数据仓库与大数据工程

Data Warehouse and Big Data Engineering

第一部分 绪论

北京交通大学计算机与信息技术学院

本部分内容提纲

- 1.1 从企业信息化到数据利用
- 1.2企业中的决策与决策支持
- 1.3 决策支持系统的演化及数据仓库
- 1.4 互联网+时代企业信息系统架构演化与大数据
- 1.5 数据仓库+大数据的决策支持平台新范式
- 1.6 数据仓库/大数据工程的定义

1. 组织与企业

- 组织—Organization
  - 从广义上说，组织是指由诸多**要素**按照一定方式相互**联系起来**的**系统**
  - an entity comprising multiple people, such as an institution or an association, that has **a collective goal** and is linked to **an external environment**.
- 企业—Enterprise, Company
  - 一般是指**以盈利为目的**，运用各种**生产要素**（土地、劳动力、资本、技术和企业家才能等），向**市场提供商品或服务**，实行自主经营、自负盈亏、独立核算的法人或其他社会经济组织。
- 企业、政府、高校、医院等都是组织

2. 企业信息化

信息化

- 日本学者梅棹忠夫：信息化是指**通讯现代化**、**计算机化**和**行为合理化**的总称。
- 林毅夫等指出：“所谓信息化，是指建立在IT产业发展与IT在社会经济各部门扩散的基础之上，**运用IT改造传统的经济、社会结构的过程**”。

企业信息化

- 指企业**以业务流程的优化和重构**为基础，在一定的深度和广度上利用计算机技术、网络技术和数据库技术，控制和集成化管理企业生产经营活动中的各种信息，实现企业内外部信息的共享和有效利用，以提高企业的经济效益和市场竞争力。

3. 数据、信息与知识

数据 → 信息 → 知识

数 + 据 = 数据

符号

凭据

记录下来当成**凭据**的符号

4. 数据的定义

- 数据
  - 广义
    - 数据是针对社会生产生活的**记录结果**，是对**客观事物的符号表示**。
  - 狭义
    - 在计算机科学中，数据是指所有输入到计算机中并被计算机程序处理的符号的总称。
- 两点常识
  - 对于企业或社会而言，**没有计算机系统，不等于没有数据，不等于没有信息系统。**
  - 有了现代的信息系统，有了信息化，**不等于企业或社会的所有数据都进入了信息系统**。

Definitions

- **Data** is **measured**, **collected** and **reported**, and **analyzed**, whereupon it can be **visualized** using graphs, images or other analysis tools.
- Data as a general concept refers to the **fact** that some existing information or knowledge is **represented or coded** in some form suitable for **better usage or processing**.
- **Raw data** ("unprocessed data") is a collection of numbers or characters before it has been "cleaned" and corrected by researchers
- **Field data** is raw data that is collected in an uncontrolled "in situ" environment.
- **Experimental data** is data that is generated within the context of a scientific investigation by observation and recording.

数据产生的必要条件

- 具有待观测或记录的对象、事件或状态
- 具有观测和记录设备
  - 笔、纸、传感器、仪器设备、计算机、录入录面、网络、…
- 有记录的必要
  - 有许多状态、事件不存在记录的必要
- 问题
  - 没有计算机以前有数据吗?
  - 什么是信息系统?
  - 没有计算机以前有信息系统吗?
  - 计算机在信息系统中的地位是什么?

5. 信息—Information

- 信息论奠基人香农（Shannon）：信息是用来消除随机不定性的东西
- 意大利学者朗高在《信息论：新的趋势与未决问题》中认为信息是反映事物的形成、关系和差别的东西，它包含于事物的差异之中，而不在事物本身。
- 信息是物质存在的一种方式、形态或运动形态，也是事物的一种普遍属性，一般指数据、消息中所包含的意义，可以使消息中所描述事件中的不定性减少。

6. 知识—Knowledge

- 你**相信的东西**就是知识吗?
- 你**知道的东西**就是知识吗?
- 知识是对某个主题**确信**的认识，并且这些认识拥有潜在的能力为**特定目的而使用**。
- **柏拉图**给出的知识的经典定义：一条陈述能称得上是知识必须满足三个条件，它一定是**被验证过的**，**正确的**，而且**被人们相信的.**

Justified True Belief

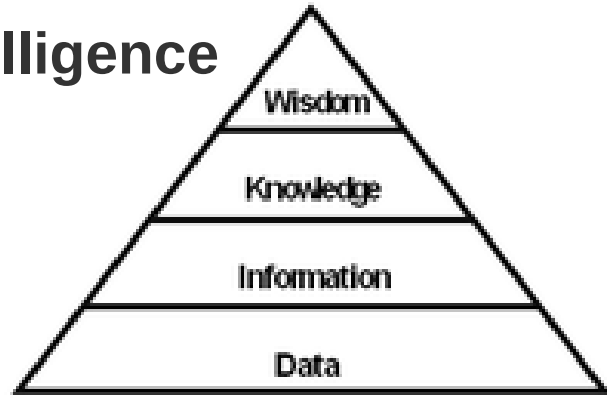
但是，有人不这么认为，相对于本体的知识

7. 知识获取

- Knowledge acquisition involves complex **cognitive processes**: perception, communication, and reasoning
- 知识获取涉及到复杂的**认知过程**
  - 感知、交互（通信、传播、交流）、推理
- 请思考如下问题
  - 小孩是如何习得知识的?
  - 人是怎么学会打球的?
  - 如何在日常工作或生活去总结得有一些有用的知识?
  - 如何去辨析真伪?

8. 数据、信息、知识、智慧或智能

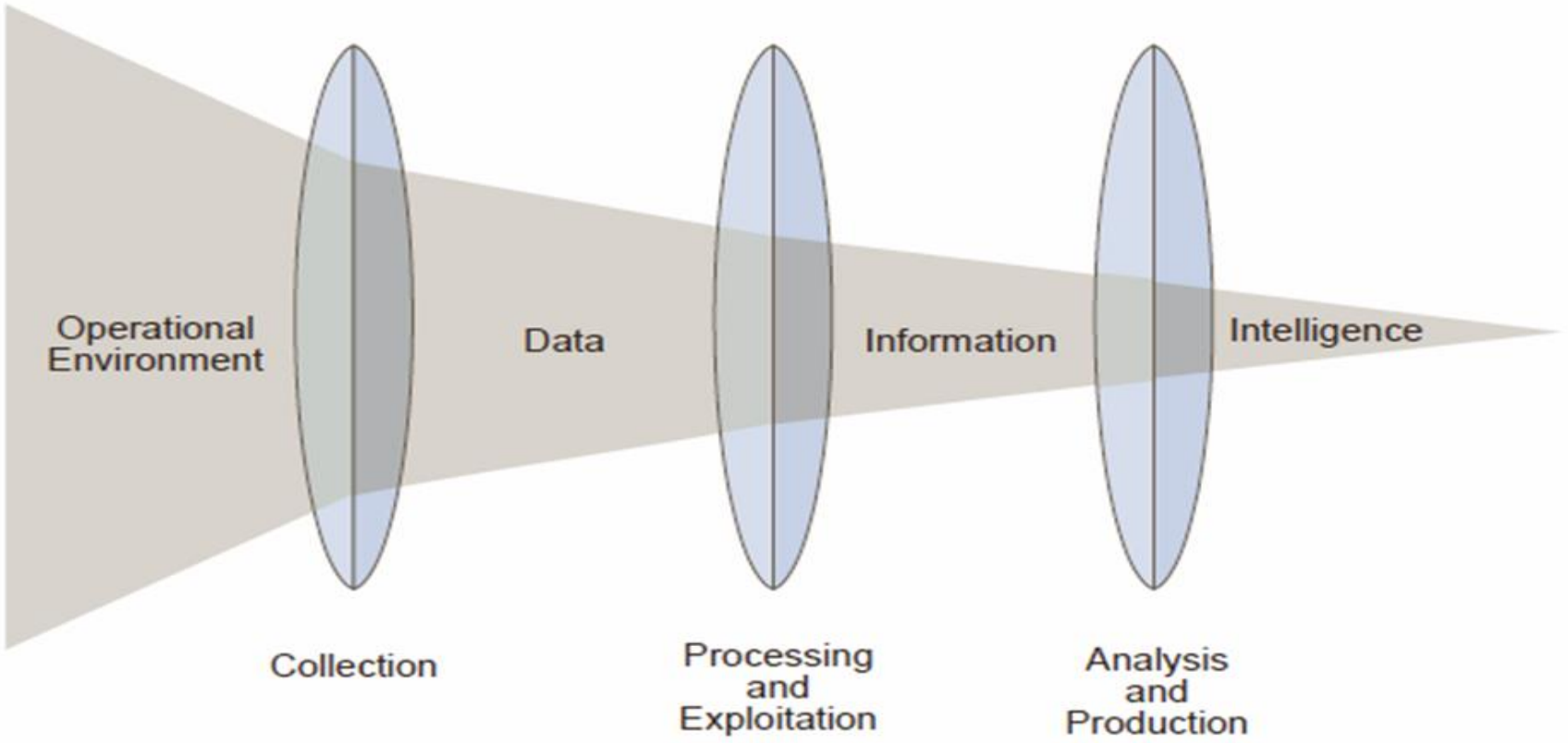
- DIKW架构
  - Data → Information → Knowledge → Wisdom
  - Data → Information → Knowledge → Intelligence
  - DIKW Pyramid or DIKW Hierarchy



- 智慧
  - 有了知识要有行动，要**有意识去利用**知识，**服务日常生活或业务中。**
- 智能
  - 有了**知识**，要能**利用**，要**有能力去行动**，通过组织或系统去利用知识，服务于日常生活或业务。

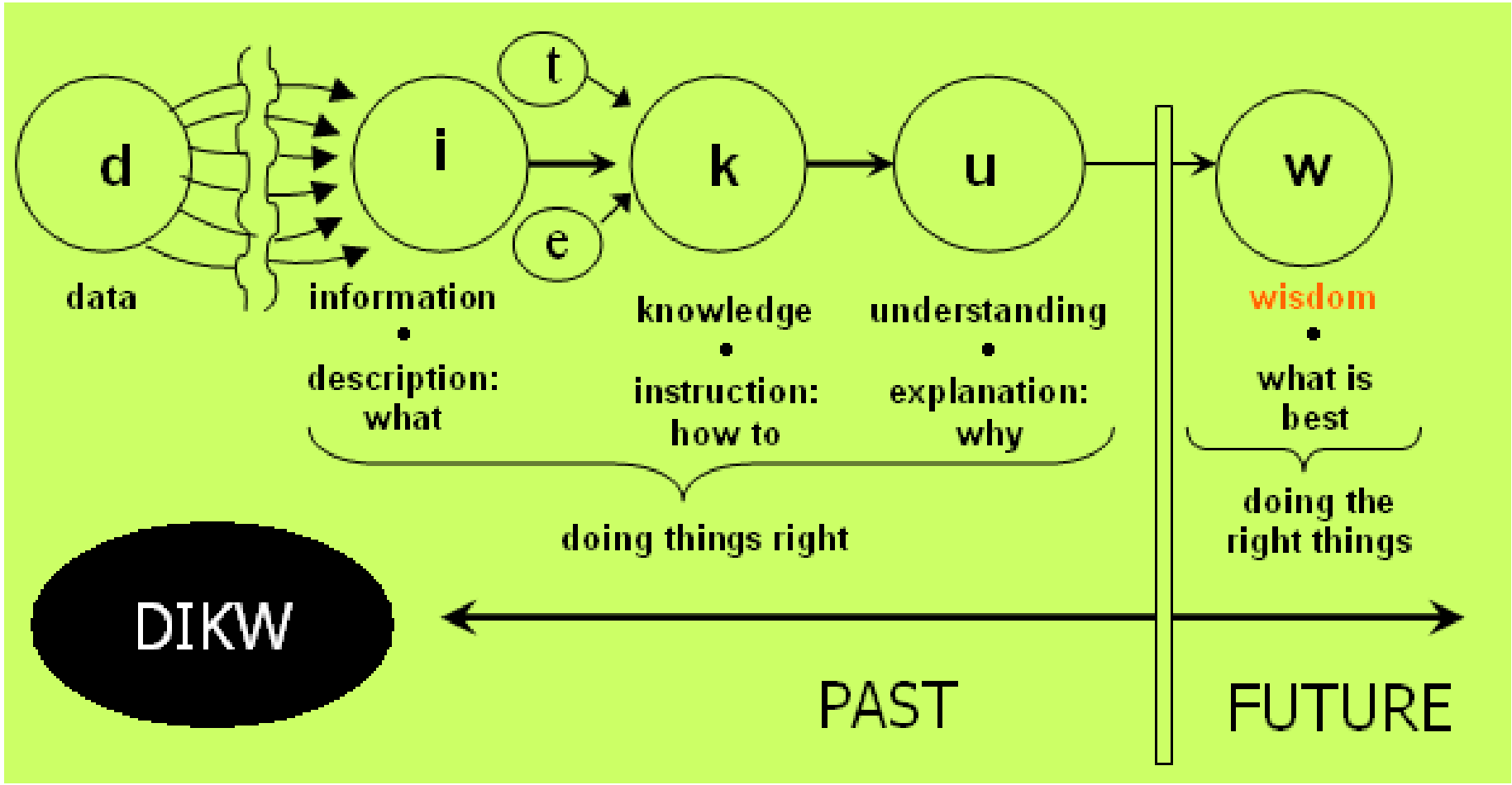
业务环境、数据、信息和智能的关系

Relationship of Data, Information and Intelligence



Source: Joint Intelligence / Joint Publication 2-0 (Joint Chiefs of Staff)

DIKW架构



以史为鉴、学以致用

本部分内容提纲

- 1.1 从企业信息化到数据利用
- 1.2 企业中的决策与决策支持
- 1.3 决策支持系统的演化及数据仓库
- 1.4 互联网+时代企业信息系统架构演化与大数据
- 1.5 数据仓库+大数据的决策支持平台新范式
- 1.6 数据仓库/大数据工程的定义

企业中的决策与决策支持

- 决策相关基础概念
- 决策支持的概念
- 决策支持系统
- 决策支持系统的一般性架构

1. 决策相关定义

- 决策—Decision making
  - 指个人、集体或自动系统为解决某个问题，借助一定的科学手段和方法，从若干备选方案中选择或综合成一个满意合理的方案，并付诸实施的过程。
  - 在心理学中，决策被作为一个解决问题的认知过程，每一个决策过程都需要从多个可能选项中确定一个最终选项。
- 决策者或决策主体 – Decision maker
  - 在特定场景中需要进行决策的人员或自动系统
- 决策问题
  - 决策者需要解决的问题

2. 决策主体分类

- 企业或组织机构中的人
  - 高级、中层、低层管理人员
  - 基础业务人员
- 日常生活中的自然人
- **自动决策程序或智能体**
  - 实时：在线推荐系统
  - 近实时：Alpha GO
  - 非实时决策：离线

3. 决策的重要性或层级

- 企业或组织中**决策层次**
  - 不同层次的业务人员所承担的工作性质与范围的不同，决定了他们各自所需承担的决策的性质和范围各不相同。
- 自然人决策问题的重要性分类
  - 重要、一般决策、不重要
- 自动决策程序或智能体
  - 目前主要服务于**细节层、游艺类**的非至关重要的一些系统业务环节
  - 个性化在线广告推送、棋类游戏

不同层次的决策举例

- 晚上去哪里吃饭?
- 现在起不起床?
- 下一步走哪里?
- 是否需要开建一条新的客运专线?
- 是否设立新的铁路局?
- 黄金周期间是否增开一些临时旅客列车，是否停开一些货运列车?
- 某天机票当前应该出什么价格?
- 如何优化运力配置?
- 是否需要进某种货?
- 需要拜访一些重要客户?
- 如何优化运输计划?
- 用户打开某个页面后给他推荐什么商品?
- 用户点击某个商品后，给他弹出什么广告?
- …

4. 决策的合理性与决策依据

- 具体决策是否合理取决于许多因素，合理的决策离不开**科学的决策方法**与**有效的决策依据**
- 决策合理性与决策依据的相关问题
  - 合理性如何评估?
  - 决策依据如何选择?
  - 由谁来做决策?
  - 决策过程如何?
  - 如何提高决策的效率?
  - 如何提高决策的合理性?

5. 企业或机构做决策的模式

- 针对企业运营过程中的某一项需要做决策的业务，根据企业和外部环境的情况，结合决策者自身的知识，作出决策。
- 决策相关因素
  - 决策者或机构
  - 企业内外部情况：资金、人员、库存、销售情况、产品质量、竞争对手、企业战略、市场行情、用户反馈、…
  - 决策者自身的经验、判断、…
  - 外部影响力
  - …

6. 数据与企业决策

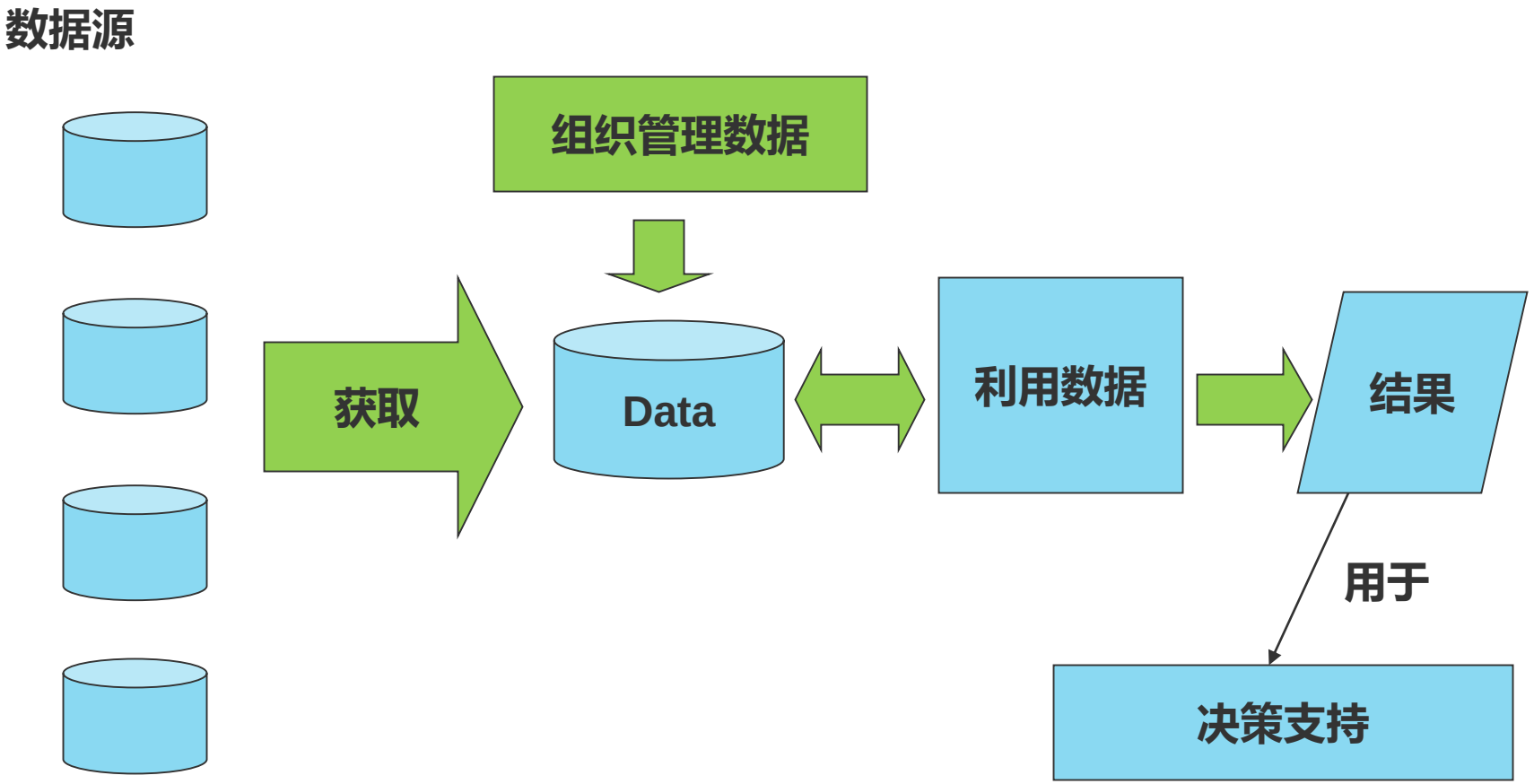
- 企业中的数据
  - 大中型企业信息化及信息系统积累了大量数据。
  - 这些数据反映了企业的业务活动的方方面面。
- 决策需要以高质量数据为依据
  - 决策不能是盲目的，必须依靠事实来说话，信息系统中的数据是企业运营事实的反映，也就成为决策的主要依据。
  - **决策支持**：为需要做决策的人提供支持的活动。
  - 如何利用企业信息系统中的数据，为决策支持提供服务，已经成为当前各企业信息利用的主要目标。

7. 决策支持系统(Decision Support System)

- DSS是企业信息系统或应用环境中的一大类重要的信息系统，这类系统**以数据为基础**，通过数据统计、分析、挖掘、展现等手段为各层次决策者**提供决策支持服务**。
- 然而，广义上看，决策支持系统或体系也不一定非得是信息系统，传统的由人与组织为决策者提供决策信息服务的**线下系统也可以看成是决策支持系统**。
- 现代企业的决策支持体系中，以信息系统形式出现的DSS起到的作用仍然是有限的，只是决策体系中的一部分。
- 问题
  - 如何实现这样的决策支持系统?

