A Survey on Data Warehousing in Health Care Service

MD. ENAMUL HAQUE, United International University, Bangladesh
DEWAN AFSANA PARVIN, United International University, Bangladesh
IMTIAAZE MAHMOOD, United International University, Bangladesh
MD. ANISUR RAHMAN, United International University, Bangladesh
MD. SADDAM HOSSAIN, United International University, Bangladesh

A healthcare data warehouse is a centralized repository of data related to healthcare domain. Generally healthcare data warehouse integrates data from ERP, EHR/EMR, CRM, claims management system, pharmacy management systems, routine health information system, disease management system, surveillance system or any other system directly or indirectly related to healthcare domain. Data warehouse design and implementation in healthcare domain is complex in nature due to the diversity of data in healthcare. Also several individual silo systems developed in healthcare domain most of the time doesn't follow interoperability standards and architecture. We have studied some papers related to data warehouse architecture development and implementation in healthcare domain, identified some key challenges to implement a data warehouse in healthcare, summarizes the reviewed papers using different attribute and finally we propose a data warehouse framework along with a business intelligent (BI) real-time dashboard from that survey papers.

CCS Concepts: • **Data warehouse, Health Care, BI, Data ,State Indicators, OLAP, ETL.**;

Additional Key Words and Phrases: Data warehouse, Health Care, BI, Data Cube, State Indicators, OLAP, ETL.

ACM Reference Format:
Md. Enamul Haque, Dewan Afsana Parvin, Imtiaaze Mahmood, Md. Anisur Rahman, and Md. Saddam Hossain. 2022. A Survey on Data Warehousing in Health Care Service. 1, 1 (February 2022), 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Data warehousing (DW) is a way of obtaining and organizing data from many sources in order to get useful actionable insight. Commonly, a data warehouse is used to link and analyze corporate data from diverse sources. The data warehouse is the central component of a business intelligence (BI) system, which is intended to reflect and report on data.

A healthcare data warehouse is a concentrated store for data gathered from many sources, processed, and formatted for analytical querying and reporting in a healthcare institution. A data lake, machine learning, and business intelligence technologies are all integrated into the healthcare data warehouse. It combines data from a variety of sources in a healthcare system, including electronic medical records, claims, supply chains, cost accounting systems, and more. It enables healthcare companies to assess a wide range of medical conditions, care

Authors' addresses: Md. Enamul Haque, mhaque202016@mscse.uiu.ac.bd, United International University, Gulshan, Dhaka, Dhaka, Bangladesh, 1213; Dewan Afsana Parvin, dparvin193028@mscse.uiu.ac.bd, United International University, Gulshan, Dhaka, Dhaka, Bangladesh, 1213; Imtiaaze Mahmood, imahmood202039@mscse.uiu.ac.bd, United International University, Gulshan, Dhaka, Dhaka, Bangladesh, 1213; Md. Anisur Rahman, mrahman202047@mscse.uiu.ac.bd, United International University, Gulshan, Dhaka, Dhaka, Bangladesh, 1213; Md. Saddam Hossain, saddam@cse.uiu.ac.bd, United International University, Gulshan, Dhaka, Dhaka, Bangladesh, 1213.

XXXX-XXXX/2022/2-ART
https://doi.org/10.1145/nnnnnnn.nnnnnnn

delivery procedures, and operations in a thorough and systematic manner, and then provide insights that lead to improvement decisions.

The healthcare data warehouse design includes three levels of data coarseness from oriented data used in generic report production to detailed entry-level information, like hospital discharges. The data warehouse design is viewed because of the data pyramid These 3 levels of aggregation among the info warehouse performance goals that mix to fulfill a wide variety of news.

At the end of our work, We will propose an architecture and it will be implemented in a web-based analytical application to increase data analytics efficiency, accuracy, and scalability. Two layers of growth for our approach are appropriate: Clinical Data Analytics and representation. At the clinical data analytics level, new healthcare datasets are added, more types of information are identified for target users, and a systematic quality assurance process is used to assure metadata quality. Reports, Statistics, Lookup, Information gathering and algorithmic methods to extract knowledge, good functionalities to compare similar datasets, and collaborative features, such as Clinical forums that allow users to help each other and suggest healthcare clinical datasets, are all examples of data representation at the data level.

