Data mining & warehouse (Survey Report 2)

Big Data in healthcare: management, analysis and future prospects

Contributed by(*Students of BSIT 7A***):**

Hifsa Basharat 45903
Haris Anwer 45902
M Laraib Kiayani 45916

Instructor

Sir. Imran Memon

Survey Report (2019) 15/12

Bahria University
Discovering Knowledge

SURVEY PAPER

Big data in healthcare: management, analysis and future prospects

Hifsa Basharat, Haris Anwer, and Muhammad Laraib Kiayani

Correspondence:
Hifsabasharat689@gmail.com;
Harisanwer125@gmail.com
skiayani404@hotmail.com

Hifsa, Haris and Laraib contributed equally in this work
For the purpose of Submition of Semester assignment

Abstract

"Big data" is a large amount of information that can work very well. Biomedical research is also a significant part of the big data relevant to public health. This data requires good care and analysis to get the right information. Otherwise, the search for answers and the search for big data quickly depends on finding the needle in the haystack. There are various problems related to each phase of big data management that can only be transmitted using high-end big data computing solutions.

This is why, to provide adequate solutions to improve public health, health professionals must be fully equipped with adequate facilities to effectively create and evaluate big data. The correct management, analysis and interpretation of big data can change the game by opening up new avenues for modern health. This is why various sectors, including the health sector, are taking important steps to turn this opportunity into better services and financial benefits. With a strong combination of biomedicine and healthcare, modern healthcare organizations can revolutionize medical and personal medicine.

Introduction

The experience is the best way to design and innovate. As we get more information, we can be better organized to deliver the best results. A large amount of data analysis and its own results took place in many of the backbones of clinical practice as a promising parameter. Special features of large data than size (data acquisition rate), Flock (data acquisition as flood), Variables from data from different sources with different frequencies) vary their own estimates in many medical data areas. The dataset for health systems was developed as structured data: data that meets the type of data, procedures and as defined Examples of the types of data available in the health sector include sickness patterns, their diagnostic properties and information, clinical outcomes of the diagnosis, patient information.

The data were structure as of self-descriptive nature. Examples of that type of data can be the censored devices used to monitor behavior of patients. Data which is un structured does not prevail any structure, it can be the prescriptions given by medical assistant wrote down in natural human language and more. Therefore, analyzing health data to obtain valuable information from people having interests like (pharmacies, etc.) is very difficult. Idea behind developing such frameworks is an effective study of different health data sources to identify patterns and patterns of interest among large data sources. These structures promote solutions for health based on the model. which have led to active patient participation for their health care. In this paper, recent years have been explored as a final solution in the construction of healthcare facilities, healing plays a key role in transforming the knowledge that raw data is available.

Therefore, some latent information on behavior, technical, therapeutic and other types of improvements in health care. After reassessment of liver therapy - it appears that the full spectrum of Panzer behavior - related to medical expertise or alternative medicine is ongoing. A large number of analyzes of EHR data, EMR e and other med data are being done to help build a healthy ideology. Companies that provide quality assurance services to the translation. Diagnostics and clinical help make

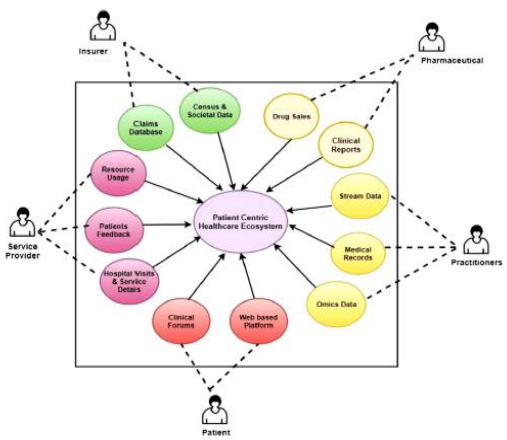


Fig. 1. A patient centric healthcare ecosystem – From the perspective of big data.

it better and more efficient. Although medical information is complex in nature, they demonstrate trust and reliability and are as important as the ease of information delivery, validation of communication across various disciplines, decision making of health research metrics. it requires special equipment and materials.