首先需要掌握数据仓库、OLAP和数据挖掘等技术
学完本课会有一些基本思路，但过程仍然不简单！

8. 决策支持系统的一般性架构



实例1-电商平台的在线推荐需求

- 电商业务
  - 顾客、产品
  - 在顾客光顾电子商城时，尽可能向顾客推荐合适的产品
- 目的
  - 产品营销—以合适价格卖出更多的商品
  - 产品规划或改进—完善产品满足需求
  - 系统改进—提升系统，吸引更多的用户
- 决策主体
  - 推荐算法
  - 产品生产部门
  - 店铺
  - 电商信息系统软件产品部门

实例2-社交媒体中的推荐

- 推荐内容
  - QQ、人人、微信中的好友或联系人推荐
  - 朋友圈或QQ群推荐
  - 微博中的热点事件或微博推荐
  - 微博中的人气或明星推荐
  - …
- 目标
  - 提高用户体验、拓展用户群、抢市场

其它实例

- 银行各种业务的客户关系管理
- 电信企业的实例
  - 业务数据如何组织
  - 用于管理决策用的数据如何组织
- 铁路货运，客运分析应用
- 保险业客户关系管理
- 民航业旅客价值分析
- 数据中心运维决策支持与生产指挥
- …

如何实现这样的系统?

本部分内容提纲

- 1.1 从企业信息化到数据利用
- 1.2 企业中的决策与决策支持
- **1.3 决策支持系统的演化及数据仓库**
- 1.4 互联网+时代企业信息系统架构演化与大数据
- 1.5 数据仓库+大数据的决策支持平台新范式
- 1.6 数据仓库/大数据工程的定义

1. 决策支持系统的发展

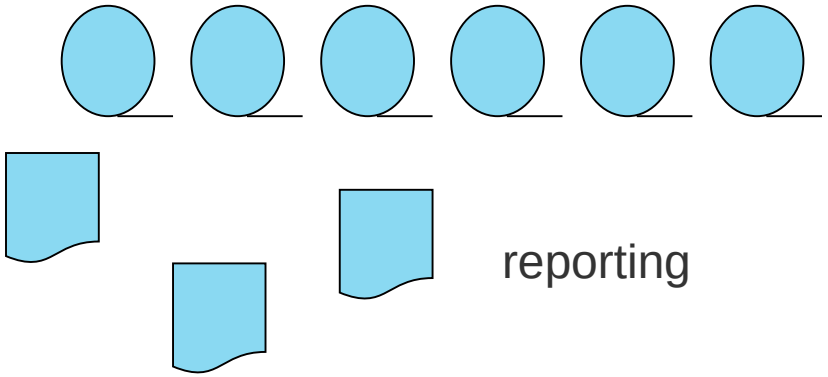
- DSS的发展历程
- 技术的发展
- 衍生出不同的系统与数据架构
- 不同架构存在的问题及阶段性的解决方案

1）Master Files—主文件

- Master files
  - 1960
  - 应用特点：报表处理和程序
  - 存储设备
    - Magnetic tape, difficult to access its data
    - Punched cards
  - 编程语言:COBOL
  - MID 1960s
    - Lots of mater files

Mater files

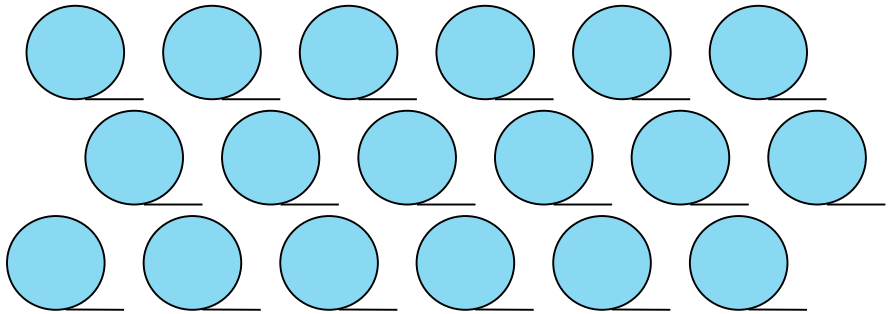
1960



Mater files

reporting

1965



许多主文件
存储在磁带上

主文件的问题-磁带

- 带来了许多的问题，成为一个巨大的障碍
  - 修改数据以后要对数据进行同步操作
  - 程序的管理很复杂
  - 开发新的程序也很复杂
  - 需要有大量的硬件来支持主文件
- 如果还在用磁带，如下各种业务系统都不会出现或难以实现
  - ATM，手机业务，铁路、航空售票，医院系统，…
- 新的存储介质DASD出现

还有人用磁带吗?

当代企业中的离线冷存储体系

许多当代企业利用磁带存储历史档案数据

磁带最大的好处：冷存储，容量大，不需要电

磁带存储缺点：顺序访问，物理可靠性不够，需要周期性维护

2）直接存取存储设备(DASD)

- Direct Access Storage Device, DASD
  - 与磁带具有本质上的不同
  - 出现时间
    - By 1970
  - DASD的出现促使数据库管理系统(DBMS)的出现
- DBMS的目的
  - 简化程序保存和访问DASD上的数据的过程

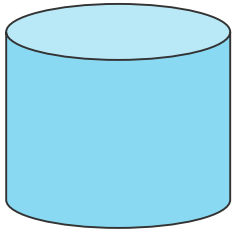
DASD的功能

- DASD所起到的功能
  - **存储**数据(Store data)
  - **索引**数据(Index data)
  - **获取**数据(Retrieve data)
  - ……
- DASA
  - 存储容量急剧攀升
  - 存储技术发展迅速
  - 存储介质、接口多样化
  - ..

3）数据库的出现

- 存储技术的发展，**数据与程序分离**的要求，理论和技术的进步，出现了数据库管理系统。
- 并使数据库成为企业中支持所有处理的**唯一的数据源。**

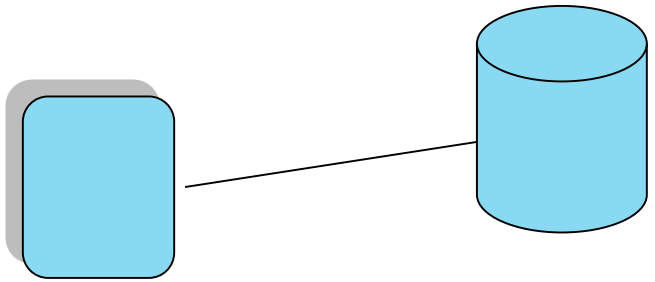
1970年代



DASD
DBMS

4）OLTP的出现

- OLTP
  - Online transaction processing
  - 出现时间
    - Mid 1970s
  - Speed of Data access and transaction process
    - 非常快
- 通过OLTP系统，计算机系统可以支持
  - 售票业务
  - 银行业务
  - 生产控制
  - …

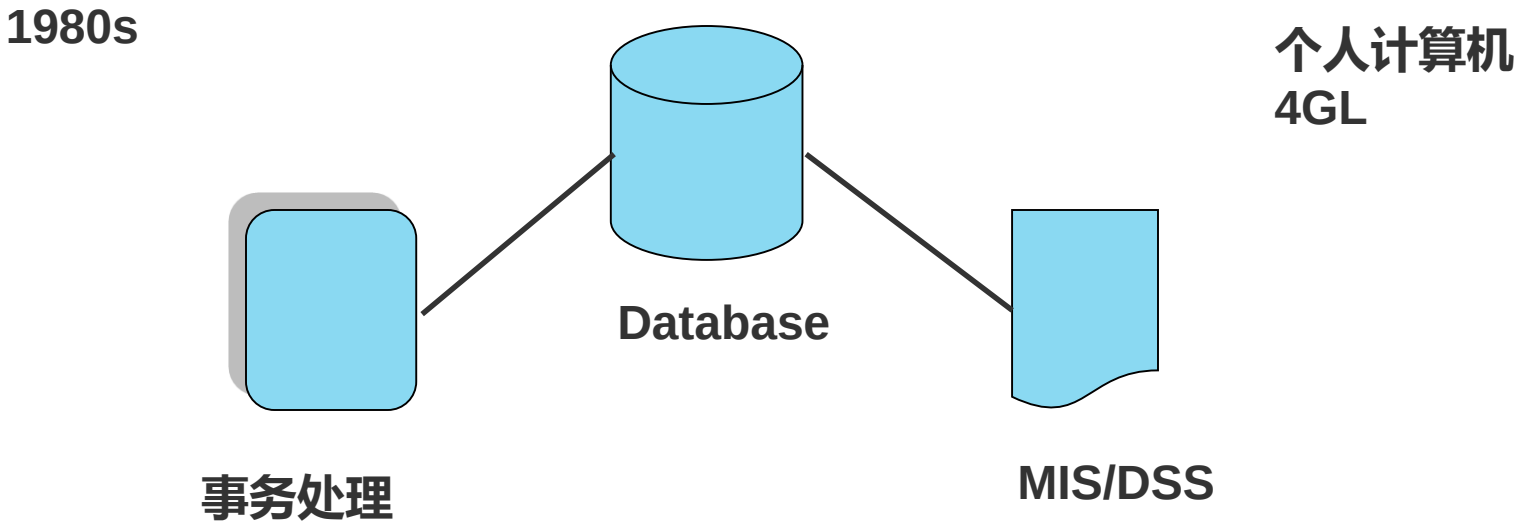


1975年左右出现高性能在线处理事务处理系统

5）PC/4GL技术的发展

- PC/4GL Technology
  - By 1980s, PC/4GL, fourth-generation languages
  - 最终用户可以直接控制数据和系统
  - 开始出现MIS(management information systems)
- MIS现在被称为DSS，早期的MIS的功能
  - 用于支持**管理决策**.
  - 由数据和技术所支持的决策都是**细节层业务层决策**。

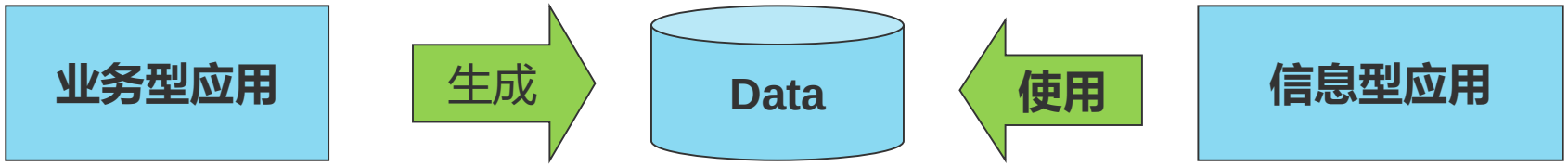
6）早期的企业信息系统架构示意图



单个数据库服务于所有目的

6）数据生成与消费关系

换个角度看问题



存在什么问题?

7）企业信息系统的类别

- **产生数据**的系统，数据源系统
  - 航空、铁路售票系统
  - 银行业务系统
  - 生产控制系统
  - …
- **利用数据**的系统
  - 基本数据处理系统，报表，统计，财务
  - 数据服务系统，利用数据，服务于其它部门或单位
  - 各类决策支持系统: CRM, BPM, …
  - …

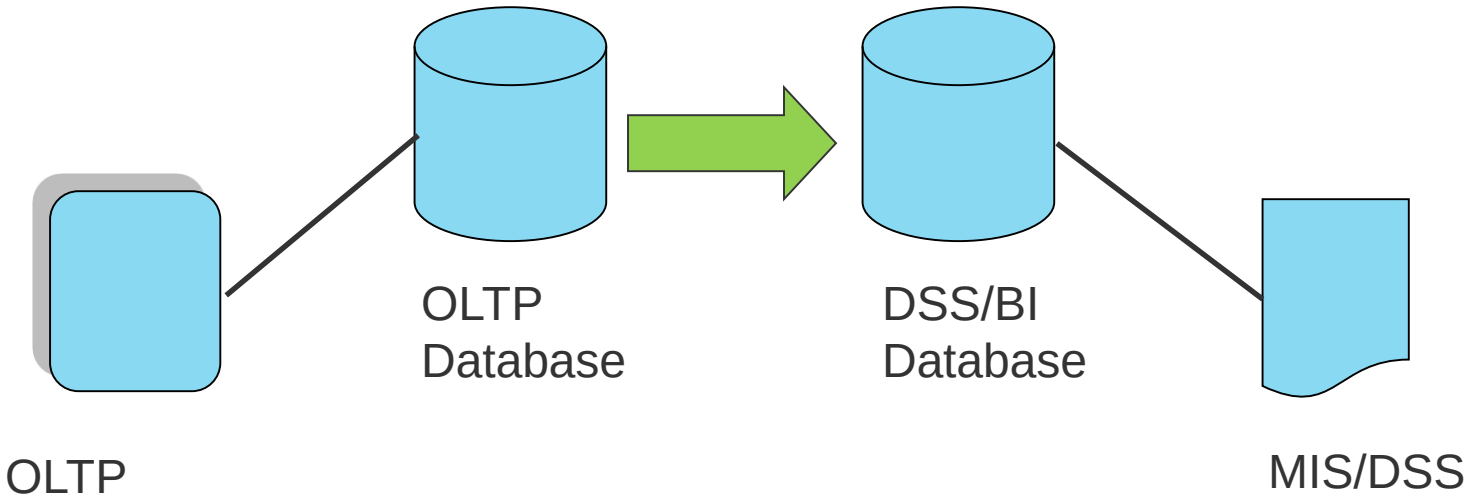
8）存在问题-产生性能冲突

- 原因
  - 一方面，OLTP系统中要求业务处理系统必须具有很高的性能，要求数据库系统的负担不能过重。
  - 另一方面，MIS或DSS系统的数据访问模式与OLTP大不相同，经常需要访问和处理大量的数据，这种不定时发生的数据处理工作对数据库系统的资源占用可能会很大。
- 解决问题的办法
  - 从OLTP系统的数据库中提取数据出来，单独构成用数据利用的系统。

2. 数据抽取程序及问题

- 抽取程序
  - 大规模OLTP系统出现后不久，就出现了进行数据抽取的程序
- 功能
  - 从文件或数据库寻找所需的数据，找到以后将找到的数据转移到其它的文件或数据库中，用于其它应用过程。

1）数据体系分离示意图



主要目的：避免性能冲突

2）抽取出现的原因

- 原因和目的
  - 避免**性能冲突**，把用于分析的数据和事务处理数据分开。
  - 终端用户拥有自己的数据，可随时进行分析利用
- 产生一个后果 (对大公司而言)
  - 过多的抽取程序和数据抽取处理

3）数据抽取及其控制问题

- 数据抽取
  - 从企业数据体系中的某个层次的数据源上获取数据，建立下一层数据存储的过程。
- 数据抽取常常会失去控制
  - 数据源多
  - 企业的部门，数据用户多，用户层次类型多
  - 数据应用类型多
  - 各种数据需求所需的数据之间存在差别，也存在交集

3. 企业组织架构和业务的发展

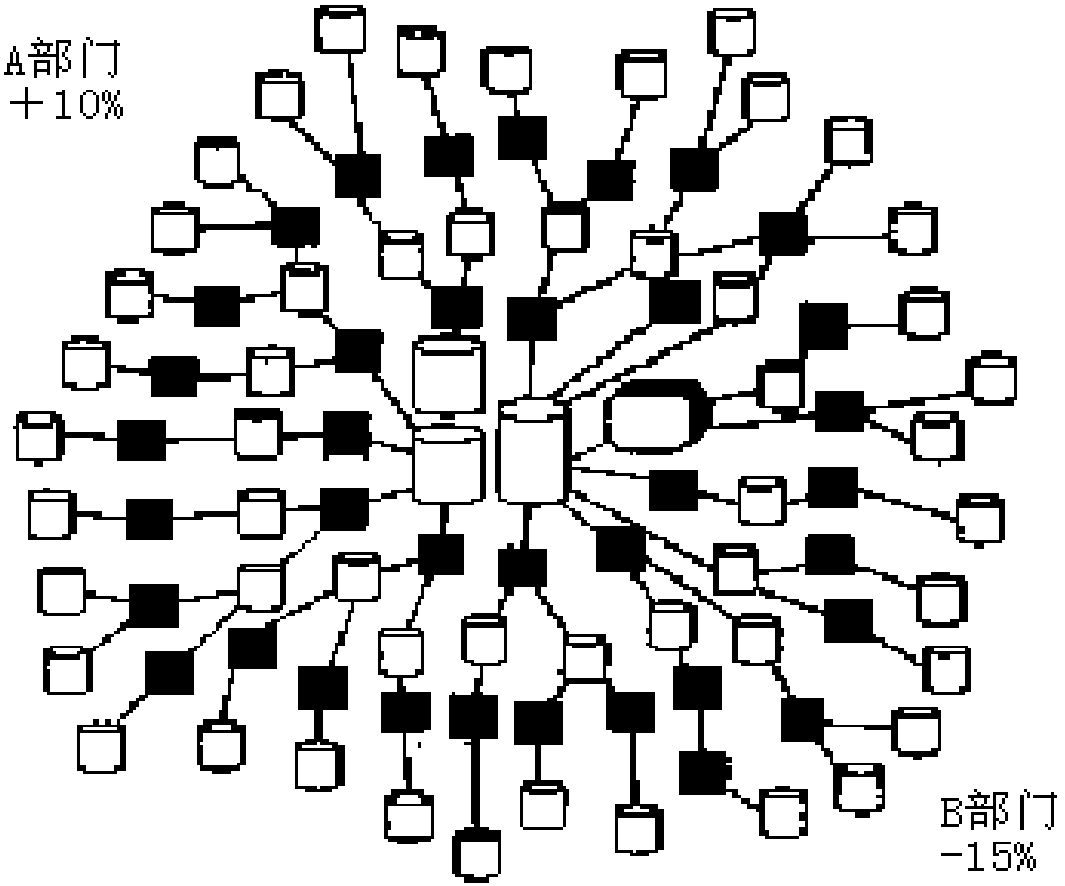
- 企业**组织机构的变化**导致信息系统的发生变化
  - 增减部门、拆分、兼并，机构职能变化
- 企业**业务变化**
  - 增加新业务，减少业务，业务流程或内容发展变化
- 企业**外部环境**的变化
  - 技术、国家政策、政治因素、管理因素
- 所有这些因素导致现在大中型企业广泛存在**信息系统零乱、结构错综复杂、数据分布广泛**的问题。

4. 企业信息体系成为蜘蛛网

- 原因
  - 组织架构与业务的发展，信息系统林立，业务交叉
  - 在不同的数据层上，存在大量的没有合理规划与控制的数据抽取程序
- 自然演化的架构或体系结构
  - 这种在企业范围内**失去控制的抽取过程**变得非常普遍，被称为 “naturally evolving architecture”.
  - 企业越大，越成熟，自然演化的架构中存在的问题就越多。

