Data Mining and Analytics Framework for Healthcare

Raghunandan Alugubelli,
University of South Florida.

Abstract

Healthcare organizations rely on data more than ever, making data collection and processing vital for any organization (Demirkan, 2013). The advancement of technology every day has led to more and more data to a level where it has become hard for an organization to manage using the current technology and approaches (AOCNP, 2015). These facts have led to big data in almost every organization that deals with data and consumers. Therefore, for any organization to meet the present and future demands of an organization, such as a healthcare organization, there is a need to develop new strategies that aim at organization and deriving meaningful information from the collected data. Especially in healthcare, it is notable that the organization produces many data rapidly, posing a challenge and advantage if the data is used (Demirkan, 2013). This research article proposes a theoretical data framework that includes the management, analysis of data in the healthcare sector.

Big data in healthcare

As suggested by the name, big data is used to represent a large amount of data that cannot be managed using traditional software or through the use of internet-based platforms (Park et al., 2014). Big data is featured in three different proportions, namely, velocity, variety, and volume. Significant data characteristics are referred to as 3Vs (Liu et al., 2015). The large volume of the data is indicative of its size, as in big data. Velocity is noted to come from the speed or rate at which the data is collected and accessible to different analyses and used in other parts. Finally, variety is used in big data regarding the different kinds of organized and unorganized data, such as data in the form of audio, transaction-level data, video, log files, and text (Salama & Shawish, 2014).

Big data has become super popular in recent years and has been an integral part of many organizations. In addition to the same, almost every sector in an organization focuses on generating and analyzing big data that it is using for various reasons (Salama & Shawish, 2014). Big data cannot be managed using traditional software. Hence, there is a need for advanced software and applications that can utilize cost-efficient and fast computational power for conducting different tasks. By implementing various artificial intelligence algorithms and the use of novel fusion, algorithms are essential. It can help extract sense from the big data in an organization (AlJarullah & El-Masri, 2013).). Big data can be pretty hazy if the organization does not have the required software and hardware to process it. However, suppose an organization can have proper tools for storage and analysis of big data. In that case, the organization can derive vital information that can be used to make the healthcare setting more aware, efficient, and interactive (Islam et al., 2015).

Methods

EHRs have been featured with many advantages in handling different information in an organization. For example, various EMR's in the market, such as Cerner, Epic, etc., can store information such as medical history, diagnoses data, labs data, surveys, questionnaires, radiology imaging, surgery data, prescriptions, demographics information (AlJarullah & El-Masri, 2013).

EHRs have helped overcome different logical errors such as illegal handwriting and mistakes in coming up with reports. The system has been featured as essential and helped enable faster recovery of information and facilitate the announcing of crucial healthcare quality depicters by organizing and improving the surveillance in public health by reporting the outbreak of any illness (Nasi, Cucciniello & Guerrazzi, 2015) immediately. The system is essential and provides relevant information regarding the care quality for employees' insurance as it helps to take charge of the costs benefits of health insurance. Lastly, EHRs provide different access to different health-related medical information to improve treatment in healthcare organizations.

A massive amount of data is generated whenever a patient visits a healthcare facility. EMR data is sitting in various data source systems, and from the systems, it can be ETL'd into the data lake.

Billing Data- We have various revenue codes that generate billing data. ICD 9 and 10 codes also do generate a lot of the data. In this data source system, all the tables should have a unique number for the patient, billing date, the procedure it was billed for, and the code of the billing data.

Registry Data- In the united states, some states require healthcare institutions to report cases from a particular disease group. So hospitals have dedicated teams working on entering these data into the database.

Labs data- General blood work and any disease-specific data

Pharmacy data- Patients get prescribed thousands of drugs. Pharmacy data systems aim to store information about drug order date, drug start date, drug end date, drug dosage information, brand name, generic name, and the status of the order.

Molecular data- We live in a world of sophisticated tests developed to detect mutations and look at aberrant expressing genes. This molecular data can be the form of the source system.

Demographics data- These are the data variables that assist in epidemiological studies. Variables such as race, ethnicity are of paramount importance. For example, demographics data play a vital role in detecting survival differences between various races for a specific disease.

