AWS Certified

DEA-C01 · Data Engineer – Associate

Quick Reference Cheat Sheet

D1 · Ingestion & Transform 34% | D2 · Data Store Mgmt 26% | D3 · Operations & Support 22% | D4 · Security & Governance 18%

D1 · Data Ingestion & Transformation

Kinesis · Glue · MSK · DMS · AppFlow · EventBridge

34%

Kinesis Data Streams (KDS)

- **Write:** 1 MB/s or 1,000 records/s **per shard**
- **Read (std):** 2 MB/s shared across all consumers / shard
- **Enhanced Fan-Out:** 2 MB/s **per consumer** per shard — HTTP/2 push
- Ordering: **per shard only** — same partition key → same shard
- Retention: 24 h default (up to 365 d at extra cost)
- IteratorTypes: TRIM_HORIZON (oldest) · LATEST · AT_TIMESTAMP

■ *EFO = dedicated throughput; use when multiple apps consume same stream simultaneously*

Kinesis Data Firehose

- Fully managed — no shards, auto-scales, no consumer code
- **Near-real-time:** minimum 60-second buffer — NOT real-time
- Destinations: S3, Redshift, OpenSearch, Splunk, HTTP endpoint
- Redshift: writes to S3 first → issues COPY command
- Built-in Lambda transformation; cannot replay data
- **Exam trap:** Firehose ≠ real-time; use KDS for <1 s latency

Amazon MSK & MSK Connect

- Managed Apache Kafka — use when you have existing Kafka code
- MSK Connect: managed Kafka Connect workers (no EC2 management)
- **Source connector:** external system → Kafka topic (ingest)
- **Sink connector:** Kafka topic → S3 / OpenSearch / Redshift (deliver)
- MSK Serverless: no capacity planning; pay per usage
- vs KDS: choose MSK when Kafka ecosystem needed; KDS for AWS-native

■ *Kafka → S3 = SINK connector (not source)*

AWS Glue ETL

- **DynamicFrame:** flexible schema — handles extra/missing/evolving fields
- **DataFrame:** better perf with fixed schema; convert via toDF() / fromDF()
- **Job Bookmark:** tracks processed S3 objects; must be explicitly enabled
- **Relationalize:** flattens nested structs/arrays → relational tables
- **Trigger types:** On-Demand · Scheduled · Conditional · EventBridge
- **Glue Flex:** 34% cheaper; start can be delayed up to 30 min

■ *Bookmark not updated on failure — failed files reprocessed on next run*

Streaming Service Comparison

Feature			
Latency	Real-time (<1 s)	Near-real (60 s+)	Real-time
Management	Shards (manual scale)	Fully managed	Managed Kafka
Data Replay	✓ (retention period)	✗ — no replay	✓ (topic retention)
Consumers	Custom code / Lambda	Auto-delivers	Kafka consumers

Feature			
AWS-native	Deep (Lambda, KDA…)	Deep	Limited

AWS DMS — Migration

- **Full Load:** copy existing data while source keeps running
- **CDC:** capture ongoing changes from transaction logs
- Cutover when CDC lag ≈ 0 — minimal downtime
- SCT: converts schema for heterogeneous migrations (Oracle→Aurora)

AWS AppFlow

- 50+ SaaS connectors: Salesforce, Zendesk, Slack, ServiceNow
- No-code; scheduled or event-triggered flows
- Destinations: S3, Redshift, Snowflake
- Use when SaaS is a supported connector — no Lambda needed

DataSync vs Transfer Family

- **DataSync:** bulk NFS/SMB/HDFS → S3/EFS; agent-based on-prem
- **Transfer Family:** SFTP/FTP/FTPS endpoints backed by S3/EFS
- DataSync = move data (migration); Transfer = expose endpoint
- DataSync up to 10× faster than open-source cp tools

D2 · Data Store Management

Redshift · S3 · DynamoDB · Lake Formation · Timestream

26%

Redshift — Distribution & Sort

- **EVEN:** round-robin across nodes — no skew, use when no dominant join key
- **KEY:** same key → same node — co-locates joins; large fact tables
- **ALL:** full copy on every node — small, slowly-changing dimension tables only
- **AUTO:** Redshift chooses based on table size
- **COMPOUND sort key:** leading column most effective; best for filtering on prefix
- **INTERLEAVED sort key:** equal weight per col; needs VACUUM REINDEX

