

数据治理能力提升专项培训

企业数据仓库体系建设

目标

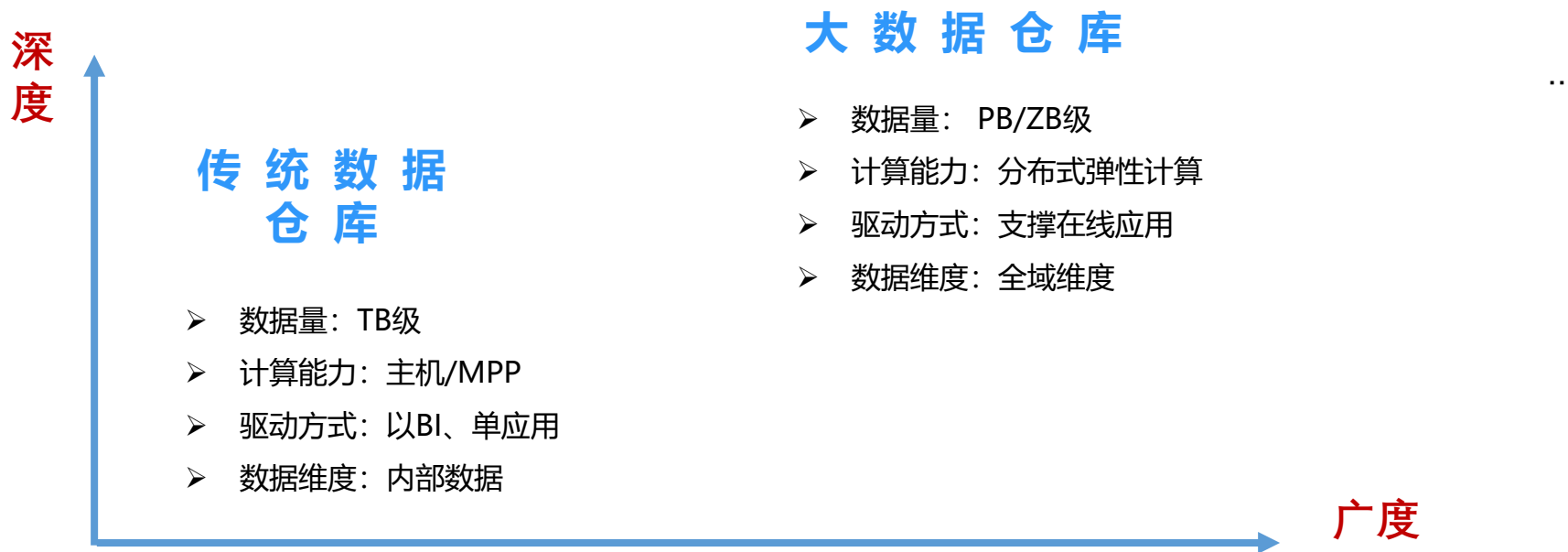
- 通过学习本课程，您将能够：
 - 了解数据仓库发展历程
 - 掌握数据仓库特点
 - 掌握数据仓库分层架构设计
 - 认识数据仓库产品解决方案

目录

1. 数据仓库概述
2. 数据仓库体系架构
3. 数据仓库建设实现

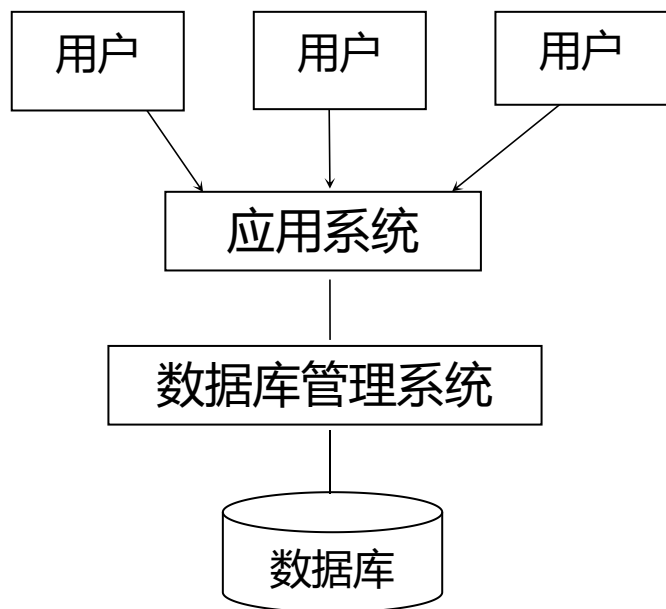
数据仓库的发展

- 数据仓库：由数据仓库之父比尔·恩门（Bill Inmon）于1990年提出，主要功能是将系统联机事务处理（OLTP）所累积的大量资料，透过数据仓库理论所特有的储存架构，做有系统的分析整理，以利用各种联机分析处理（OLAP）、数据挖掘（Data Mining），支持如决策支持系统（DSS）、商业智能（BI）的建设等。



