ISSN:2321-1156

International Journal of Innovative Research in Technology & Science(IJIRTS)

BIG DATA – OPPORTUNITIES AND CHALLENGES

Nguyễn Thị Thúy Hoài, College of technology _ Danang University

Abstract

The threading development of IT has been bringing more challenges for administrators to collect, store and analyze massive amounts of data. This data is commonly referred as "big data" which is creating new generation of decision support data management. Big data is the potential value for business and creating new jobs. In this article, the author focuses on studying the overview about big data including its definition, benefit, challenges, and pipeline.

Introduction

Nowadays, big data are becoming a new technology focus both in science and in industry. Due to such large size of data, it requires new technologies and architectures in order to extract value from it by capturing and analysis process. However, it is difficult to perform effective analysis using the exiting traditional techniques and handle such large amount of data that is growing at a huge speed. Thus, Big data can bring huge benefits to the business organizations and become relates to almost aspects of human activity from just recording events to research, design, production and digital services or products delivery to final consumer. This paper introduces the Big data technology along with its benefits, challenges and importance in the modern world, and pipeline.

The paper is organized as follows. Section II investigates Big Data and Big Data analysis definition. Section III presents the benefits and challenges of Big Data. In section IV, the author illustrates the pipeline of Big data. The paper concludes with the summary and suggestions for further research.

Overview Big Data

Definition

Big Data:In 2001 research report, Big data is defined as being three characters: Volume, Velocity, and Variety. This is the most venerable and well-known definition to describe big data. Gartner, Inc. defines big data in similar terms:

"Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making." (Gartner IT Glossary, n.d.)
Similarly, TechAmerica Foundation defines big data as follows:

"Big data is a term that describes large volumes of high velocity, complex and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management, and analysis of the information." (TechAmerica Foundation's Federal Big Data Commission, 2012)

However, later 2012, Gartner updated the definition of big data by adding a few more properties such as High Variability, High complexity, High Value.

The characters of big data below can describe as follows:

High Volume: refers to the amount or quantity of data. Because of big data has a large size, they storage in multiple terabytes and petabytes. Through a survey conducted by IBM in mid-2012, it is amazed that just over half of the 1144 respondents considered datasets over one terabyte to be big data (Schroeck, Shockley, Smart, Romero-Morales, & Tufano, 2012).

Time and the type of data are two factors that make effect on the volume of big data.

With storage capacities are more and more increase, what will happen if deemed big data today may not meet the threshold in the future? Moreover, the type of data, discussed under variety, defines what is meant by 'big'. Basing on type of data, two datasets of the same size may require different data management technologies, e.g., tabular versus video data.

High Velocity: refers to the rate at which data is created. The digital devices such as smartphones and sensors have led to a growing need for real-time analytics and evidence-based planning. Traditional data management systems are not capable of handling huge data feeds instantaneously when retailers require dealing with hundreds of thousands of streaming data sources that demand real-time analytics. This is where big data technologies come into play. They enable firms to create real-time intelligence from high volumes of 'perishable' data.

High Variety: refers to the different types of data such as structured, semi-structured, and unstructured data.

Spreadsheets or relational databases are examples of structured data. Unstructured data can be text, images, audio or video, which sometimes lack the structural organization required by machines for analysis. Extensible Markup Language (XML), a textual language for exchanging data on the Web, is a typical example of semi-structured data. XML documents contain user-defined data tags that make them machine-readable.

Table 1. Types of data

Types of data	Describe
Structured data	- data is organized in semantic chunks (entities) - similar entities are grouped together (relations or classes) - entities in the same group have the same descriptions (attributes) - descriptions for all entities in a group (schema) + have the same defined format + have a predefined length + are all present + and follow the same order
Semi-structured data	- idea predates XML but not HTML - data is available electronically in + database systems + file systems, e.g., bibliographic data, Web data + data exchange formats, e.g., EDI, scientific data - attempt to reconcile database and document "worlds" - semi-structured data + organized in semantic entities + similar entities are grouped together + entities in same group may not have same attributes + order of attributes not necessarily important + not all attributes may be required + size of same attributes in a group may differ + type of same attributes in a group may differ
Unstructured data	- data can be of any type - not necessarily following any format or sequence - does not follow any rules - is not predictable - examples include + text + video + sound + images

High Variability: In addition to the increasing velocities and varieties of data, data flows can be highly inconsistent with periodic peaks. It is more difficult to load data especially with the increase in usage of social media. Thus, seasonal and event-triggered peak data loads can be challenging to manage. Even more so with unstructured data involved.

High Complexity: Today's data comes from multiple sources and it needs to link, match, cleanse and transform data across systems. However, it is necessary to connect and correlate relationships, hierarchies and multiple data linkages or your data can quickly spiral out of control.

High Value: The potential value of Big Data is huge. All that available data will create a lot of value for organizations, societies and consumers. Big business and every industry will reap the benefits from filtered data obtained and can rank it according to the dimensions they require. Of course, data in itself is not valuable at all. Based on the analyses and process done on that data, the data is turned into information and eventually turning it into knowledge.

