

Transforming Healthcare Leveraging Data Lake

Aditya Satyadev (IIM C)

Chief Executive Officer BizAcuity

©2019 BizAcuity

Table of Contents

1. What is Data Lake?	2
2. Data Lake on Cloud.....	2
3. Architecting Data Lake for Cloud	3
4. Study of Data Lake Options Available	4
5. Data Lake for Healthcare Industry	5
6. Designing the Healthcare Data Lake on cloud	6
7. Case in Point - Healthcare Data Lake Designed for a Healthcare Client.....	8

Data Lake on Cloud – For Healthcare Industry

1. What is a Data Lake?

In today's business world, data is king. Most businesses today run on data and effective use and analysis of enterprise data. A data lake is a storage repository that holds a vast amount of raw data in its native format until it is needed. When a business question arises, the data lake can be queried for relevant data, and that smaller set of data can then be analysed to help answer the question.

Data Lake versus Data Warehouse

Typically, data was stored in data warehouses. A data warehouse is a database that is configured to analyse relational data coming from transactional systems and line of business applications. In a typical database, the data structure and schema are defined in advance, and the data can be queried using a number of SQL queries. The results of these queries are usually used for analysis and inference. The architecture of a data warehouse is typically hierarchical where data is stored data in files or folders.

A data lake is different, because it stores relational data from line of business applications, and non-relational data from mobile apps, IoT devices, and social media. The data lake uses a flat architecture to store data, which means the structure of the data or schema is not rigidly defined when data is captured. This means that the data can be stored without any biases regarding the kind of queries that could be conducted on the data. The SQL queries are not pre-defined, and hence different types of analytics such as SQL queries, big data analytics, full text search, real-time analytics, and

machine learning can be used to uncover insights.

Why do we need a Data Lake?

It has been seen that enterprises that base decisions and business value on data typically outperform those that do not. An Aberdeen survey saw organizations who implemented a Data Lake outperforming similar companies by 9% in organic revenue growth. These organizations that implemented a data lake were able to do new types of analytics like machine learning over new sources like log files, data from click-streams, social media, and internet connected devices stored in the data lake. This helped them to identify, and act upon opportunities for business growth faster by attracting and retaining customers, boosting productivity, proactively maintaining devices, and making informed decisions.

Most organization today see the need for a data lake and are hence evolving their data strategy to include data lakes, and enable diverse query capabilities, data science use-cases, and advanced capabilities for discovering new information models. Gartner names this evolution the "Data Management Solution for Analytics" or "DMSA."

2. Data Lake on Cloud

Most businesses today run on cloud. Deployment on the cloud provides performance, scalability, reliability, availability, a diverse set of analytic engines, and massive economies of scale. When it comes to Data Lakes, the advantages of being on the cloud are better security, faster time to deployment, better availability, more frequent feature/functionality updates, more elasticity, more geographic coverage, and costs linked to actual utilization.

Here are some of the benefits of having Data Lake on cloud.

Agility and dynamic processing

The biggest advantage of the cloud is its agility and flexibility. The cloud makes it possible to pay for just the configuration you use. For example, you can start with a 20-node cluster and then easily increase to 100 nodes as your requirements change. You can scale up and scale down as per need, and you pay for only what you use. Also, cloud storage provides native integration with a number of powerful services, giving the flexibility to choose the right tool to analyse data.

Cost-effective data storage and compute

In a cloud deployment a Data Lake has separate storage and compute services. This is because in the cloud, storage is cost effective and compute is expensive. This provides cost effectiveness, but also calls for careful planning of the architecture. Cloud storage provides a number of storage classes at multiple prices to suit different access patterns and availability needs, and to offer the flexibility to balance cost and frequency of data access.

Up-to-date technologies

Cloud providers add services and support that make it easier to upgrade without impacting the overall solution. For example, we have clients with cloud-based data lake architectures that were able to upgrade to a new version of Hadoop in a matter of days.

Security

Since data lakes are designed to store all types of data, enterprises expect strong access control capabilities to help ensure that their data doesn't fall into the wrong hands. Cloud Storage offers a number of mechanisms to implement secure access control over data assets.

Geographic replication for high resiliency

A big advantage to moving to the cloud is that cloud vendors have regional, cross-regional or cross-country data recovery strategies and applications in place. This means that a separate data centre is not required to ensure resiliency in case of disaster.

