美图统计分析平台架构实现

卢荣斌

2017.07.30

大纲

01 统计业务与技术碰撞

02 **美图统计平台架构实现**

03 未来规划

» 01

统计业务与技术碰撞

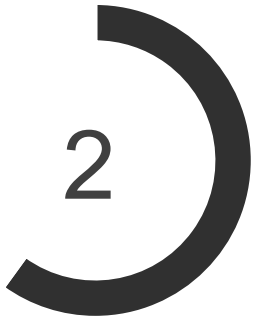
业务发展与技术迭代的过程

第一阶段：项目初期

以美拍为例，初期统计业务的特点

体量小

初期数据量比较小



统计需求较少

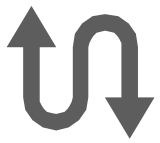
主要是一些基础的统计指标

快速响应

初期产品快速迭代，要求数据指标响应跟上迭代速度

第一阶段：初期实现方案

初期实现简单快速



Rsync

通过rsync方式收集到一个节点

Crontab

线上配置定时任务

Shell & Php

采用快速简单的脚本语言实现统计

MySQL

数据简单存放到MySQL供展示查询

第二阶段：快速发展

用户量爆发

数据量日渐增大

统计需求膨胀

VS

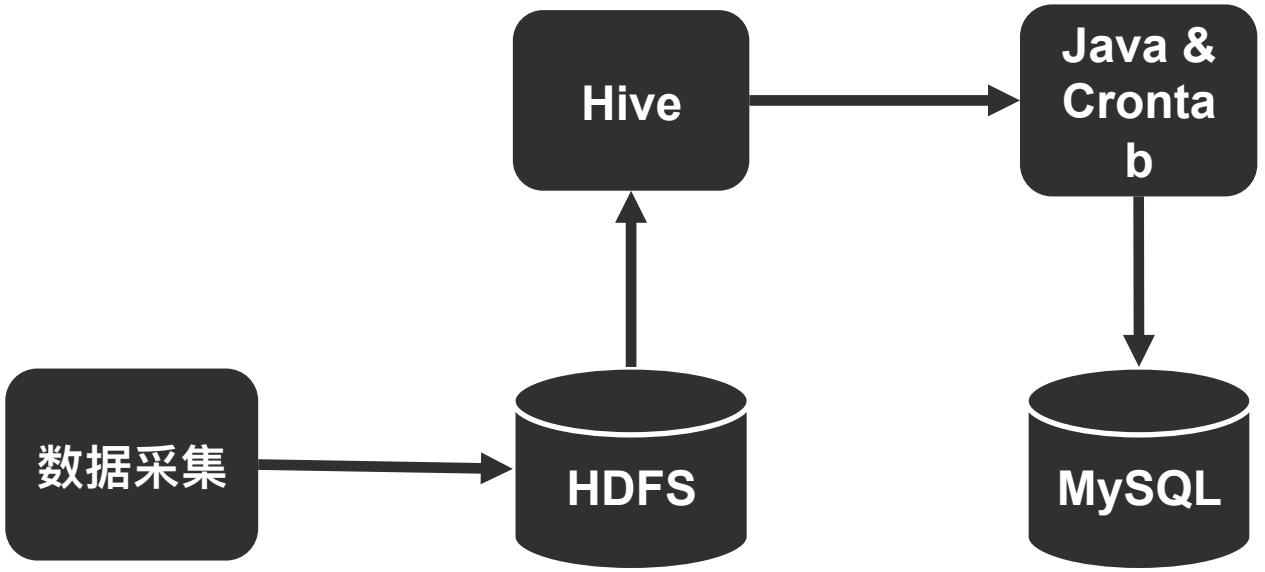
存储容量

计算

脚本维护

第二阶段：实现演进

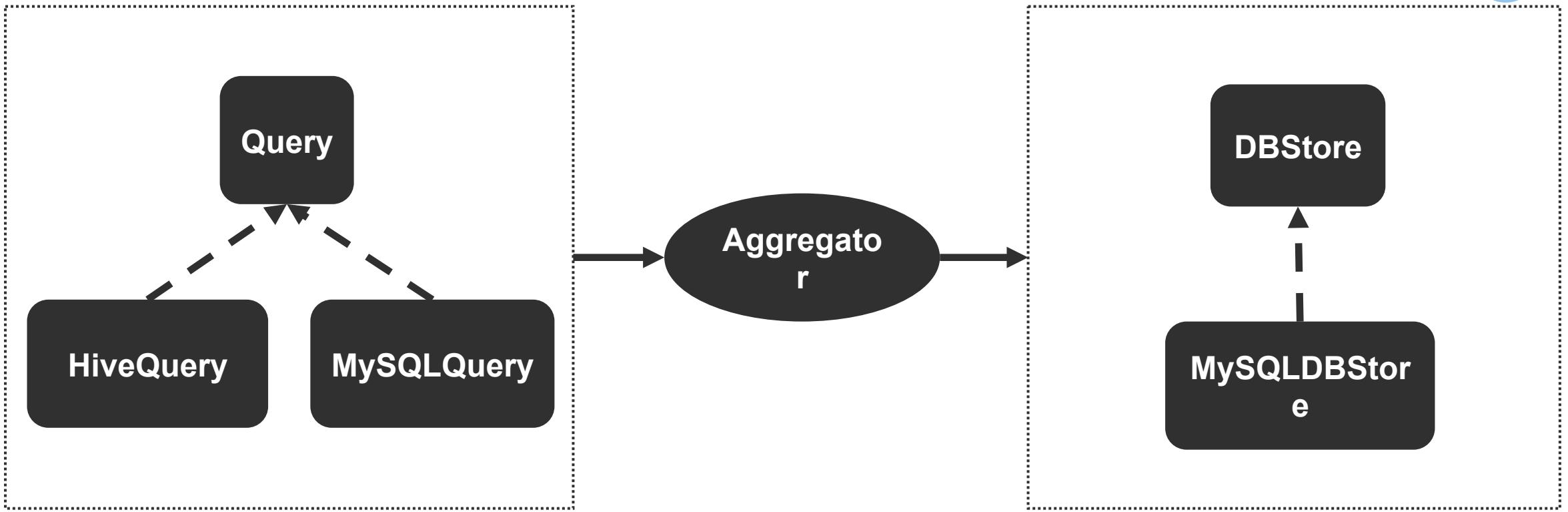
- 实现采集系统
- 搭建Hadoop分布式存储与计算
- 搭建基于Hive数据仓库
- Java替换Shell等脚本语言



