AI Pioneer

2018 AI Pioneer Conference

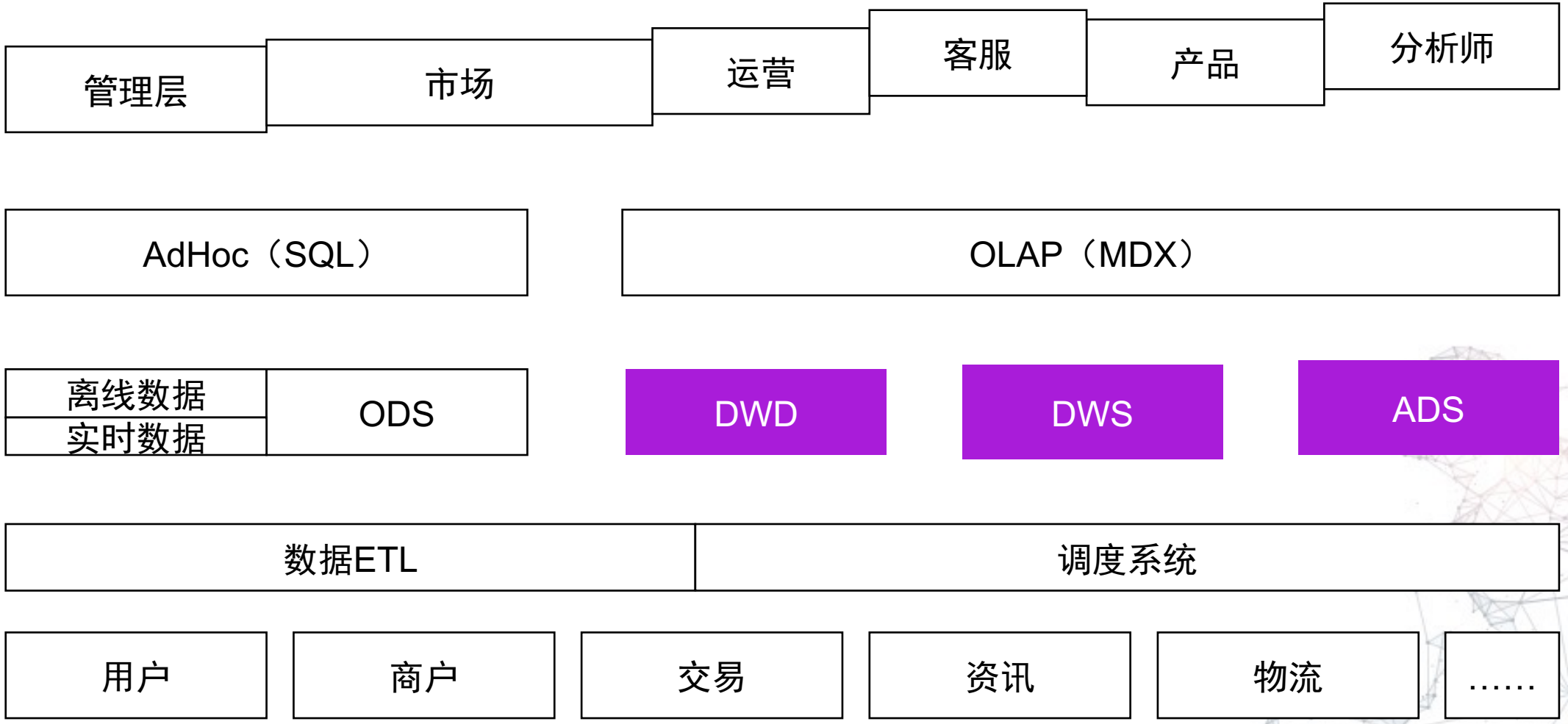
大数据指标模型治理与实践

主讲人：梁福坤

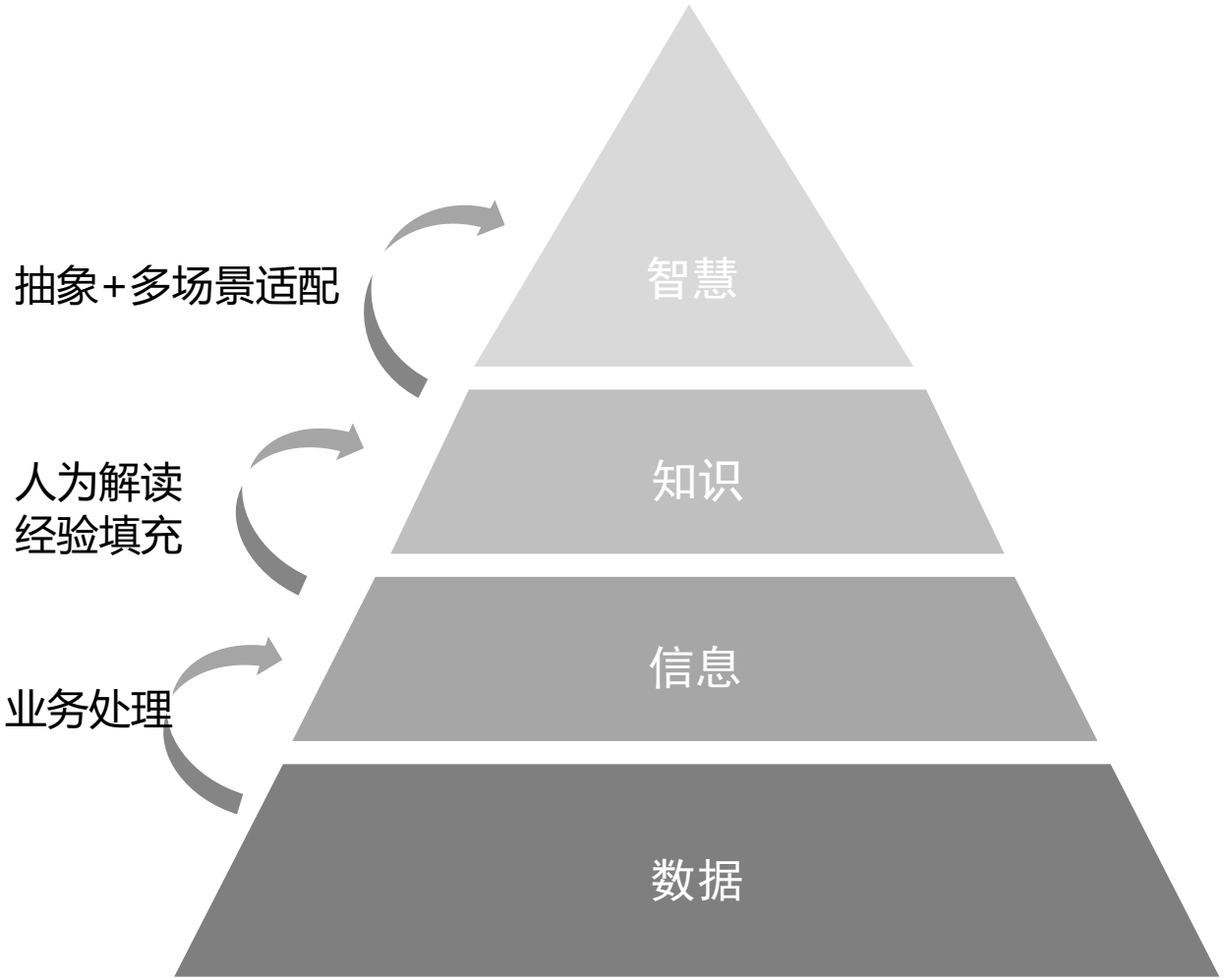
目录

目录

指标在大数据生态中的位置



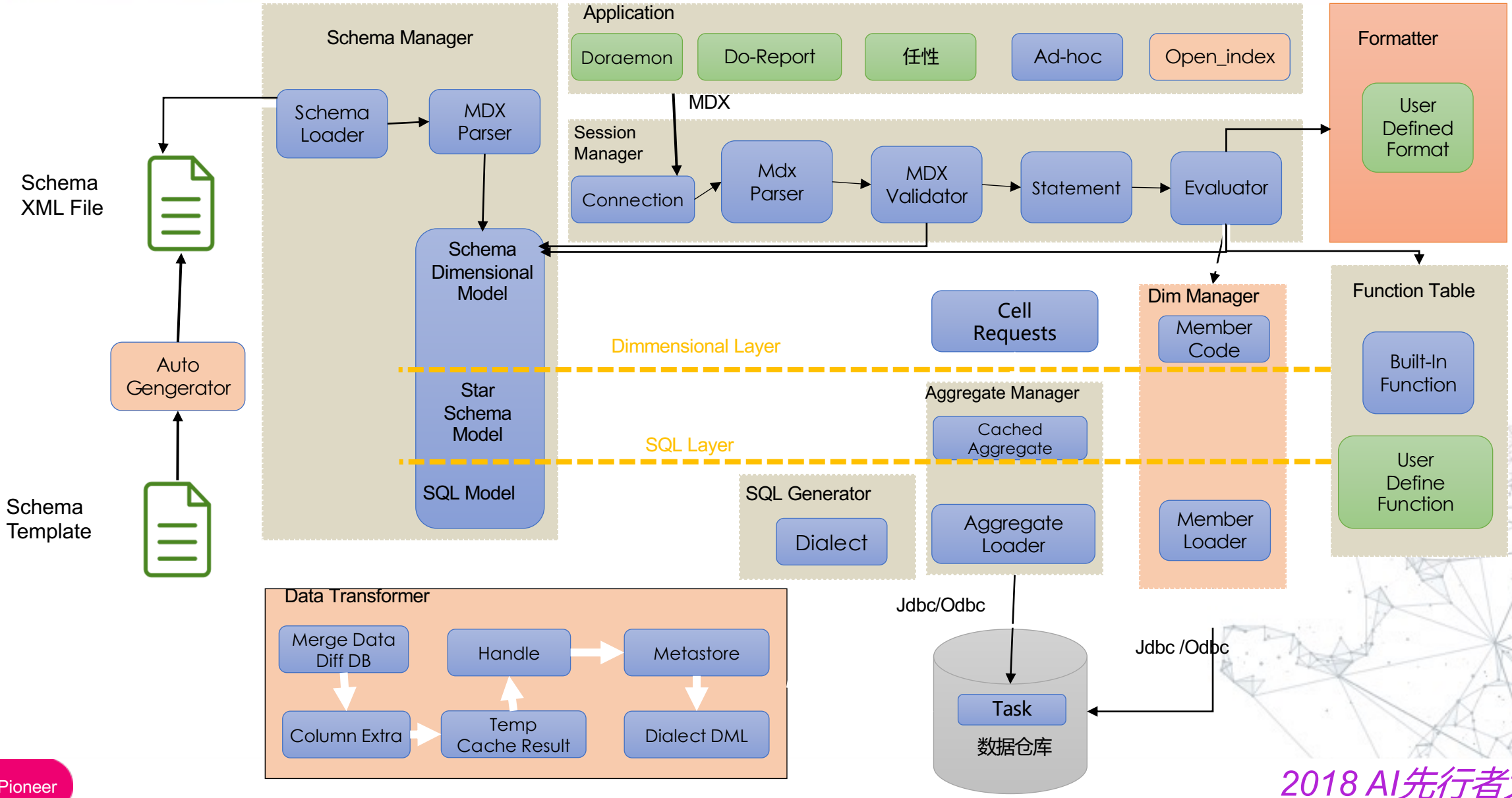
业务意义驱动



数据已经爆炸了，信息却仍稀缺~

- 从数据到信息的积累
- 多方指标定义规范统一
- 最小成本的交付业务
- 数据的多角度归属
- 公共逻辑下沉
