7000字，详解仓湖一体架构！

在了解湖仓一体化之前，我们先来看一则有关数据仓库的有趣故事吧~

沃尔玛拥有世界上最大的数据仓库系统，它利用数据挖掘方法对交易数据进行分析后发现"跟尿布一起购买最多的商品竟是啤酒！后来经过大量实际调查和分析，发现在美国，一些年轻的父亲下班后经常要到超市去买婴儿尿布，而他们中有30%～40%的人同时也为自己买一些啤酒，这是因为美国的太太们常叮嘱她们的丈夫下班后为小孩买尿布，而丈夫们在买尿布后又随手带回了他们喜欢的啤酒。

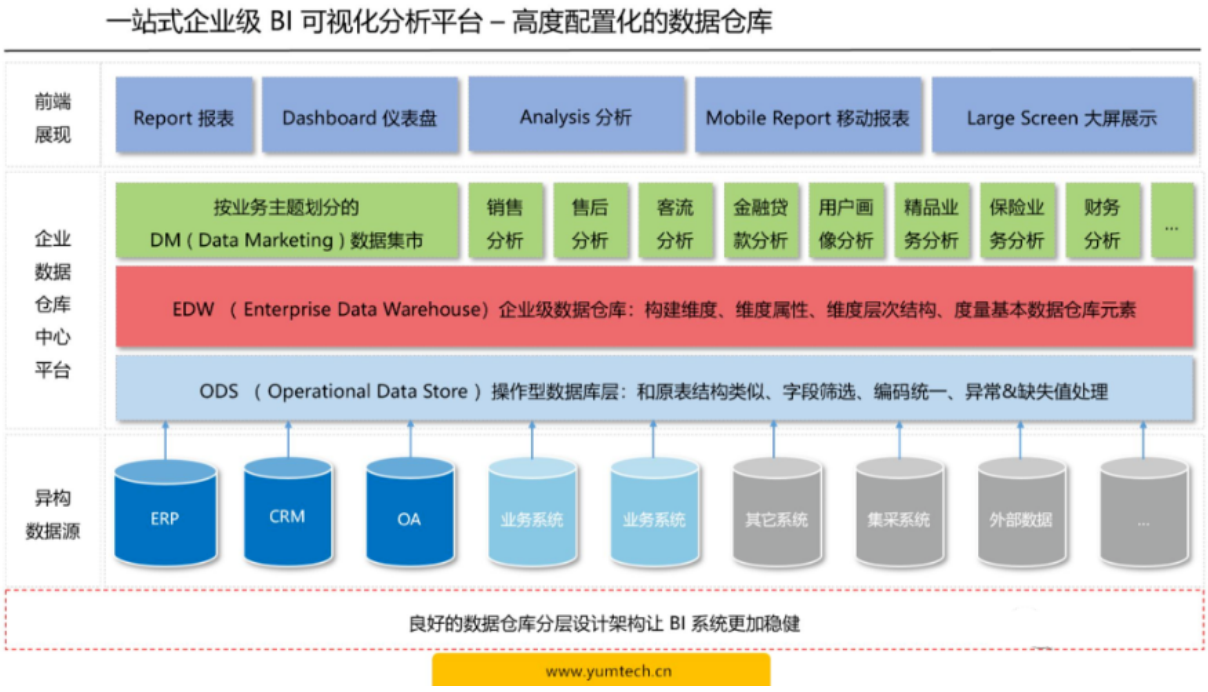
这就是大数据领域经常讲的啤酒与尿布的故事！

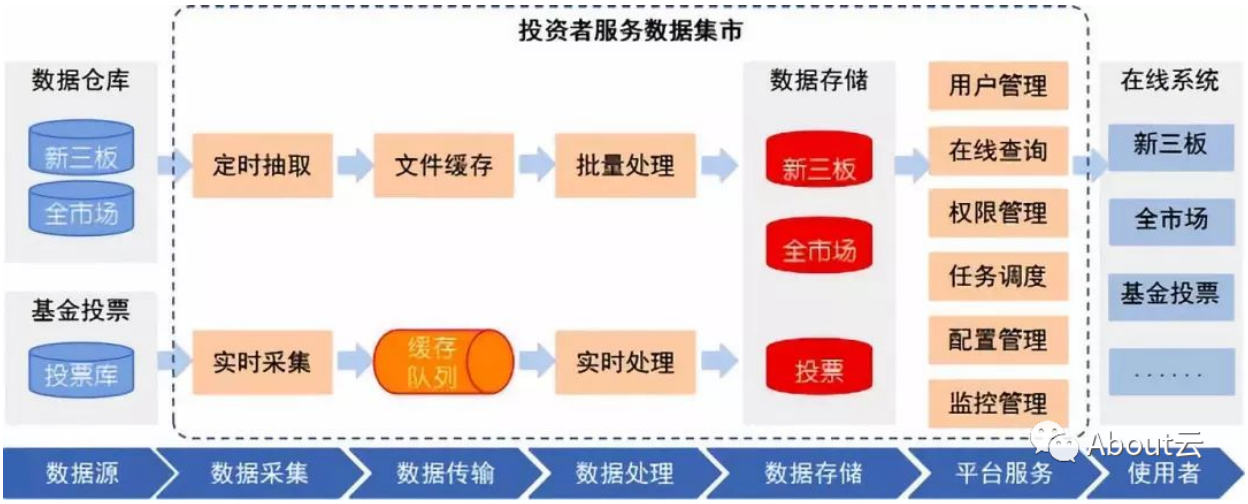
01 什么是数据仓库、数据集市和数据湖？

一、数据仓库

