亚马逊云科技

游戏行业构建数据湖的最佳实践

亚马逊云科技白皮书

在亚马逊云科技上为游戏构建数据湖的最佳实践：亚马逊云科技白皮书

目 录

在亚马逊云科技上为游戏构建数据湖的最佳实践

发布日期：**2022 年 5 月 11 日**（*文档修订（第 29 页）*）

摘要和引言

摘 要

许多客户都在亚马逊云科技上运行其数字游戏。所有这些客户都做出了独特的设计选择，使他们能够在亚马逊云科技上运行最快、最复杂且视觉效果惊人的游戏。本白皮书概述了在亚马逊云科技云上为游戏构建数据湖的最佳实践，并提供了一个参考架构来指导组织交付这些复杂的系统。

您的架构是否完善？

Amazon Well-Architected Framework 能够帮助您认识到在云上构建系统时所做决策的优缺点。通过此框架的六大要素，您可以了解设计和运行可靠、安全、高效、经济实惠且可持续的系统的架构最佳实践。使用亚马逊云科技管理控制台中免费提供的 Amazon Well-Architected Tool（需要登录），您可以通过回答每个要素的一系列问题来对照这些最佳实践检查您的工作负载。

在游戏行业剖析中，我们重点介绍在亚马逊云科技上设计、架构和部署游戏工作负载。

有关云架构的更多专家指导和最佳实践（参考架构部署、图表和白皮书），请参阅 Amazon Architecture Center。

简 介

游戏行业的竞争比以往更加激烈，每年都会发布数以千计的游戏，都在争夺玩家的时间和注意力。在当今市场上打造一款成功的游戏是一项颇具挑战性的工作，需要完成大量的工作。构建集成数据平台以有效分析玩家数据，从而帮助游戏开发人员更好地了解玩家行为和业务绩效，实现及时和数据驱动的决策，并取悦玩家。如今，越来越多的游戏开发人员正在亚马逊云科技上构建数据湖以实现这些业务成果。

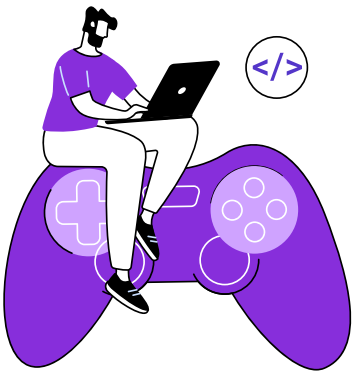
本白皮书深入讨论了在亚马逊云科技上为游戏构建数据湖的最佳实践。它旨在帮助游戏开发人员最大化其玩家数据的价值，以实现更好的游戏设计、开发、变现见解和策略。

什么是数据湖？它能为游戏开发人员带来哪些好处？

数据湖是一个集成的集中式数据平台，结合了数据存储和治理、分析、机器学习（ML）和可视化。它是一个安全、耐用且经济实惠的基于云的存储平台。利用它，您可以根据需要摄取、转换、分析和可视化结构化和非结构化数据。使用基于 Amazon Simple Storage Service（Amazon S3）的数据湖架构功能，游戏开发人员可以实现以下目标：

- 使用批处理、近实时和实时摄取方法的不同组合从各种数据源中摄取数据。
- 将摄取的数据存储在经济实惠、可靠且集中的平台中。
- 构建全面的数据目录，以便轻松访问存储在数据湖中的数据资产。
- 保护和治理存储在数据湖中的所有数据和业务元数据。
- 监控、分析及优化成本和性能。
- 直观地映射数据沿袭。
- 将原始数据资产转换为优化的可用格式并就地查询。
- 在组织内部轻松安全地共享处理后的数据集和结果。
- 使用广泛而深入的数据分析、数据科学、机器学习和可视化工具组合。
- 快速集成当前和未来的第三方数据处理工具。
- 在云中订阅第三方数据，将数据直接加载到数据湖中，并使用各种分析和机器学习服务对其进行分析。

Amazon S3 将存储与计算和数据处理分离，让构建多租户环境变得更加容易。数据湖提供的实时分析功能可以帮助您跟踪有关绩效、参与度和收入的最重要关键绩效指标（KPI），并在其生命周期的任何时候提供玩家体验及其生命周期价值（LTV）的完整视图。这使游戏开发人员能够调查和了解玩家的行为和体验，做出更好的决策，并创建和交付全新的好玩游戏。

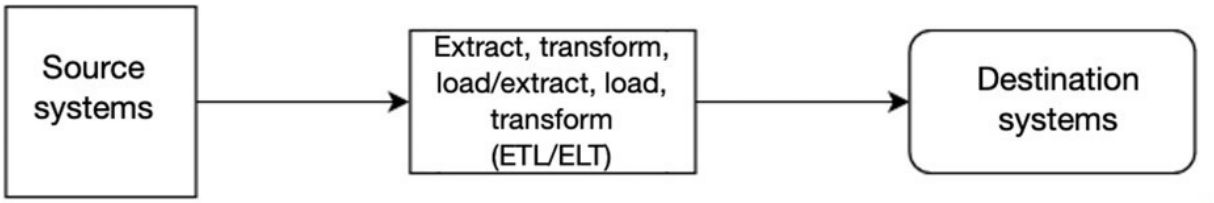


数据湖设计模式和原则

框 架

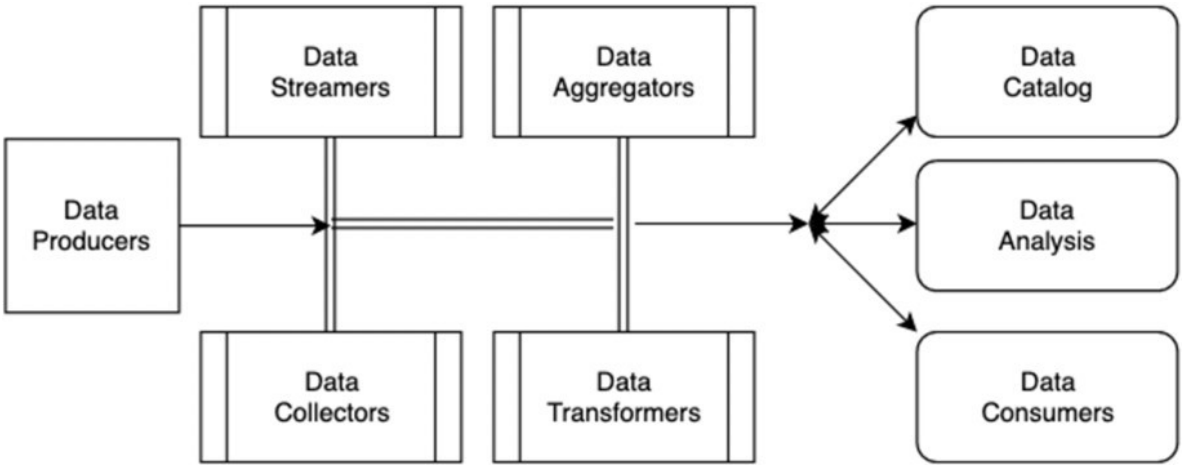
以下是在亚马逊云科技上构建数据湖的高级框架。

10,000 英尺视图



这是分析系统如何与源和目标系统配合工作的 10,000 英尺（高级）视图。

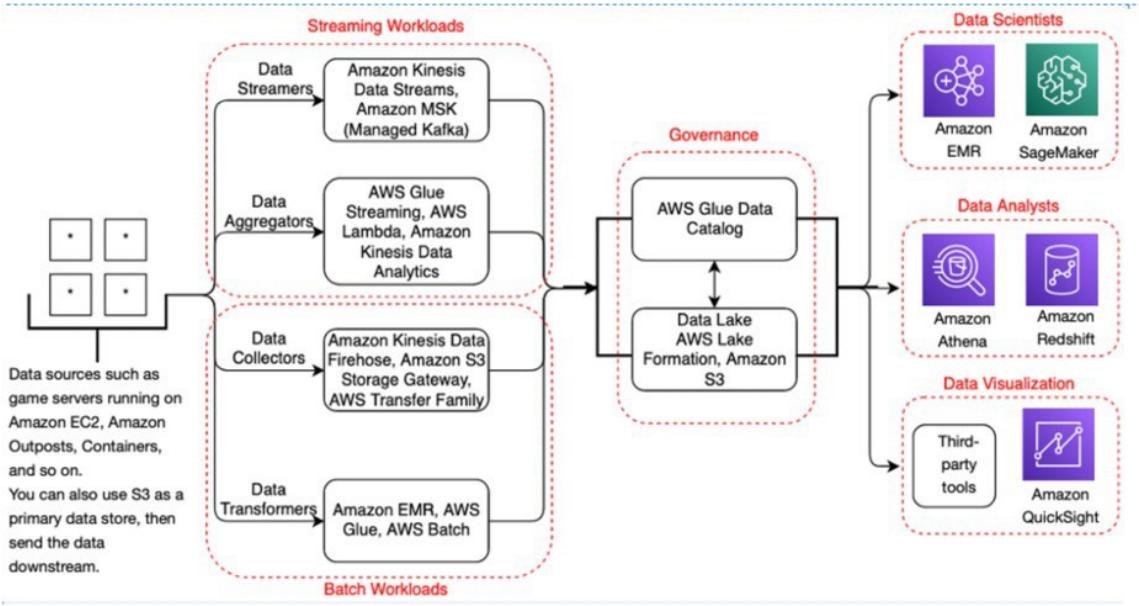
5000 英尺视图



这是分析系统如何与源和目标系统配合工作的 5,000 英尺（中级）视图。

深入到框架，其中包含数据流媒体、数据收集器、数据聚合器和数据转换器，它们从数据生产者（源）收集数据。然后根据使用场景，将数据用于分析或传输到下游使用者，并编入数据湖目录以进行受控访问。

