Big Data Mining and Semantic Technologies: Challenges and Opportunities

Ms. Yesha Mehta
Dept. of Comp. Sci.
Shree Ramkrishana Institute of Computer Education & Applied Sciences,
Surat, India
e-mail: yesha_n_mehta@yahoo.in

Dr. Sanjay Buch
Reliance Info solutions Ltd.
Naroda,
Ahmedabad, India
e-mail: drshbuch@gmail.com

***Abstract*—** Big data a term coined due to the explosion in the quantity and diversity of high frequency digital data which is having a potential for valuable insights has drawn the most attention in the area of research and development. Converting big data to actionable insights requires depth understanding of big data, its characteristics, challenges and current technological trends. A rise of big data is changing the existing data storage, management, processing and analytical mechanisms and leads to the new architecture/ecosystems to handle big data applications. This paper covers finding of our research study about big data characteristic, various types of analysis associated with it and basic big data types. First, we are presenting the big data study from data mining and analysis perspective and discuss the challenges and next, we present the result of research study on meaningful use of big data in the context of semantic technologies. Moreover, we discuss various case studies related to social media analysis and recent development trends to identify potential research directions for big data with semantic technologies.

Keywords- *Big Data, Data Mining, Semantic Technologies, Social Media, Unstructured Content*

*****

I. Introduction

A company UPS has led to savings in 2011 of more than 8.4 million gallons of fuel by cutting 85 million miles off of daily routes by leveraging big data analytics. (1). As another example, on 4 October 2012, the first presidential debate between President Barrack Obama and Governor Mitt Romney triggered more than 10 million tweets within 2 hours. These tweets are analyzed to sense the public interests and generate feedback in real time.(2) An appealing Big Data mining task is to design a Big Data mining system to predict the movement of the market in the next one or two minutes.(3)

Above examples explains the potential of big data and big data analytics in various fields. Big data analytics is emerging in fields like business, advertising, telecom, science, industry, living environment and cities, social media and networks, healthcare etc. for actionable knowledge and valuable insights. The big data wave is prominence because of the growth of the digital data creation and its constant evolvement. Every day, 2.5 quintillion bytes of data are created and 90 percent of the data in the world today were produced within the past two years (5).

This raw unstructured data is having a great potential to generate insights to support future decision making and identifying trends. Big data is about to collect, store and analyze large volumes of data and extract useful knowledge for future actions With the growth of web 2.0 user generated content, social networking sites, blogs, wikis, blogs and data mash-ups came up as prospective big data resources. Massive amounts of data are available to be harvested for competitive business advantage, sound government policies, and new insights in a broad array of applications (including healthcare, biomedicine, energy, smart cities, genomics, transportation, etc.). Yet, most of this data is in accessible for users, as we need technology and tools to find, transform, analyze, and visualize data in order to make it consumable for decision-making. (6). To analyze big data, understanding about how to use data sources like social media, web content and other offline/online resources is required. Our approach is to understand various aspects of big data, comparison of existing analytics mechanisms and to find pathway for leveraging the open big data resources to transform raw data into meaningful information.

To start with big data, literature study is done for understanding big data standards and benchmarks which covers study of i) definition, ii) characteristics iii) data types iv) architecture. Big data is discussed with various definitions in terms of V's for its characteristics. As big data is a rising field, standardization is taking place gradually with big data definition, taxonomies and architecture by NIST. NIST has also focused on big data systems architecture. The big data architecture components are weaved into categories of infrastructure, storage and management, analytics and application interface. Big data received the attention of researchers due to its characteristics where the source data is in heterogeneous formats/schema, having complex structures-relationships and it is continuously evolving with capacity to generate new knowledge and insights.

This paper surveys the literature to understand big data standards and benchmarks for i) definition, ii) characteristics iii) data types iv) big data technology. Section 3 covers the study of data from big data mining perspective and its

 ISSN: 2321-8169
4907 - 4913

comparative analysis with traditional data mining. This section covers various analytic techniques used with variety of content available for mining. Section 4 presents the challenges and idea about meaningful use of big data in context of semantics and role of semantic technology in big data analysis. Various milestones of semantic technologies are studied for exploration. Section 5 presents the review of case studies related to social media for big data analytics. Section 6 presents the summary of study and identification of research direction followed by conclusion and future work.

