GIAC | 全球互联网架构大会
GLOBAL INTERNET ARCHITECTURE CONFERENCE

阿里智能运维平台
如何助力研发应对双11挑战

运维中台　如柏

2017.12

阿里运维历程　　基础运维平台　　应用运维平台　　AIOPS

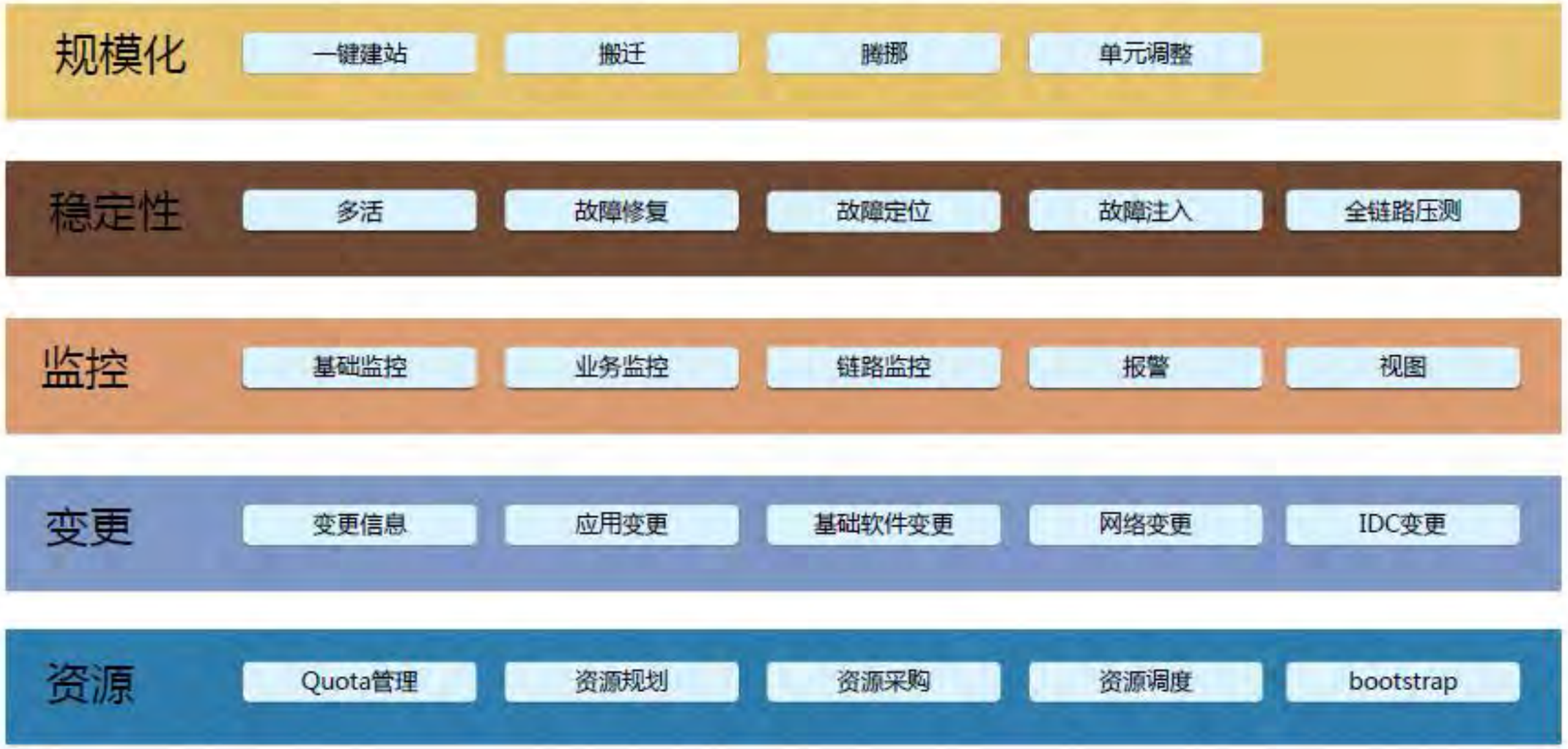
阿里运维历程

命令行工具 2007～2010
Excel，SSH，PGM

系统化工具 2010～2013
Armory，Nagios，Dragon...

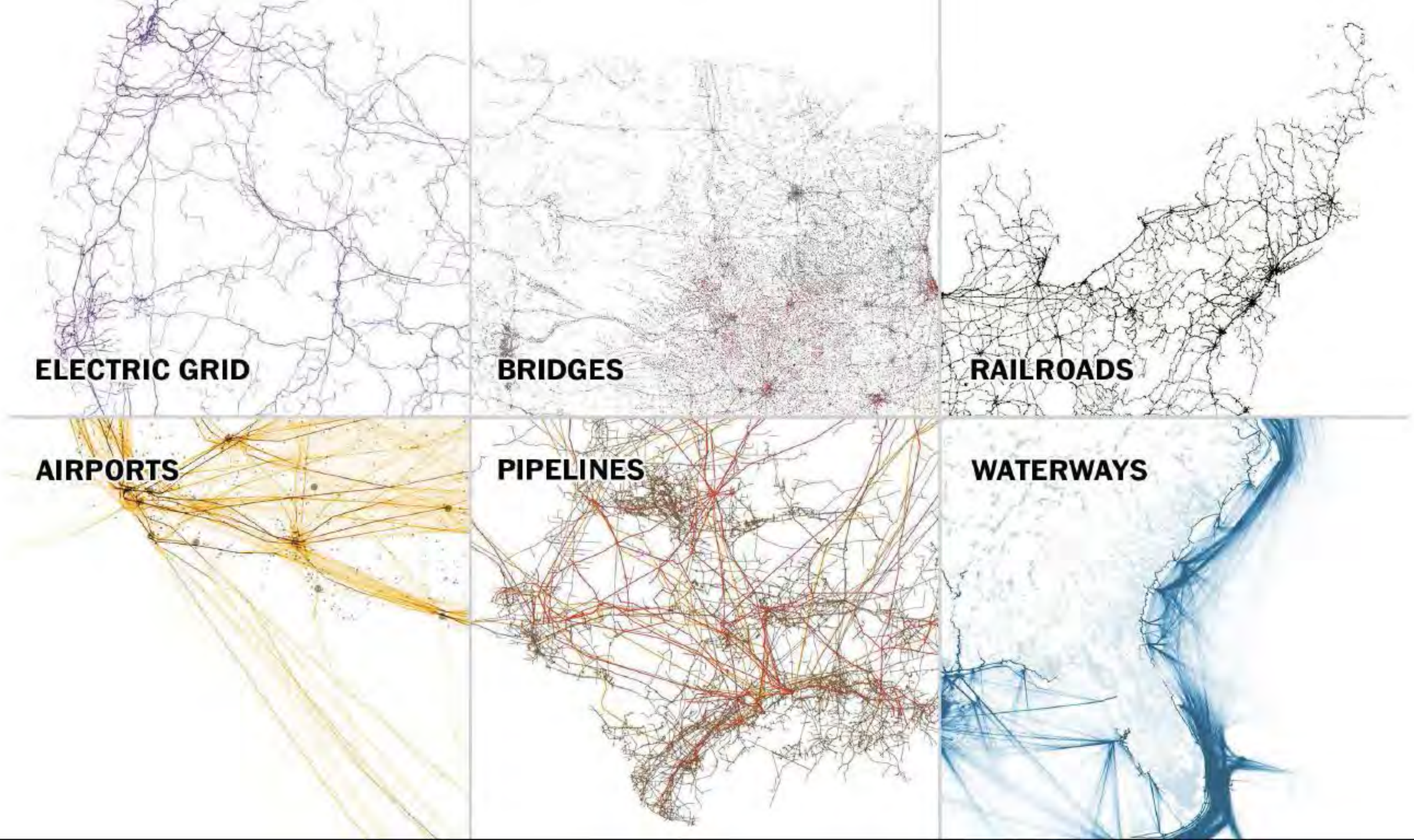
自动化平台 2013～2016
StarAgent，PSP，Alimonitor...

