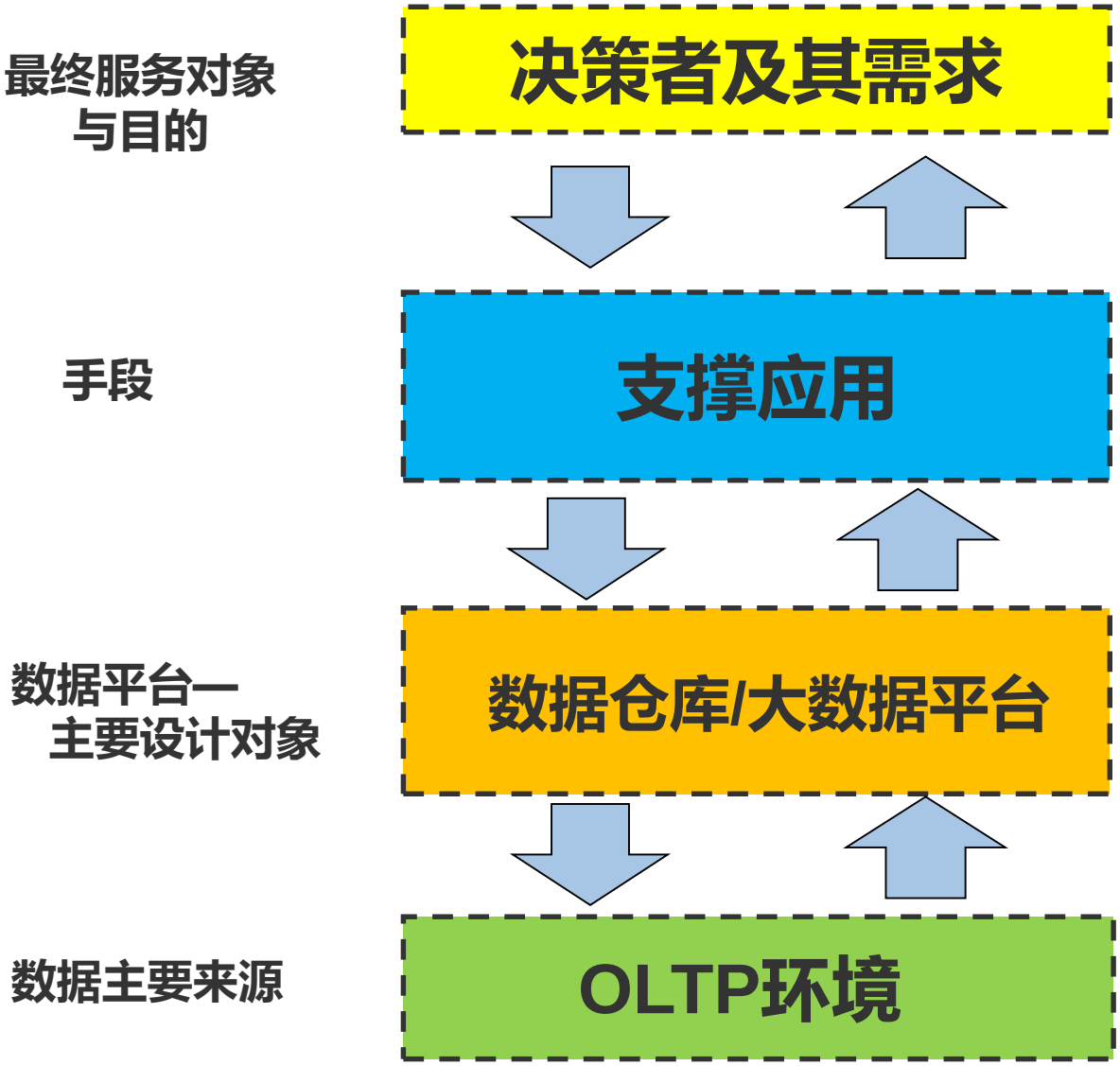
数据仓库与大数据工程

Data Warehouse and Big Data Engineering

第3部分 数据利用需求与系统架构

北京交通大学计算机与信息技术学院

为了谁？需要什么样的架构



不忘初心
方得始终

如何找需求

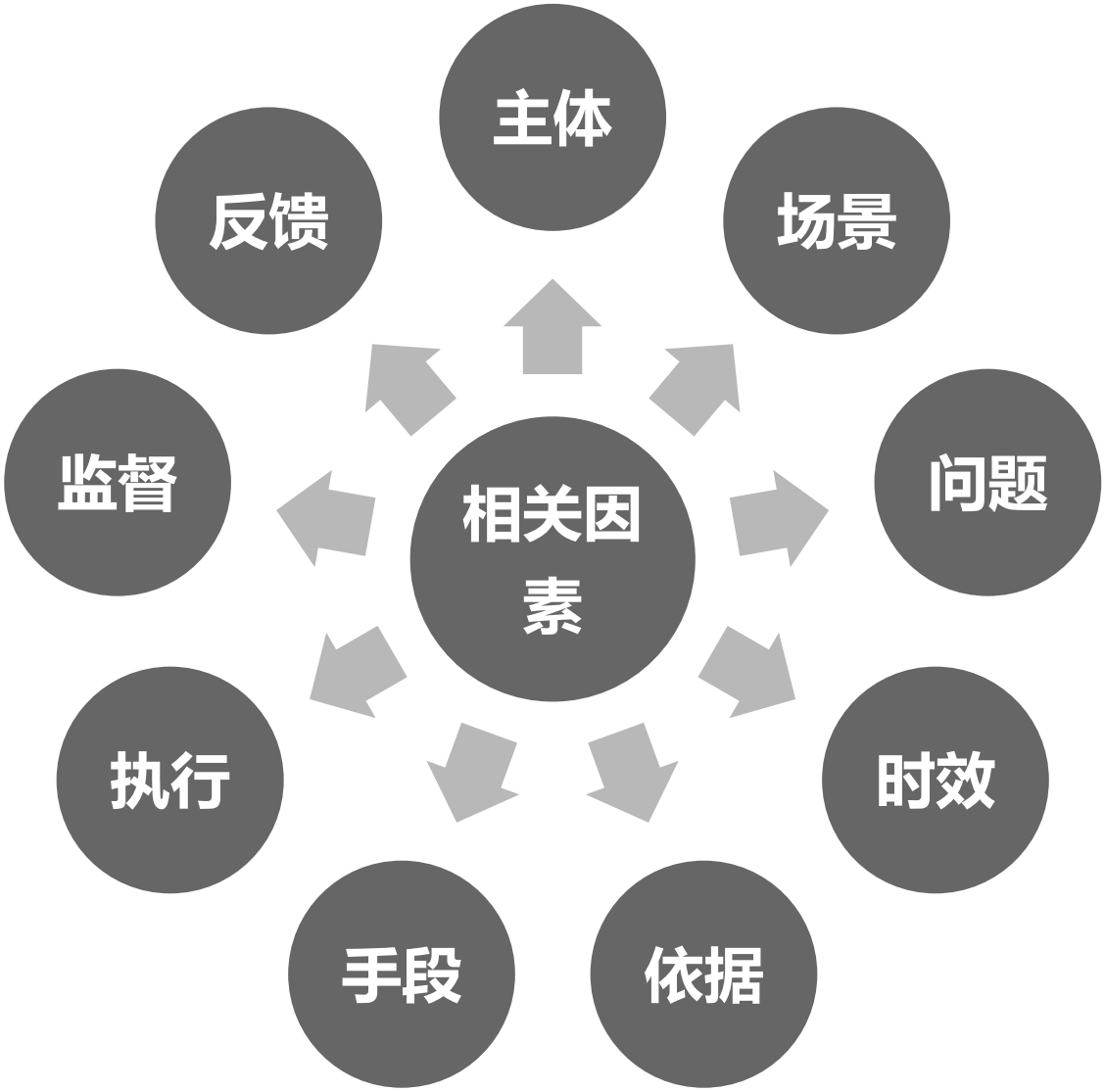
不同的需求

不同的架构

内容提纲

- 决策者的类别与决策需求
- 数据利用需求分类
- 不同应用系统架构范式
- 需求分析方法
- 不同类别应用对数据的要求
- 系统运行环境需求

1. 决策的相关因素



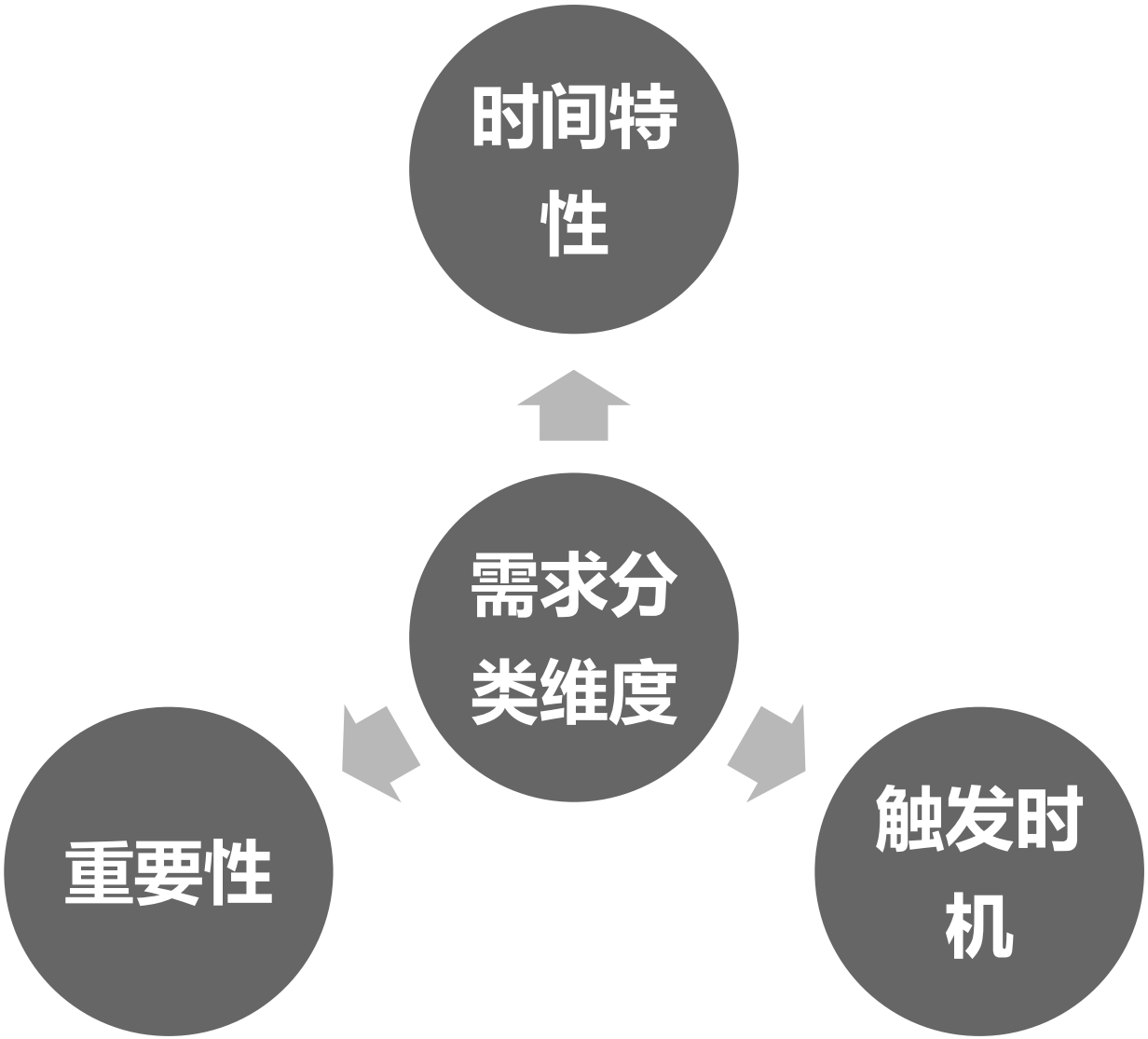
2. 回忆决策主体分类

- 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. 场景与问题—决策需求

- 主体在做决策的时候所处的大场景、状态与问题
  - 普通人上网准备看新闻 → **看什么**?
  - 学生下课后饿了 → **去哪个食堂或是叫外卖**?
  - 女主人在购买平台为孩子买衣服 → **买什么样的衣服**
  - 出行前的旅行计划：机票、火车票 → **日期、班次、车次**
  - 正在下棋中 → **走哪里**?
  - 正在开车中 → **开多快，变道，走哪儿**?
  - 招生过程划分数中 → **划在哪里**?
  - 京张高铁规划 → **首站选在哪里**?
  - 首都机场太繁忙，延误太多 → **是否要建新机场**?
  - 推荐算法发现有用户登录 → **推荐什么内容**?
  - …

4. 决策需求分类



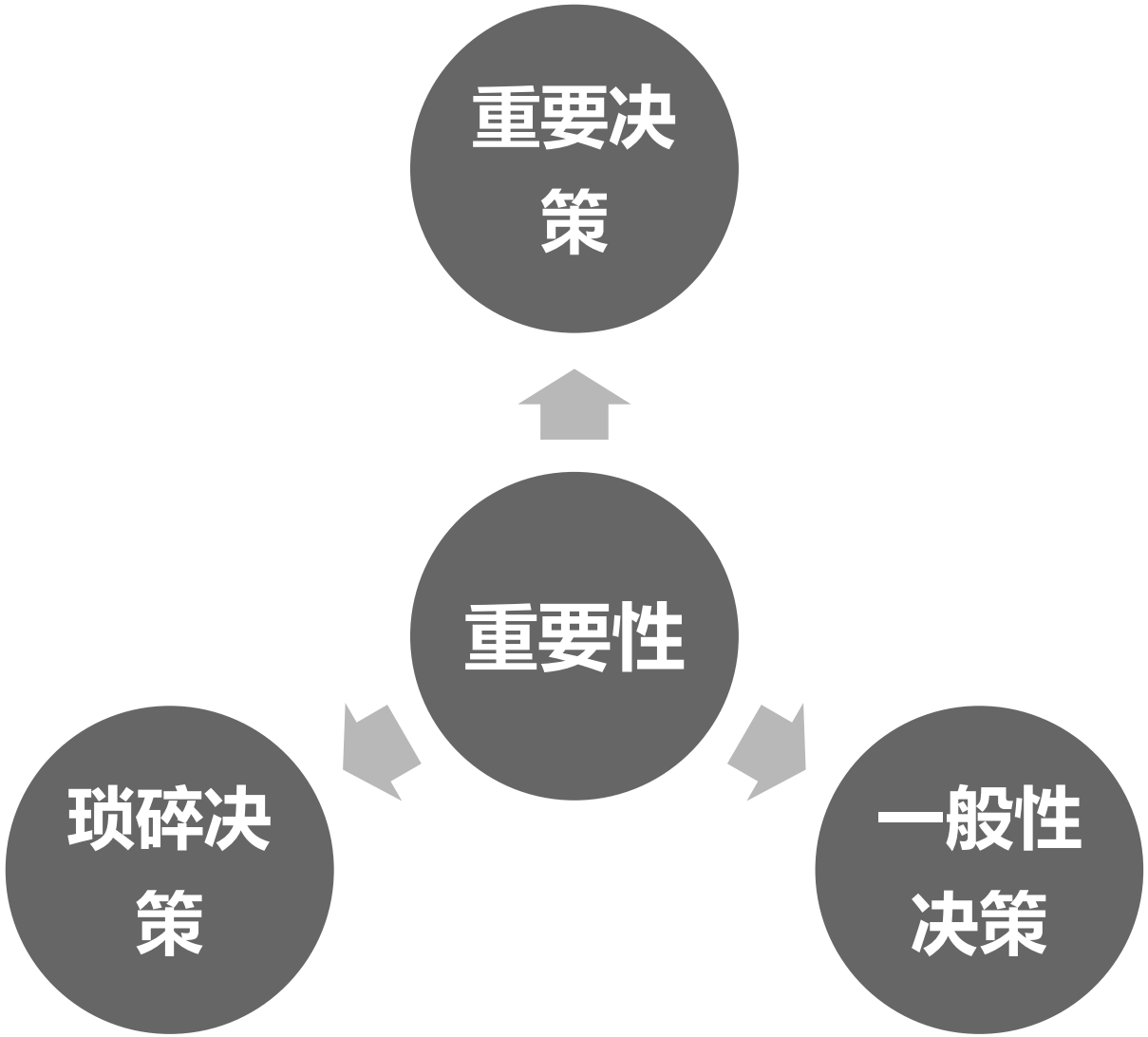
（1）决策需求时效—时间特点

- 决策有时间限制—在限定时间内完成决策
  - 实时（即时）、紧急、短时、中程、长时
- 无明确时间限制
- 决策需求时效性的相关因素
  - 时效要求的不同，对决策支持及系统的要求大不相同
  - 时效性要求一般与决策需求的重要性与紧急程度有关
  - 时效性还与涉及的历史数据、状态的时长有关
  - …