Many variations are found in the health of the computer environment such as the use of specimens from specimens with the aim of monitoring the disease, collectively, which has not been preserved by modern technology compatible with huge and massive amount of incomprehensible data. Link of image (https://www.google.com/search?q=big+data+patient+centric+framework&source=lnms&tbm=isch&sa=X&ved)

Previous Work

The born organism, like the human cell, represents the molecules and the physical elements that defect in them. Quantitative or quantitative research often collects small data and / or a simple measure to understand the complex correlations of different phenomena and events.

Consequently, more experimental data are needed to provide a more complete map of the biochemical source. This means that the more we know, the better our understanding of living events. With this in mind, modern methods have evolved faster and faster.

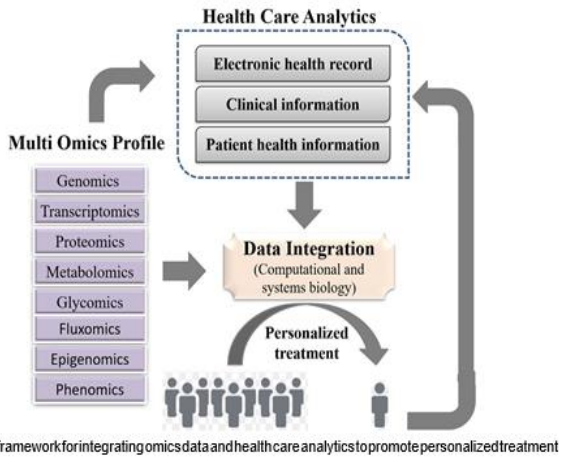
For example, much information gained from the integration of advanced technologies, such as (NGS) and (GWAS) to map the genome, can be thought of. life.

NGS provides detailed data that is yet to be published and requires a new level of testing.

Big data from omics studies

.NGS makes process easier, reduces production cost of genome duplication. Required cost to buying a complete particles of genome has dropped ($ 20 million - $ 20 million. NGS technology has been a biomedical data abounding source of genomic and transcriptomic studies. It is estimated that the total population of the 203 genome to (100- 2 million). If combine genome and metabolic data will help in gaining knowledge of patient's genome. it is called "private, special or public health".

The systematic data analysis in combination, health to improve the .treatment step by step and in the actual treatment process. The main source for producing big data on health information biomedical as well as EMRs, chemical information and insurance quotes is genome-directed events. Link of image(https://www.google.com/url?sa=i&source=images&cd=&ved=2ahUKEwjXnI345LXmAhUN2)



A framework for integrating omics data and healthcare analytics to promote personalized treatment

Although medical information is complex in nature, they demonstrate trust and reliability and are as important as the ease of information delivery, validation of communication across various disciplines, decision making of health research metrics. it requires special equipment and materials.

Quality assessment requires for integrating a robustness in data of biochemical related to different fields for optimal treatment. These type of opportunities are playing important role to compete change a lot in many patients, marketing agencies are using demographic data to help them make treatment decisions. This could be the future of sports in medicine and heal.

Therefore, some latent information on behavior, technical, therapeutic and other types of improvements in health care. After reassessment of liver therapy - it appears that the full spectrum of Panzer behavior - related to medical expertise or alternative medicine is ongoing.

A large number of analyzes of EHR data, EMR e and other med data are being done to help build a healthy ideology. Companies that provide quality assurance services to the translation. Diagnostics and clinical help make it better and more efficient. *Link for the table defined in reference section.*

Table 2
Summary of healthcare big data frameworks.

