**HealthData Insights Platform - System Design Document**

**Executive Summary**

**Project**: HealthData Insights  
**Version**: 2.0  
**Industry**: Healthcare Technology  
**Purpose**: Cloud-based platform for medical researchers to analyze anonymized patient data while maintaining HIPAA compliance  
**Users**: 5,000 researchers across 50 institutions  
**Data Volume**: 10TB patient records, 500GB daily processing  
**Criticality**: High - Contains sensitive health information

**Business Context**

**Overview**

HealthData Insights provides a secure platform for medical researchers to upload, analyze, and visualize anonymized patient datasets. The platform processes sensitive health information including diagnoses, treatments, and outcomes while ensuring patient privacy.

**Key Business Requirements**

* **Availability**: 99.9% uptime SLA (8.76 hours downtime/year)
* **Performance**: Analysis results within 5 minutes for datasets up to 1GB
* **Concurrent Users**: Support 500 concurrent researchers
* **Data Retention**: 7 years for audit logs, 2 years for analysis results
* **Recovery Objectives**: RTO: 1 hour, RPO: 15 minutes

**Financial Impact**

* **Revenue**: $50M annual from subscriptions
* **Downtime Cost**: $50,000 per hour
* **Breach Cost**: $500 per record (average 100,000 records at risk)

**Compliance & Regulatory Requirements**

**Healthcare Regulations**

* **HIPAA**: Full compliance required
  + PHI encryption at rest and in transit
  + Access controls and audit logging
  + Business Associate Agreements (BAAs) with all partners
* **HITECH Act**: Breach notification within 72 hours
* **State Laws**: California, New York, Texas privacy laws
* **GDPR**: For EU researcher access

**Industry Standards**

* **NIST 800-66**: HIPAA Security Rule compliance
* **ISO 27001**: Information security certification (in progress)
* **SOC 2 Type II**: Annual audit required

**Data Classification**

* **PHI (Protected Health Information)**: Patient records, even anonymized
* **PII (Personally Identifiable Information)**: Researcher profiles
* **Confidential**: Analysis algorithms, research results
* **Public**: Published research findings

**Architecture Overview**

**Deployment Model**

* **Primary**: AWS US-East-1 (Virginia)
* **DR Site**: AWS US-West-2 (Oregon)
* **Data Residency**: US-only, no data leaves country
* **Multi-AZ**: Deployment across 3 availability zones

**High-Level Architecture**

Internet → CloudFront → ALB → ECS Cluster → Internal Services → Data Stores

↓ ↓

WAF Protection Lambda Functions

**Detailed Components**

**External Entities**

**E1: Medical Researcher**

* **Description**: Primary platform users uploading and analyzing data
* **Access Level**: Authenticated, organization-scoped
* **Typical Actions**: Upload datasets, run analyses, download results
* **Security**: MFA required, session timeout 30 minutes
* **Risk Level**: Medium - potential data exfiltration

**E2: Organizational Admin**

* **Description**: Manages researcher accounts for their institution
* **Access Level**: Elevated privileges within organization
* **Typical Actions**: Create/disable accounts, view audit logs, set data policies
* **Security**: MFA + hardware token, privileged access management
* **Risk Level**: High - can manage multiple user accounts

**E3: Platform Administrator**

* **Description**: HealthData Insights employees managing platform
* **Access Level**: Full system access
* **Typical Actions**: System maintenance, security updates, incident response
* **Security**: MFA + VPN + bastion host + approval workflow
* **Risk Level**: Critical - unrestricted access

**E4: External Identity Provider (Okta)**

* **Description**: SAML-based authentication service
* **Integration**: OAuth 2.0 / SAML 2.0
* **Data Shared**: User identity assertions, group memberships
* **Security**: Encrypted assertions, certificate pinning
* **Risk Level**: High - authentication bypass if compromised

**E5: Malicious Actor**

* **Motivation**: Data theft, ransom, competitive advantage
* **Capabilities**: Advanced persistent threat, insider threat
* **Target Data**: PHI records, research algorithms
* **Attack Vectors**: Web vulnerabilities, credential theft, supply chain

