AWS Certified Data Engineer - Associate Cheat Sheet

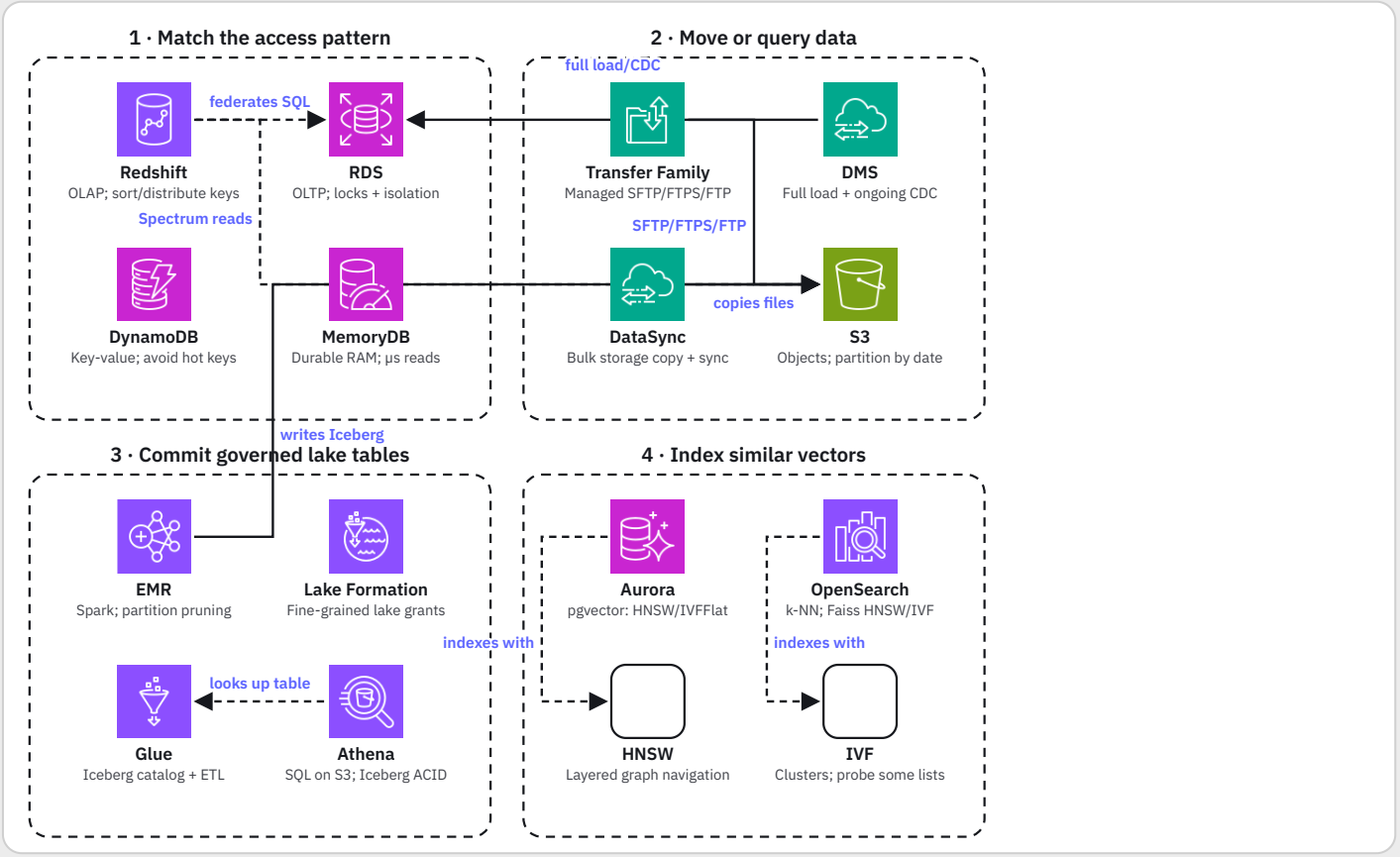
DEA-C01 · Domain 2 of 4

BlowStack

DOMAIN 2 · DATA STORE MANAGEMENT · 26% OF THE EXAM · TASK STATEMENT 2.1

Choose a data store

Match, migrate, commit, search



1 Match the access pattern

Choose by **access pattern**: Redshift for analytical scans, RDS for relational transactions, DynamoDB for key-based access, MemoryDB for durable microsecond reads. Sort/distribution keys reduce scan/shuffle cost.

