



饿了么大数据平台建设

主讲人：毕洪宇

时间：2017.08.19

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



美好生活触手可得

2015年5月
团队成立

2年





美好生活触手可得

- 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**

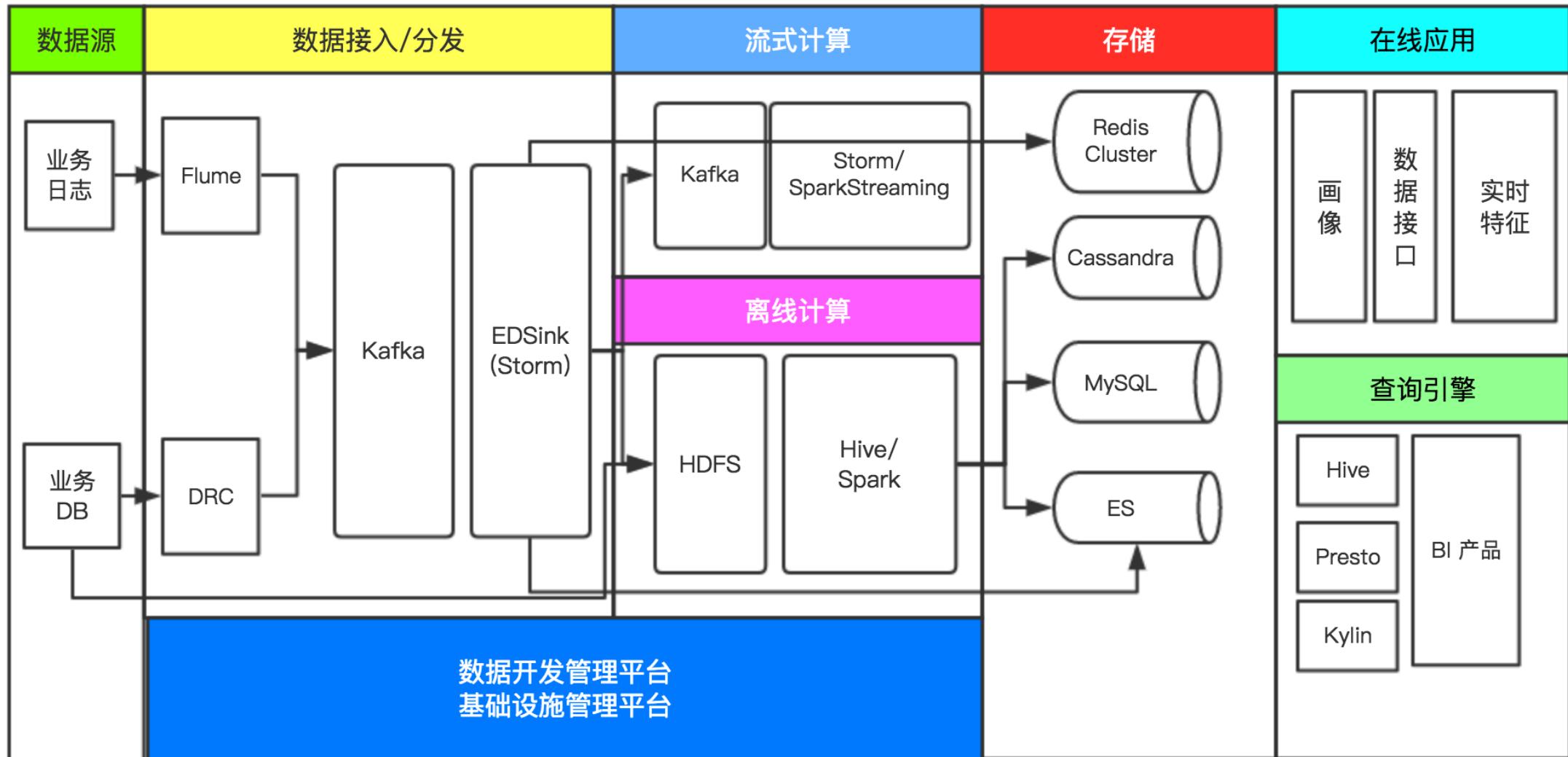


美好生活触手可得

- 实时计算

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

- 逻辑架构与数据流向



