Journal of Information & Communication Technology
Vol. 9, No. 2, (Fall 2015) 95-103

ISSUES AND CHALLENGES WITH SEMANTIC APPROACH FOR USE OF WEB-BASED DATA

Dr. Syed Saif Ur Rahman[1]
Syed Muhammad Adeel Ibrahim[2]

Department of Computer Science, SZABIST

ABSTRACT

Transition of web from syntactic to semantic technologies will create more unique data management and usage requirements. Emergence of many different data management and usage techniques, such as No-SQL databases is the sign that we will need different techniques or solutions to fulfill these requirements. An insight into the features of on-coming web technologies can help us in better understanding of these requirements and thus will be enable us to propose better solutions on-time. Further in this paper we discuss the standardize to all NoSQL databases and need to heavy consideration for the development of the system, and there are certain consideration that still to be made in the order to optimize and to develop fault tolerant Linked Data Storage system so hence it is the most important part of the whole solution. In this specimen we have to identified challenges and issues of semantic approach for use of web-based data.

Key Words: New Application Domains, Web of Data, Semantic web, Linked Data, NoSQL, Unstructured Data

INSPEC Classification : A9555L, A9630, B5270

* The material presented by the author does not necessarily portray the viewpoint of the editors and the management of the Institute of Business & Technology (IBT)

1 Dr. Syed Saif Ur Rahman : saif.rahman@szabist.edu.pk
2 Syed Muhammad Adeel Ibrahim : smadeelibrahim@yahoo.com

1. INTRODUCTION

In upcoming years, there will be a revolution of new application domains that includes social network and web of data. Web of data is considered to be a part of "semantic web" [1], or sometimes it is considered to be a "semantic web" [2]. Web of data is an approach to give a machine an ability to understand the data on web and to categories it, and to form links and relation with similar type of object and entities associated with it [3]. Web of data is an effort to create rich web application that we can understand, predict, behave and act to facilitate users. To achieve this, Web of data requires a large dataset with ability to link data with similarities [4] which requires Fast and Scalable databases/ data management System. Author would like to highlight the Issues that will be faced by databases to accommodate the changing data management needs that will emerge due to transition of syntactic web to semantic web. Since there is a rapid growth of web users and technology such as cloud computing is also accompanied creating more opportunities for medium and small entrepreneurs to extend their productive growth by implementing their businesses online. Simultaneously, new application domain such as web of data became a center of attention, this is important that whenever we talk we also need to consider it, this is another effort of creating web the way they were not, extending possibilities and approaching an artificial intelligence. This change is widely known as "Semantic Web" or "Web 3.0", the technologies of next generation. Since in recent trend NoSQL gaining large popularity in the field of data storage and it considers to be the fast, scalable and optimized solution with promising future in cloud in comparison of legacy Relation Database Management System [5] but this change took place in an uncontrolled fashion, which was not properly planned for upcoming technologies such as web of data and social networks. Authors has studied these changes and provided a relevant solution.

This research is divided into five section, In first section author have discussed the motivation behind the research in which he pointed out the existing database system model with contrast of data types (i-e. structured, unstructured and semi structured) and work load (OLTP and OLAP) and through these system model author highlighted the need of data solution that can operate on semantic web parallel to workload and different type of data favoring both [6]. In second section author discussed related concept in which he defined technical terms and NoSQL types that mainly use for storing unstructured data. In third section author have discussed related researches that relevant to this research and differentiated his research from others. In forth section author have provided statistics of the survey that he carried out and in last author have suggested a solution by comparison analysis and survey statistics [7].

2. MOTIVATION

There are certain approaches that data management system uses now days few of them are listed below

2.1 Transactional System

It is a simple model that use for normal transaction purposes, which is purely based on structure, operational data mainly on for On Line Transaction Processing (OLTP) Systems that emphasis on high query performance and Atomicity, Consistent, Isolated

Durable (ACID) Transaction [8].

