Big Data: Big Innovations in Healthcare

Jasmeen Gill
RIMT-IET, Mandi Gobindgarh

Shaminder Singh
GGI, Khanna

Devdutt Baresary
GGI,Khanna

Abstract

Big Data describes 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 analysis. Since healthcare services involve huge amounts of data, driven by record keeping, medical reports, notes, compliance & regulatory requirements, and patient care, big data will be a boon to the medical field. This research article highlights various aspects of big data like usability, security and reliability in healthcare services. Apart from this, it provides various analytical tools of big data used in healthcare.

1. Introduction

Big data is a relative term describing a situation where the volume, velocity and variety of data exceed an organization's storage or compute capacity for accurate and timely decision making [2].

Decrease in the cost of both storage and compute power has made it feasible to collect this data - which would have been thrown away only a few years ago. As a result, more and more companies are looking to include non-traditional yet potentially very valuable data with their traditional enterprise data in their business intelligence analysis. Traditional enterprise data includes customer information from CRM systems, transactional ERP data, web store transactions, and general ledger data. Machine-generated /sensor data includes Call Detail Records ("CDR"), weblogs, smart meters, manufacturing sensors, equipment logs (often referred to as digital exhaust), and trading systems data. Social data – includes customer feedback streams, micro-blogging sites like Twitter, and social media platforms like Facebook [1].

Wal-Mart handles more than a million customer transactions each hour and imports those into databases estimated to contain more than 2.5 petabytes of data. Radio frequency identification (RFID) systems used by retailers and others can generate 100 to 1,000 times the data of conventional bar code systems [2].

Private sector organizations such as Google, Twitter and Face book hold enormous data stores people across the world, and offer access to these on commercial terms [4]. Facebook handles more than 250 million photos uploads and the interactions of 800 million active users with more than 900 million objects (pages, groups, etc.) each day. More than 5 billion people are calling, texting, tweeting and browsing on mobile phones worldwide [2].

This flood of data is generated by connected devices from PCs and smart phones to sensors such as RFID reader sand traffic cams, In healthcare ,for instance ,clinical data can now come in the form of images(e.g. from X-rays, CT-scan ,and ultrasound)and videos [8,23].

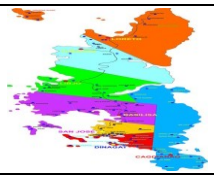
The hopeful vision of big data is that organizations will be able to harvest and harness every byte of relevant data and use it to make the best decisions. Big data technologies not only support the ability to collect large amounts, but more importantly, the ability to understand and take advantage of its full value [8, 22-23].

IBM suggests that 90% of the world's data has been generated only in the past two years. Clearly, the data volumes enterprises generate every day impact the effectiveness of how they synthesize and interpret fraud and corruption risks on a timely basis [5].Government agencies hold or have access to an ever increasing wealth of data including spatial and location data, as well as data collected from and by citizens. Experience suggests that such data can be utilized in ways that have the potential to transform service design and delivery so that personalized and streamlined services, that accurately and specifically meet individual's needs, can be delivered to them in a timely manner. The private sector holds huge amounts of data about its customers and in many cases leads the way in how this data is analyzed and used to create new business models and services. Agencies have the opportunity to learn from the innovations occurring in the private sector to operate more efficiently and deliver services more effectively while ensuring that privacy and security matters are carefully considered [4].

It is Not Just About Building Bigger Databases: Big data is not about the technologies to store massive amounts of data. It is about creating a flexible infrastructure with high-performance computing, high performance analytics and governance, in a deployment model that makes sense for the organization [2].

2. VVV (V3) structure of Big Data

The commonly accepted definition of big data comes from Gartner who define it as high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing for

Volume	Description	Model Size
Terabyte	Will fit 200,000 photos or MP3 songs on a single 1 terabyte hard drive	
Petabyte	Will fit on 16 Back blaze storage pods racked in two data center cabinets	
Exabyte	Will fit in 2,000 cabinets and fill a four story data center that takes up a city block	
Zettabyte	Will fit in 1,000 data centers	
Yottabyte	Will fit in the Island with a million data centers	

enhanced insight, decision making, and process optimization. These are known as the "three Vs" [4]. In fact, there are four key characteristics that define big data [1]:

2.1. Volume

The sheer volume of data being stored today is exploding. In the year 2000, 800,000 petabytes (PB) of data were stored in the world. Of course, a lot of the data that's being created today is not analyzed at all and that's another problem we're trying to address with Big In sights. We expect this number to reach 35 zettabytes (ZB) by 2020. Twitter alone generates more than 7 terabytes (TB) of data everyday, Facebook 10TB, and some enterprises generate Terabytes [10].

