCH TOPICS

Improve existing community detection algorithms

A well-known algorithm community detection is the *Louvain method*. It has recently been implemented in Spark, and although this allows the processing of larger graphs, its performance and correctness has not been well-researched. The challenge is to look into this in more detail.

Generating benchmark graph that are *scale-free* and have the *small world property*

Benchmarks for graph processing require that we can generate graph with certain properties of (almost) arbitrary size. The problem of generating such graphs that are both scale free ($P(k) \sim k^{-\gamma}$) and have the small-world property ($L \approx \log(N)$) is still a challenge.

LARGE-SCALE DATA-PROCESSING PLATFORMS

MOTIVATION

- Many large / big data processing platforms have emerged
- But require much expertise from the programmer
- The challenge is to make this easier / more declarative (like in DBMSs)
- Two basic approaches for high-level languages:
  - Logic-based: datalog (like Prolog, but simpeler)
  - Workflow-based: graph of tasks

LARGE-SCALE DATA-PROCESSING PLATFORMS

RESEARCH TOPICS

Implementing Extended Datalog for Graph Processing on top of Spark

We are developing a research prototype, called *DatalogRA*, implemented on top of *Spark*, where we investigate the possibilities to optimise the data processing workflow before it is executed by Spark.

Implementing a JSON query and transformation tool

We are developing a tool to query and transform JSON based on a popular query language called SQL++ that adapts SQL to deal with semistructured data.

Thank You

Questions?

mailto:jan.hidders@vub.be