Authors	Framework Name	Data Source	Analytical Capability	Application Area	Highlights
Raghupathi and Raghupathi (2014)	Generic Conceptual framework.	Geographically disparate data sources with multiple formats.	Queries and Reports.	General Healthcare architecture	Advocates a benchmarking framework for analyzing Healthcare Big Data
Chawla and Davis (2013)	A patient centric personalized healthcare framework	Electronic medical records, patient experiences, and histories	Collaborative filtering	Personalized healthcare	Data driven computational aid for personalized healthcare
Kim et al. (2014)	Big Data framework for Ubiquitous healthcare system	Continuous time series data such as ECG, respiration, and SpO2	Map-Reduce programming model	Personalized healthcare	Big data platform for processing vital signs
Fang et al. (2016)	Health informatics processing pipeline framework	Electronic Health Records, Public health, Genomic, Behavioural data	Feature selection and machine learning algorithms	Decision support system for practitioners	Systematic data processing pipeline for generic big health informatics
Youssef (2014)	Framework for secure health information system	Electronic Health Records	Map-Reduce	Security and privacy of healthcare data	Integrating mobile, cloud and big data technologies to provide personalized healthcare
Lin et al. (2015)	A cloud based framework for home diagnosis service	Patient data, patient profile and clinical data	Formal Concept Analysis (FCA)	Self-care diagnosis	Symptom-based medical record retrieval according to the user query
Legaz-Garca et al (2016)	A semantic web technology framework for managing and reusing clinical archetypes	Electronic Health Records	Ontology building through OWL.	Patient classification system based on clinical criteria	Integration of semantic resources with Electronic Health Records
Sakr and Elgammal (2016)	Smart Health Framework	Patient data from sources such as Hospital Information System, Radiology and laboratory information systems	Predictive modeling and pattern matching techniques	Data analytics for smart healthcare applications	Enhancing healthcare services by integrating sensor technologies, cloud computing and big data analytics
Mahmud et al. (2016)	Data analytics and visualization framework for health-shocks prediction	Healthcare data set focusing on context such as socioeconomic, cultural and geographical conditions	Predictive modeling using Fuzzy rule summarization	Public health services	Integrating cloud computing services with geographical information system
Jokonya (2014)	Big data integrated framework for forecasting and controlling diseases in Mining Industry	Epidemics data about infectious diseases	Predictive analytics	Disease forecasting and prevention	Holistic and structured approach to understanding the complex linkage of HIV/AIDS, tuberculosis and silicosis in mining industry
Rahman et al. (2017))	Framework for preserving privacy in RFID based healthcare systems	Data generated from RFID tags	Privacy preserving techniques	Secure healthcare services	Improved privacy in RFID based healthcare system
(Sarkar, 2017	Conceptual framework for secured distributed health information system	Electronic Health Records	Provisioning security constraints and access control mechanisms	Secure healthcare system	Distributed framework for two-level security mechanism
Hossain and Muhammad (2016)	Voice Pathology Assessment Framework	Voice Signals	Classification Techniques	To classify the signal as normal or pathological.	A cloud based big data analytical framework for handling unstructured data
Pramanik et al. (2017)	Big Data enabled smart healthcare system framework	EHR, diagnosis report, social media data, surveillance data, biometric data	Provisioning smart healthcare services through service oriented infrastructure	Smart system technologies for state of the art healthcare system	Coupling of healthcare knowledge discovery mechanism with smart service infrastructure

Existing Problems and its Solutions

Methods of data management and analysis are becoming increasingly streamlined to support real-time data, retrieval, scaling and visual information using in integrating the effective use of EMR. Approved by federal administration standard. Consequently, EHR sponsored by state, each with varying degrees of disease burden, technology interpretation, and operational efficiencies have led to problems in collaboration and data sharing.

Variables from data from different sources with different frequencies) vary their own estimates in many medical data areas. The dataset for health systems was developed as structured data: data that meets the type of data, procedures and as defined Examples of the types of data available in the health sector include sickness patterns, their diagnostic properties and information, clinical outcomes of the diagnosis, patient information.

1. *Data Storage.*

One major challenge is storing data, many benefits such as storage, access. Most health care companies have successfully utilized cloud storage solutions using IT infrastructure, due to reduced costs and increased reliability. Teams may have a hybrid system for database flexible & efficient system for storing data.

2. *Cleaning*

The data must be cleaned or folded to ensure accuracy, order, consistency, consistency and

purity after discovery. These purification techniques can be e-book or automatic using logical rules to ensure maximum accuracy and reliability. Extensively using algorithms related to .machine learning for control time, expense and block sensitive information from large sets containing data.

3. Unified format

The data generated by Patients in very huge amount is not easy to manage along the standard of EHR, since they are not already organized. Too difficult to manage huge amount of data when not having enough data centers for health officials.

