网易数帆
指标中台核心技术解析

祝联新 网易数帆 指标中台技术负责人

DataFunSummit # 2023

网易数帆 DataFun.

网易数帆 DataFun.

目录 CONTENT

01

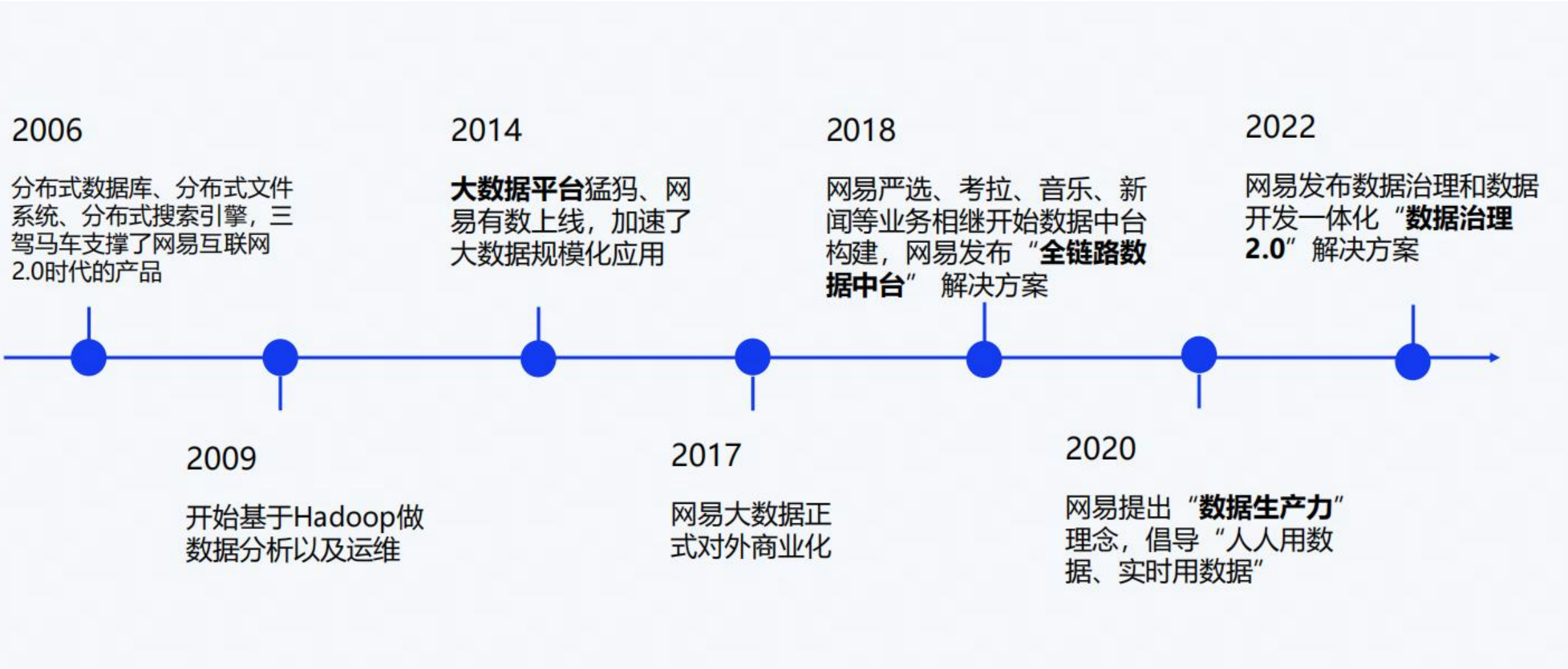
网易数帆大数据产品介绍

DataFunSummit # 2023

网易数帆 DataFun.

网易数据分析的发展历史

网易数帆 DataFun.



网易EasyData产品矩阵

网易数帆 DataFun.

02

网易数帆指标中台

DataFunSummit # 2023

网易数帆 DataFun.

为什么要构建指标中台?

网易数帆 DataFun.

指标口径不一致

- 在数据质量问题反馈中，约有31%的问题涉及指标口径

指标入口不统一

- 缺少整个企业级统一指标消费的入口，指标分布在不同的部门和分析师手上，业务人员不知道去哪里找想要的指标，缺少权威的解释

指标价值难以量化

- 难以跟踪指标的使用和量化，一张报表每个月花费3W块钱，30天内都无人查看

指标开发效率低

- 传统指标开发需要涉及跨部门，多角色协作，效率低，平均指标开发需要一周的时间，业务人员无法自助完成指标开发
- IT资源瓶颈：数据分析团队只有10个人，每周的数据分析需求有500个