2 Move or query data

Choose **CDC** for ongoing database changes after full load; Transfer Family handles managed file protocols, while DataSync bulk-copies storage. Federation queries remote data without migrating it.

3 Commit governed lake tables

Choose **Iceberg** for ACID snapshots and schema evolution on S3. Optimistic concurrency checks can reject conflicting commits; writers retry. Partition pruning reduces scans; skew and small files raise cost.

4 Index similar vectors

Choose **HNSW** for graph-based approximate nearest neighbors at higher memory/build cost. IVF trains cluster centroids and searches selected lists; more probes improve recall but increase work. Support varies by engine/version.

Exam traps

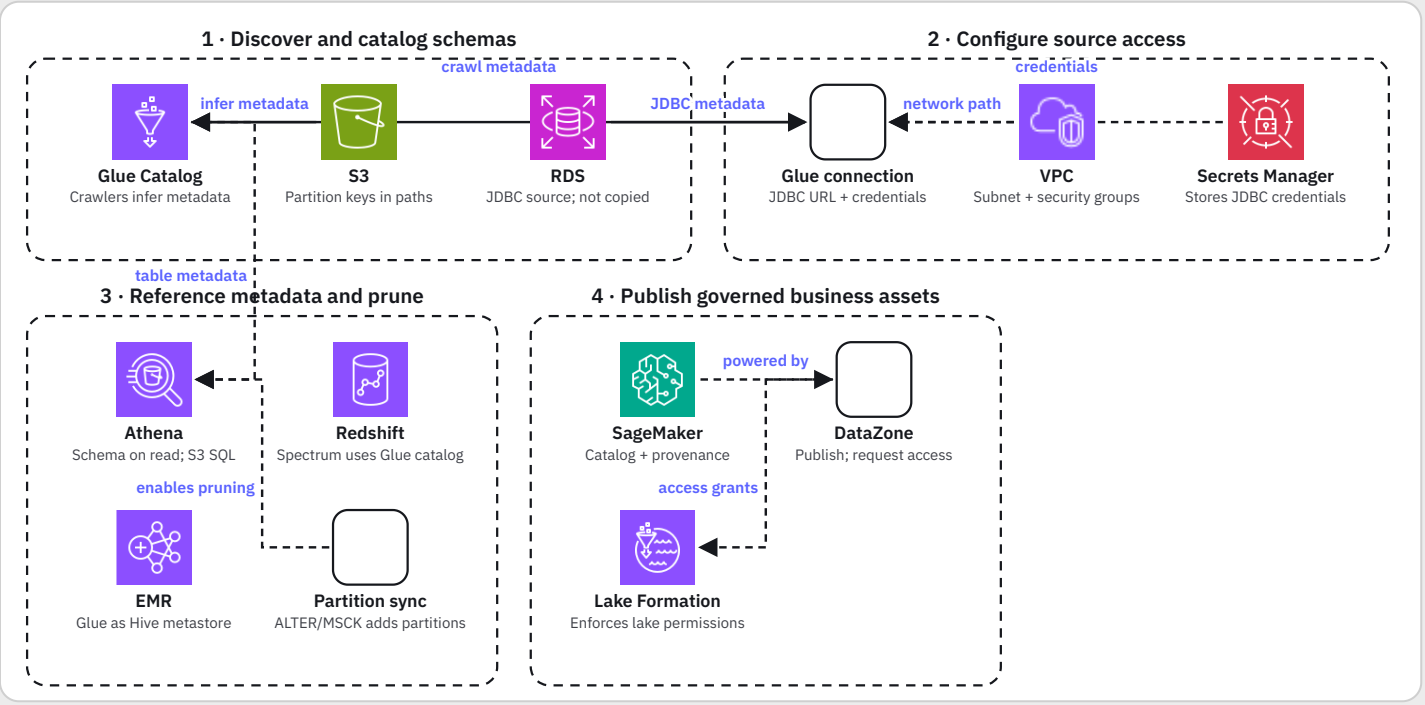
- **Federation is not materialization**: federation reads live remote databases; Spectrum reads S3; materialized views store query results and need refresh to reflect changes.
- **Locks are not permissions**: RDS/Redshift locks coordinate transactions; isolation controls visibility. Prevent deadlocks with consistent lock order and short transactions; retry aborted work.
- **Governance is not storage**: Lake Formation grants access; S3 stores lake objects. EMR supplies processing, not a durable warehouse. Iceberg supplies table semantics, not database-style row locks.