II. BIG DATA: A NEW PARADIGM FOR NOVEL DATASETS

To start with our research, first we have attempted to study big data definitions to understand the nature of big data. According to IDC ―Big data technologies describe a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling high-velocity capture, discovery, and/or analysis‖. This definition delineates the four salient features of big data, i.e., volume, variety, velocity and value." (7). In 2011, Mckinsey's report (8) defined big data as ―datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze‖ which compares traditional data systems with big data processing aspects. The standard definition developed by NIST's big data group in 2015 is, "Big data consists of extensive data sets primarily in the characteristics of volume, variety, velocity, and/or variability that require a scalable architecture for efficient storage, manipulation, and analysis." (9). According to NIST big data revolution is due to horizontal scaling, means use of commodity hardware to act as single integrated system. With time big data definition is getting standardize, but all definitions leads to the three main big data characteristic of volume, variety and velocity. As traditional data base management systems are not compatible with these characteristics of big data; scalable, interoperable and integrated architecture is required to engineer big data. Existing enterprise middle ware, business logic, data warehouses, OLAP and data mining tools, these technologies that has been using widely and invested heavily in terms of resources. A challenge here is to create big data applications using existing infrastructure and tools with novel concepts. From the studied definition we derived needs of big data architecture according to its characteristics.

Big data definitions talks about novel datasets which are large in volume, different in storage formats and continuously generating new data. With consideration of all three aspects of volume-velocity-variety, datasets require processing speed, provision to accommodate growth of data and interoperations among the data gathered from various resources. Big data definitions focus on generating value from the available data, which needs integrated mechanisms for data storage, pre-processing, analysis and mining.

TABLE I. REQUIREMENTS OF BIG DATA ARCHITECTURE

Characteristics	Requirements
Volume	Scalability, Parallel Processing, High performance computing
Velocity	Storage, Stream Processing
Variety	Data Models, Interoperability
Value	Management, Transformation, Integration

Big data definitions itself reveals the characteristics of big data in i.e. Volume, Variety and Velocity. To manage volume of big data a distributed computing and parallel programming technologies are implemented with Hadoop and Mapreduce framework. Hadoop is becoming a big data analytics benchmark to store and process massive datasets. (7)

Variety is about different data formats like structured data (RDBMS), video, audio, image, email, PDF, unstructured text and finding association between this data items. To get insights from big data methods of data mining, text mining, sentiment analysis, machine learning, NLP and semantic technologies are targeted by the researchers and practitioners.

Big data characteristics are also given by HACE theorem [Data mining with Big Data] which represents big data sources as heterogeneous, autonomous, complex and evolving. A various big data sources are diverse in context of schema, data storage model and representation. A data expression needs to be coherent for the data coming from various data sources. Large database sources are having decentralized control and variation in their data items according to regions.

The underlying data types for big data are categorized as structured, semi-structured and unstructured data. The big data companies like Walmart, Amazon, Google, Facebook, AT & T, and SDSS have data sources of these types. Structured data concerns all data which can be stored in database SQL in table with rows and columns. They have relational key and can be easily mapped into pre-designed fields. Today, those data are the most processed in development and the simple way to manage information. Semi-structured data is information that doesn't reside in a relational database but that does have some organizational properties that make it easier to analyze. With some process we can store them in relational database. CSV, XML and JSON documents are example of semi structured documents. NoSQL databases are considered as semi structured.(10). Unstructured data represent around 80% of data. It often includes text and multimedia content. Examples include e-mail messages, word processing documents, videos, photos, audio files, presentations, webpages, tweets, tags, comments of social sites and many other kinds of business documents. Following table shows the comparison of data types and their associated source application and technology. Unstructured data dominates the big data and efficient ways to process and analyze open unstructured content is the key demand of big data frameworks.

TABLE II. BIG DATA TYPES AND ASSOCIATED SOURCES

Data Types	Sources	Formats
Structured	Business Applications	RDBMS, data warehousing, OLAP, and BPM
Semi-Structured	Web Applications	HTML, RDF, XML,CSV, Linked Open Documents, JSON
Unstructured	Network Applications	User generated Text Content (tweets, blogs, emails, news, wikis, micro blogs)

To manage all three types of data NoSQL technologies are being storage benchmarks because they are distributed, open-source, schema-free and scalable. NoSQL data models are key-value stores, column-oriented databases, and document databases. To process raw data a NoSQL based data models are preferred as they leads from diverse schema to consistent data representation. An integrated framework of big data will need support for both SQL and NoSQL based data for the mining purpose and an additional layer to convert these heterogeneous format to homogenous expression.

