安全数据湖的敏捷数据原则

NTT DATA
Trusted Global Innovator

MARYMOUNT UNIVERSITY

CSA GCR cloud security alliance®
GREATER CHINA REGION

CSA cloud security alliance®

最高威胁工作组官网地址是：

https://cloudsecurityalliance.org/research/working-groups/top-threats/

联盟简介

云安全联盟(Cloud Security Alliance,CSA)是中立、权威的全球性非营利产业组织,于2009年正式成立,致力于定义和提高业界对云计算和下一代数字技术安全最佳实践的认识,推动数字安全产业全面发展。

云安全联盟大中华区(Cloud Security Alliance Greater China Region, CSA GCR)作为CSA全球四大区之一,2016年在香港独立注册,于2021年在中国登记注册,是网络安全领域首家在中国境内注册备案的国际NGO,旨在立足中国,连接全球,推动大中华区数字安全技术标准与产业的发展及国际合作。

我们的工作

联盟会刊下载地址
了解联盟更多信息

加入我们

CSA大中华区官网
(https://c-csa.cn)

点击会员

加入联盟

填写相关申请信息

成为CSA会员

JOIN US

致谢

《安全数据湖的敏捷数据（An Agile Data Doctrine for a Secure Data Lake）》由 CSA 工作组专家编写，CSA 大中华区秘书处组织翻译并审校。

中文版翻译专家组（排名不分先后）：

翻译组：胡鑫　牛媛媛

研究协调员：赵鹏

感谢以下单位的支持与贡献：

杭州美创科技股份有限公司

原文英文版编写专家组

主要作者：Dr. Dianne Murphy　Theresa Kushner　Oliver Forbes

CSA 分析师：Sean Heide　Claire Lehnert (graphic design)

Stephen Lumpe (cover illustration)　John Yeoh

在此感谢以上专家。如译文有不妥当之处，敬请读者联系CSA GCR秘书处给予雅正! 联系邮箱research@c-csa.cn；国际云安全联盟CSA公众号。

序言

数据湖作为目前正在广泛运用的一种技术，是一种集中式的数据仓库，用于接收、存储和保护大量结构化和非结构化的数据，有效集成与分析各类数据源，为相应的组织增加了巨大的价值。虽然这是对于数据聚合的一个关键解决方案，但光有数据湖技术是不够的，需要企业范围的数据原则来确保数据是可信的。而且，在整个数据社区中最大限度地实现其价值并确保遵守越来越多的法规要求至关重要。

报告考察了当数据湖作为一种主要数据管理技术而呈现的数据原则的基本参数，包括确保数据的安全和隐私，并且为数据湖的整体保护和执行提供了架构。

希望为组织提供一套能够贯穿始终的数据原则思路。建立包含敏捷思维的整体数据战略，实现数据安全和有效的隐私控制，在最大程度降低总成本的同时，提高数据湖中数据资产的价值。

李雨航 Yale Li

CSA 大中华区主席兼研究院院长

目录

介绍

数据是横向和纵向跨越多个行业和组织的重要资产。无论是对政府、企业还是非营利组织，资产的使用和安全都至关重要。然而，随着数据在规模性、多样性、高速性和变化性（即所谓的“大数据”）方面的增长，在收集、存储和检索数据时要求保障安全性、隐私性的需求以及性价比更高的技术也在不断发展。

一种目前正在广泛运用的技术是数据湖。具体来说，数据湖是一种集中式的数据库，用于接收、存储和保护大量结构化和非结构化的数据。虽然数据湖是数据聚合的一个关键解决方案，但光有数据湖技术是不够的，需要企业范围的数据原则来确保数据是可信的。而且，在整个数据社区中最大限度地实现其价值并确保遵守越来越多的法规要求至关重要。

本文观点考察了当吧数据湖作为一种主要数据管理技术而呈现出的数据原则的基本参数，包括确保数据的安全和隐私，并且不基于任何供应商的解决方案，而是考虑到数据湖的整体保护，并为这些原则的执行提供了架构。

问题

数据湖通过提供对广泛的结构化和非结构化数据源进行有效集成与分析，为相应组织增加巨大的价值。一般而言，数据湖以原始形式存储从多个数据源收集的海量数据，直到需要数据时才定义数据结构和用户需求。如果组织要使用数据并确保其安全和隐私，数据湖必须遵守严格的数据原则政策和相关程序。且需要得到数据社区（DC）中所有人的信任。为了从数据湖中获取最高的灵活性和性能，保障DC中的任何用户能轻而易举的获取正确的数据，在适当的时候，跨多个数据集进行可靠的数据集成。执行此操作时，必须确信数据是安全、准确和最高质量的。

