Agenda

- Why benchmarking
- Presentation of LDBC
  - Remarks
  - Who is who
  - Project overview: WP and Task Forces
  - TUC, Technical Users Community
- Benchmarking RDF and GDB
  - Common issues
  - Open questions
Why benchmarking

• Two main objectives:
  – Allow final users to assess the performance of the software they want to buy
  – Push the technology to the limit to allow for progress

• Main effort in DB benchmarking up to now
  – TPC: Transaction Processing Performance Council
    • Relational DBs: Transactional and DSS
LDBC

• Objectives:
  – Benchmarks for the emerging field of RDF and Graph database management systems (GDBs)
  – Spur industry cooperation around benchmarks
  – Create LDBC foundation during 1Q 2013.

  – Become a technology push effort, making improvements measurable
  – Become the de-facto research benchmark, usable, interesting and open to inputs.
Preliminary remarks

• Nature:
  – LDBC is different from other EC projects.
  – The objective for LDBC is to survive after the end.

• Opportunity:
  – Have a benchmarking effort sponsored by EC.
  – Focal point for the vendor community.
  – Showcase to the user community.

• Collaboration:
  – LDBC should lead to great achievements and world recognition, with the help of all the community: industry, technologists and users.
Who is involved

• FORTH, research centre, Greece
• TUM, research centre, Germany
• UIBK, technology centre, Austria
• Neo Technologies, Graph management, Sweeden
• OGL, RDF management, UK
• ONT, RDF management, Bulgaria
• VUA, research institution, Netherlands
• DAMA-UPC, research institution, Spain
Project Overview: the WPs

Maximizing real world impact
- WP5: Community Building and Dissemination
- WP6: Exploitation

Scenarios, workloads

Management
- WP7: Management

Research
- WP2: Query Processing
- Choke Point Analysis
- WP3: Graph Choke Point Analysis
- WP4: Semantic Choke Point Analysis

Integrated benchmarks

Fundamentals & Integration
- WP1: Common Benchmark Methodology

Benchmark definitions, data sets

Benchmark principles, methods, tools
Matrix Organization

- EU project reporting activities (WPs)
- LDBC benchmark task force activities (TFs)
Task forces

• Focusses on specific benchmarking effort (i.e. transactional, analytical, integration for RDF/GDB)
• Decides on the Use Case to be used
• Procedure:
  – Designs and implements data generation (characteristics, scale, etc.)
  – Incorporates the generic methodology
  – Designs specific workload
    • Choke points specific for the effort
    • Incorporating the needs from users
    • Incorporating the opinions from industry
Technical User Community (TUC)

- It will be the driving force for LDBC:
  - To help understand users needs and decide use cases
  - To decide the type of problem/scenarios to be tackled, i.e. task forces to be deployed
  - To provide typical queries placed to RDF and GDBs

- First TUC meeting, Barcelona 19-20 Nov.
  - Start with an on-line questionnaire:
    http://goo.gl/PwGtK
  - The outcomes will determine important directions
Common issues

• Use case for RDF and GDBs:
  • Social Network Analysis
  • Semantic Publishing, specific for RDF (SP)

• Methodology:
  • Audited benchmarks
  • Specific rules, similar to TPC

• Workload for RDF
  – Throughput, concurrency
  – Traversals
  – Reasoning
  – Data updates
  – Integration: LOD, geonames, etc.
  – SP: semantic annotation support, relationship btwn ontologies and instances, links to other content, text and metadata.

• Workload for GDBs
  – Throughput, concurrency
  – Traversals, shortest paths, pattern matching, clustering algorithms
  – Data updates, transactionality
  – Update semantics (serializable/acid vs delayed commit vs batch)
  – Raw traversal speed, use of indexes
Open questions

- **Use cases:**
  - How realistic would you see synthetic data generation?
  - Use of real data like twitter, Facebook or Open Ontologies?
  - Any suggestions for use cases?
  - Any suggestion for scenarios: analytical, transactional, integration, others?

- Would it make sense to propose open benchmark scenarios?

- The tight rules of TPC:
  - Are those against the realism of benchmarks?
  - Can we solve this in any flexible way?

- Will RDF and GDB move towards the same technology/solutions?

- GDBs: no standard language. How to proceed?