The need to document all clinical information is appropriate for communication, payment purposes and clinical analysis. Therefore, the treatment of coding systems such as procedural terminology (CPT) and international coding codes for classified diseases (ICD) have been representing most important continental information. Therefore, the sets are limited.

4. Correctness

It is found that proportion in data of EMR or EHR patient is less accurate but, partly because of poor EHR quality, limited internal consistency and misunderstanding why it is valuable. save all big data. These are all things that can describe the big data that is fun with both of his life.

The EHR aims to make quality of data much better. Writing style can be improved by using label questions that are not appropriate for the patient's symptoms.

5. Processing the Images

It is identified that many body types leading the translation of types of data & misinterpret ate from artificial distortions involving many noise create having different forms, lead to image defects such as, which can lead to the destruction of internal organs because the nerves do not correspond to the real case. Noise reduction, elimination of holes, adjustment of the image contrast obtained and resolution of the post-mistreatment image are just some of the steps that can be introduced to help the lens.

6. Security

There are so many guarantees, cases, phishing attacks and software that state if data protection a top priority for organizations working as healthcare. Following monitoring of the level of difficulty, a list of professional safety information was obtained with information on (.PHI), HIPAA's. Safety, helping many centers to maintain, transfer, validate and regulate proteins, integrity and monitoring. Safety precautions such as the use of updated anti-virus software, wall art, data encryption and multi-faceted can save many problems.

7. Visualization

A clear and concise view of data and data, a heat map and a histogram showing conflicting images and an organized display of information to reduce the risk of misunderstandings, makes easy to obtain information/data to correctly use it. Like bags in pots, ticket packs and glossaries that have their own way of displaying data. Although medical information is complex in nature, they demonstrate trust and reliability.

Analytical avenues in a typical patient-centric healthcare system.

Stakeholders	Sources of big data	Nature of analytics	Underlying analytical techniques
		Classification of patient communities	Machine learning algorithms, statistical modeling techniques, sentiment analysis
Patients	Clinical forums and Telemedicine platforms	Health text analytics	Information retrieval techniques such as text mining
		Construction of health ontologies	Crawler based algorithms, semantic analytics
		Patient network analysis	Recommendation techniques based on collaborative filtering
	Genomic database	Genomic sequence analysis	Association rule mining and visualization techniques
Medical Practitioners	Electronic Health Record (EHR)	Categorization of patients based on personal statistics	Classification techniques based on machine learning algorithms
	Personal Health Record of patients	Identification of high-risk patient for specialized treatments, designing patient specific treatment plans, offering evidence based guidelines	Optimization and simulation techniques based on prescriptive analytics
	Device Generated data	On-demand real-time health analysis of patients	Analysing data streams by sampling and filtering techniques
Hospital operators	Electronic Health Record of patients, treatment plans	Prediction of inpatient duration of stay	Resource management and optimization techniques based on regression models and artificial neural networks
	Manpower utilization reports	Discovering strategies for resource planning and utilization	Optimization techniques such as evolutionary algorithms, PSO algorithms
	Patient's feedback	Computation of patient satisfaction scores based on demographical characteristics	Approaches based on user based and service based collaborative filtering
	Instrument usage log, calibration details	Detection and prediction of faults in medical devices.	Decision tree induction, Bayesian Classifier
Pharma and clinical researchers	Daily drug sales reports	Identification of usage and purchase patterns of drugs	Association rule mining, Time series analysis, Outlier mining
	Clinical reports, Patients Health Records	Amino acid sequence analysis, structure prediction of proteins for effective drug design	Deep learning techniques based on big data frameworks such as Hadoop and Spark

disciplines, decision making of health research metrics. it requires special equipment and materials. *Link for the table defined in reference section.*

Future Directions

The systematic data analysis in combination, health to improve the .treatment step by step and in the actual treatment process. The main source for producing big data on health information biomedical as well as EMRs, chemical information and insurance quotes is genome-directed events. Although medical information is complex in nature, they demonstrate trust and reliability and are as important as the ease of information delivery, validation of communication across various disciplines, decision making of health research metrics. it requires special equipment and materials. There is a need to analyze data complexity from the reproductive system. Goal behind it the conversion of huge amount of data into information. Implement the bioengineering approach to transfer biomedical and genomic data into an imaging and control system known as bioinformatics.