- 业务权限统一

OLAP对外服务整体架构图（Mondrian extend）

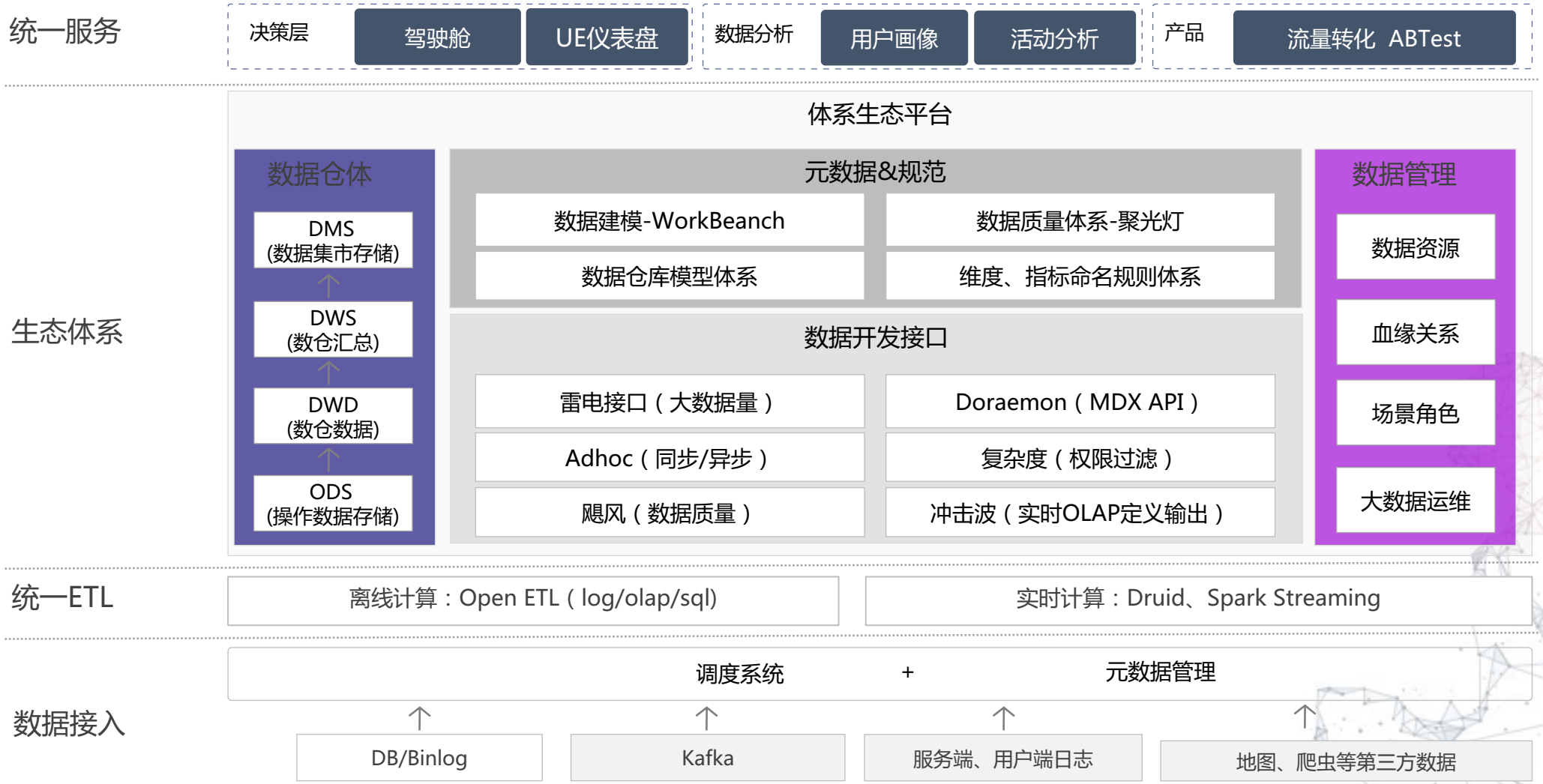


目录

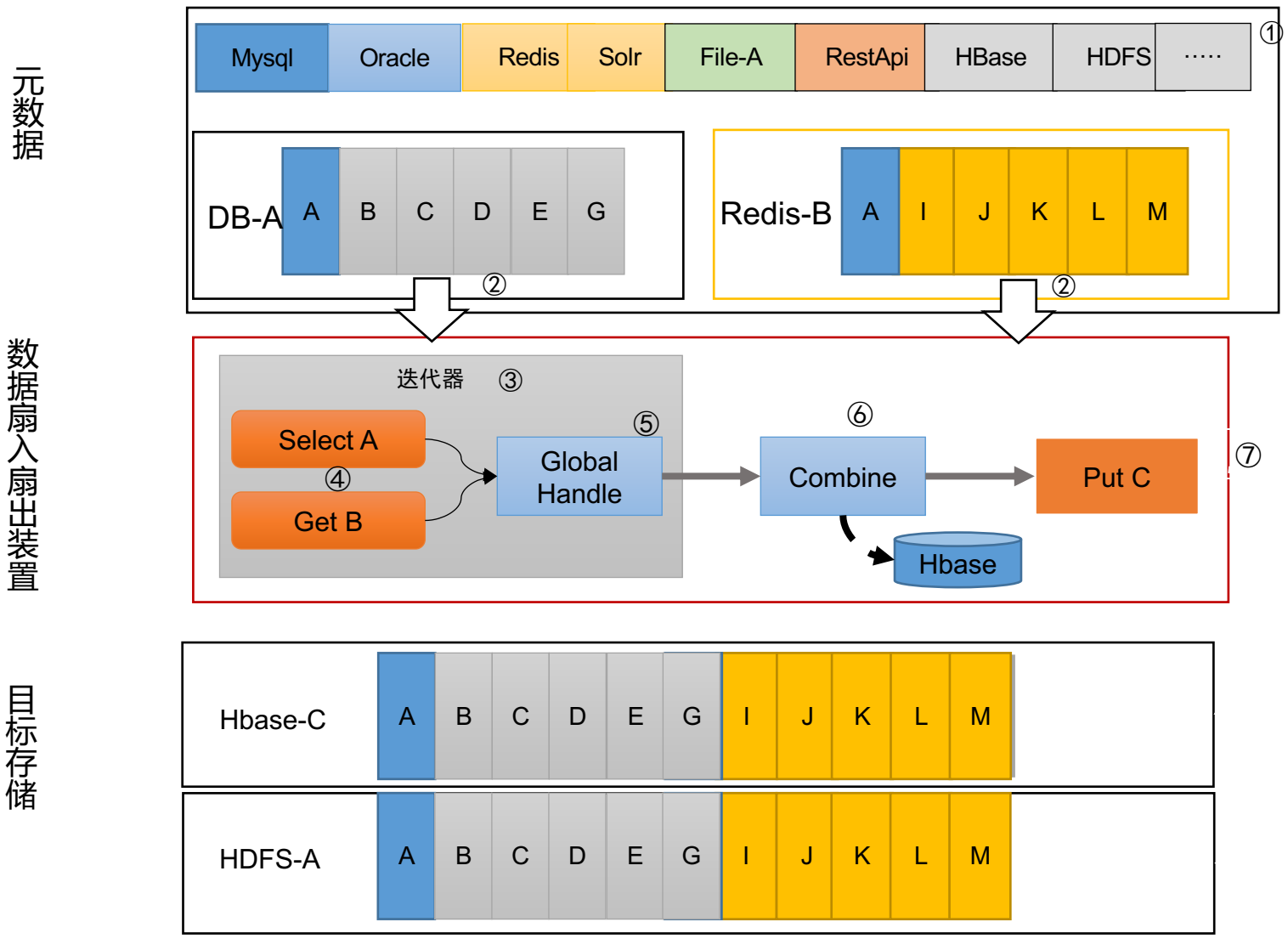
数据要秉承宁缺毋滥的准入机制

在数据采集之初完善元数据管理，在数据仓库阶段统筹业务含义，在数据流转过程中跟踪血缘关系！

以工具化生态支撑


统一服务
决策层
驾驶舱
UE仪表盘
数据分析
用户画像
活动分析
产品
流量转化 ABTest
生态体系
体系生态平台
数据仓体
DMS
(数据集市存储)
DWS
(数仓汇总)
DWD
(数仓数据)
ODS
(操作数据存储)
元数据&规范
数据建模-WorkBeanch
数据质量体系-聚光灯
数据仓库模型体系
维度、指标命名规则体系
数据开发接口
雷电接口（大数据量）
Doraemon（MDX API）
Adhoc（同步/异步）
复杂度（权限过滤）
飓风（数据质量）
冲击波（实时OLAP定义输出）
数据管理
数据资源
血缘关系
场景角色
大数据运维
统一ETL
离线计算：Open ETL（log/olap/sql）
实时计算：Druid、Spark Streaming
数据接入
调度系统
+
元数据管理
DB/Binlog
Kafka
服务端、用户端日志
地图、爬虫等第三方数据