智能化平台 2016 ～
StarAgent，蜻蜓，Normandy，Sunfire...


规模化
一键建站
搬迁
腾挪
单元调整
稳定性
多活
故障修复
故障定位
故障注入
全链路压测
监控
基础监控
业务监控
链路监控
报警
视图
变更
变更信息
应用变更
基础软件变更
网络变更
IDC变更
资源
Quota管理
资源规划
资源采购
资源调度
bootstrap


Web / API / SDK / CLI / IDE / 掌上运维 / 大屏
流程
规范
红线
故障
管理
应用运维平台
业务A
运维平台
业务B
运维平台
业务C
运维平台
算法
平台
基础运维平台
公有云
专有云
混合云
物理机
虚拟机
Docker
网络
存储
数据库
中间件
…

基础运维平台


ELECTRIC GRID
BRIDGES
RAILROADS
AIRPORTS
PIPELINES
WATERWAYS

阿里堡垒机

账号 / 权限 / 密码 / 密钥 / Root
访问控制
操作日志
高危拦截
审计 / 行为分析
合规 / 认证
操作录屏 / 回放
SSH非法入侵审计

1.高承载，可达5000人同时在线；对管理对象没有限制
2.经历多种认证合规最佳实践考验（sox404，iso27001，DSS-PCI）

运维通道 - StarAgent

稳定性 / 性能

- 去数据库依赖
- 单元化

插件平台

- 低耦合、高可扩展性

安全

- 高危命令审计 / 拦截
- 命令映射
- 全链路加密签名

自动化运维 95%

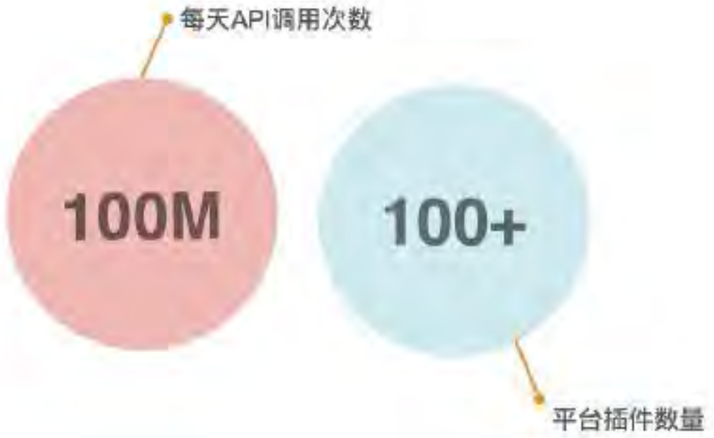
- IP段自动关联
- 自动扩缩容 / 自动负载均衡
- Agent / 插件版本自动更新
- Agent自检 / 自愈

50万

每分钟API调用

99.995%

系统稳定性



其他功能

- 基础信息采集
- Web 终端
- 分布式定时任务


服务器生命周期
应用生命周期
资产核对
OS安装
环境配置
应用发布
安全、监控、采集
日常运维

命令通道

同步命令

异步命令

查询

数据通道

核心系统指标配置

核心系统数据采集

数据通道

插件平台

静态：命令、脚本

动态：常驻进程

Portal

主机管理

主机账号

Web 终端

分布式定时任务

插件市场

文件分发

API

cmd

Plugin

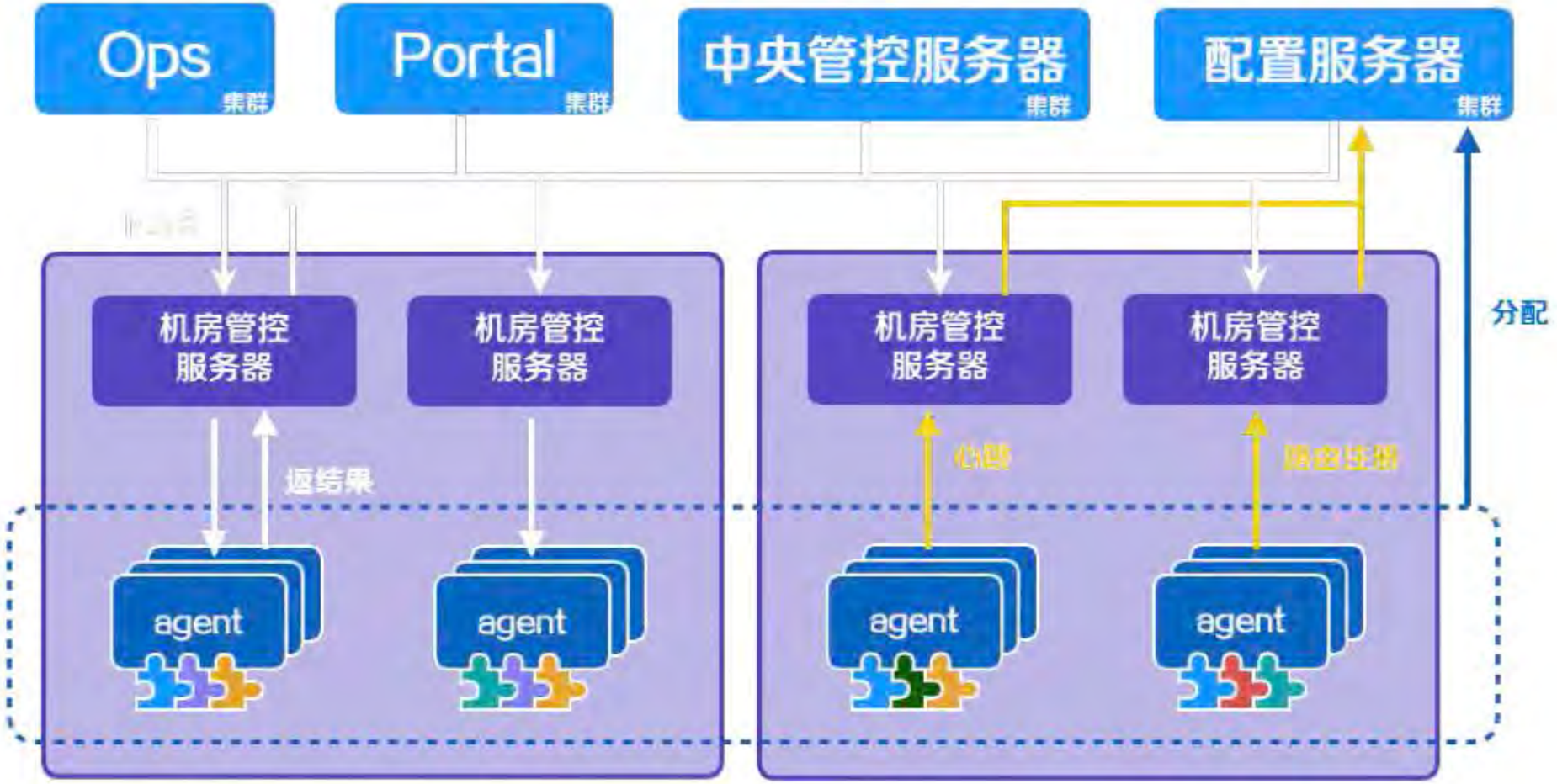
Cron

File Store

Action

SDK

移动端


Ops
集群
Portal
集群
中央管控服务器
集群
配置服务器
集群
分配
机房管控
服务器
机房管控
服务器
机房管控
服务器
机房管控
服务器
返结果
agent
agent
agent
agent

