数字转型 架构演进

SACC

2019 中国系统架构师大会

SYSTEM ARCHITECT CONFERENCE CHINA 2019

IT168.com ChinaUnix.net ITPUB

2019年10月31-11月2日 | 北京海淀永泰福朋喜来登酒店

搜狗大数据中台建设实践

sogou　申贤强

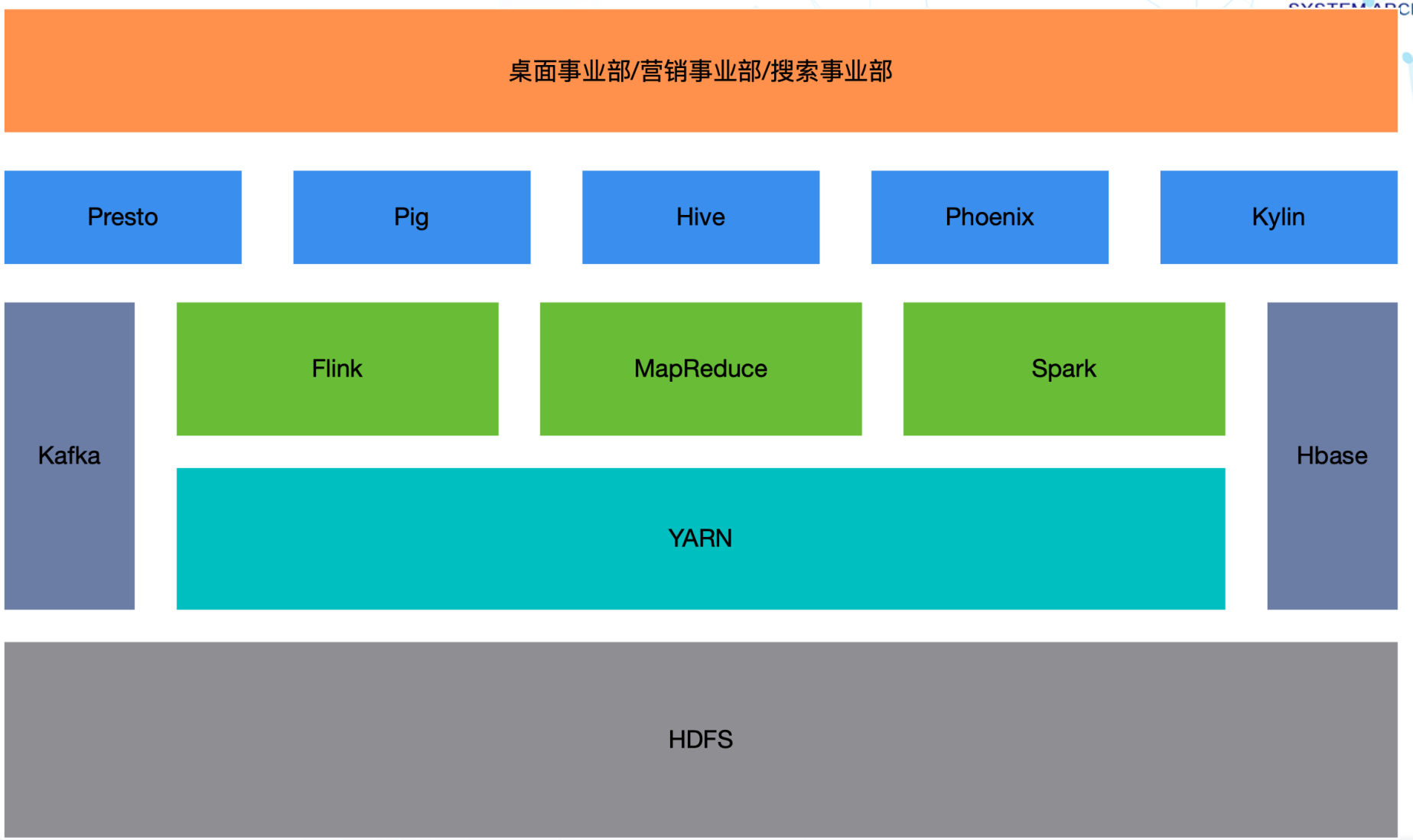
关于我们

- 来自搜狗大数据平台部
- 基于Apache Hadoop生态，建设搜狗海量数据存储和计算平台
- 提供稳定高效的数据分析系统，为搜狗各类型大数据应用，提供一站式数据处理服务
- 每天数十亿的数据增量，数以百万计的数据计算流程，使数据的价值得到充分利用
- 最前沿技术落地及推进开源技术的发展