数据血缘关系基础



离线场景：

抽象服务的取存场景

批量、批次的设计

暂停点的缓存周期

异构数据转换

引擎方言兼容

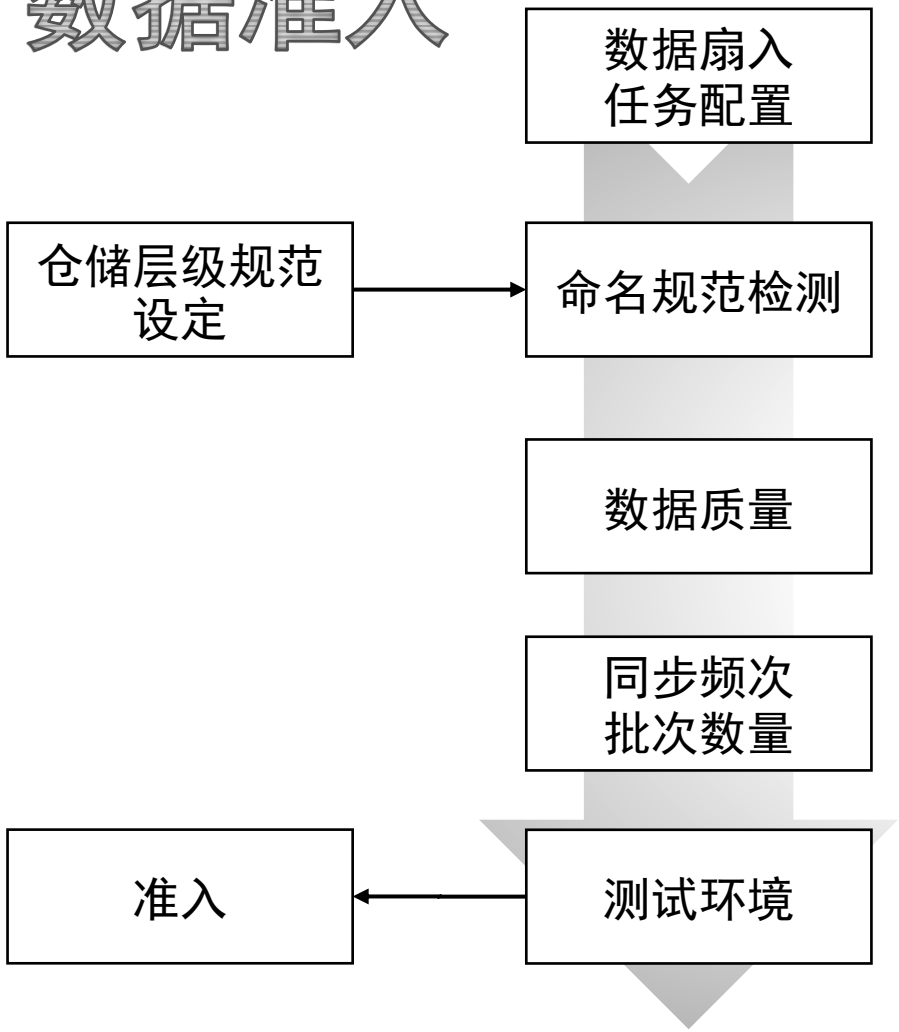
血缘关系：

细粒度的上下游依赖

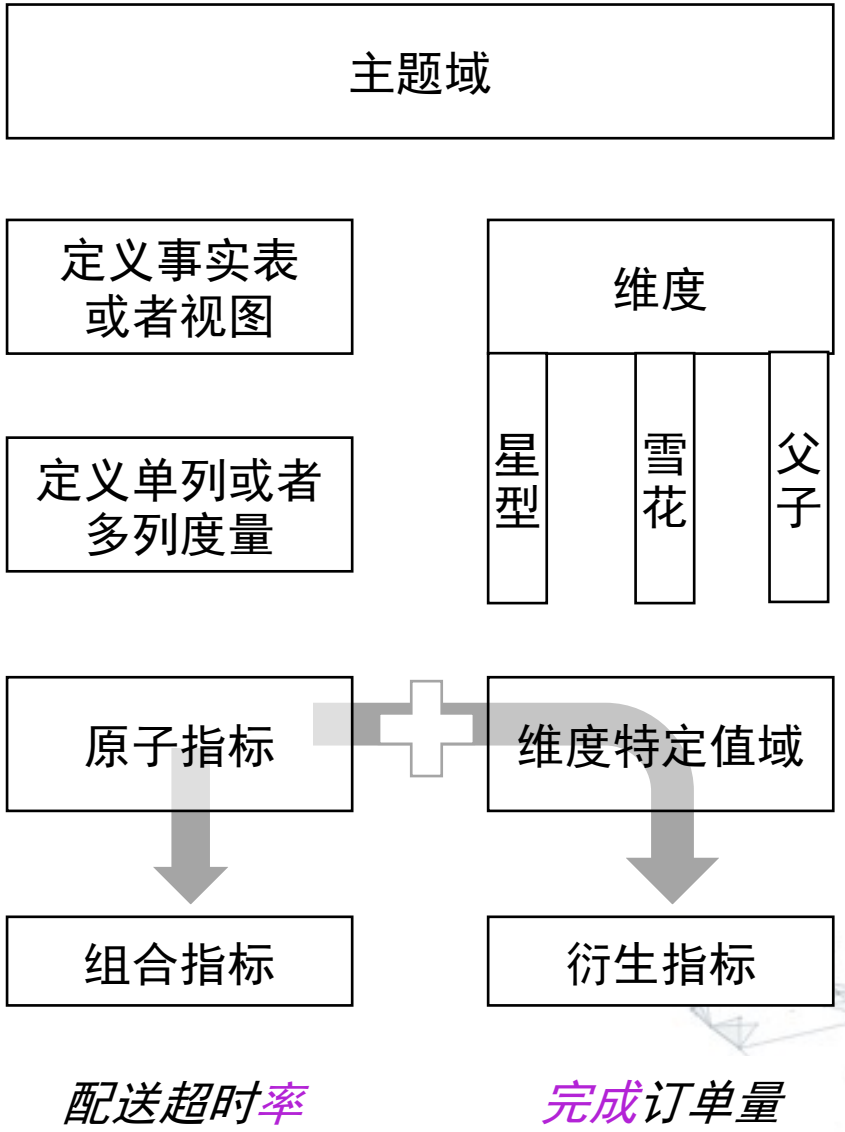
调度系统DAG

指标体系规范

数据准入



指标体系



目录

建模工具

- Workbench
- PowerDesigner

数据转换工具

- 开放式Log
- 开放式SQL
- 开放式OLAP
- 血缘关系分析

元数据管理

数据展示工具

- 数据集市
- Saiku

元数据存储

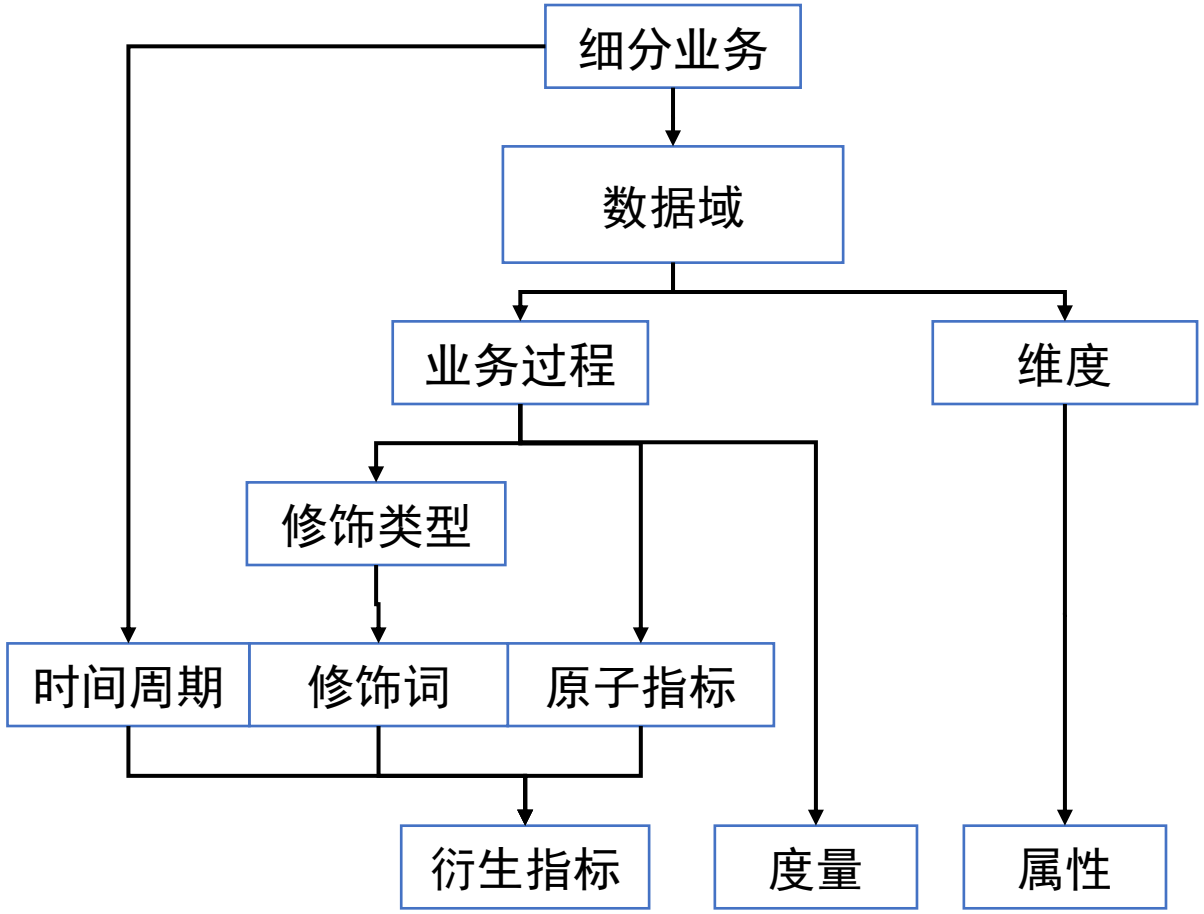
- HiveMeta
  - 主题域
  - 指标库
  - 字典库

元数据是大数据价值的基础

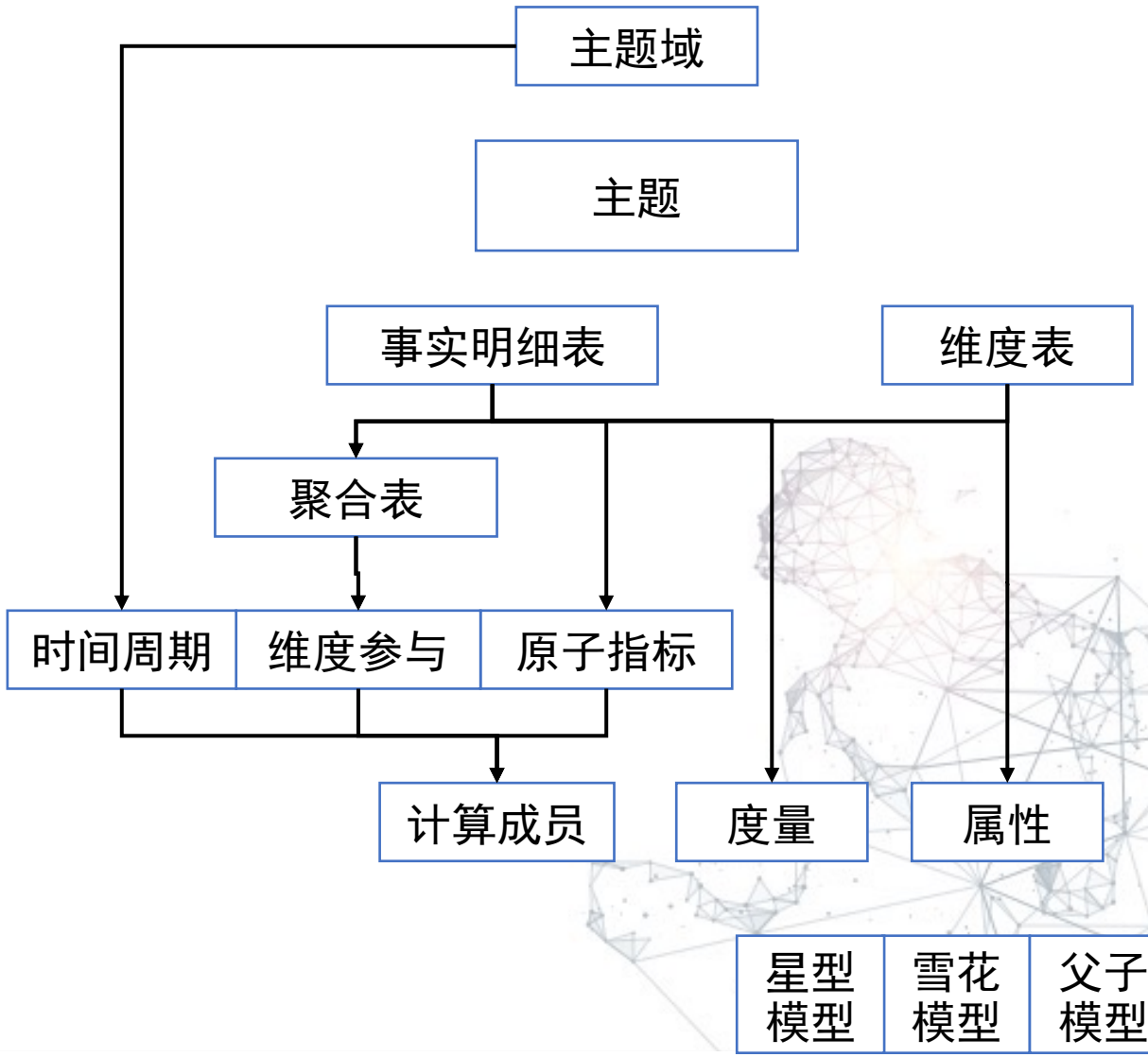
- 先有元数据后有数据生态
- 从数据到信息必需
- 质量保证的关键
- 数据流转语义可追溯

分析定义与实现

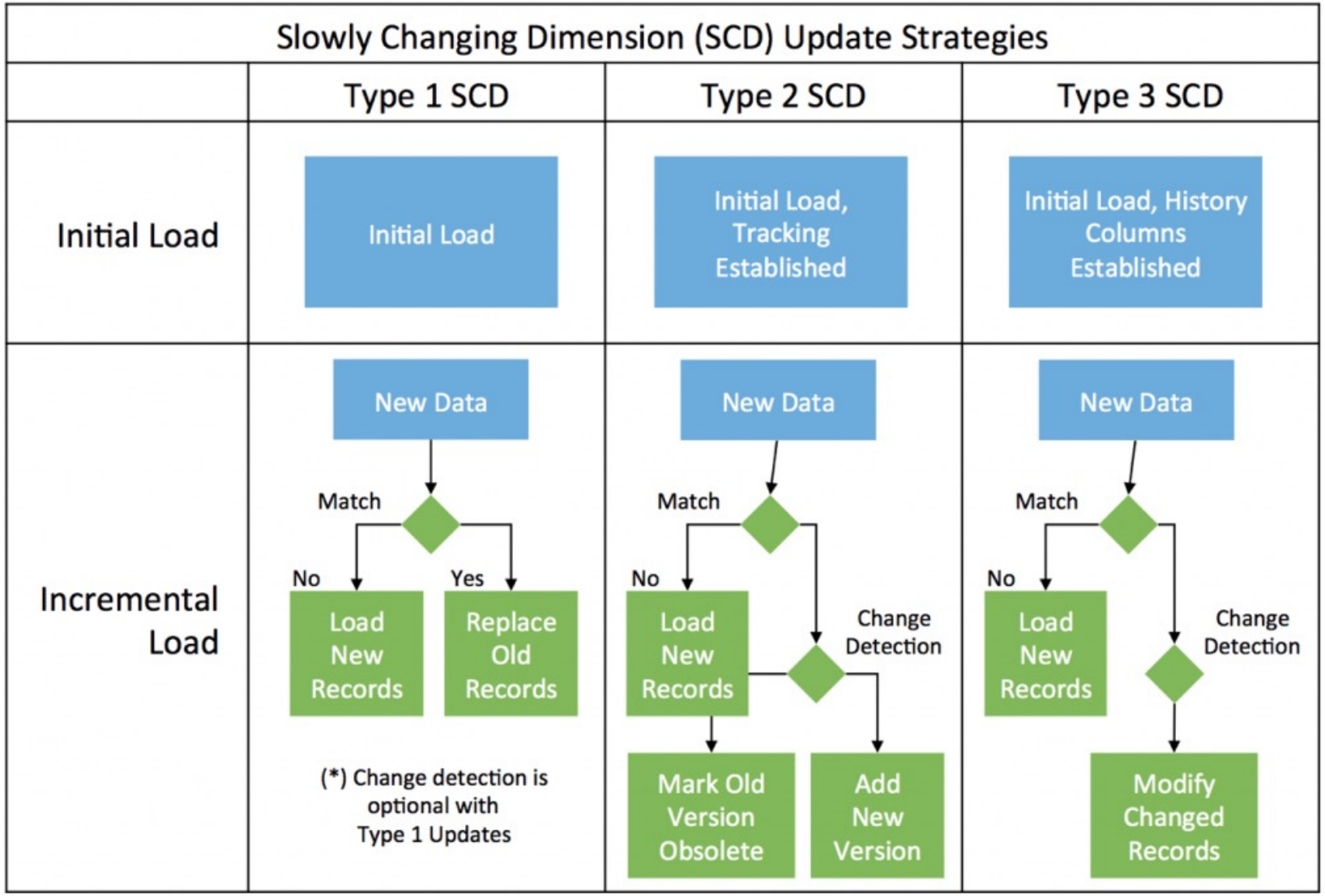
业务定义



分析实现



缓慢变化维



- 类型一：直接更新列
- 类型二：新增变更记录
- 类型三：更新列+Previous记录

引用自：https://hortonworks.com/blog/update-hive-tables-easy-way-2/

面向主题立方体

```
<Schema>
        <Cube name="Sales">
                <Table name="sales_fact_1997"/>
                <Dimension foreignKey="customer_id" name="Gender">
                        <Hierarchy allMemberName="All Genders" hasAll="true" primaryKey="customer_id">
                                <Table name="customer"/>
                                <Level column="gender" name="Gender" uniqueMembers="true"/>
                            </Hierarchy>
                    </Dimension>
                <Dimension foreignKey="time_id" name="Time">
                        <Hierarchy hasAll="false" primaryKey="time_id">
                                <Table name="time_by_day"/>
                                <Level column="the_year" name="Year" type="Numeric" uniqueMembers="true"/>
                                <Level column="quarter" name="Quarter" uniqueMembers="false"/>
                                <Level column="month_of_year" name="Month" type="Numeric" uniqueMembers="false"/>
                            </Hierarchy>
                    </Dimension>
                <Measure aggregator="sum" column="unit_sales" formatString="#,###" name="Unit Sales"/>
                <Measure aggregator="sum" column="store_sales" formatString="#,###.##" name="Store Sales"/>
                <Measure aggregator="sum" column="store_cost" formatString="#,###.00" name="Store Cost"/>
                <CalculatedMember dimension="Measures" formula="[Measures].[Store Sales] - [Measures].[Store Cost]" name="Profit">
                        <CalculatedMemberProperty name="FORMAT_STRING" value="$#,##0.00"/>
                </CalculatedMember>
        </Cube>
</Schema>
```

