2019

分享主题：企业级数据仓库介绍

《木东居士》

一个专注数据科学的公众号，分享数据相关的技术干货、思考感悟和工作经验

木东居士不属于任何培训机构，分享嘉宾均是各个岗位上的资深工程师，我们将不定期在公众号推送相关分享内容，你可以通过扫描如下二维码关注我们，快来加入我们吧！

目录

Contents

数据仓库的三个阶段

第一阶段

使用大量成熟的开源框架、主要以离线批处理为主，外围系统自研能力较弱，数据量和集群资源比较少

第二阶段

使用开源+自研方式，有自己的方法论和建模体系，有比较完善的元数据管理、数据质量监控，能有效支持离线实时需求

第三阶段

有自研的通用的一站式大数据处理平台、有完善的数仓理论基础和外围工具，有完善的数据共享机制和权限管理

趋势：

工具越来越智能，平台越来越完善

实时和离线一体化，技术不再是障碍

数据膨胀速度比以往任何时候都快，吞噬了大量的计算资源

数据仓库的痛点

痛点一：临时取数需求占用数仓人员大部分时间

痛点二：数仓规范和流程不一致，跨部门合作困难

痛点三：指标口径不一致导致数据可信度下降

痛点四：烟囱式开发形成的数据孤岛与重复计算

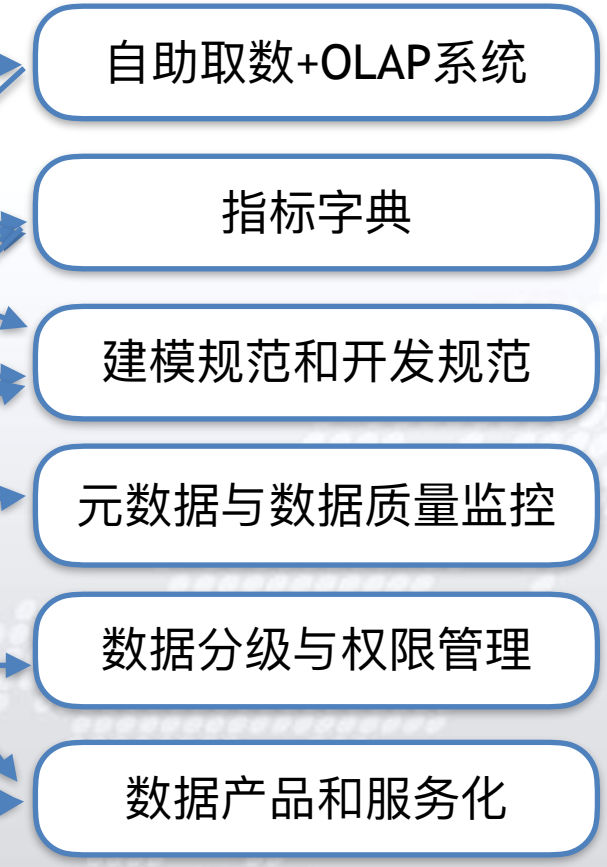
痛点五：数据膨胀导致计算资源紧张，出数时间得不到保障

痛点六：异常排查时间和修复时间长

痛点七：数据安全和数据共享矛盾不可调和

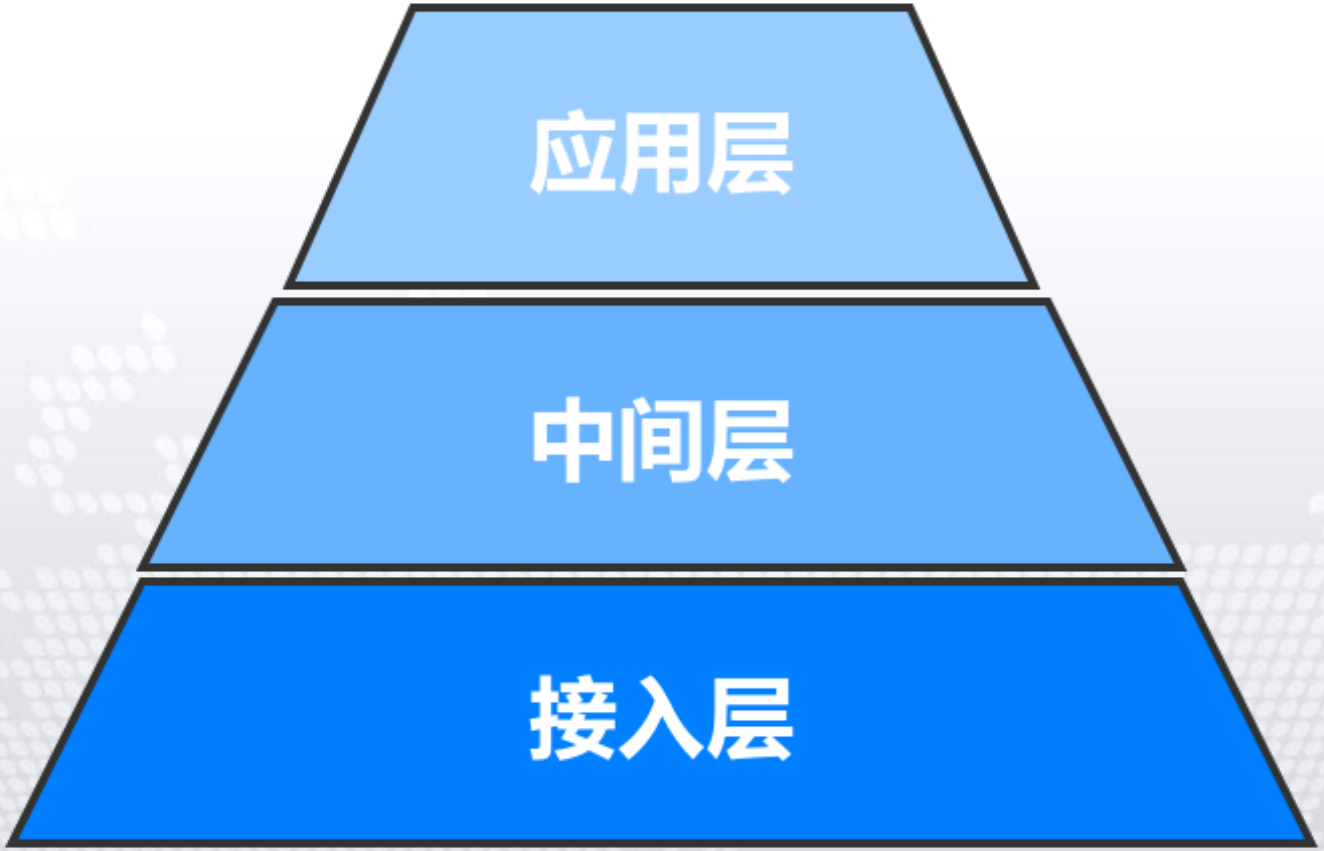
痛点八：产出形式单一

痛点九：业务需求响应不及时

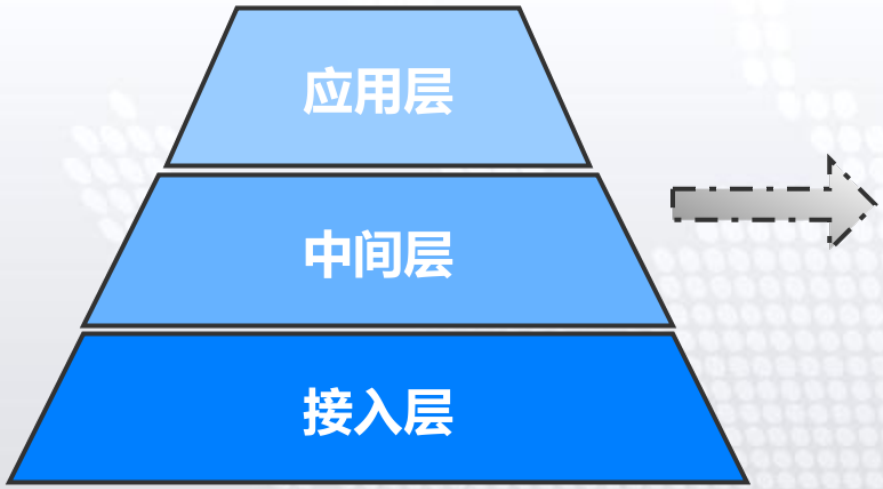


目录

Contents


应用层
中间层
接入层

数仓模型



APP数据应用层(Application Model)

TMP临时表

DWM数据集市层(Data Warehouse Market)

DWS数据汇总层(Data Warehouse Summary)

DWD数据明细层(Data Warehouse Detail)

