# AWS 上的数据湖和分析

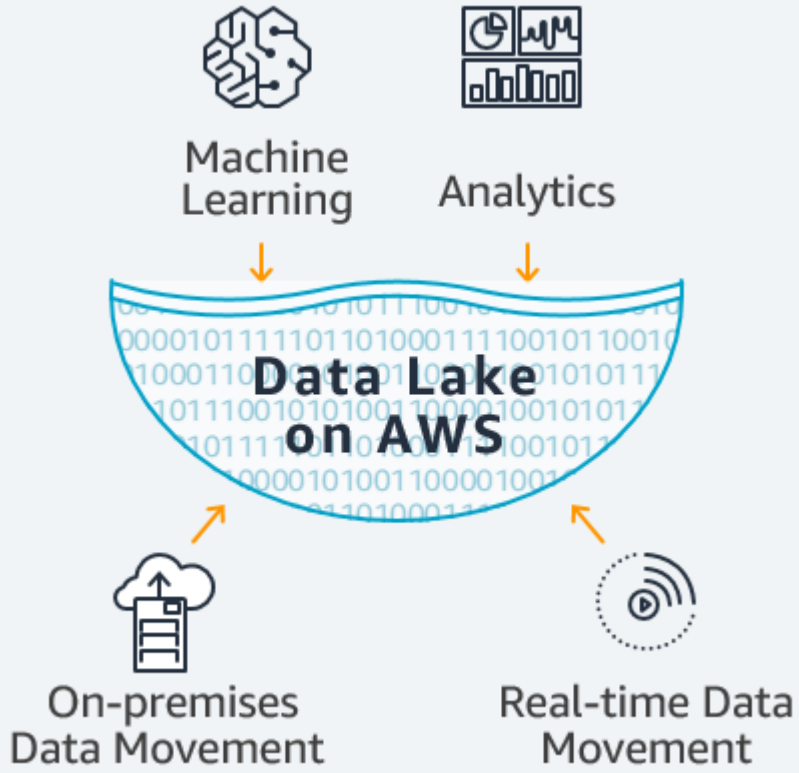
最全面、安全、可扩展且经济高效的服务组合，旨在帮助您构建数据湖和分析解决方案

AWS 供应的集成服务套件可提供快速而轻松地构建和管理数据湖以进行分析所需的每个要素。AWS 支持的数据湖可以采用传统数据孤岛和数据仓库无法胜任的方法，处理组合不同类型的数据和分析方法以获得更深层次的见解所必需的扩展性、敏捷性和灵活性。AWS 为客户提供最广泛的分析阵列和机器学习服务，以轻松访问所有相关数据，而不会妨碍安全性或监管。

AWS 上拥有数据湖和分析的组织数量比任何其他地方都多。像 NASDAQ、Zillow、Yelp、iRobot 和 FINRA 这样的客户都委托 AWS 运行其关键业务分析工作负载。

# AWS 上的数据湖和分析

为了构建您的数据湖和分析解决方案，AWS 提供了最全面的服务集来移动、存储和分析您的数据。




数据移动

从本地实时导入数据。

### 数据湖

安全地存储任何类型的数据，规模从千兆字节到艾字节不等。

### 分析

利用最齐全的分析服务选项分析您的数据。

### Machine Learning

预测未来成果并指定快速响应举措。

## 数据移动

在 AWS 上构建数据湖的第一步是将数据移动到云中。带宽和传输速度的物理限制在不会造成重大中断、高成本和长时间的前提下，限制了移动数据的能力。为了使数据传输变得简单灵活，AWS 提供了最广泛的选项来支持将数据传输到云。

要为您的数据湖构建 ETL 作业和 ML 转换，应当了解 AWS Lake Formation。

### 本地数据移动

AWS 提供了多种方法将数据从数据中心移动到 AWS。要在您的网络和 AWS 之间建立专用网络连接，可以使用 AWS Direct Connect。要使用物理设备将 PB 级到 EB 级数据移动到 AWS，您可以使用 AWS Snowball 和 AWS Snowmobile。 要使本地应用程序将数据直接存储到 AWS，您可以使用 AWS Storage Gateway。

### 实时数据移动

AWS 提供了多种方法来提取通过新来源（如网站、移动应用程序和连接互联网的设备）生成的实时数据。为了简化流数据或 IoT 设备数据的捕获和加载，您可以使用 Amazon Kinesis Data Firehose、Amazon Kinesis Video Streams 和 AWS IoT Core。

# 数据湖

数据可移动到云中后，AWS 便可以安全、大规模地利用 Amazon S3 和 Amazon Glacier 存储任何格式的数据。为了使最终用户能够轻松发现要在分析中使用的相关数据，AWS Glue 会自动创建一个可供用户搜索和查询的目录。

