数据湖分享

# 企业数据困扰

1.数据孤岛

企业发展到一定阶段，出现多个事业部，每个事业部都有各自数据，事业部之间的数据往往都各自存储，各自定义。每个事业部的数据就像一个个孤岛一样无法(或者极其困难)和企业内部的其他数据进行连接互动。”我们把这样的情况称为数据孤岛。简单说就是数据间缺乏关联性，数据库彼此无法兼容

2.非结构化数据

数据仓库可以处理来自关系型数据库中的结构化数据（行和列）、半结构化数据（如CSV、日志、XML、JSON）、无法处理非结构化数据（如email、文档、PDF等）和二进制数据（如图像、音频、视频）。

3.原始数据

# 数据湖概念

一种支持统一的数据存储，存放原始的数据。支持任意结构的数据存储，包括结构化、半结构化、非结构化。支持多种计算分析，适用多种应用场景的一种大规模存储系统架构

## ****数据湖的核心：****

1.存储海量的原始数据

2.支持任意的数据格式（结构化、半结构化、非结构化）

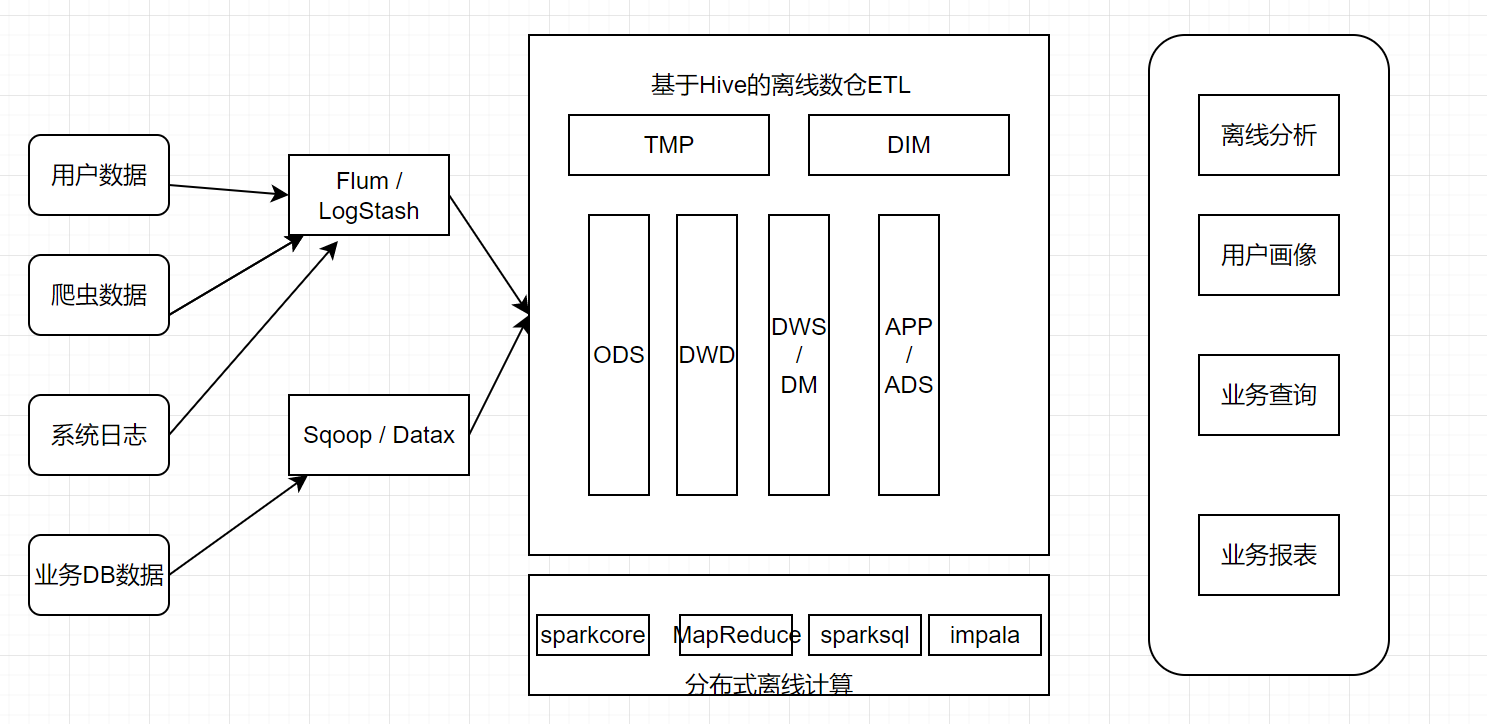
3.较强的分析和处理能力

# 数据仓库和数据湖对比

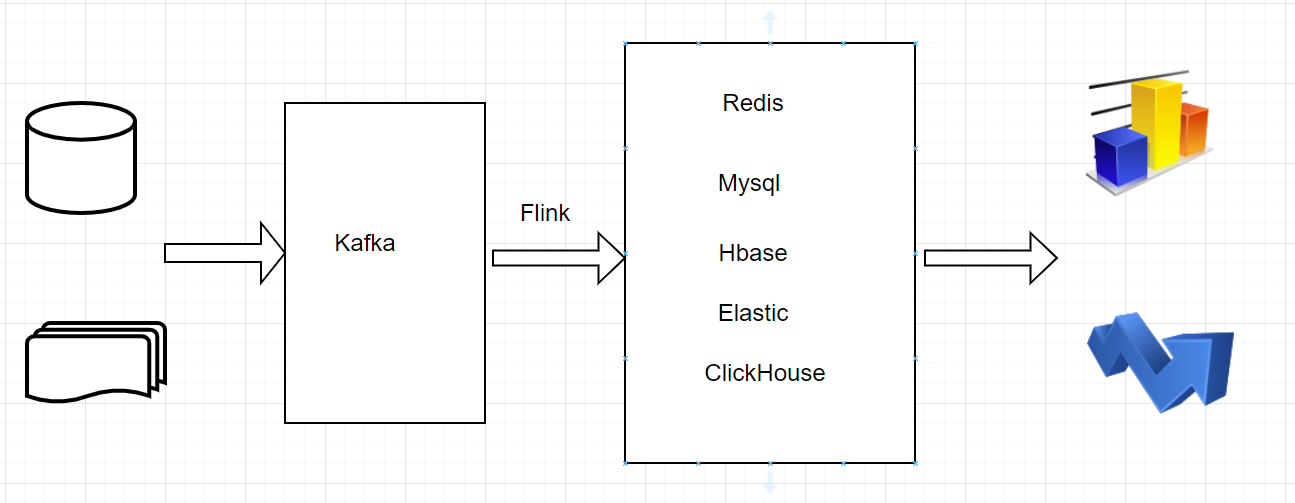


# 架构图

离线数仓

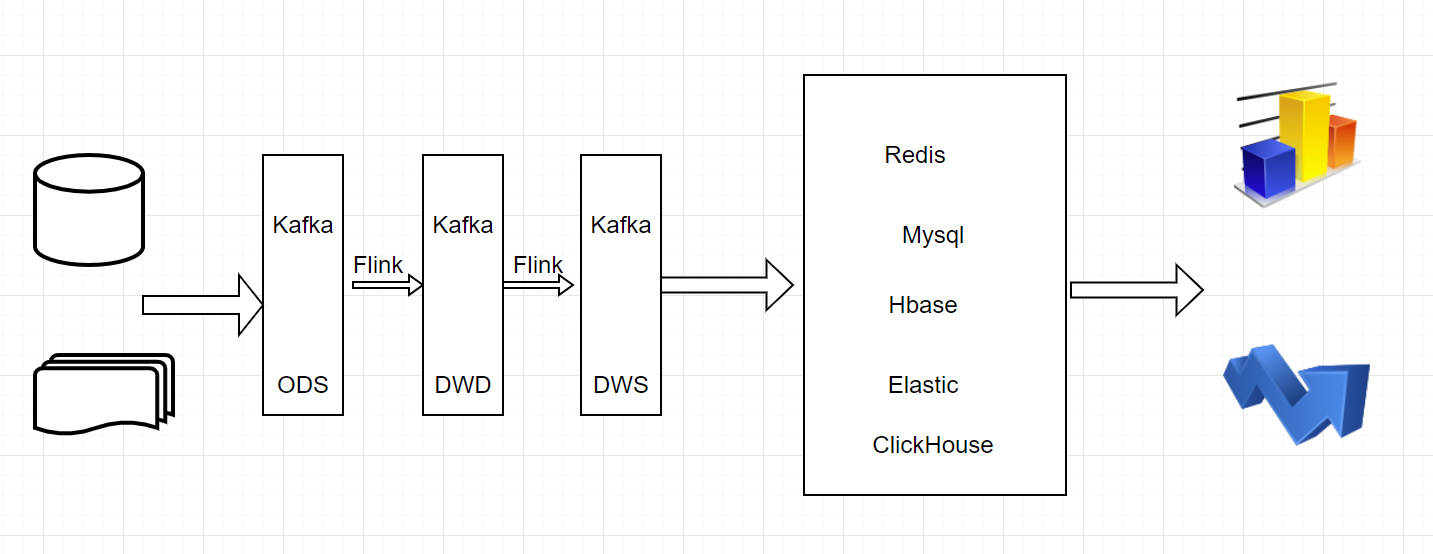


实时数仓



优点：开发简单

缺点：没有模型数据不能复用



优点：数据模型可复用，低延迟

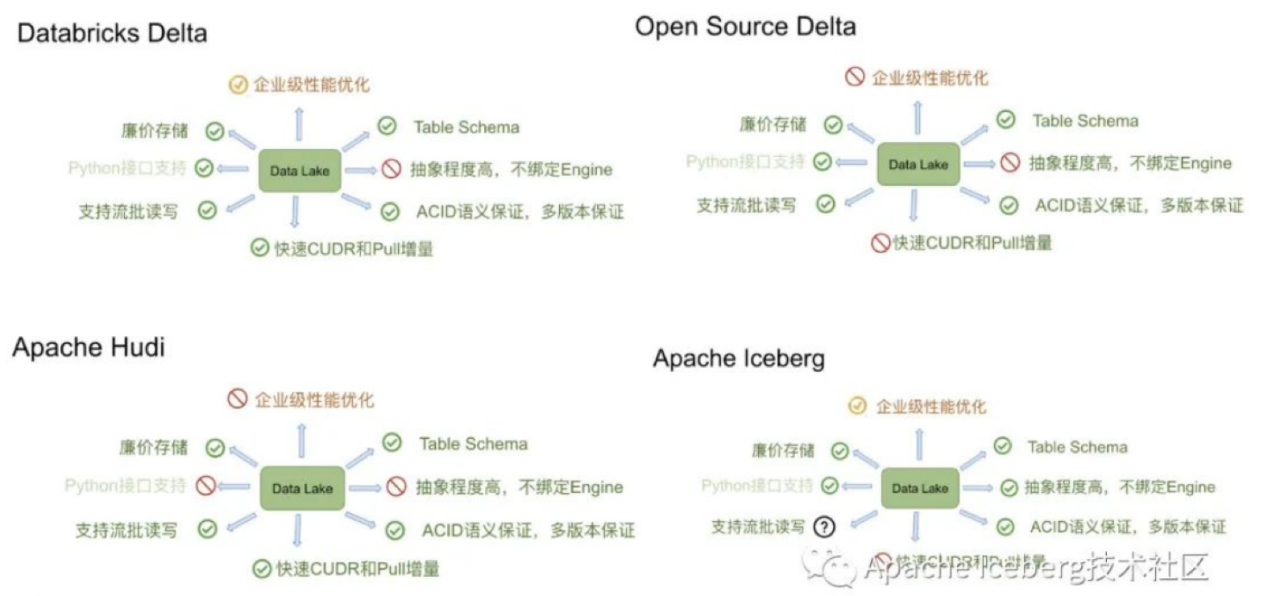
缺点：1.kafka无法支持海量数据

2.kafka 无法进行中间模型分析

数据湖架构



# 数据湖开源三大技术实现 Hudi 、iceberg 、delta Lake



# 应用场景

Hudi的应用场景：

你使用各种查询引擎，并且需要灵活管理变更的数据集。请注意，支持工具和整体的开发者体验可能较差。尽管有可能，但为实际的大规模生产工作负载安装和调试Hudi也需要一定的操作开销。如果你正在使用AWS管理服务，如Athena，Glue或EMR - Hudi已经预装和配置，并由AWS支持。

Iceberg的应用场景

你的主要痛点不是对现有记录的更改，而是在对象存储上管理巨大的表（超过10k个分区）的元数据负担。采用Iceberg将缓解与S3对象列表或Hive Metastore分区枚举相关的性能问题。相反，对删除和变更的支持还是初步的，而且涉及到数据保留的操作开销。

Delta lake的应用场景

对Spark支持友好，并且期望相对较低的写入吞吐量。如果你也已经是Databricks的客户，Delta Engine为读写性能和并发量都带来了显著的改善，对他们的生态系统进行双重开发是很有意义的。对于其他Apache Spark发行版来说，要明白Delta Lake虽然是开源的，但很可能会一直落后于Delta Engine，以起到产品差异化的作用