AWS Certified Data Engineer - Associate Cheat Sheet

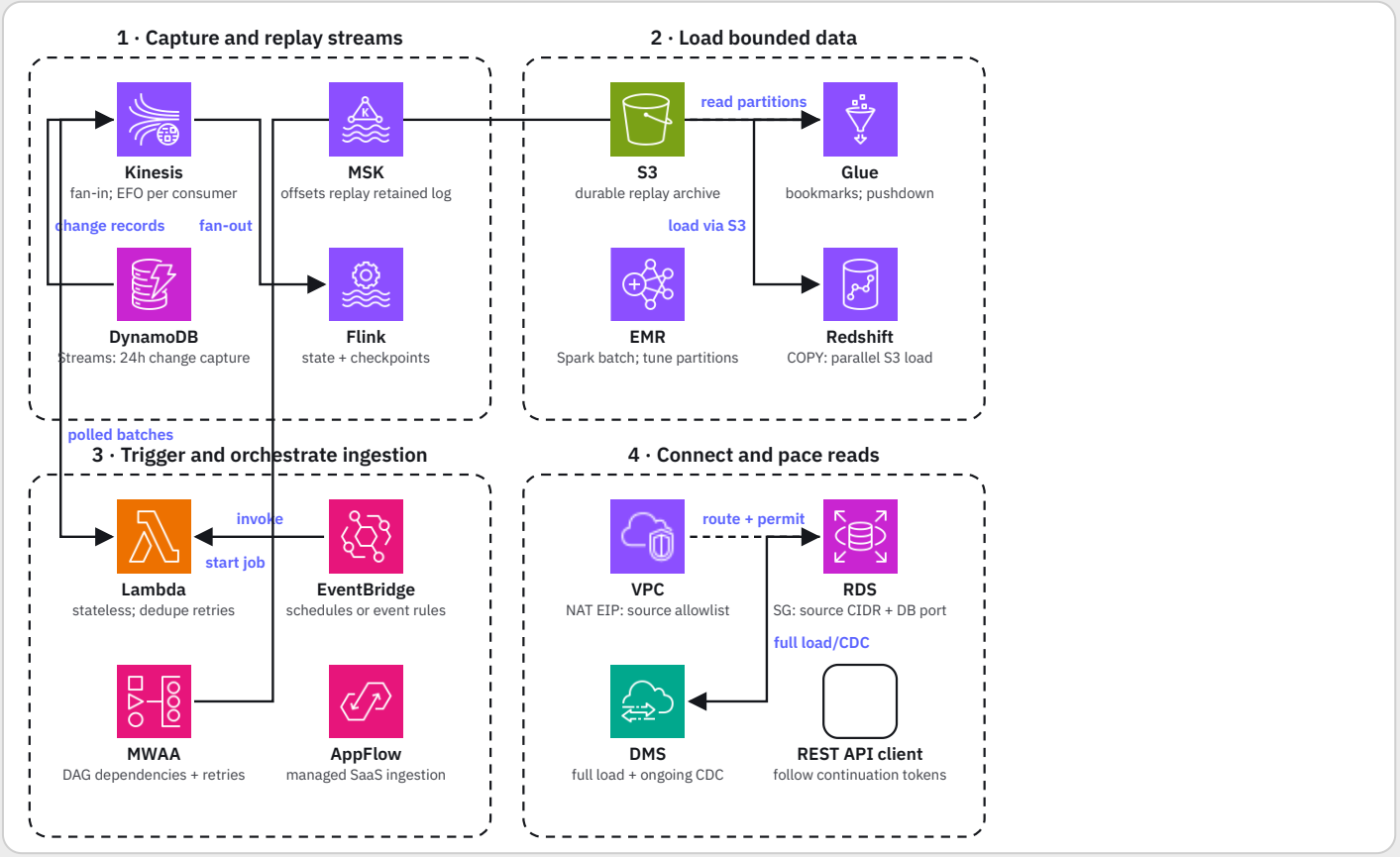
DEA-C01 · Domain 1 of 4

BlowStack

DOMAIN 1 · DATA INGESTION AND TRANSFORMATION · 34% OF THE EXAM · TASK STATEMENT 1.1

Perform data ingestion

Capture, load, trigger, connect



1 Capture and replay streams

Replay needs retained input plus a source position; checkpoints cannot restore expired records. Kinesis enhanced fan-out gives consumers dedicated read throughput; shared consumers divide read capacity.

