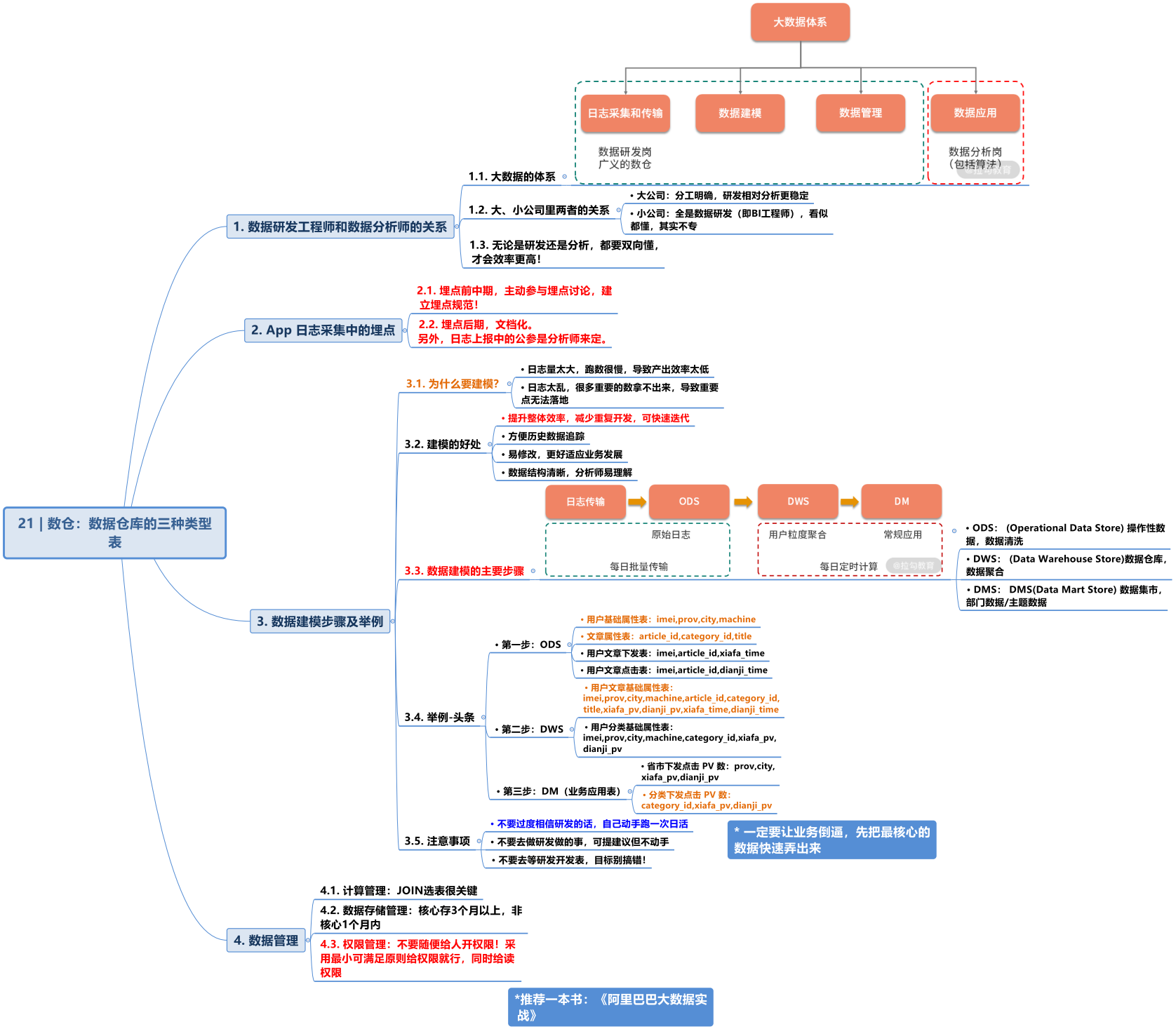
21 | 数仓：数据仓库的三种类型表
1. 数据研发工程师和数据分析师的关系
1.1. 大数据的体系
大数据体系
日志采集和传输
数据建模
数据管理
数据应用
数据研发岗
广义的数仓
数据分析岗
（包括算法）
1.2. 大、小公司里两者的关系
• 大公司：分工明确，研发相对分析更稳定
• 小公司：全是数据研发（即BI工程师），看似都懂，其实不专
1.3. 无论是研发还是分析，都要双向懂，才会效率更高！
2. App 日志采集中的埋点
2.1. 埋点前中期，主动参与埋点讨论，建立埋点规范！
2.2. 埋点后期，文档化。另外，日志上报中的公参是分析师来定。
3. 数据建模步骤及举例
3.1. 为什么要建模？
• 日志量太大，跑数很慢，导致产出效率太低
• 日志太乱，很多重要的数拿不出来，导致重要点无法落地
3.2. 建模的好处
• 提升整体效率，减少重复开发，可快速迭代
• 方便历史数据追踪
• 易修改，更好适应业务发展
• 数据结构清晰，分析师易理解
3.3. 数据建模的主要步骤
日志传输
ODS
DWS
DM
原始日志
每日批量传输
用户粒度聚合
常规应用
每日定时计算
@拉勾教育
• ODS：（Operational Data Store）操作性数据，数据清洗
• DWS：（Data Warehouse Store)数据仓库，数据聚合
• DMS：DMS(Data Mart Store) 数据集市，部门数据/主题数据
3.4. 举例-头条
• 第一步：ODS
• 用户基础属性表：imei,prov,city,machine
• 文章属性表：article_id,category_id,title
• 用户文章下发表：imei,article_id,xiafa_time
• 用户文章点击表：imei,article_id,dianji_time
• 第二步：DWS
• 用户文章基础属性表：imei,prov,city,machine,article_id,category_id,title,xiafa_pv,dianji_pv,xiafa_time,dianji_time
• 用户分类基础属性表：imei,prov,city,machine,category_id,xiafa_pv,dianji_pv
• 第三步：DM（业务应用表）
• 省市下发点击 PV 数：prov,city,xiafa_pv,dianji_pv
• 分类下发点击 PV 数：category_id,xiafa_pv,dianji_pv
3.5. 注意事项
• 不要过度相信研发的话，自己动手跑一次日活
• 不要去做研发做的事，可提建议但不动手
• 不要去等研发开发表，目标别搞错！
* 一定要让业务倒逼，先把最核心的数据快速弄出来
4. 数据管理
4.1. 计算管理：JOIN选表很关键
4.2. 数据存储管理：核心存3个月以上，非核心1个月内
4.3. 权限管理：不要随便给人开权限！采用最小可满足原则给权限就行，同时给读权限
*推荐一本书：《阿里巴巴大数据实战》