Proceedings of
INTERNATIONAL CONFERENCE ON COMPUTING, COMMUNICATION AND ENERGY SYSTEMS (ICCCES-16)
In Association with IET, UK & Sponsored by TEQIP-II
29th -30th, Jan. 2016

Paper ID:CSEIT14

PROCESSING LINKED MULTIDIMENSIONAL DATA ON THE SEMANTIC WEB

Mr. Karan Gupta
ME Computer, JSCOE, SPPU
Prof. Poonam Lambhate (Doshi)
ME Computer, JSCOE, SPPU
Dr. Emmanual M.
HOD IT Dept, PICT, SPPU

Abstract— The Semantic Web has grown tremendously with more and more information on the web being available in the form of the Resource Descriptor Framework (RDF). This has paved the way for the RDF Data Cube Vocabulary (QB), which also became a W3C recommendation, and allowed for publishing statistical and linked data in the form of web data cubes. The RDF QB Vocabulary has since then been extended in the form of several vocabularies, many of which aim to provide analytics on top of linked datasets from disparate sources. Several governments have released statistically linked data freely which can be converted into RDF based Online Analytical Processing (OLAP) cubes allowing BI Tools and BI like operations on them. However, as the related technologies grow, a unifying framework needs to be defined. The aim of this paper is to serve as roadmap for future research and development to promote linked open data which can be published as data cubes on the Semantic Web. It presents a conceptual framework that enables publishing of multidimensional data from various heterogonous sources, combining and linking them, and allowing for processing of large volumes of information for analytics, without having the need to store them in traditional data warehouses.

Index Terms—Linked Multidimensional Data, OLAP, Semantic Web.

I. INTRODUCTION

Raw data is the basic necessity to arrive to a meaningful decision. This is the base of any analytical work-flow, where raw data is transformed and cleaned using complex Extract-Transform-Load (ETL) processes and stored in a data warehouse. This data allows for multidimensional analysis and aggregation which then provides key business facts, and allows to make meaningful decisions. Data cubes allow for efficient analytical queries on large volumes of data by aggregating fact information called measures at any desired granularity or detail. As information on the web continues to grow, a lot of effort has been put in relating data and to provide it with a meaning. The Semantic Web [1] as the name suggests allows for processing web data using the semantics or the meaning of the data. It does this by using a collection of technologies [2] like Extensible Markup Language (XML), Resource Description Framework (RDF), RDF Schema and Web Ontology Language (OWL) [6]. Additional information relating to a data sources has become increasingly essential to provide the big picture. More and more efforts are being put to make open data available in RDF format [3]. RDF data can be queried using SPARQL [11], just like an OLAP cube can be queried using SQL for analytical purposes. Data on the Semantic Web can be extensively linked based on the semantics and origins, but requires a standardized format that must be used by various technologies. The Semantic Web looks to provide linked data using relationships just like in a relational database. The Semantic Web does this using the RDF triples: subject, predicate and object [4]. By linking data on the web, various disparate sources can be connected to extract valuable information. RDF Data Cube Vocabularies like QB [7] and its extensions like QB4OLAP allow integrating and querying statistical multidimensional data directly available in RDF format.

II. SEMANTIC WEB TECHNOLOGY OVERVIEW

A. Architecture

The Semantic Web achieves linking of the data and integration of data by remove ambiguities using Ontologies. "An ontology is defined as a formal, explicit specification of a shared conceptualization". The architecture of the Semantic Web [1] is illustrated in Fig 1.