早期系统采用数据库来存放管理数据，但是随着大数据技术的兴起，大家想要通过大数据技术来找到数据之间可能存在的关系，所以大家设计了一套新的数据存储管理系统，把所有的数据全部存储到数据仓库，然后统一对数据处理，这个系统叫做数据仓库。而数据库缺少灵活和强大的处理能力。

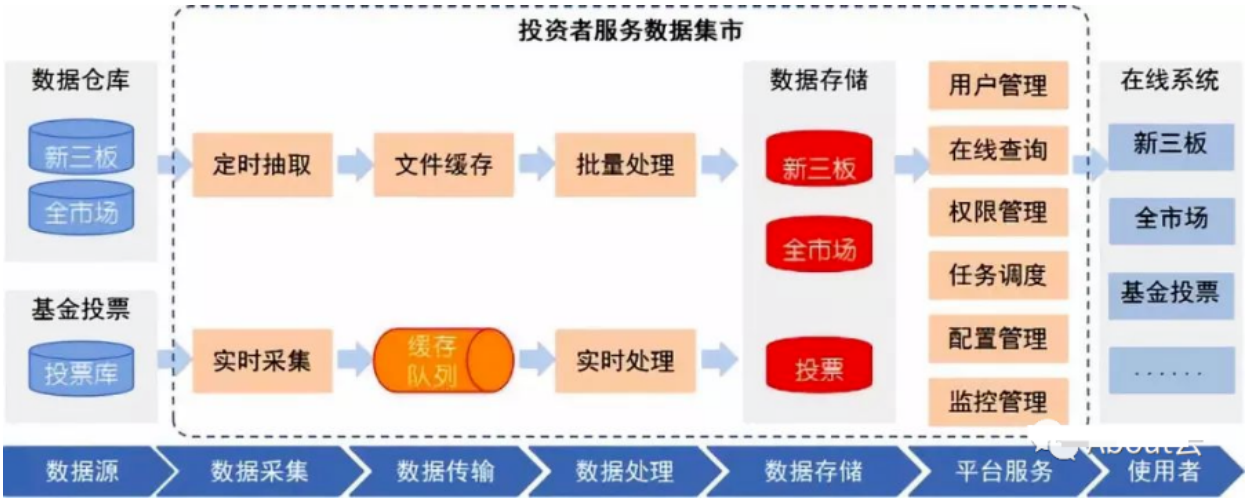
在计算机领域，数据仓库（英语：data warehouse，也称为企业数据仓库）是用于报告和数据分析的系统，被认为是商业智能的核心组件。数据仓库是来自一个或多个不同源的集成数据的中央存储库。数据仓库将当前和历史数据存储在一起，以利各种分析方法如在线分析处理(OLAP)、数据挖掘(Data Mining)，帮助决策者能快速从大量数据中，分析出有价值的信息，帮助建构商业智能(BI)。 尽管仓库非常适合结构化数据，但是许多现代企业必须处理非结构化数据，半结构化数据以及具有高多样性、高速度和高容量的数据。数据仓库不适用于许多此类场景，并且成本效益并非最佳。





