

AWS Certified Data Engineer – Associate

DEA-C01

34%	26%	22%	18%
-----	-----	-----	-----

1. (Ingestion)

Kinesis Data Streams (KDS)	1MB/s · 1,000 rec/s, 2MB/s() / Enhanced Fan - Out 2MB/s. 24h~365 가 On - demand	1 (replay)
Kinesis Data Firehose	(1~128MB / 60~900) : S3, Redshift, OpenSearch, Splunk, HTTP. Lambda JSON → Parquet/ORC 60	S3/DW
Amazon MSK	Apache Kafka. MSK Connect, MSK Serverless. Glue	Kafka ()가
DynamoDB Streams	(24h). Lambda Kinesis Data Streams for DynamoDB.	DynamoDB CDC,
Amazon SQS	Standard(1 ,) / FIFO(1 ,). DLQ.	()

Kinesis Data Streams

- : ProvisionedThroughputExceededException → 가(re-sharding: split),
- : GetRecords.IteratorAge(Milliseconds) 가 가 → , EFO
- : → /
- : KPL(), Kinesis Agent, SDK PutRecord/PutRecords. : KCL(DynamoDB), Lambda(), Firehose, Managed Service for Apache Flink.
- Lambda : (shard 10), bisect · DLQ(SQS/SNS)
- : KMS(SSE), HTTPS.
- CDC
- AWS DMS: / DB + CDC(). Full load → CDC. S3 · Kinesis · Redshift · OpenSearch 가 → CDC . DMS Serverless. ().
- AWS SCT: / (Oracle → Aurora PostgreSQL), 가 . SCT data extraction agent.
- AWS AppFlow: SaaS(Salesforce, SAP, Zendesk...) → AWS , , /
- AWS Transfer Family: SFTP/FTPS/FTP/AS2 → S3 · EFS
- S3 : S3 Event Notifications → SNS/SQS/Lambda/EventBridge. EventBridge
- Redshift : COPY(), Auto-copy from S3, Streaming ingestion(Kinesis/MSK), Zero-ETL(Aurora).

2. (Transform)

	/
AWS Glue (Spark ETL)	Spark. ETL, (1).
Amazon EMR	Hadoop/Spark, Hive/Presto/HBase/Flink/Hudi/Iceberg, Spot
AWS Lambda	(≤ 15) , /
Amazon Redshift (ELT)	SQL (, MERGE). DW , Spectrum S3
Glue DataBrew	250+ 가 ,
Managed Service for Apache Flink	SQL/Java,

AWS Glue

- : Spark(), Spark Streaming(Kinesis/MSK micro - batch), Python Shell(, pandas), Ray().
- : G.1X(4 vCPU/16GB), G.2X, G.4X, G.8X, G.025X(). DPU - (, 1). Auto Scaling, Flex().
- Job Bookmarks: → (). bookmark reset.
- DynamicFrame: (), ResolveChoice , relationalize, `toDF()` Spark DataFrame .
- : , push - down predicate(`push\_down\_predicate`), , (groupFiles/groupSize), .
- : → . / . classifier(Grok/JSON/XML/CSV). (/), .
- Glue Workflows: + DAG, (on - demand/scheduled/event(EventBridge)/conditional).
- Glue Data Quality: DQDL , ETL , → CloudWatch/EventBridge, ,
- Glue (Connection): VPC JDBC(RDS) , Secrets Manager .

Amazon EMR

- : Primary(), Core(HDFS +), Task(, HDFS → Spot).
- : Instance Fleets(Spot ,) vs Instance Groups().
- : HDFS(, =), EMRFS(S3 = ,). S3 (EMRFS consistent view).
- : Transient(, ↓) vs Long - running(). EMR Serverless(), EMR on EKS.
- Managed Scaling(/), Spot .
- : Apache Hudi / Iceberg / Delta (upsert, time travel,).

Spark

- . (`spark.sql.shuffle.partitions`), (), → salting.
- → coalesce/repartition , + .
- /persist , predicate/projection pushdown, .

3.

	/
AWS Step Functions	(ASL JSON). Standard(1 , 1 ,) vs Express(5 , 1). Catch · Parallel · Map(). Glue/EMR/ECS/Lambda/SNS 가 .
Amazon MWAA (Managed Airflow)	Python DAG , Airflow .
AWS Glue Workflows	Glue DAG. Glue .
Amazon EventBridge	: () → Scheduler(cron/rate,). Pipes(→ → →).
Amazon SNS / SQS	SNS: (/HTTP/SQS/Lambda). SQS: . DLQ.

- : S3 → EventBridge → Step Functions → Glue → → .
- : EventBridge Scheduler / Airflow cron / Glue .
- : , DLQ, (SNS).

4.

Parquet		, predicate pushdown, splittable. Athena/Spectrum/Spark .
ORC		Hive , , .
Avro		, / .
JSON / CSV	() 가 ↑, (). CSV .	