为了更快地构建一个安全的数据湖，请了解更多有关 AWS Lake Formation 的信息。

## 对象存储

## Amazon S3

Amazon S3 是一种安全、高度可扩展、持久的对象存储，具有毫秒级的数据访问延迟。S3 专为从任意位置存储任意类型的数据而构建，包括来自网站和移动应用程序、公司应用程序的数据以及来自 IoT 传感器或设备的数据。它专为存储和检索任何数量的数据而构建，具有无与伦比的可用性，并且不依赖其他服务，可提供 99.999999999%（11 个 9）的持久性。S3 Select 专注于数据读取和检索，可将响应时间缩短多达 400%。S3 提供了全面的安全性和合规性功能，可满足最严格的监管要求。

## 备份与存档

## Amazon Glacier

Amazon Glacier 是一种安全、持久且超低成本的存储，适用于长期备份和存档，允许用户在短短几分钟内访问数据，同样 Glacier Select 也只读取和检索需要的数据。它能够提供 99.999999999%（11 个 9）的持久性以及全面的安全与合规功能，可以帮助满足最严格的监管要求。客户能以每月每 GB 低至 0.004 USD 的价格存储数据，与本地解决方案相比，显著降低了成本。

## 数据目录

### AWS Glue

AWS Glue 是一种完全托管的服务，提供数据目录以使数据湖中的数据可被发现，并且能够执行提取、转换和加载 (ETL) 以准备数据进行分析。数据目录会自动创建为所有数据资产的持久元数据存储，支持在一个视图中搜索和查询所有数据。

## 分析

AWS 提供了在数据湖上运行的最广泛、最具成本效益的分析服务集合。每项分析服务都专门为广泛的分析用例而构建，例如交互式分析、使用 Apache Spark 和 Hadoop 的大数据处理、数据仓库、实时分析、运营分析、控制面板和可视化。

为了管理对数据湖中的数据进行的安全自助访问，请了解更多有关 AWS Lake Formation 的信息。

### 交互式分析

### Amazon Athena

对于交互式分析，Amazon Athena 可以使用标准 SQL 查询直接在 S3 和 Glacier 中分析数据。Athena 属于无服务器服务，因此无需设置或管理基础设施。您可以立即开始查询数据，在几秒钟内获得结果，并且仅需为您运行的查询付费。只需指向您存储在 Amazon S3 中的数据，定义架构并使用标准 SQL 开始查询即可。多数结果可在数秒内获取。

### 大数据处理

### Amazon EMR

对于使用 Spark 和 Hadoop 框架的大数据处理，Amazon EMR 提供了一种托管服务，可以轻松、快速且经济高效地处理海量数据。Amazon EMR 支持 19 种不同的开源项目，包括 Hadoop、Spark、HBase 和 Presto，通过托管 EMR Notebooks 进行数据工程、数据科学开发和协作。每个项目都会在一个版本发布后的 30 天内在 EMR 中更新，轻松地确保您拥有来自社区的最新、最好的内容。

## 数据仓库

### Amazon Redshift

对于数据仓库，Amazon Redshift 提供了针对 PB 级结构化数据运行复杂分析查询的功能，并且包括 Redshift Spectrum，可直接针对 S3 中的 EB 级结构化或非结构化数据运行 SQL 查询，而无需执行不必要的数据移动。Amazon Redshift 的成本不到传统解决方案的十分之一。您可以从小规模开始，最初每小时仅需支持 0.25 USD，并以每 TB 每年 1000 USD 的价格扩展到 PB 级数据。

## 实时分析

### Amazon Kinesis

对于实时分析，Amazon Kinesis 可以轻松收集、处理和分析流数据，如 IoT 遥测数据、应用程序日志和网站点击流。这让您可以对传入数据湖的数据进行实时处理和分析并做出响应，无需等到收集完全部数据后才开始进行处理。

## 运营分析

### Amazon Elasticsearch Service

对于运营分析（如应用程序监控、日志分析和点击流分析），Amazon Elasticsearch Service 允许您近乎实时地搜索、浏览、过滤、聚合和可视化数据。Amazon Elasticsearch Service 可以提供各种易于使用的 Elasticsearch API 和实时分析功能，还可以实现生产工作负载需要的可用性、可扩展性和安全性。

## 控制面板和可视化

### Amazon QuickSight

对于控制面板和可视化，Amazon QuickSight 为您提供快速，基于云的商业分析服务，使您可以轻松构建可从任何浏览器或移动设备访问的精致可视化效果和内容丰富的控制面板。

## Machine Learning

对于预测分析用例，AWS 提供了一系列广泛的机器学习服务，以及在 AWS 上的数据湖上运行的工具。我们的服务基于亚马逊长期积累的知识和能力，其中 ML 支持着亚马逊的推荐引擎、供应链、预测、履约中心和容量规划。

