SHAKUDO

Modern Data Stack for Healthcare

A Guide to Building a Scalable and Agile Modern Data Stack for Businesses in the Healthcare Industry

shakudo.io

Modern Data Stack for Healthcare

A Guide to Building a Scalable and Agile Modern Data Stack for Businesses in the Healthcare Industry

SHAKUDO

Introduction

From patient registries, administrative records, and insurance claims to imaging and laboratory test results, companies in the healthcare industry go through vast volumes of data on a daily basis. To extract valuable insights from such a data deluge, an agile and scalable data pipeline that can efficiently process, transform, and analyze data at scale needs to be in place.

Compared to traditional data management systems that often fall short when asked to process extensive amounts of heterogeneous healthcare data all at once, modern data stacks are much easier to scale due to their cloud-based architecture. Since each component within the data stack can work independently or interact with others to share data, companies can group these tools into customized categories, each corresponding to a specific aspect within the data workflow, such as data ingestion, cataloging, analysis, or storage. This allows companies to add, remove, or replace data components when existing systems have reached their capacity or when new tools of enhanced capabilities come out in the market.

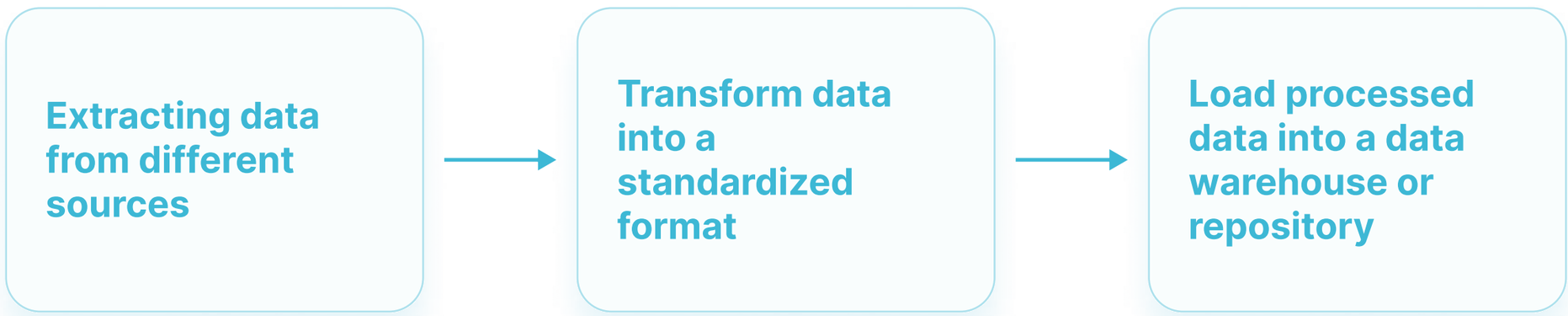
This white paper explores how companies can leverage a scalable and agile data stack to enhance their clinical data queries, secure in-house knowledge bases, and address challenges related to interoperability and data security. Since data in healthcare is often highly private, sensitive, and dynamic, having a scalable data pipeline that not only supports efficient data processing but also ensures security and compliance becomes essential. A well-designed data stack streamlines data processing, improves data quality, and facilitates real-time analytics, ultimately allowing companies to make informed decisions that enhance the patient care system.

Stage 1: Data Ingestion

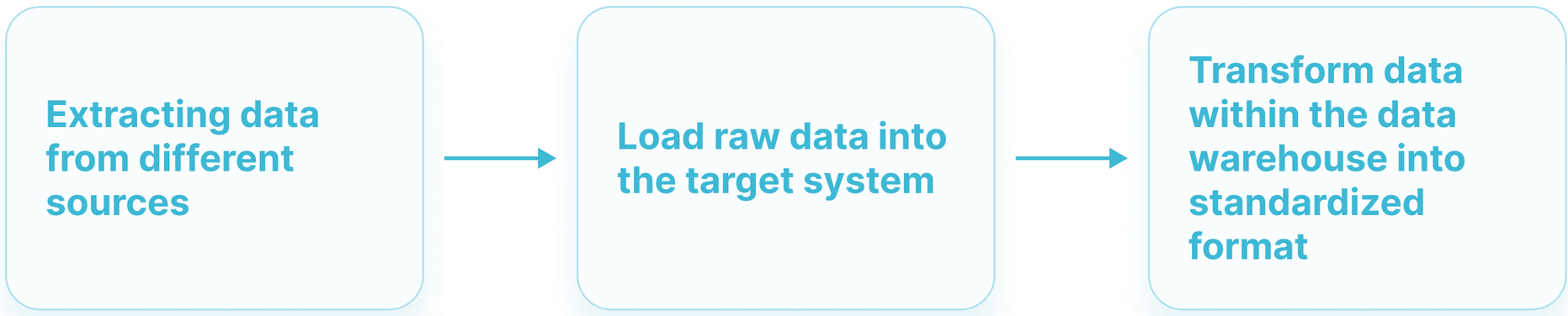
Data ingestion occurs at the initial stage of the data management process, where incoming data and information are collected from various sources and transformed into a centralized database for further analysis. Healthcare data often comes from disparate sources such as Electronic Health Records (EHRs), insurance claims, patient registries, surveys, and laboratory test results. Since these data sources can be extremely fragmented, choosing the right data tool means adopting a system capable of processing diverse data types promptly and safely.

