New Design Principles for Effective Knowledge Discovery from Big Data

Anjana Gosain
USICT
Guru Gobind Singh Indraprastha University
Delhi, India

Nikita Chugh
USICT
Guru Gobind Singh Indraprastha University
Delhi, India

ABSTRACT
Big data is creating hype in IT industry. Knowledge discovery from big data can allow organizations to have deeper insights, look at the bigger picture and project big returns. There are various principles that have been presented for knowledge discovery from big data by ORNL (Oak Ridge National University), USA. These are: (i) support a variety of analysis methods, (ii) one size doesn't fit all, (iii) make data accessible. However timeliness and security still pose great challenges in the knowledge discovery process. Timely analysis of big data is essential because data is being produced at a very high velocity. Security of big data is difficult to ensure since big data solutions were not developed with security in mind. In this paper, we give a view of various big data dimensions and present two new principles based on security and timely analysis for knowledge discovery from big data.

Keywords
Big Data, Knowledge Discovery, No Sql, Security, Sql, Timeliness

1. INTRODUCTION
Big Data refers to large volumes of high variety data being generated at high velocity. It is these three V's (volume, variety, velocity) that make big data different from traditional large volumes of data. Big data is being produced every second from online transactions, emails, videos, mobile phone, social networking, search queries etc. It is the decline in the cost of storage and processing power that have made it feasible to collect this data which was thrown away only a few years ago [1]. Big Data has changed the way people or organizations look at the data being produced rapidly and in large volumes. It is this term that is changing the course of new emerging technologies.

Big data analytics refers to capturing and analyzing big data for discovering interesting patterns and relationships in it. ORNL in its paper titled "Design Principles for Effective Knowledge Discovery from Big Data" has given three design principles [2]. These principles focus on special architectural requirements that an organization must look upon for Big Data Analytics. The first principle states to "support a variety of analysis methods". This is because big data can serve different purposes for different users. The second principle states that "one size doesn't fit all". It focuses on the need of various different data stores for storing and processing data at all the stages of the pipeline [2]. The third principle, "make data accessible", states architectural needs for making results accessible and easy to understand.

In nutshell, these three principles provide solutions to big data problems of heterogeneity, scale and accessibility respectively. But there are problems that still remain unaddressed like timeliness, privacy, security, cost, real time analysis etc.

Big data is being generated at very high velocity. We need to capture these streams of data and analyse them in least possible time. Timeliness is very essential in cases where real time analytics is needed. Also there are certain big data applications like credit card fraud detection where if timely analysis of data is not done it can cause severe losses.

Security is another major concern in handling big data. Big data is not limited to the boundaries of an organisational network. Rather it is available through and on internet. There are various security issues that need to be looked upon in big data analytics. Traditional security solutions cannot ensure big data security due to its complex nature.

In this paper we describe various dimensions of big data like its features, objectives, architecture, sources etc. and present new principles for timely analysis and security in big data. These principles focus on the need to perform timely analysis and ensuring security of big data.

The paper is worded as follows: Section 2 gives an overview of various big data dimensions. Section 3 describes the existing principles for knowledge discovery from big data. Section 4 introduces two new principles that are of significant importance in knowledge discovery from big data. Section 5 concludes the work and discusses future opportunities.

2. BIG DATA DIMENSIONS
Big data can be studied along various dimensions. They give a deeper and bigger picture to what exactly big data is. Fig. 1 summarizes the dimensions of big data.

