Request For Information

RFI

iMinds - Ghent University

Date: May 19th, 2016

Contents

1. Confidentiality

All information included in this RFI is confidential and only intended for the recipient.
No information included in this document or in discussions connected to it may be disclosed to any other party.

2. Background

2.1. iMinds - Ghent University

Our research group is part of iMinds - Ghent University and has a particular focus on machine learning, information retrieval, multimedia computing, exploratory data mining, data modelling and visualization, knowledge and reasoning, coding and information theory, robotics and control, and the use of high performance computing for these tasks.

We have world-leading expertise in the entire data value chain, from data acquisition, storage, representation and coding, to mining and learning from data, and finally valorization. Reflecting this, we are engaged in a wide range of activities, including fundamental/basic research, applied research, and contract-based research as well as consultancy for industrial partners.

2.2. Purpose

With this RFI, we request information regarding the optimal configuration of your database product for a benchmark study and (optionally) a written consent to mention your brand name in publications about the benchmark results. The written consent is required prior to the actual benchmarking process and will cover the full set of benchmark results. A written consent can be denied in which case the benchmark results will be anonymized to not violate any licensing agreements.

The same information will be gathered from different RDF database vendors and will be used to publish a current state of the art in RDF database systems.

3. Terminology and baseline

- **Default Configuration**
  - The default configuration is the configuration as described via a public online resource.
  - For store **Virtuoso** the default configuration is defined by the following resource/set-up guide:
    - **http://virtuoso.openlinksw.com/dataspace/doc/dav/wiki/Main/VirtPayAsYouGoEBSBackedAMI**
  - The default configuration for the store will use in particular:
    - **DirsAllowed = ., ../vad, ../data**
    - **MaxResults => 1 000 000 000**
    - **BufferSizes adapted to instance memory**

- **Watdiv**
  - The Waterloo SPARQL Diversity Test Suite(http://dsg.uwaterloo.ca/watdiv/) is a benchmarking system for RDF data management systems which performs a wide range of SPARQL queries which differ selectivity and structural properties.
  - It consists of two components: an RDF dataset generator and a query (template) generator. It runs a benchmark

- **Ontoforce**
  - Ontorce is a company (http://www.ontoforce.com/) which offers a semantic search platform called DISQOVER. Ontoforce provides their internal data for a second benchmark run on life science data.

- **AWS**
  - Amazon Web Services is a cloud provider. In this research AWS EC2 service is used to provides linux virtual machines for benchmarking.

- **PAGO instance**
  - Pay-as-you-go instances are Amazon Machine Images (ami) offered on Amazon Marketplace. These machine images are pre-installed on AWS EC2 instances.
  - PAGO link: **https://aws.amazon.com/marketplace/pp/B011VMCZ8K**

- **SPARQL Query Benchmarker**
  - SPARQL Query Benchmarker is a tool for running benchmarks against HTTP- accessible SPARQL systems.

4. Scope

4.1. Hardware for benchmarks with optimized configurations.

Database instances are run on AWS EC2 instances:
https://aws.amazon.com/ec2/instance-types/

- Instance types: r3.2xlarge
- 1 TB EBS Volume
- Provisioned IOPS for EBS volume (3000 IOPS)

SPARQL query benchmarker:

- 1 c3.2xlarge instance
- Ubuntu 14.04 LTS ami
- Query timeout parameter set to 300s
- 5 benchmark runs
- 1 warm up run
- 1 run = 1 random mix of 400 fixed queries from 20 templates (Watdiv)
- 8 query threads
- The most deviating run is considered an outlier and not used for calculating query runtime averages.

4.2. All benchmark setups

The benchmarks are conducted while exploring a space of parameters as defined below:

(Dataset, AWS hardware, Default/optimal configs, Number of nodes)

- *Dataset*: Watdiv datasets: 10M, 100M, 1000M and 4000M RDF triples. One nondisclosed dataset of Ontoforce of approximately 2500M triples

- *AWS hardware*: initial benchmarks are conducted on r3.xlarge instances, benchmarks with optimized configurations will be conducted on r3.2xlarge instances.