1. ETL V.S. ELT

There are two predominant methodologies for data integration: ETL and ELT. While both involve extracting and loading data into a designated data warehouse, the key distinction lies in the location and timing of their data transformation process. Take a look at the graphs below:



ETL processes data before it reaches the warehouse, reducing the risk of sensitive data exposure while ensuring that the data coming into the warehouse has been transformed into a standardized format for effective management. This process also allows data to be masked or encrypted before storage, further mitigating the risk of data breaches. ETL tools are well-suited when data consistency, quality, and security are among your priorities. However, since all data extracted must be transformed to meet analysis requirements, the process can be time-consuming, making it less efficient when handling large-scale data sources.



ELT tools, on the other hand, allow data to be extracted and loaded directly into a designated data warehouse in its raw form. Standardization of the data format only occurs after data is being stored, offering systems more time and space for data to be processed, which can be particularly advantageous when dealing with complex or evolving data needs.

2. Data Connectors

Data connectors serve as the bridge between diverse, siloed data sources and a unified healthcare record center, allowing for seamless data synchronization across different platforms. Here's a breakdown of some of the most frequently used data connectors and their applications:

Direct Database Connectors

Direct database connectors integrate seamlessly with existing EHR systems, allowing for direct data transfer between databases. These connectors can provide faster data access without the extra layer of network communication, enabling quicker data retrieval for performance-centric applications.

Application Connectors

Application connectors facilitate data exchange between different software applications. They are crucial for integrating applications that serve business functions, like enterprise resource planning (ERP), customer relationship management (CRM), and marketing automation.

Cloud-Based Connectors

Cloud-based connectors are often scalable and secured; they simply workflow for transferring and integrating healthcare data from various sources.

On-Premises Connectors

On-premises connectors are crucial for integrating data from local systems within an organization’s infrastructure to other on-premise or cloud-based systems. They are commonly used when information needs to be shared between legacy data systems and modern applications.

File Connectors

File connectors are designed to manage files, documents, and directories on a file system. These connectors support syncing data from cloud storage platforms such as Google Cloud and OneDrive to on-premises sources like FTP (File Transfer Protocol) and SFTP (Secure File Transfer Protocol). They are essential for integrating file-based data into broader data workflows.

Event Connectors

Event connectors such as Broadcom, Azure Event Hubs, and Confluent come in handy when data streaming is time-sensitive–for example, data sourced from patient monitoring or fraud detection in billing systems.

API Connectors

API connectors use Application Programming Interfaces to connect and transfer data between different software applications. They are often used to extract data from SaaS platforms, CRMs, and other web services.

HIPAA EDI Transaction Sets

While not connectors themselves, these standardized formats facilitate data exchange between different healthcare systems as mandated by the Health Insurance Portability and Accountability Act (HIPAA). Key HIPAA transaction sets include Health Care Claim, Benefit Enrollment and Maintenance, Eligibility Inquiry and Response, etc. These transaction sets are integral to maintaining efficient and compliant electronic communication in healthcare, ensuring that data is exchanged accurately and securely across systems and organizations.

Shakudo comments

Having the capacity to handle large volumes of data efficiently while ensuring its timely analysis is among the top priorities when it comes to medical data management in the healthcare industry. To establish an effective data pipeline that can quickly address errors with minimal downtime, Shakudo recommends adopting a modern data stack that simplifies and streamlines the process. When selecting the right tools for data ingestion, make sure they are capable of:

Collecting massive volumes of medical and administrative datasets in disparate formats, such as doctor records, test results, and patient surveys within a short period of time;

Detecting anomalies during data transformation to identify potential fraud and prevent costly false positive alerts;

Keep in mind that although the speed of ingestion by ELT tools generally surpasses that of ETL due to reduced data movement, the overall effectiveness is also determined by factors like the complexity of the transformation and the capability of the data warehouse.