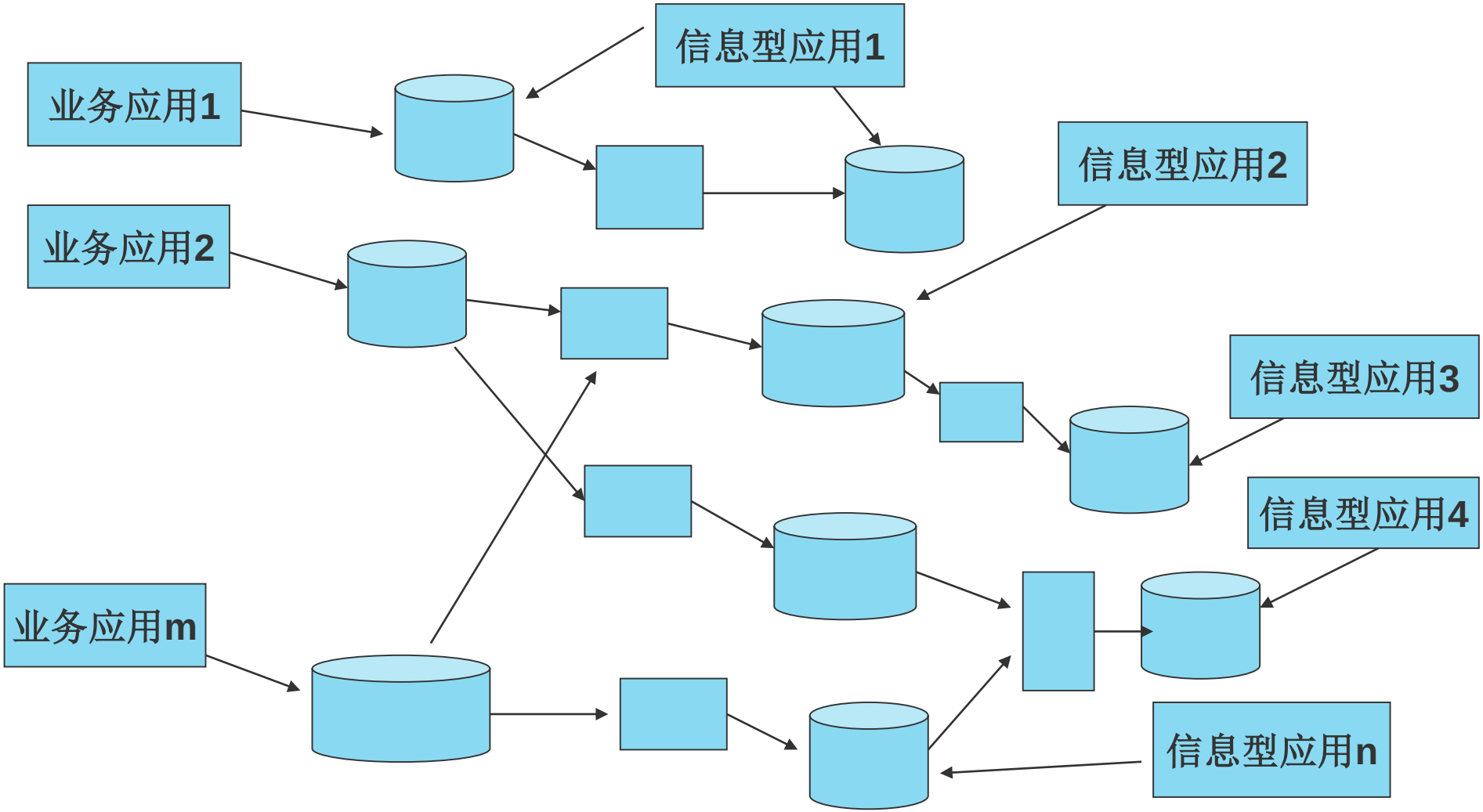
蜘蛛网示意网

- 缺共同时基
- 算法偏差
- 抽取层次不同
- 外部数据
- 数据源不同

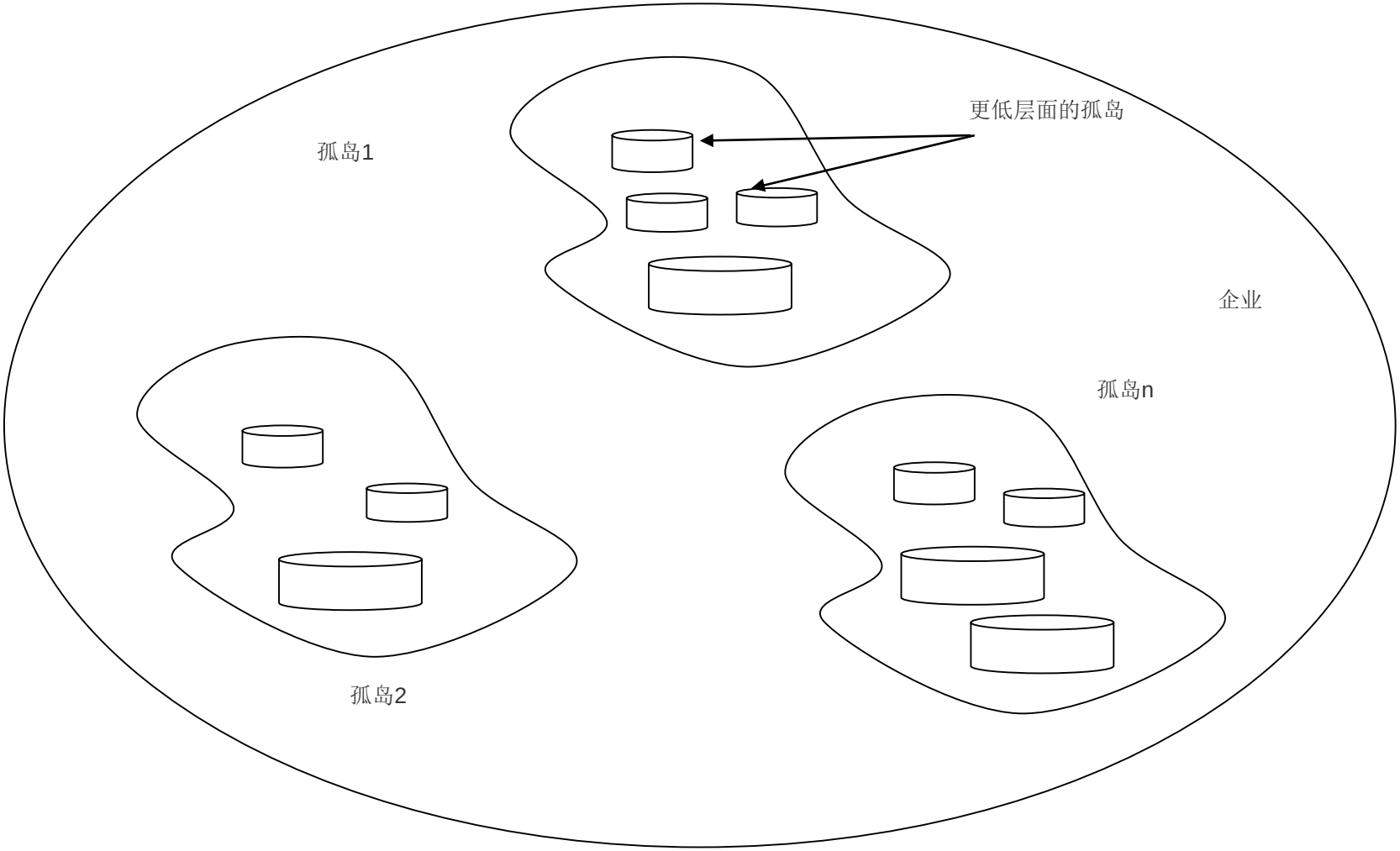


在许多企业环境中，蜘蛛网式的环境已经发展到了不可想象的复杂程度，数据的可信度低

蜘蛛网或半蜘蛛网结构

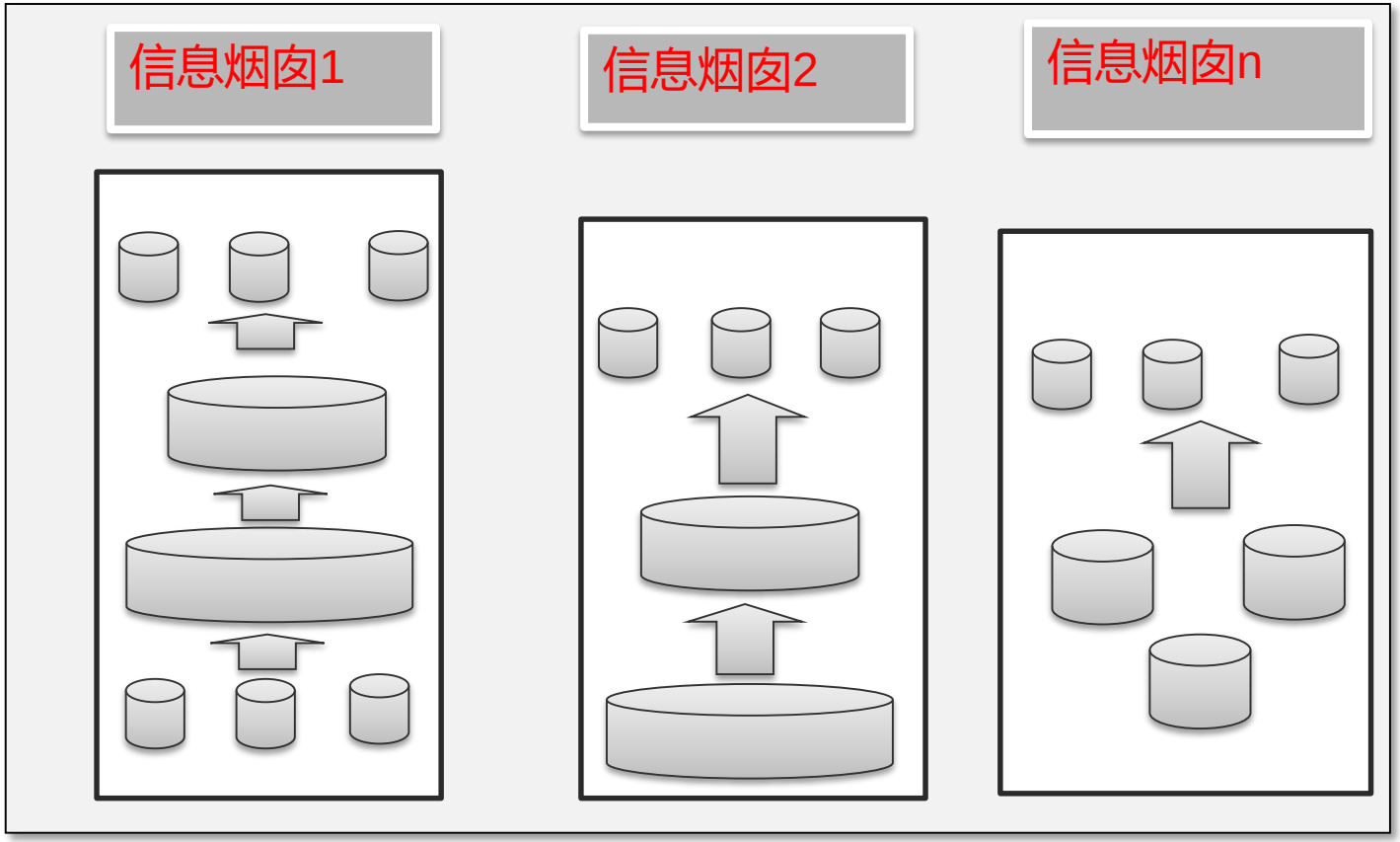


5. 企业信息架构中的信息孤岛



6. 每个信息孤岛的演化→烟囱式架构

长得高成为烟囱



7. 自然演化的架构中存在的问题

- Data Credibility
  - 数据的可信度
- Productivity
  - 生产效率，开发新信息型应用的效率
- Transform data into information
  - 将数据转化成有用的信息

1）数据的可信度低

- 部门之间数据不一致，不同步
- 例如
  - 对于某项业务指标, 一个部门说下降了15%, 另外一部门说提高了10%。
  - 除非在开展各项业务时做了很好的文档描述，对业务背景、数据来源、数据时间、数据选择条件、计算规则都进行了详细的说明。否则很难进行部门之间的调解。
- 结果：
  - 部门之间打架，管理层难以判定谁对谁错。

问题出现的原因

- 这种现象非常普通，主要有如下原因
  - 各个来源的数据时间基准不同
  - 不同数据处理算法所面向的数据不一样
  - 数据所处的抽取层次不同
  - 参考的外部数据源不完全相同
  - 最初的数据源就不同
  - …

2）应用开发效率低

- 例如
  - 设有一个具有一定历史的企业，具有大量的数据
  - 管理层希望IT部门制作一个企业级报表（corporate report），这种报表涉及多年来积累下来的许多文件和数据。
- 需要做的开发工作
  - 找到该报表需要的数据
  - 根据报表要求对数据进行编辑、处理
  - 组织资源（programmer/analyst resources）完成这些开发任务

数据定位存在的问题

- 系统存储平台相当多
  - Oracle, DB2, SQL Server, MySQL, foxbase, excel, access, …
- 名字相同的列表示不同的意思
  - Amount: 金额，数额
  - XM: 姓名，项目
- 名字不同的列表示的是相同的意思
  - Gender, Sex, xb, xingbie→性别

数据编辑处理的问题

- 需要**写很多的程序**
  - 从不同的数据源获取需要的数据
- 每个程序都要进行**定制**(customized)
  - 按这个报表的格式要求和功能要求进行数据处理
  - 每个程序都是一个小项目
- 这些程序涉及到企业业务系统中的**各种技术**
  - 访问DB2，访问Oracle，访问磁带库
  - Windows, Sco Unix, Linux, Solaris, …
  - 网络环境

新的分析任务需要进行新的数据抽取

- 面向某次特定分析需求的数据抽取不能满足下次可能的新分析任务。
- 总而言之，在蜘蛛网式的复杂构架中，对信息的访问是非常昂贵的，得到企业报表需要很多的时间，成本很高，存在重复性的成本。

3. 从数据到信息转化困难

- 假设有如下问题
  - 今年的账户活动情况与前五年各有什么不同?
- 要回答这个问题，必须到现有系统中获取必要的数据，可需要涉及许多不同的系统
  - 储蓄账号，信贷账户，信用卡，转账账户
- 如何去跟这些各不相关的系统打交道呢?
  - 这些系统和数据可能并没有集成在一起，

回忆信息的本质：差异性，关联，相比较才信息

问题

- 不同系统下所存的历史数据各不相同
  - 半年
  - 1年
  - 1年半
  - 2年
  - 5年
- 对于DSS分析人员来说，因为各个系统时基不一样，到现有系统中获取数据并不是一个可行的选择。
- 问题
  - 去哪儿去找这些数据?

数据→信息的结论

- 现状
  - 现有系统所产生的**数据缺少集成**，
  - 不同系统所保存数据的**时间跨度的不同**
  - 各个系统的可用数据时间跨度**无法满足DSS对数据时间跨度的要求**
- 以现有的、分离的、缺少数据集成的平台为基础，要将数据转化成有用的信息存在很大的问题。

面对问题，怎么办？

- 问题
  - 数据可信度低
  - 信息型应用开发成本高
  - 许多情况下难以将数据转化成有用的信息
- 解决办法
  - 从**数据应用构架的方法论**角度，我们需要作出调整。
  - 出现了合理架构的**数据仓库**

8. 企业架构中的两类数据

- 在企业的应用架构中，对应于两大类应用，存在两大类的数据
  - Primitive data (Operational Data)，原始数据，原始业务数据，操作型数据，业务型数据
  - Derived data (DSS Data)，导出数据，派生数据，决策支持数据

两类数据之间的区别

Operational data	DSS Data
Application oriented	Subject oriented
Detailed	Summarized, otherwise refined
Accurate, as of the moment access	Represents values over time, of snapshots
Serves the clerical community	Serves the managerial community
Can be updated	Is not updated
Run repetitively	Run heuristically
Requirements for processing understood a priori	Requirements for processing not understood a priori
Compatible with the SDLC	Completely different life cycle
Performance sensitive	Performance relaxed
Accessed a unit at a time	Accessed a set at a time

两类数据之间的区别(续)

Operational data	DSS Data
Transaction driven	Analysis driven
Control of update a major concern in terms of ownership	Control of update no issue
High availability	Relaxed availability
Managed in entirely	Managed by subsets
Nonredundancy	Redundancy is a fact of life
Static structure; variable contents	Flexible structure
Small amount of data used in a process	Large amount of data used in a process
Supports day-to-day operations	Supports managerial needs
High probability of access	Low, modest probability of access

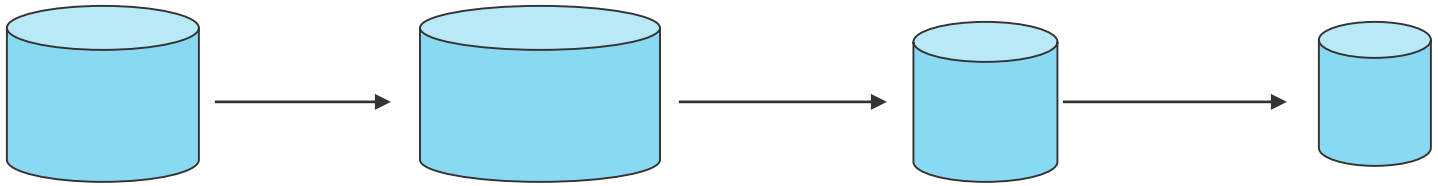
两类数据区别小结

- 原始数据是**细节层**的，用来支持企业的日常业务运作的数据。
- 原始数据**可以被修改**，导出数据可以被**重新计算但不能被直接修**改。
- 原始数据主要的数据值是**当前有效**的数据，许多导出数据反映的是数据的**历史取值**。
- 操纵原始数据的程序一般都是**简单的、不断重复执行的程序**，使用导出数据的程序一般都以**启发式的、非重复的方式运行**。
- 原始数据的用户一般是**普通业务人员**，导出数据用户一般是**管理层**。

一个问题

- 原始数据和导出数据能够同一个数据库中和平共处吗?
- 答案
  - 在有些情况下，可以!
  - 但是这些数据的类型差别如此大，对中大型企业来说，它们不能同一个数据库中，甚至不能同处于一个硬件环境中。