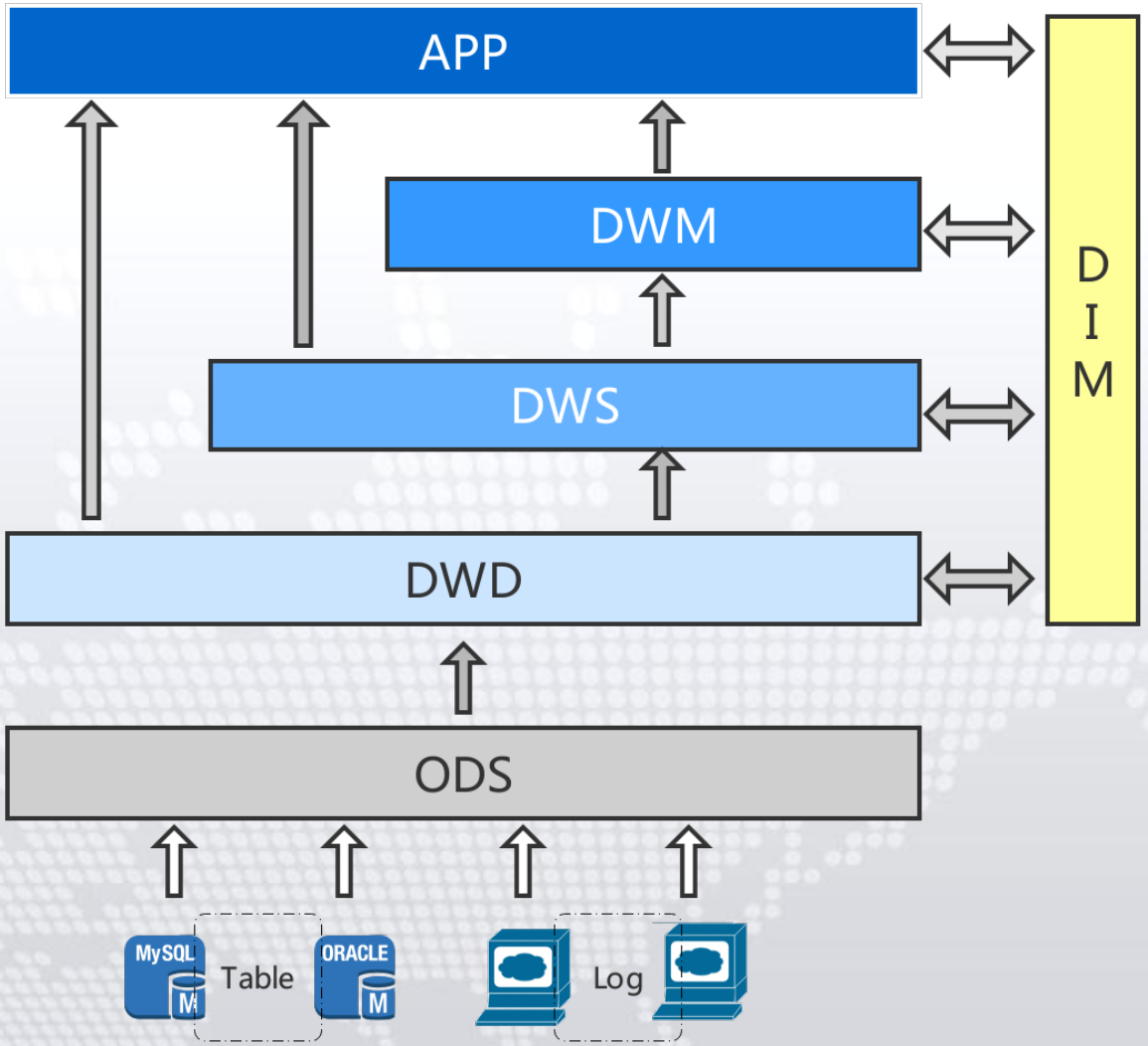
DIM维表层

ODS数据接入层(Operational Data Store)

数仓模型

APP数据应用层(Application Model)	个性化指标加工、基于应用的数据组装
DWM数据集市层(Data Warehouse Market)	宽表集市、跨过业务场景、行为数据组装
DWS数据汇总层(Dara Warehouse Summary)	单业务场景、行为数据组装、提升公共指标的复用
DWD数据明细层(Data Warehouse Detail)	标准化、维度补全、异常处理
DIM维表层(Dimension)	一致性维度建设
ODS数据接入层(Operational Data Store)	数据同步、基本保持与源数据格式一致，不过过多校验