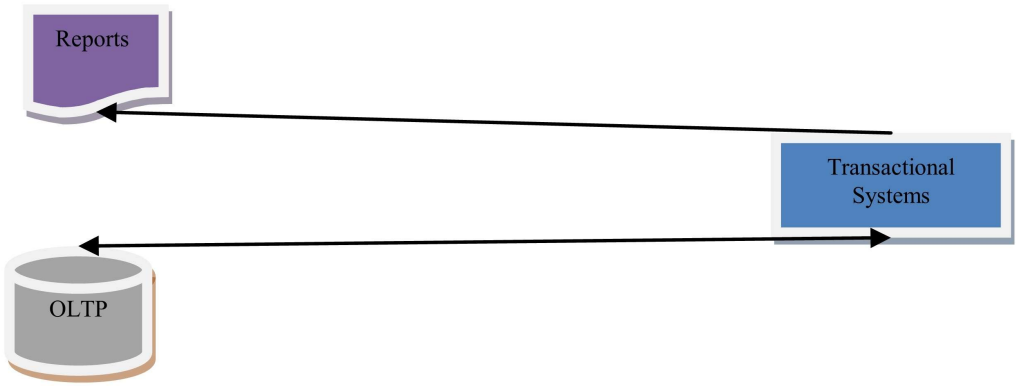

Figure 1: Durable (ACID) Transaction

Issues that mainly with Transaction Systems is that strongly schema bind and In generating report requires a good knowledge of query language and increasing complexity. Secondly increasing the performance requires scaling up and if storage gets high more computing power and network connectivity is required.

2.2 Analytical System

As reporting requirements become complex therefore new Analytical system was designed that store information in Cubes and typically focuses on the Business Intelligence, for reporting purposes this system was mainly based for On Line Analytical Processing (OLAP).

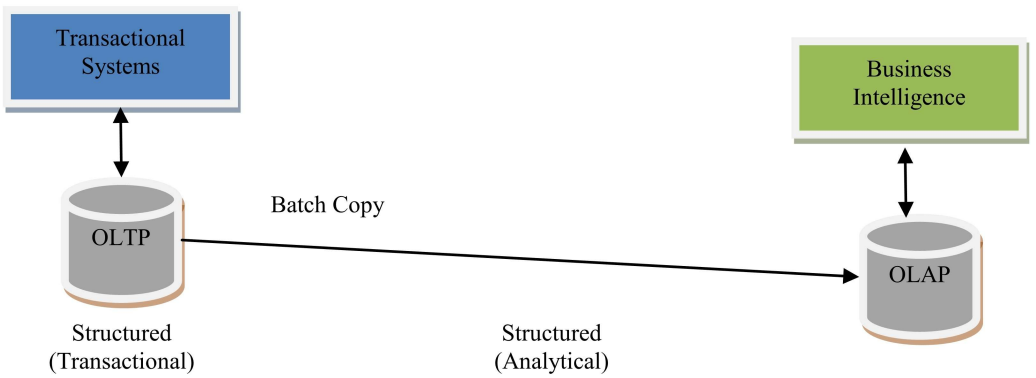

Figure 2: On Line Analytical Processing (OLAP)

Down side of the system is that it only operates on historical data that is not currently operational.

2.3 Combined System

This approach is very common now days and it is a combination of both Analytical and Transactional system in which Real time implementation is being done. However this approach is found to be productive but this can only be implemented on the structured data, while unstructured data is still left unplanned.

Figure3: Structured Data Implementation

3.4 Modern System

In modern systems the smart people are using NoSQL databases for the storage of the unstructured data along with Transactional and analytical System.

