NoSQL Systems for Big Data Management

CONFERENCE PAPER · JUNE 2014

DOI: 10.1109/SERVICES.2014.42

3 AUTHORS:

Venkat N Gudivada
East Carolina University
83 PUBLICATIONS **1,504** CITATIONS
SEE PROFILE

Dhana Rao
Marshall University
25 PUBLICATIONS **556** CITATIONS
SEE PROFILE

Vijay v Raghavan
University of Louisiana at Lafayette
272 PUBLICATIONS **3,496** CITATIONS
SEE PROFILE

Data Management Issues in Big Data Applications

Venkat N. Gudivada
Weisberg Division of Computer Science
Marshall University
Huntington, WV, USA
e-mail: gudivada@marshall.edu

Subbaiyan Jothilakshmi
Department of CS and Engineering
Annamalai University
Chidambaram, TN, India
e-mail: jothi.sekar@gmail.com

Dhana Rao
Department of Biological Sciences
Marshall University
Huntington, WV, USA
e-mail: raod@marshall.edu

***Abstract*—Big Data has the potential for groundbreaking scientific discoveries, business innovation and increased productivity. It provides as many challenges as the number of new opportunities it ushers in. However, several problems need solutions before the full potential of Big Data is realized. In this paper, we provide an overview of Big Data problems from databases perspective and elaborate on security aspects. We expect that this overview will help the reader to quickly obtain a panoramic view of research challenges in Big Data and contribute to this fast evolving discipline.**

Keywords—Big Data Management; NoSQL Databases; Database Security.

I. Introduction

Wireless sensor networks, earth-orbiting satellites, social media applications, supercomputers and supercolliders, and smart phones are generating unprecedented volumes of data. In 2014, the White House commissioned a study to examine how Big Data will transform the way we live and work [1]. The report examines new business opportunities, privacy concerns, and the potential of Big Data analytics to usurp long standing civil rights of citizens. It outlines recommendations related to preserving citizens' privacy, responsible educational innovation, preventing discrimination, and judicious use in law enforcement and national security. This study attests to the role of Big Data in impacting people across the board. Big Data is a double edged sword and entails enormous implications.

It is not just the *volume* that makes this data unparalleled. Other aspects such as *velocity*, *variety*, *veracity*, and *value* bestow this data the title *Big Data*. Velocity refers to the speed at which the data is produced. For example, detecting financial fraud and real-time monitoring of cyber security requires analysis of high velocity data. Variety refers to data heterogeneity. It is often comprised of unstructured, semi-structured, and structured data of disparate types. During its life cycle, the data goes through several transformations. It is essential to be able to trace the history of such transformations to establish veracity of data. The term data provenance is used to refer to this aspect of data. Finally, collecting, cleansing, transforming, storing, analyzing, interpreting, visualizing, and querying data require substantial resources. The data should provide value and actionable information for organizations to justify investments in Big Data.

According to EMC Digital Universe with Research & Analysis by International Data Corporation (IDC) [2], data is growing at an annual rate of 40% into the next decade. What is significant is that the smart devices that are connected to the Internet – Internet of Things (IoT) – will contribute significantly to the data volumes, velocity, and heterogeneity. Furthermore, data is doubling in size every two years. The data volume will reach 44 zettabytes by 2020, from 4.4 zettabytes in 2013 [2].

Currently, much of the Big Data, especially that originating from the social media and businesses are not looked at or analyzed more than once. However, this situation is likely to change. Data becomes more useful if it is enhanced by adding meta-data and semantic annotations. According to IDC [2], by 2020, more than 35% of all data could be considered useful due to increased production of it from IoT devices.

Big Data provides challenges as well as opportunities. There are numerous challenges from databases perspective, which are discussed in Section II. The opportunities lie in creatively integrating and analyzing heterogeneous data from multiple sources to drive innovation and produce value. Big Data is also creating a new paradigm for scientific research and related applications – data-driven science. For example, many problems in natural language understanding, computer vision, and predictive data analytics are ill-posed for solution using exact algorithms. In such cases, statistical models are used to deal with the problem complexity. In his 1960 article titled *The Unreasonable Effectiveness of Mathematics in the Natural Sciences*, Wigner discusses how mathematical models developed in one context were found to be equally applicable in totally unrelated contexts [3]. In a recent article [4], Halevy, Norvig, and Pereira argue that the accurate selection of a mathematical model ceases its

importance when compensated by *big enough* data.

