

AWS Certified

Data Analytics – Specialty

DAS-C01 · QUICK REFERENCE CHEAT SHEET

18%

Domain 1 Collection

22%Domain 2 Storage & Data
Mgmt**24%**

Domain 3 Processing

18%Domain 4 Analysis &
Visualization**18%**

Domain 5 Data Security

65 Questions

50 scored + 15 unscored

170 Minutes

Exam duration

750 / 1000

Minimum passing score

Multiple Choice

+ Multiple Response

No Penalty

For unanswered questions

Scaled Score

100–1000 range

DOMAIN 1 Collection
18%

Kinesis Data Streams (KDS) vs Kinesis Data Firehose

Feature	Kinesis Data Streams	Kinesis Data Firehose
Shard-based throughput	1 MB/s in, 2 MB/s out per shard	Auto-scales, no shards
Retention	24 hrs default (up to 365 days)	No retention — delivers & forgets
Consumers	Custom (KCL, Lambda, KDA)	S3, Redshift, OpenSearch, Splunk, HTTP
Latency	Real-time (~70 ms)	Near real-time (60s or 1 MB buffer)
Use case	Custom processing, replay needed	Managed delivery, no consumer code
Ordering	Per-shard ordering guaranteed	No ordering guarantee

Kinesis Data Analytics (KDA) / Apache Flink

- Real-time SQL / Flink on streaming data from KDS or Firehose
- Studio notebooks: interactive Flink apps with Apache Zeppelin
- Auto-scaling, no infrastructure to manage
- Use for: sliding windows, tumbling windows, sessionization, anomaly detection

MSK (Amazon Managed Streaming for Apache Kafka)

Component	Detail
Broker nodes	Managed Kafka brokers; you choose instance type & AZs
ZooKeeper	Managed ZooKeeper cluster included
Encryption	TLS in-transit, KMS at-rest; MSK Serverless also available
MSK Connect	Run Kafka Connect connectors (Debezium, S3 sink) as managed service
vs KDS	Use MSK when: existing Kafka clients, topic partitions > 200, need Kafka ecosystem

AWS IoT Core & Data Transfer

Service	Key Points
IoT Core	MQTT/HTTP ingestion; Rules Engine routes to KDS, S3, Lambda, DynamoDB
IoT Greengrass	Edge compute; run Lambda/ML models on IoT devices locally
AWS DMS	Migrate relational DBs to cloud; CDC mode for ongoing replication
AWS DataSync	NFS/SMB/HDFS/S3 to S3/EFS/FSx data transfer; up to 10x faster than open-source
Snow Family	Snowcone (8 TB), Snowball Edge (80 TB), Snowmobile (100 PB) for offline ingestion
Direct Connect	Dedicated private link; stable throughput for bulk historical loads

DOMAIN 2 Storage & Data Management
22%

Amazon S3 — Data Lake Foundation

Class	Key Facts
S3 Standard	Frequent access; 11 nines durability; millisecond retrieval
S3 Standard-IA	Infrequent access; lower storage, retrieval fee; great for staging
S3 One Zone-IA	Single AZ; 20% cheaper than Standard-IA; re-createable data
S3 Intelligent-Tier	Auto-moves between access tiers; no retrieval fees; monitoring cost
S3 Glacier Instant	Archive; ms retrieval; same latency as Standard-IA
S3 Glacier Flexible	1-12 hr retrieval; lowest archive cost for bulk queries
S3 Glacier Deep	12-48 hr; cheapest tier; compliance/long-term retention

AWS Lake Formation & Glue Data Catalog

Component	Key Points
Lake Formation	Centralized data lake setup; fine-grained column/row-level security
Glue Data Catalog	Hive-compatible metastore; stores table schemas, partitions, locations
Glue Crawlers	Auto-discover schemas from S3, RDS, DynamoDB; update catalog
LF Permissions	Grant/revoke on databases, tables, columns; integrates with IAM + Athena + Redshift Spectrum
Governed Tables	ACID transactions on S3; row-level security; automatic compaction
LF Tags (LFT)	Attribute-based access control; tag resources then grant by tag

