Oracle Database: Benchmarking a Trillion Edges RDF Graph

Oracle White Paper

June, 2020

Public

PURPOSE STATEMENT

This document provides an overview of how Oracle conducted a one trillion edges LUBM benchmark (LUBM 4400k) in September 2014 with RDF Graph in Oracle Database on an Oracle Exadata Database Machine.

DISCLAIMER

This document in any form, software or printed matter, contains proprietary information that is the exclusive property of Oracle. Your access to and use of this confidential material is subject to the terms and conditions of your Oracle software license and service agreement, which has been executed and with which you agree to comply. This document and information contained herein may not be disclosed, copied, reproduced or distributed to anyone outside Oracle without prior written consent of Oracle. This document is not part of your license agreement nor can it be incorporated into any contractual agreement with Oracle or its subsidiaries or affiliates.

This document is for informational purposes only and is intended solely to assist you in planning for the implementation and upgrade of the product features described. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. The development, release, and timing of any features or functionality described in this document remains at the sole discretion of Oracle.

Due to the nature of the product architecture, it may not be possible to safely include all features described in this document without risking significant destabilization of the code.

TABLE OF CONTENTS

INTRODUCTION

One trillion is a really big number. What could you store with one trillion facts?

- 1000 tweets for every one of the 1 Billion Twitter users.
- 770 facts about every one of the 1.3 Billion Facebook users.
- 10 facts from 107 Billion sensors, located somewhere on the planet.
- 400 metabolic readings for each of the 2.5 Billion heart beats over an average human life time.
- 12 facts about every one of the 86 Billion neurons in the human brain.
- 5 facts about every one of the 200 Billion stars in the Milky Way Galaxy.
- 7 facts about every one of the 150 Billion galaxies in the universe.
- 6,350 facts about each of the 158 Million books in the Library of Congress, the largest in the world.
- 10 facts about each of the 107 Billion people who ever lived

Resource Description Framework (RDF) graphs and the analytics they permit are becoming central to big data applications for social networks and linked data. These applications are often found in public sector, healthcare and life sciences, finance, media, and intelligence communities. The World Wide Web Consortium (W3C) [1] defines RDF and the Web Ontology Language (OWL) graph standards for representing and defining semantic data and rules, and SPARQL, a pattern matching query language designed specifically for graph analysis.

The basic nature of an RDF graph facilitates identification, integration, and discovery:

- **RDF data elements are globally unique**. They are defined using Uniform Resource Identifiers (URIs) that enable a consistent metadata layer for integration of disparate data sources.
- **RDF data elements are linked to form a graph**. Elements are used to make statements in the form of subject-predicate-object triples. Predicates (edges) link the subject and object (nodes) and can describe any relationship or property. The object can be another subject to link triples together to form a graph or a literal that is an attribute of the subject. The triples can be further qualified with a fourth named graph component, which are referred to as RDF quads.
- **The RDF model allows easy, dynamic schema evolution**. Adding a new schema element is as easy as inserting a triple with a new predicate.
- **RDF and SPARQL support ad hoc queries.** Queries may not be known when the schema is designed.
- **The RDF model makes an Open World Assumption that can facilitate discovery**. It assumes that what is unknown is undefined, rather than false, as is the case with relational technology. It also has technologies that help discover missing results.
- **RDF embeds semantics (meaning) directly in the data.** Entities are categorized with classes, predicates are properties or relationships, and they are all part of the data, unlike column headers, foreign keys, or constraints in relational data.
- **RDF supports machine-driven inferencing for discovery**. The OWL semantic language and rules used to define the predicates in triples are based on formal Description Logics that enable automatic discovery, such as identifying "same-as" relationships between different terms with the same meaning in two applications. The set of inferred triples (conclusions that can be drawn) is referred to as an entailment.
- **The OWL language can unify an enterprise's dictionaries, vocabularies, and taxonomies**. All of the terms used by the applications in an enterprise can be related to each other and form concepts. Concepts are managed as one or more domain-specific ontologies and stored in RDF graphs. Ontologies are linked to the asserted instance data in graphs and used for inferencing and querying. This is another capability that facilitates creating a consistent metadata layer for data integration.

[1] http://www.w3.org/RDF/

The Lehigh University Benchmark[2] (LUBM) is a de facto industry standard benchmark for evaluating RDF graph store product performance. It is used by RDF graph store vendors to characterize the load, inference, and query performance of their product. Vendors post results on the W3C Large Triple Stores page[3]. End users use LUBM benchmark results as part of their evaluation of an RDF Graph store product. The benchmark includes a W3C OWL-based university ontology, a data generator to create a graph of any size, and fourteen test queries.

