International Journal of Engineering Research and Modern Education (IJERME)
ISSN (Online): 2455 - 4200
(www.rdmodernresearch.com) Volume I, Issue I, 2016

BIG DATA NEW CHALLENGES, TOOLS AND TECHNIQUES

Vaikunth Pai
Department of Information Technology, Srinivas Institute of Management Studies, Mangalore, Karnataka

Abstract:

Big data is a term for huge data sets having large, varied and complex structure with challenges, such as difficulties in data capture, data storage, data analysis and data visualizing for further processing. It requires new technologies and architectures so that it becomes possible to extract valuable data from it by capturing and analysis process. Big Data is a collection of massive data sets with a great diversity of types and it is difficult to process by using traditional data processing platforms. We analyze the challenges, tools and techniques for big data analysis and design.

Index Terms: Big Data, Hadoop, NoSQL & Data Privacy

1. Introduction:

Recent technological advances in IT, such as sensors, smart mobile devices, cloud systems, data analytics, and social networks, are making possible to capture, process, and share massive amounts of data and to extract useful knowledge such as patterns and predict future trends and events. Big data is making possible tasks that were impossible before, like preventing disease spreading, personalizing healthcare, quickly identifying business opportunities, managing emergencies and so on [1]. Big data refers to data volumes in the range of exabytes (10^{18}) and more. Such volumes exceed the capacity of current online storage systems and processing systems. Data, information and knowledge are being created and collected at a rate that is rapidly approaching the exabyte/year range.

Industry is using big data for business intelligence and decision support to make prediction and further actions. The use of big data is important for obtaining actionable knowledge. Governments are also interested in using big data and predictive analytics to improve decision making and transparency, to engage citizens in public affairs and to improve national security [10]. Healthcare represents another major area to which big data may offer novel opportunities [2]. Big Data Samplesare available in are astronomy, atmospheric science, genomics, biogeochemical, biological science and research, life sciences, medical records, scientific research, natural disaster and resource management, private sector, military surveillance, financial services, retail, social networks, web logs, text, document, photography, audio, video, click streams, search indexing, call detail records, POS information, RFID, mobile phones, sensor networks and telecommunications .

This paper describes the literature review in section 2, new challenges are reviewed in section3. In section4 important tools and techniques are described and section5 concludes the work.

2. Literature Review:

Intel IT Center of Big Data Analytics survey reports that there are several challenges for big data which includes data growth, data infrastructure, data governance/policy, data integration, data velocity, data variety, and data compliance/regulation and data visualization. Stonebreaker and Hong [3] says that the design of systems and components that work well with big data would require an good understanding of both the requirements of the users and the technologies that can be used to solve the problem being investigated and the end users will not often be the system designers, and this presents an additional design challenge.

The Departments of Defense and Energy, and the Defense Advanced Research Projects Agency announced a joint R&D initiative in March 2012 that will invest more than $200 million to develop new big data tools and techniques. Its goal is to advance our "…understanding of the technologies needed to manipulate and mine massive amounts of information" [4].

Sagiroglu, S.; Sinanc, D.(May 2013),"Big Data: A Review" [5]describe the big data content, its scope, methods, samples, advantages and challenges of Data. Using Knowledge Discovery from the big data it is easy to get the information from the complicated data sets.

Campus Technology Survey 2013[6] reports, lack of skilled users to work with big data toolset. Real Time Literature Review 2013, facebook has 1.11 billion people active accounts from which 751 million using facebook from a mobile. Flicker having feature of Unlimited photo uploads (50MB per photo), Unlimited video uploads (90 seconds max, 500MB per video), the ability to show HD Video, Unlimited storage, Unlimited bandwidth. Flickr had a total of 87 million registered members and more than 3.5 million new images uploaded daily.

The 2014 IDG Enterprise Big Data research reports, in the next 12-18 months, organizations plan to invest in skill sets necessary for big data deployments, including data scientists (27%), data architects (24%), data analysts (24%), data visualizers (23%), research analysts (21%), and business analysts (21%). Organizations are seeing exponential growth in the amount of data managed with an expected increase of 76% within the next 12-18 months [7].