1000 英尺视图



这是分析系统如何与源和目标系统配合工作的 1,000 英尺（详细）视图。

深入到框架，添加了一些亚马逊云科技服务作为示例来显示数据流。这种布局是亚马逊云科技观察到的在其客户中的常见模式。

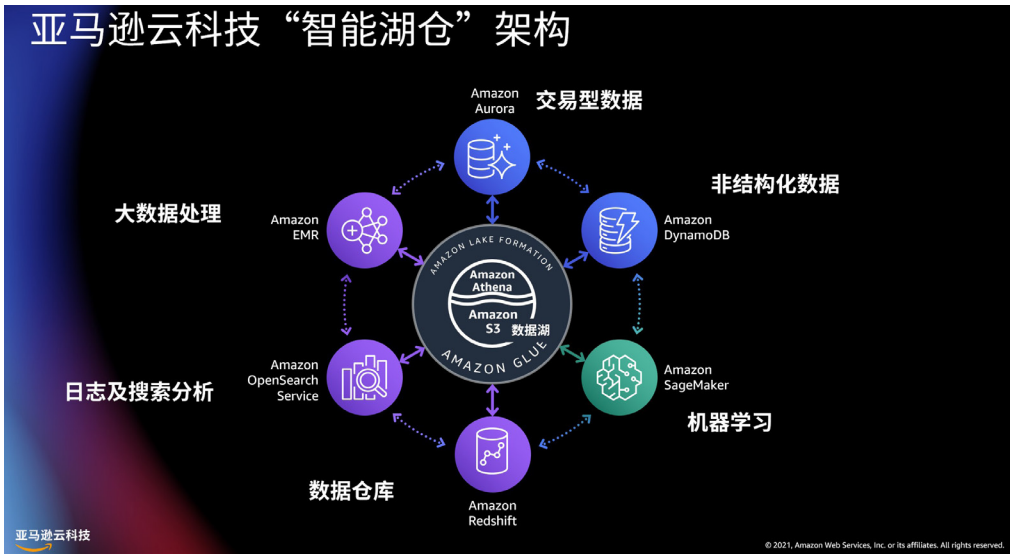
智能湖仓一体架构

游戏开发人员经常配合使用数据仓库与数据湖。数据仓库可以为处理本地数据的 SQL 查询提供更低的延迟和更好的性能。这就是为什么游戏分析中数据仓库的其中一个常见使用场景是构建日常聚合以从商业智能（BI）解决方案中使用。游戏可以生成大量数据，甚至记录下到击键的活动。这可能导致每天必须处理数 TB 的数据，并且数据可能驻留在或需要加载到不同的数据存储中。这些数据存储可以是缓存（例如 Amazon ElastiCache for Redis）、关系数据库（例如 Amazon Aurora）、NoSQL 数据库（例如 Amazon DynamoDB）以及用于存储日志数据的 Amazon S3。之后的挑战是，必须管理这些数据，并找到方法以在整个存储库中获得有意义的见解。

例如，DynamoDB 在特定使用场景（例如以个位数毫秒延迟读取和写入数据）方面表现良好。但它不应该用作分析查询的来源，因为没有一种工具可以完美地适用于每项工作。拥有智能湖仓一体架构使客户能够快速、安全、轻松地将数据移入和移出其数据存储。此外，这还使客户能够使用与许多亚马逊云科技服务集成的 Amazon Glue 数据目录将他们的数据湖连接到数据库和数据仓库。

您可以使用技术将数据湖与数据仓库集成，而不是构建孤立的数据仓库。例如，使用 Redshift Spectrum 直接从 S3 数据湖查询数据，或使用 Amazon Redshift COPY 命令以并行方式将数据从 S3 直接加载到 Amazon Redshift。许多客户都不想在很少需要查询时将 10 年或 20 年的数据加载到数据仓库中。在这种情况下，将数据仓库扩展到数据湖是一个不错的选择，因为您可以将历史（冷）数据保留在数据湖中以帮助节省成本，并在必要时使用数据仓库查询数据湖。

有关更多详细信息，请参阅从亚马逊云科技现代数据中获取见解，并参阅在亚马逊云科技上构建智能湖仓一体架构博客文章进行深入了解。



亚马逊云科技智能湖仓架构

数据摄取

游戏开发人员从各种来源收集和处理不同类型的事件。典型示例包括来自游戏、第三方服务（单击、安装、印象）以及游戏内事件的营销数据。在转换和分析数据湖中的这些数据之前，需要将其摄取到数据湖的原始区域中。本章讨论不同的数据摄取机制和设计选择。

数据收集 - 第三方服务或自助

有时，游戏开发人员会开发自己的解决方案来生成和摄取事件。例如，他们可以开发自己的移动跟踪器并拥有生成和摄取事件的代码。在这种情况下，他们可以使用自己选择的技术堆栈和数据摄取方法。然而，有时游戏开发人员依赖合作伙伴服务来收集数据。示例包括移动归因服务（例如 AppsFlyer）或第三方移动跟踪器（例如 Amplitude 或 Google Analytics）。摄取解决方案取决于此类第三方服务提供的数据导出选项。最常见的是批量导出到 S3 存储桶，您可以通过批量提取、传输和加载（ETL）过程从其中提取文件。

流式处理或批处理

可以流式处理事件或批量加载它们。流式处理意味着游戏或合作伙伴服务在事件生成时对其进行摄取。Amazon Kinesis Data Streams 或 Apache Kafka 之类的流式处理存储服务存储这些事件，使用者可以按照到达的顺序读取它们。

然而，摄取每一个事件效率非常低，会产生过多的网络流量，并且可能导致高成本，具体取决于所使用的流式处理存储和服务。这就是为什么在流式处理中经常使用某种程度的批处理来提高性能。例如，在游戏中的每场战斗之后，或者每五分钟一次，将事件发送到流中。

批量摄取意味着在服务器或合作伙伴服务中累积一组事件，然后作为包含多个事件的单个文件上传。您可以直接上传到数据湖的原始区域，或者上传到单独的存储桶并在复制到数据湖时应用某种转换。后者通常用于从第三方服务导入数据时。

在大多数情况下，流式处理是一个不错的选择。流式处理的优点包括：

- 降低了在发生客户端崩溃、网络中断（这对于手机游戏来说是正常现象）等情况时丢失事件的风险。
- 缩短了报告时间，无需等待数小时即可在数据湖中获得下一批可用的新事件。
- 使用 Apache Flink 或 Spark Streaming 等流式处理框架的实时分析功能。
- 并行支持多个独立使用者。例如，可以将数据保存在数据湖中，同时将其写入时间序列数据库。
- 当我们处理大量数据（流处理或批处理）时，成本成为一个因素。请参阅相应亚马逊云科技服务的定价页面，了解与每个流、数据摄取、数据检索和数据存储相关的成本。

有时，批处理是一种选择。批处理的优点包括：

- 能够将来自具有各种数据格式的多个数据源的数据拼接在一起。
- 能够处理大型或小型数据集。根据动态要求扩展集群大小（例如，使用 Amazon Glue、Amazon EMR Serverless 和具有 Amazon Fargate 的 Amazon Batch 的无服务器功能）。
- 通过业务服务水平协议（SLA）限制数据处理时间和资源，以实现经济高效的工作负载。

客户端或服务器

根据架构，您可以选择从游戏客户端（例如移动设备）和 / 或游戏服务器摄取游戏事件。这两种方法都很常用，并且经常一起使用，服务于不同的使用场景。用户与游戏的交互等事件可以从移动设备流式处理，而战斗结果等事件可以从服务器流式处理。

从移动设备进行流式处理时，可以使用 Amazon Pinpoint 之类的现成软件开发工具包（SDK），或自行开发可在游戏之间重复使用的软件开发工具包（SDK）。

此类 SDK 主要有三项职责：

- 对事件进行批处理以提高性能，并在设备上提供临时存储和重试，确保在未连接设备时不会丢失事件。
- 摄取到云中或本地的流，包括身份验证。
- 自动填充人口统计数据（操作系统系列、设备型号）并自动收集基本事件（如会话开始、会话停止等）。

从客户端流式处理的另一种常用选择是使用第三方 SDK，该 SDK 未连接到流，而是连接到第三方提供的后端（例如 Amplitude 或 Google Firebase），并从其后端批量导出事件。这种方法的优点包括后端服务中现成可用的控制面板以及 SDK 和后端的轻松设置。缺点包括缺少流式处理和实时功能。

REST API 或直接摄取到流

如果您选择为流式处理事件构建自定义后端，则可以使用客户端 SDK 直接将游戏与流式处理存储（如 Amazon Kinesis 或 Apache Kafka）集成，或者在流的前面实施通用 REST API。

REST API 的优点有：

- 将特定的流式处理存储技术与接口分离。不需要将客户端或服务器与流集成，也不需要在客户端上安装用于 Amazon Kinesis 的 Amazon SDK，这对于移动游戏或游戏机等平台可能很重要。
- 轻松支持目前不支持直接流式处理的游戏 — 您可以为它们构建 API 以模仿传统数据摄取解决方案。
- 更多身份验证和授权选项。

但是，此类 API 需要额外的开发工作，并且可能比直接摄取成本更高。游戏开发人员通常使用混合方法 — 他们直接从较新游戏中摄取到流（Amazon Kinesis Data Streams 或 Amazon Kinesis Data Firehose），并为旧游戏实施 REST API 层。

其他源

有时您可能需要来自其他来源的数据，例如运营数据库。您可以通过 Amazon Athena Federated Queries 等服务直接查询它们，或者使用 Amazon Database Migration Service（Amazon DMS）提取数据并将其存储在您的数据湖中。Amazon DMS 可以通过多种方式进行配置。一种是将数据库中的数据分载到 Amazon S3 上的数据湖中。这可以通过多种方式完成，包括完全加载、完全加载 + 更改数据捕获（CDC）以及仅 CDC。有关更多详细信息，请参阅 Amazon Database Migration Service 文档。

