Modern Data Platform for MeddGGenius:

A Cloud-Based Data Lakehouse Solution

BUAN 6335.502 Organizing for Business Analytics Platforms
Professor Mandar Samant

Group 2
Adithya Ramakrishnan, Eunsung Choi, Jurgen Wulur, Kartik Shah,
Prasanth Chowdary Yanamandala, Richard Yang, Tanmay Raju Shedge

1. Introduction and Problem Statement

› Executive Summary

MeddGGenius, a regional medical center, faces increasing competition and operational inefficiencies due to outdated legacy systems reliant on distributed RDBMS and Teradata warehouses. These systems struggle with growing data volumes, especially unstructured data like medical images and IoT sensor streams, resulting in data silos, high maintenance costs, and poor scalability.

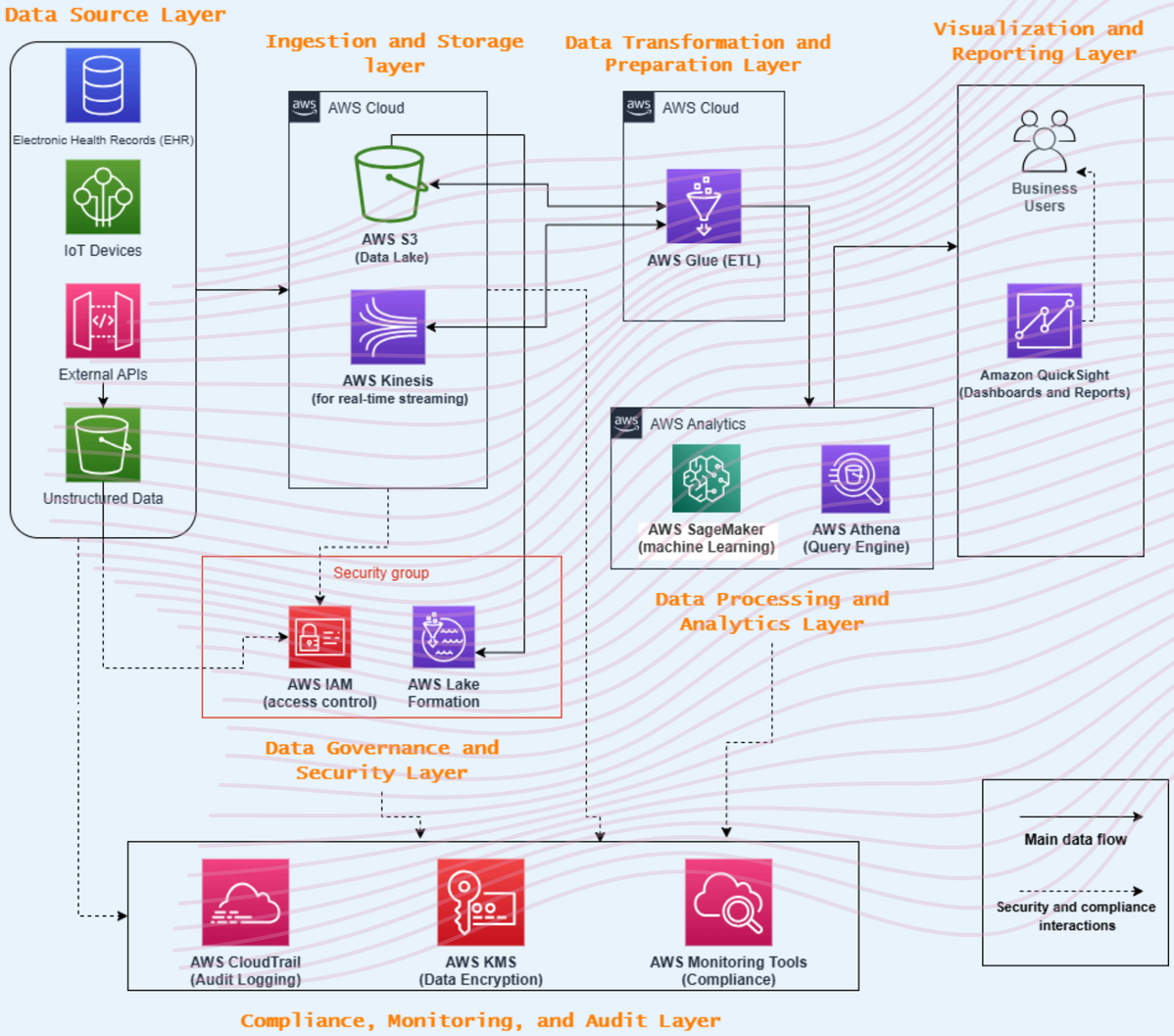
› Challenges

- High costs for maintaining legacy systems
- Inefficient integration across platforms
- Performance issues with slow, outdated systems
- Lack of real-time processing and predictive analytics capabilities