- *Default/optimal configs*: Optimal configs are based on the scripts or configuration files as provided by the RDF database vendor. Default configuration is defined in section 4.

- *Number of nodes:* benchmarks are run on 1 instance or 3 instances if this is supported by the RDF database.

Since an exhaustive exploration of the benchmark parameter space is not feasible due to resource constraints the following benchmarks will be / have been conducted and will be published after passing a peer-review process:

- Baseline Benchmarks:
  - (10M, r3.xlarge, default, 1)
  - (100M, r3.xlarge, default, 1)
  - (1000M, r3.xlarge, default, 1)
  - (1000M, r3.xlarge, default, 3) (only if supported)

- Vertical scaling Benchmark
  - (1000M, r3.2xlarge, default, 1)

- Enterprise Supported Benchmarks (see 6. statement of need)
  - (1000M, r3.2xlarge, optimal, 1)
  - (1000M, r3.2xlarge, default, 3) (only if supported)

- Life Science Benchmark
  - (Ontoforce data 2500M, r3.2xlarge, optimal, 1 or 3)

- Horizontal scaling Benchmark
  - (4000M, r3.2xlarge, optimal, 1 or 3)

NOTE: For all RDF databases inferencing is turned off.

  - (https://sourceforge.net/projects/sparql-query-bm/)

5. RFI procedure

5.1. Information for internal reproduction of results

Queries and datasets for public benchmarks:

- **Watdiv10M**:
    - Dataset: https://s3.amazonaws.com/watdiv/watdiv.10M.nt.gz
    - Querylist: https://s3.amazonaws.com/watdiv/queries10M/listfiles.txt
    - Query Files: https://s3.amazonaws.com/watdiv/queries10M/templated.tar.gz

- **Watdiv100M**:
    - Dataset: https://s3.amazonaws.com/watdiv/watdiv.100M.nt.gz
    - Querylist: https://s3.amazonaws.com/watdiv/queries100M/listfiles.txt
    - Query Files: https://s3.amazonaws.com/watdiv/queries100M/templated.tar.gz

- **Watdiv1000M**:
    - Dataset: https://s3.amazonaws.com/watdiv/watdiv.1000M.nt.gz
    - Querylist: https://s3.amazonaws.com/watdiv/queries1000M/listfiles.txt
    - Query Files: https://s3.amazonaws.com/watdiv/queries1000M/templated.tar.gz

- **Watdiv4000M**:
    - Dataset: https://s3.amazonaws.com/watdiv/watdiv.4000M.nt.gz
    - Querylist: https://s3.amazonaws.com/watdiv/queries4000M/listfiles.txt
    - Query Files: https://s3.amazonaws.com/watdiv/queries4000M/templated.tar.gz

5.2. How to deliver the answer

Send an email containing

1. A written consent for publication of all benchmark results
2. A set of configuration scripts or files to automate the deployment of your database product on AWS, if ailable using the PAGO AMI.

5.3. Contacts

For questions regarding this RFI, you are welcome to contact:

Dieter De Witte
+32 477 220 978
drdwitte@gmail.com

5.4. Timeframe

This is the timeframe for the RFI and an eventual coming project

May 19, 2016	– The RFI is sent out
June 2, 2016	– Last date for questions
June 5, 2016	– 23:59:59 GMT+2 is the deadline to deliver the requested information
June 6, 2016	– Benchmarking starts
July 31, 2016	– Benchmark Delivery Deadline
Q3, 2016	– Submission of results for Peer-Review

6. Statement of need

In the context of academic conferences and academic journal papers relevant to the semantic web, computer science and information systems we plan to compare the default configuration against an optimized configuration.

1. Written consent for publishing the results.

2. You as RDF database vendor are asked to provide a configuration file or a script (which can be made public) to run each benchmark under optimal conditions.

Example
One such optimization parameter is query timeouts on the server side.
Client-side timeout of Sparql benchmarker is 300s, stores might anticipate this.
Time-outs less than 300s are **not** allowed. The benchmark will expose if a store has a lower timeout configured.