AWS Certified Data Engineer - Associate Cheat Sheet

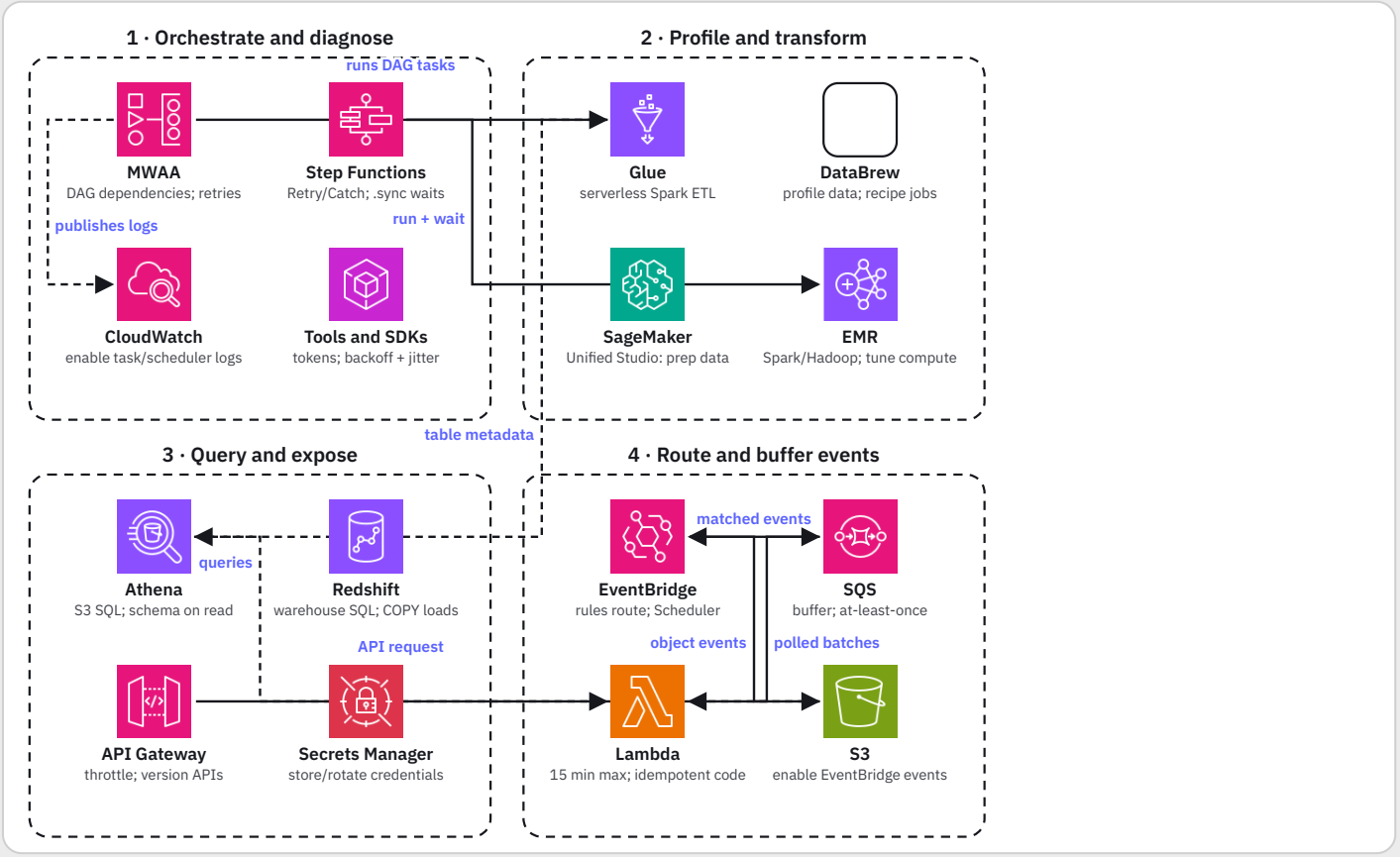
DEA-C01 · Domain 3 of 4

BlowStack

DOMAIN 3 · DATA OPERATIONS AND SUPPORT · 22% OF THE EXAM · TASK STATEMENT 3.1

Automate data processing by using AWS services

Orchestrate, prepare, query, react



1 Orchestrate and diagnose

DAGs express Airflow task dependencies; Step Functions coordinates service states. SDKs can use AWS Security Token Service (AWS STS) temporary credentials; pagination requires following continuation tokens.