The primary goal of this paper is to provide a unified view of the various issues associated with Big Data management at a conceptual level. The intent is to help the reader quickly gain an understanding of the challenges involved in harnessing the power of Big Data. The issues we consider are data quality, data streams, dynamically evolving data, data heterogeneity and modeling, multi-model databases, client and query interfaces, data compression, data encryption, privacy, access control and authorization, and deployment on cloud-hosted cluster computers. Though not all issues are applicable to every Big Data application, often they have implications indirectly through cross interactions. For example, the complexity of client and query interfaces is directly impacted by the multi-data model.

The rest of the paper is organized as follows. Section II discusses various challenges inherent to Big Data management. One such challenge, security, is elaborated in Section III. Conclusions and future research directions are provided in Section IV.

II. Big Data Management Challenges

The challenges we discuss in this section include data quality, data streams, dynamically evolving data, data heterogeneity and data modeling, multi-model databases, client and query interfaces, data compression, data encryption, access control and authorization, and deployment on cloud-hosted cluster computers. One task that crosscuts all of the above challenges is identifying a subset of Big Data that has high value. This requires separating the data that is contaminated by spam, noise, and bias from that which is uncontaminated.

A. Data Quality

In addition to internally generated data, many organizations acquire massive datasets from diverse data vendors. Typically, the data acquired from vendors is produced without any specific application or analysis context. However, the perceived meaning of the data varies with the intended purpose [5]. This necessitates defining data validity and consistency in the context of intended data use. A related issue is inconsistency between the vendor supplied data and the same data which has been modified to conform to intended use-specific validity and consistency checks. Another issue is the need for maintaining data validity and consistency across the recent and older datasets given the long data life cycles in Big Data context.

B. Data Streams

Continuous data streams are the norm in applications, such as security surveillance, sensor networks, clickstream monitoring, and network-operations monitoring. Current approaches to data stream processing focus on application specific solutions rather than generic frameworks and approaches. For example, Najmabadi et al. [6] extracting connected component labels from image and video streams using fine grain parallel field programmable gate arrays. A low-power, many-core architecture for data stream mining applications is discussed by Kanoun et al. in [7]. Yang et al. [8] describe a cloud-based, multi-filter strategy for querying streaming XML Big Data.

Data streams pose special problems given the limited memory and CPU-time resources [9]. Unlike the *one-time database queries*, streaming data queries are *long-running and continuous*. Integrating data from multiple heterogeneous streams, mining streaming data through clustering and other unsupervised machine learning techniques, dealing with data quality issues, and real-time processing of fast moving data streams are open research issues.

C. Dynamically Evolving Data

Credit card fraud detection applications critically depend on real-time and current data. Detecting fraud in applications, such as United States (US) government sponsored health care programs Medicare and Medicaid [10] requires modeling and processing of dynamically changing data. The US Congressional Office of Management and Budget estimates that improper payments in Medicare and Medicaid programs in 2010 amounts to $50.7 billion.

Time-evolving graphs are used to model, store, process, analyze, visualize, and mine dynamically evolving data [11]. Since these graphs tend to be large and require low-latency, special hardware is used. For example, YarcData's Urika appliance was used to detect this type of fraud in real-time. The appliance memory can scale to 512 terabytes, which is shared by up to 8,192 CPUs.

D. Data Heterogeneity and Modeling

One problem that almost all organizations face is dealing with disparate data mediums and quality and extent of semantic annotations. Not all data can be stored using the relational data model and yet meet the stringent latency requirements for data access. For this reason, several data models have emerged during the last few years [12][13].

The new data models include key-value [14], column-oriented relational [15], column-family [16], document-oriented [17], and graph-based [18][19]. Furthermore, object-oriented and XML databases are reemerging in the Big Data context. Resource Description Framework (RDF) [20] data model is increasingly used for knowledge representation. The data modeling challenge in Big Data context is how to model heterogeneous data which requires multiple data models.

E. Multi-model Databases

Big Data value primarily comes through integrating massive heterogeneous datasets. As discussed in Section II-D, it is unlikely that a single model can capture the essential characteristics of heterogeneous data. It is more natural and practical to model the heterogeneous data using a collection of data models.

