北京師範大學
BEIJING NORMAL UNIVERSITY

数据价值链路实践 第三讲

数据重组

主讲人：蓝昕

2021.07.31 ~ 08.28

每周六 晚7:00

大纲

1. 数据中台
2. 数据仓库：基本概念（重点）
3. 数据仓库：开发流程

数据中台

中心化分发

中台架构

左图：多张图表对比。右图：常规用法

中台与数据仓库

数据中台

- 完整的中心化解决方案
- 通过API分发数据
- 基础：多个核心系统组成，包括数据仓库、元数据管理、质量保障、任务调度等
- 汇聚各种数据源

VS

数据仓库

- 提供有序组织、干净的数据
- 中心化仓库
- 主题化、集成、时变、非易失
- 汇聚各种数据源

北京師範大學
BEIJING NORMAL UNIVERSITY

数据仓库：基本概念

基础知识

业务视角：更快更好的交付

分析报告产出更早

ETL开发和拓展耗时更少

数据交付质量更高

技术视角：根据需求层层递进

基本

集成数据，进行清洗，保证数据干净完整

【产出有效分析的最低要求】

进阶

细分主题，数据解耦重组。业务抽象清晰，便于维护拓展

【传统数仓的基本要求，数据直观反映业务活动】

优化

优化表设计，强化业务中心，降低联结深度，反范式化等

【提升查询效率，让写SQL变成享受】

拓展

