

1.3.8 Lesson Review

Date: 2/26/2026, 7:32:17 PM

Time Spent: 03:15

Score: 100%

Passing Score: 80%



Question 1

 Correct

A security operations team deploys an AI system that watches for irregular activity in data pipelines. The system alerts the team when it sees unexpected changes in dataset composition or sudden shifts in how frequently data is ingested.

What is the primary benefit of applying behavioral analytics and continuous monitoring in this scenario?

It replaces the need for traditional logging, auditing, and access-control mechanisms for sensitive data repositories

It guarantees that every change to a dataset is blocked unless it is explicitly approved by a human reviewer

It automatically corrects corrupted records and rewrites them into a clean, verified format without human oversight

- It enables early detection of abnormal data patterns that may signal contamination or misuse of datasets  Correct

Explanation

Enabling early detection of abnormal data patterns that may signal contamination or misuse of datasets is the correct answer. Behavioral analytics and continuous monitoring reveal abnormal data patterns in real time, helping the team quickly spot possible contamination, misuse, or unauthorized changes.

Guaranteeing that every change to a dataset is blocked is incorrect. Monitoring can detect suspicious changes but does not inherently block all modifications.

Automatically correcting corrupted records and rewriting them into a clean, verified format is incorrect. These tools flag anomalies; they do not automatically repair or rewrite corrupted data.

Replacing the need for traditional logging, auditing, and access-control mechanisms for sensitive data repositories is incorrect. Monitoring complements, but does not replace logging, auditing, and access control.

Related Content

 1.3.5 Data Handling Techniques

resources\questions\q_data_handling_techniques_07.question.xml

Question 2

 Correct

Which description BEST distinguishes data lineage from data provenance in an AI data pipeline?

Data lineage records who has read access to a dataset, while data provenance records who has write access to the same dataset

Data lineage defines how long data must be retained, while data provenance defines how frequently data backups must be created and tested

Data lineage describes how data is encrypted at rest, while data provenance documents which key management system is used

- Data lineage tracks every transformation a dataset undergoes, while data provenance documents the dataset's origin, licensing, and consent terms  Correct

Explanation

Data lineage tracks every transformation a dataset undergoes, while data provenance documents the dataset's origin, licensing, and consent terms is the correct answer. Data lineage focuses on the "journey" of data, capturing how it moves and is transformed from source to final use, while data provenance records where the data came from, who created it, and under what legal or consent conditions it can be used.

Data lineage describes how data is encrypted at rest, while data provenance documents which key management system is used is incorrect. Encryption and key management are security controls that protect data confidentiality; they do not describe data lineage or provenance.

Data lineage records who has read access to a dataset, while data provenance records who has write access to the same dataset is incorrect. Read and write access are access-control concepts that govern permissions on a dataset, not its movement through systems or its origin and usage rights.

Data lineage defines how long data must be retained, while data provenance defines how frequently data backups must be created and tested is incorrect. Retention periods and backup schedules relate to storage and continuity planning, not data lineage or provenance.

Related Content

-  1.3.5 Data Handling Techniques
resources\questions\q_data_handling_techniques_03.question.xml

Question 3

 Correct

A threat intelligence analyst plans to share a large set of firewall logs and malicious hash values with an external research partner to improve joint detection capabilities.

The partner needs enough detail to build useful analytics, but the organization is concerned that the data might reveal internal defensive methods and operational patterns.

What is the MOST appropriate way for the analyst to prepare this structured data before sharing it?

Replace all malicious indicators with generic placeholders but leave timestamps, IP addresses, and detection fields unchanged.

Share the raw logs and hash lists as-is so the partner has full context and can perform the most accurate analysis possible.

- Sanitize the dataset to remove or aggregate sensitive operational details, while preserving the fields necessary for meaningful security analysis.  Correct

Strip timestamps from the logs but keep all detection rules, signatures, and internal classification fields in their original form.

Explanation

Sanitizing the dataset to remove or aggregate sensitive operational details, while preserving the fields necessary for meaningful security analysis, is the correct answer. Sanitizing the data removes sensitive operational details while keeping enough structure and content for effective joint analysis.

Sharing the raw logs and hash lists as-is is incorrect. This exposes internal tactics, detection coverage, and timing patterns that attackers could exploit.

Replacing all malicious indicators with generic placeholders is incorrect. Masking indicators but keeping detailed timestamps and network fields still reveals internal behavior and monitoring patterns.

Stripping timestamps from the logs is incorrect. Hiding timestamps but exposing detailed detection logic reveals how threats are identified and prioritized, helping attackers evade defenses.

Related Content

 1.3.2 Data Security Considerations for AI

resources\questions\q_data_security_considerations_for_ai_08.question.xml

Question 4

 Correct

Why does a retrieval-augmented generation system require strong protections around its vector storage?

