简介–数据复杂性的四个阶段

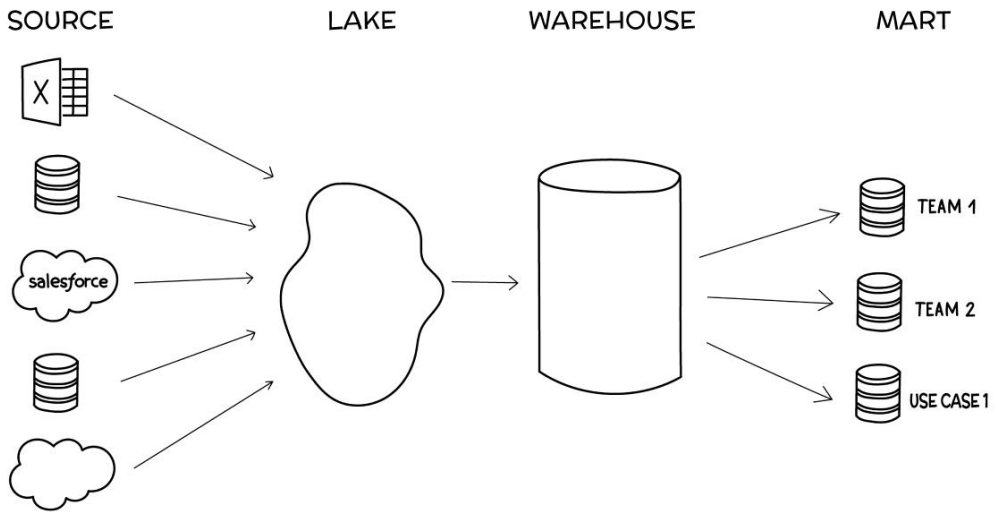
Last modified: November 01, 2019 • Reading Time: 4 minutes

数据复杂性的四个阶段

根据我们与众多组织合作的经验，我们认识到成功的公司需要经历四个不同的数据复杂化阶段。这些阶段恰好与每个阶段所需的新数据堆栈相关联，因此我们命名了这些阶段。

本书分为几部分，分别涵盖以下 4 个顺序：

1. 源数据（Source）
2. 湖（Lake）
3. 仓库（Warehouse）
4. 集市（Mart）

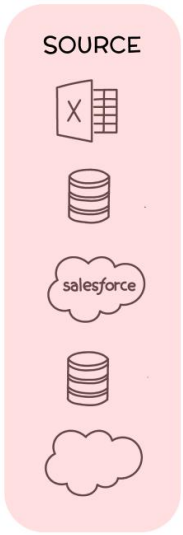


上图所描绘的每个垂直阶段都是有效的阶段，可根据您的资源，大小，重要性和组织内数据的需求进行操作。每个阶段都有其独特的优势，陷阱和最佳实践，我们将逐步介绍它们。

您的公司可能还不需要本书的全部内容，但是随着公司的数据需求不断增长，将其从各个阶段一直推进到*数据集市(Mart)*阶段将具有不可估量的价值，甚至是至关重要的。

我们将从每个概述开始：

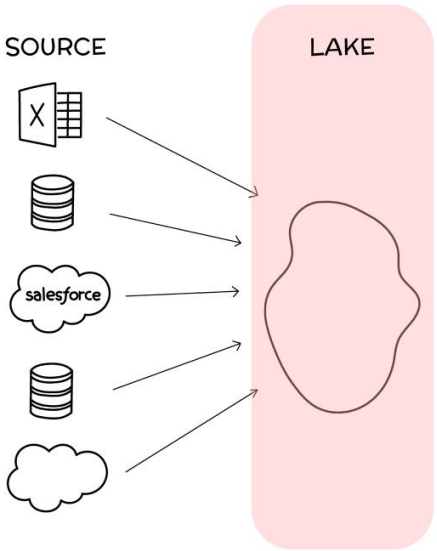
阶段 1. 源数据



当您开始使用数据时，可能只有几个感兴趣的源数据。早期的两个常见来源是 Google Analytics（分析）和您的应用程序数据在产品运行的任何 PostgreSQL 或 MySQL 数据库中。如果您公司中只有少数几个人需要使用这些资源，则可以将其设置为具有直接访问权限；对于他们来说，直接直接处理数据更加简单和敏捷。