满足额外需求，如记录变化，构建特定报表等

【根据业务需要进行扩充】

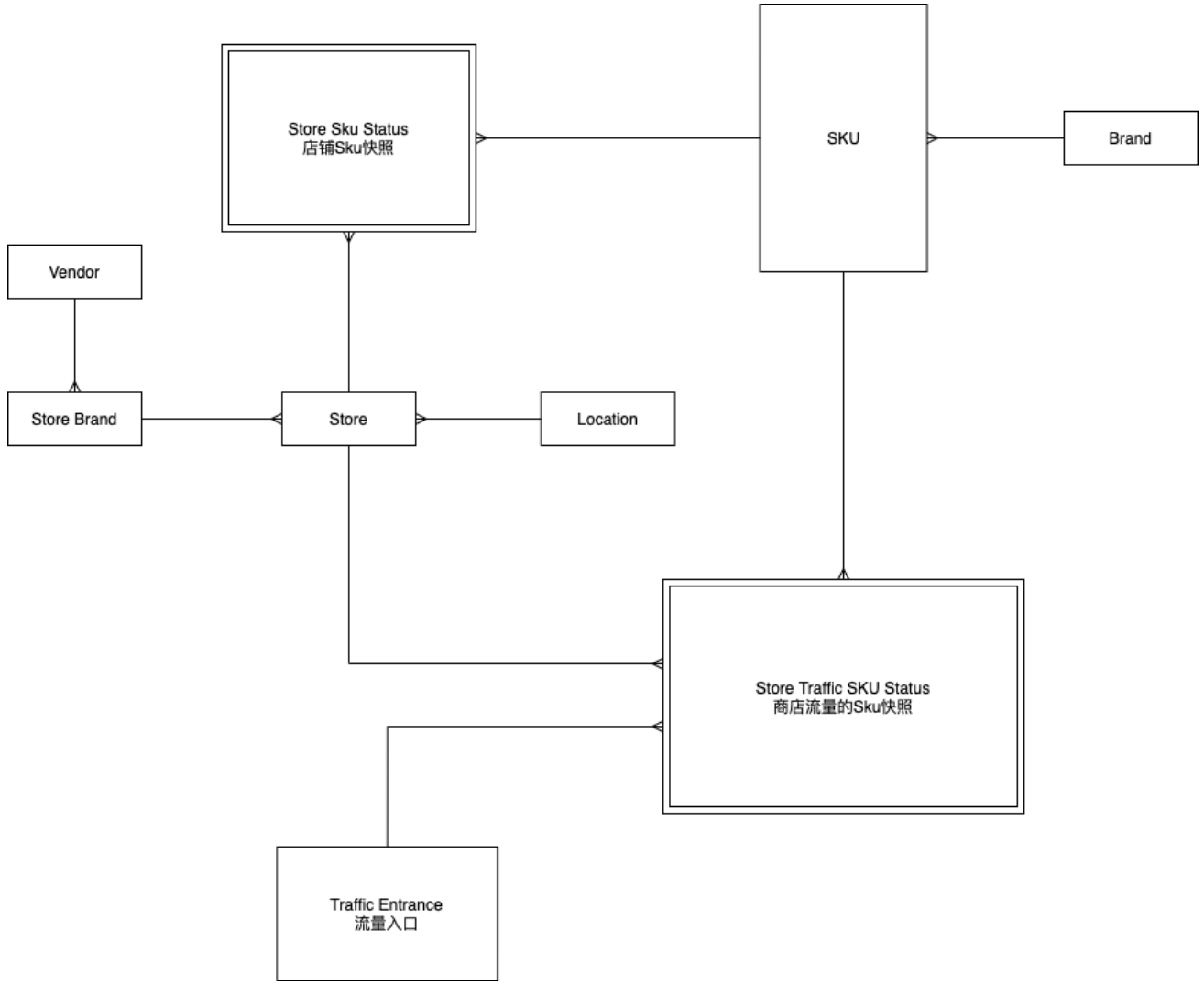
业务模型抽象：实体与关系

实体

- 数据概念抽象
- 数据属性的集合

关系

- 实体之间的关联
- 将数据连接到一起



product_id	price	snapshot_time
1	10	2020-08-01 00:00:00
1	12	2020-08-01 01:00:00
2	30	2020-08-01 00:00:00
2	40	2020-08-01 01:00:00

product_id	price	snapshot_time
1	10	2020-08-01 00:00:09
1	12	2020-08-01 01:04:05
2	30	2020-08-01 00:01:49
2	40	2020-08-01 01:03:59

左侧两张表有什么不同？

snapshot_time的口径（小时/秒）

- **定义**：数据粒度是数据主键的业务表述
- **表述**：能够准确描述即可。如每小时商品快照，也可简写为商品+小时
- 数据的粒度会影响统计分析可用的口径
- 同实体的两份数据，粒度一致才能联结
- 不同实体的两份数据，在粒度上存在包含关系（1对1或1对多）才能联结

