# 背景

# 概述

## 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。

## 数据仓库

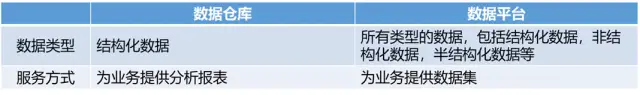
数据仓库(Data Warehouse)，也称为企业数据仓库，它是一个面向主题的、集成的、相对稳定的、反映历史变化的数据集合存储系统，它将来自不同来源的结构化数据聚合起来，用于业务智能领域的比较和分析，数据仓库是包含多种数据的存储库，并且是高度建模的。

**数据仓库 VS 数据湖**

相较而言，数据湖是较新的技术，拥有不断演变的架构。数据湖存储任何形式（包括结构化和非结构化）和任何格式（包括文本、音频、视频和图像）的原始数据。根据定义，数据湖不会接受数据治理，但专家们一致认为良好的数据管理对预防数据湖转变为数据沼泽不可或缺。数据湖在数据读取期间创建模式。与数据仓库相比，数据湖缺乏结构性，而且更灵活，并且提供了更高的敏捷性。值得一提的是，数据湖非常适合使用机器学习和深度学习来执行各种任务，比如数据挖掘和数据分析，以及提取非结构化。

## 数据湖

## 数据平台

由于数据仓库具有历史性的特性，其中存储的数据大多是结构化数据；而数据平台的出现解决了数据仓库不能处理非结构化数据和报表开发周期长的问题。  


通过以上的论述，我们发现数据平台和数据湖好像存在诸多相似性，这二者之间的区别，从个人角度理解上分析应该是数据加工的角度不同，数据湖更着重于对原始数据的存储，而数据平台则同数据仓库一样，需对原始数据进行清洗、转换等数据处理后按照统一的标准规范进行存储。

## 数据中台

数据中台通过对企业内外部多源异构的数据采集、治理、建模、分析和应用，使数据对内优化管理提高业务价值，对外进行数据合作让业务价值得到释放，使之成为企业数据资产管理中枢。数据中台建立后，会形成数据API服务，为企业和客户提供高效各种数据服务。

数据中台对一个企业的数字化转型和可持续发展起着至关重要的作用。数据中台为解耦而生，企业建设数据中台的最大意义就是应用与数据之间的解藕，这样企业就可以不受限制地按需构建满足业务需求的数据应用。

构建了开放、灵活、可扩展的企业级统一数据管理和分析平台，将企业内、外部数据随需关联，打破了数据的系统界限。

利用大数据智能分析、数据可视化等技术，实现了数据共享、日常报表自动生成、快速和智能分析，满足企业各级部门之间的数据分析应用需求。

深度挖掘数据价值，助力企业数字化转型落地。实现了数据的目录、模型、标准、认责、安全、可视化、共享等管理，实现数据集中存储、处理、分类与管理，建立大数据分析工具库、算法服务库，实现报表生成自动化、数据分析敏捷化、数据挖掘可视化，实现数据质量评估、落地管理流程。

**数据仓库 VS 数据中台**

数据仓库和传统的数据平台，其出发点为一个支撑性的技术系统，即一定要先考虑我具有什么数据，然后我才能干什么，因此特别强调数据质量和元数据管理；而数据中台的第一出发点不是[数据](http://mp.weixin.qq.com/s?__biz=MzIxMTE0ODU5NQ==&mid=2650243693&idx=1&sn=b721eaccd31b1f81cff1a63075c33d69&chksm=8f5ae831b82d61275943777963e4691848985ba4743f8418a1479af7baabd9af4abd3032314f&scene=21" \l "wechat_redirect" \t "https://mp.weixin.qq.com/_blank)而是业务，一开始不用看你系统里面有什么数据，而是去解决你的业务问题需要什么样的数据服务。

在具体的技术处理环节，二者也有明显不同，数据的预处理流程正在从传统的ETL结构向ELT结构转变。传统的数据仓库集成处理架构是ETL结构，这是构建数据仓库的重要一环，即用户从数据源抽取出所需的数据，经过数据清洗，将数据加载到数据仓库中去。而大数据背景下的架构体系是ELT结构，其根据上层的应用需求，随时从数据中台中抽取想要的原始数据进行建模分析。



# 原理

# 框架

## Iceberg

## Hudi

## DeltaLake

## TiDB

# 未来规划