2 RELATED WORKS

Healthcare data warehousing represents distinctive challenges. The trade is rife with medical reports and coding schemes, several of which are incompatible and need careful data format. Healthcare data comes from several sources and is delivered in several forms, as well as revealed books, individual spreadsheets, and several other data formats. Here, we have a tendency to ask to focus on the analysis and development decisions created in constructing a data warehouse, with a stress on the important topic's data staging and quality assurance.[1]

This article presents the findings of a research on how healthcare organizations have approached the construction of information systems for a broader range of reporting than typical financial data. To that purpose, a few basic concepts related to management information in healthcare organizations are gone through. Second, a rundown of various business intelligence tools that could be used in the construction of a computer-based management information system. [2]

The importance of Business Intelligence (BI) in healthcare and its critical aspects are discussed in this study. It is an attempt to distinguish between typical BI approaches and those required in healthcare. It also discusses the distinct character of clinical data. Data Warehouse (DW) is an essential component of Business Intelligence [3]

3 METHODOLOGY

3.1 Selection Criteria

At first, the search keywords are decided which was used for selecting literatures for our review paper. The 'AND' and 'OR' syntax are used while searching. 'AND' here defines the word that was surely used for searching and 'OR' describes any of the word can be chosen from the selected words.

3.2 Search Keywords

- "Data" AND "Warehouse"
- "Healthcare" AND "Data" AND "Warehouse"
- "Data" AND "Warehouse" AND "Business" AND "Intelligence"
- "Healthcare" AND "Business" AND "Intelligence"

After searching for the literatures, Some criteria for both inclusion or exclusion were selected. We have selected the criteria based on availability, best approaches, and better output and overall which make our path easier for working on this topic. The exclusion criteria saved time and improved the focus for which we were actually working on.

TABLE: Inclusion and exclusion criteria for selection of literature

Inclusion and Exclusion Criteria			
IC1	The studies which are focused on dedicated healthcare based explanations	EC1	Duplicate articles obtained
IC2	The studies which are based on healthcare data warehouse	EC2	Exclude the articles which are on data warehouse but did not include healthcare
IC3	Related studies must be in English language	EC3	Excludes the studies on other languages
IC4	The articles or data must be from after 2000 and till present	EC4	Excludes the studies not containing structural procedures

We will at first discuss about the basic framework on data warehousing. A data warehouse architecture is a way of thinking about data mining, data processing with ETL and OLAP, and calculating the results of data sources, data models, and statistical reports.

The following section will continue describing different types of healthcare data such as:

- Clinical Data
- Patient-Generated Data
- Cost and Utilization Data
- Public Health Data

A small table containing the framework and data model description will be provided for better understanding. Then after providing a comparison table and gap analysis of different reference papers, We will approach for our proposed model.

Our proposed model will be represented with a proposed architecture and schema diagram of data flow for understanding the overall procedure.

4 TYPES OF HEALTHCARE DATA

Acquiring data, converting it using sophisticated analytics to determine what important, and alerting clinicians and other stakeholders what and how to improve are all steps in the process of creating a successful clinical registry.

4.1 Clinical Data

Many registries rely heavily on clinical data. It has a strong capacity to synthesize clinical encounters and collect data that may be used to guide healthcare advancement. Demographics, family history, comorbidities, procedure and treatment history, and results are all used to determine who a patient is. Clinical data's breadth and depth

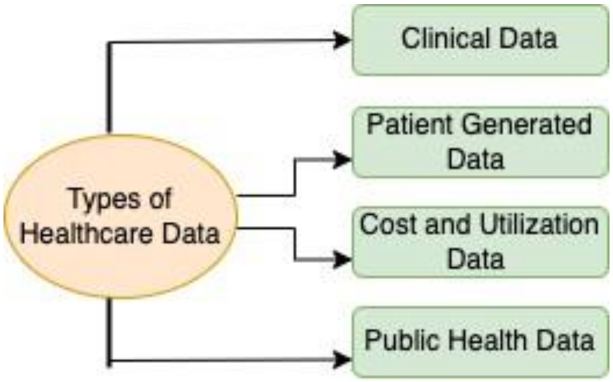

Fig. 1. Healthcare Data Type

provide the potential for quality improvement, research, registry-based studies and virtual trials, and other stakeholder activities.

