International Journal of Electrical and Computer Engineering (IJECE)
Vol. 5, No. 2, April 2015, pp. 340~343
ISSN: 2088-8708

Big Data: Challenges, Opportunities and Cloud Based Solutions

Hamid Bagheri *, Abdusalam Abdullah Shaltooki**
* University of Kurdistan, Iran
** University of Human Development, Iraq

Article Info

Article history:

Received Nov 20, 2014
Revised Dec 31, 2014
Accepted Jan 22, 2015

Keyword:

Big Data analytics
Cloud computing
Hadoop
MapReduce

ABSTRACT

We are living in an era of information explosion. There are challenges with large and complex amount of data generated every day by social networks, wikis, blogs, emails, traffic system, bridges, airplanes and engine, satellites and weather sensors. 90% of current data in the world has been created in the last two years. Our smart planet becomes more and more intelligent. Besides the challenges posed by such vast amount of data including storage, search, sharing, analysis, and visualization, there are also much opportunities for the world as it becomes more and more digitalized. This study presents Big Data and highlights its key concepts and state-of-the-art implementation as well as research challenges and suggests research directions for future. IT log analytics, Fraud detection pattern, social media pattern and modeling and management patterns are some of opportunities. Hadoop is a cloud based and open source solution for Big Data Analytics which has been written by java. Hadoop solution is currently still immature. In this paper, three topics are suggested for research direction: Security issues in Big Data, context-aware information retrieval, and integrating ontology with Big Data.

Corresponding Author:

Hamid Bagheri,
Information Technology,
Kurdistan University,
Iran, Kurdistan, Sanandaj, Pasdaran street.
Email: h.bagheri@uok.ac.ir

1. INTRODUCTION

We are living in an era of information explosion which large scale amount of data is getting increasingly larger because of virtual worlds, wikis, blogs, e-mail, online games, VoIP telephone, digital photos, instant messages (IM), tweets, traffic system, bridges, airplanes and engine, satellites, weather sensors. 90% of current data in the world has been created in the last two years [1]. There are challenges of managing such vast amount of heterogeneous data for example data variety and volume and analytical complexity. Big data analytics has been grown in the last years [2-3].

FaceBook accumulates huge amounts of data with about 800 million users and billions of page views every day which cause many challenges to storing and processing all these data. FaceBook needs analytics tool to mine and manipulate large amount of data (about 15 terabytes) every day in different languages, different times, from different locations and from different platforms.

In this section big data characteristic, four types of analytics and opportunities to create business value will be discussed.

1.1. Big Data characteristic

Most definitions of big data focus on the size of data in storage but there are other important attributes of big data: data variety and data velocity [2]. These three Vs of big data (Volume, Variety and Velocity) are shown in figure 1.

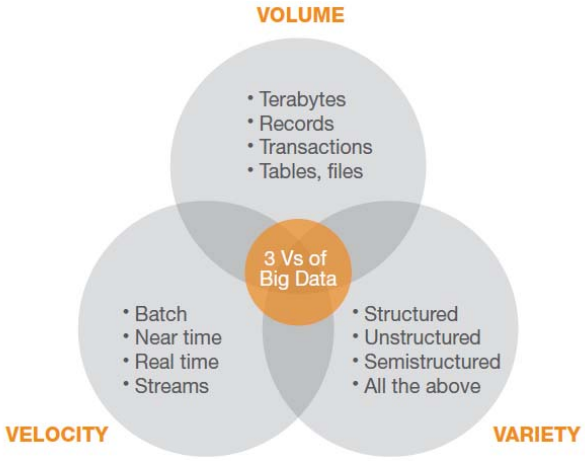

Figure 1. Characteristics of Big Data [1]

There is variety of sources in big data for example web sources including social media and logs which make it complex. Unstructured data (for example audio, video, Text) and semi-structured data (for instance XML, RSS feed) is now joined with structured data. According to Microsoft Over 85 percent of data captured is unstructured [4]. Velocity or speed for example video camera scanning in a crowd for recognizing specific face is another characteristic of big data.

1.2. Four Types of Analytics

New analytics application for example Video and audio application are needed to process streaming big data. Sound monitor to predict earthquakes and satellite images to recognize cloud patterns are another example which have to be analyzed.