重复计算大量消耗资源

- 从BI报表出发，计算使用很多中间表和计算资源，中间表和数据利用率低

指标质量差

- 指标问题溯源困难，上层依赖17层模型，指标问题的排查效率低下
- 任务开发导致的数据质量问题占比60%，90%的问题都是业务先发现

商业客户对指标中台的诉求

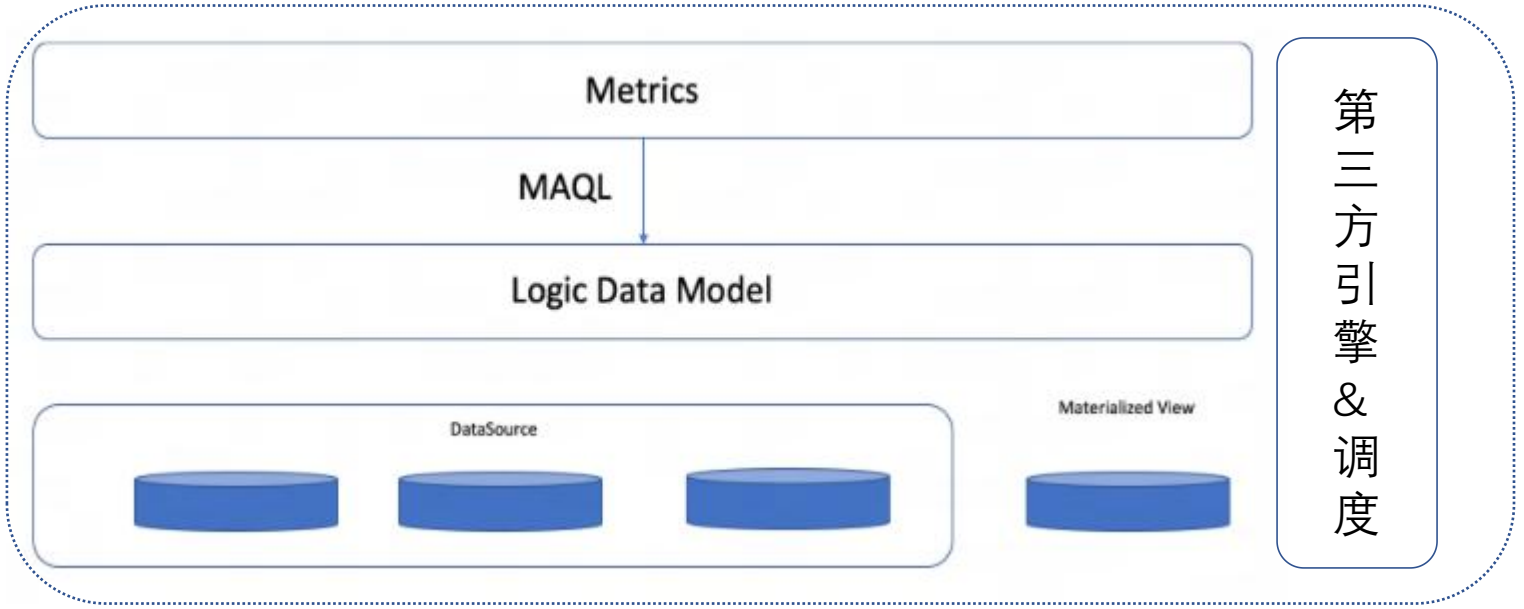
网易数帆 DataFun.

物化到业务库	指标目标管理	统一调度引擎
• hive的原始数据直接通过指标中台计算后直接物化到业务库供下游业务系统使用	• 通过指标中台可以完成业务指标目标的管理，方便查看业务目标和关联指标的进展情况	• 采用统一的调度系统，方便指标上下游任务统一管理，将指标原始数据的产出纳入依赖管理

网易数帆 指标中台解决方案

网易数帆 指标中台（Metrics Stores）介于数据中台和下游的数据应用（主要是BI）之间，提供了指标的标准化能力。通过“一次定义，多次复用”（Build once，use many），为业务决策和管理提供单一、可信的数据来源，解决长期以来指标口径不一致的问题。指标中台，构建在跨数据源（Catalog）的统一逻辑语义模型层（Logic Data Model）之上，内置了一套指标定义语言MAQL（Metrics Analysis Query Language），提供了独立于第三方计算引擎和调度系统的指标自动计算的能力，通过配置化的方式可以自动完成指标的开发，大幅度提高了指标开发的效率。

