AWS Certified Machine Learning Engineer - Associate Exam Guide

# Section 1.3 — Ensure Data Integrity & Prepare for Modeling (In-Depth)

This section focuses on ensuring high-quality, unbiased, and secure data is ready for ML training. It includes bias detection, compliance, quality validation, and AWS services to support these tasks.

## 1. Data Integrity & Quality

• Validate schema consistency: enforce column types, constraints

• Detect missing or corrupted records

• Use AWS Glue DataBrew / Glue Data Quality for profiling and validation

• Ensure correct splits: training, validation, test sets with stratification

## 2. Bias Detection & Mitigation

• Types of bias: selection bias, measurement bias, label bias

• Metrics: Class Imbalance (CI), Difference in Proportions (DPL)

• Strategies: resampling (over/under-sampling), synthetic data generation (SMOTE), stratified splits

• SageMaker Clarify: detect pre-training dataset bias and prediction bias

## 3. Security & Compliance

• Encryption at rest: SSE-S3, SSE-KMS, CSE

• Encryption in transit: TLS

• Anonymization/masking for sensitive data (PII/PHI)

• Data residency compliance (ensure regional storage for regulated datasets)

## 4. Preparing Data for Modeling

• Ensure balanced datasets to reduce prediction bias

• Shuffle and stratify to avoid leakage and imbalance

• Augmentation: text/image augmentation to improve generalization

• Configure loading methods: S3, Amazon EFS, Amazon FSx for Lustre (high-performance)

## 5. AWS Tools for Data Integrity & Preparation

• AWS Glue DataBrew → profile and detect anomalies

• AWS Glue Data Quality → enforce validation rules

• SageMaker Clarify → bias detection, explainability reports

• Amazon Macie → detect and protect sensitive data (PII/PHI)

• AWS Config & CloudTrail → compliance/audit tracking

## ✅ Exam Tips

• Clarify is the go-to tool for bias detection and explainability

• Stratified sampling ensures balanced splits for imbalanced datasets

• Use FSx for Lustre when you need high-performance model training with large datasets

• Macie = PII/PHI detection, KMS = encryption, IAM = access control

• Always validate and monitor datasets for corruption or schema mismatch