III. BIG DATA FROM DATA MINING AND ANALYTICS PERSPECTIVE

As big data needs are different from traditional data warehousing and mining systems, the underlying framework for data mining has changed with the aspect of computing and mining platforms. The analysis of big data involves multiple distinct phases like acquisition, extraction, processing and analysis which introduce challenges. Big data is all about heterogeneous and unstructured data which cannot be processed by traditional data warehousing and mining systems. To overcome limitations of traditional data mining and to introduce novel concepts for data processing in context of big data is necessary to realize potential of big data. Comparison between traditional data and big data in terms of volume, structure, data source, data store and access clears the distinction between two different approaches of data science. A two paradigm of big data 1) Streaming and 2) Batch processing are summarized. As batch processing paradigm is widely adopted batch-processing- based big data platforms are explained from architectural point of view. Data analytics research is classified into six critical technical areas: structured data analysis, text analytics, web analytics, multimedia analytics, network analytics, and mobile analytics. This classification is intended to highlight the data characteristics; however, a few of these areas may leverage similar underlying technologies. For different domains some common methods like data visualization, statistical analysis and data mining are useful for analysis of the data. (4)

Big data mining is divided into data, model and knowledge level. At the data level, data statistics based on the local data sources are generated, at the model or pattern level, pattern discovery and aggregation, at the knowledge level, correlations and causations are identified for the analysis and decision making. The analytics taxonomy is classified as structured, text, multimedia, network, web and mobile analytics which uses various mining approaches of statistical analysis, web mining, text mining, sentiment analysis, event detection, summarization and location based mining.

All branches of analytics has evolved such as structured, text, web and multimedia mining with their own algorithms and technique where big data needs integration of all these techniques and tools on a single platform. At present industry is using relevant technologies with plug and play approach to rapidly analyze data thus no single platform fulfill all the demands of big data mining model. A customized, scalable and integrated architecture is need of big data mining.

For mining big data at three levels processing is suggested in the research, i) To process heterogeneous data types including structured data, semi-structured data, and unstructured data where new data model is needed to incorporate all the data types. ii) Various web content and objects published on the web is having relationship, identification of such intrinsic semantic association (20) is a challenge to handle from data mining perspective, iii) There are social relationships between individuals forming complex social networks, such as big relationship data from Facebook, Twitter, LinkedIn, and other social media which is constantly changing. New techniques like structure-and-evolution, crowds-and-interaction, and information-and-communication are introduced to analyze such relationships.

Unstructured data play major role in big data and text mining techniques are used to analyze it. Limitations of text analytics in context of big data are, i) application of different tool to different content where best tool to analyze customer feedback differs from NLP tools of scientific documents. ii) Extracting

ISSN: 2321-8169
4907 - 4913

business entities from text and social media sentiment analysis techniques varies. iii) Results are unpredictable. As text analytics techniques changes from case to case, a common data processing approach is difficult to decide. To address these challenges smart data approach with semantic technology standards are adopted.

One of the ways for mining big data is to make ETL processes better and intelligent. We have studied semantic based ETL framework to address the variety characteristics of big data. In big data processing, ETL part of mining requires the major attention as the underlying data is heterogeneous, sparse and uncertain and unstructured. Researchers are proposing semantic based architectures and framework for the synchronized and optimized ETL process, limited to specific data sources. These data sources are published RDF data available on the web which consist meaning of the data. Semantic Technology based data pre-processing and mining with graph databases are new initiatives in big data mining process.

Following table relates big data mining phases with its associated architectural components and desired outcomes. For big data mining data will be pre-processed in ETL, knowledge inferences will be generated in Analytics and it will be presented in the form of visulaization.

