MedGGenius Data-Transformation Report

BUAN 6335: Group 5

Authors: Abhigna Alagoni, Steven Evans, Sampreethi Bokka, Monish Ravishankar, Muhammad Ayas, Daniel Li, Venkata Boojala

Table of Contents

1. **Problem Introduction and Cloud Platform Selection ......................................................**
2. **Streaming Pipeline Description..........................................................................................**
3. **Structured Pipeline Description.........................................................................................**
4. **Unstructured Pipeline Description.....................................................................................**
5. **Management, Security and Governance............................................................................**
6. **ML Use Cases.......................................................................................................................**
7. **Implementation Roadmap..................................................................................................**
8. **Challenges.............................................................................................................................**
9. **Final Architecture Diagram………………………………………………………………**
10. **References.............................................................................................................................**

**Introduction**

Given our task of coming up with a common data platform for MeddGGenius there are many factors to consider. Our approach is to break the task into 4 different components and describe how they manage and handle the hospital's needs. The first three parts described in this report consist of three pipelines for various data types and that handle the transition of MeddGGenius’s legacy Teradata system to a modern cloud-based architecture through AWS and handle the end-to-end usage of streaming, structured and unstructured data. We will also cover how each of these pipelines plug into analytics and machine learning for business and managerial users. The last component describes how the data will be managed and governed to determine who gets access to what and how that data will be secured across pipelines.

**Why AWS?**

AWS was chosen as the foundational architecture due to the wealth of options it offers to meet MeddGGenius’ complex needs and to serve as a scalable foundation as the system evolves. Given the absence of budgetary constraints and MeddGGenius’ rapid growth, we wanted to create a plan that supports diverse options and constraints, while making the flow of data from producers to consumers easy to implement and understand. While mixing various cloud platforms for different services could reduce costs, we prioritized the quality of implementation, particularly concerning the need to stream data from patient sensors and in-house systems. The quality of the connections between these sources and their users is paramount to MeddGGenius, so we want to ensure the platform remains stable. The AWS platform provides many resources that enable flexible and straightforward implementation with the right guidance.

**Streaming Data Pipeline:**

**Data Sources:**

First, we will address the different streaming sources for MeddGGenius and the solutions we have developed to ensure security and ease of delivery for the various end users. The primary users of the streaming data within MeddGGenius are internal hospital personnel, such as operators and clinicians.

The streaming data to be ingested will come from three primary sources:

* In-house medical devices
* Patient monitoring devices
* Social media/online network sources to support the usage and improvement of devices and software

To manage the ingestion and delivery of the first source (in-house devices), we propose AWS IoT Greengrass as the primary processing and delivery method.

**Benefits and intended uses of AWS GreenGrass:**

* Supports the secure connection and data preprocessing from in-house devices that are firewalled and/or not primarily connected to internal or external internet to meet HIPPA security regulatory compliance (FDA, 2021). Below is a list of devices MeddGGenius uses that will need this secure connection:
  + Infusion pumps.
  + Ventilators.
  + Non-Wireless Patient Monitoring Systems
  + Anesthesia Machines
  + Radiology Equipment (CT Scanners and MRI machines)
  + Laboratory Instruments
  + Automated Medication Systems

These devices also are firewalled for security and privacy reasons to meet HIPPA regulations. GreenGrass allows for the intake and security of ingestion and preprocessing for these devices.

* Uses Lambda internally for local preprocessing of data without internal or external internet connectivity.
* Uses GreenGrass connectors for out of box integration for these devices when they are new or currently in usage.
* Uses docker containers to support the usage of custom internal or 3rd party applications for monitoring and clinician support.
* Pushes in house data securely to AWS cloud platforms for further analytics and machine learning possibilities.
* Specifically, for OLTP, GreenGrass provides PKI services that provide further security during transactions by encrypting and decrypting using different keys.

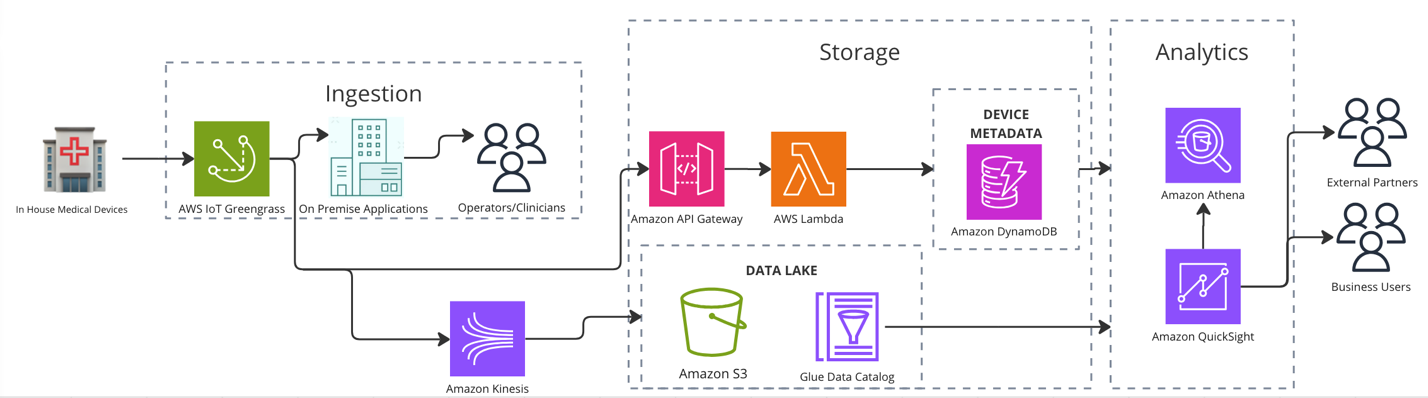
**GreenGrass Downstream connections:**

* Pushes data to existing on premise applications for fast delivery to clinicians and other on-premise staff.
* Pushes data downstream to Amazon Kinesis for further processing and analysis before storage.
* Pushes data downstream via an Amazon API gateway to support secure clearance of firewalled device data.