Amazon Redshift

Feature	Detail
Architecture	Leader node + compute nodes (MPP); columnar storage; zone maps
Distribution styles	EVEN (default), KEY (join col), ALL (small dims), AUTO
Sort keys	Compound (range scans), Interleaved (multi-col predicates)
Redshift Spectrum	Query S3 directly from Redshift without loading; share Glue catalog
RA3 nodes	Separate compute from storage; managed storage on S3-backed SSD
Concurrency Scaling	Add read capacity automatically; charged per second of use
AQUA	Advanced Query Accelerator — hardware-accelerated cache layer
Data Sharing	Share live data across Redshift clusters without copying

Amazon DynamoDB

Feature	Detail
Keys	Partition key (hash) + optional Sort key; composite key for range queries

Read modes	Eventually consistent (default), Strongly consistent, Transactional
GSI / LSI	GSI: different partition key; created anytime. LSI: same PK diff sort; at creation only
DynamoDB Streams	Change capture; 24hr retention; triggers Lambda for ETL/analytics
DAX	In-memory cache; microsecond reads; write-through; for read-heavy workloads
Export to S3	Point-in-time export to S3 without consuming RCUs; use Athena to query

DOMAIN 3 Processing
24%

AWS Glue ETL vs Amazon EMR

Feature	AWS Glue	Amazon EMR
Type	Serverless managed ETL	Managed Hadoop/Spark cluster
Runtime	Apache Spark (PySpark/Scala)	Spark, Hive, Presto, Flink, HBase
Management	No infra; auto-scale workers	You provision/manage cluster
Cost model	DPU-hours (Data Processing Units)	EC2 instance-hours
Catalog	Native Glue Data Catalog	Can use Glue catalog or own Hive
Streaming	Glue Streaming ETL (KDS/Kafka)	Spark Structured Streaming
Use when	Straightforward ETL, serverless preferred	Complex ML, custom Hadoop, cost at scale

EMR Cluster Architecture

Component	Key Points
Master/Primary	Coordinates cluster; runs resource manager; single point of failure
Core nodes	Run tasks AND store HDFS data; losing one affects HDFS
Task nodes	Run tasks only; no HDFS; safe to use Spot; scale in/out freely
Instance fleets	Mix On-Demand + Spot across instance types per fleet
EMR on EKS	Submit Spark jobs to EKS clusters; share infra with other workloads
EMR Serverless	No cluster management; per vCPU-hour billing; for intermittent jobs

AWS Lambda & Step Functions

Topic	Key Points
Lambda limits	15 min timeout, 10 GB RAM, 10 GB /tmp; 250 MB deployment package
Lambda + KDS	Event source mapping; bisect-on-error; parallelization factor (up to 10)
Lambda + S3	Trigger on s3:ObjectCreated; fan-out via SNS for multi-consumer
Step Functions	Visual workflow; Standard (audit log, 1yr) vs Express (5min, high-volume)
SF States	Task, Choice, Wait, Parallel, Map (inline or distributed), Pass, Fail, Succeed
SF + Glue	Orchestrate multi-step ETL pipelines; retry/error handling built in
EventBridge	Cron/rate schedules + event patterns to trigger Glue, Lambda, Step Fn

Data Formats & File Optimization

Format/Technique	Key Facts
Parquet	Columnar; best for Athena/Spectrum/Redshift Spectrum; predicate push-down

ORC	Columnar; optimized for Hive; lightweight indexes; good Hive/Spark choice
Avro	Row-based; best for write-heavy streaming (Kafka schema registry)
JSON	Human-readable; verbose; use for APIs; convert to Parquet for analytics
CSV/TSV	Simple; no schema; no compression metadata; poor for analytics at scale
Snappy	Fast compression; splittable with Parquet/ORC; lower ratio than GZIP
GZIP	Higher compression ratio; NOT splittable by default (EMR handles with splits)
Partitioning	Partition by year/month/day on S3 to prune files; reduces scanned data