4.2 Patient-Generated Data

Data from outside the exam room or treatment facility is frequently needed to drive high-value healthcare. Patient-generated data may provide your registry with a wealth of healthcare information. Patient-reported outcomes are a major source of patient-reported data (PRO). PRO are patient-provided health data that have not been interpreted or altered by a doctor or anybody else.

4.3 Cost and Utilization Data

Many healthcare and clinical registry projects are motivated by value-based care initiatives, which aim to improve outcomes while keeping costs low. These initiatives need comparing the cost of delivering health results against the cost of delivering those outcomes. Health insurers, governmental entities, and public payers are all important sources of cost and use data. They include public datasets from organizations such as the Centers for Medicare and Medicaid Services (CMS) and the Agency for Healthcare Research and Quality (AHRQ), as well as claims data on patient treatments (AHRQ).

4.4 Public Health Data

Clinical treatment and patient health behavior are simply two of the many elements that have a substantial impact on healthcare outcomes. Community and population health variables are responsible for up to 50 percent of health outcomes. As a result, attempts to improve health care need a comprehensive understanding of all factors that impact health outcomes. This perspective is essential for comprehending the paths that will lead to the advancement of health care.

5 DATA WAREHOUSE ARCHITECTURES

A data warehouse architecture is an approach of con- cern with analysing of data mining, data processing using ETL and OLAP and computing the result of data source, data model and represent data to statistical reports.

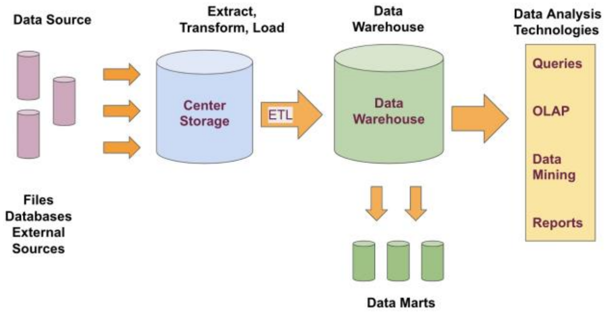

Fig. 2. Data Warehouse Architecture

6 ANALYSIS

Gaps identified and some challenges to implementing data warehouse in healthcare:

Seq	Ref	Author	Methodology	Design	De	System Backup	Security
1	[10	Nicolas		Top–Down	1		Y
2	[9	lai	Y	Top–Down		Y	Y
3	[11	Joh	Y	Top–Down	1		
4	[12	Lekha	Y	Top–Down			
5	[13	Kislaya	Y	Top–Down	7	Y	Y
6	[14	Christine		Top–Down			Y
7	[15	Nicolas		Top–Down			
8	[16	Barrett		Top–Down			

Ref	Data				ETL Tool	Purpose		
	Size	Availabi	Privacy	Quality		Administra	Managemen.	Res
[10]	>15m		Y	Y		Y	Y	Y
[9]	>100m	Y	Y	Y	ODI	Y	Y	Y
[11]		Y		Y		Y	Y	Y
[12]		Y			SSIS	Y	Y	Y
[13]	>3m	Y	Y	Y		Y	Y	Y
[14]	>99 m	Y	Y	Y	i2b2	Y	Y	Y
[15]					AT	Y	Y	Y
[16]	>4.4m	Y	Y	Y		Y	Y	Y

- In healthcare analytics, there are no well-defined matrices. This makes the DW design tedious and ambiguous.
- In most of the studied papers, no example of a real-time dashboard has been presented.
- Traditional hospitals and specialized centers differ a lot in measuring matrices or facts.
- An additional layer of complexity in terms of data confidentiality exists in healthcare data. Table: Methodology and System Perspectives
- Data Integration issue
- Data Integration issue is a big challenge as the healthcare domain is heterogeneous in nature.
- Data Integrity is a challenge.
- Handling unstructured data is a challenging and complicated task to make better sense of it. Semantic analysis is the way to handle unstructured data and extract relevant and useful information. Semantic technology has always been about the meaning of data, its context, and the relationships between pieces of information.

7 OUR PROPOSED ARCHITECTURE

The proposed architecture with the snowflake model consists of 4 major components:

- Data source (Clinical databases, healthcare datasets)
- ETL process and (Extraction, Transformation, Loading and Refresh)
- Data Warehouse
- Presentation of data.

