《大数据湖最佳实践》笔记

AUTHOR: 彭玲 TIME: 2023/1/17

数据湖概念

数据湖这个词由 Pentaho 的 CTO，James Dixon 发明，他在博客中首次提出这个概念：“如果你把数据集市看作是一家售卖干净的、规整包装的、便于消费的瓶装水的商店，那么数据湖就是更自然状态下的一大片水域。数据湖的内容从一个源头流入，各类用户可以前来检查、探索或取样。”

关键词：

- 数据处于它的原始形式和格式（自然的、原始的数据）。

- 数据被各类用户使用，比如已经或可以被大量用户获取到。

数据治理与**数据安全**是自助服务的一大挑战。

数据湖的目的：尽可能多地搜集原始数据。数据湖像储蓄罐，只是把数据存储起来，且不需要过早地对数据进行转换、处理。

数据湖成熟度

数据量与 IT 参与度上的变化：数据水洼 -> 数据池 -> 数据湖 -> 数据洋

数据沼泽

数据不充分使用（不可用），或者根本没人使用。

数据湖实施路径

1. 建设好基础设施（搭建好 Hadoop 集群，并保证正常运行）。
2. 组织好数据湖的各个区域（给不同的用户群创建好各种区域，并导入数据）。
3. 设置好数据湖的自助服务（创建数据资产的目录，设置好访问机制，准备给分析师使用的工具）。
4. 将数据湖开放给用户。

组织和管理数据

数据质量工具

数据质量包括：定义质量规则，将这些规则应用于数据以检测“异常”的违规行为，以及修复这些异常。

数据质量规则有多种形式和粒度。

数据建模工具

数据建模工具用于创建关系 schema。

元数据仓库

元数据可以手动收集，也可以通过集成各种其他工具收集，如 ETL 工具、BI 工具等。

元数据仓库的三个主要使用场景：

- 搜索数据资产
- 追踪血缘
- 影响分析

建立数据湖

建设数据湖的目的是提供一种存储企业数据的方案，旨在将分析师和数据科学家所需数据的易用性和可用性提升到极致。

Apache Hadoop 是建设数据湖常用的开源项目。

Hadoop 是一个大规模存储和并行执行的平台，它可以自动化处理一些在构建高扩展性和高可用性集群时会遇到的技术难点。

数据转换

对于那些我们暂时不需要处理的数据，我们可以将它们先存放于数据湖中，直到我们需要时再进行处理，而不是抛弃它们。

ETL 数据迁移的一种形式：将业务数据转换为数据仓库所需的分析模式。

自助服务优化

只有决策者能够将其行为建立在数据的基础上，才能体现数据的价值。

企业应该如何重新考虑收集、标记和共享数据以实现自助服务模式来赋能业务方。

一个典型的业务分析师工作流程：查找并理解所需的数据 -> 预置（获得以可用的形式和格式组织的数据） -> 预处理（组合、过滤、聚合、修复数据质量问题等） -> 分析（使用数据发现和可视化工具进行分析）。

发现和理解数据

分析师希望使用他们熟悉的**业务术语**来搜索数据，而数据集和字段通常使用模糊的**技术术语。**这使得分析师很难找到并理解数据。为了填补这一空白，许多企业寄希望于用**数据目录**把业务术语（或标签）跟数据集（及其字段）相关联，让分析人员使用标签快速查找数据集，并通过查看与每个字段关联的标签来了解这些数据集。

基于领域专家和分析师提供的标签，自动化工具利用人工智能（AI）和机器学习自动为数据集打上标签、添加注释。

建立信任

一旦分析师找到相关的数据集，下一个问题就变成了数据是否可信。信任通常基于三个维度：

- 数据质量：数据集的完整性和整洁性。
- 血缘（aka 起源）：数据来自哪里。
- 管理员：谁创建了数据集，以及为什么创建。

数据质量

在实践过程中，质量可以被定义为数据是否符合规范，其范围可以从简单到复杂。最常见的数据质量规则有：

- 完整性：字段不为空。
- 数据类型：字段的类型正确。
- 范围：字段的值位于指定范围内。
- 格式：字段具有特定格式。

- 基数：字段具有特定数量的唯一值。
- 专一性：该字段的值具有唯一性（例如，客户 ID 在客户列表中应该是唯一的）。
- 参照完整性：该字段的值位于引用值集中。

检查数据质量的最常用方法称为数据剖析。这种方法涉及读取每个字段中的数据并计算指标，如空字段数（完整性），唯一值数（基数）和唯一值百分比（专一性），以及检查数据类型和范围，格式化和执行引用完整性检查。

