D1 · Data Prep 28% | D2 · Model Dev 26% | D3 · Deploy 22% | D4 · Monitor & Security 24%

## D1 · Data Preparation for ML

### Data Wrangler vs Processing Jobs

- Wrangler: GUI, 300+ transforms, Studio, medium data
- Connectors: S3, Athena, Redshift, Snowflake, SalesForce
- Exports: Processing Job, Pipeline step, Feature Store
- Processing Jobs: PySpark/Sklearn, large-scale, prod

**TRAP:** Wrangler cannot scale to multi-TB. Export recipe to Processing Jobs for production large-scale pipelines.

### SageMaker Feature Store

- Online: single-digit ms — real-time inference lookups
- Offline: S3 Parquet — training + batch inference
- Same pipeline feeds both = no training-serving skew
- Time-travel: event_time filter prevents data leakage

**TRAP:** Fit scalers ONLY on training split; transform val/test with training statistics — not the other way around.

### Ground Truth Workforces

- Private: sensitive (medical, legal, proprietary) data
- Mechanical Turk: general tasks, large volume, cheap
- Active learning: auto-labels confident examples (~70% savings)
- Output: augmented manifest JSON Lines — SageMaker native

### Feature Engineering Reference

| Technique | Use Case |
|---|---|
| Min-Max norm | Neural nets, KNN, SVM |
| Z-score std | Linear/logistic regression |
| Log transform | Right-skewed features |
| One-hot encode | Low-cardinality categorical |
| Target encode | High-cardinality + tree models |
| Median impute | Skewed distributions |
| SMOTE | Minority class oversampling |
| PCA | Dimensionality reduction |

### Training Input Modes

- File: copy S3 to EBS — simple, small datasets
- Pipe: stream from S3 as FIFO — large, fast start
- FastFile: FUSE-mount S3 — random access

**TRAP:** Pipe + RecordIO-protobuf = fastest for large dataset built-in algo training.

## D2 · ML Model Development

### Built-In Algorithm Selector

| Task | Algorithm | Input |
|---|---|---|
| Tabular cls/reg | XGBoost | CSV/libsvm |
| Large reg | Linear Learner | CSV/RecordIO |
| Time-series | DeepAR | JSON Lines |
| Anomaly detect | Random Cut Forest | CSV/RecordIO |
| Text cls/Word2Vec | BlazingText | Plain text |
| Recommendation | Factorization Machines | RecordIO sparse |
| Clustering | K-Means | CSV/RecordIO |
| Dim. reduction | PCA | CSV/RecordIO |
| IP anomaly | IP Insights | CSV |

**TRAP:** Factorization Machines = sparse RecordIO only. Use XGBoost for dense tabular data.

### Hyperparameter Tuning (AMT)

- Bayesian: learns from prior runs — best continuous spaces
- Random: fast, independent samples, good baseline
- Grid: exhaustive — small discrete spaces only
- Hyperband: early stopping for deep learning
- Warm start: IDENTICAL_DATA_AND_ALGORITHM / TRANSFER_LEARNING

### Managed Spot Training

- Up to 90% savings vs On-Demand instances
- Checkpointing required: save state to S3 — resume after interruption

### Distributed Training

- SDP (data parallel): full model per GPU, AllReduce
- SMP (model parallel): split model across GPUs

**TRAP:** Use SMP when model exceeds single GPU memory. SDP when model fits on 1 GPU.

### Evaluation Metrics

| Task | Metrics |
|---|---|
| Binary cls | AUC-ROC, F1, Precision, Recall |
| Multiclass | F1-macro, Accuracy, Log Loss |
| Regression | RMSE, MAE, R-squared |
| Imbalanced | F1, AUC-ROC (NOT accuracy) |

## D3 · Deployment & Orchestration

### Inference Endpoint Comparison

| Type | Latency | Payload | Best For |
|---|---|---|---|
| Real-time | ms | 6 MB | Interactive |
| Serverless | ms / cold: s | 4 MB | Infrequent/spiky |
| Async | sec–min | 1 GB | Large payloads |
| Batch Tx | min–hrs | Unlimited | Offline S3 |

