基于大数据的数据仓库建设

认识数据仓库

目录

认识数据仓库

01 什么是数据仓库

02 数据仓库的发展史

03 基于大数据数仓构建特点

04 数据仓库的应用范围与前景

05 就业前景与发展方向

什么是数据仓库

什么是数据库？

1.数据库(Database)是按照数据结构来组织、存储和管理数据的建立在计算机存储设备上的仓库

2.数据库是长期储存在计算机内、有组织的、可共享的数据集合。数据库中的数据指的是以一定的数据模型组织、描述和储存在一起、具有尽可能小的冗余度、较高的数据独立性和易扩展性的特点并可在一定范围内为多个用户共享

那么，数据仓库是？

认识数据仓库

什么是数据仓库

定义：面向主题的，集成的，相对稳定的，反映历史变化的数据集合，用于支持管理决策。

面向主题 → 在较高层次上将企业信息系统的数据综合归并进行分析利用的抽象的概念。每个主题基本上对应一个相应的分析领域。

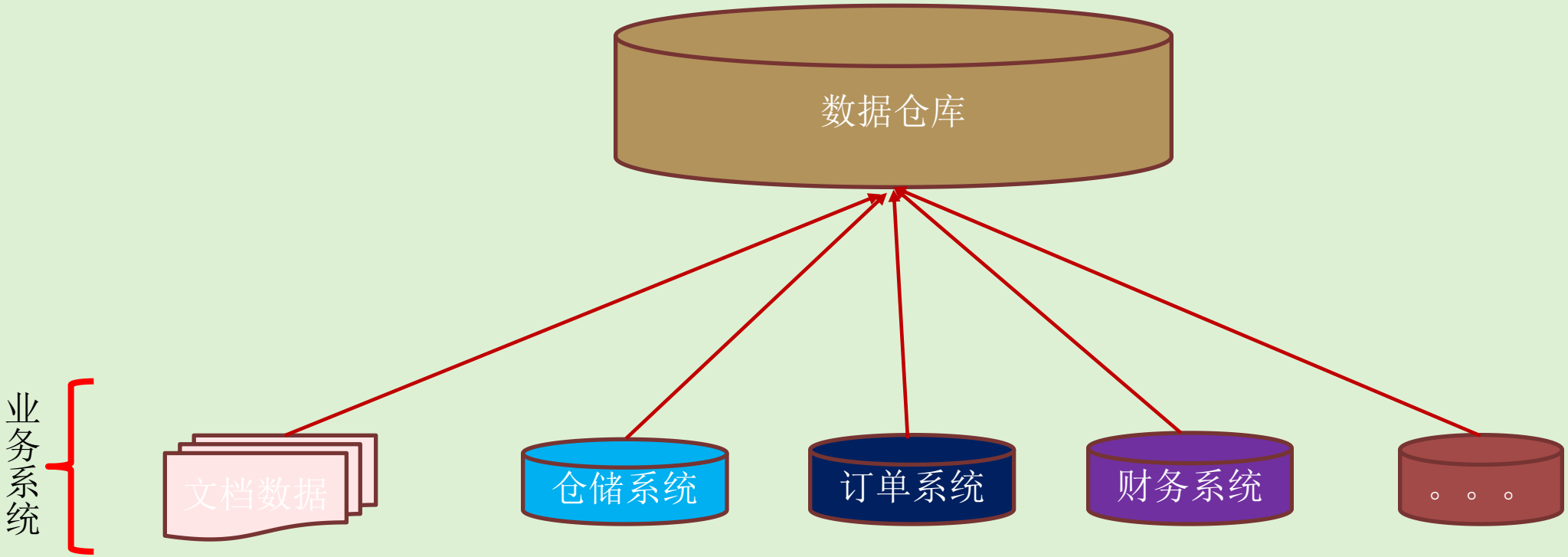
集成的 → 企业级数据，同时数据要保持一致性、完整性、有效性、精确性

稳定的 → 从某个时间段来看是保持不变的，没有更新操作、删除操作，以查询分析为主

变化的 → 反应历史变化

认识数据仓库

实现集成、稳定、反应历史变化
有组织有结构的存储数据的集合



认识数据仓库

功能	数据仓库	数据库
数据范围	存储历史的、完整的、反应历史变化的	当前状态数据
数据变化	可添加、无删除、无变更的、反应历史变化	支持频繁的增、删、改、查操作
应用场景	面向分析、支持战略决策	面向业务交易流程
设计理论	违范式、适当冗余	遵照范式(第一、二、三等范式)、避免冗余
处理量	非频繁、大批量、高吞吐、有延迟	频繁、小批次、高并发、低延迟