二、数据集市

每个部门自身也有对业务数据进行处理分析统计的需求，但不涉及到和其他数据，不希望在数据量大的数据仓库进行操作（因为操作慢，而且可能影响到其他人处理数据），所以建立一个新的存储系统，把数据仓库里关联自己的数据存储到这个系统，本质上算是数据仓库的一个子集。这个系统叫做数据集市。 例如公司里的某一个部门想对投资者服务数据进行分析，于是他们建立一个投资者服务数据的数据集市，其中数据从数据仓库中抽取：



三、数据湖

随着当前大量信息化发展和电子设备产品普及，产生大量的照片、视频、文档等非结构化数据，人们也想通过大数据技术找到这些数据的关系，所以设计了一个比数据仓库还要大的系统，可以把非结构化和结构化数据共同存储和做一些处理，这个系统叫做数据湖。 数据仓库的成长性很好，而数据湖更灵活。**数据仓库支持的数据结构种类比较单一，数据湖的种类比较丰富，可以包罗万象。**数据仓库更加适合成熟的数据当中的分析和处理，数据湖更加适合在异构数据上的价值的挖掘。

数据湖虽然适合存储数据，但缺少一些关键功能：它们不支持事务处理，不保证数据质量，并且缺乏一致性/隔离性，从而几乎无法实现混合追加和读取数据，以及完成批处理和流式作业。由于这些原因，数据湖的许多功能尚未实现，并且在很多时候丧失了数据湖的优势。

02 数据湖+数据仓=湖仓一体？

在湖仓一体出现之前，数据仓库和数据湖是被人们讨论最多的话题。

正式切入主题前，先跟大家科普一个概念，即大数据的工作流程是怎样的？这里就要涉及到两个相对陌生的名词：数据的结构化程度和数据的信息密度。前者描述的是数据本身的规范性，后者描述的是单位存储体积内、包含的信息量的大小。

一般来说，人们获取到的原始数据大多是非结构化的，且信息密度比较低，通过对数据进行清洗、分析、挖掘等操作，可以排除无用数据、找到数据中的关联性，在这个过程中，数据的结构化程度、信息密度也随之提升，最后一步，就是把优化过后的数据加以利用，变成真正的生产资料。

简而言之，大数据处理的过程其实是一个提升数据结构化程度和信息密度的过程。在这个过程中，数据的特征一直在发生变化，不同的数据，适合的存储介质也有所不同，所以才有了一度火热的数据仓库和数据湖之争。

数据仓库是一个面向主题的、集成的、相对稳定的、反映历史变化的数据集合，主要用于支持管理决策和信息的全局共享。简单点说，数据仓库就像是一个大型图书馆，里面的数据需要按照规范放好，你可以按照类别找到想要的信息。

就目前来说，对数据仓库的主流定义是位于多个数据库上的大容量存储库，它的作用在于存储大量的结构化数据，为管理分析和业务决策提供统一的数据支持，虽然存取过程相对比较繁琐，对于数据类型有一定限制，但在那个年代，数据仓库的功能性已经够用了，所以在2011年前后，市场还是数据仓库的天下。

到了互联网时代，数据量呈现“井喷式”爆发，数据类型也变得异构化。受数据规模和数据类型的限制，传统数据仓库无法支撑起互联网时代的商业智能，随着Hadoop与对象存储的技术成熟，数据湖的概念应用而生，在2011年由James Dixon提出。

相比于数据仓库，**数据湖是一种不断演进中、可扩展的大数据存储、处理、分析的基础设施。**它就像一个大型仓库，可以存储任何形式（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，数据湖通常更大，存储成本也更为廉价。但它的问题也很明显，数据湖缺乏结构性，一旦没有被治理好，就会变成数据沼泽。

从产品形态上来说，数据仓库一般是独立标准化产品，数据湖更像是一种架构指导，需要配合着系列周边工具，来实现业务需要。换句话说，数据湖的灵活性，对于前期开发和前期部署是友好的；数据仓库的规范性，对于大数据后期运行和公司长期发展是友好的，那么，有没有那么一种可能，有没有一种新架构，能兼具数据仓库和数据湖的优点呢？

于是，湖仓一体诞生了。

依据DataBricks公司对Lakehouse 的定义，**湖仓一体是一种结合了数据湖和数据仓库优势的新范式，在用于数据湖的低成本存储上，实现与数据仓库中类似的数据结构和数据管理功能。**湖仓一体是一种更开放的新型架构，有人把它做了一个比喻，就类似于在湖边搭建了很多小房子，有的负责数据分析，有的运转机器学习，有的来检索音视频等，至于那些数据源流，都可以从数据湖里轻松获取。

