上海财经大学
SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

数据库和数据仓库

饶艳超 副教授

上海财经大学会计学院

raoyanchao@qq.com

上海财经大学
SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

3

数据仓库的发展

数据仓库的发展

- **数据仓库的发展趋势**

✓早期数据仓库构建主要指的是把**企业的业务数据库**如ERP、CRM、SCM等数据按照决策分析的要求建模并汇总到数据仓库引擎中，其应用以报表为主，目的是支持管理层和业务人员决策（中长期策略型决策）。

✓随着IT技术走向互联网、移动化，数据源变得越来越丰富，在原来业务数据库的基础上出现了非结构化数据，比如**网站log，IoT设备数据，APP埋点数据**等，这些数据量比以往结构化的数据大了几个量级，对ETL过程、存储都提出了更高的要求；

✓互联网的在线特性也将业务需求推向了实时化，随时根据当前客户行为而调整策略变得越来越常见，比如**大促销过程中库存管理，运营管理**等（既有中远期策略型，也有短期操作型）；同时公司**业务互联网化之后**导致**同时服务的客户剧增**，有些情况人工难以完全处理，这就需要**机器自动决策**。比如欺诈检测和用户审核。

数据仓库的发展

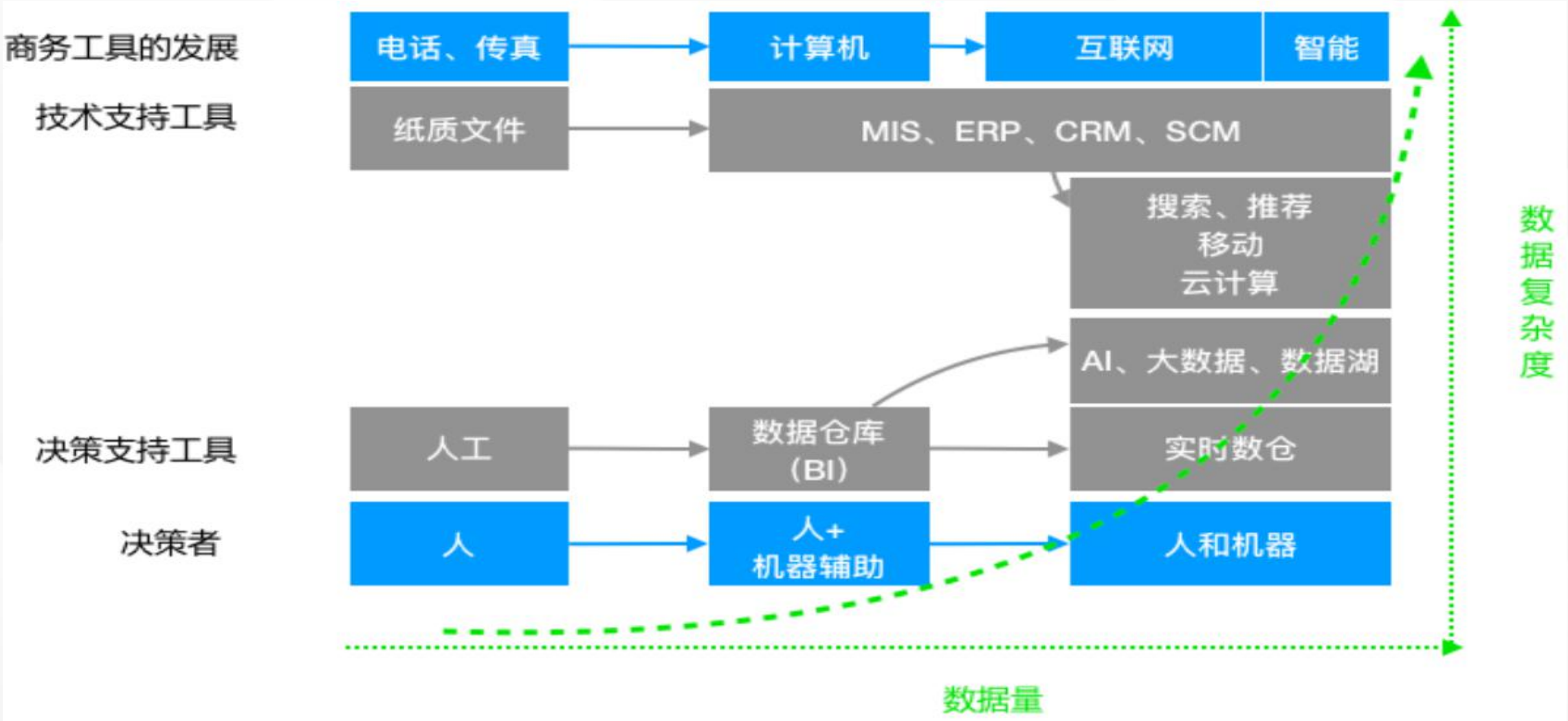
- 数据仓库的发展趋势

✓实时数据仓库以满足实时化&自动化决策需求；

✓大数据&数据湖以支持大量&复杂数据类型（文本、图像、视频、音频）；

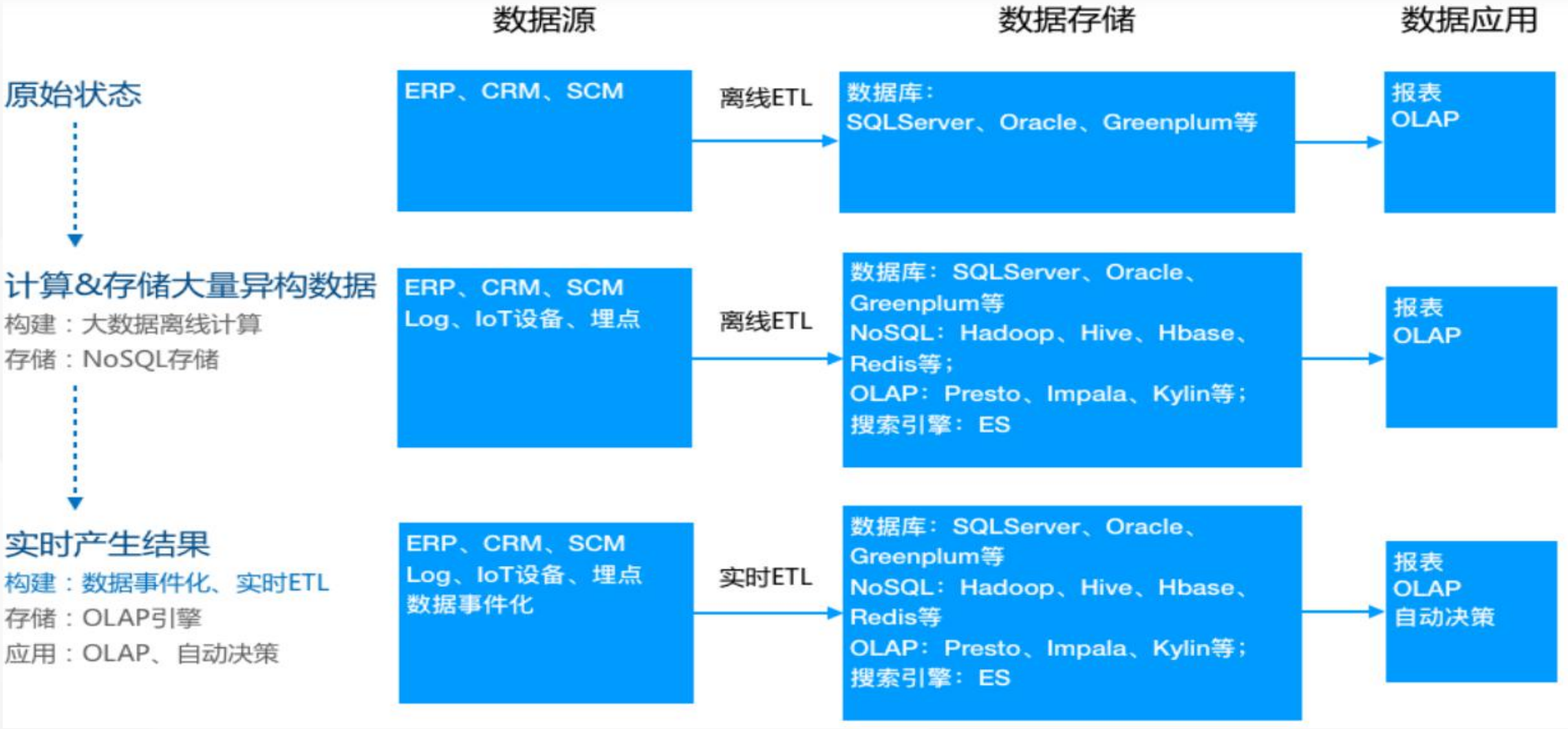
数据仓库

- 数据仓库的发展趋势



数据仓库

- 数据仓库的发展趋势



数据仓库

- **菜鸟仓配实时数据仓库：整体设计**

✓数据来源：基于业务系统的数据；

✓数据模型：采用中间层的设计理念，建设仓配实时数仓；

✓计算引擎：选择更易用、性能表现更佳的实时计算作为主要的计算引擎；

✓数据服务：选择天工数据服务中间件，避免直连数据库，且基于天工可以做到主备链路灵活配置秒级切换；

✓数据应用：围绕大促全链路，从活动计划、活动备货、活动直播、活动售后、活动复盘五个维度，建设仓配大促数据体系。

