Audit Report For German Election 2025 Data Lakehouse Project

Audited Company	Nguyen Auditing Services
Reason for Audit	Deployment at client sites
Date	17/07/2025
Auditor	Khuong Nguyen
Approval: (signature of lead auditor)	K. Nguyen
Date of approval	20/07/2025

Table of Contents

1. Overview
2. Audit Outcome
    - 2.1 Audit Scope and Methodology
    - 2.2 Overall Compliance Status
    - 2.3 Critical Risk Assessment
    - 2.4 Key Security Vulnerabilities
    - 2.5 Regulatory Compliance Concerns
    - 2.6 Business Impact Assessment
    - 2.7 Strategic Recommendations Overview
3. Recommendations
    - 3.1 Phase 1: Immediate Critical Remediation (0-30 days)
    - 3.2 Phase 2: Security Framework Development (30-90 days)
    - 3.3 Phase 3: Governance and Compliance Implementation (60-120 days)
    - 3.4 Phase 4: Operational Security Enhancement (90+ days)
    - 3.5 Implementation Timeline and Resource Planning
    - 3.6 Success Metrics and Validation
    - 3.7 Long-term Security Maintenance

1. Overview

About the Company

The **German Election 2025 Data Lakehouse** project is designed to collect, process, and analyze survey data related to the 2025 German federal elections. Utilizing a combination of Docker, Apache Spark, Apache Airflow, and PostgreSQL, this project establishes an efficient Extract, Load, Transform (ELT) pipeline. The architecture ensures seamless data ingestion, transformation, and storage, facilitating comprehensive analysis and visualization of election survey results. For more details about the architecture, visit: https://nbkhuong.github.io/german-election-2025/.

---

2. Audit Outcome

This comprehensive security audit evaluated a personal data analytics project consisting of an on-premise lakehouse infrastructure processing German election survey data, a public-facing visualization website, and associated GitHub repository with CI/CD pipeline. The assessment was conducted against established security frameworks including ISO 27001, NIST 800-53, and GDPR compliance requirements.

2.1 Audit Scope and Methodology

The audit examined six primary categories of controls:

1. **On-Premise Lakehouse Infrastructure** - Docker-based architecture with Airflow, Spark, MinIO, and PostgreSQL components
2. **Public-Facing Website** - Static site hosted on GitHub Pages with Google Analytics integration
3. **GitHub Repository and CI/CD Pipeline** - Source code management and deployment processes
4. **Security Controls** - Authentication, logging, and incident response capabilities
5. **GDPR Compliance** - Data protection and privacy requirements
6. **Risk Management** - Governance framework and risk assessment processes

2.2 Overall Compliance Status

Out of 36 controls assessed, the project demonstrates the following compliance posture:

- **Compliant:** 7 controls (19.4%) - Primarily related to appropriate technology architecture choices, physical security for home deployment, secure HTTPS deployment, and proper handling of public election data
- **Partially Compliant:** 6 controls (16.7%) - Basic capabilities exist but lack comprehensive implementation (PostgreSQL logging, backup processes, incident logging, centralized log management, security code review)
- **Non-Compliant:** 23 controls (63.9%) - Significant gaps requiring immediate remediation
- **Not Applicable:** 6 controls (16.7%) - Controls not relevant due to single-user deployment and public data nature

2.3 Critical Risk Assessment

The security risk analysis identified three distinct risk tiers based on likelihood and impact assessments:

High Risk (Score 16) - 3 Controls:

- PostgreSQL database access controls lacking role-based enforcement
- Absence of multi-factor authentication enforcement status for database access
- Missing MFA implementation across all critical systems (Airflow, MinIO, PostgreSQL)

Medium Risk (Score 9-12) - 8 Controls:

- Network segmentation gaps between public and private boundaries
- Firewall and endpoint protection deficiencies
- Data classification scheme not implemented for complex mixed datasets
- CI/CD pipeline security gaps including branch protection and automated scanning

- Risk management framework documentation missing

Low to Medium Risk (Score 4-8) - 12 Controls:

- GDPR compliance gaps in privacy policy and consent mechanisms
- Incident response procedures not formally documented
- Repository access management policies undefined

2.4 Key Security Vulnerabilities

Authentication and Access Control Failures:
The most critical vulnerabilities center around authentication and access management. Default credentials are used across PostgreSQL databases without custom role configuration, and multi-factor authentication is entirely absent from all critical system components. This creates immediate risk of unauthorized access and potential data compromise.