3. Architecting Data Lake for Cloud

While building a Data Lake on Cloud, here are some of the aspects that need to be considered:

- Data ingestion
- Data mining and exploration
- Processing and analytics
- Design and deploy workflows

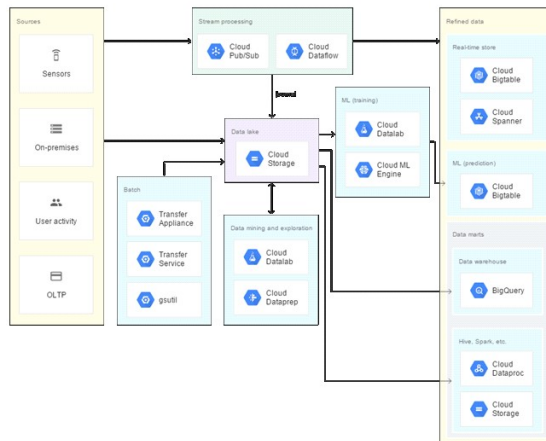
Data ingestion

A data lake architecture must be able to ingest varying volumes of data from different sources such as Internet of Things (IoT) sensors, clickstream activity on websites, online transaction processing (OLTP) data, and on-premises data, etc. Some of the ways enterprises can set up data ingestion as part of the Data Lake architecture include:

- Cloud Pub/Sub and Cloud Dataflow: Real-time data can be ingested and stored directly into Cloud Storage, scaling both in and out in response to data volume.
- Storage Transfer Service: This provides a robust framework to move large amounts of data, schedule periodic data transfers, synchronize files between source and sink, or move files selectively based on filters.
- Open source tools: For one-time or manually initiated transfers, open source tools such as gsutil can be used. It

supports multi-threaded transfers, processed transfers, parallel composite uploads, retries, and resumability.

- **Transfer Appliance:** Depending on the network bandwidth, to migrate large volumes of data to the cloud for analysis, the migration can be done offline using the Transfer Appliance.



Data mining and exploration

Data Lakes have the data stored in raw formats. Therefore, a large part of this data is not ready for immediate consumption and must be mined for use and analysis.

There are various tools available for data mining and exploration and depending on the business use of data and cost-benefit analysis, the appropriate tool can be selected.

Processing and analytics

Once data has been ingested and stored, it needs to be made available for analysis. Like in the case of a data warehouse, here too the Data Lake can have a well-understood schema that lends itself querying the database. This can be done using SQL queries.

However, the primary feature of a Data Lake vis-à-vis a data warehouse is that it may not always have a well-structured schema. The data in Data Lake is typically stored in raw formats as is, and hence SQL queries may not always work. In such cases, for each kind of analysis required, the workflow can range from simple to complex. The following diagram provides a high-level overview.

Design and deploy workflows

As a final step in getting the Data Lake ready for use, workflows need to be designed that would essentially define the way the raw data is mined, extracted, analysed and transformed into a format that downstream processes and users can consume. These workflows vary based on the nature of the data and the types of analytics used.

4. Study of Data Lake Options Available

Today data lake options for cloud are available from three leading cloud computing vendors - Amazon Web Services, Microsoft Azure and Google Cloud.

Amazon Web Services

Amazon Web Services (AWS) is the leader in this market. It is the first choice for cloud IaaS for the past several years. One of the reasons for its popularity is the vast tool set it provides and the massive scope of operations that it offers. AWS has a huge and growing array of available services, as well as the most comprehensive network of worldwide data centres.

The challenge with AWS continues to be its cost structure that can be confusing. Enterprises find it challenging to plan for AWS implementation while running a high volume of workloads on the service.

However, the strengths that AWS brings far outweighs this challenge, and organizations of all sizes continue to use AWS for a wide variety of workloads.

Microsoft Azure

With the vast enterprise background and Windows support that Microsoft has, Azure is a very close competitor to AWS with an exceptionally capable cloud infrastructure.

The biggest advantage of Azure is the integration with Windows and other Microsoft software. Because Azure is tightly integrated with these other applications, enterprises that use a lot of Microsoft software often find that it also makes sense for them to use Azure.

The challenge with Azure has been on the technical support and training, and ISV partner ecosystem. Organizations find it difficult to get the necessary support that they expect from an enterprise platform.

Google Cloud Platform

Google has deep technical expertise, and industry-leading tools in deep learning and artificial intelligence, machine learning and data analytics. And these lend themselves to the Google Cloud Platform (GCP) and make it a strong competitor.

GCP specializes in high compute offerings like Big Data, analytics and machine learning. It also offers considerable scale and load balancing since Google knows data centres and fast response time.

The challenge with GCP is that it does not offer as many different services and features as AWS and Azure. It also does not have as many global data centres as AWS or Azure, although it is quickly expanding.

5. Data Lake for Healthcare Industry

The healthcare industry uses a huge amount of data. In fact, the wheels of the industry run primarily on data and analysis of data. The data could be pertaining to clinical data or claims data. Clinical data usually pertains to important and critical information about patient diagnoses, claims and medical history. Claims data contains data about the patient care, reimbursements, claims, etc. Typically, the two types of data come in from different sources and are also meant for different purposes, and therefore a Data Lake is a good mechanism to store the data. Data Lakes are well suited for healthcare because it stores all the data in a central repository and only maps it as the need arises.

