**数据库、数据湖、数据仓库、湖仓一体、智能湖仓，分别都是什么鬼？**

<https://mp.weixin.qq.com/s/UGQi_1vjEK7CTdYBNCCJNA>

**数据仓库**

数据仓库相当于一个集成化数据管理的平台，从多个数据源抽取有价值的数据，在仓库内转换和流动，并提供给BI等分析工具来输出干货。

因为分析型业务需要大量的“读”操作，所以数据仓库通过“Denormalized”化的方式优化表结构，减少表间联接，牺牲空间来换取读性能。（一张表里的冗余数据增加了，但查询起来却更快了），并使用列式存储优化，来进一步提高查询速度、降低开销。

再结合面向分析场景的Schema设计，数据仓库就可以高效率、全方位、多维度的扛起“联机分析”重任了。



**数据湖**

企业希望把生产经营中的所有相关数据，历史的、实时的，在线的、离线的，内部的、外部的，结构化的、非结构化的，都能完整保存下来，方便“沙中淘金”。

**数据库和数据仓库都干不了这活儿，怎么办呢？**

挖个大坑，修个湖，把各种数据一滚脑灌进去囤起来，而且要持续灌，持续囤。**这就是数据湖啦！**

数据湖的本质，是由“➊**数据存储架构+➋数据处理工具**”组成的**解决方案**，而不是某个单一独立产品。

**➊数据存储架构**，要有足够的扩展性和可靠性，要满足企业能把所有原始数据都“囤”起来，存得下、存得久。

一般来讲，各大云厂商都喜欢用对象存储来做数据湖的存储底座，**比如 A**mazon **W**eb **S**ervices（亚马逊云科技），**修建“湖底”用的“砖头”，就是S3云对象存储。**

**➋数据处理工具，则分为两大类↓**

**第一类工具，解决的问题是**如何把数据“搬到”湖里，包括定义数据源、制定数据访问策略和安全策略，并移动数据、编制数据目录等等。

如果没有这些数据管理/治理工具，元数据缺失，湖里的数据质量就没法保障，“泥石俱下”，各种数据倾泻堆积到湖里，最终好好的数据湖，慢慢就变成了**数据沼泽**。

因此，在一个数据湖方案里，数据移动和管理的工具非常重要。

比如，**A**mazon **W**eb **S**ervices提供“**Lake Formation**”这个工具，帮助客户自动化地把各种数据源中的数据移动到湖里，同时还可以调用**Amazon Glue**来对数据进行ETL，编制数据目录，进一步提高湖里数据的质量。

**第二类工具，就是要从湖里的海量数据中“淘金”。**

数据并不是存进数据湖里就万事大吉，要对数据进行分析、挖掘、利用，比如要对湖里的数据进行查询，同时要把数据提供给机器学习、数据科学类的业务，便于“点石成金”。

我们继续拿**A**mazon **W**eb **S**ervices来举例子，基于**Amazon Athena**这个服务，就可以使用标准的SQL来对S3（数据湖）中的数据进行交互式查询。

再比如使用**Amazon SageMaker**机器学习服务，导入数据湖中的数据进行模型训练，这些都是常规操作。

小结一下，数据湖不只是个“囤积”数据的“大水坑”，除了用存储技术构建的湖底座以外，还包含一系列的数据入湖、数据出湖、数据管理、数据应用工具集，共同组成了数据湖解决方案。