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

DOMAIN 2 · DATA STORE MANAGEMENT · TASK STATEMENT 2.2

Understand data cataloging systems

Discover, connect, prune, govern



1 Discover and catalog schemas

Crawlers infer schemas and register tables and partitions, not copies of source data. Schema evolution requires configured schema-change policies; discovery does not guarantee downstream compatibility.

2 Configure source access

Glue connections hold source or target access settings. Private JDBC access requires subnet routing, security groups, and an authorized Glue role; storing a secret does not establish network connectivity.

3 Reference metadata and prune

Partition pruning needs key filters and current partitions: crawler, ALTER TABLE ADD PARTITION, or MSCK REPAIR TABLE. Athena partition projection computes paths without registering partitions in Glue.

4 Publish governed business assets

Business catalogs add searchable descriptions, ownership, and provenance to technical assets. SageMaker Catalog uses DataZone capabilities; supported subscription workflows provision approved access through Lake Formation.

Exam traps

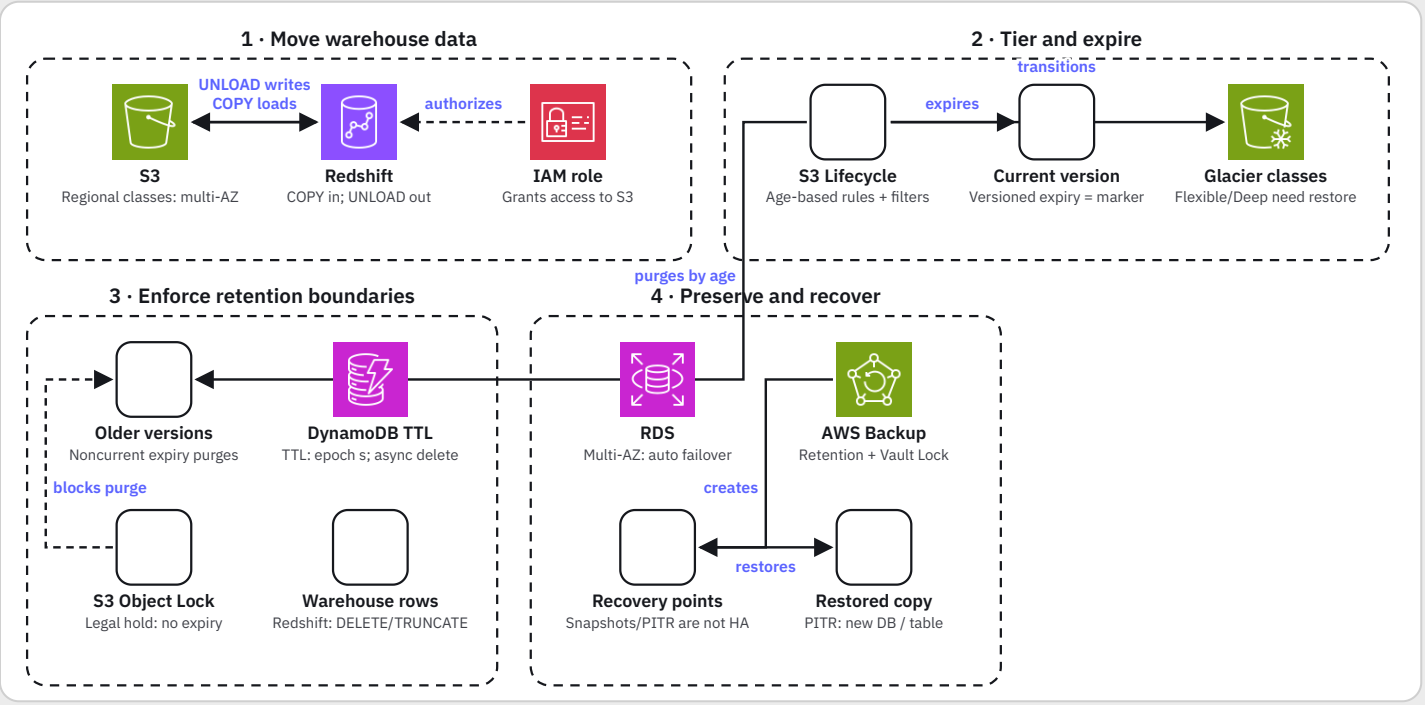
- **Schema on read** interprets source data using table metadata; changing a catalog column type does not rewrite files or repair incompatible values.
- **MSCK REPAIR TABLE** discovers Hive-style key=value partitions; it does not remove stale entries. ALTER TABLE ADD PARTITION supports explicit locations, including non-Hive layouts.
- **Catalog discovery is not authorization**: finding or publishing a business asset does not itself grant data access. Subscription approval and underlying permissions remain separate controls.