TABLE III. BIG DATA MINING PHASES AND EXPECTED OUTCOMES

Phases	Architectural Components	Desired Outcomes
Extract, Transfer and Load	Data sources and Sinks, ETL Processing	Synchronized Data Model and Statistics
Analytics	Operational Analytics	Patterns, Co-relations, Causation
Visualizations	Application User Interface	Story,Infographics, Spatial -Temporal Visualization

IV. MEANINGFUL USE OF BIG DATA: SEMANTIC TECHNOLOGIES AS BENCHMARK

Big Data is available for various fields via social networking, sensors, machine-generated information and in the form of linked data. Yet, most of this data is inaccessible for users, as we need technology and tools to find, transform, analyze, and visualize data in order to make it consumable for decision-making (6). Susan Etlinger is an industry analyst at Altimeter Group discusses the barriers about getting insights from social data. She talks about how data meaning plays an important role to human as well as machine. [12] As big data is vast and majority unstructured to extract specific information from it becomes very difficult as there is no context or meaning available with raw data. Researchers also confirm that big data becomes usable only when it is engineered meaningfully.

When we directed our study towards data and its meaning, concepts of semantic web and semantic technologies are explored. The concept of the semantic web has been around since 1999, when Tim Berners-Lee expressed his vision of a future web in which computers could understand the context of human speech and thought, to be able to "understand" meaning of information. The term ―Semantic Web‖ refers to W3C's vision of the Web of linked data. Semantic Web technologies enable people to create data stores on the Web, build vocabularies, and write rules for handling data. Linked data are empowered by technologies such as RDF, SPARQL, OWL, and SKOS.[13] RDF is a standard model for data interchange on the Web. RDF has features that facilitate data merging even if the underlying schema differ, and it specifically supports the evolution of schema over time without requiring all the data consumers to be changed. Semantic Web technology stack includes RDF, RDFS, OWL and SPARQL to represent and manage rich and complex knowledge about things, groups of things, and relations between things. These semantic web standards have features which are compatible with existing big data needs. Semantic web features of data interchange, knowledge representation and querying helps to resolved big data issues of variety and analytics.

The most important technical challenge today in managing big data is variety (heterogeneity of data and diversity of data sources) (14). To handle heterogeneity semantic approach is preferred by researchers. To understand, relate and interpret big data, explicit meaning of the data is required which is given by harnessing semantic technologies such as RDF, OWL for knowledge and insights from big data. NIST also prefers data expressivity standards as Xml/RDF and OWL. There is a need of domain knowledge to process big data; such background knowledge is also available in the form of linked data which is around 31 billion RDF triples.(11). A need of identifying critical data points and data regions for big data analytics can be satisfied by understanding meaning of big data using semantic technologies like RDF/OWL. Data mining, Machine Language, Natural Language processing are aggressively used by industry for big data projects where gradually semantic approaches are adopted for gaining actionable insight. (14) Semantic models and technologies can be, and in many cases are being, used to address various problems associated with big data. We overcome volume by enabling abstraction to achieve semantic scalability for decision making. We defined two operations – explanation and discrimination – that underlie the semantics of machine perception, and showed how they can be implemented efficiently on resourced-constrained devices. Variety challenges can be overcome using a continuum of light-weight semantics to achieve semantic integration and interoperability. (15). These initiatives are started by research community of taking the advantage of semantic technologies in resolving big data's all four technical challenges. By studying various literatures about semantic technologies and big data we have analyzed that semantic technologies will play an important role to convert big data into big insights.

For data processing context and domain knowledge of data should be available for accurate decision making. (20) Compared to other technologies, semantic technologies provide background knowledge for data context, interoperability, scalabilty, integration and accepted as data expressivity standard.

The semantic web time line shows milestones of semantic technology which are Linked Open Data (LOD), RDB2RDF, GLD (Government Linked Data) and LDP (Linked Data Platform. Utilization of these semantic standards with big data is the new opportunity for harnessing the power of semantic knowledge bases. There are closed and open world assumption related to knowledge representation, RDF based analysis support open world assumption where explicit unknown knowledge is not considered false and new inference can be derived based on contradictory knowledge too.Linked data model is adopted by big data community where RDF is suitable data model for addressing variety, big data sets can be interlinked and data access and integration will be simplified. For our research, we have considered semantic approach as benchmark for the big data processing and analytics. There are various successful use cases which implements semantic technologies to solve big data problems. We have studied few use cases related to semantic technologies and big data.