It typically stores embeddings derived from internal content in a way that, if misconfigured, could allow cross-team queries to surface data beyond a user's authorization

- It centralizes large volumes of sensitive documents as embeddings that attackers could target  Correct

It may be replicated across multiple environments for performance, increasing the number of locations where sensitive representations are stored

It often becomes the main knowledge source backing answers, so unauthorized access could expose internal information at scale

Explanation

Centralizing large volumes of sensitive documents as embeddings that attackers could target is the correct answer. This is true. Even though the data is numeric, it still reflects confidential content, so compromise of this store can expose a large portion of an organization's knowledge.

Often becoming the main knowledge source backing answers is incorrect. Vector storage often underpins answers, but that is a consequence of its use in retrieval. The core security issue is the concentration of sensitive embeddings themselves, which requires strong protection.

It may be replicated across multiple environments for performance is incorrect. Replication can increase the attack surface, but this depends on deployment choices. Protections are always needed because the embeddings centralize sensitive information.

It typically stores embeddings derived from internal content in a way that, if misconfigured, could allow cross-team queries to surface data is incorrect. Misconfiguration is mainly an access-control problem. Strong protections are required because the index stores sensitive embeddings.

Related Content

[resources\questions\q_data_security_considerations_for_ai_06.question.xml](#)

Question 5

 Correct

How can sharing structured security data, such as firewall logs and IOC lists, unintentionally weaken an organization's defenses?

It can prevent analysts from correlating events across different time zones and regions

It can force all security tools to normalize to a single, less flexible log format

It can permanently erase older threat indicators from internal monitoring systems

- It can reveal operational patterns and defensive techniques that adversaries can study

 Correct

Explanation

It can reveal operational patterns and defensive techniques that adversaries can study is the correct answer. Structured data, if shared without sanitization, can expose when systems are monitored, how alerts are generated, and which techniques are closely watched, allowing attackers to tailor campaigns to avoid or bypass those defenses.

It can permanently erase older threat indicators from internal monitoring systems is incorrect. Sharing data with outside parties does not delete or overwrite internal indicators. Retention and removal are governed by internal log policies and storage settings.

It can force all security tools to normalize to a single, less flexible log format is incorrect. Systems can map between schemas as needed. The main risk lies in what the content discloses, not in forcing uniform log formats.

It can prevent analysts from correlating events across different time zones and regions is incorrect. Analysts can normalize timestamps across sources. The greater concern is that shared logs may reveal internal timing and behavior patterns.

Related Content

-  1.3.2 Data Security Considerations for AI
resources\questions\q_data_security_considerations_for_ai_02.question.xml

Question 6

 Correct

Which concept involves intentionally generating additional training examples, such as by rotating images or adding noise, so that a model learns to generalize rather than memorize?

Data Lineage

Data Cleansing

Data Balancing

- Data Augmentation ✓ Correct

Explanation

Data Augmentation is the correct answer. Data augmentation creates new training samples by transforming existing data so the model learns general patterns instead of memorizing specific examples.

Data Balancing is incorrect. Data balancing adjusts the proportion of classes, such as over-sampling rare events or under-sampling common ones, but it does not create transformed versions of existing records.

Data Cleansing is incorrect. Data cleansing fixes errors, removes duplicates, and resolves inconsistencies to improve data quality rather than expanding the dataset with modified copies.

Data Lineage is incorrect. Data lineage tracks how data moves and changes through systems for traceability, not to generate extra training examples.

Related Content

 1.3.5 Data Handling Techniques

resources\questions\q_data_handling_techniques_05.question.xml

Question 7

 Correct

Which security goal is most directly supported by verifying the integrity of data before it is used to train an AI model?

- Preserving the accuracy and trustworthiness of AI model outputs ✓ Correct

Reducing storage costs by compressing large training datasets

Ensuring confidentiality of user identities through anonymization

Improving the speed of AI inference in production environments

Explanation

Preserving the accuracy and trustworthiness of AI model outputs is the correct answer. Integrity checks ensure data has not been altered or poisoned before it reaches the model. This helps keep AI outputs accurate and trustworthy.

Ensuring confidentiality of user identities through anonymization is incorrect. Anonymization protects privacy and confidentiality, not integrity. It does not verify that data remains unchanged or unmanipulated.

Improving the speed of AI inference in production environments is incorrect. Inference speed is a performance concern, not an integrity goal. Verifying data integrity focuses on correctness and trust, not on how fast the model runs.

Reducing storage costs by compressing large training datasets is incorrect. Compression is unrelated to detecting unauthorized changes.

Related Content

[resources\questions\q_data_security_related_to_ai_01.question.xml](#)

Question 8

 Correct

A threat intelligence team is deploying an AI system that consumes firewall logs, JSON-based threat feeds, packet payloads, and analyst chat transcripts from multiple regions.

After the first pilot, they notice three problems:

- Some model outputs appear biased toward traffic from a single data center.
- A few alerts reference internal hostnames and employee names in ways never approved for use.
- A recent model update coincides with unusually high false negatives for lateral movement.

