# 一、平台体系

常见步骤



知识结构



# 二、常见名词

热数据

是需要被计算节点频繁访问的在线类数据。

冷数据

是对于离线类不经常访问的数据，比如企业备份数据、业务与操作日志数据、话单与统计数据。

两个不同的访问频次，就导致了在数据库搭建的各自不同，有一句话简单明了：热数据就近计算，冷数据集中存储。

所以，热数据因为访问频次需求大，效率要求高，所以就近计算和部署；冷数据访问频次低，效率要求慢，可以做集中化部署，而基于大规模存储池里，可以对数据进行压缩、去重等降低成本的方法。



温数据

从数据分析的层面来看，不仅有冷热两种数据，还有温数据。

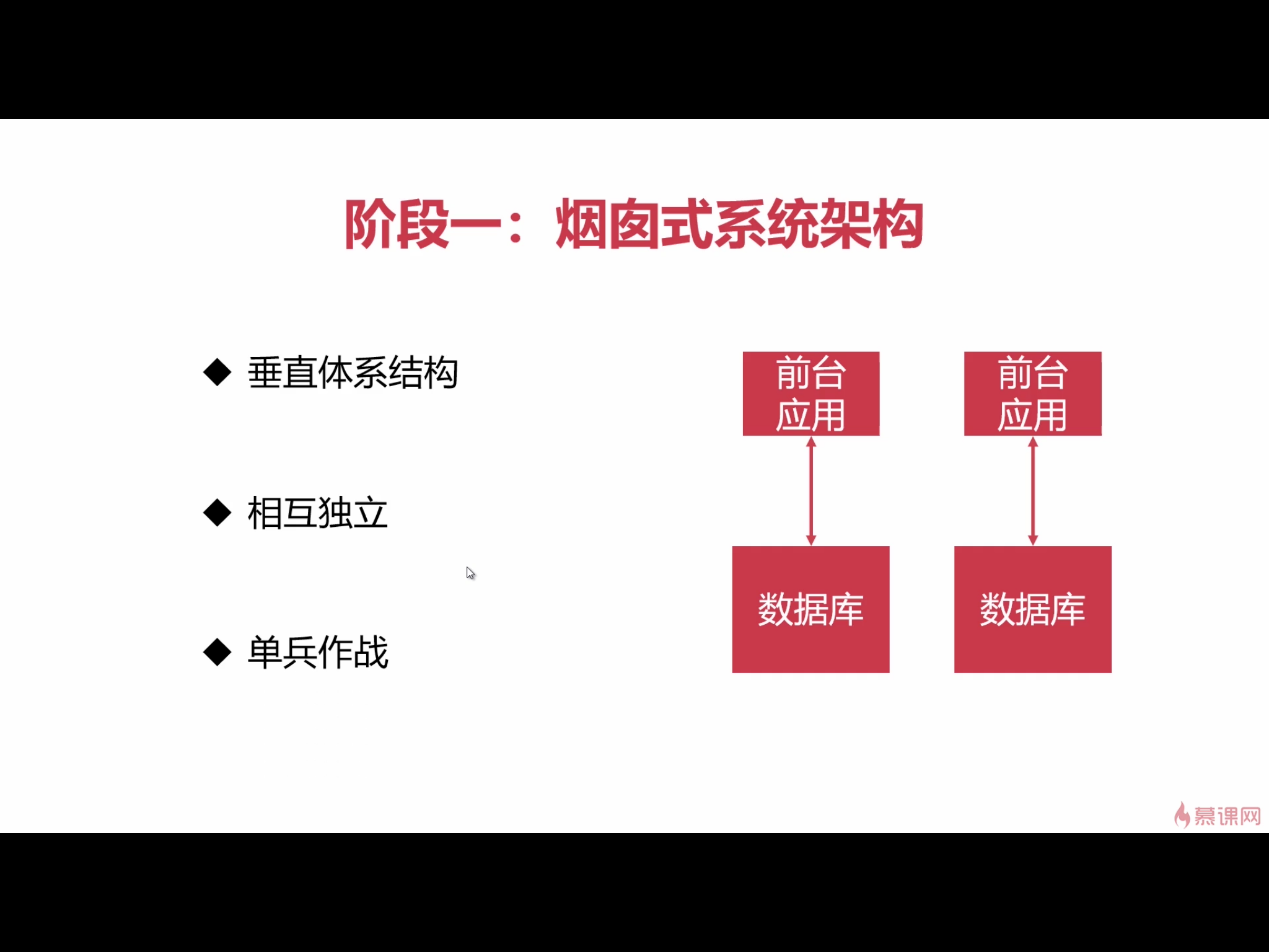
近期活跃应用、近期去过的地方等具有一定时效性的行为数据，表征“最近对什么感兴趣”。