数据仓库

菜鸟仓配实时数据仓库

整体设计

数据仓库

- **菜鸟仓配实时数据仓库：整体设计**

✓数据模型：采用中间层的设计理念，建设仓配实时数仓；

✓不管是从计算成本，还是从易用性、复用性、一致性……，都必须避免烟囱式的开发模式，而是以中间层的方式建设仓配实时数仓。

✓实时中间层分为两层。

数据仓库

- 菜鸟仓配实时数据仓库：整体设计——数据模型

✓第一层DWD公共实时明细层

✓实时计算订阅业务数据消息队列，

✓然后通过数据清洗、多数据源join、流式数据与离线维度信息等的组合，将一些相同粒度的业务系统、维表中的维度属性全部关联到一起，增加数据易用性和复用性，得到最终的实时明细数据。

✓这部分数据有两个分支：一部分直接落地到ADS，供实时明细查询使用，一部分再发送到消息队列中，供下层计算使用。

数据仓库

- 菜鸟仓配实时数据仓库：整体设计——数据模型

✓第二层DWS公共实时汇总层

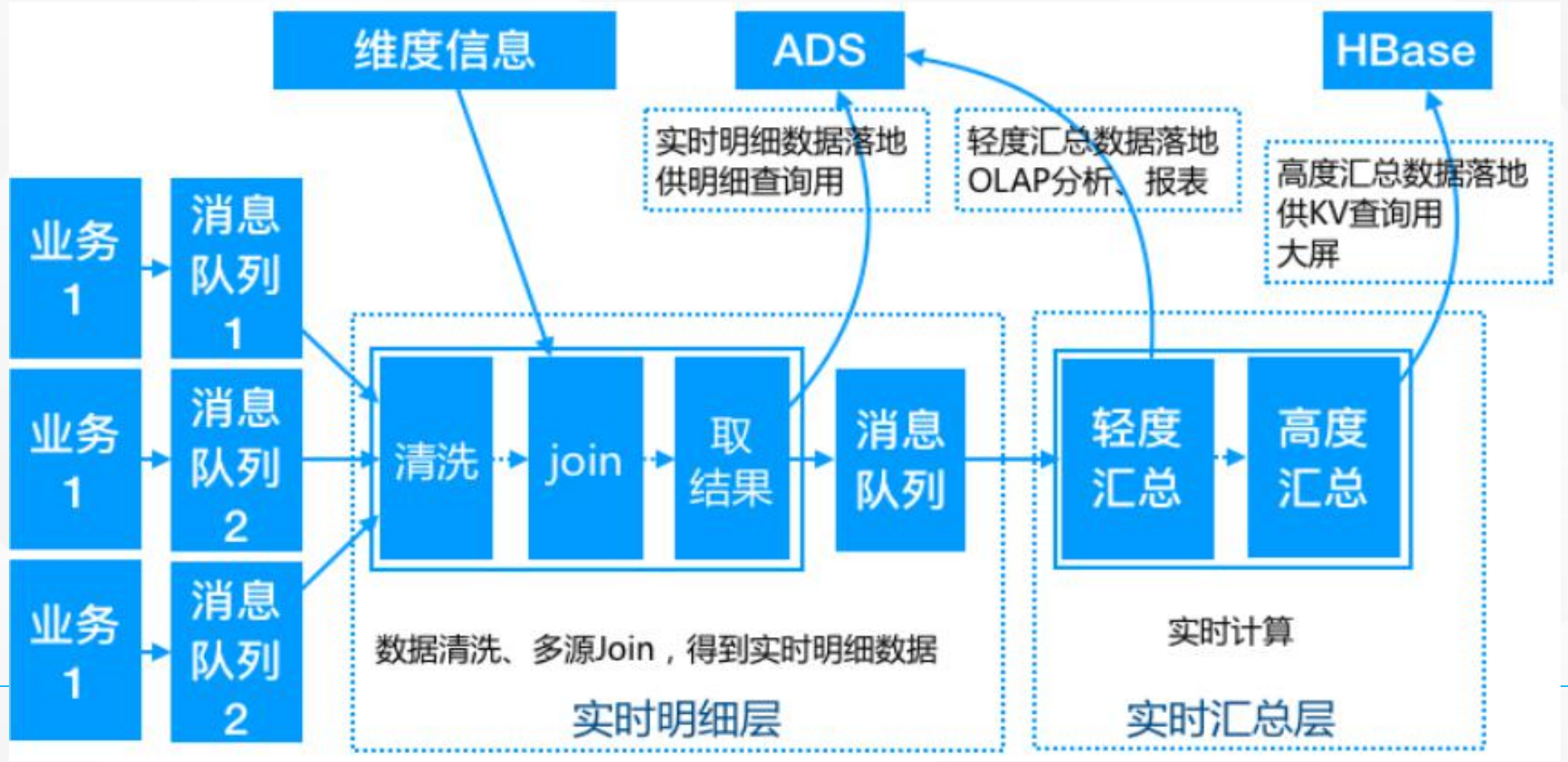
✓以数据域+业务域的理念建设公共汇总层，与离线数仓不同的是，这里汇总层分为轻度汇总层和高度汇总层，并同时产出。