Network Security Deficiencies:
The infrastructure lacks proper network segmentation, with all services exposed on localhost ports without firewall rules configured in the docker-compose setup. This relies solely on host-level protection, creating potential attack vectors.

Governance and Compliance Gaps:
The project operates without formal risk management frameworks, documented incident response procedures, or comprehensive GDPR compliance measures despite processing public data and utilizing analytics tracking.

Development Security Practices:
The CI/CD pipeline lacks fundamental security controls including branch protection, automated dependency scanning, secrets scanning, and pull request requirements, allowing direct commits to the main branch without review.

2.5 Regulatory Compliance Concerns

While the project processes only public election survey data without personal information, several GDPR compliance gaps exist:

- No privacy policy published on the public website
- Absence of consent banner for Google Analytics tracking

- Lack of documented data processing activities

These gaps, while not immediately critical given the public nature of the data, represent regulatory compliance risks that could escalate if the project scope expands to include personal data processing.

2.6 Business Impact Assessment

Given the personal project nature, the immediate business impact is limited. However, the identified vulnerabilities could result in:

- **System Compromise:** Unauthorized access through weak authentication controls
- **Data Integrity Issues:** Potential manipulation of analytics data and visualizations
- **Regulatory Violations:** Non-compliance with GDPR requirements for web analytics
- **Reputation Risk:** Security incidents affecting the public-facing website
- **Development Risks:** Insecure CI/CD practices potentially introducing vulnerabilities

2.7 Strategic Recommendations Overview

The audit findings necessitate a phased remediation approach prioritizing immediate security control implementation, followed by governance framework establishment and long-term compliance maintenance. The recommended approach focuses on:

1. **Immediate Security Hardening** - Authentication, access controls, and network security
2. **Governance Framework Development** - Risk management, incident response, and documentation
3. **Compliance Implementation** - GDPR requirements and privacy controls
4. **Sustainable Security Practices** - Automated scanning, monitoring, and regular assessments

---

3. Recommendations

3.1 Phase 1: Immediate Critical Remediation (0-30 days)

Priority 1 - Authentication and Access Control Implementation

Control: PostgreSQL Database Access Controls

- **Action Required:** Eliminate all default credentials immediately and implement custom role-based access control
- **Implementation Steps:**
  - Create dedicated service accounts for each application component (Airflow, Spark, custom services)
  - Configure PostgreSQL roles with principle of least privilege access
  - Update docker-compose environment variables with strong, unique passwords
  - Document all credential changes in secure password management system
- **Success Criteria:** No default credentials remain in use; each service operates with minimum required database permissions
- **Resource Requirements:** 8-12 hours of configuration time
- **Risk Mitigation:** Eliminates highest risk vulnerability (Score 16)

Control: Multi-Factor Authentication Implementation

- **Action Required:** Deploy MFA across all critical system components
- **Implementation Steps:**
  - Research and select appropriate MFA solution compatible with Docker deployment
  - Configure MFA for Airflow web interface administrative access
  - Implement authentication tokens or certificates for MinIO object storage access
  - Establish MFA for PostgreSQL administrative connections
  - Document MFA procedures and recovery processes

- **Success Criteria:** All administrative access requires multi-factor authentication
- **Resource Requirements:** 16-20 hours including solution selection and implementation
- **Risk Mitigation:** Addresses two high-risk vulnerabilities (combined Score 32)

Priority 2 - Network Security Hardening

Control: Network Segmentation and Firewall Configuration

- **Action Required:** Implement proper network boundaries and firewall protections
- **Implementation Steps:**
  - Configure docker-compose network isolation between service tiers
  - Implement firewall rules restricting port access (8080, 9090, 9091) to necessary connections only
  - Create separate networks for web services, data processing, and storage components
  - Document network architecture and access requirements
  - Consider implementing reverse proxy for external access management
- **Success Criteria:** Services operate in segmented networks with controlled access points
- **Resource Requirements:** 12-16 hours of network reconfiguration
- **Risk Mitigation:** Reduces medium-risk network vulnerabilities (combined Score 24)