DOMAIN 4 Analysis & Visualization**18%****Amazon Athena**

Feature	Detail
Engine	Presto/Trino under the hood; serverless; query S3 directly
Pricing	\$5 per TB scanned; use Parquet/ORC + partitioning to cut cost up to 90%
Workgroups	Isolate queries; set data-scanned limits; per-team cost tracking
Federated	Query RDS, DynamoDB, Redshift, on-prem via Lambda Data Source Connectors
CTAS	CREATE TABLE AS SELECT — saves results back to S3 as Parquet/ORC
Views	Logical views; no data stored; reuse complex SQL without duplication
Lake Formation	Fine-grained column/row security enforced in Athena queries
Performance	Partition projection; columnar format; coalesce small files with CTAS

Amazon QuickSight

Feature	Detail
SPICE	Super-fast Parallel In-memory Calculation Engine; cached dataset; faster queries
Data sources	Athena, Redshift, S3, RDS, Aurora, Salesforce, on-prem via VPC connector
Editions	Standard (per-user), Enterprise (AD integration, row-level security, encryption)
Row-level sec.	Upload rules file or use RLS dataset to restrict user access to rows
ML Insights	Anomaly detection, forecasting, auto-narrative via ML (no ML skill needed)
Embedded	Embed dashboards in apps via 1-click, Q embedding, or API
Q (NL queries)	Natural language Q&A; across your data (Enterprise add-on)

Amazon OpenSearch Service

Feature	Detail
Use cases	Log analytics, full-text search, real-time dashboards (Kibana/OpenSearch Dashboards)
Ingestion	Firehose (native), Logstash, Lambda, custom clients
Index tuning	Shard count = total data / 50 GB; avoid too many shards (overhead)
UltraWarm	S3-backed low-cost tier for infrequent indices; query via same API
Cold storage	Detach indices to S3; reattach when needed; lowest cost for old logs
ISM policies	Index State Management: auto-rollover, close, delete indices by age/size

Redshift Spectrum & Analytics Patterns

Pattern	Detail
---------	--------

Redshift Spectrum	Query S3 from Redshift; no data load; use Parquet for best perf
External tables	CREATE EXTERNAL TABLE references S3; partition on load date/region
QuickSight + Athena	Common BI pattern: S3 data lake → Athena → QuickSight dashboard
QuickSight + Redshift	Analytical DW pattern: Redshift SPICE import or Direct Query
Lambda + DynamoDB	Operational analytics: stream → Lambda → DynamoDB → DAX for APIs
EMR + Spark ML	Large-scale ML feature engineering → write features to S3 → SageMaker

DOMAIN 5 Data Security
18%

Encryption — At Rest & In Transit

Option	Key Points
S3 SSE-S3	Server-side encryption with AWS-managed S3 keys (AES-256); simplest
S3 SSE-KMS	Customer manages KMS CMK; audit key usage in CloudTrail; default recommended
S3 SSE-C	Customer provides key per request; AWS does NOT store the key
S3 CSE	Client-side encryption before upload; full control; highest complexity
KDS encryption	KMS CMK on Kinesis streams; per-shard envelope encryption
Redshift	AES-256 at rest; SSL/TLS in transit; KMS or CloudHSM for master keys
EMR	LUKS/EBS encryption; S3CSE or S3SSE for EMRFS; in-transit TLS
In-transit	TLS 1.2+ for all API calls; enforce via bucket policy aws:SecureTransport

AWS KMS & IAM for Analytics

Topic	Key Points
CMK types	AWS-managed (free, auto-rotate), Customer-managed (more control, \$1/mo)
Envelope enc.	Generate Data Key (DEK); encrypt data with DEK; encrypt DEK with CMK
Key policies	Resource-based policy on CMK; must grant kms:Decrypt to consumers
Least privilege	Separate IAM roles: Glue executor, Athena query, Redshift cluster role
Service roles	Glue: AWSGlueServiceRole; EMR: EMR_EC2_DefaultRole + EMR_DefaultRole
Conditions	Use s3:prefix, kms:ViaService, kms:CallerAccount for fine-grained access