The framework would require data engineers, BI engineers, and data scientists, depending on the institution's requirements. The requirements can vary widely depending on the scale of the project. The engineering team should comprise engineers who would test the quality of code in production.

Data systems:

For data systems, we have prominent vendors like Oracle and Microsoft are available in the market. Oracle SQL has one of the popular databases named My SQL. My SQL is from Oracle and even managed by it. My SQL has been in the market for so long, and there is a strong community built around it. MySQL is compatible with various languages, frameworks and has connectors with many of them. This is an open-source database. There is another database from Oracle that is available for enterprises for data storage purposes. It is costly but

can store large-scale data and has advanced programming features like pl/sql. It is compatible with many operating systems and is suitable for medium to large healthcare companies.

Microsoft has sql database, which is called sql server database. Microsoft sql database is compatible with Windows and Linux as well. Furthermore, since Microsoft manages this database, it is technically robust. Therefore, Microsoft SQL is better suited for the needs of medium to large healthcare companies.

In addition to these, the advent of no SQL databases has changed the whole dynamic. Traditional sql databases are known to be relational, but no sql databases are not relational. No SQL databases are good with unstructured data and can store data that are not only table-based. In terms of scalability, no sql databases are even more robust when compared to traditional sql servers. No sql databases can be graph or key-based, unlike sql relational databases. Some of the no SQL databases in the market are mongo DB etc.

Mongo DB has a dynamic schema, which gives flexibility whenever a database administrator changes the schema. Mongo DB is scalable, which decreases the workload. In terms of speed, no sql databases are faster when compared to relational databases. For adding new columns, Mongo DB is always flexible. Mongo DB is an excellent choice for companies with rapid growth in data and data dynamics. For healthcare tech companies that have rapid data growth, no SQL databases are a better fit.

Apache Cassandra was another system formed in the Facebook labs, but the company opened-sourced it later. Cassandra offers a sophisticated design with identical nodes, which offers superior scalability.

Apache Hbase is another no SQL database and works well with large datasets and sits on top of the HDFS. In addition, Hbase offers wonderful scalability when working with clusters.

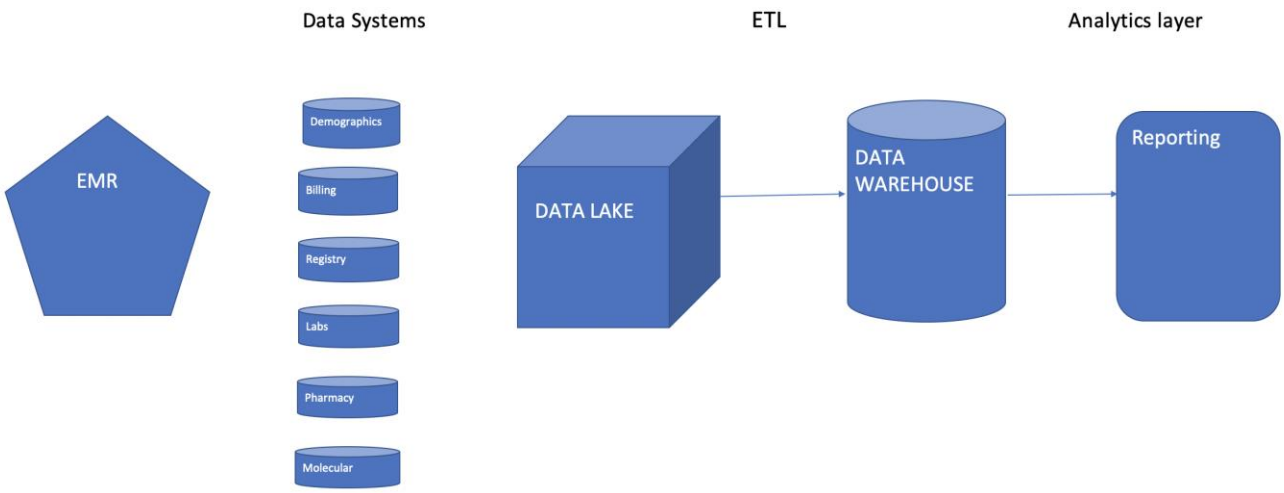
The ETL process is to extract, transform, load the data into the data warehouse. Data is usually extracted from large transactional databases. These databases function in real-time, having millions of records that store every transaction. Since these databases are not well suited for data analytics or reporting, we utilize a large ETL process to make a data warehouse and then have the reporting layers. A lot of variable transformations happen during this phase. Finally, when the data is ready, it can be loaded into the data warehouse. Much automation has occurred in this area recently. ETL has to be done to improve the data-driven decision process for a firm.