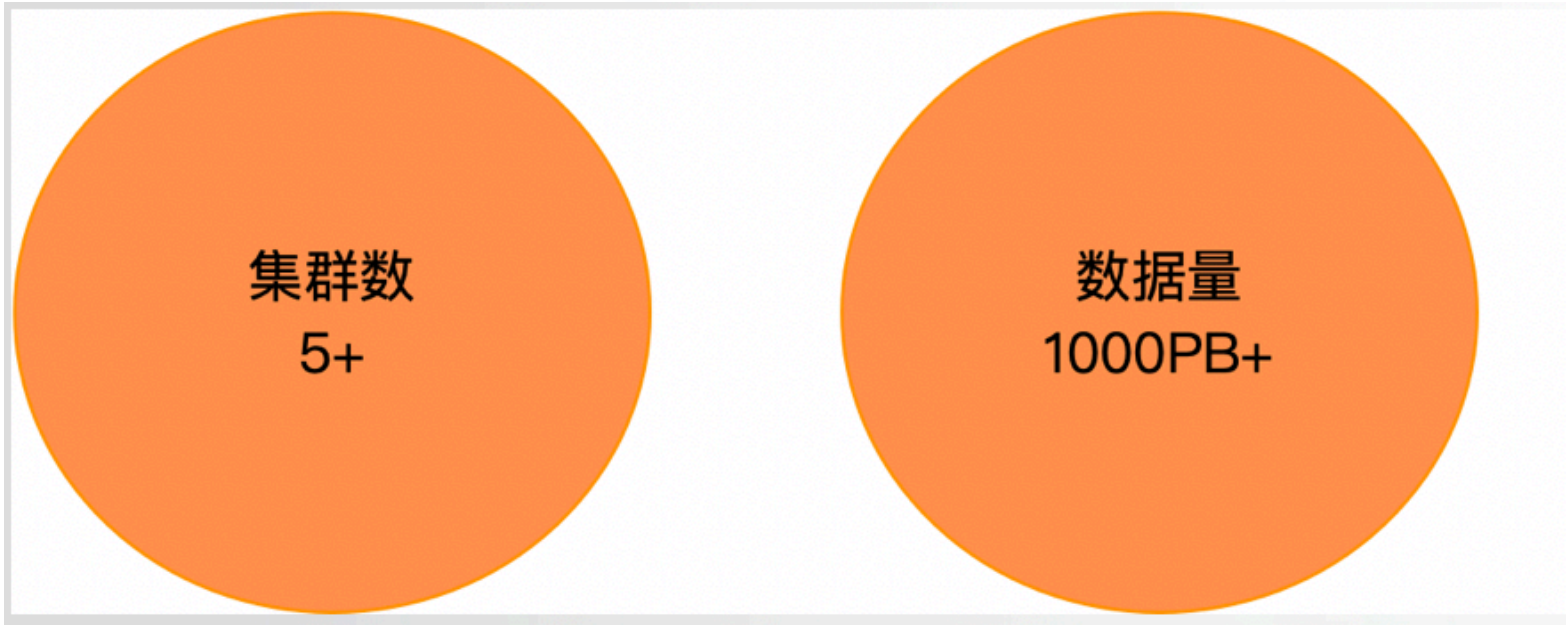
目录

1 Sogou 数据中台背景

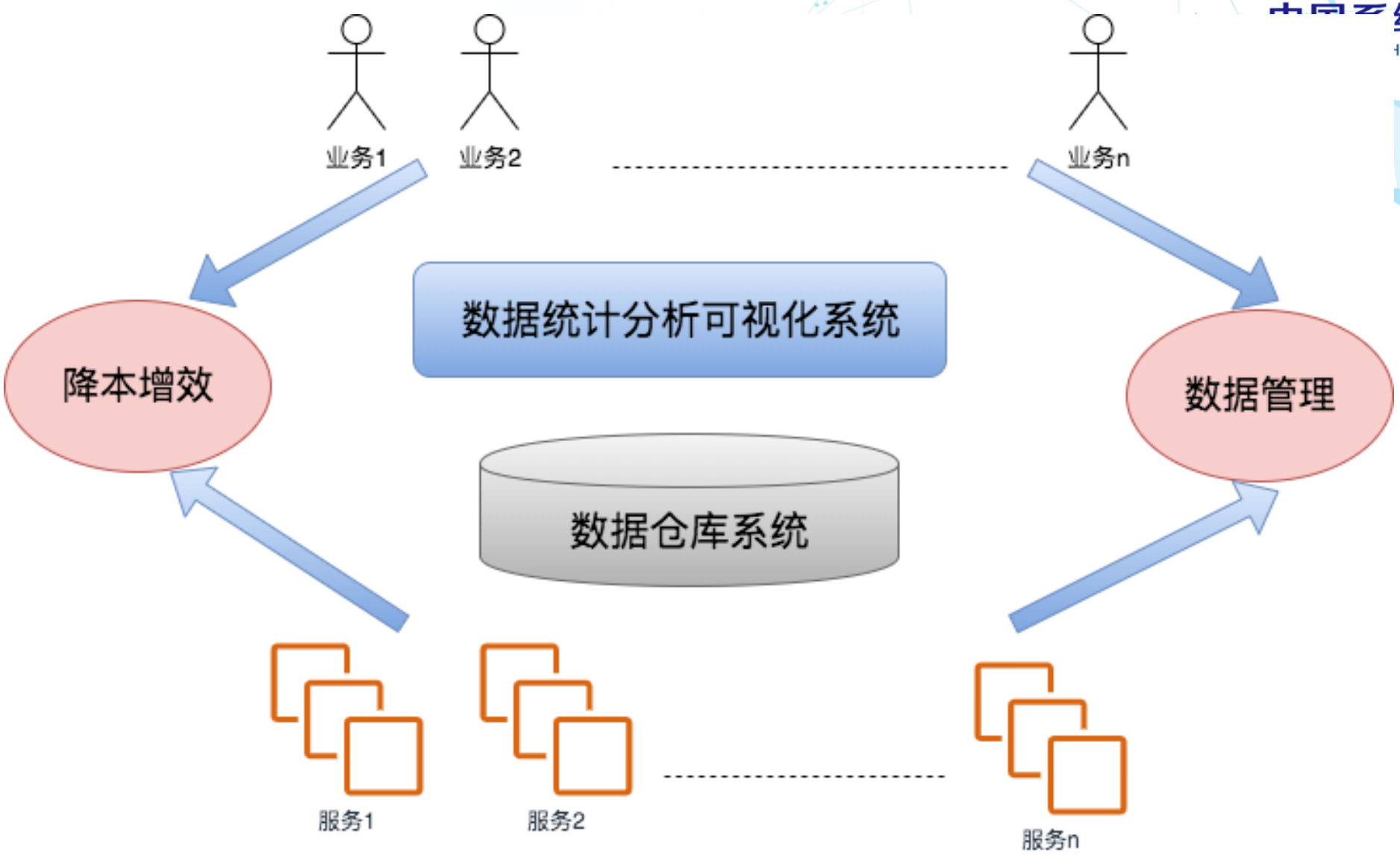
背景



背景



背景



背景

传统的数据仓库不能满足数据分析需求

- 从统计分析向预测分析转变
- 从单领域向跨领域转变
- 从被动分析向主动分析转变
- 从非实时分析向实时分析转变
- 从结构化数据向多元化转变

数据的处理架构发生了变化

以Hadoop、Spark等分布式技术和组件为核心的“计算&存储混搭”的数据处理架构，能够支持批量和实时的数据加载以及灵活的业务需求