数据转型

事件格式和更改架构

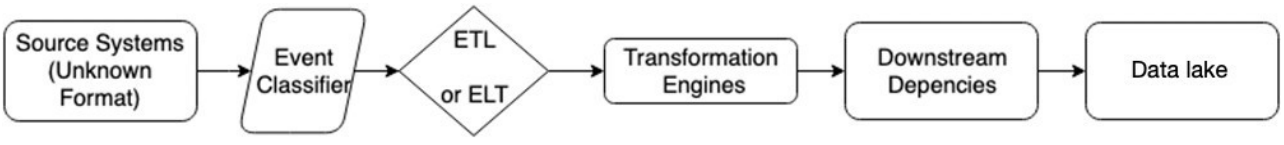
静态数据可以调整为一种格式，然而，所有的源系统都很难做到这一点。源系统生成的事件可能因使用场景和基础技术而异。建立一个可以演变并匹配如此高度差异的数据管道至关重要。例如，产品分析工作负载通常使用有限数量的字段，例如 userid、timestamp 和 productid，而游戏事件工作负载具有游戏、场景和设备类型独有的字段。

这种高级场景有两种方法：

- 亚马逊云科技中的源系统（即上游依赖项）生成和收集数据。然后，源系统使用 ETL 或提取、加载和转换（ELT）机制来修改数据。然后，为大规模数据处理构建的转换引擎处理批量和 / 或流数据。然后，我们拥有使用全部或部分数据的下游依赖项，最终进入数据湖。

从源到数据湖的数据流（数据格式已定义）

- 此方法与第一种方法类似，主要区别在于未知格式的场景。为此，可以使用数据分类器从数据存储中读取已知格式的数据，使用自定义分类器在出现新格式时读取数据。



从源到数据湖的数据流（数据格式未定义 / 未知）

传输 / 分析引擎

无论是每月生成几 GB 还是收集和存储数 PB 的用户和游戏数据，您都希望提前回答几个问题，以有效地转换数据并避免数据沼泽。

以下是一些入门问题：

- 每天和每月有多少数据？这些数据的增长情况如何？
- 数据的格式和架构是什么？
- 平均文件大小是多少？
- 有多少个源？
- 谁将成为数据的管理者？
- 需要进行哪些类型的转换？
- 使用者是谁？
- 如何保护、治理及审核数据？

以上列表并不是全部，但通过它们，您应该可以大概了解运行高效且可扩展的数据湖需要考虑的事项和做出的必要决策。在谈到提取、转换和加载数据时，没有一种工具最适合每个使用场景。相反，在具有挑战性的使用场景中，亚马逊云科技根据客户要求提供了许多工具和选项。

- Amazon Glue 是一种无服务器数据集成服务，可以轻松发现、准备和组合数据以进行分析、机器学习和应用程序开发。借助 Amazon Glue，可以使用 Amazon Glue 提供的许多不同工具进行数据发现、转换、复制和准备工作。
- 流式处理 ETL 作业 - 借助 Amazon Glue 流式处理 ETL，可以使用来自 Amazon Kinesis Data Streams 或 Apache Kafka 的数据。当需要实时转换和处理数据并加载到数据库等下游服务或 Amazon Redshift 等数据仓库时，这很有用，可用于为用户行为、欺诈活动、游戏指标等提供实时控制面板。
- 批处理 ETL 作业 - 您可能并不总是需要像使用 Amazon Glue 流式处理那样实时转换数据。相反，可以为 ETL 作业实施批处理策略，该策略可以安排为根据需要频繁或不频繁地运行。可以针对复杂类型的聚合和数据集联接运行批处理 ETL 作业。

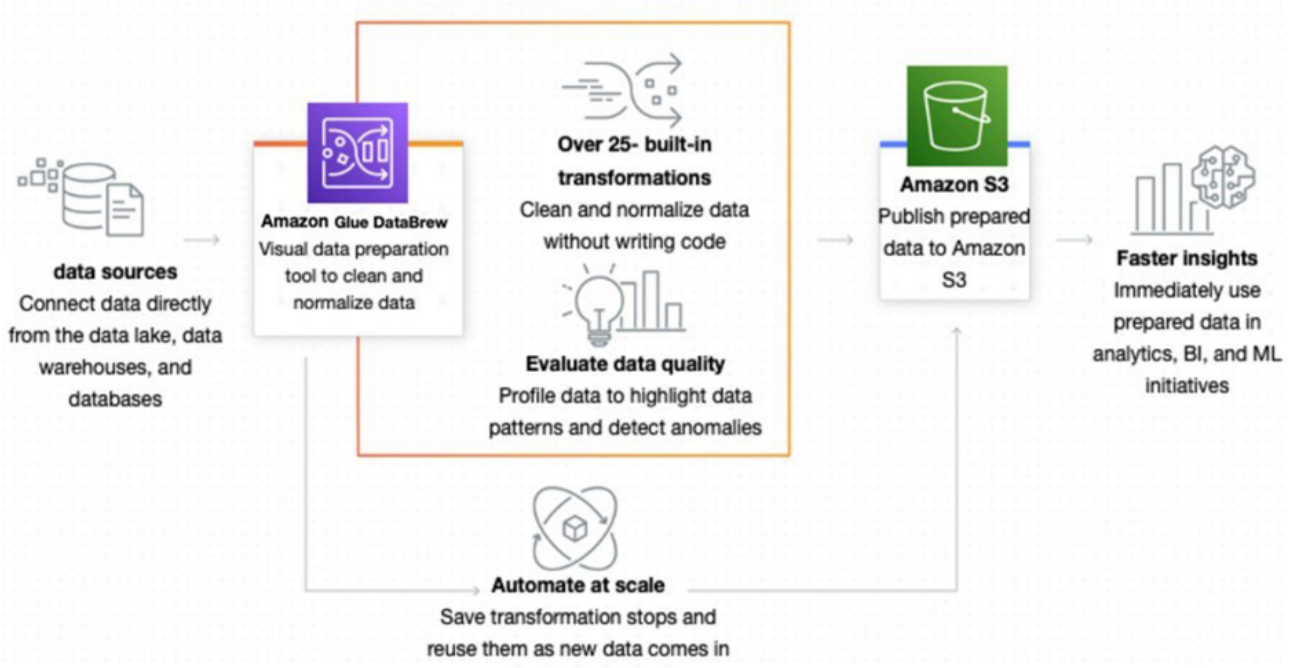
  使用批处理作业的一些常见使用场景有：

  - 将表数据从事务数据库迁移到 Amazon S3，以分析其他下游服务，例如 Amazon Athena。然后，可以使用 Amazon QuickSight 通过将 Amazon Athena 用作查询引擎来可视化这些数据。
  - 将数据从 Amazon S3 加载到 Amazon Redshift，您还可以在其中使用 Amazon Redshift 作为 BI 工具的查询引擎。
  - 预聚合和联接数据，以准备要由完全托管式机器学习服务 Amazon SageMaker 使用的数据集。

- Amazon Glue Elastic Views 让您可以轻松构建实例化视图，在多个数据存储中组合和复制数据，而无需编写自定义代码。借助 Amazon Glue Elastic Views，您可以使用熟悉的结构化查询语言（SQL） 从多个不同的源数据存储中快速创建虚拟表（实例化视图）。

  使用 Elastic Views 可以更轻松地在不同的数据存储之间复制数据和保持同步，因为您只需编写 SQL 查询即可开始在目标数据存储中创建实例化视图。您不需要再花时间开发 Amazon Lambda 函数来读取流，然后将记录写入目标数据存储，或者每五分钟左右运行一次 Amazon Glue 作业来保持数据是最新的。

- Amazon Glue DataBrew 是一种可视化数据准备工具，可让数据分析师和数据科学家轻松清理和规范数据，以便为分析和机器学习做好准备。您可以从 250 多种预构建的转换中进行选择，以自动执行数据准备任务，而无需编写任何代码。如果您想将数据集加载到可视化编辑器中，然后逐步转换数据，Amazon Glue DataBrew 是一个不错的选择。这些步骤保存在可重复使用的配方中，这样一来，你便可使用自己创建的配方一致地运行和安排 DataBrew 作业。



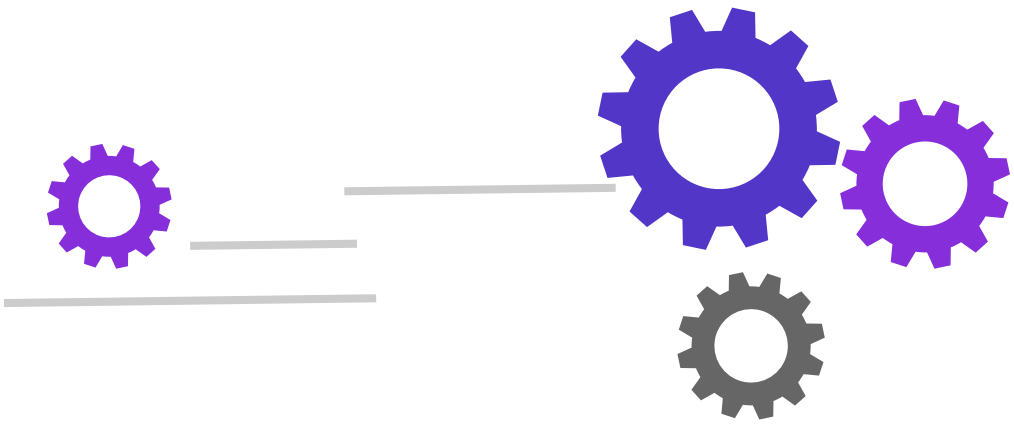
Amazon Glue DataBrew 如何帮助清理和规范数据

- Amazon EMR（以前称为 Amazon Elastic MapReduce）是一个托管集群平台，可简化在亚马逊云科技上运行大数据框架（如 Apache Hadoop 和 Apache Spark）以处理和分析大量数据。使用这些框架和相关的开源项目，您可以处理数据以用于分析目的和 BI 工作负载。Amazon EMR 还允许您将大量数据转换以及移入和移出其他亚马逊云科技数据存储和数据库，例如 Amazon S3 和 Amazon DynamoDB。如果您的使用场景涉及全天候运行的 ETL 作业，或者您需要管理其他工具（例如 Apache Hadoop、Apache HBase、Presto、Hive 等），您可能希望为 ETL 操作运行 EMR 集群。