Schema来自： https://mondrian.pentaho.com/documentation/schema.php

公共维度

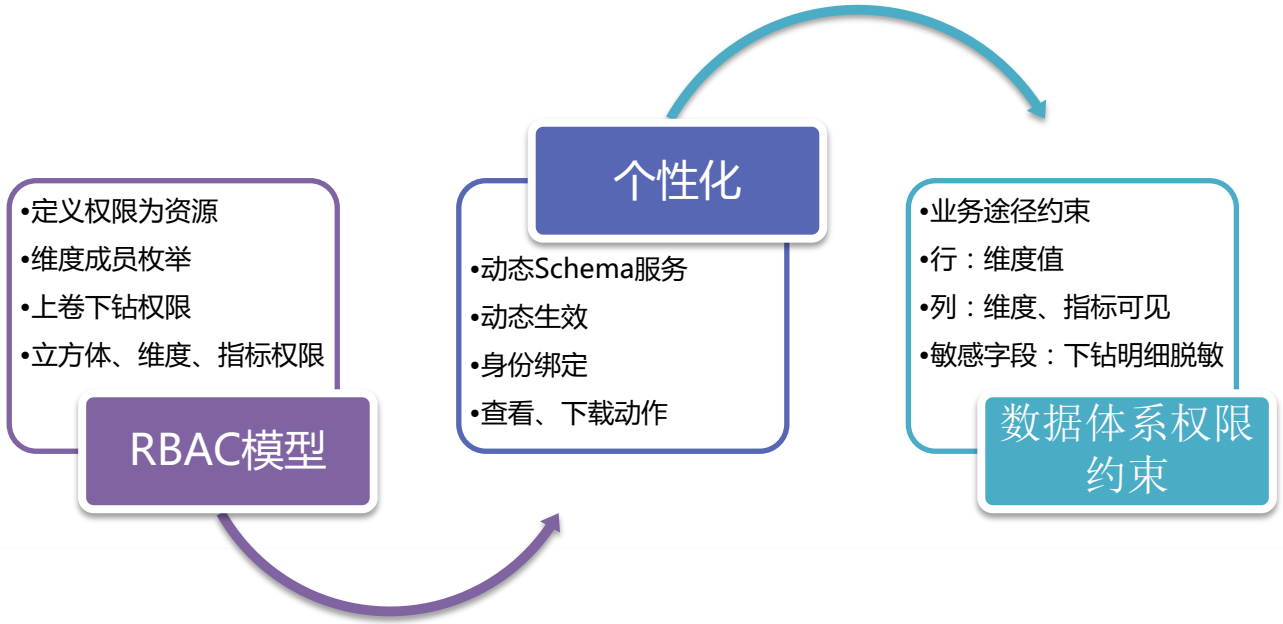
- 独立于单独立方体存在的维度
- 服务于立方体、虚拟立方体

```
<DimensionUsage source="Wuliu Aoi" name="Wuliu Aoi" caption="物流商圈" visible="true" foreignKey="wuliu_aoi_id" highCardinality="false"/>
<DimensionUsage source="Shop Source" name="Shop Source" caption="商户来源" visible="true" foreignKey="shop_id" highCardinality="false"/>
<DimensionUsage source="Shop Type" name="Shop Type" caption="商户类型" visible="true" foreignKey="shop_id" highCardinality="false"/>
<DimensionUsage source="Wuliu  City Role" name="Wuliu  City Role" visible="false" foreignKey="city_id" highCardinality="false"/>
<DimensionUsage source="KA Flag Role" name="KA Flag Role" visible="false" foreignKey="shop_id" highCardinality="false"/>
```

权限

```
<Role name="ROLE_USER">
  <SchemaGrant access="all">
    <CubeGrant cube="Order" access="all">
      <HierarchyGrant hierarchy="[City Role.Default]" rollupPolicy="partial" access="custom">
        <MemberGrant member="[City Role.Default].[City].[@DYNAMIC_CITY@]" access="all"/>
      </HierarchyGrant>
      <HierarchyGrant hierarchy="[Is Wuliu.Default]" rollupPolicy="partial" access="custom">
        <MemberGrant member="[Is Wuliu.Default].[YN].&amp;[@DYNAMIC_WULIU_FLAG@]" access="all"/>
      </HierarchyGrant>
      <HierarchyGrant hierarchy="[KA Label.Default]" rollupPolicy="partial" access="custom">
        <MemberGrant member="[KA Label.Default].[ka_label].[@DYNAMIC_KA_LABEL@]" access="all"/>
      </HierarchyGrant>
    </CubeGrant>
  </SchemaGrant>
```