数据的预处理流程正在从传统的ETL结构向ELT转变

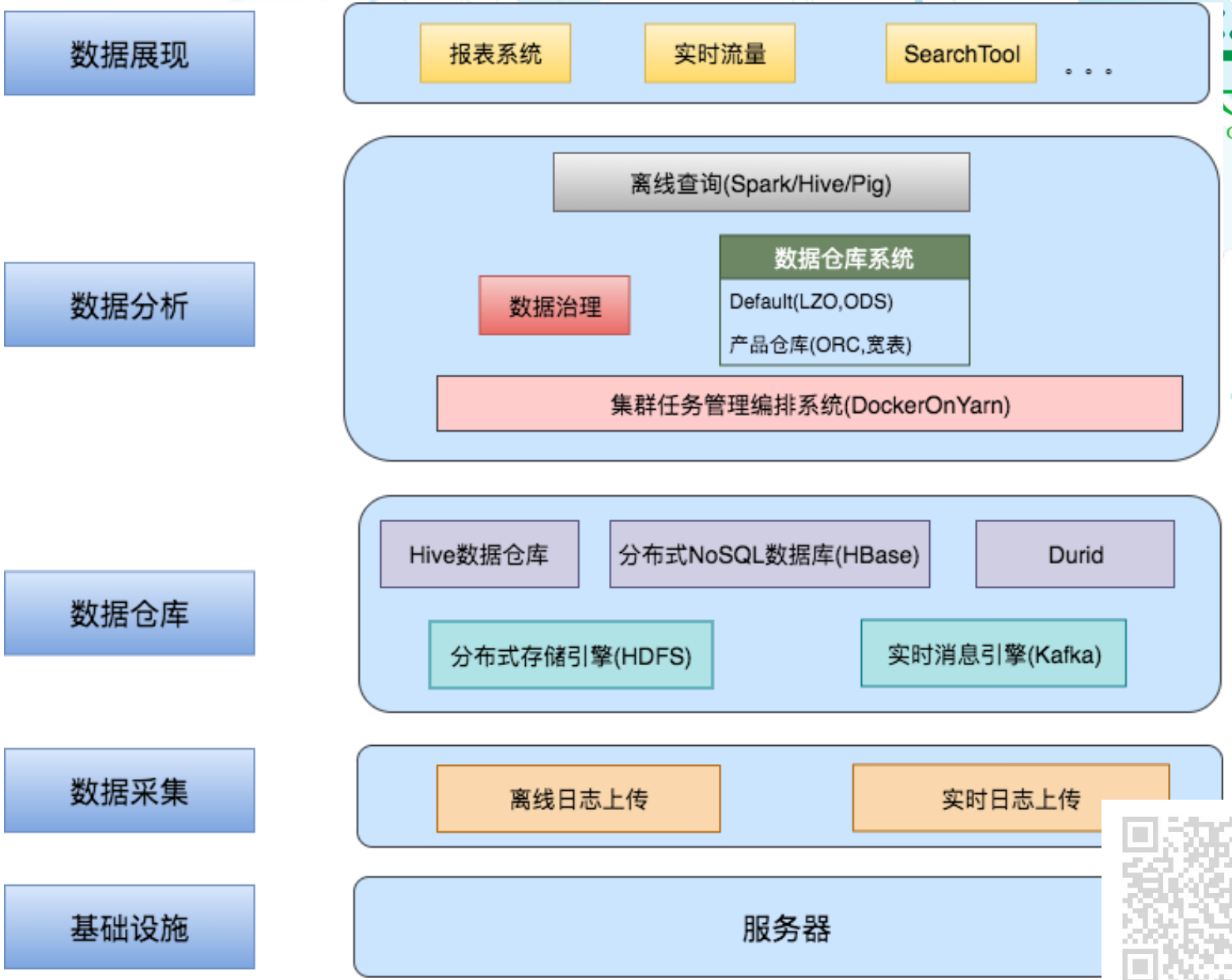
背景

之前架构

协作

效率

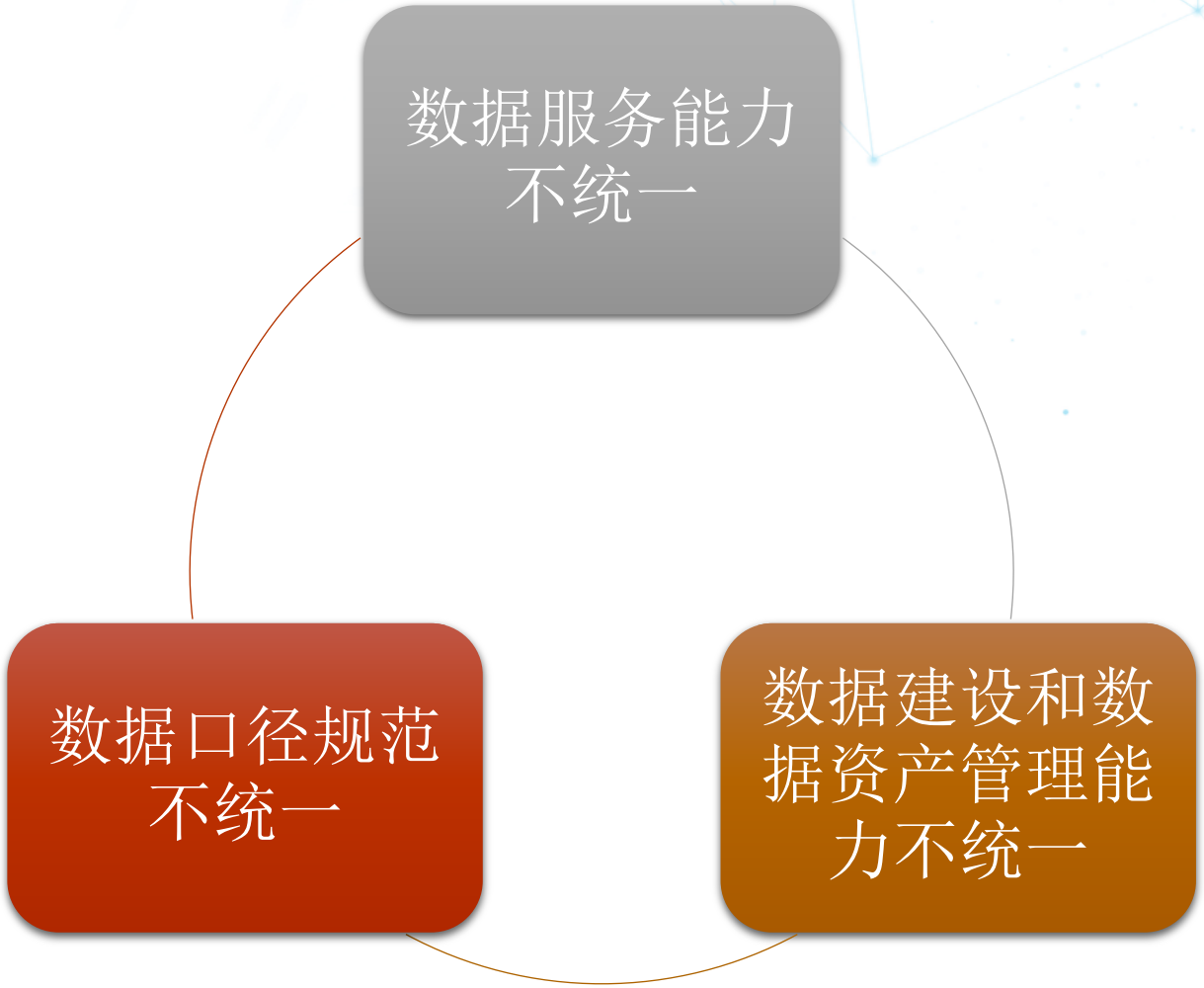
基础



背景

数据效率	数据协作	数据能力
数据仓库层次建设	权限管理弱，安全性不佳	BI分析能力
元数据管理混乱	仓库集市，数据互通难度大	业务数据转化价值能力弱
Hive/pig的计算效率	重复报表多，统计口径不一致	无数据接口