■ *Zone map pruning: Redshift skips 1 MB blocks using min/max in zone maps → sort key drives this*

Redshift COPY & Diagnostics

- COPY loads in parallel — 1 S3 file per compute slice for max throughput
- Split files into N = number of slices; use manifest for reliability
- Parquet format: no parsing overhead → fastest COPY
- COMPUPDATE OFF: skip compression analysis; ANALYZE after load
- WLM: queue slots + memory allocation per user/group
- **STL_LOAD_ERRORS:** row-level COPY failure details

■ *Redshift Spectrum: query S3 via Glue external tables — no loading needed. Data Sharing: live cross-cluster read*

DynamoDB — Keys, Indexes & Streams

- **GSI:** different PK + optional SK; own capacity; add any time; max 20
- **LSI:** same PK + different SK; shares table capacity; at creation only; max 5
- **Hot partition fix:** high-cardinality PK · write sharding · On-Demand mode
- **Adaptive Capacity:** auto-redistributes capacity to hot partitions
- **Streams:** item-level changes; 24 h retention; KEYS_ONLY / NEW_IMAGE / OLD_IMAGE / BOTH

■ *LSI cannot be added after table creation; DynamoDB Streams retention = 24 h (not configurable)*

S3 Storage Classes

Class			
Standard	ms	—	Hot data lake / active analytics
Standard-IA	ms	30 d	Infrequent access; retrieval fee
One Zone-IA	ms	30 d	Re-creatable data only (single AZ)
Glacier Instant	ms	90 d	Archive + ms retrieval (compliance)
Glacier Flexible	min–h	90 d	DR / backup archive
Deep Archive	12 h	180 d	Cheapest; rarely accessed
Intell-Tiering	auto	—	Unpredictable access patterns

Lake Formation — Fine-Grained Security

- **Column-level:** SELECT on specific columns only
- **Row-level:** data filter auto-applied per principal
- **Cell-level:** combined row + column filter
- **LF-Tags (ABAC):** tag resources; grant by sensitivity level
- **Governed Tables:** ACID transactions + time travel on S3
- Cross-account: share catalog via AWS RAM; LF enforces at query time

Amazon Timestream

- Serverless, purpose-built time-series database
- **Memory store** (hot/recent) → **Magnetic store** (cold, auto-tiered)
- Built-in: time_bin(), interpolate(), smooth()
- Integrates: Kinesis, IoT Core, Grafana, QuickSight
- Use for IoT telemetry, ops metrics — not for general OLTP

OpenSearch Index State Mgmt

- **ISM:** automates index lifecycle Hot→UltraWarm→Cold→Delete
- UltraWarm: S3-backed, read-only, lower cost than hot
- Cold: detached storage, lowest cost, rarely accessed
- Without ISM: stale indices bloat cluster → higher cost
- Object Lock Compliance: root cannot delete (SEC/FINRA)

D3 · Data Operations & Support

EMR · Athena · Step Functions · MWAA · QuickSight · Glue DQ

22%

Amazon EMR — Cluster Architecture

- **Master node:** YARN ResourceManager + HDFS NameNode — ALWAYS On-Demand
- **Core nodes:** compute + HDFS data — On-Demand recommended (Spot = HDFS loss risk)
- **Task nodes:** compute only, no HDFS — SAFE for Spot (up to 90% savings)
- **EMRFS:** store I/O in S3 (durable); HDFS for temp shuffle only
- **EMR Serverless:** no cluster management — pay per vCPU-hour + GB-hour
- **S3DistCp:** compact small Firehose output files → large files before Spark reads

■ *Core Spot = HDFS data loss risk if reclaimed mid-job; Task Spot = safe retry*

Amazon Athena — Cost & Performance

- Charged **$5 per TB scanned** — every optimization is direct cost savings
- **Parquet/ORC:** columnar → read only needed cols (90%+ scan reduction)
- **Partition pruning:** WHERE on partition cols skips non-matching partitions
- **Partition Projection:** no crawler for predictable time-series partitions
- **File sizing:** 128 MB–1 GB ideal; small files = query slowdown
- **Result cache:** identical repeated queries → zero scan cost