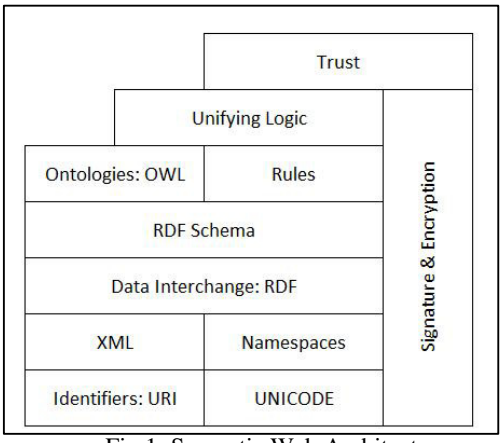

Fig 1: Semantic Web Architecture

The extensible markup language allows for proper syntax to model document contents. The Resource Description Framework (RDF) [4], gives a graph based model to describe objects and their relationships. The RDF Schema [5] provides support for a vocabulary and axioms for describing properties and classes of the RDF-based resources. OWL [6] adds additional vocabulary support for describing properties and classes in order to construct an ontology. Traditional web technologies primarily focus on representing data, whereas

the Semantic Web allows for a platform that is understood by machines, for the efficient extraction of information about web resources and relationships between heterogeneous resources. The Semantic Web thus provides a platform for applications to understand the meaning of information on the web and to correlate them.

III. RELATED WORK

A lot of research has been done to provide OLAP functionality on heterogeneous and unstructured data that is linked using RDF. Many of these works had limitations of storing the extracted Semantic Web data into a local OLAP data warehouse cube. As these approaches were semi-automatic, they were not able to adapt to the dynamic and changing nature of the of the web for real time analysis as typical Online Analytical Processing normally requires very heavy ETL work. These prompted researchers to carry out research by providing OLAP analysis without ETL, directly on the Semantic Web through frameworks based on the RDF Data Cube Vocabulary (QB) [7]. Most of the modern day vocabularies are variants of the RDF QB Vocabulary. (Fig 2.).

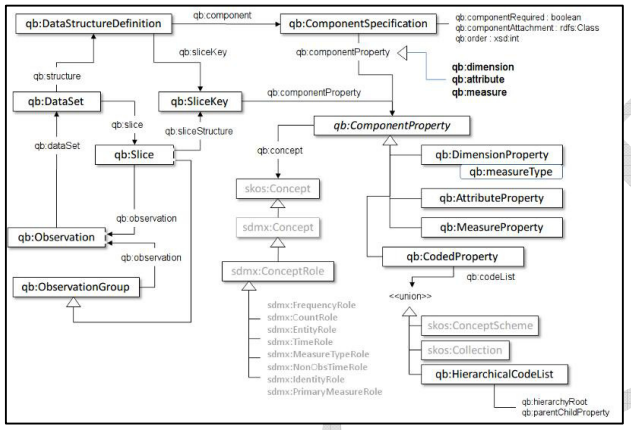

Fig 2: A pictorial representation of the QB Vocabulary

In 2011, Kämpgen et al [8] extended the QB model to represent statistical data in a multidimensional format. They presented a model to map statistical linked data that conformed to the QB vocabulary. In 2012, they demonstrated through the use of SPARQL [11] on how to generate facts from the cube through nested OLAP operations.

In 2012, Etcheverry et al. [9] introduced the Open Cubes Vocabulary (OC), a multidimensional language that supported multiple dimension hierarchies. However, the major limitation of the OC was that it was not compatible with previous QB applications. They overcame this issue by later proposing a new vocabulary [10] that extended QB to fully support OLAP models and operators. This new vocabulary was called QB4OLAP. They provided algorithms to transform cubes based on QB into an equivalent QB4OLAP cube that allowed for multidimensionalmodelling of Semantic Web data.

In 2014, they then presented an extension to the QB4OLAP vocabulary [11] to support aggregate table and a framework to translate an existing relational warehouse into its equivalent QB4OLAP schema (Fig 3.). This also demonstrated the use of SPARQL to query the QB4OLAP schema. Using SPARQL on a RDF cube modelled using the QB4OLAP vocabulary allowed for performing analytical operations such as Roll-Up and Slice on the multidimensional data.