9. 合理架构的企业信息系统环境



操作型数据

Data Warehouse

部门层

个体层

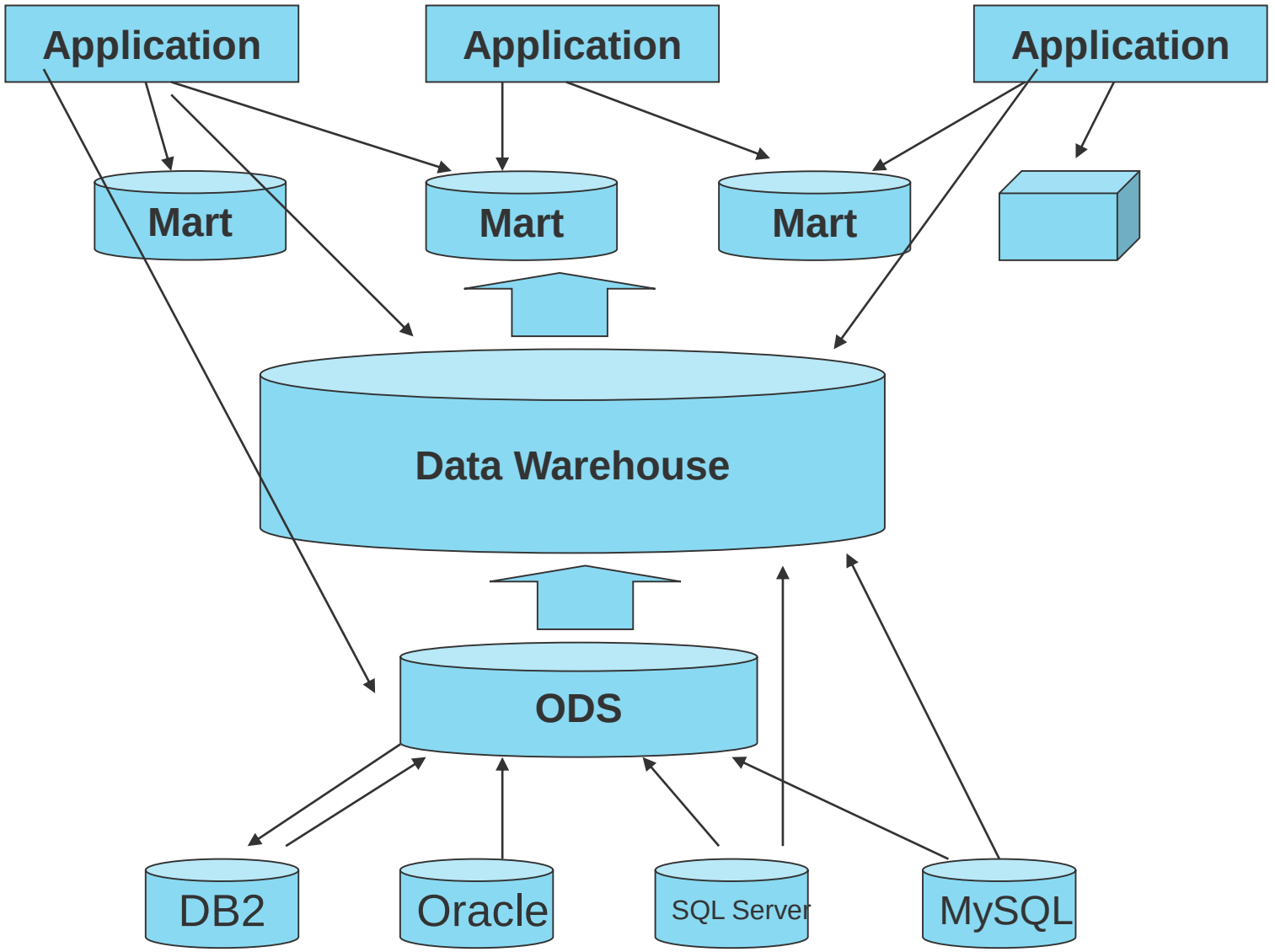
- 细节
- 每天
- 当前值
- 随时访问
- 面向应用

- 粒度化
- 包含长期时间信息
- 集成
- 面向主题
- 具有汇总型数据

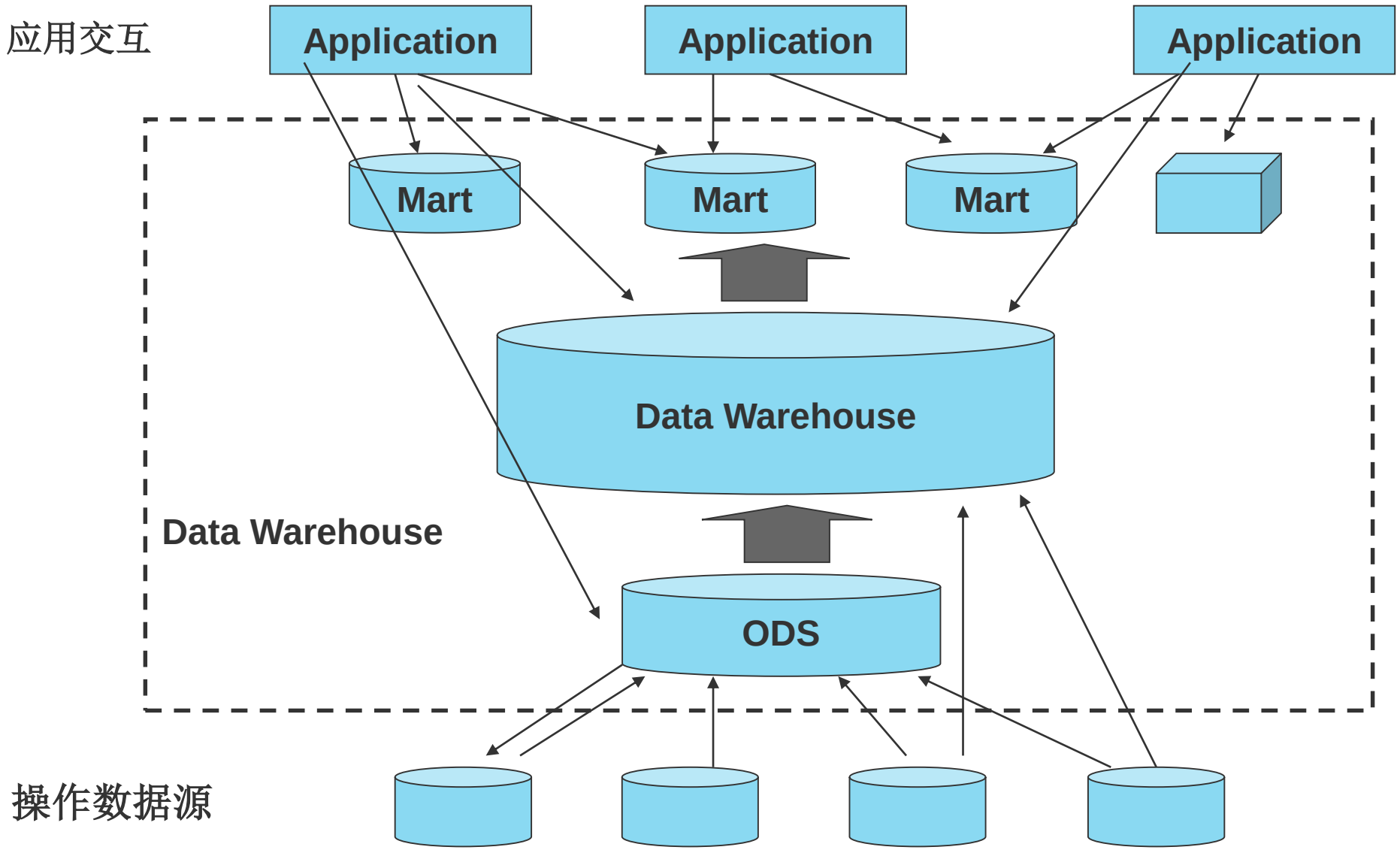
- 面向领域
- 部分导出数据
- 部分原始数据
- 典型的部门
  - 财务
  - 市场
  - 工程
  - 保险

- 临时的
- 特定目的
- 启发式
- 非重复
- 基于PC、工作站等终端

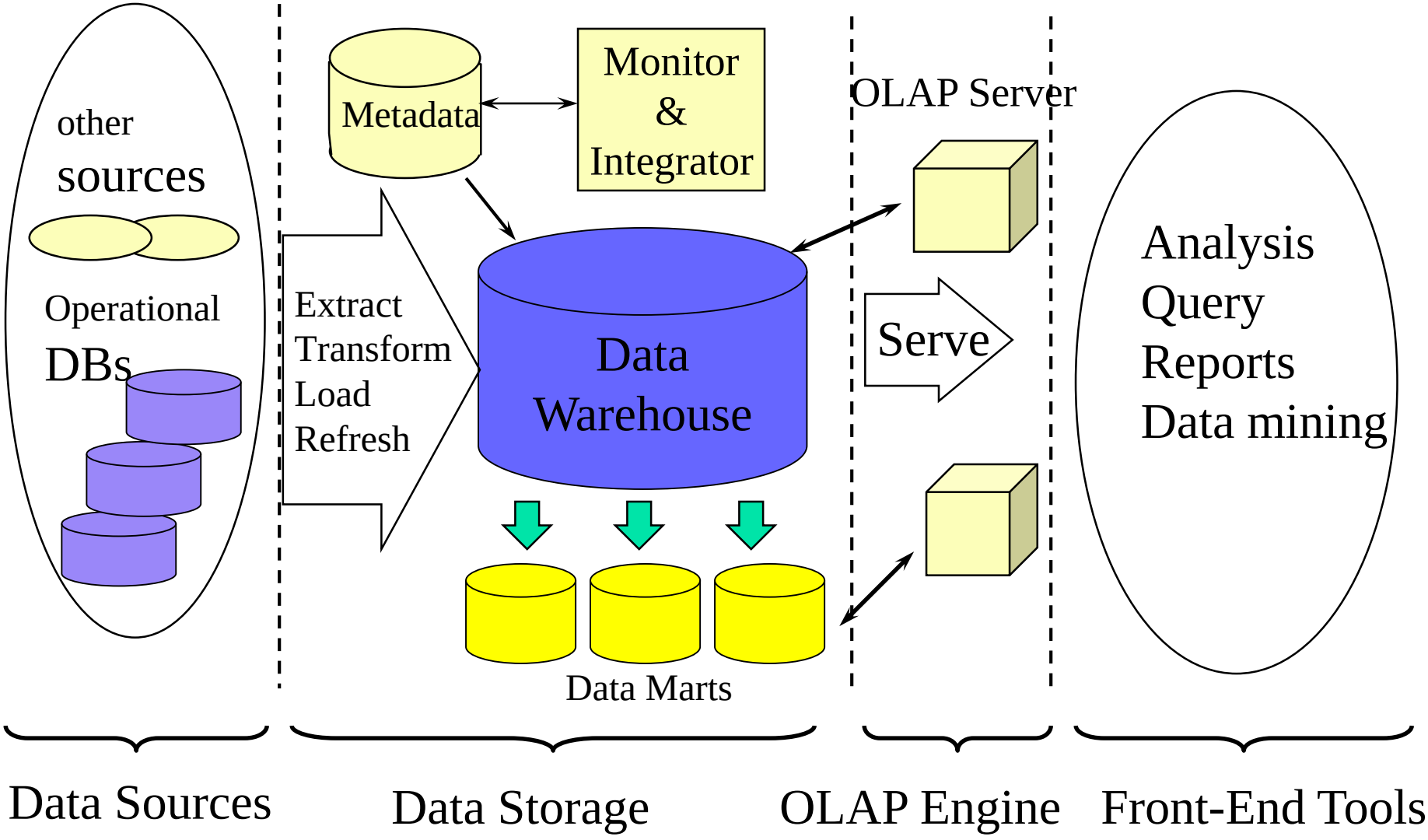
常见的数据仓库体系结构



另一种观点:内虚线框内为DW



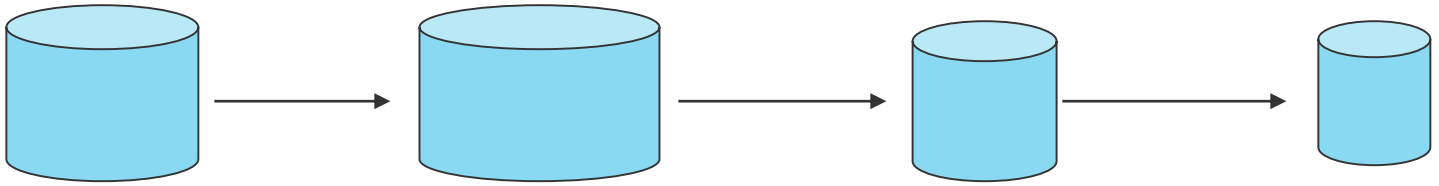
典型的数据仓库架构


other
sources
Operational
DBs
Extract
Transform
Load
Refresh
Metadata
Monitor
&
Integrator
Data
Warehouse
Data Marts
OLAP Server
Serve
Analysis
Query
Reports
Data mining
Data Sources
Data Storage
OLAP Engine
Front-End Tools

1.4.1 四层数据体系

- 操作型层(业务数据层)
  - Operational, application-oriented primitive data, high-performance transaction-processing community
- 数据仓库层
  - Data warehouse, integrated, historical primitive data, cannot be updated, and some derived data
- 数据集市层
  - Departmental, data mart, derived data, needs of the department
- 个体层
  - Individual, heuristic analysis

从用户角度的分析需求例子



Operational

J Jones
Main大街123号
信用度 - AA

Data Warehouse

J.Jones,1986-1987
High大街456号
信用度 - B

J.Jones,1987-1989
High大街456号
信用度 - A

J.Jones, 1989-今
Main大街123号
信用度 - AA

Departmental

1月 - 4101
2月 - 4209
3月 – 4175
4月 – 4215
……
……

Individual

顾客
从1982年起
账户余额>5000
信用度不低于B

临时的!

J. Jones现在的信用度是多少?

J. Jones的信用历史如何?

我们吸引的顾客是越来越多还是越来越少?

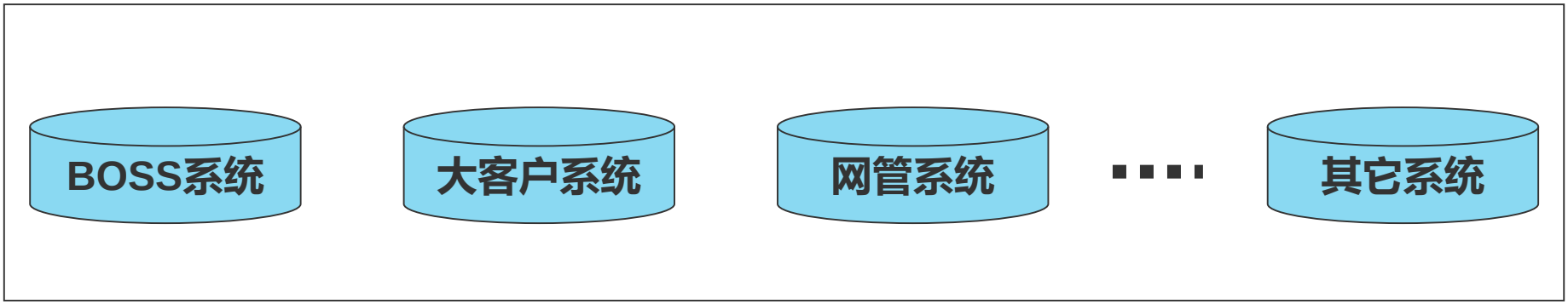
我们所分析的顾客趋势如何?

实际案例

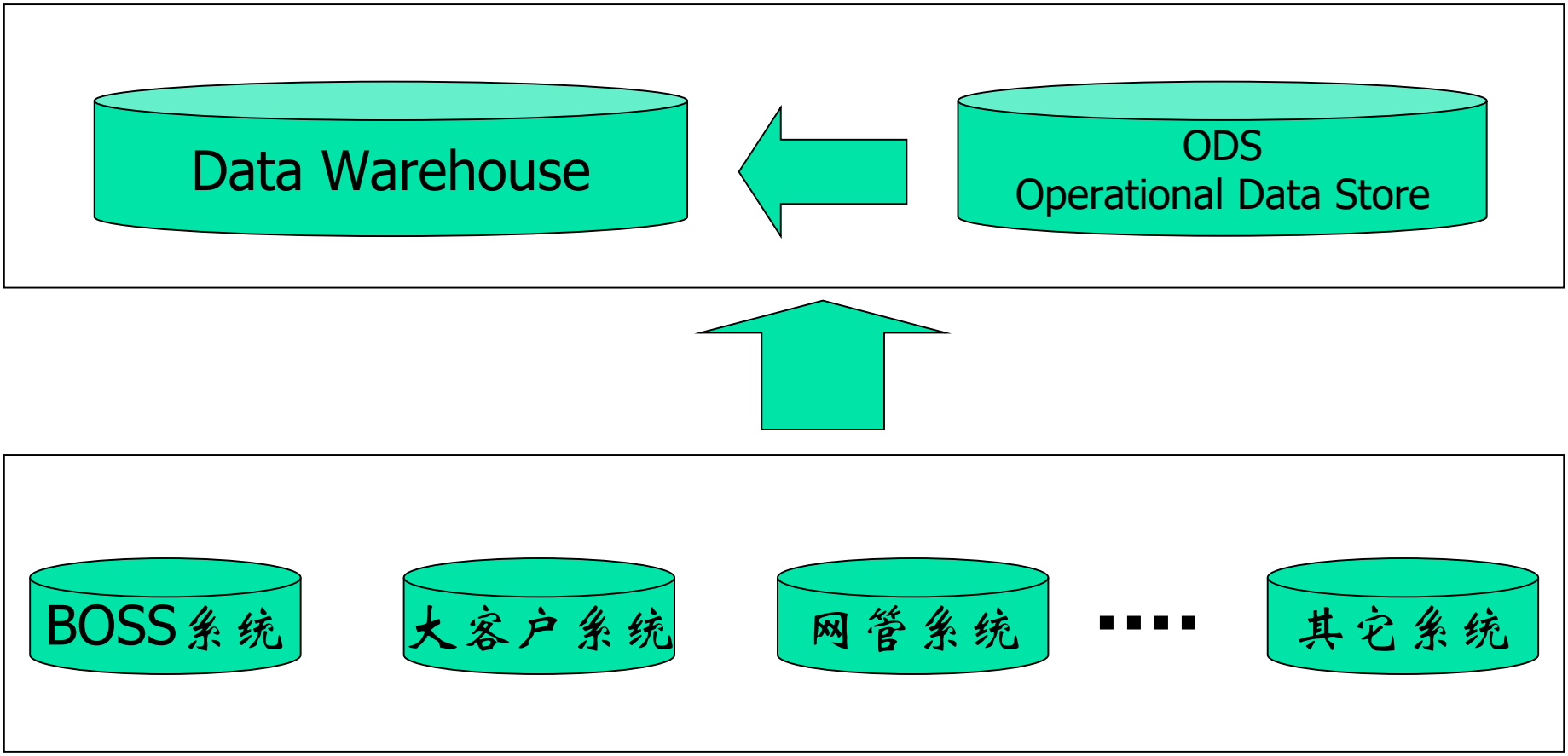
某电信公司传统的数据仓库架构

第一层 Operational Level

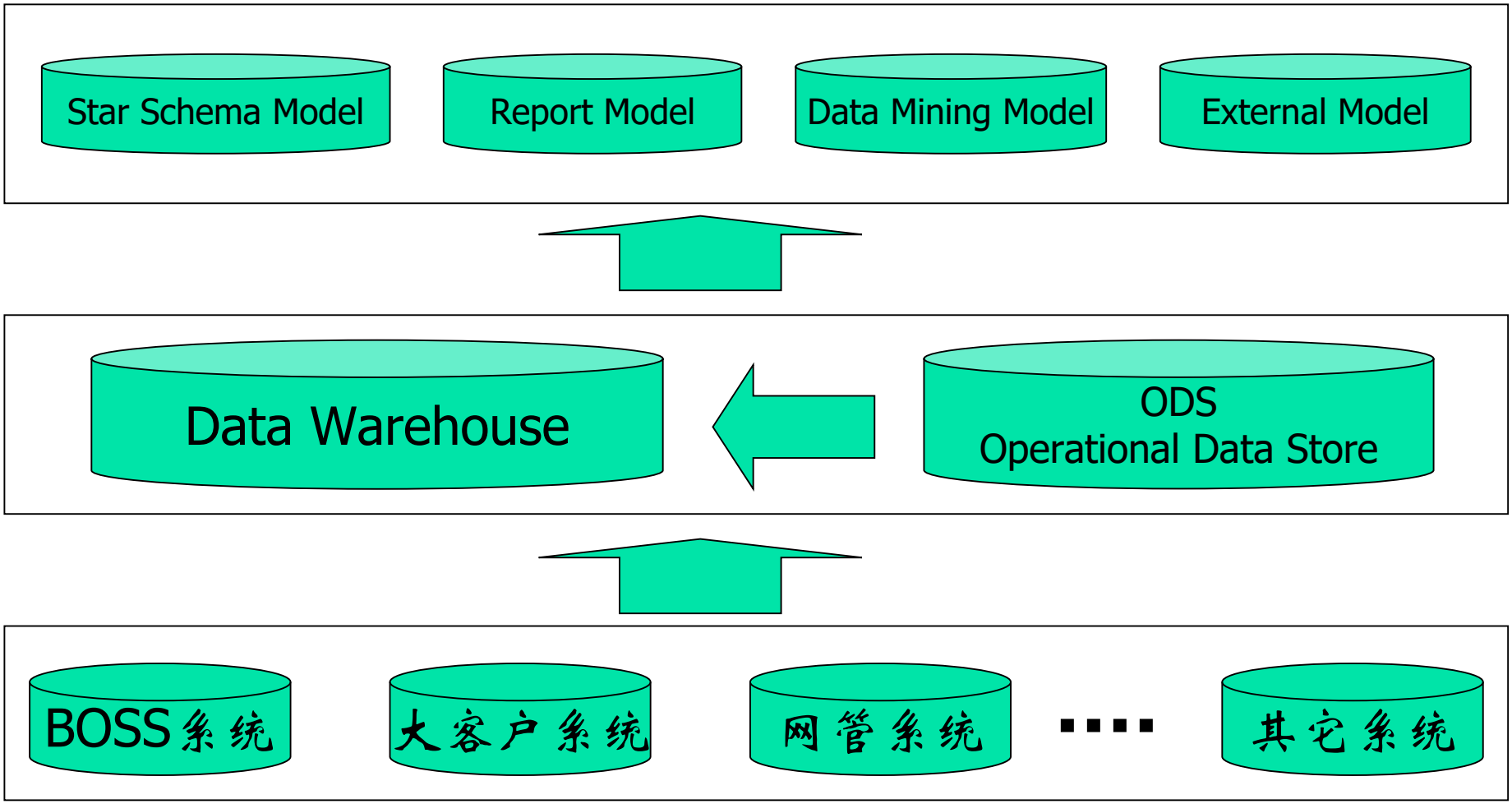
- BOSS：Business Operation Support System
  - 大客户系统
  - **网管系统：统一资源管理、统一性能管理、综合告警管理、操作维护模块**