完善错误码体系

系统错误

环境错误

用户错误

第三方依赖错误

未知错误

环境监测

监控ping报警
ssh报警
磁盘报警
syslog中机器宕机记录

去依赖

定期演练

命令通道

快车、慢车分道

同步任务异步化 - DeferedResult

心跳、注册 批量化写入，只更新需要更新的

第一重保护

核心问题：API密钥泄露、API权限过大

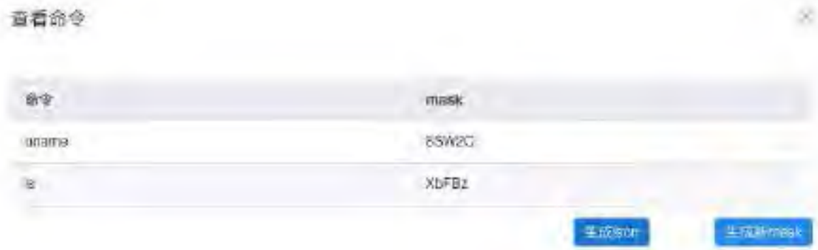
解法： API KEY生命周期管理，提供更新、重置。另外要有禁用策略

解法： 引入安全负责人、Bu稳定性负责人审批。其中降权的变更直接免批。

第二重保护

核心问题：人工命令过多且命令可被伪造

解法： 命令沙盒模块化处理，命令白名单提供主命令映射方式，以mask方式执行，参数保持不变。



第三重保护

核心问题：高危命令执行风险过大

解法：针对高危命令进行正则匹配记录拦截， 固化每个API帐号可执行高危命令的机器数量、频率，定义高危级别并实时监控预警。

第四重保护

核心问题：命令被拦截和篡改

解法：TCP Socket 全链路加密、校验

协议：启动、停止、配置、强制结束、同步执行

服务：守护、额度、部署、升级

价值：

安全：统一管理、审计

稳定性：保护服务器正常运行

成本/效率：提升agent运维效率、
避免重复建设agent、
避免重复数据采集

```
{
    "name":"plugin_name",
    "version":"1.0.1-build15",
    "schedule":"daemon",
    "daemon":{
        "exec":"[redacted]xx/bin/xxx.py",
        "user":"admin",
        "group":"admin",
        "env":{
            "PATH":"[redacted]bin:$PATH",
            "CWD":"[redacted]/${plugin.name}"
        },
        "log_dir":"${CWD}/logs"
    },
    "limit":{
        "cpu":"5%",
        "disk":"200m",
        "mem":"50m"
    }
}
```

快速复制与输出的能力

阿里云专有云

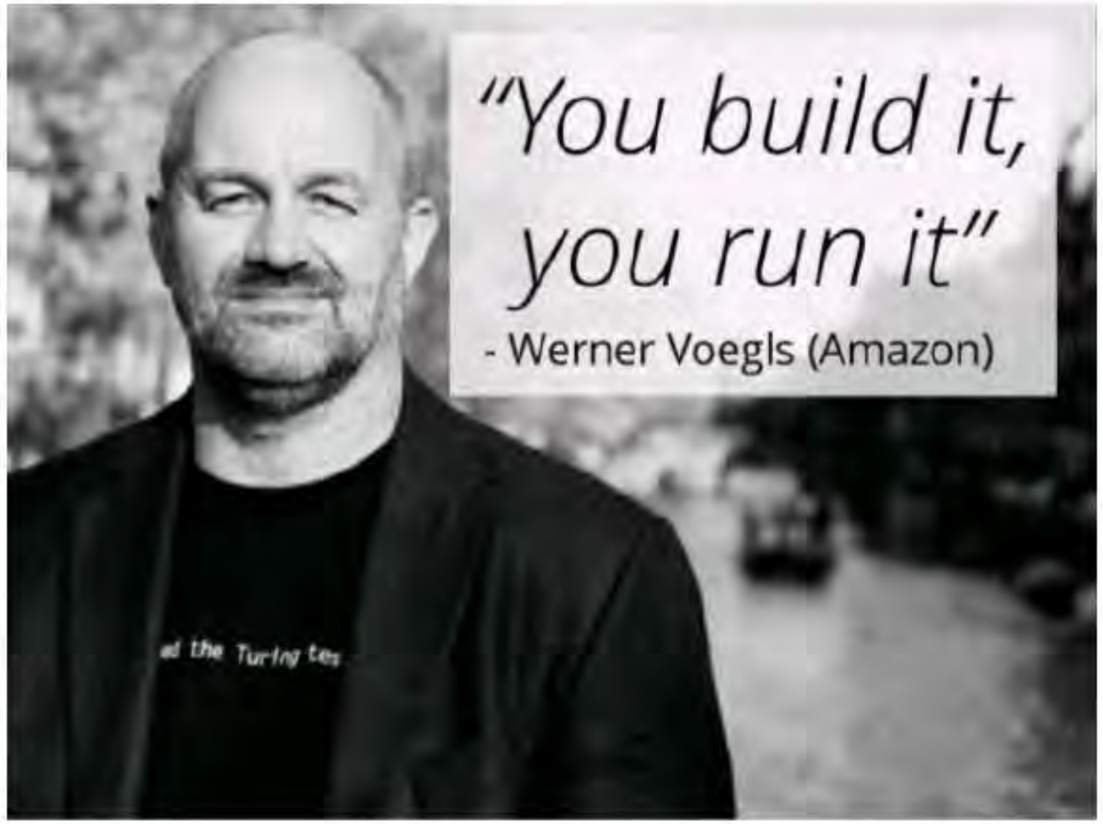
- V1/V2 40个左右的输出项目 (StarAgent 1.0)
- V3 天基 开发中...