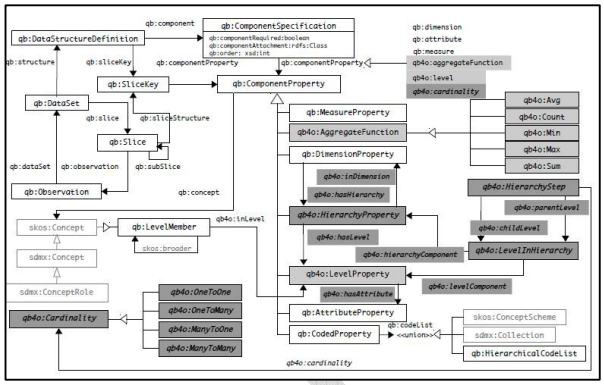

Fig 3: A pictorial representation of the QB4OLAP Vocabulary

IV. BACKGROUND

This paper aims to present a conceptual process that enables publishing of open raw data, that can then be consumed into multidimensional data cubes, and finally linked to various heterogeneous source to provide analytics on the linked open data. This will allow for a more streamlined effort to analyze existing tools and technologies to promote linked open data. A number of freely available statistical datasets have been made available which can be used as an input set for the framework. Some of these are from the World Bank, UNESCO and the European Central Bank. Combining such open multidimensional data will enrich the quality of information that can be extracted from previously thought of unrelated data sources. However, the publishing of open data doesn't necessarily provide a mechanism to link them.

The proposed framework aims at addressing this shortcoming. The framework comprises of the following steps:

(i) Publish a Data Cube from raw data
(ii) ink Other Published Open Data Cubes
(iii)Performing Analytics using QB4OLAP.

V. PROPOSED FRAMEWORK

A. Publish a Data Cube from raw data

The first step involves the use of The RDF data cube (QB) vocabulary which defines a cube through a class in the vocabulary qb: Dataset. A cube has a structure definition as qb: Data Structure Definition. This defines the structure of the cube and multiple levels qb:Observation that describe each cell of the cube. The structure is specified by the abstract property qb: ComponentProperty class, which has three sub-classes, namely qb: DimensionProperty, "qb: MeasureProperty", and qb: AttributeProperty. The first class defines the dimensions of the cube, the second class the measures or facts, while the third provides metadata.This metadata will then need to be uploaded to a common repository which can then be used locate other similar web cubes. Tools like Open Refine [18] and Fuseki server [19] can be used to assist in this process.

B. Linking Other Published Open Data Cubes

The second step is identifying compatible cubes in order to join them to facilitate recursive linking. The list of other Open Cubes needs to be maintained along with metadata information

like dimensions, measures and granularity of data. This can be done by using a common repository for the metadata of the cube. Data integration in the Web of Linked Data is facilitated by establishing an owl:same relationship. The URI references can be exploited to test of similarity as the can potentially refer to the same thing, however, additional tests are required in this area. A conceptual model of the cube data can be created using which the cubes can be mapped. The qb:DimensionProperty can be used for this purpose which specifies what observation the cubes apply to and what fact they would like aggregate on. VoID [12] can be used to create the metadata about the RDF datasets.

C. Performing Analysis by exploiting QB4OLAP

This step uses the data cubes that resulted from the previous steps in order to perform analytics using OLAP operations like Roll-Up, Slice and Dice. A parser can be built that collects information from the required web cubes and stores it in a local repository for analysis. The cubes can also be translated into QB4OLAP which allows representation of dimension levels, level members and rollup relations. It can also allow aggregate functions to work on the cubes. Both SPARQL and VoID can be used to express the multidimensional data. The SPARQL query will need to be used using the CONSTRUCT clause to create triples. These triples specify the dimension attributes and can be copied to the QB4OLAP structure. A graphical interface can then be used to show the relationships and statistical information of the linked web cubes.