2 Profile and transform

No-code profiling and reusable recipes favor DataBrew; Glue runs managed ETL, while EMR provides framework control. Profiling reveals nulls, distributions, and type mismatches before transformation.

3 Query and expose

Schema on read favors Athena for S3 SQL without warehouse loading; Glue supplies table metadata. Redshift serves warehouse analytics. API versions preserve contracts; Lambda can retrieve secrets at runtime.

4 Route and buffer events

EventBridge Scheduler handles timed invocations; rules match events, and SQS buffers bursts. Lambda polls SQS and deletes successful batches; failures reappear after visibility timeout. Code must enforce idempotency.

Exam traps

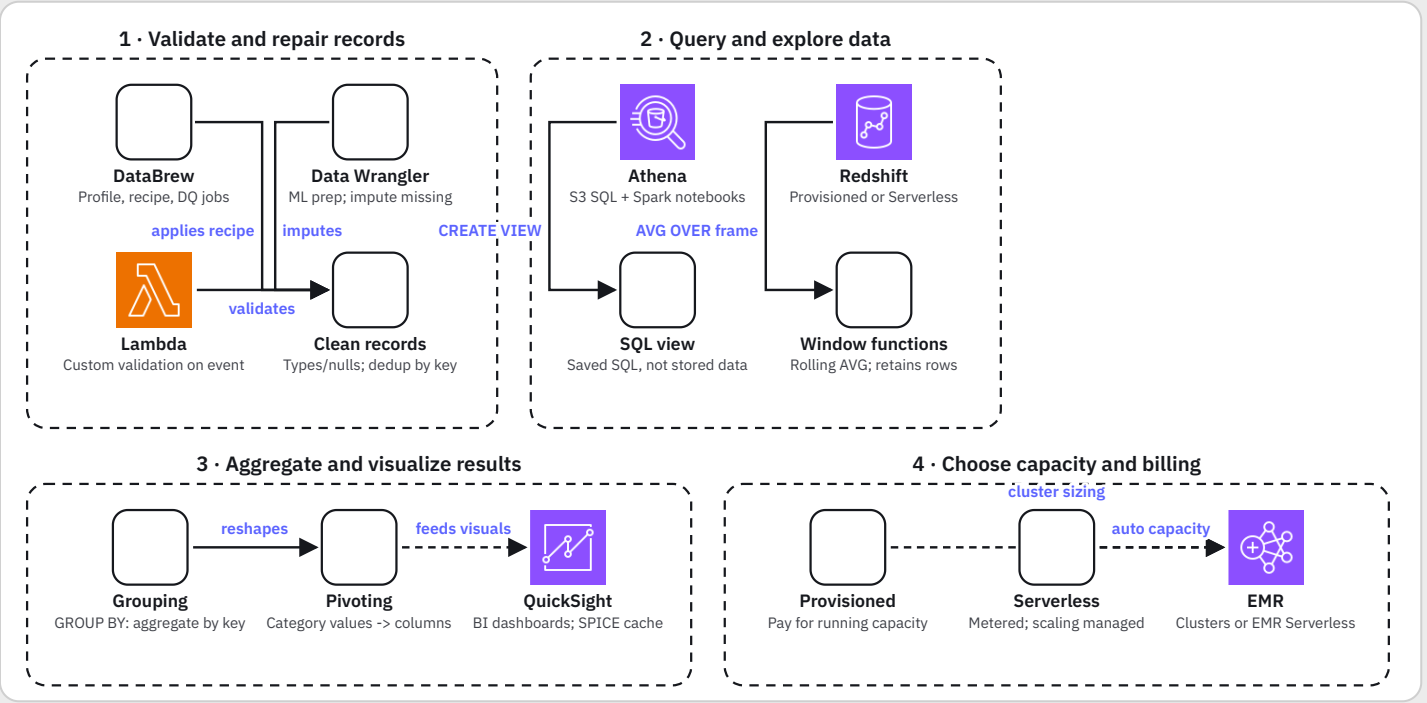
- **Task logs versus scheduler logs**: task logs diagnose execution failures; scheduler logs diagnose scheduling issues. Retries do not repair missing permissions or broken network access.
- **At-least-once delivery** requires durable deduplication, not in-memory flags. SQS visibility timeout is not retention; enabling partial batch responses lets Lambda retry only failed messages.
- **Throttling is not a quota guarantee**: API Gateway limits are best effort. SDK retries need capped exponential backoff with jitter; retrying non-idempotent writes can duplicate results.