Stage 2: Data Processing And Analysis

Data processing and analysis are integral parts of the data workflow. Once data has been fed into the system, processing tools clean and organize the data, preparing it for analysis and insight extraction. When choosing the right tool to process and analyze incoming data, consider the following:

Ensuring Data Quality

Why	How
The success of healthcare analytics tools depends heavily on the quality of the data used to train them. High-quality data leads to analytics outcomes that are much more accurate and reliable.	Implement automated data tools to assist in data profiling, cleansing, and standardization, using predefined quality regulations to filter out inconsistent, biased, and false data.

Addressing Bias

Why	How
Bias in data can lead to flawed insights, perpetuate stereotypes, and discredit your company.	Develop strategies to identify and mitigate bias in data collection and analysis processes, such as diversifying data collection and implementing preprocess methods.

Prioritizing Data Privacy

Why	How
Ensuring patient data privacy is essential for adhering to regulations such as HIPAA and maintaining patient trust.	Implement robust encryption, access controls, and compliance measures to safeguard sensitive data.

Shakudo comments

There are numerous data processing tools in the market to help process and analyze your company data, each excels in different features and capabilities. To choose the right tool specifically for your organization, conduct an in-depth evaluation and have a list of priorities: What's the capability of your existing system? What are the main characteristics of your company data? Do you have any processing requirements? What are your user demographics? Once you assess these factors, you can compare your options and select a tool that best meets your data requirements.

Here's a brief overview of a few commonly used data analytics tools to get you started:

	Hadoop	Dask	Apache Spark	IBM SPSS
Scalability	★★★★★ Highly scalable and can handle large volumes of data across platforms	☆☆☆☆ Excels at handling large datasets and can scale from a single laptop to large clusters of up to 1,000 multi-core machines	★★★★★ Known for its fast processing capabilities, especially for large-scale data analytics and machine learning tasks	☆☆☆☆ May not be as scalable as other big data platforms like Hadoop or Spark
Cost-Effectiveness	★★★★★ As an open-source platform, Hadoop is cost-effective and can run on commodity hardware	★★★★★ Open-source platform, can be run on a single machine or small cluster, avoiding the need for expensive distributed computing infrastructure	☆☆☆☆ Spark's reliance on in-memory processing can lead to higher memory consumption, which may increase hardware costs especially for large-scale data processing tasks	☆☆☆ Can be costly, especially for smaller organizations or individual users
Flexibility	☆☆☆ Built to run on specific ecosystems such as Hadoop File System (HDFS) and Yet Another Resource Negotiator (YARN)	★★★★★ Provides more flexibility in terms of data processing workflows	★★★★★ Supports a wide range of storages and cluster managers, making it a versatile tool for various data processing needs	☆☆☆☆ Offers significant flexibility in its capabilities, making it a versatile tool for statistical analysis across various fields, however less flexible than programming-based alternatives like R or Python for highly customized analyses
Ease Of Use	☆☆☆ Has a steep learning curve, requires knowledge of Jaya and MapReduce programming model and can be complex to set up and manage	★★★★★ Offers a Pandas-like API, allows easy integration with other Python libraries and tools	☆☆☆☆ Provides high-level APIs in Java, Scala, Python, and R, making it accessible to developers with different programming backgrounds, however, optimizing Spark for specific workloads can be complex and may require deep technical expertise	★★★★★ Offers a user-friendly interface for statistical analysis, making it accessible to users with limited programming skills

Stage 3: Data Storage

Once data has been processed and analyzed, it will be consolidated into a data warehouse for storage. To choose the best-suited data storing system, consider the following:

HIPAA-compliant Cloud Data Warehouses

A HIPAA-compliant cloud data warehouse is a cloud-based storage solution designed to process electronic Protected Health Information (ePHI) in accordance with the stringent privacy and security regulations set forth by HIPAA. Here are some of its essential requirements:

Security, Integrity, and Availability

The cloud data warehouse must ensure the security, integrity, and availability of ePHI, safeguarding data both at rest and in transit through different types of encryption methods. You can also refer to the National Institute of Standards and Technology (NIST) guidelines to implement access controls and physical security measures to protect ePHI.