Machine-generated data is produced in much larger quantities than non-traditional data. For instance, a single jet engine can generate 10TB of data in 30 minutes. With more than 25,000 airline flights per day, the daily volume of just this single data source runs into the Petabytes. Smart meters and heavy industrial equipment like oil refineries and drilling rigs generate similar data volumes, compounding the problem [4]. Volume specifications for big data are shown in table 1 along with the description.

Table 1. Volume Specifications in Big Data.

2.2. Velocity

Social media data streams, while not as massive as machine-generated data, produce a large influx of opinions and relationships valuable to customer relationship management. Even at 140 characters per tweet, the high velocity (or frequency) of Twitter data ensures large volumes (over 8 TB per day) [1].

2.3. Variety

The volume associated with the Big Data phenomena brings along new challenges for data centers trying to deal with it: its variety. With the explosion of sensors, and smart devices, as well as social collaboration technologies ,data in an enterprise has become complex, because it includes not only traditional relational data, but also raw , semi structured ,and unstructured data from webpages ,web log files(including click-stream data),search indexes , social media forums-mail ,documents ,sensor data from active and passive systems ,and soon .What's more, traditional systems can struggle to store and perform the required analytics to gain understanding from the contents of these logs be-cause much of the in formation being generated doesn't lend it self to traditional database technologies [10].Traditional data formats tend to be relatively well defined by a data schema and change slowly. In contrast, non-traditional data formats exhibit a dizzying rate of change [1].Up to 85 percent of an organization's data is unstructured – not numeric –but it still must be folded into quantitative analysis and decision making. Text, video, audio and other unstructured data require different architecture and technologies for analysis [2].

3. Platform Selection

The decision to choose a particular platform for a certain application usually depends on the following important factors: data size, speed or throughput optimization and model development. We will now provide more details about each of these factors [9, 26].

3.1. Data size

The size of data that is being considered for processing is probably the most important factor. If the data can fit into the system memory, then clusters are usually not required and the entire data can be processed on a single machine. The platforms such as GPU, Multicore CPUs etc. can be used to speed up the data processing in this case. If the data does not fit into the system memory, then one has to look at other cluster

options such as Hadoop, Spark etc. Again, Hadoop and Spark clusters can handle large amount of data but Hadoop has well developed tools and frameworks although it is slower for iterative tasks. The user has to decide if he needs to use off-the-shelf tools which are available for Hadoop or if he wants to optimize the cluster performance in which case Spark is more appropriate.

3.2. Speed or throughput optimization

Here, speed refers to the ability of the platform to process data in real-time whereas throughput refers to the amount of data that system is capable of handling and processing simultaneously. The users will need to be clear about whether the goal is to optimize the system for speed or throughput. If one needs to process large amount of data and do not have strict constraints on the processing time, then one can look into systems which can scale out to process huge amounts of data such as Hadoop , Peer to-Peer networks, etc.

These platforms can handle large-scale data but usually take more time to deliver the results. On the other hand, if one needs to optimize the system for speed rather than the size of the data, then they need to consider systems which are more capable of real-time processing such as GPU, FPGA etc.

3.3. Model Development

In data analytics, training of the model is typically done offline and it usually takes a significant amount of time. A model is typically applied in an online environment where the user expects the results within a short period of time (almost instantaneously).

4. Analytical tools in Big Data

Mapping this data to primary sources to their required destination required Hadoop like framework and Map Reduce like technology [8].

4.1. Hadoop

Hadoop is a new technology that allows large data volumes to be organized and processed while keeping the data on the original data storage cluster [8, 21]. Hadoop is open source framework and has two components that are HDFS and Map Reduce [8, 24]. Hadoop Distributed File System (HDFS) is the long-term storage system for web logs for example. These web logs are turned into browsing behavior (sessions) by running MapReduce programs on the cluster and generating aggregated results on the same cluster. These aggregated results are then loaded into a Relational DBMS system [8, 21]. Hadoop YARN [8] is a resource management layer and schedules the jobs across the cluster [9, 25].

