**Complications in healthcare integration models and correlated data infrastructure proposition**

Kyaw Thiha Naing

Department of Information Technology

HELP University

Metro Mall, Fraser Business Park, Kuala Lumpur

Malaysia

sam.kyaw.567@gmail.com

*Abstract: -* In healthcare systems, proper data integration models are necessary in order to provide swift treatment for patients. Without integrity and proper management of patient data, it can result in losing many lives due to unwanted delays in getting the necessary data. This study aims to solve this problem by looking at different data-related technological perspectives and discussing which is best suited for the healthcare sector. Multiple papers on different technological perspectives are reviewed to identify the underlying problems and how they can be tackled individually without getting drawbacks in return. Most impactful problems are highlighted and discussed extensively. The findings show that a data warehouse is the most viable option for tackling the highlighted problems due to its highly centralized infrastructure and data consistency. Elaborations are made on the viability of a data warehouse and how it can help healthcare systems in terms of effective data management.

*Keywords: -*Data Warehouse, Data Lake, Big Data, Database, Interoperability, Data Storage, Snowflake

**1 Introduction**

We now live in a time where quality healthcare provision is of utmost importance in the light of the Covid-19 pandemic that has left the world economy and physical accessibility in a substandard state. To ensure quality healthcare, the way medical-related data is stored and managed needs to be in an integrated system so that patients can clear out their confusion and in turn get their treatment effectively and promptly. We can think of establishing an integrated that is well-balanced and systematic by looking at data-related technological perspectives like a data warehouse, data lake, big data, and database models.

**1.1 Discussion on different technological perspectives**

If we were to look from the perspective of traditional database models, there are certain problems regarding how outdated records are managed as it causes inefficiencies in terms of the operability of the data system. There is also a highlight regarding how network communication is a problem when it comes to data management as remote access require frequent information update (Wilkes, G.J, Paul, E.S, A., P, 2003). Another point that is to be discussed is how interoperability is important in healthcare systems as healthcare services of high quality and low costs can be delivered between different countries. Furthermore, interoperability standards are necessary to be made to assure the privacy and security of the healthcare systems. The securer the healthcare systems are, the easier it is to manage them properly and effectively (Gavrilov, Vlahu-Gjorgievska & Trajkovik, 2019). The paper briefly stated how a data warehouse can be used to develop an interoperable data storage and management system. During this pandemic, it is necessary to attend to patients as swiftly as possible due to the emerging number of infected cases around the world. In order to do so, it is imperative to filter out query-relevant data out of an enormous pool of data through schemas that are built dependently on specific medical purposes. This in turn results in personalized medication in which the prescription of precise treatments and therapeutics that are well suited are given to individual patients after initially taking into consideration of all the related filtered data that influence the response to therapy to be received (Sarathkumar, Liu, Wang & Wang, 2018). Additionally, this results in less data being stored for patients which also reduces cost as more data means more storage, and more storage means more cost. The paper also continued to discuss how there are still many systems where medical data records require pre-processing before storing them which is problematic in a way that unwanted delays occur when accessing multiple data. In the era where different types of data are emerging, digitized information within the health care segment is developing greatly with information coming in from inside as well as outside sources, from versatile gadgets, wearable sensor gadgets. Electronic Wellbeing Records (EHR), Radiology pictures, recordings, clinical notes, social media, blogs, remote health monitoring gadgets, as well as newer forms of data such as imaging, sensor reading is additionally fueling the need for Big Data solutions to oversee silos of information accessible within the healthcare industry. Numerous amounts of data-structured, semi-structured and unstructured data are a characteristic that makes the health care data most challenging (Prabha & Anitha, 2015). Within the time of genomics, the volume of information being captured from biological tests and routine health care methods is developing at a phenomenal pace. This information trove has brought modern guarantees for disclosure in health care research and breakthrough medicines as well as new challenges in technology, administration, and dispersal of knowledge (Adibuzzaman, DeLaurentis, Hill & Benneyworth, 2017,2018). These two Big Data papers highlighted how there is already an overwhelming amount of data being taken in healthcare systems, therefore, resulting in more costs due to more storage and management problems due to data organization becoming more and more complicated. This paper discusses some of these problems and proposes an integration model to solve them.

