Big Data: A Review

Seref SAGIROGLU and Duygu SINANC
Gazi University
Department of Computer Engineering, Faculty of Engineering
Ankara, Turkey
ss@gazi.edu.tr, duygusinanc@gazi.edu.tr

***Abstract*—Big data is a term for massive data sets having large, more varied and complex structure with the difficulties of storing, analyzing and visualizing for further processes or results. The process of research into massive amounts of data to reveal hidden patterns and secret correlations named as big data analytics. These useful informations for companies or organizations with the help of gaining richer and deeper insights and getting an advantage over the competition. For this reason, big data implementations need to be analyzed and executed as accurately as possible. This paper presents an overview of big data's content, scope, samples, methods, advantages and challenges and discusses privacy concern on it.**

Keywords—big data, volume, variety, velocity, verification, value;

I. INTRODUCTION

Big data and its analysis are at the center of modern science and business. These data are generated from online transactions, emails, videos, audios, images, click streams, logs, posts, search queries, health records, social networking interactions, science data, sensors and mobile phones and their applications [4,13]. They are stored in databases grow massively and become difficult to capture, form, store, manage, share, analyze and visualize via typical database software tools.

5 exabytes (10^{18} bytes) of data were created by human until 2003. Today this amount of information is created in two days. In 2012, digital world of data was expanded to 2.72 zettabytes (10^{21} bytes). It is predicted to double every two years, reaching about 8 zettabytes of data by 2015 [8]. IBM indicates that every day 2.5 exabytes of data created also 90% of the data produced in last two years [17]. A personal computer holds about 500 gigabytes (10^{9} bytes), so it would require about 20 billion PCs to store all of the world's data. In the past, human genome decryption process takes approximately 10 years, now not more than a week [22]. Multimedia data have big weight on internet backbone traffic and is expected to increase 70% by 2013[9]. Only Google has got more than one million servers around the worlds. There have been 6 billion mobile subscriptions in the world and every day 10 billion text messages are sent. By the year 2020, 50 billion devices will be connected to networks and the internet [3].

In 2012, The Human Face of Big Data accomplished as a global project, which is centering in real time collect, visualize and analyze large amounts of data. According to this media project many statistics are derived. Facebook has 955 million monthly active accounts using 70 languages, 140 billion photos uploaded, 125 billion friend connections, every day 30 billion pieces of content and 2.7 billion likes and comments have been posted. Every minute, 48 hours of video are uploaded and every day, 4 billion views performed on YouTube. Google support many services as both monitorizes 7.2 billion pages per day and processes 20 petabytes (10^{15} bytes) of data daily also translates into 66 languages. 1 billion Tweets every 72 hours from more than 140 million active users on Twitter. 571 new websites are created every minute of the day [23]. Within the next decade, number of information will increase by 50 times however number of information technology specialists who keep up with all that data will increase by 1.5 times. [5].

The article is worded as follows: Section II presents substantial issues, advantages, challenges, survey results, samples, methods and knowledge discovery from big data. In Section III, the important issues in security issues are reviewed. Section IV presents benefits, potential barriers, challenges and obstacles of big data. Section V concludes the work.

II. BIG DATA

Important issues have been reviewed and discussed in this section under different subsections.

A. Important Issues

Big Data requires a revolutionary step forward from traditional data analysis, characterized by its three main components: variety, velocity and volume as shown in Figure 1 [3,8,13,17].

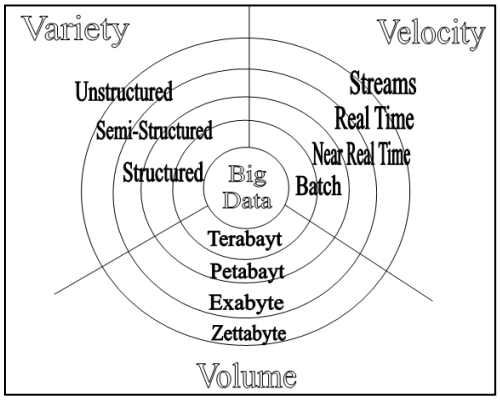

Figure 1.The three Vs of big data

Each component represented in Figure 1 is shortly expressed below.