**GreenGrass Storage:**

All the data from these on-premise devices will be stored in a primary data lake in an S3 bucket that will be the basis for all the other pipelines storage as well. However, the storage of the metadata from these devices will be stored separately. The metadata will be pushed via an amazon API gateway and further processed via amazon lambda, after which it will reach its final storage destination of Amazon Dynamo DB. The purpose of this is to ensure the secure storage of this metadata for the future use and analysis of the streaming data. In a broad sense the data from GreenGrass will flow following the pattern in figure 1 below.

***Figure 1 (ML omitted for simplicity):***



**Handling of Other Streaming Sources:**

To handle the ingestion and storage of data from patient wearable devices, social media, and network streaming platforms we propose initializing and utilizing Amazons IoT Core service platform because while patient devices are not as heavily regulated, they still contain firewalls. That does not mean we propose sacrificing security as the metadata from these devices will be stored in the same manner as those from on-prem devices.

**Benefits and Downstream integration via Amazon IoT core:**

**Benefits:**

* Allows for the ingestion and processing of telemetry data from these devices using MQTT connection pathways.
  + MQTT is a standard pathway for connection for machine-to-machine communication. This will allow on prem devices to intake and communicate with patient devices for monitoring and downstream event detection.
  + MQTT requires minimal resources as most control messages only require as little as two data bytes to send.

**Downstream Connections:**

* Patient wearable devices are instrumental in continuous health monitoring, yet they can encounter complex events that may compromise data integrity and patient safety. AWS IoT Core, Amazon Timestream, AWS IoT Events, and Amazon Simple Notification Service can collaboratively detect and address such issues.
* **Potential Complex Events in Patient Wearable Devices:**
* **Sensor Malfunctions:**
  + Sensor failures can cause incomplete, missing, or inconsistent data for these devices and can cause interruptions in service (IoT for All).

**Data Transmission Failures:**

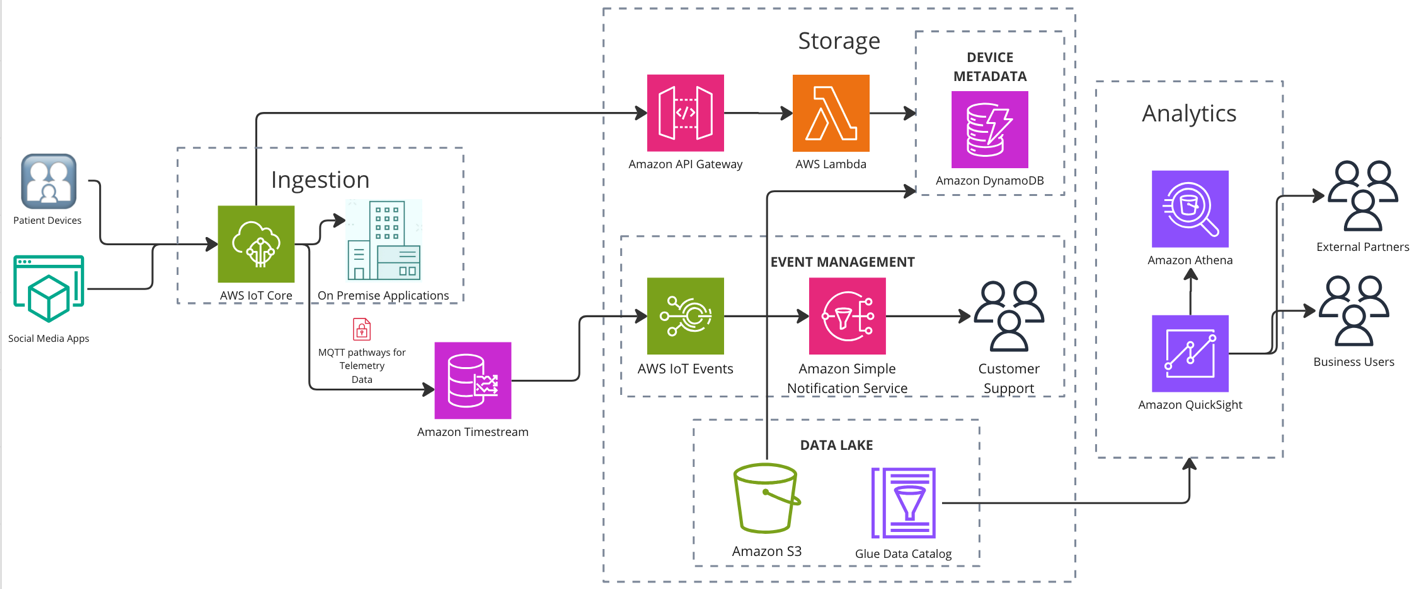
* + Issues with network connectivity can also cause interruptions and incomplete or missing data that requires prompt response by healthcare providers (MDPI, 2023).
* **Battery Depletion:**
  + Unexpected battery drain or powering down of devices perhaps during sleep monitoring can cause for unexpected interruptions in monitoring (AMA, 2023).
  + Communication to patients to change the battery or come in for service can be addressed promptly to ensure consistent tracking by healthcare providers.

**Firmware or Software Errors:**

* + Bugs and defects in patient devices can cause inconsistencies in data monitoring and require quick response by healthcare providers.