2 Load bounded data

Incremental loading uses Glue bookmarks or durable watermarks instead of rereading everything. Predicate pushdown reduces source reads; partition pruning skips partitions. Redshift COPY favors multiple balanced files.

3 Trigger and orchestrate ingestion

Dependencies favor MWAA DAGs; independent timed runs favor EventBridge or Glue schedules. Cron syntax and time zones differ. S3 notifications trigger arrivals; Kinesis event source mappings poll batches for Lambda.

4 Connect and pace reads

Allowlisting public sources uses the NAT public IP, not a private subnet CIDR. Private sources need routing and security-group access. Throttling needs bounded concurrency, backpressure, and exponential backoff with jitter.

Exam traps

- **At-least-once is not exactly-once.** Lambda stream batches can repeat; bookmarks and checkpoints do not deduplicate target writes. Idempotent keys or upserts prevent duplicate effects.
- **Stateful is not ACID.** Flink windows retain state; Lambda instances must not be the durable state store. ACID governs supported transaction boundaries, not an entire ingestion pipeline automatically.
- **Fan-out is not load balancing.** Independent Kafka consumer groups each read a topic; consumers within one group divide partitions. Multiple producers fan in, so combined throughput must fit capacity.

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

AWS Certified Data Engineer - Associate Cheat Sheet

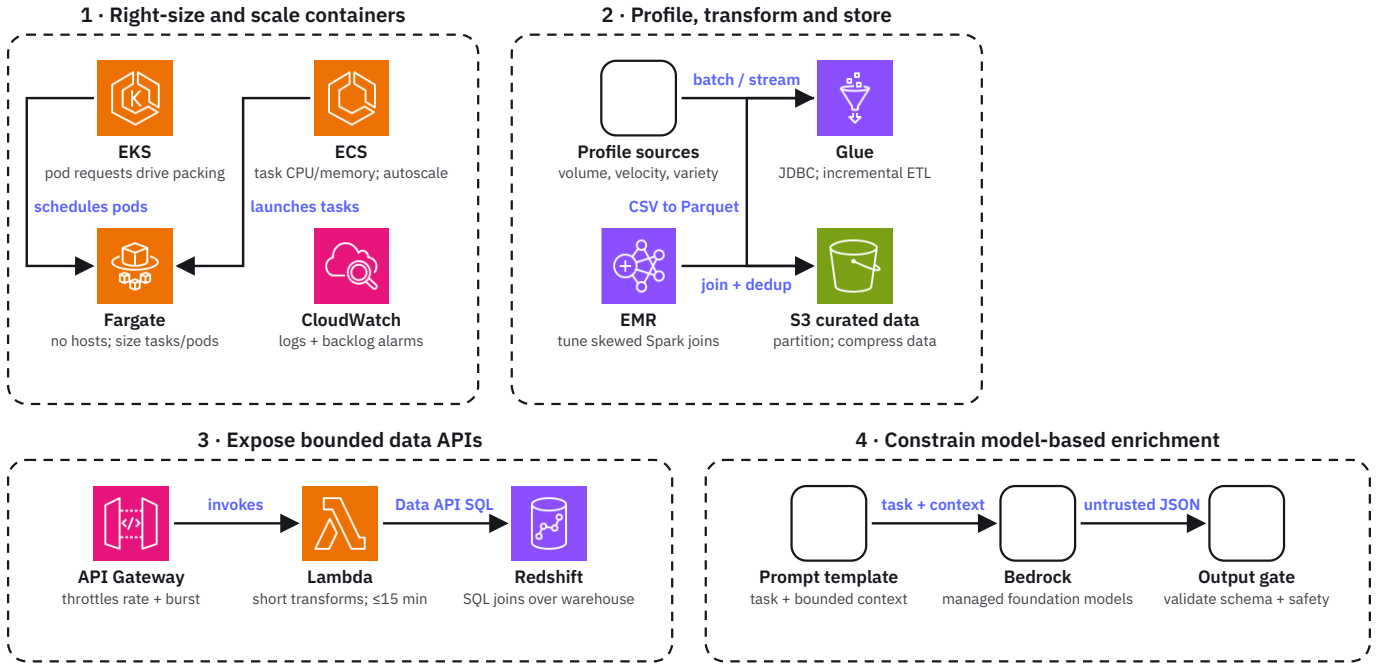
DEA-C01 · Domain 1 of 4

BlowStack

DOMAIN 1 · DATA INGESTION AND TRANSFORMATION · TASK STATEMENT 1.2

Transform and process data