数据中台



BI系统
业务系统

03

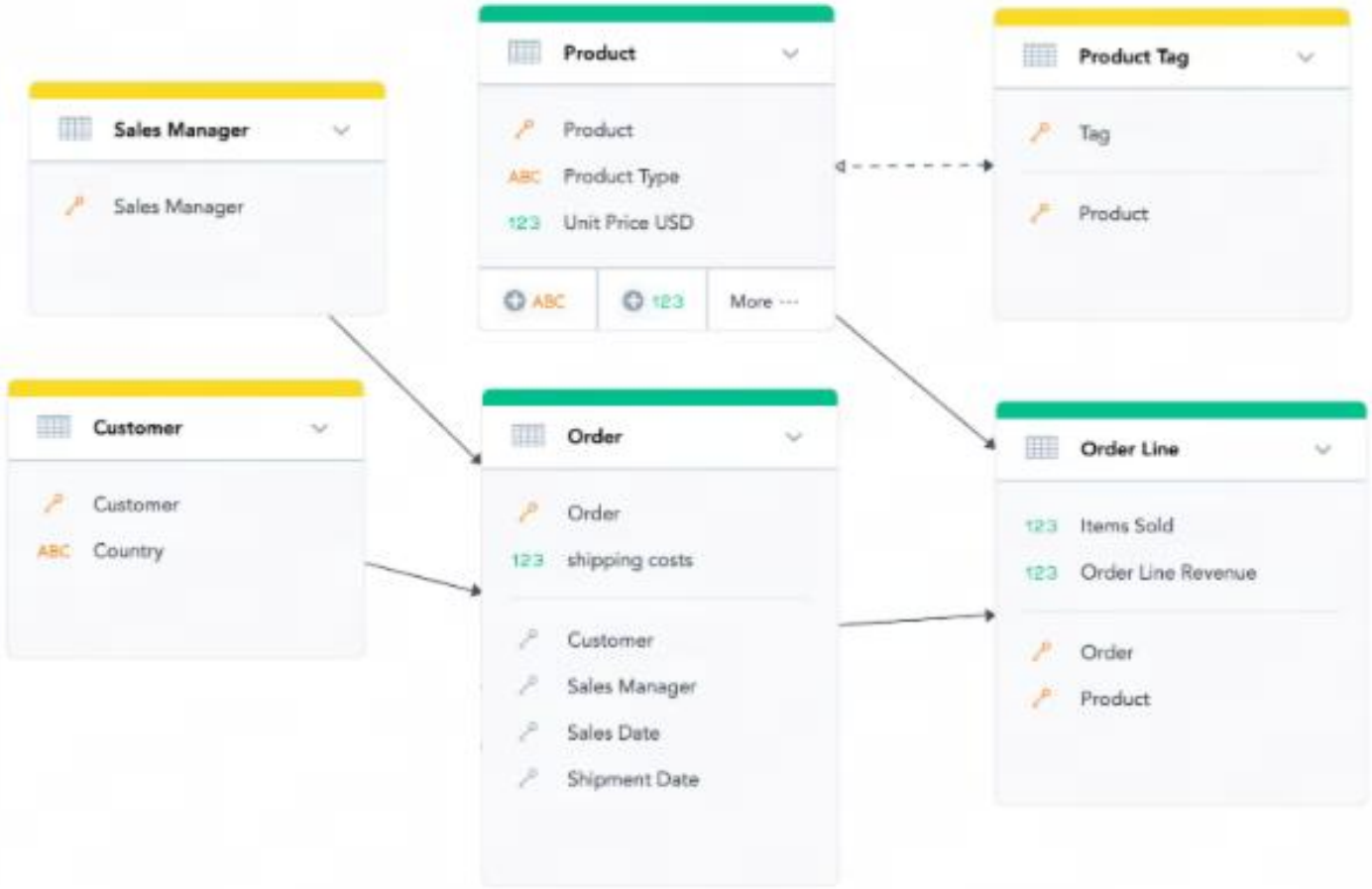
指标中台核心技术解析

DataFunSummit # 2023

网易数帆 DataFun.

构建跨数据源的统一逻辑语义层模型

LDM（Logic Data Model），也就是统一逻辑语义层模型，是构建在数据仓库和下游数据应用之间的独立层(stand-alone layer)，它可以屏蔽不同的数据来源差异，来统一应对下游的数据应用。