There are some limitations using semantic technologies with ontology. For background knowledge and context some core semantic ontology exists like WorldNet, YAGO, OpenCyc, SUMO but they do not contain all possible associations and perspective. Thus their utilization will be depended on the area of particular domain and availability of relative knowledge in ontological form. By using semantic technology it is difficult to distinguish data from noise without model. It is a challenge to create a customized semantic based model to incorporate data from external big data sources and the aggregation of data statistics from local sites in a single global model. Big data includes of highly-connected entities is not easily modeled using traditional relational schemas; instead, using graph data structures makes it easy to represent connected data and to perform rapid analysis on these large datasets. Graph data models can accommodate higher levels of variety in the data. This allows more data to be mixed and queried together. Mixing in more data adds m context and more context allows for accurate insight. Graph databases have support of SPARQL and RDF and Java which makes analytics process less complicated. Following table include the details semantic based storage of big data. (43)

Following table include the details semantic based storage of big data. (43)

TABLE IV. SEMANTIC STORAGE FOR BIG DATA

Technology	Type	Support
AllegroGrap h	Graph database	SPARQL, RDFS++, and Prolog reasoning
Sesame	Java framework	SPARQL 1.1 query,RDF/XML, Turtle, N-Triples, N-Quads, JSON-LD, TriG and TriX.
Mulgara	Java based RDF store	iTQL,SPARQL,Jena, Sesame
Neo4J	Java based Graph database	Neo4j CQL, Spring Data framework

V. USE CASES OF BIG DATA, SOCIAL MEDIA AND SEMANTIC TECHNOLOGIES

Scocial Networking sites are considerd and used as potential sources of big data and analyzed in various use cases (2) to gain new insights. Such social networking website are having complex relationships between people, places and time. To discover trends by analytics such complex and evolving relationships with respect to space, time, domain parameters are the key consideration factors. We have studied different use cases where open content of such social networking media is accessed and new trends are analyzed.

Researchers from diverse fields have analyzed Twitter content to generate specific knowledge for their respective subject domains. (18) We have studied a system "Twiteris", a semantic web application for understanding perception from social media data. It follow semantic based processing of massive amount s of event-centric data for spatio-temporal analysis. "Twitris" also covers context based semantic integration of multiple Web resources and expose semantically enriched social data to the public domain. Another system "*PoliTwi*" is designed to detect emerging political topics on twitter based on concept level sentiment analysis where web ontology and semantic networks are used as knowledge base. (46)

Semantic Web technologies enable the system's integration and analysis abilities. It has applications for studying and analyzing social sensing and perception of a broad variety of events: politics and elections, social movements and uprisings, crisis and disasters, entertainment, environment, decision making and coordination, brand management, campaign effectiveness, etc. (17). We have studied another document based project Obvio(21) which is a graph-based framework for exploring biomedical literature to facilitate Literature-Based Discovery (LBD) based on rich knowledge representations. Its broader goal is to uncover hidden and complex associations between concepts in biomedical texts. Ontotext project is use case for getting insight from heterogeneous biomedical datasets for interlinking public biomedical data sources and their discovery and exploration. (42)

Simillarly, data of social media site like linkedin, facebook and youtube is available for big data analytics in various domains. In the area of healthcae huge data documents are linked with graph based framework and role of big data and semantic technologies are examined for empowring personalized medicine. Big data has the potential in Physical, Cyber and Social Systems, in multiple domain like medical, geographic, enviormental, traffic , crime, crisis and persoanl decision making. By studying different use cases, we have observed that social media is harnessed the most for big data analytics. Twitter data is used widely from stock market prediction to sentiment analysis and crisis management. Where sites such as linkedin, youtube, flicker is yet remaining to explore. These sites can be explored for the analysis of various upcoming trends.

VI. CONCLUSION AND FUTURE WORK

We have studied big data from definitions to various use cases and derived challenges as well as opportunities of big data architecture for data mining with big data, i) Big data requirements are different than traditional data mining approach, thus scalable, interoperable and integrated architecture is required for big data analytics ii) Social sites are potential but unstructured big data resources, to analyze them a process model with consideration of space, time and domain context is desired. iii) Researchers had emphasized on meaningful big data engineering to leverage its potential for accurate insights iv) Semantic technologies are projected as knowledge representation and inference standard. It also provides background knowledge of domain and play important role to interpret, analyze and convert raw data to actionable insights. Semantics technologies are explored for getting smart data from big data. We have considered it as a benchmark for social data analytics.