就湖仓一体发展轨迹来看，早期的湖仓一体，更多是一种处理思想，处理上将数据湖和数据仓库互相打通，现在的湖仓一体，虽然仍处于发展的初期阶段，但它已经不只是一个纯粹的技术概念，而是被赋予了更多与厂商产品层面相关的含义和价值。

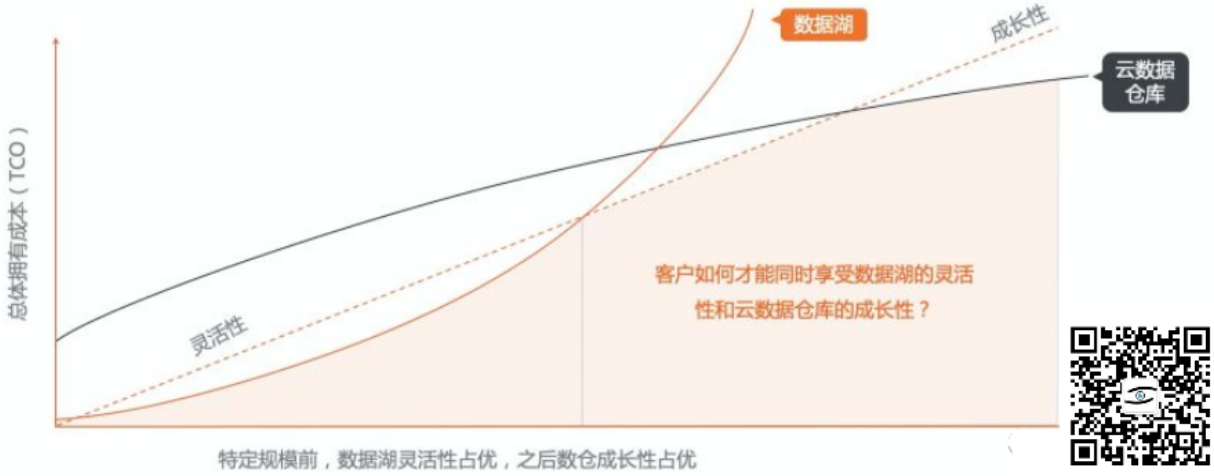
这里需要注意的是，**“湖仓一体”并不等同于“数据湖”+“数据仓”，这是一个极大的误区**，现在很多公司经常会同时搭建数仓、数据湖两种存储架构，一个大的数仓拖着多个小的数据湖，这并不意味着这家公司拥有了湖仓一体的能力，湖仓一体绝不等同于数据湖和数据仓简单打通，反而数据在这两种存储中会有极大冗余度。

