ARTICLE IN PRESS

Parallel Computing xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Parallel Computing

journal homepage: www.elsevier.com/locate/parco

Toward a data scalable solution for facilitating discovery of science resources

Jesse Weaver [a,]*, Vito Giovanni Castellana [a], Alessandro Morari [b], Antonino Tumeo [a], Sumit Purohit [a], Alan Chappell [b], David Haglin [a], Oreste Villa [c], Sutanay Choudhury [a], Karen Schuchardt [a], John Feo [a]

[a] Pacific Northwest National Laboratory Richland, WA 99354, USA
[b] Pacific Northwest National Laboratory Seattle, WA 98109, USA
[c] NVIDIA Research Santa Clara, CA 95051, USA

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Data intensive
Science metadata
Graph database
Scalability
Semantics

ABSTRACT

Data-intensive science simultaneously derives from and creates the need for large quantities of data. As such, scientists increasingly need to discover and analyze new datasets from diverse sources. Beyond the sheer volume of data, issues posed by the resultant data heterogeneity are often overlooked. We postulate that heterogeneity challenges can be solved (at least in part) with the adoption of the Resource Description Framework (RDF), a graph-based data model. In turn, this requires scalable graph query systems for discovering and analyzing data. Consequently, we investigate GEMS, a graph engine for large-scale clusters. We describe the features of GEMS that make it suitable for answering graph queries and scaling to larger quantities of data. We evaluate GEMS' ability to answer real science-based queries over real-world, curated, science metadata. We also demonstrate GEMS' ability to scale to larger datasets using a benchmark.

1. Introduction

Data-intensive science simultaneously derives from and creates the need for large quantities of data [1]. As such, scientists increasingly need to discover and analyze new datasets from diverse sources. Beyond the sheer volume of data, issues posed by the resultant data heterogeneity are often overlooked. Examples include: vocabularies in Comma-Separated Values (CSV) [2] and Resource Description Framework (RDF) [3] formats from the Global Change Master Directory (GCMD) [4]; NetCDF [5] data files from the Atmospheric Radiation Measurement (ARM) [6] climate research facility; and web pages from the International Soil Moisture Network (ISMN) [7].

Data heterogeneity challenges have been addressed (at least in part) in the Semantic Web (or Web of Data) field through various recommendations including the Resource Description Framework (RDF) [3]. RDF has been adopted in some scientific communities [8–12], including in our own previous work [13]. While it helps with the heterogeneity challenges, the adoption of RDF *exacerbates* the volume challenges. A recent benchmark study [14] suggests that the largest, surmountable RDF

* Corresponding author.
E-mail addresses: jesse.weaver@pnnl.gov (J. Weaver), vitoGiovanni.castellana@pnnl.gov (V.G. Castellana), alessandro.morari@pnnl.gov (A. Morari), antonino.tumeo@pnnl.gov (A. Tumeo), sumit.purohit@pnnl.gov (S. Purohit), alan.chappell@pnnl.gov (A. Chappell), david.haglin@pnnl.gov (D. Haglin), ovilla@nvidia.com (O. Villa), sutanay.choudhury@pnnl.gov (S. Choudhury), karen.schuchardt@pnnl.gov (K. Schuchardt), john.feo@pnnl.gov (J. Feo).

http://dx.doi.org/10.1016/j.parco.2014.08.002
0167-8191/© 2014 Published by Elsevier B.V.

datasets for most modern RDF databases are 1–10 billion RDF triples (graph edges) in size, and at those quantities, performance is significantly degraded. In our own work, a curation of science metadata has generated 1.4 billion triples, yet that constitutes only a small sample of the available science metadata. In order to handle even greater quantities of RDF triples with less cost to performance, many RDF databases (and relevant research) are moving toward parallel, cluster-based systems [15–23]. A similar trend is also occurring with generic graph databases [24,25].

