|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ITU Logo | INTERNATIONAL TELECOMMUNICATION UNION  **TELECOMMUNICATION STANDARDIZATION SECTOR**  STUDY PERIOD 2017-2020 | | FG-AI4H-G-205-A02 | |
| **ITU-T Focus Group on AI for Health** | |
| **Original: English** | |
| **WG(s):** | | Plenary | New Delhi, 13-15 November 2019 | |
| **DOCUMENT** | | | | |
| **Source:** | | Editor | | |
| **Title:** | | DEL05B: Data acquisition | | |
| **Purpose:** | | Discussion | | |
| **Contact:** | | Rajaraman Subramanian Calligo Technologies  India | | Tel: +91-9845239446 Email: [giri@calligotech.com](mailto:giri@calligotech.com) |
| **Contact:** | | Vishnu Ram India | | Email: [vishnu.n@ieee.org](mailto:vishnu.n@ieee.org) |
| **Contact:** | | Pranav Nair Calligo Technologies India | | Tel: +91-9108283995 Email: pranav.n@calligotech.com |

|  |  |
| --- | --- |
| **Abstract:** | This document contains the proposed initial structure for the FG-AI4H Deliverable 5B, “Data Acquisition”. It presents a framework for public healthcare data acquisition and management model based on standard protocol for its easy adoption by any country or international health organizations. This paper assumes basic digitization of electronic health record (EHR) at basic health facilities. There is a gap in developing an integrated and comprehensive framework that addresses the use of EHR in a standardized way for public health, privacy issue by anonymizing patient specific information, fusing multiple records with slight changes in the same information, augmenting a broad spectrum of contextual data, and so on.  This document is a draft and a work-in-progress until it is finally approved by the Focus Group. |

# Introduction

The healthcare industry historically has generated large amounts of data, driven by record keeping, compliance & regulatory requirements, and patient care *[1]*. While most data is stored in hard copy form, the current trend is towards rapid digitization of these large amounts of data. Driven by mandatory requirements and then potential to improve the quality of healthcare delivery meanwhile reducing the costs, these massive quantities of data hold the promise of supporting a wide range of medical and healthcare functions, including among others clinical decision support, disease surveillance, and population health management.

Healthcare data has seen massive growth over the last several years, with some reports estimating that healthcare data generation increases by 48% annually *[2]*. In addition, it has been estimated that the intelligent use of big data within the healthcare sector could save over $300 billion *[3]*.

While big data on Public Health are growing with the diffusion of telemedicine and e-health, and more generally with that of Internet of Things (IoT) *[4,5]* sensors and networking digital platforms, their relationship with healthcare infrastructural elements is still pioneering.

Digitized data already provide many benefits to healthcare organizations through disease prediction and surveillance, population health management and patient care improvement.

Our goal is to describe how an integrated data lake and analytics platform can be used to provide near real-time access to healthcare and biomedical research data with the ability to conduct computational healthcare research.

The rest of this document tracks the requirements that any system should fulfil and then talks about the initial design that can be implemented, that will adhere to the requirements stated below.

# Requirements

## Non-functional requirements

### Data interoperability

Interoperability is an important consideration while handling health data, especially for analytics and AI/ML use cases. As described in *[9]* different types of health information may be used as input to AI/ML mechanisms. While some of these may have standard formats *[10]*, others may require more work in standardising data formats. Openness of data formats is another important factor to achieve interoperability.

### Data availability and access

Making sure quality data is made accessible to researchers spread across the world, is a matter of paramount interest. One way to ensure that is implementing broad, in-depth and robust APIs based on open standards.

### Data security and privacy

Security and privacy are important considerations while handling health data for AI/ML use cases. The requirements for security and privacy for health data are outlined in *[11]*. This document focuses on the corresponding techniques to achieve the security and privacy goals during the process of data acquisition, processing and application. Ethical considerations and adherence to regulations on data handling is an important non-functional requirement.

Other non-functional requirements like volume, velocity, quality, performance, availability, verification considerations may also apply to specific use cases.

## Functional requirements

### Support for multiple data formats

The system should be capable of ingestion data in the form of text (prescriptions, diagnostic reports), images (DICOM-supported, camera) ,videos and even audio.

### Support for multiple types of data

When monitoring continuously a patient health condition, several types of data are generated. Medical data may include structured data like traditional Electronic Healthcare Records (EHRs), semi structured data such as logs produced by some medical devices, and unstructured data generated, for example, by biomedical imagery.

