Towards Big Linked Data: A Large-Scale, Distributed Semantic Data Storage

Bo Hu, Intelligent Society Platform Research Division, Fujitsu Laboratories of Europe Ltd (FLE), Hayes, UK

Nuno Carvalho, Intelligent Society Platform Research Division, Fujitsu Laboratories of Europe Ltd (FLE), Hayes, UK

Takahide Matsutsuka, Intelligent Society Platform Research Division, Fujitsu Laboratories of Europe Ltd (FLE), Hayes, UK

ABSTRACT

In light of the challenges of effectively managing Big Data, the authors are witnessing a gradual shift towards the increasingly popular Linked Open Data (LOD) paradigm. LOD aims to impose a machine-readable semantic layer over structured as well as unstructured data and hence automate some data analysis tasks that are not designed for computers. The convergence of Big Data and LOD is, however, not straightforward: the semantic layer of LOD and the Big Data large scale storage do not get along easily. Meanwhile, the sheer data size envisioned by Big Data denies certain computationally expensive semantic technologies, rendering the latter much less efficient than their performance on relatively small data sets. In this paper, the authors propose a mechanism allowing LOD to take advantage of existing large-scale data stores while sustaining its “semantic” nature. The authors demonstrate how RDF-based semantic models can be distributed across multiple storage servers and the authors examine how a fundamental semantic operation can be tuned to meet the requirements on distributed and parallel data processing. The authors’ future work will focus on stress test of the platform in the magnitude of tens of billions of triples, as well as comparative studies in usability and performance against similar offerings.

Keywords: Distributed Data Reconciliation, Fault Tolerance, Graph Storage, Key-Value Stores, Resource Description Framework (RDF) Store

INTRODUCTION

Nearly eight decades ago, Z1 was born with 64×22 bits of memory. Eighty years on, a modern desktop computer can easily hold an average of 500 Gigabytes and provides instant access to Petabytes, even Exabytes, of data through the Internet, exposing us to an amount much more than an individual human being can consume over her entire life time. Data, one of the fiercest “monsters” created by mankind, has overpowered its creator and yet, more and more evidently, the word “Big” seems failing to deliver a precise image of the size that we are trying to tackle – it is still growing in an unprecedented rate attribute to for example the high throughput scientific instruments (e.g. Large Hadron Collider at CERN and the

DOI: 10.4018/ijdwm.2013100102
widely used DNA microarrays). More recently, with the advances in Information Technology, the ungovernable nature of Data has started to demonstrate itself in our everyday business and personal life. Let’s take for example the overwhelming size of enterprise data accumulated by ERP and BI systems or the sheer number of emails that we send and receive on a day to day basis.

Yet, **volume** is only one challenging characteristic possessed by Big Data, along coming data **velocity** and data **variety**. In this paper, we concentrate on the **variety** aspect and envisage the merit of linking together heterogenous data sources to reveal knowledge that does not manifest in single data silos. Unveiling knowledge that lies in the interactions among data sets is, however, not easy. To many, Big interconnected Data sets are no more than a messy swap while trudging through it requires not only trained minds but also tools that can assist to weave the intertwined data into an orderly and easily-traversable knowledge network.

Linked Open Data (LOD) (Bizer et al., 2009) paradigm, an approach to cope with the variety of Big Data, has gained an increasing popularity in the past a few years and started to reach beyond academia (Hu & Svensson, 2010). LOD is underpinned by the idea of imposing a machine-readable semantic layer upon data so as to allow computers taking over some of the data analysis tasks exclusive to humans. At the heart of LOD is the Resource Description Framework, RDF\(^1\), a simple graph-based data modelling language enabling semantic mark-up of data. With RDF, LOD tries to piece together data silos and transform the current archipelagic data landscape into a connected data graph upon which complicated data analytics and business intelligence applications can be built.

The LOD vision, even though opening up great potentials, brought along with it tremendous challenges that were not previously seen in the Big Data field (Hausenblas et al., 2012). The challenges include: extending the current storage to accommodate semantics, distributing semantic data operations, automating semantic data analysis, etc. The present work aims to address the most fundamental one: large-scale semantic data (or triplified data) storage.

Large-scale semantic data storage is facilitated by large-scale data storage, while the latter has been extensively investigated by the Database, Internet, and other relevant communities (*c.f.* Toad World\(^2\) for a comprehensive survey). Effectively, thus far, the size issue is attacked by either scaling up with more high-end servers or scaling out taking advantage of low-cost commodity hardware available in massive numbers through the *cloud*. The inherent flexibility, plus low acquisition and operating costs, has rendered the latter a nearly perfect storage solution to large amounts of data. The only missing piece of the envisaged Big Linked Data jigsaw is a semantic layer that explicitly captures the meaningful relationships among data items so as to express data semantics.

The proposed storage solution accommodates storing and querying data that is encoded in RDF. We target a distributed fault tolerant system, that is able to scale in the amount of data to be managed. We use a distributed, ordered Key-value store (Seeger, 2009), where each RDF triple is regarded as the Key of the Key-value pair. This removes the necessity of handling and storing extra information such as data identifiers (widely adopted by apparently similar approaches). It also allows to query the information in a very efficient way by using range queries. Meanwhile the value part of the Key-value pairs provides a straightforward solution to associate metadata with individual triples. This gives a great potential to make the distributed data store more powerful, by for example integrating (directly in the storage system) new application specific functionalities, that would be less efficient if otherwise implemented outside the storage system.

The semantic layer also introduce a new dimension of consistency: often many data sources use different references to indicate the same real world object. A necessary and important step towards utilising LOD-enabled Big Data is to identify and reconcile multiple references for semantic consistency. Hereinafter, the term “reconciliation” is used to indicate
Related Content

Rule-Based Data Mining Cache Replacement Strategy
[www.igi-global.com/article/rule-based-data-mining-cache/75615?camid=4v1a](www.igi-global.com/article/rule-based-data-mining-cache/75615?camid=4v1a)

Online Processing of End-User Data in Real-Time Data Warehousing
M. Asif Naeem and Noreen Jamil (2015). *Improving Knowledge Discovery through the Integration of Data Mining Techniques* (pp. 13-31).
[www.igi-global.com/chapter/online-processing-of-end-user-data-in-real-time-data-warehousing/134529?camid=4v1a](www.igi-global.com/chapter/online-processing-of-end-user-data-in-real-time-data-warehousing/134529?camid=4v1a)
Spatial Data Mining: A Perspective of Big Data
[www.igi-global.com/article/spatial-data-mining/117158?camid=4v1a](www.igi-global.com/article/spatial-data-mining/117158?camid=4v1a)

Semantics-Aware Advanced OLAP Visualization of Multidimensional Data Cubes
[www.igi-global.com/article/semantics-aware-advanced-olap-visualization/1791?camid=4v1a](www.igi-global.com/article/semantics-aware-advanced-olap-visualization/1791?camid=4v1a)