The Asia HR Big Data Survey Report 2014 [8] identified that HR want to be strategic...but 80% do not have the right tools to get there, and only 23% of HR professionals know what HR Big Data is. Dell Midmarket Companies Survey 2014 [9] reports the most valuable technologies for midmarket companies running big data initiatives are real-time data processing, predictive analytics and data visualization tools.

3. Challenges:

There are three fundamental issues dealing with big data are storage issues, management issues, and processing issues.

- ✓ **Storage and Transport Issues:** The limit of Current disk technology is about 4 terabytes per disk. So, it would require 25,000 disks for 1 exabyte. Even though exabyte of data could be processed on a single computer system, it would be unable to directly attach the requisite number of disks [15]. Access to that data would overwhelm current communication networks.
- ✓ **Management Issues:** there is no perfect solution for big data management yet. This represents an important gap in the research literature on big data that needs to be filled. Jason has noted [11] that "there are no universally accepted way to store raw data … reduced data and … the code and parameter choices that produced the data".
- ✓ **Processing Issues:** For effective processing of exabytes of data requires extensive parallel processing and analytics algorithms in order to provide timely and actionable information [17].

Challenges to Big Data analysis include data inconsistency, incompleteness, scalability, timeliness and data security. Before to data analysis, data has to be well-constructed. However, considering different variety of data sets in Big Data problems, it is still a big challenge for us to propose efficient representation, access, and analysis of unstructured or semi-structured data in the further researches. The analysis process is shown In Fig.1, where the knowledge is discovered in data mining [12].

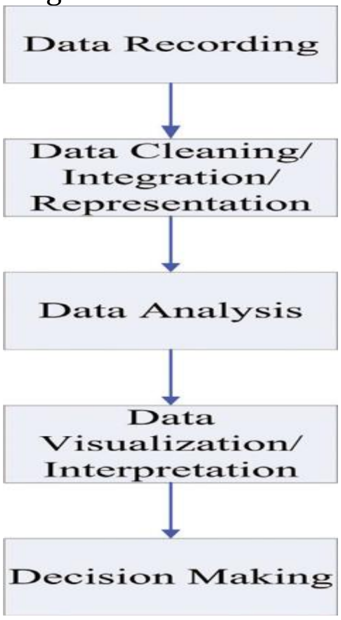

Data Recording
Data Cleaning/ Integration/ Representation
Data Analysis
Data Visualization/ Interpretation
Decision Making

Figure 1: Knowledge Discovery Process

3.1 Data Capture and Storage: Data sets grow in size because they are increasingly being gathered by different source of information-sensing mobile devices, aerial sensory technologies, remote sensing, software logs, cameras, microphones, radio-frequency identification readers, wireless sensor networks, and so on. There are 2:5 quintillion bytes of data are created every day and data size keeps on increasing exponentially. Big Data technology has changed the way we gather and store data, including data storage device, data storage architecture and data access techniques. It requires more sophisticated storage mediums with higher I/O speed to meet the challenges of big data issues. Direct-attached storage (DAS), network-attached storage (NAS), and storage area network (SAN) are the enterprise storage architectures that are commonly are in use. The existing storage architectures have severe drawbacks and limitations when it comes to large-scale distributed computing. Aggressive concurrency and per server throughput are the essential requirements for the applications on highly scalable computing clusters, and today's storage systems lack the both aspects. Optimizing data access is one of the popular ways to improve the performance of data-intensive computing; these techniques include data replication, data migration, data distribution and data access parallelism.

3.2 Data Transmission: Cloud data storage is popularly used as the development of cloud technologies. We know that the network bandwidth capacity is the bottleneck in cloud and distributed systems, especially when the volume of communication is large. On the other side, cloud storage also lead to data security problemsas the requirements of data integrity checking. Many schemes were proposed under different systems and security models.