面向业务的数据库常称作OLTP，面向分析的数据仓库亦称为OLAP

认识数据仓库

01 什么是数据仓库

02 数据仓库的发展历程

03 基于大数据数仓构建特点

04 数据仓库的应用范围与前景

05 就业前景与发展方向

认识数据仓库

数据仓库概念最早可追溯到20世纪70年代，希望提供一种架构将业务处理系统和分析处理分为不同的层次

20世纪80年代，建立TA2(Technical Architecture2)规范，该明确定义了分析系统的四个组成部分：数据获取、数据访问、目录、用户服务

1988年，IBM第一次提出信息仓库的概念：一个结构化的环境，能支持最终用户管理其全部的业务，并支持信息技术部门保证数据质量；抽象出基本组件：数据抽取、转换、有效性验证、加载、cube开发等，基本明确了数据仓库的基本原理、框架结构，以及分析系统的主要原则

认识数据仓库

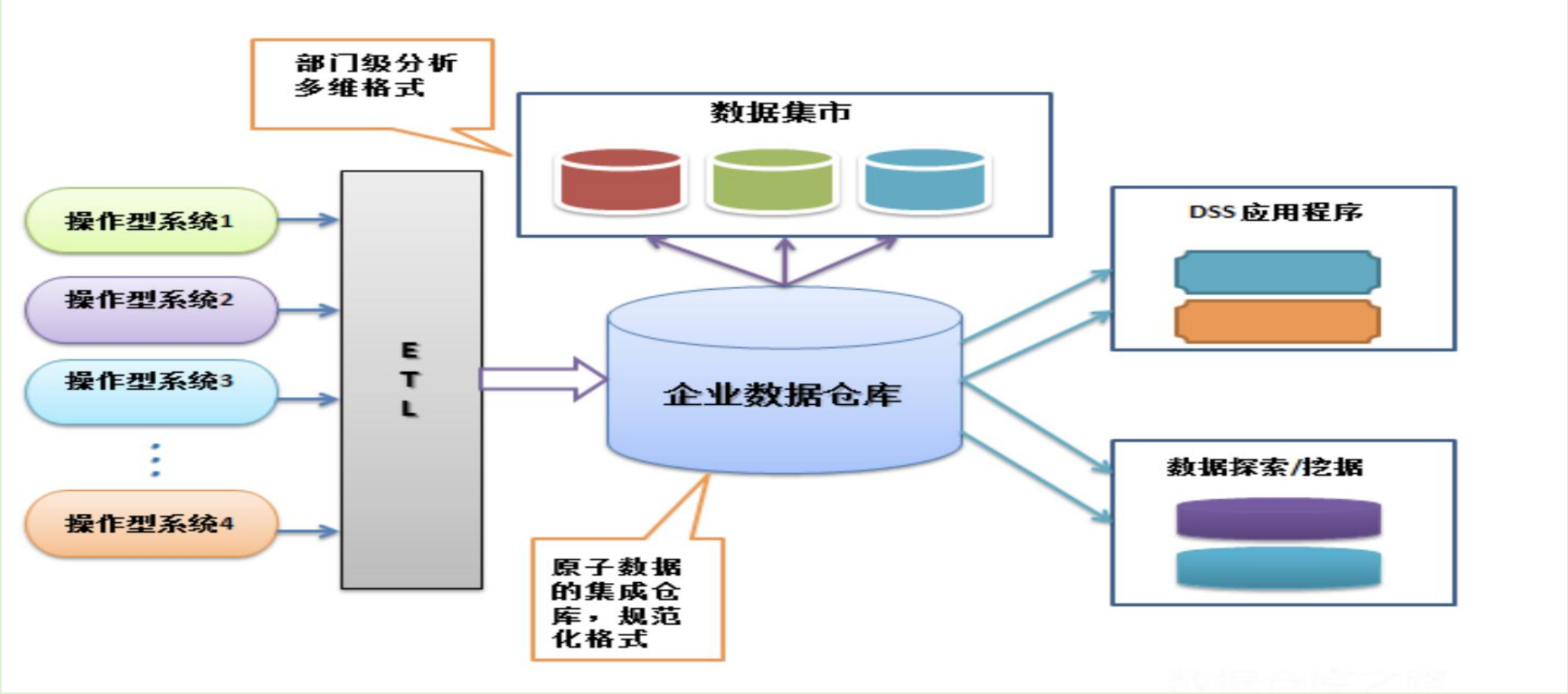
1991年，Bill Inmon出版《 Building the Data Warehouse 》提出了更具体的数据仓库原则：