传统操作型数据库架构

- ❖ 是数据库系统的主要应用
- ❖ 系统基本架构：



数据读写频率高

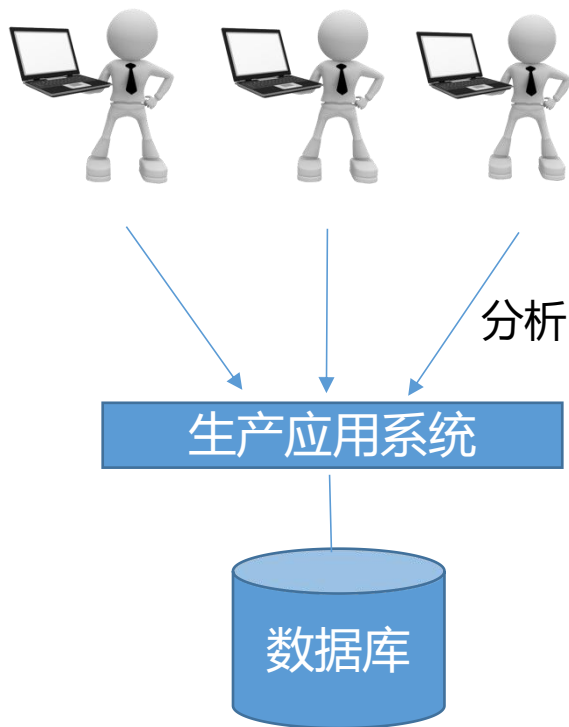
响应时间快

存储的数据量小

数据安全可靠

统计分析成为常态

OLAP和OLTP相结合



缺点

1 存储的数据量小

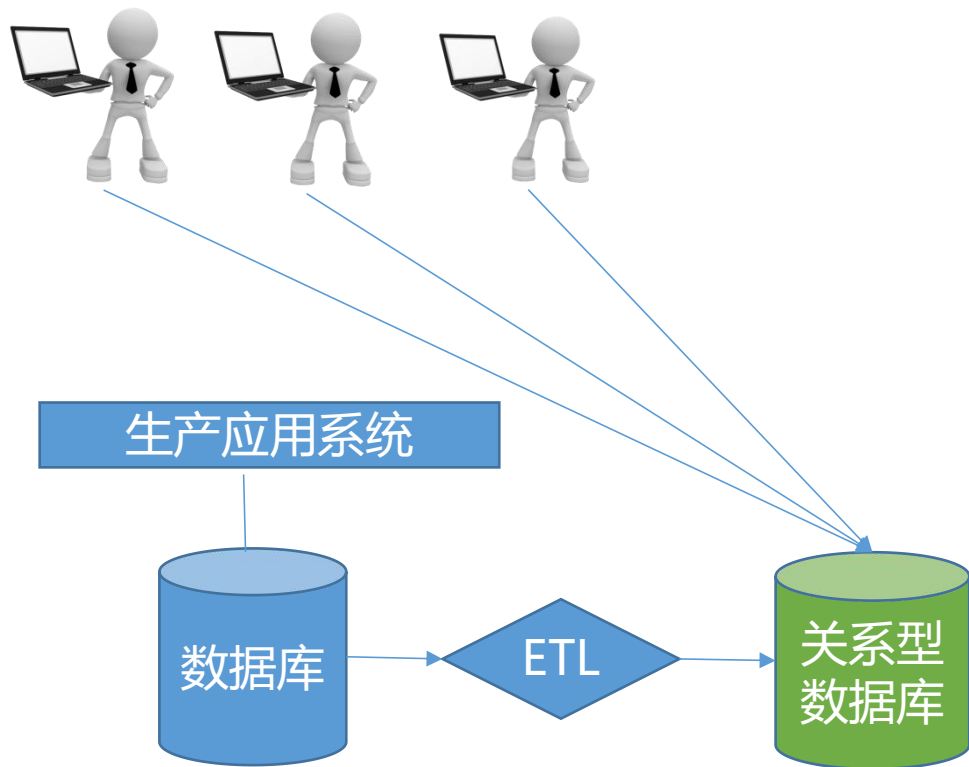
