Leveraging Semantic Web Technologies for Enterprises dealing with Big Data

Muqeem Ahmed
Assistant Prof. Department of CS
MANUU, Hyderabad, India
muqeem.ahmed@gmail.com

***Abstract*— In today's business context it is found that conventional databases are not sufficient in dealing with the data being generated for mostly two reasons one is the volume and the second is the multiple sources as well as its heterogeneous nature. Volume is a concern as size of data grows it implies more time to process it and it is more expensive to manage as well in traditional databases. Similarly the new forms of unstructured data, a source of big data, which enterprises are finding to be useful, cannot be stored or processed using existing models of RDBMs. A vast volume of data generated may not naturally fit into traditional storage systems because it is unstructured and will not fit into the conventional storage systems. Largely enterprises are realizing the importance of advances in the area of Big data. Some of the issues enterprises are facing are how to analyze these and what mathematical models to apply. Here is the where semantic web technologies coupled with Big data techniques can give powerful solutions to these problems. Big data systems do not exist in isolation and there will be need to find a path to link them with the conventional systems. There is need for mechanisms that allow seamless information flow between Big data systems and conventional systems. Semantic Web Technologies have matured and may prove essential in representing the unstructured data in a form where Big data processing can be applied.This paper is an attempt to discuss the big data challenge and opportunity, unstructured data and semantic technologies. The aim is to explore these in the context of enterprise applications and how enterprises may gain from these technologies.**

Keywords— Big data, semantic web, unstructured data

I. INTRODUCTION

The promise of data-driven decision-making is now being recognized broadly, and there is growing enthusiasm for the notion of ``Big Data.'' While the promise of Big Data is real -- for example, it is estimated that Google alone contributed 54 billion dollars to the US economy in 2009 -- there is currently a wide gap between its potential and its realization. Heterogeneity, scale, timeliness, complexity, and privacy problems with Big Data impede progress at all phases of the pipeline that can create value from data. The problems start right away during data acquisition, when the data tsunami requires us to make decisions, currently in an ad hoc manner, about what data to keep and what to discard, and how to store what we keep reliably with the right metadata. Much data today is not natively in structured format; for example, tweets and blogs are weakly structured pieces of text, while images and video are structured for storage and display, but not for semantic content and search: transforming such content into a structured format for later analysis is a major challenge. The value of data explodes when it can be linked with other data, thus data integration is a major creator of value. Since most data is directly generated in digital format today, we have the opportunity and the challenge both to influence the creation to facilitate later linkage and to automatically link previously created data. Data analysis, organization, retrieval, and modeling are other foundational challenges. Data analysis is a clear bottleneck in many applications, both due to lack of scalability of the underlying algorithms and due to the complexity of the data that needs to be analyzed. Finally, presentation of the results and its interpretation by non-technical domain experts is crucial to extracting actionable knowledge. Big data is an evolving term that describes any voluminous amount of structured, semi-structured and unstructured data that has the potential to be mined for information. Big data can be characterized by 3Vs: the extreme volume of data, the wide variety of types of data and the velocity at which the data must be must processed. Although big data doesn't refer to any specific quantity, the term is often used when speaking about peta bytes and Exabyte of data, much of which cannot be integrated easily.

II. CHALLENGE AND OPPORTUNITIES OF BIG DATA:

We are awash in a flood of data today. In a broad range of application areas, data is being collected at unprecedented scale.

Decisions that previously were based on guesswork, or on painstakingly constructed models of reality, can now be made based on the data itself. Such Big Data analysis now drives nearly every aspect of our modern society, including mobile services, retail, manufacturing, financial services, life sciences,and physical sciences. Scientific research has been revolutionized by Big Data [CCC2011a]. The Sloan Digital Sky Survey [SDSS2008] has today become a central resource for astronomers the world over. The field of Astronomy is being transformed from one where taking pictures of the sky was a large part of an astronomer's job to one where the pictures are all in a database already and the astronomer's task is to find interesting objects and phenomena in the database. In the biological sciences, there is now a well-established tradition of depositing scientific data into a public repository, and also of creating public databases for use by other scientists. In fact, there is an entire discipline of bioinformatics that is largely devoted to the curation and analysis of such data. As technology advances, particularly with the advent of Next Generation Sequencing, the size and number of experimental data sets available is increasing exponentially.