To provide a simple and unified user access to data, a *meta data model* should abstract all the underlying

data models. User queries will be specified using the meta model. Next, the query against the meta model needs to be decomposed into several queries, each of which is executed against a specific data model. Given the complexities of meta model development and query decomposition, Database-as-a-Service model [21] will simplify Big Data application development.

F. Client and Query Interfaces

Client interfaces provide programmatic access to data, whereas query interfaces are used for interactive querying and exploration of data. Efficient data access is a paramount consideration for Big Data applications. Structured Query Language (SQL) is the *de facto* standard for querying and updating relational databases. In contrast, there is no such standard query language yet which can be used to specify queries that require access to data across databases with disparate data models.

Client interfaces are typically developed by database professionals who are knowledgeable about the database schemas and how to link the data. Another issue is the number of programming languages for which client access interface is available.

The case of Statoil exploration illustrates the complexities involved in user access to Big Data [22]. Statoil is an oil and gas production company. One of its core tasks is to reduce exploration risk by developing stratigraphic models of unexplored areas using the data of previous operations at nearby sites. The data for stratigraphic modeling exceeds one petabyte and is stored under various relational database schemas. Answering queries require accessing data that is spread over 2,000 tables across different databases. As reported by Calvanese [22], answering certain queries require over four days even with the assistance of database experts. About 30 - 70% of oil and gas exploration time is spent on data gathering. Data heterogeneity will make the data access task even more time consuming.

Ontology Based Data Access (OBDA) is an approach to querying Big Data [23][24]. An ontology provides a formal representation of a domain at a conceptual level. Mappings are constructed between the ontology and data. Users specify data requests using the ontology. OBDA system translates a user data request into queries across various data sources.

Medical Literature Analysis and Retrieval System Online (MEDLINE) [25] is a bibliographic database featuring articles from academic journals covering medicine, nursing, pharmacy, dentistry, veterinary medicine, and health care. PubMed [26] stores over 24 million citations from MEDLINE and online books. Likewise, ACL Anthology is a digital archive of over 34,000 research papers in computational linguistics and natural language processing [27].

Even with advanced search tools it is often difficult to find and discover what one is looking for from large document collections. Therefore, collections like the MEDLINE and ACL Anthology need a different type of user interface for organizing, exploring, querying, visualizing, and understanding them. *Topic models* are tools just for this purpose [28]. Topic models are algorithms for discovering main themes that are present in a document collection and to organize the collection according to these themes [29].

G. Data Compression

With Big Data, secondary storage is always a premium resource. Data replication for high availability and read throughput aggravates this problem further. Compression ratios depend on the nature of the data itself as well as the compression algorithm used. Some compression algorithms are lossless, while others are lossy. If the data have been compressed using a lossless algorithm, then it is possible to recover the original data exactly from the compressed data [30]. Typically, text compression requires a lossless algorithm. Image and video data, on the other hand, may tolerate some data loss when decompressed. Application requirements dictate the choice of a compression algorithm – lossy or lossless.

Some algorithms focus on decompression speed and exploit the underlying hardware [31]. Other considerations include what additional resources are required by an algorithm. For example, some algorithms may require memory buffers for both compression and decompression [30].

Lempel-Ziv-Oberhumer (LZO) [32] is a popular, lossless data compression algorithm and its compression ratio is about 3:1. It is optimized for very fast decompression and is often used with Hadoop [33]. Relational database systems also offer options for data compression and compression ratios hover around 6:1. Gzip [34] is both a file format and a compression algorithm, whose compression ratio is about 7:1. Other products, such as RainStor database seem to provide a much higher compression ratio, 40:1 and in some cases as high as 100:1 [35].

H. Data Encryption

Big Data applications need to comply with global data security and privacy regulations to realize potential business benefits. For example, in the health care domain, patient data is made available to research and development organizations to analyze and identify emerging risks and patterns in the patient population. Health Information Portability and Accountability Act (HIPAA) and Health Information Technology for Economic and Clinical Health Act (HITECH) [36] compliance requires that the data be de-identified at the field level. If needed, the de-identified data needs to be securely re-identified for proactive treatment of the affected individuals.

Given that Big Data is stored in distributed file systems and are processed using cloud-hosted cluster computers, securing the data through encryption is extremely challenging. For example, file-system encryption is effective only for data at rest. This introduces excessive operational overhead for the continuous write encryption and read decryption. Furthermore, encryption-decryption cycle should preserve formats and referential integrity of the data. Finally, encryption/decryption should be

cost effective and not diminish operational flexibility or computational performance.