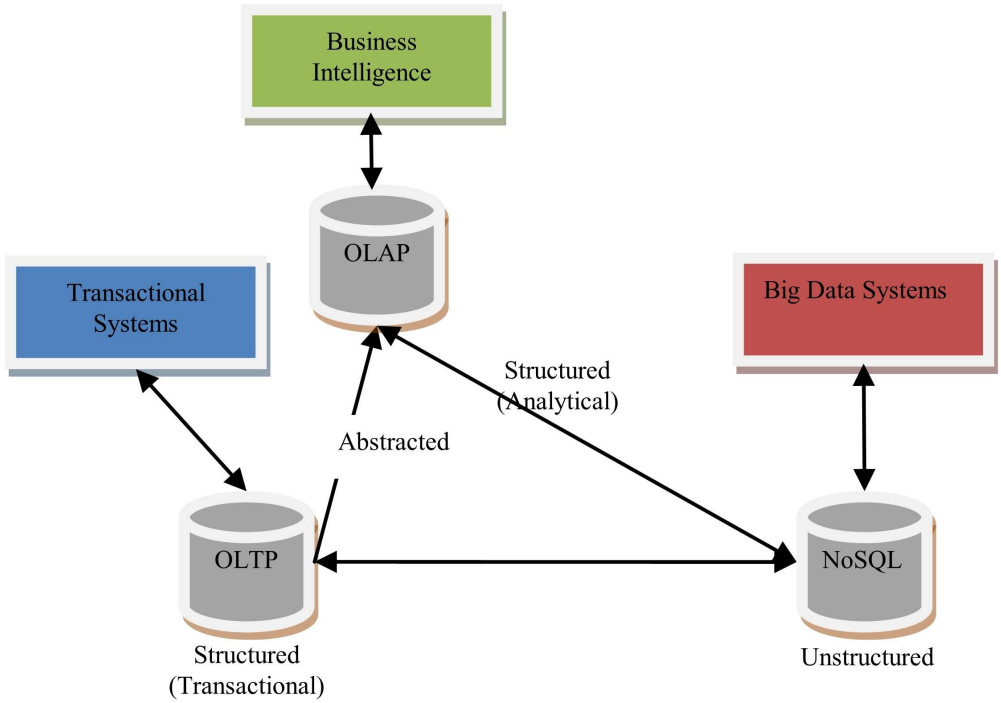

Figure 4: Transactional and analytical System

This solution is found to be useful for the systems that are presently available where the data growth is high and the size of the databases reaches to Pita Bytes and Zeta Bytes[9][10].

But there is certainly in-need of the good plan that can operate even 10 years from now onwards, when there will be new application domain such as Web of Data and Social Network. The motive behind this research was to propose a database that most suitable for that time and fully complied with Web 3.0 standards.

3. LITERARATURE REVIEW

In above mention research authors (Mark Wilson and Ian Mitchell) have elaborated different types of data, various methods storage model available now days and suggested a big data storage model that can act as an umbrella for structured, unstructured and semi-structured data storage system. In their research they have suggested a model that based on linked data. The considerations for implementing this model were clearly mentioned in the paper. Those are data integrity, integration, data management, data replication, data quality, data storage, data migration, data security and access control Through above mention research author have extracted the system model which was best fit for the solution that he have provided[10].

In above mention research authors (Michael Hausenblas and Marcel Karnstedt) have differentiated Relational Database Management System from Linked Open Data in terms of Web Data, and provided a list of applicable relational Database rules on LOD.

Above mention research is being used for supportive reasoning of the solution for identifying issues .Automatic Generation of Domain Specific Term. In this section we elaborate (the method for classifying documents) the process of automatic generation of domain specific keywords. Firstly I have built the dictionary of domain specific terms through parsing the document and generated the frequency of term related to domain through TF-IDF method specific to domain secondly we generated the frequency of term of document and then compare the term of documents with the dictionary built with the previously generated domain specific terms. The higher the terms related to the domain. That domain must relate to that domain [11][2].

4. EXPERIMENTAL SETUP

For experimental setup author have surveyed through questionnaires and on the behalf of it author have constructively formed analysis and major requirements of the system. This survey specifically taken from software development back ground interviewers; majority of them are Software developers. The sample size of the survey is 56, Since this interview is being taken in close environment on Internet therefore it only consist of people with computer back grounds but in very dispersed form, majority of interviewers are from different organization, universities within country and outside the author's country. Since the questionnaire covers a very unique domain of research that majority of people are not aware therefore author have also defined terms and explained various things within questions details.