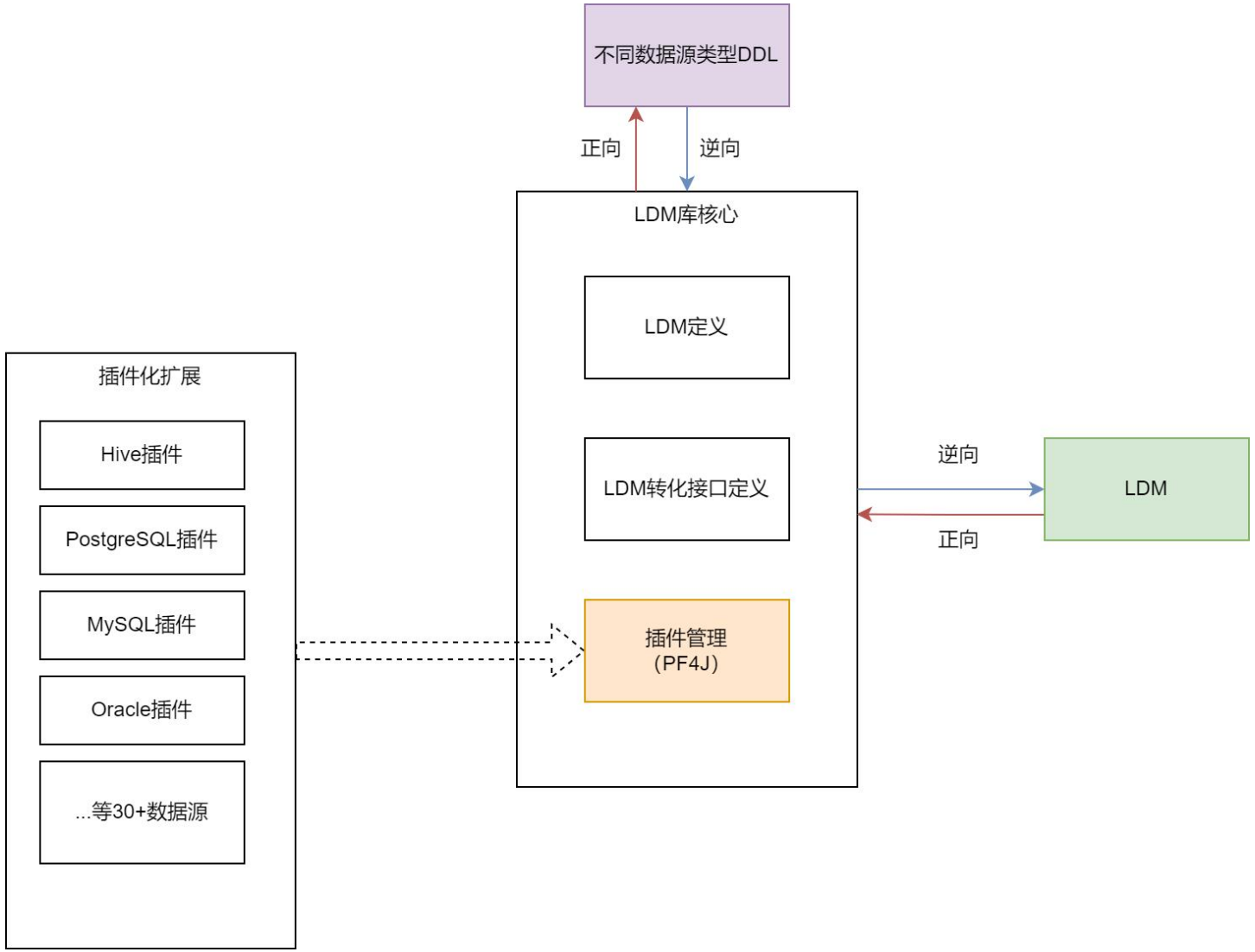


LDM 架构设计

01 屏蔽数据源差异

02 逆向建模

03 物化DDL



构建简洁高效的指标分析查询语言

传统 SQL 不像通用编程语言一样支持组合继承等能力，所以采用SQL写任务往往有大量的冗余代码，而这些冗余代码也常常会因为逻辑不一致，导致指标的计算口径不一致。我们引入MAQL（Metrics Analysis Query Language）来实现简化指标定义、实现指标的组合和复用、简化时间口径定义等目的。

01

简化指标定义

02

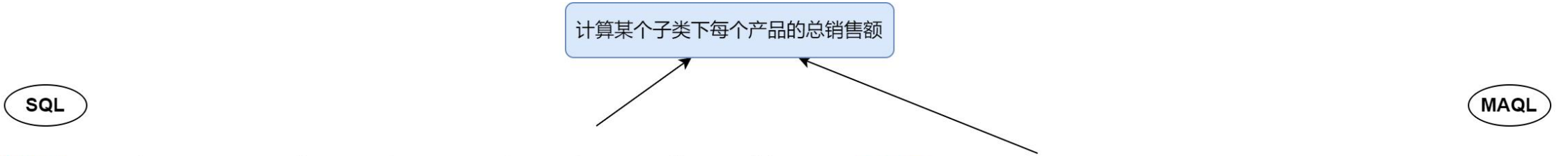
实现指标的组合和复用

03

简化时间口径定义

MAQL：简化指标定义

构建在语义模型的基础上，不需要再定义From，Join，天然支持多维分析




```sql
SELECT p.product_name, SUM(s.quantity * s.price * (1 - pm.discount))
AS total_sales
FROM xh_sales s
JOIN xh_product p ON s.product_id = p.product_id
JOIN xh_promotion pm ON s.promotion_id = pm.promotion_id
JOIN xh_category c ON p.category = c.category_name
JOIN xh_subcategory sc ON c.category_id = sc.category_id
WHERE sc.subcategory_name = "Subcategory B"
GROUP BY p.product_name;
```

```
SELECT SUM({fact/quantity}*{fact/price}*(1 - {fact/discount}))

WHERE {dim/subcategory_name} = "Subcategory B"
BY {dim/product_name}
```

