数据湖：初识篇

AUTHOR: 彭玲 TIME: 2022/3/14

What and How?

- 数据湖是什么?
- 数据湖能做什么?
- 如何建立数据湖?

数据湖总览

大数据的 4个V

- Volume（大容量）
- Variety（多形式）
- Velocity（高速率）
- Veracity（准确性）：最重要的V，我们无法判断坏的决策到底是由坏数据还是坏模型引起的。

数据湖类比

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。

- 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。

数据湖的成熟度

数据水洼

大数据技术构建的单一用途或供单一项目使用的数据集市。

数据池

数据水洼的合集，一个公用数据集市的集合。

数据湖

与数据池有两点不同：

- 支持自助服务，业务不依赖 IT 部门找到想要的数据集。
- 包含业务可能需要的数据，即使当前没有项目需要用到。

数据洋

扩展到企业的所有数据。

数据湖架构

处理引擎

- Flink
- Spark
- Beam

数据存储

- HDFS
- HBase
- Cassandra
- Kafka

交互

- Zeppelin / Spark notebook
- Tableau / Datameer (Qlik)

开源数据湖方案

目前市面上流行的三大开源数据湖方案分别为：Delta、Apache Iceberg 和 Apache Hudi。

Apache Iceberg (Netflix)

Netflix 的数据湖原先是借助 **Hive** 来构建，但发现 Hive 在设计上的诸多缺陷之后，开始转为自研 **Iceberg**，并最终演化成Apache下一个高度抽象通用的开源数据湖方案。

Apache Hudi (Uber)

Uber 的业务场景主要为：将线上产生的行程订单数据，同步到一个统一的数据中心，然后供上层各个城市运营同事用来做分析和处理。

Delta (Databricks)

Databricks 在设计 Delta 时，希望做到**流批作业**在数据层面做到进一步的统一。业务数据经过 **Kafka** 导入到统一的数据湖中（无论**批处理**，还是**流处理**），上层业务可以借助各种分析引擎做进一步的商业报表分析、流式计算以及 AI 分析等等。

数据治理方案

DataHub

DataHub 是一个现代数据目录，旨在支持端到端数据发现、数据可观察性和数据治理。这个可扩展的`元数据`平台是为开发人员构建的，让他们可以适应快速变革的数据生态系统的复杂性，并让数据从业者在其组织内利用数据的全部价值。

Amundsen

Amundsen 是一个数据发现和元数据引擎，用于提高数据分析师、数据科学家和工程师在与数据交互时的工作效率。

Atlas

Atlas 是一组可扩展的核心基础治理服务，为组织提供开放的元数据管理和治理功能，以构建其数据资产的目录，对这些资产进行分类和治理，并为数据科学家、分析师和数据治理团队提供围绕这些数据资产的协作能力。

- 能够动态创建分类 - 如 PII、EXPIRES_ON、DATA_QUALITY、SENSITIVE。
- 集成多重 Hadoop 和非 Hadoop 元数据的预定义类型。
- 血统：直观的 UI，可在数据通过各种流程时查看数据沿袭。
- 搜索发现：可按类型、分类、属性值或自由文本搜索实体，并支持DSL查询。
- 安全：与 Apache Ranger 集成，支持按分类对数据访问进行授权/数据屏蔽。