背景



目标

数据中台

前台业务

敏捷业务开发，提供战略指导
提供一站式的数据分析展现服务

数据服务

数据获取便捷，数据协作方便
降低重复建设，提升计算效率

基础平台

海量存储和计算平台，任务管理功能
提供集群成本账单，提升服务质量

定位

节约成本

1.人力成本，降低沟通成本
2.集群成本，降低重复建设

数据中台

提升效率

1.计算效率，由天->小时->秒
2.自助流程，降低需求沟通时间
3.实时性，提升数据转化价值能力

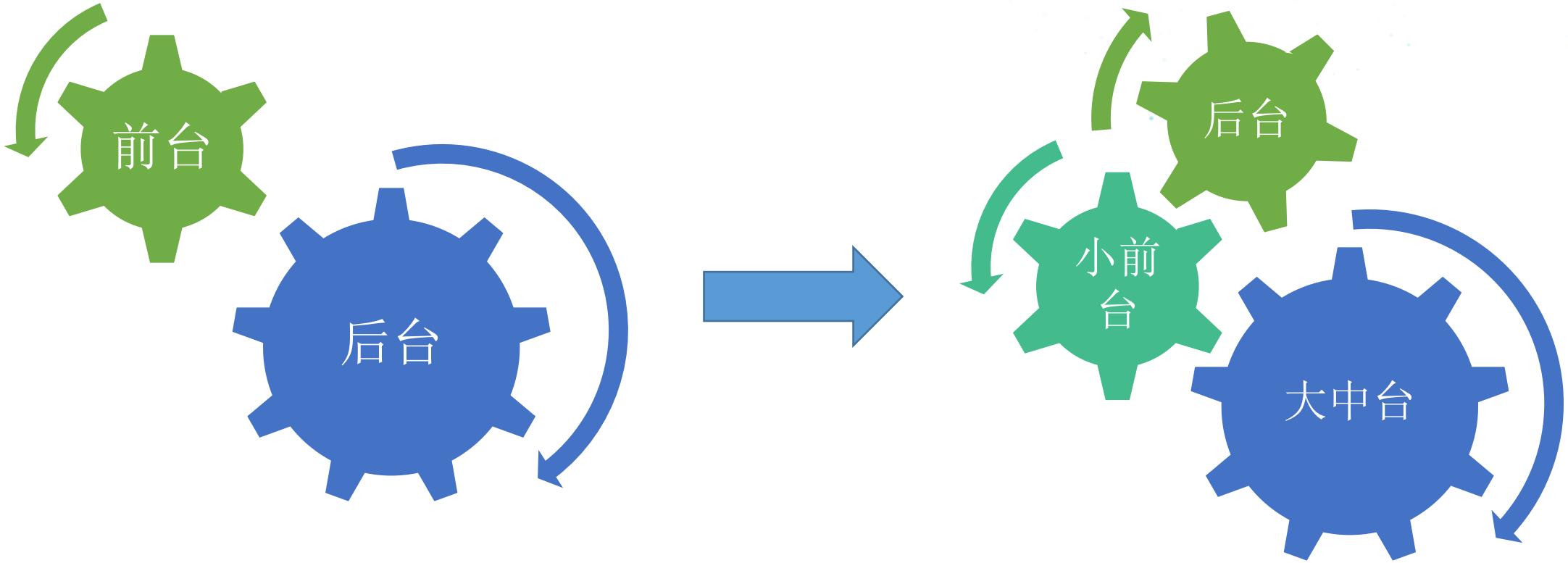
沉淀技术

1.统一化，统一数据湖入口与出口，统一技术口径
2.管理化，管理数据链路，管理元数据，数据血缘分析
3.资产化，产出数据模型，产品仓库

2 Sogou 数据中台技术历程

概述

中台首先是一种战略选择、一种组织形式，其次才是一些有形的产品支撑和实施的方法论。



企业级的能力复用

概述

数据中台

数据共享
服务共享
BI平台

数据平台

仓库管理平台
报表平台

数据仓库

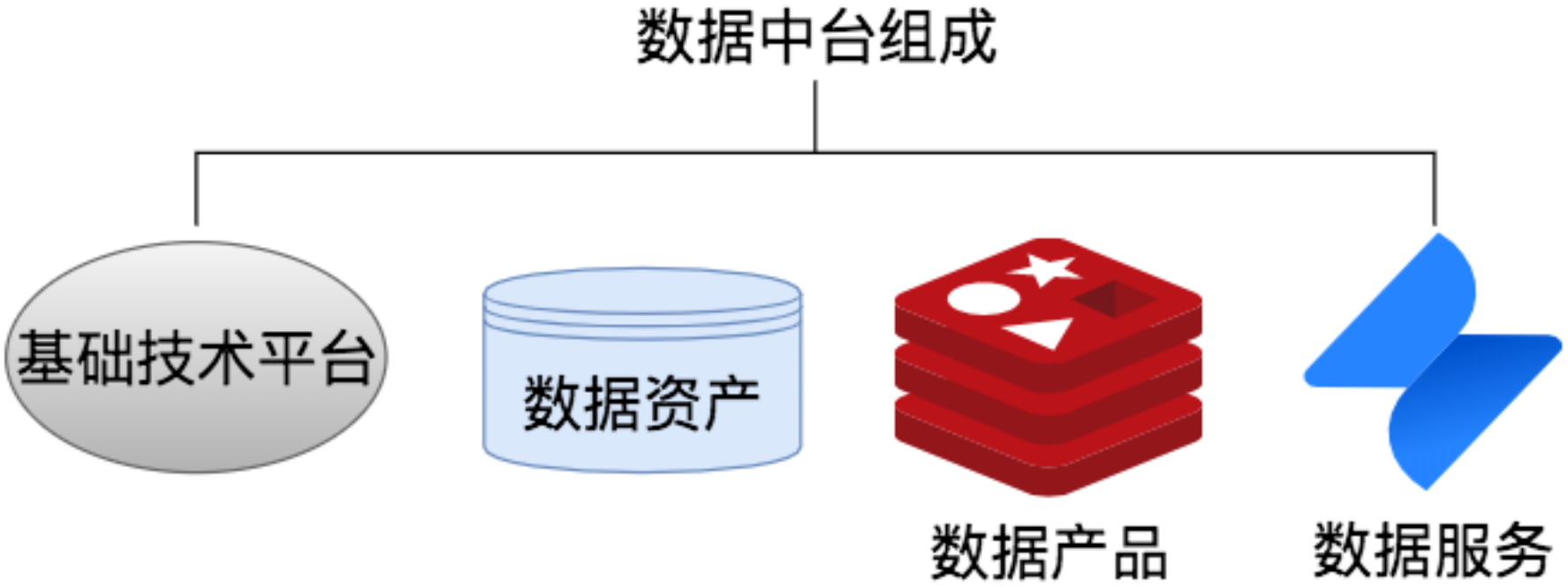
业务众多
量级偏大

数据库

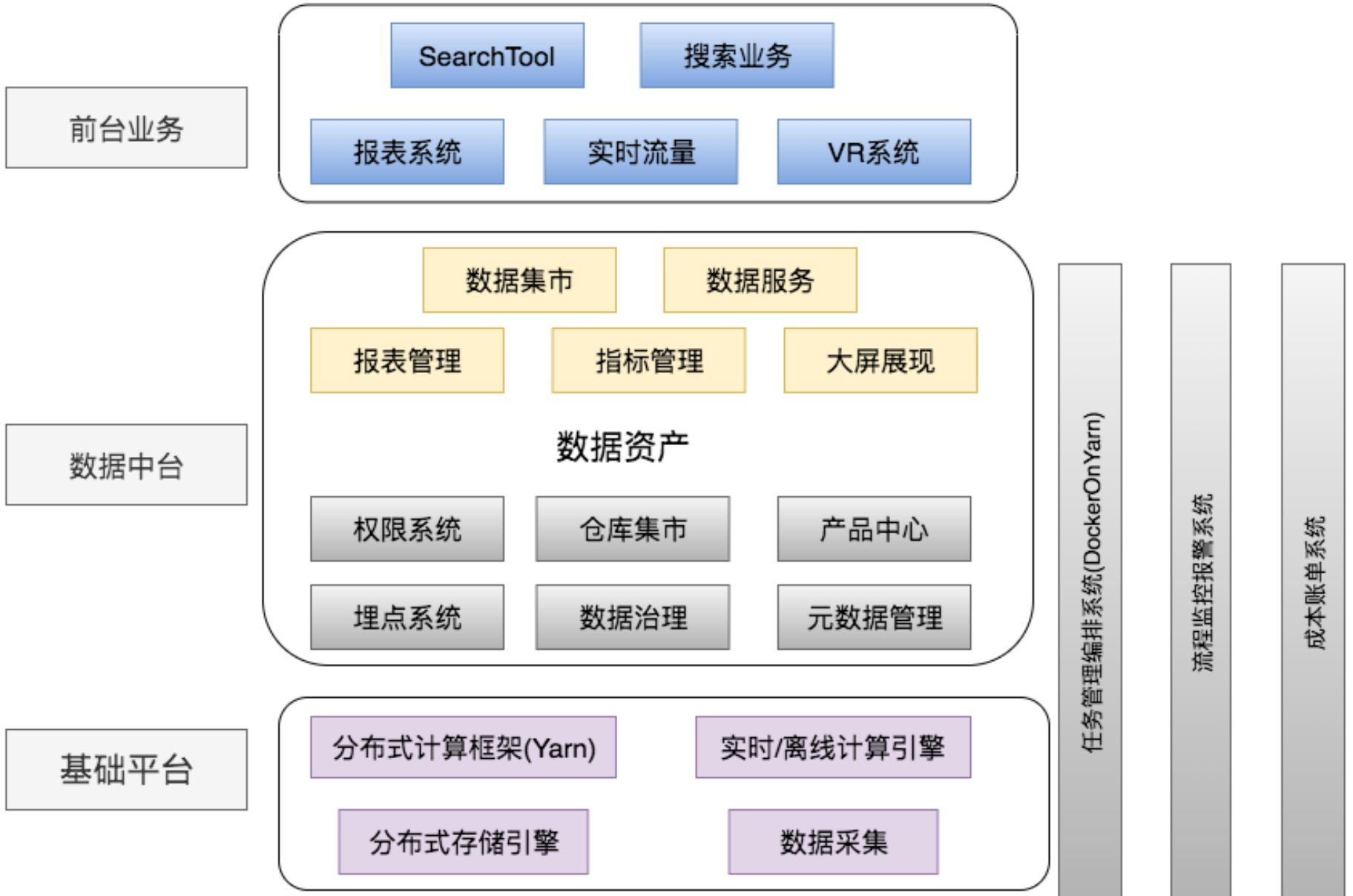
业务单一
数据量小

概述

数据中台的基本思想就是不重复造轮子，把复用共享的东西提炼出来，变成一个可以被其它业务单元引用的基本能力，为前端的业务赋能



功能设计


前台业务
SearchTool
搜索业务
报表系统
实时流量
VR系统
数据中台
数据集市
数据服务
报表管理
指标管理
大屏展现
数据资产
权限系统
仓库集市
产品中心
埋点系统
数据治理
元数据管理
基础平台
分布式计算框架(Yarn)
实时/离线计算引擎
分布式存储引擎
数据采集
任务管理编排系统(DockerOnYarn)
流程监控报警系统
成本账单系统