**数据湖和数据仓库区别在哪儿？**

这个问题其实不难回答，我们先看下面这张对比表。

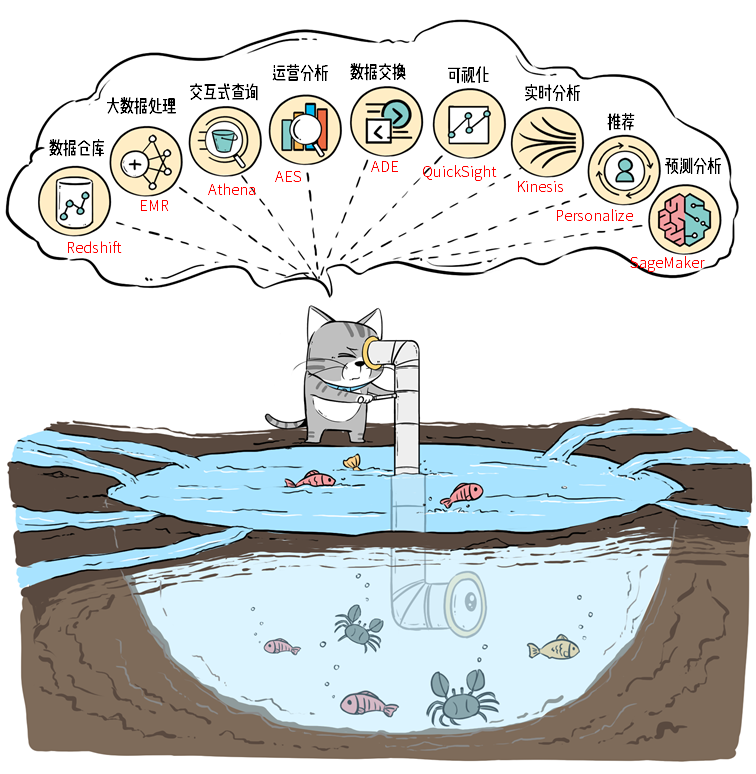


从数据含金量来比，数据仓库里的数据价值密度更高一些，数据的抽取和Schema的设计，都有非常强的针对性，便于业务分析师迅速获取洞察结果，用与决策支持。

而数据湖更有一种“兜底”的感觉，甭管当下有用没有/或者暂时没想好怎么用，先保存着、沉淀着，将来想用的时候，尽管翻牌子就是了，反正都原汁原味的留存了下来。

而从产品形态看，数据仓库可以是独立的标准化产品，拿云上数仓来举例，**Amazon Redshift**，就是一款“数仓产品”。

数据湖则是一种架构，通常是围绕对象存储为“湖底座”的大数据管理方案组合。比如，**A**mazon **W**eb **S**ervices并**没有**哪个产品叫“数据湖”，**而是以S3为基础，结合一系列数据管理工具**，帮助客户构建云上“数据湖”↓

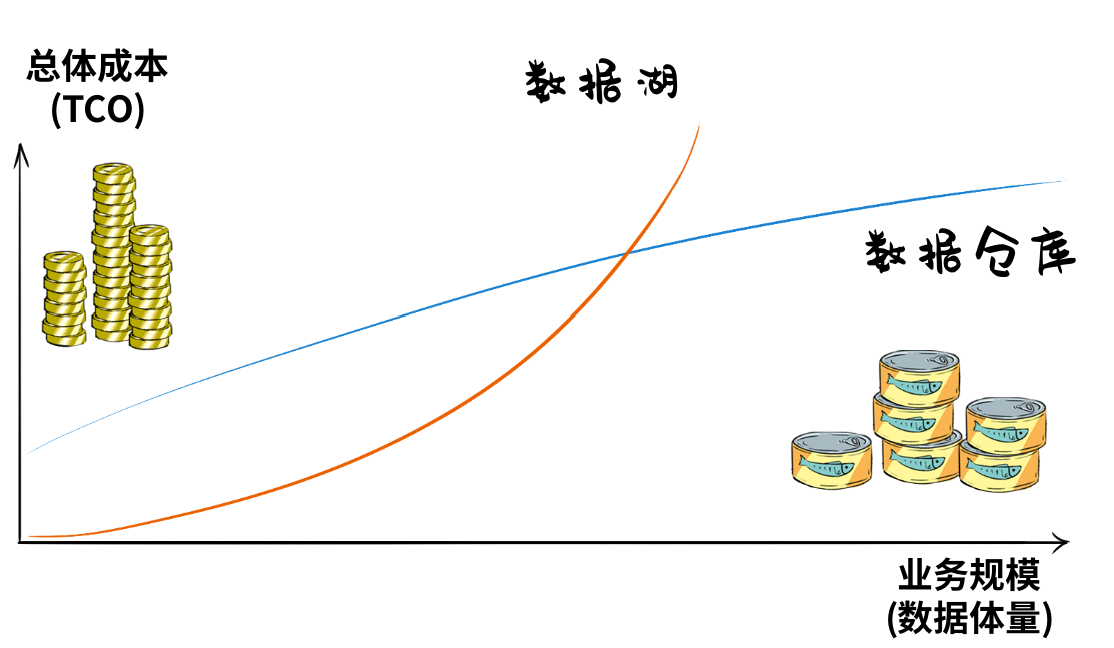


**为什么要把“湖”和“仓”糅到一起？**

曾经，数据仓库擅长的BI、数据洞察离业务更近、价值更大，而数据湖里的数据，更多的是为了远景画饼。

随着大数据和AI的上纲上线，原先的“画的饼”也变得炙手可热起来，为业务赋能，价值被重新定义。

而因为数仓和数据库的出发点不同、架构不同，企业在实际使用过程中，“性价比”差异很大。



数据湖起步成本很低，但随着数据体量增大，TCO成本会加速飙升，数仓则恰恰相反，前期建设开支很大。

总之，一个后期成本高，一个前期成本高，对于**既想修湖、又想建仓**的用户来说，仿佛玩了一个金钱游戏。

于是，人们就想，既然都是拿数据为业务服务，数据湖和数仓作为两大“数据集散地”，能不能彼此整合一下，让数据流动起来，少点重复建设呢？

**到底什么才是真正的Lake House？**

Lake House，坊间通常称之为“湖仓一体”，而**A**mazon **W**eb **S**ervices则叫做“**智能湖仓**”。

Lake House架构最重要的一点，是实现“湖里”和“仓里”的数据/元数据能够无缝打通，并且“自由”流动。