03 为什么会诞生湖仓一体化？

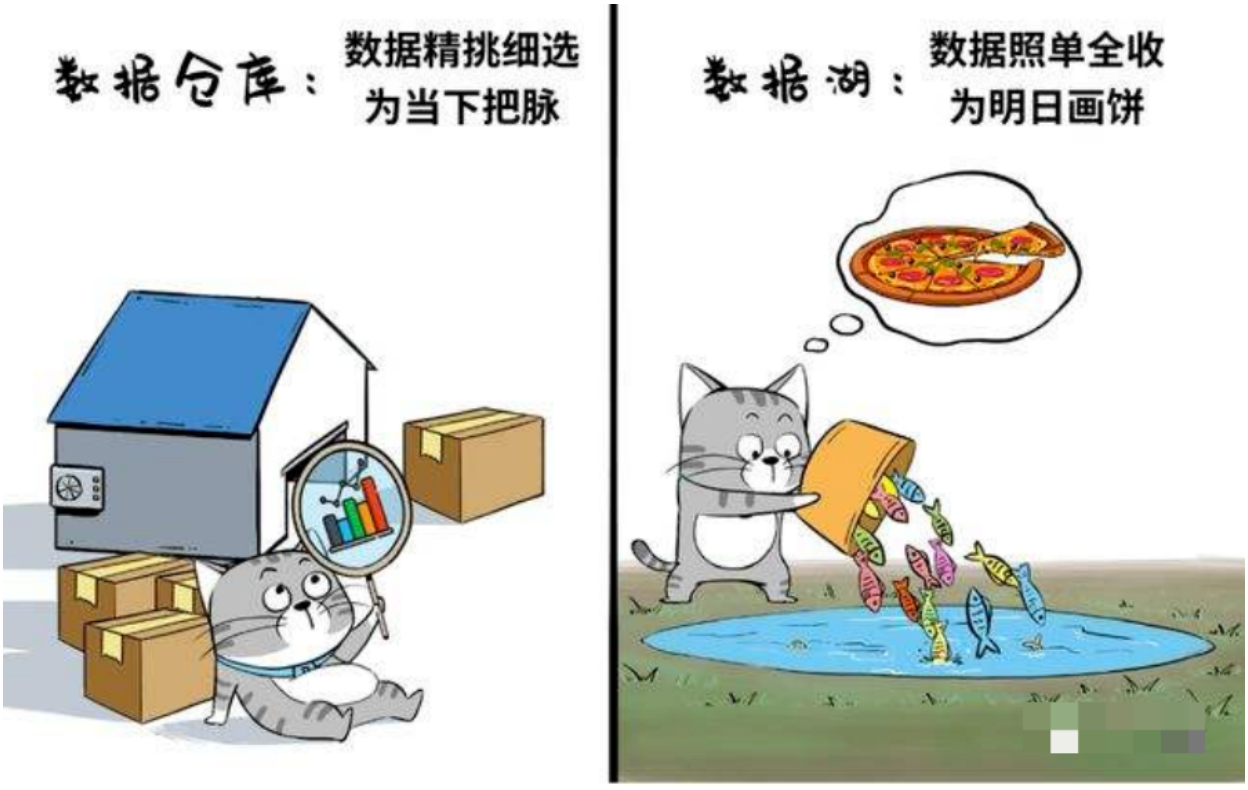
1、打通数据的存储与计算

很多公司对各类数据应用包括 SQL 分析、实时监控、数据科学和机器学习的灵活性、高性能系统的需求并未减少。AI 的大部分最新进展是基于更好地处理非结构化数据（如 text、images、video、audio ）的模型，完全纯数据仓库的二维关系表已经无法承接半/非结构化数据的处理，AI 引擎不可能只跑在纯数据仓库模型上。

一种常见的解决方案是结合数据湖和数据仓库优势，建立湖仓一体化，进而解决了数据湖的局限性：直接在用于数据湖的低成本存储上实现与数据仓库中类似的数据结构和数据管理功能。 之前的微博基于大数据的需求发展了数据仓库平台，基于AI的需求，发展了数据湖平台，这两套大数据平台在集群层面完全是割裂的，数据和计算无法在两个平台间自由流动。而使用湖仓一体，就能实现数据湖和数仓之间的无缝流转，打通了数据存储和计算的不同的层面。 **2、灵活性与成长性兼得**



通过上面这张图，可知灵活性和成长性，对于处于不同时期的企业来说，重要性不同。 当企业处于初创阶段，数据从产生到消费还需要一个创新探索的阶段才能逐渐沉淀下来，那么用于支撑这类业务的大数据系统，灵活性就更加重要，数据湖的架构更适用。 当企业逐渐成熟起来，已经沉淀为一系列数据处理流程，问题开始转化为数据规模不断增长，处理数据的成本不断增加，参与数据流程的人员、部门不断增多，那么用于支撑这类业务的大数据系统，成长性的好坏就决定了业务能够发展多远。数据仓库的架构更适用。