- 数据仓库是面向主题的
- 集成的
- 包含历史的
- 不可更新的
- 面向决策支持的
- 面向全企业的
- 最明细的数据存储
- 数据快照式的数据获取

尽管有些理论目前仍有争议，但凭借此书获得“数据仓库之父”的殊荣

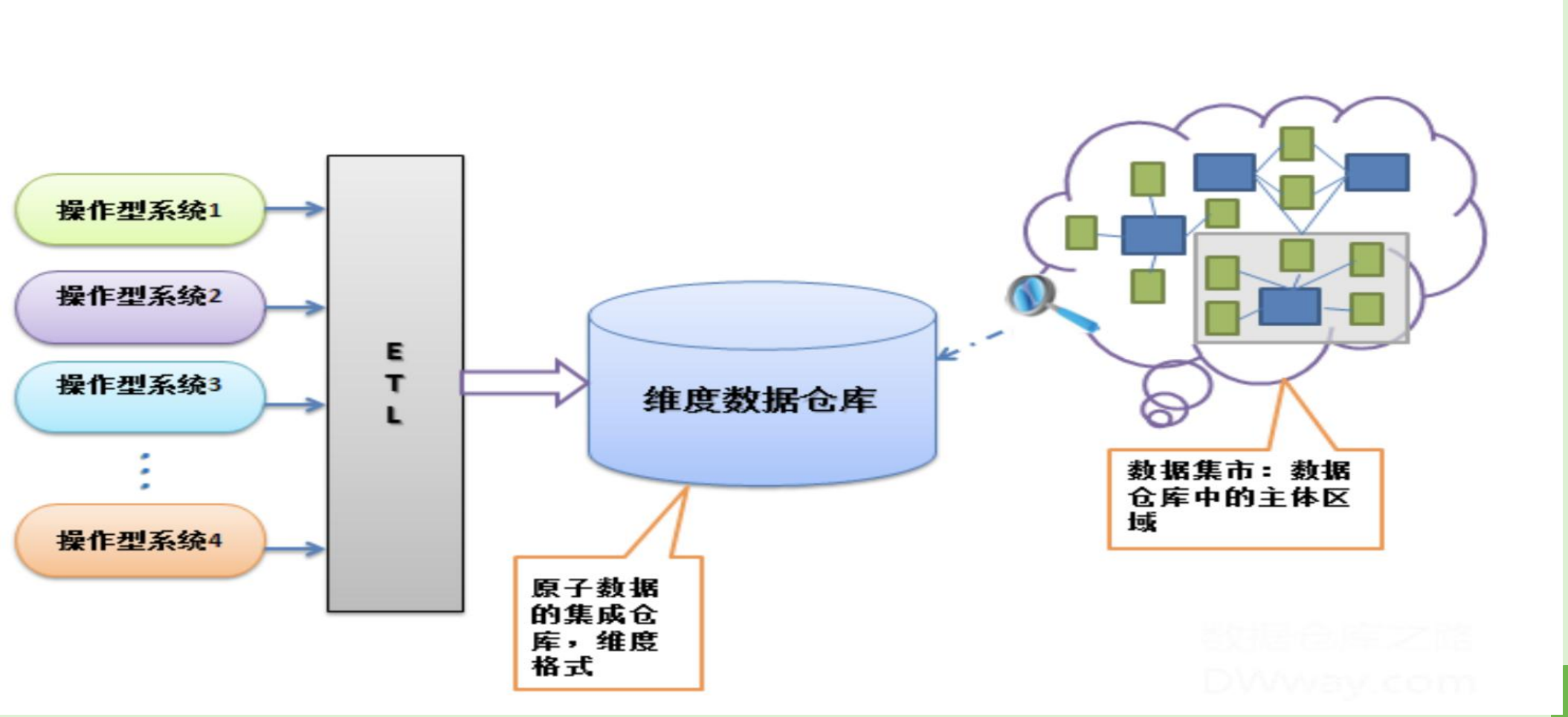
认识数据仓库

Bill Inmon主张自上而下的建设企业数据仓库，认为数据仓库是一个整体的商业智能系统的一部分。一家企业只有一个数据仓库，数据集市的信息来源出自数据仓库，在数据仓库中，信息存储符合第三范式，大致架构：



认识数据仓库

Ralph Kimball 出版《The Data Warehouse Toolkit》，其主张自下而上的建立数据仓库，极力推崇建立数据集市，认为数据仓库是企业内所有数据集市的集合，信息总是被存储在多维模型当中，其思路：



两种思路和观点在实际的操作中都很难成功的完成项目交付，直至最终Bill Inmon提出了新的BI架构CIF(Corporation information factory),把数据集市包含了进来。CIF的核心是将数仓架构划分为不同的层次以满足不同场景的需求，比如常见的ODS、DW、DM等，每层根据实际场景采用不同的建设方案，改思路也是目前数据仓库建设的架构指南，但自上而下还是自下而上的进行数据仓库建设，并未统一

认识数据仓库

01 什么是数据仓库

02 数据仓库的发展历程

03 基于大数据数仓构建特点

04 数据仓库的应用范围与前景

05 就业前景与发展方向

基于大数据的数仓构建特点

随着我们从IT时代步入DT时代，数据从积累量也与日俱增，同时伴随着互联网的发展，越来越多的应用场景产生，传统的数据处理、存储方式已经不能满足日益增长的需求。而互联网行业相比传统行业对新生事物的接受度更高、应用场景更复杂，因此基于大数据构建的数据仓库最先在互联网行业得到了尝试

尽管数据仓库建模方法论是一致的，但由于所面临的行业、场景的不同，在互联网领域，基于大数据的数据仓库建设无法按照原有的项目流程、开发模式进行，更多的是需要结合新的技术体系、业务场景进行灵活的调整，以快速响应需求为导向