**Storage:**

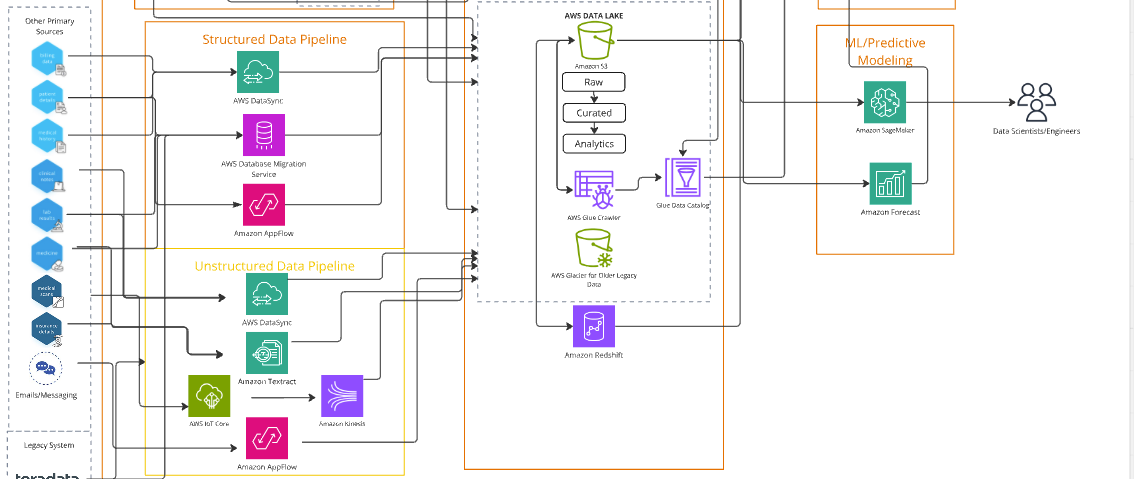
As with GreenGrass, the full set of streaming data from these sources will be sent to the data lake in S3 for quick analysis by the end business users via Amazon Athena and Amazon Quicksight and ML modeling by data engineers and scientists. The device meta data will be sent via the Amazon API Gateway and through Lambda for final storage in Dynamo DB. Figure 2 below shows the pathway these sources will follow into our full architecture.

*Figure 2 (ML Omitted for Simplicity):*

**Structured Data Pipeline:**  
 A good data structure is essential for securely processing data and information from diverse sources and achieving the best results through analysis and machine learning. Examples of information created in the record include patient demographics, appointment times, billing information, medical records, and test results.

This pipeline provides a unified platform for data ingestion, storage, transformation, and analysis by integrating with other MeddGGenius databases. The pipeline uses AWS services such as AWS DataSync, AWS Database Migration Service (DMS), Amazon AppFlow, Amazon S3, AWS Glue, and Amazon Redshift to provide secure, scalable, and efficient data and analytics.

*Figure 3: Structured and Unstructured Architecture Diagram*



The data pipeline process follows a design consisting of five main phases:

**1. Data Ingestion:**

The ingestion phase collects data from different sources to ensure seamless integration of pipelines and change security. This step combines data from local servers systems, SaaS platforms, and social media to create comprehensive data.

* **AWS DataSync**:
  + Automates the switch of historic datasets, consisting of patient medical histories and billing facts, fromlocal servers storage to Amazon S3.

Example: Moving CSV files containing legacy lab effects into the AWS information lake for further processing.

* **AWS Database Migration Service (DMS):**
  + Enables real-time replication and migration of transactional statistics from relational databases to Amazon Redshift.

Example: Synchronize ordinary updates from the sq. database containing billing information and appointment data into a pipeline.

* **Amazon AppFlow**:
  + Facilitates secure integration of structured data from SaaS packages like Salesforce.

Example: Captures patient scheduling and CRM data from Salesforce, transferring it to the pipeline for analysis.

**2. Data Storage**

Store processed records securely in repositories optimized for operational and analytical workloads.

* **Amazon S3:**
  + Acts as the important information lake for storing raw, transformed, and curated established facts.

Example: data historic patient visits records and clinical notes for compliance and regulatory purposes.

* **Amazon Redshift**:
  + Serves as the primary data warehouse for analytics and reporting. Optimized for querying dependent datasets and generating dashboards.

Example: stores curated billing and patient visit records, permitting business teams to investigate monthly revenue trends.

**3. Data Transformation**

The transformation stage cleanses, organizes, and standardizes ingested data to make it analytics prepared. This ensures that the data is consistent and reliable for downstream use.

* **AWS Glue Crawler**:
  + It automatically detects schemas for raw structured data stored in S3 and populates the Glue information Catalog.

Example: Scans billing information files saved inside the data lake and organizes them into metadata tables for querying.

* **AWS Glue ETL**:
  1. Cleans and preprocesses-based facts, dealing with operations like deduplication, normalization, and facts type standardization.

Example: cleans patient demographic records through disposing of duplicate entries and standardizing date formats.

**4. Data Analysis**

In the analysis stage, stored structured data is leveraged to generate insights for business users, clinicians, and data scientists.

* **Amazon Athena**:
  + Enables SQL-based querying of structured data stored in Amazon S3 without requiring additional infrastructure.

Example: Ad-hoc querying of archived patient demographics to identify trends in hospital admissions.

* **Amazon QuickSight**:
  + Provides dashboards and visualizations for operational and financial reporting.

Example: Generates actual-time dashboards displaying health facility occupancy quotes and monthly billing summaries.

* **Amazon SageMaker**:
  + Supports machine learning workflows, enabling predictive analytics.

Example: Builds predictive models to forecast patient readmission probabilities based on historical data.