› Goals and Objectives

- Create a unified data platform capable of handling both OLTP and OLAP workloads
- Develop a system that can efficiently manage structured and unstructured data
- Improve patient care through enhanced data-driven decision-making.
- Comply with the set regulations of HIPAA and NIST standards on cybersecurity

2. Proposed Solution: AWS Based Data Lakehouse



› High-Level Architecture:

1. **Data Sources**
2. **Ingestion and Storage**
3. **Data Governance and Security**
4. **Data Transformation and Preparation**
5. **Data Processing and Analytics**
6. **Visualization and Reporting**
7. **Compliance, Monitoring, and Audit**

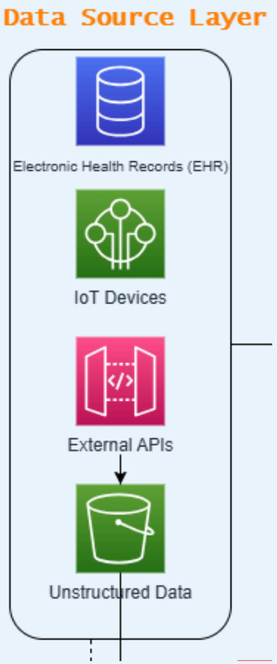
› Data Flow Overview:

1. Data ingested via Kinesis or batch processes is stored in S3.
2. AWS Glue processes and catalogs data.
3. Processed data is analyzed using Athena, SageMaker, or Comprehend Medical.
4. Insights are visualized in QuickSight.
5. Lake Formation and IAM ensure secure access controls.

2. Proposed Solution

01

Data Sources Layer



> Key Data Sources:

- EHRs: Patient data (demographics, diagnoses, treatments)
- IoT Devices: Real-time health data (e.g., heart rate, glucose levels)
- Medical Imaging: X-rays, MRIs, CT scans
- Clinical Notes: Physician notes and summaries
- External APIs: Public health data and third-party apps

> Data Ingestion:

- Real-Time: AWS Kinesis for IoT streams
- Batch: AWS S3 for central data storage
- Migration: AWS DataSync/DMS for legacy systems

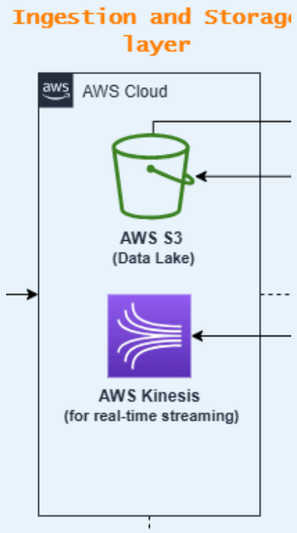
> Key Considerations for MeddGGenius:

- Data Volume & Velocity: Handle large-scale batch data and real-time IoT streams
- Data Variety: Support diverse formats (structured, unstructured, semi-structured)
- Data Veracity: Ensure accuracy with quality checks during ingestion
- Data Security & Privacy: Ensure HIPAA compliance with secure storage and access
- Interoperability: Enable seamless integration with internal systems and external partners

2. Proposed Solution

02

Ingestion and Storage Layer



> Key Components:

- AWS S3 (Data Lake):
  - Centralized, scalable storage for raw, structured, and unstructured data (e.g., JSON, CSV, images).
  - Cost-effective and ready for downstream processing.
- AWS Kinesis:
  - Real-time data streaming from IoT devices, APIs, and external systems.
  - Processes time-sensitive data (e.g., vitals, device logs) and stores it in S3.

> Data Flow:

- Real-Time Streaming: IoT devices and APIs send data to Kinesis, then to S3 (e.g., vitals, sensor logs).
- Batch Uploads: EHRs, medical images, and notes are uploaded to S3 and categorized into buckets.