Variety makes big data really big. Big data comes from a great variety of sources and generally has in three types: structured, semi structured and unstructured. Structured data inserts a data warehouse already tagged and easily sorted but unstructured data is random and difficult to analyze. Semi-structured data does not conform to fixed fields but contains tags to separate data elements [4,17].

Volume or the size of data now is larger than terabytes and petabytes. The grand scale and rise of data outstrips traditional store and analysis techniques [4,16].

Velocity is required not only for big data, but also all processes. For time limited processes, big data should be used as it streams into the organization in order to maximize its value [4,16].

During in the intensity of this information, another component is the verification of data flow. It is difficult to control large data so data security must be provided. In addition, after producing and processing of big data, it should create a plus value for the organization.

There are some questions and important answers summarized below that from the TDWI survey which is asked to the data management professionals [12].

- After the organization applied some form of big data analytics, these benefits occur: better aimed marketing, more straight business insights, client based segmentation, recognition of sales and market chances,
- While implementing big data analytics, these issues are potential barriers: inexpert stuff, cost, privation of business sponsorship, hard to designing analytic systems, lack of current database software in analytics
- Whereas significant crowd define big data now and in future is an opportunity because of exhaustive analytics, some of them see big data as problem because of managing
- Big data types that stored and using with advanced techniques today are: structured, semi structured, complex, event and unstructured data
- While replacing analytics platforms, these problems occur: cannot fit to big volumes of data, cannot support to needed analytic models, data loading is too slow, requirement of advanced analytics platform, IT cannot catch up with demands

As can be seen from the survey that big data analysis still needs more attention. Analyzing big data can require hundreds of servers running massively paralel software. That actually distinguishes big data, aside from its variety, volume and velocity, is the potential to analyze it to reveal new insights to optimize decision making.

B. Big Data Samples

Examples in the literature are available in are astronomy, atmospheric science, genomics, biogeochemical, biological science and research, life sciences, medical records, scientific research, government, natural disaster and resource management, private sector, military surveillance, private sector, 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 [20]. Organizations in any industry have big data can benefit from its careful analysis to gain insights and depths to solve real problems [8].

McKinsey Global Institute specified the potential of big data in five main topics [9]:

- Healthcare: clinical decision support systems, individual analytics applied for patient profile, personalized medicine, performance based pricing for personnel, analyze disease patterns, improve public health
- Public sector: creating transparency by accessible related data, discover needs, improve performance, customize actions for suitable products and services, decision making with automated systems to decrease risks, innovating new products and services
- Retail: in store behavior analysis, variety and price optimization, product placement design, improve performance, labor inputs optimization, distribution and logistics optimization, web based markets
- Manufacturing: improved demand forecasting, supply chain planning, sales support, developed production operations, web search based applications
- Personal location data: smart routing, geo targeted advertising or emergency response, urban planning, new business models

Web provides kind of opportunities for big data too. For example; social network analysis such as understanding user intelligence for more targeted advertising, marketing campaigns and capacity planning, customer behavior and buying patterns also sentiment analytics. According to these inferences firms optimization their content and recommendation engine [1]. Some companies such as Google and Amazon publishing articles related to their work. Inspired by the writings published, developers are developing similar technologies as open source software such as Lucene, Solr, Hadoop and HBase. Facebook, Twitter and LinkedIn are going a step further thereby publishing open source projects for big data like Cassandra, Hive, Pig, Voldemort, Storm, IndexTank.

In addition, predictive analytics on traffic flows or identify guilties and threats from different video, audio and data feeds are advantages of big data again [3].

In 2012, Obama regime announced big data initiatives of more than $200 million in research and development investments for National Science Foundation, National Institutes of Health, Department of Defense, Department of Energy and United States Geological Survey. The investments were launched to take a step forward instruments and methods for access, organize and collect findings from vast volumes of digital data [14].

C. Methods

Most enterprises are facing lots of new data, which arrives in many different forms. Big data has the potential to provide insights that can transform every business. Big data has generated a whole new industry of supporting architectures such as MapReduce. MapReduce is a programming framework for distributed computing which was created by Google using the divide and conquer method to break down complex big data problems into small units of work and process them in parallel [13]. MapReduce can be divided into two stages [10]:

- Map Step: The master node data is chopped up into many smaller subproblems. A worker node processes some subset of the smaller problems under the control of the JobTracker node and stores the result in the local file system where a reducer is able to access it.
- Reduce Step: This step analyzes and merges input data from the map steps. There can be multiple reduce tasks to parallelize the aggregation, and these tasks are executed on the worker nodes under the control of the JobTracker.