4.2. MapReduce

The programming model used in Hadoop is MapReduce [9, 26] which was proposed by Dean and Ghemawat at Google. MapReduce is the basic data processing scheme used in Hadoop which includes breaking the entire task into two parts, known as mappers and reducers. At a high-level, mappers read the data from HDFS, process it and generate some intermediate results to the reducers. Reducers are used to aggregate the intermediate results to generate the final output which is again written to HDFS.A typical Hadoop job involves running several mappers and reducers across different nodes in the cluster. Map/Reduce in distributed is currently using by Google [8, 24].

5. Fields in Big Data

Big data is performing better results in the following fields like [11].

5.1. Financial Services

Data gleaned from mobile money services can provide deep insight into spending and saving habits across sectors and regions. Digital payment histories can allow individuals to build credit histories, making them candidates for loans and other credit-based financial services.

5.2. Education

Data derived from the use of mobile value-added services can be used to improve public-sector understanding of educational needs and knowledge gaps, allowing more targeted and timely initiatives to disseminate critical information.

5.3. Agriculture

Mobile payments for agricultural products, input purchases and subsidies may help governments better predict food production trends and incentives. This knowledge can be used to ensure the availability of proper crop storage, reduce waste and spoilage, and provide better information about what types of financial services are needed by farmers. Mobile use patterns may also help governments and development organizations identify regions in distress so that targeted assistance can be directed to them. Early detection can help prevent families from leaving their land and further decreasing agricultural production.

5.4. Health

Data collected through mobile devices, whether captured by health workers, submitted by individuals, or analyzed in the form of data exhaust, can be a crucial tool in understanding population health trends or stopping outbreaks). When collected in the context of individual electronic health records, this data not only improves continuity of care for the individual, but it can be used to create massive datasets with which treatments and outcomes can be compared in an efficient and cost effective manner. Many platforms and tools are used for big data analytics in healthcare [6].

5.4.1. Hadoop. Distributed File System (HDFS) enables the underlying storage for the Hadoop cluster. It divides the data into smaller parts and distributes it across the various servers/nodes.

5.4.2. MapReduce. provides the interface for the distribution of sub-tasks and the gathering of outputs. When tasks are executed, MapReduce tracks the processing of each server/node.

5.4.3. Pig Latin. programming language is configured to assimilate all types of data (structured/unstructured, etc.). It is comprised of two key modules: the language itself, called PigLatin, and the runtime version in which the PigLatin code is executed.

5.4.4. Hive .is a runtime Hadoop support architecture that leverages Structure Query Language (SQL) with the Hadoop platform. It permits SQL programmers to develop Hive Query Language (HQL) statements akin to typical SQL statements.

5.4.5. Jaql .is a functional, declarative query language designed to process large data sets. To facilitate parallel processing, Jaql converts "'high-level' queries into 'low-level' queries" consisting of MapReduce tasks.

5.4.6. Zookeeper .allows a centralized infrastructure with various services, providing synchronization across a cluster of servers. Big data analytics applications utilize these services to coordinate parallel processing across big clusters.

5.4.7. Hbase. is a column-oriented database management system that sits on top of HDFS. It uses a non-SQL approach.

5.4.8. Lucene. project is used widely for text analytics /searches and has been incorporated into several open source projects. Its scope includes full text indexing and library search for use within a Java application.

5.4.9. Avro. facilitates data serialization services.

5.4.10. Avro. facilitates data serialization services.

5.4.11. Versioning. and version control is additional useful features.

5.4.12. Mahout. is yet another Apache project whose goal is to generate free applications of distributed and scalable machine learning algorithms that support big data analytics on the Hadoop platform.