HDFS

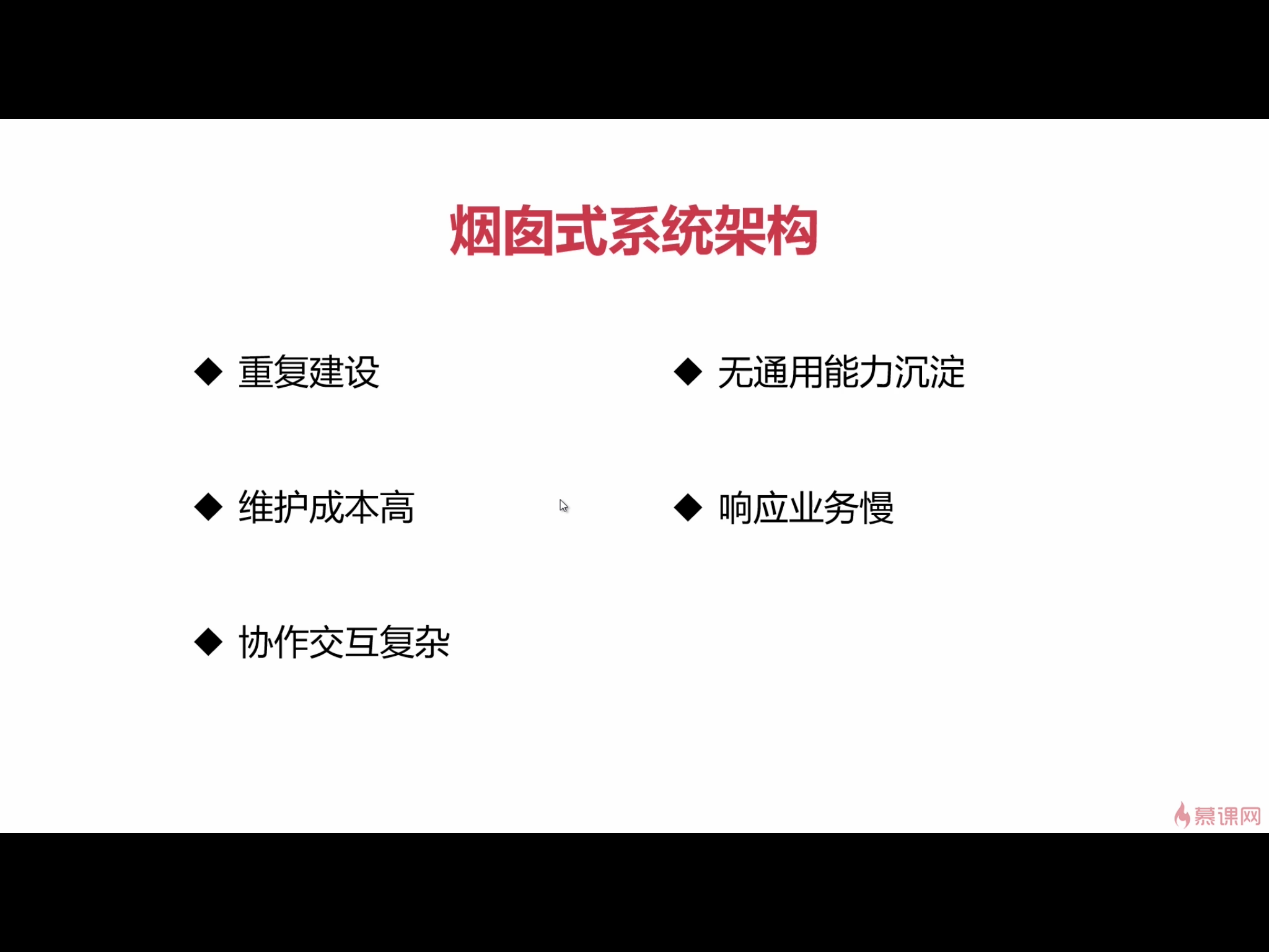
Hadoop分布式文件系统(HDFS)是指被设计成适合运行在通用硬件(commodity hardware)上的分布式文件系统（Distributed File System）。