蚂蚁金融云

- 公有云 30+ 客户；
- 专有云 / 专有域：网商银行、Alipay+、新加坡等几个 （ SSH / StarAgent 1.0 ）

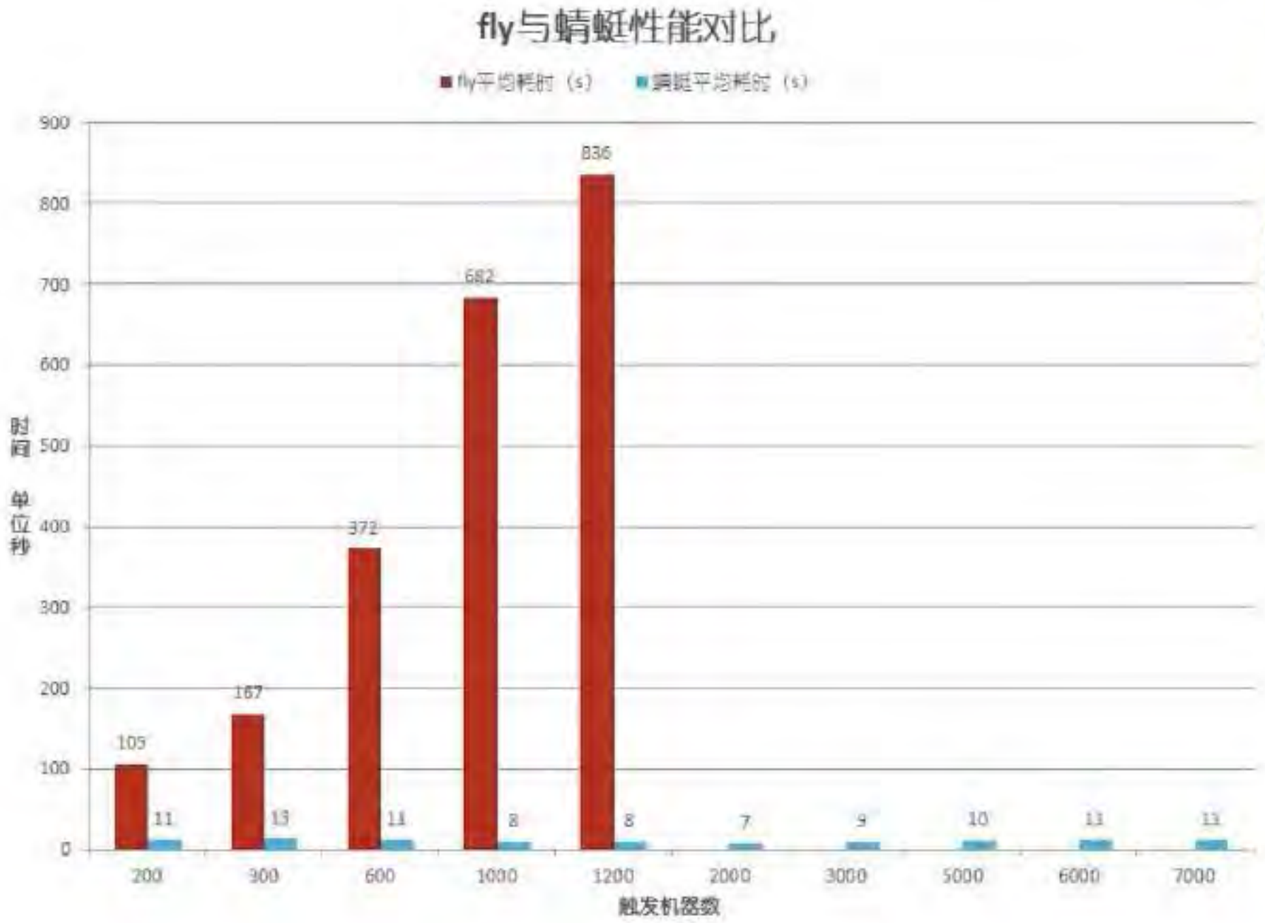
中间件EDAS

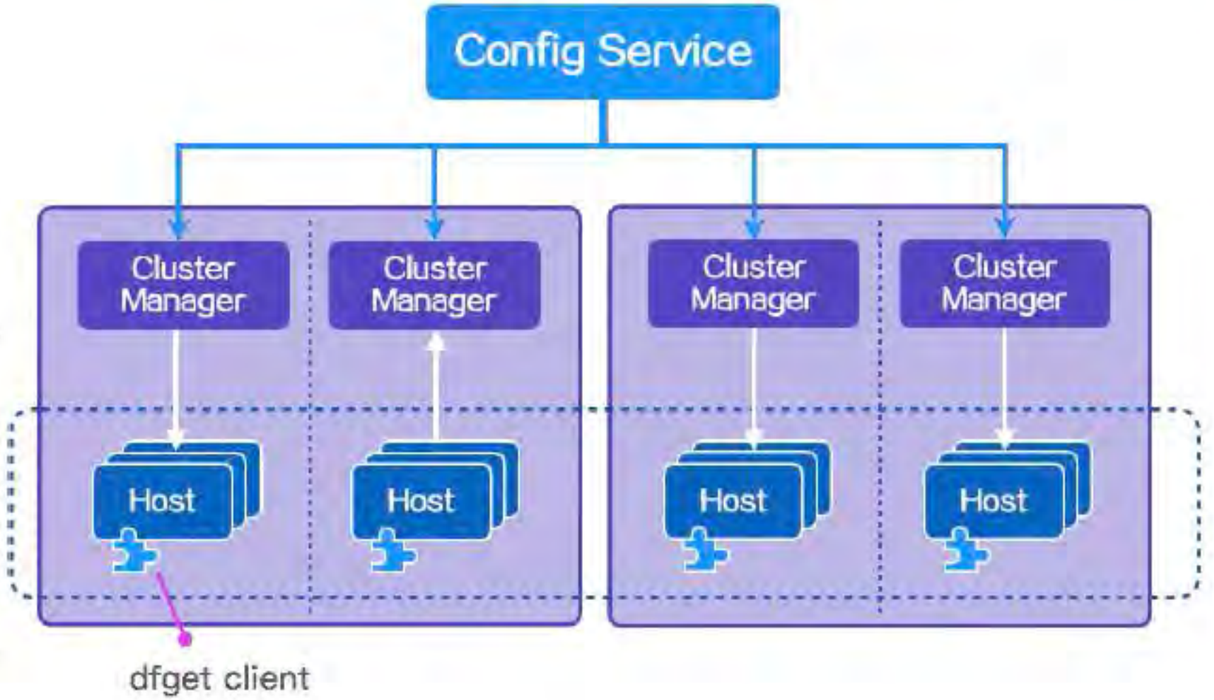
- 几百个输出项目 (StarAgent 1.0)



Werner Vogels of Amazon.com in a blog comment back in 2006:

"The best way to completely automate operations is to have to developers be responsible for running the software they develop. It is painful at times, but also means considerable creativity gets applied to a very important aspect of the software stack. It also brings developers into direct contact with customers and a very effective feedback loop starts. There is no separate operations department at Amazon: you build it; you run it."





保护数据源 Protect Source

加速分发速度 Boost Delivery Speed

节省夸IDC带宽 Save cross IDC bandwidth

节省跨国带宽 Save cross board bandwidth

:om

场景 Scenarios：

安装包
packages for deploy system

静态文件
image files, js files, CDN resources etc.

配置文件
config files

镜像
container images, VM images, Physical images etc.