Advantage of Data Lake for Healthcare industry

For healthcare industry, the robust, customizable data lake solutions provide the following advantages:

- Allow users to analyze and visualize data from various sources through a central dashboard that mines the data from its sources
- Allow users to search from entire existing healthcare data set to meet requirements
- Ability to use varied data, in any given format, so that researchers can focus on healthcare innovations and cures
- Provide data analysis that allows enterprises to build plans and business cases for research and funding

Healthcare Data Lake on Cloud

Data Lakes for healthcare are ideal to be deployed in the cloud, because the cloud provides performance, scalability, reliability, availability, a diverse set of analytic engines, and massive economies of scale.

Considering the vast amount of unstructured data coming in from various sources, storage and computing is definitely to be considered while building a healthcare data lake. And for this reason, the data lake on cloud makes more sense for healthcare since it allows for easier availability of data, more geographic coverage, better security, faster deployment time, and frequent feature and functionality updates.

6. Designing the Healthcare Data Lake on Cloud

For any data lake to be more effective, it is better to start with a business problem in mind, stay focused, and solve it and deliver results that can meet the demands of the business.

Planning for structured and unstructured data

Healthcare data is a mixture of structured and unstructured data. Patient demographic information, diagnosis and procedure codes, medication codes, and certain other data from the EHR are typically generated in a standardized, structured way. Whereas data collected by clinicians, patients, during research, during diagnosis and innovative care could be largely unstructured.

Structured data is data stored within fixed confines, such as a file. Structured data is easier to analyze and store because it has straightforward boundaries and is created and stored in a standardized format. But unstructured data has no prescribed form and hence more challenging to access and use.

Storing data in the cloud gives organizations a level of flexibility that they often can't achieve with on-premises deployments. Cloud data storage also saves organizations money by allowing them to purchase more storage space as needed, rather than investing in additional on-premise servers. Moving data to the cloud not

only gives organizations an easier way to expand, but it cuts back on the cost of hardware for on-premises servers and additional IT staff needed to manage and maintain on-premises servers.

Deciding on cloud architecture

Once the data planning is done, organizations need to decide if they wish to deploy their tools in the public cloud, private cloud, or a combination of both.

- Public cloud is the most scalable data storage solution. Storage space can be added or dropped as the size of an organization changes. This makes public cloud popular for temporary projects as well as data migration.
- Private cloud gives organizations more control over where their data resides and its accessibility to users. The private cloud gives health IT staff direct control over the contents stored in the cloud. Healthcare organizations may benefit from private cloud because they can keep a close eye on PHI.

The deciding factors between public and private cloud are budget, staff, and the amount of data that needs to be stored. Public cloud is often the less expensive option for health systems that have a lot of unstructured data and a lower budget that can't cover private cloud deployments.

Considering object storage for the data lake

Object storage manages data as objects instead of files or blocks. Objects are kept in a storage pool that does not have a hierarchical structure. Instead, object storage uses unique identifiers that allow data to be stored anywhere in the storage pool. Storing data using object storage gives healthcare organizations more possibilities

for data analytics and offers a scalable infrastructure.

Building workflows and framework

Once data use is determined, workflows and frameworks need to be built to mine and analyse the data in a format that makes it usable for business decisions. Organizing data and making it accessible when needed is a key step in making the data actionable for analytics.

Recommended cloud platform for healthcare data lakes

While all three cloud options meet the requirements of a healthcare data lake, AWS has a clear edge over the others in its suitability and adaptability for the healthcare domain. Here are some reasons why.

- **Data movement made simpler and easier** - AWS provides multiple ways to move data from your datacenter to AWS. Tools such as AWS Direct Connect, AWS Snowball and AWS Snowmobile, and AWS Storage Gateway help with this. AWS also provides multiple ways to ingest real-time data generated from new sources such as websites, mobile apps, and internet-connected devices. Amazon Kinesis Data Firehose, Amazon Kinesis Video Streams, and AWS IoT Core are tools that can be used here. For the mix of structured and unstructured data that healthcare has, these tools enable data movement and storage in a highly efficient manner.
- **Object storage, backup and archival** - Once data is ready for the cloud, AWS makes it easy to store data in any format, securely, and at massive scale with Amazon S3 and Amazon Glacier. S3 is a scalable key-based object store that is well-suited for storing and retrieving large datasets. Due to its underlying

infrastructure, S3 is excellent for retrieving objects with known keys. S3 maintains an index of object keys in each region and partitions the index based on the key name. For large datasets like healthcare data and genomics, population-level analyses of these data can require many concurrent S3 reads by many executors.