**2 Problem Formulation**

Despite the fact that there are various backgrounds described regarding the problems above, only two of the most concerning aforementioned will be discussed. These two problems directly impact the integrity of the data systems in the healthcare aspect when proper attendance to patients for their wellbeing is concerned.

**2.1 Old data record management**

As mentioned in the introduction, an overwhelming amount of data is being accessed and stored every day when it comes to healthcare systems. This can eventually lead to the surreal amount of data building up over time in the hospital data storage systems later on. The efficiency of the database system is also affected detrimentally due to the presence of built-up old data. This results in the inconsistency of the database systems due to the depletion of the informative value the old data had. On the other hand, setting a date and deleting data that is older than that specified date also results in systemic problems. It is important to acknowledge that there are still some essential data for the workflow that needs to be stored for future references and other medical purposes. Blind deletion of data can result in valuable information being destroyed and cause problems within the source system. Therefore, it is a necessity to have a system where there is a balance between saving and deleting old data (Wilkes, G.J, Paul, E.S, A., P, 2003). Relevant data needs to be filtered out of the irrelevant ones within the old data pool through a systematic procedure in which they are processed so that the invaluable ones that are left behind can be deleted without any hesitation. Furthermore, some of the old data stored can also affect how trends are studied in data analysis as being shown in Fig.1 below.

Chart, histogram

Description automatically generated

Fig.1: Comparison of original and normalized data

There can be conditions when compared to the recent data, some of the old data were not affected by certain variables that are brought about by immediate occurrences which results in miscalculations during the analysis process. Additionally, due to the lack of a systemic data integration system to selectively differentiate relevant data from irrelevant ones, more storage space is used resulting in higher costs. Maintenance of that vast pool of data also takes a share in high costs due to having to hire more people in order to withhold a fluid workflow.

**2.2 Less efficient management due to problems in network communication and security**

In hospital facilities, multiple computers often do share common information throughout their databases spread across as a communication network. Online database systems are necessary not only because of real-time data synchronization but also due to limited functionality available for offline systems. Remote access is also of importance, which needs to be irrespective of the network communication link, but the problem comes in when modifications are made to the root database system as it may require access to that particular data by another database within the same network. Moreover, shared database systems require frequent information updates in their local memory in order to function properly. Without good data integration, the frequent data updates and modifications require steady surveillance, intensive labor, and time (Wilkes, G.J, Paul, E.S, A., P, 2003). Undoubtedly, this also results in higher costs for the data management sector. With network communication comes along network security issues as data exchange between several databases can result in data leak or data loss. There needs to be a centralized control system that frequently communicates with others over the whole database network to carefully access which information to pass on and to which database sector. As long as there is a need to get online services, there will always be network-related problems, especially when dealing with bigger systems.

**3 Problem Solution**

**3.1 Essentiality of data integration**

As discussed previously in the introduction, the data usage of healthcare is rapidly growing at an alarming rate due to the fact that various methods of data attainment are emerging accordingly with technological advancements. It is important to get in control of those data as it can significantly improve clinical results, patient care, and also financial statements for the organizations related to healthcare. In order to effectively monitor and attend to patients with rare conditions, it is important for there to be old data records for referencing purposes so that the follow-up effects can be analyzed better. At the same time, deletion of unnecessary data in the hospital records can also result in more valuable information of new patients in critical conditions being stored. Better data management also means better optimization in patient scheduling and intake as well as verification of insurances and billing forms. Furthermore, physicians’ work-related information like privileges, organizations, faculty appointments, and locations can easily be tracked. Having better network communication within the data system also helps in improving call center productivity and quick checking applications and appointment information as real-time data is more accessible (Islam, Hasan, Wang, Germack, & Noor-E-Alam, 2018).

**3.2 Problems in implementing data integration**

There are several problems in taking the necessary steps for the proper implementation of data integration.

**3.21 Wrong data formatting**

Actions like data analysis and visualizations cannot be made to anomalous data that is in ambiguity or in an appropriate format. Manual data formatting, validation, and correction is tedious and takes up a big chunk of the IT professionals’ time (Sachin, Angappa, Milind & Jaswant, 2019).