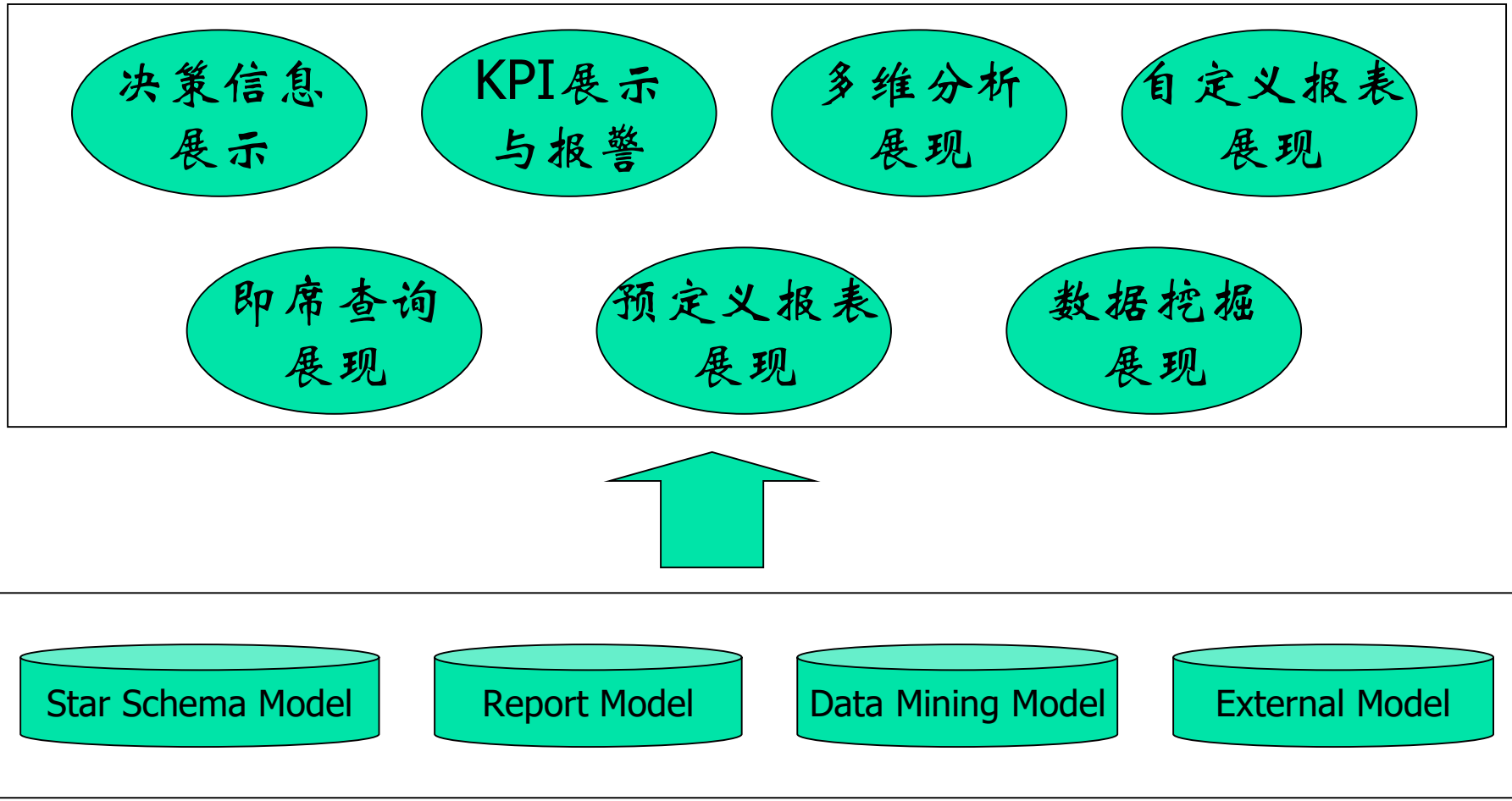
第二层 Data Warehouse



第三层 Departmental



第四层 个体层应用


决策信息展示
KPI展示与报警
多维分析展现
自定义报表展现
即席查询展现
预定义报表展现
数据挖掘展现
Star Schema Model
Report Model
Data Mining Model
External Model

10. 数据仓库架构所解决的问题

- 数据的一致性—通过集成
- 长期数据存储提升数据的利用效率
- 减轻生产环境的压力，为生产环境的改造奠定基础
- …

本部分内容提纲

- 1.1 从企业信息化到数据利用
- 1.2 企业中的决策与决策支持
- 1.3 决策支持系统的演化及数据仓库
- **1.4 互联网+时代企业信息系统架构演化与大数据**
- 1.5 数据仓库+大数据的决策支持平台新范式
- 1.6 数据仓库/大数据工程的定义

1. 互联网+时代OLTP系统的演变

- 互联网+各行各业，OLTP系统大量涌现
- OLTP系统规模越来越大与复杂程度越来越高
- OLTP系统所涉及的数据存储形态变化巨大
- OLTP系统的集成性越来越好
- 出现大批亿级以上客户服务对象的OLTP系统
- OLTP系统对数据利用的闭环需求越来越强烈