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

DOMAIN 2 · DATA STORE MANAGEMENT · TASK STATEMENT 2.3

Manage the lifecycle of data

Move, tier, delete, protect



1 Move warehouse data

Use **COPY / UNLOAD** for parallel bulk movement between S3 and Redshift. Attach an IAM role with the required S3 permissions; encrypted objects can also require KMS permissions.

2 Tier and expire

Choose **transitions** to lower storage cost, not to delete data. Account for retrieval costs and minimum storage durations. Expiration permanently deletes unversioned objects; versioned buckets need separate noncurrent rules.

3 Enforce retention boundaries

For **permanent erasure**, remove S3 versions and retained backup copies when controls permit. Redshift DELETE/TRUNCATE does not purge snapshots. Object Lock legal holds block protected-version deletion until removed.

4 Preserve and recover

Choose **Multi-AZ** for RDS or Redshift failover. RDS/DynamoDB PITR restores a new DB/table; Redshift snapshots restore a cluster. Backups support recovery from deletion; S3 One Zone classes lack multi-AZ redundancy.

Exam traps

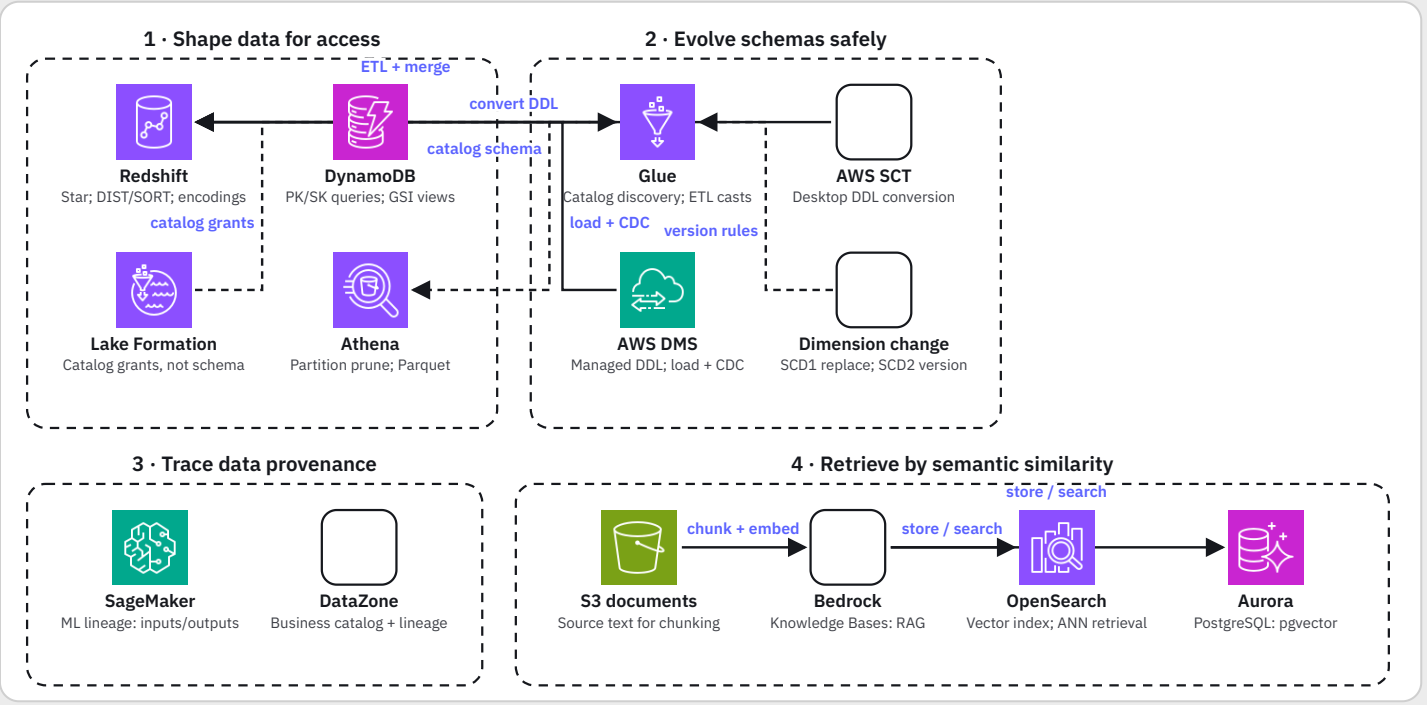
- **Suspending versioning** does not erase existing versions. A delete marker hides an object; permanent erasure requires deleting its retained versions.
- **DynamoDB TTL is asynchronous**, typically deleting expired items within a few days. Filter expired items from reads when they must disappear immediately; TTL is not a precise deletion deadline.
- **Durability is not availability**: preserved data can be temporarily inaccessible. Multi-AZ supports failover; backups support recovery. Neither replaces the other.