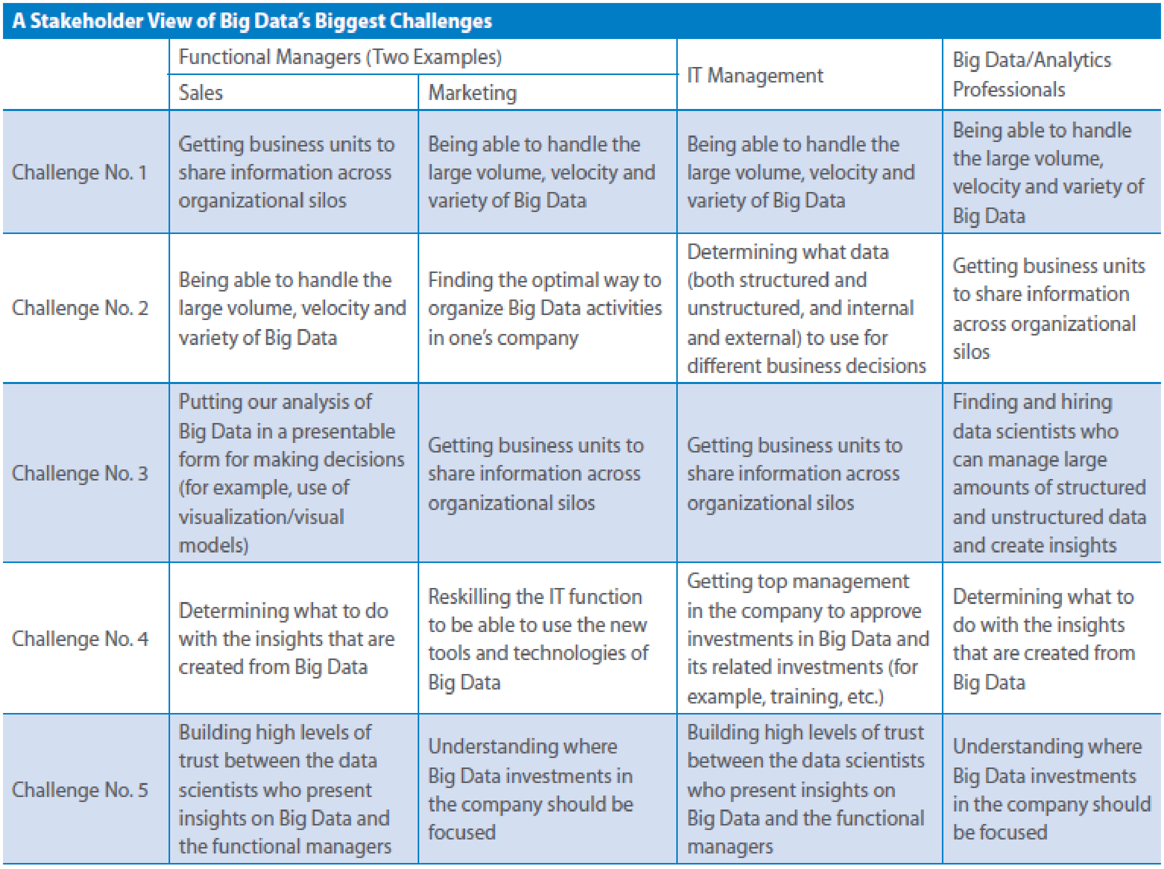
3.3 Data Curation: Data curation is aimed at data discovery and retrieval, data quality assurance, value addition, reuse and preservation over time. This field specifically involves a number of sub-fields including authentication, archiving, management, preservation, retrieval, and representation. The existing database management tools are unable to process Big Data that grow so large and complex. This situation will continue as the benefits of exploiting Big Data allowing researchers to analyze business trends, prevent diseases, and combat crime.

3.4 Data Analysis: The first impression of Big Data is its volume, so the biggest and most important challenge is scalability when we deal with the Big Data analysis tasks. In the last few decades, researchers paid more attentions to accelerate analysis algorithms to cope with increasing volumes of data and speed up processors following the Moore's Law. Real-time Big Data applications, like navigation, social networks, finance, biomedicine, astronomy, intelligent transport systems, and internet of thing, timeliness is at the top priority. How can we grantee the timeliness of response when the volume of data will be processed is very large? It is still a big challenge for stream processing involved by Big Data. It is right to say that Big Data not only have produced many challenge and changed the directions of the development of the hardware, but also in software architectures.

