饿了么大数据平台建设

主讲人：毕洪宇

时间：2017.06.11

拉扎斯网络科技（上海）有限公司
Rajax Network & Tehnology Co

About me

毕洪宇

eBay/PPTV Database Engineer
VIPSHOP Staff Database/Big Data Engineer
饿了么 Big Data Platform Director

Agenda

大数据平台现状

面临的挑战

技术选型

架构设计

稳定性

工具链

平台的一些想法

大数据平台现状

2015年5月

团队成立

2年

- # 离线计算

增量(不考虑副本) 100TB/day

集群规模 100-1000 nodes， X10 expanding

表数据 90K表　400报表

调度任务 20K

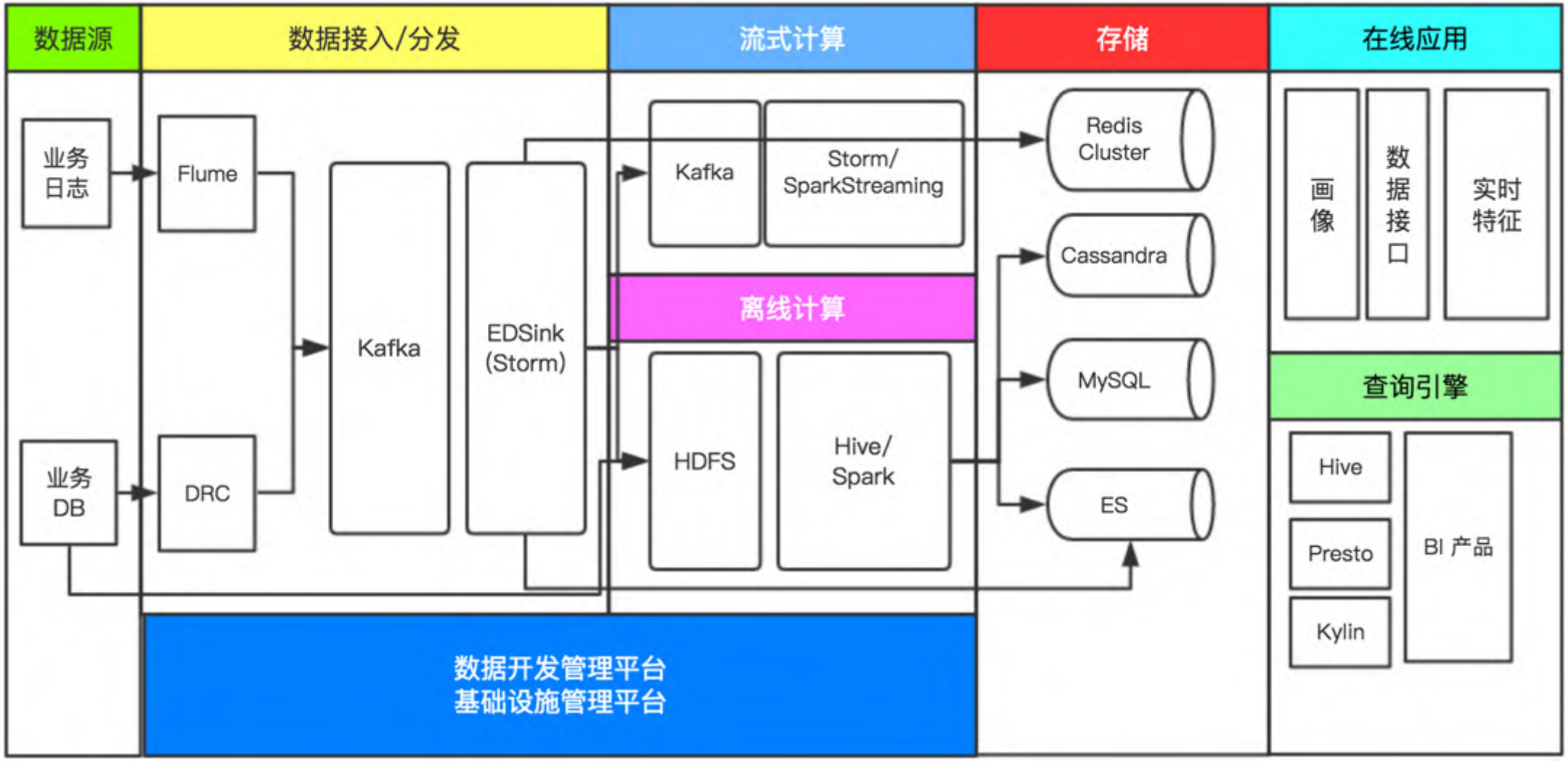
任务数 80K mapreduce/spark

计算数据吞吐 3PB/day

- 实时计算

集群规模	kafka	50+ Topology
10-100+ nodes	1M+records/s	8+ GB/s ，2M+ records/s

- 逻辑架构与数据流向


数据源
数据接入/分发
流式计算
存储
在线应用
业务日志
业务DB
Flume
DRC
Kafka
EDSink (Storm)
Kafka
Storm/ SparkStreaming
离线计算
HDFS
Hive/ Spark
Redis Cluster
Cassandra
MySQL
ES
画像
数据接口
实时特征
查询引擎
Hive
Presto
Kylin
BI 产品
数据开发管理平台
基础设施管理平台

面临的挑战

人少活多 积累不足

内在质量差不多就行

应对套路"千人千面"