数据预置

预置有两个方面：获得使用数据的权限，以及物理地获取到数据。

目录是一种敏捷的进行访问控制方法，创建元数据目录以使分析人员无需访问数据本身即可查找数据集。识别出正确的数据集以后，分析师提交权限申请，数据管理员或者所有者决定是否授权、权限有效期以及对哪部分数据开放权限。访问期限到期后，可以自动撤销访问权限或者请求延期。

数据湖元数据

数据湖的一些特性使得数据检索较为困难。

数据目录可以通过为字段和数据集打上一致的业务标签并提供类似网上商店的使用界面来解决这个问题，它使得用户可以通过业务术语来描述所需要的数据，也可以通过数据集的业务标签和描述信息来理解数据。

技术元数据

通过数据剖析获得的统计信息和表、字段名字这样的元数据统称为技术元数据。通过技术元数据，可以帮助我们理解数据。

业务元数据

业务元数据可以帮助分析师找到合适的数据。业务元数据有多种形式。

打标

为了能使用业务元数据来查找数据，需要将合适的术语和概念指定给数据集。这个过程就是“打标”，就是将业务术语指定给特定字段或者数据集，这些字段或者数据集包含这些术语所描述的数据。

自动编目

如果依赖手动打标，那么仅仅只有最常用的数据集会被打上标签，绝大部分则会被遗忘在黑暗的角落。解决这个问题的办法就是自动化。新型工具利用人工智能和机器学习对处于“暗处”的数据集进行识别、自动打标和备注（基于 SME 和分析师在其他地方提供的标签），以便分析师可以找到和使用这些数据集。

敏感数据管理和访问控制

我们将任何需要满足合规检查以及访问控制的数据统一叫作敏感数据。

为了管理敏感数据，企业首先需要为数据进行编目（知道相关数据存储在哪儿），然后通过访问控制或者脱敏技术来做数据保护。

新型安全系统不是为特定的数据集和字段定义访问和脱敏规则，而是为特定的标签定义规则，然后将其运用于所有拥有这些标签的数据集和字段。

数据质量

通过目录来组织和表达数据质量信息。

基于标签的数据质量规则，其主要思想是先定义规则，然后将规则应用于那些具有相应标签的数据字段。

建立血缘关系

目录需要回答的一个关键问题是分析师是否可以信任这些数据，并且能够说明数据来源。目录的一个任务就是展示数据资产的血缘关系，并且能够补充缺失的血缘关系。

数据预置

一旦识别出了合适的数据，用户希望通过其他工具来使用这些数据。为了支持这些能力，目录常常提供数据预置选项。数据预置可能只是简单地通过特定的工具打开数据集。另外一个预置操作涉及数据访问。用户找到了想要的数据，在使用之前必须先请求访问它，在得到允许之后将数据导入到数据湖中。

数据访问控制

授权与访问控制

授权一般是给指定分析师分配对特定数据资产（如特定文件或者表）进行特定操作的权限（如读权限与写权限）。为了简化此过程，安全管理员通常会创建角色（权限集合）并将这些角色分配给不同分组的分析师。

基于标签的控制策略

传统的访问控制基于物理文件和文件夹。例如，Hadoop 文件系统（HDFS）支持典型的 Linux 访问控制列表（ACL）。更优雅的解决方案是某些 Hadoop 发行版中采用的基于标签的安全策略。例如，Cloudera Navigator 和 Apache Ranger（作为 Hortonworks Hadoop 发行版的一部分提供）支持基于标签的策略。这些工具不是为每个文件和文件夹指定 ACL，而是允许安全管理员使用标签设置策略。这些标签可以通过 Cloudera Navigator 和 Apache Atlas 等本地目录工具设置，并通过基于策略的访问控制工具（如 Apache Ranger）自动获取。

数据脱敏

敏感数据的加密方式：

- 透明加密
- 显式加密
- 脱敏

预置数据

数据预置是建立数据湖的重要部分，由四个步骤组成：请求文件、审核请求、批准请求、预置数据。

其中，请求文件这一步骤由想要访问数据集的分析师完成。该请求通常描述以下内容：

- 需要什么数据（哪个数据集以及是否需要整个数据集或数据集的一部分）。
- 谁需要访问权限（需要访问数据的用户或组列表）。
- 项目（需要数据的项目）。
- 访问的业务理由（为什么需要数据）。

- 需要多长时间（可以访问的持续时间）。
- 如何提供数据（用户应该直接访问数据，还是复制到指定的数据库或数据湖）。

如果要复制数据，请求还应该指定：

- 应放置数据的位置。
- 是私人副本还是可以共享。
- 是一次性快照还是应该保持更新。
- 访问到期后，应该保持更新还是删除。