**Data Stores**

**DS1: Patient Database (Amazon RDS - PostgreSQL)**

* **Purpose**: Stores anonymized patient records
* **Data Types**: Medical history, diagnoses, treatments, outcomes
* **Encryption**: AES-256 at rest (AWS KMS), TLS 1.3 in transit
* **Access Control**: IAM roles, database-level permissions
* **Backup**: Automated daily, point-in-time recovery enabled
* **Size**: 10TB, growing 100GB/month
* **Criticality**: Critical - contains all PHI

**DS2: Research Results Store (Amazon S3)**

* **Purpose**: Stores analysis outputs and visualizations
* **Bucket Structure**:
  + results-private/: Researcher-specific results
  + results-shared/: Published findings
  + results-temp/: Processing intermediates (24hr lifecycle)
* **Encryption**: SSE-KMS with customer managed keys
* **Access Control**: Bucket policies, pre-signed URLs for downloads
* **Versioning**: Enabled with MFA delete protection
* **Criticality**: High - contains research IP

**DS3: User Session Store (Amazon ElastiCache Redis)**

* **Purpose**: Active user sessions and temporary tokens
* **Data Types**: Session IDs, JWT tokens, user preferences
* **Encryption**: In-transit only (TLS)
* **Access Control**: Security group restrictions
* **Persistence**: In-memory only, 30-minute TTL
* **Criticality**: Medium - temporary data only

**DS4: Audit Log Store (Amazon CloudWatch + S3)**

* **Purpose**: Comprehensive audit trail for compliance
* **Data Types**: All API calls, data access, authentication events
* **Retention**: 7 years in S3 Glacier
* **Encryption**: SSE-S3
* **Access Control**: Read-only except for compliance team
* **Immutability**: Object lock enabled
* **Criticality**: High - required for compliance

**DS5: Application Secrets (AWS Secrets Manager)**

* **Purpose**: Database credentials, API keys, certificates
* **Rotation**: Automatic 90-day rotation
* **Access Control**: IAM roles, least privilege
* **Audit**: All access logged to CloudTrail
* **Criticality**: Critical - system-wide impact if compromised

**Processing Components**

**P1: API Gateway (Amazon API Gateway)**

* **Purpose**: RESTful API endpoint management
* **Features**: Rate limiting, API key management, request validation
* **Authentication**: JWT validation
* **Logging**: All requests logged to CloudWatch
* **Security**: Input validation, SQL injection prevention
* **Criticality**: High - main entry point

**P2: Web Application (ECS Fargate - Node.js)**

* **Purpose**: Main application logic and UI serving
* **Framework**: Express.js with security headers
* **Scaling**: Auto-scaling 2-20 instances
* **Security**:
  + OWASP dependency scanning
  + Container scanning in CI/CD
  + Read-only root filesystem
* **Criticality**: Critical - core application

**P3: Authentication Service (AWS Cognito + Custom Lambda)**

* **Purpose**: User authentication and authorization
* **Features**:
  + SAML integration with Okta
  + MFA enforcement
  + Password policy (14 chars, complexity)
  + Account lockout (5 attempts)
* **Token Management**: JWT with 15-minute expiry
* **Criticality**: Critical - auth bypass = total compromise

**P4: Data Processing Pipeline (AWS Lambda)**

* **Purpose**: Anonymization and analysis jobs
* **Runtime**: Python 3.9 with pandas, numpy
* **Triggers**: SQS messages, S3 events
* **Security**:
  + No internet access (VPC Lambda)
  + Temporary credentials only
  + 5-minute max execution
* **Criticality**: High - processes PHI

**P5: Notification Service (Amazon SNS + SES)**

* **Purpose**: Email notifications for job completion
* **Security**: No PHI in notifications
* **Rate Limiting**: 100 emails/user/day
* **Criticality**: Low - convenience feature

**P6: Admin Portal (Separate ECS Service)**

* **Purpose**: Platform administration interface
* **Access**: VPN + bastion host only
* **Features**: User management, system monitoring, incident response
* **Audit**: Every action logged with video recording
* **Security**: Separate VPC, hardware token required
* **Criticality**: Critical - admin access

**Trust Boundaries**

**TB1: Internet → AWS Edge (CloudFront/WAF)**

* **Controls**:
  + AWS WAF with OWASP Core Rule Set
  + DDoS protection (AWS Shield Advanced)
  + Geographic restrictions (US/EU only)
  + Rate limiting: 1000 req/min per IP

**TB2: AWS Edge → VPC Public Subnet**

* **Controls**:
  + Security groups (port 443 only)
  + Network ACLs
  + ALB with SSL termination
  + Certificate: EV SSL, HSTS enabled