With these elements, the system design will integrate and represent information from scattered datasets, enable versatile analysis queries, and supply precise answers at an acceptable level of comprehension. The snowflake model epitome is constructed on prime of the open-sourced data warehouse for information representation.

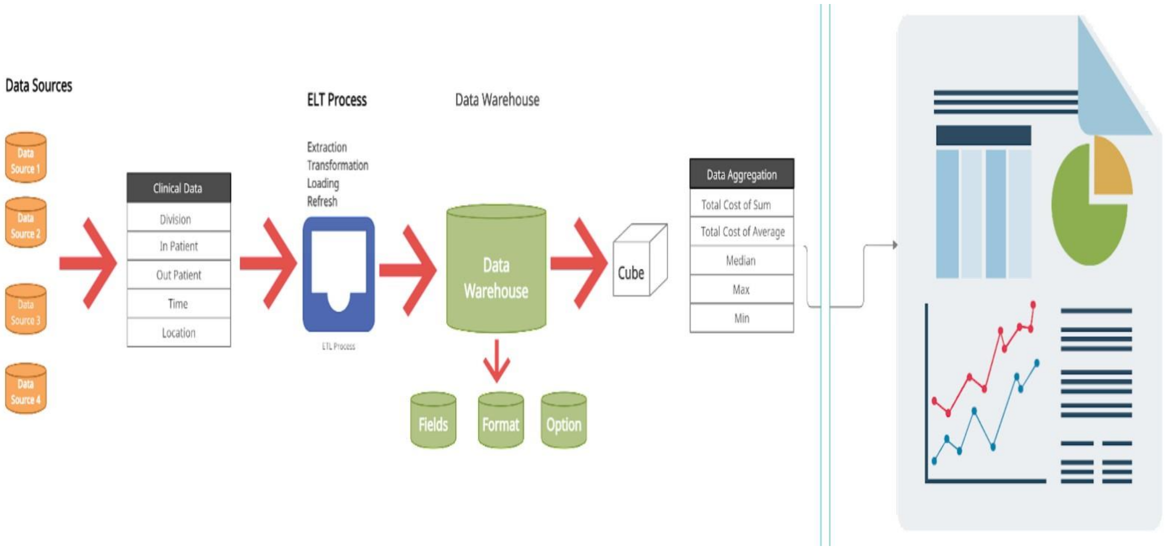

Fig. 3. Our Proposed Framework

Therefore, advantages of our proposed features like stability facing analytics data and significant traffic—and blessings of our design by exploitation properties, classes, and also the web-based open-source system will be absolutely tending to clinical analytical reports. Once the information of various datasets is extracted, Our architecture provides a platform for representation and an open-source web analytics tool.

8 DATASET

We have used a demo clinical database describing patient's information. It also states hospital's administrative cost and utilization data. The dataset describes patient's admitted department as well treatment cost and also finally patient satisfaction level. Using the data we prepared a real-time dashboard by our proposed architecture.

The dataset we have used here from flexmnoster.com. From this platform we have used their example healthcare dataset.

Division	Gender	Patient Birth Year	Patient Satisfaction	Cost
Cardiology	F	19..	Good	10000.00 BDT
Dermatology	M	19..	Good	10000.00 BDT
Oncology	F	19..	Excellent	10000.00 BDT
Neurology	M	19..	Excellent	10000.00 BDT
ophthalmology	M	19..	Negative	10000.00 BDT
Surgery	F	19..	Neural	10000.00 BDT

9 SNOWFLAKE MODEL

Our Proposed architecture is underway in a web-based analytical tool to improve the efficiency, accuracy, and scalability of data analytics. Suitable directions for our model expansion include two levels: Clinical Data Analytics and representation. The clinical data analytics level adds more healthcare datasets, identifies more types of knowledge for target users, and involves a systematic quality assurance method to ensure the quality of metadata. Representation of Data level includes Reports, Statistics, Query, Data mining and automated methods to extract knowledge, good functionalities to compare similar datasets, and collaborative features, such as Clinical forums that allow users to help each other and suggest healthcare clinical datasets.