Business Associate Agreement (BAA)

Use a service provider that is not only willing to sign a BAA–a legal document outlining the responsibilities of the service provider in protecting ePHI and ensuring HIPAA compliance–but also has a proven track record of maintaining high compliance with HIPAA.

Access Control

Implement strict access controls such as role-based access control (RBAC), multi-factor authentication, and access logs to ensure that only authorized personnel can access ePHI. Maintain a detailed audit record as a reference to potential security breaches.

Risk Management

Deploy risk management strategies to enhance disaster recovery capabilities by having redundant resources, backups, and a disaster recovery plan. This ensures that ePHI remains accessible even during service outages or emergencies. Regular risk assessments also need to be conducted on a routine basis to identify potential threats to ePHI.

NoSQL Databases

Unlike SQL, NoSQL is a database management approach that stores data in a more natural and flexible way, making it easy to scale and adapt thus well-suited for modernizing healthcare data management with better data accessibility and integration. There are mainly four types of NoSQL databases in the market today: document databases that store data in JSON-like documents, key-value databases, wide-column databases, and graph databases.

Document Database

Document databases store data in documents similar to JSON where each document contains pairs of fields and values. These values vary from numbers, arrays, etc. It offers a flexible data model to process unstructured data sets with complex relationships or hierarchies.

```
{
  "name": "Shakudo",
  "function": company,
  "email": "info@shakudo.io",
  "address": {
    "street": "312 Adelaide St",
    "city": "Toronto"
  }
}
```

Key-value Database

Key-value databases use unique keys associated with different values to store and retrieve data. They are suitable for storing data such as session management, user preferences, product recommendations, and caching because they are highly partitionable.

```
Key: user:123
Value: {"name": "Shakudo", "email": "info@shakudo.io",
"function": "company"}
```

Wide-column Databases

Wide-column databases store data in tables, rows, and dynamic columns. These databases enhance performance by reducing the storage space through column compression techniques. The wide rows and columns allow efficient retrieval of sparse data, making them ideal for scenarios involving large and sparse datasets where many cells in a table are null.

```
Row Key | Name      | Email
user:1  | Shakudo   | info@shakudo.io
user:2  | Shakudo2 | info2@shakudo.io
```

Graph Databases

A graph database organizes data as nodes and edges–nodes represent entities such as people, places, or objects, while edges define the relationships between these entities. This structure is particularly effective for managing highly interconnected data where relationships might not be immediately apparent such as social media data and location-based service data.

```
(user1)-[:MANAGER_OF]→(user2)
(user2)-[:WORKS_AT]→(Shakudo)
```

Compared to traditional relational databases, consider NoSQL databases if the below features are among your priorities:

Scalability

NoSQL databases such as MongoDB can be scaled horizontally, meaning that it can process increased loads of data by adding more servers to the cluster. This is crucial when the system is asked to process and store large volumes of incoming medical data, such as records and files while maintaining high performance because no matter how much data a single storage device is capable of handling, the capacity will be reached.

Flexibility

Relational databases often require careful model planning before any data is loaded into the database–once the data has been loaded, however, it takes much more time to modify. In this case, the dynamic schema nature of NoSQL databases provides much more flexibility, making it quick and easy to process and store diverse data types.

Fault Tolerance and Easy Accessibility

NoSQL databases often employ data replication strategies as well as automatic partitioning to copy and distribute data across different nodes. This allows the system to keep running even after a server failure or when some nodes become unavailable, ensuring that the medical applications remain operational amid network outages and hardware failures.

Compatibility

NoSQL databases integrate well with various programming languages and frameworks, making it easier for healthcare companies to develop applications that need to interact with a guide range of systems and technologies.

User-Friendly Interface

NoSQL databases are known for their user-friendly interfaces as well as ease of use. Some of the tools provide graphical user interfaces, making it easy for staff within the healthcare industry to interact with the databases.

Cost-Effectiveness

Compared to relational databases, many NoSQL databases are available as open-source software, making it a cost-effective option for healthcare organizations. This is particularly advantageous for startups and businesses during their initial stages.

Shakudo comments

When it comes to choosing the right data warehouse to store healthcare data, it is always a good idea to use a combination of tools targeted to specific storage requirements. Softwares such as Amazon Redshift, Google BigQuery, Snowflake, and Microsoft Azure are just a few examples of data tools that offer HIPAA-compliant cloud data warehouses with hardware-accelerated encryption. These tools provide real-time predictive analytics for high concurrency and accessibility. Each platform excels in unique strengths, so choose the tool based on specific organizational needs such as scalability, ease of integration, and budget. Keep in mind the global regulations and standards that are essential to compliance maintenance.