Big data is studied as an emerging trend and practices are moving from big data to smart data. Our future work will be in the direction of leveraging big data available on social media with semantic approach for converting big data to actionable knowledge. There are possibilities for designing big data architecture for identifying social media related trends and patterns by applying big data mining with relevant semantic technology and analytical methods which is our next research objective.

REFERENCES

[1] Thomas H. Davenport, Jill Dyché, Big Data in Big Companies, International Institute for Analytics, May 2013
[2] "Twitter Blog, Dispatch from the Denver Debate," http://blog.twitter.com/2012/10/dispatch-from-denver-debate.html, Oct. 2012
[3] J. Bughin, M. Chui, and J. Manyika, Clouds, Big Data, and SmartAssets: Ten Tech-Enabled Business Trends to Watch. McKinSeyQuarterly, 2010.
[4] Han Hu, Yonggang Wen, *Senior Member, IEEE*, Tat-Seng Chua, and Xuelong Li, Fellow, IEEE, Towards scalable system Towards Scalable Systems for Big Data Analytics: A Technology Tutorial June 2013
[5] "IBM What Is Big Data: Bring Big Data to the Enterprise," http://www-01.ibm.com/software/data/bigdata/, IBM, 2012
[6] Srividya K Bansal T"owards a Semantic Extract-Transform-Load (ETL) framework for Integration Big Data" (BigData Congress), 2014 IEEE International Congress
[7] Han Hu, Yonggang Wen,Senior Member, IEEE,Tat-Seng Chua, and Xuelong Li,Fellow, IEEE, Towards Scalable Systems for Big Data Analytics:A Technology Tutorial, IEEE Access June 2014
[8] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers, "Big data: The next frontier for innovation, competition, and productivity," *McKinsey Global Institute*, pp. 1–137, 2011
[9] NIST Big Data Public Working Group Definitions and Taxonomies Subgroup, Volume 1, Definitions Apr 2015 http://dx.doi.org/10.6028/NIST.SP.1500-1
[10] [Online] http://planetcassandra.org/what-is-nosql
[11] C. Bizer, P. Boncz, M. L. Brodie, and O. Erling, "The Meaningful Use of Big Data: Four Perspectives – Four Challenges," *SIGMOD Rec.*, vol. 40, no. 4, pp. 56–60, Jan. 2012.
[12] [Online] "Interview with Susan Etlinger: Developing a Data-Driven Culture for Higher Education Marketing" http://www.thelawlorgroup.com/interview-susan-etlinger-developing-data-driven-culture-higher-education-marketing/
[13] [Online] Semantic Web http://www.w3.org/standards/semanticweb/
[14] [Online] Data Semantics and Semantic Web - 2014 year-end reflections and prognosis by Amit Sheth http://amitsheth.blogspot.in/ Dec 2014
[15] Krishnaprasad Thirunarayan and Amit Sheth: Semantics-empowered Big Data Processing with Applications, In: AI Magazine, Vol. 36, No. 1, pp. 39-54, Spring 2015.
[16] Krishnaprasad Thirunarayan and Amit Sheth: Semantics-empowered Approaches to Big Data Processing for Physical-Cyber-Social Applications, In: Proceedings of AAAI 2013 Fall Symposium on Semantics for Big Data, Arlington, Virginia, November 15-17, 2013, 8 pages.

ISSN: 2321-8169
4907 - 4913

[17] Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan and Amit P. Sheth. Harnessing Twitter 'Big Data' for Automatic Emotion Identification. In Proceedings of International Conference on Social Computing (SocialCom), 2012.