数据文件
AI，Search Index，business hot data

其他文件
any other files served by http

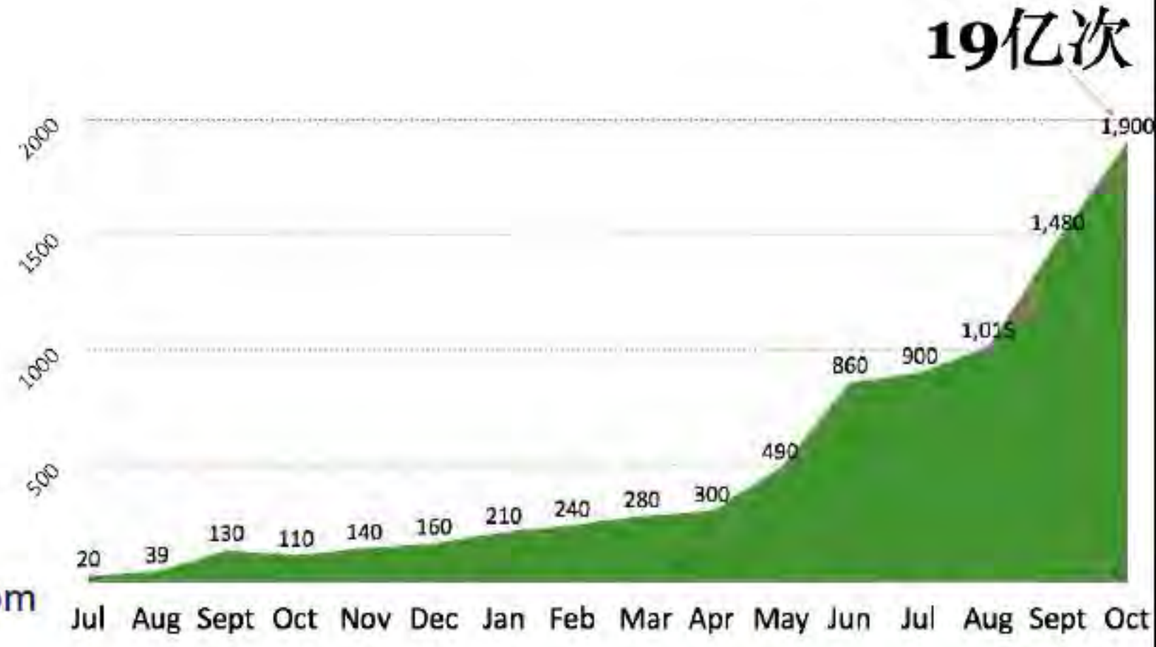
功能特性 Features:

断点续传
Resume from break point

智能网络/磁盘IO控制
Intelligent Network/Disk I/O control

智能动态压缩
Intelligent Dynamic compression **10%** best **86%** for images

镜像预热
Container Image Warmup (Pouch、Docker、Hyper) **67%** average saving **120x** max



