MAR 2016
6th ISSUE

A NEWSLETTER WRITTEN
BY COC-QU ACM-W CHAPTER

Qassim University
جامعة القصيم

BIG DATA

QASSIM UNIVERSITY
COLLEGE OF COMPUTER
IT DEPARTMENT

acm
COC-QU ACM-W
WOMAN IN COMPUTING

كلية الحاسب
COLLEGE OF COMPUTER

BIG DATA

A NEWSLETTER WRITTEN
BY COC-QU ACM-W CHAPTER

TEAM

Aseel Alawadh
Amal Alkhamees

Meznah Alquraishi
Muneerh Althbiti

Sarah Alhabeeb
Tasneem Almurad

Eatidal Almutairi
Hessah Alhudathy

Redna Almutlaq
Sara Almutairi

Haya Almohiteb

DESIGNER
Reham Alblihed

REVIEWED BY
Mrs. Samer Alduaji

TABLE OF CONTANTS

BIG DATA DEFINITION

Data is simply the fact or bit of information, but not the information itself. When data are processed, interpreted, organized, structured so as to make them meaningful or useful, they are called information. So, the information provides context for data.[1]

Big Data is a collect of very huge and complex data - both structured and unstructured - that can't be handled in one tool as it will take a long time to be analyzed and organized.

Big Data is a collect of very huge and complex data - both structured and unstructured - that can't be handled in one tool as it will take a long time to be analyzed and organized.Big data can be analyzed for insights that lead to better decisions and strategic business moves. There are so many websites that use big data, such as eBay.com, Amazon.com, google.com and more. [2]

HISTORY OF BIG DATA

The story of big data was started since 70 years ago. The old way to store data and information is building a library to collect books and papers. When a computer was used by human, it was acceptable to use it as a storage. However, the data and information were increasing day by day which caused what known as “information explosion” that mean large data is storing every day.

1989 Possibly the first use of the term Big Data (without capitalization) in the way it is used today. International best-selling author Erik Larson pens an article for Harpers Magazine speculating on the origin of the junk mail he receives. He writes: “The keepers of big data say they are doing it for the consumer’s benefit. But data have a way of being used for purposes other originally intended”.

1991 1997 Computer scientist Tim Berners-Lee announced the birth of World Wide Web. It is interconnected web of data, accessible to anyone from anywhere that where digital storage became more cost effective than paper. The web is increasing in size 10-fold each year. Much of this data that will never be seen by anyone and, therefore, yield no insight. So the Google search appearing.

1999 The term Big Data appears in Visually Exploring Gigabyte Dataset in real time, published by the Association for Computing Machinery.

HISTORY OF BIG DATA

2005 The birth of "Web 2.0" - the user-generated web where the majority of content will be provided by users of services, rather than the service providers themselves, also sees the creation of Hadoop the open source framework created specifically for storage and analysis of Big Data sets.

2008 2010 The world's servers process 9.57 trillion gigabytes of information equivalent to 12 gigabytes of information per person, per day. Eric Schmidt, executive chairman of Google tells a conference that as many data is now being created every two days as was created from the beginning of human civilization to the year 2003.

2013 The McKinsey report states that by 2018 the US will face a shortfall of between 140,000 and 190,000 professional data scientists, and states that issues including privacy, security and intellectual property will have to be resolved before the full value of Big Data will be realized.

2014 More people are using mobile devices to access digital data than office or home computers. 88% of business executives surveyed by GE working with Accenture report that big data analytic is a top priority for their business. At the end, the big data is the result of the development of computerized the data and information.[3]

CHARACTERISTICS OF THE BIG DATA

To identify the big data, there are three factors to determine if these data are big or no.

1. **Volume**

   The number of Terabyte of data that people are launching daily on the Internet around the world.

2. **Variety**

   With the increasing number of internet users, smartphones and social media, data varied either structured, semi-structured or unstructured data. Unstructured data include images, videos, messages, GPS data and so on. So, there has been a great variety of data.

3. **Velocity**

   In recent years, the speed for data occurrence is different. For example, the speed for development of tweets around the world differ from the speed for upload the images[4].

