

Chapter 93

Big Data and Healthcare Data: A Survey

Bikash Kanti Sarkar

Birla Institute of Technology, Ranchi, India

ABSTRACT

Big data and its analytics yield a lot of opportunities to make great progresses in many fields, ranging from economic and business activities to public administration, from national security to scientific researches and so on. However, the most noticeable point is that healthcare data has been recently identified as a prime example of big data. Undoubtedly, efficient use of healthcare resources has become a key factor in improving overall healthcare system. But for managing healthcare data and obtaining potential results, we need integration and sharing of data that ultimately demand the concept of distributed system. The paper in its first phase gives an overview on big data and healthcare data from different aspects. A review on the state-of-the-art distributed file system (Hadoop) is conducted in this stage too. The primary aim of this phase is to provide an overall picture on big data as well as healthcare data for non-expert readers. In the next phase, a cloud-based e-health system is proposed for the expert audiences. The expected promising characteristics as well as the managerial implications of the model are highlighted in the analysis section.

INTRODUCTION

Truly speaking, the concept of ‘big data’ is not new; however, the way it is defined is constantly changing. In practice, a data set is called ‘big’ if it ranges from a few terabytes (1TB = 2^{40} bytes) to many petabytes (1PB = 2^{50} bytes) but the term ‘big data’ technically implies that the generation rate is unprecedented. According to Philip and Zhang (2014), the speed of data-growth has already exceeded Moore’s law. In this context, IBM has reported that 90% of the data created in the world has been produced in the last two years (IBM, 2011). More than 267 million transactions are produced per day over 6 thousand stores of Wal-Mart. Till 2011, almost 3 terabytes of data are collected by the US Library of Congress. According to Bakshi and Kapil (2012), the size of digital data in 2011 was roughly 1.8 zettabytes (1.8×10^{21}).

DOI: 10.4018/978-1-7998-5339-8.ch093

bytes) and the supporting network infrastructure has to manage 50 times more information by year 2020. Obviously, the real concern is storing and processing of data. In other words, data sets are growing so rapidly that storing and processing them using traditional data management tools or applications (e.g., DBMS) become very difficult. However, the notable matter is that huge amount of data with variation significantly assist the researchers to frame effective predictive models by finding useful patterns through qualitative data analysis. In fact, decisions that were previously based on guesswork or on painstakingly constructed models of reality, can now be made based on the available data itself. Certainly, we were lacking it earlier.

Big data phenomenon is an emerging area of research at the present date. The field indeed belongs to Data Science and Data Engineering. It is interesting to note that a study on the evolution of big data (a research and scientific topic) claims that the term was found in 1970s but it has been published in 2008 (Gali & Henk, 2012). For defining big data, there exist various forms of V's such as 3V's, 4V's, 5 V's, etc. For example, Laney (2011) used volume, velocity and variety, known as 3V's to characterize the concept of big data, whereas people often extend another 'V' as per their special requirements. According to Zikopoulos and Chris (2011), the fourth 'V' can be value, variability or virtual. Thus, big data is conceptually described by the features, like volume, velocity, variety, etc. (i.e., in terms of V's) and most of us agree that big data should have four standard characteristics (called as 4V's namely volume, variety, velocity and veracity) as suggested by IBM (Kevin, 2013). More details on big data definition are unfolded in the immediate next section.

At the present date, the concept of big data yields a lot of progressive opportunities in many fields such as Information Technology, customer care service, online transaction, web-data applications, risk management, astronomy, healthcare system, etc.

In healthcare system, the information stored in health database has been increased over the last ten years, leading it to be considered 'big data'. According to Raghupathi(2010), this industry has historically generated huge amount of data driven by record-keeping and patient-care. A number of surveys and journal articles describe that the massive quantity of health data holds the promise of supporting a wide range of medical and healthcare services, including clinical decision support, sensor based health condition and food safety monitoring, disease surveillance and population health management (Dembosky, 2012; Feldman, Martin, & Skotnes, 2012; Fernandes, O'Connor, & Weaver, 2012). For evidence, Cancer is still a major challenge in research and medical-world despite enormous advances in technology. The reason is, analysis of cancer requires petabytes of data to be gathered from various dispersed clinical and analytical databases with high dimensional scaling and interpretation to identify the state of the disease and the survival potential of the patient (Dutta & Bilbao, 2012). Further, the use of information technology on healthcare data can reduce the cost of healthcare while improving its quality by emphasizing more on preventive and personalized care and basing of continuous monitoring (CCC, 2011). In this context, an estimation of savings of \$300 Billion every year in the US alone is reported in a study (James, Michael, Brad, Jacques, Richard, Charles, & Hung, 2012). According to the report prepared by IBM (2013), e-health system supported in U.S. reduces a big amount of healthcare spending and improves healthcare process to a great extent. In this regard, two evidences collected from this report are included in the Appendix.

To satisfy the aforementioned healthcare purposes, health information needs to be accessible and available to everyone involved in the delivery of patient healthcare from the researchers attempting to find causes, treatments, and cures for diseases to the patients themselves. Definitely, a high-level integration of data and its sharing is essential among different healthcare practitioners and institutions in order to

29 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:

www.igi-global.com/chapter/big-data-and-healthcare-data/275371

Related Content

A Comprehensive Report on Security and Privacy Challenges in Software as a Service

Pradeep Kumar Tiwari and Sandeep Joshi (2021). *Research Anthology on Architectures, Frameworks, and Integration Strategies for Distributed and Cloud Computing* (pp. 1714-1739).

www.irma-international.org/chapter/a-comprehensive-report-on-security-and-privacy-challenges-in-software-as-a-service/275362

Design and Implementation of Multi-Agent Online Auction Systems in Cloud Computing

Hongyan Yu, Srikanta Patnaik, Shenjia Ji, Liguojia and Tengxiao Yang (2021). *Research Anthology on Architectures, Frameworks, and Integration Strategies for Distributed and Cloud Computing* (pp. 251-269).

www.irma-international.org/chapter/design-and-implementation-of-multi-agent-online-auction-systems-in-cloud-computing/275288

Cloud Computing, Green Computing, and Green ICT

Kijpokin Kasemsap (2021). *Research Anthology on Architectures, Frameworks, and Integration Strategies for Distributed and Cloud Computing* (pp. 2083-2099).

www.irma-international.org/chapter/cloud-computing-green-computing-and-green-ict/275380

IoT Big Data Architectures, Approaches, and Challenges: A Fog-Cloud Approach

David Sarabia-Jácome, Regal Gonzalez-Usach and Carlos E. Palau (2021). *Research Anthology on Architectures, Frameworks, and Integration Strategies for Distributed and Cloud Computing* (pp. 227-250).

www.irma-international.org/chapter/iot-big-data-architectures-approaches-and-challenges/275287

Comparative Study for Different Provisioning Policies for Load Balancing in Cloudsim

Layla Albdour (2021). *Research Anthology on Architectures, Frameworks, and Integration Strategies for Distributed and Cloud Computing* (pp. 600-611).

www.irma-international.org/chapter/comparative-study-for-different-provisioning-policies-for-load-balancing-in-cloudsim/275304