5. **Governance and Compliance:**

Governance spans all stages of the pipeline to ensure data security and regulatory compliance.

* **AWS Lake Formation:**

Manages role-based access controls, ensuring only authorized personnel can view specific data. (e.g., clinicians vs. Finance staff).

Example: clinicians may be given access to patient medical records, while billing data is restricted to finance staff.

* **AWS Key Management Services (KMS):**

Encrypts data in storage (Amazon S3, Redshift) and transits data to protect sensitive information.

Example: Patient demographic data stored in S3 and queried through Athena is encrypted during storage and while accessed by an analyst.

**Benefits:**

* **Scalability**: Dynamically handles increasing data volumes using AWS’s elastic resources.
* **Compliance**: Meets healthcare standards like HIPAA through encryption and access controls.
* **Cost-Efficiency**: Uses cost-effective storage (S3) and serverless tools (Glue, Lambda) to optimize expenses.
* **Actionable Insights**: Provides real-time analytics and predictive capabilities to support clinical and operational decisions.

**Importance of the Structured Data Pipeline:**

* Handling high volumes of structured data like patient demographics, billing, and lab results.
* Ensuring HIPAA compliance for secure data processing.
* Supporting real-time operational analytics and machine learning models for predictive healthcare.

**Integration with the AWS Data Lake**

The structured data pipeline integrates seamlessly with the AWS Data Lake to manage data efficiently from ingestion to analytics. Raw data from sources like local serversystems and SaaS platforms is first sent to Amazon S3, which acts as the main storage layer. The data is processed using AWS Glue, which cleans, formats, and organizes it into a structured form. After transformation, the data is stored in a curated format, prepared for reporting and analysis.  
  
To maintain organization, AWS Glue Crawler examines the data stored in S3, identifies its structure, and updates the Glue Data Catalog with relevant metadata. This approach makes it easier to query the data using Amazon Athena or perform deeper analysis with Amazon Redshift. By integrating these tools, the pipeline enables real-time queries, advanced analytics, and predictive modeling, all while ensuring scalability and compliance.

**Unstructured Data Pipeline:**

File documents, reports, contracts have been around for decades. The importance of an unstructured data pipeline is to process these kinds of data that do not have a predefined format. Some examples of unstructured data sources for MeddGGenius and any company are medical images, social media content, and clinical notes.

The pipeline shows an unstructured data pipeline typical stages such as data ingestion, data preprocessing, data storage, data transformation, analysis, data visualization and reporting, and pipeline monitoring. We utilize a wide range of AWS services to help secure our data, make it flexible, cost-effective, and so much more.

1. **Data Ingestion:** This stage allows us to collect unstructured data and put it into the system.

* **AWS Data Sync:** Files such as PDF, text files, or scanned images are stored in a file system after using AWS DataSync and then transfers those files to Amazon S3.

Example: Some clinics/hospitals move soft copy medical data from local servers to cloud for processing them further.

* **Amazon Textract:** Takes scanned documents, forms, handwritten prescriptions, or PDFs and extracts structured data from unstructured documents. After processing the data with Amazon Textract, it can be integrated into an unstructured data pipeline for storage, processing, and further analysis.

Example: Parsing patient consent forms to store information about demographics.

* **AWS IoT Core and Amazon Kinesis:** Devices send unstructured data and route it to Amazon Kinesis for real-time streaming and analytics.

Example: Streaming data from Apple watch or other similar devices.

* **Amazon AppFlow:** Allows direct integration with many external applications, ensuring seamless ingestion.

Example: Synchronizing messages/emails having feedback data into the pipeline to perform sentiment analysis.

**Benefits of Data Ingestion:**

* Automates the process of data collection, highly reducing manual interventions.
* Works with real-time ingestion for quick analysis.

1. **Data Storage:** The objective of this step is to efficiently store unstructured data for accessibility.

* **Amazon S3:** Storage that can seamlessly store raw, curated, and analytics-ready data. We can also use this as the central data repository by organizing the data into separate layers in S3; you enable an efficient workflow for processing, querying, and analyzing diverse datasets.

Example: Storing soft copy images for future analysis.

* **Amazon Redshift**: This is a cloud-based data warehouse service perfect for analytics-ready data to enable faster queries and real-time analysis.

Example: Keeping analyzed customer data from a retailer to generate useful business insights.

**Benefits of Data Storage:**

* Easily Scalable to support handling of huge size of unstructured data
* Storage solution which is highly cost effective

1. **Data Transformation:** The goal is to transform or convert the unstructured data into semi-structured or structured format for analysis.

* **AWS Glue Crawler:** This tool helps scan data sources and catalogs metadata for all datasets. An example can be for the clinical notes, we will use Glue Crawler to catalog structured and semi-structured data extracted from lab results and notes.

Example: Analyzing metadata from scanned PDFs.

* **AWS Glacier:** AWS Glacier is an ideal solution for storing older legacy data since it does not require frequent access, but must be retained for compliance, audits, or historical value. For example, for lab results, the user is able to archive raw files such as PDF in AWS Glacier if they aren’t accessed frequently.

Example: Archiving 5-year-old student grade data to comply with university policies.

**Benefits of Data Transformation:**

* Data Standardization
* Highly Cost Effective

1. **Data Analytics/Analysis:** This is one of the most important stages, where you generate and discover insights from data which can help improve decision-making, increase efficiency, and many more.

* **Amazon Athena:** A serverless, interactive query service that allows you to run SQL and ad-hoc queries on structured or semi-structured data directly on data stored in Amazon S3.