Get practice exams and lessons for AWS Certified Data Engineer - Associate

DOMAIN 3 · DATA OPERATIONS AND SUPPORT · TASK STATEMENT 3.2

Analyze data by using AWS services

Clean, query, visualize, scale



1 Validate and repair records

Use **DataBrew** for no-code profiles/recipes; Data Wrangler for ML preparation. Imputation replaces missing values; deduplication needs a business key and survivor rule. Lambda supplies custom checks.

2 Query and explore data

Window functions retain detail rows; GROUP BY collapses groups. Rolling averages use AVG(...) OVER with ordering and a frame. Athena and Redshift support SQL views; standard views store queries, not results.

3 Aggregate and visualize results

QuickSight builds dashboards; SPICE caches imported data, while direct query reads the source. Aggregation summarizes records with SUM, COUNT, or AVG; grouping partitions those calculations by key.

4 Choose capacity and billing

Serverless fits variable demand: Athena SQL bills scanned bytes; Spark notebooks bill compute. Redshift and EMR offer provisioned and serverless options; provisioned capacity can suit steady workloads.

Exam traps

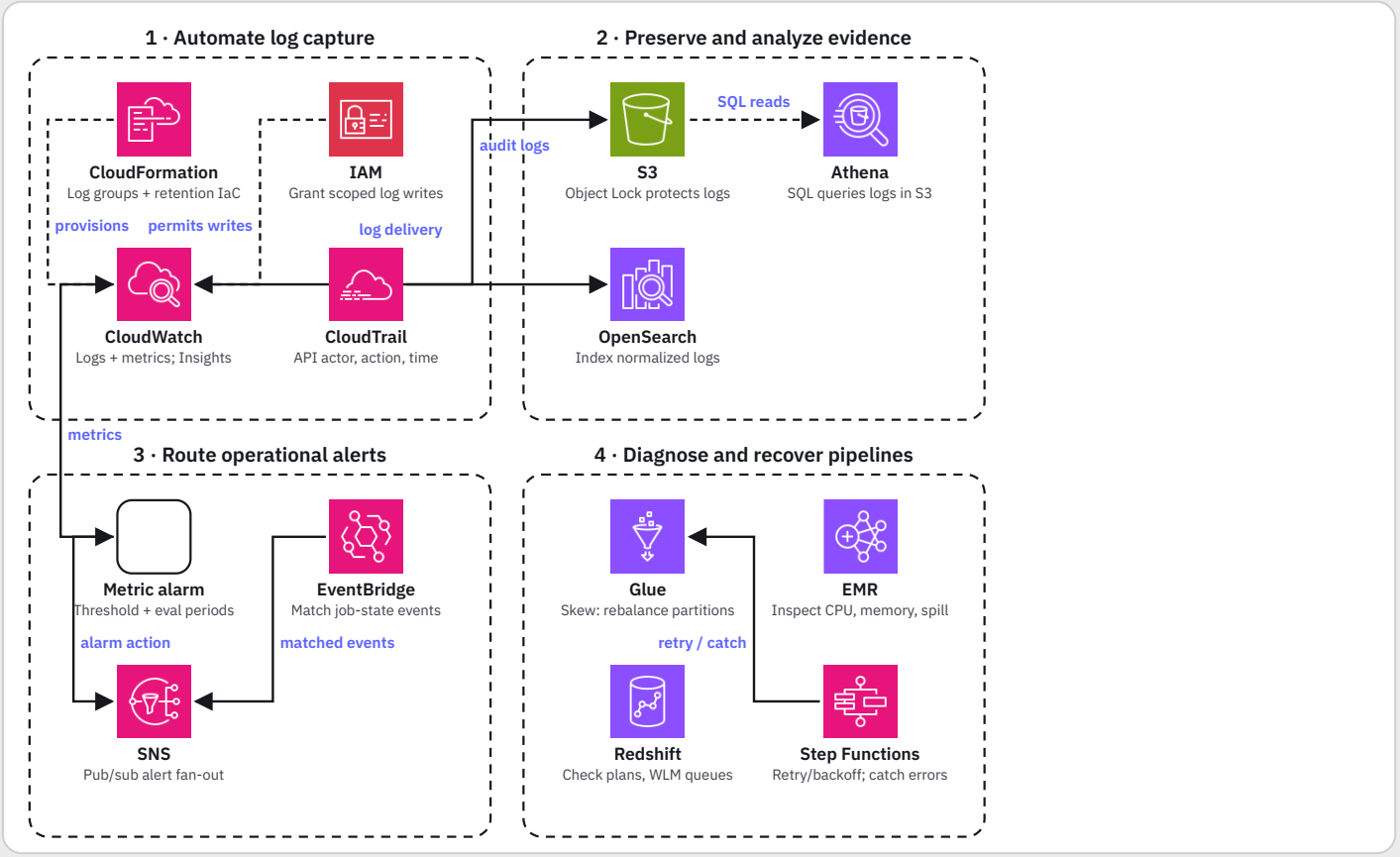
- **DISTINCT** removes identical selected rows, not every duplicate business entity. ROW_NUMBER partitioned by business key with deterministic ordering can select the intended survivor.
- **Serverless** does not mean free or always cheaper. Athena SQL scans cost money; partition pruning and columnar formats reduce scanned bytes. Athena Spark notebooks use separate compute billing.
- **Pivoting** turns category values into columns. Lateral movement is an attacker's traversal between resources or identities, not an analytics transformation.