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



美好生活触手可得

效率

质量

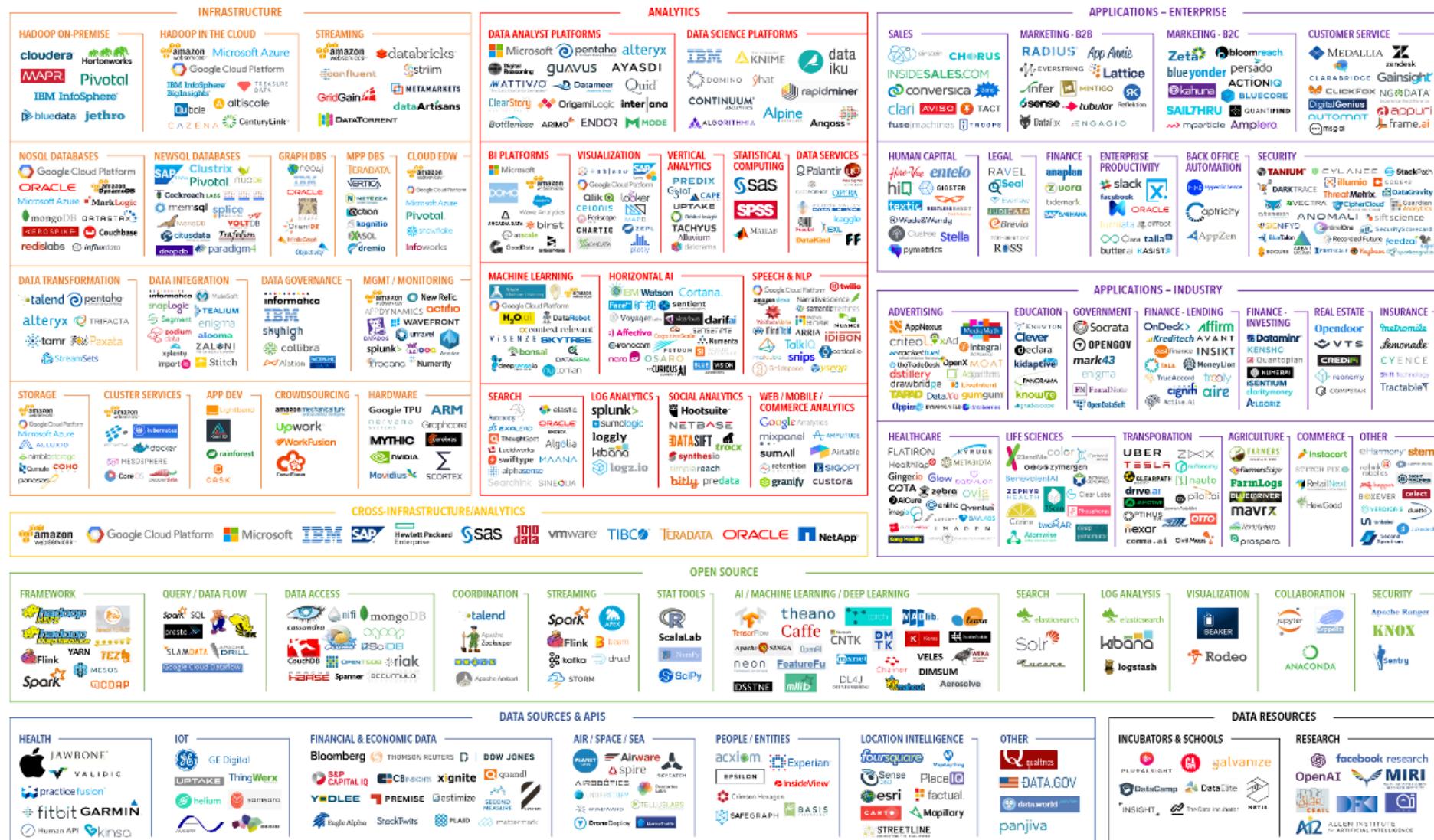
持续扩展

技术选型



美好生活触手可得

BIG DATA LANDSCAPE 2017

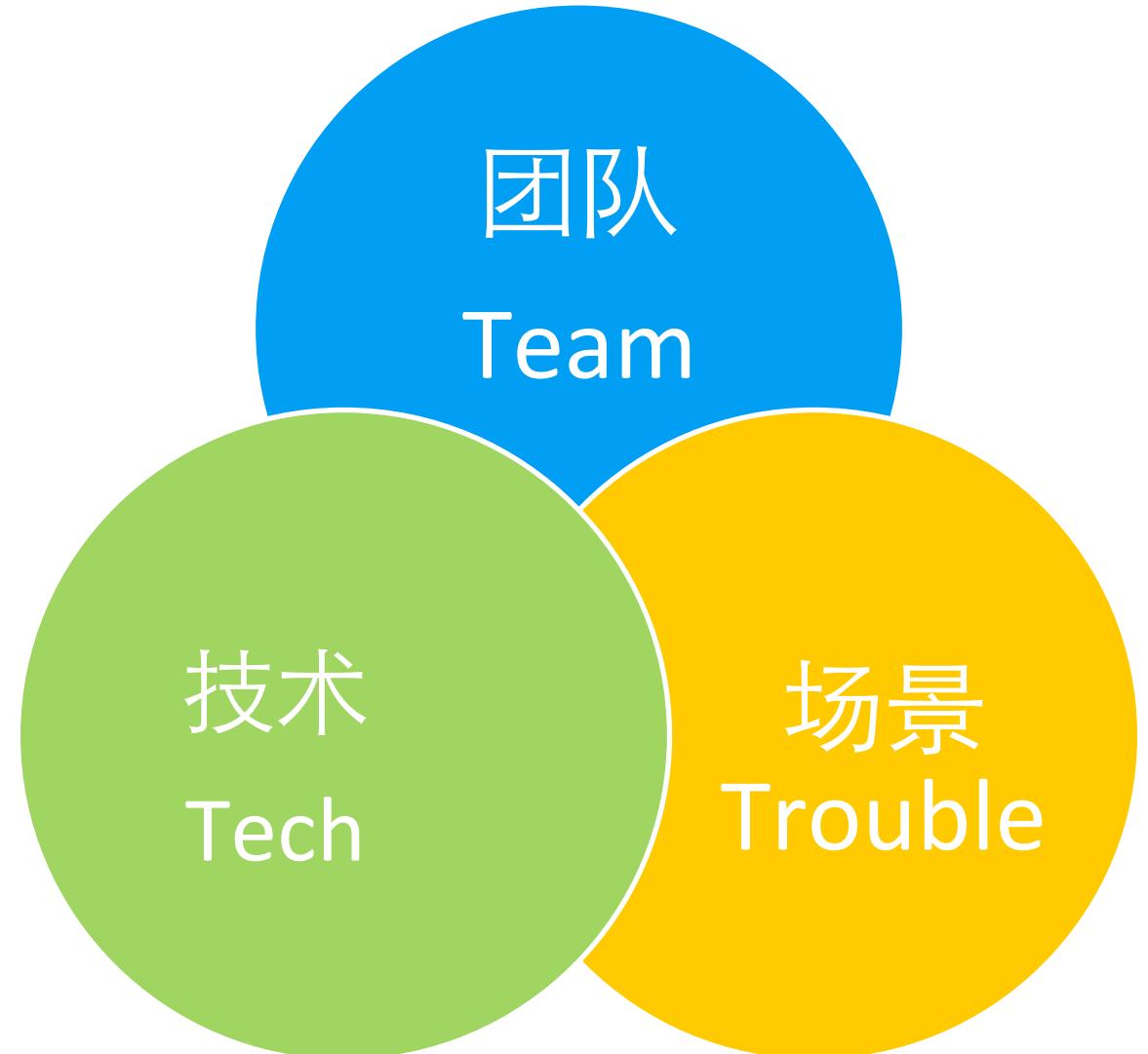


3T

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





美好生活触手可得

- 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



美好生活触手可得