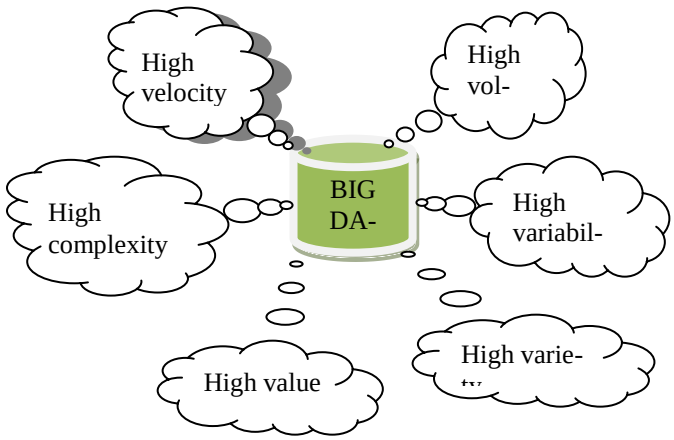

Figure 1. Characters of Big data

Big Data Analysis

Analytics is not new because there are many analytic technologies that have been available for many years such as regression analysis, simulation, and machine learning. Nowadays, new sources of data, business opportunities have created current interest and opportunities in big data analytics. It extend a new area of practice and study called 'data science' that encompasses the techniques, tools, technologies and process for making sense out of big data. Large data sets containing a variety of data types are processed by big data analytics that aim to uncover hidden patterns, unknown correlations, market trends, customer preferences and other useful business information. The results that finding after analyzing can lead to more effective marketing, new revenue opportunities, better customer service, improved operational efficiency, competitive advantages over rival organizations and other business benefits.

According to Philip Carter, Associate Vice President of IDC Asia Pacific, "Big data technologies describe a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data by enabling high-velocity capture, discovery and/or analysis" (Source; IDC. Big Data Analytics: Future Architectures, Skills and Roadmaps for the CIO, September 2011). In addition, this analysis is needed in real time or near-real time, and it must be affordable, secure and achievable.

There are three key technologies for extracting business value from Big data [2]

+ Information management for big data. Strategic of big data is managing data that is processes and controlled through big data analytics.

+ High-performance analytics for big data. With increasingly complex problems using more data, the requirement is to gain rapid insights from big data and the ability to solve that huge data.

+ Flexible deployment options for big data. There are options such as on-premises or hosted, software-as-a-service (SaaS) approaches for big data and big data analytics.

Benefits and challenges of Big Data

Big data revolution is creating new ways to gather and analyze information of varying types, size, and volume. But more important, big data brings a new set of opportunities for human resource and financial. However, due to solve the huge amount of data, there are many challenges of big data that is faced when using big data technologies.

Benefits of Big Data

Businesses are using the power of insights provided by big data to instantaneously establish who did what, when and where. The biggest value created by these timely, meaningful insights from large data sets is often the effective enterprise decision-making that the insights enable.
Extrapolating valuable insights from very large amounts of structured and unstructured data from disparate sources in different formats require the proper structure and the proper tools. To obtain the maximum business impact, this process also requires a precise combination of people, process and analytic tools. Some of the potential business benefits from implementing an effective big data insights methodology include:

+ Timely insights from the vast amounts of data. This includes those already stored in company databases, from external third-party sources, the Internet, social media and remote sensors.

Real-time monitoring and forecasting of events that impact either business performance or operation

+ Ability to find, acquire, extract, manipulate, analyze, connect and visualize data with the tools of choice (SAP HANA, SAP Sybase®, SAP Intelligence Analysis for Public Sector application by Palantir, Kapow®, Hadoop).
Convergence of the BDI solution for variety with the speed of SAP HANA for velocity

+ The capability of Hadoop (is an open source project hosted by Apache Software Foundation that is a tool of big data technology) for volumes to manage vast amounts of data, in or out of the Cloud, with validation and verification.
Identifying significant information that can improve decision quality.

+ Mitigating risk by optimizing the complex decisions of unplanned events more rapidly.

Challenges of Big Data

Privacy and Security

Due to Big data refers to the huge of digital information companies and governments, security and privacy issues are magnified by velocity, volume and variety of big data, such as large-scale cloud infrastructures, diversity of data sources and formats, streaming nature of data acquisition and high volume inter-clod migration. Moreover, the attack surface of

entire system will be quickly increased by using large-scale cloud infrastructures, with a diversity of software platforms, spread across large networks of computers.

Data access and sharing of information

It is necessary to make data open and make it available that is to be used to make accurate decisions in time. This leads to better decision-making, business intelligence and productivity improvements.

Almost companies refuse sharing of data between companies because they want to guarantee privacy, security for their clients and operations.