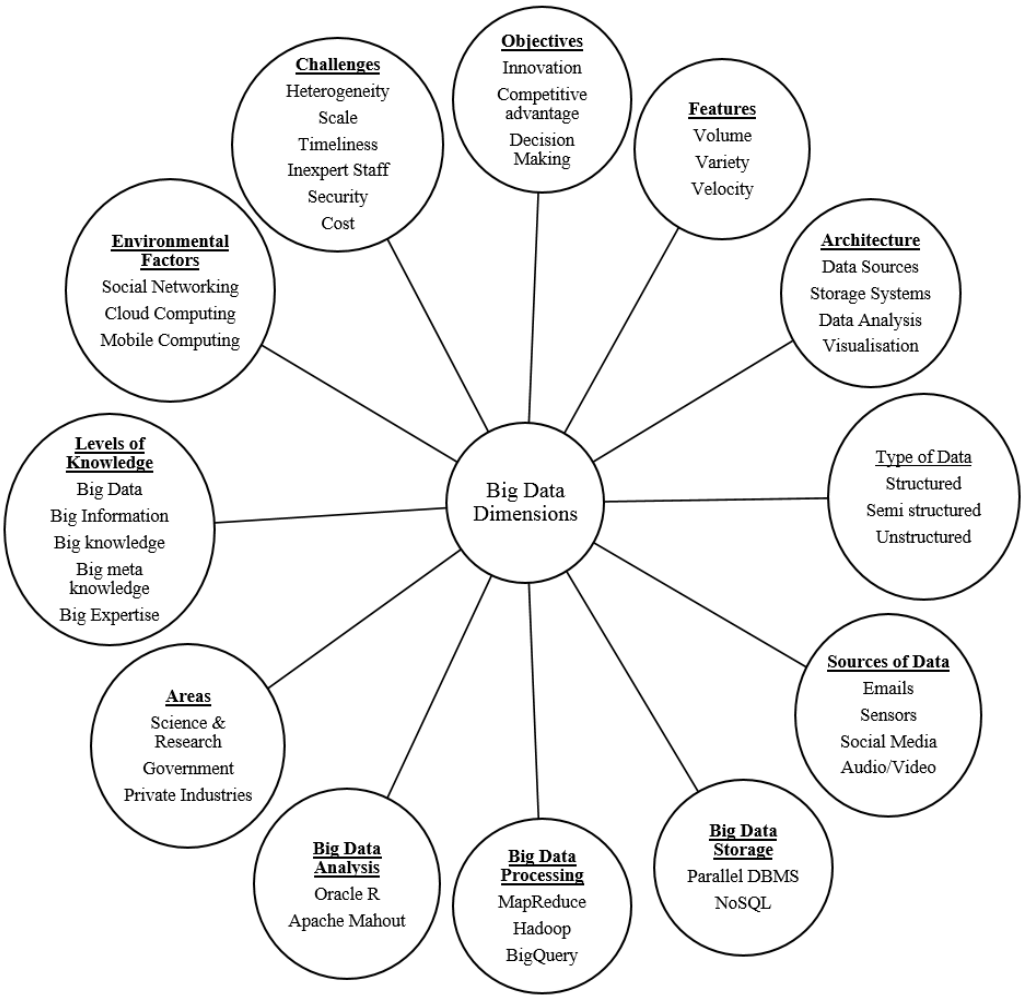

Fig 1: Big Data Dimensions

- Objectives - This dimension describes the various objectives that big data can help us to achieve. It can be used by various industries with different purposes in mind. Most industries identify big data as the most powerful resource for gaining competitive advantage. Others look at it as a treasure of information for new innovations. It can also be used for the purpose of business intelligence and decision making. It can be efficiently used to depict customer behaviours and future needs.

- Features – Big data is a term used for massive sets of data being generated at high velocity from variety of sources. It differs from traditional databases on the basis of 3 V's i.e. Volume, Variety, and Velocity [1, 3, 4]. Volume of big data refers to the size of data. The data being produced is now of the order of petabytes and above. The feature of variety indicates different forms of data. Data sources are heterogeneous both at the schema level and at the instance level [5]. Velocity indicates the rate at which data is being generated and is made available. Information is being generated at faster rate than ever before. A significant challenge here is to perform real time or near real time analysis of big data.

- Architecture - Big data can be studied from the point of view of its architectural structure. Due to unique characteristics of big data, traditional data warehouse or informational data architectures do not perform well. The Big Data architecture challenge is to meet the rapid use and rapid data interpretation requirements while at the same time correlating it with other data [6]. Oracle has proposed an architecture that bridges traditional information architecture and big data architecture [6]. As stated by Oracle, Big Data architecture starts from the point of data sources. This data is then captured by various storage systems. The data is then processed by using various technologies like MapReduce [7], Hadoop [8] etc. which are present above the storage systems. The next level of big data architecture aims at integration and analysis of data. The last level of big data architecture deals with making results accessible. It contains various tools for visualizing the results in the most efficient manner.

- Type of data - Big data encompasses structured, semi structured and unstructured data. According to TCS, big data contains structured data that can be stored in fixed fields, unstructured data that cannot be stored in fixed fields and semi-structured data that cannot be stored in fixed fields but tags can be used to categorize this data [9]. Also it was observed by TCS that 51% of data is structure, 27% of data is unstructured and 21% of data is semi structured [9].