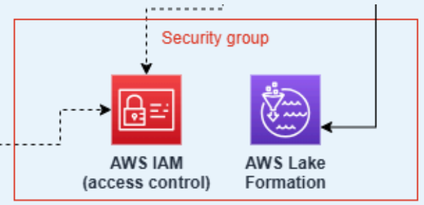
> Key Considerations for MeddGGenius:

- Scalability: Handle growing batch and real-time data seamlessly
- Diversity: Support structured, unstructured, and semi-structured data
- Quality: Ensure accuracy with automated validation
- Security: Maintain HIPAA compliance with encryption and access controls
- Integration: Enable smooth exchange with systems and APIs

2. Proposed Solution

03

Data Governance and Security Layer



> Key Components:

- AWS Lake Formation: Centralized catalog and fine-grained access controls (e.g., column/row-level).
- AWS IAM: Role-based access (RBAC) with MFA for added security.

> Data Flow

- Data in S3 is cataloged and secured by AWS Lake Formation.
- AWS IAM enforces role-based access, limiting data to authorized users only.

> Compliance:

- HIPAA: Secure handling of PHI to meet healthcare regulations
- NIST Framework: Aligns with risk management standards for data confidentiality and integrity

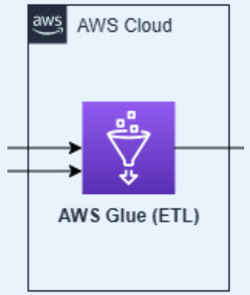
> Key Considerations for MeddGGenius:

- Access Control: Ensure only authorized personnel have access to sensitive data
- Auditability: Maintain logs for all data access and changes for compliance and monitoring
- Encryption: Protect data at rest and in transit with advanced encryption methods
- Scalability: Support growing governance needs as data volume increases
- User Training: Educate users on secure data practices to reduce human error

2. Proposed Solution

04

Data Transformation and Preparation Layer



> Key Components

- AWS Glue: Automates ETL processes, schema discovery, and data transformation.

> Flow of Operations

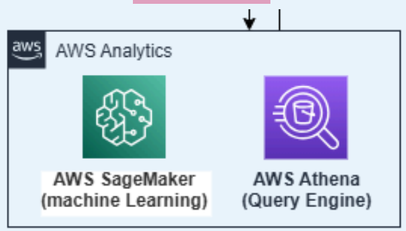
1. Data Extraction: AWS Glue retrieves raw data from S3, Kinesis, or external sources.
2. Data Transformation: Clean, deduplicate, and standardize data (e.g., map ICD-10 codes, format timestamps).
3. Data Loading: Transformed data is stored in S3 or passed to AWS SageMaker (ML) and Athena (analytics).

> Key Considerations for MeddGGenius:

- Scalability: Ensure the system scales with increasing data volumes from IoT devices, APIs, and batch uploads
- Data Quality: Maintain robust cleaning and validation processes for consistent, accurate data
- Compliance: Encrypt sensitive data and ensure adherence to HIPAA standards during transformations
- Flexibility: Support diverse data formats (structured, semi-structured, and unstructured) for compatibility with healthcare datasets
- Efficiency: Optimize transformation pipelines to minimize latency and resource usage

2. Proposed Solution

05



Data Processing and Analytics Layer

> Key Components

- AWS SageMaker: Build, train, and deploy ML models for predictive analytics
- AWS Athena: Perform SQL-based queries on S3 data for patient cohort analysis, population health management, and operational reporting.

> Use Cases

- Predictive Modeling: Detect early disease signs, optimize treatments, and forecast patient outcomes.
- SQL Querying: Identify patient cohorts, analyze trends, and report operational KPIs.

> Key Considerations for MeddGGenius

- Scalability: Ensure the system handles increasing data volumes and complex analytics workloads.
- Accuracy: Monitor and validate predictive models to ensure consistent performance over time.
- Compliance: Protect patient data with HIPAA-compliant tools and secure ML pipelines.
- Integration: Ensure seamless compatibility with other AWS services like QuickSight and Glue.
- Cost Management: Optimize usage of serverless tools to minimize costs without compromising performance.

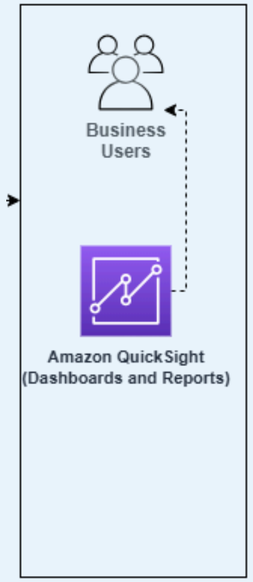
> Benefits

- Actionable Insights: Predict outcomes and identify trends for data-driven decisions.
- Efficiency: Process data cost-effectively with scalable, serverless tools.
- Improved Care: Enable personalized treatments and proactive interventions.

