AWS Certified Data Engineer - Associate (DEA-C01) 考试指南

简介

AWS Certified Data Engineer - Associate (DEA-C01) 考试旨在考查考生能否实施数据管道，以及能否根据最佳实践监控、排查和优化成本和性能问题。

本考试还考查考生能否完成以下任务：

- 在应用编程概念时，摄取和转换数据并编排数据管道。
- 选择最佳数据存储，设计数据模型，对数据架构进行编目以及管理数据生命周期。
- 运行、维护和监控数据管道。分析数据并确保数据质量。
- 实施相应的身份验证、授权、数据加密、隐私和监管。启用日志记录。

目标考生描述

目标考生应具有相当于 2-3 年的数据工程经验。目标考生应了解数量、种类和速度对数据采集、转换、建模、安全、治理、隐私、架构设计和最佳数据存储设计的影响。此外，目标考生还应具有至少 1-2 年的 AWS 服务实践经验。

一般 IT 知识推荐

目标考生应具备以下的一般 IT 知识：

- 设置和维护从摄取到目标的提取、转换和加载 (ETL) 管道
- 根据管道的要求，应用高级但与语言无关的编程概念
- 如何使用 Git 命令进行源代码控制
- 如何使用数据湖存储数据
- 联网、存储和计算的一般概念

AWS 知识推荐

目标考生应具备以下 AWS 知识：

- 如何使用 AWS 服务完成本考试指南的 “简介” 部分列出的任务
- 了解用于加密、监管、保护和记录属于数据管道的所有数据的 AWS 服务
- 能够比较 AWS 服务来了解不同服务之间的成本、性能和功能差异
- 如何构建 SQL 查询以及如何在 AWS 服务上运行 SQL 查询
- 了解如何使用 AWS 服务分析数据，验证数据质量和确保数据一致性

超出目标考生考试范围的工作任务

下表列出了不要求目标考生能够完成的相关工作任务。此列表并非详尽无遗。
以下任务超出考试范围：

- 执行人工智能和机器学习 (AI/ML) 任务。
- 证明了解编程语言特定的语法。
- 根据数据得出业务结论。

有关范围内 AWS 服务和功能的列表, 范围外的 AWS 服务和功能列表，请参阅附录。

考试内容

答案类型

本考试具有两种类型的试题：

- **单选题：** 具有一个正确答案和三个错误答案（干扰项）
- **多选题：** 在 5 个或更多答案选项中具有两个或更多正确答案

选择一个或多个最准确表述或回答试题的答案。干扰项或错误答案是知识或技能不全面的考生可能会选择的答案选项。干扰项通常是与内容领域相符的看似合理的答案。

未回答的试题将计为回答错误；猜答案不会扣分。本考试包括 50 道试题，这些试题将影响您的分数。

不计分内容

考试包括 15 道不计分试题，这些试题不影响您的分数。AWS 收集这些不计分试题的答题情况以进行评估，以便将来将这些试题作为计分试题。在考试中不会标明这些不计分试题。

考试结果

AWS Certified Data Engineer - Associate (DEA-C01) 考试结果分为及格和不及格两种。本考试按照 AWS 专业人员根据认证行业最佳实践和准则制订的最低标准进行评分。

您的考试结果换算分数为 100 – 1000 分。最低及格分数为 720 分。
您的分数表明您的总体考试答题情况以及是否通过考试。
换算评分模型有助于在难度水平可能略有不同的多种考试形式中换算分数。

您的成绩单可能包含一个分类表，其中列出您在每个部分的考试成绩。本考试采用补偿评分模型，这意味着您无需在每个部分都达到及格分数。您只需通过整体考试即可。

考试的每个部分具有特定的权重，因此，某些部分的试题比其他部分多。分类表包含一般信息，用于重点说明您的强项和弱项。在解读各个部分的反馈时，请务必小心谨慎。

内容大纲

本考试指南包括考试的权重、内容领域和任务表述。并未列出考试的全部内容。不过，每个任务表述都提供有额外的背景信息，有助于您备考。

考试中考查的内容领域和相应的权重如下：

- 领域 1： 数据摄取和转换（占评分内容的 34%）
- 领域 2： 数据存储管理（占评分内容的 26%）
- 领域 3： 数据操作和支持（占评分内容的 22%）
- 领域 4： 数据安全与治理（占评分内容的 18%）

领域 1： 数据摄取和转换

任务表述 1.1： 执行数据摄取。

掌握以下知识：

- 用于摄取数据的 AWS 服务的吞吐量和延迟特性
- 数据摄取模式（例如，频率和数据历史记录）
- 流数据摄取
- 批量数据摄取（例如，计划的摄取、事件驱动的摄取）
- 数据摄取管道的可重放性
- 有状态和无状态数据事务