第三阶段：

有追求的程序员

拒绝重复劳动，构建统计组件，提高生产力。



接下来：还有痛点

- 一定程度减少了编码量，但还是有一定的编码成本
- 每上线一个统计任务，需要发布一次包。

- ❑ 业务依赖
- ❑ 重复编码量
- ❑ 运维成本
- ❑ 个人成长

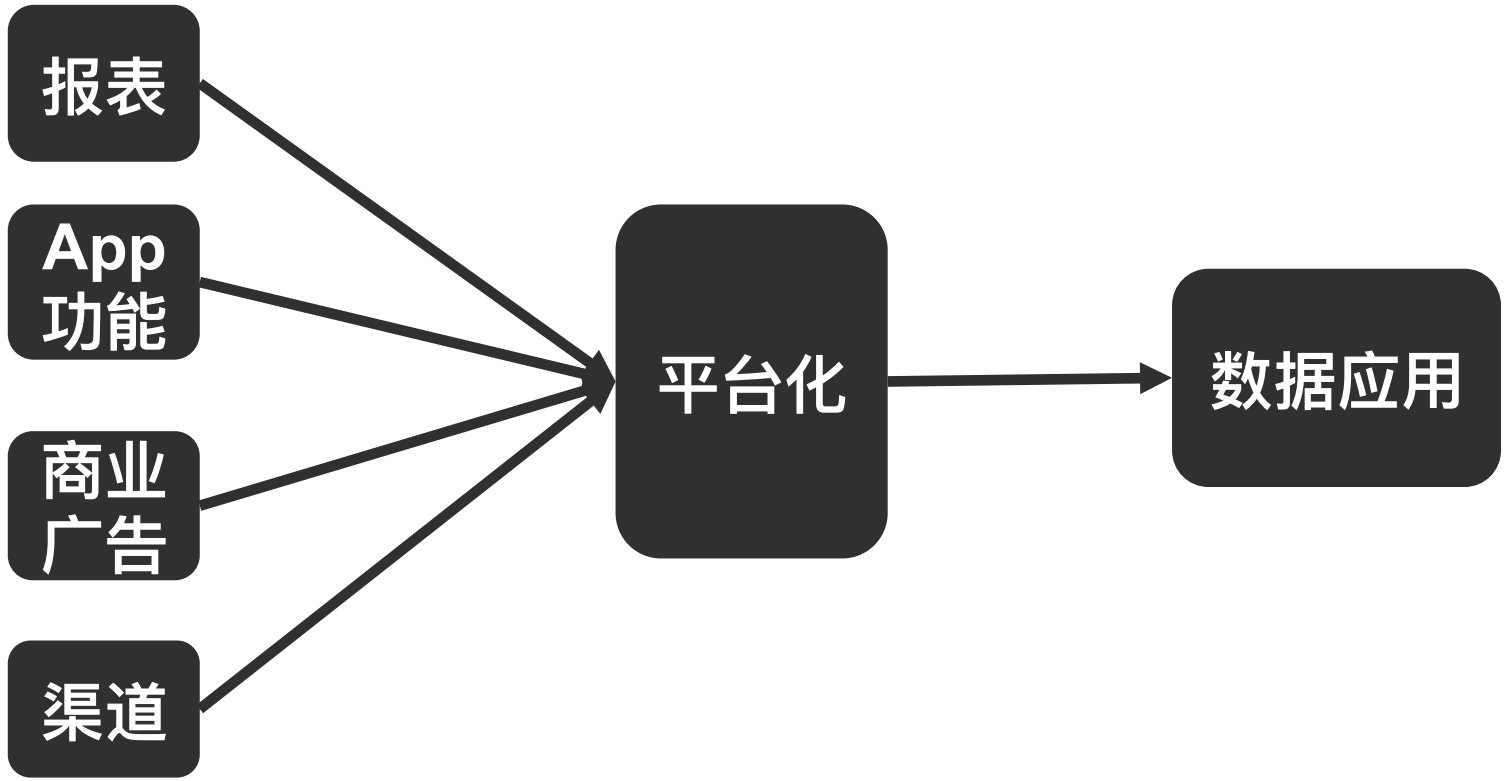