Storage and processing issues

Storing the large of data becomes the one of challenges for big data because the storage available is not enough. Cloud infrastructure may seem an option to store the rigorous demands of the big data on networks, storage and servers outsourcing the data. However, with terabytes of data, it with take large amount of time to upload in cloud and this data is changing so rapidly which will make this data hard to be uploaded in real time.

Analytical challenges

The choosing the type of analysis to be done on this huge amount of data brings along with it some huge analytical challenges. The various types of data such as unstructured, semi structured or structured data requires a large of advance skills.

Skill requirement

Due to Big data process the huge amount of data, it needs to attract organizations and youth with diverse new skill sets. These skills should extend to research, analytical, interpretive and creative ones.

Technical challenges

Fault-tolerant computing is extremely hard, involving intricate algorithms. It is true that no entire machines and software have reliable fault Torrance. Thus reducing the probability of failure to an "acceptable" level becomes the main task. However, the more we strive to reduce this probability, the higher the cost.

The scalability issue of Big data will be extremely difficult if we use old technologies because of the fact that many more hardware resources such as catch and processor memory channels are shared across a core in a single node.

Big data pipeline

Generally, each analyst's workflow varies in specific ways, however almost analyst activities are clustered into five steps [3]: 1) acquiring data, 2) choosing the architecture, 3) shaping the data to the architecture, 4) writing and editing code, 5) reflecting and iterating on the results.

Acquiring data: Finding answers for questions where the data in their big-data systems came from and how they discover sources of data becomes the first challenge. There is a verity of sources and formats that can collect data increasingly such as online databases of public statistics, private companies sell data from data market places (Microsoft's Azure Marketplace and Info chimps). However, it is not easy to combine data with information that collected themselves from sensor systems. Thus, improving standards for announcing data, helping people find data and formatting data can be more easily entered becomes very necessary.

Choosing architecture based on cost and performance: The second challenge of big data is choose architecture and building an appropriate big data solution because so many factors have to be considered such as storage, cost. Cloud is becoming a completely new part of designing a data analysis. Using cloud computing can computation, uploading/ downloading data, storage, simply pay more by buying either more machines or larger a board selection of VMs. However, clouds can double the memory or computation speed of a machine that does no meaning that double its speed. It can impose non-linear costs as communication overhead, storage, and other aspects change.

Shaping the data to the architecture: The analyst upload the data into platform after choosing a suitable architecture. They ensure that the data is uploaded in a way that is compatible with how the computation will be structured, distributed and partitioned appropriately. Moreover, data must be cleaned before data has been uploaded. This is the difficult process because it requires multiple people's expertise.

Write code: Selecting analysis will execute with an architecture selected and data in place. The analyses were articulated through code, written in C# and Microsoft's SCOPE, R, Python or PIG (a database-like language). Users must design their code and systems around the idea of separating their work into parallelizable jobs.

Debugging and iteration: Testing software is one of the activities that play an important role in the process of software. Thus, analyst must test their results after the execution run is complete. This leads to a process of debugging and looking for errors, iteration and changing code to work and visualization, in order to interpret results.

IEEE Transactions on Software English, 978-1-4799-0192-0/13, 2013.

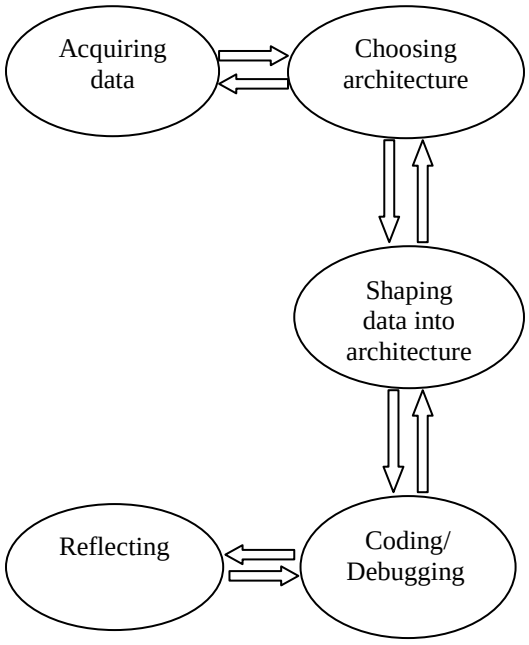

Figure 2. Big data pipeline

Conclusion

This paper described the overview of big data, its definition, benefits, challenges and pipeline. Big data will bring a lot of opportunities for human and finance if we know how to explore it. All challenges that have been described in this paper will help the business organizations to consider them right in the beginning and to find the methods to counter them.

References

[1] Hugh J.Watson, "Big data analytics: Concepts, Technologies and Applications", 2014.

[2] Mark Troester, "Big data meets big data analytics", ASME Biomed 2008 Conference, June 2008. .

[3] Danyel Fisher, Rob DeLine, Mary Czerwinski, Steven Drucker,, "Interactions with big data analytics", *Gas-troenterology*, 1997, 112:839.

[4] Avita Katal, Mohammad Wazid, R. H Goudar, "Big Data: Issues, Challenges, Tools and Good Practices,"