调用原则

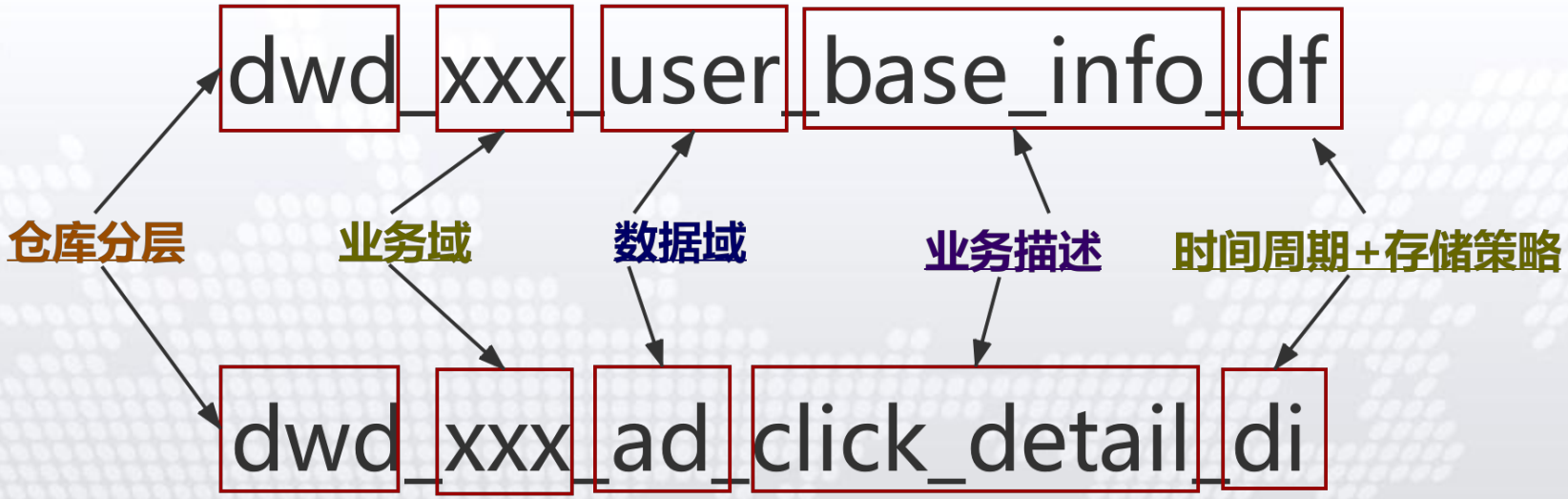


总原则：
禁止逆向调用
避免同层调用
优先使用公共层
避免跨层调用

目录

Contents

- 表命名规范
- 字段命名规范
- 需求对接规范
- 数据开发规范



属性字段	文本字段，使用通用单词即可
指标字段	修饰词+原子词+时间修饰
计数字段	<计数主体>_cnt
比例字段	<计数主体>_rate
分区字段	日/周/月/季度/年：ds 小时分区：hh 分钟分区：mm 业务分区：业务描述文本
费用字段	<计数主体>_amt
标识字段	is_<标识主体>
时间字段	<业务主体>_time
日期字段	<业务主体>_date