3.5 Data Visualization: For Big Data applications, it is particularly difficult to conduct data visualization because of the large size and high dimension of Big Data. However, current Big Data visualization tools mostly have poor performances in functionalities, scalability and response time. The main objective of data visualization is to represent knowledge more intuitively and effectively by using different graphs. To convey information easily by providing knowledge hidden in the complex and large-scale data sets, both aesthetic form and functionality are necessary.

3.6 Privacy and Security: Privacy in particular raises many concern as big data could be used to re-identify privacy-sensitive data even when this data has been anonymized [16]. Security also raises challenging issues including scalable security administration, management and integration of heterogeneous data security policies, and the security of data when hosted clouds.

TCS Big Data Global Trend Study 2013[14] comparing five Key challenges of Sales, Marketing, IT and Analytics Managers as shown in fig. 2.



A Stakeholder View of Big Data's Biggest Challenges

	Functional Managers (Two Examples)		IT Management	Big Data/Analytics Professionals
	Sales	Marketing		
Challenge No. 1	Getting business units to share information across organizational silos	Being able to handle the large volume, velocity and variety of Big Data	Being able to handle the large volume, velocity and variety of Big Data	Being able to handle the large volume, velocity and variety of Big Data
Challenge No. 2	Being able to handle the large volume, velocity and variety of Big Data	Finding the optimal way to organize Big Data activities in one's company	Determining what data (both structured and unstructured, and internal and external) to use for different business decisions	Getting business units to share information across organizational silos
Challenge No. 3	Putting our analysis of Big Data in a presentable form for making decisions (for example, use of visualization/visual models)	Getting business units to share information across organizational silos	Getting business units to share information across organizational silos	Finding and hiring data scientists who can manage large amounts of structured and unstructured data and create insights
Challenge No. 4	Determining what to do with the insights that are created from Big Data	Reskilling the IT function to be able to use the new tools and technologies of Big Data	Getting top management in the company to approve investments in Big Data and its related investments (for example, training, etc.)	Determining what to do with the insights that are created from Big Data
Challenge No. 5	Building high levels of trust between the data scientists who present insights on Big Data and the functional managers	Understanding where Big Data investments in the company should be focused	Building high levels of trust between the data scientists who present insights on Big Data and the functional managers	Understanding where Big Data investments in the company should be focused

Figure 2: Comparing Key Challenges of Sales, Marketing, IT and Analytics Managers

4. Tools and Techniques:

Organizations use many various techniques and technologies to aggregate, manipulate, analyze, and visualize Big Data. They come from various fields such as statistics, computer science, applied mathematics, and economics. Some of them have been developed intentionally and some of them have been adapted for this purpose.

To capture the value from Big Data, we need to develop new techniques and technologies for analyzing it. Until now, scientists have developed a wide variety of techniques and technologies to capture, curate, analyze and visualize Big Data.

4.1 Big Data Techniques:

Big Data needs extraordinary techniques to efficiently process large volume of data within limited run times. Reasonably, Big Data techniques involve a number of disciplines, including statistics, data mining, machine learning, neural networks, social network analysis, signal processing, pattern recognition, optimization methods and visualization approaches. There are many specific techniques in these disciplines, and they overlap with each other (illustrated as Figure 2).