I. Privacy

Privacy and security are tightly integrated aspects of Big Data. Protecting rights of privacy is a great challenge. For example, in 2013, there were more than 13 million identity thefts in the United States and it is one of the fastest growing crimes [37]. Other facets such as encryption in both hardware and software, and round the clock monitoring of security infrastructure are critical to protecting privacy. The notion of personally identifiable information is difficult to define precisely. Furthermore, as data goes through various transformations, it becomes even more difficult to identify and tag personally identifiable data. It has been shown that even anonymized data can often be re-identified and attributed to specific individuals [38]. Sadly, re-identification has become a center piece for business models employed in fields such as online behavioral advertising.

J. Access Control and Authorization

Access control refers to ways in which user access to applications and databases is controlled. Databases limit access to those users who have been authenticated by the database itself or through an external authentication service, such as Kerberos [39]. Authorization controls what types of operations can an authenticated user perform. Access control and authorization capabilities of relational database systems have evolved over a period of four decades. In contrast, data management for Big Data applications is provided by a class of systems referred to as Not Only SQL (NoSQL) systems [40].

NoSQL systems principally focus on providing near real-time reads and writes in the order of billions and millions, respectively. NoSQL systems features vary widely and there are no standards yet. They use different data models, some do not provide database transactions, while others do not use SQL. They are referred to by various names including NoSQL, NewSQL, and non-RDBMS. To avoid the misconception that NoSQL systems eschew SQL, they are also referred to as *Not only SQL*.

NoSQL systems are relatively new and are evolving very rapidly. Their access control and authorization capabilities vary widely. Some NoSQL systems provide limited capabilities and some assume that the system is operating in a trusted environment. For example, initial versions of Riak, a key-value NoSQL database, provided no authentication or authorization support [41]. We elaborate on this aspect in Section III.

K. Deployment on Cloud-hosted Cluster Computers

Though Big Data applications can be developed and tested on desktop computers on a small scale, usually they are developed, tested, and deployed on cluster computers. Installing, operating, and maintaining cluster computers require specialized technical expertise in addition to significant upfront investment in hardware. For this reason, many Big Data applications are developed using cloud-hosted, cluster-powered application hosting commercial platforms such as Amazon Web Services [42] and Heroku [43]. In contrast, XSEDE is a free supercomputer platform dedicated for advancing academic science and engineering research [44].

III. Security Challenges in Managing Big Data

Database systems security has been a topic of major research interest in the database community [45]. Database security has multiple dimensions including physical, personnel, operational, and technical. The physical dimension deals with barriers to ensure physical inaccessibility to unauthorized users. The personnel facet is related to employing trustworthy people to operate the database. Policies and procedures that govern operating and maintaining aspects of databases comprise the operational dimension. These three dimensions are external to the technical aspects of database systems.

Traditionally, security aspects addressed by the database system include protecting confidentiality, ensuring data integrity, and assuring availability. Protecting confidentially involves preventing unauthorized access to sensitive information such as health records, credit history, trade secrets, marketing and sales strategies. This may require encrypting the data, authenticating users, and providing fine granular access.

Ensuring data integrity requires that data insertions, modifications, and deletions are done by authorized users in a way that none of the database integrity constraints are violated. Attacks such as data corruption through viruses and SQL injections make data integrity assurance a difficult job. High availability requires database system's resilience to attacks such as denial of service.

Big Data ushers in several more challenges. For example, initiatives by various governments, such as Right to Information [46], Freedom of Information [47], and Open Government Initiative [48] provide access to vast amounts of data to the public at large. One of the greatest challenges is privacy-preserving data mining and analytics – ensuring that deriving personally identifiable information is impossible.

The sheer volume of data can easily overwhelm the first-generation security information and event management technologies. For example, Barclays bank generated over 44 billion security events per month in 2013 [49]. Analyzing database access logs to proactively identify security breaches is also made difficult by data volume. Identifying useful data for a given context from massive datasets is a problem in itself. This problem is often referred to as *right data* in contrast to *big data*.

