

AWS Certified Data Engineer – Associate (DEA-C01)

Course Duration ~45 Hours · Exam-Aligned Curriculum

Ingestion & Transformation — 34%

Data Stores — 26%

Operations — 22%

Security & Governance — 18%

In short: This is an instructor-led certification program mapped 1:1 to the official AWS Certified Data Engineer – Associate (DEA-C01) exam guide. The exam validates the ability to implement data pipelines and to monitor, troubleshoot, and optimize cost and performance issues per best practices — ingesting and transforming data and orchestrating pipelines while applying programming concepts; choosing an optimal data store, designing data models, cataloging schemas, and managing data lifecycles; operationalizing and monitoring pipelines and ensuring data quality; and implementing authentication, authorization, encryption, privacy, governance, and logging. You build hands-on with AWS Glue, Amazon Athena, AWS Lake Formation, Amazon Redshift, Amazon EMR, Amazon Kinesis, and Amazon S3 — plus modern topics like Apache Iceberg, vector indexes, and LLM-assisted processing — finishing with full-length mock exams and an end-to-end pipeline capstone. On completion you are ready to pass DEA-C01 and take on AWS data engineering roles.

Certification	AWS Certified Data Engineer – Associate	Offered by	Amazon Web Services (AWS)
Exam code	DEA-C01	Level	Associate
Recommended	2–3 years of data engineering experience, plus 1–2 years of hands-on experience with AWS services		
Format	65 questions (50 scored + 15 unscored); multiple choice & multiple response	Duration	130 minutes
Passing score	720 / 1000 (pass/fail)	Recertification	Every 3 years
Delivery	Pearson VUE test center or online proctored		

Domain 1: Data Ingestion and Transformation

34%

ESTIMATED TIME

15–16 Hours

1.1 Perform data ingestion

- Read from streaming sources (Amazon Kinesis, Amazon MSK, DynamoDB Streams, AWS DMS, AWS Glue, Amazon Redshift) and batch sources (Amazon S3, Glue, EMR, DMS, Redshift, Lambda, Amazon AppFlow)
- Configure batch ingestion; consume data APIs; set up schedulers (Amazon EventBridge, Apache Airflow, time-based) and event triggers (S3 Event Notifications, EventBridge)
- Handle throttling and rate limits; manage fan-in/fan-out for streaming; describe pipeline replayability; define stateful vs. stateless transactions

1.2 Transform and process data

- Optimize container usage (Amazon EKS, ECS); connect to sources via JDBC/ODBC; integrate data from multiple sources; optimize processing costs
- Implement transformation services (Amazon EMR, AWS Glue, Lambda, Redshift); convert formats (e.g., CSV to Apache Parquet); troubleshoot transformation failures
- Create data APIs; define volume/velocity/variety; integrate large language models (LLMs) for data processing

1.3 Orchestrate data pipelines

- Build workflows with Lambda, EventBridge, Amazon MWAA, AWS Step Functions, Glue workflows; design for performance, availability, scalability, resiliency, fault tolerance; maintain serverless workflows; send alerts (Amazon SNS, SQS)

1.4 Apply programming concepts

- Optimize code and Lambda concurrency; use languages/frameworks (Python, SQL, Scala, R, Java, Bash, PowerShell); apply software engineering best practices (version control, testing, logging, monitoring)
- Use Infrastructure as Code (AWS CloudFormation, AWS CDK) and AWS SAM for serverless pipelines; describe CI/CD, distributed computing, data structures, and algorithms

HANDS-ON

- Build a streaming (Kinesis) and batch (S3 + Glue) ingestion pipeline with EventBridge triggers
- Orchestrate an end-to-end workflow with Step Functions and deploy it via CDK

Assessment: Quiz + Lab

Domain 2: Data Store Management**26%****ESTIMATED TIME**

11–12 Hours

2.1 Choose a data store

- Select and configure storage for cost, performance, and access patterns (Amazon Redshift, EMR, Lake Formation, RDS, DynamoDB, Kinesis Data Streams, MSK)
- Apply services to use cases (e.g., HNSW indexing with Aurora PostgreSQL, Amazon MemoryDB for fast key/value access); manage migration and remote access (Redshift federated queries, materialized views, Redshift Spectrum); manage locks
- Manage open table formats (Apache Iceberg); describe vector index types (HNSW, IVF)