#### Electronic healthcare records (EHR)

It contains a complete patient medical history stored in a digital format; it is formed by a multitude of medical data describing the patient’s health status like demographics, medications, diagnoses, laboratory tests, doctor’s note, radiology documents, clinical information, and payment notes. Thus, EHR represents a valuable source of information for the purpose of healthcare analytics. Furthermore, EHR allows exchanging data between healthcare professionals community.

#### Biomedical images

Biomedical imaging is considered as a powerful tool regarding disease detection and care delivery. Nevertheless, processing this kind of images is challenging as they include noisy data that needs to be discarded in order to help physicians make accurate decisions.

#### Social network analysis

Performing social network analysis requires gathering data from social media like social networking sites. The next step consists of extracting knowledge that could affect healthcare predictive analysis such as discovering infectious illnesses. In general, social networks data is marked by uncertainty that makes their use in designing predictive models risky.

#### Sensing data

Sensors of different types are employed in healthcare monitoring solutions. Those devices are essential in monitoring a patient health as they measure a wide range of medical indicators such as body temperature, blood pressure, respiratory rate, heart rate, and cardiovascular status. In order to ensure an efficient health monitoring, patients living area may be full of devices like surveillance cameras, microphones, and pressure sensors. Consequently, data volume generated by health monitoring systems tends to increase tremendously which requires adopting sophisticated methods during the processing phase.

### Support for multiple data sources

#### Mobile Phone

Nowadays, mobile phone represents one of the most popular technological devices in the world. Compared to their early beginnings, mobile phones transformed from a basic communication tool to a complex device offering many features and services. They are currently equipped with several sensors like satellite positioning services, accelerometers, and cameras. Due to their multiple capabilities and wide use, mobile phones are ideal candidates regarding health data collection allowing the design of many successful healthcare applications like pregnancy monitoring *[6]*, child nutrition *[7]*, and heart frequency monitoring *[8]*.

#### Healthcare sensors

State-of -art, purpose-built sensors can be found aplenty in the healthcare market. From wrist-based devices that monitor heart rate to blood sugar sensors, these devices generate a lot of data which not only help individual diagnosis but can also be used for research. These data points are only going to explode in volume and that, in turn is going to make the challenge of efficient data collection even more critical and complex.

#### Diagnostic machines

The traditional machines and tools used in diagnostics have been the primary source of health care data for long. Their position in the realm of data collection is paramount because of the veracity and quality of data it provides.

#### Traditional Data stores (databases)

The data that is present in unconnected, private data repositories form a vast majority of the health care data that currently exists. The proposed system should be able to ingest data from such sources, so that its accessible in a uniform way and its value soars.

### Support for various modes of data collection

To support a large number of diagnostic devices that operate under various environmental conditions, it might not be possible for the devices to send data, as and when they are generated. Hence, the data acquisition module should be able to operate in stream or batch mode. Furthermore, offline operations might also be supported.

### Support for various data operations

Based on the use case under consideration, this may include identification of the source of data, the type of pre-processing, aggregation, AI/ML techniques for processing, target application points for the output from AI/ML models.

One feature essential as part of operations is to read the data gathered from healthcare sensors in several formats and then data flows through semantic module before being normalized.

#### Semantic module

It is based on ontologies, which constitute efficient tools when it comes to representing actionable knowledge in the field of biomedicine. In fact, ontologies have the ability to extract biomedical knowledge in a formal, powerful, and incremental way. They also allow automation and interoperability between different clinical information systems. Automation has a major benefit; it helps medical personnel in processing large amounts of patients’ data, especially when taking into consideration that these personnel is often overwhelmed by a series of healthcare tasks. Introducing automation in healthcare application contributes to providing assistance to human medical staff, which enhances its overall performance. It should be highlighted that automation will help humans in performing their duties rather than replacing them. Interoperability is an important issue when dealing with medical data. In fact, healthcare databases lack homogeneity as they adopt different structures and terminologies. Therefore, it is difficult to share information and integrate healthcare data. In this context, ontologies may play a determinant role by establishing a common structure and semantics, which allows sharing and reuse of data across different systems. In other words, by defining a standard ontology format, it becomes possible to map heterogeneous databases into a common structure and terminology. For instance, the Web Ontology Language (OWL) represents the standard interchange format regarding ontology data that employs XML syntax.