■ *Athena Federated Query: Lambda connector → query RDS, DynamoDB, on-prem via SQL*

Spark AQE (Adaptive Query Execution)

- Runtime re-optimization using actual data statistics
- **Coalesce partitions:** merge small post-shuffle → fewer large
- **Broadcast join:** auto-convert sort-merge → broadcast if side is small
- **Skew join:** detect and split hot partition into sub-tasks
- Default ON Spark 3.x; set spark.sql.adaptive.enabled=true

Step Functions

- **Standard:** exactly-once, up to 1 year, full execution history
- **Express:** high-volume, up to 5 min, cheaper per transition
- **Distributed Map:** 10K concurrent child executions (S3/SQS items)
- Use Standard for critical pipelines; Express for event-driven
- Integrates with Glue, EMR, Lambda, Redshift Data API

Amazon MWAA (Managed Airflow)

- Managed Apache Airflow — use for existing Python DAG code
- Sensors: S3KeySensor, ExternalTaskSensor (wait for external events)
- vs Step Functions: MWAA = Airflow expertise; SFN = JSON state machine
- MWAA for complex dependency graphs + long external waits
- Auto-scales workers based on task queue depth

QuickSight SPICE

- **SPICE:** in-memory engine — sub-second queries without hitting source
- Scheduled refresh: hourly to weekly; or manual import
- SPICE full → purchase more capacity or delete unused datasets
- Direct Query mode: always fresh; higher latency per query
- ML Insights: anomaly detection + forecasting built-in

Glue Data Quality (DQDL)

- Define rules: Completeness, Uniqueness, ColumnValues, RowCount
- On failure: raise exception to halt pipeline (prevents bad data)
- Integrates with CloudWatch for DQ metric dashboards
- Can run as post-action in Glue Catalog crawlers
- **DQDL example:** Completeness "col" > 0.95 · Uniqueness "id" = 1.0

Redshift Diagnostics & Ops

- **STL_LOAD_ERRORS:** row-by-row COPY failure details
- **STV_LOCKS:** real-time table lock + transaction info
- **STL_QUERY:** historical query execution log
- **SVL_QUERY_SUMMARY:** per-step execution stats
- STL = persisted logs; STV = real-time in-memory snapshots

D4 · Data Security & Governance

KMS · IAM · Lake Formation · Macie · VPC Endpoints · Compliance

18%

S3 Encryption & KMS

- **SSE-S3:** AWS-managed AES-256; no KMS cost; no CloudTrail key logs
- **SSE-KMS:** CMK; CloudTrail logs every key use; needs kms:GenerateDataKey + kms:Decrypt
- **SSE-C:** customer provides key per request; HTTPS required; key not stored by AWS
- **CSE:** encrypt before upload — AWS never sees plaintext
- **S3 Bucket Key:** reduces KMS API calls ~99% — always enable at scale
- **Object Lock Compliance:** root cannot delete; for SEC/FINRA regulatory data
- ■ *Object Lock Governance mode: s3:BypassGovernanceRetention permission can override*

IAM, VPC & Network Security

- Glue role: s3:GetObject (src) + s3:PutObject (dst) + kms:Decrypt + kms:GenerateDataKey
- Least privilege: grant specific S3 prefixes — never s3:*
- **VPC Gateway endpoints (free):** S3, DynamoDB — required for private subnets
- **VPC Interface endpoints ($):** Glue, KMS, STS, CloudWatch Logs
- aws:SecureTransport=false → DENY bucket policy blocks HTTP access
- Redshift cluster: encrypt in transit (SSL=true); in-rest KMS encryption

Amazon Macie — PII Detection

- ML-based PII discovery in S3 (SSN, credit cards, credentials)
- Finding → EventBridge → Lambda → quarantine/notify
- **No built-in remediation** — all response via EventBridge + Lambda
- Custom identifiers: org-specific regex patterns for PII
- Multi-account: centralized Macie admin via AWS Organizations

Lake Formation Permissions

- **Column-level:** invisible columns for restricted users in Athena
- **Row-level:** data filter auto-applied per user/role
- **LF-Tags (ABAC):** tag by sensitivity; grant by tag — scales to 1000s of tables
- Cross-account: Producer shares via RAM; consumer queries via Athena
- LF governs — S3 bucket policy alone cannot enforce column/row security