1. 互联网+时代OLTP系统的演变

- OLTP系统所涉及的数据存储形态变化巨大

多渠道的数据采集

持续50亿条记录/每天

持续35TB非结构化数据写入/每天

1. 互联网+时代OLTP系统的演变

▶互联网+各行各业，OLTP系统大量涌现

电信运营商

- 信令经营分析，CDR系统建设
- 社会关系挖掘

金融证券

- 交易历史统计，异常行为检测
- 商业决策支持

医疗

- 区域卫生医疗，临床决策支持
- 疾病模式分析，全民健康档案

国防安全

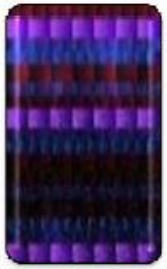
- 情报分析，网络安全、舆情分析
- 流量统计，图像，音视频分析

智慧城市

- 智慧交通
- 危险源监控
- 数字城市管理

互联网-物联网

- 社交挖掘，兴趣推荐
- 流式数据分析挖掘，实时统计

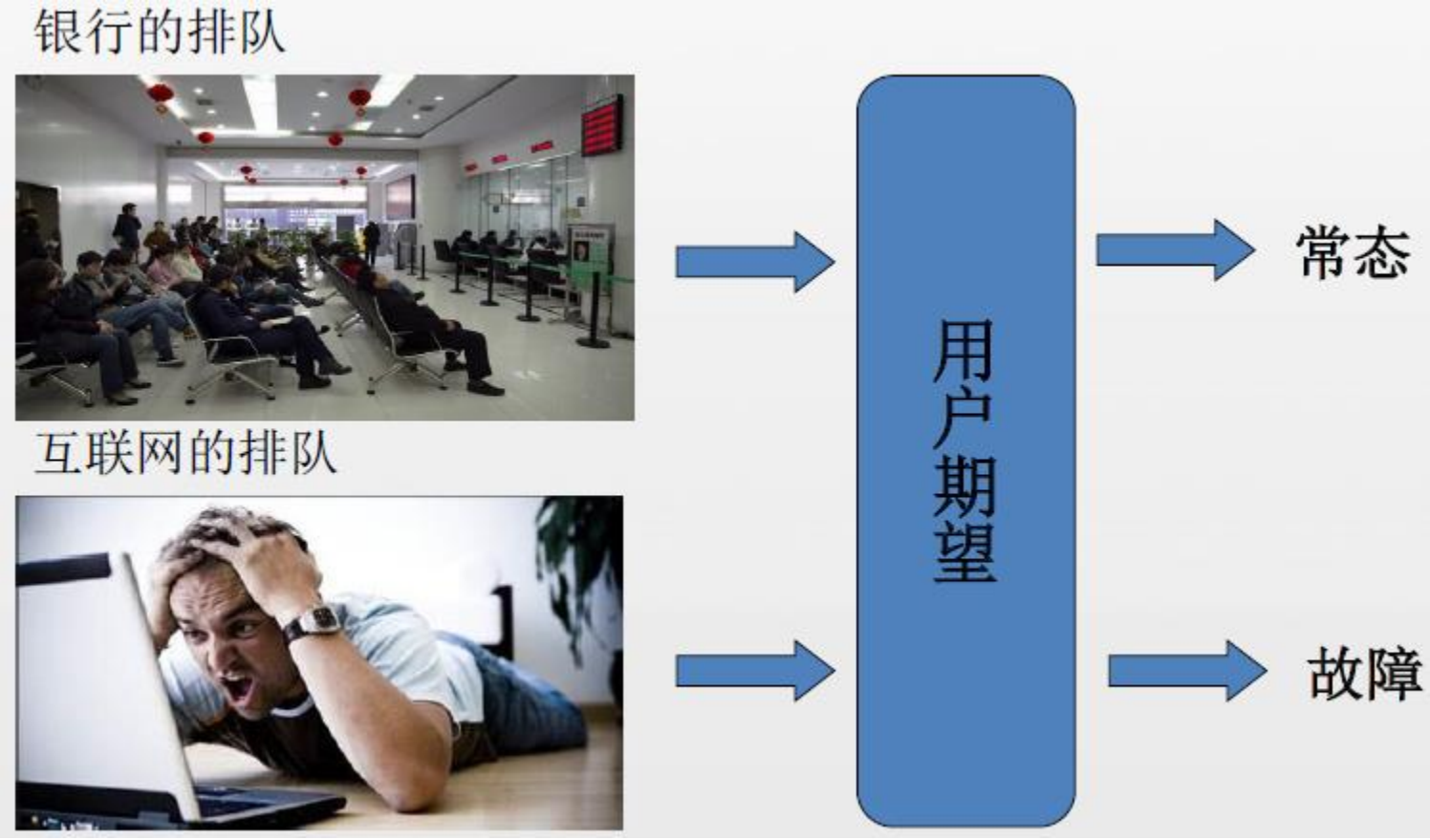


基础研究

- 理化模拟，生命科学
- 电力调度，能源勘探
- 气象气候，地球模拟

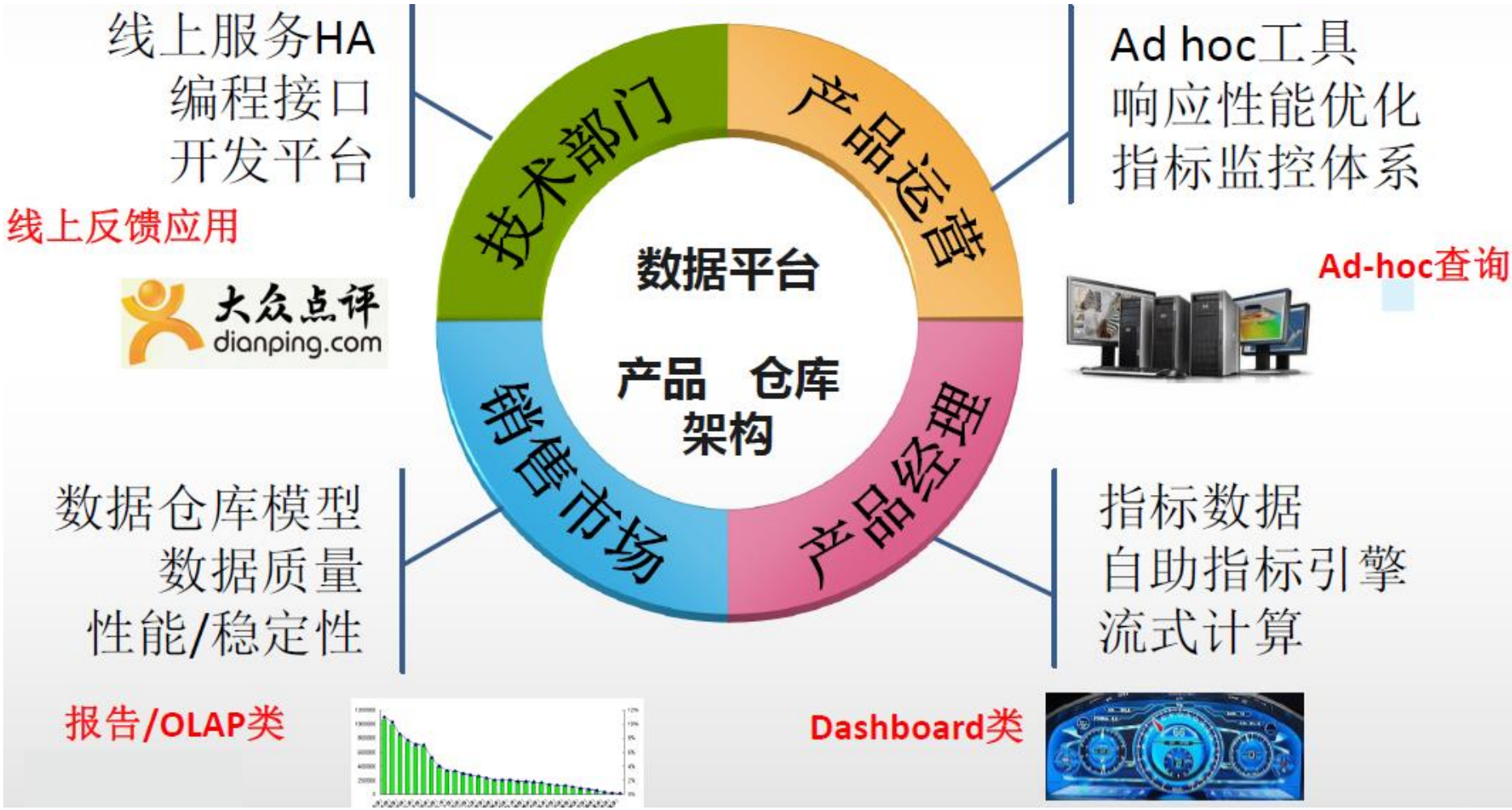
1. 互联网+时代OLTP系统的演变

- 互联网+时代用户数据体验需求



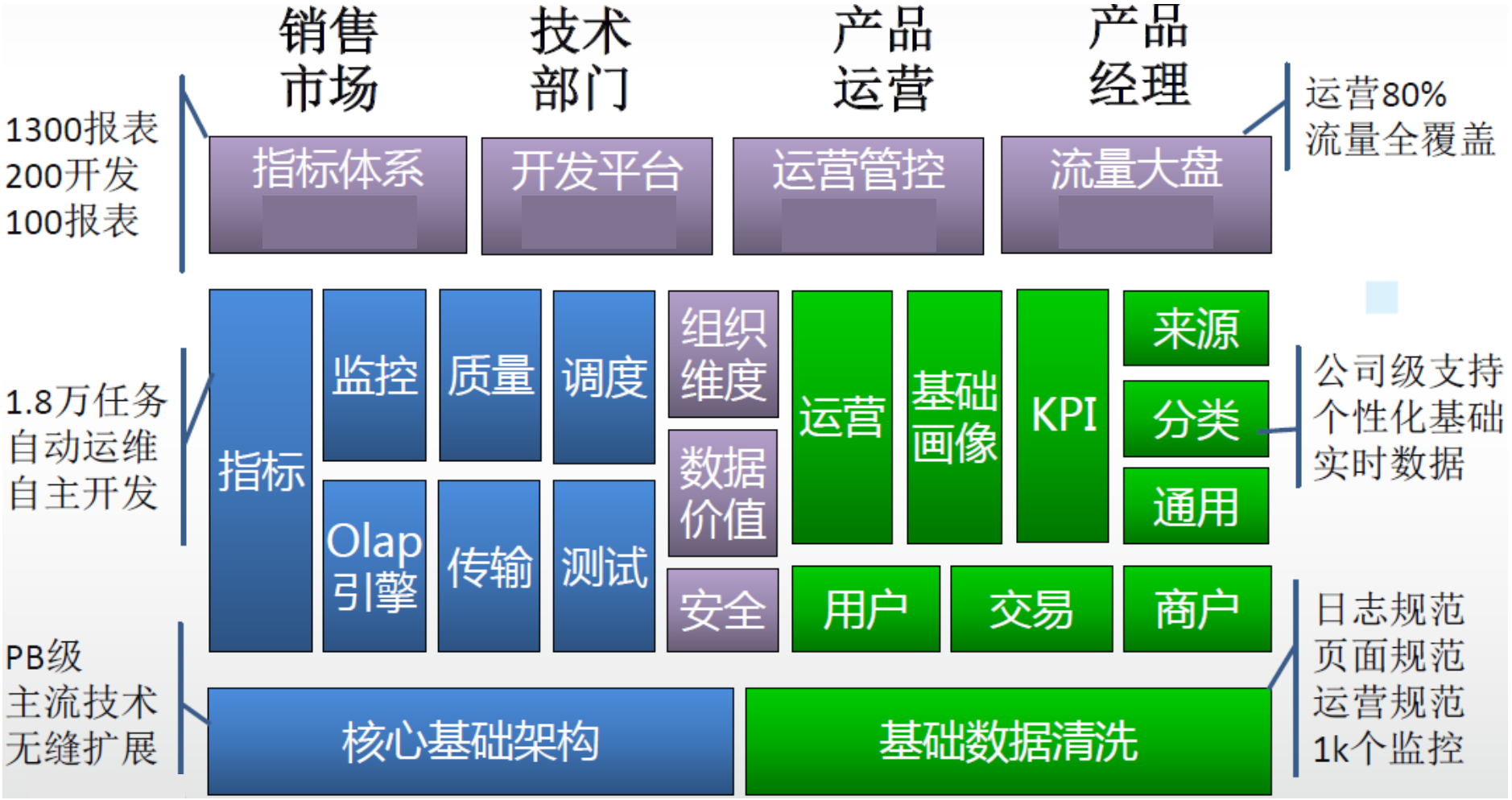
1. 互联网+时代OLTP系统的演变

▶互联网+时代数据用户



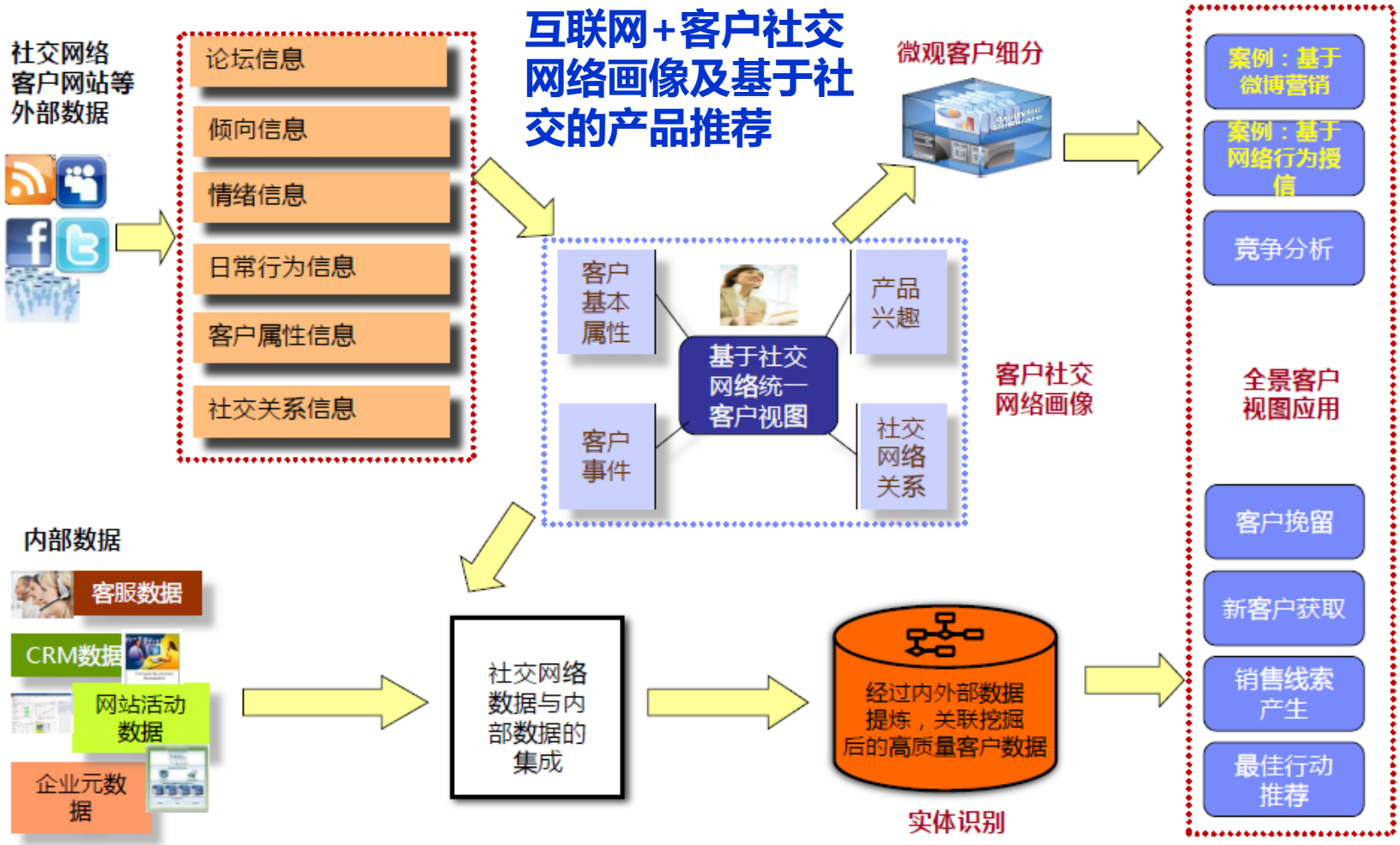
1. 互联网+时代OLTP系统的演变

- 出现大批亿级以上客户服务对象的OLTP系统



1. 互联网+时代OLTP系统的演变

▶OLTP系统与对数据利用的闭环需求越来越强烈

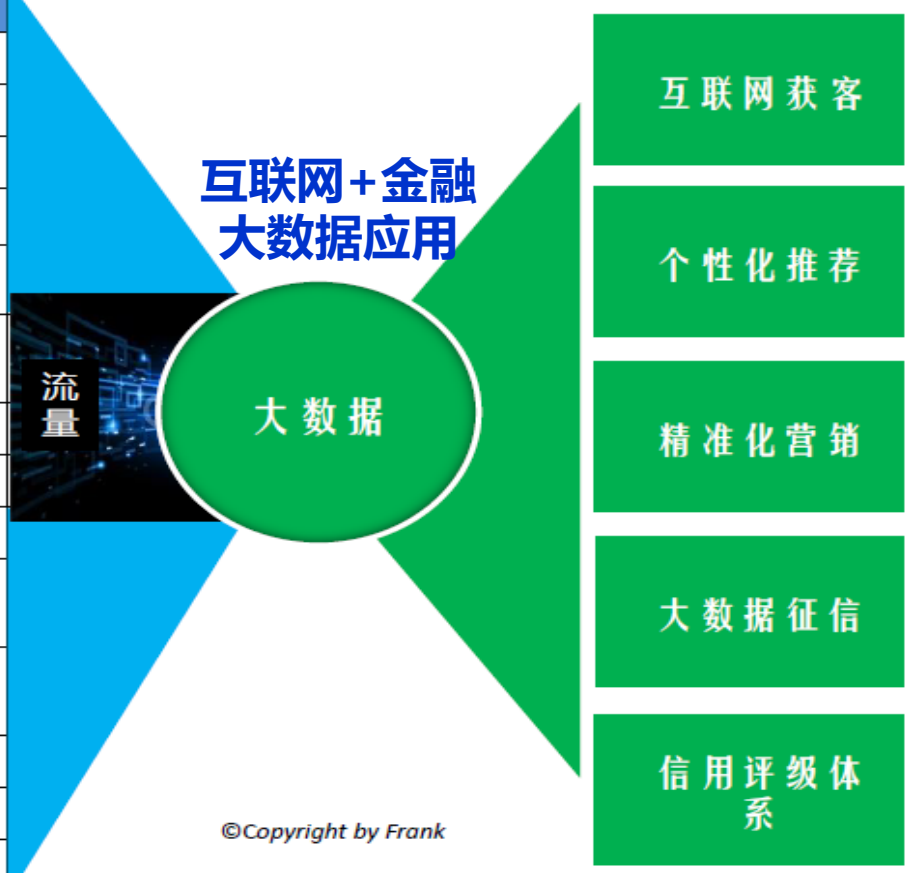


1. 互联网+时代OLTP系统的演变

▶OLTP系统的集成性越来越好

应用领域	业务模式	技术要点
支付结算	1. 第三方支付	虚拟账户
网络融资	2. P2P网贷	征信与风险评估体系
	3. 众筹融资	征信与风险评估体系
	4. 电商小贷	网络模型与信用体系
平台金融	5. 平台金融	大数据
	6. 供应链金融	大数据与信用评估模型
	7. 金融系电商	大数据
渠道创新	8. 传统电子渠道	移动互联网
	9. 金融超市	搜索、入口与流量
	10. 搜索与金融门户	搜索、入口与流量
	11. 直销银行	电子渠道、移动互联网
产品创新	12. 余额理财	第三方合作渠道
	13. 无抵押贷款	第三方合作渠道
虚拟货币	14. 虚拟货币	网络算法

系统统一**集成建设的模式**越来越普遍
系统联动、数据互通、模型标准一致
简化ETL，简化数据利用层的建模



企业级的ERP模式越来越普遍

1. 互联网+时代OLTP系统的演变

▶ 互联网+银行的在线和数字化业务模式

消费者变化

- 相互沟通频繁
- 社交网络
- 消息灵通
- 购买随处发生

业务模式变化

- 引入渠道，品牌和社区的概念
- 以客户为中心
- 统一的跨渠道销售
- 业务创新速度加快
- 交易成本显著下降
- 合规需求发生变化

IT架构变化

- 高级的用户界面，丰富的多媒体内容
- 复合型应用，云计算
- 基于上下文感知的计算
- 大数据挖掘利用，分析软件的实时洞察力

1. 互联网+时代OLTP系统的演变

▶ 互联网+水利大数据基本需求

- 提高行业服务能力
- 促进行业加快转型

水利大数据资源化

业务　技术　科学　艺术

信息数据化

数据共享化

数据服务化

1. 互联网+时代OLTP系统的演变

▶互联网+水利大数据应用

互联网+

把水利基础设备连接到互联网上，让水利设施具有计算、通信、精确控制、远程协调和自我管理的功能，实现虚拟网络控制和现实水利建设管理的融合。

物联网

基于物联网技术，充分利用MSTP和智能感知等技术深入开发和高度整合水利信息资源，实现水利信息感知、传输、应用的网络化与智能化，有效的实现水利信息的共享，提升水利工程运用和管理的效率和效能。

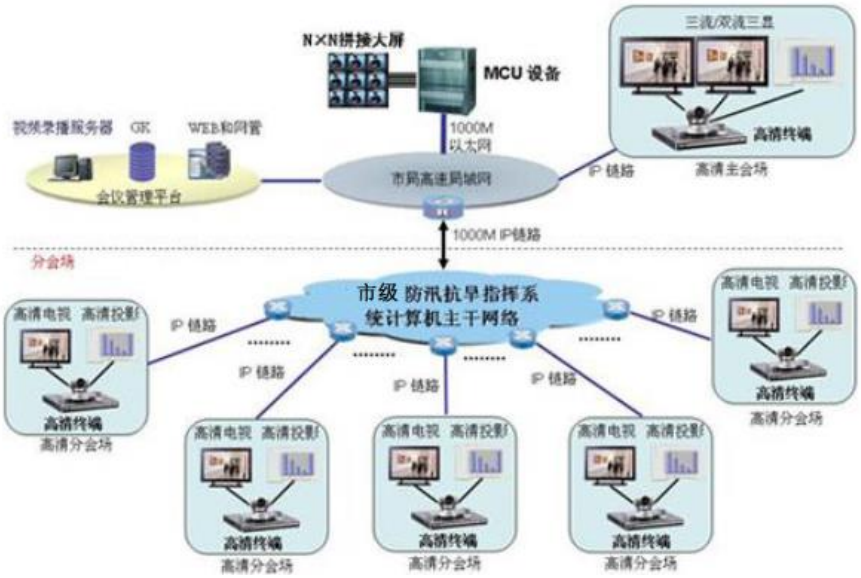
云计算

水利行业密集型计算和海量存储的需求随着防汛任务和突发水事而瞬间剧增。将云计算有效解决密集型计算和高可靠性存储需求伴随防汛任务和突发水事瞬间大幅度增长与建设投资之间的矛盾。

1. 互联网+时代OLTP系统的演变

▶互联网+水利大数据应用场景：防汛抗旱

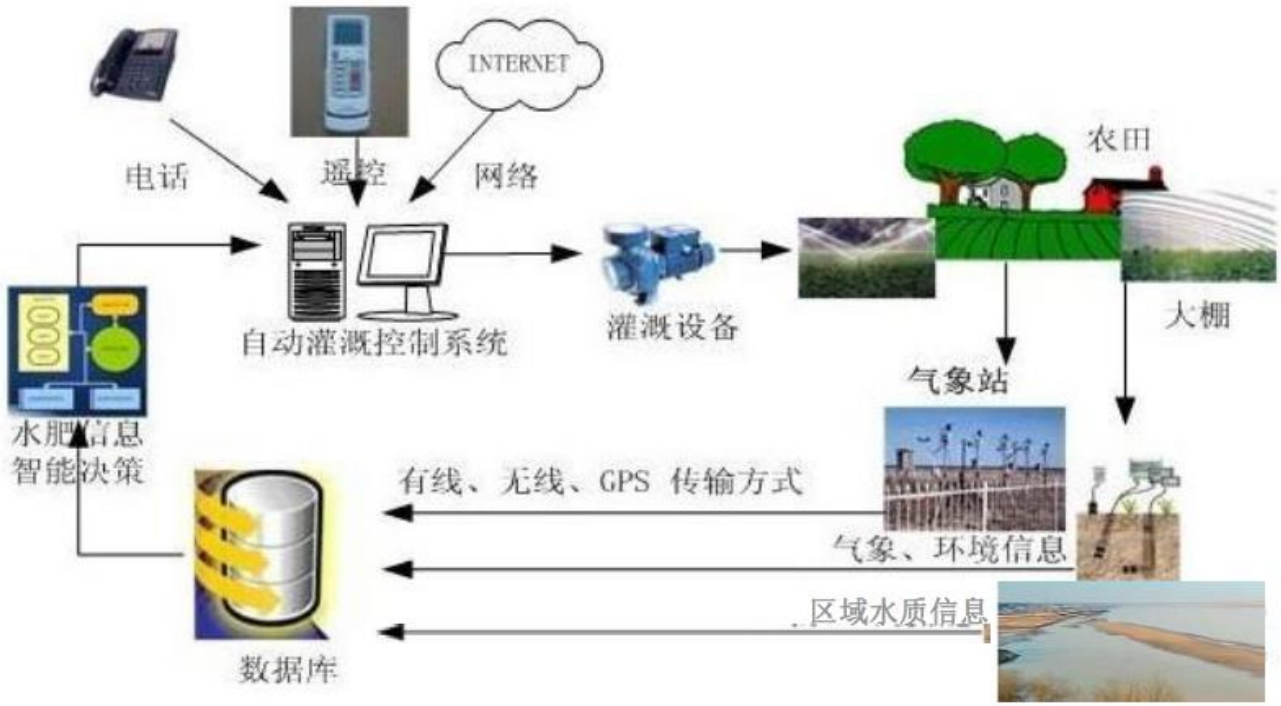
运用互联网大数据等技术手段，通过远程数据的自动采集、监控、调度、分析、预警，实现了“不到现场看现场”的效果，有效降低了人员的劳动强度，可整体掌握区域范围的水情雨情，为防汛指挥决策提供强力支撑。



1. 互联网+时代OLTP系统的演变

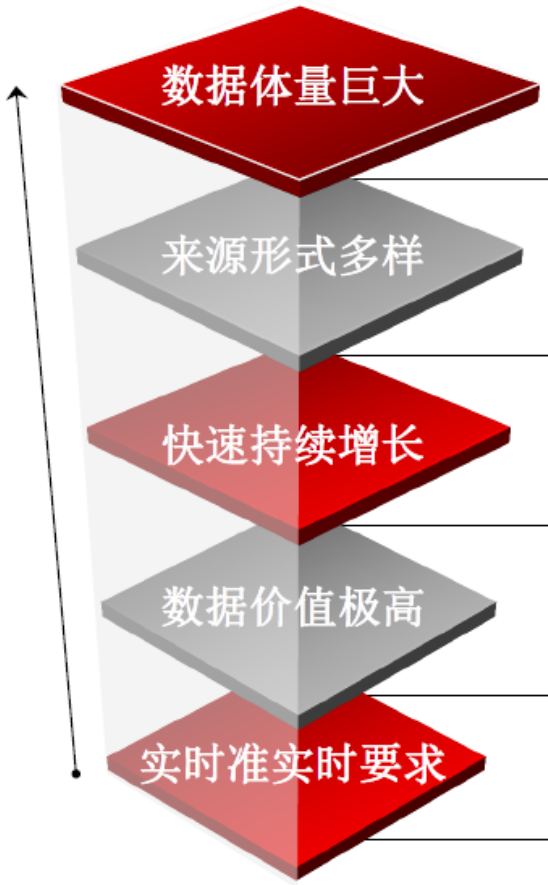
互联网+水利大数据应用场景：水利设施自动化

在水利管理中，水利部门可以根据区域水量、水位、潮位、气象、水质、蒸发量等信息进行分析，为水资源调度、农业灌溉等提供决策支持，从而实现区域内各类水利设施按需自动控制，提高效率。此外，防汛工程、山洪预警、城市排水等工程，也均可以借助来自多个部门的数据提高效率



1. 互联网+时代OLTP系统的演变

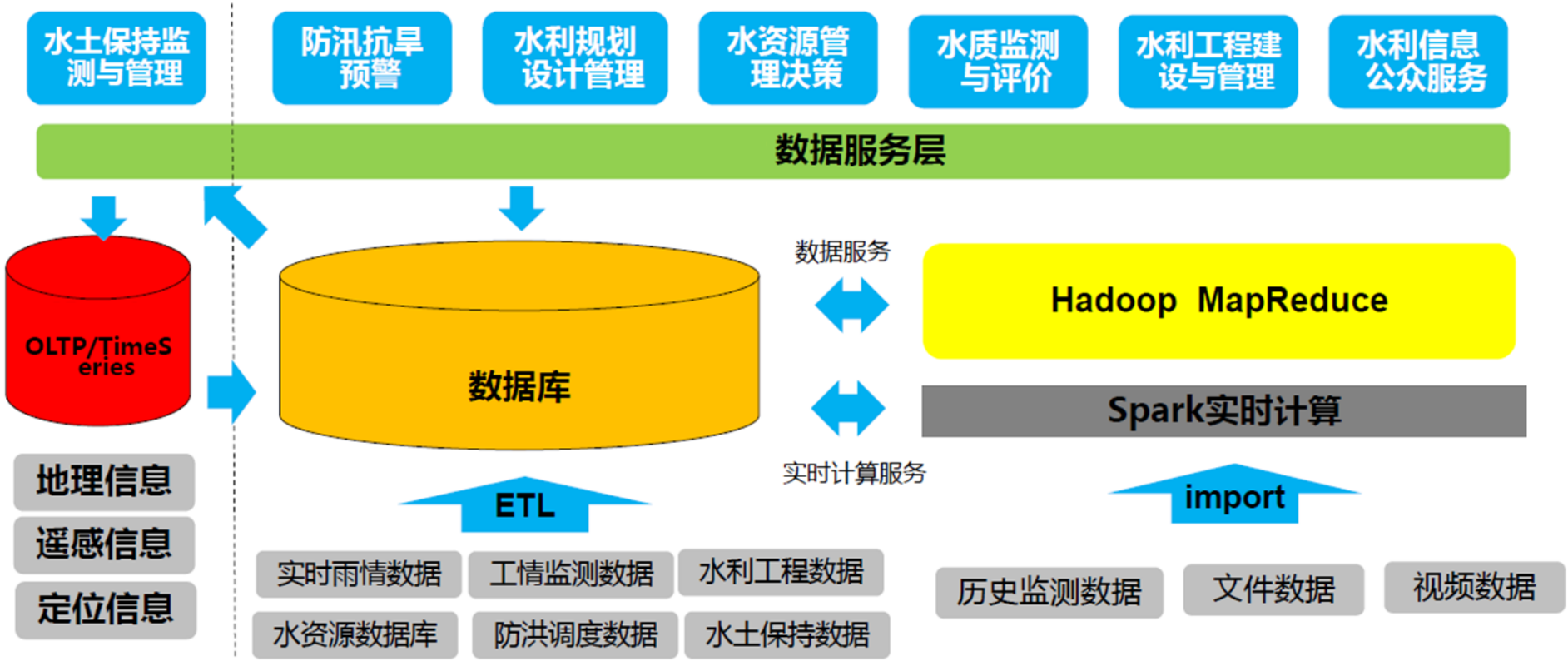
互联网+水利大数据特点



- 数百TB或PB以上
- 来源：勘测、规划、设计、施工、管理
- 结构：长系列的结构化、半结构化数据和大量非结构化数据
- 水利行业各环节信息化应用增加
- 监测密度及指标提升，数据增加速度不断加快
- 水利工程建设、管理及水行政业务处置的依据，蕴含较高的价值
- 判别应急事件，存在实时或准实时处理

1. 互联网+时代OLTP系统的演变

▶水利大数据系统架构



2. 技术条件演变与大数据的迅猛发展

- **硬件性能**提升
  - 内外存储能力大幅提升大幅降价
  - CPU处理能力
- **数据处理能力**飞速进步
  - 通信、传输、存储、分析、…
- **新的计算模式**
  - 分布存储、并行计算

2. 技术条件演变与大数据的迅猛发展

- 云计算、大数据分析、移动、社交和物联网（IoT）

2. 技术条件演变与大数据的迅猛发展

2. 技术条件演变与大数据的迅猛发展

大数据在银行业的典型使用

风险管理

- 信用风险模型
- 偿付与资本优化
- 金融风险分析
- 产品与承保优化
- 实时欺诈检测
- 大数据征信

渠道优化

- 渠道创新
- 网点转型
- 渠道有效性
- 渠道集成与管理
- 联系中心转型
- 自助服务
- 移动服务

客户管理

- 客户分析与洞察
  - 客户保留
  - 交叉销售与向上销售
  - 360度全景客户视图
  - 增强的客户细分
- 产品有效性分析
- 社交媒体/舆情分析
- 联系中心分析
- 最佳行动推荐（NBA）
- 产品组合管理
- 客户之声

运营优化

- 历史数据保存与管理
- 系统日志维护
- 系统故障分析

3. 决策支持架构的演变

- 决策支持**对象层次的下移**
  - 从面向高层的决策，向更多面向中下层决策演变
  - 从向面人的决策更多地面向机器的决策
  - 从面向内部人员的决策更多的面向外部客户的决策
- 决策支持**时效性不断提升**
  - 从偶发性决策支持近实时和实时决策支持演变
- 整体系统架构**从开环**更多地**演变到闭环**
  - 业务→数据→决策支持→决策→业务
- 从关系型+BI组件，到**混合架构**

4. Big Data—大数据基本概念

- **麦肯锡对大数据的定义**

“大数据”是指其大小超出了典型数据库软件的采集、储存、管理和分析等能力的数据集。

- **维基百科对大数据的定义**

大数据是指无法在一定时间内用常规主流软件工具对其内容进行获取、管理和处理的数据集合

- **大数据定义内涵**
  - 符合大数据标准的数据集大小是变化的，会随时间推移、技术进步而增长
  - 不同部门符合大数据标准的数据集大小会存在差别。目前，大数据的一般范围是从几个TB到数个PB（数千TB）

大数据—Big Data

- 大数据是一个包罗万象的术语，用于指任何一种量大、复杂的用传统的数据处理应用难以处理的数据集。
- 挑战包括
  - 分析、捕获、保管、搜索、共享、存储、转换、可视化、隐私保护
- 大数据很难用大多数关系型数据库管理系统和桌面分析和可视化工作，需要能在几十、几百甚至几千台服务器上跑的大规模并行软件对数据进行处理。

大数据特征

- Volume – 量
- Variety - 多样性，类别很重要
- Velocity - 速度，产生和处理数据的速度
- Variability - 可变性，数据时常会发生变化
- Veracity - 真实性，质量，数据的真实性影响分析的质量
- Complexity - 数据管理很复杂，多数据来源，linked, connected and correlated，关联很重要

OBAMA政府Big Data R&D Initiative

- **Aims:**
  - Advance state-of-the-art core technologies needed to collect, store, preserve, manage, analyze, and share huge quantities of data;
  - Harness these technologies to accelerate the pace of discovery in science and engineering, strengthen our national security, and transform teaching and learning;
  - Expand the workforce needed to develop and use Big Data technologies
- **National Science Foundation, National Institutes of Health, Department of Defense, Department of Energy，US Geological survey…**