**TRAP:** Async scales to 0 when idle. SNS notification on completion. Max payload 1 GB.

### A/B Testing — Production Variants

- Multiple models on one endpoint with traffic weights
- Champion 90% / Challenger 10% — per-variant metrics
- Shadow variant: mirror traffic, don't return response

### Multi-Model Endpoints (MME)

- Thousands of models on ONE endpoint, shared compute
- On-demand load from S3; LRU cache in container memory

**TRAP:** First request per model = slow cold load. All models must use same container.

### SageMaker Pipelines

- DAG steps: Processing, Training, Tuning, Transform
- ConditionStep: branch on metric threshold
- RegisterModel: add to registry if condition passes
- CallbackStep: pause for SQS external system signal
- Step caching: skip unchanged steps on re-run

### Model Registry Workflow

- Status: PendingManualApproval → Approved → Rejected
- Only Approved models deploy to production endpoints
- Lambda polls Approved event → triggers CD pipeline

### Event-Driven MLOps

- S3 arrival / Monitor drift → EventBridge → Lambda → Pipeline
- Pipeline: Process → Train → Evaluate → Condition → Register
- Approved model → Lambda → update endpoint

### Endpoint Auto Scaling

- Primary metric: InvocationsPerInstance (target tracking)
- Serverless scales to 0 — zero cost when idle
- Inference Recommender: benchmark instance types automatically

## D4 · Monitor & Security

### Model Monitor — 4 Types

| Monitor Type | Detects | Tool |
|---|---|---|
| Data Quality | Feature distribution drift, schema violations | Built-in |
| Model Quality | Accuracy/F1 degradation vs ground truth | Built-in |
| Bias Drift | Fairness changes (DPPL, DI, CI) | Clarify |
| Feature Attribution | SHAP value shift — feature importance | Clarify |

**TRAP:** All 4 monitor types require: data capture enabled on endpoint + baseline computed on training data first.

### Clarify Bias Metrics

- Pre-training: Class Imbalance (CI), Difference in Positive Labels (DPL)
- Post-training: DPPL, Disparate Impact (DI), Treatment Equity
- DI < 0.8 = potential discrimination (4/5 rule)
- SHAP: per-feature contribution per prediction (local + global)

### SageMaker Encryption

| Resource | Parameter |
|---|---|
| Training EBS volume | ResourceConfig.VolumeKmsKeyId |
| Model artifacts S3 | OutputDataConfig.KmsKeyId |
| S3 training data | SSE-KMS on S3 bucket |
| Inter-container traffic | EnableInterContainerTrafficEncryption |
| Feature Store online | OnlineStoreConfig.SecurityConfig.KmsKeyId |
| Feature Store offline | OfflineStoreConfig.S3StorageConfig.KmsKeyId |
| Studio EFS volume | KMS key on EFS file system |
| CloudWatch logs | KMS encryption on log group |

### VPC Mode — Required Endpoints (Private Subnet)

- S3 Gateway Endpoint: FREE — training data from S3
- ECR Interface Endpoint: pull container images
- SageMaker API Interface: SDK calls from VPC
- CloudWatch Interface: publish logs
- NetworkIsolation=True: block ALL internet from container

**TRAP:** Cannot pull ECR image in private subnet = missing ECR VPC Interface endpoint.

### IAM, Audit & Governance

- Execution role per resource: training, endpoint, pipeline

D1 · Data Prep 28% | D2 · Model Dev 26% | D3 · Deploy 22% | D4 · Monitor & Security 24%

**TRAP:** Class imbalance: accuracy is misleading. A model always predicting majority achieves high accuracy. Use F1 or AUC-ROC.