✓轻度汇总层写入ADS，用于前端产品复杂的olap查询场景，满足自助分析和产出报表的需求。

✓高度汇总层写入Hbase，用于前端比较简单的kv查询场景，提升查询性能，比如实时大屏等。

数据仓库

- 菜鸟仓配实时数据仓库：整体设计——数据模型



数据仓库

- 菜鸟仓配实时数据仓库：整体设计

✓数据保障

✓集团每年都有双十一等大促，大促期间流量与数据量都会暴增。

✓实时系统要保证实时性，相对离线系统对数据量更敏感，对稳定性要求更高。

✓所以为了应对这种场景，还需要在这种场景下做两种准备：

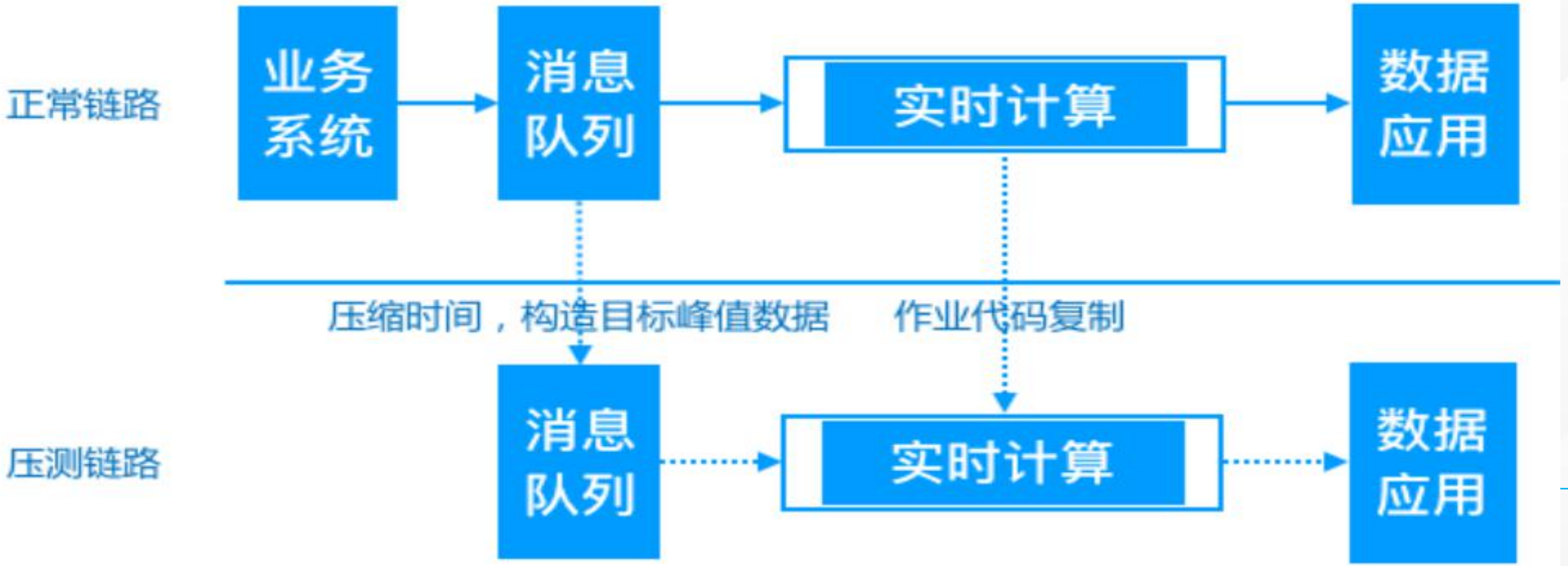
- 大促前的系统压测；
- 大促中的主备链路保障；

数据仓库

- 菜鸟仓配实时数据仓库：整体设计——数据保障

系统压测

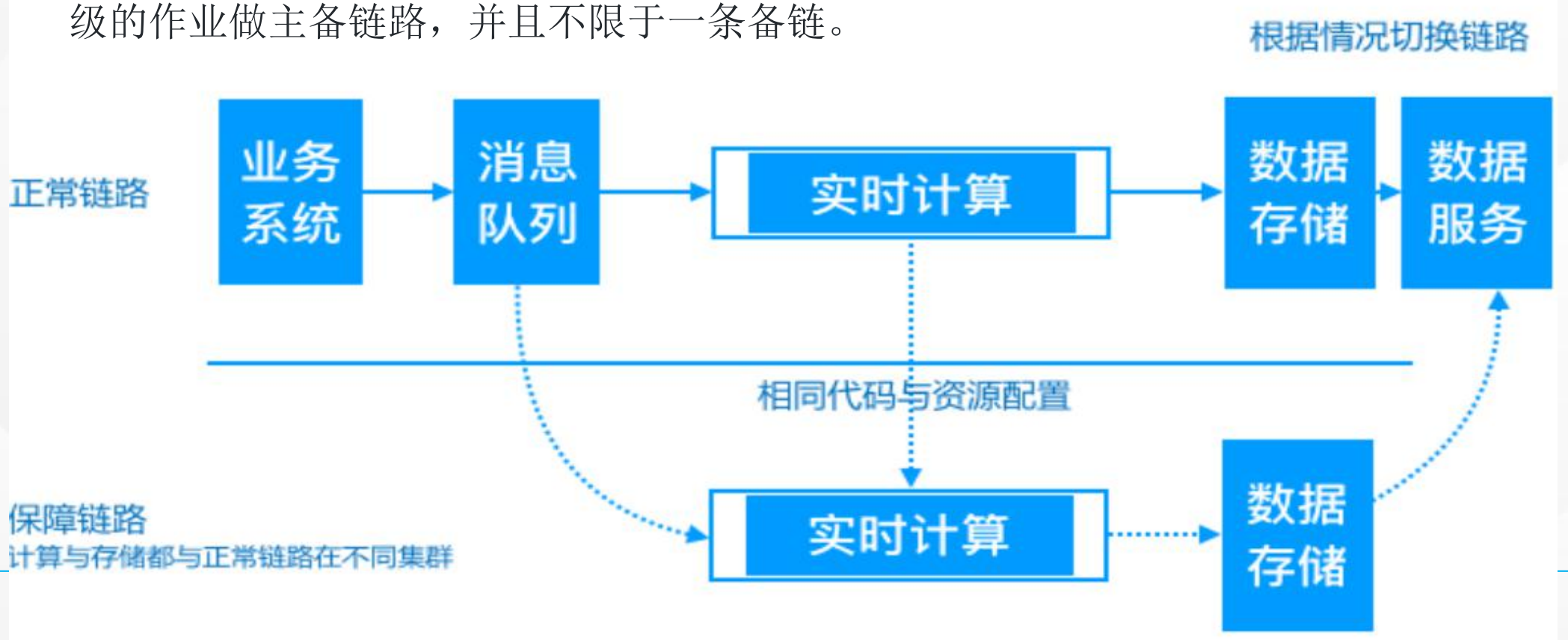
✓ 压测的主要目的是产出实时计算在大促过程所需资源及其配置。



- 菜鸟仓配实时数据仓库：整体设计——数据保障

主备链路

✓ 主备链路保障的目的是在主链出现问题时能通过备链提供服务，可以只针对高优先级的作业做主备链路，并且不限于一条备链。



数据仓库

- **实时数仓与离线数仓的对比**

✓从架构上，有比较明显的区别，实时数仓以Kappa架构为主，而离线数仓以传统大数据架构为主；Lambda架构可以认为是两者的中间态。

✓从建设方法上，实时数仓和离线数仓基本还是沿用传统的数仓主题建模理论，产出事实宽表。

✓从数据保障上，实时数仓因为要保证实时性，所以对数据量的变化较为敏感。在大促等场景下需要提前做好压测和主备保障工作，这是与离线数据的一个较为明显的区别。

数据库和数据仓库

- 思考：

✓当前的哪些变化引发了数据仓库的发展？

✓数据仓库的发展表现出了怎样一种发展趋势？

✓简要概述菜鸟仓配实时数仓的整体设计框架

✓实时数仓和离线数仓有什么区别？

上海财经大学

SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

THANK YOU