(2) 决策的触发时机

- 例行性决策
  - 周期性决策—频率问题：高频、低频
  - 非周期性决策：例行性，有一定频率
- 偶发性决策
  - 偶尔进行的决策
- 事件触发的决策
  - 由相关事件触发而驱动的决策，如用户操作

(3) 决策的重要性



4. 决策者行为模式

- 时间特点
  - 紧急决策、快速决策、沉思后决策、反复掂量、放弃决策
- 决策依据
  - 基础数据、加工后的信息、脑中知识、策略、历史案例、仿真推演
- 一般决策模式
  - 感知**状态**、利用**知识**、根据**策略**、做出**选择**、实施**行动**

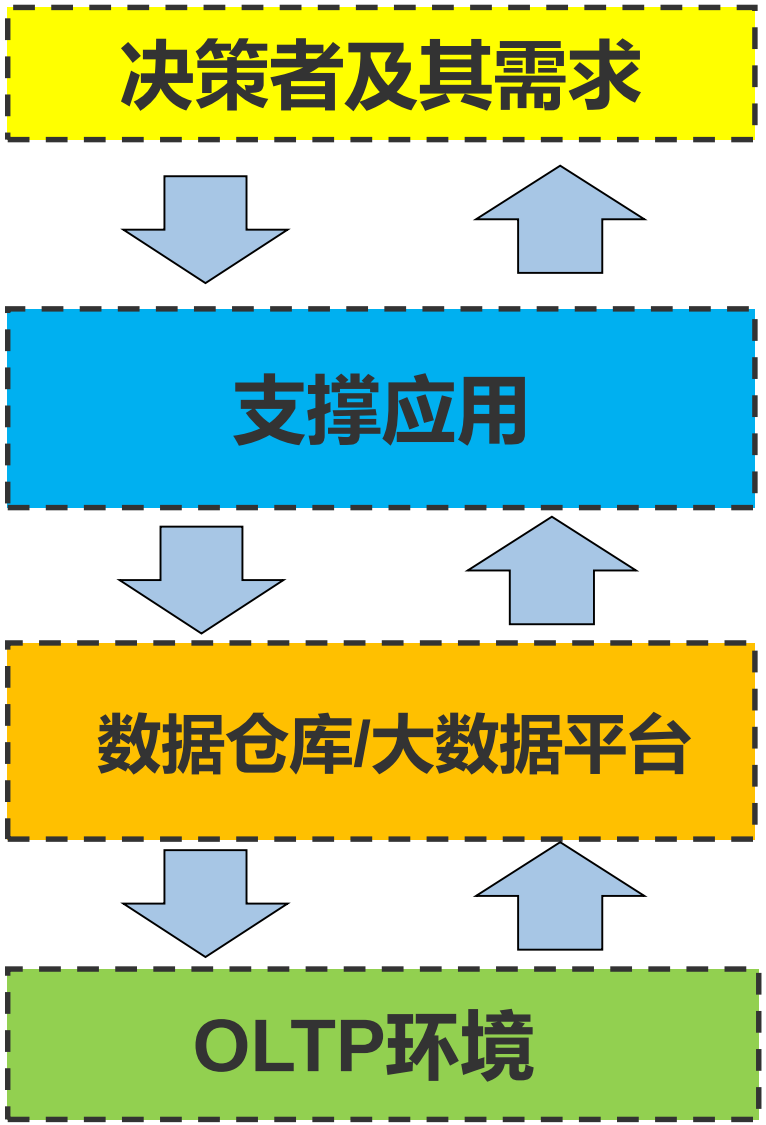
为了谁？需要什么样的架构

最终服务对象与目的

手段

数据平台—主要设计对象

数据主要来源



状态、知识、策略源自哪里？

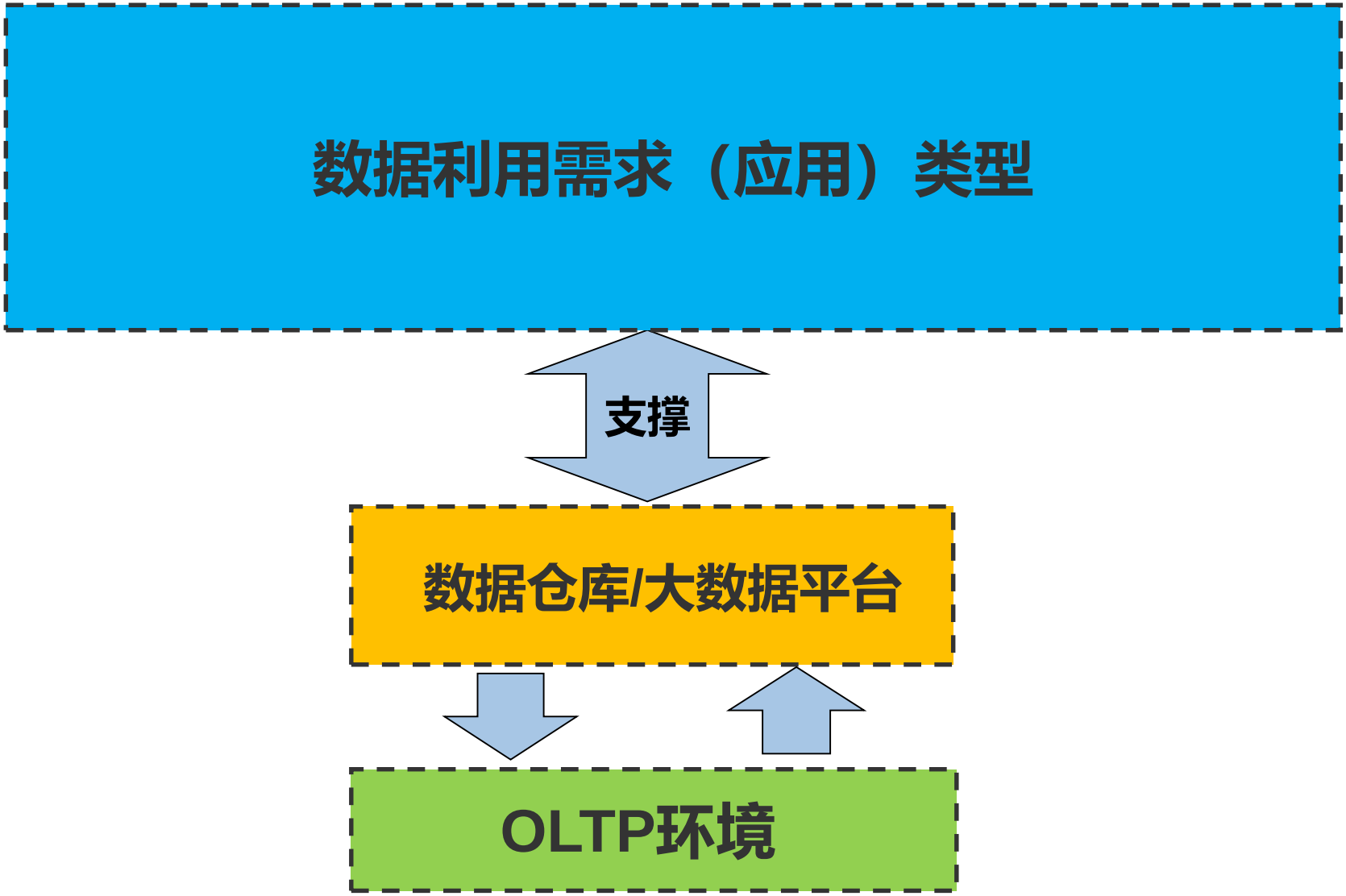
对后端平台而言，决策需求会产生数据利用需求

当前与历史数据是主要来源

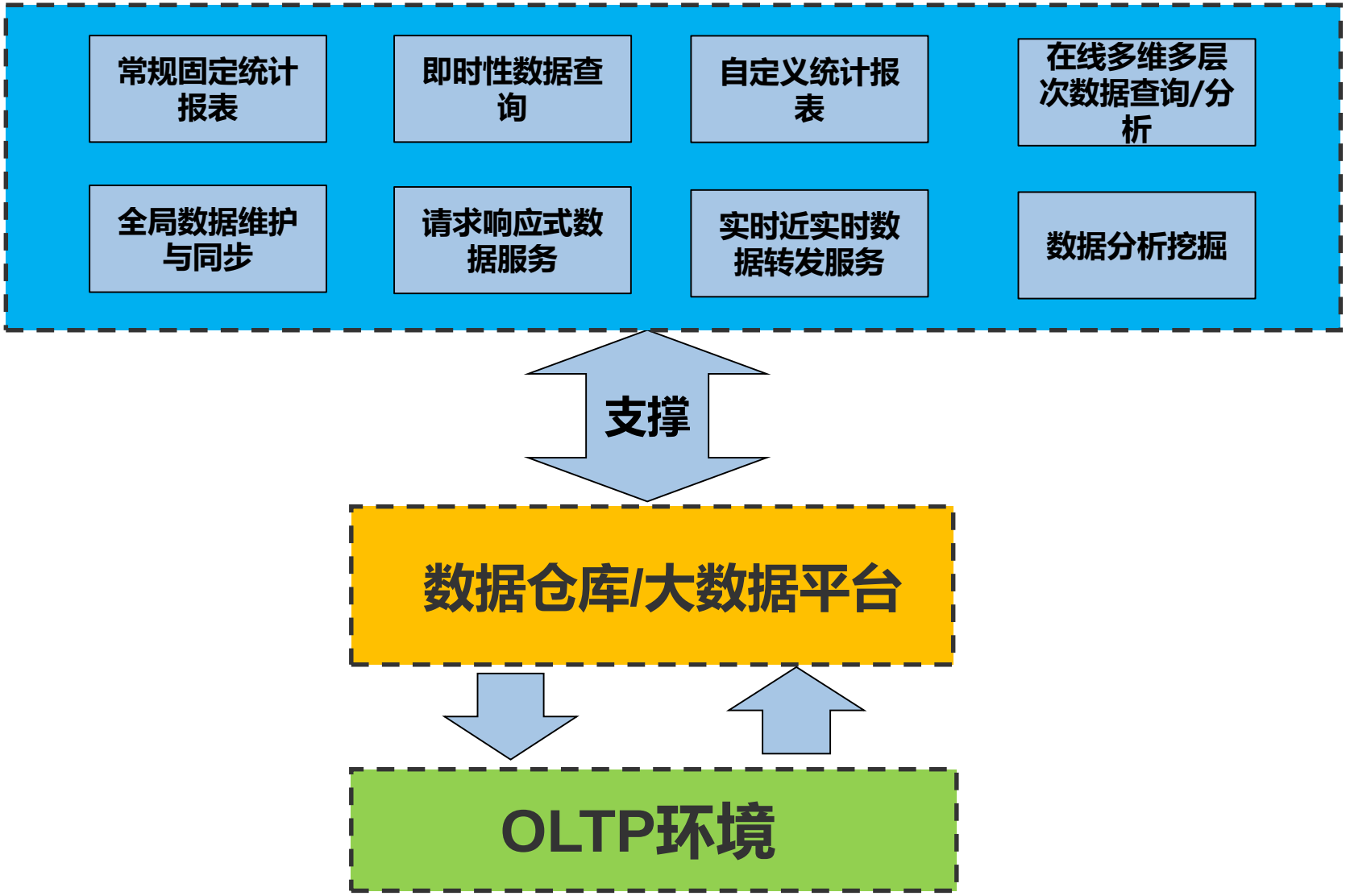
内容提纲

- 决策者的类别与决策需求
- **数据利用需求分类**
- 不同应用系统架构范式
- 需求分析方法
- 不同类别应用对数据的要求
- 系统运行环境需求

1. 数据利用需求（应用）类别示意



2. 数据利用需求（应用）类别示意



(1) 常规固定统计报表

- 或称预定义报表，预定义是指由IT工程师或业务分析人员根据最终用户的决策支持需求预先定义好一批常规性的报表。
  - 最为常见的基于数据仓库/大数据平台的决策支持类应用。在建立数据仓库/大数据平台之前，许多此类报表在现实业务中已经存在，由线下计算实现或由其他分散的系统支撑实现。
  - 在数据仓库/大数据平台建成之后，改由新平台来支撑实现，**实现时间或经济成本相应变低**。
  - 而有些报表则是建立数据仓库/大数据平台之后，因数据环境更好而新提出实现的数据视图。

(2) 即时性数据查询

- 提供即时查询的内容与业务场景有关，但是其共同的特性就是即时性
  - 即决策者希望决策环境能实时或近实时返回查询结果。
- 服务的决策者
  - 即可以是企业的各层级管理人员
  - 也可以企业外的决策者，如旅客或货主。
- 平台能否提供即时的数据查询服务是决策支持环境建设时需要回答的重要问题

(3) 自定义统计报表

- 自定义报表应用对应于**偶发式的、预定义报表无法支持**的决策支持信息需求
  - 由报表定义人员根据**当前的具体决策**支持信息需求，使用**报表交互定义工具**，以可视化交互的方式定义出报表逻辑结构与查询条件，形成一个动态产生的用户数据视图，提交给给支撑平台执行，返回查询结果。
- 在这个场景中，报表定义人员有两种
  - 最常见的是对主题数据模型较为熟悉的**业务专家**，业务专家得到结果后再反馈对决策者。
  - 也可以是熟悉数据模型和工具使用的**决策者本身**。

(4) 在线多维多层次数据查询/分析

- 这类需求对应于人们在**决策过程中不断变化**的**实时/近实时**数据支持需求
  - 典型应用模式—OLAP(联机分析处理、在线分析处理)
- **在线**：点体现了用户对查询或分析的性能要求，即要求分析尽快完成，将结果反馈给用户
- **多维**：数据视图希望在不同维度切换
  - 时间、空间、产品、…
- **多层次**：数据视图中每一个维度可能需要在不同的层次切换
  - 天—月—季，省—地区—城市—县域，…

（5）全局数据维护/同步

- 全局数据
  - 指企业或组织机构多个业务部门或系统都涉及到的数据。
  - 合理有效地做好**全局数据的维护**具有重要的意义
- 全局数据维护/同步
  - 涉及企业级OLTP业务或全局型事务处理业务
- 这种业务涉及数据修改、更新，涉及多个系统，需要保存**最新的、一致的业务数据**，特别是主数据。
- 数据仓库/大数据平台的**数据集成性**，为建立全局数据维护/同步的**奠定很好的基础**。