Data input validation and filtering, real-time regulatory compliance monitoring, and secure communications pose additional problems. Cloud-hosted, distributed cluster computing infrastructure must ensure secure computations by encrypting data during transit. Finally, data provenance [50] is an issue that received little or no attention from a security standpoint. As data goes through various transformations, metadata associated with provenance grows in complexity. The size of provenance graphs

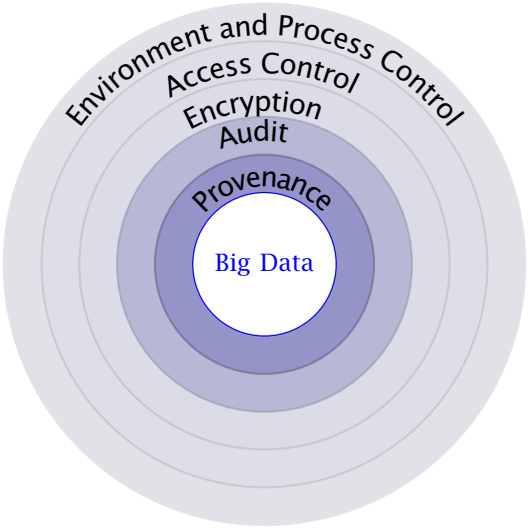

Figure: 1 Five-layer Big Data security model.

increases rapidly [51] which makes analyzing them computationally expensive.

Big Data security models entail more complexity due to their distributed nature relative to traditional database systems. We envision a five-layer security model for Big Data security as shown in Figure 1. Security controls are applied beginning with the outermost layer and sequentially progressing to the innermost layer. The outermost layer, *Environment and Process Control*, enforces physical controls in the underlying deployment environment such as firewalls, file system permissions, network configurations.

The *Access Control* layer restricts access to data through user authentication and authorization controls. The next layer provides data *encryption and decryption* services. Encryption is needed for both data-at-rest and data-in-flight. The *Provenance* layer is responsible for tracking data transformations and recording their annotations. The innermost one is the *Audit* layer which records and analyzes all database accesses in real-time to discover security breaches and to ensure compliance.

IV. Conclusion and Future Work

The ability to effectively process massive datasets has become integral to a broad range of academic disciplines. However, this does not preclude the need for deeper understanding of the theoretical foundations of scientific domains. The adage – a tool without theory is blind and a theory without tool is useless – holds in the Big Data context too. Big Data enables scientists to overcome problems associated with small data samples in ways, such as relaxing the assumptions of theoretical models, avoiding over-fitting of models to *training data*, effectively dealing with noisy training data, and providing ample *test data* to validate models.

Big Data has the potential to fundamentally affect the way we live and work. Just as a picture is worth 1000 words, Big Data analytics can help us unravel a thousand stories by analyzing and interpreting the data. Big Data offers possibilities for uncovering unexpected and hidden insights. For example, in medical and health data, these insights may lead to ground-breaking discoveries and profitable innovation. However, several problems need to be solved before this potential can be realized. As much as it sounds ironical, only technology can solve technology created problems. We expect that the research issues raised in this paper will inspire the readers to solve Big Data problems and advance this fast moving and exciting field.

Acknowledgment

S. Jothilakshmi is a postdoctoral researcher at Marshall University, USA. She is sponsored by the University Grants Commission of India under the Raman Fellowship program.

References

[1] Executive Office of the President. Big data: Seizing opportunities, preserving values. [retrieved: March, 2015]. [Online]. Available: http://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_5.1.14_final_print.pdf

[2] V. Turner. The digital universe of opportunities: Rich data and the increasing value of the internet of things. [retrieved: March, 2015]. [Online]. Available: http://www.emc.com/leadership/digital-universe/2014iview/digital-universe-of-opportunities-vernon-turner.htm

[3] E. Wigner, "The unreasonable effectiveness of mathematics in the natural sciences," *Communications in Pure and Applied Mathematics*, vol. 13, no. 1, pp. 1 – 14, February 1960.

[4] A. Halevy, P. Norvig, and F. Pereira, "The unreasonable effectiveness of data," *IEEE Intelligent Systems*, vol. 24, no. 2, pp. 8 – 12, 2009.

[5] D. Loshin. Understanding big data quality for maximum information usability. [retrieved: March, 2015]. [Online]. Available: http://www.dataqualitybook.com

[6] S. M. Najmabadi, M. Klaiber, Z. Wang, Y. Baroud, and S. Simon, "Stream processing of scientific big data on heterogeneous platforms – image analytics on big data in motion," *2013 IEEE 16th International Conference on Computational Science and Engineering*, pp. 965–970, 2013.