The term "analytics" has four types [3]: Quantitative Research and Development, Data Scientists, Operational Analytics, and Business Intelligence and Discovery. By putting big data and analytics together we will discover most significant results in business value. In the next section big data is shown as a special asset that produces big opportunities for business.

1.3. Smarter and Intelligent Planet: Big Data Opportunities

Besides the challenges posed by such vast amount of data (Big Data), there are also much opportunities for the world as it becomes more and more digitalized. For example, information derived from digital records can make doctors' job easier in accurately diagnosing and treating illnesses, and bring down healthcare costs for patients, and the overall quality and efficiency of healthcare will be improved [5-7].

Our smarter planet has become more and more intelligent. There are some big opportunities that derive from big data [6]: IT Log analytics, Fraud detection pattern, social media pattern and Modelling and management patterns.

2. CURRENT STATE AND RELEVANT TOPICS

Big Data technologies are not a replacement for current technologies; they are a complement [6]. Big Data must be integrated with the rest of enterprise infrastructure. Besides the current solutions for big data analysis, there are some new challenges, for instance need for robust statistical methods and managing missing data [8-10]. As mentioned in section 1 we have unstructured and structured data; integration between them is another challenge [11]. In this section big data analytics solution, tools and techniques will be reviewed.

2.1. Cloud Based Big Data Solution

Cloud computing provides new capabilities for performing analysis across all data in an organization. It uses new technical approaches to store, search, mine and distribute massive amounts of data [6]. Problems such as large-scale image processing, sensor data correlation, social network analysis, encryption/decryption, data mining, simulations, and pattern recognition can be solved in the cloud computing domain.

To cope with problem mentioned above about FaceBook, Cloud allows Facebook to leverage more than 8,500 Central Processing Unit (CPU) cores and petabytes of disk space to create rich data analytics on a wide range of business characteristics.

New cloud computing technologies such as Hadoop, MapReduce and BigTable are driving analytic transformation in the way organizations store, access and process massive amounts of disparate data via massively parallel and distributed IT systems. Cloud application architectures are based on two principles:

- Elasticity: only use computing resources when needed.
- Scalability: highly elastic infrastructure to response changing condition such as data volumes.

Researches are driving across the cloud ecosystem for some reasons. First of all, cloud produces the new analytic capabilities of big data. Second, it provides massively scalable analytics and third reason is all facilities listed above are combined with the security and financial advantages of switching to a cloud computing environment.

There are challenges when dealing with big data of volumes greater than 10 terabytes. Although relational database models are capable of running in a Data Cloud, many current relational database systems fail in the Data Cloud in two important ways:

- Many relational database systems cannot scale to support petabytes or greater amounts of data storage.
- When complex data is normalized into a relational table format impedance mismatch happens. When data is collected, often the first step is to transform the data, normalize the data, and insert a row into a relational database. Next, users query data based on keywords or pre-loaded search queries and wait for the results to return. Once returned, users sift through results.

2.2. Hadoop: Open Source Heart of Big Data and Cloud Oriented Approach

Hadoop is a Top level Apache project open sourcesoftware framework that's written in java programming language [9]. It enables applications to work with thousands of computational independent computers and petabytes of data. Hadoop was derived from Google's MapReduce and Google File System (GFS). Hadoop has two parts: a file system (Hadoop Distributed File System or HDFS) and a programming paradigm (MapReduce). Tasks such as sorting, data mining, image manipulation, social network analysis, inverted index construction and machine learning are prime jobs for MapReduce.

HDFS is a distributed, scalable, and portable file system written in Java for the Hadoop framework. HDFS stores large files across multiple machines. It achieves reliability by replicating the data across multiple hosts. HDFS was designed to handle very large files.

Forrester regards Hadoop as the most significant part of the next-generation Enterprise Data Warehousing (EDW) in the cloud [7]. Hadoop implements the core features that are at the heart of most modern EDWs: cloud-facing architectures, in-database analytics, mixed workload management and a hybrid storage layer.