WHY IS BIG DATA IMPORTANT?

"You can't manage what you don't measure."

There is much wisdom in that saying, which has been attributed to both W. Edwards Deming and Peter Drucker. It explains why the recent explosion of digital data is so important. Simply because managers will be able to measure big data, and hence, know radically more about their businesses and directly translate that knowledge into improved decision making and performance. Consider retailing, Booksellers in physical stores could always track which books sold and which did not. If they had a loyalty program, they could tie some of those purchases to individual customers. And that was about it.[4]

Once shopping moved online, the understanding of customers increased dramatically. Online retailers could track not only what customers bought, but also what else they looked at; how they navigated through the site; how much they were in influenced by promotions, reviews, and page layouts; and similarities across individuals and groups. Before long, they developed algorithms to predict what books of customers would like to read next—algorithms that performed better every time the customer responded to or ignored a recommendation. Traditional retailers simply could not access this kind of information, let alone act on it in a timely manner.

Big data is changing how executives approach decisions. There is a lot less decision making on gut feelings, experience, and intuition. Decisions are also made faster.[5]

HOW IT WORKS

Big data generally comes from three sources:

1.Streaming Data

which run to the IT systems from a network of connected devices.

2.Social Media Data

This source has an increasingly set of information, especially for marketing, sales and support functions. That includes tweets, videos, pictures, etc.

3.Publicly Available Sources

Huge amounts of data that are available through open data sources, such as the US government's data.gov.[2]

After knowing where the data comes from, decisions should be made regarding the following questions:

How to store and manage the big data?
How much of it to analyze?
How to use any insights you uncover?

Then, a decision should be taken about which technology, or tool, will be used to make the most of the big data analytics. The following factors will be considered:

- Cheap, huge storage.
- Faster processors.
- Open source distributed big data platforms.
- Parallel processing, clustering, virtualization, large grid environments.
- Cloud computing and other flexible resource allocation arrangements.

BIG DATA ANALYTICS

WHAT IS BIG DATA ANALYTICS?

Big data tools can analyze high-volume, high-velocity, and high-variety information assets far better than conventional tools and relational databases. The later tools struggle to capture, manage, and process big data within a tolerable elapsed time and at an acceptable total cost of ownership. [6]

WHAT ARE THE BENEFITS OF BIG DATA ANALYTICS TOOLS?

It is used to improve customer retention, help with product development and gain a competitive advantage. Whether an organization is looking to boost sales and marketing results, uncover new revenue opportunities, improve customer service, optimize operational efficiency, reduce risk or drive other business results, big data insights can help.

TOP BIG DATA TOOLS

Many tools are used for data analysis, some are open source tools.and others are not. Bellow you will find the top four tools for big data analytics. [7]

Apache Hadoop

Hadoop is an open source software framework originally developed by Doug Cutting and Mike Cafarella in 2006. It has been built to handle very large data sets. Hadoop is made up of two main parts: the Hadoop Distributed File System (HDFS)which is the storage part of the tool and MapReduce which is the processing engine of Hadoop. The processing of data is done on parallel. For more information about this tool visit hadoop.apache.org.

Apache Spark

Apache Spark is an open source framework for cluster computing. It can be used as an alternate to Hadoop's MapReduce due to its ability to analyze data up to 100 times faster for certain applications. Apache Spark is commonly used for streaming data, machine learning and interactive analysis. For more information about this tool visit spark.apache.org.

Apache Hive

Apache Hive is a data warehouse infrastructure. It is a SQL-on-Hadoop data processing engine that utilizes a query language called HiveQL, based on SQL, but does not strictly follow the SQL-92 standard. For more information about this tool visit hive.apache.org

NoSQL Databases

NoSQL databases, Not Only SQL databases, are not bound by traditional schema models allowing them to collect unstructured datasets. The flexibility of NoSQL databases like MongoDB, Cassandra, and HBase make them a popular option for big data analytics. For more information about this tool visit http://nosql-database.org/

PARADOXES OF BIG DATA

There are Three Paradoxes of Big Data:

1.The Transparency Paradox
There is the upcoming transformation of the Internet of things to the Internet for everything. Cisco estimates that smart devices connected to the Internet will reach 37 billion in 2020.These devices and the data they produce will contribute to making the world more transparent, but the collection is invisible and techniques are opaque. Big data promises to use this data to make the world more transparent.