Big Data has the potential to revolutionize not just research, but also education [CCC2011b]. A recent detailed quantitative comparison of different approaches taken by 35 charter schools in NYC has found that one of the top five policies correlated with measurable academic effectiveness was the use of data to guide instruction [DF2011]. Imagine a world in which we have access to a huge database where we collect every detailed measure of every student's academic performance. This data could be used to design the most effective approaches to education, starting from reading, writing, and math, to advanced, college-level, courses. We are far from having access to such data, but there are powerful trends in this direction. In particular, there is a strong trend for massive Web deployment of educational activities, and this will generate an increasingly large amount of detailed data about students' performance. It is widely believed that the use of information technology can reduce the cost of healthcare while improving its quality [CCC2011c], by making care more preventive and personalized and basing it on more extensive (home-based) continuous monitoring. McKinsey estimates [McK2011] a savings of 300 billion dollars every year in the US alone.In a similar vein, there have been persuasive cases made for the value of Big Data for urban planning (through fusion of high-fidelity geographical data), intelligent transportation (through analysis and visualization of live and detailed road network data), environmental modeling (through sensor networks ubiquitously collecting data) [CCC2011d], energy saving (through unveiling patterns of use), smart materials (through the new materials genome initiative [MGI2011]), computational social sciences(a new methodology fast growing in popularity because of the dramatically lowered cost of obtaining data) [LP+2009], financial systemic risk analysis (through integrated analysis of a web of contracts to find dependencies between financial entities) [FJ+2011], homeland security (through analysis of social networks and financial transactions of possible terrorists), computer security (through analysis of logged information and other events, known as Security Information and Event Management , and so on. In 2010, enterprises and users stored more than 13 exabytes of new data; this is over 50,000 times the data in the Library of Congress. The potential value of global personal location data is estimated to be $700 billion to end users, and it can result in an up to 50% decrease in product development and assembly costs, according to a recent McKinsey report [McK2011]. McKinsey predicts an equally great effect of Big Data in employment, where 140,000-190,000 workers with "deep analytical" experience will be needed in the US; furthermore, 1.5 million managers will need to become data-literate. Not surprisingly, the recent PCAST report on Networking and IT R&D [PCAST2010] identified Big Data as a "research frontier" that can "accelerate progress across a broad range of priorities." Even popular news media now appreciates the value of Big Data as evidenced by coverage in the Economist [Eco2011], the New York Times [NYT2012], and National Public Radio [NPR2011a, NPR2011b]. While the potential benefits of Big Data are real and significant, and some initial successes have already been achieved (such as the Sloan Digital Sky Survey), there remain many technical challenges that must be addressed to fully realize this potential. The sheer size of the data, of course, is a major challenge, and is the one that is most easily recognized. However, there are others. Industry analysis companies like to point out that there are challenges not just in Volume, but also in Variety and Velocity [Gar2011], and that companies should not focus on just the first of these. By Variety, they usually mean heterogeneity of data types, representation, and semantic interpretation. By Velocity, they mean both the rate at which data arrive and the time in which it must be acted upon. While these three are important, this short list fails to include additional important requirements such as privacy and usability.

III. Semantic web can take big data smarter

Over the past few years, major enterprises have shown interest in combining semantic web technology with big data for added value. Let's take a look at what enterprises are seeking and why they think semantic web can make big data smarter.

IV. Key Benefits

A. Provides end-users increased ability to self-manage data from varied sources

Users need to be able to search, access, aggregate, curate, filter, visualize, analyze, collaborate and create reports. They need to combine extracted or analyzed data from big data stores with data from documents, emails, spreadsheets, the web and other databases to get further insights.

By providing self-help, IT is no longer the bottleneck to business analysis and action. However IT needs to continue to manage access, security, data lineage, back-up and other much-desired enterprise IT support and governance functions. Smart data layers and smart data solutions using unified information based

on semantic technology can address user self-help needs while providing the IT support and governance functions.

B. Addresses varying user needs and changing business environments

In traditional big data IT solutions, the data model and the IT solutions are designed to address specific business needs and to handle specific data types and data sources. As the business needs and data sources change, the IT solutions no longer work and new data marts and new solutions must be built.

Semantic-based solutions have data models that can evolve in run time. This allows the solutions to evolve with user customization requirements and changing business environments. When building a solution with semantic technology, a user can start with something quick and then evolve the solution, adding new datasets as needed, saving significant support time and expenses.

