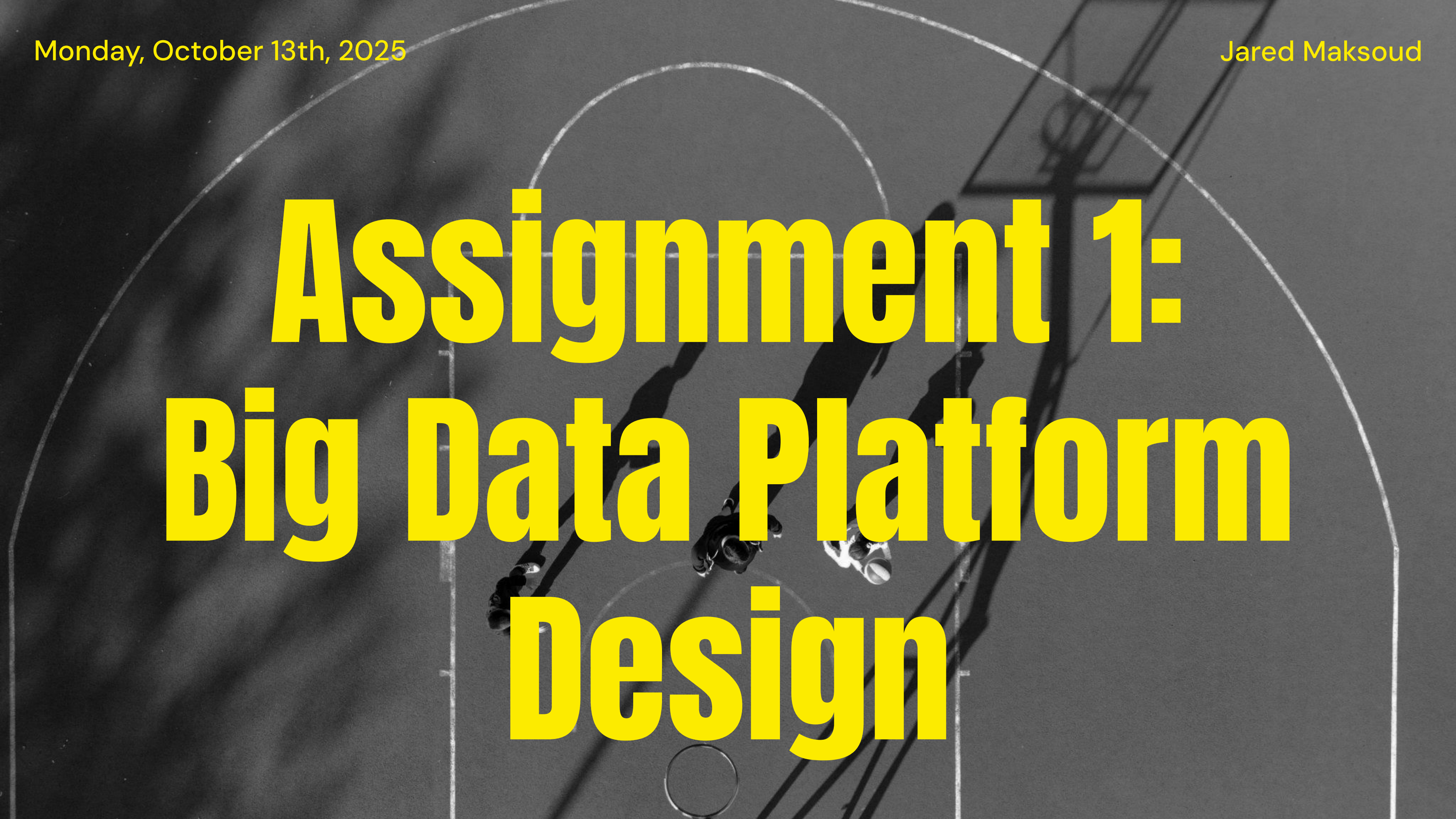
Monday, October 13th, 2025

Jared Maksoud

# Assignment 1: Big Data Platform Design



| Data Source | Description | Update Frequency |
|---|---|---|
| Electronic Medical Records (EMR) | Demographics, diagnosis codes, prescriptions, lab tests, treatments, hospitalization, etc. | Continuously updated |
| Administrative Data | Hospital discharge data and other operational records used for government reporting | Updated daily |
| Claims Data | Billable interactions between patients and healthcare system, third-party services, insurance firms | Updated daily to weekly |
| Patient and Disease Registries / Health Surveys | Clinical systems tracking key metrics of condition symptoms and nationwide hospital survey data | Updated quarterly |
| Bioinformatics from Health Apps and Wearables | Data collected from personal or clinical wearable devices that monitor health indicators | Updated daily |
| IoT / Telehealth Data | Real-time data streams from connected at-home medical devices and telehealth consultations. | Continuously updated |
| Imaging Data (X-Ray, CT, MRI, Mammogram) | Large visual data files generated by radiology departments | Updated on-demand |