- 权限体系完整
- 动态化、个性化
- 以业务语言定义



File View Tools Help

Opened

admin

Saiku Analytics

Cubes

SteelWheelsSales

Dimensions

- Customers
- Markets
- Order Status
  - (All)
  - Type
- Product
- Time

Measures

- Measures
  - Quantity
  - Sales

Columns: Sales, Type

Rows: Country

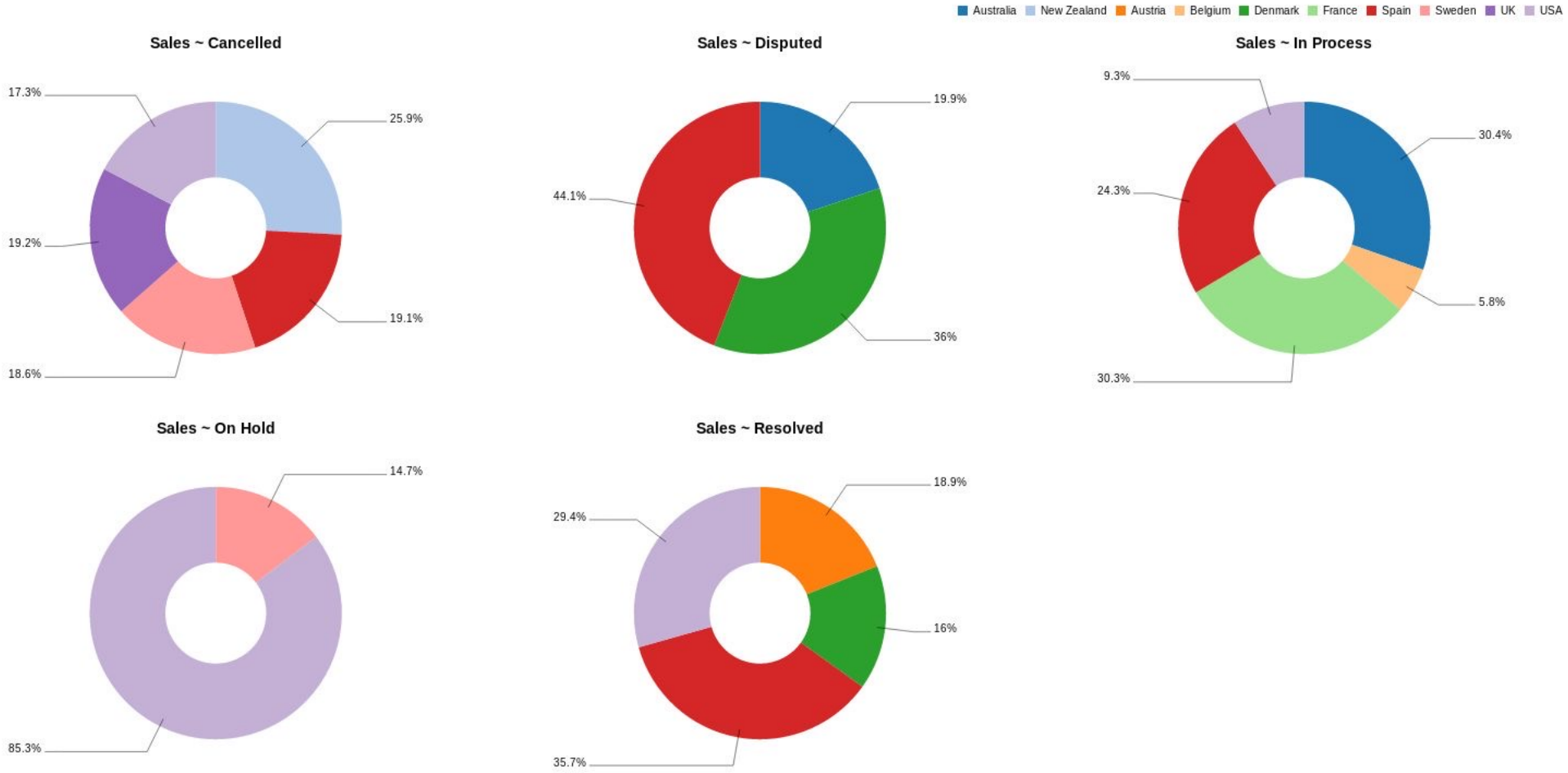
Filter

Info: 11:52 / 6 x 12 / 0.05s

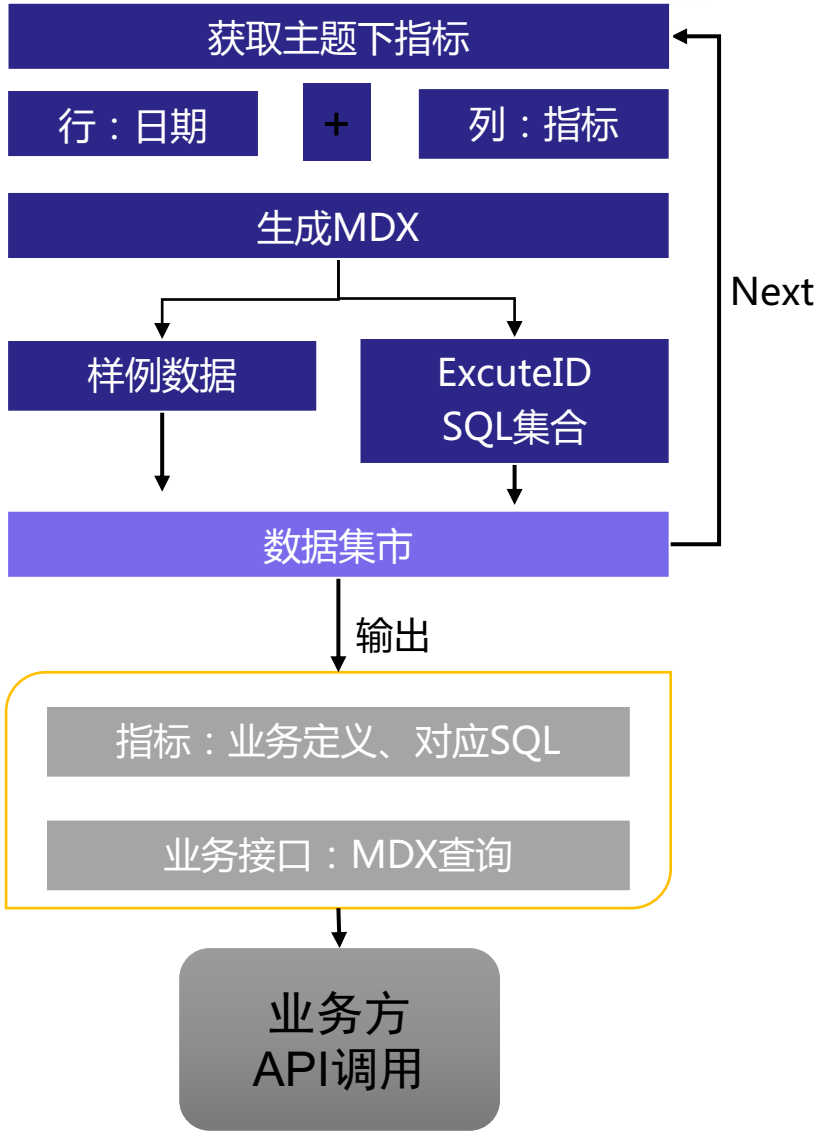
Export

Treemap

Sunburst



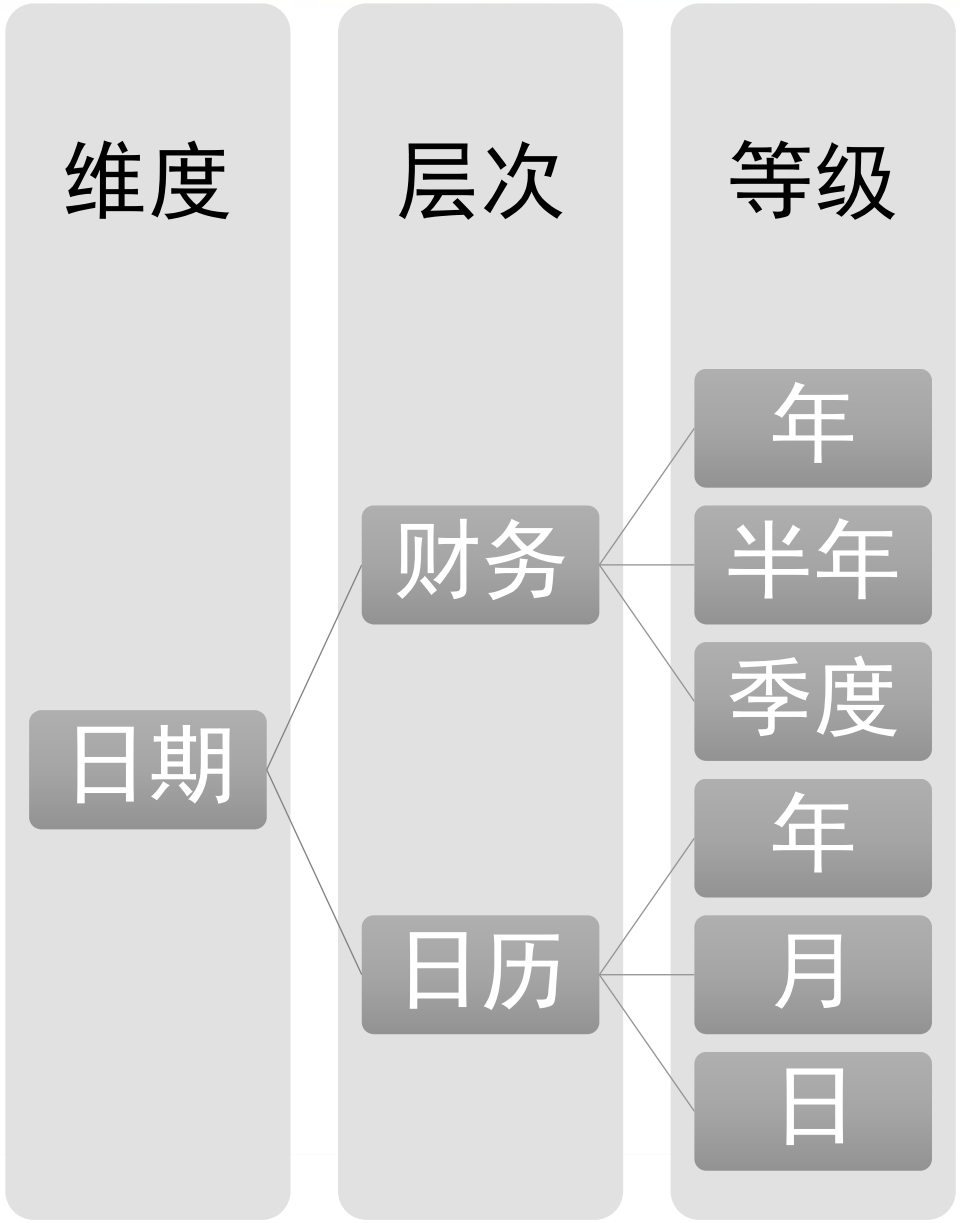
指标业务建模统一

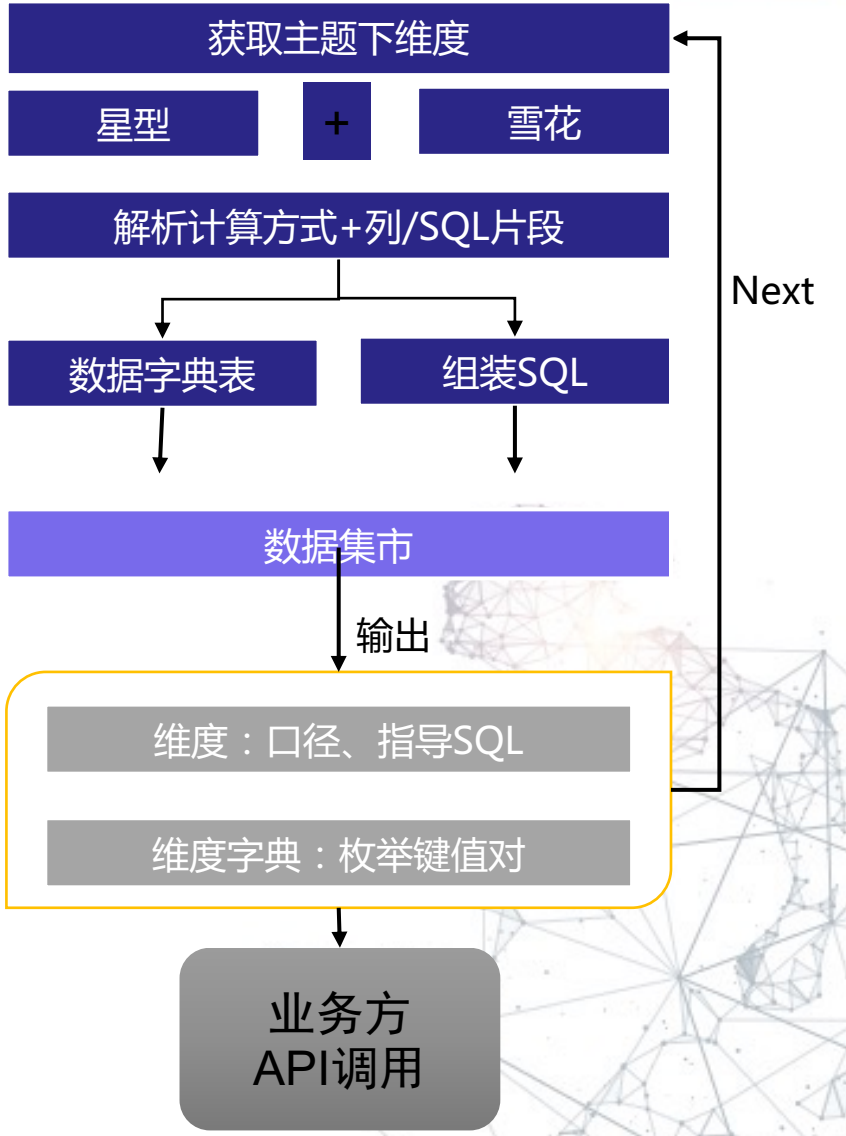


一处定义处处引用

- 业务沉淀以建模后作为输出
- 指标输出指导SQL
- 提供API调用业务语言，摒弃多变得SQL

维度的建模和统一


维度
层次
等级
日期
财务
日历
年
半年
季度
年
月
日


获取主题下维度
星型
+
雪花
解析计算方式+列/SQL片段
数据字典表
组装SQL