Fire and Forget

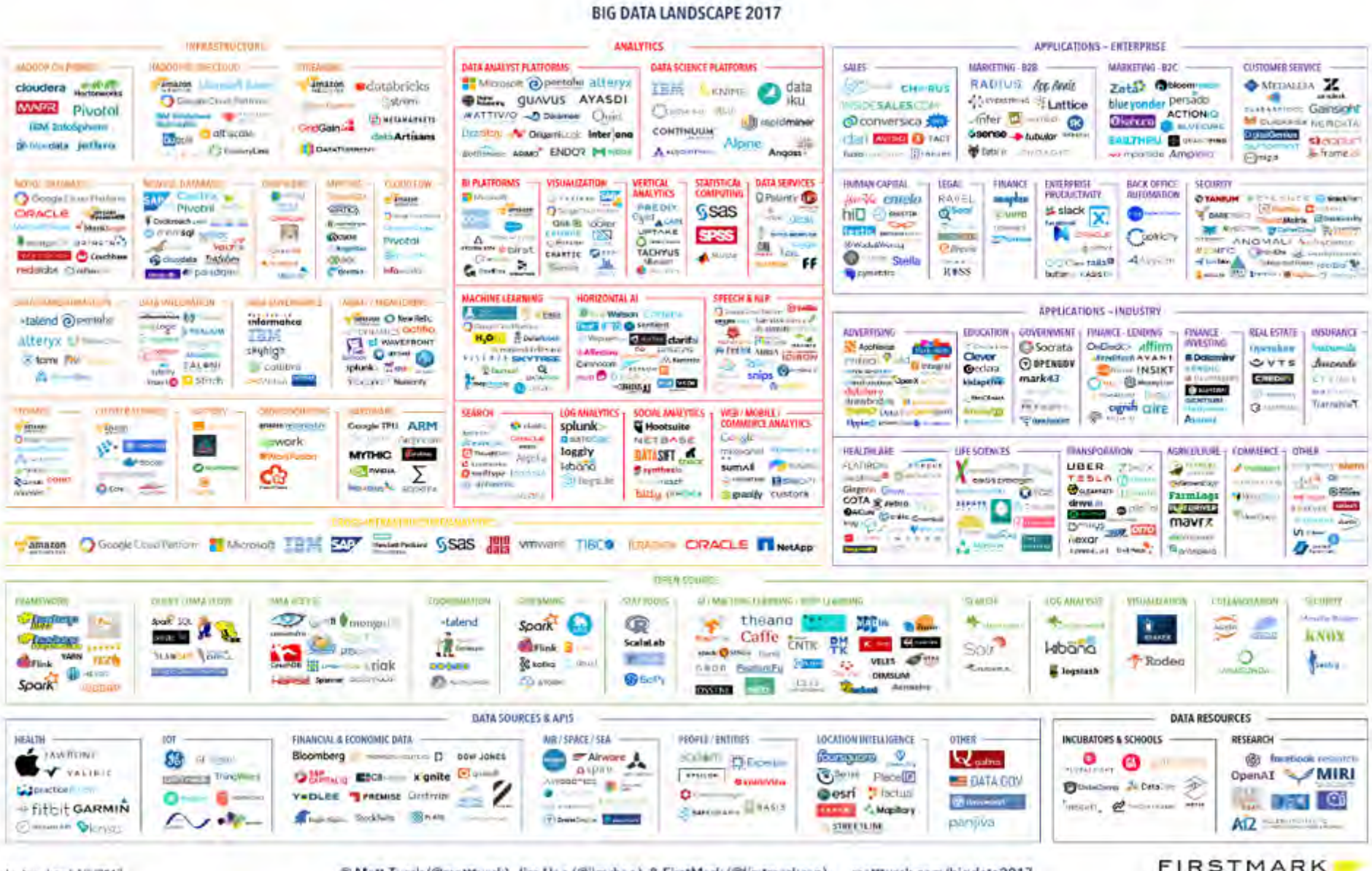
效率

质量

持续扩展

技术选型

4.1

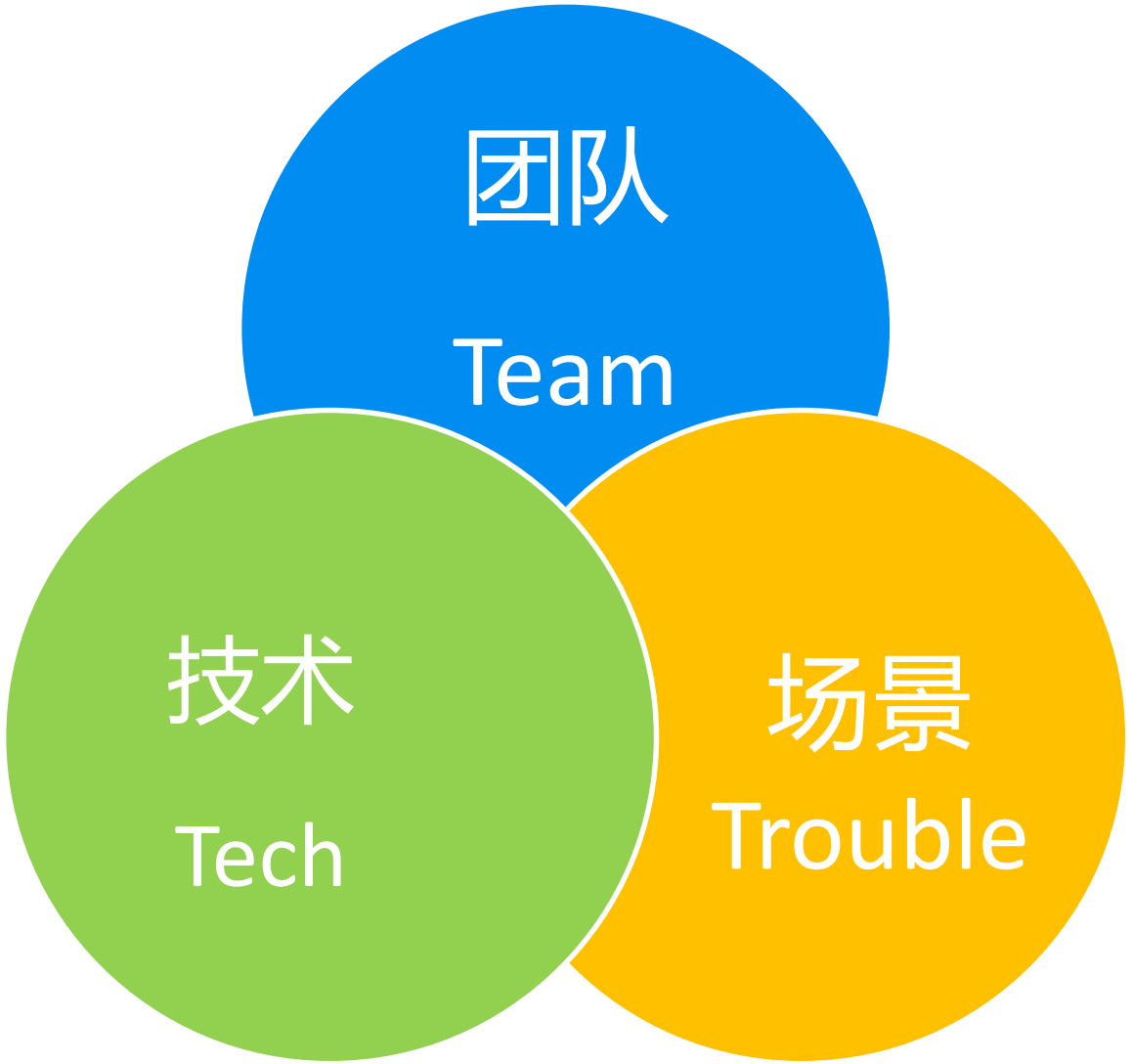

BIG DATA LANDSCAPE 2017
INFRASTRUCTURE
ANALYTICS
APPLICATIONS - ENTERPRISE
APPLICATIONS - INDUSTRY
DATA SOURCES & APIS
DATA RESOURCES
Last updated 4/5/2017
© Matt Turck (@mattturck), Jim Hao (@jimrhao), & FirstMark (@firstmarkcap)
mattturck.com/bigdata2017
FIRSTMARK

3T

Trouble：解决什么问题

Tech：哪些合适的技术，生态和社区的状态

Team：熟悉程度 学习成本 使用成本 运维成本



- # Presto VS Hive VS Spark SQL

场景

海量数据 Ad-hoc查询

团队

Presto略熟

技术

社区方面 Spark

稳定性 Presto

使用成本 Spark SQL

- # HBase VS Cassandra

场景

海量存储/批量更新/K-V(Object)访问

团队

不熟悉 学习成本差不多

技术

社区方面 国内HBase VS 国外Cassandra

运维成本 Cassandra较低

使用成本 Cassandra功能方便

- # Storm VS Spark Streaming VS Flink

场景

实时计算引擎(ms/sec/mins)

团队

Storm>Spark Streaming>Flink

技术

社区方面 Spark Streaming
运维成本 Storm
使用成本 Spark SQL

- 选型的一些心得

关于feature

优点决定是否要谈恋爱

不过就算优点再好

缺点接受不了

如何在一起幸福地过一辈子

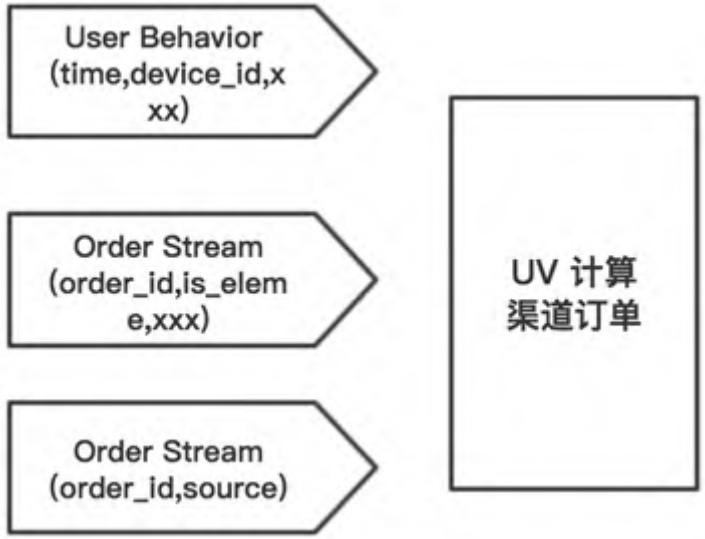
关于决策

喜欢是放肆

爱是克制

架构设计

• 可扩展-同步&一致

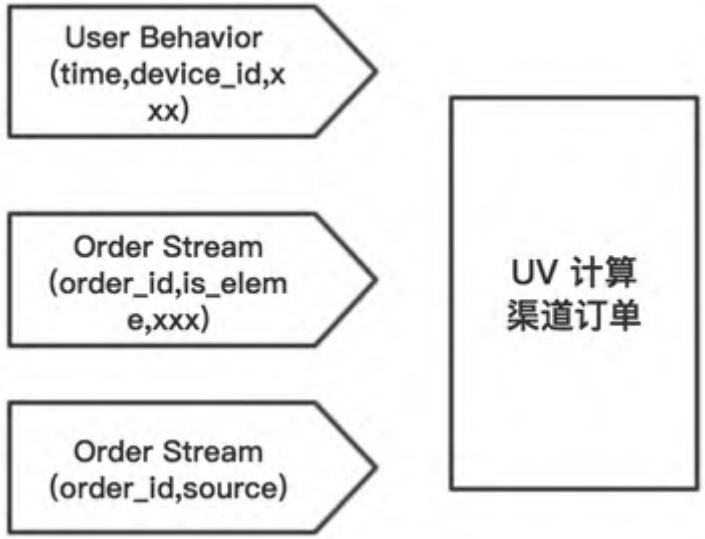


如何多流合并计算：

constraints:
乱序到达
不可预知

When Trigger Action Within an acceptable Time Window.