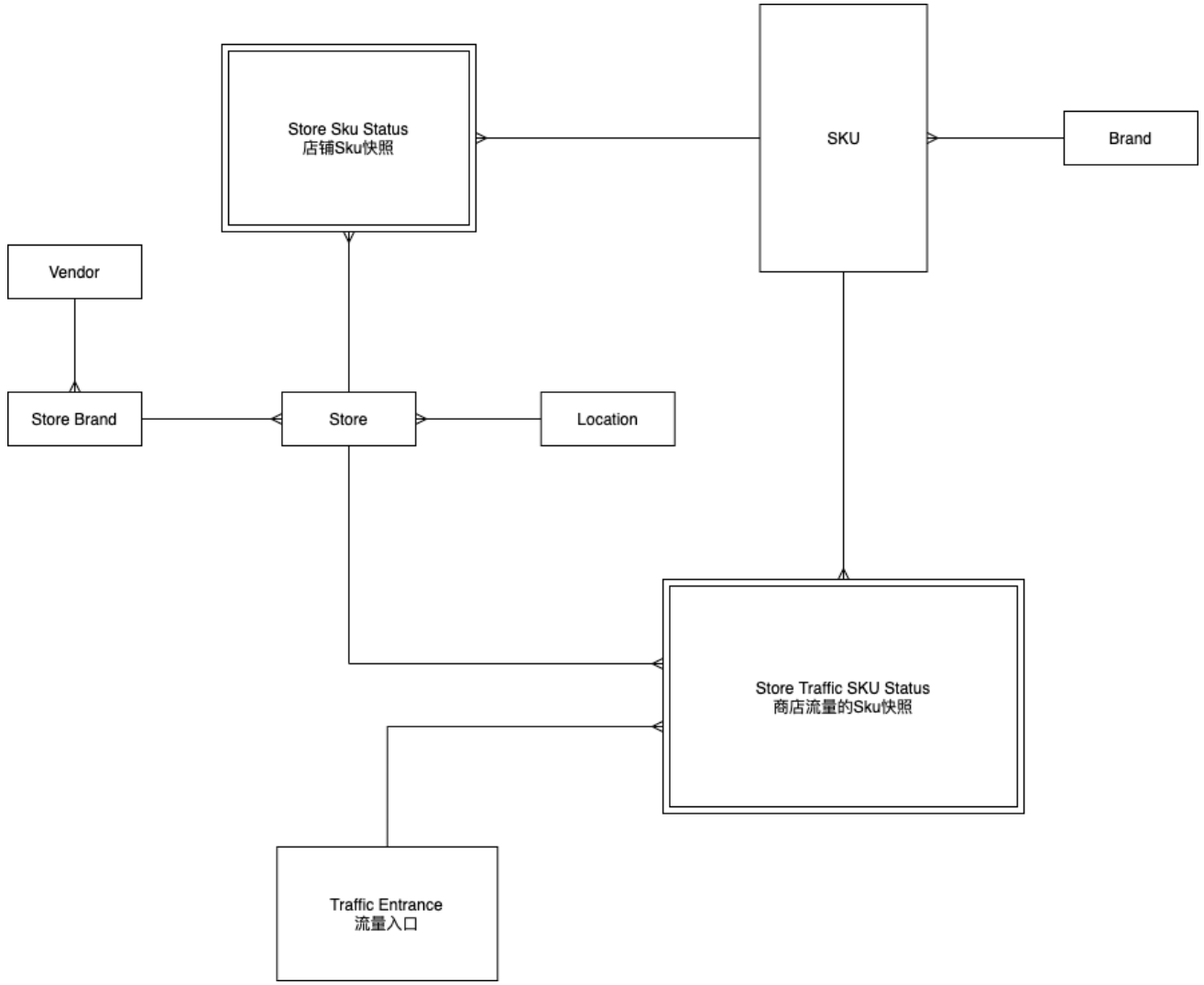
表与属性的分类

事实

- 业务活动的信息
- 对比分析-指标
- 度量—衡量业务成绩

维度

- 业务活动的环境
- 对比分析-分组
- 属性—对比不同环境



常见数据仓库模型

E-R模型

- 3NF范式
- 最少的数据冗余
- 常作为底层模型
- ETL开发便利

维度模型

- 业界最流行
- 分析端友好
- 以业务为中心

其它

- 适用特定业务场景