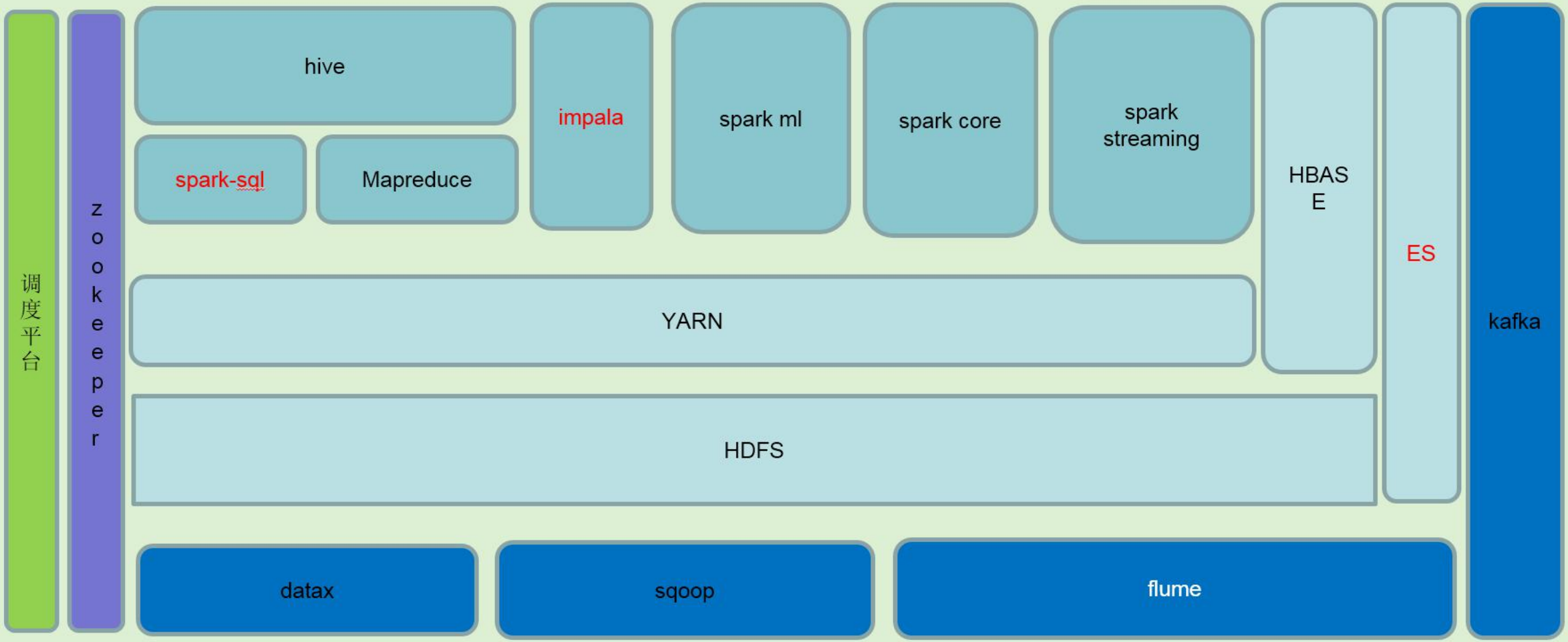
应用场景广泛

- 1，传统的数仓建设周期长，需求稳定，面向DSS、CRM、BI等系统，时效性要求不高
- 2，基于大数据的数据仓库建设要求快速响应需求，同时需求灵活、多变，对实时性有不同程度的要求，除了面向DSS、BI等传统应用外，还要响应用户画像、个性化推荐、机器学习、数据分析等各种复杂的应用场景

技术栈更全面、复杂

- 传统数仓建设更多的基于成熟的商业数据集成平台，比如Teradata、Oracle、Informatica等，技术体系比较成熟完善，但相对比较封闭，对实施者技术面要求也相对专业且单一，一般更多应用于银行、保险、电信等“有钱”行业
  - 基于大数据的数仓建设一般是基于非商业、开源的技术，常见的是基于hadoop生态构建，涉及技术较广泛、复杂，同时相对于商业产品，稳定性、服务支撑较弱，需要自己维护更多的技术框架

技术栈转变



数仓模型设计更灵活

- 传统数仓有较为稳定的业务场景和相对可靠的数据质量，同时也有较为稳定的需求，对数仓的建设有较为完善的项目流程管控，数仓模型设计有严格的、稳定的建设标准
- 在互联网行业：
  - ✓ 行业变化快、业务灵活，同时互联网又是个靠速度存活的行业
  - ✓ 源数据种类繁多：数据库、Nginx log、用户浏览轨迹等结构化、非结构化、半结构化数据
  - ✓ 数据质量相对差，层次不齐