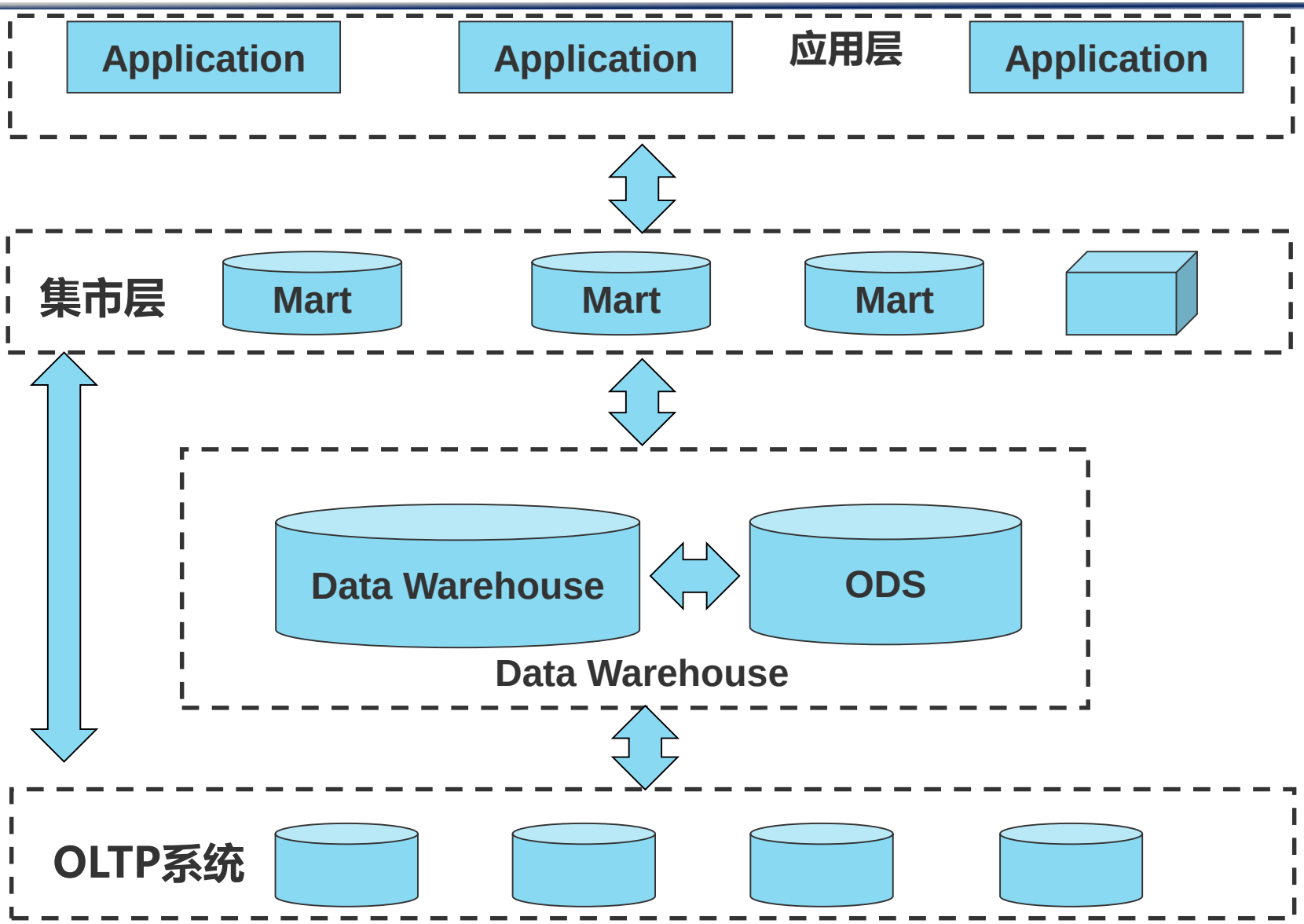
OBAMA政府Big Data Initiative

- Department of Defense—Data to Decisions: a big bet on big data, $250 million,… TO:
  - Harness and utilize massive data in new ways and bring together sensing, perception and decision support to make truly **autonomous systems** that can maneuver and make decisions on their own.
  - Improve situational awareness to help warfighters and analysts and provide increased support to operations,…

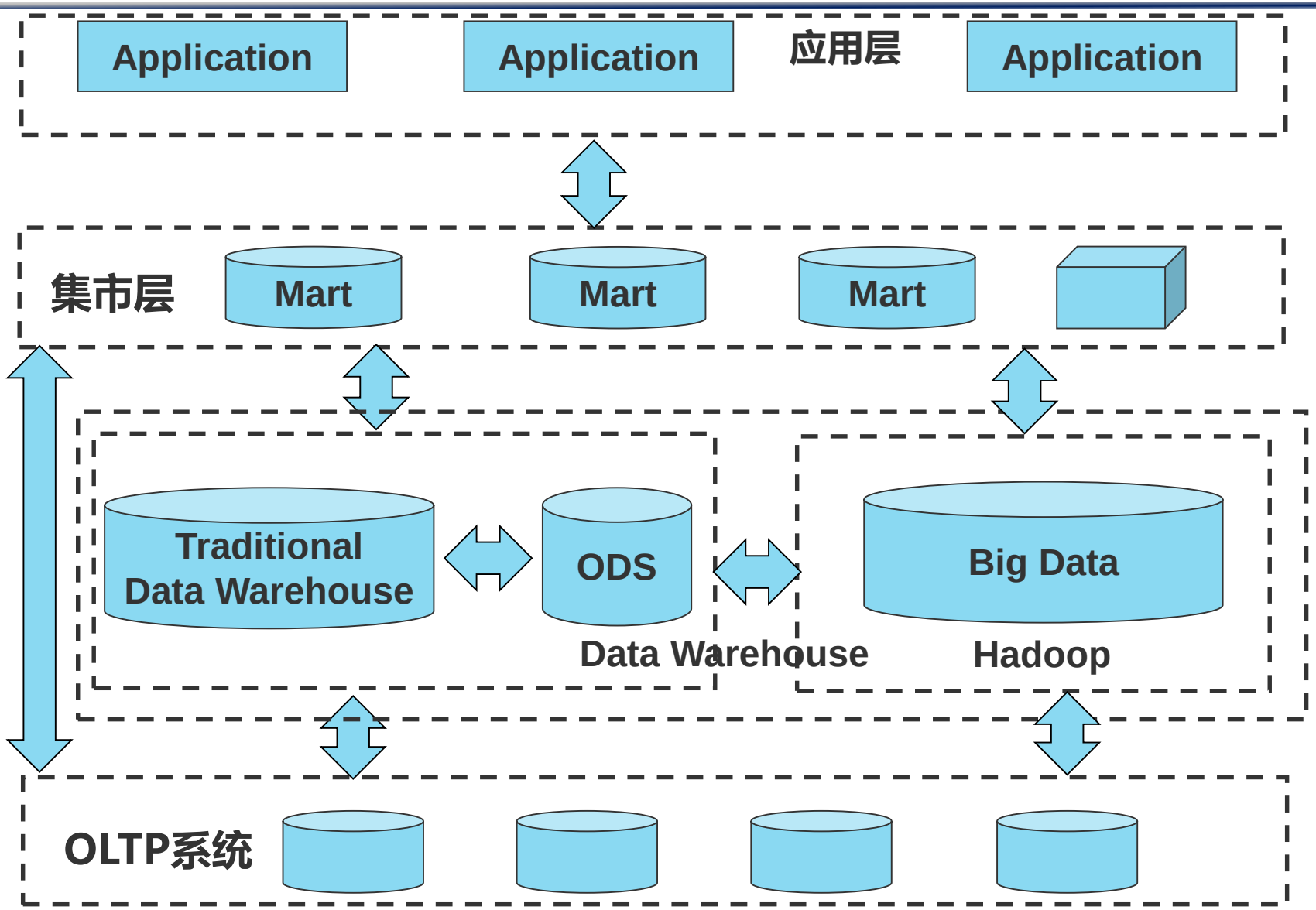
本部分内容提纲

- 1.1 从企业信息化到数据利用
- 1.2 企业中的决策与决策支持
- 1.3 决策支持系统的演化及数据仓库
- 1.4 互联网+时代企业信息系统架构演化与大数据
- **1.5 数据仓库+大数据的决策支持平台新范式**
- 1.6 数据仓库/大数据工程的定义

1. 传统四层式数据仓库数据架构



2. 数据仓库+大数据架构


Application
Application
应用层
Application
集市层
Mart
Mart
Mart
Traditional
Data Warehouse
ODS
Data Warehouse
Big Data
Hadoop
OLTP系统

3. 用户行为模式及影响

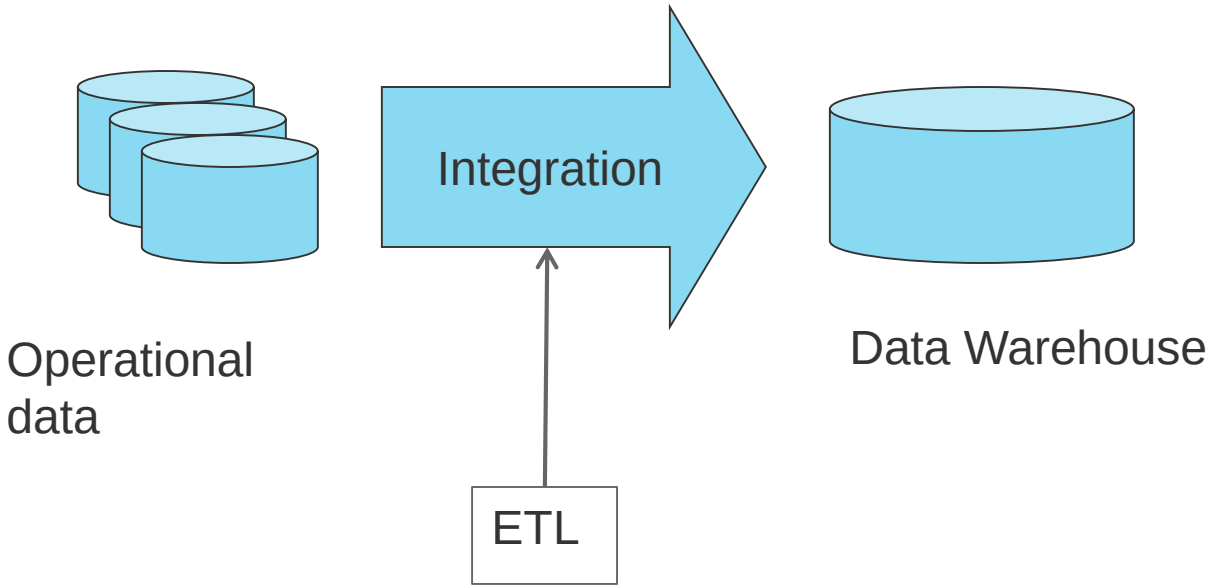
- 用户—DSS 分析人员或系统
  - 业务人员：Business person first and foremost
  - 技术人员：Technician second.
  - 自动系统：Automatic Agent
- DSS分析人员的主要任务
  - To define and discover information used in corporate decision-making.
- DSS分析人员的思维模式
  - Give me what I say I want, then I can tell you what I really want.
  - 先把我要的数据给我，然后我会告诉你我真正需要的数据，工作于发现模式。

用户思维模式的影响

- DSS分析人员的这种思维习惯是很重要，也是合理的、普通的。
  - 这种模式对平台的**数据提供能力**具有很高的要求
  - 这种模式对数据仓库的开发和数据仓库之上的**应用开发具有非常大的影响**。
  - 要求系统具有**快速反应能力**
- **数据集成性**要求高
- 对支撑平台的**硬件环境要求**与OLTP平台不尽相同
- 典型的系统开发周期(SDLC)在某种程度上不再适用于有些环节的开发工作

4. 数据仓库环境中的数据集成问题

- 数据集成是实现数据仓库数据**企业级视图**的关键，不可缺少。
- 数据集成通过**ETL软件**或程序完成



否则，无法为分析主题提供全面的数据内容，分析应用研发将变得繁杂而不易管理。

数据集成

- 如何集成
  - 自主编程
    - Java, C++，有大量的Open Source tool可以参考。
    - 数据库存储过程
    - 是一项系统性研发工作
  - 采用现有ETL 商业软件
    - Data integrator,
    - 特点：可通过工作流(workflow)作业式设计，形成ETL任务，自动定时完成数据集成工作。
  - 编程 or 购买ETL商业软件?
    - 能否满足你的需求，以及投入/产出考虑。
- 数据集成一般一次性完成，但却是数据仓库能否成功的关键之一。

数据集成例子(1)：商业领域

把各业务系统的部分数据整合成具有信息关联的完整数据

人寿保险

张三, 女
1965年7月20日
……

汽车保险

张三
去年有两张罚单
一次大事故, ……

房屋保险

张三
学院路123号
已婚, ……

健康保险

张三, 1个孩子
高血压
……

集成

张三
女
1965年7月20日
去年有两张罚单
一次大事故
学院路123号
已婚
1个孩子
高血压
……

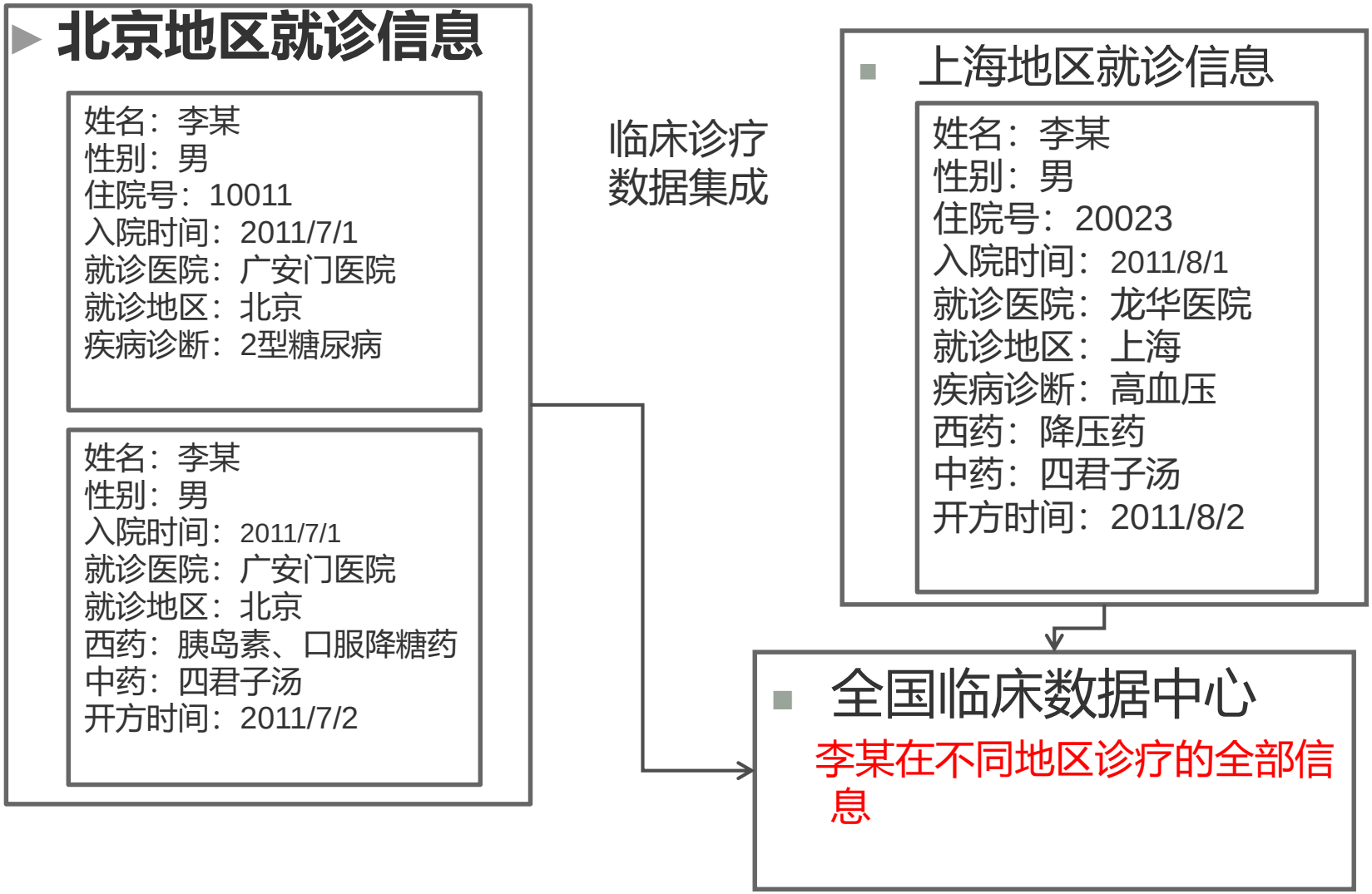
完整视图，全面把握情况

不同的业务关注的角度不一样

数据集成(2)：通信领域

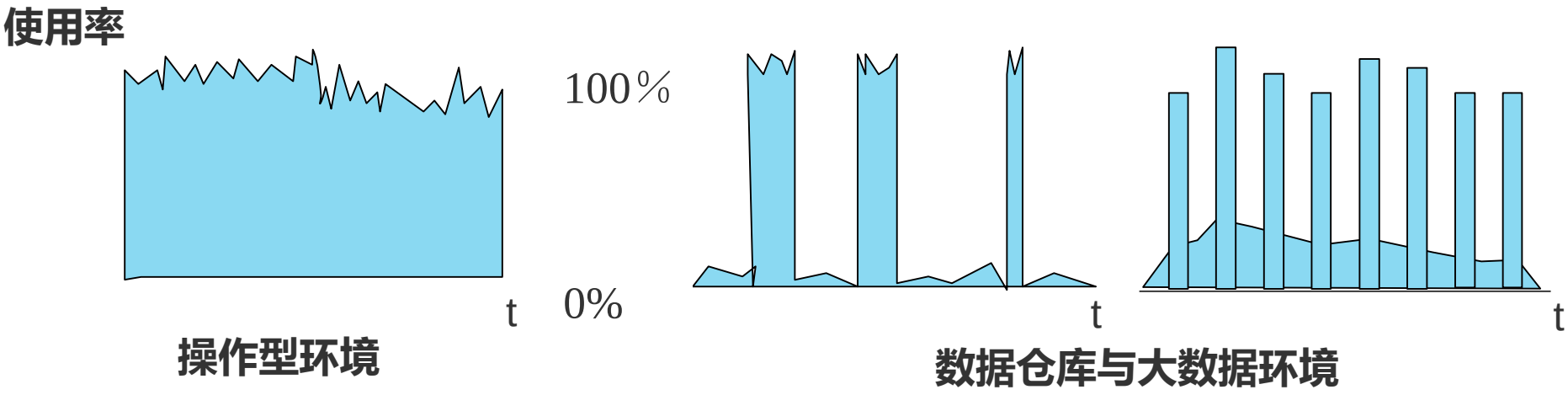
- 手机用户套餐推荐
- 用户行为信息集成
  - 语音通话行为数据
  - 短信行为数据
  - 上网行为数据
- 套餐数据

数据集成(3)：医疗领域


北京地区就诊信息
姓名：李某
性别：男
住院号：10011
入院时间：2011/7/1
就诊医院：广安门医院
就诊地区：北京
疾病诊断：2型糖尿病
姓名：李某
性别：男
入院时间：2011/7/1
就诊医院：广安门医院
就诊地区：北京
西药：胰岛素、口服降糖药
中药：四君子汤
开方时间：2011/7/2
临床诊疗
数据集成
上海地区就诊信息
姓名：李某
性别：男
住院号：20023
入院时间：2011/8/1
就诊医院：龙华医院
就诊地区：上海
疾病诊断：高血压
西药：降压药
中药：四君子汤
开方时间：2011/8/2
全国临床数据中心
李某在不同地区诊疗的全部信息