具备以下技能：

- 从流数据源（例如，Amazon Kinesis、Amazon Managed Streaming for Apache Kafka [Amazon MSK]、Amazon DynamoDB Streams、AWS Database Migration Service [AWS DMS]、AWS Glue、Amazon Redshift）读取数据
- 从批量数据源（例如，Amazon S3、AWS Glue、Amazon EMR、AWS DMS、Amazon Redshift、AWS Lambda、Amazon AppFlow）读取数据
- 为批量摄取实施相应的配置选项
- 使用数据 API
- 使用 Amazon EventBridge、Apache Airflow 或基于时间的任务和爬网程序计划设置调度器
- 设置事件触发器（例如，Amazon S3 事件通知、EventBridge）
- 从 Amazon Kinesis 中调用 Lambda 函数
- 为 IP 地址创建允许列表来允许连接到数据源
- 实施限流和解决速率限制问题（例如，DynamoDB、Amazon RDS、Kinesis）
- 管理流数据分配的扇入和扇出

任务表述 1.2： 转换和处理数据。

掌握以下知识：

- 根据业务需求创建 ETL 管道
- 数据数量、速度和种类（例如，结构化数据、非结构化数据）
- 云计算和分布式计算
- 如何使用 Apache Spark 处理数据
- 中间数据暂存位置

具备以下技能：

- 根据性能需求优化容器使用情况（例如，Amazon Elastic Kubernetes Service [Amazon EKS]、Amazon Elastic Container Service [Amazon ECS]）
- 连接到不同的数据源（例如，Java 数据库连接 [JDBC]、开放式数据库连接 [ODBC]）
- 整合来自多个来源的数据
- 在处理数据时优化成本
- 根据要求实施数据转换服务（例如，Amazon EMR、AWS Glue、Lambda、Amazon Redshift）
- 在不同格式之间转换数据（例如，从 .csv 转换到 Apache Parquet）
- 对常见的转换失败和性能问题进行故障排除和调试
- 创建数据 API，通过 AWS 服务向其他系统提供数据

任务表述 1.3： 编排数据管道。

掌握以下知识：

- 如何集成各种 AWS 服务来创建 ETL 管道
- 事件驱动型架构
- 如何根据计划或依赖项为数据管道配置 AWS 服务
- 无服务器工作流

具备以下技能：

- 使用编排服务为 ETL 数据管道构建工作流（例如，Lambda、EventBridge、Amazon Managed Workflows for Apache Airflow [Amazon MWAA]、AWS Step Functions、AWS Glue 工作流）
- 构建数据管道来提高性能、可用性、可扩展性、恢复能力和容错能力
- 实施和维护无服务器工作流
- 使用通知服务发送警报（例如，Amazon Simple Notification Service [Amazon SNS]、Amazon Simple Queue Service [Amazon SQS]）

任务表述 1.4： 应用编程概念。

掌握以下知识：

- 持续集成和持续交付 (CI/CD)（实施、测试和部署数据管道）
- SQL 查询（用于数据源查询和数据转换）
- 用于可重复部署的基础设施即代码 (IaC)（例如，AWS Cloud Development Kit [AWS CDK]、AWS CloudFormation）

- 分布式计算
- 数据结构和算法（例如，图形数据结构和树数据结构）
- SQL 查询优化

具备以下技能：

- 优化代码来减少数据摄取和转换的运行时间
- 配置 Lambda 函数来满足并发性和性能需求
- 执行 SQL 查询来转换数据（例如，Amazon Redshift 存储过程）
- 构建 SQL 查询来满足数据管道要求
- 使用 Git 命令执行创建、更新、克隆和分支存储库等操作
- 使用 AWS Serverless Application Model (AWS SAM) 打包和部署无服务器数据管道（例如，Lambda 函数、Step Functions、DynamoDB 表）
- 从 Lambda 函数中使用和挂载存储卷

领域 2： 数据存储管理

任务表述 2.1： 选择数据存储。

掌握以下知识：

- 存储平台及其特性
- 满足特定性能要求的存储服务和配置
- 数据存储格式（例如，.csv、.txt、Parquet）
- 如何将数据存储与数据迁移要求保持一致
- 如何为特定访问模式确定相应的存储解决方案
- 如何管理锁定来防止访问数据（例如，Amazon Redshift 和 Amazon RDS）