C. Manages terminology, concepts and relationships while connecting diverse data from varied data sources

Different data sources can define the same entity, concept or term differently. For example, IBM may be called Big Blue or International Business Machines. There is a need not only to have a glossary of terms and entities but also to manage the relationships between different data and meta-data so that search, data lineage and other actions can be performed. Moreover, as data leaves its application, metadata must travel with the data so that the data does not lose its meaning. Semantic technology addresses these and other data relationships and meta-data management needs. If the smart data layer is placed over big data store and other existing data stores, the smart data layer can manage relationships across all these varied sources.

D. *Industry Group Adoption*

Leading industry groups such as OMG, EDM Council, CDISC and HL7 understand that big data and semantic web technology are ideal complements and have been building industry standard data models based on semantic technology that can be used with big data. Many of these groups are working with regulatory bodies to use these standards for government compliance and risk management. These standards will drive enterprise adoption.

V. Making Big Data Smarter

As we create a semantic layer over your big data initiative, be sure to include the following elements:

1. **Flexible, universal data model based on industry standards:** Using standard industry models with a semantic platform, allows for big data solution developers to quickly create industry or company specific solutions that can be used with big data stores and where the solutions can evolve as data needs evolve.
2. **Use of semantic RDF standards to make the data "self-describing":** By using semantic RDF standards, instance data and meaning (meta-data) travel together so that both humans and machines can understand and use the data. Use of platforms or solutions built on standards also means that the solution built will be inter-operable with other technologies using the standard.
3. **Graph representation and management of data:** Big data is just a large bucket of key/value pairs, with little if any relationships between the data. By using a graph representation, big data gets contextualized with entity and relationships that can be used for search and analysis. To understand the value, look at what value Facebook's open graph provides to the Facebook social media solution.
4. **Service-Oriented Architecture (SOA) infrastructure:** A SOA infrastructure over big data and existing data stores allows in run time to bring in data into the big data store as necessary. It can also be used to extract data in run time to create sandbox data marts for combining data from varied sources for user manipulation.
5. **Post-ingestion data characterization:** Big data is all about collecting data without worrying about schemas and data descriptions but the problem is that usually the data never gets any sort of description so it stays "dumb" and of limited utility. But as you use and understand the data the Semantic layer should automatically classify the data, associate relationships and find new relationships. This is done by using OWL — the Web Ontology Language — in the semantic layer

VI. Observation and analysis

During the last 35 years, data management principles such as physical and logical independence, declarative querying and cost-based optimization have led, during the last 35 years, to a multi-billion dollar industry. More importantly, these technical advances have enabled the first round of business intelligence applications and laid the foundation for managing and analyzing Big Data today. The many novel challenges and opportunities associated with Big Data necessitate rethinking many aspects of these data management platforms, while retaining other desirable aspects. We believe that appropriate investment in Big Data will lead to a new wave of fundamental technological advances that will be embodied in the next generations of Big Data management and analysis platforms, products, and systems.

Big Data and the Semantic Web are on a track to intersect. And businesses that want to be on track to profit from the explosion in data should start looking a little more closely at that intersection, and soon.

We believe that these research problems are not only timely, but also have the potential to create huge economic value in the economy for years to come. However, they are also hard, requiring us to rethink data analysis systems in fundamental ways. A major investment in Big Data, properly directed, can result not only in major scientific advances, but also lay the foundation for the next generation of advances in science, medicine, and business.

VII. Conclusion

Big data is an evolving term that describes any voluminous amount of structured, semi-structured and unstructured data that has the potential to be mined for information. Big data can be characterized by 3Vs: the extreme volume of data, the wide variety of types of data and the velocity at which the data must be must processed. Although big data doesn't refer to any specific quantity, the term is often used when speaking about peta bytes and Exabyte of data, much of which cannot be integrated easily.

Largely enterprises are realizing the importance of advances in the area of Big data. Some of the issues enterprises are facing are how to analyze these and what mathematical models to apply. Here is the where semantic web technologies coupled with Big data techniques can give powerful solutions to these problems. Big data systems do not exist in isolation and there will be need to find a path to link them with the conventional systems. There is need for mechanisms that allow seamless information flow between Big data systems and conventional systems. Semantic Web Technologies have matured and may prove essential in representing the unstructured data in a form where Big data processing can be applied.

REFERENCES

[1] Auer S., Bizer C., Kobilarov G., Lehmann J., Cyganiak R. and Ives Z. G. (2007). Dbpedia: A nucleus for a web of open data. In ISWC/ASWC, volume 4825 of Lecture Notes in Computer Science, p. 722–735: Springer.