Example: Searching for similar keywords in a customer data

* **Amazon** **QuickSight:** A business intelligence service that allows you to perform visualization in BI tools and create dashboards and reports.

Example: Generating real time dashboards and graphs of revenue/sales.

**Benefits of Data Analytics/Analysis:**

* Helps leadership teams with business and recommended insights.
* Visual graphic insights support quick decision making.

1. **Machine Learning/Predictive Modeling:** A field of artificial intelligence that enables computers to learn and analyze data and even make predictions.

* **Amazon Sagemaker:** This service contains a variety of machine learning frameworks that enables the user to utilize advanced ML techniques and use cases like image processing, natural language processing, classification, regression, clustering, recommendation systems, and anomaly detection for unstructured data.

Example: Training an NLP model to identify customer type through his face detection.

**Benefits of Machine Learning/Predictive Modeling:**

* Supports predictive analytics
* Automates super complex tasks

**Conclusion:** It is important to have an unstructured data pipeline in any organization’s data-driven environment since it enhances scalability, improves data organization and accessibility, and facilitates real-time data processing. Since most of data comes from images, PDFs, emails, and many others with an unstructured format that needs converting, having a well-defined pipeline is now essential more than ever.

**Data Management, Security, and Governance:**

To ensure robust data management, security, and governance in the healthcare sector it is important to maintain compliance, operational efficiency and very importantly, patient's trust and faith. For MeddGGenius, the digital transformation initiative led by CDO aims to create a secure, compliant, and scalable data infrastructure. The CDO is tasked with addressing the current pain points and addressing them by integration of disparate data sources, protecting the sensitive information of patients and ensuring data quality and accessibility across different departments.

MeddGGenius must focus on complying with HIPAA compliance, NIST cybersecurity standards and advanced data governance while moving to modern data architecture. Given the variety of structured, semi-structured, and unstructured data in MeddGGenius’s ecosystem, the following strategies will be implemented:

**Unified Data Storage:**

The new data platform will adopt a lakehouse architecture, combining the best aspects of data lakes and data warehouses. This can allow both structured and unstructured data storage like Electronic Health Records (EHR), X-rays, MRIs, CT scans, IoT data. Data will be stored in Amazon S3 for scalability provided by the services, where metadata is managed by AWS Glue.

**Master Data Management:**

To ensure single source of Truth, MeddGGenius will employ a Master Data Management strategy to compile all the data from inventory, staff to the patients. MDM helps us in not having duplicate, consistency in records of patients and inconsistency across systems.

We have chosen AWS as our suite for building the new architecture because it addresses all the critical aspects of data security, compliance, and governance.

*Figure 4: Proposed Amazon MSG Applications*



Given the strict nature of the healthcare industry in general, and considering HIPAA compliance, among other data integrity and confidentiality concerns, these services provide an end-to-end solution for security management, threat detection, and governance. Below, we outline each chosen AWS service and why it is best suited for the MeddGGenius platform.

**AWS Identity and Access Management (IAM):**

AWS IAM is fundamental for controlling access to resources within the MeddGGenius platform. Given the sensitivity of patient data, ensuring that only authorized personnel can access specific data is critical for HIPAA compliance.

**Amazon GuardDuty:**

It is a threat detection service that continuously monitors for malicious or unauthorized behavior. Given the healthcare sector's vulnerability to cyberattacks, proactive threat detection is essential.

**AWS Config:**

AWS Config is necessary to track the changes in the configuration across the platform. Maintaining a consistent and compliant configuration is critical, given the need for compliance with HIPAA and NIST standards.

**AWS Key Management Services:**

AWS Key Management Service (KMS) is a managed service provided by Amazon Web Services (AWS) that allows companies to create, control and manage the cryptographic keys that encrypt and protect their data. KMS keys can be used to encrypt/decrypt data encryption keys used to encrypt PHI in a customer’s applications or in AWS services that use AWS KMS. AWS KMS can be used in conjunction with a HIPAA account, but PHI can only be processed, stored, or transmitted in HIPAA Eligible Services. AWS KMS is normally used to generate and manage keys for applications running in other HIPAA Eligible Services.

**AWS WAF:**

AWS WAF protects web applications and APIs from common web threats, such as SQL injection, cross-site scripting (XSS), and DDoS attacks.

**Amazon Macie:**

Amazon Macie is a data security and privacy service that utilizes machine learning to identify sensitive data, such as personally identifiable information. In healthcare, data privacy is not an option. Amazon Macie discovers sensitive data using machine learning and pattern matching, provides visibility into data security risks, and enables automated protection against those risks. Macie stands out for its ability to secure sensitive data with minimal manual intervention. By leveraging machine learning, it provides automated discovery and classification of data, enabling organizations to focus on strategic security measures rather than routine data audits.

**AWS CloudTrail:**

AWS CloudTrail provides detailed logging of user activities and API calls in the AWS environment. This is needed for auditing, security investigations, and tracking accountability.

**ML Use-Cases:**

**Staff Scheduling Optimization:**

With recent advancements in machine learning for predictive analytics and optimization, improving staff scheduling has become more practical than ever. By equipping data scientists with relevant data to forecast staffing demands and create efficient schedules, hospitals can enhance operational efficiency, lower costs, and boost employee satisfaction.

Using a combination of Reinforcement Learning (RL), Linear Programming (LP), and Gradient Boosted Decision Trees (GBDT), it is possible to optimize efficiency, reduce costs and improve employee satisfaction.

* RL dynamically adapts to changing demands by learning optimal scheduling policies through sequential feedback, effectively handling unexpected patient surges.
* LP ensures in meeting the constraints like staff availability, Labor laws and patient-to-staff ratios.
* GBDT predicts the staffing needs by using historical data of patient volumes, increasing accuracy.