全面：支持pouch、docker、hyper等各种容器镜像

无缝：0侵入，无需daemon改造

极速：支持镜像预热，push时完成并行分层预热

稳定：支持30GB甚至更大的镜像分发

高效：支持内存文件系统，极致的资源利用和极致的效率提升

极简：完美替代mirror，并在窄带、跨洋、远距离传输等极端条件下具备世界级竞争力

动态压缩提升整体速度

开源：

Dragonfly
community

Unwatch 32 Star 219 Fork 33

https://github.com/alibaba/dragonfly

协议 Apache 2.0

https://github.com/alibaba/dragonfly

功能：P2P文件分发，容器镜像分发、局部限速、磁盘容量预检

企业版：

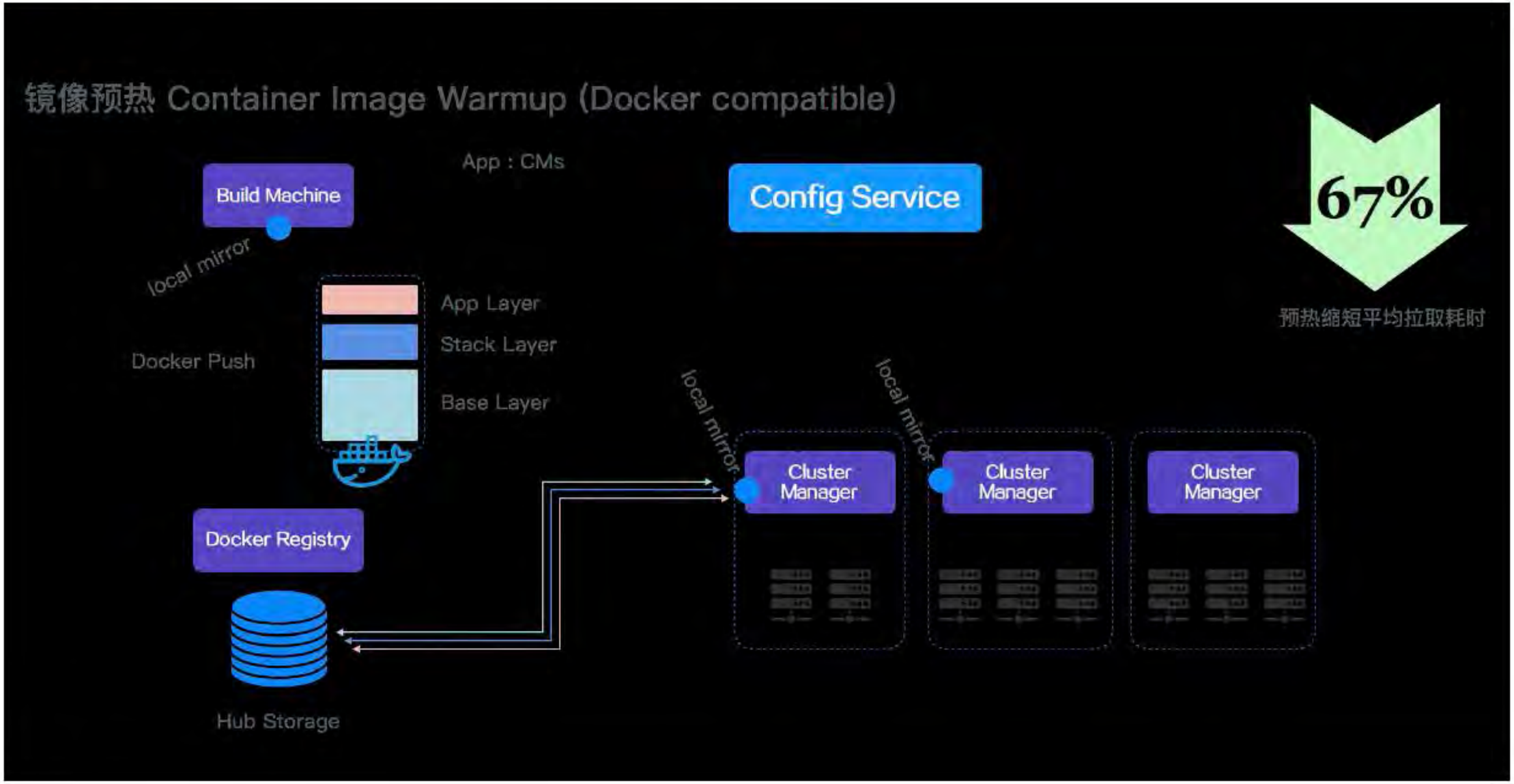
Dragonfly
enterprise

功能：断点续传、全局限速、镜像预热、支持内存文件系统、
智能网络流控、智能动态压缩、智能调度策略

渠道：云效
阿里云容器服务（公有云、专有云）

https://www.aliyun.com/product/containerservice

https://www.aliyun.com/solution/dedicatedcloud

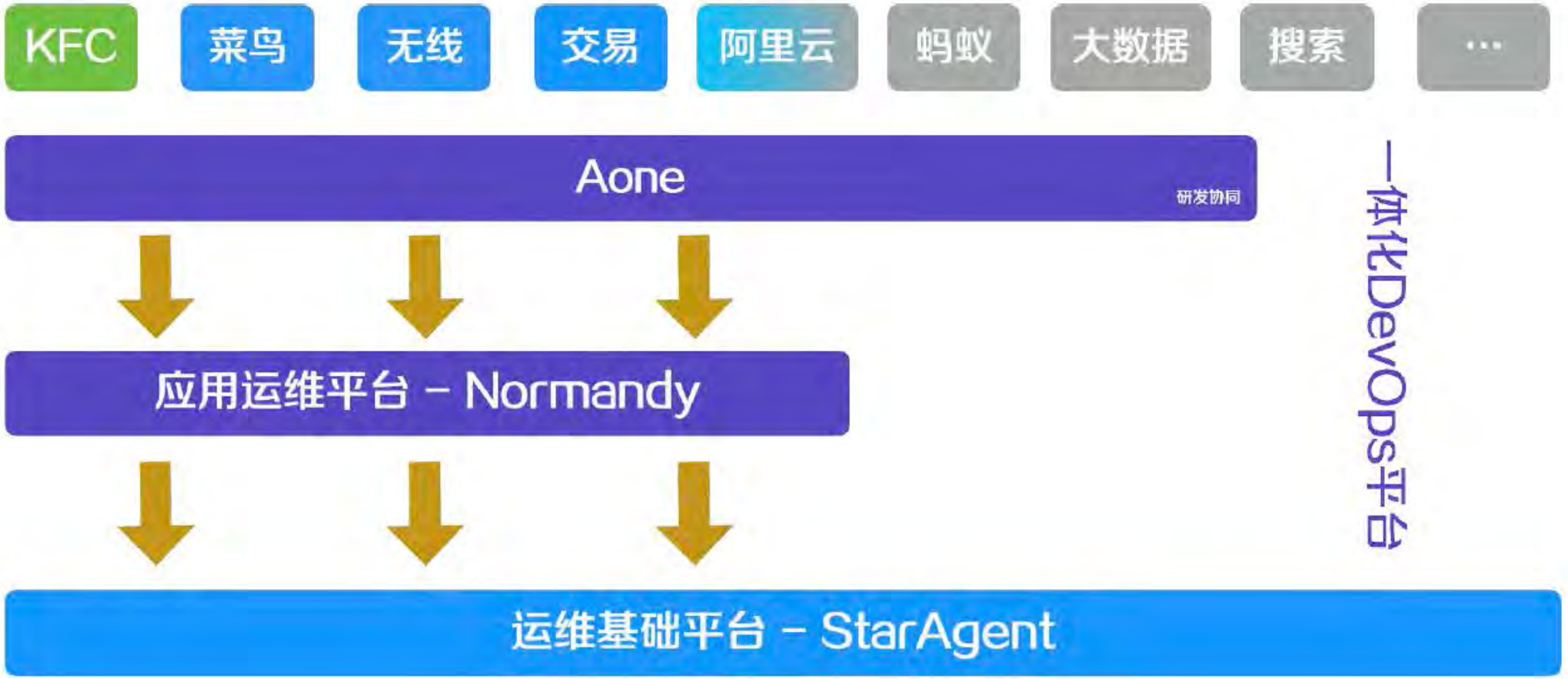

镜像预热 Container Image Warmup (Docker compatible)
App : CMs
Build Machine
local mirror
Config Service
67%
预热缩短平均拉取耗时
App Layer
Stack Layer
Base Layer
Docker Push
local mirror
local mirror
Cluster Manager
Cluster Manager
Cluster Manager
Docker Registry
Hub Storage

大促期间（11.10~11.11）共执行文件分发任务7600W次，其中镜像分发17.6W次。总流量118TB。

10点23分触发万台服务器下载2G的预热文件，总计耗时111秒。100%下载成功

应用运维平台

业务关系


KFC
菜鸟
无线
交易
阿里云
蚂蚁
大数据
搜索
...
Aone
研发协同
应用运维平台 - Normandy
运维基础平台 - StarAgent
一体化DevOps平台

环境类型

生产环境　预发环境　测试环境

复制环境

正式	正式	beta	灰度	
机器数量 10394	机器数量 10394	机器数量 10394	机器数量 10394	新建环境
当前版本 7579146	当前版本 7579146	当前版本 7579146	当前版本 7579146	（建设中）
基线一致性 99%	基线一致性 99%	基线一致性 99%	基线一致性 99%	
应用重启　资源变更	应用重启　资源变更	应用重启　资源变更	应用重启　资源变更	
发布　查看拓扑	发布　查看拓扑	发布　查看拓扑	发布　查看拓扑	

我的告警

最近12时告警数 21 条

时间	级别	摘要	状态	智能推荐操作
11-04 00:02	Warning	buy2 共有1235台机器触发，cpu利用率>70%，查看详情	运行中	机器重启
11-04 00:02	Critical	buy2 共有1235台机器触发，cpu利用率>70%，查看详情	待处理	清理磁盘
11-04 00:02	Critical	buy2 共有1235台机器触发，cpu利用率>70%，查看详情	成功	Java dump
11-04 00:02	Warning	buy2 共有1235台机器触发，cpu利用率>70%，查看详情	进行中	暂无推荐
11-04 00:02	Warning	buy2 共有1235台机器触发，cpu利用率>70%，查看详情	失败	机器重启