(6) 请求响应式数据服务

- 由**数据需求方驱动**的服务过程
- 平台收到请求后，需要审核服务请求的合法性，审核通过后调用相关预定义服务功能为请求者提供服务，并返回服务结果。
- 根据结果内容规模的差异，请求响应式服务可以分成
  - **批量式数据请求服务**
  - **单条数据请求服务**。

(6) 请求响应式数据服务

- **批量**数据服务
  - 经常以类似ftp的文件服务形式提供，需求方与服务主体间一般只有松散或偶发的业务协作需要，业务上只会产生偶发性批量数据服务请求。
- 系统间的**单条数据**请求响应式服务
  - 一般发生于**数据验证式或状态判断式**应用场景。
  - 行业数据仓库/大数据平台一般会有全局性业务数据，可以对企业内部或外部系统提供数据验证服务
  - 此类服务目前也越来越多。常见一个经典案例就是现在许多的行业大数据平台提供的针对**公民的信用、资质等**提供实时查询服务。

(7) 实时近实时数据转发服务

- 数据仓库/大数据平台提供一种常规性数据服务，一般发生**业务上有紧密关联的单位**之间。
  - 我国铁路总公司级的各数据仓库/大数据平台与铁路局相应业务平台之间的实时/近实时数据转发服务。
  - 在民航业中，中国航信数据服务部运营的数据仓库/大数据平台也需要向各个航空公司系统提供类似服务。
- 在早期，这种业务上紧密合作的单位之间的服务一般是以**天为单位**提供数据转发服务。然而，随着用户需求与服务能力的提升，这种服务的提供在时效上越来越**向近实时甚至是实时的**服务方式发展演变。

(8) 数据分析挖掘需求

- 快速数据扫描分析
- 近实时分析、挖掘、学习
- 长期趋势性分析、挖掘、学习
- 科学实验

其它需求

- 需要支持哪些应用主题，建设顺序
- 性能需求、一致性需求
- 数据保存期需求
- 这些需求必定要用于指导数据仓库和大数据平台的数据设计
  - 如粒度设计，面向应用的查询优化设计，数据分区，数据维护周期，数据集成，数据清理，数据间一致性处理等，这些方面都与用户的需求有关。

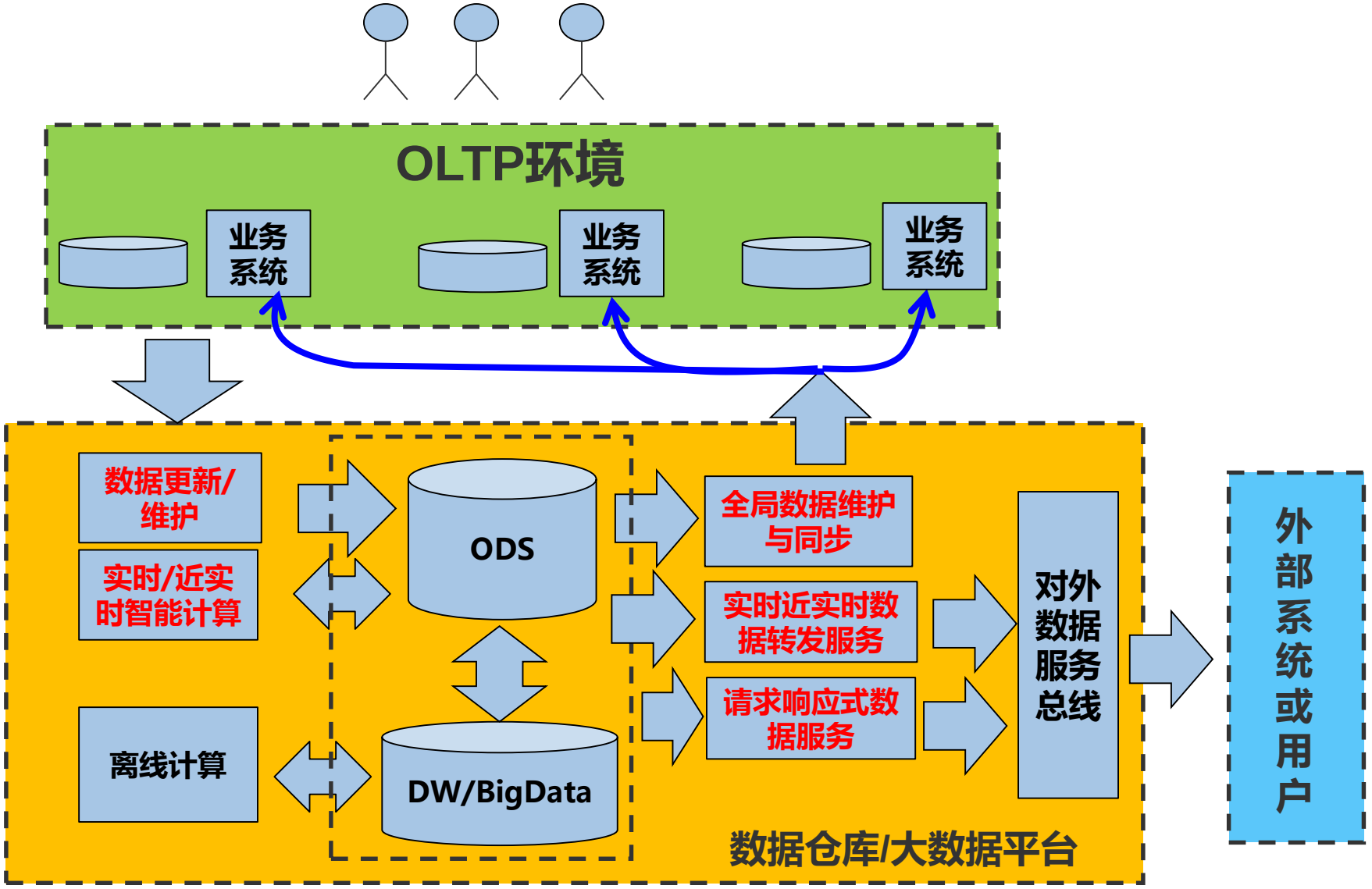
关于需求类别小结

- 对于数据仓库和大数据平台，必然存在用户的需求。同时，根据用户的需求，得到各个子系统的设计需求，才有可能实现整个系统。
- 在数据应用过程，可以把握应用的数据需求，从而动态调整主题设计。
- 即数据需求的变化，产生了主题数据调整的需求。

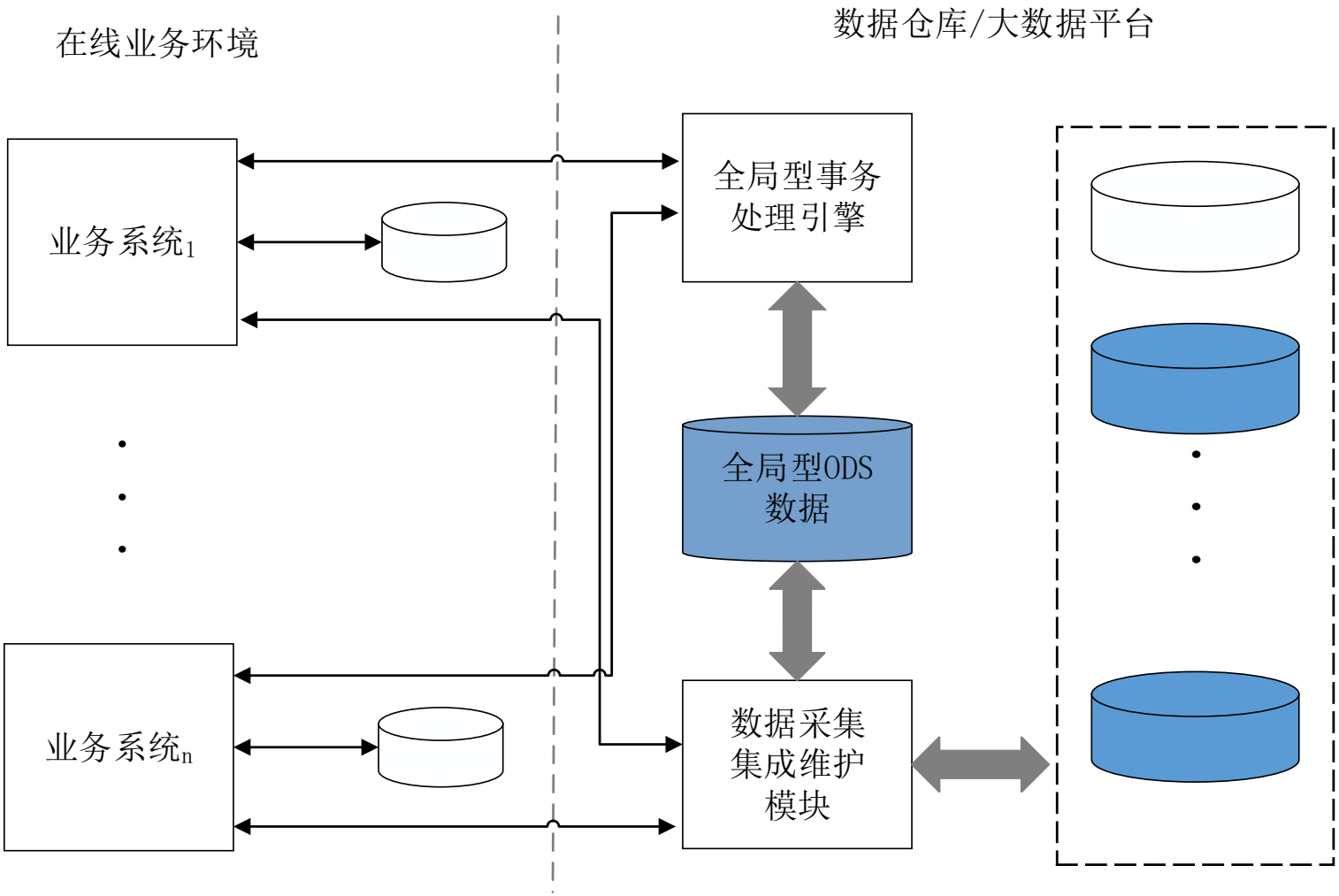
内容提纲

- 决策者的类别与决策需求
- 数据利用需求分类
- **不同应用系统架构范式**
- 需求分析方法
- 不同类别应用对数据的要求
- 系统运行环境需求

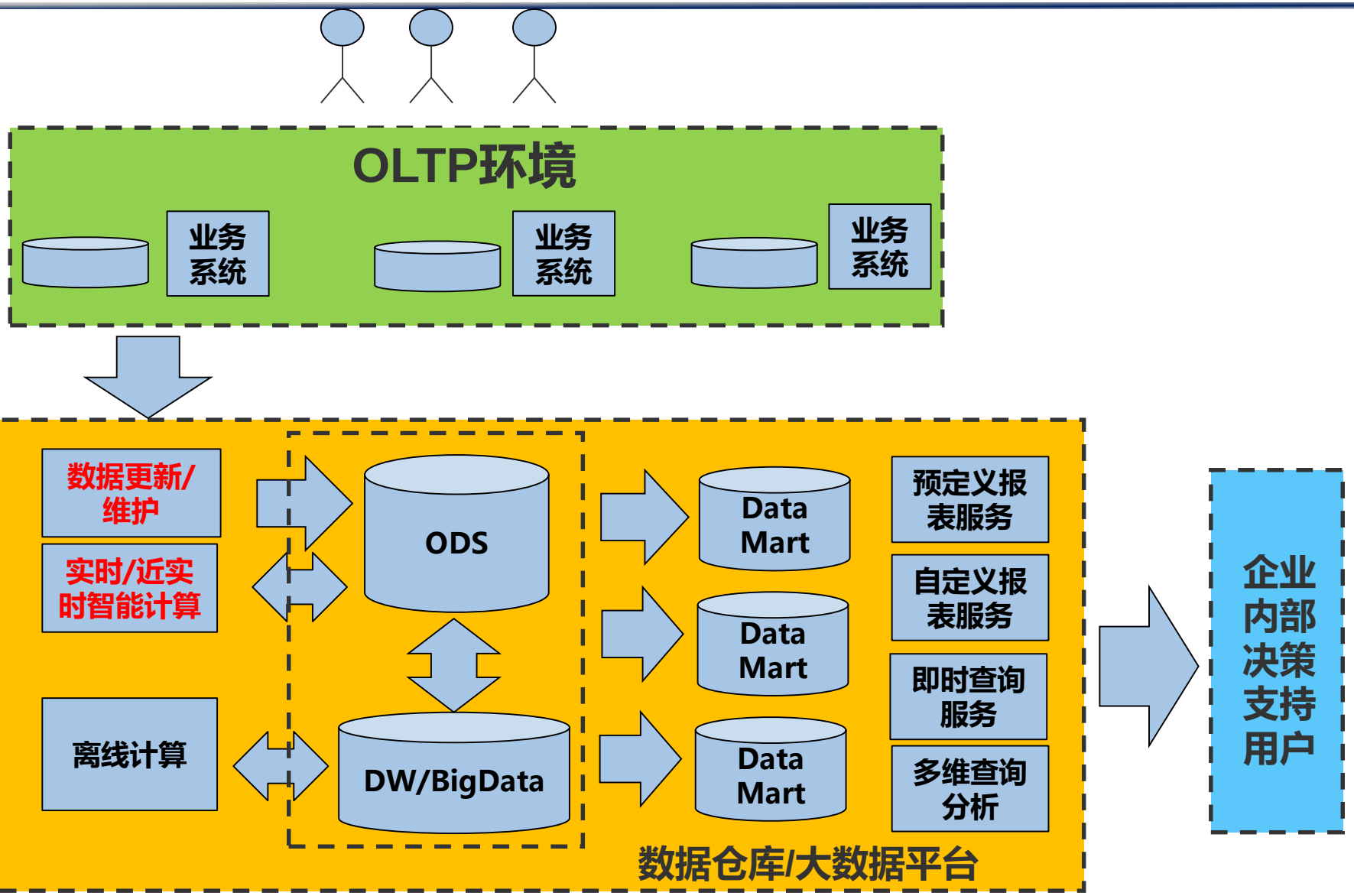
1. 对外数据服务类架构


OLTP环境
业务系统
业务系统
业务系统
数据更新/维护
实时/近实时智能计算
离线计算
ODS
DW/BigData
全局数据维护与同步
实时近实时数据转发服务
请求响应式数据服务
对外数据服务总线
数据仓库/大数据平台
外部系统或用户