- Amazon Athena 是一种交互式查询服务，可使用此服务通过标准 SQL 在 Amazon S3 中轻松分析数据。Athena 是一种无服务器服务，因此您无需管理任何基础设施，且只需为您运行的查询付费。您可以使用 Athena 联合查询来查询 S3 以外的数据存储，例如 RDS 数据库、本地数据库、Amazon DynamoDB、ElastiCache for Redis、Amazon Timestream 等。借助此功能，您可以使用 Athena 组合和移动来自许多数据存储的数据。Athena 是一款出色的临时数据探索工具，在根据用户查询模式构建数据湖时受益最大。

  如果您不确定查询模式是什么，一个常见的模式和良好的起点是将您的数据划分为年、月和日。通过这样做，您可以使用 Athena 仅筛选需要访问的分区，这有助于降低成本并加快每个查询的结果。如果您需要通过使用 Create Table as Select（CTAS）来处理较小的数据子集，Athena 还可以用作轻量级 ETL 工具，您可以在其中将原始数据格式化、压缩和分区为优化的格式，以供下游引擎使用。

- Amazon Redshift 是速度最快、使用最广泛的云数据仓库。Amazon Redshift 与数据湖集成，性价比是任何其他数据仓库的 3 倍。当您拥有长时间运行且需要聚合大量数据的复杂查询时，Amazon Redshift 可以帮助您为数据湖提供支持。您可以使用 Amazon Glue、Amazon EMR、Amazon Database Migration Service（Amazon DMS）、Amazon Kinesis Data Firehose 等集成工具将数据移入和移出 Amazon Redshift。



数据编录

数据目录已成为现代数据管理和治理的核心组件和核心技术。它们对于数据共享和自助分析支持、增加数据和分析的商业价值至关重要。根据 Gartner 研究报告：

“随着组织继续努力寻找、清点和分析分布广泛且多样化的数据资产，对数据目录的需求正在飙升。数据和分析领导者必须调查并采用机器学习增强的数据目录，作为其整体数据管理解决方案战略的一部分。”

成功的数据目录实施使组织能够不断提高数据分析的速度和质量，并在整个组织内更好地实现数据大众化和数据的有效使用。因此，在亚马逊云科技上为游戏设计数据湖时，选择正确的数据目录技术来实施是一个重要的决定。

数据目录是元数据的集合，结合了数据发现和管理工具，提供了跨所有数据源的数据资产清单。数据目录可帮助组织内的数据使用者更高效地发现、理解和使用数据。它可以帮助组织打破数据湖采用的障碍。妥善维护的数据目录使数据分析师和用户能够以自助服务模式工作，以快速发现可信赖的数据，评估并明智地决定使用哪些数据集，以及高效而自信地执行数据准备和分析。

在数据湖环境中，可以部署多种框架、工具和技术用于数据摄取、转换、可视化和访问控制。实现这一目标的最有效方法是维护一个中央数据目录，并在 Amazon Glue、Amazon EMR、Amazon Athena、Apache Hadoop、Apache Spark、Hive、Impala 和 Presto on Amazon 等各种框架中使用它。这使得确保元数据完整性和应用数据治理策略的工作相对容易。

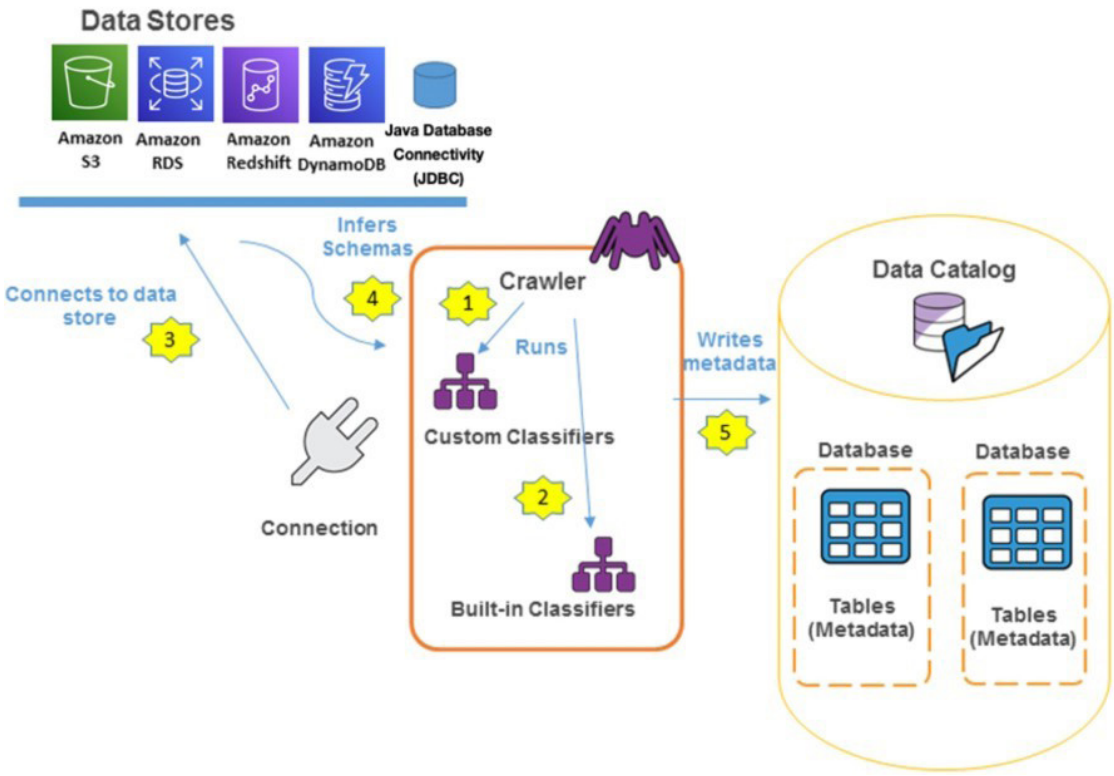
将 Amazon Glue 数据目录用作数据湖的元数据存储

Amazon Glue 数据目录是一个完全托管式持久性元数据存储，允许您存储、注释和共享元数据。它提供一个覆盖各种数据源和数据格式的统一元数据存储库，并且与 Amazon EMR、Amazon RDS、Amazon Redshift、Redshift Spectrum、Amazon Athena、Amazon Lake Formation 以及任何与 Apache Hive 元存储兼容的应用程序相集成。

使用 Amazon Glue 数据目录，您可以存储和查找元数据，跟踪数据孤岛中的数据，并使用该元数据来查询和转换数据。Amazon Glue 数据目录还通过架构更改跟踪和数据访问控制提供全面的审核和治理功能，允许您审核对数据架构的更改。这有助于确保数据不会被不当修改或无意共享。

Amazon Glue 数据目录可以扩展以满足您的许多数据编目要求和需求。Amazon Glue 数据目录表的源可以包括 Amazon S3、Amazon Kinesis、Amazon DocumentDB、Amazon DynamoDB、Amazon Redshift、MongoDB、Apache Kafka、Java Database Connectivity（JDBC）等。可以手动或通过自动化将自定义数据库和表描述以及表属性添加到目录中。例如，您可以将数据所有者、数据描述和数据敏感性添加到 Amazon Glue 表。

Amazon Glue 数据目录是数据的位置、架构和运行时指标的索引。可以使用数据目录中的信息来创建和监控 ETL 作业。每个亚马逊云科技账户可以在每个亚马逊云科技区域拥有一个 Amazon Glue 数据目录。数据目录中的信息存储为元数据表，其中每个表指定一个数据存储。通常，您运行爬网程序来清点数据存储中的数据，但还有其他方法可以将元数据表添加到数据目录中。有关如何使用 Amazon Glue 数据目录的信息，请参阅填充 Amazon Glue 数据目录。



Amazon Glue 爬网程序如何与数据存储和其他元素交互以填充数据目录

在配置 Amazon Glue 爬网程序以发现 Amazon S3 中的数据时，可以选择完全扫描（每次爬网程序运行时都会处理给定路径中的所有对象）或增量扫描（仅处理新添加文件夹中的对象）。

当对表的更改是不确定的并且可以影响任何对象或分区时，完全扫描很有用。将新分区或文件夹添加到表中时，增量爬取很有用。对于大型、频繁更改的表，可以增强增量爬取模式，以减少爬网程序确定更改了哪些对象所需的时间。

鉴于支持将 Amazon S3 事件通知作为 Amazon Glue 爬网程序源，游戏开发人员可以将 Amazon S3 事件通知配置为发送到 Amazon Simple Queue Service（Amazon SQS）队列，爬网程序使用该队列来识别新添加或删除的对象。每次运行爬网程序时，都会检查 SQS 队列是否有新事件，如果没有找到，爬网程序将停止。如果在队列中找到这些事件，则爬网程序将检查它们各自的文件夹并处理新对象。这种新模式减少了爬网程序更新大型和频繁更改的表所需的成本和时间。