### Autopilot + JumpStart + Experiments

- Autopilot: AutoML — preprocesses, selects algos, tunes HPO
- Output: leaderboard + 2 explainability notebooks
- JumpStart: Llama 2, BERT, Stable Diffusion fine-tuning
- Experiments: Experiment > Trial > Trial Component hierarchy

- Training: s3:GetObject/Put, ecr:*, kms:Decrypt/GenerateDataKey
- Role Manager: Data Scientist, MLOps Engineer personas
- SCPs: enforce VPC mode + network isolation org-wide
- Macie: PII detection in captured inference data + training data
- CloudTrail: all SM API calls — enable data events for S3
- Lineage: dataset → training → model → endpoint graph

## Master Quick Reference — Key Facts & Exam Decision Rules

| Service / Concept | Key Facts | Exam Trap or Decision Rule |
|---|---|---|
| Data Wrangler | GUI, 300+ transforms, Studio, medium datasets | Cannot scale to multi-TB; export recipe to Processing Jobs for production |
| Feature Store | Online (ms) + Offline (S3); same pipeline = no skew | Offline for training; online for real-time inference; time-travel prevents leakage |
| Processing Jobs | Managed PySpark/Sklearn; input/output via S3 | Use for large-scale feature engineering Data Wrangler cannot handle |
| Ground Truth Active Learn | Auto-labels confident; humans handle ambiguous only | Reduces labeling effort ~70%; output = augmented manifest JSON Lines |
| Training Input Modes | File: copy EBS; Pipe: stream FIFO; FastFile: FUSE | Pipe + RecordIO = fastest for large datasets with built-in algorithms |
| XGBoost HPO | num_round, max_depth, eta, subsample, lambda, alpha | Lower max_depth + higher lambda = less overfitting; use early_stopping_rounds |
| Managed Spot Training | Up to 90% savings; requires S3 checkpointing | checkpoint_s3_uri required; job resumes from last checkpoint after interruption |
| SDP / SMP Distributed | SDP: full model per GPU + AllReduce; SMP: split model | SMP when model exceeds single GPU memory; SDP when model fits on 1 GPU |
| DeepAR | JSON Lines input; one model for MULTIPLE time series | Outputs probabilistic forecasts with quantiles (p10/p50/p90) |
| Factorization Machines | High-dim sparse input; RecordIO-protobuf only | For recommendation systems NOT dense tabular — use XGBoost for tabular |
| HPO Bayesian vs Random | Bayesian learns from prior runs; Random is independent | Bayesian = best continuous spaces; warm start = IDENTICAL_DATA_AND_ALGORITHM |
| Inference Endpoint Types | RT: ms/6MB; Serverless: cold-start; Async: 1GB; Batch: S3 | Async for >6MB payloads; Serverless for infrequent traffic; Batch for offline |
| Multi-Model Endpoint | Thousands of models on 1 endpoint; on-demand S3 load | First request per model = slow cold load; all models need same container |
| Pipelines ConditionStep | Branch on metric threshold; gates RegisterModel step | CallbackStep = pause pipeline for SQS external API signal; caching skips unchanged steps |
| Model Registry | Versioned catalog; Pending → Approved → Rejected | Only Approved models deploy; Lambda detects approval event for CD trigger |
| Model Monitor 4 Types | Data Quality, Model Quality, Bias Drift, Feat Attribution | All 4: data capture + baseline required; Bias + Attribution use Clarify/SHAP |
| DI (Disparate Impact) | Ratio positive predictions: favored vs disfavored group | DI < 0.8 = potential discrimination (4/5 rule); regulatory compliance concern |
| SHAP Explainability | Per-feature contribution per prediction (local + global) | Required for Feature Attribution Drift Monitor and GDPR right-to-explanation |
| VPC Endpoints for SM | S3 Gateway (free); ECR + SM API + CW as Interface | Missing ECR endpoint = training job fails to pull Docker image in private subnet |
| SageMaker Encryption | VolumeKmsKeyId (EBS); OutputDataConfig.KmsKeyId (S3) | Feature Store needs 2 KMS keys: one for online store, one for offline (S3) |
| IAM Execution Role | Per resource; S3 + ECR + CW + KMS permissions | Least privilege: specific S3 prefixes not s3:*; SCPs enforce org-wide guardrails |
| Macie + CloudTrail | Macie: PII detection in S3; CloudTrail: API audit | Enable CloudTrail data events for S3 object-level audit; query logs with Athena |