Oracle conducted a one trillion edges LUBM benchmark (LUBM 4400k) in September 2014 with Oracle Database 12.1.0.1 standard installation on Exadata Database Machine and achieved two record-setting accomplishments:

- Oracle believes its benchmark is the largest complete LUBM benchmark in the industry to date.
- The combined load, inference, and query results are the fastest RDF graph performance numbers reported; this is especially significant for a benchmark of this scale and complexity.

The details for this benchmark, including results, configuration, and best practices are discussed in the next section of this paper.

A TRILLION EDGES RDF GRAPH BENCHMARK ON ORACLE DATABASE

As big data graphs grow from billions to trillions of relationships it becomes increasingly important to characterize product performance. Oracle conducted an RDF graph LUBM 4400k benchmark. It involved loading, inferencing, and querying over one trillion edges with RDF Graph in Oracle Database on an Oracle Exadata Database Machine. The LUBM environment was used to generate data about universities and their departments. The data was created and ordered into 4.4 million named graphs by expanding the triples into quads. There was one named graph per university. The overall graph included 605.4 billion unique asserted quads and an entailment of another 475.6+ billion quads.

The Results

The RDF Graph LUBM 4400k benchmark on Oracle Database achieved the following results:

- **Data Loading Performance:** 1.420 million Quads Loaded and Indexed per Second.
  - 605.4 Billion Quads were loaded and two indexes were created in 115.2 hours.
  - Note: Graph loading in Oracle Database is unique in the industry for checking that quads are well formed and for removing duplicates.
- **Inference Performance:** 1.527 million Triples Inferred and Indexed per Second.
  - 475.6 Billion Triples and two indexes were created in 86.5 hours.
- **SPARQL Query Performance:** 1.130 Million Query Results per Second.
  - 92.5 Billion Answers were generated in 22.5 hours.

A Trillion triples graph

Asserted	Inferred	Total	Answers
605.4 Billion Quads	475.6 Billion Triples	1.081 Trillion Quads	92.5 Billion

The Configuration

The market-leading performance of this benchmark was due to the combination of the native RDF graph store capabilities of RDF Graph in Oracle Database on the balanced configuration of an Oracle Exadata Database Machine X4-2. The unique capabilities of the Exadata Database Machine that assisted benchmark query performance include:

- **Smart Scan** that reduces data movement between storage servers (cells) and database server by pushing queries down to the storage cell,
- **storage indexes** used by the storage cell to read only regions of storage that have relevant data, and
- **InfiniBand fabric** that provides fast transfer (40 Gb/second) of relevant bytes back to the database server to complete the execution of a query.

[2] http://swat.cse.lehigh.edu/projects/lubm/

[3] https://www.w3.org/wiki/LargeTripleStores

The Oracle Exadata Database Machine X4-2 High capacity full rack was configured as follows:

- 8 database nodes and 14 storage nodes for a total of 168 CPU cores
- 2 TB total RAM and 44.8 TB Flash Cache
- ZS3-2 storage with 2 controllers and 8 trays of disks
- Software: Oracle Database 12.1.0.1 standard installation on Exadata Database Machine.

Best Practices Used

The best practices fall into two categories, database settings and tuning.

Database settings:

- SGA_TARGET=132GB
- PGA_AGGREGATE_TARGET=100G
- Open cursors=1000
- Processes=1000
- 32K blocksize given to all graph tablespaces
- a TEMP group created with 3 bigfile tablespaces
- Use of the auto-allocate option for allocation of tablespace extents coupled with a large, 8 million bytes extent size. This reduced the number of waits caused by HV enqueue contention; that is, waits on a lock that is used to alter the high-water mark in a tablespace. As a result, contention among multiple processes requesting tablespace expansion could be avoided.
- DOP settings (296, 256, 192) for automatic degrees of parallelism used in loading, inferencing, and querying.
- Use of additional compression beyond basic table compression during inferencing provided by the Hybrid Columnar Compression feature of Oracle Exadata Database Machine.

Tuning:

- Oracle Enterprise Manager provided specific performance insights into operations for tuning. The methodology used is documented in the Oracle Database Performance Tuning Guide.[4]

CONCLUSION

RDF graphs provide unique, standards-based, big data capabilities for metadata integration, and discovery to support social networks and linked data applications in a variety of industries. RDF Graph demonstrated industry-leading scalability and performance for loading, inference, and querying a one trillion edges RDF graph managed in Oracle Database. The LUBM 4400k RDF graph benchmark benefited from the balanced hardware configuration of an Oracle Exadata Database Machine X4-2. The best practices settings used to achieve these benchmark results are also generally applicable to real-world applications on Oracle Exadata Database Machine and other balanced hardware configurations.

[4] http://docs.oracle.com/database/121/TGDBA/toc.htm

CONNECT WITH US

Call +1.800.ORACLE1 or visit oracle.com.
Outside North America, find your local office at oracle.com/contact.

blogs.oracle.com facebook.com/oracle twitter.com/oracle

Oracle Database: Benchmarking a Trillion Edges RDF Graph
June 2020