2. Proposed Solution

06

Visualization and Reporting Layer



> Key Component

- AWS QuickSight:
  - Interactive Dashboards: Real-time monitoring of metrics like vitals, readmission risks, and operations.
  - AWS Integration: Connects with S3, SageMaker, and Athena for seamless visualization.
  - Custom Visualizations: Supports healthcare-specific charts and IoT data integration for continuous insights.

> Flow of Operations

- Data Integration: Combine processed data from SageMaker (ML outputs), Athena (SQL queries), and S3 (structured data).
- Visualization: Use QuickSight to create real-time dashboards and static reports.
- Real-Time Updates: Optimize large-scale queries for live updates and timely decisions.

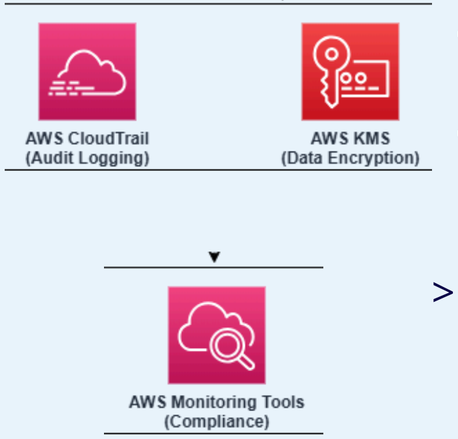
> Key Considerations for MeddGGenius

- User Accessibility: Ensure dashboards are intuitive and accessible to stakeholders with role-based access
- Real-Time Performance: Maintain low-latency updates for critical metrics like patient vitals and staffing
- Customization: Provide tailored dashboards to meet diverse healthcare and operational needs
- Data Security: Protect sensitive healthcare data with encryption and HIPAA-compliant access controls
- Scalability: Ensure the system scales to handle growing data sources and complex visualizations

2. Proposed Solution

07

Compliance, Monitoring, and Audit Layer



> Key Components

- AWS CloudTrail: Logs all API calls and system events, providing a detailed audit trail for HIPAA compliance.
- AWS Key Management Service (KMS): Encrypts data at rest and in transit with fine-grained access controls to secure sensitive information.
- AWS CloudWatch: Monitors infrastructure and application performance, providing real-time alerts for unusual activity or potential breaches.

> Data Flow

1. Real-Time Logging: CloudTrail records all data interactions and system events.
2. Encryption: KMS applies robust encryption algorithms to secure data at rest and in transit.
3. Monitoring: CloudWatch analyzes access patterns and system behaviors, sending alerts for potential threats.

> Key Considerations for MeddGGenius

- Regulatory Adherence: Ensure full compliance with HIPAA and NIST standards
- Data Transparency: Maintain a detailed audit trail for accountability
- Access Control: Use fine-grained encryption key management to secure sensitive data
- Proactive Security: Leverage real-time monitoring and alerting tools for threat detection
- Scalability: Ensure monitoring tools can handle increased data interactions seamlessly

3. Roadmap for Execution

› Data Migration and Management Strategy

Objective: Ensure smooth, secure, and efficient migration of healthcare data to the new system.
Steps:

- Assessment Phase: Identify existing data quality issues and gaps.
- ETL Pipelines: Extract, Transform, and Load (ETL) data incrementally to reduce risks.
- Validation & Governance: Perform rigorous validation and enforce governance to maintain data integrity and security.

› Data Retention Policy

Objective: Align with compliance standards like HIPAA while optimizing costs.
Steps:

- Define retention periods for datasets based on their use and legal requirements.
- Leverage AWS S3 storage tiers:
  - Active data → Standard Tier
  - Less frequently accessed data → Infrequent Access Tier
  - Long-term archival → Glacier Tier
- Establish automated secure disposal processes to safeguard patient privacy.

› Progress Tracking and Feedback Loops

Objective: Ensure continuous improvement and alignment with goals.
Steps:

- Regularly review performance metrics (e.g., ETL speed, data quality).
- Gather feedback from stakeholders at defined intervals.
- Adapt processes based on lessons learned to refine future phases.