架构-优化后

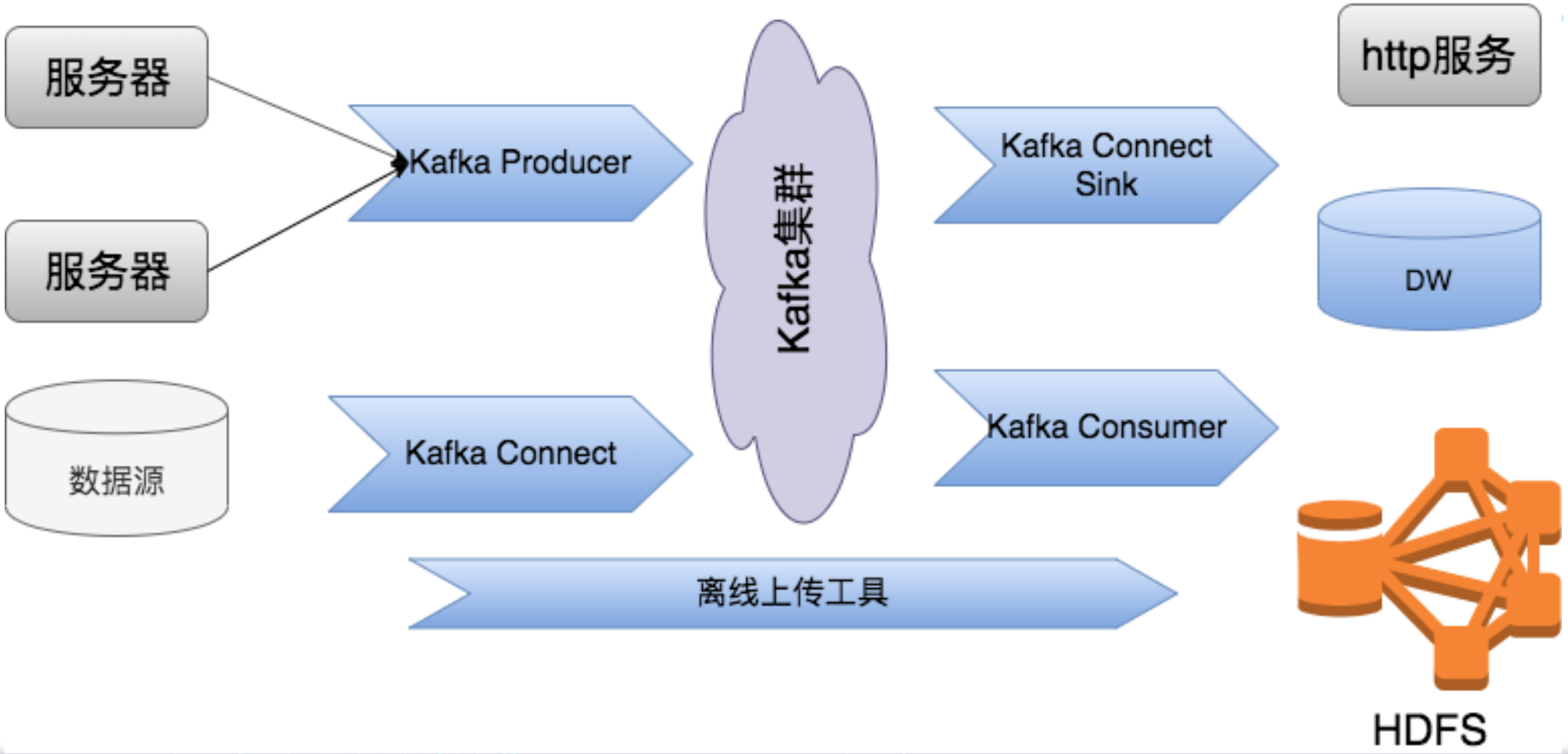
协作

效率

基础

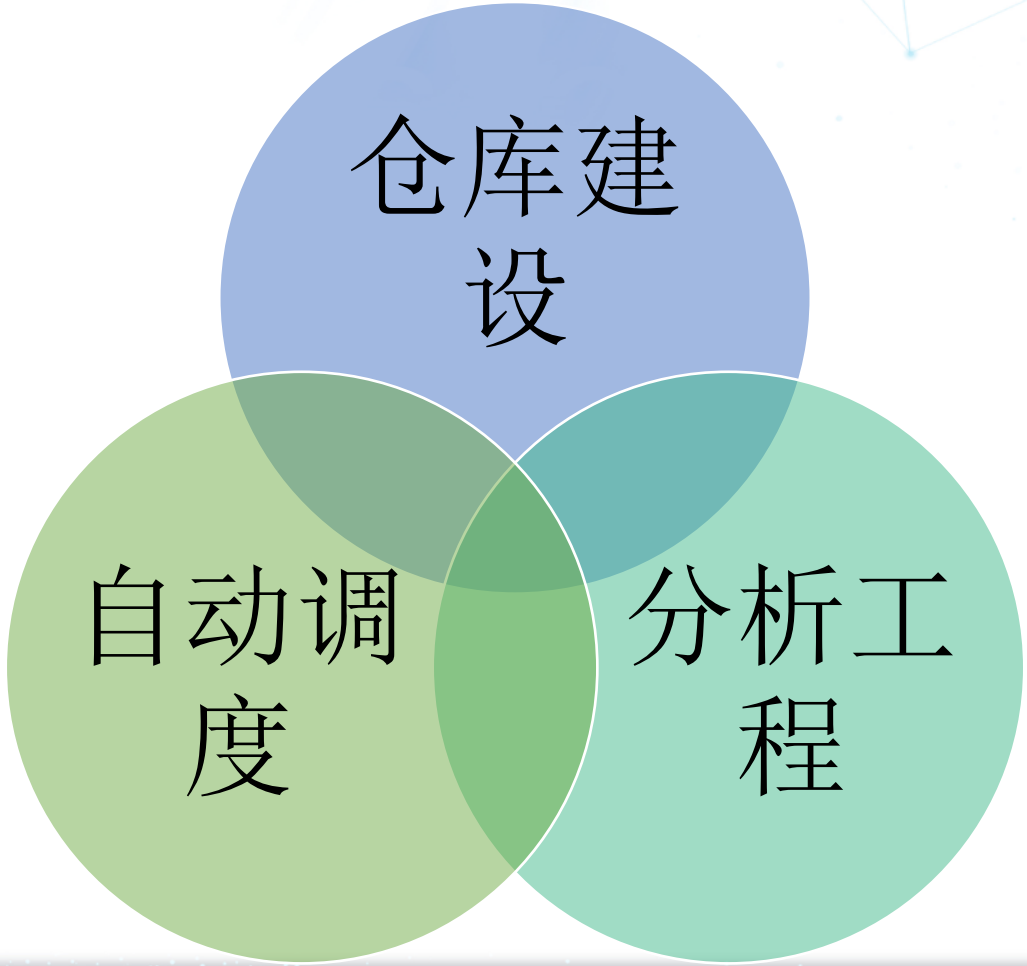
架构-数据采集

https://github.com/sogou/tail2kafka
https://github.com/sogou/hdfs

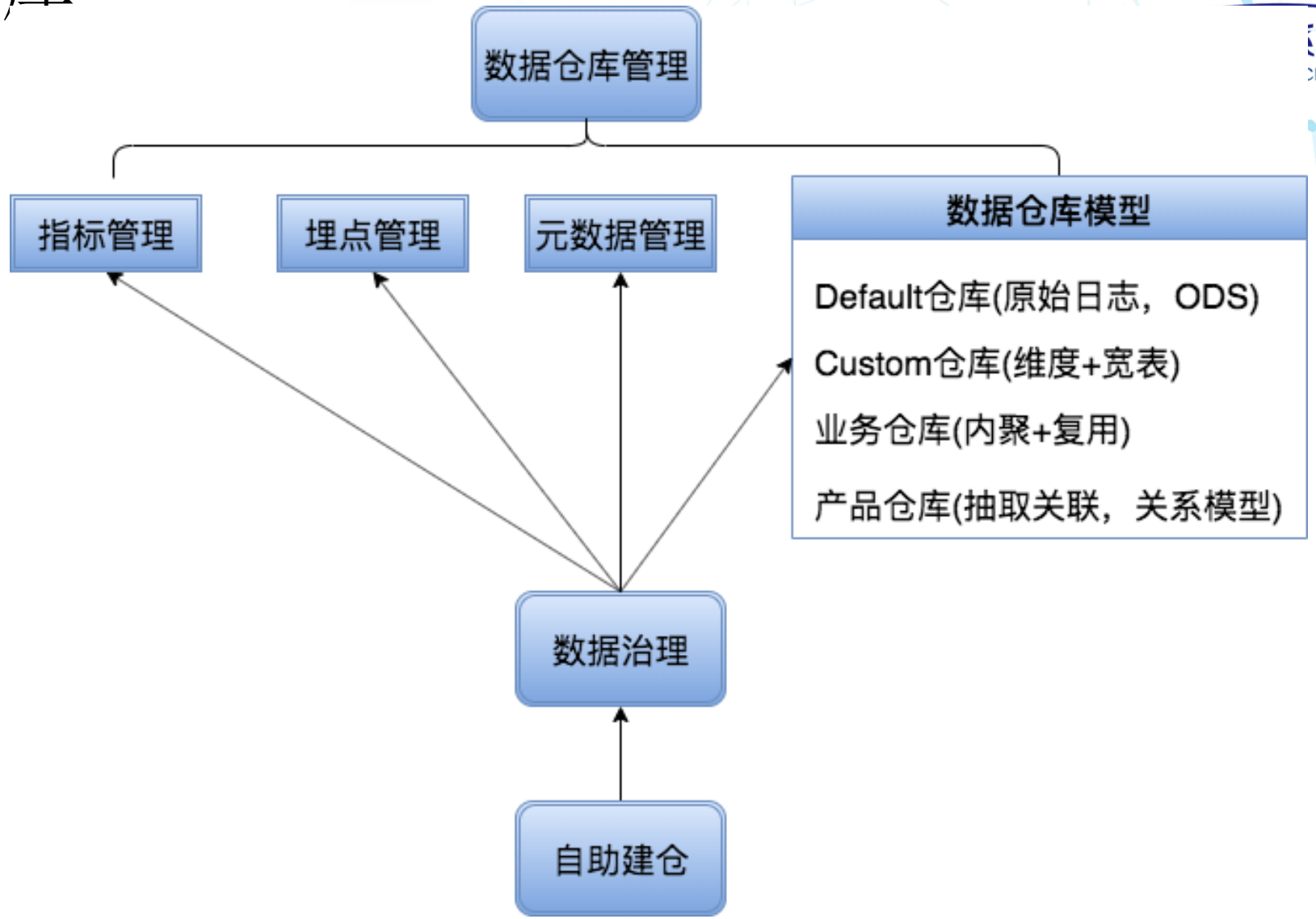


架构-数据仓库

数据治理（Data Governance）是指对数据湖中的数据进行存取、处理、分析及传输。



架构-数据仓库

