12 Data Security & Governance

Data Security & Governance: Access Control, Secrets Management, Encryption, Compliance & Unity Catalog

Basic Level Questions (1-10)

1. What are the different authentication methods available in Databricks? *Focus on: Azure AD integration, personal access tokens, service principals, SSO, MFA*

2. Explain the concept of Databricks workspaces and how they provide isolation. *Focus on: Workspace boundaries, resource isolation, user management, network isolation*

3. What is Azure Key Vault and how does it integrate with Databricks for secrets management? *Focus on: Secret storage, retrieval mechanisms, access control, rotation capabilities*

4. What are the basic principles of data encryption at rest and in transit in Databricks? *Focus on: Storage encryption, network encryption, key management, compliance requirements*

5. What is Unity Catalog and what problems does it solve in data governance? *Focus on: Centralized metadata, access control, data discovery, lineage tracking*

6. How do you create and manage secret scopes in Databricks? *Focus on: Scope creation, permission management, secret retrieval, best practices*

7. What are the different types of access control models in Databricks? *Focus on: Workspace-level, cluster-level, notebook-level, table-level access control*

8. What is the difference between Volumes and traditional DBFS storage in Unity Catalog? *Focus on: File system abstraction, access control, governance, performance*

9. How do you implement basic row-level security in Databricks? *Focus on: Filtering mechanisms, policy enforcement, user context, dynamic masking*

10. What are the key compliance frameworks that Databricks supports? *Focus on: SOC 2, GDPR, HIPAA, PCI DSS, regional compliance requirements*

Intermediate Level Questions (11-20)

11. How would you implement a comprehensive data classification system using Unity Catalog? *Focus on: Tagging strategies, automated classification, policy enforcement, metadata management*

12. Explain how to set up fine-grained access control for different user roles accessing the same dataset. *Focus on: Role-based access, attribute-based access, dynamic permissions, principle of least privilege*

13. How do you implement data masking and anonymization techniques in Databricks? *Focus on: Dynamic masking, static masking, tokenization, differential privacy*

14. What are the best practices for managing service principal authentication in production environments? *Focus on: Service principal lifecycle, credential rotation, scope limitation, monitoring*

15. How would you implement audit logging and compliance reporting for data access patterns? *Focus on: Audit events, log analysis, compliance dashboards, access pattern monitoring*

16. Explain how to implement cross-workspace data sharing while maintaining security controls. *Focus on: Delta Sharing, cross-workspace permissions, federation, governance boundaries*

17. How do you handle PII data discovery and protection in large datasets? *Focus on: Automated PII detection, data classification, protection mechanisms, compliance workflows*

18. What strategies would you use to implement data lineage tracking across multiple workspaces? *Focus on: Lineage capture, cross-system tracking, metadata correlation, visualization*

19. How do you implement secure data sharing with external partners while maintaining governance? *Focus on: External sharing mechanisms, access controls, monitoring, compliance verification*

20. Explain how to set up comprehensive monitoring for security events and policy violations. *Focus on: Security monitoring, anomaly detection, alert systems, incident response*

Advanced/Difficult Level Questions (21-30)

21. Design a comprehensive data governance framework for a multi-tenant Databricks environment with strict regulatory requirements. *Focus on: Tenant isolation, policy enforcement, compliance automation, audit capabilities*

22. How would you implement a zero-trust security model for Databricks in a hybrid cloud environment? *Focus on: Network security, identity verification, continuous monitoring, threat detection*

23. Design a solution for automated data classification and protection that scales across petabytes of data. *Focus on: ML-based classification, automated policy application, performance optimization, accuracy maintenance*

24. How would you implement end-to-end encryption for sensitive data processing workflows while maintaining query performance? *Focus on: Format-preserving encryption, searchable encryption, key management, performance optimization*

25. Design a comprehensive secrets management strategy that handles rotation, distribution, and emergency revocation at enterprise scale. *Focus on:*

Automated rotation, distribution mechanisms, emergency procedures, dependency management

26. How would you implement a data sovereignty solution that ensures data residency compliance across multiple regions? *Focus on: Data localization, cross-border restrictions, regulatory compliance, architecture design*

27. Design a privacy-preserving analytics solution that enables data collaboration while protecting individual privacy. *Focus on: Differential privacy, federated learning, secure multi-party computation, privacy budgets*

28. How would you implement a comprehensive data breach detection and response system? *Focus on: Anomaly detection, behavioral analysis, automated response, forensic capabilities*

29. Design a governance solution that automatically enforces data retention policies and implements right-to-be-forgotten capabilities. *Focus on: Automated deletion, compliance tracking, data lifecycle management, audit trails*

30. How would you implement a unified security and governance layer across multiple data platforms (Databricks, Synapse, Power BI)? *Focus on: Platform integration, unified policies, cross-platform monitoring, governance federation*

Compliance & Regulatory Scenarios

Real-World Compliance Challenges

Scenario A: Your organization needs to comply with GDPR while maintaining high-performance analytics. How would you architect data processing pipelines to ensure compliance?

Scenario B: You're implementing a healthcare analytics platform that must comply with HIPAA. What security controls and governance measures would you implement?

Scenario C: A data breach has been detected in your Databricks environment. Walk through your incident response procedure.

Scenario D: You need to implement data sharing between organizations in different countries with varying data sovereignty laws.

Advanced Security Patterns

Enterprise Security Architecture

Identity & Access Management:

- Federation with enterprise identity providers
- Conditional access policies implementation
- Multi-factor authentication enforcement
- Privileged access management

Data Protection:

- Column-level encryption strategies
- Homomorphic encryption for computation on encrypted data
- Secure enclaves and confidential computing
- Hardware security module integration

Network Security:

- Private endpoints and VNet integration
- Network traffic inspection and filtering
- Micro-segmentation strategies
- API gateway security patterns

Governance Implementation

Operational Governance

Policy Management:

- Policy as code implementation
- Automated policy testing and validation
- Policy version control and rollback
- Impact analysis for policy changes

Metadata Management:

- Business glossary implementation
- Data quality metrics and monitoring
- Schema evolution tracking
- Relationship mapping and impact analysis

Compliance Automation:

- Automated compliance checking
- Regulatory reporting automation
- Policy violation detection and remediation
- Compliance dashboard and metrics