• 可扩展-避免热点

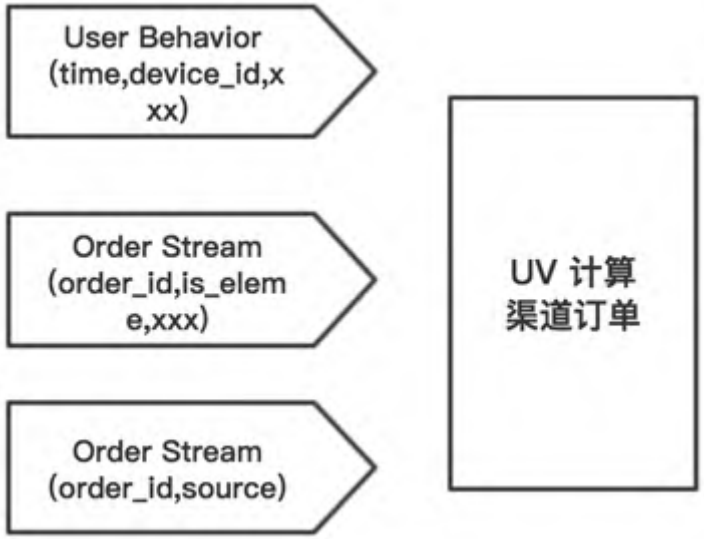


如何进行去重计算：

key: 20170725_uv　set{ device_id }

key: < device_id >_20170725 value: placeholder

• 可扩展-成本



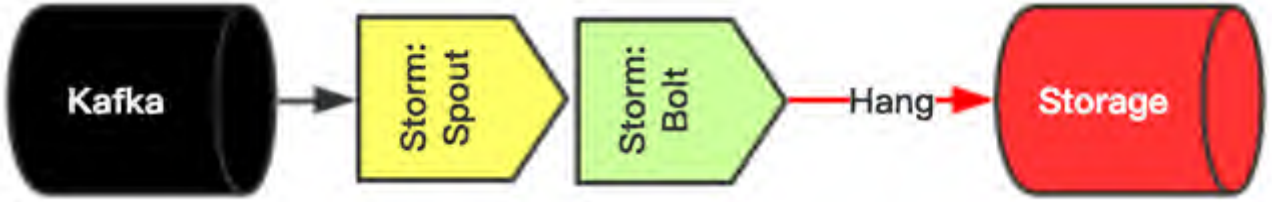
如何进行多维度UV计算：

K-V方案：写入吞吐高，内存占用高

Redis HLL方案：写吞吐降不下来

JVM HLL方案：幂等 & 本地计算

- 可扩展- self stabilization



数据不可丢：enable ACK

why not back-pressure