阶段 2. 数据湖（Data Lake）

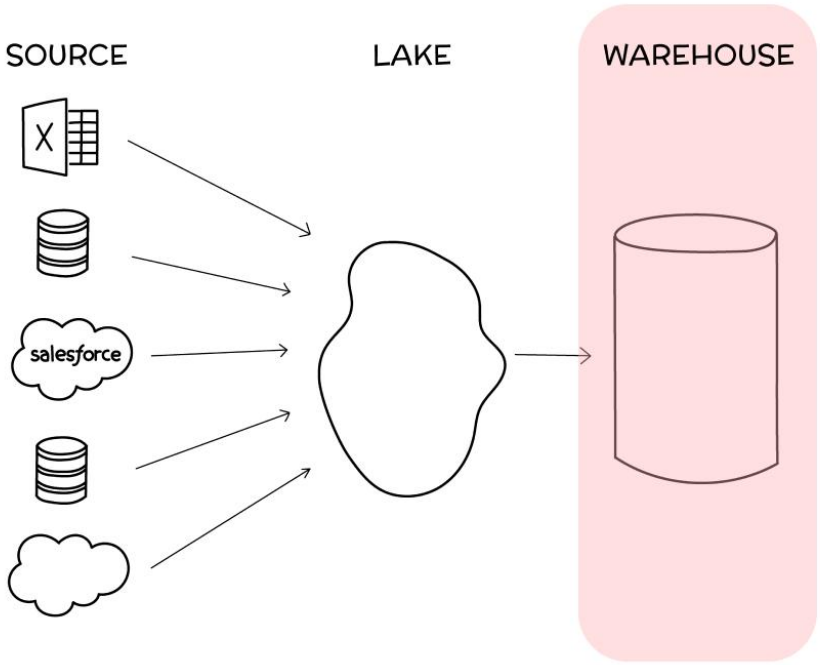
当您开始依赖更多的数据源，并且更频繁地需要混合数据时，您将需要构建一个*数据湖(Data Lake)*-将所有数据集中在一个统一的高性能数据源。



特别是当您需要使用来自 Salesforce，Hubspot，Jira 和 Zendesk 等应用程序的数据时，您将需要为这些数据创建一个单一仓库，以便可以使用一个 SQL 语句访问所有数据，而不是许多不同的 API。

阶段 3. 数据仓库（单一事实来源）

在*数据湖(Data Lake)*阶段，引入更多人使用数据时，您必须向他们解释每种模式的特殊之处，哪些数据在哪里以及需要在每个表中进行过滤的特殊条件，来获取适当的结果。这变得很繁重，并且会导致您经常遇到完整性问题。最终，您将需要开始将数据清理到单一的、干净的事实来源。

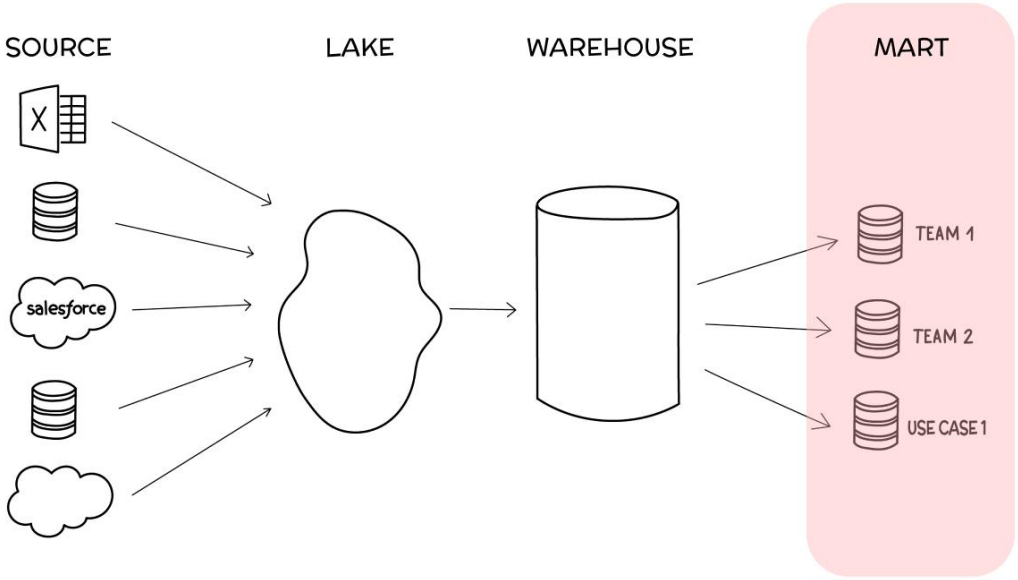


在过去，这个阶段（创建数据仓库 Warehouse）一直是一场噩梦，并且有许多著作着眼于如何最好地建模数据以进行分析处理。但是如今，这并不困难-不仅使您不必向新的团队成员解释所有模式的怪异，而且还可以节省个人的时间，不必重复，编辑和维护自己的乱七八糟的查询。

阶段 3. 数据集市 Data Marts

当您拥有干净的数据并在其上拥有良好的 BI 产品时，您应该开始注意到公司中的许多人都能够回答他们自己的问题，并且越来越多的人参与其中。这是个好消息：您的公司越来越了解信息，业务和生产力结果也应显示出来。您也不必担心完整性问题，因为您已经对数据进行了建模，并且不断地将其维护为干净，清晰的事实来源。

但是，最终，在该事实来源中将有数百个表，并且当用户尝试查找与他们相关的数据时，他们将不知所措。您可能还会发现，根据团队，部门或用例的不同，不同的人希望使用以不同方式构造的同一数据。由于这些原因，您将要开始推出*数据集市(Data Marts)*。



*数据集市(Data Marts)*是为某个团队或某个查询目的设定的，更小更具体的事实来源。例如，销售团队可能只需要主仓库中的 12 个左右的表，而营销团队可能需要 20 个表-其中一些是相同的，但有些不同。

Written by: Dave Fowler
Reviewed by: Matt David