为什么要建立数据湖

Last modified: October 29, 2019 • Reading Time: 5 minutes

什么是数据湖（Data Lake）？

*数据湖（Data Lake）*是在单个位置中包含多个原始数据源的存储库。在云中，这些通常存储在云 c-store 数据仓库或 S3 buckets 中。数据可以采用多种格式，可以是结构化，半结构化，非结构化甚至是二进制的。

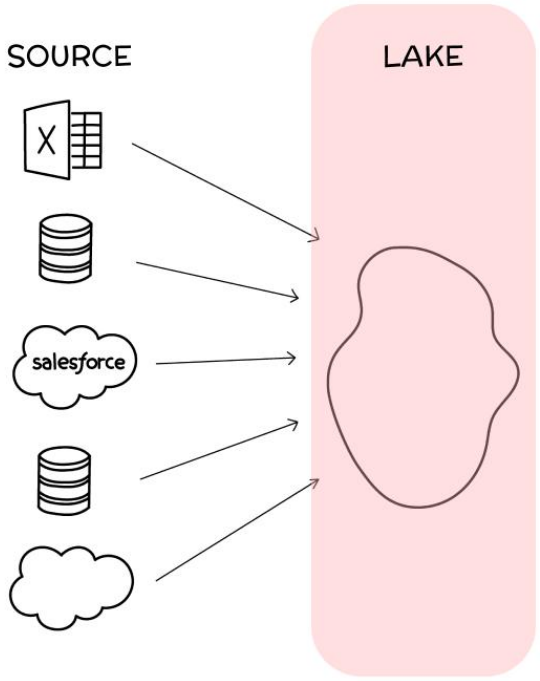
*数据湖（Data Lake）*一词，来源于油湖（精炼前油），有别于描述了有序，隔离和提炼过数据的*数据集市（Data Mart）*，该术语描述一种将所有数据都放在一个位置可以更轻松地处理的大型数据集，便于数据建模过程的早期就开始获得对数据的洞察。

如果满足以下条件，则此阶段适合您：

- 对于 Salesforce 等云应用程序源，您需要唯一或组合的图表/仪表板。
- 您的图表和仪表板将由一组核心人员创建，这些人员都将能够了解凌乱数据结构的来龙去脉。
- 您对数据建模感到恐惧（但是不必恐慌-这就是我们拥有这本书的原因）。
- 您甚至不能分出时间进行简单的数据建模，而且感到现在还行，虽然承担一些技术债务。
- 您拥有大量数据，并且需要更多高性能查询。

如果满足以下条件，则您已经超出了这个阶段：

- 更多的人将使用此数据集。
- 您想要建一个全公司的*唯一事实来源*。
- 您不喜欢与数据完整性问题作斗争。
- 您需要将数据的结构与始终变化的事务源分开。
- 您不喜欢重复自己（DRY）



建立数据湖（Data Lake）的四大理由

1）数据是合并的

随着数据需求的扩展，使用保存在多个不同孤岛中的数据变得越来越难。从产品的角度来看，将流量数据存储在 Google Analytics（分析）中，将销售记录存储在 Salesforce 中，将产品试用参与数据存储在某个数据库中，是合理的。但是，当您需要分析渠道和归因模型时，就需要同时使用它们。

在源数据阶段，我们讨论了混合选项，但是由于混合将所有预联接结果加载到 BI 产品中，因此这些模型在可联接的数据量方面受到极大限制，并且不是可伸缩的解决方案。

在*数据湖（Data Lake）*中，所有数据都可以合并在一起，以便可以一起分析。这使获得洞察变得更加容易，并为数据探索提供了更多的深度。

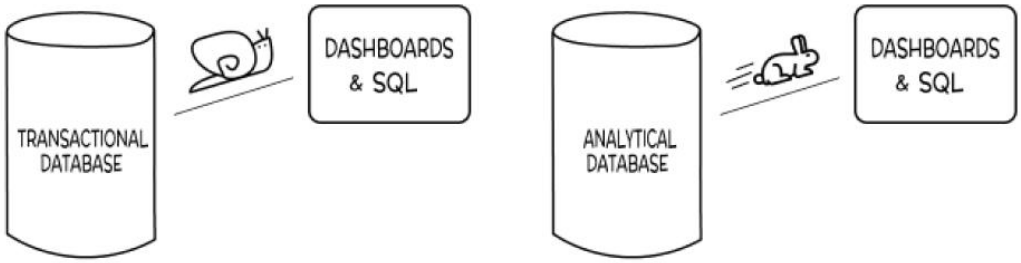
2）完整的查询入口

您的企业使用的应用程序可能仅提供对数据的事务性 API 访问。它们不是为报表而设计的，因此，除非将数据导出并放入您可以轻松查询的格式，否则最终将限制您提取的内容。如果将这些 API 直接用于报表，它们也可能变得非常昂贵。

如果您使用 ELT 产品提取这些 API 数据并将其加载到*数据湖（Data Lake）*中，则将拥有 SQL 或您使用的任何 BI 产品的所有功能和灵活性-并且每个图表的成本都不会大幅增加。

3）性能

源数据可能来自实际的生产数据库，这可能会影响正在运行的应用程序的性能。**事务数据库不是为需要大量数据（例如聚合）的查询而优化的**。



*数据湖*旨在独立于生产环境来处理这些类型的临时分析查询。您可以在*数据湖*上扩展资源，以更快地查询数据。

4）进展

将数据集中到一个点是进入其他阶段的必要步骤。它使处理数据变得非常容易，以至于许多 BI 产品都需要这个阶段-因为它们仅连接到单个仓库源。

在*数据仓库*阶段，您将能够在*数据湖*之上实施适当的建模。通过建模，数据将变得更干净，从而使更多的人可以使用它，减少错误，并减少重复工作。

Written by: Tim Miller, Matt David
Reviewed by: Dave Fowler