**Medical Imaging Analysis:**

Recent advancements in machine learning, particularly in image classification and deep learning, have significantly improved the speed and accuracy of medical imaging analysis. With these technologies, hospitals can achieve more precise diagnoses, reduce the risk of human error, and deliver faster results for patients. By working with comprehensive medical image data, data scientists can create models that help doctors interpret imaging scans more quickly and reliably, ultimately improving patient care and outcomes.

Medical Imaging analysis can be used to achieve faster diagnosis and reduce human error by leveraging Convolutional Neural Networks, Transfer Learning, and AutoML.

CNN helps in finding patterns in patient’s reports such as X-rays, MRIs and CTscans. This can help in accurately detecting fractures, tumors, or infections.

Transfer learning is a method that can facilitate acceleration of model training by making use of pre-trained models fine-tuned for specific image-based analysis, this can improve the performance of models on smaller datasets.

In addition, AutoML for Vision simplifies this process by automating the selection and optimization of models, thus making advanced image analysis accessible to clinicians without extensive ML expertise.

These techniques will enhance diagnostic accuracy, accelerate the analysis process, and reduce the cognitive burden on radiologists to improve patient outcomes and workflow efficiency.

**Disease outbreak Prediction:**

As we all saw in the wake of the COVID-19 pandemic, hospital and healthcare systems can be put under massive strain when disease outbreaks occur. These events cannot always be prevented, so being able to predict them and reduce their impact is critical to hospital infrastructure and patient care. ML models offer good predictive methodologies when trying to make these predictions such as:

* Time Series forecasting to detect early changes in patient inflow from new and past diseases. This can provid MeddGGenius with the capability to quickly detect and respond to rapid increases in patient demand of resources and care with the proper resource application.
* Recurrent Neural Network implementation to track existing patient health documentation and critical response to outbreaks to prevent further spreading among MeddGGenius’ hospital systems.

**Surgical Outcome Prediction:**

Given the improvement of ML technologies in image classification and deep learning tasks, predicting and improving surgical outcomes is more feasible than it has been in the past. Enabling Data Scientists with the necessary data to predict and plan surgeries can lead to improved patient outcomes and better healthcare strategies. During the pre-planning phase of surgeries models can be trained to:

* Predict anomalies in diagnostic imaging before surgeries occur so that healthcare professionals are better prepared for operating procedures and know exactly what the patient needs before going into surgery.
* Predict patient recovery times post-surgery given their prior history to improve and optimize patient outcomes. This will result in more cost savings due to less frequent errors during and after surgery based on more accurate diagnosis.

**Operational Shortage Detection:**

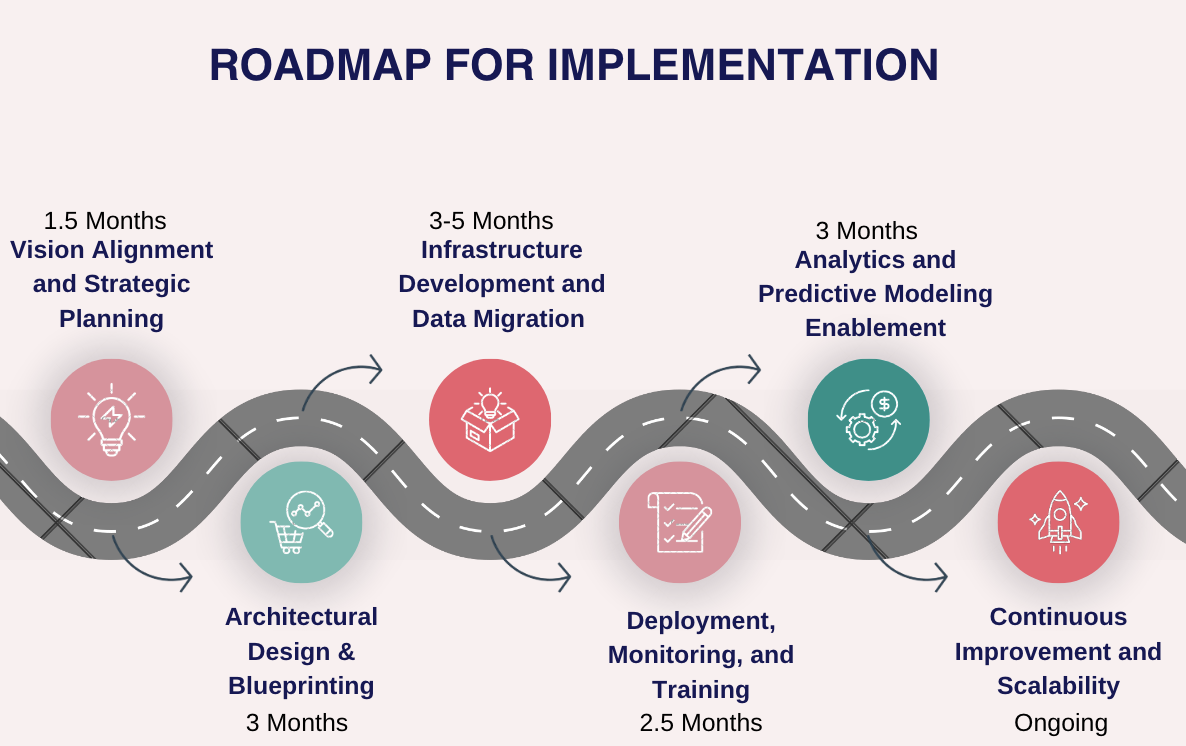
Operational shortages are a problem in many industries and in healthcare it is crucial to patient and staff operations. When data scientists are given the necessary data in a streamlined manner it can lead to better demand forecasting for more cost-efficient purchases. Some methods that can be used here are:

* Time Series forecasting to detect seasonal changes in demand of hospital resources and systems so that managers can improve decision making and cost savings.
* Efficient anomaly detection of shortages so that managers and employees can more rapidly respond to these challenges and reduce downtime.