### 框架和接口

对于专家级机器学习从业者和数据科学家，AWS 提供了 AWS Deep Learning AMI，支持使用 ML 和 DL 优化型 GPU 实例轻松构建深度学习模型和集群。AWS 支持所有主流机器学习框架，包括 Apache MXNet、TensorFlow 和 Caffe2，以便您可以引入或开发您选择的任何模型。这些功能提供了深度学习和机器学习工作负载所需的无与伦比的功能、速度和效率。

### 平台服务

对于想要深入了解 ML 的开发人员，Amazon SageMaker 这项项平台服务通过提供连接到训练数据、选择和优化最优算法和框架，以及在 Amazon EC2 的自动扩展集群上部署模型所需的一切，让构建、训练和部署 ML 模型的整个过程得以简化。SageMaker 还包含托管的 Jupyter 笔记本，可以轻松浏览和可视化在 Amazon S3 中存储的训练数据。

### 应用程序服务

对于希望在其应用程序中插入预构建的 AI 功能的开发人员，AWS 提供面向解决方案的 API，支持计算机视觉和自然语言处理。这些应用程序服务允许开发人员在不必开发和训练自己模型的前提下，为其应用程序添加智能。

# 基于 AWS 构建的数据湖与分析远超过基于其他平台构建的数量


FINRA


airbnb


Nasdaq


amazon


21ST
CENTURY
FOX


Sysco


Expedia


EPIC
GAMES
FORTNITE


HEARST


Zillow


yelp


AMGEN®





# 为什么要使用 AWS 上的数据湖和分析？

## 灵活性高，选择丰富

AWS 提供一系列最广泛的分析工具和引擎，使用开放格式和开放标准分析数据。您可以将数据存储在您自己选择的基于标准的数据格式中，例如 CSV、ORC、Grok、Avro 和 Parquet，并且可以灵活地通过多种方式分析一天的情况，例如数据仓库、交互式 SQL 查询、实时分析和大数据处理。您可以为 AWS 中的数据使用广泛的分析服务，确保满足您目前和未来的分析用例需求得到满足。

## 无与伦比的可扩展性和可用性

Amazon S3 专为存储和检索任何数量的数据而构建，具有无与伦比的可用性，并且从头开始构建，可提供 99.999999999%（11 个 9）的持久性。它是唯一一款具有以下特点的存储产品：可以将您的数据存储在单个 AWS 区域内三个可用区中的多个数据中心之中，一旦单个数据中心出现问题，即可保证无与伦比的恢复能力；而且它也是唯一能在任何区域之间无缝复制数据的存储产品。

## 高度安全

S3 是唯一允许您在账户和对象级别应用访问、日志和审计策略的云存储平台。S3 提供自动服务器端加密，可使用 AWS Key Management Service (KMS) 管理的密钥执行加密，还可使用您管理的密钥执行加密。在跨区域复制时，S3 会对传输中的数据执行加密，并允许您对源和目标区域使用单独的账户以防止恶意内部删除。为了在攻击的早期阶段主动监测到攻击，Amazon Macie 这一由 ML 支持的安全服务可以监控数据访问活动异常，并在检测到未经授权的访问风险或意外数据泄漏时发出详细警报。

## 经济高效

基于 AWS 构建的数据湖具有最高的成本效益。不经常使用的数据可以移动到 Amazon Glacier，以超低成本提供长期备份和存档。Amazon S3 管理功能可以分析对象访问模式，以便根据需要将不常用的数据移动到 Glacier，或者使用生命周期策略自动移动这些数据。您可以通过低至查询每 GB 数据 0.005 USD 的价格开始使用 Amazon Athena 查询数据。 其他分析和机器学习服务的定价也采用即用即付的方式，根据您使用的资源收费。

## 高速性能

Amazon Redshift 和 Amazon Athena 等 AWS 分析服务旨在实现快速交互式查询性能，以支持大量并发交互式查询。使用 Amazon S3 Select 运行 AWS 广泛的分析和机器学习服务组合时，由于仅返回对象内必要的数据子集，因此查询速度最多可以提高 400%，成本也会显著降低。Glacier Select 提供了类似的功能，允许您更快地检索存档数据，并允许您扩展数据湖的分析功能，以包括存档存储。

## 规模最大的合作伙伴网络

AWS 合作伙伴网络 (APN) 的合作伙伴整合数量是其他任何企业的两倍之多，拥有来自全球各地的数万家合作伙伴，包括咨询服务和独立软件供应商。这使您可以轻松地使用和集成您当今使用和喜爱的许多工具。由 AWS 解决方案架构师和合作伙伴开发的数据湖快速入门通过几个简单的步骤，帮助您基于 AWS 安全性和高可用性最佳实践构建、测试和部署数据湖解决方案。

# 开始使用 AWS


注册 AWS 账户


在数天内构建安全的数据湖


开始使用 AWS 进行构建