[18] Xin Chen, Mihaela Vorvoreanu, and Krishna Madhavan Mining Social Media Data for Understanding Students' Learning Experiences IEEE TRANSACTIONS ON LEARNING TECHNOLOGIES, 2013

[19] Amit Sheth, 'Transforming Big Data into Smart Data: Deriving Value via harnessing Volume, Variety and Velocity using semantics and Semantic Web,' keynote at the 21st Italian Symposium on Advanced Database Systems, June 30 - July 03 2013

[20] Xindong Wu, Fellow, IEEE, Xingquan Zhu, Senior Member, IEEE, Gong-Qing Wu, and Wei Ding, Senior Member, IEEE Data Mining with Big Data, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING Vol 26 Jan 2014

[21] [Online] Obvio http://wiki.knoesis.org/index.php/Obvio

[22] Yuri Demchenko, Peter Membrey " Defining Architecture Components of the Big Data Ecosystem" Collaboration Technologies and Systems (CTS), 2014 International Conference IEEE

[23] Souripriya Das, Oracle, Seema Sundara, Oracle, Richard Cyganiak, DERI, National University of Ireland, Galway R2RML: RDB to RDF Mapping Language [Online]. Available: http://www.w3.org/TR/r2rml/

[24] J. Kelly, "Taming big data," 2013, [Online].Available: http://wikibon.org/blog/taming-big-data/.

[25] KRISH KRISHNAN Data Warehousing in the AGE of Big Data

[26] P. Vassiliadis, A. Simitsis, and E. Baikousi, "A Taxonomy of ETL Activities," in *Proceedings of theACM Twelfth International Workshop on DataWarehousing and OLAP*, New York, NY, USA, 2009,pp. 25–32.

[27] Unstructured Data Management, Hitachi Data System, 2013. [online] Available: http://www.hds.com/solutions/it-strategies/unstructured-data-management.html

[28] P. Hayes and B. McBride, "Resource descriptionframework (RDF)," 2004. [Online]. Available:http://www.w3.org/TR/rdf-mt/.

[29] Jure LeskovecStanford Univ.Anand RajaramanMilliway LabsJeffrey D. Ullman Stanford Univ. Mining of Massive Datasets

[30] "SPARQL Query Language for RDF." [Online].Available: http://www.w3.org/TR/rdf-sparql-query/.[Accessed: 28-Feb-2014].

[31] "Using semantic enrichment to enhance big data solutions" [Online] Available: http://www-01.ibm.com/software/ebusiness/jstart/semantic/

[32] Ioannis Kopanas, Nikolaos M. Avouris, and Sophia Daskalaki "The Role of Domain Knowledge in a Large Scale DataMining Project" 2002

[33] David Wang, MIT Automated Semantic Correlation betweenMultiple Schema for Information Exchange Thesis, MIT Sloan School of Management 2000

[34] R. Cattell, "Scalable sql and nosql data stores," *SIGMOD Rec.*, vol. 39,no. 4, pp. 12–27, 2011.

[35] J. Dean and S. Ghemawat, "Mapreduce: simplified data processing on large clusters," *Commun. ACM*, vol. 51, no. 1, pp. 107–113, 2008.

[36] T. White, *Hadoop: The definitive guide*. O'Reilly Media, Inc., 2012.

[37] J. Gantz and D. Reinsel, "Extracting value from chaos," *IDC iView,pp*. 1–12, 2011.

[38] E. B. S. D. Divyakant Agrawal, Philip Bernstein and Others, "Challenges and opportunities with big data - a community white paper developed by leading researchers across the united states," The ComputingResearch Association, CRA report, Feb. 2012.

[39] C. Bizer, T. Heath, and T. Berners-Lee, "Linked data the story so far," *International journal on semantic web and information systems*, vol. 5, no. 3, pp. 1–22, 2009

[40] "SPARQL Query Language for RDF." [Online]. Available: http://www.w3.org/TR/rdf-sparql-query/.

[41] NIST Big Data Interoperability Framework: Volume 5, Architectures White Paper Survey http://dx.doi.org/10.6028/NIST.SP.1500-5

[42] [Online]Marin Dimitrov, Semantic Technologies for Big Data 2012http://www.slideshare.net/marin_dimitrov/semantic-technologies-for-big-data

[43] [Online] Srinath Srinivasa, Big Data and the Semantic Web: Challenges and Opportunities http://www.slideshare.net/srinaths/big-data-and-the-semantic-web-challenges-and-opportunities

[44] [Online] http://www.tutorialspoint.com/neo4j/

[45] [Online] http://franz.com/agraph/allegrograph/

[46] Sven Ril, Dirk Reinel, Jorg Scheidt, Roberto V. Zicari, "PoliTwi : Early Detection of emerging political topics on twitter and the impact of concept level sentiment analysis, Knowledge Based Systems ELSEVIER, May 2014