[2] Breslin J., Harth A., Bojars U. and Decker S. (2005). Towards Semantically-Interlinked Online Communities. In ESWC.

[3] Baget, J.; Corby, O.; Dieng-Kuntz, R.; Faron-Zucker, C.; Gandon, F.; Giboin, A.; Gutierrez, A.; Leclère, M.; Mugnier, M. and Thomopoulos (2008), R. Griwes: Generic Model and Preliminary Specifications for a Graph-Based Knowledge Representation Toolkit Proc. of the 16th International Conference on Conceptual Structures (ICCS'2008).

[4] Bojars, U., Breslin, J.G, Finn, A., Decker, S. (2008): "Using the Semantic Web for linking and reusing data across Web 2.0 communities", J. Web Sem. 6: 21-28.

[5] Corby, O., Dieng-Kuntz, R., and Faron-Zucker, C. (2004),Querying the semantic web with the corese search engine. ECAI/PAIS2004.

[6] Corby, O (2008): Web, Graphs & Semantics, Proc. of the 16th International Conference on Conceptual Structures (ICCS'2008)

[7] Erétéo, G., Buffa, M., Gandon, F., Grohan, P., Leitzelman, M., Sander, P. (2008): A State of the Art on Social Network Analysis and its Applications on a Semantic Web, SDoW2008 (Social Data on the Web), workshop at the 7th International Semantic Web Conference.

[8] Freeman, L.C.: Centrality in social Networks: Conceptual Clarification. Social Networks 1, 215-239, 1979.

[9] Gruber T. (1993). A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition, 5(2), 199–220.

[10] Gruber, T. (2005): Ontology of folksonomy: A mash-up of apples and oranges. MTSR2005.

[11] Gruber, T. (2008): Collective knowledge systems: Where the Social Web meets the Semantic Web J. Web Sem, 6, 4-13

[12] Kim, H., Yang, S., Song, S., Breslin, J. G., Kim, H. (2007): Tag Mediated Society with SCOT Ontology. ISWC2007.

[13] Hendler, J., Goldbeck, J. (2008): Metcalfe's law, web 2.0 and the Semantic Web. J. Web Sem. 6(1):14-20.

[14] Limpens F., Gandon F. and Buffa M. (2008). Bridging Ontologies and Folksonomies to Leverage Knowledge Sharing on the Social Web: a Brief Survey. In Proc. 1st International Workshop on Social Software Engineering and Applications (SoSEA), L'Aquila, Italy.

[15] Martin A.C. (2005). From high maintenance to high productivity: What managers need to know about Generation Y, Industrial and Commercial Training, Vol 37, n°1 pp 39-44

[16] Mika, P. (2005): Ontologies are us: A unified model of social networks and semantics. In Gil, Y., Motta, E., Benjamins, V. R. and Musen, M. A., (2005): Eds., The Semantic Web. Proceedings of the 4th International Semantic Web Conference, ISWC2005, volume 3729 of Lecture Notes in Computer Science, p. 522–536: Springer.

[17] Morville, P. (2004): "Ambient findability", Digital Web Magazine, http://www.digital-web.com/articles/ambient_findability/

[18] Newell, A. (1982): The Knowledge Level Artificial Intelligence, 18, 87-127

[19] PARK J. & HUNTING S. (2002). XML Topic Maps: Creating and Using Topic Maps for the Web. Addison-Wesley Professional.

[20] Passant, A., Laublet, P. (2008): Meaning Of A Tag: A Collaborative Approach to Bridge the Gap Between Tagging and Linked Data. LDOW2008.

[21] Veres C. (2006). The language of folksonomies: What tags reveal about user classification. In Natural Language Processing and Information Systems, volume 3999/2006 of Lecture Notes in Computer Science, p. 58–69, Berlin / Heidelberg: Springer.

[22] Vuorikari, R., Manouselis, N., Duval, E. (2007): Metadata for social recommendations: storing, sharing and reusing evaluations of learning resources, Social Information Retrieval Systems: Emerging Technologies and Applications for Searching the Web Effectively, (Goh, D.H. and Foo, S., eds.), Idea Group Inc., pp.87-107.

[23] Wellman, B. (1996): "For a social network analysis of computer networks: a sociological perspective on collaborative work and virtual community", Proceedings of the ACM SIGCPR/SIGMIS conference on Computer personnel research, Denver, Colorado, United States p 1 – 11.