2 只能针对某一个应用数据进行分析

3 影响正常业务使用效率

4 无历史数据

OLAP剥离成为当时主流趋势

OLAP和OLTP剥离



优点

1

提高了业务系统性能

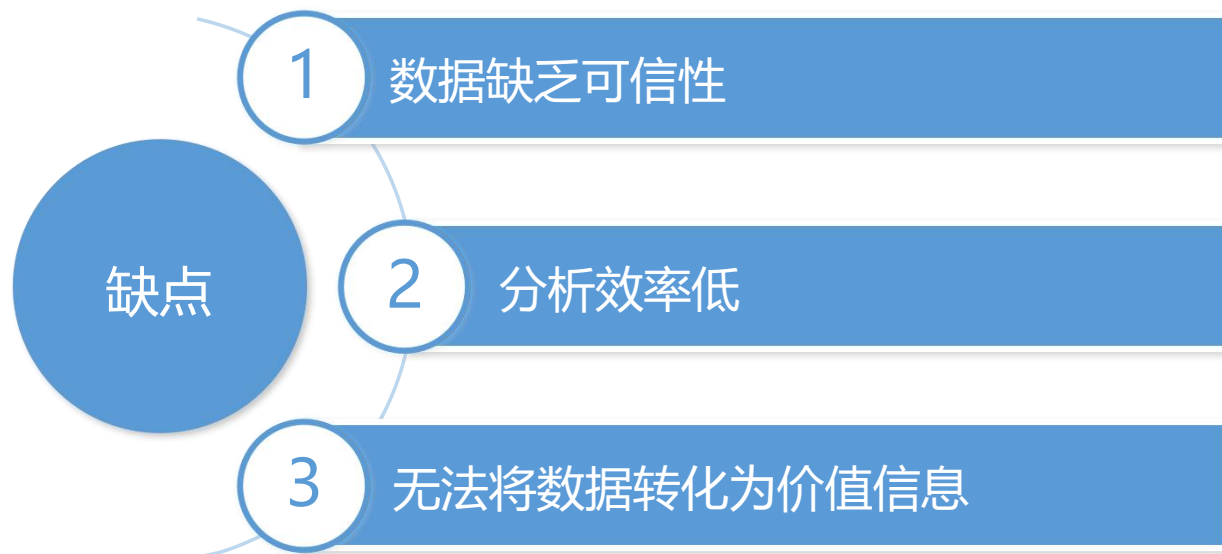
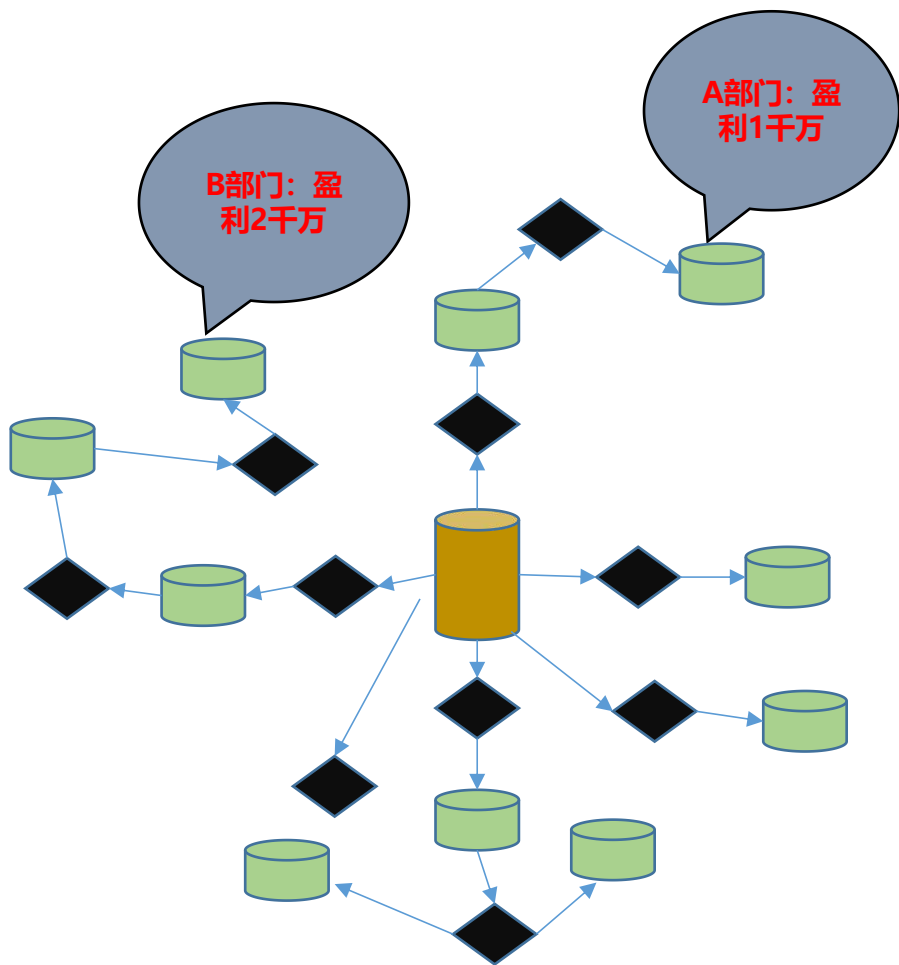
2