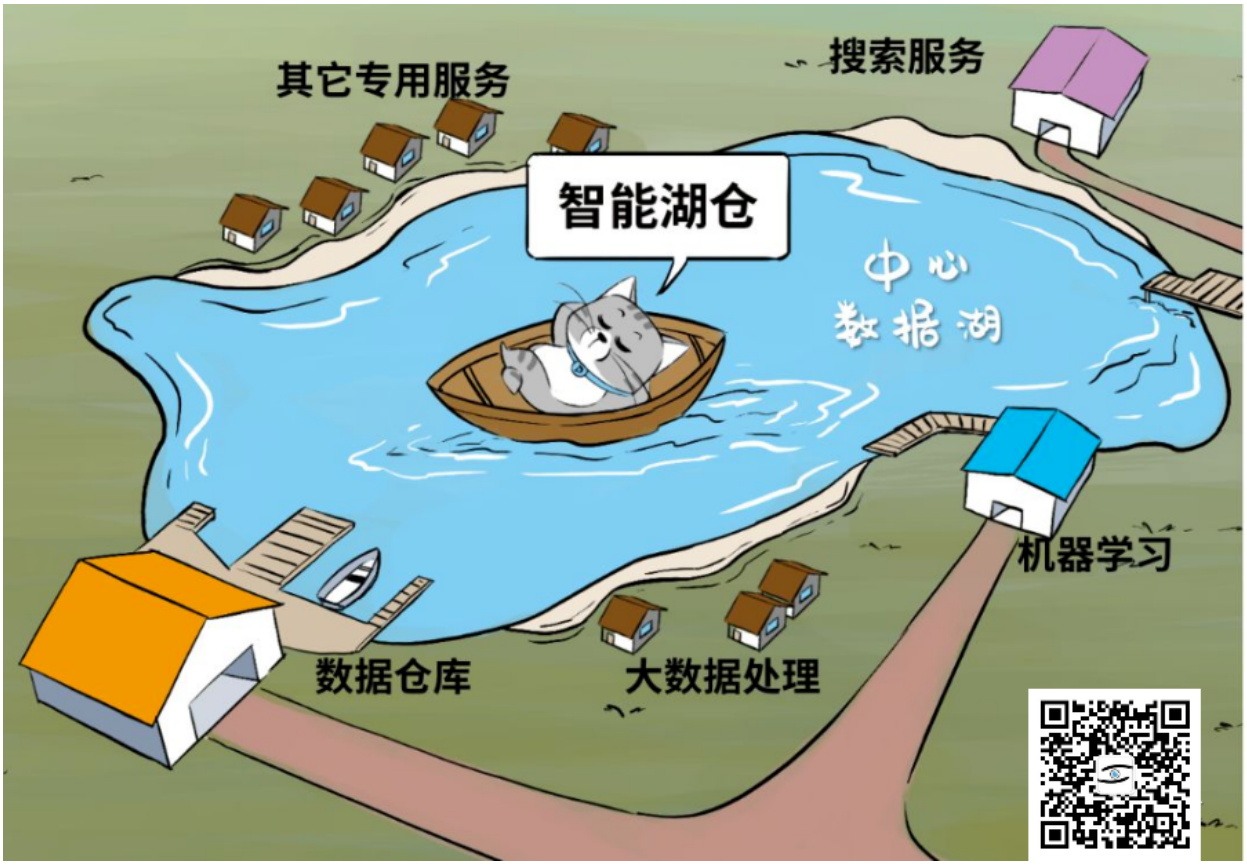


经过对数据湖和数据仓库的深入阐述和比较，可以发现：数据湖和数据仓库一个面向初创用户友好，一个成长性更佳。对企业来说，数据湖和数据仓库是否必须是一个二选一的选择题？是否能有一种方案同时兼顾数据湖的灵活性和云数据仓库的成长性，将二者有效结合起来为用户实现更低的总体拥有成本？那么湖仓一体化就是答案！

04 什么是湖仓一体化？

随着当前大数据技术应用趋势，企业对单一的数据湖和数仓架构并不满意。越来越多的企业开始融合数据湖和数据仓库的平台，不仅可以实现数据仓库的功能，同时还实现了不同类型数据的处理功能、数据科

学、用于发现新模型的高级功能。 湖仓一体是一种新型开放式架构，将数据湖和数据仓库的优势充分结合，它构建在数据湖低成本的数据存储架构之上，又继承了数据仓库的数据处理和管理功能，打通数据湖和数据仓库两套体系，让数据和计算在湖和仓之间自由流动。作为新一代大数据技术架构，将逐渐取代单一数据湖和数据仓库架构。 **有人把“湖仓一体”做了形象的比喻，就好像湖边搭建了很多小房子，有的可以负责数据分析，有的来运转机器学习，有的来检索音视频等等，而这些数据源流，都可以从数据湖里轻松取得。**



05 湖仓一体Data Lakehouse介绍

Data Lakehouse（湖仓一体）是新出现的一种数据架构，它同时吸收了数据仓库和数据湖的优势，数据分析师和数据科学家可以在同一个数据存储中对数据进行操作，同时它也能为公司进行数据治理带来更多的便利性。那么何为Data Lakehouse呢，它具备些什么特性呢？

一直以来，我们都在使用两种数据存储方式来架构数据：

数据仓库：数仓这样的一种数据存储架构，它主要存储的是以关系型数据库组织起来的结构化数据。数据通过转换、整合以及清理，并导入到目标表中。在数仓中，数据存储的结构与其定义的schema是强匹配的。