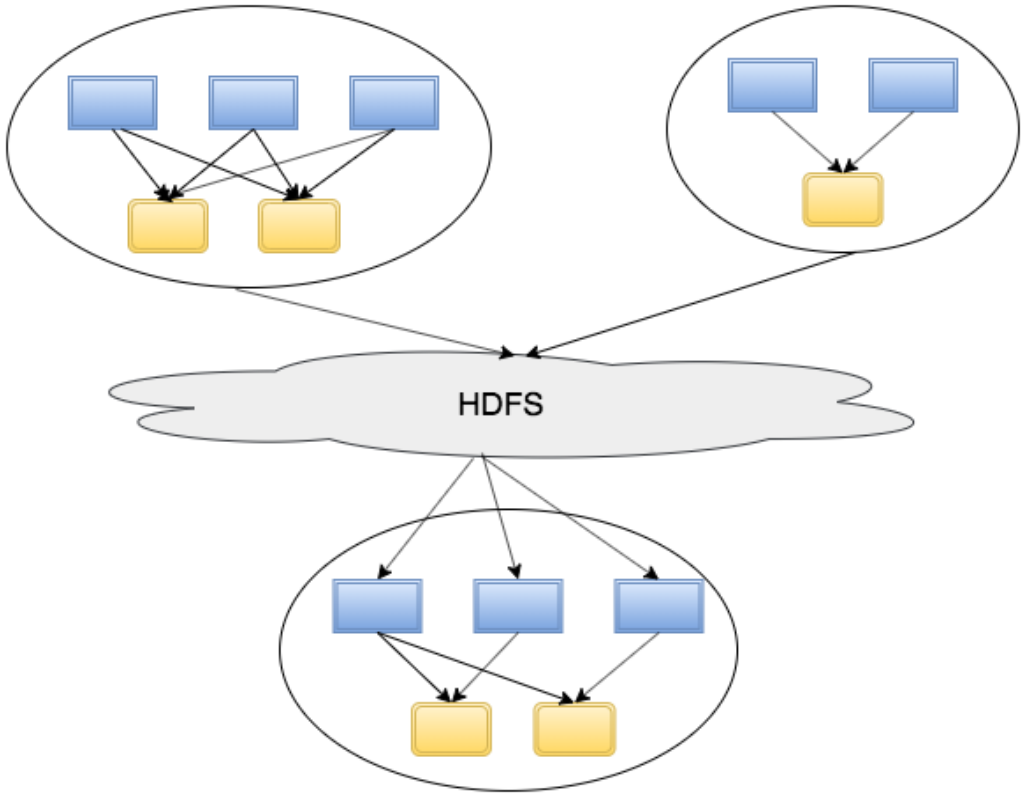

数据仓库管理
指标管理
埋点管理
元数据管理
数据仓库模型
Default仓库(原始日志，ODS)
Custom仓库(维度+宽表)
业务仓库(内聚+复用)
产品仓库(抽取关联，关系模型)
数据治理
自助建仓

架构-数据统计

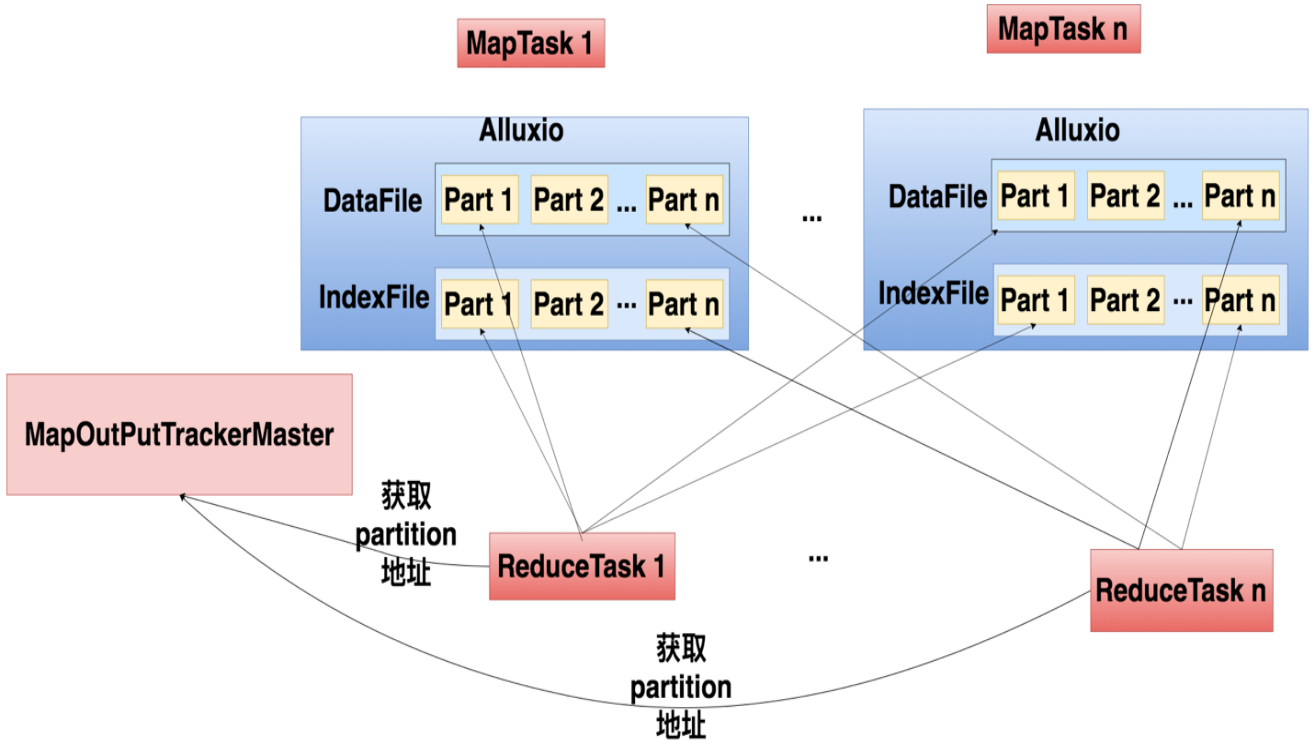
统计分析引擎的对比

	Hive	SparkSQL	Phoenix	Presto
稳定性	优	良	中	差
查询性能	差	良	良	优
并发性	优	良	差	差
扩展性	优	优	优	优
SQL兼容性	良	良	差	良