The healthcare industry contains a large amounts of data, driven by record keeping, medical reports, notes, compliance & regulatory requirements, and patient care [6,12].Mostly data which is used in the form of hard copy, the current trend is toward digitization of these large amounts of data. Driven by mandatory requirements and the potential to improve the quality of healthcare delivery meanwhile reducing the costs, to speed up the system, these massive quantities of data (known as 'big data') hold the promise of supporting a wide range of medical and healthcare functions, including among others clinical decision support, disease surveillance, and population health management [6, 13-16].Reports say data from the U.S. healthcare system alone reached, in 2011, 150 exabytes. At this rate of growth, big data for U.S.healthcare will soon reach the zettabyte (1021 gigabytes) scale and, not long after, the yottabyte (1024 gigabytes) [6, 17].To manage this very large amount of data with traditional software/ hardware/ data management tools is very difficult or impossible [6, 18].So this big data amount of data require advanced techniques and technologies to enable the capture, storage, distribution, management and analysis of the information" [6, 17].

There are a wide range of benefits given by big data to the medical field like these analytical techniques can be applied to the vast amount of patient-related health and medical data to reach a deeper understanding of outcomes, which then can be applied at the point of care. These analytical techniques analyze disease patterns and tracking disease outbreaks and transmission to improve public health surveillance and speed response also provide faster development of more accurately targeted vaccines, e.g., choosing the annual influenza strains; and, 3) turning large amounts of data into actionable information that can be used to identify needs, provide services, and predict and prevent crises, especially for the benefit of populations [6,20].These also can contribute to Evidence-based medicine: Combine and analyze a variety of structured and unstructured data-EMRs, financial and operational data, clinical data, and genomic

data to match treatments with outcomes, predict patients at risk for disease or readmission and provide more efficient care; These techniques can be used to capture and analyze in real-time large volumes of fast-moving data from in-hospital and in-home devices, for safety monitoring and adverse event prediction; The analytic techniques can be applied to patient profiles (e.g., segmentation and predictive modeling) for identify individuals who would benefit from proactive care or lifestyle changes, for example, those patients at risk of developing a specific disease (e.g., diabetes) who would benefit from preventive care [6,19].

6. Challenges in Big Data

Big data is not only tabular; it also includes documents, e-mails, pictures, videos, sound bites, social media extracts, logs and other forms of information that is difficult to fit into the nicely organized world of traditional database tables (rows and columns) .Companies that tackle big data as a technology-only initiative will only solve a single dimension of the big data mandate. There are sheer volumetric issues, such as billions of rows of data that need to be solved. While tried-and-true technologies (partitioning) and newer technologies (MapReduce, etc.) permit organizations to segment data into more manageable chunks, such an approach does not deal with the issue that rarely used information is clogging the pathway to necessary information. Traditional lifecycle management technolo -gies will alleviate many of the volumetric issues, but they will do little to solve the non-technical issues associated with volumetrics [7].Here we can say that there are some challenges in big data migration in cloud namely

(i) Scalable Data Management
(ii) Data Management for Large Applications
(iii) Large Multitenant Databases
(iv) Large Databases security issues for cloud compu -ting, MapReduce and Hadoop environment [8].

7. Conclusion

It is deduced from the survey that big data allows huge amount of data storage and hence referred to as huge data set having really huge magnitude. Three characteristics define Big Data: volume, variety, and velocity. Data analytics in healthcare is evolving into a promising field for providing insight from very large data sets and improving outcomes while reducing costs. Likewise, business performance can be improved via data driven decision making with big data technologies. The analytical tools: Hadoop and Map Reduce technology have made it possible to analyze and manage huge amount of data resolving scalability issues.

The hopeful vision of big data is that organizations will be able to harvest and harness every byte of relevant data and use it to make the best decisions. Big data technologies are foreseen as to not only support the ability to collect large amounts, but more importantly, the ability to understand and take advantage of its full value.

8. References

[1] Dijcks, J.P. , "Big Data for the Enterprise", *An Oracle White Paper*, June 2013, pp. 1-15.

[2] Troester, M., "Big Data Meets Big Data Analytics", *SAS, white paper, SAS Institute Inc*., pp.1-13.

[3] Provost, F., and Fawcett, T., "Data Science and its relation -ship to Big Data and Data Driven Decision Making", *MARY ANN LIEBERT, INC*., 1, 1, March 2013, pp.51-59.

[4] "Big Data Strategy- issues papers", March 2013, pp.1-12,

[5] "Big Data Survey-Real-Time Stream Processing and Cloud -Based Big Data Increasing in Today's Enterprises", *Giga Spaces Technologies*, 2012, pp.1-5.