MAQL：实现指标组合和复用

创建和存储指标来指定数据的“基本事实分析”(即单一事实来源)。存储的指标可以在其他指标中重复使用

网龄千元橙分期目标完成进度

```
SELECT SUM({metrics/cnt_cfq_htd}/{metrics/may_qjpd_prom_pg})
by {dim/mt_xy_week_report_fct.bureau_name}
```

组合

组合

今年网龄千元橙分期完成数

```
SELECT SUM({metrics/cnt_cfq})
WHERE {date/mt_xy_cfq_fct.dat_date} = {period/cy} BY
{dim/mt_xy_week_report_fct.bureau_name},
{date/mt_xy_cfq_fct.dat_date}
```

今年网龄千元橙分期目标完成数量

```
SELECT SUM({fact/may_qjpd_prom_pg})
BY {dim/mt_xy_week_report_fct.bureau_name}
```

派生

网龄千元橙分期完成数

```
SELECT SUM({fact/cnt_cfq})
BY {dim/mt_xy_week_report_fct.bureau_name},
{date/mt_xy_cfq_fct.dat_date}
```

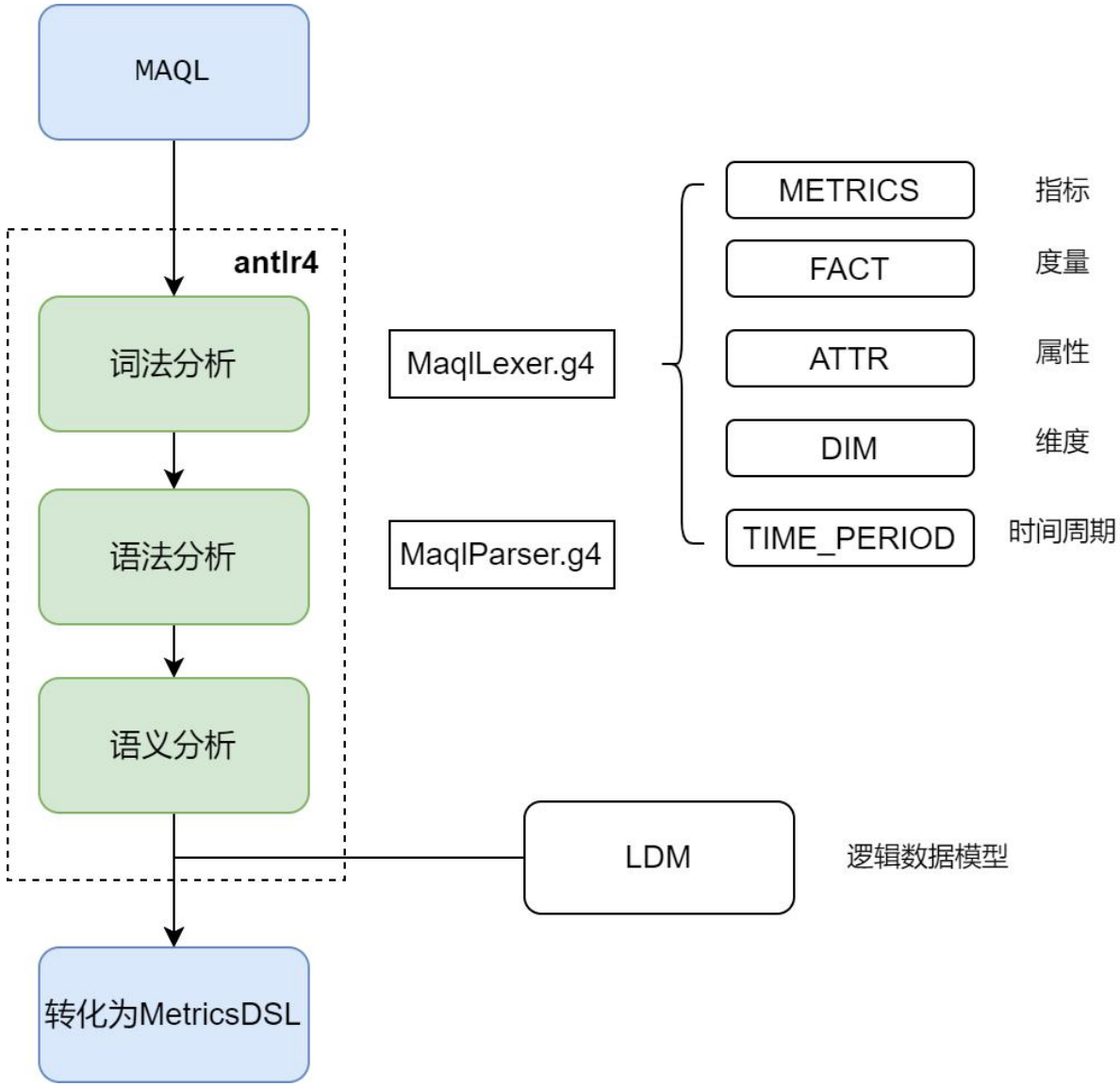
MAQL：简化时间口径定义

定义指标时经常需要包含时间口径，我们可以通过简单的语法直接使用时间周期，从而实现时间口径的统一，并且支持上一个交易日等证券行业特色的时间周期，来满足客户的特定时间口径需求

```
# 在指标中台中创建了上个月的时间周期 "lastMonth"
# 查询上个月销售额
# 使用时间周期时对应的字段类型必须为时间维度，操作符必须为"="
select sum{fact/price} where {date/create_time} = {period/lastMonth}
```

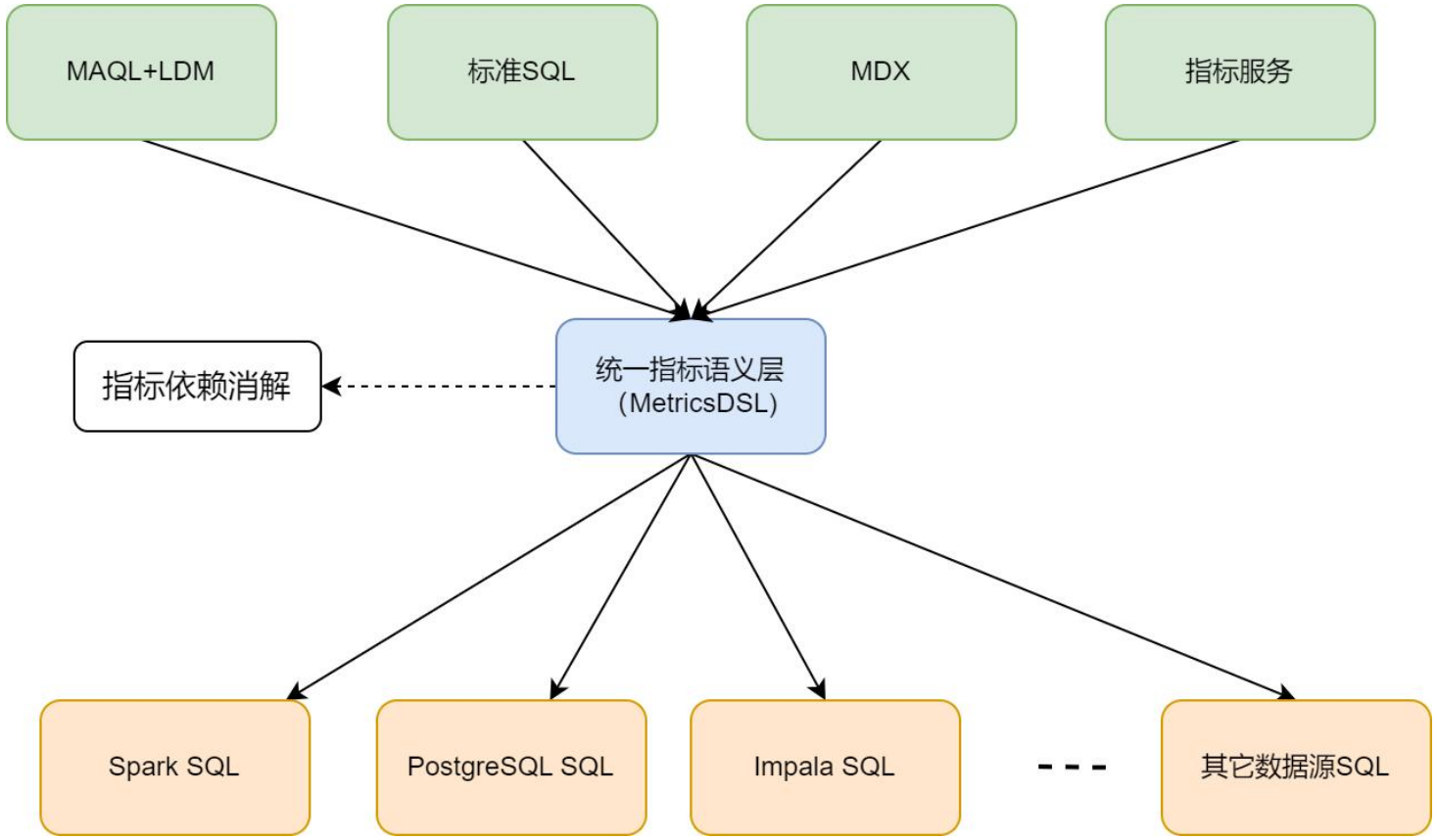
MAQL 实现

- 支持包括AVG、COUNT、SUM等7种聚合函数
- 支持包括AND、OR等10种逻辑操作函数
- 支持包括+、-、*、%、ABS等11种常用数据处理函数
- 支持直接使用时间周期