Audit, Compliance & Config

- CloudTrail data events: off by default; enable for s3:GetObject audit trail
- Query CloudTrail logs in S3 with Athena — cost-effective at scale
- **AWS Config rules:** s3-bucket-public-access-prohibited · s3-bucket-ssl-requests-only
- NON_COMPLIANT → auto-remediation via SSM Automation runbook
- **Redshift DDM:** mask at query time; DDM = masked visible; CLS = hidden

Master Quick Reference — Key Facts & Exam Decision Rules

Service / Concept		
KDS — Throughput	Write 1 MB/s or 1K rec/s per shard; Read 2 MB/s shared; EFO: 2 MB/s per consumer	EFO = dedicated per consumer — use when multiple apps consume same stream concurrently
Kinesis Firehose	Fully managed; min 60 s buffer; destinations: S3, Redshift, OpenSearch, Splunk	NOT real-time (60 s+); Redshift loads via S3 COPY; no replay capability
Glue Job Bookmark	Tracks processed S3 objects for incremental loads; must be explicitly enabled	Bookmark NOT updated on failure — failed files reprocessed on next run
MSK Connect	Managed Kafka Connect; source = ingest to Kafka; sink = deliver from Kafka	Kafka → S3 = SINK connector (not source)
File Formats	Parquet/ORC: columnar, best for Athena/Spark; Snappy: fast; GZIP: smaller	CSV → Parquet reduces Athena scan cost by 90%+; always use Parquet for analytics
AWS DMS + CDC	Full Load (existing) + CDC (ongoing) = near-zero downtime cutover	SCT required for heterogeneous migrations (Oracle → Aurora etc.)
Redshift Dist Style	EVEN: no skew; KEY: co-locate joins; ALL: small dims; AUTO: Redshift chooses	KEY = large fact tables with frequent joins on same column
Redshift COPY	1 file per slice = max parallelism; Parquet = fastest; COMPUPDATE OFF	STL_LOAD_ERRORS for debugging; run ANALYZE after every COPY load
DynamoDB GSI vs LSI	GSI: any PK, own capacity, add any time; LSI: same PK, table capacity, at creation only	LSI cannot be added after table creation; max 5 LSIs, 20 GSIs per table
Lake Formation	Column/row/cell security; LF-Tags (ABAC); Governed Tables = ACID on S3	S3 bucket policy alone cannot enforce column/row security — need LF + Glue catalog
S3 Object Lock	Compliance: nobody can delete (even root); Governance: bypass permission exists	Requires bucket versioning; use Compliance for SEC/FINRA regulatory data
EMR Node Types	Master: always On-Demand; Core: On-Demand (has HDFS); Task: safe for Spot	Core Spot = HDFS data loss if instance reclaimed mid-job
Athena Cost	$5/TB scanned; Parquet+partitions = 90%+ scan reduction; result cache = 0 scan	Partition Projection eliminates crawler for predictable time-series partitions

Service / Concept		
Step Functions	Standard: exactly-once, 1 yr; Express: high-volume, 5 min, cheaper per transition	Distributed Map: 10K concurrent child executions for S3/SQS fan-out workloads
Glue Flex Execution	34% cheaper; start can be delayed up to 30 min (spare capacity)	NOT for SLA-bound pipelines; use for nightly/weekend batch jobs
SSE-KMS Permissions	Write: kms:GenerateDataKey; Read: kms:Decrypt — both required on CMK	S3 Bucket Key reduces KMS API calls ~99% — enable for cost at scale
Amazon Macie	ML-based PII detection in S3; findings published to EventBridge	NO built-in remediation — wire EventBridge → Lambda for quarantine/notify
CloudTrail Data Events	S3 object-level ops (Get/Put/Delete); OFF by default; extra cost to enable	Enable for compliance; query with Athena for cost-effective log analysis
VPC Endpoints	Gateway (free): S3, DynamoDB; Interface ($): Glue, KMS, STS, CloudWatch	Always need S3 Gateway endpoint for private-subnet Glue jobs; S3GW is free
aws:SecureTransport	Condition key false = HTTP request (not HTTPS); use in bucket policy DENY	Blocks unencrypted access: {Bool: {aws:SecureTransport: false}} in Condition