2.The Identity Paradox
Each of us wants to protect his personal identity while accessing to big data collection such as Internet browsing history, history of purchases, posts on social networks and others.This poses a threat to this identity and converts it into a set of projections on data.

3.The Power Paradox
The individuals and users don't gain much benefits from big data as the companies, that own or use, the analysis tools gain.
Here is an example of this power : the companies discover that a teenage girl pregnant because of changing purchasing habits they know even before her father knew of her pregnancy!
So, if we do not build privacy, transparency, autonomy, and identity protections into big data from the outset, the power paradox will diminish big data's lofty ambitions. We need a healthier balance of power between those who generate the data and those who make inferences and decisions based on it so that one doesn't come to unduly revolt or control the other. [8]

BIG DATA APPLICATIONS

Starting from government to sports and entertainment, big data analytics are used in different fields and for many uses. Here are some examples of where this technology can be used.

Science and Research

Google's DNA Stack compiles and organizes DNA samples of genetic data from around the world to identify diseases and other medical defects. These fast and exact calculations eliminate any 'friction points,' or human errors that could be made by one of the numerous science and biology experts working with the DNA.

Sports

Big data can be used to improve training and understanding competitors. Besides, it is possible to predict winners in a match using big data analytics.Future performance of players could be predicted as well.

Technology

eBay.com uses two data warehouses at 7.5 petabytes and 40PB as well as a 40PB Hadoop cluster for search, consumer recommendations, and merchandising. [9]
Amazon.com handles millions of back-end operations every day, as well as queries from more than half a million third-party sellers.
Facebook handles 50 billion photos from its user base.[10]

Healthcare

By providing personalized medicine and prescriptive analytics, clinical risk intervention and predictive analytics.[11]

BIG DATA APPLICATIONS

Marketing

- campaign management and optimization.
- Micro segmentation of customers and markets.
- Location-based marketing
- Cross-selling and up-selling.
- Sentiment analysis.

Finance

- Risk management.
- Fraud detection and prevention.
- Wealth mangment.
- Anti-money laundering.
- Credit risk, Scoring and analysis.
- Trade surveillance.

Government

- Fraud and threat prediction and detection.
- cyber security.
- Complic and regulatory analysis.

Insurance

- Risk assessment and avoidance.
- Claims Fraud detection
- Call center workload analysis.
- Telematics-optimized underwriting.
- Customer values management.
- Catastrophic planning.

Healthcare

- Patient care quality and outcome analysis.
- Reimbursement model.
- Public health reporting .
- Clinical data transparency.
- Public health surveillance and response.
- Clinical trial design and analysis.

Retail

- Merchandising and market basket analysis.
- Supply chain management and analytics.
- Loyalty program management.
- Event/behavior-based targeting
- Cross-channel customer service optimization.

Telecommunications

- Customer churn prevention .
- Call detail record analysis .
- Network planning and optimization .
- Mobile user location analysis .
- New product research and

WHERE ARE BIG DATA

In 2015, bigger, faster and different data had been created. It has been noticed that there is a spread of self-service data analytics with the adoption of cloud services, Hadoop, NoSQL and other technologies.[12]

Therefore in 2016, the researchers believe through the use of both AI and Machine learning, it will become possible for these huge amounts of information to be processed, stored and mined without needing human interactions to do so. This will result in a scrutiny increase on how data is dealt with and protected. It is likely to see more automated platforms through many technologies and trends to track -machine learning, AI, advanced analytics, predictive analytics, real-time analytics, Hadoop, Spark, other Apache Foundation projects, open source, cloud-based-as-a-service offerings, self-service, and more.[13]

AHOOD ALAMER

8th Level of IT, College of Computer
CoC-QU ACM-W student chapter's Member
Qassim University

My name is Ahood i'm from alqassim born on 1992. My mother give me nice name ' Ahood' I love my name i'm thankful for my mother. Iam interested in Information technology, draw and design in photoshop, I i'm looking to develop other skills. I love everything related to computer, I see IT all my dream and I feel happy when i'm programming. I love helping people whenever I can and I like positive people in their lives. Finally, (everything you need, will come to you at the perfect time). I'm lucky because I will graduate from information technology in Qassim University.

