

**[浅谈数据仓库和大数据](https://www.cnblogs.com/lxbmaomao/p/9525487.html)**

**前言**

数据仓库是今年来适应利用数据支持决策分析的强烈需求而发展起来的数据库应用技术，诚然，数据仓库以数据库为基础，但是他在需求、客户、体系结构与运行机制等方面与数据库存在重大的不同，Kimball说:"我们花了二十年的时间往数据库中加入数据，现在该是拿出来使用的时候了。"    ---摘自数据仓库工具箱:维度建模(第二版)

**1.数据仓库理解**

根据笔者自己的理解，数据仓库是一个抽象的概念，而实现的载体则是我们常见的各种数据库表。比如传统行业中用到的Oracle、Teradata(简称TD)、GreenPlum(简称GP)，互联网行业中用到的Hive、Spark。它的一个主要应用点的体现就是我们企业中建设的数据平台

**2.数据平台&数据仓库的组成**

先宽泛的讲述下数据平台和数据仓库的组成

数据平台可划分为如下子平台：数据同步平台、数据仓库、调度平台、元数据管理平台、即席查询平台、数据可视化平台、数据质量检测系统

数据仓库由于是一个抽象的概念所以可以简单的理解为不同粒度的数据层，比如：数据缓冲层(存放当日增量数据)、数据明细层(存放最全的明细数据)、数据模型层(轻粒度的数据汇总以及模型设计，这个时候需要设计相应的主题)、数据集市层(一般就是一些宽表，包含多维度和指标，方便用来做多维分析)、数据应用层(主要是开放给业务侧使用，多存放粗粒度的数据报表)

每个公司对每层的命名不一样，所以就不用什么ods，dw等来标识各个层级。

**3.数仓和大数据的关系**

在写对数仓的理解的时候其实已经点出了数仓和大数据的关系，现在讲一下我个人理解的数仓演变历史以及为什么要用大数据来实现数仓。

**3.1 数仓的过去**

数据仓库在很早的时候就被数据仓库之父Inmon 提出来了，按我的理解，很早的时候主要是对自己企业内部业务数据的一个分析决策，用一些传统的关系型数据库为载体，加上Kettle、Informatica、DataStage等ETL工具以及Biee、smartBi等报表工具来支撑企业自己的数据仓库建设。其服务目标是部分企业高管、市场人员、运营人员等。

**3.2 数仓的现在**

数据仓库在现在的应用已经远非以前的应用范围，其区别主要体现在一下几点：

1.数据源多样化

   以前的数据源可能多数是来自企业的交易数据，但是现在则会由以下来源组成，比如：交易数据、行为数据、竟对数据、CRM、 财务数据等

2.数据量暴涨

   由于以前数据来源单一，数据量相比较现在增长不会那么快，多数的中小型企业使用一个高配的小机+关系型数据库即可满足数仓的计算资源需求。但是现在由于数据源的多样化，导致数据量暴涨，单机的计算已经满足不了现有资源，但是MPP架构的数据库又太贵(比如TD)，而且也无法满足企业处理行为日志数据的需求。这个时候大数据技术应运而生，以hadoop+hive的技术处理开始占据主流。尤其是互联网行业的迅速发展，对其有一定的带动性。

3.服务对象的不同

   个人说的以前偏向于传统行业，他们的服务对象偏向于高管、运营、财务等职能人员。而现在则会有各种各样的服务对象，比如：高管、运营、财务等职能人员关注的报表必不可少，除此之外还有产品、运营等人关注的竟对分析以及内部风控，反欺诈等线上系统以及推荐等系统模块的接口调用

**3.3 数仓的未来**

目前在国外已经有人提出了Data Lake和Data Vault的概念。让我们拭目以待是否可以在国内普及应用~

**总结**

之所以出现hadoop、spark等大数据处理技术，数据多样化以及对数据时效性越来越高占主要因素，成本也是其中的一个因素，但是个人认为占比很小，因为区别仅仅在于到底是把钱花在了机器和系统上还是花在了人力上~

下面简单的对传统行业以及互联网行业建设数据仓库的时候用到的一些工具做些总结

1.传统行业

数据库：Oracle、DB2、TD(MPP结构，列式存储)、GP(MPP结构，列式存储)、SybaseIQ(MPP结构，列式存储)、MySql Inforbright、MsSql、等

ETL工具：Informatica、DataStage、Kettle、Automation(调度平台)、SSIS、企业内部调度工具等

可视化工具：Biee、tableau、Congons、Power BI等

2.互联网行业

离线仓库架构：Sqoop+hadoop+hive/spark+mysql/hbase+echarts/tableau/highchars

实时架构:flume+kafka+storm/spark streaming+hbase/redis+echarts/tableau/highchars