[7] K. Kanoun, M. Ruggiero, D. Atienza, and M. van der Schaar, "Low power and scalable many-core architecture for big-data stream computing," *2014 IEEE Computer Society Annual Symposium on VLSI*, pp. 468–473, 2014.

[8] C. Yang, C. Liu, X. Zhang, S. Nepal, and J. Chen, "Querying streaming xml big data with multiple filters on cloud," *2013 IEEE 16th International Conference on Computational Science and Engineering*, pp. 1121–1127, 2013.

[9] L. Golab and M. T. Özsu, "Data stream management," *Synthesis Lectures on Data Management*, vol. 2, no. 1, pp. 1–73, 2010.

[10] US Government. Centers for medicare & medicaid services. [retrieved: March, 2015]. [Online]. Available: http://www.cms.gov/

[11] V. V. Raghavan. Visual analytics of time-evolving large-scale graphs. [retrieved: March, 2015]. [Online]. Available: http://grammars.grlmc.com/bigdat2015/courseDescription.php

[12] solid IT. Knowledge base of relational and NoSQL database management systems. [retrieved: March, 2015]. [Online]. Available: http://db-engines.com/en/ranking

[13] A. Schram and K. M. Anderson, "Mysql to nosql: Data modeling challenges in supporting scalability," in *Proceedings of the 3rd Annual Conference on Systems, Programming, and Applications: Software for Humanity*, ser. SPLASH '12. New York, NY, USA: ACM, 2012, pp. 191–202.

[14] R. Gandhi, A. Gupta, A. Povzner, W. Belluomini, and T. Kaldewey, "Mercury: Bringing efficiency to key-value stores," in *Proceedings of the 6th International Systems and Storage Conference*, ser. SYSTOR '13. New York, NY, USA: ACM, 2013, pp. 6:1–6:6.

[15] Z. Liu, S. Natarajan, B. He, H.-I. Hsiao, and Y. Chen, "Cods: Evolving data efficiently and scalably in column oriented databases," *Proc. VLDB Endow.*, vol. 3, no. 1-2, pp. 1521–1524, Sep. 2010.

[16] A. Lakshman and P. Malik, "Cassandra: A structured storage system on a p2p network," in *Proceedings of the Twenty-first Annual Symposium on Parallelism in Algorithms and Architectures*, ser. SPAA '09. New York, NY, USA: ACM, 2009, pp. 47–47.

[17] P. Murugesan and I. Ray, "Audit log management in mongodb," *2014 IEEE World Congress on Services*, pp. 53–57, 2014.

[18] R. Angles, "A comparison of current graph database models," *2014 IEEE 30th International Conference on Data Engineering Workshops*, pp. 171–177, 2012.

[19] I. Robinson, J. Webber, and E. Eifrem, *Graph Databases*. O'Reilly, 2013.

[20] Z. Kaoudi and I. Manolescu, "Rdf in the clouds: A survey," *The VLDB Journal*, vol. 24, no. 1, pp. 67–91, Feb. 2015.

[21] D. Agrawal, A. El Abbadi, F. Emekci, and A. Metwally, "Database management as a service: Challenges and opportunities," in *Data Engineering, 2009. ICDE '09. IEEE 25th International Conference on*, March 2009, pp. 1709–1716.

[22] D. Calvanese. End-user access to big data using ontologies. [retrieved: March, 2015]. [Online]. Available: http://grammars.grlmc.com/bigdat2015/courseDescription.php

[23] D. Calvanese, G. D. Giacomo, D. Lembo, M. Lenzerini, A. Poggi, and R. Rosati, "Ontology-based Database Access," in *Sistemi Evoluti per Basi di Dati*, 2007, pp. 324–331.

[24] A. Poggi, D. Lembo, D. Calvanese, G. D. Giacomo, M. Lenzerini, and R. Rosati, *Linking Data to Ontologies*, 2008, vol. 10.

[25] U.S. National Library of Medicine. Medical Literature Analysis and Retrieval System Online (MEDLINE). [retrieved: March, 2015]. [Online]. Available: http://www.nlm.nih.gov/pubs/factsheets/medline.html

[26] NCBI. PubMed, National Center for Biotechnology Information. [retrieved: March, 2015]. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed

[27] ACL, "ACL Anthology: A Digital Archive of Research Papers in Computational Linguistics," http://www.aclweb.org/anthology/, [retrieved: March, 2015].