具备以下技能：

- 根据特定成本和性能要求实施相应的存储服务（例如，Amazon Redshift、Amazon EMR、AWS Lake Formation、Amazon RDS、DynamoDB、Amazon Kinesis Data Streams、Amazon MSK）
- 根据特定访问模式和要求配置相应的存储服务（例如，Amazon Redshift、Amazon EMR、Lake Formation、Amazon RDS、DynamoDB）
- 将存储服务应用于相应的使用案例（例如，Amazon S3）
- 将迁移工具集成到数据处理系统（例如，AWS Transfer Family）
- 实施数据迁移或远程访问方法（例如，Amazon Redshift 联合查询、Amazon Redshift 物化视图、Amazon Redshift Spectrum）

任务表述 2.2： 了解数据编目系统。

掌握以下知识：

- 如何创建数据目录
- 根据要求对数据进行分类
- 元数据和数据目录的组成部分

具备以下技能：

- 通过数据目录使用数据源中的数据
- 构建和引用数据目录（例如，AWS Glue 数据目录、Apache Hive 元存储）
- 查找架构并使用 AWS Glue 爬网程序填充数据目录
- 将分区与数据目录同步
- 创建新的源或目标连接进行编目（例如，AWS Glue）

任务表述 2.3： 管理数据的生命周期。

掌握以下知识：

- 利用相应的存储解决方案来满足冷热数据要求
- 如何根据数据生命周期优化存储成本
- 如何删除数据来满足业务和法律要求
- 数据留存策略和归档策略
- 如何使用相应的恢复能力和可用性保护数据

具备以下技能：

- 执行加载和卸载操作以在 Amazon S3 和 Amazon Redshift 之间移动数据
- 管理 S3 生命周期策略来更改 S3 数据的存储层
- 使用 S3 生命周期策略使数据在到达特定期限时过期
- 管理 S3 版本控制和 DynamoDB TTL

任务表述 2.4： 设计数据模型和架构演变。

掌握以下知识：

- 数据建模概念
- 如何使用数据沿袭确保数据的准确性和可信度
- 索引编制、分区策略、压缩和其他数据优化技术的最佳实践
- 如何为结构化数据、半结构化数据和非结构化数据建模
- 架构演变技术

具备以下技能：

- 为 Amazon Redshift、DynamoDB 和 Lake Formation 设计架构
- 解决数据特性变化问题
- 执行架构转换（例如，使用 AWS Schema Conversion Tool [AWS SCT] 和 AWS DMS Schema Conversion）
- 使用 AWS 工具（例如，Amazon SageMaker ML Lineage Tracking）确定数据沿袭

领域 3： 数据操作和支持

任务表述 3.1： 使用 AWS 服务自动处理数据。

掌握以下知识：

- 如何维护数据处理和排除故障来获得可重复的业务结果
- 用于数据处理的 API 调用
- 哪些服务接受脚本（例如，Amazon EMR、Amazon Redshift、AWS Glue）

具备以下技能：

- 编排数据管道（例如，Amazon MWAA、Step Functions）
- 故障排除 Amazon 托管工作流故障
- 通过代码调用 SDK 来访问 Amazon 功能
- 使用 AWS 服务功能处理数据（例如，Amazon EMR、Amazon Redshift、AWS Glue）
- 使用和维护数据 API
- 准备数据转换（例如，AWS Glue DataBrew）
- 查询数据（例如，Amazon Athena）
- 使用 Lambda 自动处理数据
- 管理事件和调度器（例如 EventBridge）

任务表述 3.2： 使用 AWS 服务分析数据。

掌握以下知识：

- 权衡预置的服务和无服务器服务的利弊
- SQL 查询（例如，带有多个限定符或 JOIN 子句的 SELECT 语句）
- 如何将数据可视化来进行分析
- 何时以及如何应用清理技术
- 数据聚合、滚动平均值、分组和透视

具备以下技能：

- 使用 AWS 服务和工具（例如，AWS Glue DataBrew、Amazon QuickSight）对数据进行可视化
- 验证和清理数据（例如，Lambda、Athena、QuickSight、Jupyter Notebooks、Amazon SageMaker Data Wrangler）
- 使用 Athena 查询数据或创建视图
- 使用通过 Apache Spark 查找数据的 Athena 笔记本

任务表述 3.3： 维护和监控数据管道。

掌握以下知识：

- 如何记录应用程序数据
- 性能优化的最佳实践
- 如何记录对 AWS 服务的访问
- Amazon Macie、AWS CloudTrail 和 Amazon CloudWatch

