# 背景

# 概述

数据湖是目前比较热的一个概念，许多企业都在构建或者计划构建自己的数据湖。但是在计划构建数据湖之前，搞清楚什么是数据湖，明确一个数据湖项目的基本组成，进而设计数据湖的基本架构，对于数据湖的构建至关重要。关于什么是数据湖？有不同的定义。

Wikipedia上说数据湖是一类存储数据自然/原始格式的系统或存储，通常是对象块或者文件，包括原始系统所产生的原始数据拷贝以及为了各类任务而产生的转换数据，包括来自于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）、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）。

AWS定义数据湖是一个集中式存储库，允许您以任意规模存储所有结构化和非结构化数据。

微软的定义就更加模糊了，并没有明确给出什么是Data Lake，而是取巧的将数据湖的功能作为定义，数据湖包括一切使得开发者、数据科学家、分析师能更简单的存储、处理数据的能力，这些能力使得用户可以存储任意规模、任意类型、任意产生速度的数据，并且可以跨平台、跨语言的做所有类型的分析和处理。

关于数据湖的定义其实很多，但是基本上都围绕着以下几个特性展开。

1. 数据湖需要提供足够用的数据存储能力，这个存储保存了一个企业/组织中的所有数据。
2. 数据湖可以存储海量的任意类型的数据，包括结构化、半结构化和非结构化数据。
3. 数据湖中的数据是原始数据，是业务数据的完整副本。数据湖中的数据保持了他们在业务系统中原来的样子。
4. 数据湖需要具备完善的数据管理能力（完善的元数据），可以管理各类数据相关的要素，包括数据源、数据格式、连接信息、数据schema、权限管理等。
5. 数据湖需要具备多样化的分析能力，包括但不限于批处理、流式计算、交互式分析以及机器学习；同时，还需要提供一定的任务调度和管理能力。
6. 数据湖需要具备完善的数据生命周期管理能力。不光需要存储原始数据，还需要能够保存各类分析处理的中间结果，并完整的记录数据的分析处理过程，能帮助用户完整详细追溯任意一条数据的产生过程。
7. 数据湖需要具备完善的数据获取和数据发布能力。数据湖需要能支撑各种各样的数据源，并能从相关的数据源中获取全量/增量数据；然后规范存储。数据湖能将数据分析处理的结果推送到合适的存储引擎中，满足不同的应用访问需求。
8. 对于大数据的支持，包括超大规模存储以及可扩展的大规模数据处理能力。

综上，个人认为数据湖应该是一种不断演进中、可扩展的大数据存储、处理、分析的基础设施；以数据为导向，实现任意来源、任意速度、任意规模、任意类型数据的全量获取、全量存储、多模式处理与全生命周期管理；并通过与各类外部异构数据源的交互集成，支持各类企业级应用。

## 数据湖 vs 数据仓库

## 优点

# 原理

# 框架

## Iceberg

## Hudi

## DeltaLake

## TiDB

# 未来规划