STORM-1949

数据可丢：circuit break

稳定性

- # 执行计划

回滚方案 回滚方案 回滚方案

灰度 灰度 灰度

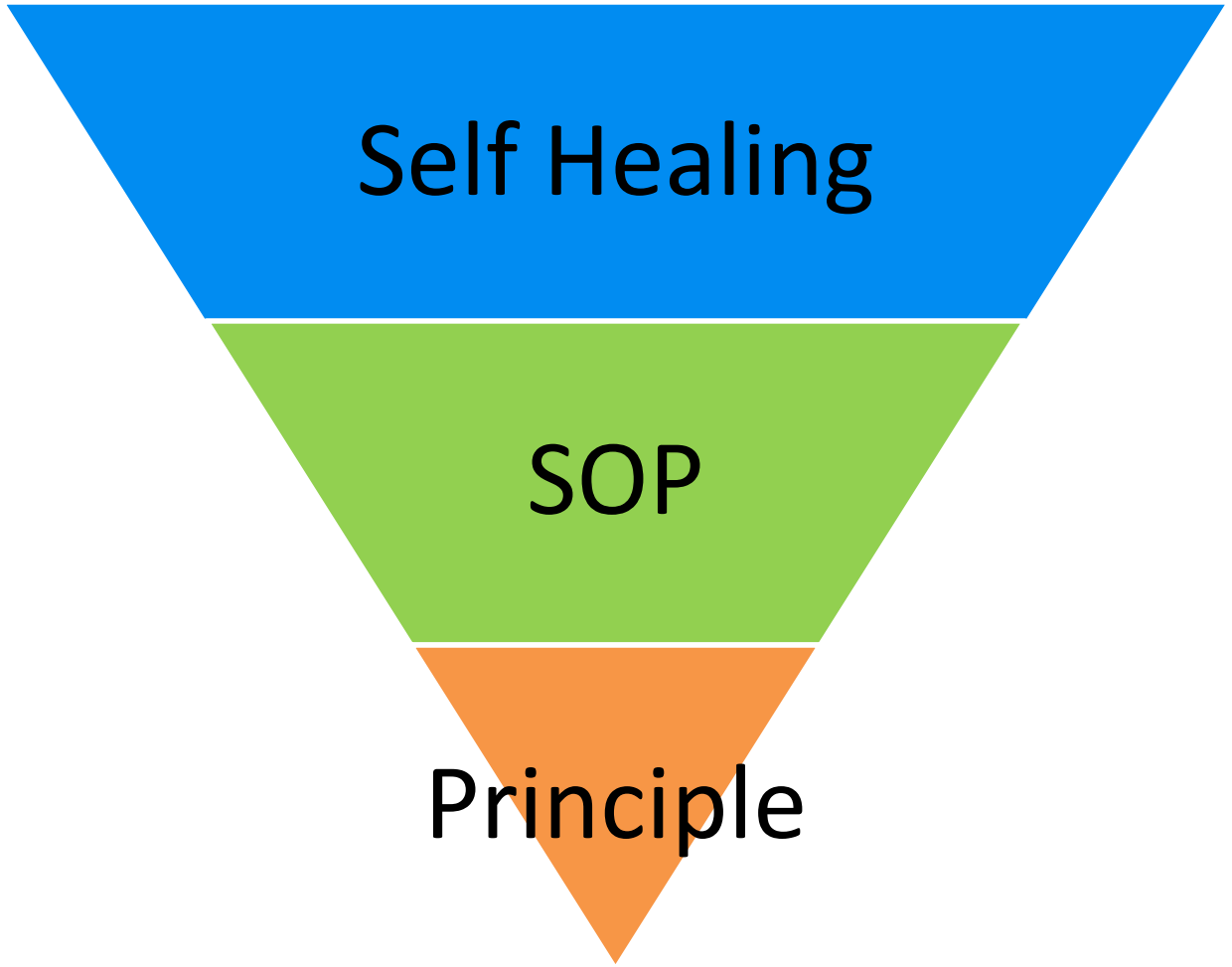
考虑异常流程

时间估算很重要

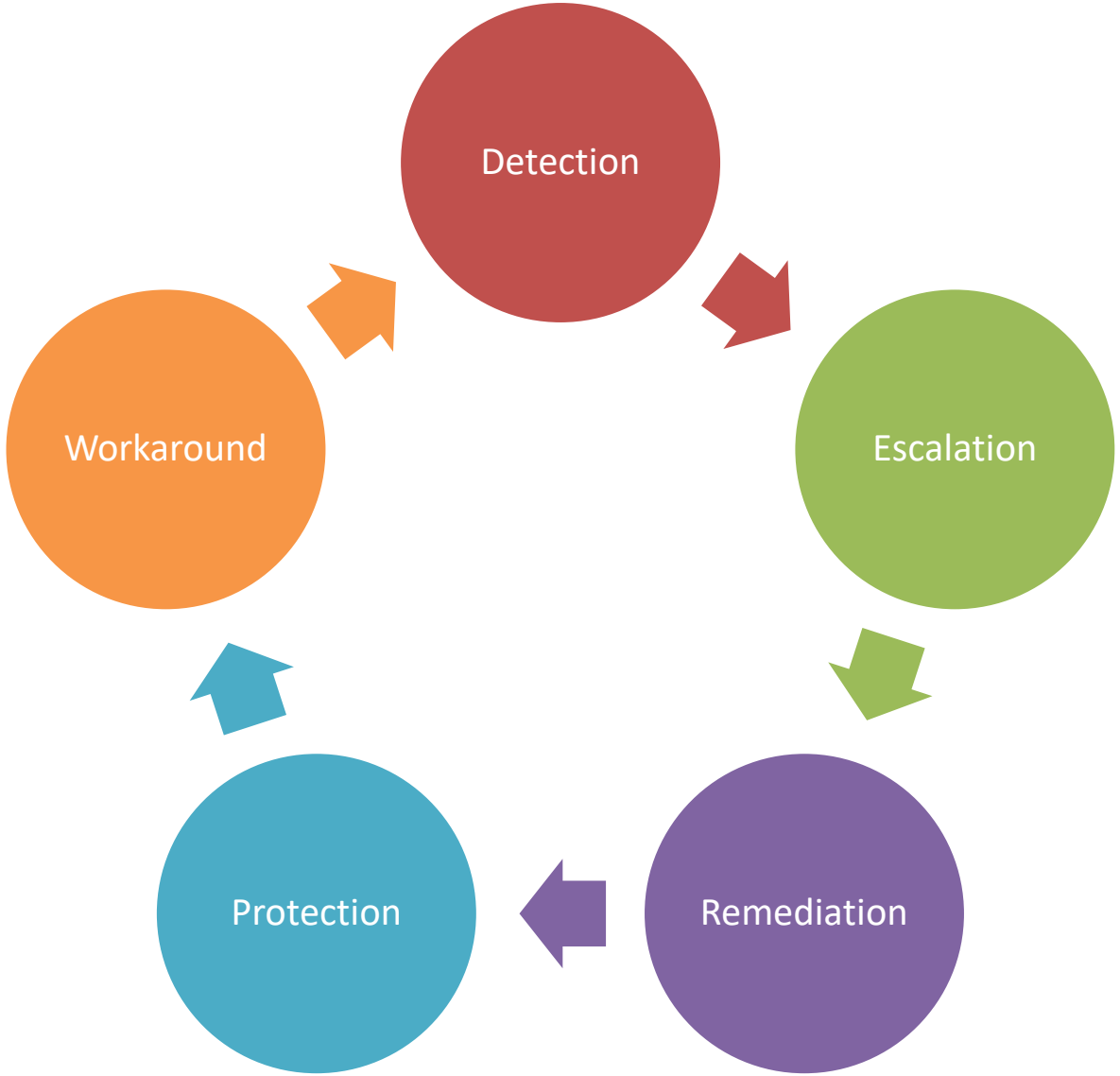
MTTR=告警响应响应+介入时间+处理时间

监控 ≠ 告警
监控 = metrics+trigger+action

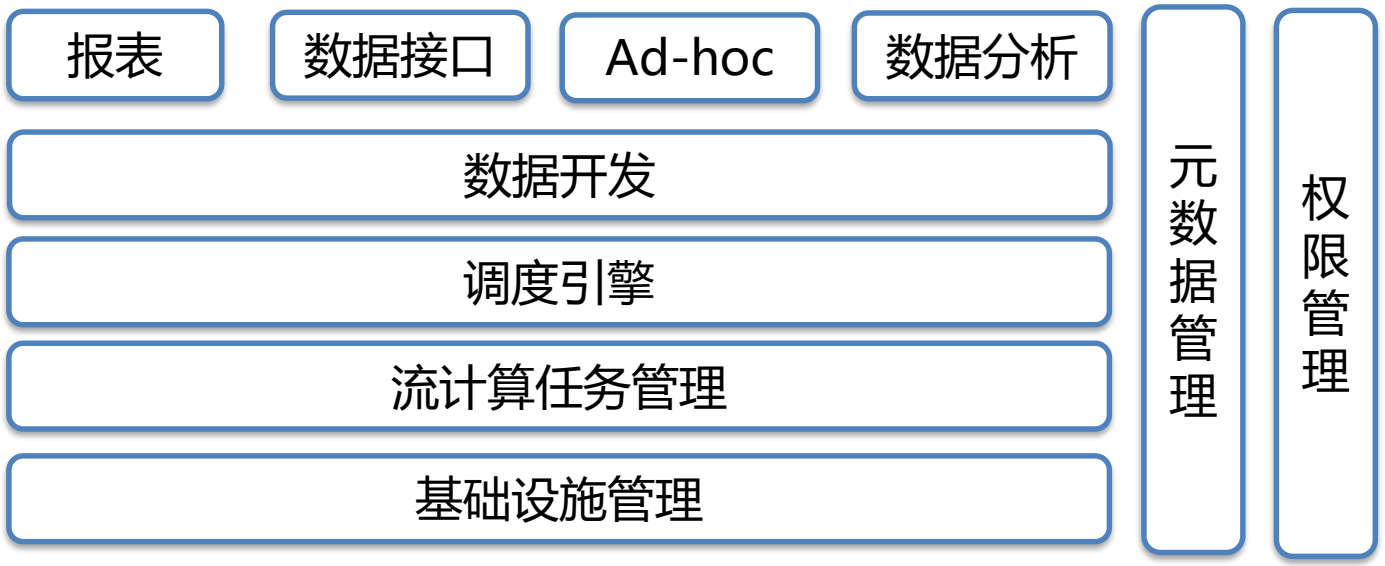
如何控?

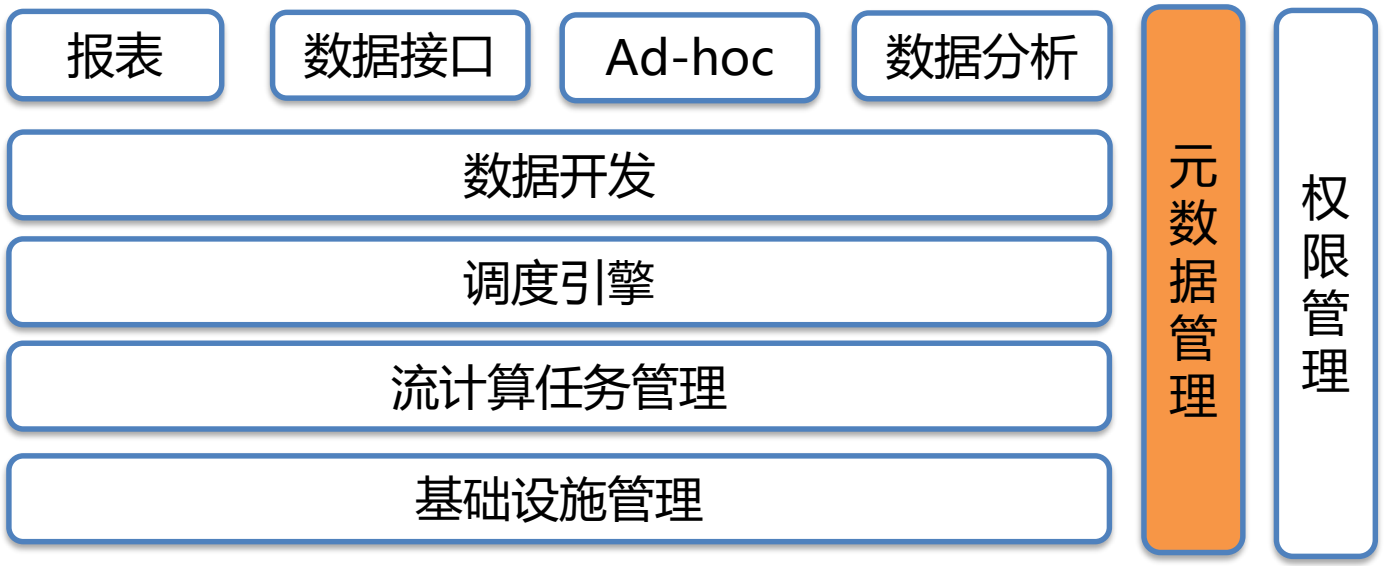


- 复盘原则 DERP(W)