- 不建议过多学习

基本概念

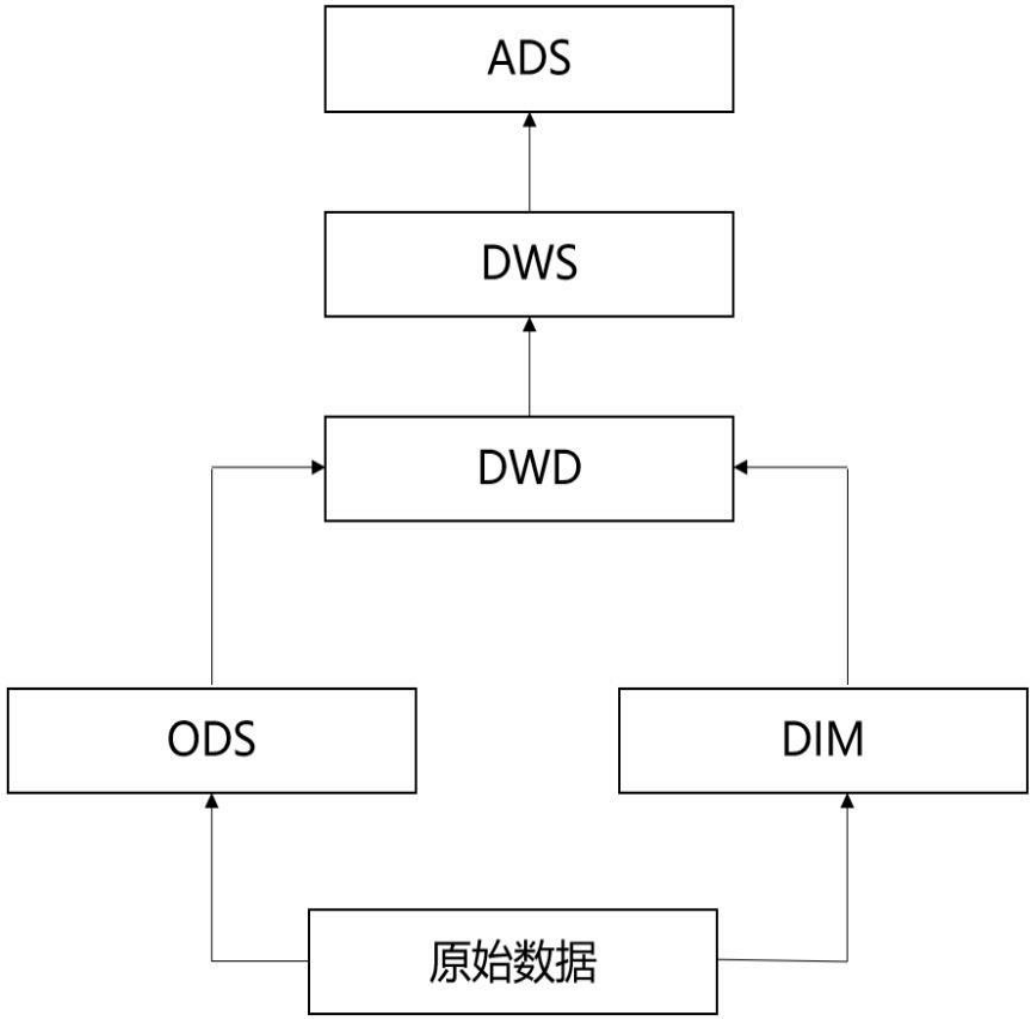
应用场景

实体关系

数据粒度

事实维度

模型理论



分层模型

将数据分成若干层，将数据处理过程解耦

- ODS：操作数据层，业务原始数据
- DWD ： 明细数据层，数仓最细粒度的业务明细信息
- DWS ： 汇总数据层，轻度聚合或宽表
- ADS： 应用数据层，供下游应用服务直接使用
- DIM： 维度表