Consequently, we are investigating a distributed, graph-based query system being developed at Pacific Northwest National Laboratory (PNNL) called the Graph Engine for Multithreaded Systems (GEMS) [26]. Postulating that the RDF data model is sufficient for solving the heterogeneity problems, we seek to evaluate how well GEMS can handle the problem of ever-growing volumes of data (i.e., data scaling [27]). GEMS stores RDF triples in distributed, in-memory indexes in such a way as to quickly and naturally support parallel graph walking, the fundamental operation for graph queries. It is built atop our custom Global Memory and Threading (GMT) [28] runtime system for clusters, designed from the ground up to tolerate the distributed, random data accesses that naturally occur when walking a distributed graph. GEMS is a work in progress, but our results herein demonstrate GEMS' capability and potential for scaling queries to large graphs.

As an extension of [13], the novel contributions of this work are as follows.

- An extended and more detailed description of the real-world, data-intensive, science-based use case first introduced in [13].
- An updated description of the GEMS software stack previously introduced in [13,29]. Specifically, our graph data structure has changed radically, and we no longer use distributed hashmaps for storing results.
- A more in-depth evaluation using GEMS to answer queries on science metadata than we were previously able to provide in [13], also using a dataset that is over ten times larger.
- An evaluation of GEMS' scalability using generated benchmark datasets. Unlike in [26], this paper evaluates on individual query times rather than (parallel) query throughput.
- A performance comparison between GEMS and Urika [30].

The rest of the paper is organized as follows. In Section 2, we describe an example of curating science metadata into the RDF data model so as to provide background for the kind of data over which we wish to query. We briefly introduce the RDF data model, the SPARQL query language [31], give a concrete example of the curated metadata, and provide example queries. Section 3 describes GEMS at a high level with emphasis on the characteristics of GEMS that improve data scalability of graph walking, the fundamental operation for answering graph-based queries over the curated science metadata introduced in Section 2. Section 4 presents experimental results for: (1) the performance of GEMS to answer queries over science metadata; (2) the scalability of GEMS to handle larger datasets using a benchmark; and (3) a performance comparison with Urika [30]. Section 5 covers related works, and conclusions and future work are discussed in Section 6.

2. Curated science metadata from RDESC

The investigation herein is motivated by ever-growing, heterogeneous, science metadata. In this section, we describe efforts in curating (i.e., coping with the heterogeneity of) science metadata as part of the Resource Discovery for Extreme Scale Collaboration (RDESC) project. We include this description for two reasons: (1) to establish and give an example of data heterogeneity in science; and (2) to provide background for the RDESC metadata that will be used to evaluate GEMS in Section 4.2.

2.1. RDF for metadata

Facilitating discovery of science resources requires semantically meaningful integration of otherwise disparate, heterogeneous metadata. These metadata are available in different binary or syntactic formats (e.g., variables and attributes in NetCDF [5] files, attributes in HDF [32] files, "hidden" web services, embedded in HTML pages, or even simple text) and usually do not have clear semantic interoperability (e.g., measurements for the same property can implicitly have different units, or variations in definitions of altitude). In order to integrate the metadata, we first establish a precise understanding of the metadata and then attempt to capture those semantics in an ontology (see Fig. 1). To integrate the metadata, we convert it from its native form to *RDF triples* — a graph-based data model that is flexible enough for nearly (if not truly) any kind of metadata.

As the term implies, RDF triples consist of three (ordered) parts: *subject*, *predicate*, and *object*. This simple data model constitutes a labeled, directed multigraph. A triple `S P O` uniquely identifies a directed edge in the graph going from the vertex uniquely identified by `S` to the vertex uniquely identified by `O`. `P` is effectively the edge label.

At present, the RDESC metadata consists of nearly 1.4 billion RDF triples, mostly metadata about ARM [6] data streams and GeoNames [33] locations, although it includes to a lesser proportion descriptions of Global Change Master Directory (GCMD) [4] locations, keywords, and datasets, as well as metadata from the International Soil Moisture Network (ISMN) [7]. Specifically, the RDESC metadata contains descriptions of 781,750 data resources, which can be broadly broken down

Please cite this article in press as: J. Weaver et al., Toward a data scalable solution for facilitating discovery of science resources, Parallel Comput. (2014), http://dx.doi.org/10.1016/j.parco.2014.08.002