数据中台

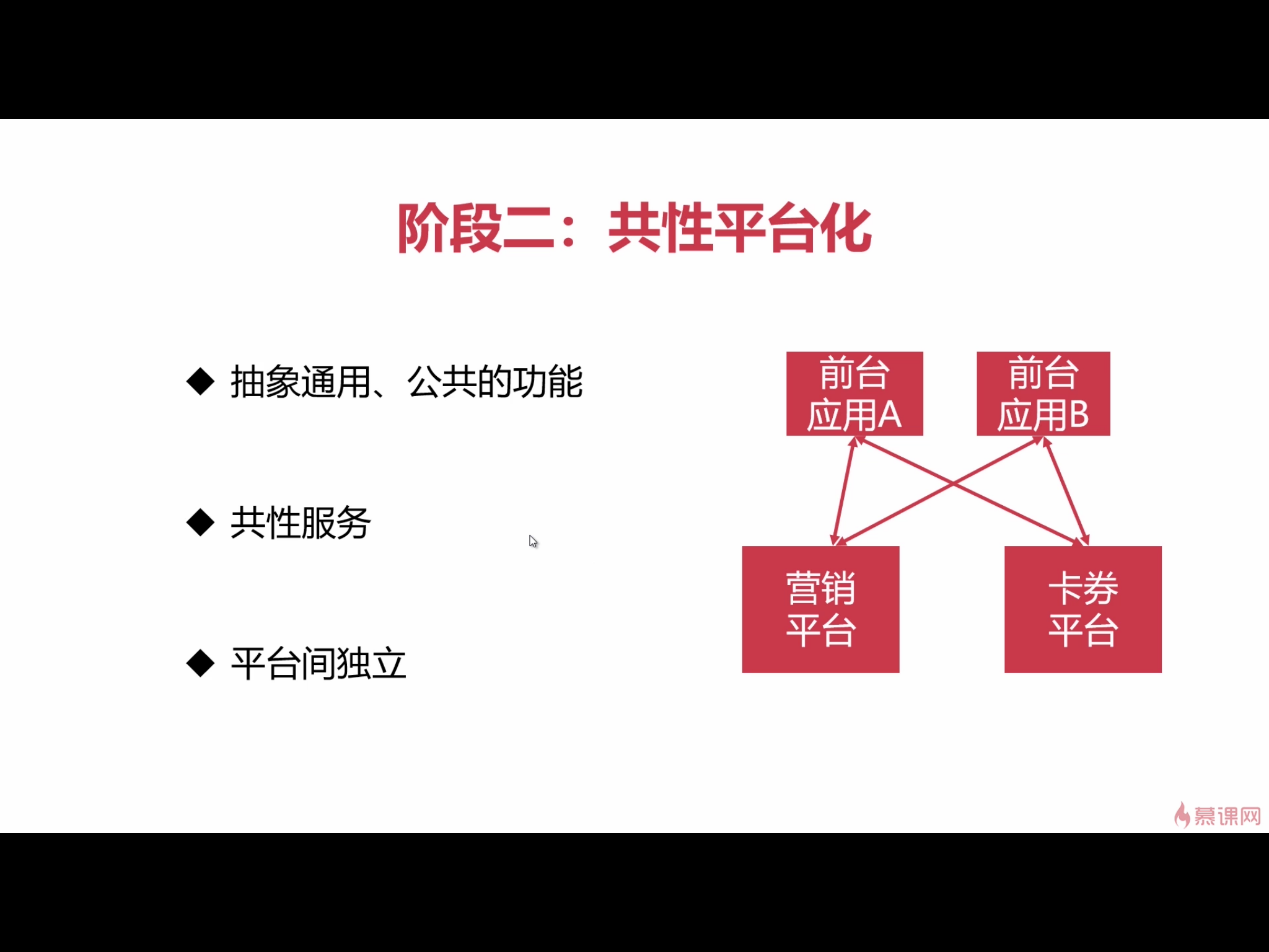
问题来源：

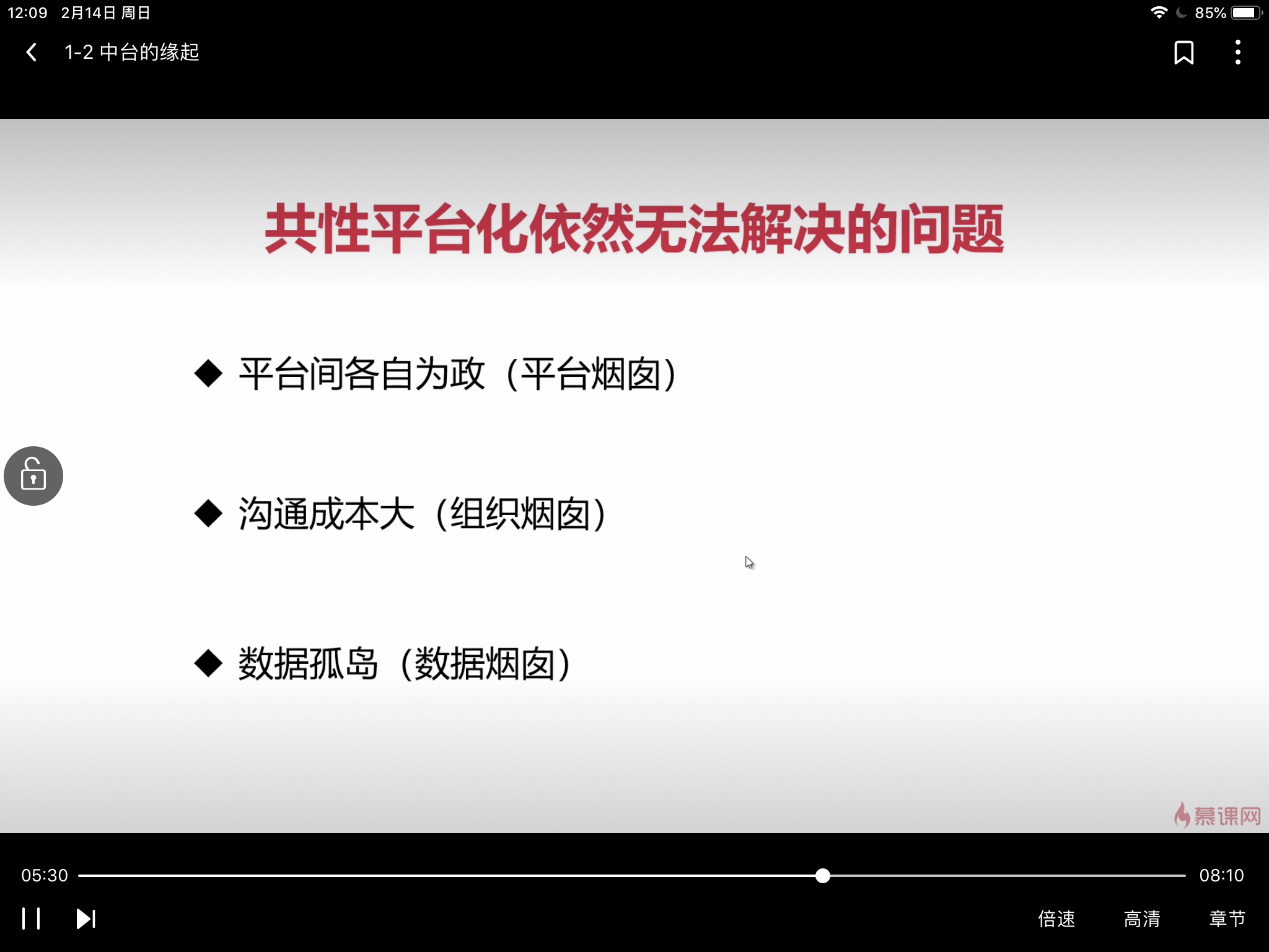


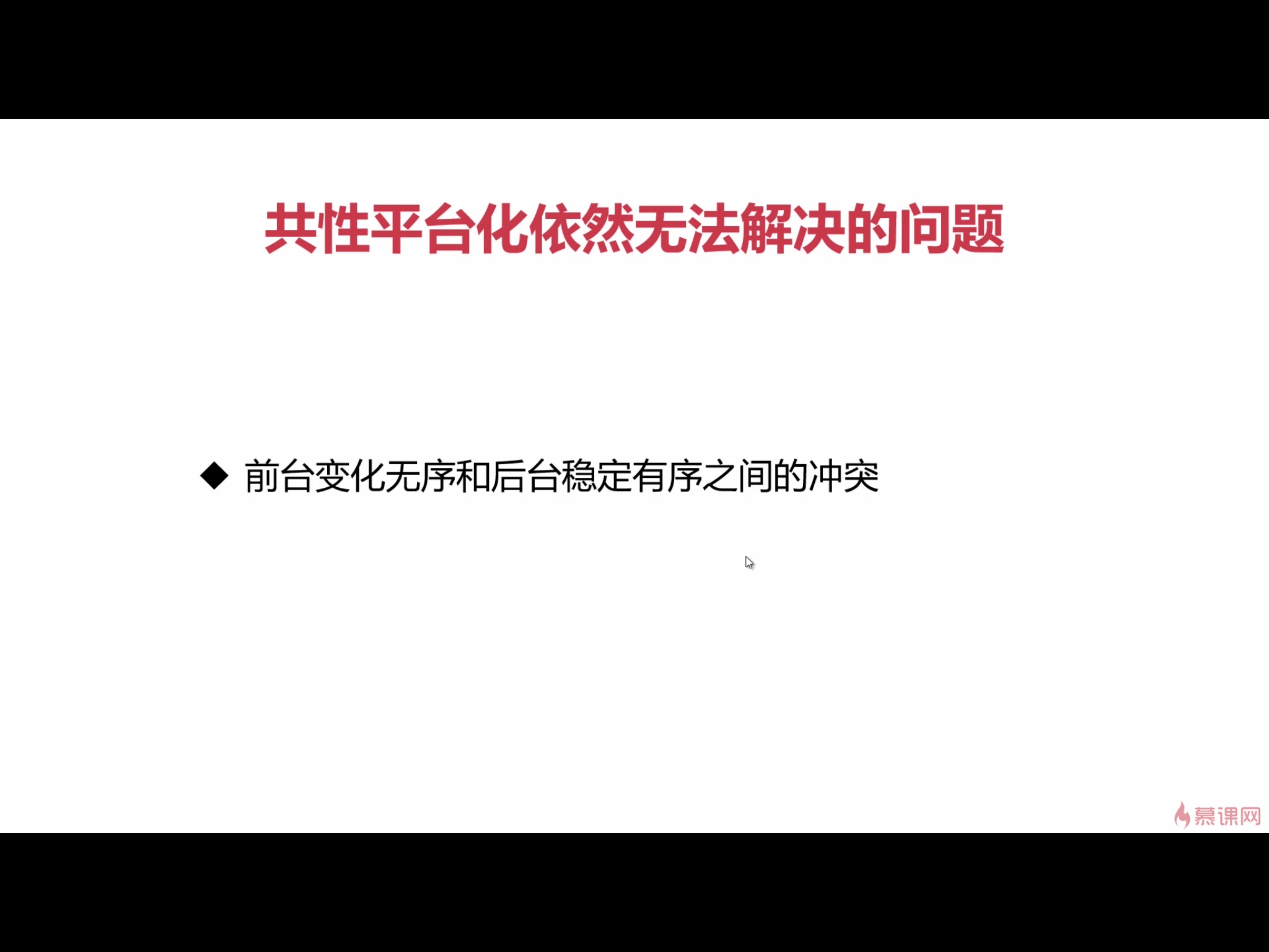
