

简译版

## 部署“数据湖”的最佳实践

非官方中文译文·安天技术公益翻译组 译注

文档信息			
原文名称	Best Practices for Deploying Data Lakes		
原文作者	John Gray	原文发布日期	2019 年 8 月 9 日
作者简介	John Gray 是 Infiniti 咨询集团的首席技术官。		
原文发布单位	Information Week		
原文出处	<a href="https://www.informationweek.com/best-practices-for-deploying-data-lakes/a/d-id/1335332">https://www.informationweek.com/best-practices-for-deploying-data-lakes/a/d-id/1335332</a>		
译者	安天技术公益翻译组	校对者	安天技术公益翻译组
分享地址	请浏览创意安天论坛 <a href="https://bbs.antivy.cn">bbs.antivy.cn</a> 安天公益翻译板块		
免责声明	<ul style="list-style-type: none"> <li>本译文译者为安天实验室工程师，本文系出自个人兴趣在业余时间所译，本文原文来自互联网的公共方式，译者力图忠于所获得之电子版本进行翻译，但受翻译水平和技术水平所限，不能完全保证译文完全与原文含义一致，同时对所获得原文是否存在臆造、或者是否与其原始版本一致未进行可靠性验证和评价。</li> <li>本译文对应原文所有观点亦不受本译文中任何打字、排版、印刷或翻译错误的影响。译者与安天实验室不对译文及原文中包含或引用的信息的真实性、准确性、可靠性、或完整性提供任何明示或暗示的保证。译者与安天实验室亦对原文和译文的任何内容不承担任何责任。翻译本文的行为不代表译者和安天实验室对原文立场持有任何立场和态度。</li> <li>译者与安天实验室均与原作者与原始发布者没有联系，亦未获得相关的版权授权，鉴于译者及安天实验室出于学习参考之目的翻译本文，而无出版、发售译文等任何商业利益意图，因此亦不对任何可能因此导致的版权问题承担责任。</li> <li>本文为安天内部参考文献，主要用于安天实验室内部进行外语和技术学习使用，亦向中国大陆境内的网络安全领域的研究人士进行有限分享。望尊重译者的劳动和意愿，不得以任何方式修改本译文。译者和安天实验室并未授权任何人士和第三方二次分享本译文，因此第三方对本译文的全部或者部分所做的分享、传播、报道、张贴行为，及所带来的后果与译者和安天实验室无关。本译文亦不得用于任何商业目的，基于上述问题产生的法律责任，译者与安天实验室一律不予承担。</li> </ul>		

## 部署“数据湖”的最佳实践

John Gray

2019 年 8 月 9 日

**随着数据的迅速变化和增长，从数据中获取价值的需求变得越来越迫切。本文提出了一些关于如何处理“数据湖”的建议。**

尽管“数据湖”(data lake)是一个新兴的术语，但随着数据日益成为现代业务的基础，数据湖越来越受 IT 团队的认可。作为减少数据扩张和数据孤岛的解决方案，数据湖从“数据仓库”(data warehouse)行业中脱颖而出，其目标是解决 IT 团队在创建战略数据仓库时遇到的困难，并基于此制定关键业务决策。数据湖的用途包括：数据分析、更好地了解客户需求、人工智能(AI)、解决实时挑战等等。

从很多方面来看，数据湖是数据仓库的演变。许多数据仓库项目以失败告终，原因包括：成本太高、耗时太长、只能实现预期目标的一小部分等等。随着数据的迅速变化和增长，快速获取数据价值的需求变得越来越迫切。花费数月或数年来分析数据并建模，以供业务使用是不现实的。这是因为，当数据仓库中的数据可用时，业务需求早已发生了变化。

与数据仓库类似，“数据集市”(data mart)的概念出现了，包含了特定用途的数据或按特定属性分类的数据（例如，营销部门数据）。鉴于人们可以更好地理解数据的使用，且数据可以更快地生效，因此数据集市更加成功。但是，数据集市的划分性质，使得它们对拥有大量数据、需要跨职能、跨多方使用数据的企业而言不那么有用。

出于这个原因，数据湖的概念横空出世，旨在满足大规模业务需求。它们能够加快速度，使数据更容易用于以前未定义的需求。凭借巨大的低成本计算能力和几乎无限的存储，真正的大规模云计算出现了，使数据湖方法变得可行。

数据湖仍然是一个相当新的概念，尚未完全被市场接受。因此，早期采用者将从数据湖中获得最大的价值，如在日常业务中利用数据湖为 AI 赋能。除了已经采用数据湖的企业，还有许多企业的 IT 团队正在评估数据湖，希望为其业务找到合适的解决方案。企业该如何正确部署数据湖呢？以下是三个最佳实践。

## 1. 有明确目的使用数据湖

创建数据湖的核心原因是：将数据湖中的数据用于某种目的。虽然理论上来说，数据湖应该服务于许多尚未定义的用途，但企业最好先了解一下如何使用其中的数据。企业应考虑，除了用于存储目的之外，他们还能如何从数据湖中获取哪些价值。与任何 IT 计划一样，首先将数据湖的部署与具体策略相匹配，这一点非常重要。该策略不仅要与 IT 目标保持一致，还要与长期业务目标保持一致。