[6] Raghupathi, W., and Raghupathi, V., "Big data analytics in healthcare: promise and Potential", *Health Information Science and Systems*, 2, 3, 2014.

[7] Albala, M., "Making Sense of Big Data in the Petabyte Age", *cognizant 20-20 insights*, June 2011, pp.1-7.

[8] Manekar, A., and Pradeepini, G., "A Review on cloud based big data analytics", *ICSES- Journal on computer networks and communication, 1, 1, 2015, pp.6*-9.

[9] Singh, D., and Reddy, C., K., "A survey on platforms for big data analytics*", journal of big data: a springer open journal*, 1, 8, 2014, pp. 1-20.

[10] Zikopoulos, P., C., Eotan C., Deroos, D., Lapis, G., "Under -standing Big Data", *McGraw-Hil*, 2012.

[11] "Big Impact: New Possibilities for International Develop -ment", *The World Economic Forum*, 2012, pp.1-10.

[12] Raghupathi, W., "Data Mining in Health Care", *In Health care Informatics: Improving Efficiency and Productivity.* Edited by Kudyba S. Taylor & Francis, 2010, pp.211–223.

[13] Burghard, C., "Big Data and Analytics Key to Account -able Care Success", *IDC Health Insights*, 2012.

[14] Dembosky, A., "Data Prescription for Better Healthcare." *Financial Times*, December 2012, pp. 19, Available from: http://www.ft.com/intl/cms/s/2/55cbca5a-4333-11e2-aa8f-00144feabdc0.html#axzz2W9cuwajK.

[15] Feldman, B., Martin, E., M., and Skotnes, T., "Big Data in Healthcare Hype and Hope." October 2012. http://www.west-info.eu/files/big-data-inhealthcare.pdf.

[16] Fernandes, L., Connor, M., and Weaver, V., "Big data, bigger outcomes", *J AHIMA*, 2012, pp.38–42.

[17] "Transforming Health Care through Big Data Strategies for leveraging big data in the health care industry", IHTT, 2013. http://ihealthtran.com/wordpress/2013/03/iht%C2%B2-releases-big-data-research-reportdownload-today/.

[18] "Drowning in Big Data? Reducing Information Techno – logy Complexities and Costs for Healthcare Organizations", Frost & Sullivan, http://www.emc.com/collateral/analyst-reports/frost-sullivan-reducing-information-technologycomplexities-ar.pdf.

[19] "IBM: IBM big data platform for healthcare.", *Solutions Brief*,2012.http://public.dhe.ibm.com/common/ssi/ecm/en/ims14398usen/IMS14398USEN.PDF.

[20] Manyika, J., Chui, M., Brown, B., Buhin, J., Dobbs, R., Roxburgh, C., and Byers, A., H., "Big Data: The Next Frontier for Innovation, Competition, and Productivity*, USA, McKinsey Global Institute*, 2011.

[21] Abouzeid,A.,Pawlikowski,K.,B.,Abadi,D.,J.,Rasin,A., and Silberschatz, A., "HadoopDB:An Architectural Hybrid of Map Reduceand DBMS Technologies for Analytical Workloads " , *PVLDB*,2,1,2009,pp.922–933.

[22] Agrawal, D., Das, S., and Abbadi, A., E., "Big data and cloud computing: New wine or just new bottles?", *PVLDB*, 3, 2, pp. 1647–1648.

[23] "Solution Brief Big Data in the Cloud: Converging Tech nologies, How to Create Competitive Advantage Using Cloud-Based Big Data Analytics ", *White paper on Big Data*.

[24] "Hadoop Distributed File System: Architecture and Design", http://hadoop.apache.org/common/docs/r0.18.2.

[25] Vavilapalli, V., K., Murthy, A., C., Douglas, C., Agarwal, S., Konar, M., Evans, R., Graves, T., Lowe, J., Shah, H., and Seth, S., "Apache hadoop yarn: Yet another resource negotiator", in *Proceedings of the 4th annual Symposium on Cloud Comp -using*, 2013, pp. 5.

[26] Dean, J., and Ghemawat, S., "MapReduce: simplified data processing on large clusters", Commun. *ACM*, 51, 1, 2008, pp. 107–113.