2.2 Understand data cataloging systems

- Build and reference technical data catalogs (AWS Glue Data Catalog, Apache Hive metastore); use Glue crawlers to discover schemas and populate catalogs; synchronize partitions; create and manage business catalogs (Amazon SageMaker Catalog)

2.3 Manage the lifecycle of data

- Load/unload between S3 and Redshift; manage S3 Lifecycle policies and storage tiering; expire aged data; manage S3 versioning and DynamoDB TTL; delete data for legal/business requirements; protect data with resiliency and availability

2.4 Design data models and schema evolution

- Design schemas for Redshift, DynamoDB, and Lake Formation; handle changing data characteristics; perform schema conversion (AWS SCT, AWS DMS Schema Conversion); establish data lineage; apply indexing, partitioning, and compression best practices; describe vectorization concepts (Amazon Bedrock knowledge base)

HANDS-ON

- Catalog a data lake with Glue crawlers and query it via Redshift Spectrum
- Design an Iceberg-backed table with lifecycle and partitioning strategies

Assessment: Quiz + Lab

Domain 3: Data Operations and Support**22%****ESTIMATED TIME**

9–10 Hours

3.1 Automate data processing

- Orchestrate pipelines (Amazon MWAA, Step Functions); troubleshoot managed workflows; call SDKs; process data with EMR, Redshift, Glue; prepare data (AWS Glue DataBrew, Amazon SageMaker Unified Studio); query with Amazon Athena; automate with Lambda and EventBridge

3.2 Analyze data

- Visualize data (DataBrew, Amazon QuickSight); verify and clean data (Lambda, Athena, QuickSight, Jupyter, SageMaker Data Wrangler); use SQL in Redshift and Athena; use Athena Spark notebooks; compare provisioned vs. serverless; define aggregation, rolling averages, grouping, pivoting

3.3 Maintain and monitor data pipelines

- Extract logs for audits; deploy logging and monitoring; send alerts; troubleshoot performance; use AWS CloudTrail and Amazon CloudWatch Logs; analyze logs (Athena, EMR, Amazon OpenSearch Service, CloudWatch Logs Insights)

3.4 Ensure data quality

- Run data quality checks; define quality rules and investigate consistency (DataBrew); describe data sampling techniques; implement data skew mechanisms

HANDS-ON

- Automate a Glue/Athena pipeline with MWAA and alerting
- Build data quality rules in DataBrew and monitor with CloudWatch Logs Insights

Assessment: Quiz + Lab

Domain 4: Data Security and Governance**18%****ESTIMATED TIME**

7–8 Hours

4.1 Apply authentication mechanisms

- Update VPC security groups; create and manage IAM groups, roles, endpoints; rotate credentials (AWS Secrets Manager); set up IAM roles for access; apply IAM policies (S3 Access Points, AWS PrivateLink); use SageMaker Unified Studio domains and projects

4.2 Apply authorization mechanisms

- Create custom IAM policies; store credentials (Secrets Manager, Systems Manager Parameter Store); manage database access (Redshift); manage permissions via AWS Lake Formation; apply role-based, tag-based, and attribute-based authorization following least privilege

4.3 Ensure data encryption and masking

- Apply data masking and anonymization; encrypt/decrypt with AWS KMS; configure encryption across account boundaries; enable encryption in transit

4.4 Prepare logs for audit

- Track API calls with AWS CloudTrail; store logs in CloudWatch Logs; use CloudTrail Lake for centralized queries; analyze logs (Athena, CloudWatch Logs Insights, OpenSearch)

4.5 Understand data privacy and governance

- Grant data sharing permissions (Redshift); identify PII (Amazon Macie with Lake Formation); prevent data replication to disallowed Regions; track configuration changes (AWS Config); maintain data sovereignty; manage access via SageMaker Catalog; describe governance frameworks and data sharing patterns

HANDS-ON

- Secure a data lake with Lake Formation permissions, KMS encryption, and masking
- Audit pipeline activity with CloudTrail Lake and Athena

Assessment: Quiz + Lab

Exam Readiness & Mock Exams**ESTIMATED TIME**

3–4 Hours

TOPICS

- Domain-by-domain review keyed to the official weightings; pacing for 65 questions in 130 minutes
- Service-selection reasoning: cost, performance, and access patterns; question strategy and elimination techniques
- Recertification path: 3-year validity

HANDS-ON

- Full-length timed mock exams with detailed review
- Capstone: end-to-end pipeline — ingest, transform, catalog, govern, and monitor

Assessment: Capstone + Mock Exams