查看更多

变更记录

buy2
预发环境发布预发环境发布预发...
2017-11-09 13:05:22 春子

buy2
预发环境发布预发环境发布预发...
2017-11-09 13:05:22 春子

buy2
预发环境发布预发环境发布预发...
2017-11-09 13:05:22 春子

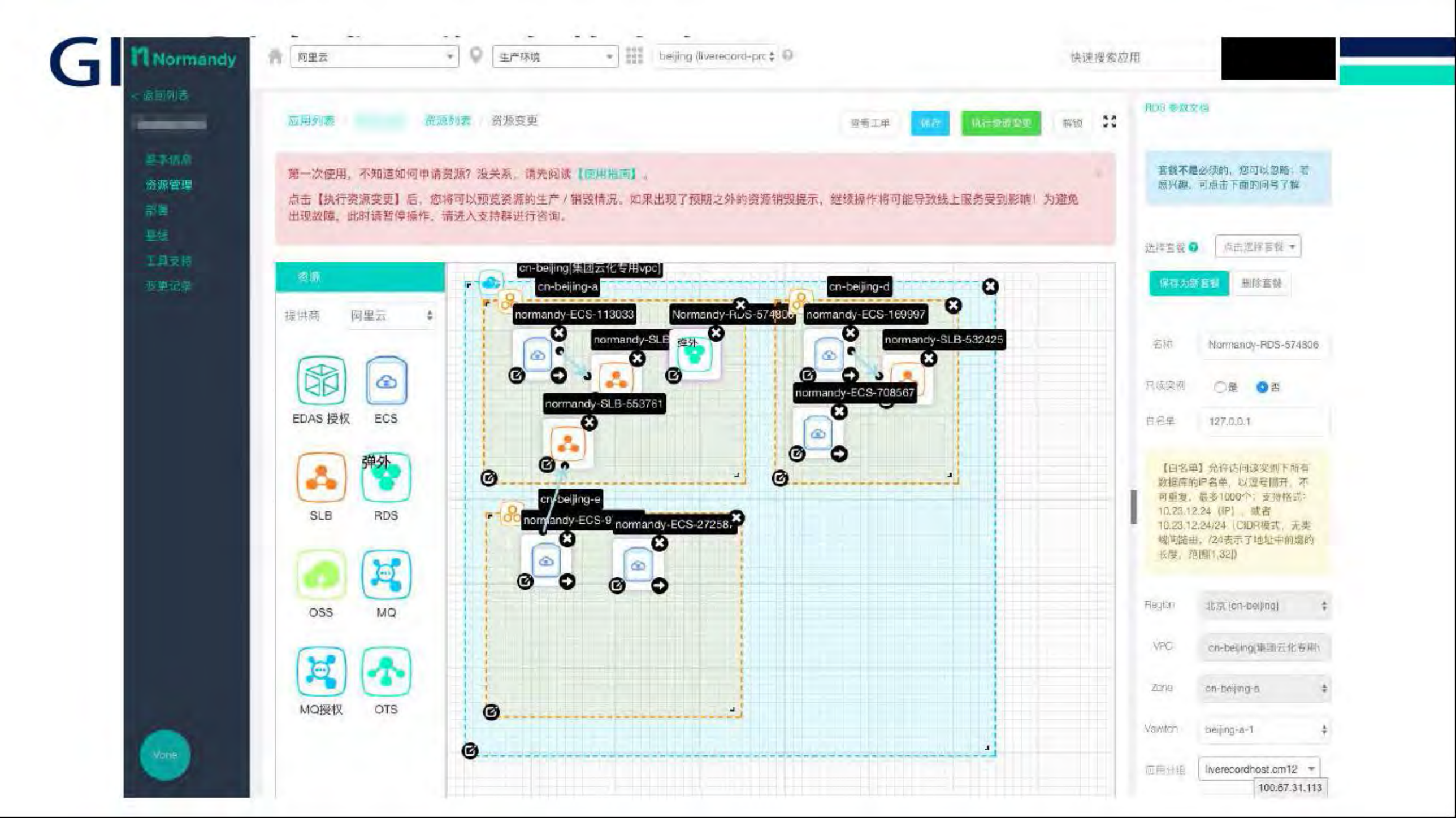
基础设施即代码

混合云一站式资源申请

资源基线

自动化

弹性伸缩


Normandy
阿里云
生产环境
beijing (liverecord-prc
快速搜索应用
< 返回列表
基本信息
资源管理
部署
监控
工具支持
变更记录
应用列表
资源列表
资源变更
查看工单
保存
执行资源变更
解锁
第一次使用，不知道如何申请资源？没关系，请先阅读【使用指南】。
点击【执行资源变更】后，您将可以预览资源的生产 / 销毁情况。如果出现了预期之外的资源销毁提示，继续操作将可能导致线上服务受到影响！为避免出现故障，此时请暂停操作，请进入支持群进行咨询。
资源
提供商
阿里云
EDAS 授权
ECS
SLB
RDS
OSS
MQ
MQ授权
OTS
cn-beijing|集团云化专用vpc
cn-beijing-a
cn-beijing-d
normandy-ECS-113033
Normandy-RDS-574806
normandy-ECS-169997
normandy-SLB-532425
normandy-SLB-553761
normandy-ECS-708567
cn-beijing-e
normandy-ECS-27258
RDS 参数文档
套餐不是必须的，您可以忽略；若感兴趣，可点击下面的问号了解
选择套餐
保存为新套餐
删除套餐
名称
Normandy-RDS-574806
只读实例
是
否
白名单
127.0.0.1
【白名单】允许访问该实例下所有数据库的IP名单，以逗号隔开，不可重复，最多1000个；支持格式：10.23.12.24（IP），或者10.23.12.24/24（CIDR模式，无类域间路由，/24表示了地址中前缀的长度，范围[1,32]）
Region
北京 (cn-beijing)
VPC
cn-beijing(集团云化专用)
Zone
cn-beijing-a
Vswitch
beijing-a-1
应用分组
liverecordhost.cm12
100.67.31.113
Vone

GIAC | 全球互联网
GLOBAL INTERNET ARCHITE

创建发布单

*选择应用 ali-pad-pushcenter

环境级别 预发 测试 正式

发布信息

*环境 杭州预发 上海预发 山东预发 美国预发 日本预发 更多

*发布类型 Pouch Image Diamond RPM安装包 压缩文件 Docker Image

*分批方式 机房均分 分组均分

*暂停策略 第一批暂停 每批暂停

*分批数量 3

*计划发布时间 2017-11-27 15:30

版本信息

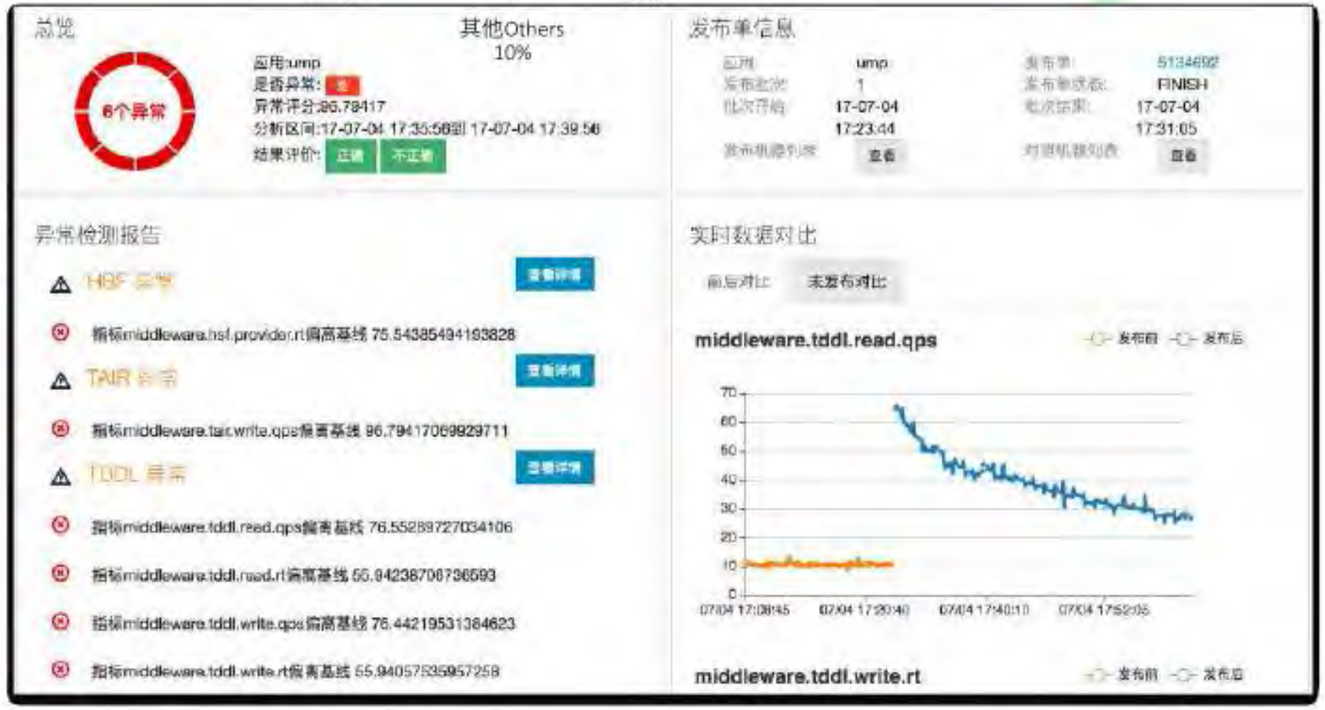
V-1.6.8	猫子	2017-1-23 18:00:00	V-1.6.5	猫子	2017-1-23 18:00:00
V-1.6.7	猫子	2017-1-23 18:00:00	V-1.6.4	猫子	2017-1-23 18:00:00
V-1.6.6	猫子	2017-1-23 18:00:00	查看更多版本		