D1 · Data Prep 28% | D2 · Model Dev 26% | D3 · Deploy 22% | D4 · Monitor & Security 24%

| Service / Concept | Key Facts | Exam Trap or Decision Rule |
|---|---|---|
| Data Wrangler | GUI, 300+ transforms, Studio, medium datasets | Cannot scale to multi-TB; export recipe to Processing Jobs for production |
| Feature Store | Online (ms) + Offline (S3); same pipeline = no skew | Offline for training; online for real-time inference; time-travel prevents leakage |
| Processing Jobs | Managed PySpark/Sklearn; input/output via S3 | Use for large-scale feature engineering Data Wrangler cannot handle |
| Ground Truth Active Learn | Auto-labels confident; humans handle ambiguous only | Reduces labeling effort ~70%; output = augmented manifest JSON Lines |
| Training Input Modes | File: copy EBS; Pipe: stream FIFO; FastFile: FUSE | Pipe + RecordIO = fastest for large datasets with built-in algorithms |
| XGBoost HPO | num_round, max_depth, eta, subsample, lambda, alpha | Lower max_depth + higher lambda = less overfitting; use early_stopping_rounds |
| Managed Spot Training | Up to 90% savings; requires S3 checkpointing | checkpoint_s3_uri required; job resumes from last checkpoint after interruption |
| SDP / SMP Distributed | SDP: full model per GPU + AllReduce; SMP: split model | SMP when model exceeds single GPU memory; SDP when model fits on 1 GPU |
| DeepAR | JSON Lines input; one model for MULTIPLE time series | Outputs probabilistic forecasts with quantiles (p10/p50/p90) |
| Factorization Machines | High-dim sparse input; RecordIO-protobuf only | For recommendation systems NOT dense tabular — use XGBoost for tabular |
| HPO Bayesian vs Random | Bayesian learns from prior runs; Random is independent | Bayesian = best continuous spaces; warm start = IDENTICAL_DATA_AND_ALGORITHM |
| Inference Endpoint Types | RT: ms/6MB; Serverless: cold-start; Async: 1GB; Batch: S3 | Async for >6MB payloads; Serverless for infrequent traffic; Batch for offline |
| Multi-Model Endpoint | Thousands of models on 1 endpoint; on-demand S3 load | First request per model = slow cold load; all models need same container |
| Pipelines ConditionStep | Branch on metric threshold; gates RegisterModel step | CallbackStep = pause pipeline for SQS external API signal; caching skips unchanged steps |
| Model Registry | Versioned catalog; Pending → Approved → Rejected | Only Approved models deploy; Lambda detects approval event for CD trigger |
| Model Monitor 4 Types | Data Quality, Model Quality, Bias Drift, Feat Attribution | All 4: data capture + baseline required; Bias + Attribution use Clarify/SHAP |
| DI (Disparate Impact) | Ratio positive predictions: favored vs disfavored group | DI < 0.8 = potential discrimination (4/5 rule); regulatory compliance concern |
| SHAP Explainability | Per-feature contribution per prediction (local + global) | Required for Feature Attribution Drift Monitor and GDPR right-to-explanation |
| VPC Endpoints for SM | S3 Gateway (free); ECR + SM API + CW as Interface | Missing ECR endpoint = training job fails to pull Docker image in private subnet |
| SageMaker Encryption | VolumeKmsKeyId (EBS); OutputDataConfig.KmsKeyId (S3) | Feature Store needs 2 KMS keys: one for online store, one for offline (S3) |
| IAM Execution Role | Per resource; S3 + ECR + CW + KMS permissions | Least privilege: specific S3 prefixes not s3:*; SCPs enforce org-wide guardrails |
| Macie + CloudTrail | Macie: PII detection in S3; CloudTrail: API audit | Enable CloudTrail data events for S3 object-level audit; query logs with Athena |