数据湖：数据湖这样的一种数据存储结构，它可以存储任何类型的数据，包括像图片、文档这样的非结构化数据。数据湖通常更大，其存储成本也更为廉价。存储其中的数据不需要满足特定的schema，数据湖也不会尝试去将特定的schema施行其上。相反的是，数据的拥有者通常会在读取数据的时候解析schema（schema-on-read），当处理相应的数据时，将转换施加其上。

现在许多的公司往往同时会搭建数仓、数据湖这两种存储架构，一个大的数仓和多个小的数据湖。这样，数据在这两种存储中就会有一定的冗余。

Data Lakehouse的出现试图去融合数仓和数据湖这两者之间的差异，通过将数仓构建在数据湖上，使得存储变得更为廉价和弹性，同时lakehouse能够有效地提升数据质量，减小数据冗余。在lakehouse的构建中，ETL起了非常重要的作用，它能够将未经规整的数据湖层数据转换成数仓层结构化的数据。

Data Lakehouse概念是由Databricks提出的，在提出概念的同时，也列出了如下一些特性：

- **事务支持：**Lakehouse可以处理多条不同的数据管道。这意味着它可以在不破坏数据完整性的前提下支持并发的读写事务。
- **Schemas：**数仓会在所有存储其上的数据上施加Schema，而数据湖则不会。Lakehouse的架构可以根据应用的需求为绝大多数的数据施加schema，使其标准化。
- **报表以及分析应用的支持：**报表和分析应用都可以使用这一存储架构。Lakehouse里面所保存的数据经过了清理和整合的过程，它可以用来加速分析。同时相比于数仓，它能够保存更多的数据，数据的时效性也会更高，能显著提升报表的质量。
- **数据类型扩展：**数仓仅可以支持结构化数据，而Lakehouse的结构可以支持更多不同类型的数据，包括文件、视频、音频和系统日志。
- **端到端的流式支持：**Lakehouse可以支持流式分析，从而能够满足实时报表的需求，实时报表在现在越来越多的企业中重要性在逐渐提高。
- **计算存储分离：**我们往往使用低成本硬件和集群化架构来实现数据湖，这样的架构提供了非常廉价的分离式存储。Lakehouse是构建在数据湖之上的，因此自然也采用了存算分离的架构，数据存储在一个集群中，而在另一个集群中进行处理。
- **开放性：**Lakehouse在其构建中通常会使Iceberg，Hudi，Delta Lake等构建组件，首先这些组件是开源开放的，其次这些组件采用了Parquet，ORC这样开放兼容的存储格式作为下层的数据存储格式，因此不同的引擎，不同的语言都可以在Lakehouse上进行操作。

Lakehouse的概念最早是由Databricks所提出的，而其他的类似的产品有Azure Synapse Analytics。Lakehouse技术仍然在发展中，因此上面所述的这些特性也会被不断的修订和改进。

06 湖仓一体化有什么好处？

湖仓一体能发挥出数据湖的灵活性与生态丰富性，以及数据仓库的成长性与企业级能力。帮助企业建立数据资产、实现数据业务化、进而推进全线业务智能化，实现数据驱动下的企业数据智能创新，全面支撑企业未来大规模业务智能落地。其主要优势主要有以下几个方面： **数据重复性：**如果一个组织同时维护了一个数据湖和多个数据仓库，这无疑会带来数据冗余。在最好的情况下，这仅仅只会带来数据处理的不高效，但是在最差的情况下，它会导致数据不一致的情况出现。湖仓一体的结合，能够去除数据的重复性，真正做到了唯一。 **高存储成本：**数据仓库和数据湖都是为了降低数据存储的成本。数据仓库往往是通过降低冗余，以及整合异构的数据源来做到降低成本。而数据湖则往往使用大数据文件系统和Spark在廉价的硬件上存储计算数据。湖仓一体架构的目标就是结合这些技术来最大力度降低成本。 **报表和分析应用之间的差异：**数据科学倾向于与数据湖打交道，使用各种分析技术来处理未经加工的数据。而报表分析师们则倾向于使用整合后的数据，比如数据仓库或是数据集市。而在一个组织内，往往这两个团队之间没有太多的交集，但实际上他们之间的工作又有一定的重复和矛盾。而当使用湖仓一体架构后，两个团队可以在同一数据架构上进行工作，避免不必要的重复。 **数据停滞：**在数据湖中，数据停滞是一个最为严重的问题，如果数据一直无人治理，那将很快变为数据沼泽。我们往往轻易的将数据丢入湖中，但缺乏有效的治理，长此以往，数据的时效性变得越来越难追溯。湖仓一体的引入，对于海量数据进行治理，能够更有效地帮助提升分析数据的时效性。 **潜在不兼容性带来的风险：**数据分析仍是一门兴起的技术，新的工具和技术每年仍在不停地出现中。一些技术可能只和数据湖兼容，而另一些则又可能只和数据仓库兼容。湖仓一体的架构意味着为两方面做准备。