**TB3: Public Subnet → Private Subnet**

* **Controls**:
  + No direct internet route
  + NAT Gateway for outbound only
  + Security groups with least privilege
  + VPC Flow Logs enabled

**TB4: Application Layer → Data Layer**

* **Controls**:
  + Database security groups (app-specific)
  + IAM database authentication
  + SSL/TLS for all connections
  + Connection pooling with limits

**TB5: AWS Account → External Services**

* **Controls**:
  + VPC endpoints for AWS services
  + Private link for Okta
  + Egress filtering
  + API rate limiting

**Data Flows**

**Flow 1: Researcher Login**

* **Path**: Researcher Browser → CloudFront → ALB → Auth Service → Okta → Auth Service → Session Store
* **Data**: Username, password, MFA code, SAML assertion
* **Classification**: Confidential (credentials)
* **Protocol**: HTTPS (TLS 1.3) throughout
* **Authentication**: Certificate pinning, SAML signatures
* **Rate Limit**: 5 attempts per 15 minutes
* **Logging**: All attempts logged

**Flow 2: Dataset Upload**

* **Path**: Researcher Browser → CloudFront → ALB → Web App → S3 (via presigned URL)
* **Data**: Anonymized patient datasets (CSV/JSON)
* **Classification**: PHI - Critical
* **Size Limit**: 1GB per file
* **Validation**:
  + File type checking
  + Virus scanning (ClamAV)
  + Schema validation
  + PHI detection scan
* **Encryption**: Client-side encryption option available

**Flow 3: Analysis Execution**

* **Path**: Web App → SQS → Lambda → Patient Database → Lambda → Results S3
* **Data**: Analysis parameters, patient queries, statistical results
* **Classification**: PHI/Confidential
* **Processing**: Isolated Lambda environment
* **Query Limits**: 100K records max, 5-minute timeout
* **Audit**: All queries logged with researcher ID

**Flow 4: Results Retrieval**

* **Path**: Web App → Results S3 → CloudFront → Researcher Browser
* **Data**: Analysis results, visualizations
* **Classification**: Confidential research data
* **Access Control**: Pre-signed URLs, 1-hour expiry
* **Download Limit**: 10GB/day per researcher
* **Watermarking**: PDF results include researcher ID

**Flow 5: Admin User Management**

* **Path**: Admin Browser → VPN → Bastion → Admin Portal → Auth Service → Audit Logs
* **Data**: User CRUD operations, permission changes
* **Classification**: Confidential
* **Approval**: Two-admin approval for deletions
* **Audit**: Video session recording
* **Rollback**: All changes reversible for 30 days

**Flow 6: Database Replication**

* **Path**: Primary RDS → AWS Backbone → DR RDS Instance
* **Data**: Full patient database
* **Classification**: PHI - Critical
* **Encryption**: TLS + AWS backbone encryption
* **Frequency**: Synchronous replication
* **Monitoring**: 1-minute lag alerting

**Flow 7: Audit Log Streaming**

* **Path**: All Services → CloudWatch → Kinesis Firehose → S3 Glacier
* **Data**: API calls, data access, errors
* **Classification**: Confidential
* **Retention**: 7 years
* **Immutability**: Object lock enabled
* **Access**: Break-glass procedure required

**Flow 8: External Identity Sync**

* **Path**: Okta → Private Link → Auth Service → User Database
* **Data**: User attributes, group memberships
* **Classification**: PII
* **Frequency**: Real-time SCIM provisioning
* **Validation**: Schema enforcement
* **Error Handling**: Dead letter queue

**Security Controls**

**Preventive Controls**

* **Network Security**:
  + VPC with public/private/database subnets
  + Security groups with least privilege
  + NACLs as defense in depth
  + VPC endpoints for AWS services
* **Identity & Access**:
  + MFA mandatory for all users
  + IAM roles for service accounts
  + Temporary credentials only (STS)
  + Regular access reviews
* **Data Protection**:
  + Encryption at rest (AES-256)
  + Encryption in transit (TLS 1.3)
  + Key rotation every 90 days
  + Data masking in non-prod

**Detective Controls**

* **Monitoring**:
  + CloudWatch metrics and alarms
  + VPC Flow Logs
  + AWS GuardDuty
  + Custom SIEM integration
* **Alerting Thresholds**:
  + Failed logins: 5 in 15 minutes
  + Large downloads: >1GB
  + Unusual API patterns
  + Geographic anomalies