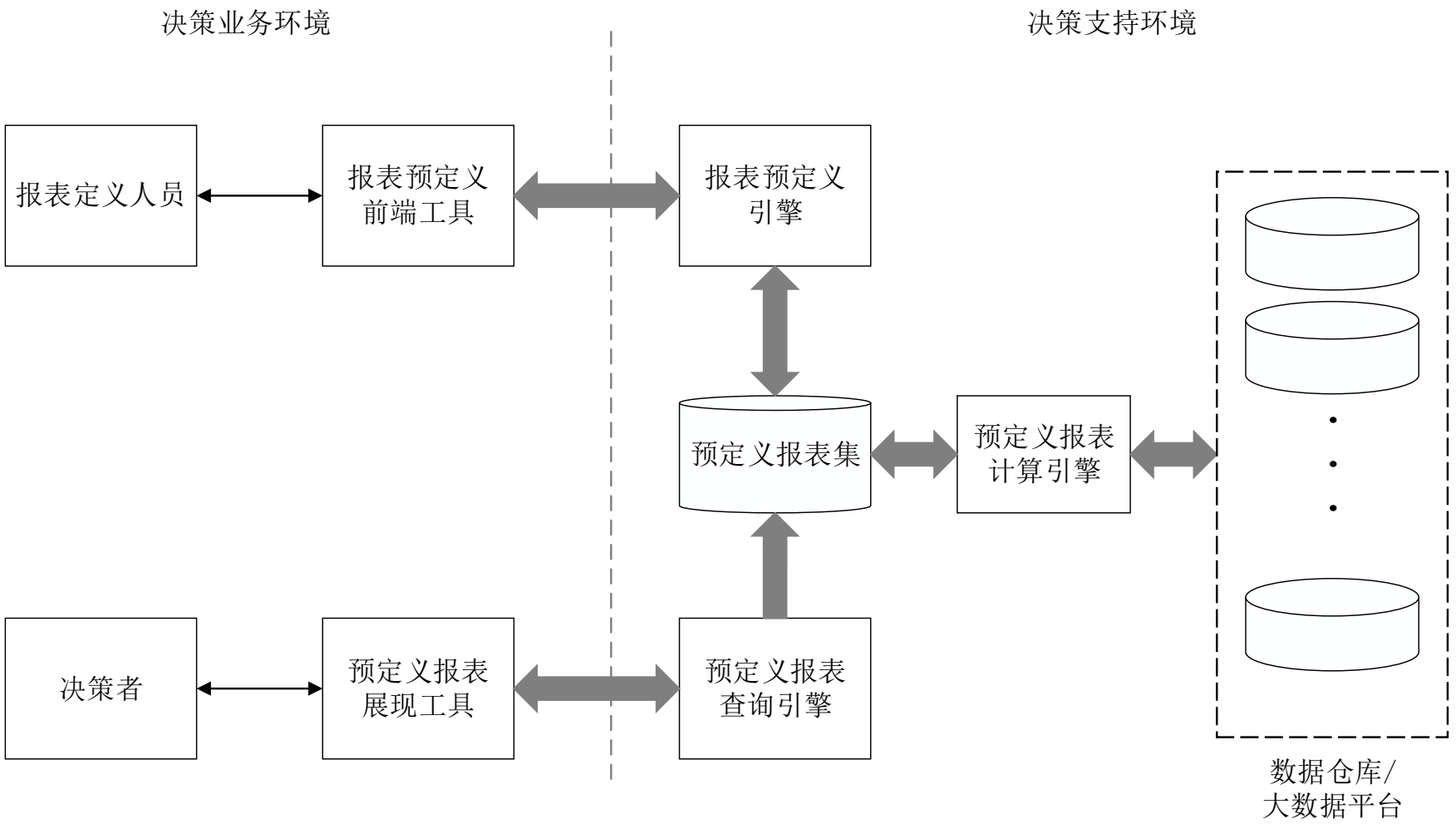
2. 全局型事务处理：数据维护、同步



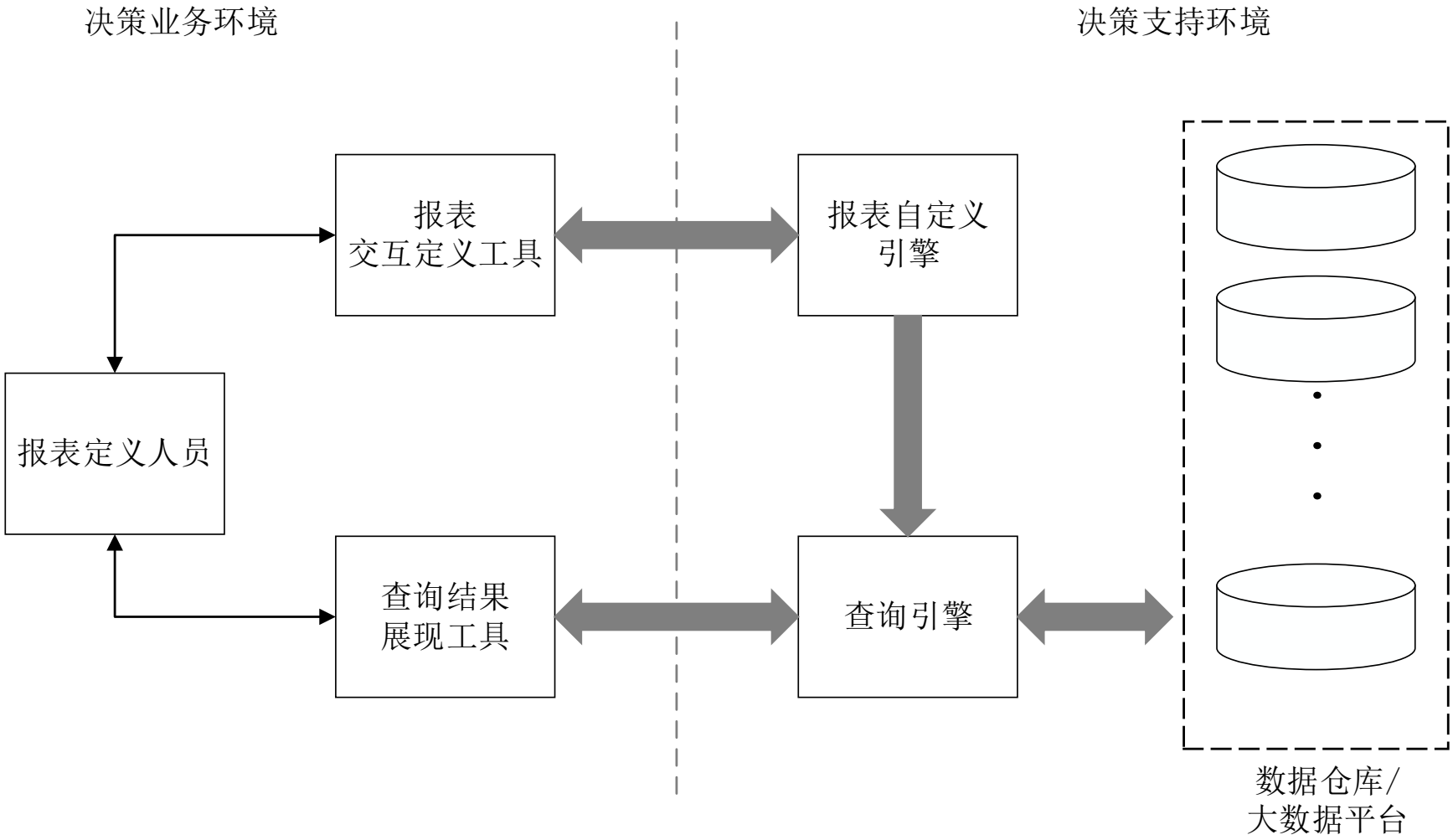
3. 传统的面向企业内部决策支持服务类系统架构



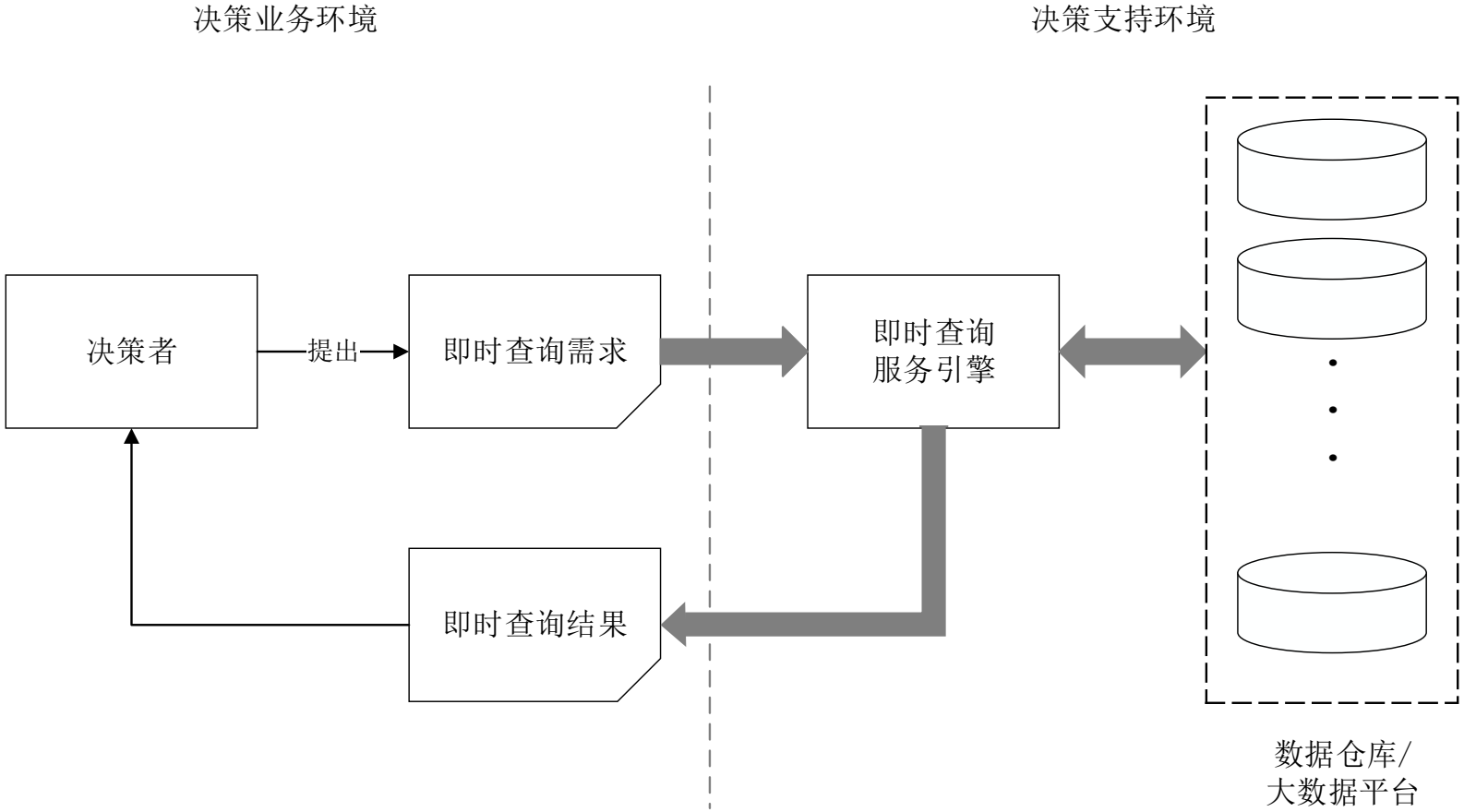
(1) 预定义报表服务实现架构



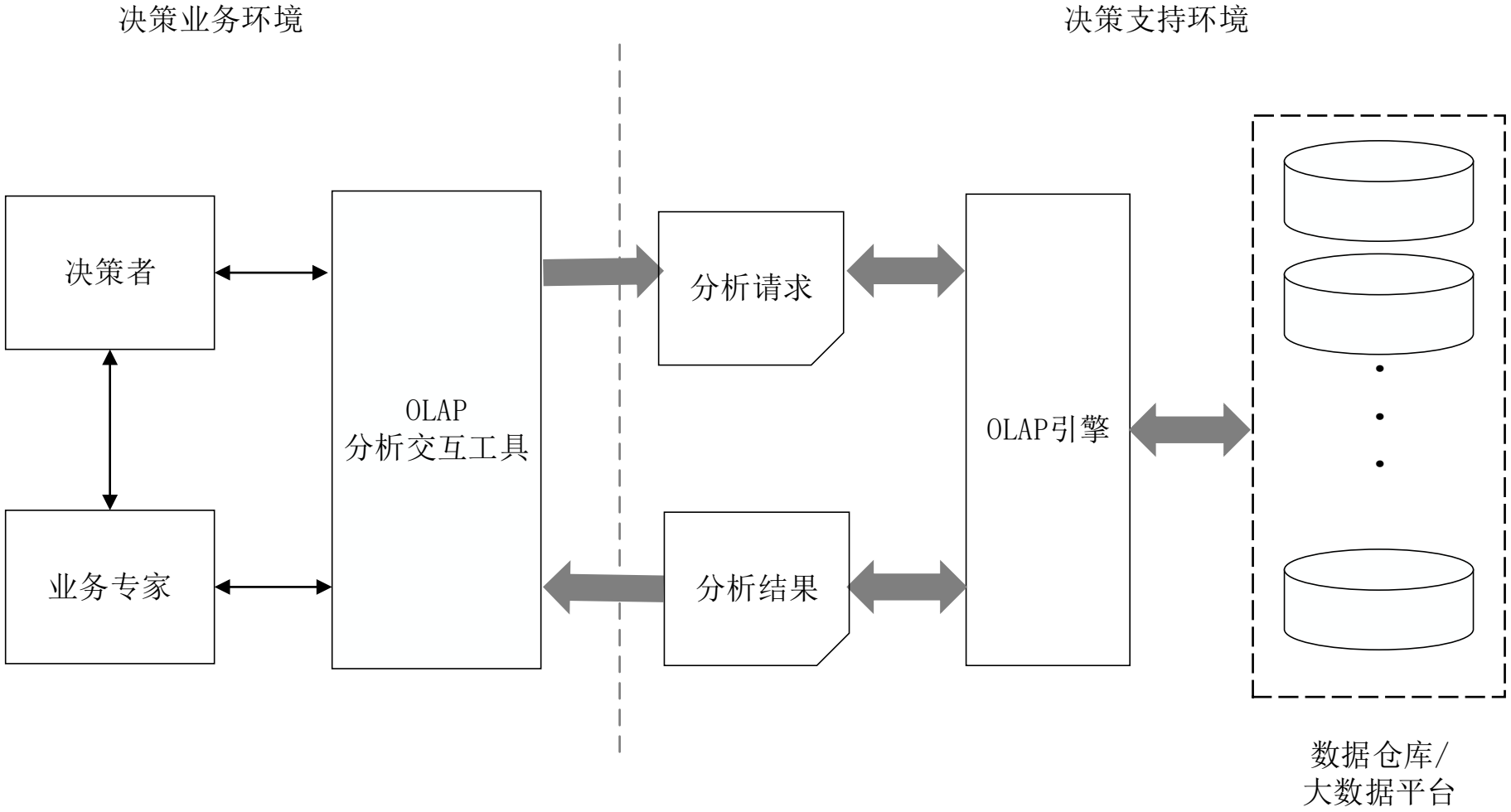
(2) 自定义报表服务实现架构



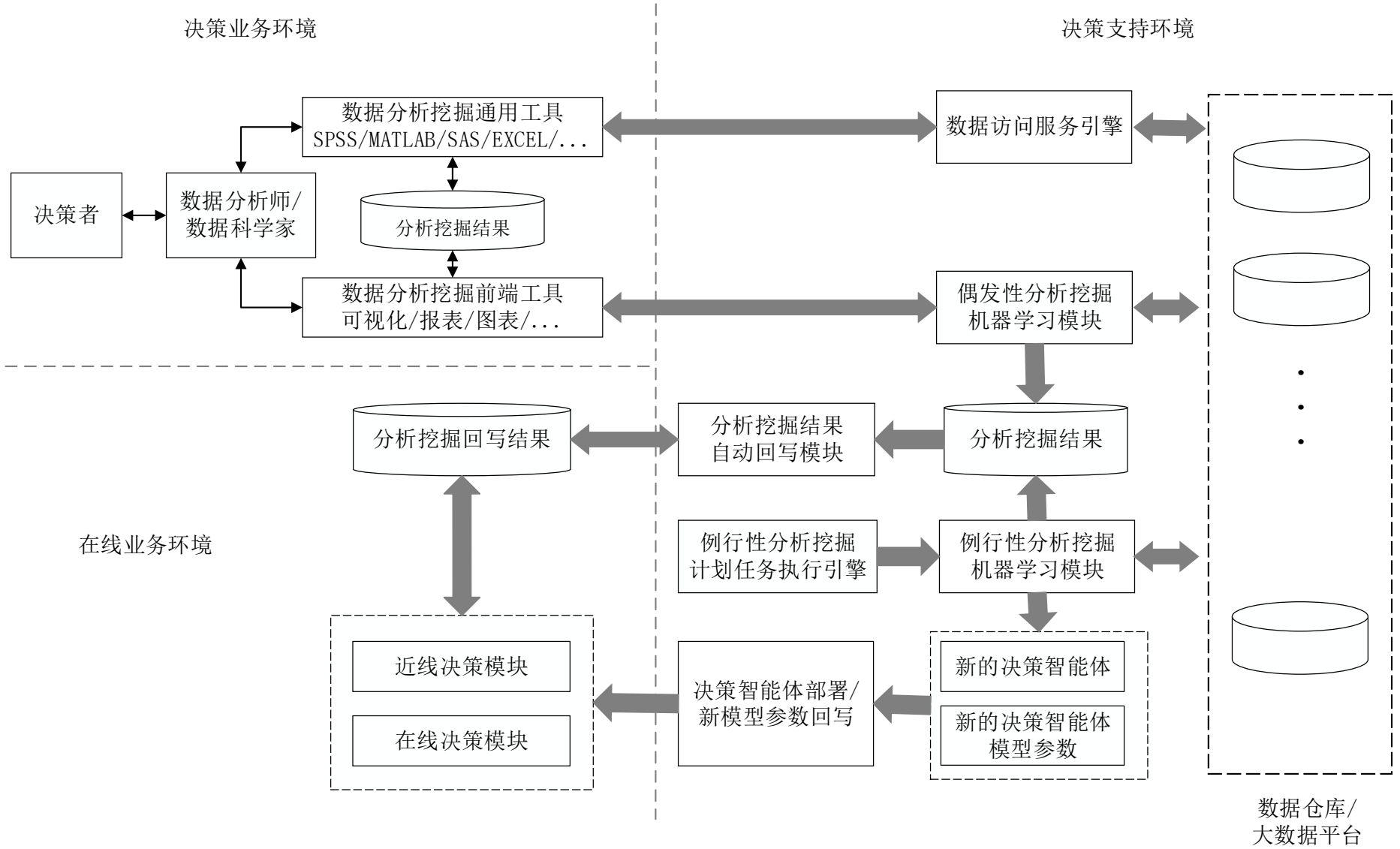
(3) 即时查询服务实现架构



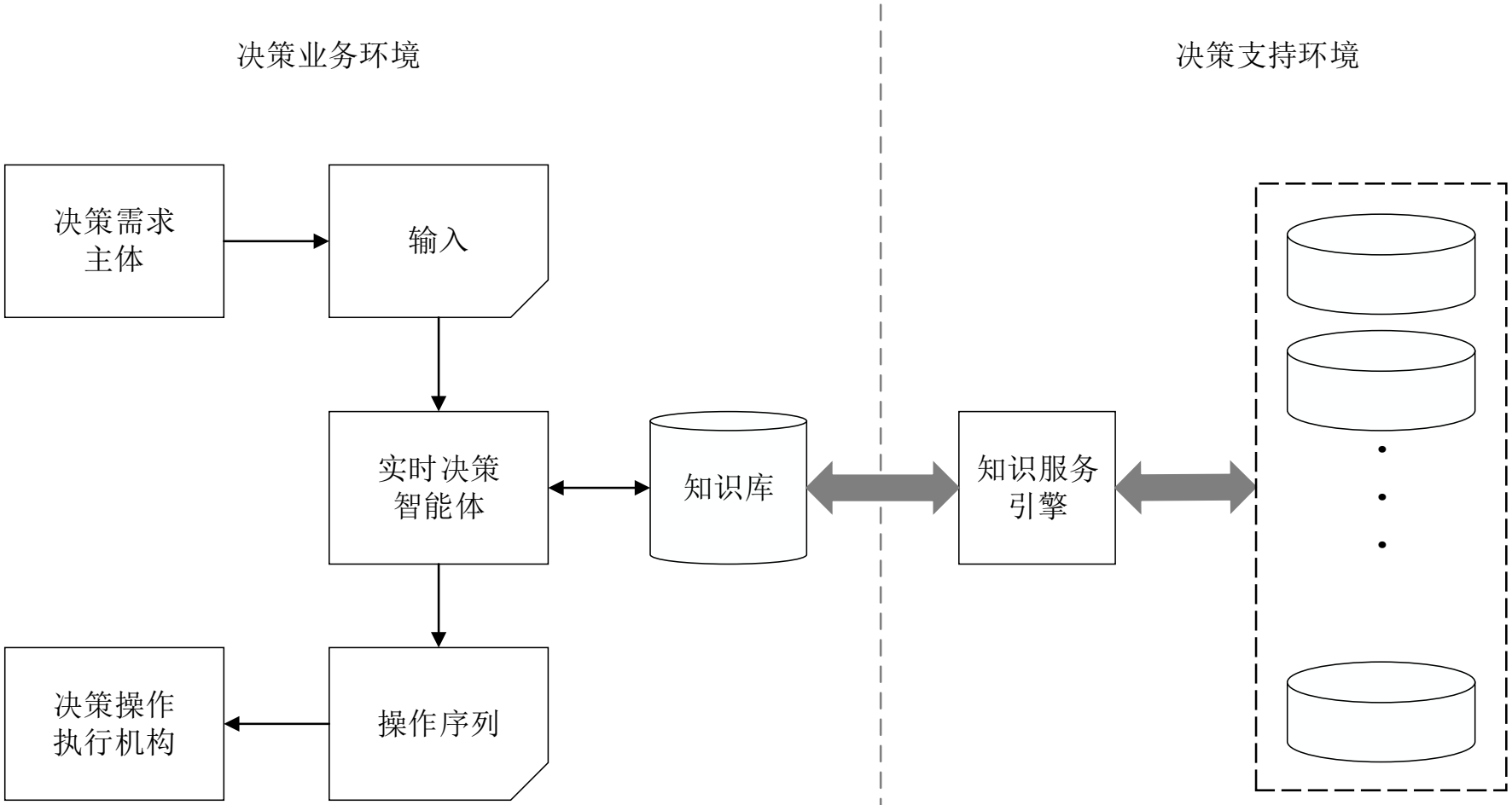
(4) 多维查询/分析处理实现架构



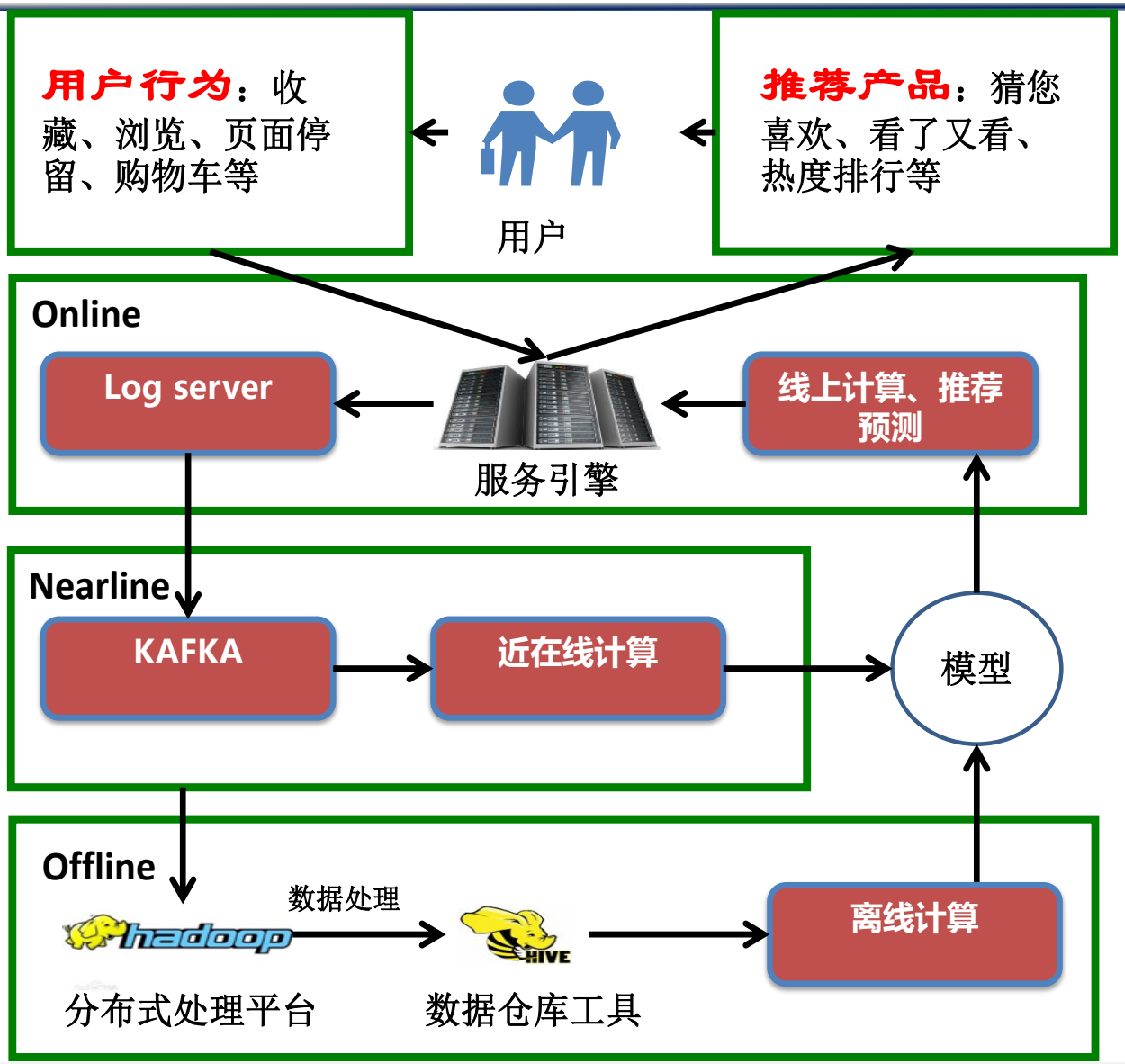
4. 数据挖掘类应用实现架构



5. 实时自动决策服务应用架构



某企业近实时自动决策支持系统架构



内容提纲

- 决策者的类别与决策需求
- 数据利用需求分类
- **不同应用系统架构范式**
- 需求分析方法
- 不同类别应用对数据的要求
- 系统运行环境需求

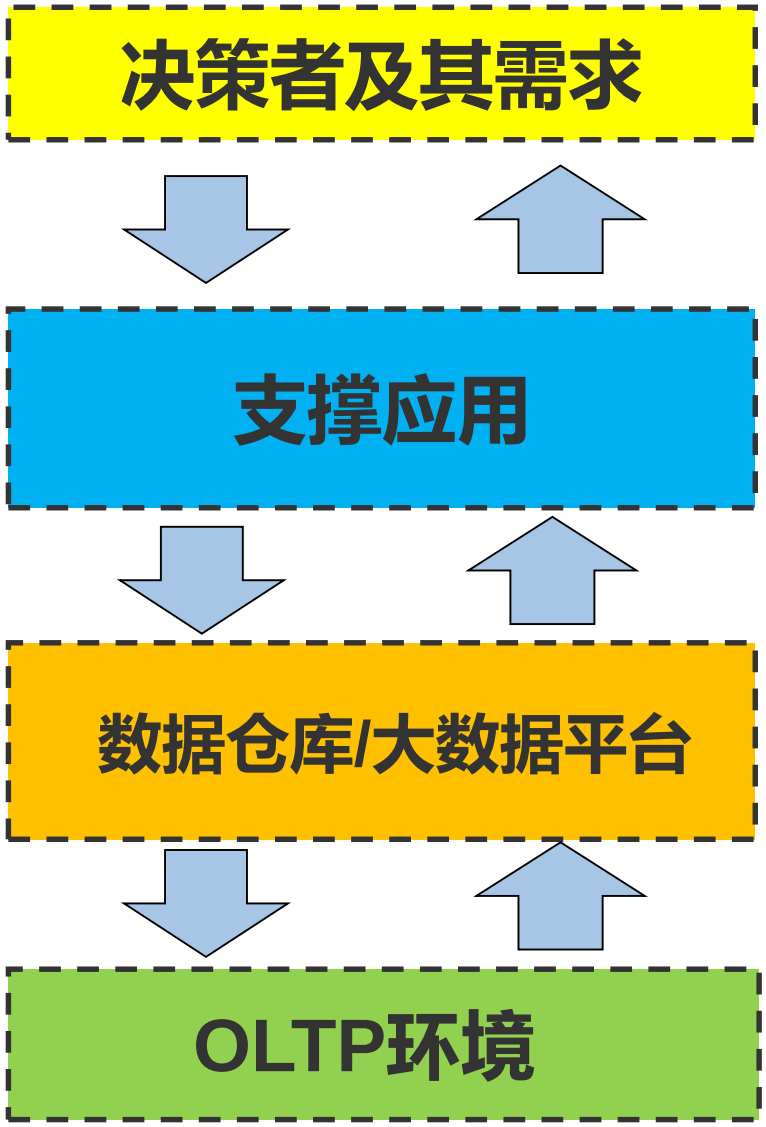
为了谁？需要什么样的架构

最终服务对象
与目的

手段

数据平台—
主要设计对象

数据主要来源



如何找需求，
如何分析并确
定需求？

企业和组织机构领导非常头痛的问题

- 我们单位的数据仓库或大数据平台该如何建
  - 为了什么而建
    - 支撑什么应用
    - 应用场景是什么
  - 我的平台里装什么数据
  - 要建多大的平台
  - 要投入多少钱
  - 能有收获没有

系列经典的问题

1. 数据仓库与大数据平台的需求分析方法

- 设计需求是指用以指导系统设计的明确的建设要求
- 需求分析方法
  - 根据企业的战略，如何利用**各种资源**，收集或发现的建设需求，对需求进行分析并最终明确出建设需求的方法
- 可用资源
  - 内外部专家、企业员工、需求分析人员
  - 企业战略目标
  - 现有系统设计文档
  - 资金