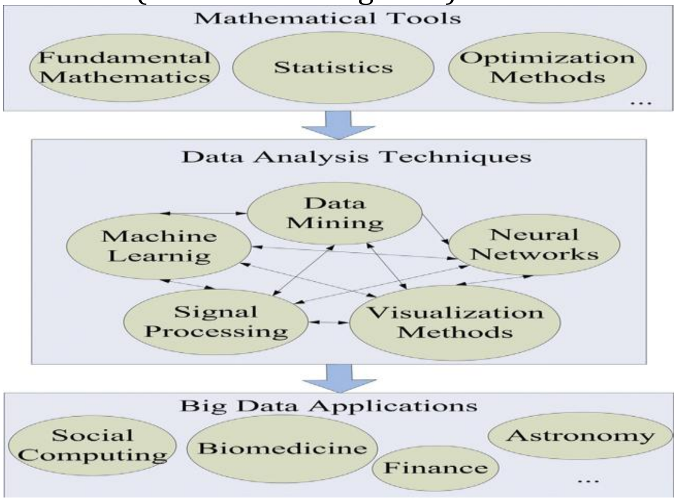


Figure 3: Big Data Techniques

- ✓ **Optimization Methods** have been applied to solve quantitative problems in a lot of fields, such as physics, biology, engineering, and economics.
- ✓ **Statistics** is the science to collect, organizes, and interprets data. Statistical techniques are used to exploit correlation ships and causal relationships between different objectives. Numerical descriptions are also provided by statistics. However, standard statistical techniques are usually not well suited to manage Big Data.
- ✓ **Data Mining** is a set of techniques to extract valuable information (patterns) from data, including clustering analysis, classification, regression and association rule learning.
- ✓ **Machine Learning** is an important subjection of artificial intelligence which is aimed to design algorithms that allow computers to evolve behaviors based on empirical data. The most obvious characteristic of machine learning is to discovery knowledge and make intelligent decisions automatically.
- ✓ **Artificial Neural Network** (ANN) is a mature technique and has a wide range of application coverage. Its successful applications can be found in pattern recognition, image analysis, adaptive control, and other areas.
- ✓ **Visualization Approaches** are the techniques used to create tables, images, diagrams and other intuitive display ways to understand data.

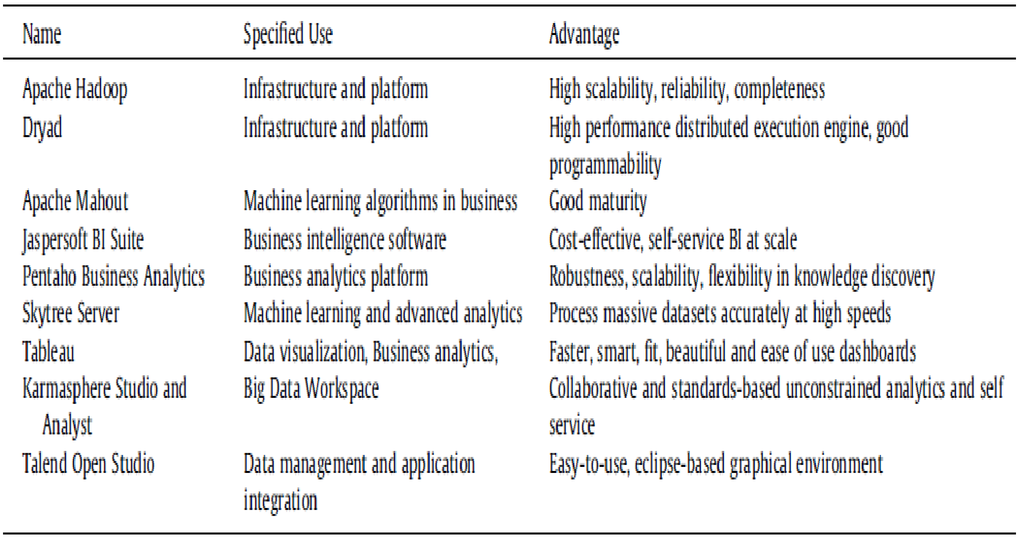
- ✓ **Social Network Analysis** (SNA) which has emerged as a key technique in modern sociology, views social relationships in terms of network theory, and it consists of nodes and ties.
- ✓ Higher level Big Data technologies include distributed file systems, distributed computational systems, massively parallel-processing (MPP) systems, data mining based on grid computing, cloud-based storage and computing resources, as well as granular computing and biological computing.

4.2 Big Data Tools:

Current Big Data tools concentrate on three classes, namely, batch processing tools, stream processing tools, and interactive analysis tools.