4. Pros and Cons of Our Approach

› Pros

- Unified Data Management: Consolidates diverse healthcare data into a centralized repository, streamlining data management and improving accessibility.
- Improved Patient Care: Leverages advanced analytics and machine learning to uncover valuable insights for better diagnoses, personalized treatments, and improved patient outcomes.
- Enhanced Operational Efficiency: Automates tasks, reduces processing time, and improves the efficiency of healthcare operations, allowing for better resource allocation.
- HIPAA Compliance and Security: Ensures compliance with HIPAA regulations and meets NIST cybersecurity standards to protect patient data.
- Scalability and Flexibility: Offers a scalable and flexible platform to support future growth and innovation in healthcare delivery.
- Advanced Analytics: Facilitates predictive analytics and machine learning capabilities for proactive identification of patient needs and improved decision-making.

› Cons

- Cost: Implementing and maintaining the solution can be expensive.
- Complexity: The architecture can be complex, requiring specialized skills.
- Data Migration: Migrating data from legacy systems can be time-consuming and resource-intensive.
- Data Security: Despite security measures, there is always a risk of data breaches.

5. Challenges in Our Approach

› Challenges

- Data Quality
  - Healthcare data comes from various sources and can have inconsistencies, errors, and missing values.
  - Stringent data quality checks and validation rules are needed throughout the data lifecycle.
- Legacy System Integration
  - Integrating with existing RDBMS and Teradata systems can be complex.
  - Careful planning and appropriate tools are needed for a smooth transition.
- Real-Time Data Management
  - Efficiently managing high volume and velocity of real-time data requires scalable infrastructure.
  - Seamless integration and optimization of AWS Kinesis are crucial.
- Technology Updates
  - Staying current with new technologies and trends in data analytics and machine learning is essential.
  - Continuous learning and adaptation are key to remaining effective and innovative.
- Data Security
  - The risk of data breaches requires vigilance and proactive security measures.
  - Compliance with HIPAA and robust security protocols are essential.

6. Machine Learning Use Cases

› Case 1 - Medical Imaging and Analysis

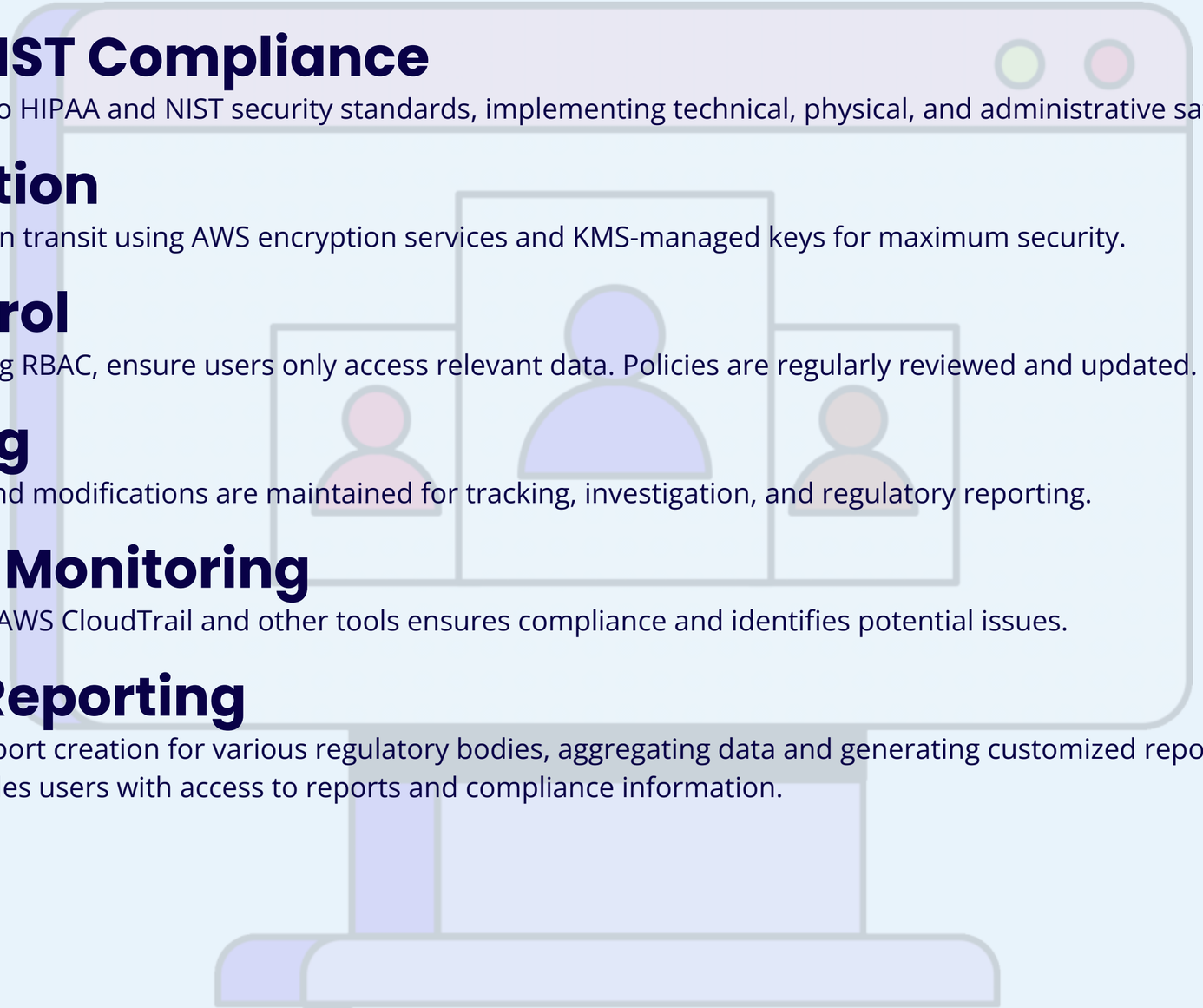
- Deep learning models and neural networks to analyze imaging data and detect signs of conditions based on X-ray vision.
- Diagnose patients using imaging data, assisting in preventive care.
- Neural networks and deep learning capabilities in AWS SageMaker can be pre-tarained and detect conditions with high accuracy.
- Computer viision is trained on large datasets of medical images to process and analyze x-ray scans and detect conditions.
- AWS SageMaker would be used train these models and AWS Athena can provide aggregated reports ti track accuracy of diagnosis or critical conditions.