ETL usually consolidates data from various source systems and puts them in a manner that businesses can utilize. There has been tremendous pressure in the united states from regulatory agencies on healthcare institutions to comply with regulations and handle this. It would require excellent tracking of all the metrics. Monitoring of these metrics would only be possible if the organizations have enterprise-level data warehouses and dashboards.

To implement a successful ETL strategy, the healthcare organization needs to develop a road map, which outlines the strategy about infrastructure, data system, end data warehouse structure, processes, resources, and timelines. Data consolidation and integration is the critical factor in ETL. Bringing the data from multiple data source systems such as financial, medical records, pathology, etc., and mapping all the essential variables to each table is very important. Making this data integrate from various sources and creating views is a vital step in the ETL process. However, there are some challenges with this process which makes overall ETL a bit difficult. Accessing data from various systems is a challenge owning to various databases that are stored. Different emr's have different backend tied to, and if a healthcare institution has external documentation/records coming from different emr, it poses a challenge for data integration. Another challenge with the ETL process is data quality issues. Blob files, text files, and free text fields are complicated to deal with, especially in terms of querying and making a discrete variable for the data. Particularly non-discrete data is known to be not much use for reporting purposes. As the author, I feel that there is good business potential for organizations to capitalize on this free text data and converting them into discrete variables or to the tune of data that can be utilized for business reporting. We recently saw the advent of the NLP algorithm implemented in healthcare to etl these clinical notes, unstructured data into structured data. Even in these cases, the problem is that different physicians, nurse practitioners write the notes in multiple ways, and comprehending these various scenarios into the NLP algorithms is challenging. Critical components of the ETL process should allow incremental loading of the data into warehouses, supporting real-time data analytics processes. Also, the ETL teams need to ensure the logging of these pipelines to ensure hot bug fixes. Having low latency would be ideal for any ETL process to succeed.

Even after getting the data organized into databases, there might be a need for data warehouses in a healthcare firm. Datawarehouse is vital for the implementation of business intelligence programs in healthcare. Business intelligence solutions are necessary to draw insights, make data-driven decisions, and make insightful dashboards. For an organization, data warehousing costs money to employ personnel to work on the data engineering aspects, but in many cases, the benefits outweigh the costs. Data warehouse hosts critical data elements required for enterprise analytics, and often data houses will have data elements coming from a wide array of data source systems. Data warehouses usually provide access to a lot of historical data, which is helpful for data analytics. DW also may host metadata in some cases and essentially involves the extract, transform load (ETL) process. ETL needs to be done from various database systems. We have varisoftwarere's in the market to handle etl processes. Dw involves various data transformations such as normalization, reverse coding the variables, etc.

Few of the vendors in this space include Informatica, Datastage from IBM, SAS data managemenCognosnos data manager. With the latest advancement in cloud-based technology, we have cloud-based data warehouses operating in the market. Enterprises are opting for cloud solutions because of scalability, increase performance, and costs. The cloud is beneficial when the data operations require a large amount of computing power. Cloud-based data warehousing solutions also use remote resources, which leads to cutting down on expenses.



Snowflake is one of the leading providers of cloud-based data warehousing solutions. Snowflake can support semi-structured data types such as arrays, objects, and the data can be loaded without worrying about schemas.

With snowflake, a company can choose cluster size relative to the size of relevant data operations, and they can scale up or down as per requirements. Snowflake also offers more automated database maintenance.