- Sources of data - Big data is generated from almost every digital process. Some of the major sources that generate big data are mobile devices and personal computers, social media, online transactions, emails, videos, audios, images, sensors etc. From the point of view of a company, the sources of big data can be internal or external. Internal sources include sources within the company like employee records, production records, visits to company website etc. External sources include sources outside the company like social media websites, government regulations etc.

- Big data storage - This dimension identifies various technologies for storage and management of big data. The two most common technologies to store large data are parallel DBMS and NoSQL [10]. They both are capable of scaling out for handling big data.

- Big data processing - The next dimension after big data storage is big data processing. In big data processing a problem is divided into several small operations, and results are then combined [10]. The key point of Big data processing is Divide and Conquer. Commonly used technologies for big data processing are MapReduce [7], Hadoop [6, 8], BigQuery [11].

- Big Data analysis - This aspect focuses on interactive data analysis with real-time answers. Two commonly used technologies for big data analysis are Oracle R [10, 12] and Apache Mahout [10, 13].

- Areas - Big data analysis can provide better returns in every sector of society. It is extensively being used in the field of science and research. For example large geospatial data is being produced and stored for the purpose astronomical analysis. Big data analytics has also helped governments to utilise the hidden patterns in data for the purpose of planning and decision making. One of the striking examples of big data involvement in the government sector is the announcement of Big Data Research and Development Initiative by the Obama Administration in 2012 [14]. Private industries such as healthcare, telecommunication, finance, marketing, utilities etc. are also using big data to a great extent. According to IBM data [15], in Healthcare there was a 20% decrease in patient mortality by analysis of streaming patient data.

- Levels of Knowledge - In the depth of knowledge, there are layers of knowledge content [16]. These levels of knowledge contain data at the lowest level and expertise at the highest level. As we move above from less structured, simple semantics data to more structured semantic rich data the various levels of knowledge are data, information, knowledge, meta-knowledge and expertise.

- Environmental Factors - Big data cannot survive alone without an array of supporting environmental factors. It is with factors like large processing technology, cloud computing, machine learning, mobile computing, and social networking etc. that the power of big data can be harnessed properly. Factors like mobile computing, cloud computing, social networking have emerged as the ones that are supporting as well as producing big data. SMAC - social, mobile, analytics and cloud together support each other to maximize their effects.

- Challenges with big data - Big data due to its distinct features velocity, variety and volume possesses significant challenges. The most commonly faced problems with big data are heterogeneity, scale, timeliness, complexity, and privacy. These problems appear right from the beginning of data acquisition phase and continue till the end of analysis process [17]. Timeliness comes into play in situations where the result of the analysis is required immediately. Real time analytics of huge varied data requires specialized tools and techniques. Another major concern is cost of ownership of big data [18]. Huge costs are incurred to organizations in storing and analysing big data. There is a trade-off between the costs incurred and the returns obtained. Also a major problem faced by organisations is availability of human resources to deal with big data. There has been a great gap between the demand and availability of data scientists in the industry. All these problems pose significant challenges in big data analytics. Despite these problems, the power of big data analytics cannot be ignored. Hence, there is a great need for advancement in technology to overcome all these problems.

3. EXISTING PRINCIPLES

Based on real world big data projects, ORNL has presented three system design principles for discovery of new knowledge. These principles aim to inform organisations about improving their big data analytics process. These are concerned with enabling researchers to handle big data in an easy and efficient manner. A brief description of these principles is given below.

3.1. Support a Variety of Analysis Methods

There are various users that perform operations on Big Data. Limiting them to using a set of tools is not efficient as some of them may be unfamiliar with the given set of tools. Also different users experience differences in the level of comfort in using a tool. For example, programmers may use java while statisticians may feel comfortable using R, SAS etc [2]. Therefore, the architecture must support a variety of methods and analysis environments [2]. It should support a variety of statistical, data mining, machine learning and visual analysis tools.

3.2. One Size does not Fit All

This principle discards the traditional approach of using a single storage mechanism. This is because the data being captured is diverse in variety and hence demands different storage architectures. Relying on single large relational databases can create problems in scaling up to highly vary big data. For example, processing unstructured or semi structured data using relational techniques can pose great problems. This is because not all data can be easily modelled using relational techniques [2]. Hence, there is a need for specialized data management systems.