5. METHODOLOGY

By reference to the system model of "Linked data connecting and exploiting big data by Mark Wilson, Ian Mitchell" author came to a conclusion of defining Linked Data storage for the given system. The proposed system of the given paper clearly states about the use of linked data for high scalability and performance. This solution is considered to be best as it emphasis on the structured and unstructured, and big data storage concerns for the both types of data. Furthermore this system emphasis on linking data so that existing schema of OLAP and OLTP could be maintain; to use linked data to maintain the relation between databases and to scale out multiple, small and compute nodes.

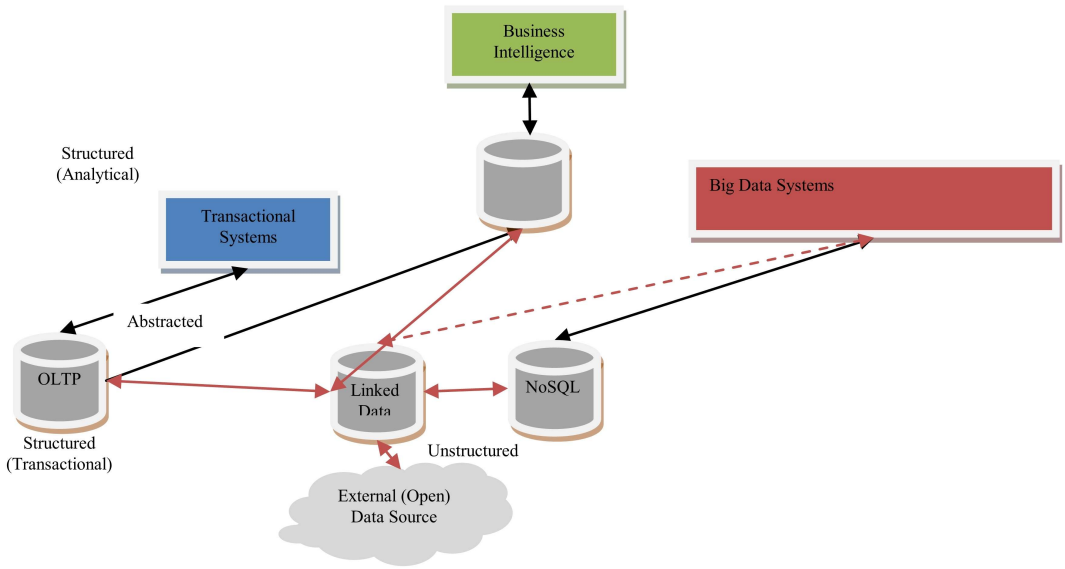

Figure 5: the Relation between databases and to scale out multiple

Since author's research focuses on the semantic aspect of the data storage therefore author have concerned architecture of Virtuoso server that enables storage of linked data through multiple source which includes various type of RDBMS, web services and existing web content RDF and non RDF. This System provides clear separation of Open and Closed Linked Data with linked data services for clients. This system can be considered to accommodate in above mention solution in above figure. But there are certain barriers, that i would like to address, In virtuoso there are no such component that is being used to translate specialize OLAP cubes for linked data, and also an absence of support for NoSQL causes unstructured data translation for Linked data impossible. Therefore author has suggested a system that fairly based on Virtuoso architecture but have some features favoring unstructured data and OLAP. In below figure author have reestablished an architecture that based on Virtuoso Server, but the vital deference between them is that, proposed architecture supports OLTP, OLAP along with NoSQL for Big data linking.