“拥有”了抽取出来的数据

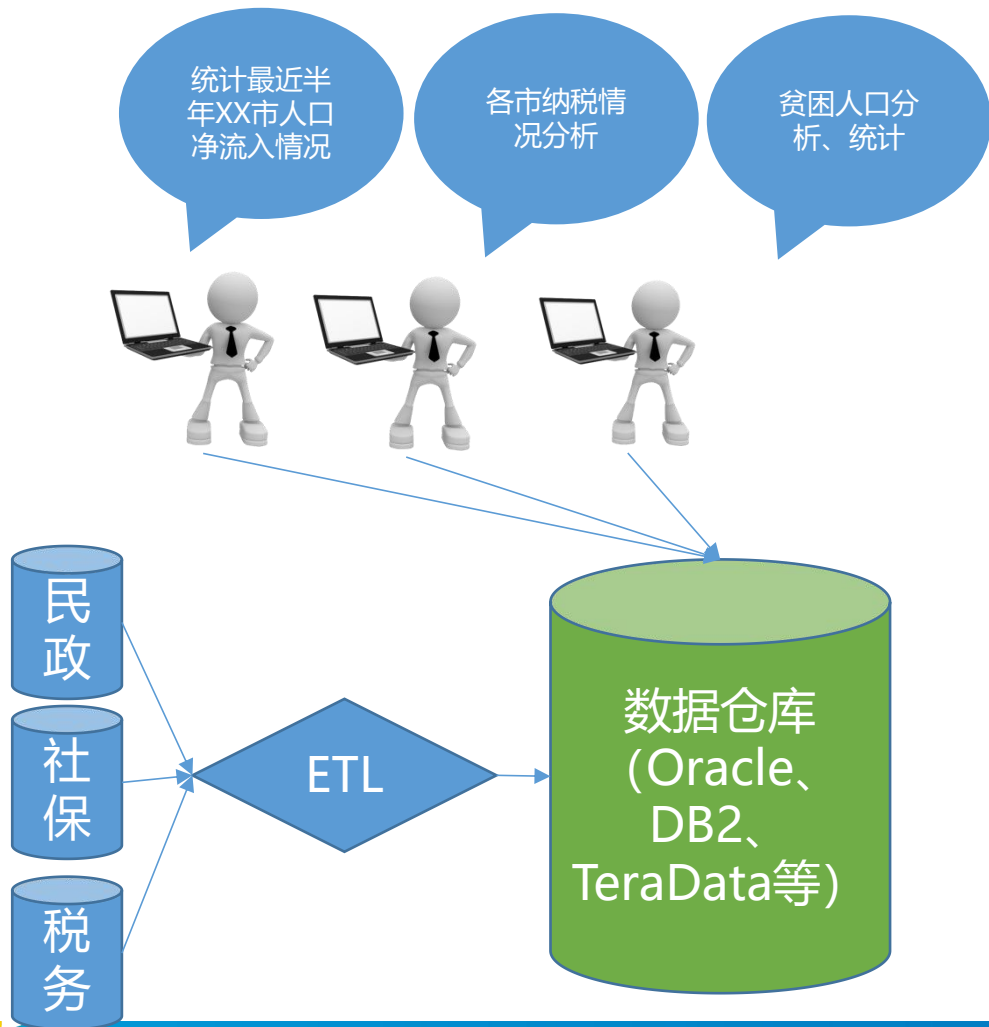
3

可以针对分析应用进行特殊配置

蜘蛛网架构问题凸显



基于关系型数据库构建的数据仓库



优点



数据被集成



包含历史数据



标准化

缺点

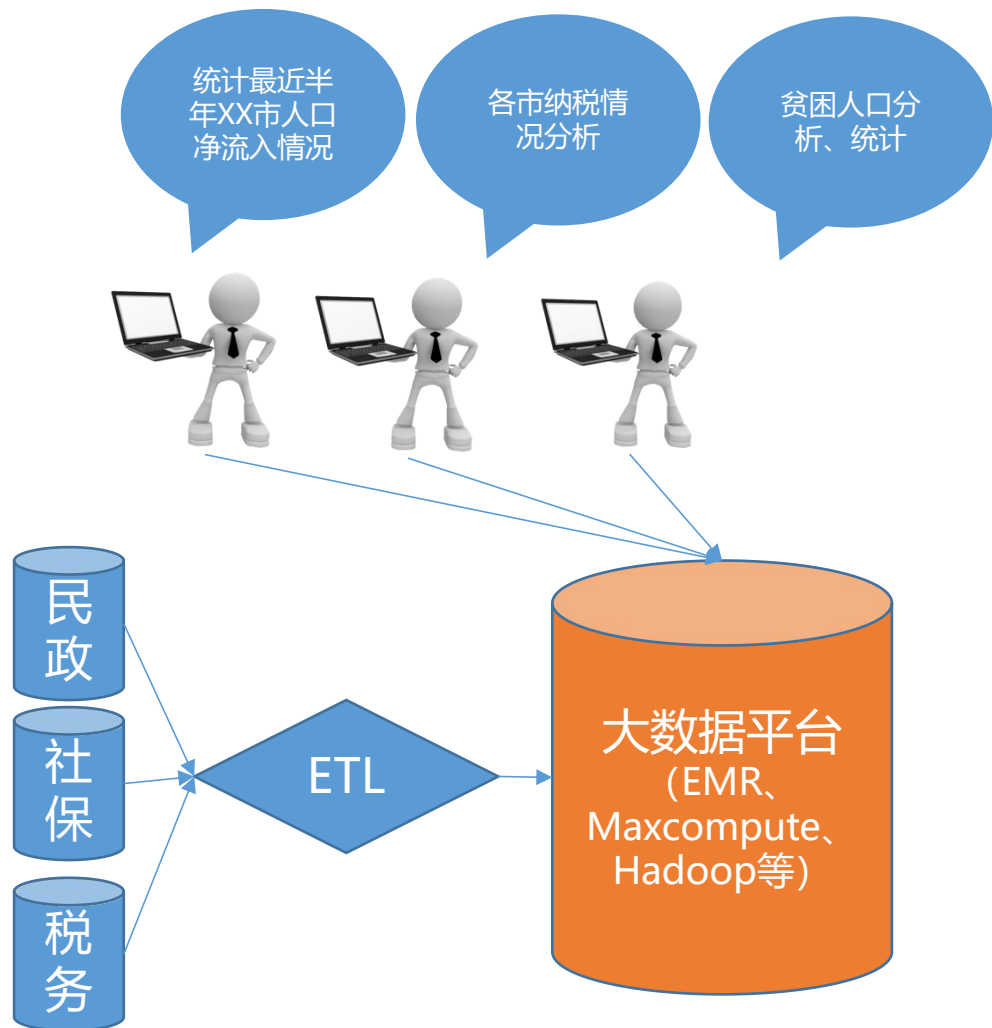


大数据量存不下



大数据量算不动

大数据数据仓库



优点



存得下



算得动



安全保障



机器学习

数据仓库的特点