Hadoop created to inspire by BigTable which is Google's data storage system, Google File System and MapReduce [6]. Hadoop is Java based framework and heterogeneous open source platform. It is not a replacement for database, warehouse or ETL (Extract, Transform, Load) strategy. Hadoop includes a distributed file system, analytics and data storage platforms and a layer that manages parallel computation, workflow and configuration administration [8,22]. It is not designed for real time complex event processing like streams. HDFS (Hadoop Distributed File System) runs across the nodes in a Hadoop cluster and connects together the file systems on many input and output data nodes to make them into one big file system [4,13,19].

As seen Fig. 1 and Fig. 2, Hadoop offers [21]:

- HDFS: A highly fault tolerant distributed file system that is responsible for storing data on the clusters.
- MapReduce: A powerful parallel programming technique for distributed processing on clusters.
- HBase: A scalable, distributed database for random read/write access.
- Pig: A high level data processing system for analyzing data sets that occur a high level language.
- Hive: A data warehousing application that provides a SQL like interface and relational model.
- Sqoop: A project for transferring data between relational databases and Hadoop.
- Avro: A system of data serialization.
- Oozie: A workflow for dependent Hadoop jobs.
- Chukwa: A Hadoop subproject as data accumulation system for monitoring distributed systems.
- Flume: A reliable and distributed streaming log collection.
- ZooKeeper: A centralized service for providing distributed synchronization and group services

HPCC (High Performance Computing Cluster) Systems distributed data intensive open source computing platform and provides big data workflow management services. Unlike Hadoop, HPCC's data model defined by user. The key to complex problems can be stated easily with high level ECL basis. HPCC ensure that ECL is executed at the maximum elapsed time and nodes are processed in parallel. Furthermore HPCC Platform does not require third party tools like GreenPlum, Cassandra, RDBMS, Oozie, etc [22].

The three main HPCC components are [22]:

- HPCC Data Refinery (Thor) is a massively parallel ETL engine that enables data integration on a scale and provides batch oriented data manipulation.
- HPCC Data Delivery Engine (Roxie) is a massively parallel, high throughput, ultra fast, low latency, allows efficient multi user retrieval of data and structured query response engine.
- Enterprise Control Language (ECL) is automatically distributes workload between nodes, has automatic synchronization of algorithms, develop extensible machine learning library, has simple usage programming language optimized for big data operations and query transactions.

Figure 2 indicates comparisons between HPCC Systems Platform and Hadoop in terms of architecture and stacks. According to reference [22], some differences summarized below:

- HPCC clusters can be exercised in Thor and Roxie. Hadoop clusters perform with MapReduce processing.
- In HPCC environments ECL is primary programming language. However, Hadoop MapReduce processes are based on Java language.
- HPCC platform builds multikey and multivariate indexes on Distributed File System. Hadoop HBase procures column oriented database.
- Data warehouse abilities used in HPCC Roxie for structural queries and analyzer applications on the other hand Hadoop Hive provide data warehouse abilities and allow data to be loaded into HDFS.
- On the same hardware configuration a 400-node system, HPCC success is 6 minutes 27 seconds and Hadoop success is 25 minutes 28 seconds. This result showed that HPCC faster than Hadoop for this comparison.