  - 现有技术条件
  - …

2. OLTP系统建设需求特点

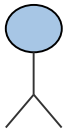
- OLTP数据库设计需求特点
  - 有确定的业务场景下的一组具体应用需求
  - 需求能以明确的物流 、数据流和数据处理流形式表示
  - 这些需求成为数据库系统设计和开发的出发点和基础
- OLTP系统的设计需求收集分析相对容易，方法也成熟
  - 一般都具有非常明确的应用场景
  - 目标明确，需求相对来说比较容易辨识与定义出来

3. 规划与需求分析的艰巨性

- 与OLTP系统相比，行业数据仓库或大数据平台规划设计过程并不简单，要做好相关的设计与过程管理对设计人员的要求很高。
- 主要原因
  - 系统应用需求收集比较困难
  - 涉及的数据模型范围广，涉及企业或部门全局数据模型
  - 企业战略与平台的关系比较难以明确
  - 平台的建设需要具有较大的投入

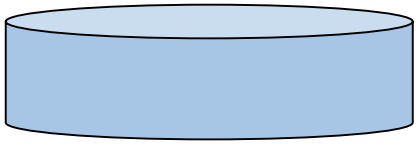
需求发现的难度

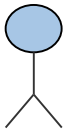
大数据的几点典型特征



高层领导

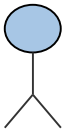
大数据战略
资金、资源
哪个点上可以做?
感觉许多地方可做，是否





部门领导

大致知道有哪些数据，有什么用
可以怎么做
需要投入多少
有问题要解决，能用什么解决



基层员工

熟悉数据，业务场景，不知道别人有什么数据
时不时有些想法
决策能力不够
有供应商介绍解决方案

4. 关于数据仓库建设需求的一种观点

- 针对数据仓库的设计需求，有一种观点认为
  - 数据的分析处理的**需求更灵活，没有固定的模式**
  - 在进行数据仓库设计之前，很难**提前准确把握用户的真实需求**
  - 因此，**不可能从用户需求出发来进行数据仓库的设计**
- 原因：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.
  - 工作于发现模式，其任务在于不断地定义和寻找**企业决策**中所需要的信息
  - **数据需求太灵活，潜在业务需求场景太多**