**3.22 Data duplication**

Making sure that data duplication is not present and at the same time establishing a foundation for a well-integrated data system to ensure that data duplications occur less in the future is also one of the challenges IT professionals face. The presence of multiple data duplications can result in missed sale opportunities as a lot of time is wasted on reaching out to contacts that are no longer present in their companies (Boskova, & Stadler, 2020).

**3.23 Low-quality data**

Managing data quality is also one of the challenges as data with poor quality data can lead to lost revenue, damage in reputation, and missed insights.

Accurate business decisions can only be made through good data quality management (Sachin, Angappa, Milind & Jaswant, 2019).

**3.24 Data misplacement**

IT professionals have to manually curate data from contrasting sources and combine them which takes a lot of time. This creates inefficiency as that time should instead be well spent on data insight analysis and reassessing business practices of value (Sachin, Angappa, Milind & Jaswant, 2019).

**3.25 Lack of understanding of data**

When sharing data between technical and business teams, some of the data definitions are not thoroughly understood, causing miscommunication which can disrupt the workflow. Implementing data governance and data stewardship in the process of data integration to solve those problems is one of the big challenges that IT professionals face (Sachin, Angappa, Milind & Jaswant, 2019).

**3.3 Advantages and Disadvantages of data integration in healthcare systems**

**3.1 Advantages**

**(1) Higher efficiency**

As a lot of important information is collected from different patients and the time of doctor and the patient meeting is limited, specific data filtering services provided by data integration can allow doctors to stay focused on each individual patient (Michael, 2020).

**(2) Error reduction**

Data integration can result in computerized orders from doctors which reduces the errors caused by misunderstanding doctor’s handwriting as well as mistakes in transcriptions.

**(3) Early Detection**

A well-integrated data system can be helpful in the early detection of medical issues due to predictive assessments and algorithmic processes (Michael, 2020).

**(4) Providing lifesaving information**

Data integration can provide doctors with relevant information from past records to help them determine what is the best possible treatment for patients having similar symptoms.

**3.2 Disadvantages**

**(1) Increased workload**

Data integration brings about advancements that are beneficial but at the same time increases the workload of the medical organizations. Frequent data updates need to be managed properly to avoid orders being stalled due to lack of items in stock (Michael, 2020).

**(2) High cost**

Establishing a well-integrated data system is costly as organizations not only have to buy the necessary software but also train the staff on how to use. (Michael, 2020).

**(3) Technical problems**

As inevitable it is for data integration computer systems to have a breakdown, it is very concerning for the healthcare sector as patients’ lives are at stake and the provision of medication will get problematic.

**3.4 Proposition of Technical Architecture**

The technical architecture to be proposed is based on the Snowflake schema of the data warehouse. The business scenario to be exemplified is keeping data records for patients that are discharged from a hospital. Each row of the fact table represents a patient discharged from a certain branch on a certain date by a certain specialist in the hospital. The dimensions for this model were created by going through business questions that are shown in Fig.2 below.

Text

Description automatically generated

Fig.2: Business questions for dimension tables

The fact measures were also determined through the questions shown in Fig.3 below.

Graphical user interface, text, application

Description automatically generated

Fig.3: Fact measures for the fact table

These sequential processes led to the creation of the snowflake schema shown in Fig.4 below that is based on a typical hospital-related business scenario which is keeping records of discharged patients.

Diagram

Description automatically generated

Fig.4: Snowflake schema (zoom in)

The relationships are also displayed denoting the interconnectivity of the dimensions and the fact table in the schema. There are sub-dimensions that split off from patient and specialist dimensions which increases their data quality due to complementary background information provided and in turn making the fact table more sensible. The schema also provides a lot of options when it comes to joining tables for a filtered view to get data based on varying business requirements like an example shown below in Fig.5 where selective information about a specific patient discharged is shown by executing a complex multi-table join query.