- 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

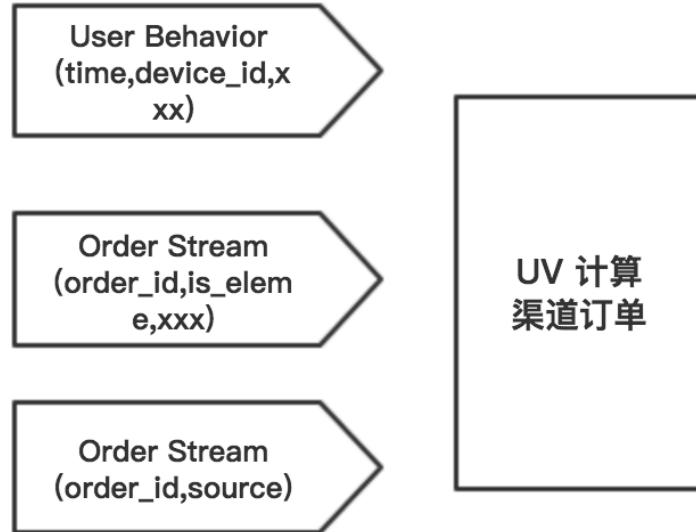
关于决策

喜欢是放肆

爱是克制

架构设计

- 可扩展-同步&一致



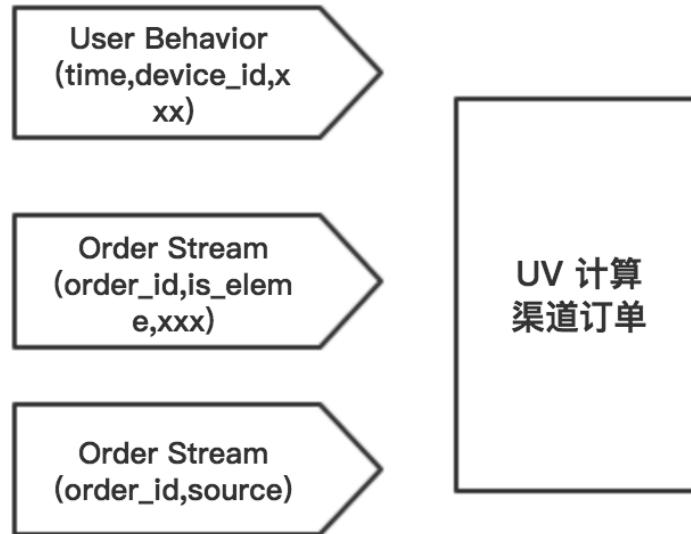
如何多流合并计算：

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

User Behavior
(time,device_id,x
xx)

Order Stream
(order_id,is_ele
ment,xxx)

Order Stream
(order_id,source)