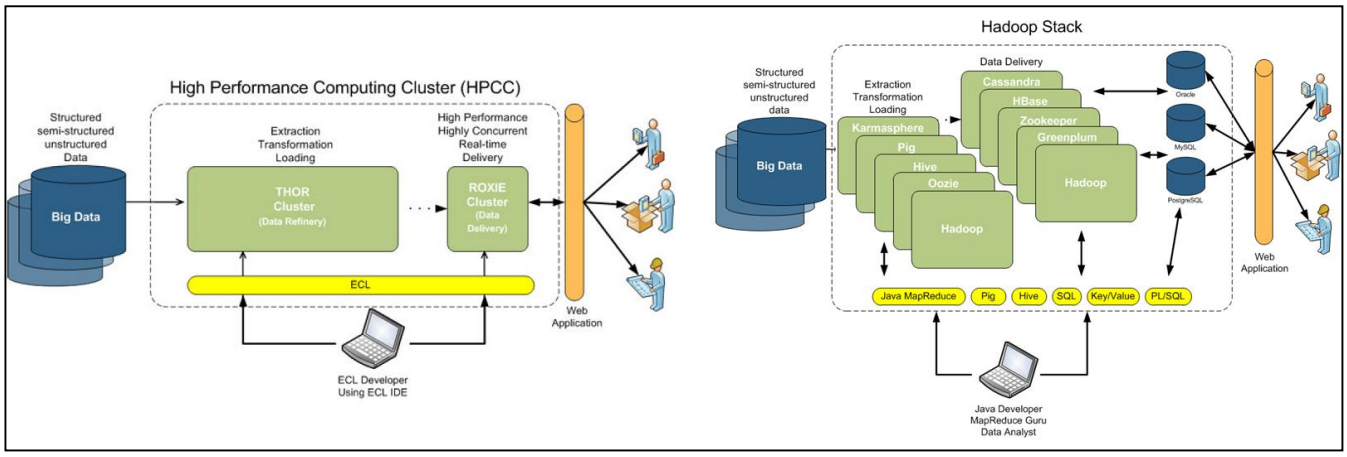

Figure 2. Comparison between HPCC Systems Platform and Hadoop architecture [22]

D. Knowledge Discovery from Big Data

Knowledge Discovery from Data (KDD) entitle as some operations designed to get information from complicated data sets [6]. Reference [18] outlines the KDD at nine steps:

1. Application domain prior to information and defining purpose of process from customer's perspective.
2. Generate subset data point for knowledge discovery.
3. Removing noise, handling missing data fields, collecting required information to model and calculating time information and known changes.
4. Finding useful properties to present data depending on purpose of job.
5. Mapping purposes to a particular data mining methods.
6. Choose data mining algorithm and method for searching data patterns.
7. Researching patterns in expressional form.
8. Returning any steps 1 through 7 for iterations also this step can include visualization of patterns.
9. Using information directly, combining information into another system or simply enlisting and reporting.

Reference [6] analyzes knowledge discovery from big data in three principles using Hadoop. These are:

1. KDD includes a variety of analysis methods as distributed programming, pattern recognition, data mining, natural language processing, sentiment analysis, statistical and visual analysis and human computer interaction. Therefore architecture must support various methods and analysis techniques.
    - Statistical analysis interested in summarizing massive datasets, understanding data and defining models for prediction.
    - Data mining correlate with discovering useful models in massive data sets by itself, machine learning combine with data mining and statistical methods enabling machines to understand datasets.
    - Visual analysis is developing area in which large datasets are serviced to users in challenging ways will be able to understand relationships.
2. A comprehensive KDD architecture must procure to keep and operate process line.
    - Preparation of data and batch analytics are made, for proper troubleshooting with errors, missing values and unusable format.
    - Processing structured and semi structured data
3. It is cardinal that making results accessible and foolproof. For this reason following approaches are used to overcome this issue.
    - Using open source and popular standards
    - Use WEB based architectures
    - Publicly available results

III. PRIVACY AND SECURITY ISSUES

In May 2012, Intel IT Center surveyed 200 IT managers in large companies to find out how they were approaching big data analytics [7]. They asked IT managers what standards they would like to see addressed for big data analytics and the answers were: data security, technology to keep customers' data private, data transparency, performance benchmarking, data and system interoperability. There were answers concerns via third party cloud vendors regarding; data security and privacy concerns, company policy prevents me from outsourcing data storage and analytics, overall costs and I'm doing my data management/ analytics in house don't plan to outsource. According the survey apprehensions ordinarily about security.