具备以下技能：

- 提取日志来进行审核
- 部署日志记录和监控解决方案以便于审核和追溯
- 在监控期间使用通知发送警报
- 故障排除性能问题
- 使用 CloudTrail 跟踪 API 调用
- 对管道进行故障排除和维护（例如，AWS Glue、Amazon EMR）
- 使用 Amazon CloudWatch Logs 记录应用程序数据（侧重于配置和自动化）
- 使用 AWS 服务（例如，Athena、Amazon EMR、Amazon OpenSearch Service、CloudWatch Logs Insights、大数据应用程序日志）分析日志

任务表述 3.4： 确保数据质量。

掌握以下知识：

- 数据采样技术
- 如何实施数据偏斜机制
- 数据验证（数据完整性、一致性和准确性）
- 数据分析

具备以下技能：

- 在处理数据时，运行数据质量检查（例如，检查空字段）
- 定义数据质量规则（例如，AWS Glue DataBrew）
- 调查数据一致性（例如，AWS Glue DataBrew）

领域 4： 数据安全性和监管

任务表述 4.1： 应用身份验证机制。

掌握以下知识：

- VPC 安全联网概念
- 托管服务和非托管服务之间的差异
- 身份验证方法（基于密码、基于证书和基于角色）
- AWS 托管策略和客户托管策略之间的差异

具备以下技能：

- 更新 VPC 安全组
- 创建和更新 IAM 组、角色、终端节点和服务
- 创建和轮换凭证来管理密码（例如，AWS Secrets Manager）
- 设置 IAM 角色来进行访问（例如，Lambda、Amazon API Gateway、AWS CLI、CloudFormation）
- 将 IAM 策略应用于角色、终端节点和服务（例如，S3 访问点、AWS PrivateLink）

任务表述 4.2： 应用授权机制。

掌握以下知识：

- 授权方法（基于角色、基于策略、基于标签和基于属性）
- 适用于 AWS 安全性的最低权限原则
- 基于角色的访问控制和预期的访问模式
- 保护数据来防止在服务中进行未经授权访问的方法

具备以下技能：

- 在托管策略不满足需求时创建自定义 IAM 策略
- 存储应用程序和数据库凭证（例如，Secrets Manager、AWS Systems Manager Parameter Store）
- 在数据库中为数据库用户、组和角色提供访问权限和授权（例如，适用于 Amazon Redshift）

- 通过 Lake Formation 管理权限（适用于 Amazon Redshift、Amazon EMR、Athena 和 Amazon S3）

任务表述 4.3： 确保数据加密和脱敏。

掌握以下知识：

- AWS 分析服务（例如，Amazon Redshift、Amazon EMR、AWS Glue）中提供的数据加密选项
- 客户端加密和服务器端加密之间的差异
- 保护敏感数据
- 数据匿名化、脱敏和密钥加盐

具备以下技能：

- 根据合规法律或公司策略应用数据脱敏和匿名化
- 使用加密密钥加密或解密数据（例如，AWS Key Management Service [AWS KMS]）
- 配置跨 AWS 账户边界的加密
- 为数据启用传输中加密功能。

任务表述 4.4： 准备日志进行审核。

掌握以下知识：

- 如何记录应用程序数据
- 如何记录对 AWS 服务的访问
- 集中式 AWS 日志

具备以下技能：

- 使用 CloudTrail 跟踪 API 调用
- 使用 CloudWatch Logs 存储应用程序日志
- 使用 AWS CloudTrail Lake 进行集中式日志记录查询
- 使用 AWS 服务（例如，Athena、CloudWatch Logs Insights、Amazon OpenSearch Service）分析日志
- 集成各种 AWS 服务来执行日志记录（例如，在具有大量日志数据时集成 Amazon EMR）

任务表述 4.5： 了解数据隐私和监管。

掌握以下知识：

- 如何保护个人身份信息 (PII)
- 数据主权

具备以下技能：

- 授予数据共享权限（例如，Amazon Redshift 数据共享）
- 实施 PII 识别（例如，将 Macie 与 Lake Formation 一起使用）
- 实施数据隐私策略来防止将数据备份或复制到不允许的 AWS 区域
- 管理在账户中发生的配置更改（例如 AWS Config）

附录

考试范围内的 AWS 服务和功能

下表列出了考试范围内的 AWS 服务和功能。此列表并非详尽无遗，并且可能会更改。AWS 产品/服务的类别与产品/服务的主要功能一致：

分析：

- Amazon Athena
- Amazon EMR
- AWS Glue
- AWS Glue DataBrew
- AWS Lake Formation
- Amazon Kinesis Data Firehose
- Amazon Kinesis Data Streams
- Amazon Managed Service for Apache Flink
- Amazon Managed Streaming for Apache Kafka (Amazon MSK)
- Amazon OpenSearch Service
- Amazon QuickSight