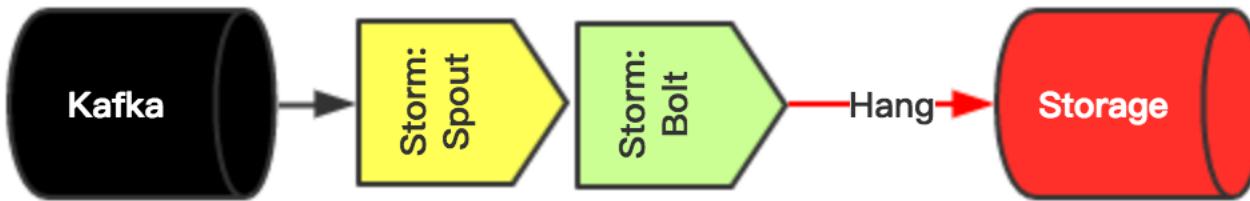
如何进行多维度UV计算：

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



美好生活触手可得

- 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

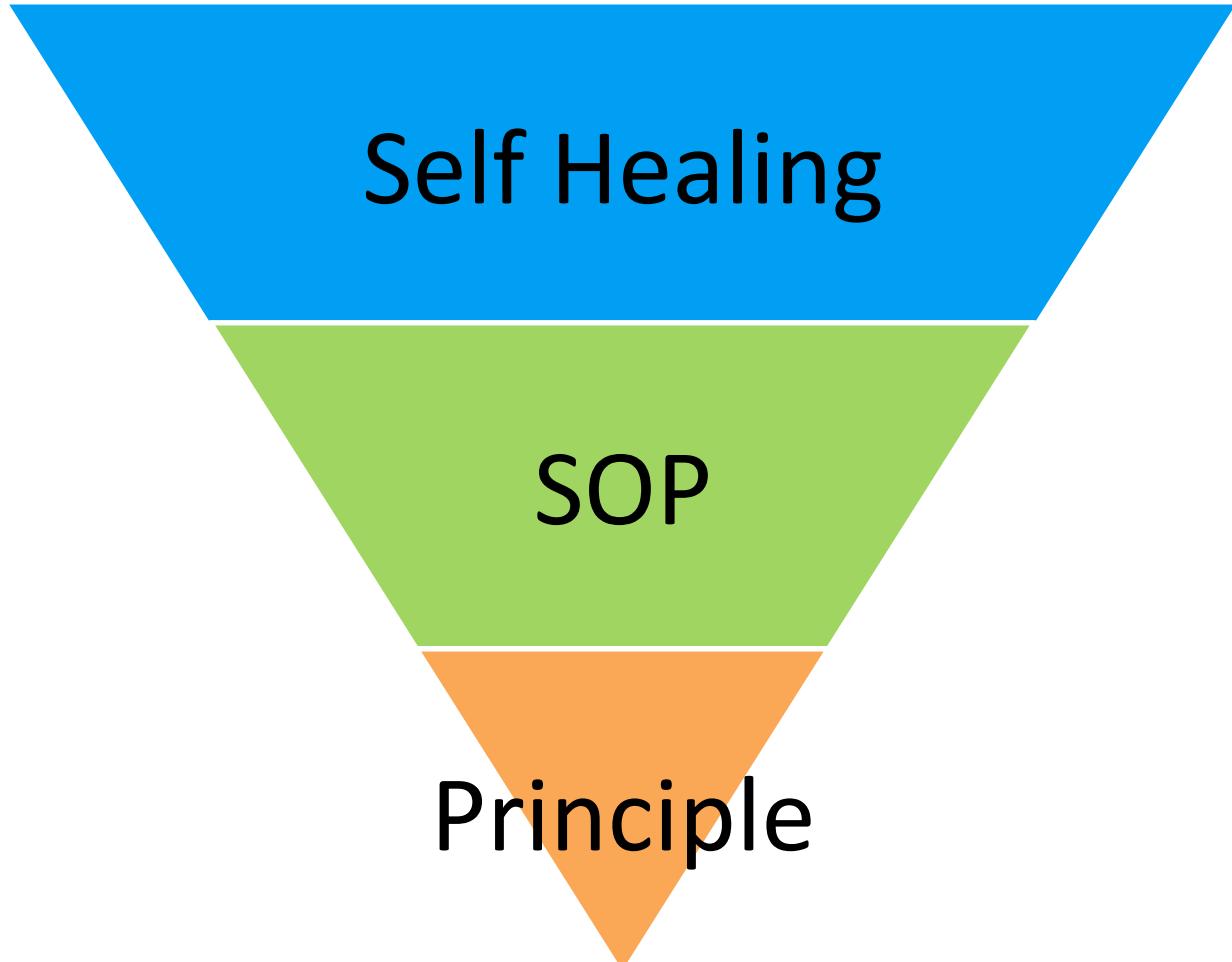
时间估算很重要

MTTR=告警响应响应+介入时间+处理时间

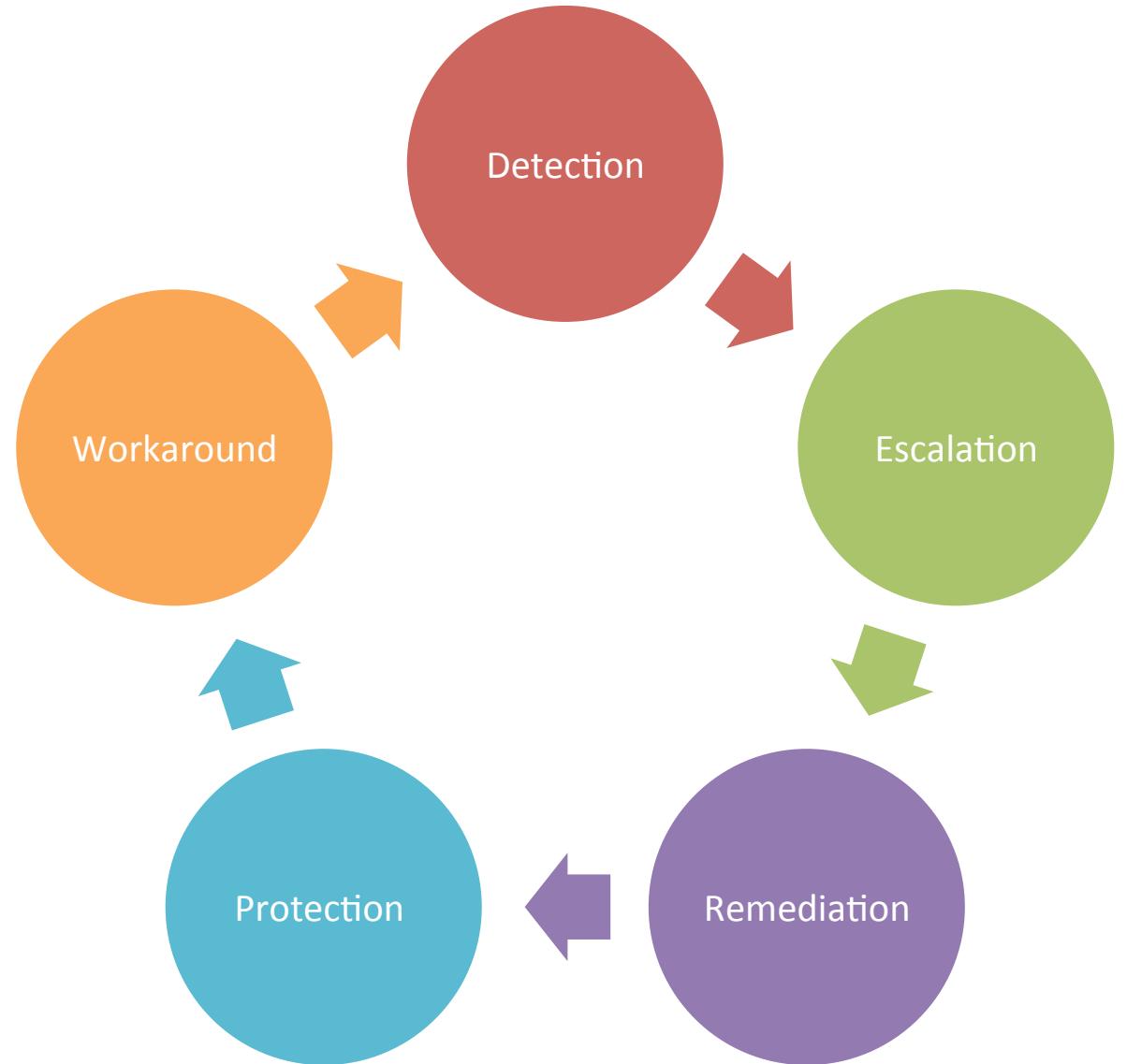
监控 ≠ 告警

监控 = metrics+trigger+action

如何控?



- 复盘原则 DERP(W)