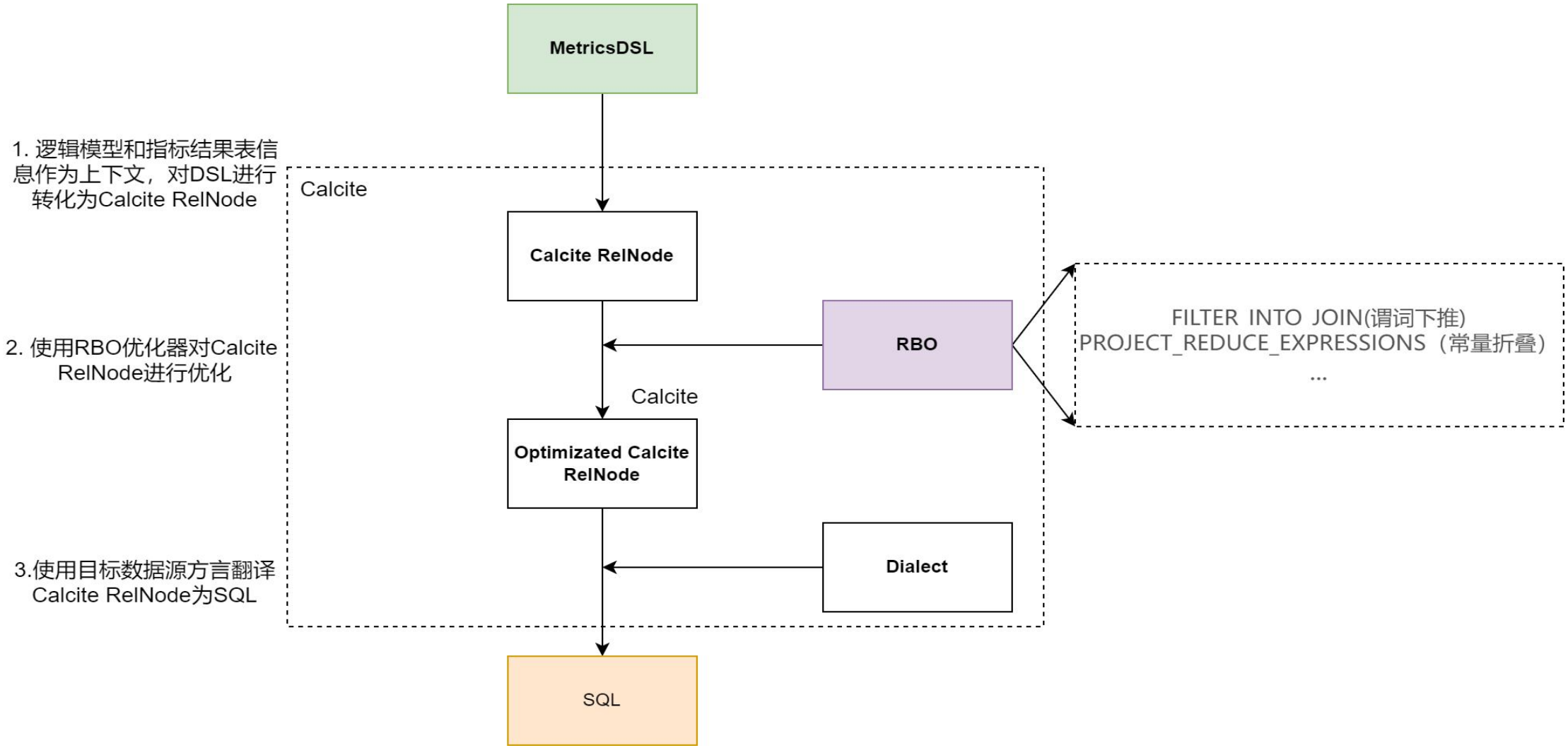


构建统一指标查询语义层

通过构建统一的指标查询语义层 MetricsDSL，将指标的查询需求做进一步抽象，并且在语义层实现指标依赖消解等工作，屏蔽底层不同数据源的SQL语法差异，并且隔离不同的指标查询需求（不同语法）对于底层数据源的入侵和影响，从而方便分别扩展不同的指标查询语法和对接不同的数据源目标。