使用 Amazon EMR，可以将 Spark SQL 配置为使用 Amazon Glue 数据目录作为其元存储。当您需要持久元存储或由不同集群、服务、应用程序或亚马逊云科技账户共享的元存储时，亚马逊云科技建议使用此配置。适用于 Apache Hive 元存储的 Amazon Glue 数据目录客户端是 Amazon EMR 集群上 Apache Hive 元存储客户端的开源实施，它使用 Amazon Glue 数据目录作为外部 Hive 元存储。它用作构建连接到 Amazon Glue 数据目录的 Hive 元存储兼容客户端的参考实施。它可以移植到其他 Hive 元存储兼容平台，例如其他 Hadoop 和 Apache Spark 发行版。可以从 Apache Hive 元存储迁移到 Amazon Glue 数据目录。有关更多信息，请参阅 GitHub 上的 Hive 元存储与 Amazon Glue 数据目录之间的迁移。

由于许多亚马逊云科技服务将 Amazon Glue 数据目录用作其中央元数据存储库，因此可能需要查询数据目录元数据。为此，可以在 Athena 中使用 SQL 查询。您可以使用 Athena 查询 Amazon Glue 目录元数据，例如数据库、表、分区和列。有关更多信息，请参阅查询 Amazon Glue 数据目录。

可以使用 Amazon Identity and Access Management（IAM）策略来控制对 Amazon Glue 数据目录管理的数据源的访问。这些策略使企业中的不同组能够安全地将数据发布到更广泛的组织，同时保护敏感信息。IAM 策略让您可以清晰一致地定义哪些用户可以访问哪些数据，无论其位于何处。

数据目录的其他选项

如果 Amazon Glue 数据目录不能满足您对数据编目的所有业务和技术要求，亚马逊云科技 Marketplace 上还有其他企业级解决方案可供您评估，例如 Informatica、Collibra、Alation 和 unifi。这些解决方案还可以帮助您创建和维护用于映射到基础表和列的业务词汇表。其中一些提供连接器以与原生亚马逊云科技云服务集成，例如 Amazon Glue 数据目录、S3、Athena、DynamoDB 和 Amazon Redshift。

此外，还有许多开源元数据管理解决方案可用于提高数据使用者的生产力并加快获得见解的时间。例如，Apache Atla、Amundsen、Metacat、Herd 和 Databook。

数据生命周期管理

大多数游戏开发人员在亚马逊云科技上设计和构建数据湖时都选择使用 Amazon S3 作为其主要存储平台。Amazon S3 旨在实现 99.999999999%（11 个 9）的数据持久性，并提供行业领先的可扩展性、数据可用性、安全性和性能。Amazon S3 还可以自动免费提供强大的读写一致性，并且不会更改性能或可用性。使用 Amazon S3 作为数据湖可减少过度预置、打破数据孤岛并提供无限规模。

Amazon S3 存储类专为不同的使用场景而设计，并可灵活、自动地管理您的成本。

它们包括：

- S3 Standard，适用于频繁访问数据的通用存储
- S3 Intelligent-Tiering，访问模式未知或不断变化的数据
- S3 Standard-Infrequent Access 和 S3 One Zone-Infrequent Access，适用于长期存在但访问频率较低的数据
- S3 Glacier 和 S3 Glacier Deep Archive，适用于长期存档和数字保存
- S3 on Outposts，适用于现有亚马逊云科技区域无法满足数据驻留要求时，在本地存储 S3 数据

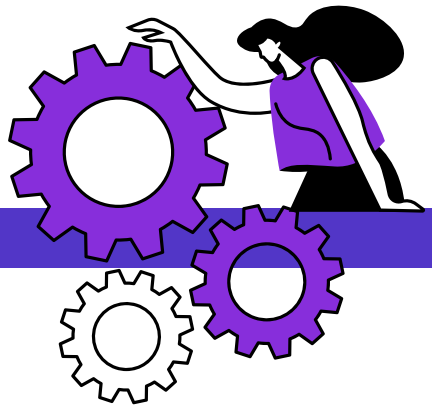
借助这种使用范围广泛且经济实惠的存储类，游戏开发人员能够在不牺牲性能或对其应用程序进行更改的情况下节省成本。可以使用 S3 存储类分析来分析存储访问模式，以帮助决定何时将正确的数据转换到正确的存储类以降低成本，并配置 S3 生命周期策略以完成转换和到期操作。

如果数据的访问模式不断变化或未知，则可以使用 S3 Intelligent-Tiering 根据不断变化的访问模式对对象进行分层，并自动节省成本，且无需运营开销。您每月只需支付少量的监控和自动分层费用。没有生命周期费用和检索费用。S3 Intelligent-Tiering 的工作原理是监控访问模式，然后将连续 30 天未访问的对象移到 Infrequent Access 层。激活一个或两个存档访问层后，S3 Intelligent-Tiering 会自动将连续 90 天未访问的对象移到 Archive Access 层，然后将连续 180 天未访问的数据移到 Deep Archive Access 层。

您可以配置 Amazon S3 生命周期，以便大规模有效地管理和优化 S3 存储成本。生命周期配置是一组规则，定义 Amazon S3 对一组对象应用的操作。

操作可分为两种类型：

- 转换操作
- 到期操作



随着对象数量的增长、多租户存储桶的创建以及工作负载大小的增加，创建和管理许多所需的生命周期配置规则可能成为一项非常复杂的任务。使用对象标记可以减少您需要管理的规则数量，从而帮助简化此过程。对象标记的目的是对存储进行分类，每个标记是一个键值对。您可以在上传新对象时对它们添加标记，也可以对现有对象添加标记。

S3 生命周期规则包含对象标记筛选器，以指定符合特定生命周期操作的对象，无论是将对象移到更经济实惠的存储类，还是在预定义的保留期后删除它们。每个规则可以包含一个或一组对象标记。然后，该规则应用于具有特定标记的对象子集。您可以指定基于多个标记的筛选器。有关更多实施详细信息，请参阅通过将对象标记与 Amazon S3 生命周期结合使用来简化数据生命周期。

Amazon S3 Storage Lens 提供跨 Amazon S3 存储的使用情况和活动的单一视图。它分析存储指标以提供上下文建议，从而帮助您优化存储成本并应用最佳实践来保护数据。您可以使用 S3 Storage Lens 生成摘要见解并确定潜在的成本节约机会。例如，您可以找到可以转换为成本更低的存储类并更好地利用 S3 存储类的对象；您可以识别并减少不完整的分段上传字节；并且可以减少保留的非当前版本的数量。

工作流编排

ETL 操作是数据湖的支柱。ETL 工作流通常涉及编排和监控许多顺序和并行数据处理任务的执行。随着数据量的增长，游戏开发人员发现他们需要迅速采取行动来处理这些数据，以确保他们做出更快、更明智的设计和业务决策。为了大规模处理数据，游戏开发人员需要弹性预置资源，以管理来自越来越多不同来源的数据，并且通常最终构建复杂的数据管道。

亚马逊云科技托管编排服务 [例如 Amazon Step Functions 和 Amazon Managed Workflows for Apache Airflow（MWAA）] 是托管工作流编排服务，有助于简化涉及多种技术的 ETL 工作流管理。这些服务提供成功管理数据处理工作流所需的可扩展性、可靠性和可用性。

Amazon Step Functions

Amazon Step Functions 是一种无服务器编排服务，可让您结合 Amazon Lambda 函数和其他亚马逊云科技服务，使用状态机构建可扩展的分布式应用程序。Step Functions 基于状态机和任务。状态机是一个工作流。任务是工作流中的一种状态，代表另一个亚马逊云科技服务执行的单个工作单元。工作流中的每一步都是一个状态。使用 Step Functions 内置控件，您可以检查工作流中每个步骤的状态，以确保您的数据处理作业按预期运行。

Step Functions 可水平扩展并提供容错工作流。您可以使用并行转换或动态并行运算更快地处理数据。
并且，它可以让您轻松地重试失败的转换，或者选择一种特定的方式来处理错误，而无需管理复杂的过程。Step Functions 为您管理状态、检查点和重启，以确保您的工作流按顺序运行。Step Functions 可以与多种亚马逊云科技服务集成，包括：

- Amazon Lambda
- Amazon Fargate
- Amazon Batch
- Amazon Glue
- Amazon Elastic Container Service (Amazon ECS)
- Amazon Simple Queue Service (Amazon SQS)
- Amazon Simple Notification Service (Amazon SNS)
- Amazon DynamoDB 等

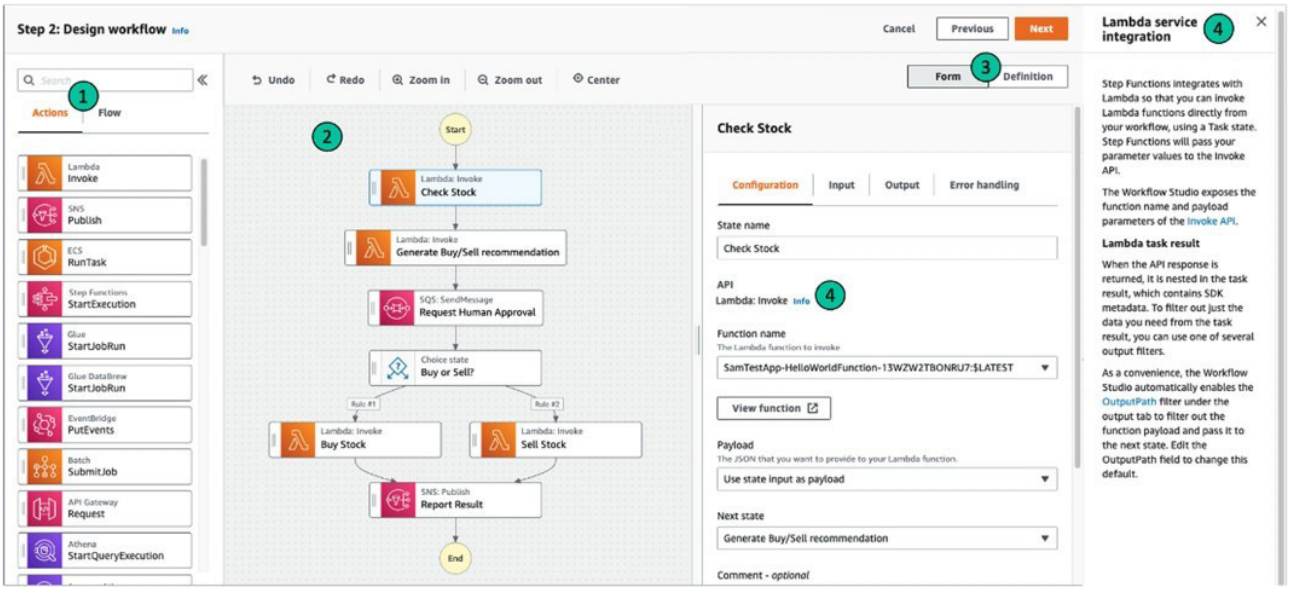
Step Functions 有两种工作流类型：

- **标准工作流**具有精确一次的工作流转换，最长可运行一年。
- **快速工作流**具有至少一次的工作流转换，最长可运行五分钟。

标准工作流非常适合长期运行、可审核的工作流，因为它们可显示执行历史和可视化调试。快速工作流非常适合高事件率工作负载，例如流数据处理。您的状态机转换的行为会有所不同，具体取决于您选择的类型。有关详细信息，请参阅标准与快速工作流。