Another leading cloud-based data warehouse provider in the market is Amazon redshift. Redshift has capabilities of machine learning etc. Redshift has a large customer base too. These two cloud data warehouses support parallel processing and have some excellent sophisticated architecture

In addition to the regular data, since we live in a data explosion generation, we have big data generated at a tremendous speed. The primary purpose of collecting these data is to optimize customer services rather than consumer consumption (De Domenico et al., 2015). Managing big data is challenging. Make it structured and valuable for business intelligence, data science, and data analytics teams in the organization. (Narrod, Zinsstag & Tiongco, 2012).

Big data management needs sophisticated high-end computing software and architecture, which needs to be governed by a scientific research computing team. The scientific research team can comprise directors, software engineers for building these big data frameworks, and administrators for maintaining these clusters. In addition, a scientific computing team would be needed to initiate, maintain the cluster environment. The big data engineering team can work hand in hand with the scientific computing team to achieve the goal. These experts' main tasks are to ensure that they integrate, annotate and present the complex data to understand more appropriately in the absence of an appropriate way of solving them.

Big data is large and has been known to be challenging to deal with. Big data is often hosted over many platforms, necessitating migration projects to have the data in one place. These migration projects are often tricky, involving a large team of engineers, architects, administrators, consume more time and money. Data mapping is one of the critical components of the migration project to see data is consistent across all the platforms. Data quality is another component that makes the management of big data cumbersome. Big data is sometimes cumbersome because of the semi-structured or raw data. Data mostly comes from various platforms, which makes the ETL process a significant step in migration. Data governance is another vital step in maintaining extensive data systems.

Different programming languages are needed to work on big data, such as python and R. In healthcare organizations, the commonly used platforms known for big data include Hadoop and Apache Spark.

Hadoop ecosystem

HDFS is one of the popular open-source frameworks that is used for ingesting big data. HDFS is a data ingestion ecosystem that comprises several other components that support the process of data ingestion. Data in hdfs resides in the form of clusters. The essential aim of the Hadoop ecosystem is to process massive amounts of complex data in a short amount of time. Data size is massive; hence, it demands various computing machines needed for distribution and finishing the processing within a limited time. For instance, when working with different nodes, an individual must focus on handling problems such as achieving parallelized computation, distributing data, and handling all the system failures (Peek, Holmes & Sun, 2014).

In HDFS data management component is made of oozie, zookeeper etc. Data processing components are hive, pig etc. Data storage can be done in hdfs or hbase. Pig is developed in yahoo labs. Pig uses a querying language similar to SQL. Pig is a platform in Hadoop for data processing and etl processes. Map-reduce uses parallel algorithms to process the data, and it is made of two essential functions. The map does the filtering and sorting of the data—map essential outputs the critical value, which is an input for reduction. Reduce does the job as the name suggests – it aggregates the data and may combine several datasets into one or vice versa. The reduced operation aims at combining all the values that are said to have shared the same access (Stephens et al., 2015). The method is efficient in providing parallelism of the computation, handling different system failures, and scheduling further communication between machines across large-scale and small-scale clusters of other machines. The hive component of Hadoop uses a query language that is similar to SQL. Hive is used for querying and creating large tables. The majority of the SQL data types are supported in the hive. Mahout is another component of Hadoop and is a machine learning component. Through Hadoop Distributed File systems (HDFS), it is noted that one can quickly provide both scalability, efficiency, and replica based on how data is stored according to the nodes of the different clusters (Peek, Holmes & Sun, 2014).

Researchers have used Hadoop to apply data sets that are said to be impossible to handle. Hadoop and other significant ecosystems might significantly help support machine learning and bioinformatics departments

of the healthcare research setting. For example, genomic and mutation data are primarily known to be massive and cumbersome. Hadoop can help in streamlining the data processing. The top vendors that provide the Hadoop ecosystem are Horton works, Cloudera.