- 数据仓库是一个面向主题的、集成的、时变的、非易失的数据集合，用于支持管理决策。



- 1.在较高层次对数据综合、归类，针对某一分析领域所抽象出来的分析对象



- 1.数据来源的格式类型不同
- 2.编码、命名格式、属性单位不一致
- 3.对原有数据进行综合、计算



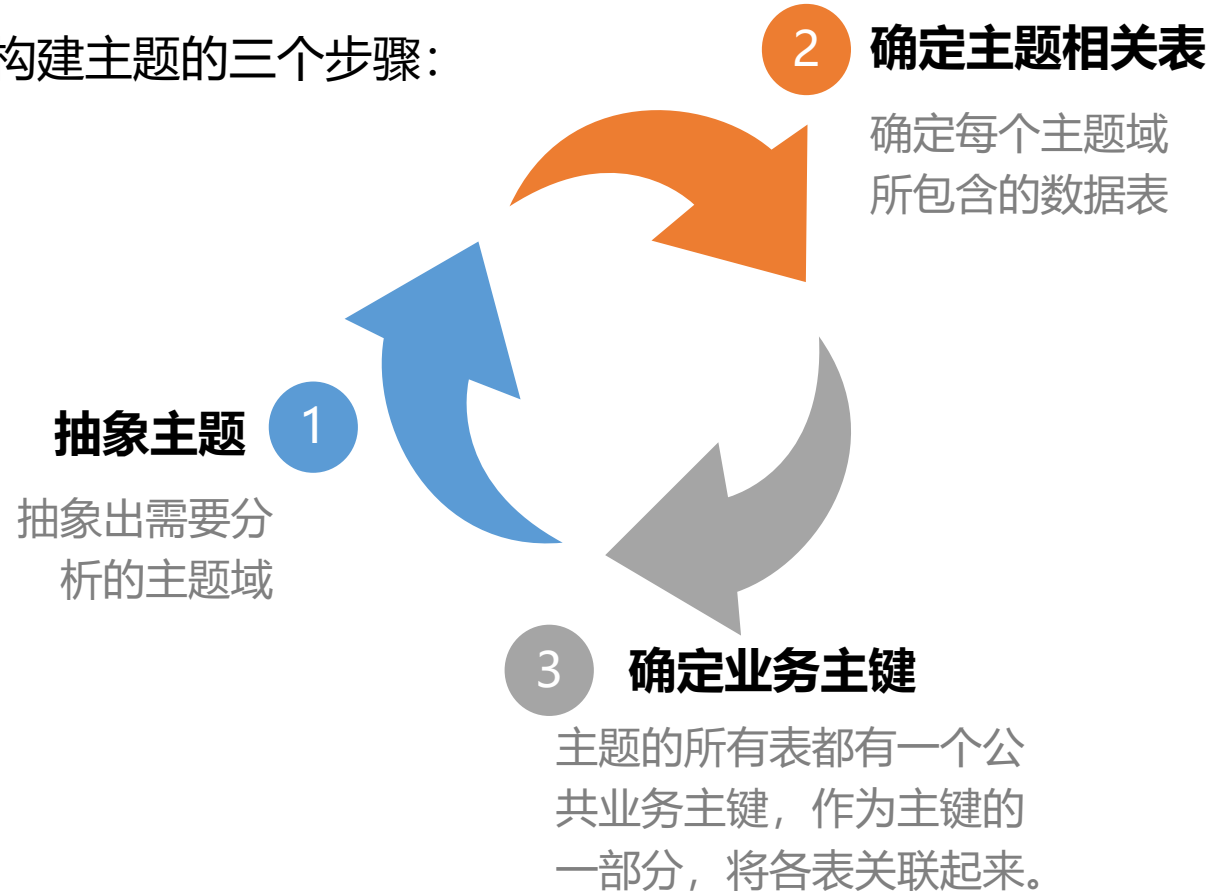
- 1.数据仓库的数据是不同时间的数据集合
- 2.随时间变化不断增加、删除(老化)、综合数据
- 3.数据仓库表结构中一般都带有时间字段