## Related work

Some general guidelines for handling data may be derived from [ITU-T Y.3600], [ITU-T Y.3172] and [ITU-T Y.3174].

NOTE – While these ITU Recommendations may provide general guidelines especially in communication networks, it is needed to customize these for specific use cases in the area of health.

# Solution framework

## Data lake

Data lake is a repository that holds a vast amount of raw data in its native (structured or unstructured) format until the data is needed. Storing data in its native format enables you to accommodate any future schema requirements or design changes.

A data lake is ideal when valuable data sources are dispersed among on-premises data centers, software providers, partners, third-party data providers, or public datasets. A data lake offers a foundation for storing on-premises, third-party, and public datasets at low prices and high performance.

Additionally, an array of descriptive, predictive, and real-time agile analytics built on this foundation can help meet a company’s most important business needs, such as forecasting service delivery and service utilization patterns, evaluating effectiveness and cost of service delivery, and analysing financial performance comparisons against estimated expenses to support federal and/or state reporting.

## Understanding the functions of data lake

1. **Data Submission**  
   The first function of a data lake is to receive large amounts of data. This data can be submitted to the data lake as either batch uploads or streaming data.
2. **Data Processing**  
   The data lake then validates the data, adds any necessary metadata, and indexes the data accordingly.
3. **Data Management**  
   Next, the data is transformed and aggregated into a format that is ready to be stored long term, and made available to the analytics applications.
4. **Searching**  
   The data lake catalogues the indexed data and provides a query tool for interacting with the data in a read-only manner.
5. **Publishing**  
   The final function of a data lake is to publish the analysed data in meaningful charts and graphs via a BI tool or data visualizer.

## How is Data Lake different from a data warehouse?

It’s been said that a data warehouse is like “a store of bottled water – cleansed and packaged and structured for easy consumption – the data lake is a large body of water in a more natural state. The contents of the data lake stream in from a source to fill the lake, and various users of the lake can come to examine, dive in, or take samples.”

That’s why many healthcare organizations are shifting to a data lake architecture. A data lake is an architectural approach that allows you to store massive amounts of data into a central location, so that it’s readily available to be categorized, processed, analysed and consumed by diverse groups within an organization. Since data can be stored as-is, there is no need to convert it to a predefined schema and you no longer need to know what questions you want to ask of your data beforehand.

A data lake can have all kinds of data, structured or unstructured, and offers the agility to reconfigure the underlying schema on the fly. The same cannot be said for data warehouses. Additionally, the raw data stored in data lakes is never lost – it is stored in its original format for further analytics and processing.

A data lake can actually complement and extend your existing data warehouse. If you’re already using a data warehouse, or are looking to implement one, a data lake can be used as a source for both structured and unstructured data.

## Why a data lake is necessary?

Healthcare providers and health plans are asking the question, “Do we need an enterprise data warehouse, a data lake, or both as part of our overall data architecture?” This question is top of mind with healthcare executives who are challenged to improve medical outcomes for patients, drive patient and member engagement, and bend the cost curve. In order to do so, these organizations are dependent upon the ability to rapidly ingest and analyse large volumes of data in batch or real-time from an extensive range of sources in a variety of formats. This guide takes a deeper dive into the role data lakes play in healthcare and will:

1. **Explore** the data lake concept and how it differs from a more conventional enterprise data warehouse approach
2. **Discuss** how a data lake typically co-exists with an enterprise data warehouse to enable advanced analytics that are difficult to attain with a traditional enterprise data warehouse architecture
3. **Present** use cases that are best served by a data lake environment, and the typical starting points for a data lake effort and associated architecture

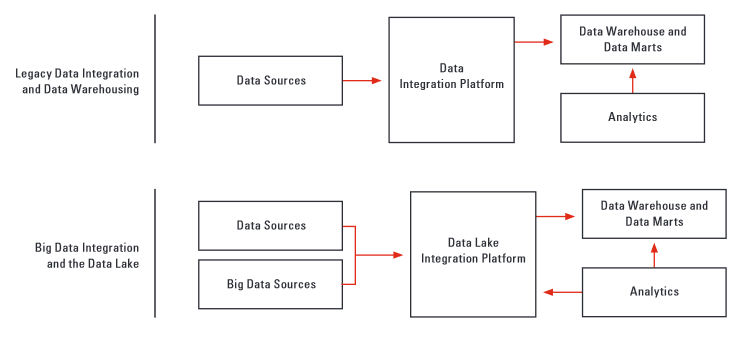


Figure 1 – The data lake platform of healthcare information systems

In summary, a data lake has the following characteristics:

1. Centralized data integration, with a schema-at-read, indexed-for-search optimized store that is based on usage, as well as historical archive and key value data retrieval
2. Optimized storage with relevant latency, redundancy, isolation, and durability (memory, file system, NoSQL, or database)
3. Mechanism for rapid ingestion of data, distributed compute, and workflow control
4. Ability to integrate and map data across multiple data types and sources
5. Provides access to users, Enterprise data warehouse (EDW), data marts, and analytical applications
6. Incorporates metadata exchange and governance
7. Catalogued and indexed for rapid search and data retrieval
8. Ability to manage security, permissions, and provide data masking on sensitive patient information
9. Operationally managed and centrally controlled
10. Supports self-provisioning of compute nodes, data, and analytic tools without IT intervention

## Storage Requirements

Healthcare professionals need fast and secure access to sensitive files, including protected health information (PHI). Healthcare IT needs to store, manage, and protect a never-ending stream of data, especially unstructured data. These challenges needs to be taken care of, while maintaining the highest level of data privacy and compliance. This need to be addressed separately as a sub-topic within this module.

## Legacy System Support

As touched upon earlier, the ability to successfully interact with popular legacy systems, is really crucial for world-wide acceptance of the proposed system. This in turn, will avoid data silos and effectively ensure all the health care data can be stored and used under the aegis of the proposed system.

## Privacy preservation

Needless to say, data privacy is one of the foundational pillars upon which any system should be designed. Keeping in mind the sensitivity of health data, multiple contingencies should be kept in place so that no data which can be used to identify individuals, is ever leaked into the public domain.

## Scalability

Availability of the proposed system is crucial if we are to see a widespread acceptance of the services provided. To make sure the proposed system is online and always available, we need to track usage metrics which are specifically designed to track system usage. Based on further analysis, we should be able to scale up the storage and supported components to accommodate the exploding volume of data that gets generated and uploaded.

# References

1. Raghupathi W; Data Mining in Health Care. In Healthcare informatics: Improving Efficiency and Productivity. Edited by Kudyba S, Taylor & Francis; 2010211-233
2. EMC. The Digital Universe Driving Data Growth in Healthcare. Secondary The Digital Universe Driving Data Growth in Healthcare. 2014. <https://www.emc.com/analyst-report/digital-universe-healthcare-vertical-report-ar.pdf>.
3. Manyika J, Chui M, Brown B, et al. Big data: The next frontier for innovation, competition, and productivity. 2011.
4. <http://literature.cdn.keysight.com/litweb/pdf/5992-3024EN.pdf>
5. Gunasekaran Manogaran, R.Varatharajan, Daphne Lopez et al A new architecture of Internet of Things and big data ecosystem for secured smart healthcare monitoring and alerting system. <https://doi.org/10.1016/j.future.2017.10.045>
6. IBM: Harvard Medical School: 2011. http://publicdhe.ibm.com/common/ssi/ecm/en/imc14685usen/IMC14685USEN.PDF
7. Raghupathi W, Kesh S: Interoperable electronic health records design: towards a service-oriented architecture. e-Service Journal 2007,5:39-57
8. Borkar VR, Carey MJ, Chen L: Big data platforms what’s next? ACM Crossroads 2012. 19(1)44-49.
9. <http://med.stanford.edu/content/dam/sm/ehr/documents/SM-EHR-White-Papers_v12.pdf>
10. <https://insights.datamark.net/unstructured-data-in-electronic-health-record-systems-challenges-and-solutions/>
11. <https://link.springer.com/article/10.1186/s40537-017-0110-7>
12. Regenstein, M., and D. Sickler. 2006. *Race, ethnicity, and language of patients: Hospital practices regarding collection of information to address disparities in health care*. Princeton, NJ: Robert Wood Johnson Foundation.
13. Romano, P. S., J. J. Geppert, S. Davies, M. R. Miller, A. Elixhauser, and K. M. McDonald. 2003. A national profile of patient safety in U.S. Hospitals. *Health Affairs* 22(2):154-166.
14. Siegel, B., M. Regenstein, and K. Jones. 2007. *Enhancing public hospitals' reporting of data on racial and ethnic disparities in care*. New York: The Commonwealth Fund.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_