Conclusion Implementation Time and Roadmap:

Given the complexity of such a large data platform migration and challenges that can occur along the way, we propose a gradual strategic implementation. Figure 5 below shows an image of what our proposed roadmap looks like in a general sense.

*Figure 5: Implementation Roadmap*



**Vision Alignment and Strategic Planning (1.5 months):**

Outside of our team, prior to full implementation we propose having a meeting with the designated team stakeholders outside of the data pipeline to fully assess the needs of each part of the organization. This step is crucial to the overall development of the required pipelines and given our proposal, changes can be made as necessary to accommodate the vision alignment and strategic initiatives of MeddGGenius. Cost, capacity and consistency are critical at this juncture, and we want to make sure all bases are covered prior to beginning the transition. Given the importance of cross-functional alignment, the meetings will serve to identify and necessary adjustments early in the process.

To allow sufficient time for scheduling, discussions, and feedback incorporation we suggest completing this phase over the next six weeks. We also want to propose that each meeting keeps adequate documentation, stakeholder input and a risk assessment so that all bases are covered prior to beginning implementation.

**Architectural Design and Blueprinting (3 months):**

After management has time to review our proposal and come up with the necessary changes to ensure all storage, processing, cost, and infrastructural needs are met, we propose a final planning to get a final architectural plan created prior to implementation. This proposal is a draft, and we want to ensure that changes are made to ensure a smooth transition from our current Terradata platform. Another key reason we propose this portion is so that all teams that will be involved in the data transition can be ready and equipped with the necessary diagrams and infrastructural changes that will be made along with potential challenges that could arise in the process. We propose submitting a proof of concept alongside the documentation passed throughout the organization to increase the likelihood of bigger buy-in and understanding across the key departments such as data engineering, analytics, customer service, on premise staff and managers. It will also give HR proper notice to ensure that any new staffing is prepared for when new analytical possibilities and ML options are fully implemented. **For a full image of our proposed architecture diagram see figure 6 at the conclusion of the report.**

**Infrastructure Development and Data Migration (3-5 months):**

Once the final blueprints are drawn up, we propose that the initial stages of the new pipelines be set up and brought in via an experimental process. Given the critical importance of the use of streaming data within MeddGGenius data infrastructure, we propose that this pipeline be experimented with and implemented **within the first month of this stage**. The experimental design of implementation during this phase can be carried out in a manner that keeps the necessary scaling in mind. Also, given the size and complexity of the hospital systems and current infrastructure we suggest starting out with the hospitals that have more stable systems already in place where micro adjustments can be made and scaled out until the pipelines can be fully implemented. We want to ensure that during this phase, any needed adjustments are made so that future implementations at other hospitals can go more smoothly and be scaled out and up efficiently.

Structured and unstructured data ingestion will begin concurrently to support the transition from Terradata without immediate disruption. A parallel run, where both the new and existing systems operate simultaneously, will allow us to validate consistency and identify any discrepancies in real-time.

Testing and validation will be critical during this phase. Regular checkpoints with stakeholders will ensure issues are addressed promptly. Additionally, staff training will begin here to familiarize users with the new systems and minimize disruptions during the full migration.

Additionally, we propose a dedicated phase for monitoring the performance of the new pipelines and making the necessary adjustments. This phase will consist of end-to-end performance validation and optimization of processing times and collecting of user feedback. Any identified challenges will be rectified to ensure the system meets the needs of MeddGGenius’ long-term goals.

**Deployment, Monitoring and Training (2.5 months):**

Once infrastructure development and data migration are complete, we propose initiating a full deployment phase. This phase focuses on transitioning all operational pipelines into a production environment. We want this phase to ensure a consistent and full transition of all systems to the usage of the new pipelines and infrastructure.

To achieve this, deployment will occur in stages, starting with smaller, less critical environments to provide validation under production conditions. Any necessary adjustments identified during this stage will be taken care of before deploying across the rest of the organization. Collaboration with stakeholders will ensure alignment with the operational readiness of the organization.

Following deployment, continuous monitoring of the new pipelines will ensure system stability, data accuracy, and performance optimization. During this phase, any errors, bottlenecks, or discrepancies will be addressed promptly. A robust monitoring framework will be established to provide insights into pipeline health and processing efficiency.

Simultaneously, training sessions will be conducted for staff and stakeholders who will interact with the new system. These sessions will include tailored programs for different user groups, such as analytics teams, customer service representatives, and hospital managers, to ensure they can effectively leverage the system’s new capabilities. Regular feedback loops will be established to refine both system operations and user training materials.

**Analytics and Predictive Modeling Enablement (2 months):**

Once the system is fully enabled, we will focus on giving data engineering, analytics, and science teams the necessary tools for prescriptive and predictive modeling capabilities. This phase involves integrating machine learning tools and analytical frameworks to obtain the potential of the infrastructure.

Key deliverables include:

* Developing and deploying predictive models tailored to organizational needs.
* Improving reporting and visualization tools to improve decision making across the organization.

This phase will require much cross-functional collaboration among various technical and business teams such as data engineering, IT, management, and operations to ensure user and client needs are met and decision-making capabilities are improved.

**Continuous Improvement and Scalability (Ongoing):**

This final phase focuses on the long-term improvement and scalability of the existing system to adapt to the rapidly changing healthcare industry.