- : Snappy(, Parquet , splittable) · gzip(↑, -splittable) · bzip2(splittable,) · Zstd() · LZO.
- Athena/Spectrum = (Parquet) + + + (≈ 128MB~1GB).
- : S3 `year=2024/month=01/day=05/` (Hive). 가
- Athena Partition Projection: → .
- : → ↑. : compaction , Firehose , Glue groupFiles, Hudi/Iceberg .
- : Glue , Athena CTAS / `INSERT INTO`, Firehose , EMR.

5. S3

	Amazon S3 (+ Glue Catalog / Lake Formation)
DW · BI	Amazon Redshift (RA3 / Serverless)
S3 SQL ()	Amazon Athena
- / TPS	Amazon DynamoDB (+ DAX)
OLTP	Amazon RDS / Aurora
	Kinesis Data Streams / MSK
	Amazon OpenSearch Service
	ElastiCache / MemoryDB / DAX
/ /	Neptune / Timestream / QLDB

Amazon S3

- : Standard → Standard - IA / One Zone - IA () → Intelligent - Tiering(,) → Glacier Instant / Flexible / Deep Archive(가 ,).
- Lifecycle : N , M (), , .
- + MFA Delete, Object Lock(WORM: /) = . .
- S3 Select: SQL . Storage Lens: . 가 .
- Cross - Region Replication(CRR)/SRR: DR . (RTC).
- : prefix , - (2020~). , 가 .

DynamoDB TTL / S3 , Redshift ↔ S3 COPY/UNLOAD .

6. Amazon Redshift

- : RA3(= ,) / DC2(SSD,) / Serverless(RPU ,). MPP + + .
- (DISTSTYLE): KEY(→ -) · ALL(→) · EVEN() · AUTO(, Redshift가).
- (SORTKEY): Compound(, ·) / Interleaved(가 , ↑).
- COPY: S3/EMR/DynamoDB/SSH , gzip/bzip2/zstd , , COMPUPDATE/STATUPDATE.
- UNLOAD: S3 (Parquet , 가).
- WLM: . Auto WLM() vs Manual. SQA(), Concurrency Scaling(가), QMR().
- Spectrum: S3 (Glue Catalog), , Redshift .
- Federated Query: RDS/Aurora(PostgreSQL · MySQL) . Materialized View: , .
- Data Sharing(RA3): . AWS Data Exchange for Redshift.
- : VACUUM(,), ANALYZE() — .
- : Zero - ETL(Aurora→Redshift), Streaming Ingestion(Kinesis/MSK), Auto - copy(S3), Redshift ML(SageMaker), , (RLS).
- : KMS , VPC · VPC , IAM · IAM Identity Center, → S3/CloudWatch, STL/SVL/SVV .

7. Amazon DynamoDB

- **Capacity Mode**: () + (). 가 .
- **Capacity Mode**: On - demand(가 .) vs Provisioned(+ Auto Scaling, 가 . ↓).
RCU=4KB / (=2), WCU=1KB / .
- **Capacity Mode** / : 3,000 RCU · 1,000 WCU . : , write sharding(), / , adaptive capacity .
- **GSI**: . , , 20 . 가 .
- **LSI**: . , , 가 , 10GB .
- **DAX**: (, read - through/write - through).
- **TTL**: (, 48h). Streams .
- **Streams** → Lambda / Kinesis Data Streams for DynamoDB(.).
- **Global Tables**: - (). PITR(35), .
- **S3** : PITR , → Athena/Glue . Import from S3.
- : , , GSI, . (TransactWrite/Read).

8. Athena · Glue Data Catalog · Lake Formation

Amazon Athena

- Trino/Presto(SQL) + Spark. S3 Glue Catalog
- : TB (\$5). Athena for Spark DPU
- Workgroup: / , , CloudWatch
- CTAS / INSERT INTO: () → ETL. UNLOAD
- Partition Projection, Query result reuse(), Prepared statements, (Lambda).
- Federated Query: Lambda DynamoDB · RDS · Redshift · CloudWatch ·
- ACID · 가 Apache Iceberg

AWS Glue Data Catalog

- (Hive). Athena · EMR · Redshift Spectrum · Glue가
- : Database → Table(, , SerDe,).
- / : `MSCK REPAIR TABLE` / `ALTER TABLE ADD PARTITION` /
- (Avro/JSON/Protobuf) —

AWS Lake Formation

- : DB/ / . IAM
- LF - Tags(): . Data filters(/).
- : Athena, Redshift Spectrum, EMR, Glue, QuickSight.
- Blueprints(), , (). Hybrid access mode.

9.