4.2.1 Big Data Tools Based on Batch Processing:

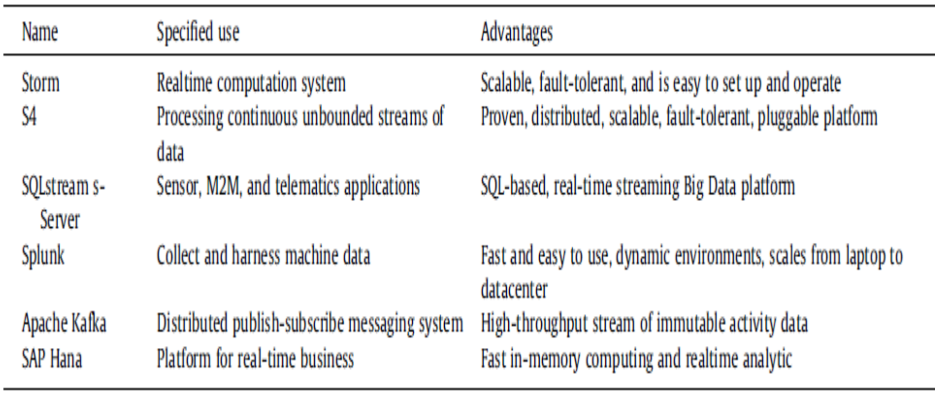
One of the most famous and powerful batch process-based Big Data tools is Apache Hadoop. It provides infrastructures and platforms for other specific Big Data applications [13].



Name	Specified Use	Advantage
Apache Hadoop	Infrastructure and platform	High scalability, reliability, completeness
Dryad	Infrastructure and platform	High performance distributed execution engine, good programmability
Apache Mahout	Machine learning algorithms in business	Good maturity
Jaspersoft BI Suite	Business intelligence software	Cost-effective, self-service BI at scale
Pentaho Business Analytics	Business analytics platform	Robustness, scalability, flexibility in knowledge discovery
Skytree Server	Machine learning and advanced analytics	Process massive datasets accurately at high speeds
Tableau	Data visualization, Business analytics,	Faster, smart, fit, beautiful and ease of use dashboards
Karmasphere Studio and Analyst	Big Data Workspace	Collaborative and standards-based unconstrained analytics and self service
Talend Open Studio	Data management and application integration	Easy-to-use, eclipse-based graphical environment

Table 1: Big Data tools based on batch processing

4.2.2 Stream Processing Big Data Tools:



Name	Specified use	Advantages
Storm	Realtime computation system	Scalable, fault-tolerant, and is easy to set up and operate
S4	Processing continuous unbounded streams of data	Proven, distributed, scalable, fault-tolerant, pluggable platform
SQLstream s-Server	Sensor, M2M, and telematics applications	SQL-based, real-time streaming Big Data platform
Splunk	Collect and harness machine data	Fast and easy to use, dynamic environments, scales from laptop to datacenter
Apache Kafka	Distributed publish-subscribe messaging system	High-throughput stream of immutable activity data
SAP Hana	Platform for real-time business	Fast in-memory computing and realtime analytic

Table 2: Big Data tools based on stream processing

Hadoop does well in processing large amount of data in parallel. It provides a general partitioning mechanism to distribute aggregation workload across different machines. Nevertheless, Hadoop is designed for batch processing. It is a multi-purpose engine but not a real-time and high performance engine, since there are high throughout latency in its implementations. Stream Big Data has high volume, high velocity and complex data types. Indeed, when the high velocity and time dimension are concerned in applications that involve real-time processing, there are a number of different challenges to Map/Reduce framework. Therefore, the real-time Big Data platforms, such as SQL stream, Storm and Stream Cloud, are designed especially for real-time stream data analytics.

4.2.3 Big Data Tools Based on Interactive Analysis:

The interactive analysis presents the data in an interactive environment, allowing users to undertake their own analysis of information. Users are directly connected to the computer and hence can interact with it in real time. The data can be reviewed, compared and analyzed in tabular or graphic format or both at the same time.

- ✓ **Google's Dremel** in 2010, Google proposed an interactive analysis system, named Dremel [16], which is scalable for processing nested data. Dremel has a very different architecture compared with well-known Apache Hadoop, and acts as a successful complement of Map/Reduce-based computations. It has capability to run aggregation queries over trillion-row tables in seconds by means of combining multi-level execution trees and columnar data layout.
- ✓ **Apache drill** is another distributed system for interactive analysis of Big Data. It is similar to Google's Dremel. For Drill, there is more flexibility to support a various different query languages, data formats and data sources.