Scale, transform, expose, enrich



1 Right-size and scale containers

Right-Sizing matches CPU/memory to observed demand; horizontal scaling adds tasks or pods. AWS Fargate removes host management, not resource limits. Backpressure requires buffering and sufficient consumer throughput.

2 Profile, transform and store

AWS Glue favors managed ETL; Amazon EMR favors framework control. JDBC/ODBC need compatible drivers, credentials and network access. Tables are structured, JSON semi-structured, text unstructured.

3 Expose bounded data APIs

Pagination bounds response size with continuation tokens. AWS Lambda can submit SQL through the Redshift Data API and poll asynchronous statements. API throttling protects backends; it does not paginate query results.

4 Constrain model-based enrichment

Choose **Amazon Bedrock** for managed foundation-model inference; Amazon SageMaker for custom model training/hosting. Treat source text as untrusted; validate output before writes or API actions.

Exam traps

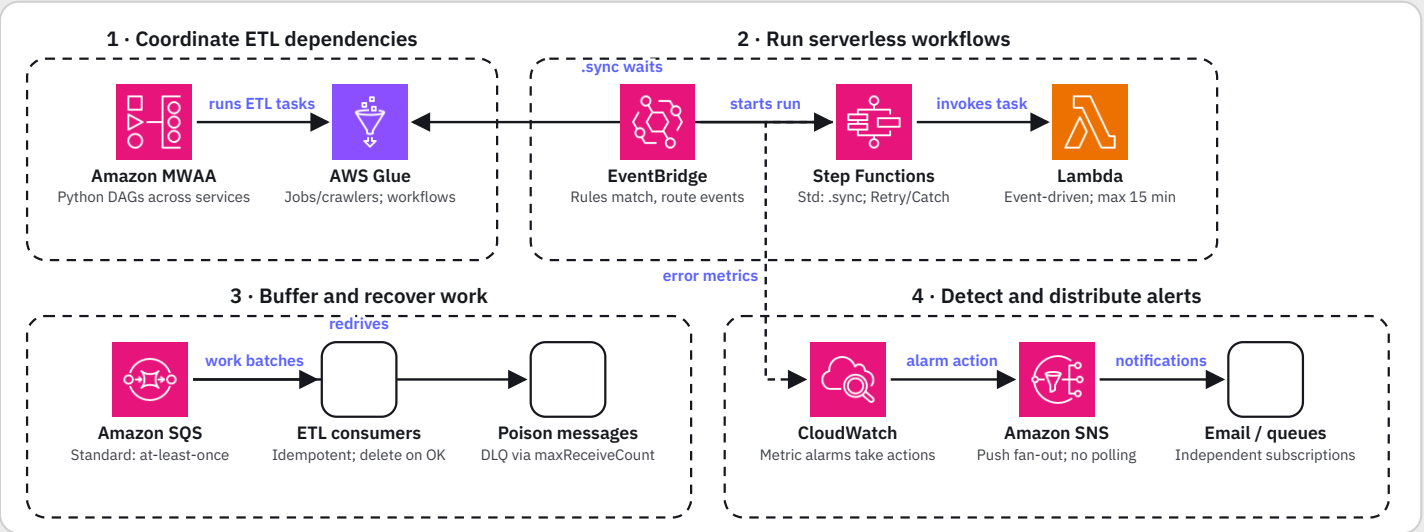
- **Glue bookmarks are not deduplication or CDC**: they track processed input for supported sources. Updates, deletes and replay-safe target writes still require explicit handling.
- **More compute does not fix skew**: inspect slow tasks and repartition or salt hot keys. For Athena cost, reduce bytes scanned with columnar Parquet, compression and partition pruning.
- **Joins can multiply rows** when keys repeat; deduplicate using business keys and ordering rules. Schema discovery does not make type changes compatible with downstream transformations.