TASNEEM AL-MURAD

6th Level of IT, College of Computer
CoC-QU ACM-W student chapter's Member
and chair of Computer Club at CoC
Qassim University

Positive, optimistic girl born in 3rd of Oct line93. Syrian born and live in Qassim. From I was young I was interested in Technologies , my goal was being Hacker in the future. I'm youngest in my family but they always ask me when they having problem in any mobile, computer, modem even TV or receiver. I choose my major IT, I love it and hope to be top in it. I like security and programming. My hobbies are making video, learning new languages and get knowledge from every majors also playing games specially strategic games "chess, sudoku, Age of Empire". I'm not perfect, I'm rule breaker, I trust my self. I want my parents proud on me. This is me "Tasneem".

References :

[1]Diffen.com. (2016). Data vs Information - Difference and Comparison | Diffen. [online] Available at: http://www.diffen.com/difference/Data_vs_Information [Accessed 21 Mar. 2016].

[2] Sas.com. (2016). What is Big Data and why it matters. [online] Available at: http://www.sas.com/en_us/insights/big-data/what-is-big-data.html [Accessed 21 Mar. 2016].

[3]Gropp, B. (2015). Papers everyone should read. ACM SIGHPC Connect, 1(4), pp.5-5.

[4]McAfee, A. and Brynjolfsson, E. (2010). Big Data: The Management Revolution. [online] Available at: http://www.rosebt.com/uploads/8/1/8/1/8181762/big_data_the_management_revolution.pdf [Accessed 21 Mar. 2016].

[5]Anon, (2016). 1st ed. [ebook] Available at: http://vivaldipartners.com/pdf/18224_HBR_SerasaExperian_Report_English.pdf [Accessed 21 Mar. 2016].

[6]Qubole. (n.d.). Big Data Analytics Tools Overview & Benefits | Qubole. [online] Available at: https://www.qubole.com/resources/articles/big-data-analytics/ [Accessed 21 Mar. 2016].

[7]Anon, (n.d.). [online] Available at: The Technical University of Denmark in course 11374 : http://www.bigdata.dtu.dk [Accessed 21 Mar. 2016].

[8]Neil M. Richards & Jonathan H. King. 2013. Three Paradoxes of Big Data. [ONLINE] Available at: http://www.stanfordlawreview.org/online/privacy-and-big-data/three-paradoxes-big-data. [Accessed 21 March 16].

[9]Admire Moyo. UN Global Pulse (2012). Big Data for Development: Opportunities and Challenges (White p. by Letouzé, E.). New York: United Nations

[10]WEF (World Economic Forum), & Vital Wave Consulting. (2012). Big Data, Big Impact: New Possibilities for International Development. World Economic Forum. Retrieved 24 August 2012, from http://www.weforum.org/reports/big-data-big-impact-new-possibilities-international-development

[11]"Big Data for Development: From Information- to Knowledge Societies", Martin Hilbert (2013), SSRN Scholarly Paper No. ID 2205145). Rochester, NY: Social Science Research Network; http://papers.ssrn.com/abstract=2205145

[12]InformationWeek. (n.d.). Big Data Predictions For 2016 - InformationWeek. [online] Available at: http://www.informationweek.com/big-data/big-data-analytics/big-data-predictions-for-2016/d/d-id/1323671 [Accessed 21 Mar. 2016].

[13]Enterprise, B., Channels, C., Type, C. and Us, A. (2016). Big Data Top Trends In 2016 | Articles | Chief Data Officer. [online] Channels.theinnovationenterprise.com. Available at: https://channels.theinnovationenterprise.com/articles/big-data-trends-2016 [Accessed 21 Mar. 2016].