3.2 Phase 2: Security Framework Development (30-90 days)

Priority 3 - CI/CD Security Enhancement

Control: GitHub Repository Security Implementation

- **Action Required:** Establish secure development and deployment practices
- **Implementation Steps:**

    - Enable GitHub branch protection rules requiring pull requests for main branch
    - Configure GitHub Actions workflows for automated dependency vulnerability scanning
    - Implement secrets scanning to detect credentials in code repositories
    - Set up automated security code analysis using GitHub security features
    - Establish pull request review requirements even for single-contributor projects
    - Create security-focused GitHub Actions workflow for deployment validation
- **Success Criteria:** All code changes require security scanning and review processes
- **Resource Requirements:** 10-14 hours of GitHub configuration and workflow creation
- **Risk Mitigation:** Addresses multiple medium-risk CI/CD vulnerabilities (combined Score 45)

Control: Infrastructure Security Hardening

- **Action Required:** Apply security configurations to computing infrastructure
- **Implementation Steps:**
    - Research and apply Spark cluster security hardening guidelines
    - Configure authentication and authorization for Spark cluster components
    - Implement secure communication protocols between Spark nodes
    - Update default configurations for enhanced security posture
    - Document security configuration changes and maintenance procedures
- **Success Criteria:** Spark cluster operates with security-hardened configurations
- **Resource Requirements:** 8-10 hours of research and configuration
- **Risk Mitigation:** Reduces infrastructure risk (Score 9)

Priority 4 - Data Management and Classification

Control: Information Asset Classification Implementation

- **Action Required:** Develop formal data classification scheme
- **Implementation Steps:**
  - Analyze current data assets including public election survey data
  - Define classification levels appropriate for project scope (Public, Internal, Restricted)
  - Create data handling procedures for each classification level
  - Document data lineage and processing workflows
  - Implement data labeling within processing pipelines
  - Establish data retention and disposal procedures
- **Success Criteria:** All data assets classified with appropriate handling procedures
- **Resource Requirements:** 6-8 hours of analysis and documentation
- **Risk Mitigation:** Addresses data classification risk (Score 9)

3.3 Phase 3: Governance and Compliance Implementation (60-120 days)

Priority 5 - Risk Management Framework

Control: Formal Risk Management System

- **Action Required:** Establish comprehensive risk management capabilities
- **Implementation Steps:**
  - Create formal risk register documenting all identified risks including MFA, network security, and consent management
  - Develop risk assessment methodology with likelihood and impact scoring
  - Establish risk mitigation planning and timeline tracking procedures
  - Implement quarterly risk register review and update cycles

    - Create risk reporting and escalation procedures
    - Document risk management policy and procedures
- **Success Criteria:** Active risk register with documented mitigation plans and regular review cycles
- **Resource Requirements:** 12-16 hours of framework development and documentation
- **Risk Mitigation:** Addresses all risk management gaps (combined Score 30)

Control: Incident Response Capability Development

- **Action Required:** Create formal incident response procedures
- **Implementation Steps:**
    - Develop incident response plan covering detection, analysis, containment, eradication, and recovery
    - Define incident classification criteria and response procedures
    - Establish incident logging and documentation requirements
    - Create communication procedures for security incidents
    - Implement annual incident response plan review and testing procedures
    - Document roles and responsibilities for incident response
- **Success Criteria:** Documented incident response plan with annual review process
- **Resource Requirements:** 8-12 hours of procedure development
- **Risk Mitigation:** Reduces incident response risks (combined Score 16)

Priority 6 - GDPR Compliance Implementation

Control: Privacy Policy and Consent Management

- **Action Required:** Implement comprehensive privacy compliance measures
- **Implementation Steps:**
    - Create detailed privacy policy covering data collection, processing, and analytics practices

    - Implement cookie consent banner on static website for Google Analytics
    - Document data processing activities including legal basis for processing
    - Establish procedures for handling data subject inquiries (though not applicable for public data)
    - Publish privacy policy prominently on website with easy access
    - Configure Google Analytics for privacy-compliant data collection
- **Success Criteria:** Website displays privacy policy and consent mechanisms for all analytics tracking
- **Resource Requirements:** 6-10 hours of legal research, policy creation, and technical implementation
- **Risk Mitigation:** Addresses GDPR compliance gaps (combined Score 16)