It is facing data-driven networks. Different types of quantitative data in preventive care, such as metadata to provide better treatment than options. treatment [25]. That's why new technologies are emerging that are needed to help analyze this digital economy. In fact, over $ 100 million of goals have been established such as "Big Data" research to improve quality data tools with better organizational practices, better access and more sensitive data analysis. *Link for the table defined in reference section.*

Table 3
Big Data Tools for Healthcare Ecosystem.

Task	Tools	Merits & its applications
Data Integration	Pentaho (2017)	Tool for performing knowledge discovery process from a scalable environment in a robust and flexible manner.
	Palantir (2017)	Assist decision makers for highlighting the process insights by uncovering treatment options and improving the standard of patient care.
	Ayata (2017)	Performs exclusive prescriptive analytics from large amount of data towards helping organizations for making smarter decisions.
	Attunity (2017)	Modern data integration platform for performing automated data pipelines at a faster rate from different architectures such as cloud and data lakes.
	Informatica (2017)	Provides enterprise information management solutions over data from diversified sources
	Jitterbit (2017)	Facilitates integration of data from Saas based cloud services and on-premise applications in an intelligent manner.
Searching and processing	Apache Lucene (2017)	High performance, full-featured text search engine for performing full-text search across different platforms.
	Google Dremel Melnik et al. (2010)	Complements Map/Reduce based computations supported by Hadoop for processing nested data with high scalability.
	Cloudera Impala (2017)	Executes low latency and high concurrency analytical queries on top of Hadoop.
	Dryad Isard et al. (2007)	Provides high performance distributed execution engine with good programming constructs.
Machine Learning	Apache Mahout (2017)	Offers distributed machine learning library for processing scalable mining algorithms.
	Skytree (2017)	Tool for performing machine learning and advanced analytics of massive data sets at high speed.
	Karmasphere (2017)	Big data workspace tool for discovering pattern insights from large volume of data stored on hadoop clusters.
	BigML (2017)	Platform for offering solutions to big data use-cases through predictive analytics.
Stream Data Processing	Apache Storm (2017)	Scalable and flexible real-time computation system for processing massive amount of data.
	S4 Neumeyer et al. (2010)	Tool for processing unbounded stream data in a distributed, scalable and fault-tolerant manner.
	SQLstream Blaze (2017)	Supports creation of distributed streaming applications that deliver data ingestion, integration, and analytics in real time.
	Splunk (2017)	Highly scalable tool to collect and harness machine data. Ability to scale from laptop to datacentre.
	Apache Kafka (2017)	A distributed, high throughput stream data processing tool for facilitating publish-subscribe messaging system.
	SAP Hana (2017)	Tool for in-memory computing of stream data for real-time analytics.

The development of 'omics' data from the large Genome Project and other population-based projects has many benefits. In the identification of an overall genetic population such as 1000 genomes, these candidates will have an incredible advantage over the amount of compressed data. Similarly, the Encyclopedia-based Bantu Genome of DNA Elements (ENCODE) aims to identify all functionalities in the human genome using bioinformatics methods. Here, we list some of the tools based on bioinformatics most used to make large data analytics than omics data.

Spark Seq is an effective and ready-to-use platform based on the structure of Apache Spark and the Hadoop library that is used to analyze genome data for collaborative analysis of genomic data with negative accuracy

SAMQA detects errors and detects large-scale genome data. This tool is based on the National Atlas Genome Cancer Program to identify and report errors including sequence/map error modules [SAM] and illiteracy. **Dist Map** is another tool that will be used in reading diffuse maps based on Hadoop gloss which aims to include various tracking features. For example, one application called a BWA mapper can produce 500,000 readings in 6 hours, about 13 times faster than a normal single node mapper.

The **Seq Ware** application engine is based on the Apache HBase database system that allows access to large genome databases by combining browsers and genome tools.