一套系统一套前后台。

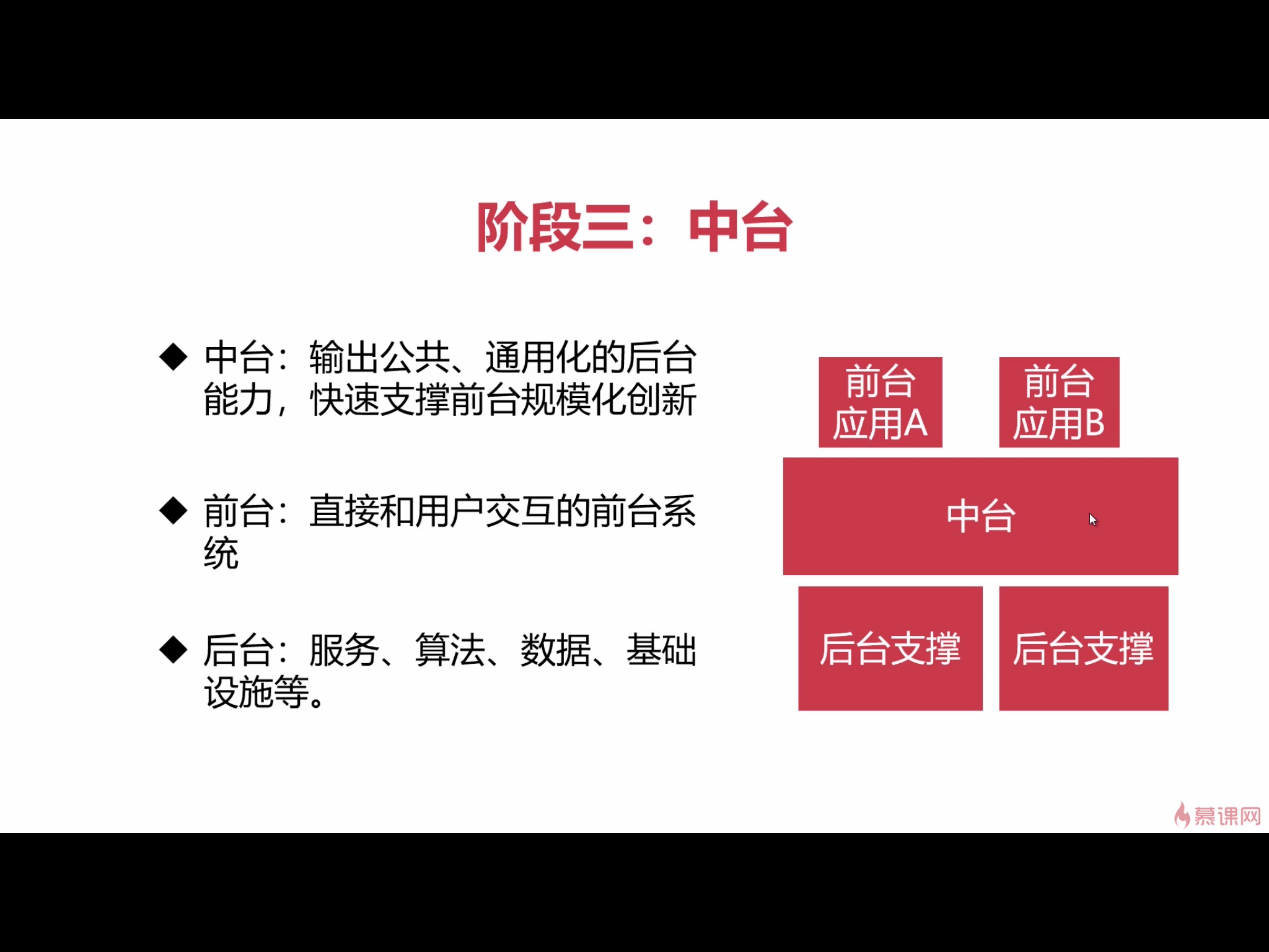


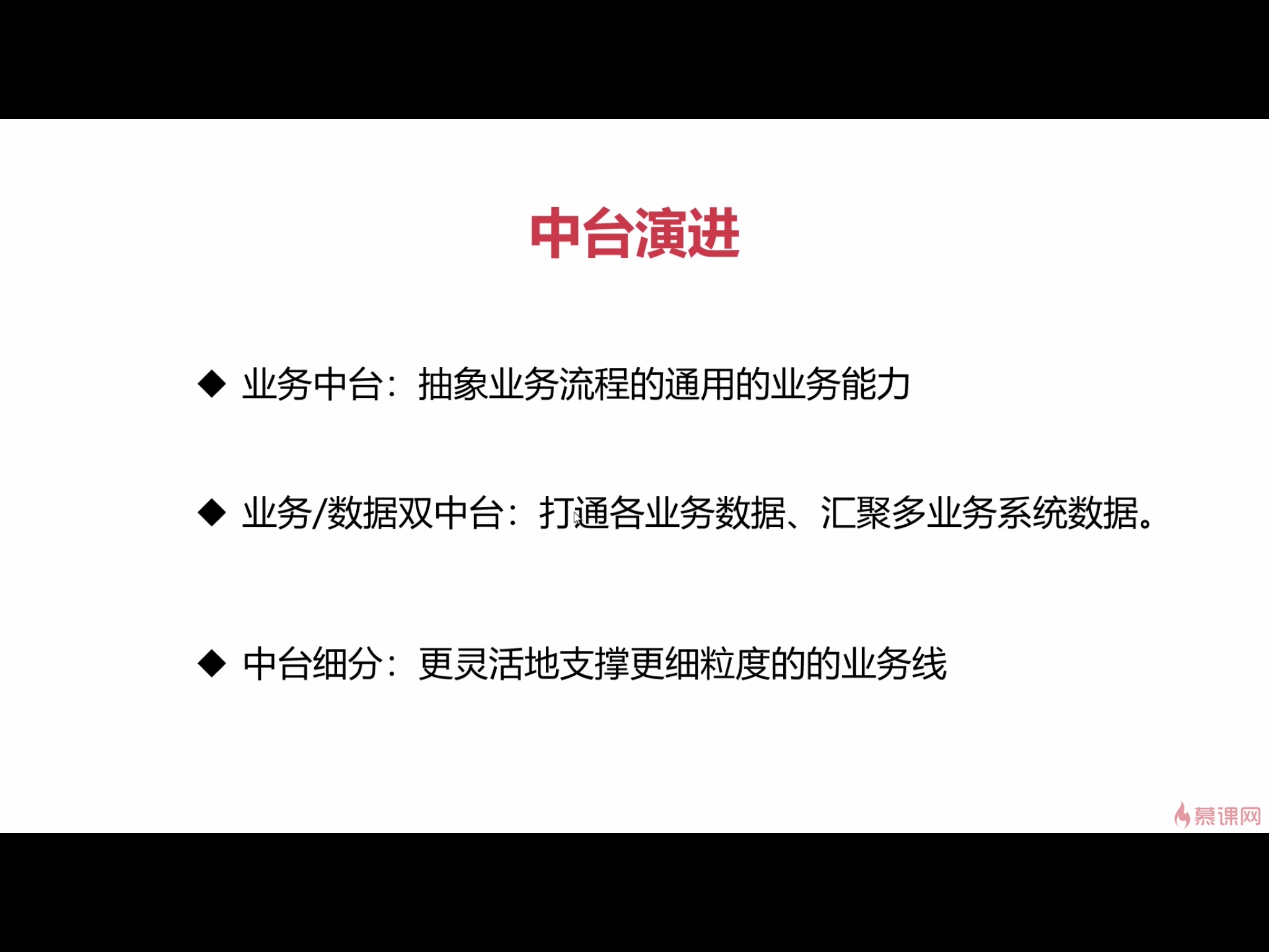
特别是后台服务重复建设，比如一套应用就一套权限管理系统，显然不合适。













数据仓库和数据湖

数据湖与数据仓库之间的异同点有很多，很容易混淆，但是最重要的区别有两个：

存储数据类型：数据仓库是存储数据，进行建模，存储的是结构化数据；数据湖以其本源格式保存大量原始数据，包括结构化的、半结构化的和非结构化的数据。在需要数据之前，没有定义数据结构和需求。

数据处理模式：在我们可以加载到数据仓库中的数据，我们首先需要定义好它，这叫做写时模式（Schema-On-Write）。而对于数据湖，您只需加载原始数据，然后，当您准备使用数据时，就给它一个定义，这叫做读时模式（Schema-On-Read）。这是两种截然不同的数据处理方法。因为数据湖是在数据到使用时再定义模型结构，因此提高了数据模型定义的灵活性，可满足更多不同上层业务的高效率分析诉求。

# 三、大数据基础平台及企业级数仓

HDFS架构原理及深度剖析