工具链



美好生活触手可得

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



美好生活触手可得

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

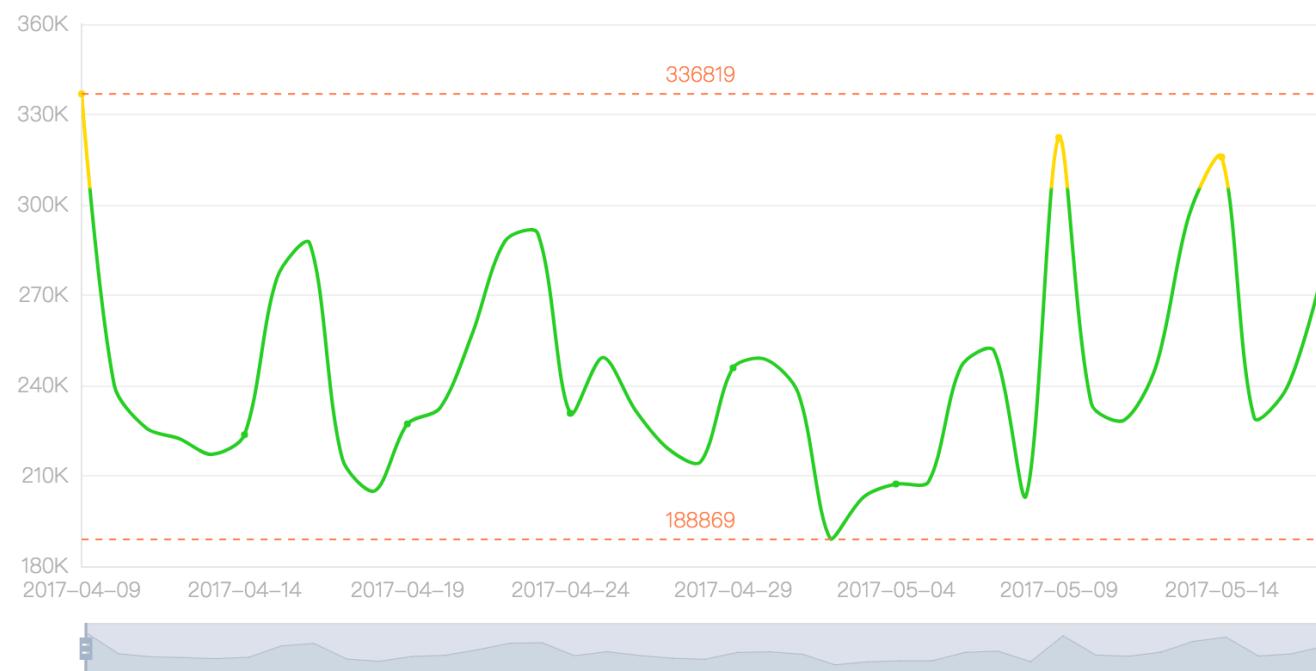
权限管理

- 数据治理工具

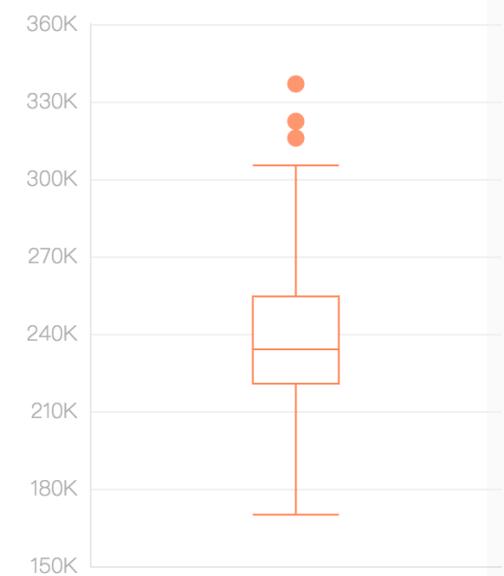
➤ **数据地图**
指标
报表
表
血缘

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	188,869	(180000, 300000) 正常	+14.3%	(-10%, +10%) +4.3%	+16.83%	(-5%, +5%) +11.83%

历史趋势

箱线图: 

➤ **数据质量**
数据源
策略
触发





美好生活触手可得

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

dm_trd_order_activity_sum

SQL 导出 ▾

一级主题: 交易

二级主题:

三级主题:

表热度: 4206

所有者: yi.xiao02

创建时间: 2017-05-25 19:13:14

表容量: 14.22G 「 9.97G 234.34% - 」 ↗

分区字段: dt string

生命周期: 三个月

字段信息

字段名/注释

+ 添加字段

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

Titan 调度平台

欢迎首页 Action管理 操作日志(12780)

Action名称: 生命周期: 运行状态: 开发者:

新增 复制 编辑 放入回收站 状态 查看血缘关系

actionId	任务名称	部门
12780	run_hive(rec_log_hotfood_api_feature_hi	运营中心
12779	run_hive(dw_tms_humercrowd_tb_inv	运营中心
12778	import_mysql(ods_talaris_crowd_tb_inv	运营中心
12777	export_mysql(st_trd_infrequent_user)	运营中心
12775	export_st_trd_multidimensional_basic_	运营中心
12774	export_sh_mysql_100_dw/st_trd_multid	运营中心
12773	export_mysql_sh_test_report(st_trd_mu	运营中心
12764	export_mysql(brelek_user)	运营中心
12755	export_mysql(st_dt_restaurant_geohash	运营中心
12754	run_hive(dw_tms_lpd_sargeras_t_comp	运营中心
12753	import_mysql(ods_lpd_sargeras_t_com	运营中心

50 | 第 1 共185页

Action关系 ActionXml

新增前置任务 新增后置任务 删除依赖 启用 部分启用

actionId 任务名称

前置任务 - 1 个任务

12778 import_mysql(ods_talaris_crowd_tb_invite_re

编辑任务

① 任务信息 ② 脚本信息 ③ 创建完成

```
1 insert overwrite table dw.dw_tms_humercrowd_tb_invite_record partition (dt= "${day}")
2 select
3     coalesce(t1.id,t2.id) id
4     ,coalesce(t1.inviting_courier_id,t2.inviting_courier_id) inviting_courier_id
5     ,coalesce(t1.invited_courier_id,t2.invited_courier_id) invited_courier_id
6     ,coalesce(t1.city_code,t2.city_code) city_code
7     ,coalesce(t1.paid_at,t2.paid_at) paid_at
8     ,coalesce(t1.bonus,t2.bonus) bonus