**Responsive Controls**

* **Incident Response**:
  + 24/7 SOC monitoring
  + 15-minute response SLA
  + Automated containment for common threats
  + Forensic tooling pre-installed

**Known Vulnerabilities & Risks**

**Technical Debt**

1. **Legacy Lambda Runtime**: Some functions still on Python 3.7 (EOL)
2. **Hardcoded Configs**: 3 services have embedded credentials
3. **Missing CORS Headers**: Admin portal lacks proper CORS config
4. **Outdated Dependencies**: 15 npm packages need updates

**Architectural Risks**

1. **Single Point of Failure**: Auth service has no redundancy
2. **Shared Database**: Research and operational data mixed
3. **Insufficient Segmentation**: Same VPC for all environments
4. **Key Management**: Some services use shared KMS keys

**Operational Risks**

1. **Backup Testing**: DR never fully tested
2. **Privileged Access**: 10 admins have production access
3. **Third-Party Risk**: No security assessment of Okta
4. **Logging Gaps**: Lambda functions lack detailed logging

**Previous Security Incidents**

**Incident 1: SQL Injection Attempt (2024 Q1)**

* **Vector**: Malformed dataset filename
* **Impact**: None - WAF blocked
* **Remediation**: Added input validation

**Incident 2: Credential Stuffing (2024 Q3)**

* **Vector**: 50K login attempts from botnet
* **Impact**: 2 accounts compromised (no MFA)
* **Remediation**: MFA mandatory, rate limiting

**Incident 3: S3 Bucket Misconfiguration (2024 Q4)**

* **Vector**: Public read on results bucket
* **Impact**: 100 files exposed for 2 hours
* **Remediation**: Bucket policies, AWS Config rules

**Threat Landscape**

**Threat Actors**

1. **Cybercriminals**: Ransomware, data theft for sale
2. **Nation States**: Medical research espionage
3. **Competitors**: Industrial espionage
4. **Insiders**: Disgruntled employees, curious researchers

**Attack Vectors**

1. **Web Application**: OWASP Top 10 vulnerabilities
2. **Supply Chain**: Compromised dependencies
3. **Social Engineering**: Phishing for researcher credentials
4. **Insider Threat**: Authorized user data exfiltration
5. **API Abuse**: Automated data scraping

**Industry Threats**

* 300% increase in healthcare breaches (2024)
* Average healthcare breach cost: $10.1M
* Ransomware targeting research institutions
* Nation-state interest in medical research

**Recommendations & Roadmap**

**Immediate (Q1 2025)**

1. Update all Lambda functions to Python 3.11
2. Implement auth service redundancy
3. Remove hardcoded credentials
4. Enable AWS Security Hub

**Short-term (Q2 2025)**

1. Separate VPCs for environments
2. Implement database activity monitoring
3. Deploy runtime application security
4. Conduct penetration testing

**Long-term (Q3-Q4 2025)**

1. Zero-trust network architecture
2. Homomorphic encryption for analysis
3. Blockchain audit trail
4. AI-powered threat detection

**Appendices**

**A. Network Diagram**

[Detailed VPC architecture diagram would be here]

**B. Data Classification Matrix**

| **Data Type** | **Classification** | **Encryption** | **Retention** | **Access** |
| --- | --- | --- | --- | --- |
| Patient Records | PHI-Critical | AES-256 | 7 years | Need-to-know |
| Research Results | Confidential | AES-256 | 2 years | Researcher-owned |
| System Logs | Confidential | AES-256 | 7 years | Security team |
| User Sessions | Internal | TLS only | 30 min | System only |

**C. Compliance Mapping**

| **Requirement** | **Implementation** | **Evidence** |
| --- | --- | --- |
| HIPAA Encryption | All PHI encrypted | AWS KMS logs |
| Access Controls | IAM + database ACLs | Access reports |
| Audit Logging | CloudTrail + custom | Immutable logs |
| Breach Response | IR playbooks | Drill reports |

**D. Key Security Metrics**

* Mean Time to Detect (MTTD): 15 minutes
* Mean Time to Respond (MTTR): 45 minutes
* Security Incidents YTD: 3
* Vulnerability Scan Coverage: 95%
* Patch Compliance: 87%