- 1.数据仓库中的数据一般只insert，不update、delete

数据仓库四大特点（面向主题）

构建主题的三个步骤：



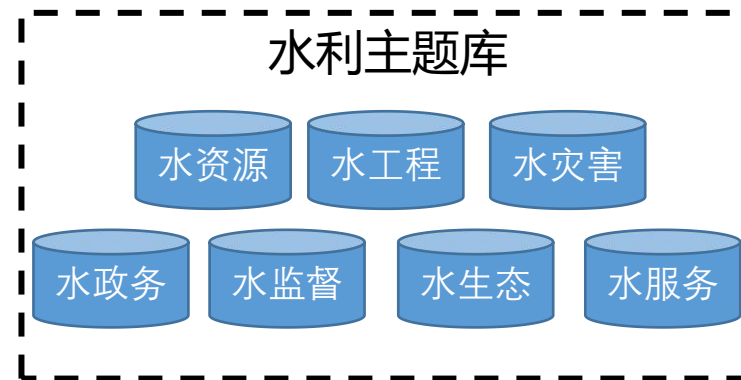
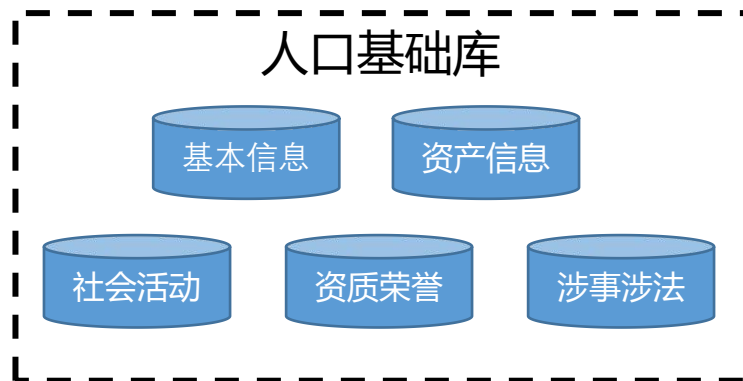
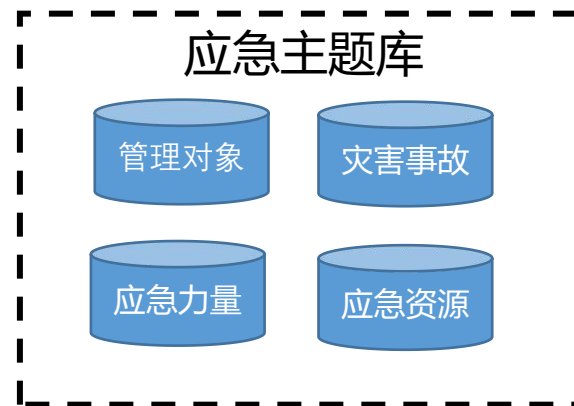
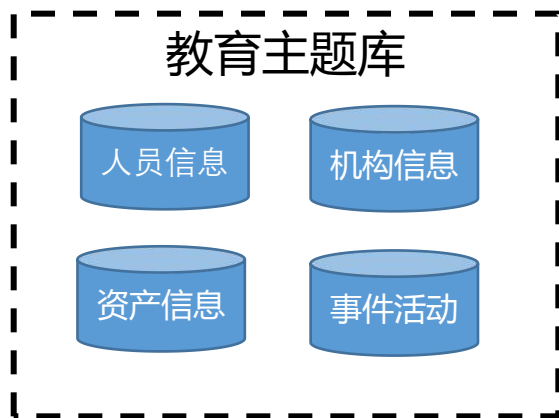
零售企业主题划分为：

- 1 商品
- 2 客户
- 3 供应商
- 4 仓库

数据仓库四大特点（面向主题）

主题名	公共码键	属性组
商品	商品号	商品固有信息：商品号，商品名，类型，颜色等 商品采购信息：商品号，供应商号，供应价，供应日期，供应量等 商品库存信息：商品号，库房号，库存量，日期等
客户	客户号	客户固有信息：客户号，客户名，性别，年龄，文化程度，住址，电话等 客户经济息：客户号，年收入，家庭总收入等