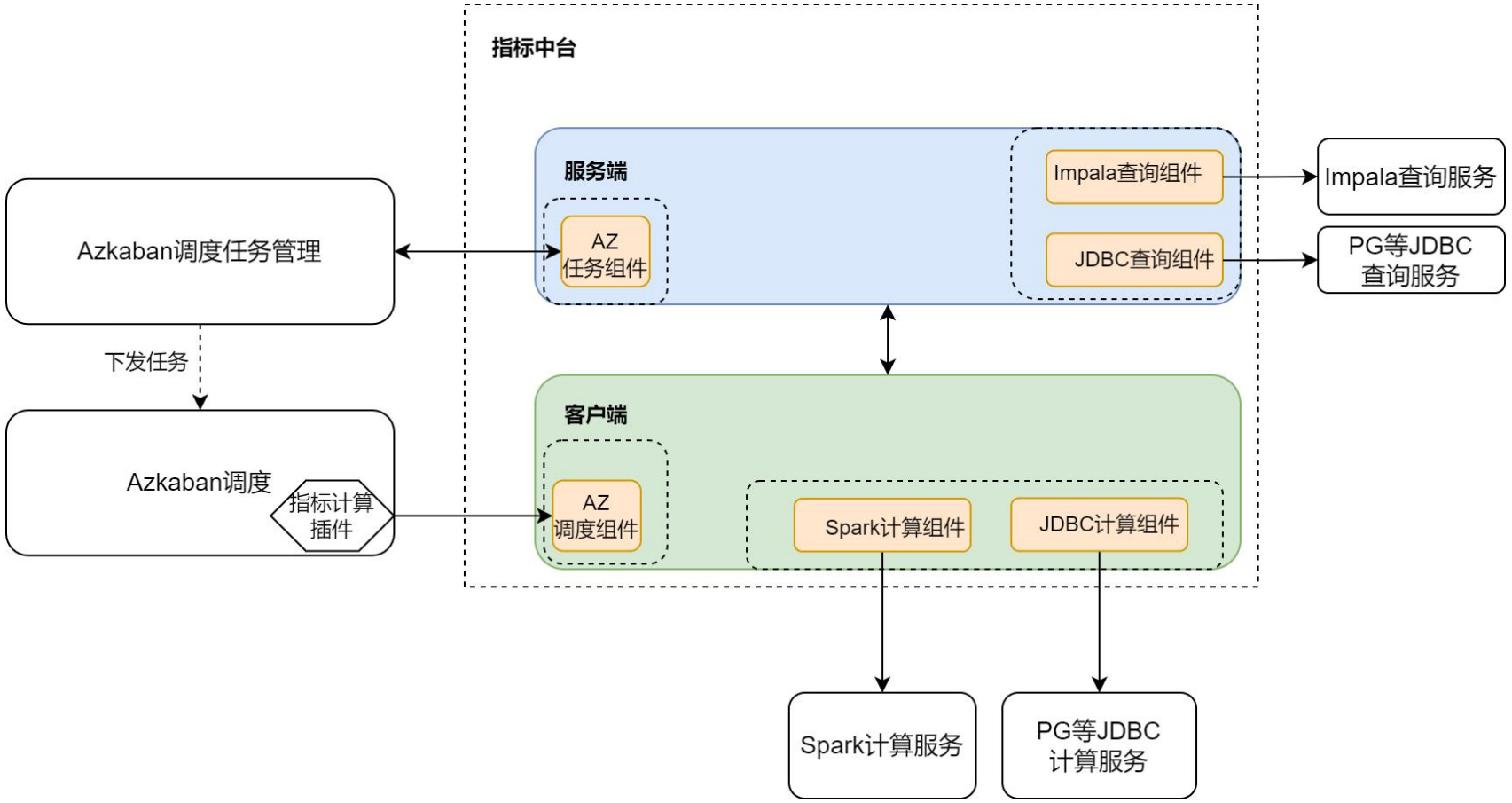


基于Calcite实现引擎SQL翻译



引擎解耦，灵活对接第三方引擎

为了满足能够灵活对接第三方引擎，包括调度引擎、查询引擎和计算引擎，我们通过适配组件来完成引擎和调度的对接，并基于抽象引擎的能力来完成指标中台核心能力的建设



04

未来规划和展望

DataFunSummit # 2023

网易数帆 DataFun.

未来规划和展望

网易数帆 DataFun.

1. 深入指标应用场景：数据洞察、仪表盘、KPI管理、指标地图。
2. 对接更多BI系统，打通生产到消费的完整链路
3. 支持更多数据源，例如doris等MPP数据源
4. 接入AIGC，实现基于自然语言的指标查询


感谢观看
网易数帆
DataFun.