Detection
Escalation
Remediation
Protection
Workaround

工具链


报表
数据接口
Ad-hoc
数据分析
数据开发
调度引擎
流计算任务管理
基础设施管理
元数据管理
权限管理


报表
数据接口
Ad-hoc
数据分析
数据开发
调度引擎
流计算任务管理
基础设施管理
元数据管理
权限管理

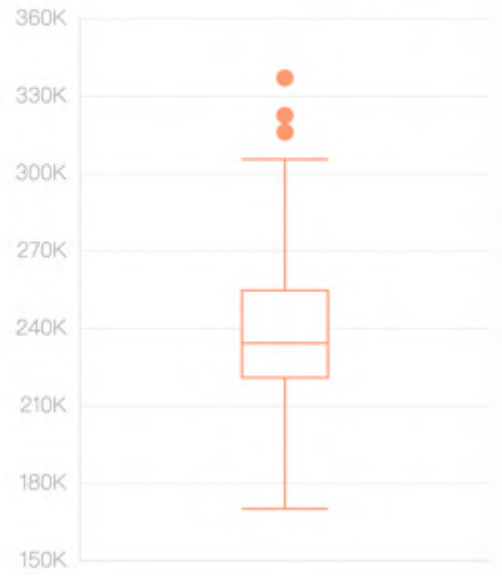
• 数据治理工具

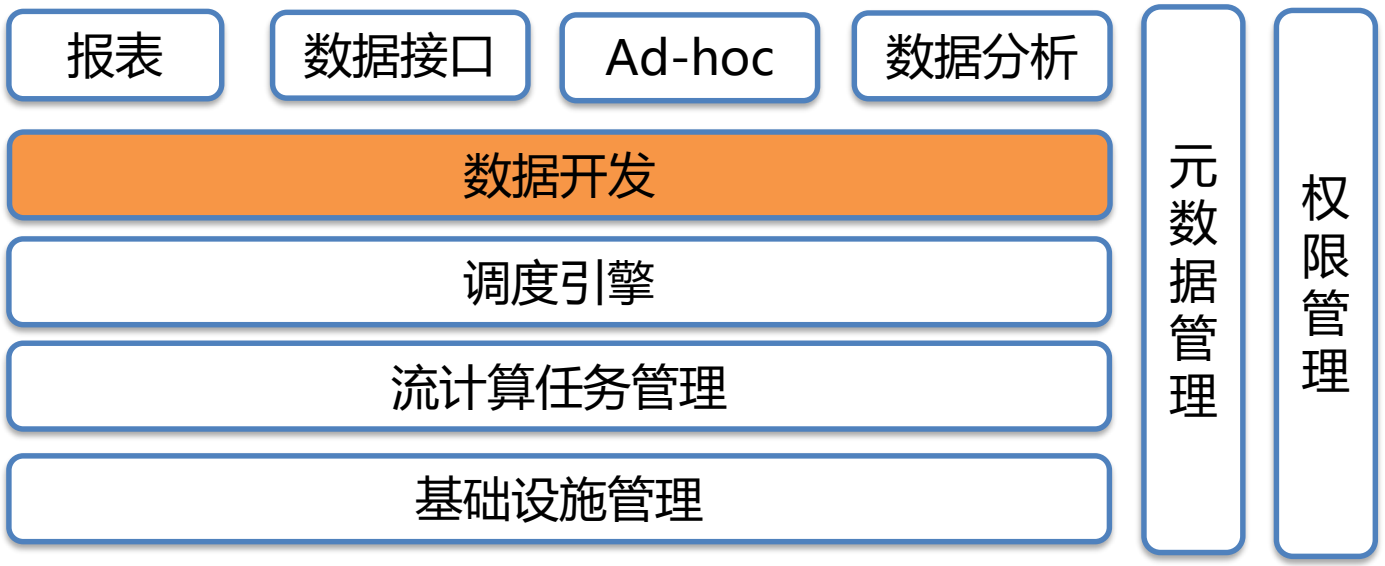
- ➢ 数据地图
  - 指标
  - 报表
  - 表
  - 血缘
- ➢ 数据质量
  - 数据源
  - 策略
  - 触发

当日值(2017/05/17)	历史最大值	历史最小值	阈值	昨日环比	昨日环比波动阈值	上周同比	上周同比波动阈值
273,363	336,819 2017/04/09	188,869 2017/05/02	(180000, 300000) 正常	+14.3%	(-10%, +10%) +4.3%	+16.83%	(-5%, +5%) +11.83%

历史趋势

箱线图:




报表
数据接口
Ad-hoc
数据分析
数据开发
调度引擎
流计算任务管理
基础设施管理
元数据管理
权限管理

- 数据开发管理平台

• 数据开发管理平台-表管理

- ➢ 静态数据

  主题/类型/分区/维度/度量

  生命周期/备份周期

  格式/存储

  是否允许删

  是否敏感/加密

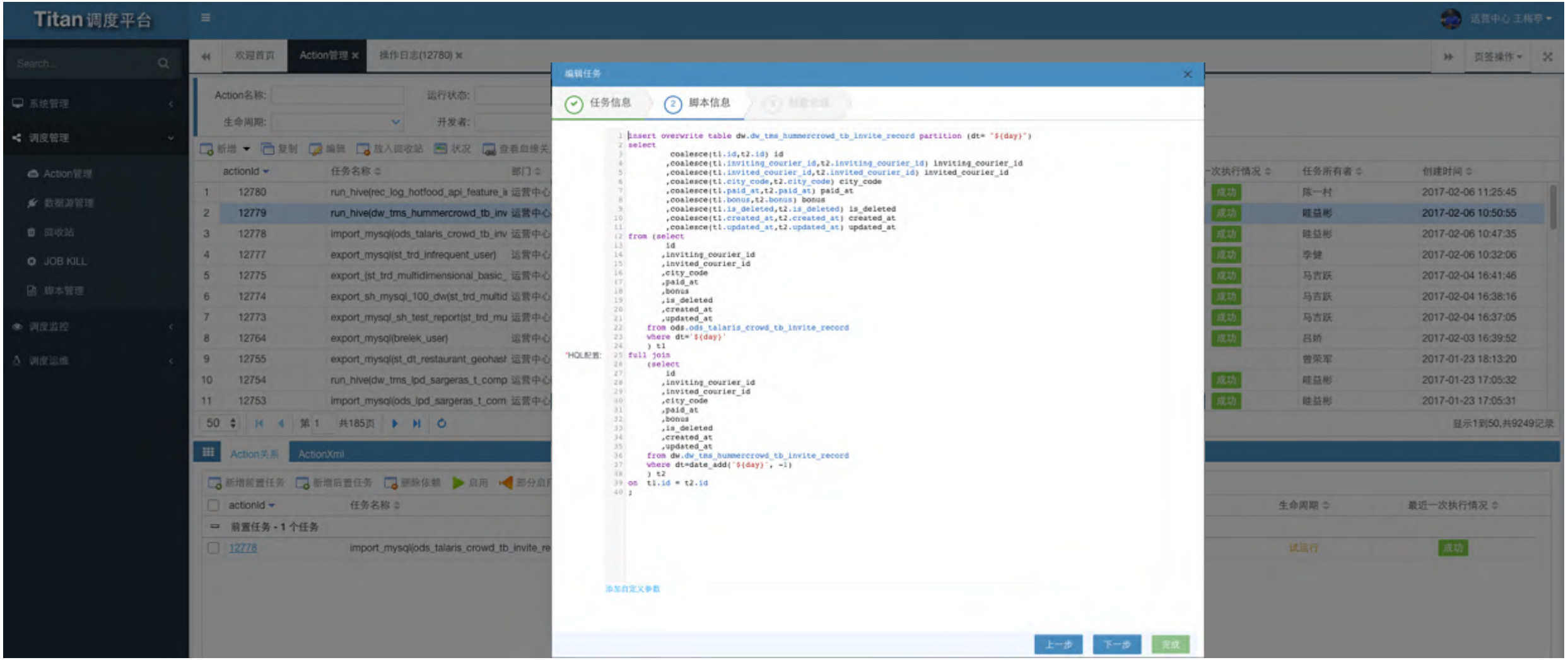
- ➢ 动态数据

  热度

  容量

字段名称	数据类型	字段类型	敏感等级	加密方式	注释	码值	操作
order_id	bigint	未知	非敏感	无	订单id		修改
order_created_t...	string	未知	非敏感	无	下单时间		修改
order_status	bigint	未知	非敏感	无	订单状态		修改
shop_id	bigint	未知	非敏感	无	商户id		修改
shop_name	string	未知	非敏感	无	商户名称		修改
city_id	bigint	未知	非敏感	无	城市id		修改
city_name	string	未知	非敏感	无	城市名称		修改
user_id	bigint	未知	非敏感	无	用户id		修改