9     ,coalesce(t1.is_deleted,t2.is_deleted) is_deleted
10    ,coalesce(t1.created_at,t2.created_at) created_at
11    ,coalesce(t1.updated_at,t2.updated_at) updated_at
12 from (select
13     id
14     ,inviting_courier_id
15     ,invited_courier_id
16     ,city_code
17     ,paid_at
18     ,bonus
19     ,is_deleted
20     ,created_at
21     ,updated_at
22     from ods.ods_talaris_crowd_tb_invite_record
23     where dt= "${day}"
24     ) t1
25 full join
26 (select
27     id
28     ,inviting_courier_id
29     ,invited_courier_id
30     ,city_code
31     ,paid_at
32     ,bonus
33     ,is_deleted
34     ,created_at
35     ,updated_at
36     from dw.dw_tms_humercrowd_tb_invite_record
37     where dt=date_add('${day}', -1)
38     ) t2
39 on t1.id = t2.id
40;
```

添加自定义参数

上一步 下一步 完成

次执行情况	任务所有者	创建时间
成功	陈一村	2017-02-06 11:25:45
成功	眭益彬	2017-02-06 10:50:55
成功	眭益彬	2017-02-06 10:47:35
成功	李健	2017-02-06 10:32:06
成功	马吉跃	2017-02-04 16:41:46
成功	马吉跃	2017-02-04 16:38:16
成功	马吉跃	2017-02-04 16:37:05
成功	吕婧	2017-02-03 16:39:52
	曾荣军	2017-01-23 18:13:20
成功	眭益彬	2017-01-23 17:05:32
成功	眭益彬	2017-01-23 17:05:31

显示1到50,共9249记录

生命周期	最近一次执行情况
试运行	成功



美好生活触手可得

- **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



美好生活触手可得

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

The screenshot displays two windows of a data application interface. The top window is a query editor titled 'hive查询' (Hive Query). It shows a SQL query in the main pane:

```
select
dt,
count(tracking_id)
from dm.dim_apollo_waybill_wide_Detail
where dt>=get_Date(-10) and is_valid=1 and shipping_state=40 and js_fml=0 and tms_source_id=1
group by dt
order by dt
```

The bottom window is a table titled 'job执行情况' (Job Execution Status) under 'RUN#1'. It lists two jobs with their details:

用户名	任务ID	开始时间	持续时间	任务占部门利用率	任务来源	当前job进度	SQL
17六	1674609	2017-05-31 23:36:03	00:03:57	0.13%	定时任务	49	SELECT a.*, b.* FR...
	0	2017-05-31 23:39:56	00:00:04	0.07%	实时查询	5	select ab.* from (se...