5. 硬件使用模式

- 操作型环境和数据仓库环境的另一个重要的不同点在于硬件的使用模式上。



操作型环境

数据仓库与大数据环境

问题：为什么会产生这样的模式差异?
注意：第1图和后两个图间的差异，以及后两个图之间的差异

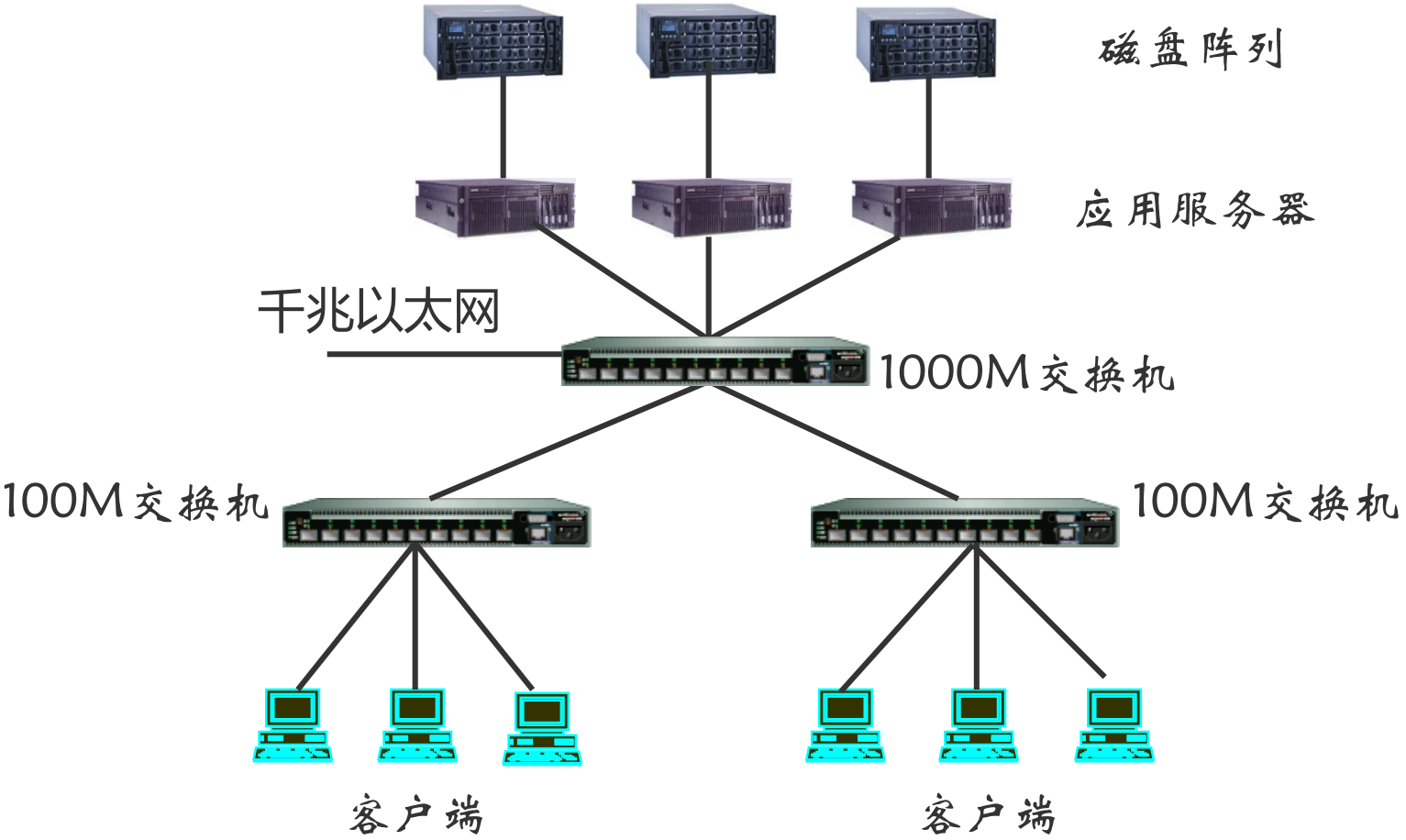
硬件利用模式不同

- 对于操作型处理来说
  - 硬件使用利用模式相对稳定，可预测。
- 对于数据仓库处理或应用来说
  - 它的硬件使用模式相当不稳定。
- 硬件使用模式的不同，说明
  - 不应将两种应用混在一起。
  - 分开以后，可以针对不同的处理，分别进行相应的优化处理。

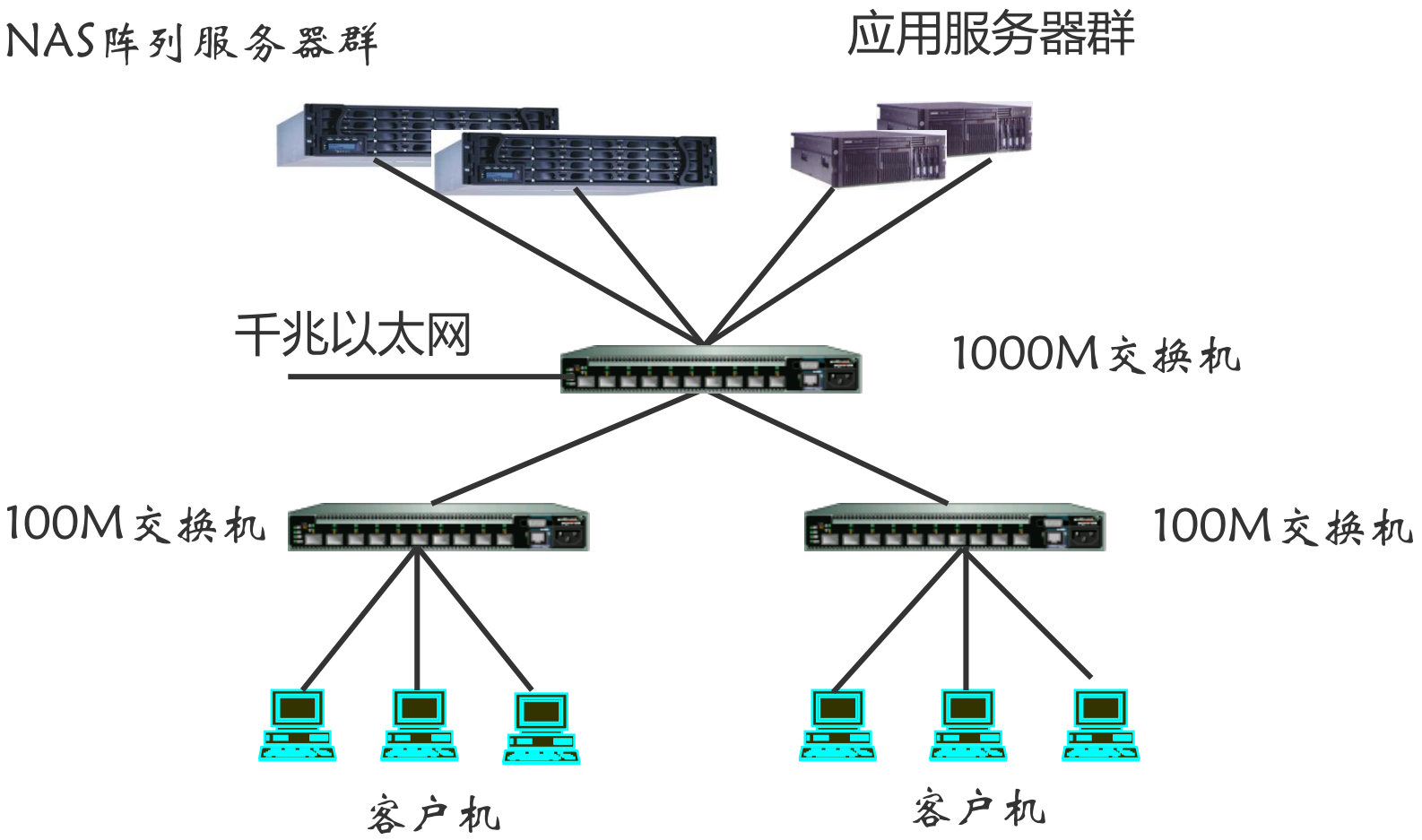
当前常见存储体系简介

- 常见存储体系
  - DAS，Direct Attached Storage
  - NAS，Network Attached Storage
  - SAN，Storage Area Networks

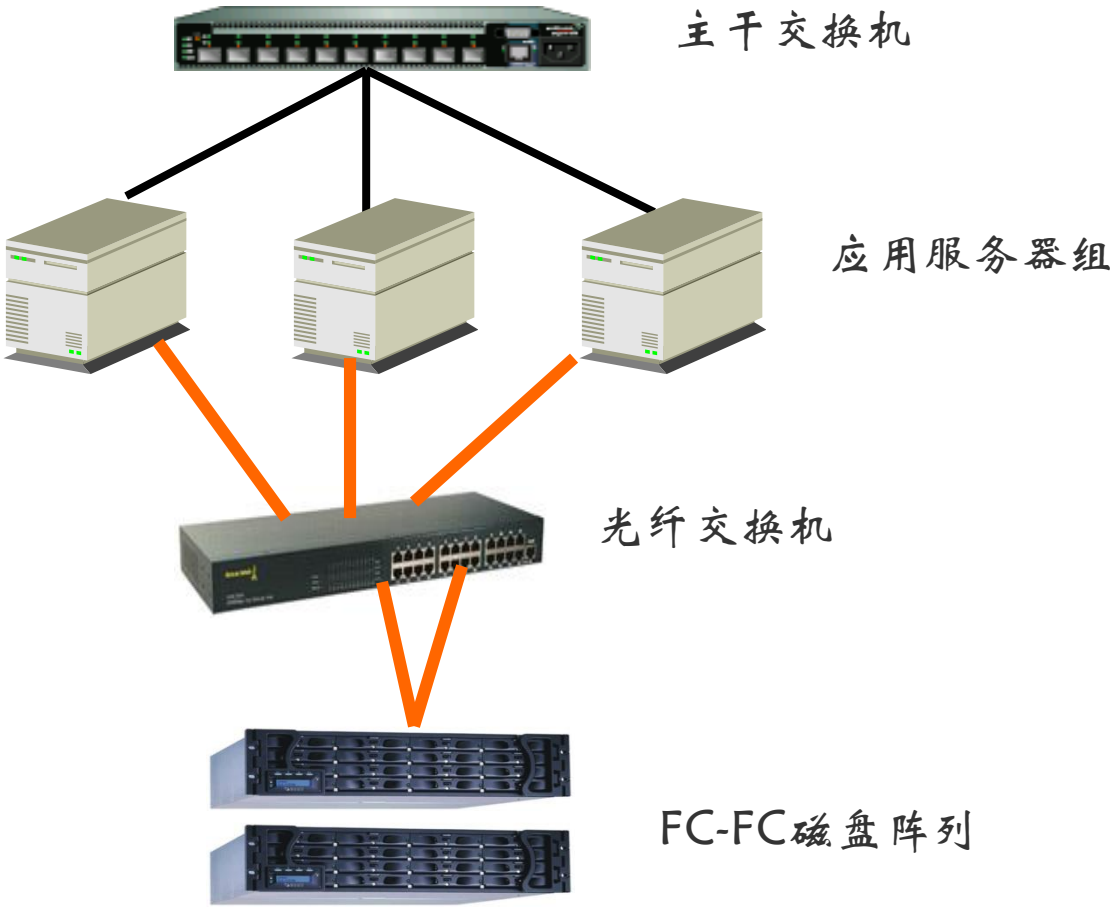
DAS存储体系结构实例图



NAS存储体系结构实例图

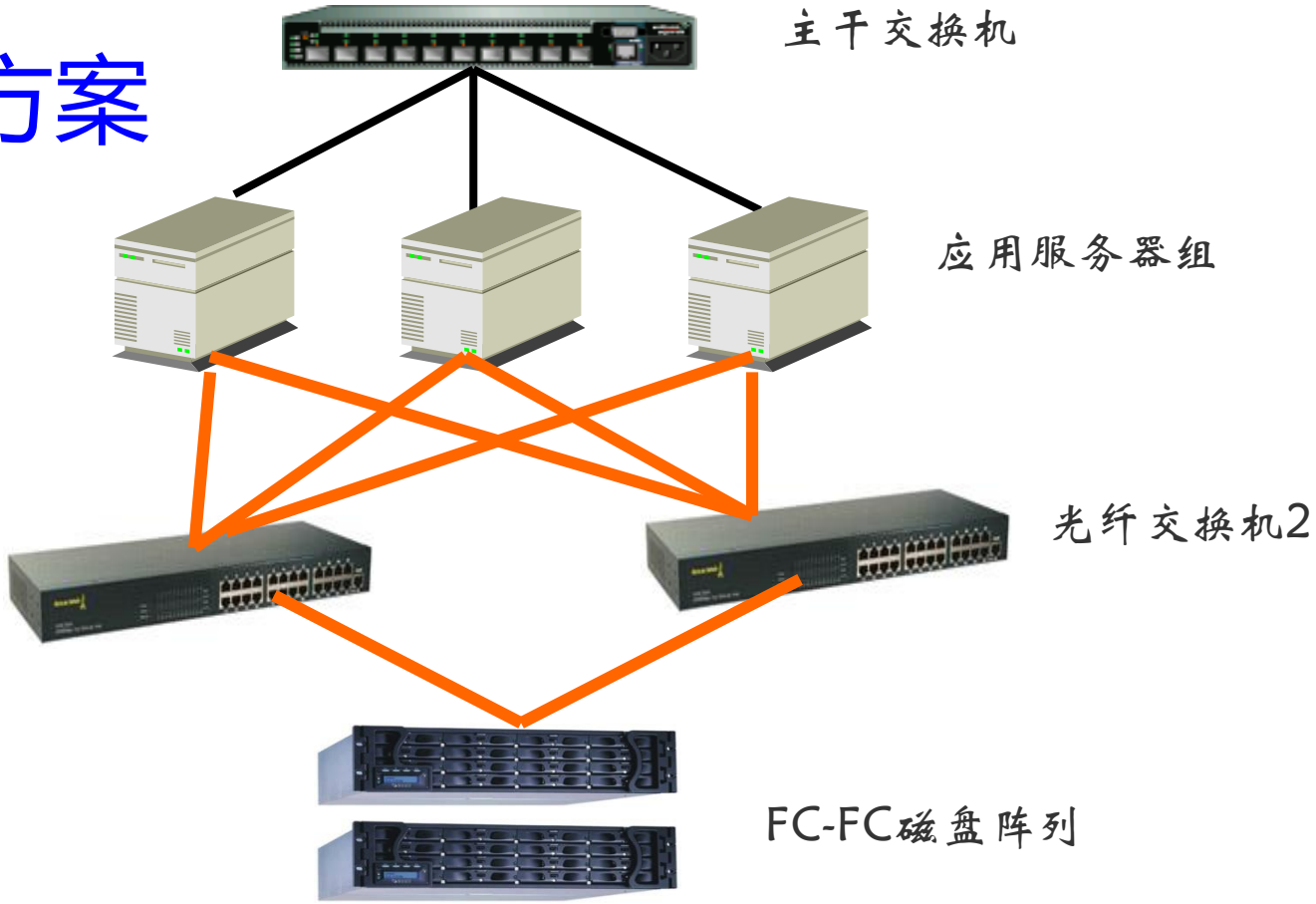


SAN存储体系结构实例图一



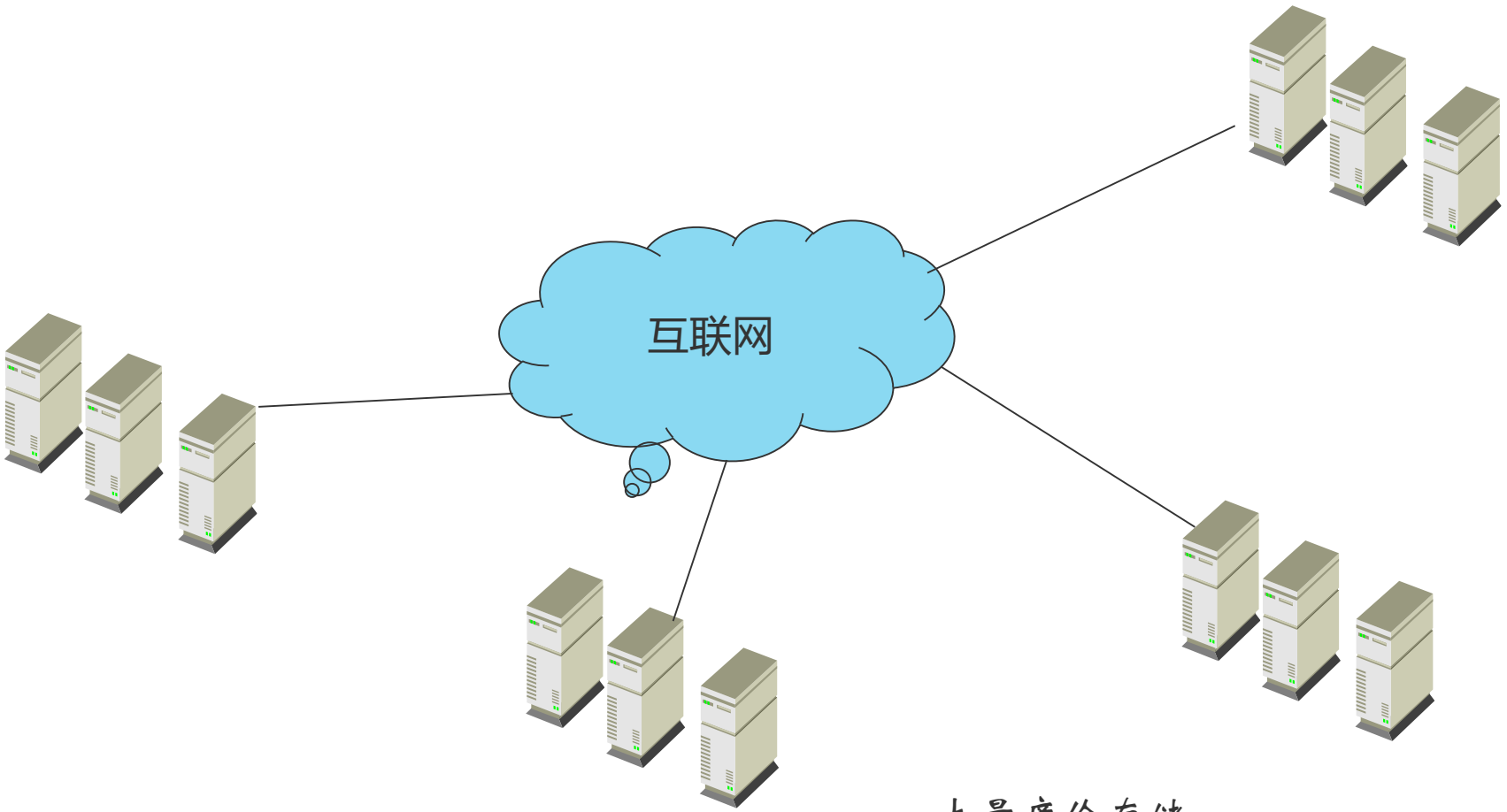
SAN存储体系结构实例图二

双冗余方案



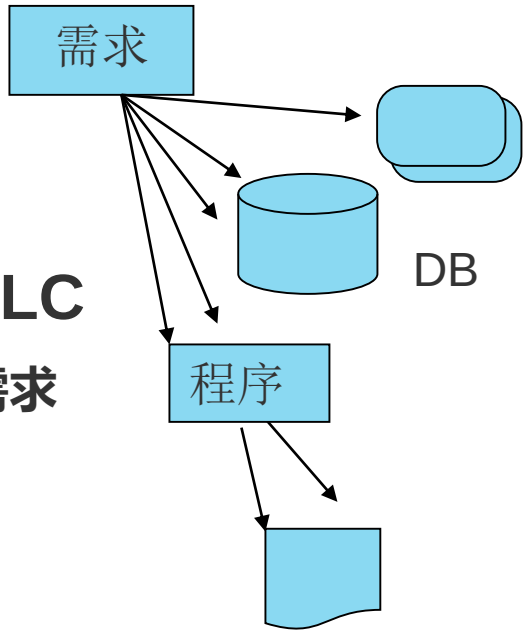
新一代分布式存储与并行计算架构

分布式存储架构



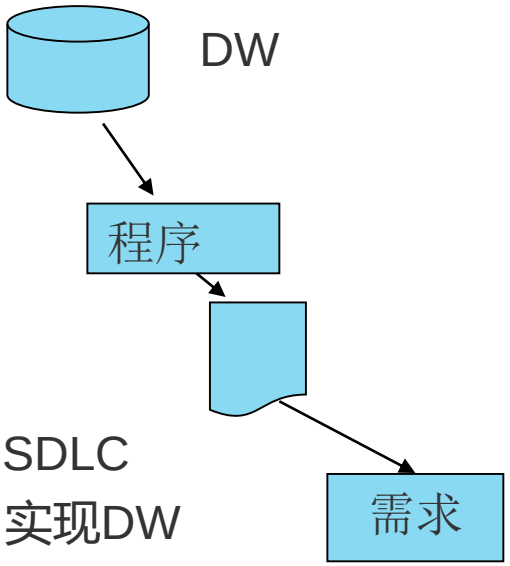
大量廉价存储
与服务设备

6. 开发生命周期



- **传统SDLC**
  - **收集需求**
  - **分析**
  - **设计**
  - **编程**
  - **调试**
  - **集成**
  - **实现**

几乎完全相反！



- DW SDLC
  - 实现DW
  - 集成数据
  - 检验偏差
  - 编程
  - 设计DSS
  - 分析结果
  - 理解需求

本部分内容提纲

- 1.1 从企业信息化到数据利用
- 1.2 企业中的决策与决策支持
- 1.3 决策支持系统的演化及数据仓库
- 1.4 互联网+时代企业信息系统架构演化与大数据
- 1.5 数据仓库+大数据的决策支持平台新范式
- **1.6 数据仓库/大数据工程的定义**

数据仓库与大数据工程

- 数据仓库与大数据平台的**规划、设计、实现和运维全生命周期工程方法论和技术**，主要涉及如下环节的核心概念、方法与技术：
  - 数据集成
  - 数据利用需求
  - 系统支撑架构
  - 数据组织与环境
  - 数据和功能模型设计
  - 系统实现
  - 部署与运行管理
- 课程内容围绕这些环节开展

本部分结束！

BEIJING JIAOTONG UNIVERSITY