› Case 2 - Preventive Care and Tracking Vitals

- Train

› Case 3 -

7. Compliance Monitoring and Reporting



› HIPAA and NIST Compliance

Our data lakehouse adheres to HIPAA and NIST security standards, implementing technical, physical, and administrative safeguards to protect PHI.

› Data Encryption

Data is encrypted at rest and in transit using AWS encryption services and KMS-managed keys for maximum security.

› Access Control

Strict access controls, including RBAC, ensure users only access relevant data. Policies are regularly reviewed and updated.

› Audit Logging

Detailed logs of data access and modifications are maintained for tracking, investigation, and regulatory reporting.

› Compliance Monitoring

Continuous monitoring using AWS CloudTrail and other tools ensures compliance and identifies potential issues.

› Regulatory Reporting

- The platform facilitates report creation for various regulatory bodies, aggregating data and generating customized reports.
- A self-service portal provides users with access to reports and compliance information.

References

1. Amazon Web Services. (n.d.). Amazon QuickSight. Retrieved November 27, 2024. https://aws.amazon.com/quicksight/?amazon-quicksight-whats-new.sort-by=item.additionalFields.postDateTime&amazon-quicksight-whats-new.sort-order=desc
2. Amazon Web Services. (n.d.). Population health applications with Amazon HealthLake, Part 1: Analytics and monitoring using Amazon QuickSight. Retrieved December 2, 2024. https://aws.amazon.com/blogs/machine-learning/population-health-applications-with-amazon-healthlake-part-1-analytics-and-monitoring-using-amazon-quicksight/
3. AWS Architecture Blog: AWS Architecture Blog (amazon.com)
4. Azure Architecture Blog. (n.d.). Azure Architecture Blog – Microsoft Community Hub. Retrieved November 21, 2024. https://techcommunity.microsoft.com/category/azure/blog/azurearchitectureblog
5. High Scalability. (n.d.). Real-world architectures: Examples like WhatsApp and Netflix. Retrieved November 25, 2024. http://highscalability.com/
6. Microsoft. (n.d.). Analytics end-to-end with Azure Synapse. Retrieved December 2, 2024. https://learn.microsoft.com/en-us/azure/architecture/example-scenario/dataplate2e/data-platform-end-to-end
7. Turck, M. (2021). Trends and technology landscape for data, AI, and ML architectures. Retrieved December 2, 2024. https://mattturck.com/data2021/

감사합니다 감사합니다 감사합니다

ధన్యవాదాలు ధన్యవాదాలు ధన్యవాదాలు

धन्यवाद धन्वाद धन्यवाद

நன்றி நன்றி நன்றி

THANK YOU

From Adithya, Eunsung, Kartik, Tanmay, Jurgen, Prasanth, and Richard

谢谢 谢谢 谢谢

આભાર આભાર આભાર

ಧನ್ಯವಾದಗಳು ಧನ್ಯವಾದಗಳು ಧನ್ಯವಾದಗಳು