- **Data cataloging** - Healthcare data often needs to be heavily catalogued. Before we can derive insights from the genomes of thousands of individuals, genomic data must first be transformed into a queryable format. AWS Glue is a tool that automatically creates a single catalog that enables it to be searchable and queryable on demand.
- **Data analytics** - AWS provides a broad, and cost-effective set of analytic services that run on the data lake. Each analytic service is purpose-built for a wide range of analytics use cases such as interactive analysis, big data processing using Apache Spark and Hadoop, data warehousing, real-time analytics, operational analytics, dashboards, and visualizations. The various tools of AWS are integrated in a way that makes most sense of healthcare data. For instance, if there is a brain scan in Amazon S3, you can use the deep learning AMI and P2 instance family to build machine learning models to identify images that represent different stages in your disorder-of-interest. Then run association analyses that combine genomics data with brain imaging models and drug response data to identify what genomic variants associate with improved treatment outcomes. You can manage these analyses via Jupyter notebooks, or you can connect with your BI tool of choice to visualize your results.
- **Predictive Analytics** - AWS also provides a set of machine learning services and tools that run on the data lake on These services come equipped with the knowledge and capability of the

domain and industry. The healthcare ecosystem has chosen a variety of tools and techniques for working with big data, but one tool that is most effective in this is Spark on Amazon EMR. Spark is uniquely capable in advancing genomics algorithms given the complex nature of genomics research. And this is one more advantage that AWS brings to healthcare data lakes.

Some best practices for data lake on cloud

- Do not treat data lake on cloud as a cheaper option of implementing a data warehouse or on-premises data lake. There needs to be a clear business need and data reason behind the design of the data lake.
- While doing Hadoop in the cloud, design around object storage. Object storage in the cloud adds to the complexity but is more flexible, cost effective and gives better performance.
- Data should be actively and securely managed.
- Load data into staging, perform data quality checks, clean and enrich it, steward it, and run reports on it completing the full management cycle.
- Collecting, storing, and using data produced by patients is a major challenge, and organizations often have data they can't sort through and use efficiently. Healthcare data lakes contain valuable information that can be used to improve patient care, but organizing the data and making it available takes significant IT infrastructure planning.
- Organizations preparing to make collected health data actionable should create a roadmap that will enable them to leverage data to improve workflow and patient care.

7. Case in Point – Healthcare data lake designed for a healthcare client

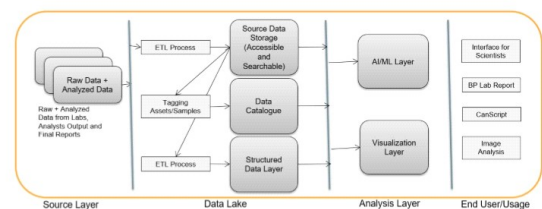
A global leader in advancing personalized oncology treatment approached us to build a data lake to meet their data requirements. The client was actively involved in supporting more effective and efficient cancer drug development, and wanted to build a data lake platform to consolidate clinical and research data generated while working on various samples and research processes.

Study of data and design of storage

Based on the types of data existing at the client site, the following were considered while designing the data storage:

- Use of an AWS S3 system, along with IAM and Access Security Policies
- Building of necessary connectors for different sources
- Building / enabling ETL hooks and API connectors to access the various pieces of data
- Import test data of different types as per need

Data cataloging and searchability



In order to collate and enable data access according to search criteria and need, the following were done:

- Build the search engine
- Ensure metadata tagging and searchability of all data
- Implement a governance model/rules and process

Designing the Sematic Layer

To set up the data lake on the cloud, the following was also done:

- Develop schemas, queries, stored procedures, etc.
- Test with sample datasets to validate
- Ensure records from clinical and pharma are being captured
- Data store in DW Logic approved / agreed to able to search and retrieve data quickly and also for analytics
- Redshift was chosen for the data warehouse services. Data stored in data warehouse was easy to search and retrieve for business intelligence reports and analytics

Designing the workflows and framework

As a final stage in the data lake setup, the following was done to ensure data was mined and reported as per the requirement:

- Developed the DevOps process
- Monitored/alerts/scheduling for failures/success
- ETL Framework based on AWS Glue to discover data and store the associated metadata in the AWS Glue Data Catalog
- Cataloged data was searchable using Elastic Search and available for various business application through various SQL queries.

Benefits

The following are the advantages to the user when we design and deploy the data lake for cloud:

- Automated data tagging and metadata-oriented search
- Catalog driven data governance
- Catalog metadata governance
- Schema definition
- Data population
- Query structured data