• 数据开发管理平台-ETL



数据开发管理平台-ETL

前置检测：数据源检测
后置action：cleanup or post-validation

模板SQL
依赖识别/依赖强度
代码版本

多执行引擎
压力感知/资源调度感知

多数据存储推送

数据开发管理平台-任务SLA&分析

- 链路分析

  关键节点(出度)分析

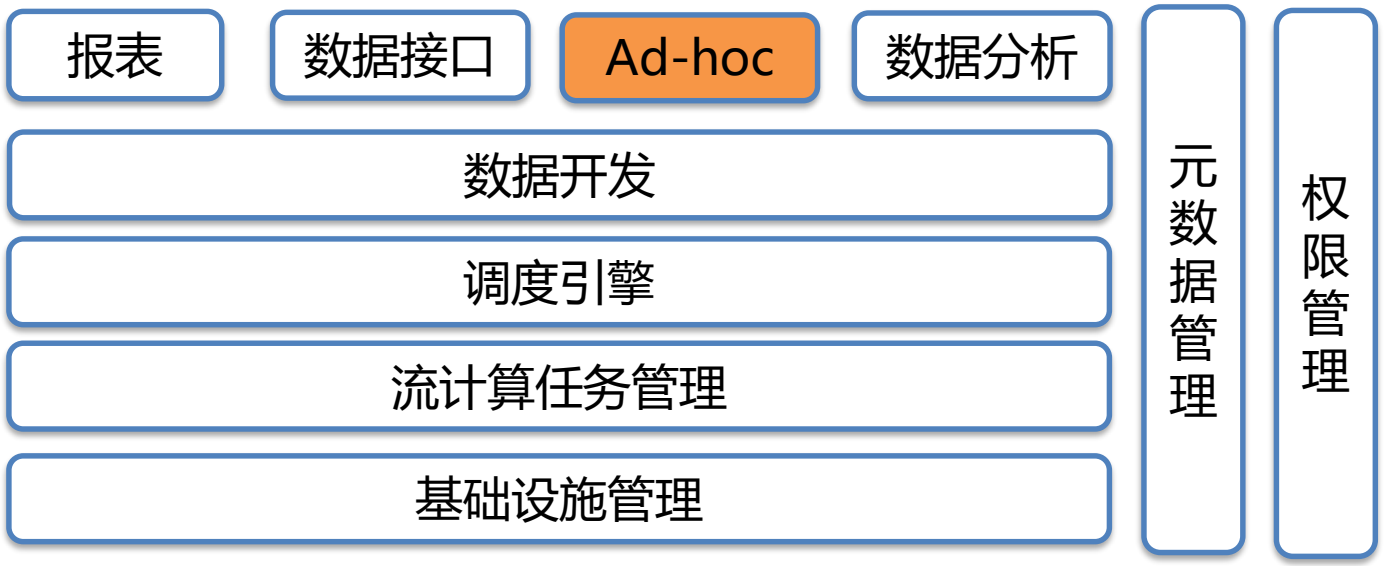
  运行趋势(运行时间/完成时间/数据量)分析

  运行日志

  step tagging：倾斜/小文件/资源/参数

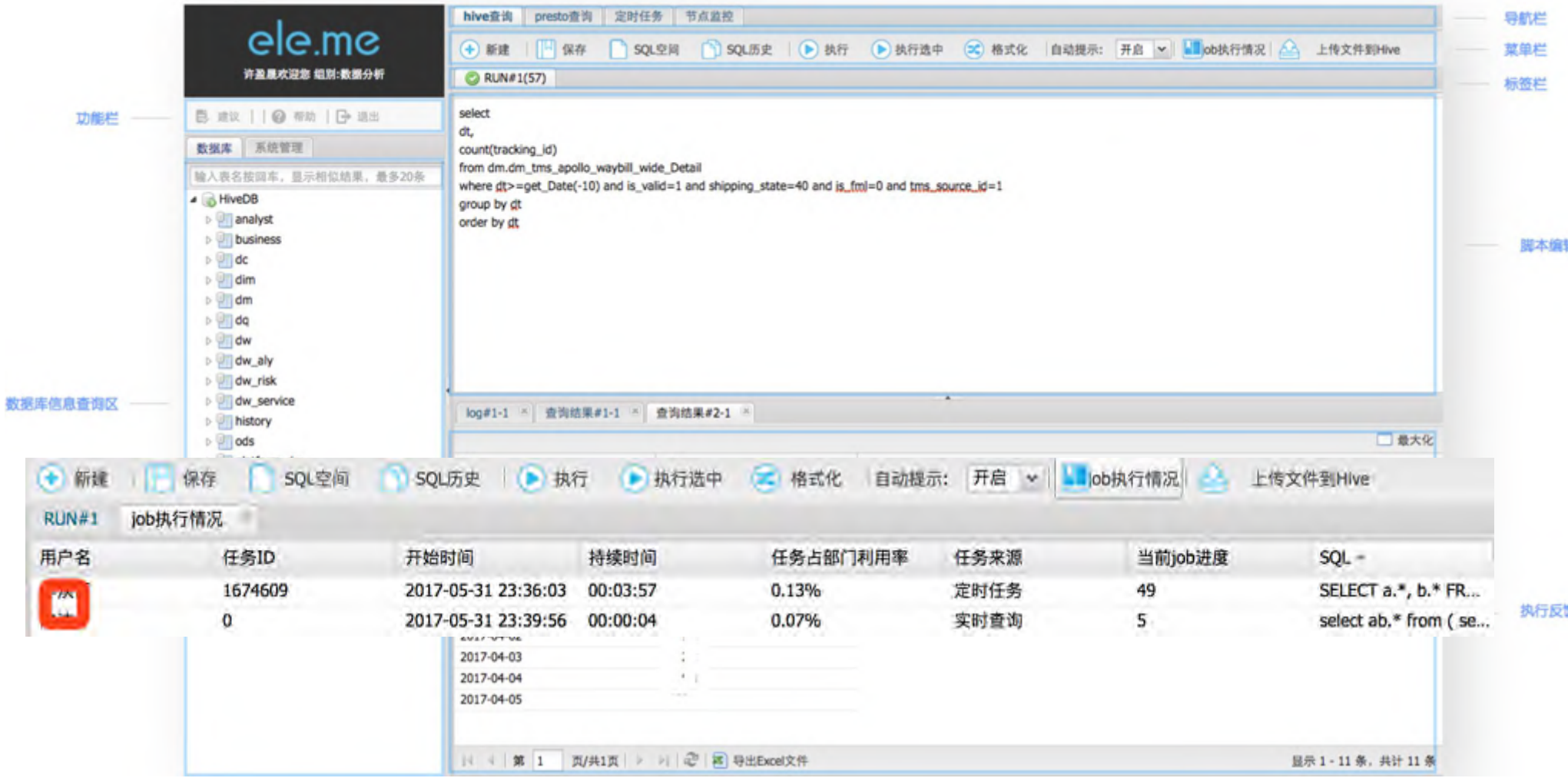
  任意下钻

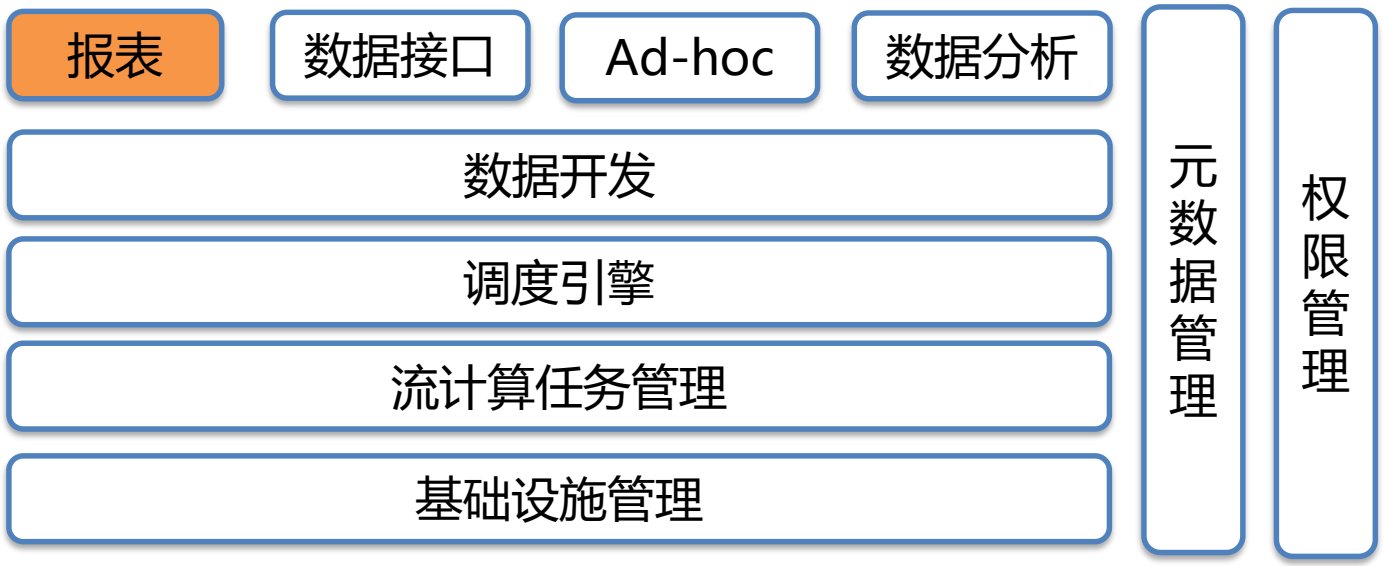
- 报警


报表
数据接口
Ad-hoc
数据分析
数据开发
调度引擎
流计算任务管理
基础设施管理
元数据管理
权限管理

• 数据应用-即席查询

- ➢ 功能性
  - 权限
  - 即席/定时
  - 历史记录
  - 运行状态
- ➢ 非功能性
  - 资源隔离
  - 前置参数



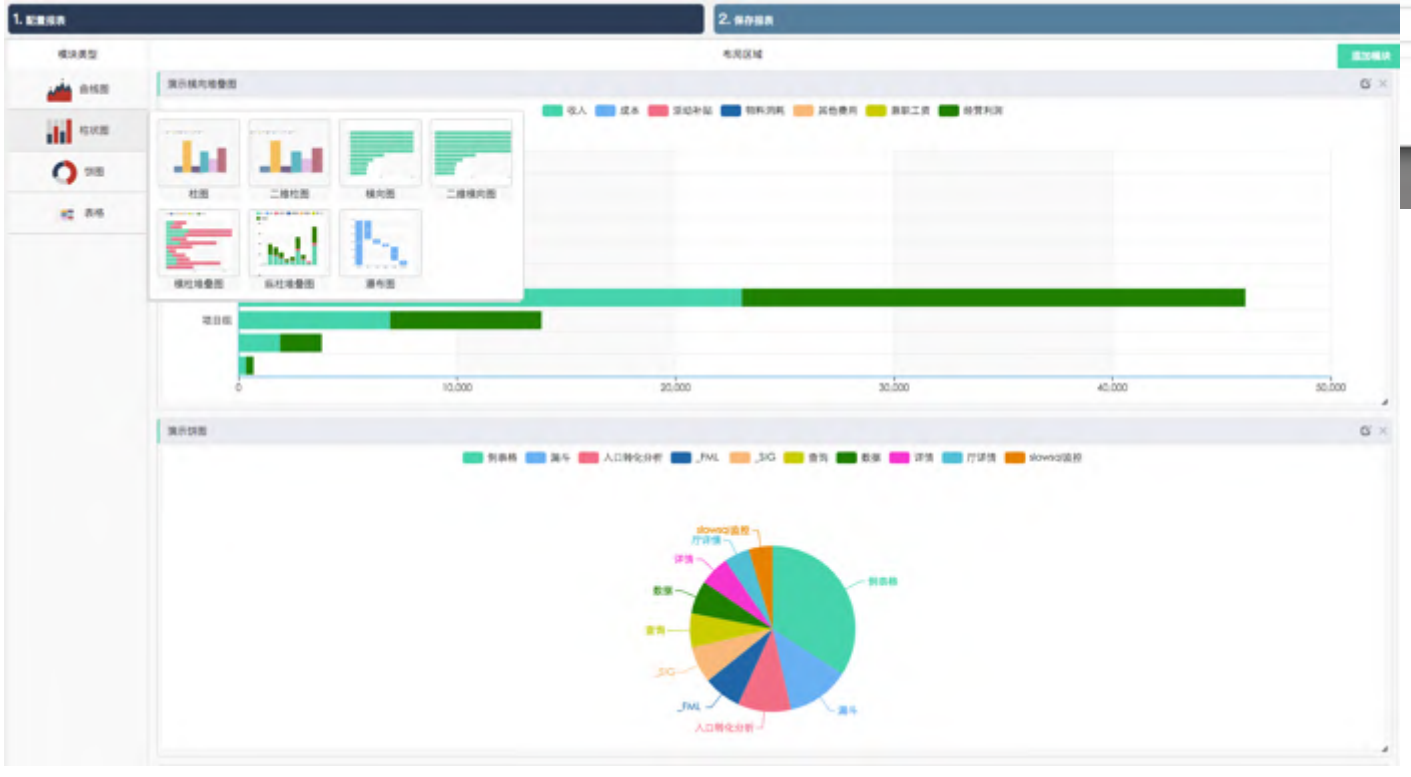