The ruining of traditional defensive environments united with attackers' abilities to survive traditional security systems requires organizations to adopt an intelligence driven security model that is more risk aware, contextual and agile. Intelligence driven security relies on big data analytics. Big data involve both the breadth of sources and the information depth needed for programs to specify risks accurately, to defend against illegal activity and advanced cyber threats. A big data driven security model has the following characteristics [15] :

- Internal and external data sources that multiply in value and create a synergistic learning effect.
- Automated tools that collect diverse data types and normalize them.
- Analytics engines manage to process massive volumes of fast changing data in real time.
- Advanced monitoring systems that continuously analyse high value systems, resources and make considerations based on behavior and risk models.
- Active controls such as need additional user authentication, blocking data transfers or simplification analysts' decision making.
- Centralized warehouse where all security related data is made available for security analysts to query.
- Standardized views into demonstrations of compromise that are created in machine readable form and can be shared at scale by trusted sources.
- N-tier infrastructures that create scalability across vectors and have ability to process large and complex searches and queries.
- High degree of integration via security and risk management tools to facilitate detailed investigations of potential problems.

Reference [5] states how developing a holistic and confident approach for big data is:

- To begin of a management project, companies need to place and describe data sources origination, created and access authorizations.
- To categorize discovered as its importance.
- To guarantee that records are archived and protected according to standards and regulations.
- To develop policies related data processing, such as defining stored data types, store time, storehouse and data accessed types.

By keeping data in one place, it occurs a target for attackers to sabotage the organization. It required that big data stores are rightly controlled. To ensure authentication a cryptographically secure communication framework has to be implemented. Controls should be using principle of reduced privileges, especially for access rights, except for an administrator who have permission data to physical access. For effective access controls, they should be continuously observed and switched as change employees organization roles so employees do not aggregate immoderate rights that could be misused. Other security procedures are needed to capture and analyze network traffic such as metadata, packet capture, flow and log information.. Organizations should guaranteed investments in security products using agile technologies based analytics not static equipments. Another problem is associated with organizing compliance of data protection laws. Organizations have to consider legal branching for storing data [5,11,15].

However, big data have security advantages. When organizations categorize knowledge, they control data according to specified by the regulations such as imposing store periods. This allows organizations to select data that has neither little value nor any need to be kept so that it is no longer available for theft. Another benefit is massive data can be mined for threats such as evidence of malware, anomalies, or phishing [5].

IV. Overall Evaluation

The amount of data has been increasing and data set analyzing become more competitive. The challenge is not only to collect and manage vast volume and different type of data, but also to extract meaningful value from it [10]. Also needed, managers and analysts with an excellent insight of how big data can be applied. Companies must accelerate employment programs, while making significant investments in the education and training of key personnel [2].

Through TDWI Big Data Analytics survey, benefits of big data are: better aimed marketing, more straight business insights, client based segmentation, recognition of sales and market chances, automated decision making, definitions of customer behaviors, greater return on investments, quantification of risks and market trending, comprehension of business alteration, better planning and forecasting, identification consumer behavior from click streams and production yield extension [12].

In addition, TDWI array potential barriers to implementing big data analytics like: inexpert stuff and cannot find to hire big data experts, cost, privation of business sponsorship, hard to designing analytic systems, lack of current database software in analytics and fast process time, scalability problems, incapable to make big data usable for end users, data load cannot fast enough in current database software, lack of compelling business case [12].

According to the, Intel IT Center Big Data Analytics survey, there are several challenges for big data: data growth, data infrastructure, data governance/policy, data integration, data velocity, data variety, data compliance/regulation and data visualization [7].

In addition, Intel IT Center specify obstacles of big data as: security concerns, capital/operational expenses, increased network bottlenecks, shortage of skilled data science professionals, unmanageable data rate, data replication capabilities, lack of compression capabilities, greater network latency and insufficient CPU power [7].

In spite of potential barriers, challenges and obstacles of big data, it has great importance today and in the future.

V. CONCLUSION

In this article, an overview of big data's content, scope, samples, methods, advantages and challenges and discusses privacy concern have been reviewed. The results have shown that even if available data, tools and techniques available in the literature, there are many points to be considered, discussed, improved, developed, analyzed, etc. Besides, the critical issue of privacy and security of the big data is the big issue will be discussed more in future.

Although this paper clearly has not resolved the entire subject about this substantial topic, hopefully it has provided some useful discussion and a framework for researchers.