Annotations on the right side of the interface identify various components: '功能栏' (Function Bar), '菜单栏' (Menu Bar), '标签栏' (Tab Bar), and '脚本编辑器' (Script Editor).



美好生活触手可得

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

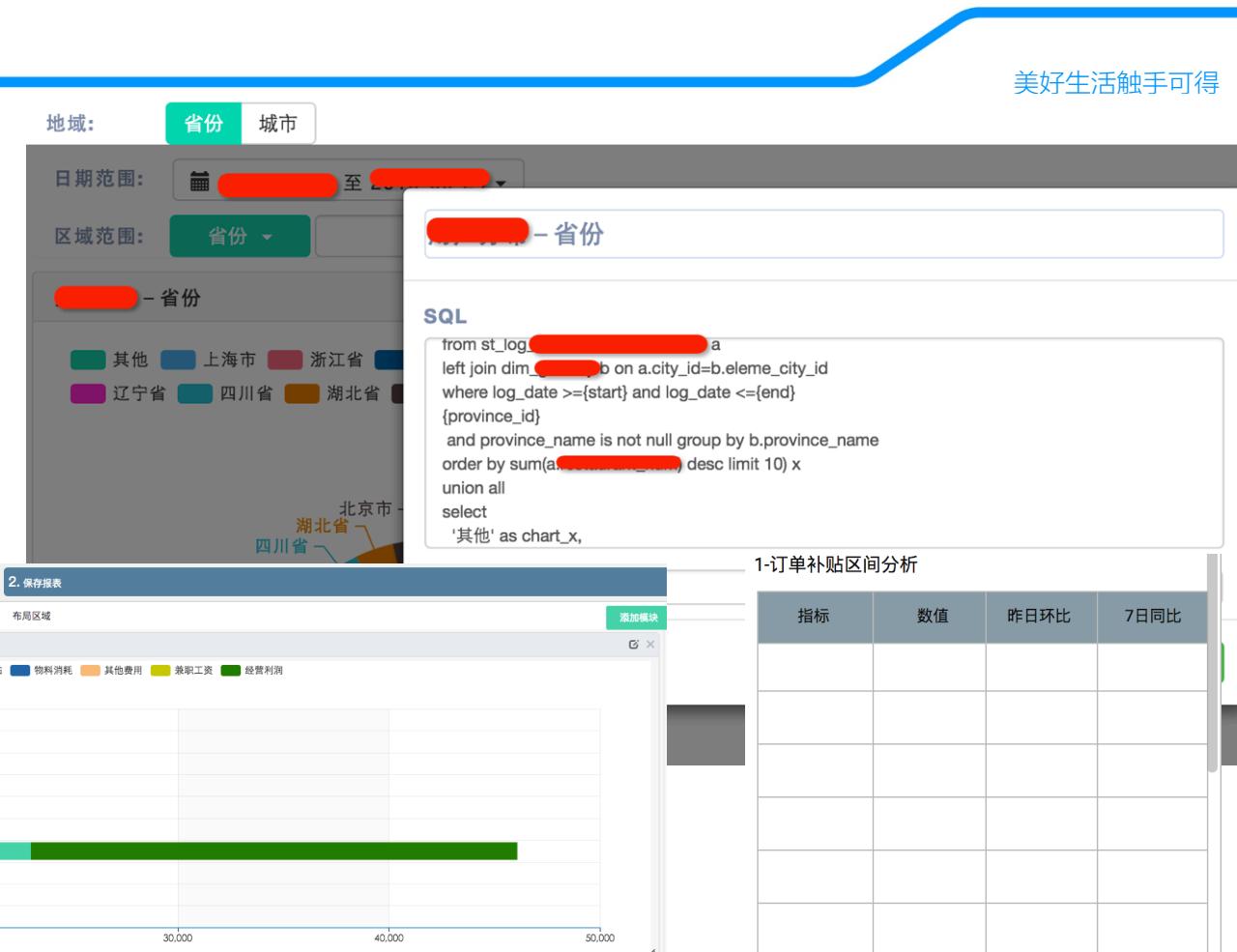
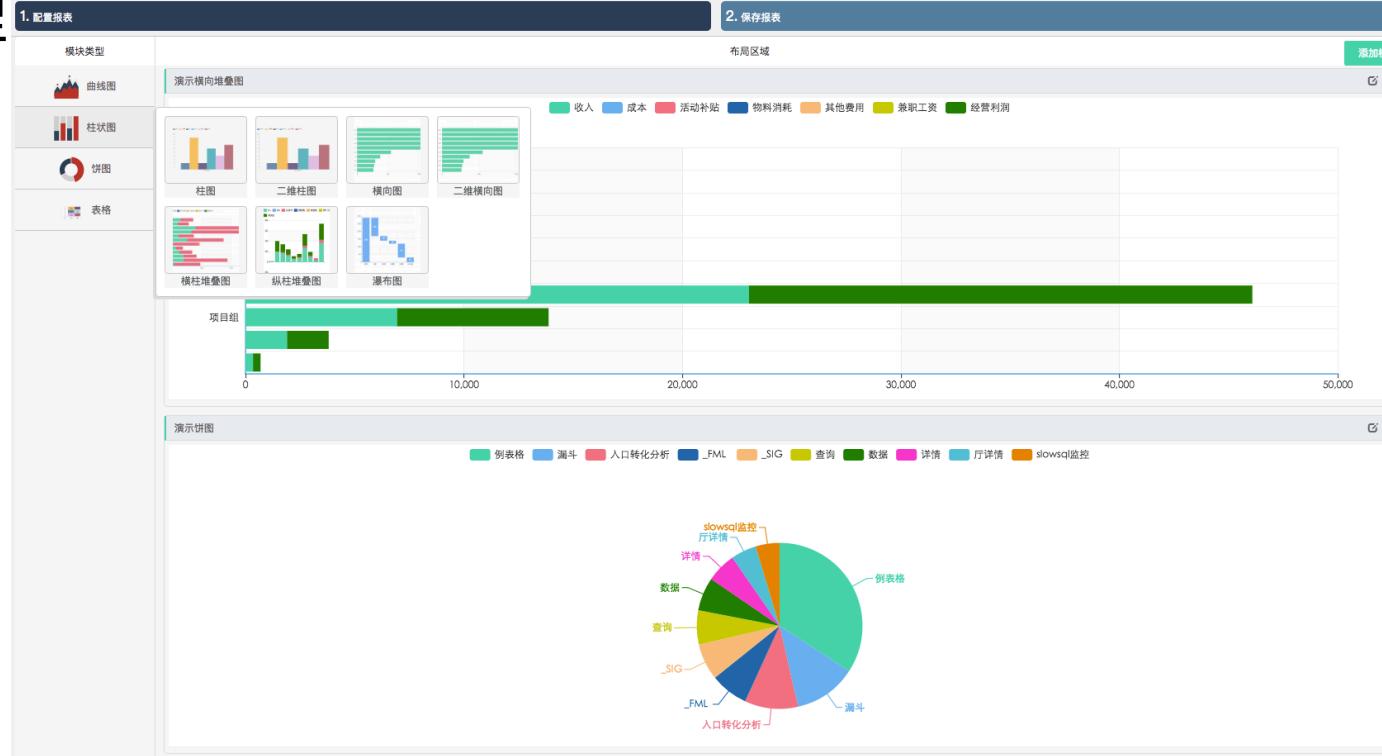
权限管理