# Part 1: Descriptive and Predictive Analytics Dashboard

## Descriptive

**Amazon EMR Dashboard**

- Patient performance…
  - Previous discharge outcomes, vitals, treatments, and medication history, comorbidities, and predispositions

## Predictive

**SageMaker Neural Network**

- Readmission rate prediction, retrained daily on incoming patient data
- Makes prediction using all data sources

## Data Sources

- EMR Data, lab reports, history (updated continuously upon record changes)
- Admission and insurance claim data (updated daily)

## Data Storage + Processing

- Store raw and static data in **Amazon S3**
- Query and aggregate trends in **Amazon Athena**
- Stream real-time patient data, and potential alerts, through **AWS Kinesis**

# Part 2: Patient Engagement and Remote Monitoring Platform

Collect data from patients, outside of the hospital, in real-time, to enhance patient dashboard and better prevent readmission

## Data Ingestion

- Continuous streams from IoT-connected wearables, home medical devices, and telehealth applications ingested through **AWS IoT Core**
- Real-time data flows into **AWS Kinesis Data Streams** for immediate processing and alert generation

## Data Processing

- **AWS Glue** performs ETL to clean, normalize, and enrich the sensor and app data
- Extracted features standardized using patient IDs and timestamps for integration with hospital data

## Data Storage

- Store raw and static data in **Amazon S3**
- Processed and aggregated data is joined with existing analytics dashboard data in **Amazon Redshift**

## Machine Learning

- **The SageMaker NN model** used in the hospital's predictive dashboard will be retrained nightly to predict patient readmission risk over the next week to determine if urgent treatment is necessary. If so, patient will receive an alert on their telehealth app
- Real-time inference combining both clinical (EMR) and behavioral (IoT) data, alongside all other available patient data, to predict readmission risk more accurately to provide them more enhanced and informed treatments

# Cloud Provider Selection

- Scalable data storage → **AWS S3** data lake
  - Highly durable (99.9999999%), HIPAA compliant, low-cost, scalable, data is continuously available
  - S3 data is losslessly compressed into parquet files
- **Amazon Redshift** data warehouse for analytics dashboard
  - Quick, large scale data querying and aggregation
- Using **AWS IoT Core** to enable medical-grade IoT data ingestion
- **AWS Kinesis Streams** for handling continuous IoT and wearable data
- **AWS Glue** for data combination, cleaning, and normalization
  - Prepares EHR, IoT, claim, etc. data as EMR visualization software input
- To allow SQL querying on data, using **Amazon Athena** for data aggregation
- Using **Amazon SageMaker**, a cloud ML service, to deploy a NN prediction model
- **Amazon EMR** to provide visual analytics dashboard using Spark, including NN results

# Data Sizing

## A few assumptions:

- The average patient produces ~80 megabytes each year in imaging and EHR data.
- Healthcare industry data growth is compounded annually at a rate of 36% through 2025.
- There will be new patients

Static Data Size Averages

- X-Ray: 10MB
- CT: 150MB-1GB
- MRI: 50-250MB
- Mammogram: 400MB
- 80% of all patients have at least ONE static medical image on their file, and every patient, on average, has three studies in their lifetime

**Relational Data:**

500,000 patients * 80MB = 40,000,000 MB = 40TB

**Static Data:**

400,000 patients * 100MB * 3 Studies = 120TB

**Wearables and Telehealth Data Streams:**

500,000 patients * 2MB/patient/year = 1TB

**Calculations:**

Year 1: 161.00TB
Year 2: 161.00TB * 1.36 = 219.96TB
Year 3: 219.96TB * 1.36 = 297.78TB
Year 4: 297.78TB * 1.36 = 405.00TB
Year 5: 405.00TB * 1.36 = 550.79TB
+ 50 TB buffer for scaling

**~ 600.000 TB required for S3**

# Capacity Sizing

## Amazon S3

Projected Year 1: 40 TB →
Projected Year 2: 137 TB
+ 50 TB for scalability = 187 TB