02

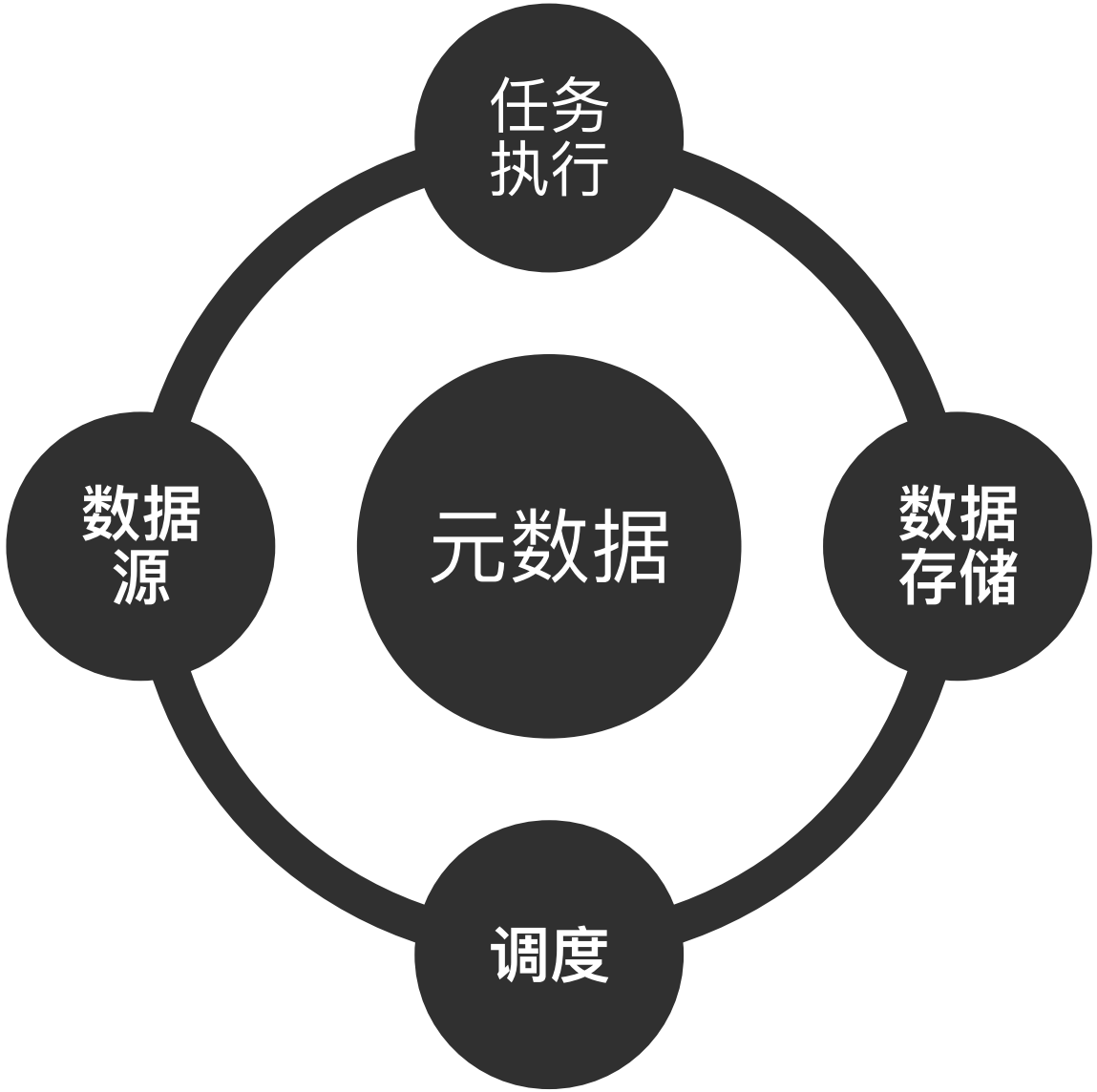
统计平台架构实现

构建平台，提供服务。

统计平台化

提供平台，解耦业务依赖





统计平台化

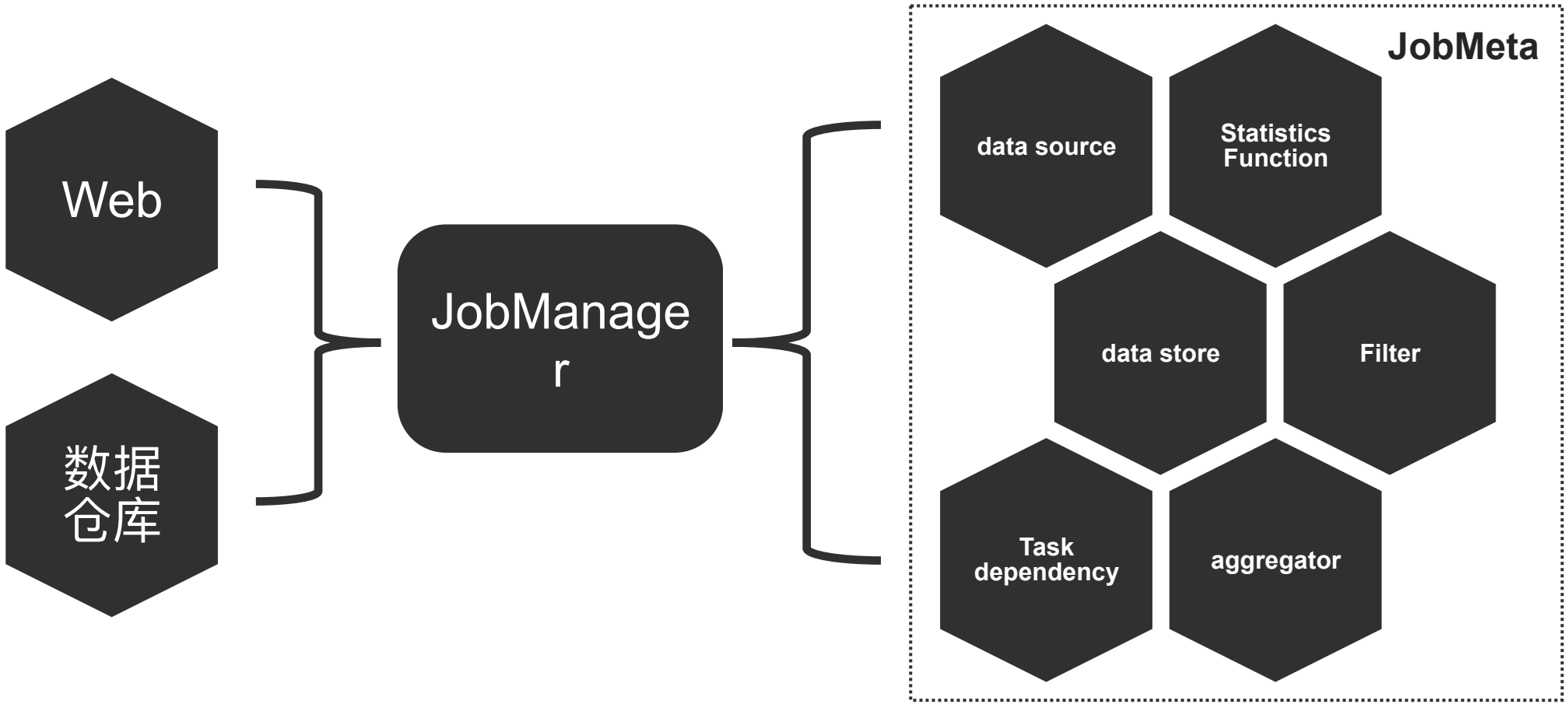
抽象与模块化

«