数据仓库：开发流程

实战

开发流程

需求分析

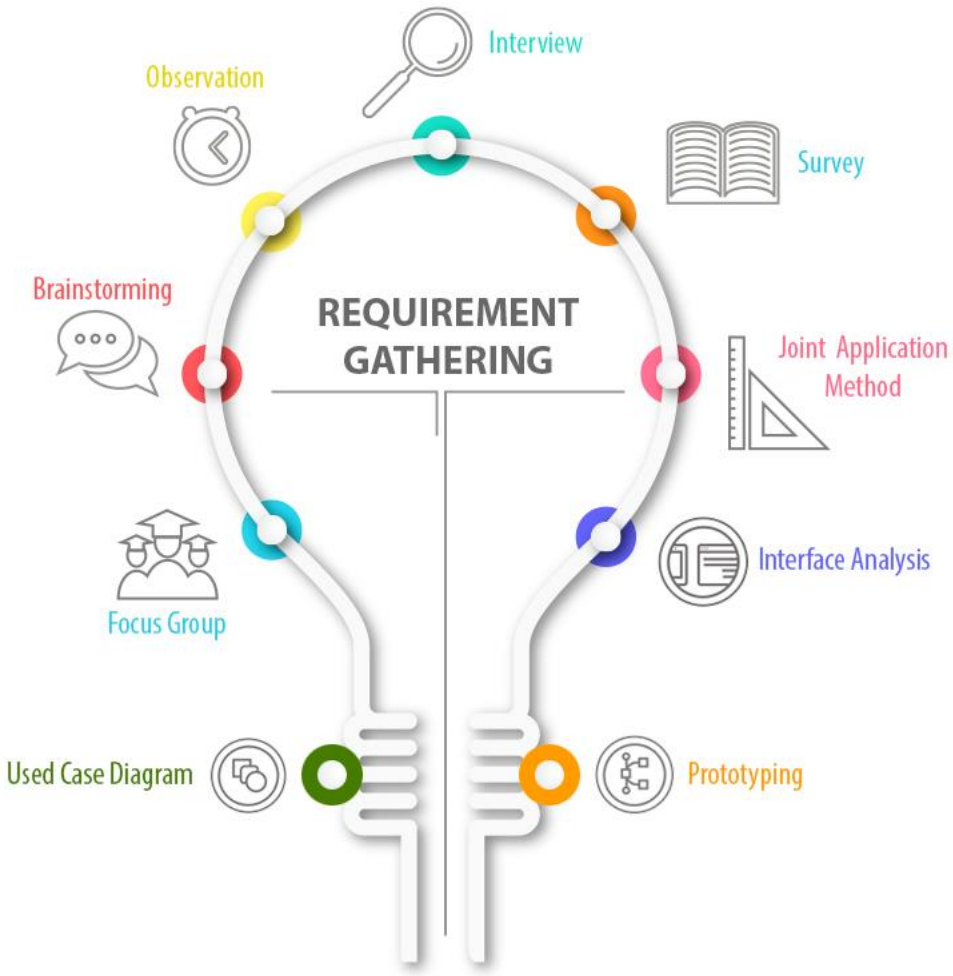
模型设计

后续流程

需求分析

理解业务，收集需求，划定开发范围

- 理解业务：通过阅读文档，试用app等方法。理解业务是如何运转的
- 收集需求：向业务端的数据使用者了解他们对数据的需求
- 划定范围：在一个周期的项目中要完成的相应工作