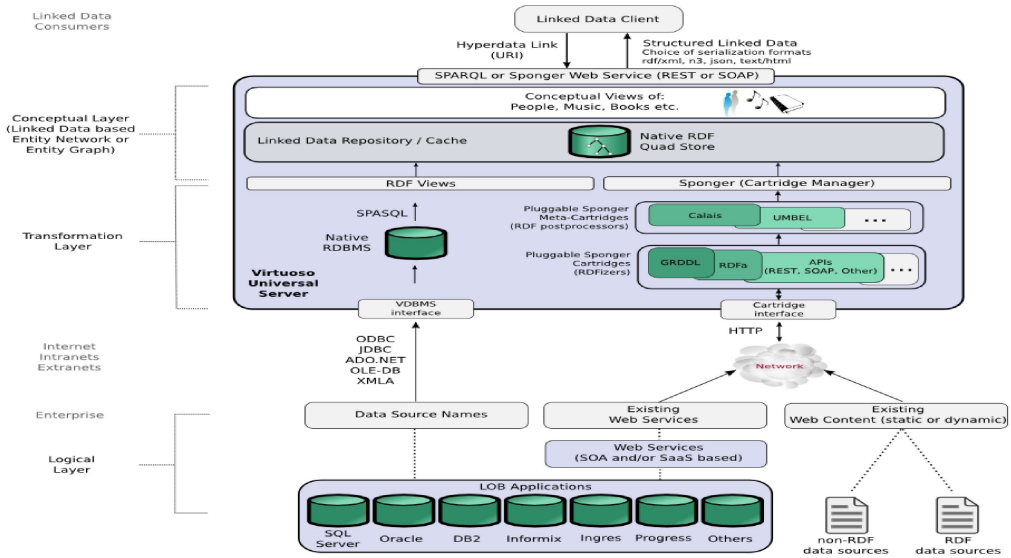

Figure 6: NoSQL for Big data linking

Top of the architecture there is linked data client this can be any client that uses open Linked data service. Moving below from client it has services such as SPARQL or Sponger Web service that are used for retrieving Linked Data, this service open for the Linked data clients for utilization. Below this layer it has views that separate open Linked data from Closed Linked Data and provide security to Quad Store that founds just below this layer. Quad Stores is the most important part of the solution that stores RDF in form of quads, when we talk about quads it means tuples found in form of relational graphs, therefore the database that is considered here is Graph.

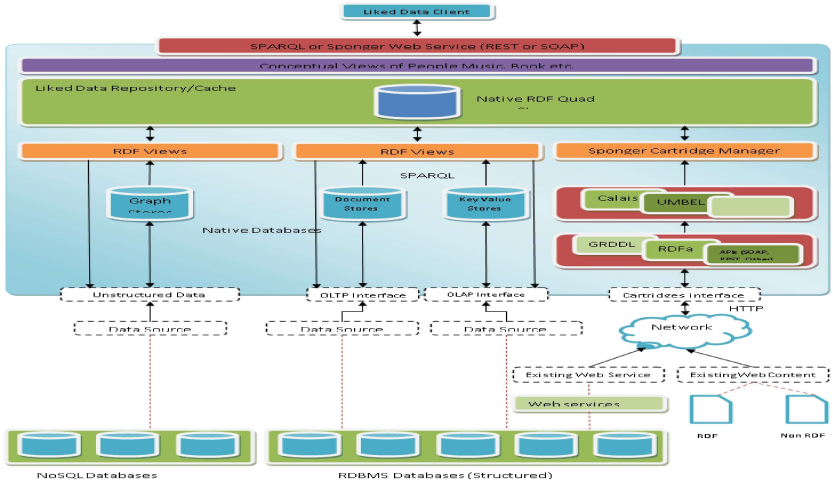

Figure 7: RDF in appearance of quads

Moving below from quad stores there are three views, two of them are of RDF and the other one is for sponger cartridge. In Sponger cartridge are the external service data that comes from HTTP Internet cloud in form of existing web service or existing web content (including RDF and Non-RDF). However this part of architecture is repetition of Virtuoso Server Architecture.
One of the views of RDF is solely facilitating unstructured data, while other is facilitating structured data (OLTP and OLAP Systems). To form a RDF views author have selected various NoSQL approaches for high performance and scalable. For Linked Data formation of unstructured data, author have selected Native Graph, there are certain modes on which graph database can be operated as key value store, document stores, RDBMS stores. Due to flexible architecture of Graph stores author have considered it for native unstructured data. However it is unidirectional connected to RDF stores while unstructured data interface can directly call RDF views from Linked data but cannot insert tuples directly, this is to maintain security with high data retrieval speed.