01

指定库和队列

声明使用的库

指定本账号所属的队列

02

表创建

外部表、ORC

\t分割、Location指定

03

分区加载

删除分区

加载分区

指定分区Location

04

UDF/临时表创建

删除UDF/临时表

创建UDF/临时表

05

数据写入

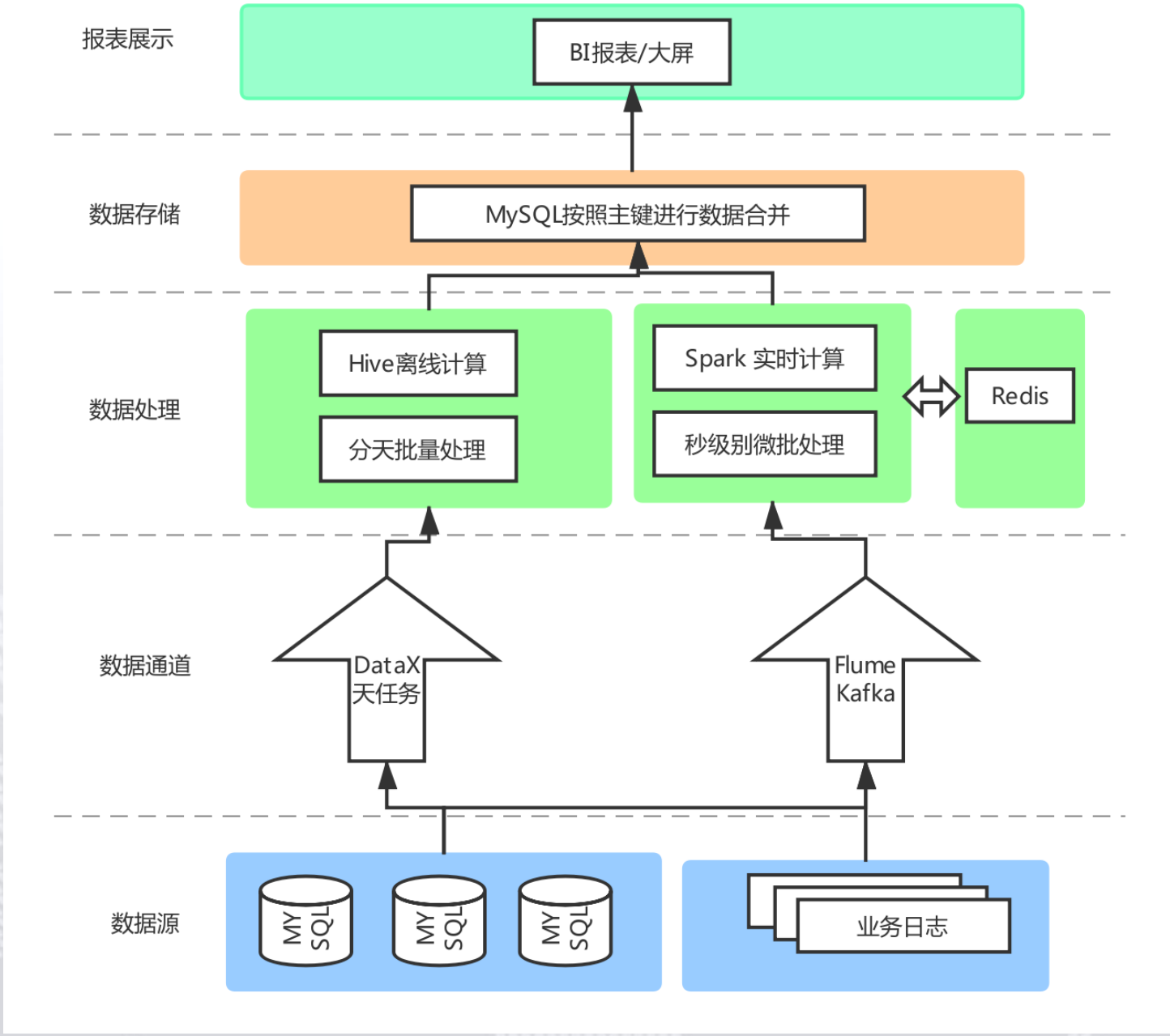
通过覆盖的方式写入

06

临时表删除

清理临时表

Lambda架构


报表展示
BI报表/大屏
数据存储
MySQL按照主键进行数据合并
数据处理
Hive离线计算
分天批量处理
Spark 实时计算
秒级别微批处理
Redis
数据通道
DataX
天任务
Flume
Kafka
数据源
MY SQL
MY SQL
MY SQL
业务日志

目录

Contents

外围系统建设

01 调度系统

02 元数据管理系统

03 通用离线和实时计算平台

04 数据质量监控

元数据管理系统

元数据管理系统是外部了解数仓的门户入口，一个好的元数据系统至少应包含如下信息：

1、**表信息**：包括表英文名、中文注释、表状态(在线&下线)
2、**字段信息**：包括字段类型、英文名、中文名、字段注释、保密级别(机密/保密/一般)、统计逻辑说明
3、**负责人信息**：业务/开发负责人名超链接、所在部门
4、**分区信息**：分区名、分区大小、分区记录条数、生成分区的时间
5、**血缘信息**：表上游、下游节点信息
6、**代码信息**：生成该表对应的代码地址超链接
7、**存储信息**：总表大小、波动情况
8、**热度信息**：标识被下游依赖的多寡
9、**权限信息**：申请访问超链接、权限审批到单人单表单字段粒度，不同保密级别字段对应不同审批流程
10、**使用注意事项QA**等