[28] H. Soleimani and D. J. Miller, "Parsimonious topic models with salient word discovery," *IEEE Transactions on Knowledge and Data Engineering*, vol. 99, no. PrePrints, p. 1, 2014.

[29] D. M. Blei, "Probabilistic topic models," *Commun. ACM*, vol. 55, no. 4, pp. 77–84, Apr. 2012.

[30] K. Sayood, *Introduction to Data Compression*, 4th ed. Morgan Kaufmann, 2012.

[31] A. Ozsoy, "Culzss-bit: A bit-vector algorithm for lossless data compression on gpgpus," in *Proceedings of the 2014 International Workshop on Data Intensive Scalable Computing Systems*, ser. DISCS '14. Piscataway, NJ, USA: IEEE Press, 2014, pp. 57–64.

[32] S. Navqi, R. Naqvi, R. A. Riaz, and F. Siddiqui, "Optimized rtl design and implementation of lzw algorithm for high bandwidth applications," *Electrical Review*, no. 4, pp. 279–285, April 2011.

[33] A. Holmes, *Hadoop in Practice*. Manning Publications Co., 2012.

[34] J.-l. Gailly, *gzip: The data compression program*. iUniverse, 2000.

[35] RainStor. Industry leading compression translates to huge cost savings. [retrieved: March, 2015]. [Online]. Available: http://rainstor.com/products/rainstor-database/compress/

[36] T. L. Murray, M. Calhoun, and N. C. Philipsen, "Privacy, confidentiality, hipaa, and hitech: Implications for the health care practitioner," *The Journal for Nurse Practitioners*, vol. 7, no. 9, pp. 747–752, October 2011.

[37] United Credit Service. Identity theft; will you be the next vitcim? [retrieved: March, 2015]. [Online]. Available: https://ucscollections.wordpress.com/2014/03/06/identity-theft-will-you-be-the-next-victim/

[38] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," *2013 IEEE Symposium on Security and Privacy*, vol. 0, pp. 111–125, 2008.

[39] S. T. F. Al-Janabi and M. A. S. Rasheed, "Public-key cryptography enabled kerberos authentication," in *Developments in E-systems Engineering (DeSE), 2011*. IEEE Computer Society, Dec 2011, pp. 209–214.

[40] V. Gudivada, D. Rao, and V. Raghavan, "Renaissance in data management systems: Sql, nosql, and newsql," *IEEE Computer, forthcoming*.

[41] solid IT. Current data security issues of nosql databases network defense & forensics insights. [retrieved: March, 2015]. [Online]. Available: http://www.fidelissecurity.com/files/NDFInsightsWhitePaper.pdf

[42] AWS. Amazon web services. [retrieved: March, 2015]. [Online]. Available: http://aws.amazon.com/

[43] Heroku. Cloud application platform. [retrieved: March, 2015]. [Online]. Available: https://www.heroku.com/

[44] XSEDE. Advanced cyberinfrastructure. [retrieved: March, 2015]. [Online]. Available: http://www.xsede.org

[45] E. Bertino and R. Sandhu, "Database security – concepts, approaches, and challenges," *IEEE Transactions on Dependable and Secure Computing*, vol. 2, no. 1, pp. 2–19, 2005.

[46] Government of India. Right to information. [retrieved: March, 2015]. [Online]. Available: http://righttoinformation.gov.in/

[47] USA Federal Government. Freedom of information act. [retrieved: March, 2015]. [Online]. Available: http://www.foia.gov/

[48] The White House. Open government initiative. [retrieved: March, 2015]. [Online]. Available: http://www.whitehouse.gov/Open/

[49] B. Glick. Information security is a big data issue. [retrieved: March, 2015]. [Online]. Available: http://www.computerweekly.com/feature/Information-security-is-a-big-data-issue

[50] U. Braun, A. Shinnar, and M. Seltzer, "Securing provenance," in *Proceedings of the 3rd Conference on Hot Topics in Security*, ser. HOTSEC'08, 2008, pp. 4:1–4:5.

[51] Y.-W. Cheah, "Quality, retrieval and analysis of provenance in large-scale data," Ph.D. dissertation, Indianapolis, IN, USA, 2014.