Cloud Burst is the only computational model used in genome mapping to try to solve large-scale sequencing data boundaries.

Hydra uses a distributed Hadoop computer system to validate a large peptide database and a spectrum of protein databases. This instrument is able to produce 27 billion peptide scores in less than 60 minutes on the Hadoop cluster.

Blue SNP is an R package based on the Hadoop platform used for analysis of association studies (GWAS) at the genome level, which focuses on quantitative data to find significant correlations between genotype-phenotype data. The effectiveness of this instrument was estimated to evaluate 1000 phenotypes at 106 SNPs in 120 people over a half hour period.

Myrna, based on cloud-based pumps, provides information on gene expression numbers, including read alignment, data standardization, and sample numbers. In recent years, entertainment platforms have increased the amount of disease data. For example, the Array Express Support Archive for the Genomics data warehouse contains information from over 30,000 light sources and over one million assets.

Supplementary data requires better and more efficient bioinformatics packages to analyse and interpret the information obtained. This has created some tools for evaluating measures data. Below are some of the most popular advertising platforms for analysing big data.

Discussion

Health services require strong combination of different sourced data provide poor care. This perspective is so interesting that even a lot of patient genomic data has to be written, but marketing and dealing with suppliers make therapeutic decisions. This could be a turning point in medicine and health in the future. Although medical information is complex in nature, they demonstrate trust and reliability and are as important as the ease of information delivery, validation of communication across various disciplines, decision making of health research metrics.

It requires special equipment and materials. Variables from data from different sources with different frequencies) vary their own estimates in many medical data areas. The dataset for health systems was developed as structured data: data that meets the type of data, procedures and as defined Examples of the types of data available in the health sector include sickness patterns, their diagnostic properties and information, clinical outcomes of the diagnosis, patient information.

Data management and analysis methods have become leaner to support data, recovery, resize, scaling, analysis in real time in helping integrating effective use of EMR health care. .

Conclusion

At present, various biomedical and health care systems such as app sensors provide large amounts of data. Therefore, some latent information on behavior, technical, therapeutic and other types of improvements in health care. After reassessment of liver therapy - it appears that the full spectrum of Panzer behavior - related to medical expertise or alternative medicine is ongoing. A large number of analyzes of EHR data, EMR e and other med data are being done to help build a healthy ideology. Companies that provide quality assurance services to the translation. Diagnostics and clinical help make it better and more efficient.

These companies' overall goals include reducing drug costs, developing a cynical decision support system (CDS), providing a good plan for treatment, identify and preserve huge data frauds. Although, everyone faces problems with federal resources such as processing, distribution, and data security.

Although medical information is complex in nature, they demonstrate trust and reliability and are as important as the ease of information delivery, validation of communication across various disciplines, decision making of health research metrics. it requires special equipment and materials.

The combined dataset of health research and biomedical organizations has led to better monitoring, identification and treatment of various diseases. It also helps to build a health system that is better and healthier. Modern healthcare professionals have seen the big data opportunity and so on have taken a major leap forward in analyzing data in behavioral health. Supercomputers on a large number of computers help extract good info relevant to huge data in a very short period. Providing exact and relevant information to improve existing health conditions, researchers have entered big data despite construction

problems. Clinical trials, e-commerce tests and insurance together, acquiring nano markers is part of the vision and creative process for analyzing key health data. Variables from data from different sources with different frequencies) vary their own estimates in many medical data areas. The dataset for health systems was developed as structured data: data that meets the type of data, procedures and as defined Examples of the types of data available in the health sector include sickness patterns, their diagnostic properties and information, clinical outcomes of the diagnosis, patient information. Large-scale data analysis adds errors to the internal structure of a planned and unplanned system. Moving to an open dashboard is an obstacle that is known to be overcome.

In short, the core information of a huge amount information, which can extract a lot of information from these data and can predict future events.

It seems well-led by healthcare leaders and healthcare companies that the broader health market is designed to grow even more. In fact, in the brief we see a series of studies that are used that show the most influential decision-making results and the health component to be executed. The sheer volume of clinical data in various fields has forced computer experts to develop methods for analyzing and interpreting such large amounts of data over a period of time.