Get practice exams and lessons for AWS Certified Data Engineer - Associate

DOMAIN 3 · DATA OPERATIONS AND SUPPORT · TASK STATEMENT 3.3

Maintain and monitor data pipelines

Capture, preserve, alert, recover



1 Automate log capture

Repeatable capture uses CloudFormation for log groups, retention and IAM permissions. Applications must emit logs; CloudWatch Logs Insights queries them. CloudTrail records API activity, not application internals.

2 Preserve and analyze evidence

Audit evidence needs retention and integrity: validate CloudTrail digests to detect tampering. Athena queries S3; OpenSearch requires ingestion. Normalize timestamps and request IDs to correlate events across sources.

3 Route operational alerts

Threshold breach selects CloudWatch alarms; matching job-state events selects EventBridge rules. Both can publish to SNS for subscriber fan-out. Email subscriptions require confirmation.

4 Diagnose and recover pipelines

Safe recovery combines checkpoints with idempotent writes. Retry transient faults with capped exponential backoff and jitter. Rising lag signals downstream pressure; inspect skew and resource utilization before scaling.

Exam traps

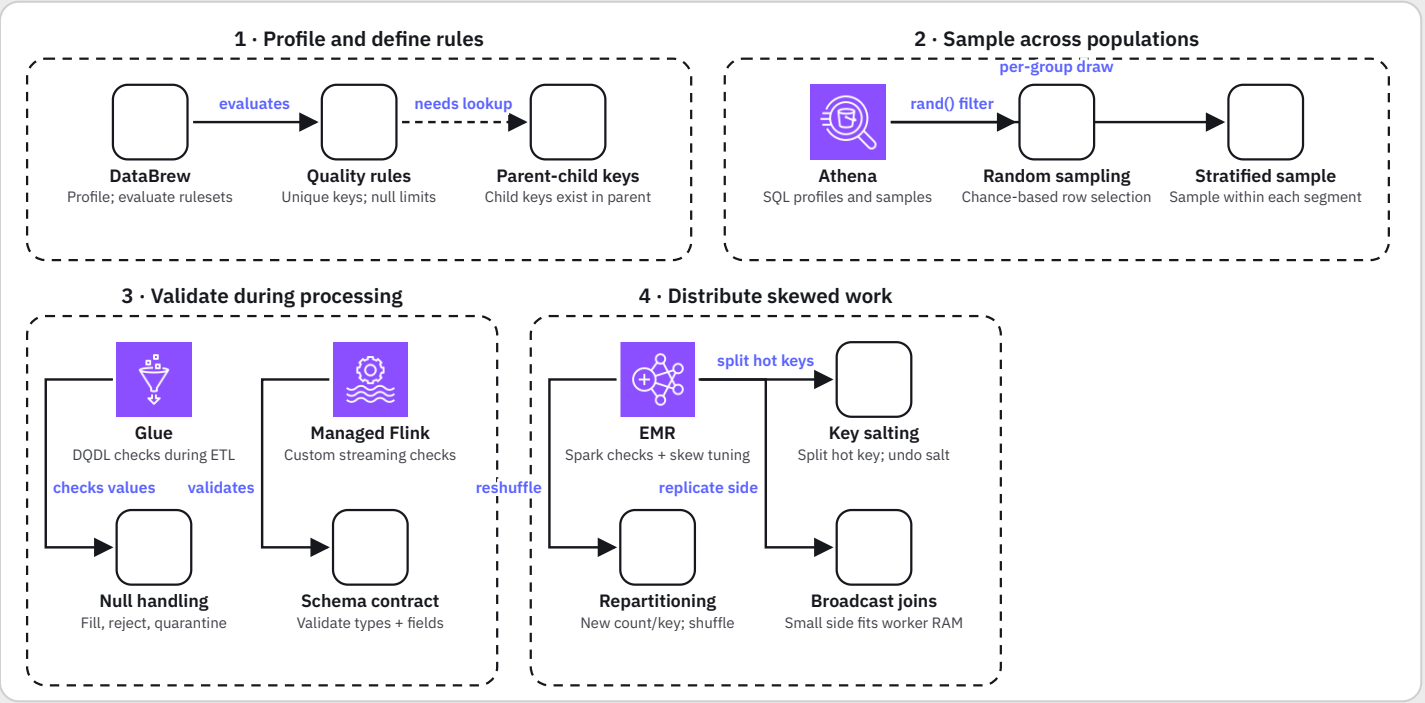
- CloudTrail Event history provides 90 days of management events, not application logs. **S3 object data events** require explicit selection; a trail delivers ongoing audit logs to S3.
- **Integrity validation detects tampering**; it does not prevent deletion. S3 Object Lock protects retained object versions. CloudWatch log retention does not configure S3 retention.
- **More workers do not fix a hot key**: skew can leave one task overloaded. Checkpoints reduce replay, but do not alone prevent duplicate output; retries still require idempotent writes.