Apache Spark

Another popular framework in addition to Hadoop is Spark. Apache spark was developed at the University of California, Berkely. Apache spark is considerably faster than Hadoop owing to its processing speed of big data. Spark engine is unified and used for distributed data processing known for the inclusion of higher-level libraries, essential for supporting the different SQL queries, machine learning, streaming data, and graph processing (Song & Ryu, 2015). These libraries are necessary and are known for increasing the developers' productivity as the programming interface is featured using efforts to code. That is said to be combined to create more complex computations (Forkan et al., 2015). Spark core has in-memory computing. Spark uses structured query language, which in turn helps for working with tabular data. Spark has a streaming component to support the needs of companies that have real-time streaming data, such as Netflix or any other entertainment company. In streaming, data can be ingested from many sources like Kafka, Flume, etc. Spark is compatible with multiple languages such as scala, python. Spark supports real-time streaming applications, and this is a point that makes it superior to the Hadoop ecosystem.

Resilient Distributed Datasets (RDDs) can be implemented in the memory processing supported, making Spark to be about 100x faster than Hadoop, which is implemented in various pas analytics, known for using small datasets (Song & Ryu, 2015). This is noted to be necessary if the data involved is small compared to the available memory. This is indicated to show that big data processing uses Apache Spark and is known for using Apache Spark that needs a large memory. It is a fact that the memory cost tends to be high as compared to the hard drive, MapReduce is essential, and it is supposed to provide a cost-effective approach when used for big datasets than Apache Spark (Khazaei et al., 2015). Spark also has machine learning libraries embedded in it, making data science analytics feasible with the spark ecosystem.

Conclusion

Different healthcare and various biomedical tools such as sensors used for mobile biometric, genomics, and smartphone applications have been essential in generating enormous data. Hence, it is necessary to develop different methods that can access and use this data. For instance, healthcare can get a clear insight into different methods, skills, medical, and other needed improvements in the healthcare setting through additional data analysis. Therefore, big data analytics of EHRs, EMRs, and other medical data is a continuous process for building a better prognostic framework. The main goal of coming up with the framework is to reduce the cost used for analysis, develop an adequate Clinical Decision Support (CDS) system, and better platforms used to provide treatment.

References

AlJarullah, A., & El-Masri, S. (2013). A novel system architecture for the national integration of electronic health records: a semi-centralized approach. *Journal of medical systems*, *37*(4), 1-20.

Alugubelli, R. (2016). Exploratory Study of Artificial Intelligence in Healthcare. International Journal of Innovations in Engineering Research and Technology, 3(1), 1–10.

Banos, O., Villalonga, C., Garcia, R., Saez, A., Damas, M., Holgado-Terriza, J. A., ... & Rojas, I. (2015). Design, implementation, and validation of a novel open framework for agile development of mobile health applications. *Biomedical engineering online*, *14*(2), 1-20.AOCNP, D. (2015).

The evolution of the electronic health record. *Clinical journal of oncology nursing*, *19*(2), 153.

De Domenico, M., Nicosia, V., Arenas, A., & Latora, V. (2015). Structural reducibility of multilayer networks. *Nature communications*, *6*(1), 1-9.

Demirkan, H. (2013). A smart healthcare systems framework. *It Professional*, *15*(5), 38-45.

Forkan, A. R. M., Khalil, I., Ibaida, A., & Tari, Z. (2015). BDCaM: Big data for context-aware monitoring—A personalized knowledge discovery framework for assisted healthcare. *IEEE transactions on cloud computing*, *5*(4), 628-641.

He, W., Wu, H., Yan, G., Akula, V., & Shen, J. (2015). A novel social media competitive analytics framework with sentiment benchmarks. *Information & Management*, *52*(7), 801-812.

Herland, M., Khoshgoftaar, T. M., & Wald, R. (2014). A review of data mining using big data in health informatics. *Journal of Big data*, *1*(1), 1-35.

Islam, S. R., Kwak, D., Kabir, M. H., Hossain, M., & Kwak, K. S. (2015). The internet of things for health care: a comprehensive survey. *IEEE access*, *3*, 678-708.

Khazaei, H., McGregor, C., Eklund, J. M., & El-Khatib, K. (2015). Real-time and retrospective health-analytics-as-a-service: a novel framework. *JMIR medical informatics*, *3*(4), e36.

Kulev, I., Vlahu-Gjorgievska, E., Trajkovik, V., & Koceski, S. (2013). Development of a novel recommendation algorithm for collaborative health: Care system model. *Computer Science and Information Systems*, *10*(3), 1455-1471.