Stage 4: Data Visualization And Reporting

Data visualization and reporting in healthcare play a crucial role by transforming complex data into easily digestible graphs and visuals that can be distributed across the organization. Several tools are used for healthcare data visualization, including Tableau, Power BI, QlikView, Looker, and Plotly. They offer functionalities like drag-and-drop interfaces, integration capabilities, and customization options to identify trends, patterns, and anomalies, transforming complex medical data into predictive analytics for informed decision-making.

Microsoft Power BI, for example, allows you to create visuals and dashboards from a wide range of custom visuals catering to different presentation needs. Tableau also offers an intuitive interface with strong visualization capabilities and real-time data analytics that allow users to create complex visualizations with minimal effort, making it suitable for enterprise-level applications. Amazon QuickSight is another cloud-powered BI platform that enables users to create interactive visualizations, reports, and dashboards from a variety of data sources.

Shakudo comments

Aside from utilizing data visualization and reporting tools to help interpret and present data, you can also implement machine learning models for predictive analytics, such as predicting patient readmission or identifying high-risk patients. Another approach is to apply Natural Language Processing (NLP) techniques to analyze unstructured data from clinical notes or patient feedback. Make sure that healthcare-specific standards such as HL7, FHIR, and DICOM are all integrated as part of your strategy.

Stage 5: Security And Compliance

Apart from establishing strict procedures, policies, and conducting regular risk assessments, rigid access control measurements such as Role-Based Access Control and Multi-Factor Login should also be implemented along with encryption strategies to safeguard medical data against unauthorized access and data breaches.

Shakudo comments

Safeguarding data against cyber threats is an ongoing topic that needs to be addressed across the board, regardless of industries and businesses. In the healthcare industry, in particular, companies should implement targeted data governance strategies to ensure the compliance of highly sensitive and private medical data. To learn more about how to build a robust data governance program that guarantees the reliability and safety of data, check out our comprehensive white paper on the essential steps to building an effective data governance framework.

Stage 6: User Accessibility And Experience

The final step to making an agile and scalable modern data stack accessible is to enhance its overall user experience–that is, to adopt data tools that are user-friendly and easy to understand so that everyone in the organization, regardless of their technical background, can quickly retrieve and extract useful information from the data collected. Tools such as Apache Superset, for example, provide no-code interfaces for building charts and diagrams, enabling users to create visualizations without programming skills.

On the other hand, user accessibility also extends beyond healthcare providers to patients. When patients have access to their health information through intuitive platforms, they are encouraged to actively engage in their healthcare journey, improving the overall quality of the treatment. Methods such as leveraging clinical data to enhance patient experiences through mobile applications and websites are helpful in ensuring that the information is not only accessible but also easy to digest for patients. Consider patient-centric applications such as Medidata Patient Cloud and Patient Centric Solutions (PCS) to provide patients with easy access to health information, appointment scheduling, and communication with healthcare providers.

Shakudo comments

Building an agile and scalable modern data stack for businesses in the healthcare industry can be rather challenging, especially when your company is short-handed with fewer tech-savvy employees. In such cases, having access to a comprehensive cluster of data tools that are effective, up-to-date, and easy-to-use can not only simplify your data management process but also improve its overall cost-efficiency.

To make sure that the data stack you have is not only best-of-breed but most up-to-date, Shakudo acts as an operating layer that manages and monitors the data and AI tools that run inside your cloud's VPC or on-prem. With over 160 top-tier data tools available on a single unified platform, Shakudo enables companies to get to production-grade outcomes with a reliable, performant, and cost-effective solution and tech stack while keeping your team focused on building core product capabilities.

To learn more about the dynamic data tools available on the Shakudo platform and how we can help deploy, monitor, and optimize your data stack, schedule a call with our Shakudo expert.

SHAKUDO

ABOUT SHAKUDO

Shakudo creates compatibility across the best-of-breed data tools for a more reliable, performant, and cost effective data and AI operating system. As an operating layer on top of your cloud Shakudo allows you to pick the best-of-breed data tools for your needs, while providing a platform with fully automated DevOps experience. This combines the best of both worlds in data stack practices so you can focus on delivering business value with data.

Shakudo is the most **easy**, **secure**, **cost-effective**, **scalable** way to bring the most advanced data and AI tools to your data. Find out more at **shakudo.io.**