07 湖仓一体落地路径与成本

A：现在大多数企业都已经有了自己的一套大数据架构，他们如何基于已有的架构落地湖仓一体？有哪些可行的落地路径？成本可能主要会来自哪里？

Q：现在有一部分企业已经有了自己的大数据架构，这些企业相对来说可能诞生的比较早，大多数都是选的 Hadoop 体系，或是自建的 Hadoop 体系，或是使用云上托管的 Hadoop 体系。这些企业可以有很多选择，他可以选择像 Databricks 那样的方案，也可以选择像 MaxCompute 这样的方案。

这两条路径都相对可行，那怎么选？这通常要看企业是不是希望在大数据技术栈上做更多投入。如果企业觉得没必要在基础设施上投很多资源，而是要把更多资源放在业务上，那选一个更偏全托管版的湖仓一体解决方案更有价值。反之，如果企业技术人员很多，希望底层基础设施足够灵活并且是自己可控的，就可以选择在湖上建仓的模式。

还有一些比较新的企业，比如过去三年内成立的，它们有很多都处于高速增长阶段。这些企业其实天生就长在云上，甚至一开始选的大数据架构就已经是云数仓的架构，这类企业基于现有的架构向前演进相对比较简单。只要尽量使用云基础设施，开通几个云服务就能形成一套湖仓一体架构了，这是一个简单直接且相对单一化的路径。

那成本主要来自哪里？如果企业选择全托管的湖仓一体解决方案，则成本主要来自于对当前数据，比如数仓迁移、数据整理等一次性开支，一旦这部分工作做完，后续在数据治理上形成正循环，整体成本不会太高。如果企业选择自己维护一套湖仓一体架构，则成本主要来自持续维护和调优整套基础设施的人力成本和硬件成本。

A：根据您的了解，当前企业尝试落地湖仓一体的时候遇到的问题和挑战主要有哪些？现在是采用湖仓一体的好时机吗？

Q：现在大多数企业都还没有用到湖仓一体的新架构，他们要么选择了数据湖方案，要么选择了数仓方案。湖仓一体作为一个新兴架构，很多企业目前还在早期探索阶段。有些企业在把数据放到数据湖上之后，发现在数据湖上做好数据治理或者数据管理相对比较困难，这个时候再去采用湖仓一体模式，在现有相对更灵活但不够管理化的数据上，再抽象一层数仓层和治理层，对数据做更好的管理和治理。对于数仓的用户，如果采用的数仓系统支持湖仓一体架构，直接挂载数据湖就好了。

企业尝试落地湖仓一体时会遇到的问题和挑战主要有几点。首先，如果团队没有足够好的数据治理或数据管理经验，挑战会比较大。这也是为什么我们推出的方案几乎都在向全托管或全服务的 SaaS 模式走，就是希望能够降低门槛。

其次，对于自建湖仓一体的企业，他们会遇到的挑战主要是湖仓一体的高复杂度，特别是湖仓之间如何协同的问题，这里面涉及到两套系统存储打通的问题、元数据一致性问题、湖和仓上不同引擎之间数据交叉引用的问题，以及带宽问题、安全问题，等等。另外，由于湖仓一体架构底层是一个二元体系，那向上面向用户的时候，用户是不是能看到两个体系？如果用户能够看到两个体系的话，如何区分和引导？如果用户看不到的话，那底下开发需要做什么样的封装？这些都是自建湖仓体系会遇到的问题。

总之，如果企业并不是一定要大力投入做基础设施的话，直接采用全托管版本的湖仓一体的架构会简单很多。

最后，湖仓一体还是一个新兴的方向，很多问题还在探索中，比如哪些数据放在数仓 / 数据湖？更适合有一定探索和创新意愿的企业。