Lake Formation Security

Feature	Detail
Column-level	Grant SELECT on specific columns; omitted columns hidden from query
Row-level	LF row filters; condition expressions like WHERE dept = session_dept
Cell-level	Combine column + row filters for cell-level (Enterprise feature)
Data location	LF manages S3 access; remove direct S3 permissions; use LF location grants
Cross-account	RAM share Glue catalog; grant LF permissions to external account principal

Monitoring, Auditing & Network Controls

Control	Key Points
CloudTrail	API audit log; data events for S3/Lambda (extra cost); 90-day default
Amazon Macie	ML-based PII/sensitive data discovery in S3; generate findings for remediation

VPC Endpoints	Gateway endpoints: S3, DynamoDB (free). Interface endpoints: KMS, Glue, Kinesis
PrivateLink	Keep analytics traffic on AWS backbone; no internet exposure
Security Groups	Stateful; apply to Redshift, EMR, Lambda in VPC; use SG-referencing rules
S3 Bucket Policy	Enforce https (aws:SecureTransport), restrict to VPC endpoint, deny non-encrypted uploads
Config Rules	Track compliance: S3 public access, KMS key rotation, CloudTrail enabled

MASTER SERVICE QUICK REFERENCE — MOST TESTED

Know These Cold

Service	Category	Remember This
Kinesis Data Streams	Real-time streaming	Shards, 24hr retention, replay, custom consumers
Kinesis Firehose	Managed delivery	S3/Redshift/OpenSearch; buffer 60s or 1 MB; no code
Kinesis Analytics	Real-time SQL/Flink	Sliding/tumbling windows; anomaly detection
MSK	Managed Kafka	Kafka ecosystem; use when Kafka clients needed
AWS Glue	Serverless ETL	PySpark; Glue Catalog; Crawlers; DPU billing
Amazon EMR	Hadoop/Spark cluster	Core=HDFS; Task=Spot ok; EMR Serverless available
Amazon Athena	Serverless SQL on S3	\$5/TB; Parquet+partition saves 90%; Federated
Redshift	Data warehouse	MPP columnar; RA3; Spectrum; Concurrency Scaling
Amazon S3	Data lake storage	11 nines; partitioning; lifecycle to Glacier
Lake Formation	Data lake governance	Column/row-level security; Glue catalog; LFT
QuickSight	BI dashboards	SPICE cache; row-level sec; ML insights; embed
OpenSearch	Search & log analytics	Firehose ingest; UltraWarm; Cold storage; ISM
DynamoDB	Serverless NoSQL	GSI; Streams→Lambda; DAX; Export to S3
AWS Lambda	Event processing	15 min; KDS source mapping; S3 trigger; Step Fn
Step Functions	Workflow orchestration	Standard vs Express; Map/Parallel states; Glue ETL
AWS Glue DataBrew	Visual data prep	No-code transforms; 250+ transforms; profile data
Amazon Macie	S3 data privacy	PII detection; ML-based; findings to Security Hub
AWS KMS	Key management	Envelope encryption; CMK audit; kms:Decrypt grants
VPC Endpoints	Private access	Gateway: S3/DynamoDB; Interface: KMS/Glue/Kinesis
CloudTrail	API audit log	Who/what/when; data events for S3; 90-day free

EXAM DAY REMINDERS

750/1000 to pass | 65 questions (50 scored + 15 unscored) | 170 minutes | No penalty for guessing — answer every question!
 Firehose = managed delivery (no replay) | KDS = replay + custom consumers | Parquet/ORC + partitioning = cheapest Athena queries
 Glue = serverless ETL | EMR = full cluster control | Lake Formation = column/row-level security on Glue catalog
 RA3 = separate compute + storage | Redshift Spectrum = query S3 from Redshift | SPICE = QuickSight in-memory cache
 SSE-KMS = audit key usage in CloudTrail | VPC Gateway Endpoint for S3 is FREE | Macie = PII in S3