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

DOMAIN 1 · DATA INGESTION AND TRANSFORMATION · TASK STATEMENT 1.3

Orchestrate data pipelines

Coordinate, trigger, buffer, alert



1 Coordinate ETL dependencies

Choose **Glue workflows** for Glue job/crawler dependencies; choose Amazon MWAA for Python Airflow DAGs across services. Managed infrastructure does not define DAG logic or retry policies.

2 Run serverless workflows

Choose **Standard .sync** to wait for Glue job completion; EventBridge routes events, Step Functions tracks workflow state, and Lambda runs bounded tasks. Retry/Catch handles configured errors, not every failure.

3 Buffer and recover work

For **bursty work**, SQS buffers across multiple AZs; scale consumers within downstream limits. Set visibility timeout for processing, delete only on success, and use idempotency plus a DLQ for safe recovery.

4 Detect and distribute alerts

Use **SNS fan-out** for alerts to people or multiple consumers; CloudWatch alarms publish notifications on configured state changes. Add SQS subscriptions when consumers need durable buffering.

Exam traps

- **Express workflows** have a five-minute maximum and do not support .sync integration. Standard workflows support long-running orchestration and waiting for Glue job completion.
- **Visibility timeout** hides a message; it does not delete it. Failed or slow processing can cause redelivery. Even FIFO deduplication does not make external ETL side effects exactly-once.
- **EventBridge** routes events; it does not track multi-step dependencies or wait for ETL completion. Lambda's 15-minute limit makes it unsuitable for waiting on long-running jobs.