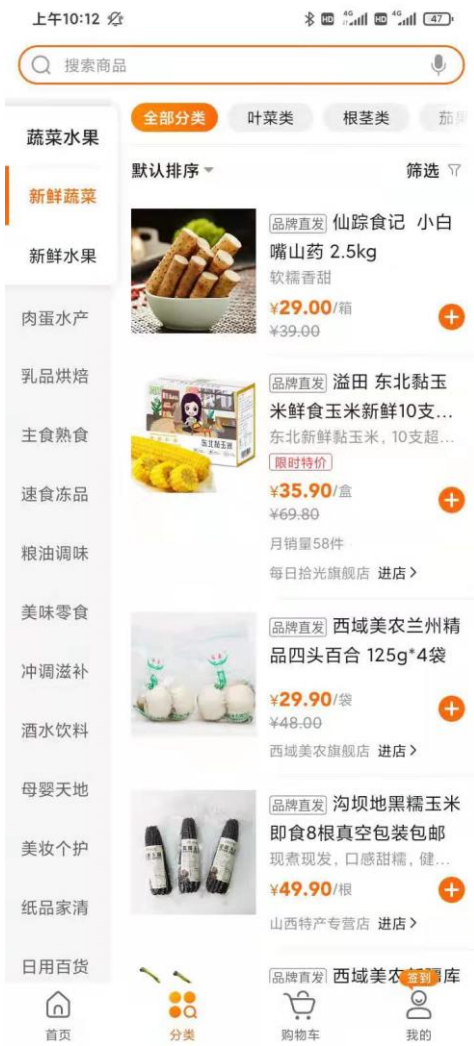
开发流程

需求分析

模型设计

后续流程

快速理解业务



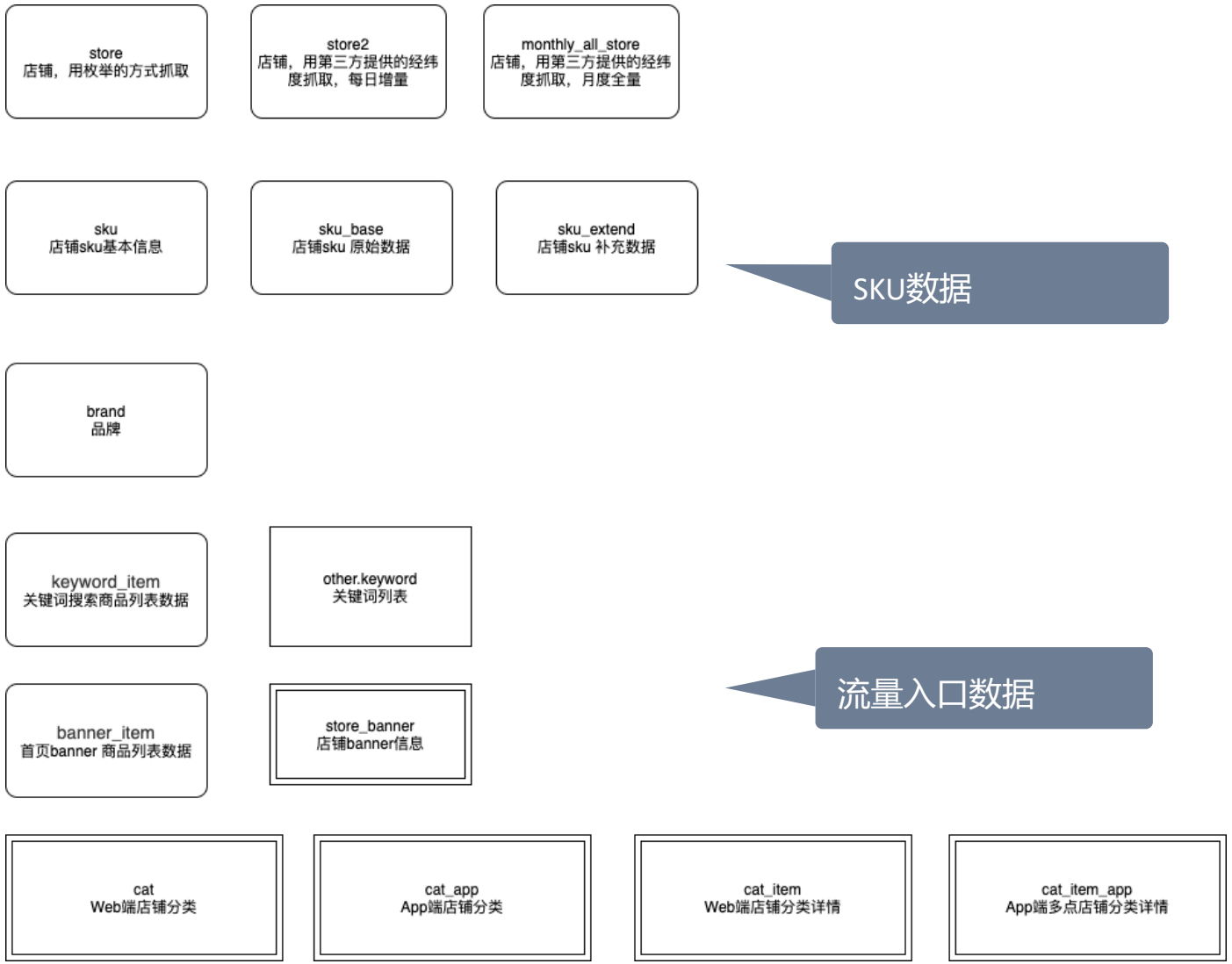
开发流程

需求分析

模型设计

后续流程

数据源



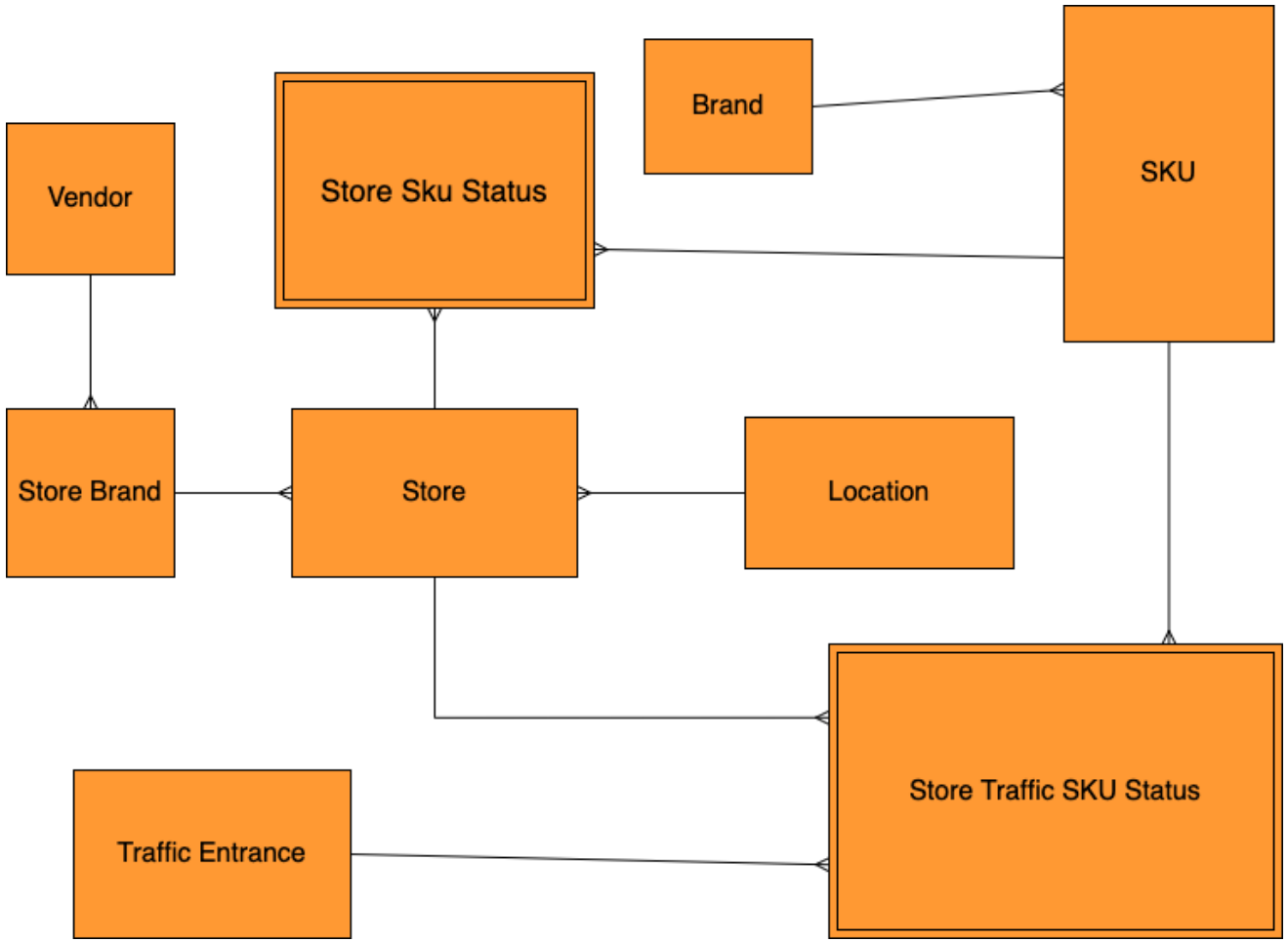
开发流程

需求分析

模型设计

后续流程

E-R模型



ER图 + 细化表设计

Store	
PK	id