根据数据处理需求，Step Functions 直接与亚马逊云科技提供的其他数据处理服务集成，例如用于批处理的 Amazon Batch、用于大数据处理的 Amazon EMR、用于数据准备的 Amazon Glue、用于数据分析的 Athena 和用于计算的 Amazon Lambda。使用 Amazon Redshift（Lambda、Amazon Redshift 数据 API）运行 ETL/ELT 工作流演示了如何使用 Step Functions 和 Amazon Redshift 数据 API 来运行将数据加载到 Amazon Redshift 数据仓库中的 ETL/ELT 工作流。管理 Amazon EMR 作业演示了 Amazon EMR 和 Amazon Step Functions 的集成。

Workflow Studio for Amazon Step Functions 是一种低代码可视化工作流设计器，可让您通过编排亚马逊云科技服务来创建无服务器工作流。它使游戏开发人员可以轻松构建无服务器工作流，并使游戏开发人员能够专注于构建更好的游戏玩法，同时减少为工作流定义编写配置代码和构建数据转换所花费的时间。使用拖放来创建和编辑工作流，控制如何筛选或转换每个状态的输入和输出，以及配置错误处理。当您创建工作流时，Workflow Studio 会验证您的工作并生成代码。



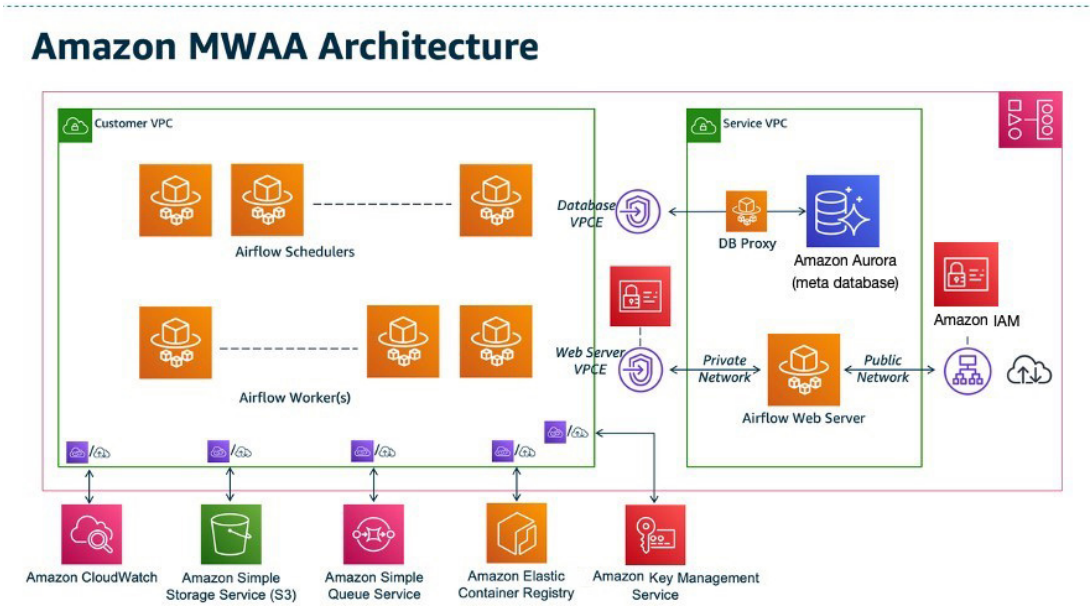
设计工作流示例

Amazon Managed Workflows for Apache Airflow (MWAA)

Apache Airflow 是一种开源工具，用于以编程方式创作、调度和监控工作流。Amazon Managed Workflows for Apache Airflow（MWAA）是一种适用于 Apache Airflow 的托管编排服务，可以更轻松地构建、操作和扩展端到端数据管道，且无需管理底层基础设施就可以实现可扩展性、可用性和安全性。Amazon MWAA 可以帮助降低运营成本和工程开销。

MWAA 的弹性伸缩机制会自动增加 Apache Airflow 工作线程的数量以响应排队的任务，并在没有更多任务排队或运行时处理多余的工作线程。您可以直接在 MWAA 控制台上配置任务并行度、弹性伸缩和并发设置。

Amazon MWAA 使用以 Python 编写的有向无环图（DAG）来编排和安排您的工作流。要在 Amazon MWAA 环境中运行 DAG，请将文件复制到 Amazon S3，然后让 Amazon MWAA 知道您的 DAG 和支持文件在 Amazon MWAA 控制台上的位置。Amazon MWAA 负责在工作线程、调度程序和 Web 服务器之间同步 DAG。



MWAA 部分中包含的所有组件在您的账户中都显示为单个 Amazon MWAA 环境

Amazon MWAA 支持与 Amazon Athena、Amazon Batch、Amazon CloudWatch、Amazon DynamoDB、Amazon DataSync、Amazon EMR、Amazon Fargate、Amazon EKS、Amazon Kinesis Data Firehose、Amazon Glue、Amazon Lambda、Amazon Redshift、Amazon SQS、Amazon SNS、Amazon SageMaker 和 Amazon S3 的开源集成，并支持数百个内置和社区创建的运算符和传感器以及第三方工具（例如 Apache Hadoop、Presto、Hive 和 Spark），用于执行数据处理任务。

代码示例可用于更快的集成。例如，将 Amazon MWAA 与 Amazon EMR 结合使用演示了如何使用 Amazon EMR 和 Amazon MWAA 启用集成。使用 Apache Hive 和 Hadoop 创建自定义插件将引导您完成在 Amazon MWAA 环境中使用 Apache Hive 和 Hadoop 创建自定义插件的步骤。

数据安全和治理

数据湖内的安全性在两个方面得到解决。

- 静态数据安全
- 传输中数据安全

静态数据 - 使用亚马逊云科技，您有许多选项可以潜在地满足您的加密需求。在数据湖框架内，数据主要驻留在主数据存储（例如 Amazon S3）中，在某些使用场景中则驻留在辅助数据存储（例如非主要区域中的 S3 或 Outposts）中。**有关最新详细信息，请参阅最新的亚马逊云科技文档：**

- 对于 S3，请参阅使用服务器端加密保护数据 - Amazon Simple Storage Service
- 对于 Outposts，请参阅 Amazon Outposts 中的数据保护 - Amazon Outposts
- 有关加密的重要性，请参阅亚马逊云科技安全博客上的加密的重要性以及亚马逊云科技如何提供帮助

传输中数据 - 为了保护传输中数据，亚马逊云科技鼓励客户使用多级方法。亚马逊云科技数据中心之间的所有网络流量都在物理层透明地加密。使用支持的 Amazon EC2 实例类型时，VPC 内和跨区域的对等 VPC 之间的所有流量都在网络层透明地加密。在应用程序层，客户可以选择是否以及如何使用传输层安全性（TLS）之类的协议进行加密。所有亚马逊云科技服务终端节点都支持 TLS 以创建安全的 HTTPS 连接来发出 API 请求。**有关最新详细信息，请参阅我们的最新文档：**

- 数据加密的逻辑分离
- 实施安全密钥和证书管理
- 提醒意外的数据访问
- 使用 TLS 对经过 EC2 或 EKS 的所有网络流量进行身份验证
- 安全性支柱 - Amazon Well-Architected Framework

如何规范数据（GDPR、CCPA 和 COPPA）

GDPR、CCPA、COPPA 等法规要求运营商必须实施遗忘权。这意味着必须具有在指定天数内根据请求删除特定用户的个人身份信息（PII）的技术能力。对于以读取优化格式（如 Apache Parquet）存储在数 TB 数据湖中的分析数据，这通常是一个挑战。原因有两点：

首先，您需要扫描数据湖中的所有数据，以识别包含具有您需要的用户 ID 的记录的分区。

其次，您不能从 Parquet 文件中删除单个记录或更新分区内的单个 Parquet 文件 - 必须重新计算和重新写入整个分区。

在大数据湖上，这两种操作都非常耗时且耗费资源，因为它们涉及许多元数据操作（例如 S3 LIST），并且它们必须扫描所有数据。

解决这个问题的主要方法有两种：避免将 PII 存储在数据湖中，或者实施额外的元数据层以减少操作次数和扫描数据量以搜索特定用户 ID，并删除相应的数据。

遵守监管框架的最安全方法是不在数据湖中存储 PII。根据您需要的分析类型，这可能可行，也可能不可行。例如，您可以摆脱数据湖中任何形式的用户标识符，但这将使分析通常需要的任何类型的每用户聚合成为不可能。