The combination of universal computerization techniques from both burners and doctors has proven very useful. As such, the focus will be on building a globally collected with "human" technology. A concept can improve the conditions of patients suffering from diseases by using tools. We have to obtain the best genome data, including the residual and confidential results from experiments and genetic experiments. Consequently, there is potential at every step of the multi-tasking process to prevent the development of trends in investment.

References

1. Laney D. 3D data management: ontrolling data volume, velocity, and variety, Application delivery strategies. Stam- ford: META Group Inc; 2001.
2. Mauro AD, Greco M, Grimaldi M. A formal definition of big data based on its essential features. Libr Rev. 2016;65(3):122–35.
3. Gubbi J, et al. Internet of Things (IoT): a vision, architectural elements, and future directions. Future Gener Comput Syst. 2013;29(7):1645–60.

4. Doyle-Lindrud S. The evolution of the electronic health record. Clin J Oncol Nurs. 2015;19(2):153–4.
5. Gillum RF. From papyrus to the electronic tablet: a brief history of the clinical medical record with lessons for the digital Age. Am J Med. 2013;126(10):853–7.
6. Reiser SJ. The clinical record in medicine part 1: learning from cases*. Ann Intern Med. 1991;114(10):902–7.
7. Reisman M. EHRs: the challenge of making electronic data usable and interoperable. Pharm Ther. 2017;42(9):572–5.
8. Murphy G, Hanken MA, Waters K. Electronic health records: changing the vision. Philadelphia: Saunders W B Co; 1999. p. 627.
9. Shameer K, et al. Translational bioinformatics in the era of real-time biomedical, health care and wellness data streams. Brief Bioinform. 2017;18(1):105–24.
10. Service, R.F. The race for the $1000 genome. Science. 2006;311(5767):1544–6.
11. Stephens ZD, et al. Big data: astronomical or genomical? PLoS Biol. 2015;13(7):e1002195.
12. Yin Y, et al. The internet of things in healthcare: an overview. J Ind Inf Integr. 2016;1:3–13.
13. Moore SK. Unhooking medicine [wireless networking]. IEEE Spectr 2001; 38(1):
14. Nasi G, Cucciniello M, Guerrazzi C. The
15. role of mobile technologies in health care processes: the case of cancer supportive care. J Med Internet Res. 2
16. 1.5.3. 2017.Shvachko K, et al. The hadoop distributed file system. In: Proceedings of the 2010 IEEE 26th symposium on mass storage systems and technologies (MSST). New York: IEEE Computer Society; 2010. p. 1–10.
17. Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters. Commun ACM. 2008;51(1):107–13.
18. Zaharia M, et al. Apache Spark: a unified engine for big data processing. Commun ACM. 2016;59(11):56–65.
19. Gopalani S, Arora R. Comparing Apache Spark and Map Reduce with performance analysis using K-means; 2015.
20. Ahmed H, et al. Performance comparison of spark clusters configured conventionally and a cloud servicE. Procedia Comput Sci. 2016;82:99–106.
21. Saouabi M, Ezzati A. A comparative between hadoop mapreduce and apache Spark on HDFS. In: Proceedings of the 1st international conference on internet of things and machine learning. Liverpool: ACM; 2017. p. 1–4.

PLAGAIRISM REPORT

Found 1% match (on links of images inserted), checked through Turnitin

turnitin

Digital Receipt

This receipt acknowledges that Turnitin received your paper. Below you will find the receipt information regarding your submission.

The first page of your submissions is displayed below.

Submission author:	HIFSA BASHARAT
Assignment title:	Fyp 3
Submission title:	Survey Report Assignment2
File name:	2.docx
File size:	849.11K
Page count:	13
Word count:	3,801
Character count:	21,368
Submission date:	15-Dec-2019 01:07AM (UTC+0500)
Submission ID:	1234507572

Assignment Inbox: Fyp 3

Assignment Title	Info	Dates	Similarity	Actions
Fyp 3	ⓘ	Start 13-Dec-2019 11:04PM Due 25-Dec-2019 11:59PM Post 21-Dec-2019 12:00AM	1%	Resubmit View