3.3. Make Data Accessible

Merely presenting results of big data analytics is not enough. The results should be accompanied with relevant information like nature of inputs, assumptions while deriving these results etc. Also the results should be made available in a way that they are easy to understand and interpret. Three approaches that have been used to accomplish this are using open, popular standards, adoption of light weight architectures, and exposing results via web-based API [2].

4. PROPOSED PRINCIPLES

The principles proposed by ORNL provide a direction towards handling big data. They focus on big data problems of heterogeneity, scale and accessibility. But as the rate at which data is being produced is enormously high, we need new solutions to perform timely analysis of big data. Another major concern that needs to be looked upon is security. This aspect aims to protect value hidden in big data from intruders. In this section, we present two new principles (i) Perform timely analysis (ii) Ensure big data security. Our principles aim to add value to the already existing principles. These principles may help organisations to explore, analyze and interact with big data in the most productive ways.

4.1. Perform Timely Analysis

We suggest a new principle based on the problem of timeliness. Timeliness refers to the rate of data acquisition and data analysis. The acquisition phase is one of the major phases big data pipeline. As big data consists of data streams of higher velocity and higher variety, the architecture handling big data must deliver low, predictable latency in both capturing data and in executing short, simple queries [1]. Also there are cases when results of analysis are required immediately. The infrastructure required for analyzing big data must deliver faster response times.

Traditional SQL solutions do not address the need of timely analysis in big data. Data needs to be first parsed and transformed as per schema and then can be stored in SQL datastores. This may take a lot of time. The problems are getting complex. The scalability, low latency needs are greater. Hence, we cannot rely only on SQL solutions. NoSQL solutions can help solve some of the problems of traditional SQL solutions.

NoSQL databases are specialized databases for big data world. They are optimized for fast query capture and simple query patterns [1]. They provide faster capture because data is captured without modelling it according to schema.

However due to schema less nature of NoSQL, they cannot handle complex queries because additional intelligence is required to interpret the stored data. Due to this problem, NoSQL solutions cannot be used alone. They need SQL solutions for handling complex queries.

Therefore, a viable solution to realize the benefits of NoSQL solutions is that they must be integrated with SQL solutions into a single infrastructure supporting big data analysis. Hybrid strategy allows each database to do what it is best at [19].

Some of the possible ways to split data between SQL and NoSQL databases can be [19]:

- Place the high-read data needing high-availability in a NoSQL and historical, reporting data into a relational database.
- Keep the data that needs to be queried and reported in traditional SQL database and the data that needs best performance in a fast, distributed NoSQL database.
- For static data, NoSQL offers speed and efficiency. For dynamic, data entry with concurrent users, relational databases are more suitable.

This principle can also be seen as an extension of the already stated "one size doesn't fit all" principle since this also demands need of specialized databases.

4.2 Ensure Big Data Security

Big Data is an important asset that can allow organisations to have better return on investments. With the advent of supporting technology data is now being considered as important as labour and capital. But as big data model is built on web, security and privacy becomes a great issue. The more the data is available; the better is the analytics, but more it is vulnerable to threats.

The intelligence derived from big data is not only valuable to businesses and customers, but also to hackers [20]. Not ensuring big data security may cause financial and reputational losses to an organisation. E.g. Exposing credit card numbers or location of customers may undermine trust in the organisation.

Traditional information architectures were easier to secure as they consisted of a central repository. However in case of big data, there are various nodes among which data is distributed. Just having a firewall to protect our network will not work in this case. We need to protect both data and its value.

Another major issue in big data security is that this data is subject to various legal jurisdictions. Various countries have formulated their own privacy laws. Big data professionals need to keep these legislations in mind. A legal department must be involved in the formulation of policies regarding what to store, what to share, who can access what etc.

Some of the methods that organisations should adopt to ensure security of big data are:

- Use enterprise security solutions available for big data like HP ArcSight ESM [20], Zettaset Orchestrator [21] etc. This is because big data processing solutions like hadoop were not built with security in mind [21]. Hence they need supporting solutions to ensure security and privacy of big data. These solutions continuously monitor big data and its usage patterns to detect any malicious activity.

- Impose access control to ensure that right user gets right access to right data at the right time [22]. This prevents cyber criminal from having access to sensitive information. Controls should be set in accordance with the principle of least privilege.

- Weed off data that is of little value or no longer needed [23]. This reduces the risk of data breach as hacker cannot get access to data that is no longer stored and is already disposed. This requires careful formulation of company policies regarding what type of data is to be stored and for how long.