Similarly structured data also planned this way but differentiated by workload mechanism which causes two additional interface for OLTP and OLAP each and two native stores, document and key value. Since this solution is designed for Linked data storage therefore author has no interest in considering it for storing OLAP and OLTP data for processing purpose on it therefore NoSQL storage has been used to store data. Since OLTP is dynamic and continuously changing process therefore author have used Document stores as native storage which are known for high performance mechanism, which on the other hand OLAP is slowly changing or never changing process therefore author considered Key Value stores as native storage which is high scalable and flexible to use with RDF.

CONCLUSION
In this specimen author has identified challenges and issues of semantic approach for use of web-based data, and proposed a solution in form of system architecture based on survey and literature search. However author has proposed architecture based on his research, but there are more considerations that would be made before implementing the solution. Furthermore the big challenges that need to be sort out is native unstructured data storage, it is still not standardize to all NoSQL databases and may need heavy consideration for the development of the system, and there are certain consideration that still to be made in order to optimize and to develop fault tolerant Linked Data Storage system, hence it is the most important part of the whole solution.

ACKNOWLEDGMENT
First of all with a profound gratitude, we are thankful to Almighty Allah forgiving us success, knowledge and understanding without which we would not been capable of completing this research paper.
We are also profoundly grateful to all our family members whose endurance and understanding have played a significant role in our success by sacrificing the important family time and supporting us all over the research work.
We are finally thankful to the editor, reviewers and IBT specially who provided us with the opportunity to publish our research paper in this esteemed journal.

REFERENCES

[1] J-S. Brunner, L. Ma, C. Wang, L. Zhang, Y. Pan, K. Srinivas, "Explorations in the use of Semantic Web Technologies for Product Information Management", In Proc. of WWW, (2007), pp. 747 - 756.

[2] C. Murray, N. Alexander," Oracle Spatial Resource Description Framework (RDF)", 10g Release 2 (10.2), (2005),pp.44.

[3] J. Melton. SQL, "XQuery, and SPARQL: Making the Picture Prettier", IEEE conference, (2006),pp.52-55.

[4] L. Ma, L. Zhang, J-S. Brunner, C. Wang, Y. Pan, Y. Yu. ," SOR: A Practical System for OWL Ontology Storage", Reasoning and Search In Proc. of VLDB,(2007),pp.2-5.

[5] A. Ranganathan, Z. Liu, "Information Retrieval from Relational Databases using Semantic Queries", In Proc. of ACM ,(2006), pp. 820 – 821.

[6] J. Dolby, A. Fokoue, A. Kalyanpur, A. Kershenbaum, L. Ma, E. Schonberg, K. Srinivas. Scalable Semantic Retrieval Through Summarization and Refinement, In Proc. of AAAI, (2007), pp. 299 - 304.

[7] S. Harris and N. Gibbins, "Store: Efficient Bulk RDF Storage," Proc. 1st International Workshop on Practical and Scalable Semantic Systems, (2003) pp.1-15

[8] E. Sirin, B. Parsia and J. Hendler, "Filtering and Selecting Semantic Web Services with Interactive Composition Techniques," IEEE Intelligent Systems, vol.19, no.4, (2004), pp.42-49,

[9] S. Wong, V. Tan, W. Fang, S. Miles and L. Moreau, "Grimoires: Grid registry with metadata oriented interface: Robustness, efficiency, security," IEEE Distributed Systems Online, vol.6, no.10,(2005),pp.25.

[10] V. Haarslev and R. Möller, "Racer: A Core Inference Engine for the Semantic Web," Proc. 2nd International Workshop on Evaluation of Ontology-based Tools, (2003), pp. 27-36.

[11] I. Foster and C. Kesselman, "The Grid: Blueprint for a New Computing Infrastructure," second ed., Morgan Kaufmann, (2004), pp.42-46.

[12] I. Sheth, C. Ramakrishnan and C. Thomas, "Semantics for the Semantic Web: The Implicit, the Formal and the Powerful," J. Semantic Web and Information Systems (IJSWIS), vol.1, no.1, (2005), pp.118-120.