Graphical user interface, text, application, email

Description automatically generated

Fig.5: Data of a specific discharged patient

This alone shows how connected the dimensions and the fact table are in terms of relationship and perfectly fulfill the purpose of the business scenario mentioned. Finally, the snowflake schema is incorporated into the full technical architecture shown in Fig.6.

Chart

Description automatically generated with low confidence

Fig.6: Final technical architecture after data warehouse incorporation

**4 Conclusion**

In conclusion, for the long-term, in case there will be data warehouse storage problems, they can be solved by incorporating a cloud system from the start. The cost saved by not needing to buy physical hardware can be used to buy a lot of storage space on the cloud. Furthermore, to justify the technical architecture being proposed, it perfectly tackles and solves the problems described in the problem statement section. First of all, old data record management is no longer a problem and can still be kept due to the snowflake schema taking less storage. Additionally, due to normalization, data redundancy is also greatly reduced. Network security can also be strengthened by creating policies and permissions for selected users in hospitals. The user privileges can also be customized and granted accordingly in the form of roles by the database administrator. Due to restrictions being made on data transfers, data loss or leak during the transfer process is greatly reduced. The centralization of the snowflake schema also contributes to better connectivity of the interconnected network of computers in hospitals.

***References:***

*Adibuzzaman M, DeLaurentis P, Hill J,*

*Benneyworth BD. Big data in*

*healthcare – the promises, challenges and*

*opportunities from a research*

*perspective: a case study with a model database.*

*AMIA Annu Symp Proc.*

*2018;2017:384–92*

*Big Data solutions in Healthcare: Problems and*

*perspectives*

*2015 . Prabha Susy Mathew, Anitha S. Pillai .*

*10.1109/iciiecs.2015.7193211 . 2015*

*International Conference on Innovations in*

*Information, Embedded and Communication*

*Systems (ICIIECS)*

*Boskova, V., & Stadler, T. (2020). PIQMEE:*

*Bayesian Phylodynamic Method for Analysis of*

*Large Data Sets with Duplicate Sequences.*

*Molecular Biology and Evolution, 37(10), 3061–*

*3075.* [*https://doi.org/10.1093/molbev/msaa136*](https://doi.org/10.1093/molbev/msaa136)

*Healthcare data warehouse system supporting*

*cross-border interoperability*

*2019 . Goce Gavrilov, Elena Vlahu-Gjorgievska,*

*Vladimir Trajkovik .*

*10.1177/1460458219876793*

*. Health Informatics Journal*

*Islam, M., Hasan, M., Wang, X., Germack, H., &*

*Noor-E-Alam, M. (2018). A Systematic Review*

*on Healthcare Analytics: Application and*

*Theoretical Perspective of Data Mining.*

*Healthcare, 6(2), 54.*

*https://doi.org/10.3390/healthcare6020054.*

*Michael Sony (2020) Pros and cons of implementing*

*Industry 4.0 for the organizations: a review and*

*synthesis of evidence, Production &*

*Manufacturing Research, 8:1, 244-*

*272, DOI:*[*10.1080/21693277.2020.1781705*](https://doi.org/10.1080/21693277.2020.1781705)

*Sachin S. Kamble, Angappa Gunasekaran, Milind*

*Goswami & Jaswant Manda (2019) A systematic*

*perspective on the applications of big data*

*analytics in healthcare management,*

*International Journal of Healthcare*

*Management, 12:3, 226-240,*

*DOI:10.1080/20479700.2018.1531606*

*Scalable Architecture for Personalized Healthcare*

*Service Recommendation Using Big Data Lake*

*2018. Sarathkumar Rangarajan, Huai Liu, Hua*

*Wang, Chuan-Long Wang .*

*10.1007/978-3-319-76587-7\_5 .*

*Service Research and Innovation*

*Wilkes, G. J., Paul, E. S., A., & P. (2003).*

*HEALTHCARE DATABASE MANAGEMENT*

*OFFLINE BACKUP AND*

*SYNCHRONIZATION SYSTEM AND METHOD*

*(US 2003/0204420 A1).*

*U.S. Patent and Trademark Office.*

*https://patentimages.storage.googleapis.com/1a/b0/8d/a9a29f5bd62c45/US20030204420A1.pdf*