数据集市
输出
维度：口径、指导SQL
维度字典：枚举键值对
业务方
API调用
Next

常见问题1：明细表和汇总表的抉择

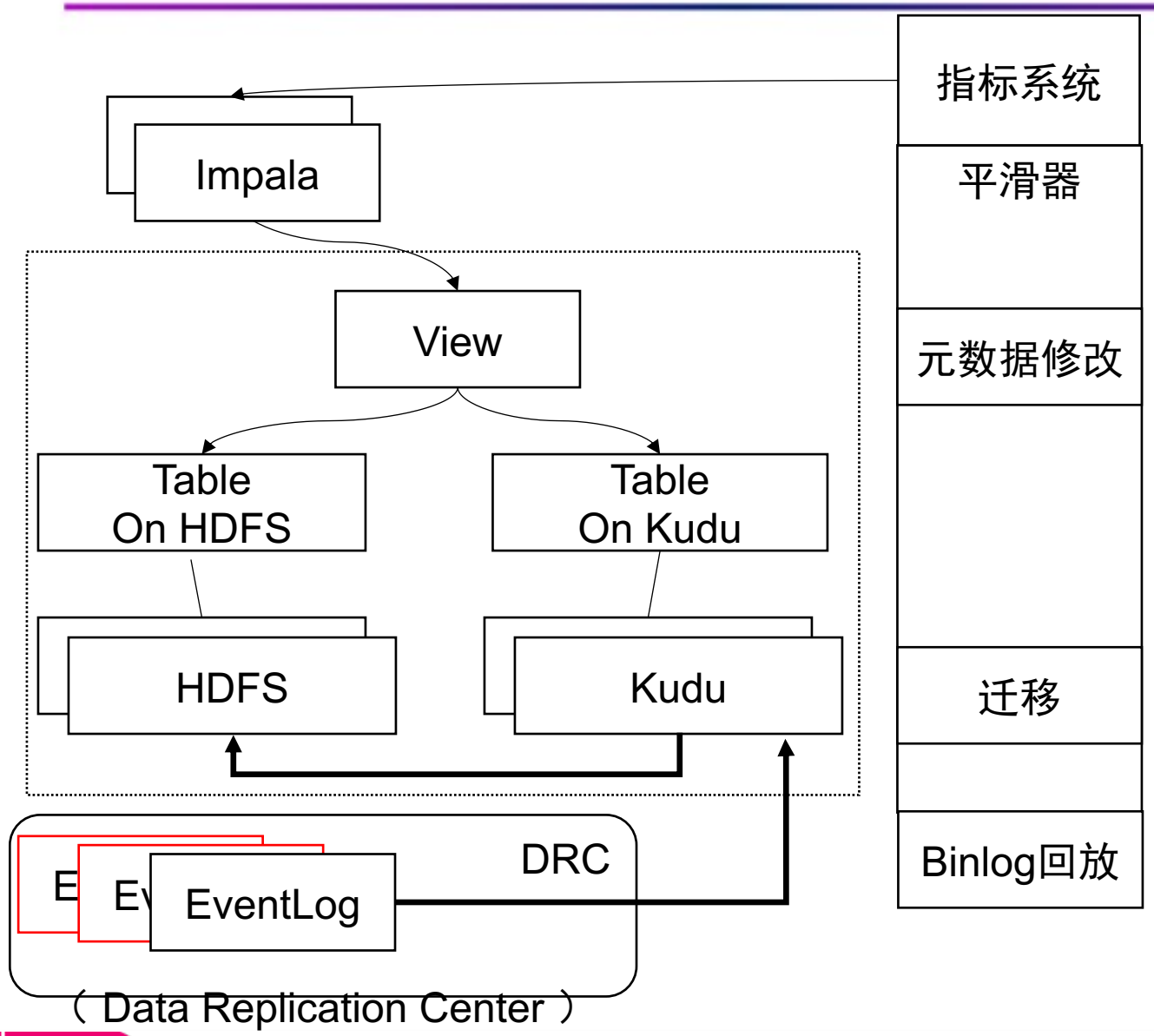
评测项目	明细表	汇总表
数据变更	灵活，通用，拥抱变化	增加指标不影响 增加维度数据洗牌
时间周期	采集即所见 起步最爱	依赖前置ETL流程 提早布局，覆水难收
回溯级联更新	效率高 建模层同步清除维度缓存	按照血缘关系定向更新 技术门槛
查询效率	较慢，参照明细规模	较快，适合汇聚后的二次度量
权限设定	较为繁琐	汇聚后业务层次脱敏
Count(Distinct)	支持	不支持，曲线救国，参照Kylin方案

规范

- 汇总表基于主题建设
- 提前布局汇总表对业务释放
- 明细表授权对业务方慎重
- 扩展汇总维度工具化、流程化

常见问题2：增加维度后数据回溯

常见问题3：实时OLAP指标和离线指标融合



融合

- 基于视图对外交付
- 冷热交替数据通过调度定期变更
- 数据交替期间多存储
- 元数据变更与广播
- 迁移完Kudu清理

常见问题4：指标、数据的加减法

数据
- 表、字段访问热度
- 数据上下游任务编排

指标与维度
- OLAP报表访问热度
- 字典访问统计

日报任务
- 签到任务倒计时
- 业务场景阶段性

瘦身

- 数据时间收益陨落
- 大数据量迁移高成本
- 任务繁琐注入

- 做减法同样重要
- 数据流转周期均干预
- 建立数据存放周期等级
- 技术+人工并进

THANKS