Activities include:

* Assessing system performance and incorporating new technologies.
* Establishing a feedback loop for continuous and ongoing improvement of infrastructure and analytics capabilities.
* Scaling the infrastructure to support new hospitals, departments, or use cases as needed.
* Revisiting data governance, security, and compliance protocols to ensure they remain up to date with HIPPA standards and regulations.

This phase positions MeddGGenius with the tools to adapt and flexibly meet the needs and demands of its users, employees, and regulatory agencies.

**Advantages:**

1. **Real-Time Patient Monitoring**

Wearable fitness gadgets provide continuous updates on affected person vitals consisting of heart price and blood pressure. Those devices ship on the spot alerts to medical doctors if irregularities are detected, allowing for activate medical action. for example, an unexpected drop in oxygen levels ought to cause an alert, assisting save you critical headaches.

1. **Efficient Hospital Operations**

By using real-time dashboards, hospital staff can track bed availability, emergency room occupancy, and resource usage. This ensures that administrators can act quickly to optimize patient flow and resource allocation during peak times.

1. **Streamlined Data from Clinical Systems**

Appointment schedules and other patient data from systems like Salesforce are automatically updated across all departments. This reduces delays in communication between staff, ensuring that everyone has the latest information.

1. **Predictive Maintenance of Equipment**

Using connected devices, hospitals can monitor the health of critical equipment like ventilators or dialysis machines. Predictive tools help schedule maintenance before failures occur, ensuring uninterrupted patient care.

**Challenges:**

1. **Delays in Accessing Data Across Regions**

If patient records are stored in one geographical region and accessed from another, there can be delays. This becomes an issue during emergencies when immediate access to patient history is crucial.

1. **Compatibility with Older Devices**

Some hospital equipment may not integrate easily with modern cloud platforms. For example, older machines may need additional systems or custom solutions to send data in real time, which can increase setup time and costs.

1. **Dependence on Cloud Services**

The reliance on cloud-based tools like Amazon Kinesis or IoT Events can pose a risk if there are outages. A failure in these systems might delay critical alerts, such as vital sign irregularities, potentially endangering patients.

1. **Meeting Strict Healthcare Regulations**

Real-time data must remain secure and compliant with healthcare regulations like HIPAA. This requires constant vigilance to ensure all data handling processes, from transfer to storage, meet regulatory standards. Any misconfiguration could result in fines or breaches of patient trust.

-x-

*Figure 6: Full Architecture Diagram*

A diagram of a software company

Description automatically generated with medium confidence

**References:**

* + - 1. **FDA - Medical Device Cybersecurity:** U.S. Food and Drug Administration. (2021). *Medical device cybersecurity: What you need to know*. Retrieved from <https://www.fda.gov/consumers/consumer-updates/medical-device-cybersecurity-what-you-need-know>
      2. **AWS - IoT Medical Devices:** Amazon Web Services. (n.d.). *AWS architecture for medical devices*. Retrieved from <https://d1.awsstatic.com/architecture-diagrams/ArchitectureDiagrams/iot-medical-devices-ra.pdf?did=wp_card&trk=wp_card>
      3. **Keyfactor - Public Key Infrastructure (PKI):** Keyfactor. (n.d.). *What is PKI?* Retrieved from <https://www.keyfactor.com/education-center/what-is-pki/#:~:text=practice%20PKI%20programs.-,What%20are%20Common%20Challenges%20that%20PKI%20Solves?,Email%20and%20data%20encryption>

1. **AWS - Streams Introduction:** Amazon Web Services. (n.d.). *Introduction to Amazon Kinesis Data Streams*. Retrieved from <https://docs.aws.amazon.com/streams/latest/dev/introduction.html#:~:text=Accelerated%20log%20and%20data%20feed,Streams%20applications%20and%20data%20streams>.
2. **AWS - Using Athena and Quicksight:** Amazon Web Services. (2021). *Use Amazon Athena and Amazon QuickSight in a cross-account environment*. Retrieved from <https://aws.amazon.com/blogs/big-data/use-amazon-athena-and-amazon-quicksight-in-a-cross-account-environment/>
3. **IoT For All. (n.d.). How IoT devices transform healthcare with real-time data collection.** Retrieved from [https://www.iotforall.com](https://www.iotforall.com/)
4. **MDPI. (2023).** Lack of connectivity and disruptions in IoT health devices. *MDPI Journal of Diagnostics, 14*(18), 2037. Retrieved from [https://www.mdpi.com](https://www.mdpi.com/)
5. **American Medical Association. (n.d.). 4 mistakes your patients should avoid with wearables.** *AMA Insights.* Retrieved from [https://www.ama-assn.org](https://www.ama-assn.org/)
6. **Therapeutic Goods Administration (TGA). (2022). Actual and potential harm caused by medical software.** *TGA Report.* Retrieved from [https://www.tga.gov.au](https://www.tga.gov.au/)
7. **Cleio Insights. (n.d.). Wearable medical devices: Challenges in design.** Retrieved from [https://cleio.com](https://cleio.com/)
8. **Amazon Web Services. (n.d.). Replication instance: AWS Database Migration Service.** Retrieved December 7, 2024, from <https://docs.aws.amazon.com/dms/latest/userguide/CHAP_ReplicationInstance.html>
9. **Amazon Web Services. (n.d.). Connectors available in Amazon Athena.** Retrieved December 7, 2024, from <https://docs.aws.amazon.com/athena/latest/ug/connectors-available.html>
10. **Amazon Web Services. (n.d.). AWS AppFlow.** Retrieved December 7, 2024, from <https://aws.amazon.com/appflow/>