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

DOMAIN 2 · DATA STORE MANAGEMENT · TASK STATEMENT 2.4

Design data models and schema evolution

Model, evolve, trace, retrieve



1 Shape data for access

Choose a **star schema** for analytics; denormalize DynamoDB around access patterns. Redshift distribution can colocate joins; sort keys skip blocks. Diverse DynamoDB partition keys prevent hot partitions.

2 Evolve schemas safely

Nullable/defaulted additions support **backward compatibility**; validate type changes. SCT is desktop conversion; DMS Schema Conversion is managed. Unsupported objects need manual fixes; SCD2 keeps dated versions.

3 Trace data provenance

Use **data provenance** to trace sources, transformations, and downstream impact. SageMaker ML Lineage Tracking links artifacts and executions; DataZone accepts lineage events. Coverage depends on captured metadata.

4 Retrieve by semantic similarity

For **semantic search**, embed chunks and queries with a compatible model; retrieve nearby vectors, then supply text to generation. OpenSearch or Aurora PostgreSQL can store vectors; metadata filters constrain retrieval.

Exam traps

- **Crawler discovery is not migration**: updating catalog metadata does not rewrite old files or guarantee Athena compatibility. Renames, drops, and incompatible type changes require explicit handling.
- **Indexes are engine-specific**: Amazon RDS commonly uses B-tree indexes; Redshift uses distribution, sort keys, and encodings. DynamoDB GSIs add access paths but consume storage and write capacity.
- **RAG is not retraining**: retrieved passages become model context, not model weights. Changed documents require ingestion/index updates; vector similarity alone does not guarantee a correct answer.

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

DOMAIN 2 · DATA STORE MANAGEMENT · WHAT TO REMEMBER

Concepts and services of Data Store Management

Concepts to know

Data Partitioning

Partitioning divides data by keys such as date or tenant. **Match keys to access patterns** to enable pruning and parallel reads; avoid skew, hot keys, and excessive small partitions.

Schema Evolution

Schema evolution manages changes to fields, types, and table structure. **Preserve reader compatibility** when changing datasets; updating catalog metadata alone does not convert existing files.

Data Retention

Retention defines how long data remains available before deletion. **Distinguish expiration from permanent deletion**: object versions, backups, and legal holds can preserve data beyond its normal lifecycle.

Core services



Amazon Redshift

Use **Spectrum for S3 queries**, federated queries for supported operational databases, COPY for bulk ingestion, and materialized views for stored results that require refresh.



Amazon S3

Use **Lifecycle transitions for storage tiering** and expiration for retention. In versioned buckets, current-version expiration usually adds a delete marker; noncurrent versions need separate rules.



AWS Glue

Use **crawlers to discover schemas** and populate the Data Catalog. Catalog tables and partitions describe source data; cataloging does not copy it or enforce data quality.

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