数据原则包括角色和责任，是数据湖“良好”持续数据管理的先决条件，并决定了整个DC所需的决策，从而确保有效管理和安全。

什么是数据社区？

社区定义为在特定领域具有共同利益的个人的统一团体；广义来说，是指将分散在一个更大的社会或团体的人员和专家，由共同的兴趣和政策联系在一起形成一个共同的团体。因此，数据社区是由平时分散在组织中投入研究数据的个人，通过常见的数据政策（如隐私政策）联系在一起而组成的统一团体。

确保组织中的DC正常运转是数据治理人员或者团队的职责。《数据治理：如何设计、部署和维持有效的数据治理计划》一书的作者John Ladley将数据治理定义为信息管理、企业信息管理和数据架构的一个组成部分。正如财务会计师使用一个流程和一套原则管理公司的财务一样，负责数据资产的数据经理也需要流程和规则。这些原则由数据治理建立，并由信息管理团队实施。Ladley指出，“数据治理不是由管理信息的人员履行的职能。”相反，数据治理是一种管理监督，有助于为数据管理过程提供秩序和框架。

数据社区是围绕关键数据集建立的。例如，一个社区可能负责财务数据管理，另一个负责客户数据。这些社区确保其影响范围内的数据准确、及时、一致、安全，并提供给其他社区。

任何社区的首要价值是为其成员提供归属感和安全感。数据社区也不例外。因此，确保数据安全应该是社区章程的一部分，为数据用户提供了对收集和管理的数据的安全访问。

理论与应用

数据原则、政策和实施规则必须平衡数据本身的管理和控制及整个数据社区数据湖潜在用户的需求，确保：

1. 信任：用户必须知道，从数据湖开发的可视化和报表是基于可识别和可靠来源的准确数据（“真相”），以便组织避免根据“不良数据”的分析结果做出错误的决定或行动
2. 数据血源可追溯性：用户必须能够通过数据生命周期中的多次跃程轻松识别数据的血缘和安全性，包括数据的原始来源、任何中间转换过程中流程的细节和算法的使用、安全控制以及复制数据转换路径中使用的任何数据的能力。
3. 使用情况跟踪：组织必须知道谁访问了数据，包括特权用户、他们访问了什么、是否进行了更改以及更改的方式和时间，以便组织能够满足任何法规要求，如《健康保险流通与责任法案（HIPAA）》或《塞班斯-奥克斯利法案（SOX）》。
4. 启用发现：数据湖的内容可能很复杂，因此必须对数据有效标记，包括添加元数据，并且文件需要被开发，例如开发数据目录，以便可以使用与用户群体相关的术语识别相关的数据集。
5. 数据安全：系统必须确保数据只能由授权用户访问，尊重数据的机密性和隐私（例如，个人身份信息或PII），并允许及时识别任何利用外部或内部数据的威胁及其补救措施。此外，还必须考虑高度机密的数据应该存入指定的数据湖中。支付卡信息或SSN等数据可能永远不存在于数据湖中，这是非常合理的。
6. 数据混淆：为了最大限度地提高可用性，当多个用户可以检索同一数据记录时，必须保护敏感和个人数据元素的机密性，这些用户可能只有权查看某些指定的数据元素，而不能查看其他数据元素（如工资信息）。这包括数据脱敏、加密和标记化等技术。
7. 可审计性：必须确保内部和外部审计师可根据需要获得所有数据血缘和数

据访问信息，促进对欺诈、风险和合规性的审计。

8. 性能：数据原则流程不得对数据湖的性能产生重大不利影响，包括从多个来源获取数据和发现数据的过程。

9. 故障排除：必须有现成的技术解决由数据的安全性、可信度或质量引起的问题，并确保及时采取适当的纠正措施。

10. 保留、归档和处置：数据湖必须遵守数据生命周期中的所有法律保留要求，确保遵守高于组织之上的任何条件（例如，执法授权）。

11. 监控：必须建立数据湖监控规范，并且必须围绕数据分类分级和组织合规报告的要求，配置基于数据分类分级和使用模式（如下载数据、修改数据、删除数据等）的警报通知。

其中许多要求并不新鲜，但组织在实施“良好”安全和隐私的同时，通常为了创造或保持竞争优势，越来越多的关注于从其数据资产中获得最大价值。长期以来，包括数据治理在内的数据原则一直被认为是从数据中获得业务价值的关键成功因素。尽管如此，也许是时候在新技术的背景下重新审视数据治理的原则，比如数据湖，以及对安全和隐私的日益关注。