一种常见的方法是用替代标识符替换真实的用户标识符，将这些映射存储在数据湖之外的单独表中，并避免存储除这些内部 ID 之外的任何其他类型的用户信息。这样，您仍然可以在不知道用户是谁的情况下分析用户行为。在这种情况下，数据湖本身不存储任何 PII。从外部表中删除特定用户的映射就足以使数据湖中的数据记录与用户无关。请咨询您的法务部，了解这是否足以满足您案件的监管要求。

还有许多技术可以模拟数据，例如屏蔽、散列、模糊、以随机百分比修改数字等。如果您选择将可识别个人身份的数据存储在数据湖中，则可以使用这些技术来模拟这些数据。有一些第三方产品，例如 Dataguise 或 Collibra（在亚马逊云科技 Marketplace 上提供），可以自动执行此操作。

另一种方法是优化用户 ID 查找和删除性能，通常通过实施额外的元数据层来实现。如果法务部告诉您模拟或替代 ID 不足以满足监管要求，则可能需要这样做。您可以构建自己的解决方案来通过 user_id 维护索引，或者在数据湖中使用 Apache Hudi 等开放数据格式。

您可以构建自己的索引解决方案。例如，您可以按用户 ID 维护数据湖中的文件索引，并在使用 Lambda 函数添加新文件时对其进行更新。然后删除作业只能直接重写那些文件。如何在亚马逊云科技数据湖中删除用户数据博客文章中描述了此类解决方案的示例。

对于提供额外好处（例如版本控制、事务和改进性能）的替代方法，请对数据湖使用 Apache Hudi、Delta Lake 或 Apache Iceberg 数据格式中的一种。此类解决方案在 Parquet 或 AVRO 等开放文件格式之上维护额外的元数据，以启用更新插入 / 删除。使用这些解决方案时，您可以使用 SQL 查询或 Spark 作业按 ID 删除特定用户数据。这在读取方面仍然很昂贵，但在写入方面会快得多。通常，即使在大数据湖中，此类删除作业也可以在合理的时间内工作。

缺点是您需要使用支持此类格式的分析引擎和 ETL 工具。亚马逊云科技服务的支持程度因格式而异。Hudi 可能是一个不错的起点。Amazon Redshift、Athena、Amazon Glue 和 EMR 都支持它。特定功能支持也可能因服务而异。请务必查看特定服务的文档。

数据发现

数据发现是指对存储在亚马逊云科技上的数据中的模式进行整体识别。Amazon Macie 等服务提供托管数据安全和数据隐私服务，该服务使用机器学习和模式匹配来发现和保护您在亚马逊云科技中的敏感数据。

数据发现的使用场景包括：

- 从业务 SLA 中存储的数据提取价值
- 防止敏感数据类型被摄取到数据湖中

有关发现敏感数据的示例，请参阅使用 Macie 发现敏感数据作为自动化数据管道的一部分。

有关将 Amazon RDS 作为数据存储的信息，请参阅使用 Macie 为 Amazon RDS 数据库启用数据分类。

有关在 DynamoDB 中检测敏感数据的信息，请参阅使用 Macie 在 DynamoDB 中检测敏感数据。

数据治理

数据治理是指从组织内部数据的可用性、质量、可用性、沿袭和安全性方面对数据资产进行整体管理。

数据治理在很大程度上取决于业务策略，通常涵盖以下方面：

- **数据所有权和责任**
  - 有适当的结构来确定原始、策划和处理格式的数据的权限和角色。
  - 能够监控在亚马逊云科技内进行的所有 API 调用，这对于审核您的亚马逊云科技账户中的任何可疑操作至关重要。

- **实施策略和规则**
  - 根据组织结构制定有关数据将用于什么以及谁可以访问它的策略。
  - 为具有适当权限的用户实现数据共享、数据质量、警报和更快数据访问的自动化。

- **技术流程、工具和实践**
  - Amazon Lake Formation 是一种集成的数据湖服务，可让您轻松提取、清理、编目、转换和保护您的数据，并使其可用于分析和机器学习。Lake Formation 为您提供了一个中央控制台，您可以在其中发

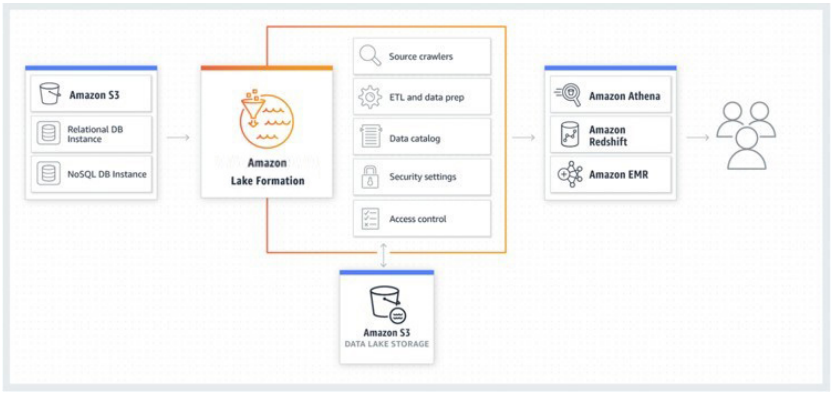
现数据源、设置转换作业以将数据移动到 Amazon S3 数据湖、删除重复和匹配记录、编目数据以供分析工具访问、配置数据访问和安全策略，以及审核和控制来自亚马逊云科技分析和机器学习服务的访问。Lake Formation 提供的功能可让您轻松治理和管理数据湖中的数据：

• Governed 表使您的数据湖能够支持原子性、一致性、隔离性、持久性（ACID）事务，您可以在其中可靠地添加和删除 S3 对象，同时保护数据目录的完整性。

• 行级安全性提供表、列和行级别的安全性，并直接从 Lake Formation 进行管理。您可以将行级安全性应用于基于 S3 的表，例如 Amazon Glue 数据目录、Amazon Redshift 数据共享、Governed、Apache Hive、Apache Hudi 等。

• Amazon Glue 数据目录是一个持久的元数据存储，包含有关您的数据的信息，例如格式、结构、大小、数据类型、数据字段、行数等。您可以在与许多原生亚马逊云科技服务集成的 Amazon Glue 数据目录中注册数据库和表。

• Amazon CloudTrail 用于审核和监控在您的亚马逊云科技账户中进行的 API 调用，并直接内置于 Lake Formation。



Lake Formation 管理此处显示的所有任务，并与亚马逊云科技数据存储和服务集成

• Amazon Glue DataBrew 为您提供可视化界面来帮助您以可重复的自动化方式准备、分析数据和跟踪沿袭。DataBrew 可以获取传入的原始数据并运行您为其设置的自动化配方，例如删除列、对一些值或标题进行格式设置，然后将处理后的数据写入 Amazon S3 以供下游的其他分析或机器学习服务使用。Amazon Glue DataBrew 还可用于数据质量自动化和警报。

• Apache Atlas、Deequ 和 Apache Ranger 等开源工具也很受客户的欢迎，这些客户希望能够管理和设置基础设施以及获取所需配置，因为这些工具本身并不与 Athena 或 Glue 等其他亚马逊云科技服务集成。

与数据管理不同的是，数据治理的主要关注点是在制定合理的业务决策时使用数据，数据治理关注的是您在整个组织中管理和使用数据的纪律性。如果没有数据治理和适当的维护机制，数据湖就有可能成为数据沼泽，并成为断开连接的数据孤岛的集合。

数据可视化

Amazon QuickSight 是一种无服务器 BI 工具，使用户能够根据您的数据创建可视化效果和控制面板。QuickSight 可以连接到许多不同的数据存储，例如 Amazon S3、本地数据库、Salesforce、Amazon RDS、Amazon Redshift 等。使用 QuickSight，您可以为许多特定于游戏的使用场景创建控制面板。QuickSight 将机器学习见解直接内置到平台中，它可以潜在地检测欺诈或玩家活动中的异常情况。QuickSight 还可以与 Amazon CloudTrail 结合使用，以帮助监控、可视化和审核您的数据湖操作。

有关更多信息，请参阅使用 Amazon CloudTrail 和 Ingest 在 Amazon QuickSight 上运行使用情况分析以及摄取和可视化游戏的流数据。

监控

概括地说，您的数据湖中有两种类型的监控：

- 监控资源
- 监控这些资源中的数据质量

此方面包括几个选项：

- **监控亚马逊云科技资源** - Amazon CloudWatch 根据您的环境收集和跟踪指标、监控日志、设置阈值并触发警报。CloudWatch 可以监控亚马逊云科技资源，例如 Amazon EC2 实例、Amazon S3、Amazon EMR、Amazon Redshift、Amazon DynamoDB 和 Amazon Relational Database Service (RDS) 数据库实例，以及其他数据湖应用程序和服务生成的自定义指标。通过 CloudWatch，您可以全面了解系统范围内的资源利用率、应用程序性能和运行状况。您可以利用这些见解主动应对问题并保持数据湖应用程序和工作流平稳运行。

  有关更多信息，请参阅监控和优化数据湖环境以及亚马逊云科技上的数据湖实施指南。

  Datadog、New Relic 和 Splunk 等亚马逊云科技合作伙伴提供与 Amazon CloudWatch 类似的功能，并且可以作为我们客户的选项。

- **监控数据质量** - 客户每天都会在亚马逊云科技上积累数 PB 的数据。面对这种持续不断的数据流，这些客户很难维持他们管理的数据质量。幸运的是，亚马逊云科技在这方面有一些选择。例如，在数据层，您可以将 Deequ 与 Amazon Glue、Amazon Amplify 和 DynamoDB 结合使用，以创建数据质量和分析框架。在机器学习方面，您可以使用 SageMaker 在生产环境中自动监控机器学习模型，并在出现数据质量问题时获得通知。