## Amazon Redshift

2 nodes of ra3.4xlarge (32 TB each) → If one fails, the other can continue serving queries

## Amazon IoT Core

20 devices to provide data stream to all twenty, on average, departments in the hospital

## Amazon Kinesis

14 shards to properly handle all IoT data to handle over 1 TB monthly

## Amazon Glue

10 DPUs (40 vCPUs, 160 GB memory) → ample computing power for all 500,000 patients

## Amazon Athena

32 DPUs of reserved capacity to aggregate data, powering Amazon EMR dashboard

## Amazon EMR

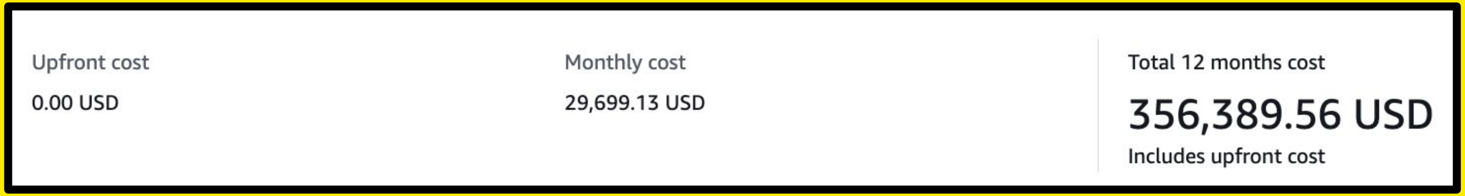
4 r6g.4xlarge core nodes → Scalable Spark cluster for dashboard

## Amazon Sagemaker

4 instances of ml.c5.large, containing the neural network model, ensure low latency predictions for incoming data streams

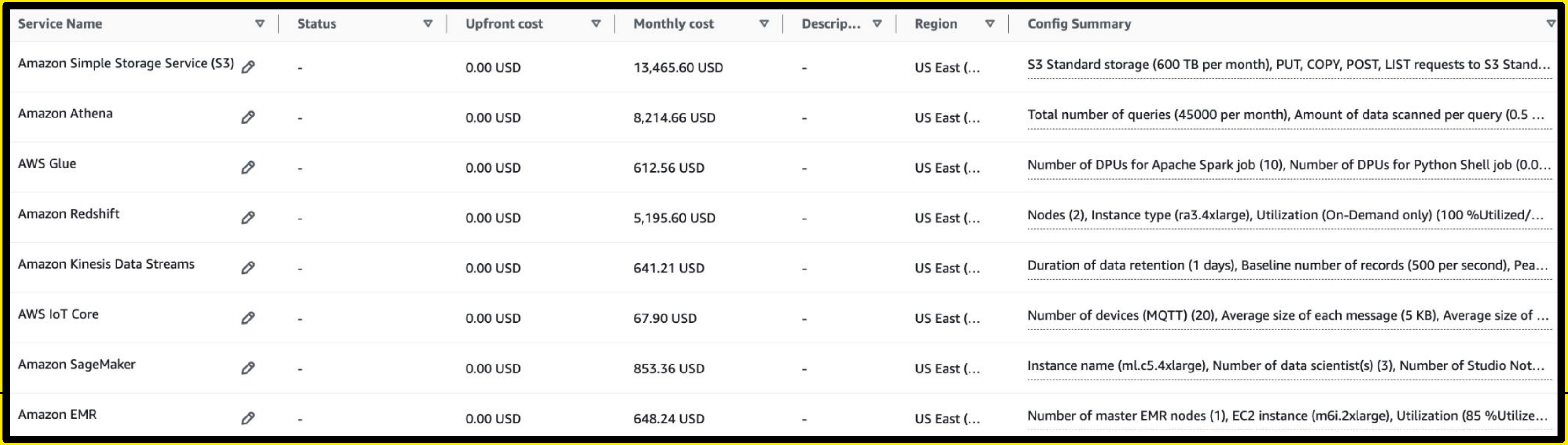
# Cost Summary

Estimate Summary:



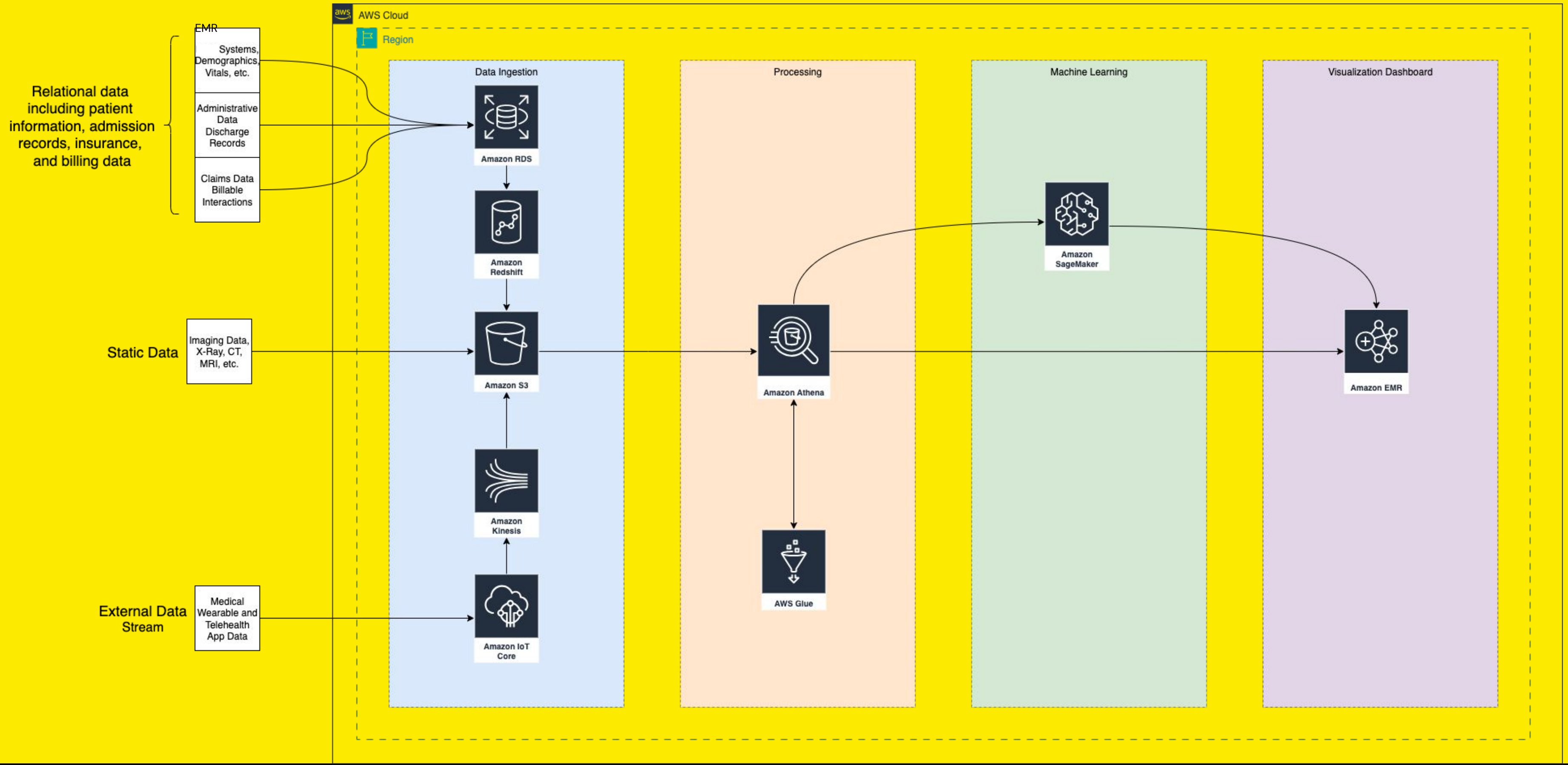
| Upfront cost | Monthly cost | Total 12 months cost |
|---|---|---|
| 0.00 USD | 29,699.13 USD | 356,389.56 USD<br>Includes upfront cost |