REFERENCES

[1] A. Vailaya, "What's All the Buzz Around "Big Data?"", IEEE Women in Engineering Magazine, December 2012, pp. 24-31,

[2] B. Brown, M. Chui and J. Manyika, "Are you Ready for the era of 'Big Data'? " McKinsey Quarterly, McKinsey Global Institute, October 2011

[3] B.Gerhardt, K. Griffin and R. Klemann, "Unlocking Value in the Fragmented World of Big Data Analytics", Cisco Internet Business Solutions Group, June 2012,
http://www.cisco.com/web/about/ac79/docs/sp/Information-Infomediaries.pdf

[4] C. Eaton, D. Deroos, T. Deutsch, G. Lapis and P.C. Zikopoulos, Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data, Mc Graw-Hill Companies, 978-0-07-179053-6, 2012

[5] C. Tankard, "Big Data Security", Network Security Newsletter, Elsevier, ISSN 1353-4858, July 2012

[6] E. Begoli and J. Horey, "Design Principles for Effective Knowledge Discovery from Big Data", Software Architecture (WICSA) and European Conference on Software Architecture (ECSA) Joint Working IEEE/IFIP Conference on, Helsinki, August 2012

[7] Intel IT Center, "Peer Research: Big Data Analytics", Intel's IT Manager Survey on How Organizations Are Using Big Data, August 2012,
http://www.intel.com/content/dam/www/public/us/en/documents/reports/data-insights-peer-research-report.pdf

[8] Intel IT Center, "Planning Guide: Getting Started with Hadoop", Steps IT Managers Can Take to Move Forward with Big Data Analytics, June 2012
http://www.intel.com/content/dam/www/public/us/en/documents/guides/getting-started-with-hadoop-planning-guide.pdf

[9] 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, 2011, http://www.mckinsey.com/~/media/McKinsey/dotcom/Insights%20and%20pubs/MGI/Research/Technology%20and%20Innovation/Big%20Data/MGI_big_data_full_report.ashx

[10] K. Bakshi, "Considerations for Big Data: Architecture and Approach", Aerospace Conference IEEE, Big Sky Montana, March 2012

[11] M. Smith, C. Szongott, B. Henne and G. Voigt , "Big Data Privacy Issues in Public Social Media", Digital Ecosystems Technologies (DEST), 6th IEEE International Conference on, Campione d'Italia, June 2012

[12] P. Russom, "Big Data Analytics ", TDWI Best Practices Report, TDWI Research, Fourth Quarter 2011,
http://tdwi.org/research/2011/09/best-practices-report-q4-big-data-analytics/asset.aspx

[13] R.D. Schneider, Hadoop for Dummies Special Edition, John Wiley&Sons Canada, 978-1-118-25051-8, 2012

[14] R. Weiss and L.J. Zgorski, "Obama Administration Unveils "Big Data" Initiative:Announces $200 Million in new R&D Investments", Office of Science and Technology Policy Executive Office of the President, March 2012

[15] S. Curry, E. Kirda, E. Schwartz, W.H. Stewart and A. Yoran, "Big Data Fuels Intelligence Driven Security", RSA Security Brief, January 2013
http://www.emc.com/collateral/industry-overview/big-data-fuels-intelligence-driven-security-io.pdf

[16] S. Madden, "From Databases to Big Data", IEEE Internet Computing, June 2012, v.16, pp.4-6

[17] S. Singh and N. Singh, "Big Data Analytics", 2012 International Conference on Communication, Information & Computing Technology Mumbai India, IEEE, October 2011

[18] U. Fayyad, G. Piatetsky-Shapiro and P. Smyth, "From Data Mining to Knowledge Discovery in Databases", American Association for Artificial Intelligence, AI Magazine, Fall 1996, pp. 37- 54

[19] V. Borkar, M.J. Carey and C. Li, "Inside "Big Data Management": Ogres, Onions, or Parfaits?", EDBT/ICDT 2012 Joint Conference Berlin Germany, 2012

[20] http://en.wikipedia.org/wiki/Big_data , last access 11.03.2013

[21] http://hadoop.apache.org/ , last access 11.03.2013

[22] http://hpccsystems.com/ , last access 11.03.2013

[23] http://www.humanfaceofbigdata.com/ , last access 11.03.2013