统计平台化-模块化

统计平台化-任务元数据管理

整合数据仓库、Web、任务元数据管理

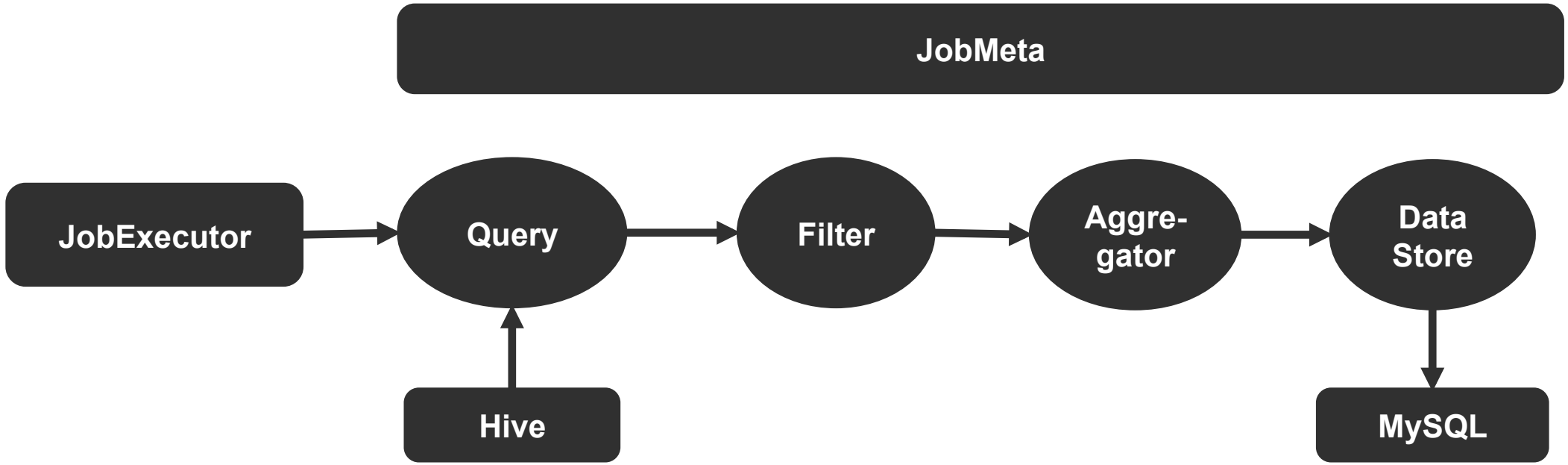


统计平台化：任务调度

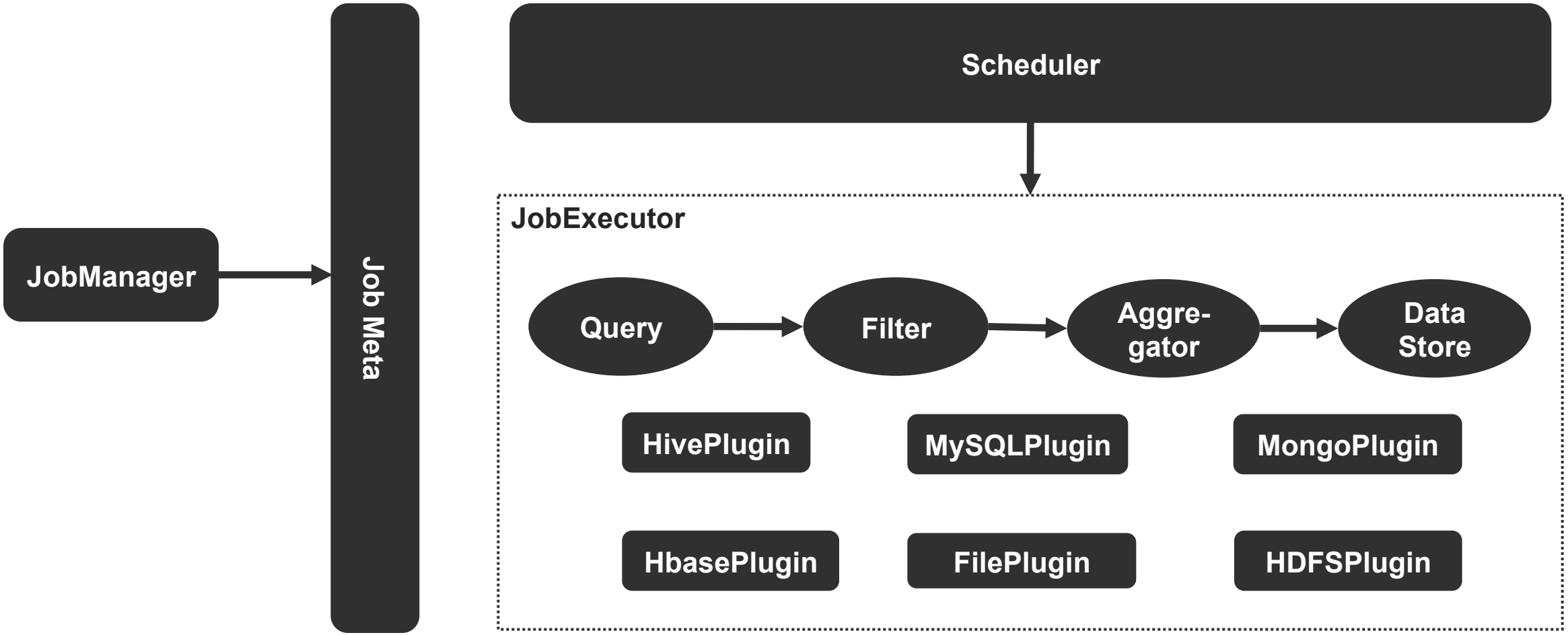
相对简单，满足需求

统计平台化：任务执行

流程、插件化



统计平台化：基础架构


Scheduler
JobManager
Job Meta
JobExecutor
Query
Filter
Aggre-
gator
Data
Store
HivePlugin
MySQLPlugin
MongoPlugin
HbasePlugin
FilePlugin
HDFSPlugin