报表
数据接口
Ad-hoc
数据分析
数据开发
调度引擎
流计算任务管理
基础设施管理
元数据管理
权限管理

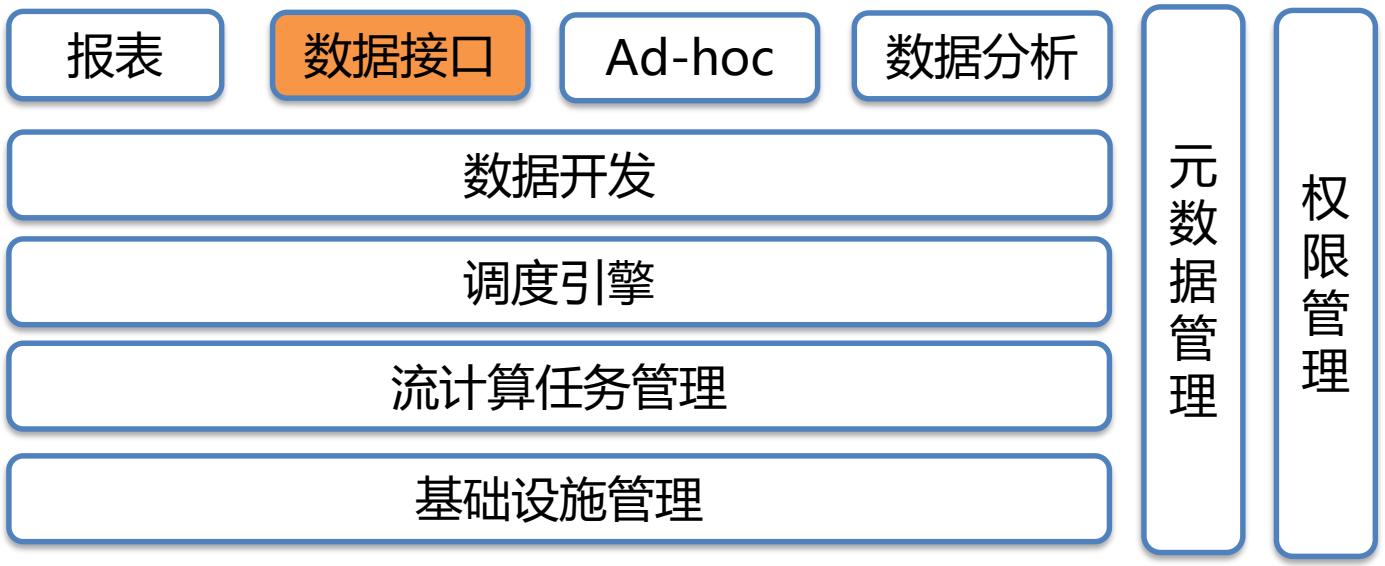
• 数据应用-报表开发

SQL as Report

Drag&Drop

多屏

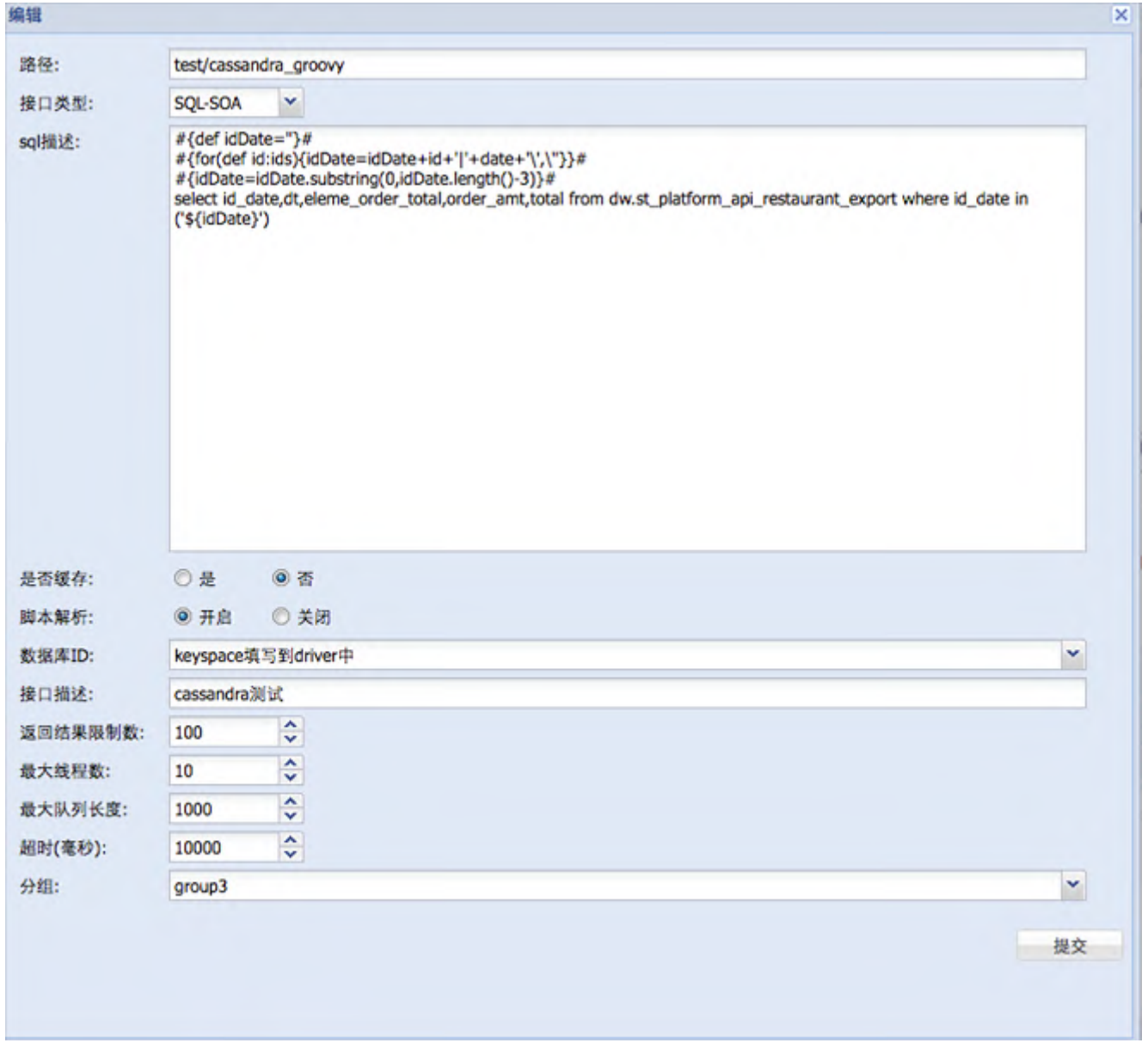


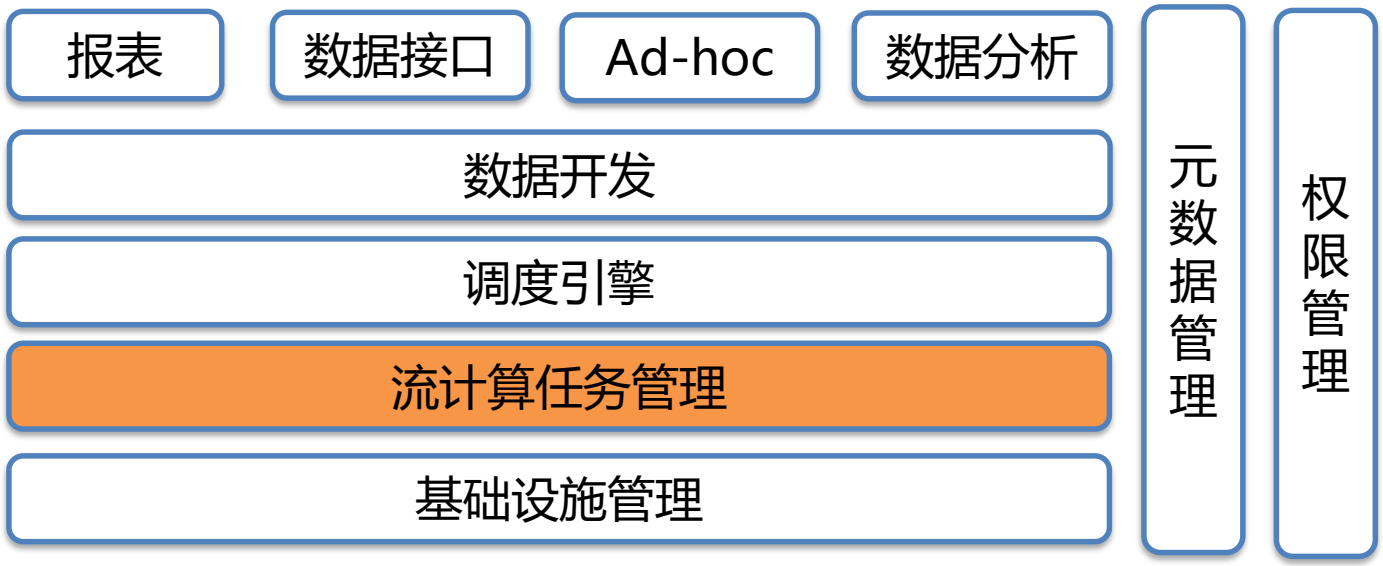

报表
数据接口
Ad-hoc
数据分析
数据开发
调度引擎
流计算任务管理
基础设施管理
元数据管理
权限管理

数据应用-接口开发

SQL as Service

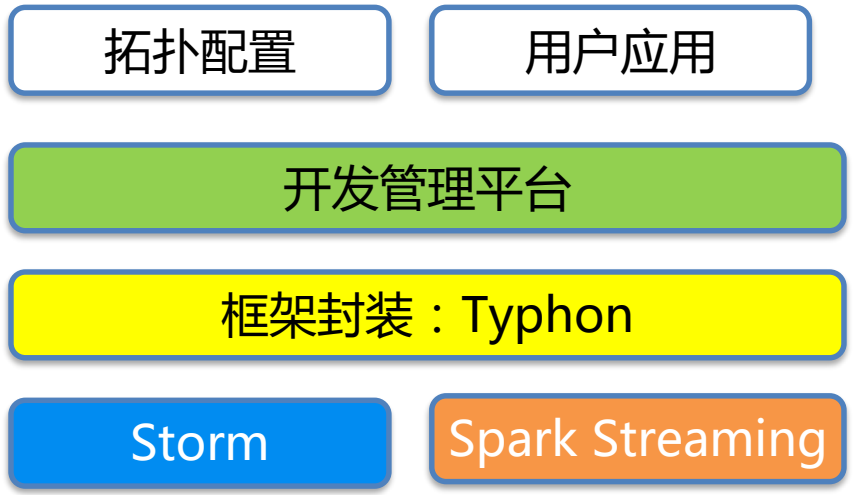
隔离
缓存
条件判断

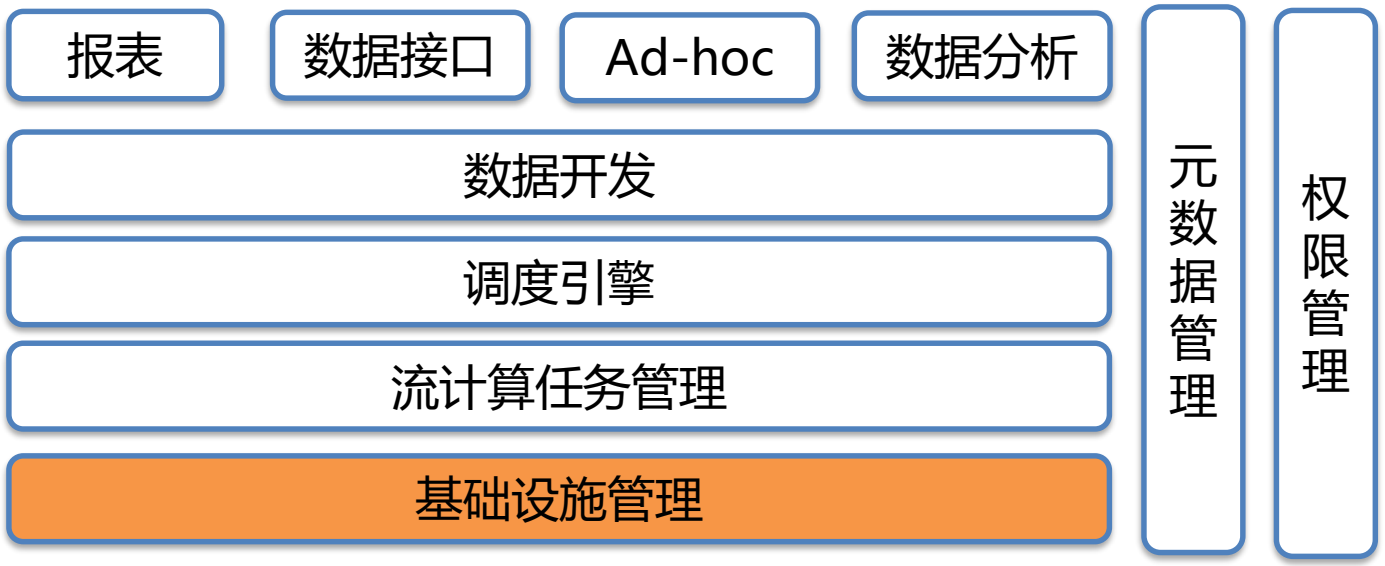



报表
数据接口
Ad-hoc
数据分析
数据开发
调度引擎
流计算任务管理
基础设施管理
元数据管理
权限管理

实时开发管理平台

- 场景
  各种Dashboard
  实时特征计算
  POI感知
- 平台
  数据源接入 & topic申请
  封装框架细节
  可配置拓扑