5. Conclusion:

Big Data has the potential to revolutionize not just research, but also implementation practice and learning. There are frameworks like Hadoop, and mechanisms like NoSQL and new programming platforms to handle BD in development, testers are having a big time in finding optimized solutions, tools and frameworks to test the BD. The advanced techniques and technologies for developing Big Data science is with the purpose of advancing and inventing the more sophisticated and scientific methods of managing, analyzing, visualizing, and exploiting informative knowledge from large, diverse, distributed and heterogeneous data sets. The ultimate aims are to promote the development and innovation of Big Data sciences, finally to benefit economic and social evolutions. Big Data techniques and technologies should stimulate the development of new data analytic tools and algorithms and to facilitate scalable, accessible, and sustainable data infrastructure so as to increase understanding of human and social processes and interactions.

What was once called as "Garbage Data" is today termed as "Big Data". Nothing is wasted, nothing is deleted or removed. Everything is important for the business, for decision making and for the future of the organization. The future is not far; it is tomorrow or may be even today.

6. References:

1. E. Bertino, Data Protection from Insider Threats. Morgan & Claypool, 2012.
2. T. Murdoch, A. Detsky, "The Inevitable Application of Big Data to Health Care", JAMA, 2013, 309(13):1351-1352.
3. Stonebraker, M. and J. Hong. 2012."Researchers' Big Data Crisis; Understanding Design and Functionality", Communications of the ACM, 55(2):10-11
4. Mervis, J.2012. "Agencies Rally to Tackle Big Data", Science, 336(4):22, June 6, 2012.
5. S. Salmin, E. Bertino, "A Comprehensive Model for Provenance", Invited Paper, Proceedings of the First International Workshop on Modeling Data-Intensive Computing (MoDIC 2012), Florence, Italy, October 15-18, 2012, LNCS 7518, Springer.

6. http://campustechnology.com/articles/2013/12/04/survey-identifies-top-3-challenges-to-big-data-adoption.aspx
7. IDG Enterprise Big Data research 2014, available at http://www.idgenterprise.com/report/big-data-2.
8. The Asia HR Big Data Survey Report 2014, available at http://hrboss.com/asia-hr-big-data-survey-report-2014.
9. Dell Survey: Midmarket Companies Aggressively Embrace Big Data Projects 2014, available at http://www.dell.com/learn/us/en/uscorp1/press-releases/2014-04-28-dell-software-big-data-midmarket-survey.
10. Elisa, Bertino, Cyber Center, CERIAS and CS Department, Purdue University, West Lafayette, Indiana(USA).-" Big Data - Opportunities and Challenges", 37th Annual Computer Software and Applications Conference, 2013 IEEE.
11. JASON. 2008. "Data Analysis Challenges", The Mitre Corporation, McLean, VA, JSR-08-142
12. Jiawei Han, Micheline Kamber, Data Mining: Concepts and Techniques, Diane Cerra, second ed., 2000.
13. C.L. Philip Chen , Chun-Yang Zhang Department of Computer and Information Science, Faculty of Science and Technology, University of Macau, Macau, China - "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data"
14. TCS 2013 Global Trend Study on "The Emerging Big Returns on Big Data" available at http://sites.tcs.com/big-data-study/big-data-infographic-2/.
15. Stephen Kaisler, Frank Armour, J. Alberto Espinosa- American University, William Money George Washington University- "Big Data: Issues and Challenges Moving Forward"2013 46th Hawaii International Conference on System Sciences- 1530-1605/12 $26.00 DOI 10.1109/HICSS.2013.645, © 2012 IEEE.
16. Seref SAGIROGLU and Duygu SINANC, Gazi University, Department of Computer Engineering, Faculty of Engineering Ankara, Turkey- "Big Data: A Review"- 978-1-4673-6404-1/13/$31.00 ©2013 IEEE.
17. AvitaKatal, Mohammad Wazid, R H Goudar- Department of CSE, Graphic Era University, Dehradun, India. "Big Data: Issues, Challenges, Tools and Good Practices"- 978-1-4799-0192-0/13/$31.00 ©2013 IEEE.