统计平台化：功能扩展

丰富功能，解决更多业务场景

临时取数

- HQL语法解析
- 语句校验

多数据源

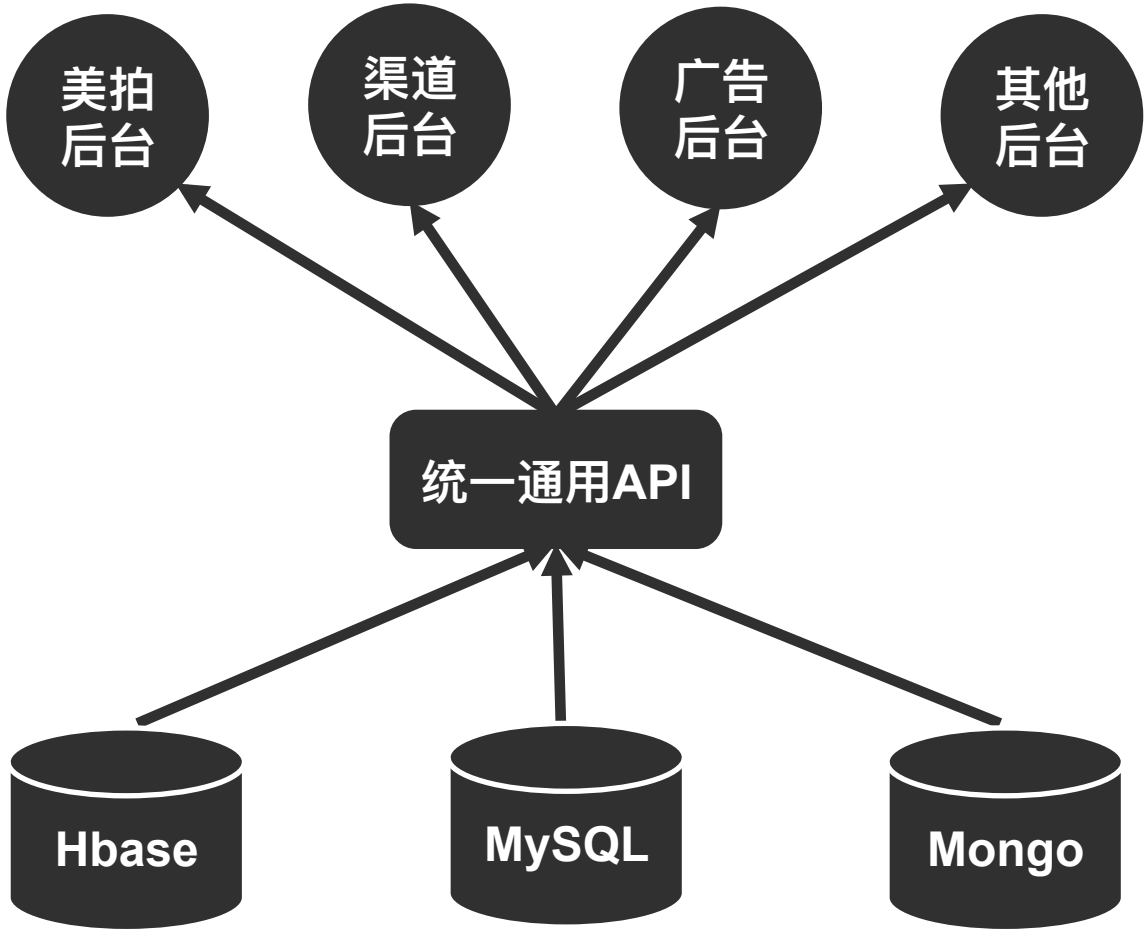
- Hive
- MySQL
- Bitmap

多存储

- MongoDB
- HDFS
- File
- MySQL

统计算子

- 去重
- 数组
- TopN
- 自定义UDF



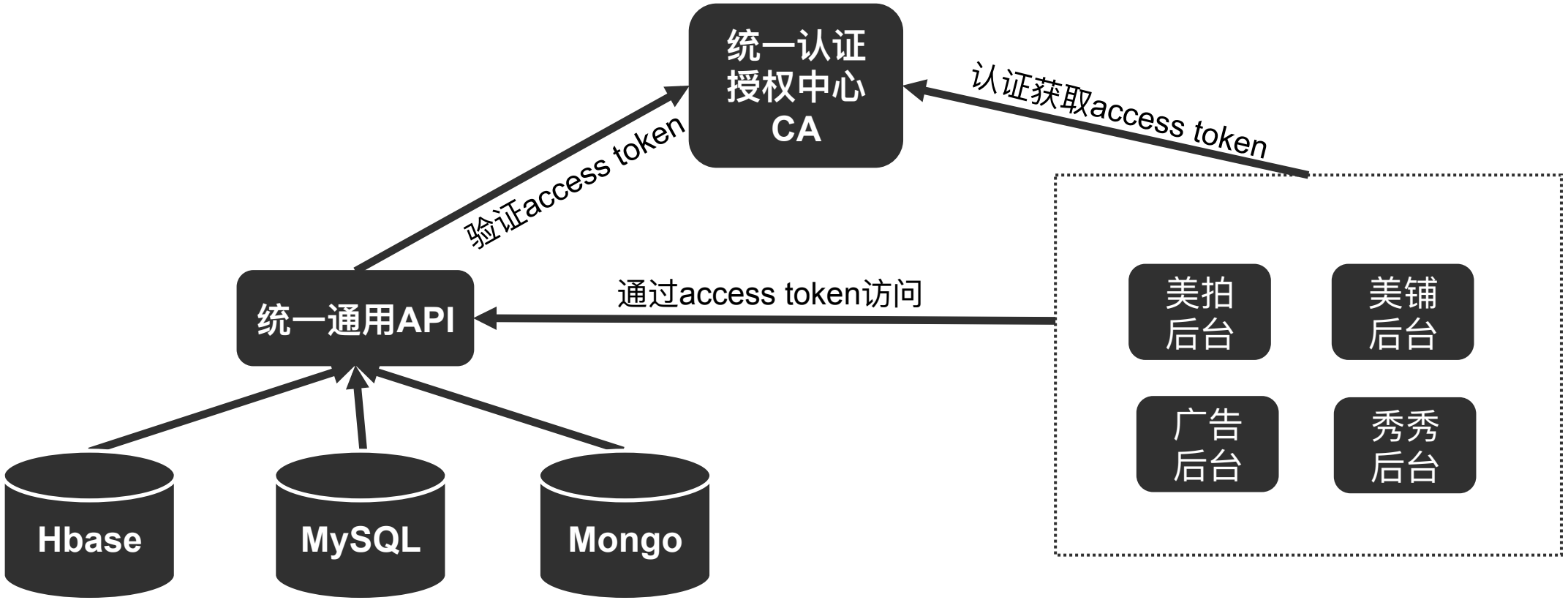
统计平台化：可视化

抽象存储层，统一规范API

