Domain 1: Data Engineering

Core Competency
- Collecting, preparing, and transforming data for ML use on AWS
- Using reliable, scalable, and secure data engineering pipelines

1.1 Data Ingestion
- Batch Ingestion
  - Use Amazon S3 for storing raw datasets
  - AWS Glue: Managed ETL service for large-scale data transformation
  - AWS Data Pipeline or Step Functions: Orchestration of scheduled jobs
- Streaming Ingestion
  - Amazon Kinesis Data Streams or Kinesis Firehose: Real-time streaming data
  - Amazon MSK (Managed Kafka): Apache Kafka for stream ingestion
  - AWS DMS: Change Data Capture from databases
  - Lambda: Lightweight processing during ingestion

1.2 Data Transformation and Preparation
- AWS Glue
  - Serverless ETL for:
  - Data cleaning
  - Format conversion (CSV → Parquet, etc.)
  - Partitioning
  - Glue DataBrew: No-code transformation for data wrangling
- SageMaker Data Wrangler
  - Drag-and-drop UI for:
  - Visual transformations
  - Feature engineering
  - Data validation and schema checking
  - Integration with SageMaker Pipelines
- Common Transformations
  - Missing value imputation: Pandas, Data Wrangler
  - Categorical encoding: One-hot, label encoding
  - Normalization / Standardization: Scikit-learn, Data Wrangler
  - Text cleaning: SageMaker Processing, Lambda
  - Image resizing/cropping: OpenCV, TensorFlow preprocessing, SageMaker

1.3 Feature Engineering
- Feature Types
  - Numerical: Use scaling/normalization
  - Categorical: Encoding (Label, One-Hot)
  - Text: Tokenization, stemming, embedding (e.g., Word2Vec, BERT)
  - Time Series: Lag features, rolling averages
  - Images: Pixels, embeddings from CNN
- Tools for Feature Engineering
  - SageMaker Processing: Custom scripts (Sklearn, Pandas)
  - Glue: Large-scale ETL
  - SageMaker Feature Store: Centralized store for feature reuse

1.4 SageMaker Feature Store
- Purpose
  - Store, retrieve, share features across teams & pipelines
- Components
  - Offline store: S3, for training
  - Online store: DynamoDB, for real-time inference
- Benefits
  - Feature reuse
  - Feature lineage tracking
  - Decouples feature engineering from modeling
  - Supports versioning, time-travel queries

1.5 Data Labeling and Management
- Amazon SageMaker Ground Truth
  - Managed data labeling service with:
  - Human annotators (private/public workforce)
  - Active learning
  - Built-in UI templates for images, text, video
  - Supports:
  - Bounding boxes
  - Text classification
  - Semantic segmentation
  - Output: Manifest file (JSONL format)
- Active Learning in Ground Truth
  - Model makes predictions on unlabeled data
  - Low-confidence samples prioritized for human review
  - Improves labeling efficiency and accuracy

1.6 Data Storage Best Practices
- Storage Options
  - Amazon S3: Primary ML data lake (scalable, durable)
  - S3 Glacier: Long-term archival
  - Amazon RDS / Aurora: Structured tabular data
  - Amazon DynamoDB: Online features, key-value lookup
  - Amazon Redshift: Analytical data warehouse
- Security and Compliance
  - Use SSE-KMS for encryption
  - Apply IAM roles and bucket policies
  - Enable CloudTrail + Access Analyzer for audit
- Data Versioning
  - Use S3 object versioning
  - Track versions with metadata (timestamps, hash)
  - Use manifest files to log dataset inputs

Sample Exam Scenarios
- Prepare structured dataset from multiple CSV files in S3 for ML training:
  - Use Glue job or SageMaker Data Wrangler
- Reuse engineered features across multiple models and ensure consistency:
  - Store features in SageMaker Feature Store
- Label thousands of product images for object detection task:
  - Use SageMaker Ground Truth with bounding box UI
- Ensure streaming customer interactions are stored and retrievable for real-time predictions:
  - Ingest via Kinesis, store in Feature Store (online)

Summary Checklist
- Ingest data: S3, Glue, Kinesis, DMS
- Clean/transform: Glue, DataBrew, Wrangler
- Feature engineering: SageMaker Processing, Feature Store
- Store features: Online/offline store
- Label data: Ground Truth, Manifest format
- Manage versions: S3 + Metadata tagging
- Secure storage: IAM, SSE-KMS, bucket policies

Recommended Docs & Resources
- SageMaker Data Wrangler Docs
- SageMaker Feature Store Docs
- Ground Truth Labeling Guide
- Glue and Glue DataBrew
- AWS Data Engineering on ML Lens

Presented with xmind AI