数据质量监控

数据质量监控系统主要基于规则判断达到数据监控的目的，系统建设一般分为三个阶段

1、表级别的监控：主要为表的总条数、总大小、分区数据、各分区条数，各分区大小，条数/大小同环比，日增长情况等

2、字段级别监控：枚举值异常判断、特殊值判断、范围判断等

3、全链路数据监控：主要依赖于上下游血缘分析，自动判断跟踪故障点，并及时告知相关负责人

其中，表级别和字段级别的监控是比较常规且易实现的监控方式，全链路数据监控比这两者要复杂很多，涉及到从源数据–>数据通道–>数据ETL–>数据展示的全过程

目录

Contents

数据产品化：

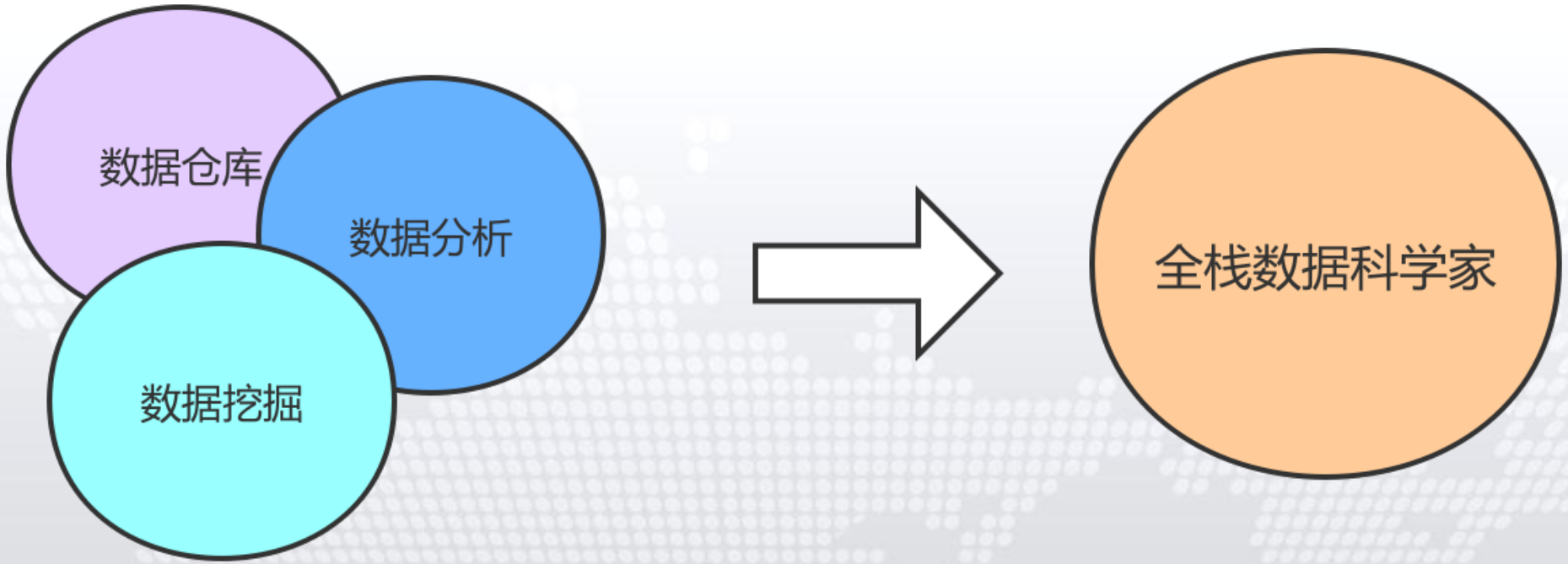
- 面向管理层的宏观经营分析系统
- 面向运营人员的业务监控报表系统
- 面向广告以及营销的一体化数据营销平台

其中营销平台涉及用户圈选+用户触达+日志回流+效果分析(实验组+对照组)

数据服务化：

- 数据以接口方式直接服务于线上业务
- 数据以共享平台方式提供基础标签服务

趋势二：单一技能变多项技能



《木东居士》

一个专注数据科学的公众号，分享数据相关的技术干货、思考感悟和工作经验

木东居士不属于任何培训机构，分享嘉宾均是各个岗位上的资深工程师，我们将不定期在公众号推送相关分享内容，你可以通过扫描如下二维码关注我们，快来加入我们吧！

THANK YOU FOR YOUR GUIDANCE.

谢谢