Estimate Breakdown::



| Service Name | Status | Upfront cost | Monthly cost | Descrip... | Region | Config Summary |
|---|---|---|---|---|---|---|
| Amazon Simple Storage Service (S3) | - | 0.00 USD | 13,465.60 USD | - | US East (... | S3 Standard storage (600 TB per month), PUT, COPY, POST, LIST requests to S3 Stand... |
| Amazon Athena | - | 0.00 USD | 8,214.66 USD | - | US East (... | Total number of queries (45000 per month), Amount of data scanned per query (0.5 ... |
| AWS Glue | - | 0.00 USD | 612.56 USD | - | US East (... | Number of DPUs for Apache Spark job (10), Number of DPUs for Python Shell job (0.0... |
| Amazon Redshift | - | 0.00 USD | 5,195.60 USD | - | US East (... | Nodes (2), Instance type (ra3.4xlarge), Utilization (On-Demand only) (100 %Utilized/... |
| Amazon Kinesis Data Streams | - | 0.00 USD | 641.21 USD | - | US East (... | Duration of data retention (1 days), Baseline number of records (500 per second), Pea... |
| AWS IoT Core | - | 0.00 USD | 67.90 USD | - | US East (... | Number of devices (MQTT) (20), Average size of each message (5 KB), Average size of ... |
| Amazon SageMaker | - | 0.00 USD | 853.36 USD | - | US East (... | Instance name (ml.c5.4xlarge), Number of data scientist(s) (3), Number of Studio Not... |
| Amazon EMR | - | 0.00 USD | 648.24 USD | - | US East (... | Number of master EMR nodes (1), EC2 instance (m6i.2xlarge), Utilization (85 %Utilize... |

# System Architecture Diagram


Relational data including patient information, admission records, insurance, and billing data
EMR
Systems, Demographics, Vitals, etc.
Administrative Data Discharge Records
Claims Data Billable Interactions
Static Data
Imaging Data, X-Ray, CT, MRI, etc.
External Data Stream
Medical Wearable and Telehealth App Data
AWS Cloud
Region
Data Ingestion
Amazon RDS
Amazon Redshift
Amazon S3
Amazon Kinesis
Amazon IoT Core
Processing
Amazon Athena
AWS Glue
Machine Learning
Amazon SageMaker
Visualization Dashboard
Amazon EMR

# Works Cited

Bali, Amit et al. “Management of medical records: facts and figures for surgeons.” Journal of maxillofacial and oral surgery vol. 10,3 (2011): 199-202. doi:10.1007/s12663-011-0219-8

BMD Software. “Still Using Gzip? Tackling Data Compression in Modern Medical Imaging Systems.” BMD Software, 16 Feb. 2022, www.bmd-software.com/news/tackling-data-compression-in-modern-medical-imaging-systems/.

Das, Subrata Kumar, and Mohammad Zahidur Rahman. “A secured compression technique based on encoding for sharing electronic patient data in slow-speed networks.” Heliyon vol. 8,10 e10788. 29 Sep. 2022, doi:10.1016/j.heliyon.2022.e10788

Eastwood, Brian. “How to Navigate Structured and Unstructured Data as a Healthcare Organization.” Technology Solutions That Drive Healthcare, 17 June 2025, healthtechmagazine.net/article/2023/05/structured-vs-unstructured-data-in-healthcare-perfcon.

“HCAHPS: Patients’ Perspectives of Care Survey.” CMS.Gov, www.cms.gov/medicare/quality/initiatives/hospital-quality-initiative/hcahps-patients-perspectives-care-survey. Accessed 13 Oct. 2025.

Kaylin. “Medical Record Data Storage & Retrieval: Healthcare Patient Data Storage.” Healthcare Data Management Software & Services | Harmony Healthcare IT, 4 Aug. 2020, www.harmonyhit.com/health-data-volumes-skyrocket-legacy-data-archives-rise-hie/.

Kholodenko, Alex. “Data Storage in Healthcare - Solutions and Installation Guides.” CodeIT, 25 July 2025, codeit.us/blog/data-storage-in-healthcare.

Li, Shunlei, et al. “Towards Scalable Medical Image Compression Using Hybrid Model Analysis - Journal of Big Data.” SpringerOpen, Springer International Publishing, 22 Feb. 2025, journalofbigdata.springeropen.com/articles/10.1186/s40537-025-01073-1.

“Library Guides: Data Resources in the Health Sciences: Clinical Data.” Clinical Data - Data Resources in the Health Sciences - Library Guides at University of Washington Libraries, guides.lib.uw.edu/c.php?g=99209&p=642709#12207216. Accessed 13 Oct. 2025.

Liu, Feng et al. “The Current Role of Image Compression Standards in Medical Imaging.” Information (Basel) vol. 8,4 (2017): 131. doi:10.3390/info8040131

Sachdeva, Sonali et al. “Unraveling the role of cloud computing in health care system and biomedical sciences.” Heliyon vol. 10,7 e29044. 2 Apr. 2024, doi:10.1016/j.heliyon.2024.e29044