Kuziemsky, C. E., Monkman, H., Petersen, C., Weber, J., Borycki, E. M., Adams, S., & Collins, S. (2014). Big Data in Healthcare–Defining the Digital Persona through User Contexts from the Micro to the Macro: Contribution of the IMIA Organizational and Social Issues WG. *Yearbook of medical informatics*, *9*(1), 82.

Li, S., Zhang, T., Gao, J., & Park, Y. (2015, March). A sticky policy framework for big data security. In *2015 IEEE First International Conference on Big Data Computing Service and Applications* (pp. 130-137). IEEE.

Liu, C., Wang, F., Hu, J., & Xiong, H. (2015, August). Temporal phenotyping from longitudinal electronic health records: A graph based framework. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 705-714).

Narrod, C., Zinsstag, J., & Tiongco, M. (2012). A one health framework for estimating the economic costs of zoonotic diseases on society. *EcoHealth*, *9*(2), 150-162.

Nasi, G., Cucciniello, M., & Guerrazzi, C. (2015). The role of mobile technologies in health care processes: the case of supportive cancer care. *Journal of medical Internet research*, *17*(2), e26.

Park, Y., Shankar, M., Park, B. H., & Ghosh, J. (2014, March). Graph databases for large-scale healthcare systems: A framework for efficient data management and data services. In *2014 IEEE 30th International Conference on Data Engineering Workshops* (pp. 12-19). IEEE.

Peek, N., Holmes, J. H., & Sun, J. (2014). Technical challenges for big data in biomedicine and health: data sources, infrastructure, and analytics. *Yearbook of medical informatics*, *9*(1), 42.

Salama, M., & Shawish, A. (2014, March). A Novel Mobile-Cloud based Healthcare Framework for Diabetes. In *HEALTHINF* (pp. 262-269).

Salas-Vega, S., Haimann, A., & Mossialos, E. (2015). Big data and health care: challenges and opportunities for coordinated policy development in the EU. *Health Systems & Reform*, *1*(4), 285-300.

Song, T. M., & Ryu, S. (2015). Big data analysis framework for healthcare and social sectors in Korea. *Healthcare informatics research*, *21*(1), 3.Buchanan, W., & Woodward, A. (2017). Will quantum computers be the end of public-key encryption?. *Journal of Cyber Security Technology*, *1*(1), 1-22.

Soudris, D., Xydis, S., Baloukas, C., Hadzidimitriou, A., Chouvarda, I., Stamatopoulos, K., ... & Ghia, P. (2015, July). AEGLE: A big bio-data analytics framework for integrated healthcare services. In *2015 International Conference on Embedded Computer Systems: Architectures, Modeling, and Simulation (SAMOS)* (pp. 246-253). IEEE.

Steinberg, G. B., Church, B. W., McCall, C. J., Scott, A. B., & Kalis, B. P. (2014). Novel predictive models for metabolic syndrome risk: a "big data" analytic approach. *Am J Manag Care*, *20*(6), e221-e228.

Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., ... & Robinson, G. E. (2015). Big data: astronomical or genomic?. *PLoS biology*, *13*(7), e1002195.

Wallace, P. J., Shah, N. D., Dennen, T., Bleicher, P. A., & Crown, W. H. (2014). Optum Labs: building a novel node in the learning health care system. *Health Affairs*, *33*(7), 1187-1194.Adler-milstein, J., & Pfeifer, E. (2017). Information blocking: is it occurring, and what policy strategies can address it?. *The Milbank Quarterly*, *95*(1), 117-135.

Yuan, B., & Herbert, J. (2014). Context-aware hybrid reasoning framework for pervasive healthcare. *Personal and ubiquitous computing*, *18*(4), 865-881.

Zhang, F., Cao, J., Khan, S. U., Li, K., & Hwang, K. (2015). A task-level adaptive MapReduce framework for real-time streaming data in healthcare applications. *Future generation computer systems*, *43*, 149-160.

Zhang, Y., Sun, L., Song, H., & Cao, X. (2014). Ubiquitous WSN for healthcare: Recent advances and future prospects. *IEEE Internet of Things Journal*, *1*(4), 311-318.