Get practice exams and lessons for AWS Certified Data Engineer - Associate

DOMAIN 3 · DATA OPERATIONS AND SUPPORT · TASK STATEMENT 3.4

Ensure data quality

Profile, sample, validate, rebalance



1 Profile and define rules

Use **DataBrew rulesets** for recurring null/uniqueness checks; profiles expose anomalies. Parent-child consistency requires explicit cross-dataset checks in Glue or Athena.

2 Sample across populations

Choose **stratified sampling** to cover rare segments; random samples can miss them. Weight strata for population estimates. Representativeness depends on coverage; samples cannot prove full-dataset quality.

3 Validate during processing

Use **in-flight checks** in Glue ETL or Flink streams to catch missing values and incompatible schemas before publication. Glue rules require configuration; Flink validation and failure routing require application logic.

4 Distribute skewed work

For **hot-key skew**, salt keys to spread work. Repartitioning changes distribution but costs a shuffle. Broadcast a small join side only when it fits each worker; these Spark techniques apply to Glue and EMR.

Exam traps

- **Rules are not constraints**: quality checks detect violations; lake uniqueness and foreign keys are not automatically enforced. Failure handling must be configured.
- **More partitions may not fix skew**: the same hot key can still overload one partition. Salted joins need matching keys; salted aggregations need a final merge.
- **NULL is not an empty string**: test both explicitly. SQL = NULL does not detect nulls; use IS NULL. Missing fields and incompatible field types also require schema validation.

Get practice exams and lessons for AWS Certified Data Engineer - Associate

DOMAIN 3 · DATA OPERATIONS AND SUPPORT · WHAT TO REMEMBER

Concepts and services of Data Operations and Support

Concepts to know

Data Quality

Quality rules test completeness, validity, uniqueness, and relationships. Choose **validation before downstream publication** to reject or quarantine failures; profiling discovers patterns but does not enforce correctness.

Idempotency

Repeated processing produces the same intended result without duplicate effects. Choose **deduplication keys or idempotent writes** for retried tasks and duplicate events; retries and checkpointing alone do not guarantee duplicate-free output.

Data Skew

Uneven partitions create stragglers. Choose **key salting for hot keys**, repartitioning for uneven distribution, or broadcast joins when the small side fits worker memory; adding workers alone may not resolve skew.

Core services



AWS Glue

Choose **managed ETL with integrated quality checks** for recurring transformations. Job bookmarks track processed inputs; they do not guarantee record-level deduplication.



Amazon CloudWatch

Choose **metrics, alarms, and application logs** to detect pipeline failures and investigate performance. Logs Insights queries operational logs; AWS CloudTrail records API activity.



Amazon Athena

Choose **serverless SQL over Amazon S3** for ad hoc exploration and validation without loading a warehouse. Partition pruning and columnar formats reduce scanned data and scan-based query costs.

Get practice exams and lessons for AWS Certified Data Engineer - Associate