  任务管理&监控
  多引擎支持




报表
数据接口
Ad-hoc
数据分析
数据开发
调度引擎
流计算任务管理
基础设施管理
元数据管理
权限管理

• 基础设施管理平台

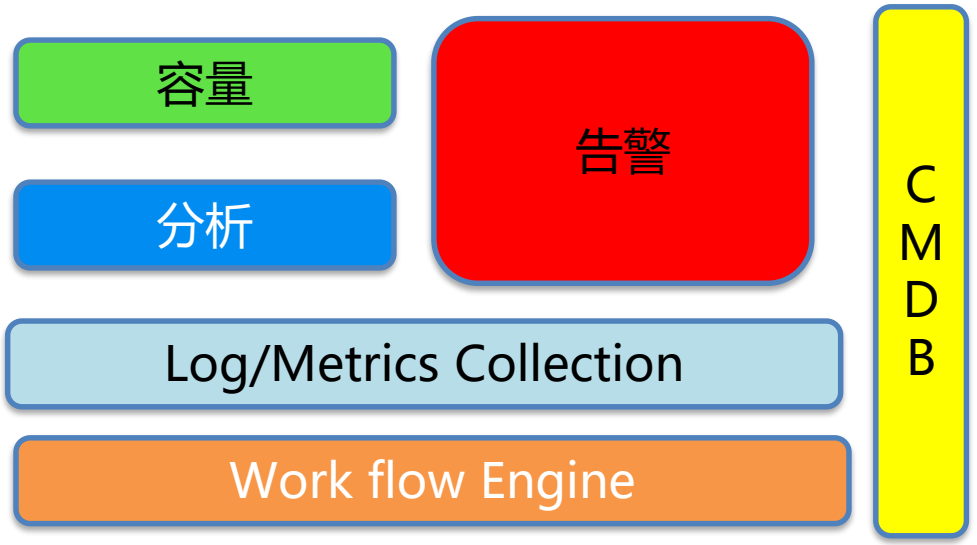
cmdb：自动化的基石

ops：重复的操作构建到workflow

alert：告警管理和分析

capacity：访问热度，空间增量/速，计算增量

analysis：统一的任务/链路性能分析



平台的一些想法

- 沟通和协调是最大的成本
  Do not take things personally
  1.自助化 / 自解释 or document
  2.SOP / 一键 or 自动 / 联动

- What gets measured gets fixed

- 设计
  1.Less is more
  2.Think about future， design with flexibility，but only implement for production

- 选型：最合适的，而不是最先进的

- 面向业务效率：产品不足服务凑

- 预期管理：Do not assume anything

THANKS !