Which action BEST demonstrates an analysis of the situation that addresses the most likely root causes across the data pipeline?

Disable ingestion of unstructured data such as packet payloads and chat transcripts, relying on structured firewall logs and threat feeds to reduce noise and simplify analysis.

Increase model complexity and retrain with a deeper neural network so it can better learn subtle patterns from the combined structured and unstructured sources.

- Introduce data verification with cryptographic hashes at ingestion, track data lineage and provenance for each dataset, and implement data balancing to correct skew toward traffic from one data center.  Correct

Centralize all data sources into a single, high-performance data lake, enable full-text indexing on every field, and postpone integrity checks until after model training is complete.

Explanation

Introducing data verification with cryptographic hashes at ingestion, track data lineage and provenance for each dataset, and implementing data balancing to correct skew toward traffic from one data center is the correct answer. This option links each symptom to a data issue: verification and lineage help investigate tampering tied to the update, and data balancing addresses bias toward one data center while provenance supports control over sensitive content.

Increasing model complexity and retraining with a deeper neural network is incorrect. Making the model deeper ignores the underlying data integrity, governance, and imbalance problems, and may actually reinforce poisoned or skewed patterns.

Disabling ingestion of unstructured data, such as packet payloads and chat transcripts, is incorrect. Removing unstructured data avoids analyzing the real causes—lack of verification, lineage, and balancing—while throwing away useful context that could improve detection.

Centralizing all data sources into a single, high-performance data lake is incorrect. Centralizing and indexing data without early integrity checks allows poisoned or skewed data to shape the model before problems are detected, and does not address provenance or sensitive-data leakage.

Related Content

[resources\questions\q_data_security_related_to_ai_04.question.xml](#)

Question 9

 Correct

Which scenario BEST illustrates the purpose of data cleansing in an AI security pipeline?

Applying digital signatures and recording dataset hashes on a ledger so that any unauthorized modification can be detected later

Encrypting log files in transit between regional offices and the central analytics platform using strong transport protocols

Aggregating firewall logs from multiple offices into a central data lake for long-term storage and historical reporting

Removing duplicate records, filling in missing values, and

- correcting inconsistent log formats before training a detection model

 Correct

Explanation

Removing duplicate records, filling in missing values, and correcting inconsistent log formats before training a detection model is the correct answer. Data cleansing improves data quality by removing duplicates, fixing inconsistent formats, and handling missing values so the model trains on accurate, consistent information.

Aggregating firewall logs from multiple offices into a central data lake is incorrect. Centralizing logs is data aggregation which, helps collection and storage, but does not correct errors, gaps, or inconsistencies in the data.

Encrypting log files in transit between regional offices and the central analytics platform is incorrect. Encrypting logs protects data in transit, focusing on confidentiality.

Applying digital signatures and recording dataset hashes on a ledger is incorrect. Digital signatures and hashes verify that data has not changed, but they do not fix quality problems such as duplicates or malformed entries.

Related Content

 1.3.5 Data Handling Techniques

resources\questions\q_data_handling_techniques_01.question.xml

Question 10

 Correct

A security engineer is designing a new AI-driven intrusion detection system that ingests firewall logs, packet captures, and threat intelligence feeds from multiple regions in real time.

Leadership wants the system to quickly spot lateral movement while ensuring attackers cannot tamper with the training data as it flows through the environment.

Which approach BEST applies data security practices to protect the AI data pipeline end-to-end?

Rely on the AI model's anomaly detection to identify any tampered or malicious training data, without verification or encryption steps.

- Implement data verification at ingestion using cryptographic hashes, encrypt all data in transit and at rest, and restrict access to the pipeline with strong access controls and logging.  Correct

Use semi-structured formats like JSON for all logs and packet data, since a standardized format reduces the risk of data poisoning.

Focus primarily on improving the detection model's accuracy by collecting as many logs as possible.

Explanation

Implementing data verification at ingestion using cryptographic hashes, encrypting all data in transit and at rest, and restricting access to the pipeline with strong access controls and logging is the correct answer. This combines integrity (hash-based verification), confidentiality (encryption in transit and at rest), and access control, directly securing the AI data pipeline.

Focusing primarily on improving the detection model's accuracy is incorrect. More and faster data can help accuracy, but without encryption, integrity checks, and access control, the central data store remains exposed to data poisoning and data theft.

Using semi-structured formats like JSON for all logs and packet data is incorrect. Using a common format like JSON improves processing, not security. Sensitive values can still be exposed, and data can be altered unless protected by encryption and integrity controls.

Relying on the AI model's anomaly detection to identify any tampered or malicious training data is incorrect. This is risky because a poisoned dataset can corrupt the model's behavior and hide the manipulation it is supposed to detect.

Related Content

resources\questions\q_data_security_related_to_ai_03.question.xml

Copyright © CompTIA, Inc. All rights reserved.