应用程序集成：

- Amazon AppFlow
- Amazon EventBridge
- Amazon Managed Workflows for Apache Airflow（Amazon MWAA）
- Amazon Simple Notification Service (Amazon SNS)
- Amazon Simple Queue Service (Amazon SQS)
- AWS Step Functions

云财务管理：

- AWS Budgets
- AWS Cost Explorer

计算：

- AWS Batch
- Amazon EC2
- AWS Lambda
- AWS Serverless Application Model (AWS SAM)

容器：

- Amazon Elastic Container Registry (Amazon ECR)
- Amazon Elastic Container Service (Amazon ECS)
- Amazon Elastic Kubernetes Service (Amazon EKS)

数据库：

- Amazon DocumentDB（与 MongoDB 兼容）
- Amazon DynamoDB
- Amazon Keyspaces（适用于 Apache Cassandra）
- 适用于 Redis 的 Amazon MemoryDB
- Amazon Neptune
- Amazon RDS
- Amazon Redshift

开发工具：

- AWS CLI
- AWS Cloud9
- AWS Cloud Development Kit (AWS CDK)
- AWS CodeBuild
- AWS CodeCommit
- AWS CodeDeploy
- AWS CodePipeline

前端 Web 和移动：

- Amazon API Gateway

机器学习：

- Amazon SageMaker

管理和监管：

- AWS CloudFormation
- AWS CloudTrail
- Amazon CloudWatch
- Amazon CloudWatch Logs
- AWS Config
- Amazon Managed Grafana
- AWS Systems Manager
- AWS Well-Architected Tool

迁移和传输：

- AWS Application Discovery Service
- AWS Application Migration Service
- AWS Database Migration Service (AWS DMS)
- AWS DataSync
- AWS Schema Conversion Tool (AWS SCT)
- AWS Snow Family
- AWS Transfer Family

联网和内容分发：

- Amazon CloudFront
- AWS PrivateLink
- Amazon Route 53
- Amazon VPC

安全性、身份和合规性：

- AWS Identity and Access Management (IAM)
- AWS Key Management Service (AWS KMS)
- Amazon Macie
- AWS Secrets Manager
- AWS Shield
- AWS WAF

存储：

- AWS Backup
- Amazon Elastic Block Store (Amazon EBS)
- Amazon Elastic File System (Amazon EFS)
- Amazon S3
- Amazon S3 Glacier

超出考试范围的 AWS 服务和功能

下表列出了超出考试范围的 AWS 服务和功能。此列表并非详尽无遗，并且可能会更改。与考试的目标工作职责完全无关的 AWS 产品/服务被排除在此列表之外：

分析：

- Amazon FinSpace

业务应用程序：

- Alexa for Business
- Amazon Chime
- Amazon Connect
- Amazon Honeycode
- AWS IQ
- Amazon WorkDocs
- Amazon WorkMail

计算：

- AWS App Runner
- AWS Elastic Beanstalk
- Amazon Lightsail
- AWS Outposts
- AWS Serverless Application Repository

容器：

- AWS 云端 Red Hat OpenShift 服务 (ROSA)

数据库：

- Amazon Timestream

开发工具：

- AWS Fault Injection Simulator (AWS FIS)
- AWS X-Ray

前端 Web 和移动：

- AWS Amplify
- AWS AppSync
- AWS Device Farm
- Amazon Location Service
- Amazon Pinpoint
- Amazon Simple Email Service (Amazon SES)

物联网 (IoT)：

- FreeRTOS
- AWS IoT 1-Click
- AWS IoT Device Defender
- AWS IoT Device Management
- AWS IoT Events
- AWS IoT FleetWise
- AWS IoT RoboRunner
- AWS IoT SiteWise
- AWS IoT TwinMaker

机器学习：

- Amazon CodeWhisperer
- Amazon DevOps Guru

管理和监管：

- AWS Activate
- AWS Managed Services (AMS)

媒体服务：

- Amazon Elastic Transcoder
- AWS Elemental Appliances and Software
- AWS Elemental MediaConnect
- AWS Elemental MediaConvert
- AWS Elemental MediaLive
- AWS Elemental MediaPackage
- AWS Elemental MediaStore
- AWS Elemental MediaTailor
- Amazon Interactive Video Service (Amazon IVS)
- Amazon Nimble Studio

迁移和传输：

- AWS Mainframe Modernization
- AWS Migration Hub

存储：

- EC2 Image Builder

调查问卷

本考试指南对您有帮助吗？请通过调查问卷，反馈您的意见。