解决方案

随着数据湖对组织的价值越来越高，组织对有效数据治理和管理的需求也在不断增加。为了应对这种情况，必须实施一套数据原则框架，该框架包括安全的数据管道，确保在整个企业数据社区需要时，能以正确的格式提供最新的、可信的数据。该数据原则框架必须是灵活的，因为随着数据爆炸性的增长，数据生态系统在不断发展变化，增加了数据源的可变性，导致会增加更多的数据管道，产生新的数据架构，进而也增加了对数据隐私和合规性的关注。与此同时，高明的黑客、民族国家活动、恶意内部员工和意外数据丢失都给数据安全带来了越来越大的挑战。此外，随着执法力度的加大，制定了新的隐私法规，数据隐私已成为社会关注的问题。最后，数据也越来越分散，没有清晰的边界：在核心（内置好的开箱即用和多云部署情况）和边缘的多源头接收、存储、运算上，用户的终端要做更多处理。

敏捷数据原则框架必须提供一个灵活的正式结构控制数据的定义、收集、存储、处理和使

用，管理风险并确保数据的质量和可用性，从而支持整个企业数据社区的数据驱动决策。

传统的数据治理主要是防御性的，主要侧重在减轻风险，延长数据可用性，因为投入成本高而被诟病。然而，与其他敏捷方法一样，敏捷数据原则应该涉及整个数据社区（技术、商业和管理），应该对持续的变化和提升做出反应，并且应该嵌入到数据社区所有人员的日常活动中，同时确保数据质量、数据可用性、数据安全和数据隐私。

敏捷数据湖威胁模型 框架需要首先让组织的数据社区定义以下治理因素，重点关注角色和职责：

a. 为了维护和提高数据质量，并在不断发展的数据湖环境中维护安全和隐私，敏捷数据治理需要哪些必要的政策和标准

b. 在整个数据社区中，哪些术语应该被定义好（例如数据质量、及时性、访问和授权的定义）？

c. 数据社区中存在哪些数据资产？谁是数据的所有者？如何将持续不断的识别出的数据集放到资产列表中?

d. 元数据是数据湖的基本优先事项：现在有哪些描述性信息（例如，数据目录）？谁负责持续创建和维护元数据？确保技术标签能够翻译以供使用？

e. 如何管理数据管道和储存？从而可以确保遵守政策时，在监管要求下，不影响数据的可用性，保证数据可访问。

f. 数据管道的哪些部分需要自动化，以及如何自动化，包括机器学习等先进技术？

g. 如何将数据治理标准和政策集成到工作流程中，避免处理延迟或数据可用性受到限制？在数据使用中，时效是很重要的。

h. 数据原则政策和标准如何与正在使用的进程、工具、技术和平台（包括云平台和内部设施）集成？

i. 谁管理数据湖（角色和职责），包括访问控制、元数据管理、接收监控警报和数据源选择？

j. 如何建立和维护数据社区，包括用户社区如何分享个人在数据检索和分析过程中获得的知识？

k. 如何执行管理变更，包括变更要求、进度记录和变更决定，确保数据不会出现过期而影响正常使用？

l. 包括治理政策在内的数据原则以及产生的变化如何传达给整个数据社区？

m. 谁负责执行？

n. 如何衡量成功，如何定义成功的指标？

o. 使用什么技术，如何支撑数据原则的安全实施？

一旦计划好，数据原则就必须作为数据湖管理计划的组成部分坚决实施。为了满足当下的敏捷环境，自动化是确保最小化持续投入成本，最大化保证数据的安全性和隐私性的必要条件。

结语

NTT DATA相信，在最大程度降低总成本的同时，提高数据湖中数据资产的价值，确保安全性和隐私性不仅是专业人员的责任，也需要数据社区中包括所有数据用户在内的每个人的共同努力。敏捷技术已经成功地实现了软件开发的现代化，也为不断发展的数据生态系统做出了类似的承诺。此外，数据操作在实施过程中可能具有一定价值，但在此之前，组织需要先仔细考虑，形成一套能够贯穿始终的数据原则。数据原则包括确定关于数据和数据湖的基本决策，并建立包含敏捷思维的整体数据战略。数据安全和有效的隐私控制至关重要，约定好整个数据社会的参与必须始终处于数据治理追求的前沿。无论如何，我们的努力必须植根于一种结构稳固、面向框架的基本方法——数据原则。

CSA GCR cloud security alliance GREATER CHINA REGION ®

Cloud Security Alliance Greater China Region

扫码获取更多报告

云安全联盟大中华区
官网：https://c-csa.cn 电话：0755-86548359
地址：上海市浦东新区东育路255弄5号3FB30（前滩国际经济组织集聚区）/
广东省深圳市南山区粤海街道北京航空航天大厦1号楼905