3. PROPOSED RESEARCH DIRECTION

In this section three research directions are proposed:

3.1. Suggested Topic 1: Context-Aware Information Retrieval (IR)

Search and Retrieval with a huge amount of structured and unstructured data are affected by Context in many ways. For example information in big data considered as a something dynamic over time and changing circumstances. For supporting these dynamic situation context must be applied to search and IR and new framework should be applied.

3.2. Suggested Topic 2: Big Data Security Challenges

Today, database management systems only support security policies at fine grain level [12] from inappropriate access; while due to the less structured and informal nature of big data current software has no such safeguards.

The future of big data will be in the cloud but these solutions also come with some challenges such as security. In Big Data Analysis on cloud, some researches about Access Control, encryption for tackling security problem and enforcing security policies must be done. For defining new models and methods we can follow "Data Security as a Service (DaS) "approach.

3.3. Suggested Topic 3: Integrating Ontology with Big Data Analytics

With a huge amount of data collected by Web 2.0, there is another field for research. Ontology is the structural framework for organizing information. Today, big data is not just about size of data. The most

important interest is digging and analyzing unstructured data. For taking advantages of opportunities mentioned in previous section, Big Data might benefit from ontology technology and Ontology-based analysis.

4. CONCLUSION

There are challenges with large and complex amount of data generated every day by so many different sources and from different platforms. According to [1] about 90% of world's data has been created in the last two years. Our smart planet becomes more and more intelligent. Besides the challenges posed by such vast amount of data, there are also much opportunities for the world as it becomes more and more digitalized. This study presents Big Data and highlights its key concepts and current approaches as well as research challenges and suggests three research directions for future. IT log analytics, Fraud detection pattern, social media pattern and modeling and management patterns are some of opportunities. Hadoop is a cloud based and open source solution for Big Data Analytics which is still immature. In this paper, three topics are suggested for research direction: Security issues in Big Data, context-aware information retrieval, and integrating ontology with Big Data.

REFERENCES

[1] Big Data, for better or worse: 90% of world's data generated over last two years. ScienceDaily. Retrieved August 23, 2013, from http://www.sciencedaily.com- /releases/2013/05/130522085217.htm
[2] T Sutikno, D Stiawan, IMI Subroto. Fortifying Big Data infrastructures to Face Security and Privacy Issues. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*. 2014; 12(4): 751-752.
[3] Neil Raden, Big Data Analytics Architecture, 2012, Hired Brains, Inc
[4] Microsoft Big Data, solution brief www.microsoft.com
[5] Michael Farber, MikeCameron, Christopher Ellis , Massive Data Analytics and cloud, Booz Allen Inc, 2011
[6] C.Zikopoulis, Ch.Eaton, D.deRoos, Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data, The McGraw-Hill Companies, 2012.
[7] The Forrester Wave™: Enterprise Hadoop Solutions, Q1 2012
[8] H. Demirkan, D. Delen, Leveraging the capabilities of service-oriented decision support systems: Putting analytics and big data in cloud, Decis. Support Syst. (2012), doi:10.1016/j.dss.2012.05.048
[9] http://hadoop.apache.org/
[10] Jianqing Fan, Han Liu, Statistical Analysis of Big Data on Pharmacogenomics, *Advanced Drug Delivery Reviews* (2013), doi: 10.1016/j.addr.2013.04.008
[11] Managing Data in Motion: Data Integration Best Practice Techniques and Technologies, First Edition (2013) 125-128. doi:10.1016/B978-0-12-397167-8.00018-2
[12] Big Data: What It Is and Why You Should Care. Richard L. Villars. Carl W. Olofson. Matthew Eastwood. June 2011

BIOGRAPHIES OF AUTHORS

Hamid Bagheri received the B.S. and M.S. degrees in Software engineering from the Shahid Beheshti University, Tehran, in 2011.
Since 2009, he has been working in Information Technology in Kurdistan University. His research interests include Service Oriented Arhcitectureand, Big Data and Ultra large Scale Systems.

Abdusalam Abdullah Shaltooki received his M.S. degree in software engineering from the Sulaimniah University, in 2011. His research interests include software Engineering and Big data Analysis. He is working as a lecturer in the University of Human Development, Iraq.