数据仓库四大特点（面向主题）



数据仓库四大特点（集成的）



数据仓库四大特点（时变）

数据仓库的数据是不同时间的数据集

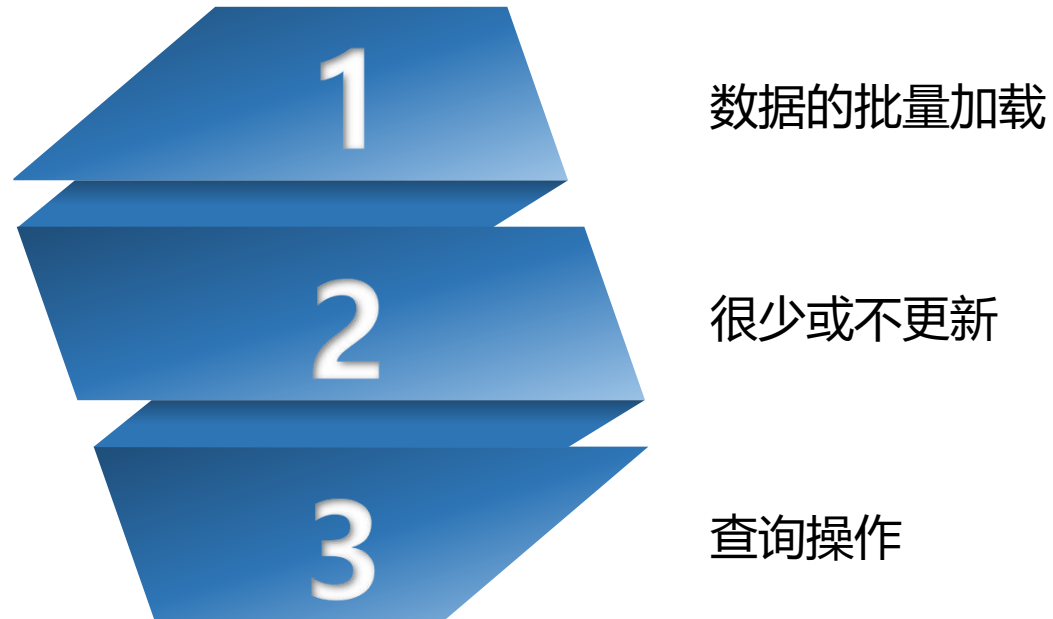
* 数据仓库的关键结构都隐式或显式地包含时间元素



VDEF6	自定义选项6	VARCHAR(1)
VDEF7	自定义选项7	INTEGER
VDEF8	自定义选项8	INTEGER
XGSHR	下个赎回日	varchar(8)
ISLOCAL	ISLOCAL	VARCHAR(1)
FXSTYLE	发行手续费率(%)	DECIMAL(12, 6)
FXTOTAL	FXTOTAL	DECIMAL(20, 4)
FXCASH	FXCASH	DECIMAL(20, 4)
SRC_DT	源系统日期	VARCHAR(8)
ETL_DT	平台转换日期	VARCHAR(8)

数据仓库四大特点（非易失）

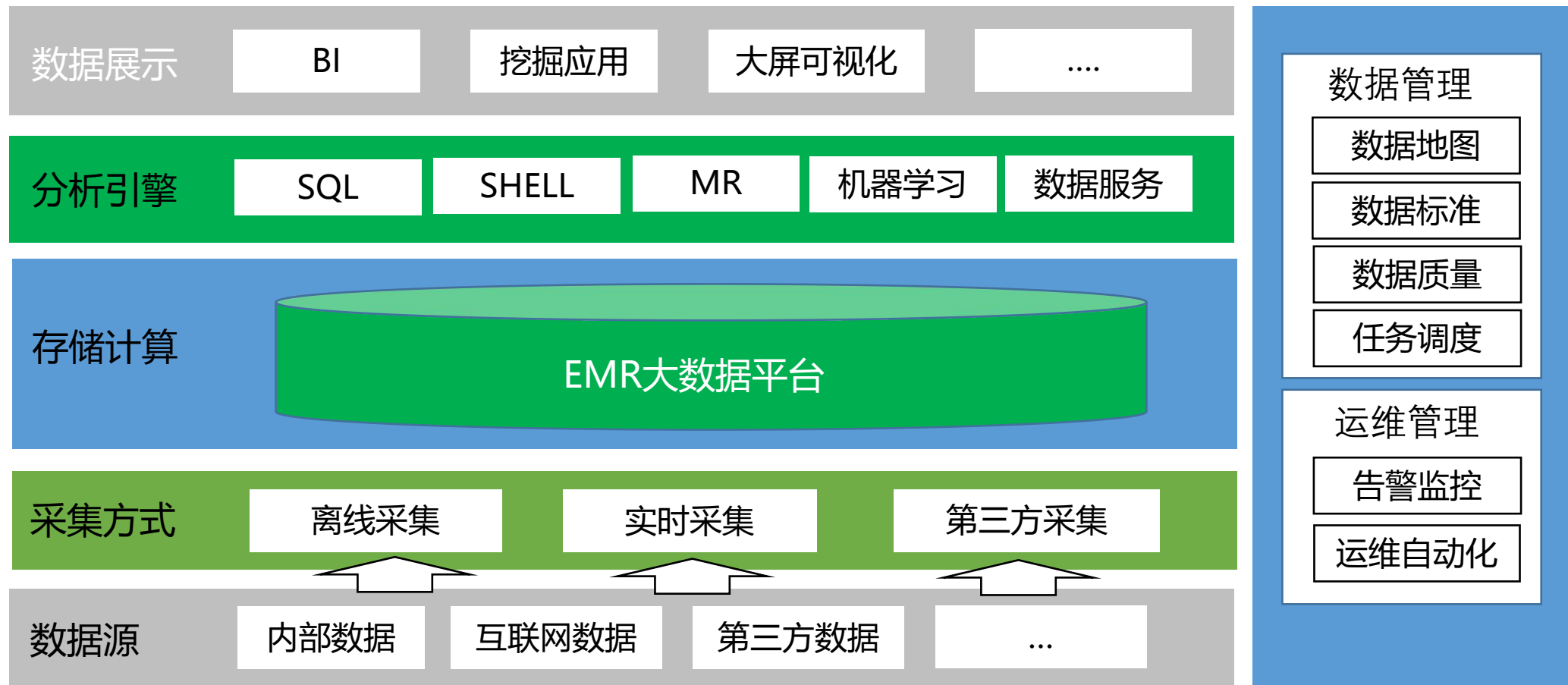
数据是经过抽取而形成的分析型数据（静态快照），不具有原始性，主要供企业决策分析之用，执行的主要是‘查询’操作，一般情况下不执行‘更新’操作。