所以，在互联网领域，数仓模型的设计更关注灵活、快速响应和应对多变的市场环境，更加以快速解决业务、运营问题为导向，快速数据接入、快速业务接入，更不存在一劳永逸

认识数据仓库

01 什么是数据仓库

02 数据仓库的发展史

03 基于大数据数仓构建特点

04 数据仓库的应用范围与前景

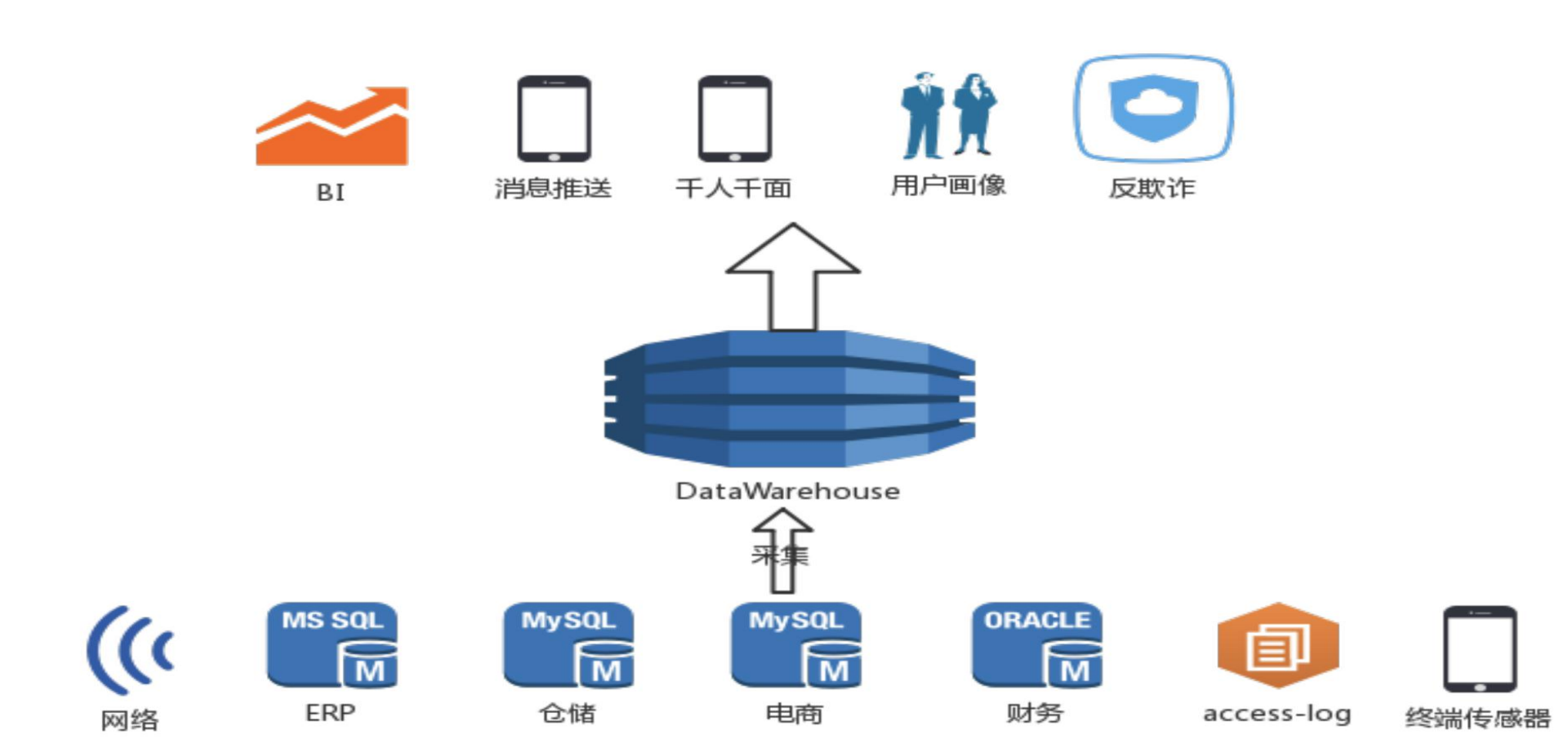
05 就业前景与发展方向

数据仓库的应用范围与前景

- 数仓存在的意义

工程治理

1，基于大数据的数据仓库在互联网行业主要的应用



未来更广泛的应用场景

- 数据分析、数据挖掘、人工智能、机器学习、风险控制、无人驾驶
- 数据化运营、精准运营
- 广告精准、智能投放

认识数据仓库

行业就业、薪资

数据仓库 [雍和宫] 1天前发布
20k-40k 经验3-5年 / 本科
电商 高级 数据分析 数据挖掘

数据仓库工程师 [朝阳区] 12:19发布
30k-45k 经验1-3年 / 本科
中级 数据分析 算法 数据挖掘 机器学习 深度学习

数据仓库/数据研发工… [双井] 15:12发布
25k-50k 经验3-5年 / 本科
高级 数据分析 数据挖掘 建模 数据开发

数据仓库工程师 [北下关] 10:53发布
20k-35k 经验3-5年 / 本科
高级 大数据 数据挖掘 hadoop

大讲台
dajiangtai.com

THANKS

小讲老师：84985152

助教老师：484166349