架构-数据统计

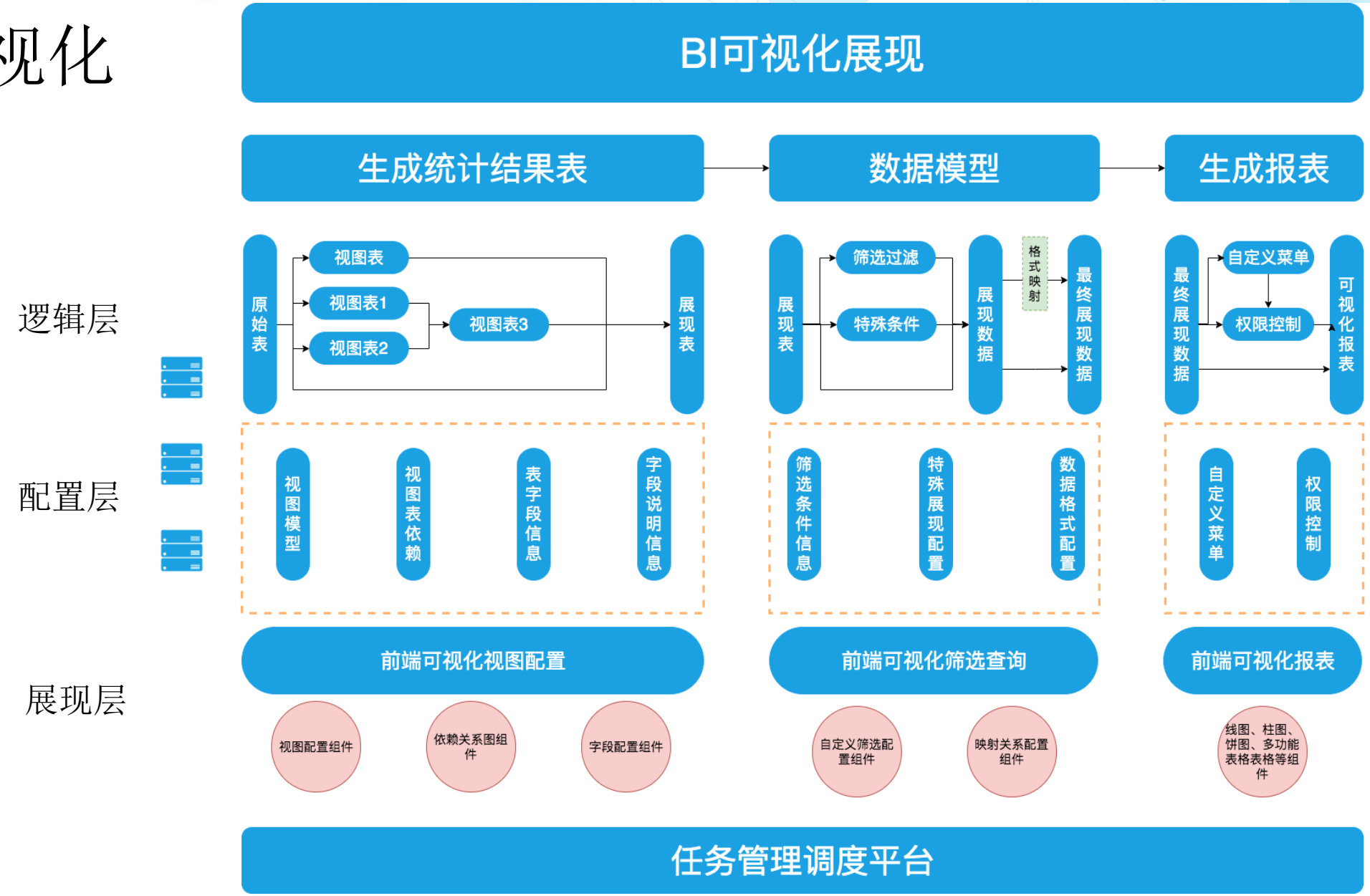


Hive任务执行



SparkOnAlluxio多轮迭代执行

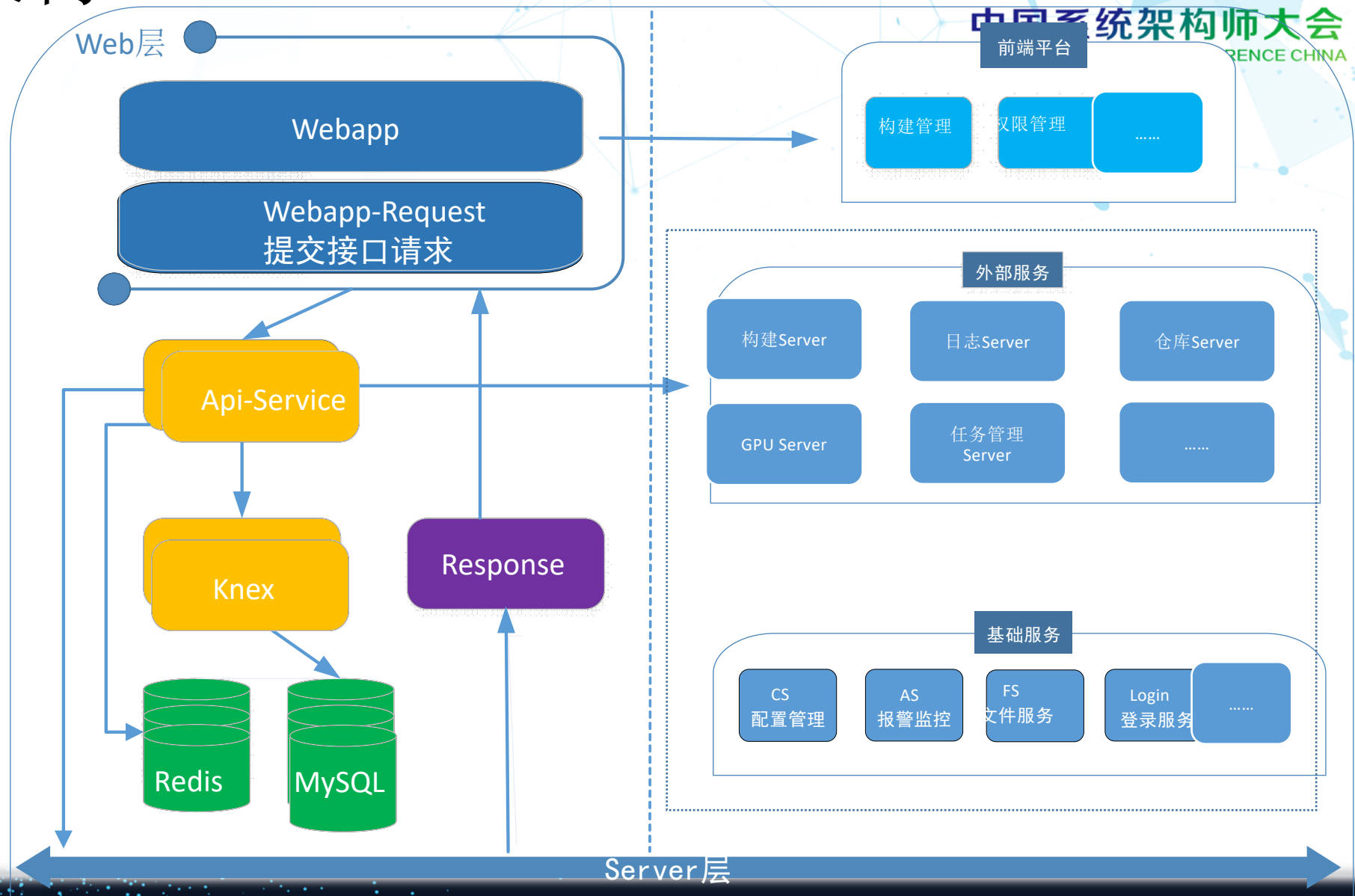
架构-BI可视化


BI可视化展现
生成统计结果表
数据模型
生成报表
逻辑层
原始表
视图表
视图表1
视图表2
视图表3
展现表
展现表
筛选过滤
特殊条件
展现数据
格式映射
最终展现数据
最终展现数据
自定义菜单
权限控制
可视化报表
配置层
视图模型
视图表依赖
表字段信息
字段说明信息
筛选条件信息
特殊展现配置
数据格式配置
自定义菜单
权限控制
展现层
前端可视化视图配置
前端可视化筛选查询
前端可视化报表
视图配置组件
依赖关系图组件
字段配置组件
自定义筛选配置组件
映射关系配置组件
线图、柱图、饼图、多功能表格表格等组件
任务管理调度平台

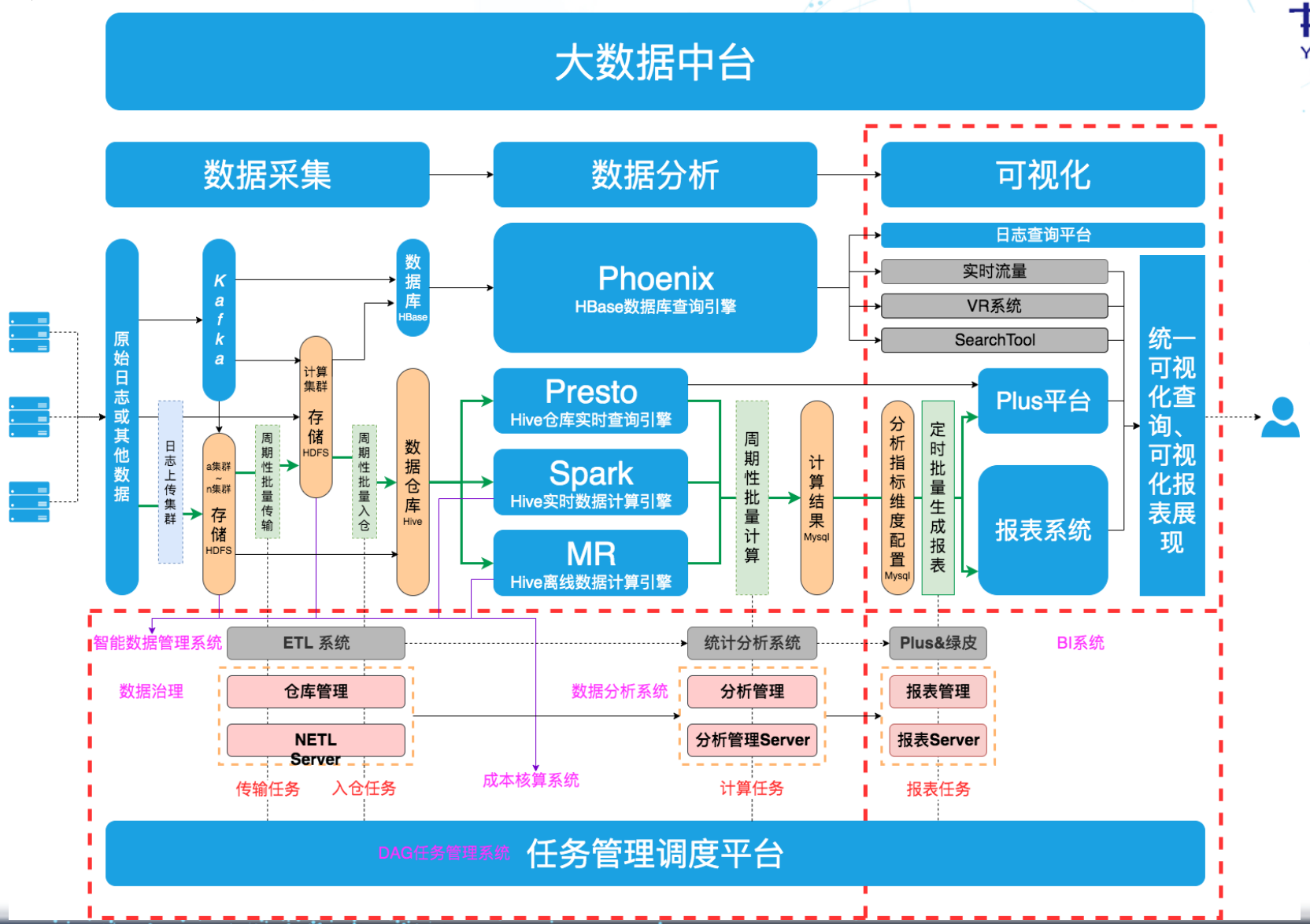
3 Sogou 数据中台架构

Sogou数据中台架构

大数据平台的业务具有多流程、多作业、业务低耦合的特性，因此从技术实现上来说，使用MVVM这种开发模式比较适合，以此达到前后端彻底分离，各个业务模块只需要提供API即可