5. 需求与设计方法的相关观点--Inmon

- Inmon的观点
  - 数据仓库的需求只有**当部分的填入数据**以后才能知道
  - 以往适用的OLTP系统的设计方法对于数据仓库并不一定适合
  - 数据仓库**以启发式、迭代式的**方式构建
  - 数据仓库的整个生命周期内**存在反馈循环**

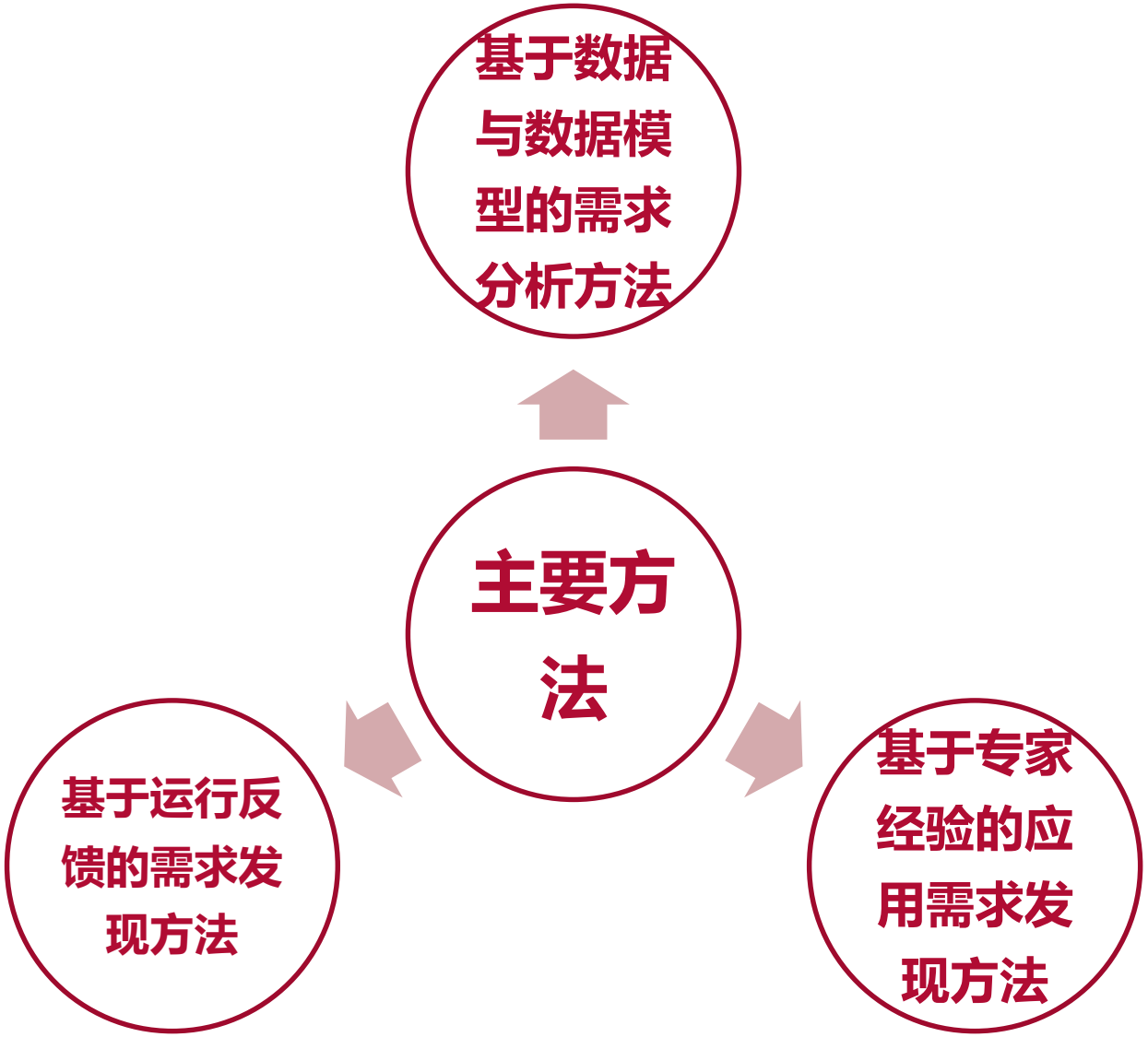
6.这种观点所带来的困惑

- 不知道用户需要什么数据，数据仓库/大数据平台**从何开始设计，凭什么去设计**
- 如何开展一个数据仓库/大数据平台建设项目，**合同标的怎么界定**
- 应用需求很可能会发生变化，如何**适应需求的变化**

这种观点所带来的困惑

- 这个说法虽然不尽合理，但也反映两层意思。
  - 首先，数据仓库或大数据平台的应用需求比较难以明确，并且需要有**层次较高、经验丰富、对企业本身战略目标、业务诉求及信息系统架构比较清楚的人**去协助规划需求；
  - 另一方面，数据仓库和大数据平台的需求集**需要不断完善和迭代**。
- Inmon的结论
  - 数据仓库不能采用传统的需求驱动的设计方法
  - 对需求进行预测仍然是必要的
  - 现实中的方法处在这两者之间

7. 需求收集分析工作的主要方法



(1) 基于数据与数据模型的需求分析方法

- 基于数据模型的需求发现方法
  - 基础工作，出发点
    - 企业信息系统架构分析、企业数据模型梳理分析为基础
  - 参与人员
    - 架构师、行业专家、系统分析师、业务人员等人员
  - 根据企业数据模型，提出潜在需求

企业数据模型的稳定性

- Enterprise Data Model
  - 反映企业的现有业务应用和数据体系的数据模型
- 企业数据模型是具有稳定性的
  - 不可能超越这个模型在企业内部获得该模型无法提供的数据，即数据具有边界
  - 企业数据模型结构基本是稳定的

(2) 基于专家经验的应用需求发现方法

- 基于专家经验的应用需求发现主要
  - 出发点：专家经验
  - 是指行业专家或咨询专家
  - 根据相类似行业或其他行业的经验
  - 结合数据现状及行业经验
  - 提出潜在需求，反复交流核实
  - 形成战略目标和项目需求

(3) 基于运行反馈的需求发现方法

- 基于运行反馈的需求发现
  - 是指在前期平台建设完成之后
  - 在运行过程中发现新的需求。
    - 存在的问题，完善
    - 新的需求，不断丰富

7. 大数据平台需求工程咨询项目

- 为了有效开展需求工程，大型数据仓库与大数据平台的需求工程有时需要企业启动一些企业内部需求工程项目，甚至启动一些咨询项目协助企业开展
  - 系统规划
  - 企业数据模型整理分析
  - 需求发现
  - 需求建议
  - 需求整理
  - 最终形成咨询报告。

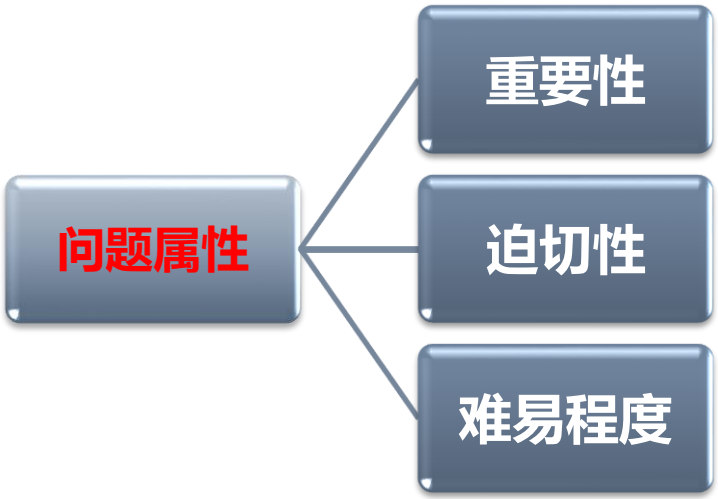
8. 需求整理

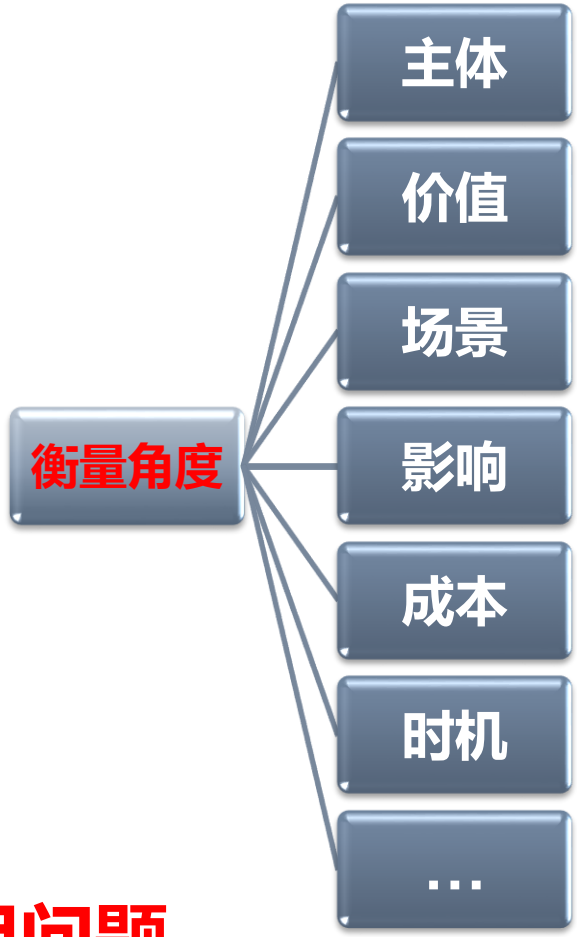
- 通过需求工程所收集的需求将形成一定的**需求集合**，企业应**持续维护企业决策支持应用需求**，结合企业战略目标，完善描述**平台发展蓝图**。
- 同时将需求集从适用场景、面向的决策主体、重要性、迫切性、实现难度、成本、相关数据等不同的角度进行梳理归类。
- 形成不断维护完善的**需求分类集合**，作为后续设计工作的输入。
- 问题：需求点如何选择?

应用需求选择原则

- 一定要结合具体应用场景
- 做好痛点和关键问题分析
- 对需求进行排序，确定实践战略

大数据应用成功因素：发现痛点、关键问题





做大数据的人应**学会分析应用问题**

在业务流程场景或生态圈中去寻找关键问题

大数据处理最常见的场景模式

1、面向业务系统：大数据+智能技术

正确知识+现状+智能策略实时执行

目标：信息系统的智能→安全和高效

2、面向人与组织：大数据分析挖掘支撑决策

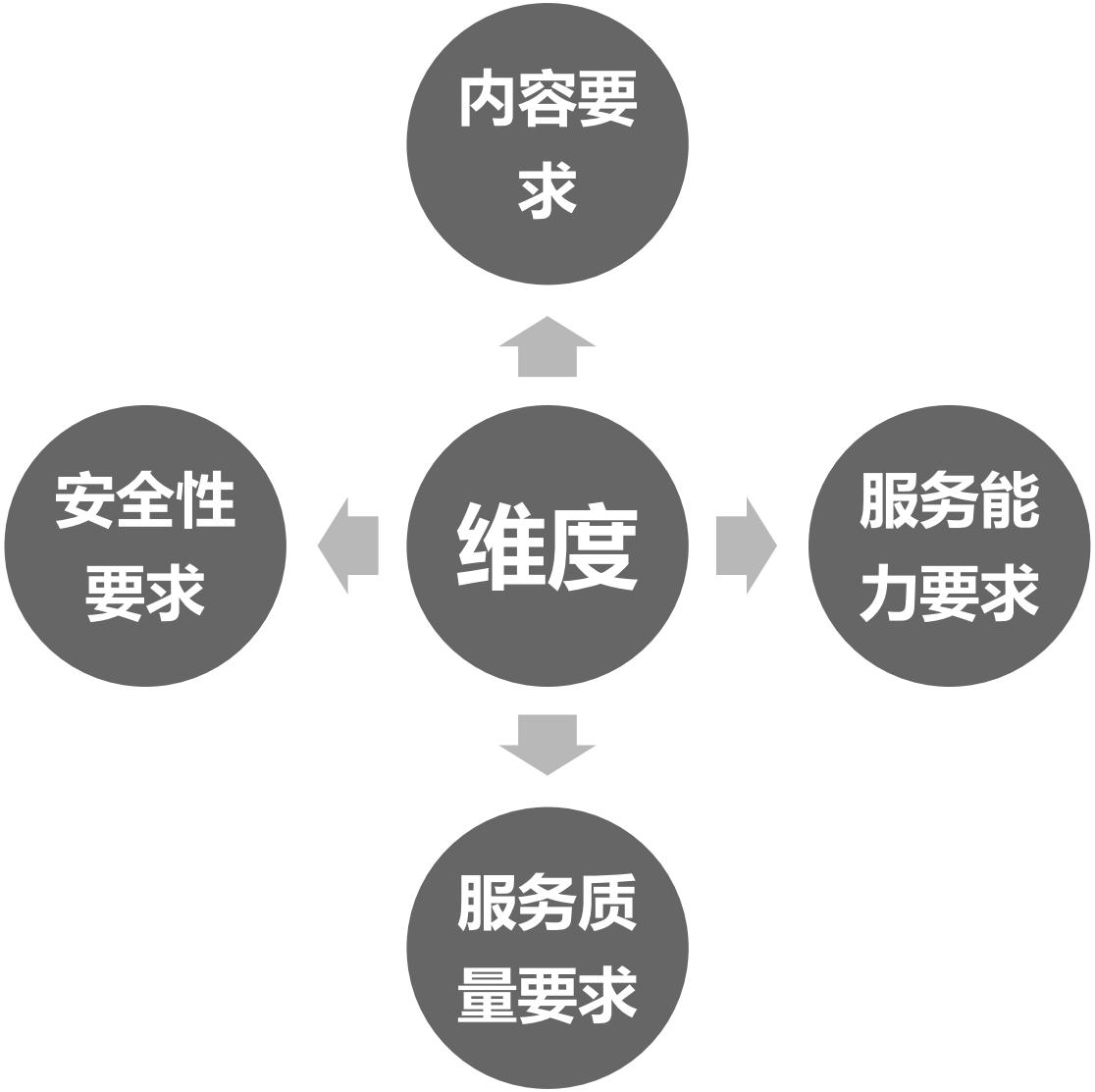
知识+现状+合理决策

目标：人与组织的智慧

内容提纲

- 决策者的类别与决策需求
- 数据利用需求分类
- 不同应用系统架构范式
- 需求分析方法
- **不同类别应用对数据的要求**
- 系统运行环境需求

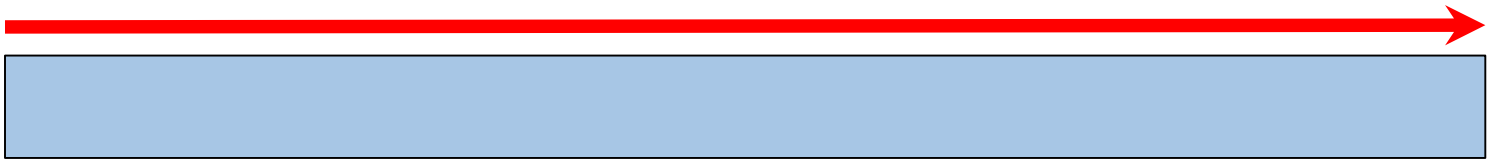
对数据平台的要求



1. 数据内容要求

- 数据时间跨度需求
  - 当前数据
  - 近期数据
  - 远期数据
- 数据集成性需求
  - 综合性应用—企业级视图
- 数据的全局一致性需求
  - 数据标准问题
  - 数据发布控制问题