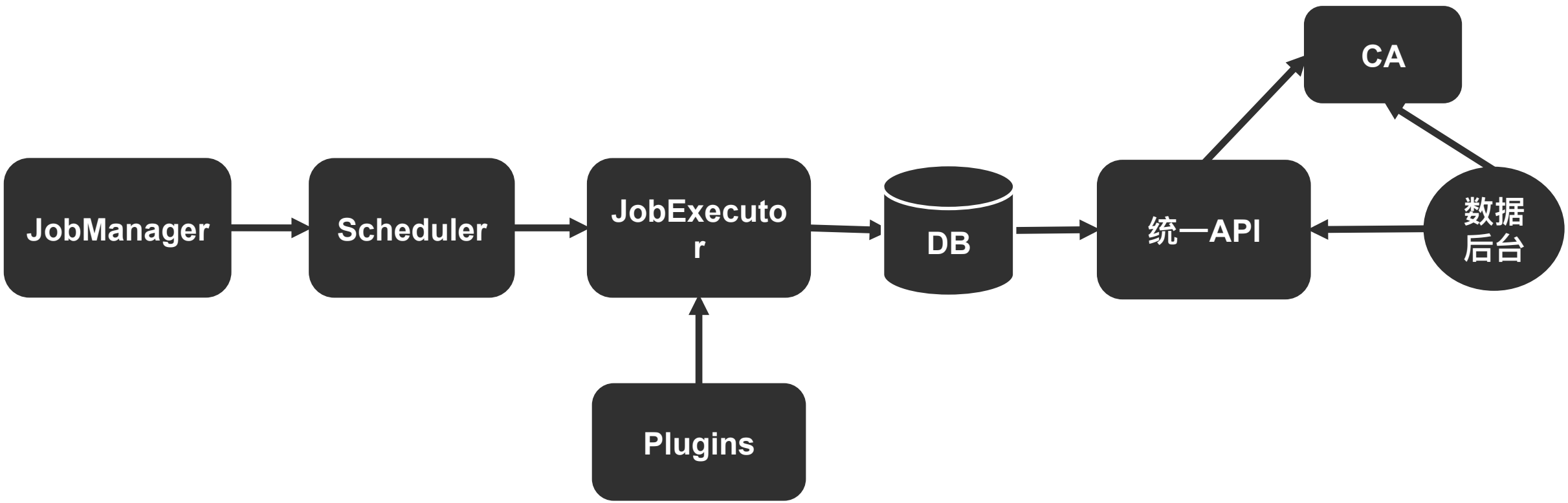
«

统计平台化：安全

对接内部服务统一认证授权中心



统计平台化-架构总结


CA
JobManager
Scheduler
JobExecutor
DB
统一API
数据后台
Plugins

» 03

未来规划

在做以及未来即将开展的功能

未来规划

分布式调度

支持任务与资源的分布式调度

OLAP

针对数据分析人员的即席查询

数据可视化

统一可视化平台

实时统计

能比较快速便捷地接入实时统计

Thanks

Q&A