- : (OLTP) vs / (+ , DW). SCD Type 1() · Type 2(가) · Type 3().
- Redshift: =DISTKEY()+SORTKEY(), =DISTSTYLE ALL.
- DynamoDB: . S3 : (raw → cleansed → curated), Parquet+ .
- : Avro(가), Parquet(가 OK), Iceberg/Hudi/Delta(, time travel).
Glue , (BACKWARD/FORWARD/FULL).
- : AWS SCT, DMS Schema Conversion(DB).
- : SageMaker ML Lineage Tracking, DataZone, Glue(), DataBrew .
- : , (Redshift AZ64/ZSTD), (, RDS),

10.

- Amazon CloudWatch: . Logs(), Logs Insights(), Metric Filter(→), Contributor Insights, EMR/Glue/Kinesis .
- AWS CloudTrail: API () + (S3 · Lambda ,). CloudTrail Lake(SQL , 10).
- : Kinesis GetRecords.IteratorAge(), Write/ReadProvisionedThroughputExceeded. DynamoDB ThrottledRequests · ConsumedCapacity. Redshift CPUUtilization · PercentageDiskSpaceUsed · QueueLength. Glue job DPU .
- Glue : Continuous logging, Spark UI/History Server, Job Run Insights(), .
- EMR: CloudWatch , YARN · Spark History · Persistent UI, S3 .
- : S3 → Athena/Glue , OpenSearch , CloudWatch Logs Insights. EMR.
- : CloudWatch → SNS. → EventBridge → (Lambda/Step Functions).
- : . / · Spot.

11.

- : (), (), (), (freshness),
- AWS Glue Data Quality: DQDL (: `Completeness "col" > 0.95`, `IsUnique`, `ColumnValues in [...]`) ETL → EventBridge/CloudWatch,
- Glue DataBrew: (), 250+
- Deequ(EMR/Spark): ,
- : / / .

12. (가)

- IAM: vs (), , ABAC(). AWS vs
- : (), Redshift/RDS IAM · IAM Identity Center 가
- Secrets Manager: + (Lambda), RDS/Redshift/DocumentDB , vs SSM Parameter Store(, , SecureString KMS).

SSE - S3	S3 AES - 256 . 가 ().
SSE - KMS	KMS → . CloudTrail . CMK . (Bucket Key KMS ↓)
DSSE - KMS	KMS ().
SSE - C	, AWS .
CSE	(). AWS .

- : HTTPS/TLS (S3 `aws:SecureTransport`), Redshift SSL, JDBC SSL, VPC
- KMS: (), + IAM+grant, (Redshift/EMR/Glue/Kinesis/Athena) KMS .
- : VPC Gateway Endpoint(S3 · DynamoDB,) / Interface Endpoint(PrivateLink)(,), · NACL, S3 Access Points(), +NAT, Glue/EMR/Redshift VPC
- 가: , RBAC/ABAC, Lake Formation , Redshift GRANT(. .), S3 · ACL ().

13.

- PII : Amazon Macie(S3 → ML, findings → EventBridge/Security Hub), Glue (PII detection transform,), Comprehend, Macie + Lake Formation .
- : Redshift . RLS, Lake Formation , Athena/Glue .
- : SCP, S3 (), KMS , AWS Config .
- AWS Config: () (remediation).
- : Redshift datashare, Lake Formation , AWS Data Exchange, S3 Access Grants, Clean Rooms(,).
- : S3 Object Lock(WORM), Glacier Vault Lock, (AWS Backup), .
- : CloudTrail(→ S3), CloudTrail Lake, CloudWatch Logs , Athena/OpenSearch .

14.

- → S3() : Kinesis Data Firehose (+ Lambda , Parquet) .
- 1 + + : Kinesis Data Streams (+ Enhanced Fan - Out).
- 가 : IteratorAge → / , EFO.
- DB (CDC) : DMS → S3/Kinesis. : SCT.
- Spark ETL + : AWS Glue. . Spot · Hudi/Iceberg : EMR.
- Glue : Job Bookmark.
- S3 SQL, : Athena + Workgroup . ↓ : Parquet+ + .
- Redshift S3 : Spectrum. RDS : Federated Query. : Data Sharing.
- Redshift - : DISTKEY. : DISTSTYLE ALL. : SORTKEY.
- Redshift : Concurrency Scaling + Auto WLM.
- DynamoDB / : + write sharding, on - demand .
- DynamoDB : S3 Export() → Athena/Glue.
- : MWAA. : Step Functions. Glue : Glue Workflows.
- : EventBridge Scheduler. → → : EventBridge Pipes.
- S3 : Macie. ETL PII : Glue .
- KMS . : SSE - KMS(+ Bucket Key). : Secrets Manager.
- S3 · DynamoDB () : Gateway VPC Endpoint. : Interface Endpoint(PrivateLink).
- : SCP + S3 + AWS Config .
- : Glue Data Quality(DQDL). : DataBrew.
- : Lake Formation (LF - Tags / Data Filters).
- : compaction / Firehose ↑ / Glue groupFiles / Iceberg .