（1）数据时间跨度需求



当前数据

近期数据

远期数据

对于每一种具体应用**都需要**
理清数据的时间跨度需求

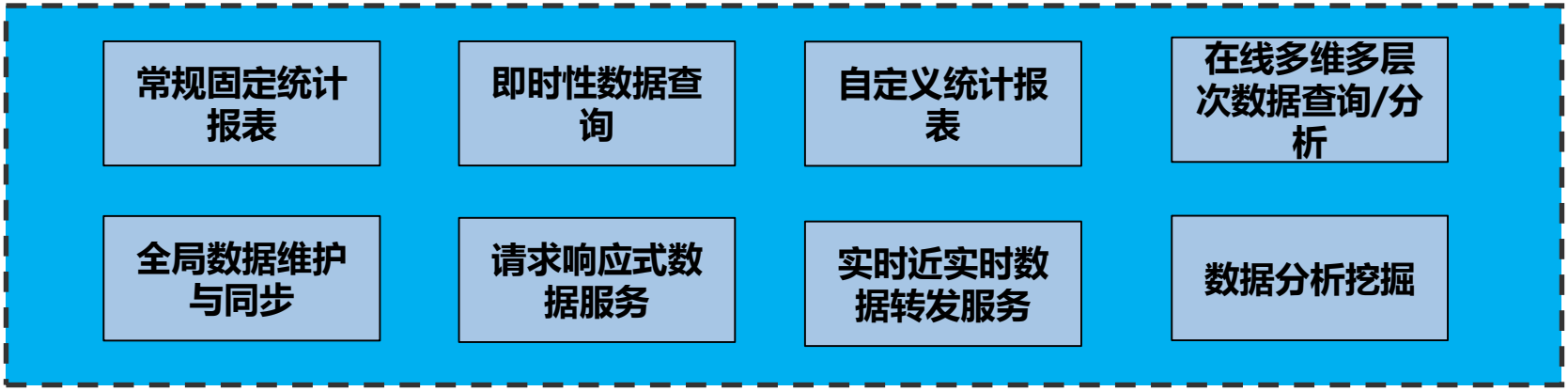
常规固定统计报表	即时性数据查询	自定义统计报表	在线多维多层次数据查询/分析
全局数据维护与同步	请求响应式数据服务	实时近实时数据转发服务	数据分析挖掘

（2）应用对数据集成性需求

企业总体数据模型范围

对于每一种具体应用**都需要理清所涉及的数据模型范围**

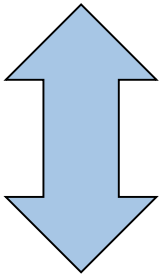
第五部将介绍数据模型范围的概念



（3）数据全局一致需求

企业总体数据模型

数据发布控制需求：数据发布的内容范围、受众范围、发布时机、内容格式标准等一致性需求



数据模型、接口及内容标准制定需求

常规固定统计报表	即时性数据查询	自定义统计报表	在线多维多层次数据查询/分析
全局数据维护与同步	请求响应式数据服务	实时近实时数据转发服务	数据分析挖掘

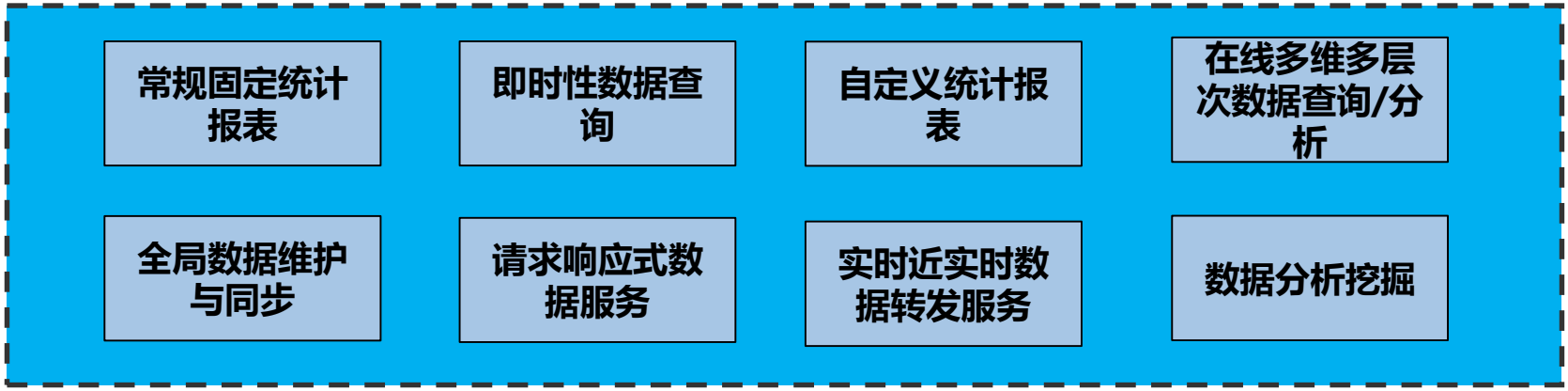
2. 数据服务能力需求

- 快速访问处理数据的能力
  - 索引
  - 缓存
  - 内存计算
  - 预计算
  - 并行分布式计算
- 批量计算的能力
- 吞吐量
- 数据动态性

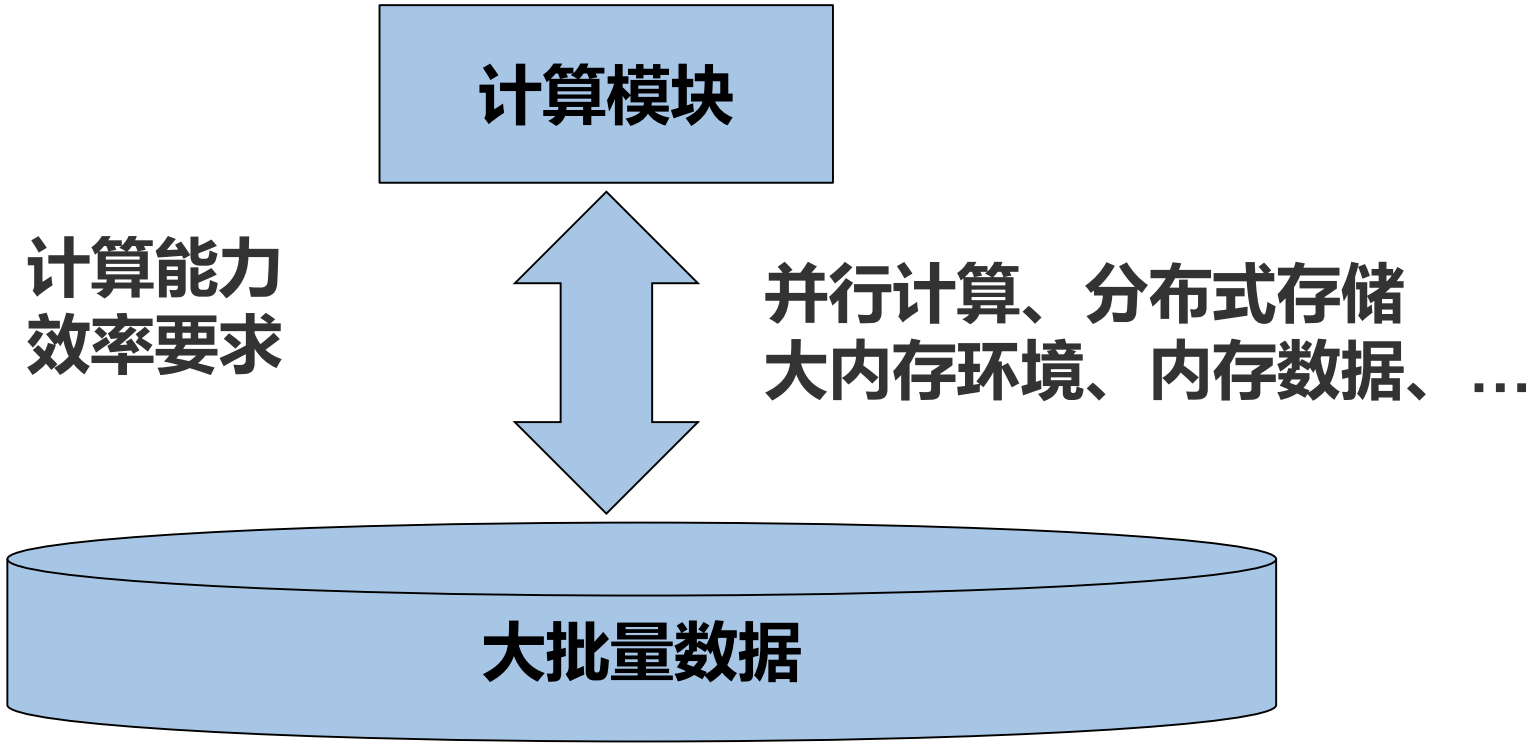
(1)快速访问处理数据的能力

从处理**速度**的角度，明确提出对**不同类别数据处理的效率**要求。

由此带来对索引、缓存、内存计算、预计算、并行分布式计算的设计及技术环境需求

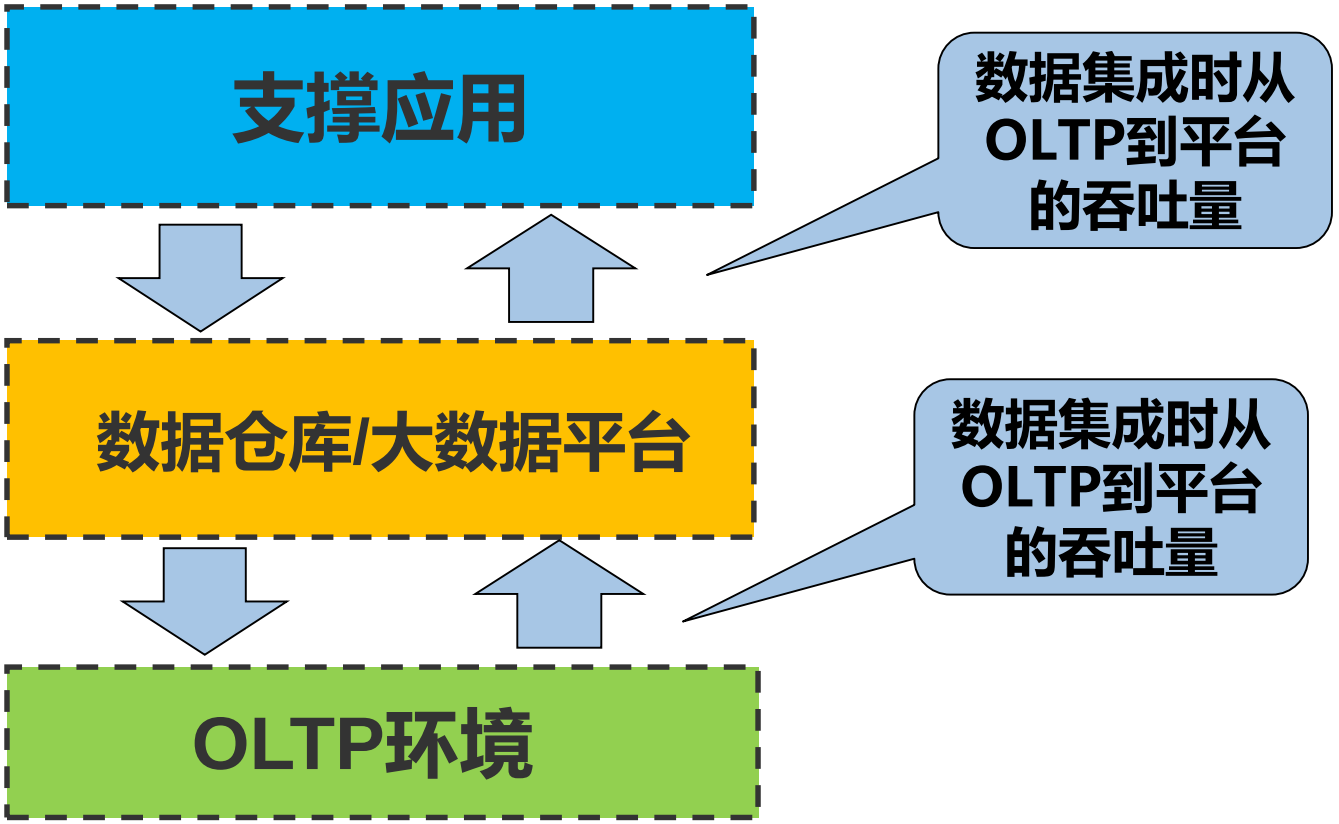


(2)批量计算的能力



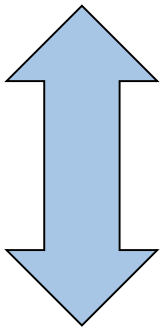
给定计算要求，单位时间内能够处理的数据规模
给定时长的一定量数据和计算要求，给定硬件环境，指定最短完成时间
给定时长的一定量数据和计算要求，指定最短完成时间，设计硬件环境