- Use attribute relationships for ensuring big data security [24, 25].This is another method of securing big data based on relationships among attributes present in big data. The attributes that need to be protected are identified based on type of big data and company policies.

5. CONCLUSION AND FUTURE WORK

Big Data can be looked upon as future of IT industry. In this paper we have presented various dimensions which characterize Big Data from various point of views. These dimensions represent the essential points that one should keep in mind while dealing with big data. Also, we have presented two new principles for timeliness and security in big data. These principles may help organisations to achieve better results of big data analytics. Following these principles potential benefits of big data can be realised in the most effective manner. However, there are still various issues that need to be addressed such as cost, real time analysis etc. Our future work will address such issues of Big Data.

6. REFERENCES

[1] Dijcks, J.P. 2012. Oracle: Big Data for Enterprise. White Paper. Oracle Corporation.

[2] Begoli, E. and Horey, J. 2012. Design Principles for Effective Knowledge Discovery from Big Data. In Proceedings of the Joint Working IEEE/IFIP Conference on Software Architecture (WICSA) and European Conference on Software Architecture (ECSA).

[3] Sagiroglu, S. and Sinanc, D. 2013. Big Data: A Review. In Proceedings of the International Conference on Collaboration Technologies and Systems.

[4] Demchenko, Y., Grzsso, P., De Laat, C., Membrey, P. 2013. Addressing Big Data Issues in Scientific Data Infrastructure. In the proceedings of the International Conference on Collaboration Technologies and Systems.

[5] Dong, X. L. and Srivastava, D. 2013. Big Data Integration. In the proceedings of the IEEE 29th International Conference on Data Engineering (ICDE).

[6] Sun, H. and Heller, P. 2012. Enterprise Information Management: Oracle Information Architecture: An Architect's Guide to Big Data. White Paper. Oracle Corporation.

[7] Dean, J. and Ghemawat, S. MapReduce: Simplifed Data Processing on Large Clusters. http://static.googleusercontent.com/media/research.google.com/en//archive/mapreduce-osdi04.pdf

[8] Jablonski, J. Introduction to Hadoop. White Paper. Dell

[9] Big Data Study – The 10 Key Findings. http://sites.tcs.com/big-data-study/big-data-study-key-findings/

[10] Commercial and Open Source Big Data Platforms Comparison. http://architects.dzone.com/articles/commercial-and open-source-big

[11] Sato, K. An Inside Look at Google BigQuery. White Paper. Google Inc.

[12] R Technologies from Oracle. http://www.oracle.com/technetwork/topics/bigdata/r-offerings-1566363.html

[13] What is Apache Mahout?. https://mahout.apache.org/

[14] Big Data is a Big Deal. http://www.whitehouse.gov/blog/2012/03/29/big-data-big-deal

[15] Big Data in Action.

http://www-01.ibm.com/software/in/data/bigdata

/industry.html

[16] Zhang, D. 2013. Inconsistencies in Big Data. In the proceedings of the 12th IEEE International Conference on Cognitive Informatics & Cognitive Computing (ICCI&CC).

[17] Challenges and Opportunities with Big Data. http://www.cra.org/ccc/files/docs/init/bigdatawhitepaper.pdf

[18] 2013. Reducing the Total Cost of Ownership of Big Data. White Paper. Impetus Technologies Inc.

[19] Sasirekha, R. NoSQL, the database for the Cloud. White Paper. Tata Consultancy Services Limited.

[20] Big security for big data. Business White Paper. Hewelett-Packard Development Company.

[21] The Big Data Security Gap: Protecting the Hadoop Cluster. White Paper. Zettaset. www.zettaset.com/info-center/datasheets/zettaset_wp_security_0413.pdf

[22] Kindervag, J., Balaouras, S., Hill, B. W., Mak, K. The Future of data Security and Privacy: Controlling Big Data. Technical report. Forrester Research Inc.

[23] Tankard, C, "Big Data Security", Journal of Network Security, vol. 2012 (Issue 7), July 2012.

[24] Kim, S. H., Kim, N. U., Chung, T. M. 2013. Attribute Relationship Evaluation Methodology for Big Data Security. In the proceedings of the International Conference on IT Convergence and Security (ICITCS).

[25] Kim, S. H., Eom, J. H., Chung, T. M. 2013. Big Data Security Hardening Methodology Using Attributes Relationship. In the proceedings of the International Conference on Information Science and Applications (ICISA).