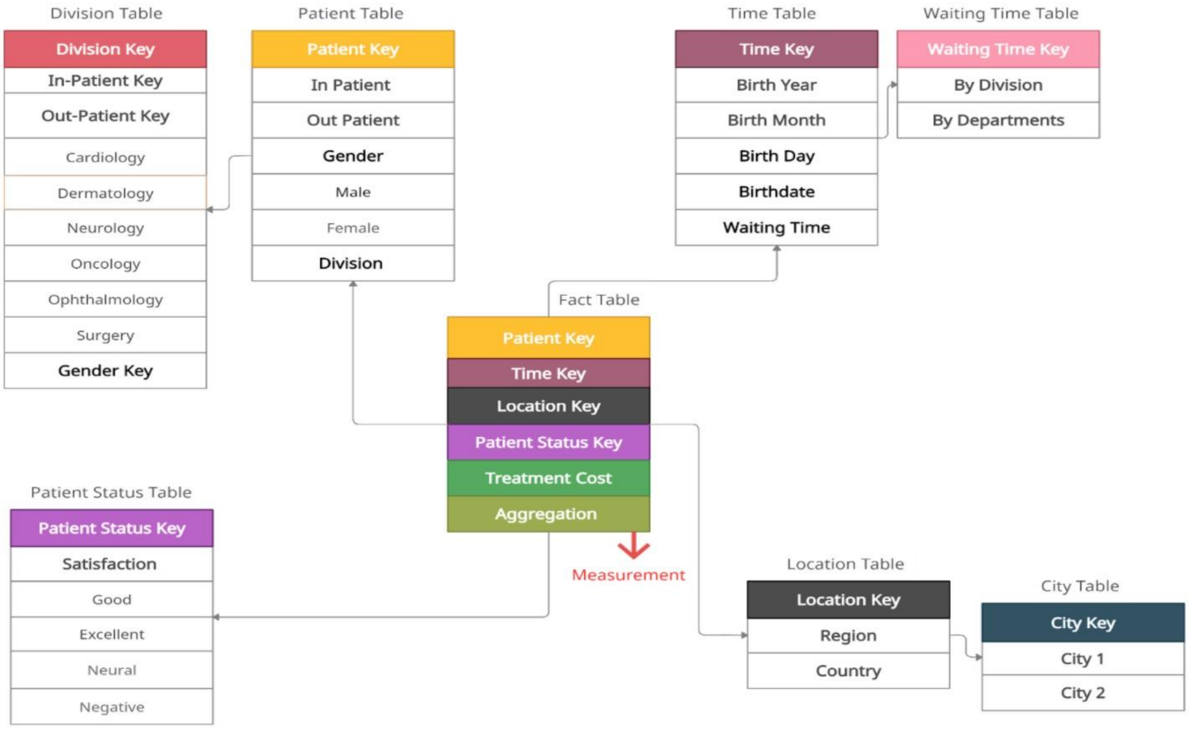

Fig. 4. Our Proposed Snowflake Schema Model

10 OPPORTUNITIES AND CHALLENGES

Lack of access to data Inability to produce timely reports Incomplete data Integration and Interoperability issues Volume of data

11 CONCLUSION

The healthcare system is one of the most aggregate structures of data warehouse and business Intelligence. Nowadays, the health system has become more competitive to the trade or industrial world. The data warehouse and BI is an analytical and statistical way to improve the healthcare system. Yet now, Healthcare organizations have not been able to exploit the full potential of big data analytics because of various challenges.Dearth of people with desired skills and expertise; complexity of healthcare domain; lack of knowledge of the benefits of technology; and resistance to adoption are some of the other challenges that impede the implementation of data warehouse and business intelligence. The elemental components required for successful healthcare data warehouse and business intelligence includes: Data inputs, Functional elements (data preparation, data processing, analytical model, visualization), Human element, and Security elements (network security, data protection, access control, policy formulation). Fundamental steps of data warehouse and business intelligence program for healthcare organizations comprises of:creating a strategic roadmap, clear definition of business use case, determining and procuring necessary resources, explicit definition of roles and responsibility of big data team, designing appropriate big data architecture, acquiring suitable tools and technology, framing data governance plan, migration from traditional technology to big data analytics, backing by management, culture change, analysis of data and visualization; and generation of insights and interpretation for intelligent application. We have analyzed frameworks, models from different articles, journal papers. Among these, we tried to filter out models and frameworks mostly used in the healthcare domain. We also tried to identify some gaps in existing study and finally we proposed a business intelligence dashboard solution. The dataset we used is imaginary data of a hospital.