3.4 Phase 4: Operational Security Enhancement (90+ days)

Priority 7 - Logging and Monitoring Implementation

Control: Enhanced Audit Logging and Monitoring

- **Action Required:** Improve security monitoring and audit capabilities
- **Implementation Steps:**
    - Configure comprehensive PostgreSQL audit logging for all database activities
    - Implement centralized log management system collecting logs from all Docker containers
    - Establish log retention policies and integrity protection measures
    - Create security monitoring dashboards and alerting for suspicious activities
    - Document log analysis procedures and security event investigation processes
    - Implement automated log analysis for security event detection
- **Success Criteria:** Centralized logging with security monitoring and automated alerting

- **Resource Requirements:** 16-20 hours of logging infrastructure setup and configuration
- **Risk Mitigation:** Enhances partially compliant controls and improves overall security visibility

Control: Backup and Disaster Recovery**

- **Action Required:** Formalize data protection and recovery capabilities
- **Implementation Steps:**
    - Document comprehensive backup procedures for all data assets and configurations
    - Implement automated backup scheduling for critical data and system configurations
    - Establish backup testing and validation procedures
    - Create disaster recovery procedures including system restoration processes
    - Implement backup integrity verification and off-site storage considerations
    - Document recovery time objectives and recovery point objectives
- **Success Criteria:** Formal backup and disaster recovery procedures with regular testing
- **Resource Requirements:** 10-14 hours of procedure development and automation setup
- **Risk Mitigation:** Addresses partially compliant backup processes

Priority 8 - Repository Access Management

Control: Formal Access Management Policies

- **Action Required:** Establish access control governance for development resources
- **Implementation Steps:**
    - Create formal access management policies for repository and system access

    - Document access provisioning and deprovisioning procedures
    - Establish access review procedures for potential future collaborators
    - Create access control documentation and audit procedures
    - Implement access logging and monitoring for administrative activities
- **Success Criteria:** Documented access management policies with audit capabilities
- **Resource Requirements:** 4-6 hours of policy development
- **Risk Mitigation:** Addresses low-risk repository management gaps (Score 4)

3.5 Implementation Timeline and Resource Planning

Total Implementation Effort: Approximately 140-200 hours across all phases

Recommended Timeline: 4-6 months for complete implementation

Critical Path: Authentication and network security controls must be completed before other enhancements

Phase 1 (Immediate - 30 days): 36-48 hours

- Authentication controls: 24-32 hours
- Network security: 12-16 hours

Phase 2 (30-90 days): 34-42 hours

- CI/CD security: 18-24 hours
- Infrastructure hardening: 8-10 hours
- Data classification: 6-8 hours

Phase 3 (60-120 days): 26-38 hours

- Risk management: 12-16 hours
- Incident response: 8-12 hours
- GDPR compliance: 6-10 hours

Phase 4 (90+ days): 30-44 hours

- Logging and monitoring: 16-20 hours
- Backup and recovery: 10-14 hours
- Access management: 4-6 hours

3.6 Success Metrics and Validation

Security Posture Improvement:

- Achieve 80%+ compliance rate across all assessed controls
- Eliminate all high-risk vulnerabilities (Score 16)
- Reduce medium-risk vulnerabilities by 75%
- Implement automated security monitoring and alerting

Operational Effectiveness:

- Establish documented procedures for all critical security processes
- Implement regular security review and update cycles
- Create maintainable security configuration management
- Enable rapid incident detection and response capabilities

Compliance Achievement:

- Full GDPR compliance for website analytics and data processing
- Alignment with ISO 27001 and NIST 800-53 security frameworks
- Documented risk management and governance processes
- Regular audit and assessment capabilities

3.7 Long-term Security Maintenance

Quarterly Activities:

- Risk register review and update
- Security configuration validation
- Access control review and validation
- Incident response plan testing

Annual Activities:

- Comprehensive security assessment
- Incident response plan review and update
- Security policy and procedure updates
- Penetration testing or security review

Ongoing Monitoring:

- Automated security scanning and alerting
- Log analysis and security event investigation
- Vulnerability management and patching
- Security awareness and training (if expanding team)

This comprehensive remediation plan addresses all identified non-compliant controls while establishing sustainable security practices for long-term maintenance and potential project expansion.