目录

1. 数据仓库概述
2. 数据仓库体系架构
3. 数据仓库建设实现

大数据数仓架构体系图



数据仓库分层体系

目前数梦工场广泛应用于各行业的数仓分层体系有三种，如下：

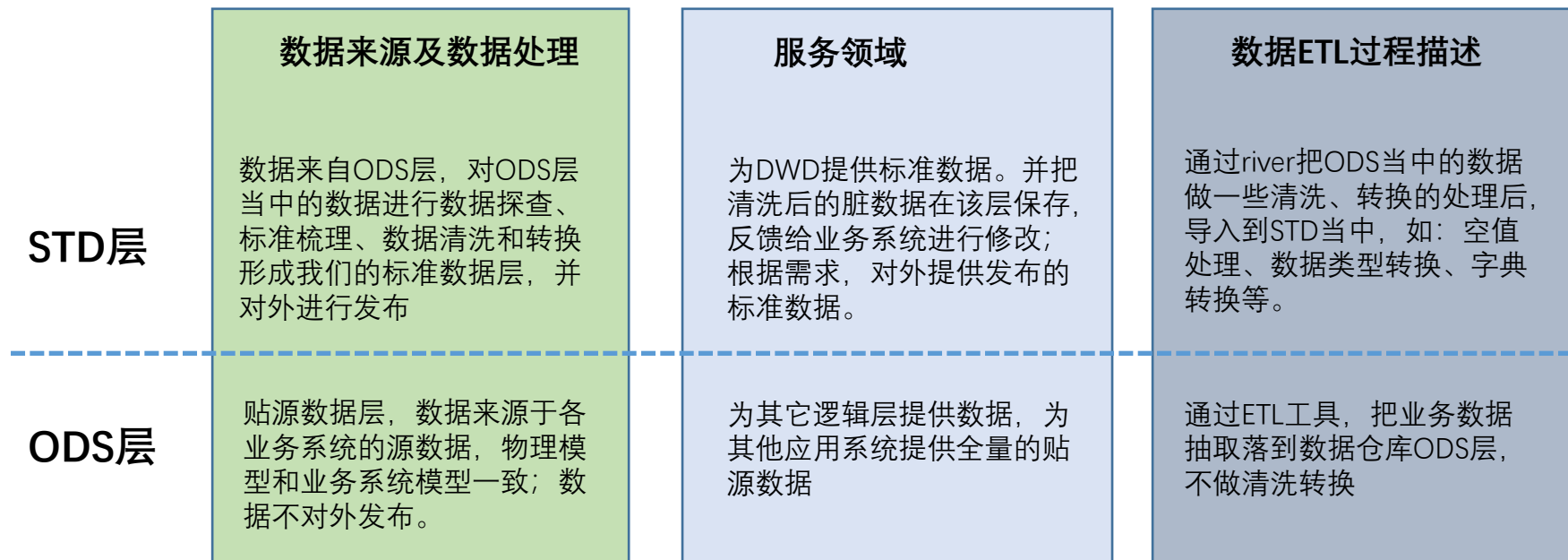


数据仓库分层体系一



- ✓ 清晰数据结构
- ✓ 屏蔽原始数据的异常
- ✓ 屏蔽业务的影响

数据仓库分层模式—架构介绍



数据仓库ODS层介绍



功能

- ❖ ODS保存所有操作数据
- ❖ 在业务系统和数据仓库之间形成一个隔离，源系统数据结构的变化不影响其他数据分层
- ❖ 转移一部份业务系统查询细节数据的功能，减轻业务系统负担
- ❖ 减轻业务系统被反复抽取的压力，由ODS统一进行抽取和分发

建表方式及处理原则

- ❖ 保留所有数据
- ❖ 根据源业务系统表的情况以增量或全量方式抽取数据
- ❖ ODS层以流水表和快照表为主，按日期对数据进行分区保存，不使用拉链表
- ❖ ODS层的数据不做清洗和转换。数据的表结构和数据粒度与原业务系统保持一致

数据仓库STD层介绍

STD标准数据层

ODS贴源数据层

功能

- ❖ 为业务系统层提供经过清洗、转换后的明细数据，满足对标准化数据需求

建表方式及处理原则

- ❖ 数据结构与ODS层一致，但可以对表结构进行裁减、合并等操作
- ❖ 对数据做清洗、转换等处理如：日期时间格式转换、字段合并、空值处理、脏数据处理、命名规范。
- ❖ STD层不会一直保存数据，具体保存周期根据业务情况而定。
- ❖ STD层保存清理出来的脏数据、有问题的数据

数据仓库分层体系二



- ✓ 清晰数据结构
- ✓ 数据血缘追踪
- ✓ 减少重复开发
- ✓ 把复杂问题简单化
- ✓ 屏蔽原始数据的异常
- ✓ 屏蔽业务的影响

以上内容仅为本文档的试下载部分，为可阅读页数的一半内容。如要下载或阅读全文，请访问：<https://d.book118.com/766040100102010105>