Our future plan is to incorporate a national level dataset from Bangladesh public healthcare context and build a realtime dashboard to generate a forecast of public health issues.

REFERENCES

[1] Khan, S. I., Hoque, A. S. M. L. (2015). Development of national health data warehouse for data mining. Database Systems Journal, 6(1), 3-13.

[2] Spil, Ton Stegwee, Robert Teitink, C.J.A.. (2002). Business intelligence in healthcare organizations. 9 pp.. 10.1109/HICSS.2002.994108.

[3] Berndt, Donald J., John W. Fisher, Alan R. Hevner, and James Studnicki. ”Healthcare data warehousing and quality assurance.” Computer 34, no. 12 (2001): 56-65.

[4] Donald J. Berndt,Alan Hevner of University of South Florida ”Healthcare data warehousing and quality assurance” 2002.

[5] Thien-An Ngoc Nguyen and M-Tahar Kechadi, Arsalan Shahid (2021) “Big DataWarehouse for Healthcare-Sensitive Data Appli cations”

[6] Hamoud, Alaa, Ali Salah Hashim, and Wid Akeel Awadh. ”Clin ical data warehouse: a review.” Iraqi Journal for Computers and Informatics 44, no. 2 (2018)

[7] Gavrilov, Goce, Elena Vlahu-Gjorgievska, and Vladimir Trajkovik. ”Healthcare data warehouse system supporting cross-border inter operability.” Health informatics journal 26, no. 2 (2020): 1321- 1332.

[8] Poenaru, Cristina Elena, Daniel Merezeanu, Radu Dobrescu, and Eugenie Posdarascu. ”Advanced solutions for medical information storing: Clinical data warehouse.” In 2017 E-Health and Bioengi neering Conference (EHB), pp. 37-40.

[9] Karami, M., A. Rahimi, and A.H. Shahmirzadi, Clinical Data Warehouse: An Effective Tool to Create Intelligence in Disease Management. The health care manager, 2017. 36(4): p. 380-384.

[10] Garcelon, N., et al., A clinician friendly data warehouse oriented toward narrative reports: Dr. Warehouse. Journal of biomedical informatics, 2018. 80: p. 52-63.

[11] Chelico, J.D., et al. Designing a Clinical Data Warehouse Architecture to Support Quality Improvement Initiatives. in AMIA Annual Symposium Proceedings. 2016. American Medical Informatics Association. [36] Narra, L., T. Sahama, and P. Stapleton. Clinical data warehousing for evidence based decision making. in MIE. 2015.

[12] Kunjan, K., et al. A Multidimensional Data Warehouse for Community Health Centers. in AMIA Annual Symposium Proceedings. 2015. American Medical Informatics Association.

[13] Turley, C.B., et al., Leveraging a statewide clinical data warehouse to expand boundaries of the learning health system. eGEMs, 2016. 4(1).

[14] Garcelon, N., et al., Improving a full-text search engine: the importance of negation detection and family history context to identify cases in a biomedical data warehouse. Journal of the American Medical Informatics Association, 2016. 24(3): p. 607-613.

[15] Jones, B. and D.K. Vawdrey, Measuring Mortality Information in Clinical Data Warehouses. AMIA Summits on Translational Science Proceedings, 2015. 2015: p. 450.

[16] Delamarre, D., et al., Semantic integration of medication data into the EHOP Clinical Data Warehouse. Studies in health technology and informatics, 2015. 210: p. 702-706.

[17] Rinner, C., et al., A Clinical Data Warehouse Based on OMOP and i2b2 for Austrian Health Claims Data. Studies in health technology and informatics, 2018. 248: p. 94-99.

[18] Sheta, O.E.-S. and A.N. Eldeen, Building a health care data warehouse for cancer diseases. arXiv preprint arXiv:1211.4371, 2012.

[19] Khedr, Ayman, Sherif Kholeif, and Fifi Saad. "An integrated business intelligence framework for healthcare analytics." International Journal 7.5 (2017).