Sogou数据中台架构

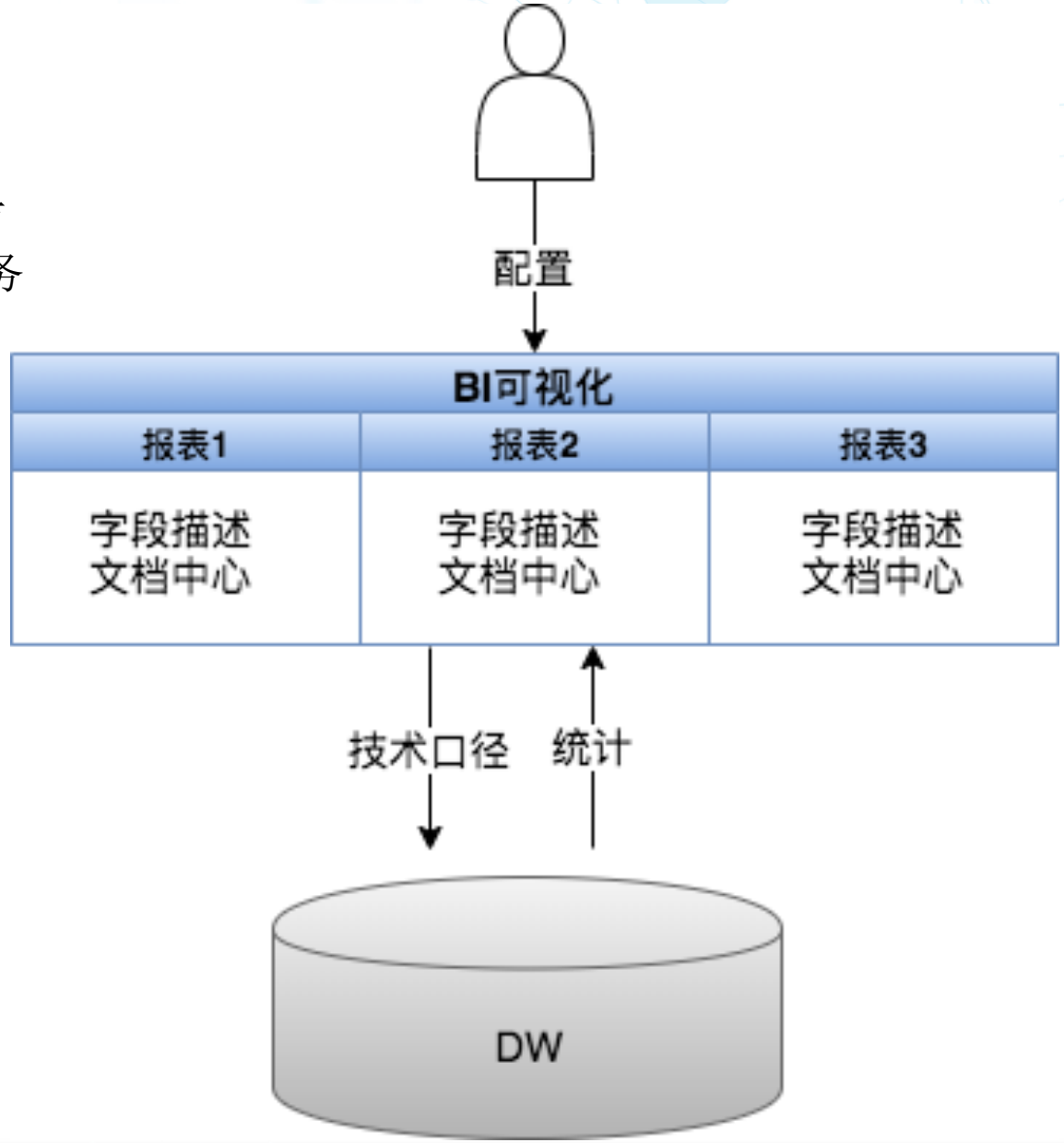


Sogou数据中台架构

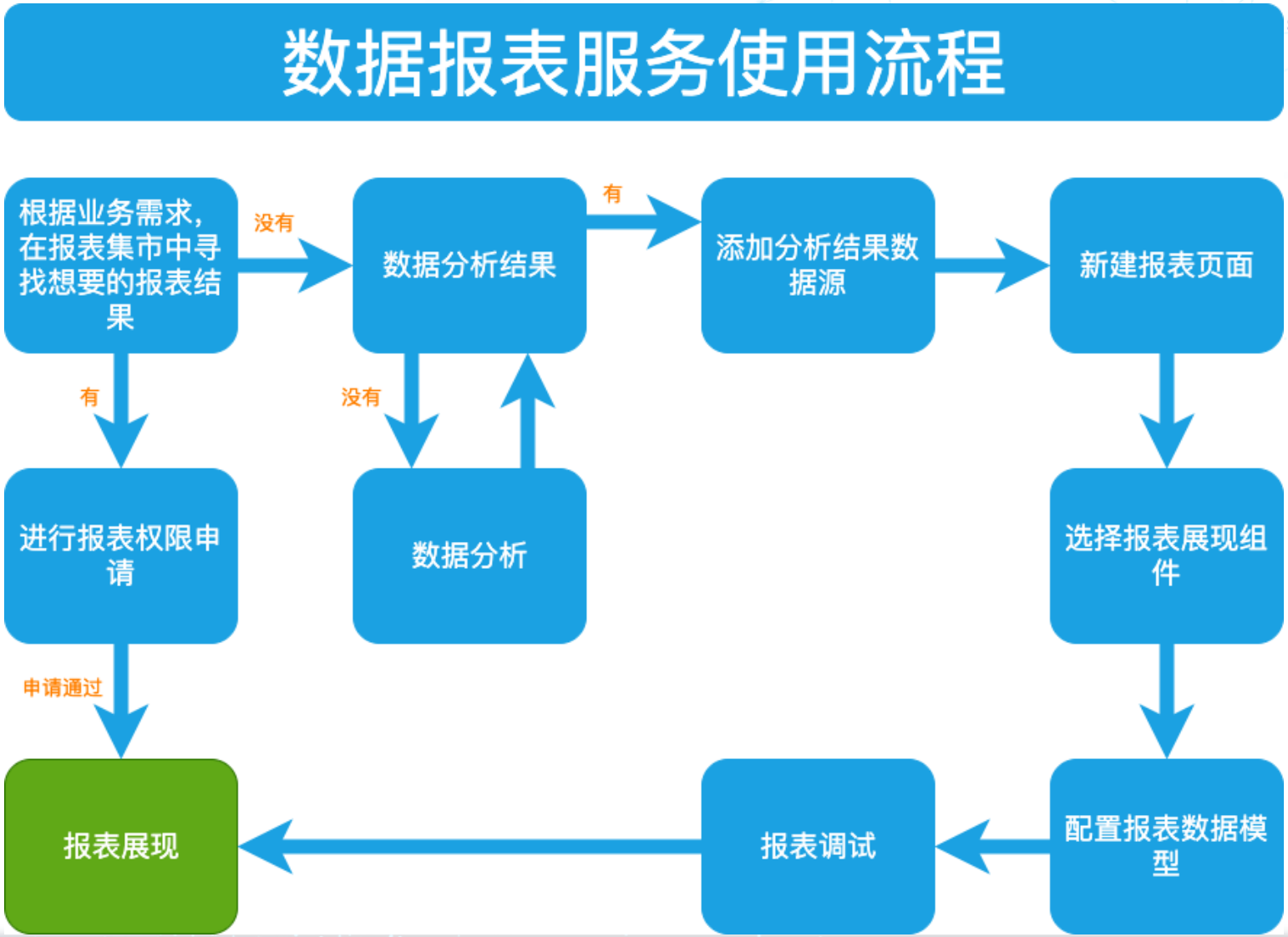
业务查看报表集市
查看/建立报表服务

统一技术口径

数据仓库建模



Sogou数据中台架构



Thanks