- 数据应用-报表开发

SQL as Report

Drag&Drop

多屏





美好生活触手可得

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

编辑

路径: test/cassandra_groovy

接口类型: SQL-SOA

sql描述:

```
#{def idDate=""}
#{for(def id:ids){idDate=idDate+id+'|'+date+'\','\'}}
#{idDate=idDate.substring(0,idDate.length()-3)}
select id_date,dt,eleme_order_total,order_amt,total from dw.st_platform_api_restaurant_export where id_date in ('${idDate}')
```

是否缓存: 是 否

脚本解析: 开启 关闭

数据库ID: keyspace填写到driver中

接口描述: cassandra测试

返回结果限制数: 100

最大线程数: 10

最大队列长度: 1000

超时(毫秒): 10000

分组: group3

提交



美好生活触手可得

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



拓扑配置

用户应用

开发管理平台

框架封装 : Typhon

Storm

Spark Streaming



美好生活触手可得

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

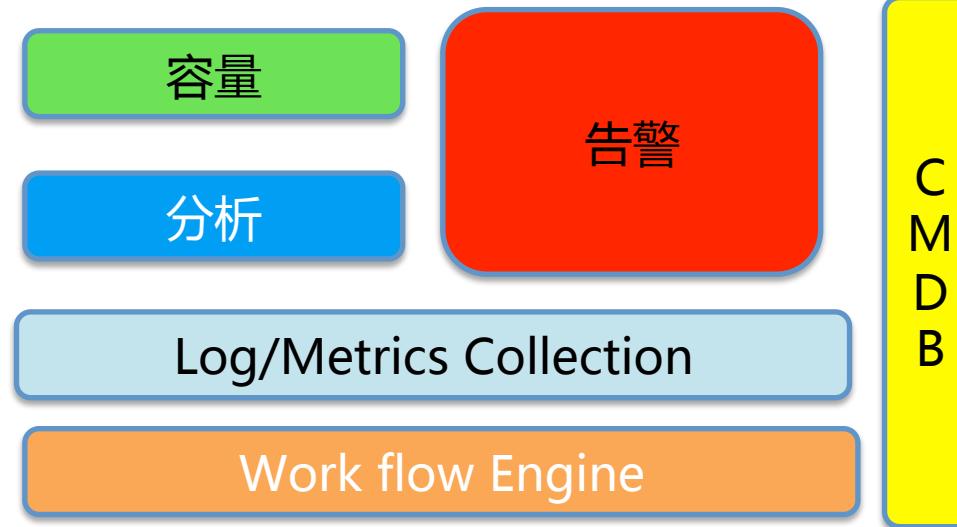
cmdb：自动化的基石

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，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

欢迎关注饿了么技术社区
We are hiring



THANKS !