FK	location_id
FK	store_brand_id
	name
	address
	phone

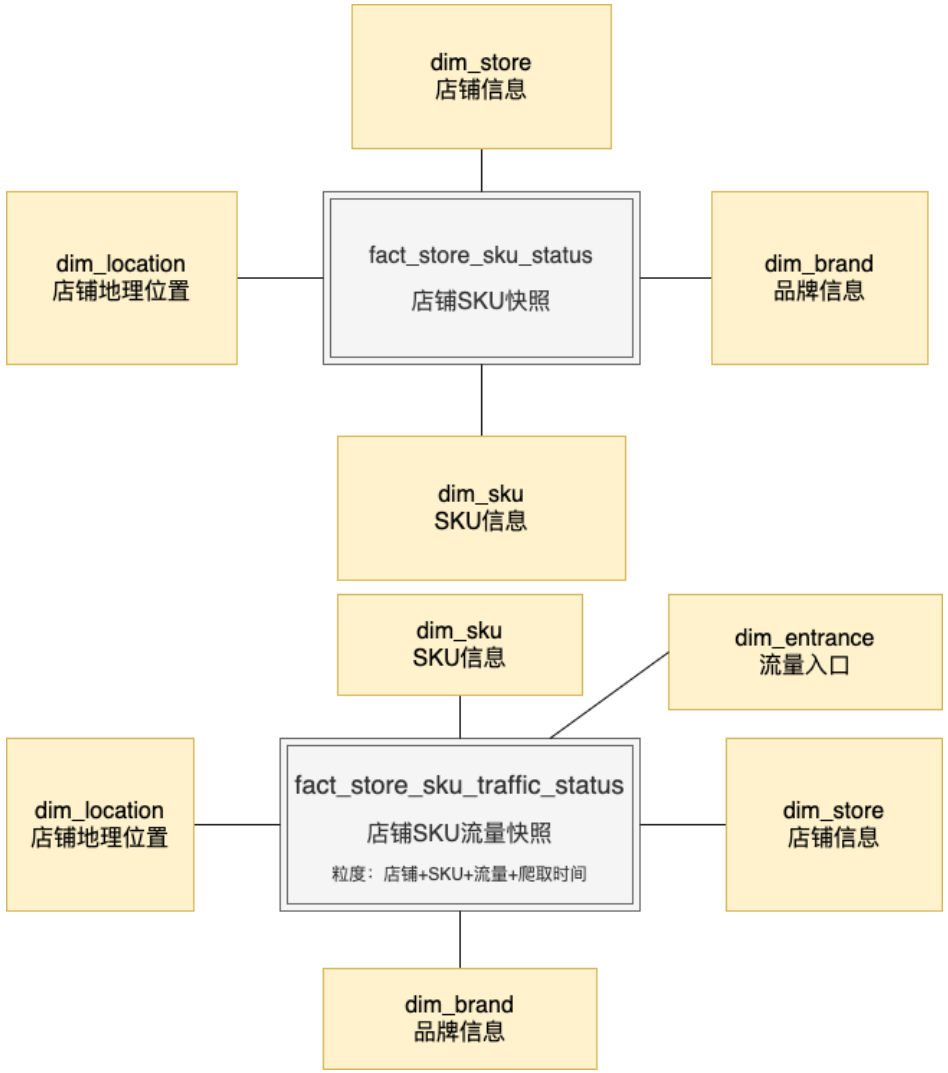
开发流程

需求分析

模型设计

后续流程

维度模型



应用维度模型理论，在ER模型的基础上优化

- 区分维度、事实表
- 事实表是分析核心，存储业务度量
- 反范式化，冗余数据

开发流程

需求分析

模型设计

后续流程

ETL开发

第二讲中涉及的内容

- 基于数仓模型设计数据流并优化
- 开发数据管道
- 自底向上开发

质量测试

交付之前需要进行质量测试，避免污染下游应用

- 基础检查应在ETL开发中完成
- 依照数据质量评价指标进行测试
- 完整性、准确性是重点
- 流程化、自动化

部署交付

将ETL流程和相应程序部署到调度系统上，持续更新

北京師範大學
BEIJING NORMAL UNIVERSITY

第四讲 算法与可视化 2021.08.21

现场弹幕答疑

复杂问题投递邮箱

xingxing_fisher@163.com