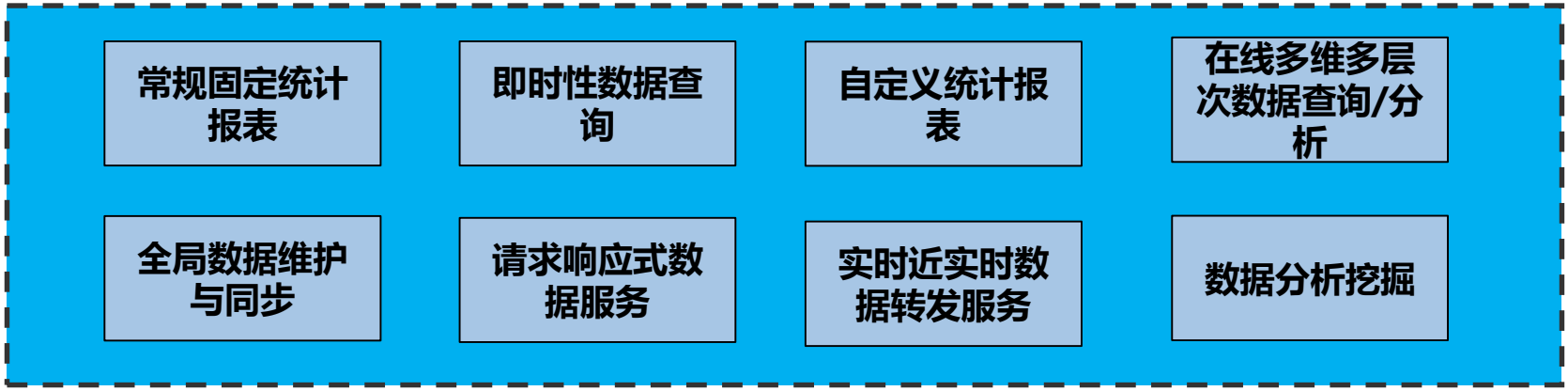
(3) 吞吐量


支撑应用
数据集成时从OLTP到平台的吞吐量
数据仓库/大数据平台
数据集成时从OLTP到平台的吞吐量
OLTP环境

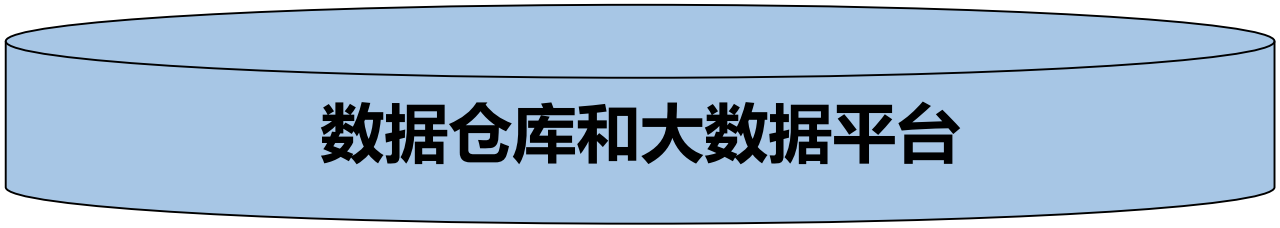
(4) 数据动态性

从数据动态、新鲜程度说明需求，如实时、近实时、小时、…





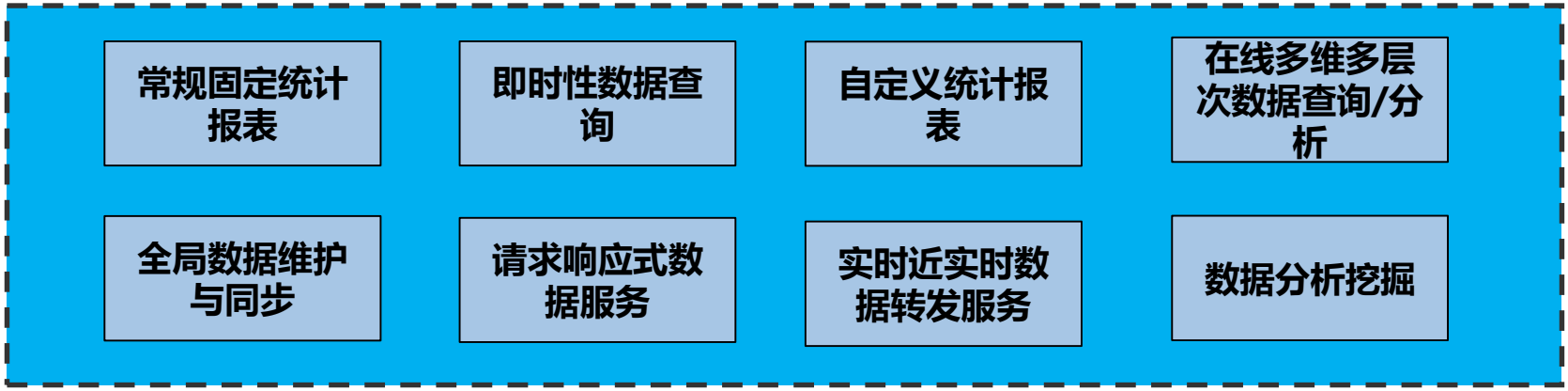
3. 服务质量需求



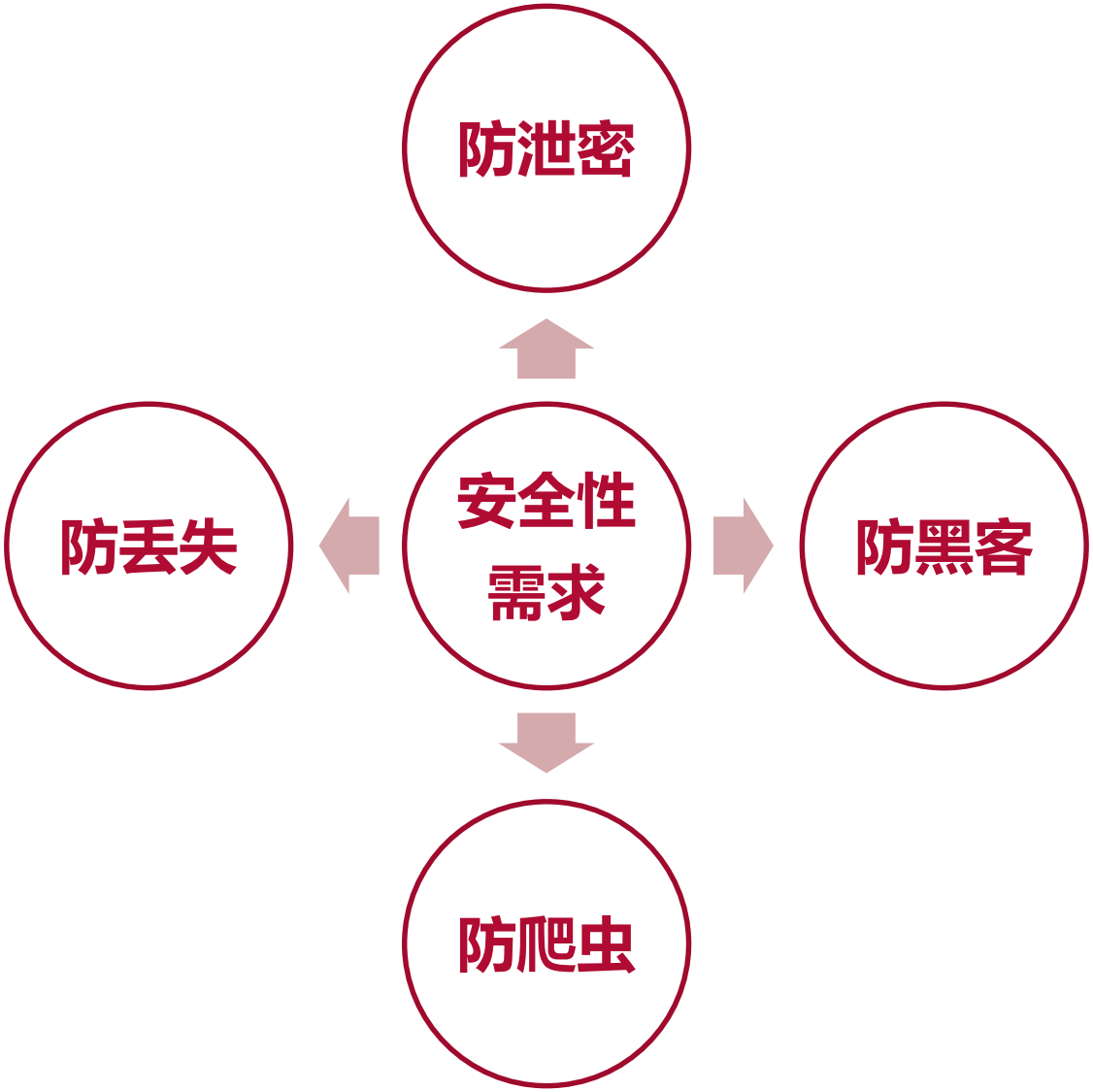
数据服务的可靠性
数据服务的响应速度

服务

数据质量的稳定性
数据服务的持续稳定性



4. 数据安全性需求



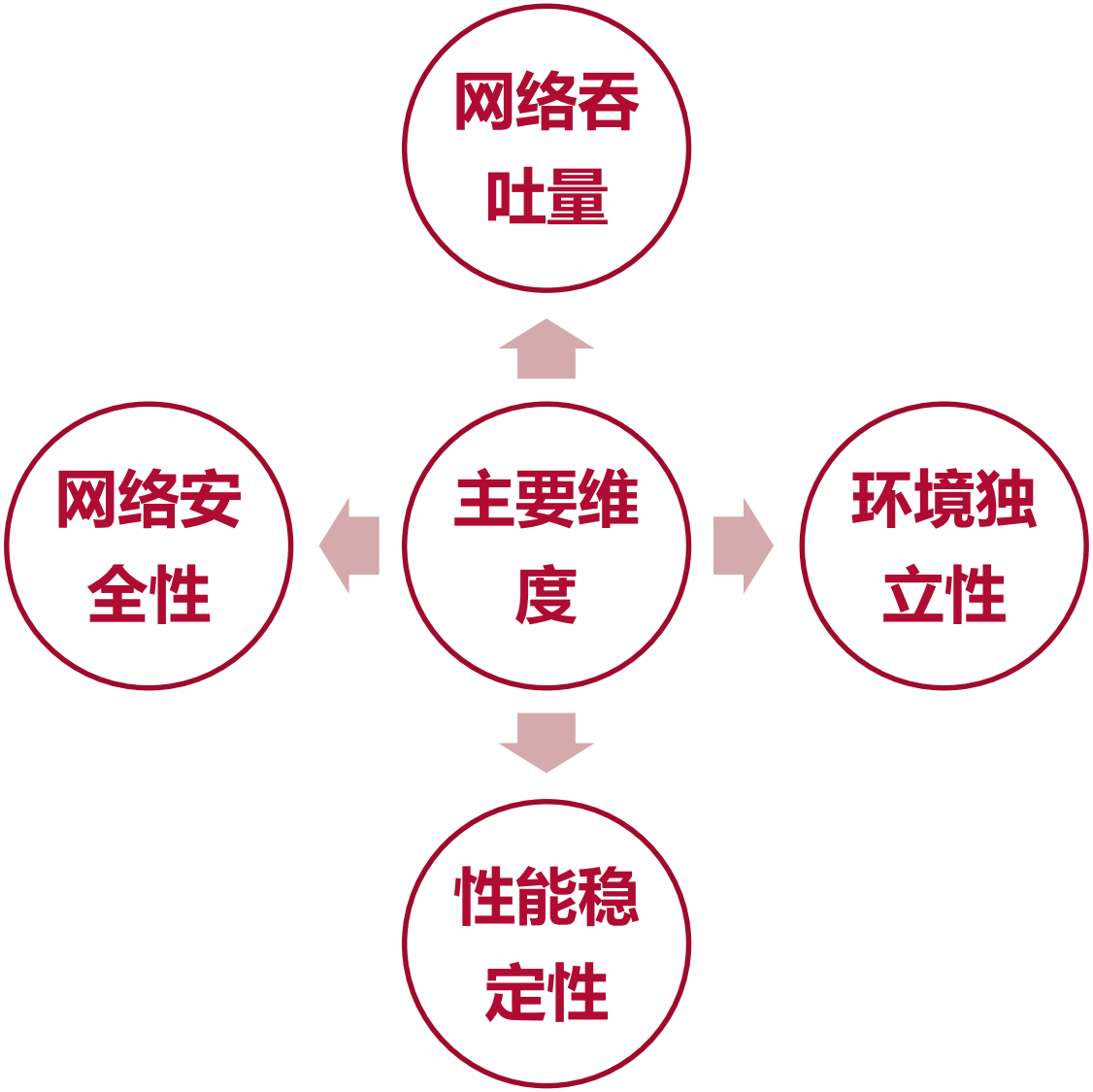
内容提纲

- 决策者的类别与决策需求
- 数据利用需求分类
- 不同应用系统架构范式
- 需求分析方法
- 不同类别应用对数据的要求
- **系统运行环境需求**

系统运行环境需求分类

- 环境需求分类
  - 网络环境需求
  - 数据集成环境需求
  - 存储环境需求
  - 应用分析环境需求
  - 数据服务环境需求
  - WEB服务环境需求
  - 数据管理与安全控制环境需求

1. 网络环境需求



2. 数据集成环境需求

- 集成所用硬件环境需求
  - ETL服务器
  - ETL服务器集群
- 数据集成软件环境需求
  - 数据采集器(定制或标准探针)
  - 采集中间件软件, (Kafka, flume, …)
  - 并行流式处理平台（spark streaming, storm, …）

3. 存储环境需求

- 数据存储容量需求
- 数据存取吞吐量
- 是否支持数据多备份
- 数据服务稳定性——容灾和恢复
- 存储容量平滑可扩展性
- 数据存储架构需求
  - 直连存储—DAS
  - 存储区域网—SAN
  - 分布式存储—HDFS

4. 应用分析环境需求

- 在线分析软硬件环境需求
- 近实时分析软硬件环境需求
- 离线分析软硬件环境需求

5. 数据服务环境需求

- 各种数据服务环境
  - 实时数据服务环境
  - 近实时数据服务环境
  - 离线数据服务环境
- 主要需求内容
  - 硬件环境
  - 软件接口

6. Web服务环境需求

- 各类常见Web服务环境
  - 基础数据服务环境
  - 应用分析服务环境
  - 系统管理服务环境
- Web服务基础软件体系
- 环境硬件配置

7. 数据管理安全控制环境需求

- 数据访问监控环境需求
- 分布式多备份存储体系
- 异地灾备——离线冷存储

某电信公司数据仓库需求实例

BEIJING JIAOTONG UNIVERSITY

1. 系统业务功能需求

- 用户发展情况分析
- 业务发展情况分析
- 收益情况分析
- 市场竞争情况分析
- 服务质量分析
- 营销管理分析
- 大客户分析
- 新业务及数据业务发展分析
- 营销渠道分析

收益情况分析

- 通过对该公司各类用户群体消费情况、欠费情况、交费情况、话费结构及话费回收等的分析，了解移动公司在产品或服务推广时在各类用户群体及其话费结构上的收益情况，为预测未来客户消费趋势，客户交费行为，有效控制欠费，加强管理，提高企业效益提供有效的依据。

大客户分析

- 通过对该公司大客户特殊群体各种情况的分析，了解大客户构成与整个客户群体的构成差异，了解影响大客户新增/流失的主要因素，了解大客户使用业务量的特征，以便在更好地保留大客户的同时发展新的大客户。

2. 系统运行环境需求

- 数据采集服务器环境要求
- 数据仓库服务器环境要求
- 数据仓库管理服务器环境要求
- 应用分析服务器环境要求
- Web服务器环境要求
- 客户端环境要求
- 网络环境要求

网络环境需求

- 数据采集服务器，数据仓库服务器，数据仓库管理服务器，应用分析服务器，Web服务器通常需要运行在一个百兆/千兆IP局域网环境，而客户端既可以直接连到该局域网，也可以通过广域网访问系统。

3. 其它非功能性需求

- 性能需求
  - 数据抽取、多维分析、自定义报表、决策信息展示和静态报表展示、专题分析性能需求
- 数据存储需求
- 访问安全性
- 数据质量需求

数据抽取需求

- 在每日凌晨开始提取昨日的增量数据（如详单、变更用户等），在每日所需的分析正式工作开始时间前完成整个处理工作，满足日分析的需要。
- 在每月出帐完成的第2日（2002-6-30日后是每月月初）开始处理月的增量数据（如帐单等），满足月分析的需求。

决策信息展示和静态报表展示性能需求

- EIS和静态报表的查询主要面向高层决策信息的查询者，速度要求相对较快，速度主要取决于展示的分析结果的数据量。
- EIS和静态报表主要是分析结果的展示，要求预先生成展示结果，预先生成的性能取决于所需分析结果的难度。

存储需求

- 业务基本信息：代码表、客户基本信息永久保存。
- 日粒度明细变动信息：根据需要保留1-3个月。
- 月粒度明细变动信息：根据需要保留3-12个月。
- 日汇总变动信息：根据需要保留3-6月。
- 月汇总变动信息：根据需要保留12-24月。
- 多维数据信息：保留二年。
- 数据保持定期的备份和过期在线数据的归档。

数据质量需求

- 数据抽取系统能及时反馈非法数据，提供不符合质量的源数据记录报告，让提高数据质量成为一项长期的目标。
- 在数据转换流程中定义检查和核对点，保证数据仓库的整体数据质量。
- 在系统设计的过程中，尽量降低低质量（如大量的未知数据、空数据）数据对业务分析的整体性影响，将低质量数据对系统的影响控制在局部的范围内。

本部分小结

- 从不同的角度叙述数据仓库与大数据平台的需求组成，需求分析方法，在现实工程项目非常实际的指导意义。

本部分结束！

BEIJING JIAOTONG UNIVERSITY