成本优化

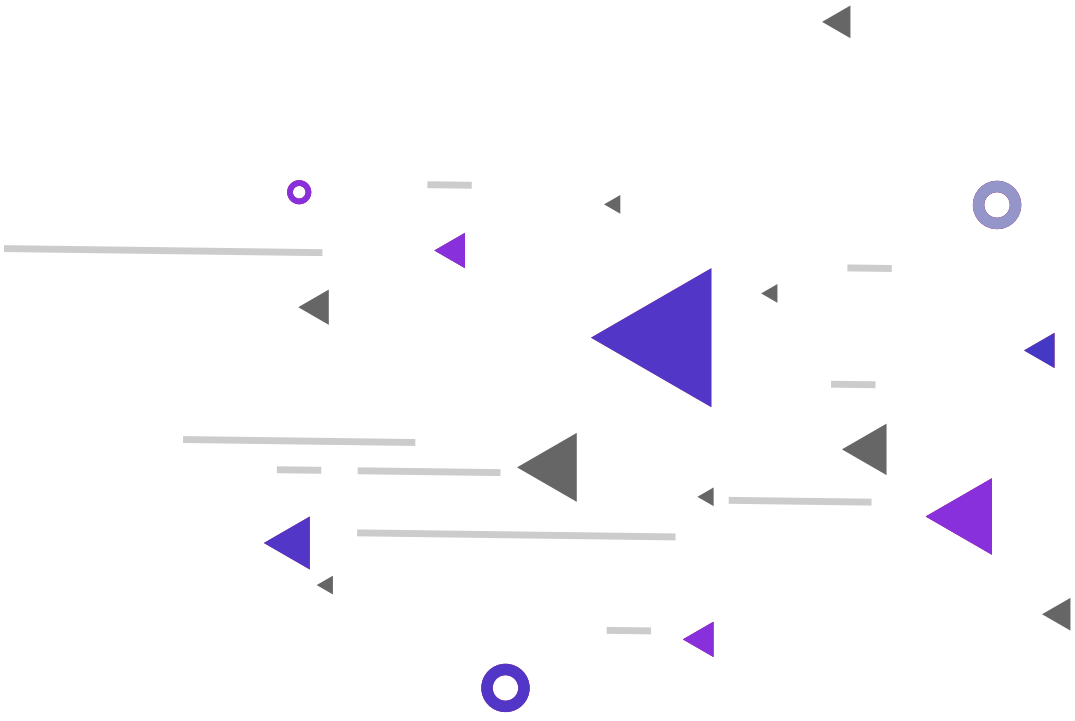
成本源自对数据湖的基础服务的使用。使用 Lake Formation 中的功能不收取额外费用，但是，在使用 Amazon Glue、Amazon S3、Amazon EMR、Amazon Athena、Amazon Redshift、Amazon Kinesis 等服务时，将收取标准使用费率。

在亚马逊云科技上构建数据湖时需要考虑的一些成本因素包括：

- **成本驱动因素** - 亚马逊云科技存在三个主要成本驱动因素：计算、存储和出站数据传输。这些因亚马逊云科技产品、定价模型和区域（数据位置）而异。
- **成本优化策略** - 基于成本的三个主要驱动因素，我们可以利用亚马逊云科技上大量经济实惠的解决方案来帮助满足正确的使用场景和预算。例如：
    - **计算：按需型实例与实惠配套（Savings Plans）与竞价型实例与预留** - 客户可以根据工作负载类型和成本选择合适的模型。对于 Amazon Redshift 和 EMR 等服务，重要的是要“调整”您的解决方案以最适合您的业务需求，然后使用预留实例进一步降低成本。请参阅 Amazon EC2 定价。
    - 大多数亚马逊云科技分析服务都是无服务器的，这意味着您只需为使用的计算付费，仅此而已。在使用无服务器服务优化成本时，有一些关键注意事项：
        - 尽可能避免重新处理数据，只处理您需要的数据。例如，这可以通过在处理数据时使用 Amazon Glue 作业书签来完成，或者在使用 Athena 时仅针对分区中需要的数据来完成。
        - 为数据处理分配正确数量的资源。
        - 在 Amazon S3 中存储数据时正确的分区、压缩、存储和生命周期策略。

- **存储：**
  - S3 层与 EBS 与 EFS 与 Amazon FSx 与 Snow 系列：— 您可以从亚马逊云科技上提供的各种存储工具中进行选择。如果您更倾向于让亚马逊云科技根据您的独特使用情况决定正确的 S3 层，也可以使用智能选项。
  - **出站数据传输** - 客户无需为跨所有区域的所有服务的入站数据传输付费。从亚马逊云科技到 Internet 的数据传输按服务收费，费率特定于源区域。请参阅每项服务的定价页面以获取更详细的定价信息。最佳实践包括：
    - 连接到亚马逊云科技服务时，避免通过 Internet 路由流量。使用 VPC 终端节点（如果可用）。
    - 考虑使用 Amazon Direct Connect 连接到本地网络。
    - 避免跨区域成本，除非立项需要。
    - 创建数据传输成本分析控制面板，以战略性地管理容量和使用情况。

- **成本计算工具 -** 有关亚马逊云科技每月成本的更详细计算，请参阅亚马逊云科技价格计算器。https://calculator.aws/ 对于您的亚马逊云科技账户上的现有基础设施，请参阅亚马逊云科技管理控制台中的亚马逊云科技成本管理控制台，以详细分析您的使用情况（需要登录）。在您的亚马逊云科技账户中订阅 Trusted Advisor 报告以进行成本优化检查，从而为您节省资金（需要登录）。例如，您的亚马逊云科技账户中可能有可以删除的未使用资源。



延伸阅读

有关更多信息，请参阅：

- 构建大数据存储解决方案（数据湖）以实现最大灵活性（亚马逊云科技白皮书）
- 使用 CDK 管道部署数据湖 ETL 作业（博客文章）
- 使用 Amazon Step Functions 和 Amazon Lambda 编排多个 ETL 作业（GitHub 上的示例）
- 使用 Amazon Step Functions 和 Apache Livy 编排 Apache Spark 应用程序（博客文章）
- 使用 Amazon MWAA、Amazon Step Functions、Amazon Glue 和 Amazon EMR 构建复杂的工作流（博客文章）
- Field Notes：如何使用亚马逊云科技云开发工具包（CDK）构建 Amazon Glue 工作流（博客文章）
- 使用 Amazon Glue DataBrew 和 Amazon Lambda 建立自动化数据质量工作流和警报（博客文章）
- 使用 Amazon Lake Formation 发现元数据：第 1 部分（博客文章）
- 使用 Amazon Lake Formation 发现元数据：第 2 部分（博客文章）
- 使用 Amazon Lake Formation 的有效数据湖，第 1 部分：受管表入门（博客文章）
- 亚马逊云科技数据和分析能力合作伙伴

贡献者

本文的贡献者包括：

- Karthik Kumar Odapally，亚马逊云科技游戏解决方案高级架构师
- Jackie Jiang，亚马逊云科技游戏解决方案高级架构师
- Eugene Krasikov，亚马逊云科技游戏解决方案高级架构师
- Michael Hamilton，亚马逊云科技 SA 分析专家

中文校对：

- 胡文静 ，亚马逊云科技合作伙伴解决方案架构师
- 黄家曦，亚马逊云科技解决方案架构师
- 唐健，亚马逊云科技解决方案架构师

文档修订

日期	描述
2022 年 5 月 11 日	首次发布
2022 年 8 月 9 日	中文校对发布

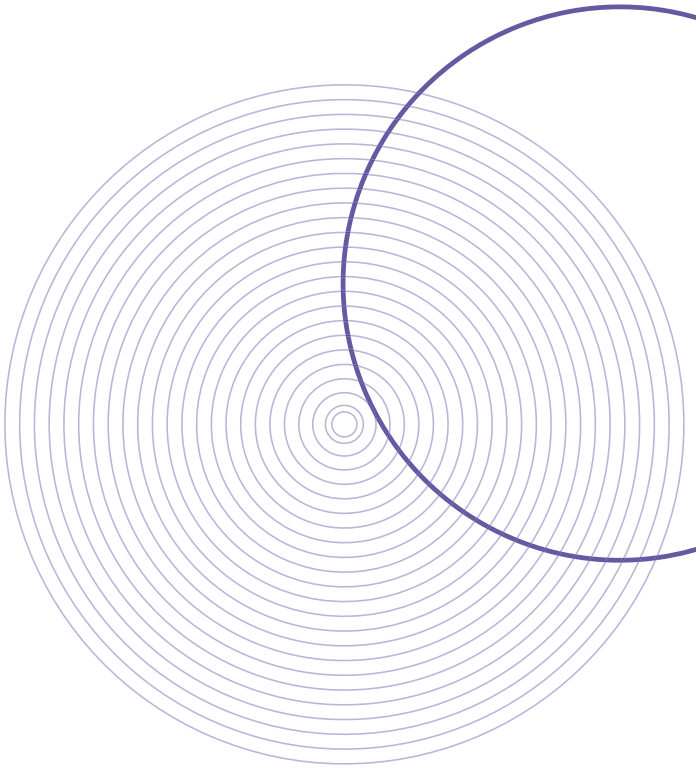
声明

客户负责对本文档中的信息进行独立评估。

本文档：(a) 仅供参考；(b) 代表亚马逊云科技当前的产品服务和实践，如有变更，恕不另行通知；(c) 不构成亚马逊云科技及其附属公司、供应商或授权商的任何承诺或保证。亚马逊云科技产品或服务均“按原样”提供，没有任何明示或暗示的担保、声明或条件。亚马逊云科技对其客户的责任和义务由亚马逊云科技协议决定，本文档与亚马逊云科技和客户之间签订的任何协议无关，亦不影响任何此类协议。

亚马逊云科技词汇表

有关最新的亚马逊云科技术语，请参阅亚马逊云科技一般参考中的亚马逊云科技词汇表。




亚马逊云科技