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

AWS Certified Data Engineer - Associate Cheat Sheet

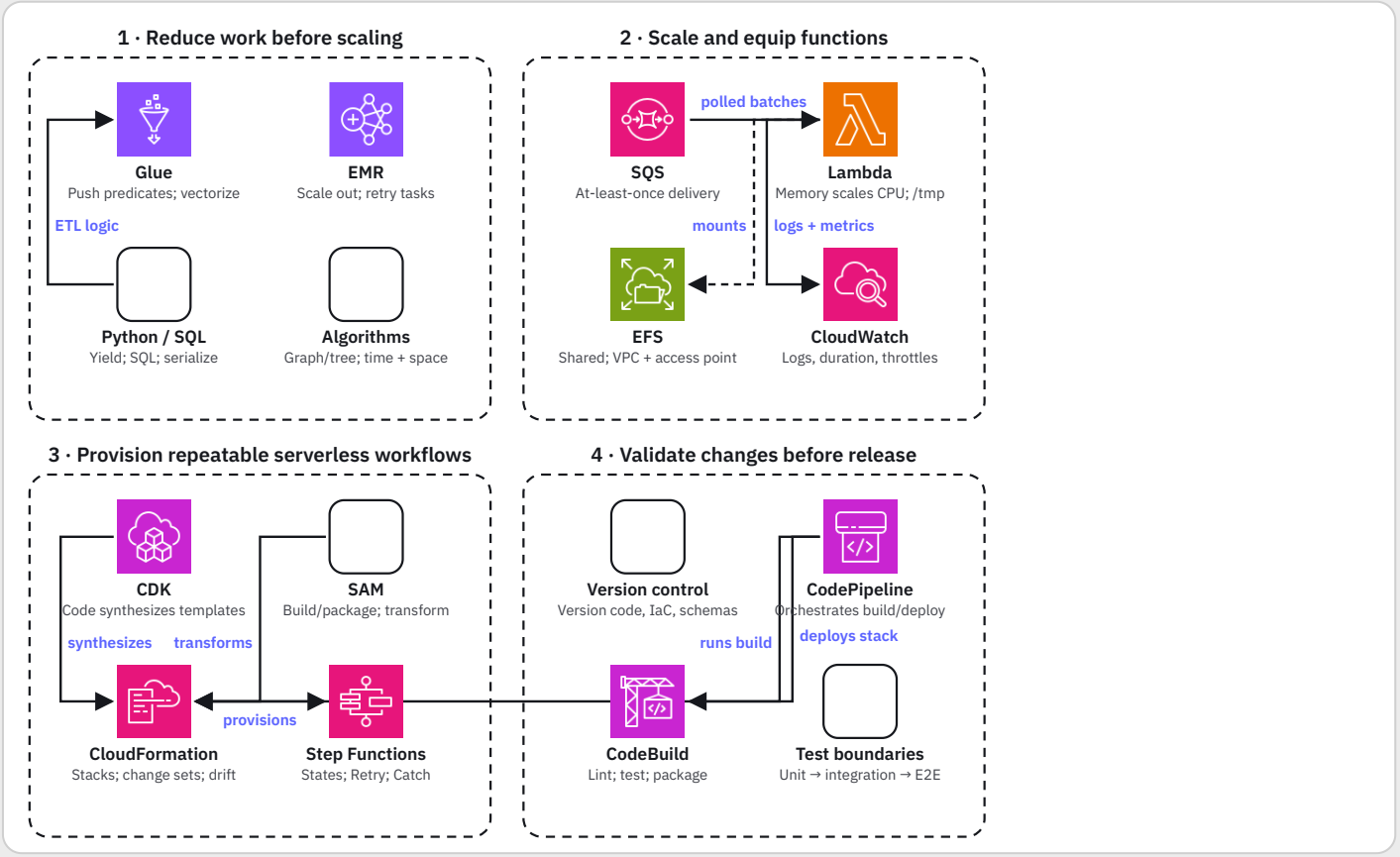
DEA-C01 · Domain 1 of 4

BlowStack

DOMAIN 1 · DATA INGESTION AND TRANSFORMATION · TASK STATEMENT 1.4

Apply programming concepts

Optimize, scale, provision, validate



1 Reduce work before scaling

Use **pushdown** to read less and vectorization to avoid scalar loops. Generators bound memory; joins can multiply rows; windows retain rows. Distributed workers process partitions; shuffles reduce data locality.

2 Scale and equip functions

Use **reserved concurrency** to reserve and cap capacity; provisioned concurrency prewarms environments. SQS event source mappings poll and batch; partial batch responses retry failed messages rather than successful ones.

3 Provision repeatable serverless workflows

Choose **SAM** for serverless functions, workflows and tables; CDK for language-based IaC. Both use CloudFormation. SAM packages code and dependencies; change sets preview infrastructure updates before execution.

4 Validate changes before release

Use **CI/CD** to validate revisions and promote artifacts. Unit tests isolate logic; integration tests exercise boundaries; E2E tests verify pipeline outcomes. Log structured context; retry transient errors, quarantine bad records.

Exam traps

- **Provisioned concurrency** reduces cold starts; it does not cap concurrency. Reserved concurrency reserves and caps it. Required concurrency ≈ requests/second × average duration in seconds.
- **/tmp is not shared storage**: Lambda provides 512–10,240 MB per execution environment; reuse is not guaranteed. EFS provides persistent shared files through a VPC mount and access point.
- **Idempotency is not drift repair**: repeated IaC deployments converge on declared state; drift detection reports supported differences without fixing them. Duplicate SQS events still need idempotent code.

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

DOMAIN 1 · DATA INGESTION AND TRANSFORMATION · WHAT TO REMEMBER

Concepts and services of Data Ingestion and Transformation

Concepts to know

Incremental Loading

Incremental loading ingests changes instead of full datasets. Choose **CDC for ongoing database changes**; watermarks or job bookmarks track eligible new data, but updates, deletes, and late arrivals require explicit handling.

Idempotency

Idempotency makes repeated processing produce the same intended result. **At-least-once delivery requires retry-safe writes**: use stable event identifiers, deduplication, or upserts so retries and replay do not duplicate results.

Predicate Pushdown

Predicate pushdown applies supported filters at the source or reader before downstream processing. Choose **filtering before data movement** to reduce scans, transfer, and compute; filtering after loading does not provide the same savings.

Core services



AWS Glue

Choose **managed Spark ETL** for distributed transformations. Crawlers discover schemas, job bookmarks track supported incremental inputs, and pushdown reduces data read.



Amazon Kinesis Data Streams

Choose **replayable streaming with independent consumers**. Retention bounds replay; partition keys affect ordering and hot shards; enhanced fan-out provides dedicated read throughput per consumer.



AWS Lambda

Choose **short, event-driven transformations** within the 15-minute limit. Tune memory, batching, and concurrency; make stream processing retry-safe and use partial batch responses to limit reprocessing.

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