企业应考虑，创建数据湖是否有助于企业利用其中的数据。长期（例如数年）保存数据的成本高昂。如果企业不打算在短期内将其数据用于特定目的，那么存储该数据就会浪费资金。

## 2. 最细粒度地存储数据并进行标记

最细粒度地存储数据，有助于企业对数据进行组合、聚合或以其他方式处理，以服务于多种目的。在将数据存储到数据湖之前，请勿聚合或汇总数据。对拥有数据湖的企业而言，只有其中的数据被使用，数据湖的价值才能体现。因此，企业最好对数据进行标记和分类，然后再将其放到数据湖中，以便在需要时，能够随时从数据湖中筛选出相应的数据。对数据进行标记（上报所需），可以帮助企业启用分析项目。此外，机器学习（ML）和 AI 可以通过筛选现有数据和创建标签，帮助实现数据标记。

此外，公司可以使用这些数据分析、ML 和 AI 项目来提高整体业务竞争力。一种工具可以为另一种工具赋能。

### 3. 制定数据销毁计划

公司往往会积累大量数据，却没有创建任何计划来销毁不必要的数据资产。特别是，如果公司负有合规义务，需要在一定时间后销毁信息（例如，GDPR 要求公司在一定时间后销毁所保存的欧盟公民数据），无数据销毁计划会成为履行这些义务的障碍。

为数据湖配备销毁计划，有助于企业检索需要销毁的内容及何时销毁。这还有助于企业跟踪所有客户数据所在的位置：只用一个存储位置能够降低成本并节省查询时间。

### 为未来做准备

随着越来越多的数据扩散到业务领域，企业需要以战略方式存储和使用这些数据。数据湖正在成为一种很好的赋能方式，可以充分挖掘企业数据的价值。在考虑数据湖解决方案时，首先要确定企业将如何使用数据，然后确定将数据置于何处。例如，由于存储成本低，云对数据湖具有极大的吸引力。如果企业对云有兴趣，可以进行调查，选择能够满足其独特基础设施需求的第三方提供商。云服务提供商或企业自己的 DevOps 团队应如何构建数据湖流程，以便根据其目标从数据湖中加载和提取数据呢？

毫无疑问，要想获取数据湖的全部价值，需要进行大量的处理过程。企业应考虑分析过程中的哪些步骤可以实现自动化。此外，企业需要熟练的人员来构建基础设施，以托管数据湖、将数据加载到数据湖中，以及从数据湖中提取数据以供使用。此外，IT 人员和业务领导之间应进行定期、开放的沟通，这是实现任何 IT 转型（如数据湖解决方案）的良好开端。

## 安天简介

安天是引领威胁检测与防御能力发展的网络安全国家队,始终坚持自主先进的能力导向,依托下一代威胁检测引擎等先进技术和赛博超脑大平台工程能力积累,研发智甲、探海、镇关、捕风、追影、拓痕等系列产品,为客户构建端点防护、流量监测、边界防护、导流捕获、深度分析、应急处置的安全基石。安天致力于为客户建设实战化的态势感知体系,依托全面持续监测能力,建立系统与人员协同作业机制,指挥网内各种防御机制联合响应威胁,实现从基础结构安全、纵深防御、态势感知与积极防御到威胁情报的有机结合,协助客户开展深度结合与全面覆盖的体系化网络安全规划与建设,支撑起协同联动的实战化运行,赋能客户筑起可对抗高级威胁的网络安全防线。

安天是全球基础安全供应链的核心赋能方,全球近百家著名安全企业、IT 企业选择安天作为检测能力合作伙伴,安天的威胁检测引擎为全球超过三十万台网络设备和网络安全设备、超过十五亿部智能终端设备提供了安全检测能力。安天的移动检测引擎获得国际知名测试机构颁发的 2013 年度权威评测奖项。

安天是中国应急响应体系中重要的企业节点,在“红色代码”、“口令蠕虫”、“心脏出血”、“破壳”、“魔窟”等重大安全威胁和病毒疫情方面,实现了先发预警和全面应急响应。安天针对“方程式”、“白象”、“海莲花”、“绿斑”等几十个高级网空威胁行为体(如 APT 组织)及其攻击行动,进行持续监测和深度解析,协助客户在“敌情想定”下形成有效防护,通过深度分析高级网空威胁行为体的作业能力,安天建立了以实战化对抗场景为导向的能力体系。

安天被行业管理机构、客户和伙伴广泛认可,已连续六届蝉联国家级安全应急支撑单位,是中国国家信息安全漏洞库六家首批一级支撑单位之一,亦是国家网络与信息安全信息通报机制技术支撑单位,国家信息安全漏洞共享平台成员单位。

2016 年 4 月 19 日,在习近平总书记主持召开的网络安全和信息化工作座谈会上,安天创始人、首席架构师作为网络安全领域发言代表,向总书记进行了汇报。2016 年 5 月 25 日,习近平总书记在黑龙江调研期间,视察了安天总部,并对安天人说,“你们也是国家队,虽然你们是民营企业”。

安天实验室更多信息请访问:

<http://www.antiy.com> (中文)

<http://www.antiy.net> (英文)

安天企业安全公司更多信息请访问:

<http://www.antiy.cn>

安天移动安全公司 (AVL TEAM) 更多信息请访问:

<http://www.avlsec.com>