VI. CURRENT LIMITATIONS

Many of the existing tools and technologies do not support the exploitation of data cubes and the ability to link them easily. Also, the tools to perform analytics on these cubes are restricted by the limitations of the vocabularies. Pre-calculated cubes that contained aggregated information are not feasible to maintain due to the dynamic nature of the web. This functionality is crucial to provide quick OLAP browsing. Tools that can identify similar cubes can again significantly reduce effort required to link multidimensional data. Complex SPARQL queries need to be written to provide roll up and drill down.

VII. CONCLUSION

As governments and organizations release large volumes of Open Data conforming to RDF standards, there is any urgent need to standardize on a methodology in order to be able to link the disparate data sources using web cubes. However, as the cube vocabulary is still evolving and tools are restricted by the vocabulary they support, the Semantic Web requires a unifying framework in the release and linking of open multidimensional data. Linked Data has the potential to unravel a wealth of information. A common understanding of the entire process is required while publishing as well as exploiting the published cubes. A conceptual framework was introduced that looked at defining the structure of the base cube and the linking them using commonalities. In order to fully exploit the information there needs to be a coordinated effort in the development of tools that can provide OLAP like operations on large volumes of information efficiently. We hope that his paper will contribute to development and evolution of adopting a standardized methodology to harness the full potential of linked multidimensional data on the Semantic Web.

ACKNOWLEDGEMENT

The authors gratefully acknowledge the contributions of the Director JSCOE, Principal JSCOE and departments of Computer and Information Technology JSCOE, Hadapsar for their help and guidance.

REFERENCES

[1] T. Berners-Lee, J, Hendler, O. Lassila, “The Semantic Web. Scientific American” 284(5), pp. 35–43, 2001.
[2] http://www.w3.org/standards/semanticweb/data.
[3] Kalampokis, E., Tambouris, E., Tarabanis, K.: A Classification Scheme for Open Government Data: Towards Linking Decentralized Data. International Journal of Web Engineering and Technology, 2011.
[4] K. Graham, C. Jeremy, “Resource Description Framework (RDF): Concepts and Abstract Syntax”, 2004.
[5] Brickley Dan, Guha RV. RDF Vocabulary Description Language 1.0: RDF Schema, 2004
[6] D. Mike, G. Schreiber, “OWL Web Ontology Language Reference”, 2004.
[7] http://www.w3.org/TR/vocab-data-cube.
[8] Kämpgen, Benedikt, Sean O’Riain, et Andreas Harth. “Interacting with statistical linked data via olap operations”, 2012.
[9] Etcheverry, Lorena, et Alejandro A. Vaisman “Enhancing OLAP Analysis with Web Cubes” Springer, 2012.
[10] Etcheverry, Lorena, et Alejandro Vaisman. QB4OLAP: A New Vocabulary for OLAP Cubes on the Semantic Web. Springer, 2012.
[11] Etcheverry, Lorena, Alejandro Vaisman, et Esteban Zimányi.
[12] http://www.w3.org/TR/void/
[13] Alarcón, R., Wilde, E.: From restful services to rdf: Connecting the web and the Semantic Web. California, 2010.
[14] Sebastian Speiser, Andreas Harth, “Integrating Linked Data and Services with Linked Data Services”. Proceedings of the 8th Extended Semantic Web Conference, Springer.
[15] Steffen Stadtmüller, Andreas Harth. “Towards Data-driven Programming for RESTful Linked Data”. Proceedings of the ISWC 2012 workshop on Programming the Semantic Web.
[16] Romero, Oscar, et Alberto Abelló, “Automating multidimensional design from ontologies. ACM Press, p. 1‑8, 2007.
[17] Nebot, Victoria, Rafael Berlanga, Juan Manuel Pérez, María José Aramburu, et Torben Bach Pedersen. “Multidimensional Integrated Ontologies: A Framework for Designing Semantic Data Warehouses”. Springer Berlin Heidelberg, p. 1‑36, 2009.
[18] http://openrefine.org/
[19] https://jena.apache.org/documentation/serving_data/
“Modeling and Querying Data Warehouses on the Semantic Web Using QB4OLAP“. Springer, 2014