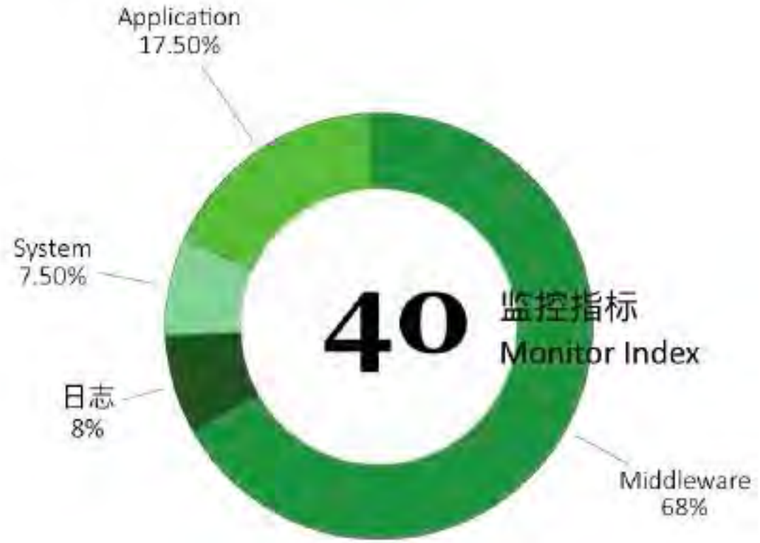
机器信息

无人值守发布变更

发布/变更故障 deploy failure

30%





System Montoring	CPU, Memory, IO	User, System, Load etc.
Container	Tomcat,JVM etc.	HTTP failure/error,thread etc.
Middleware	HSF,Notify,TDDL,MetaQ etc.	QPS,RT etc.
Application	Log	Exception cout,rate,trend etc.
Business	Login, Transaction etc.	count, rate, trend etc.

基本策略

新增的Exception

Fatal Exception （NoClassDefine, NPE...）

旧的Exception，但是Rate飙高

业务指标
系统指标

} 异常检测

趋势　同比　环比　静态阈值　分段静态阈值　动态阈值

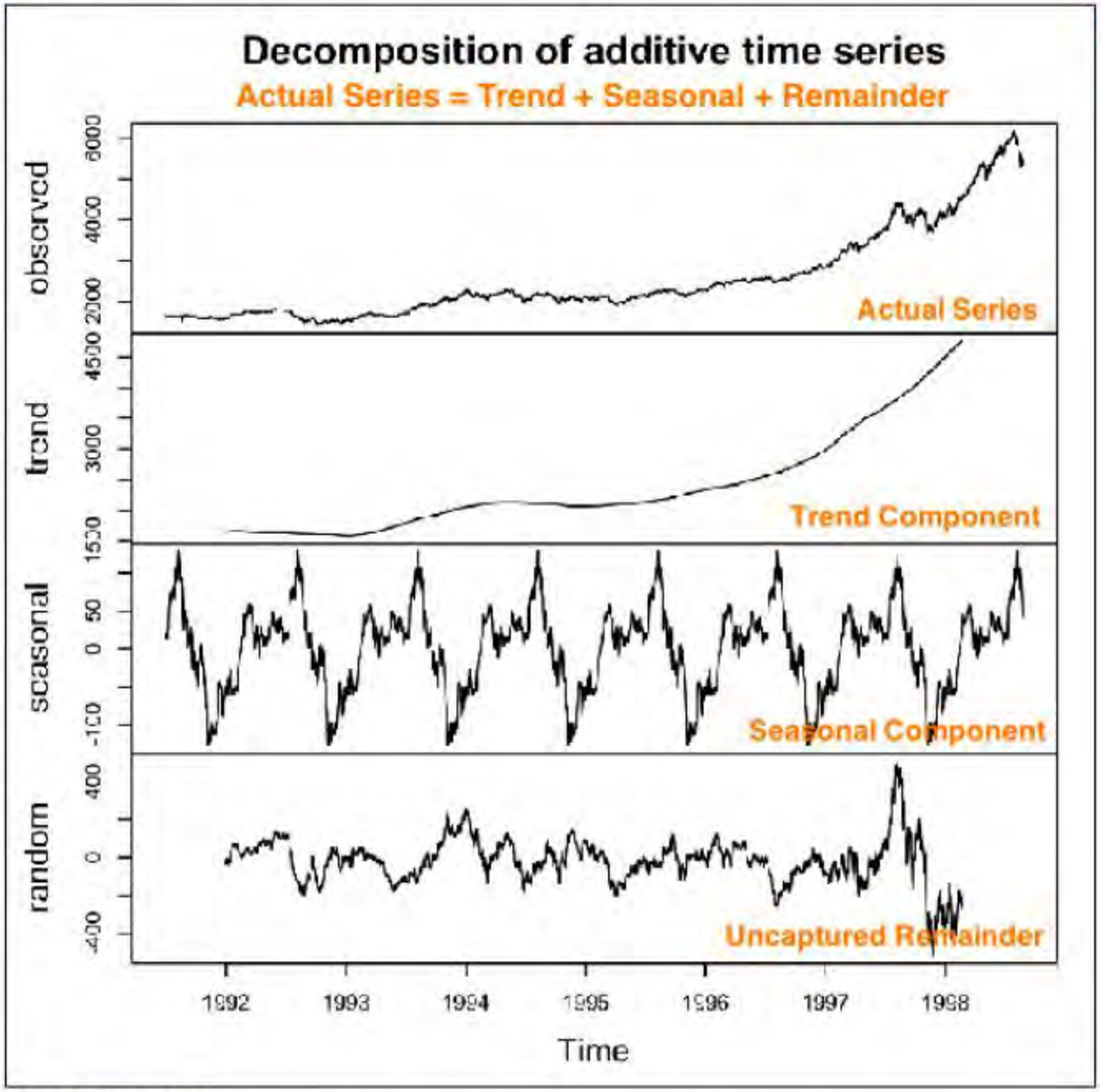
时间序列预测

分段历史平均

Arima

Holt-Winters

STL



预测：三次指数平滑Holt-Winters

适用于趋势性(Trend)和周期性(Seasonality)的时序指标

模型参数：α / β / γ　截距/斜率/周期平滑系数

参数确定:

◆ 人工配置

◆ 自动训练: 排除异常点 ->最大化拟合度

- Prediction:
  $\hat{y}_{t+1} = a_t + b_t + c_{t+1-m}$
- Baseline ("intercept"):
  $a_{t+1} = \alpha(y_{t+1} - c_{t+1-m}) + (1-\alpha)(a_t + b_t)$
- Linear Trend ("slope"):
  $b_{t+1} = \beta(a_{t+1} - a_t) + (1-\beta)b_t$
- Seasonal Trend:
  $c_{t+1} = \gamma(y_{t+1} - a_{t+1}) + (1-\gamma)c_{t+1-m}$

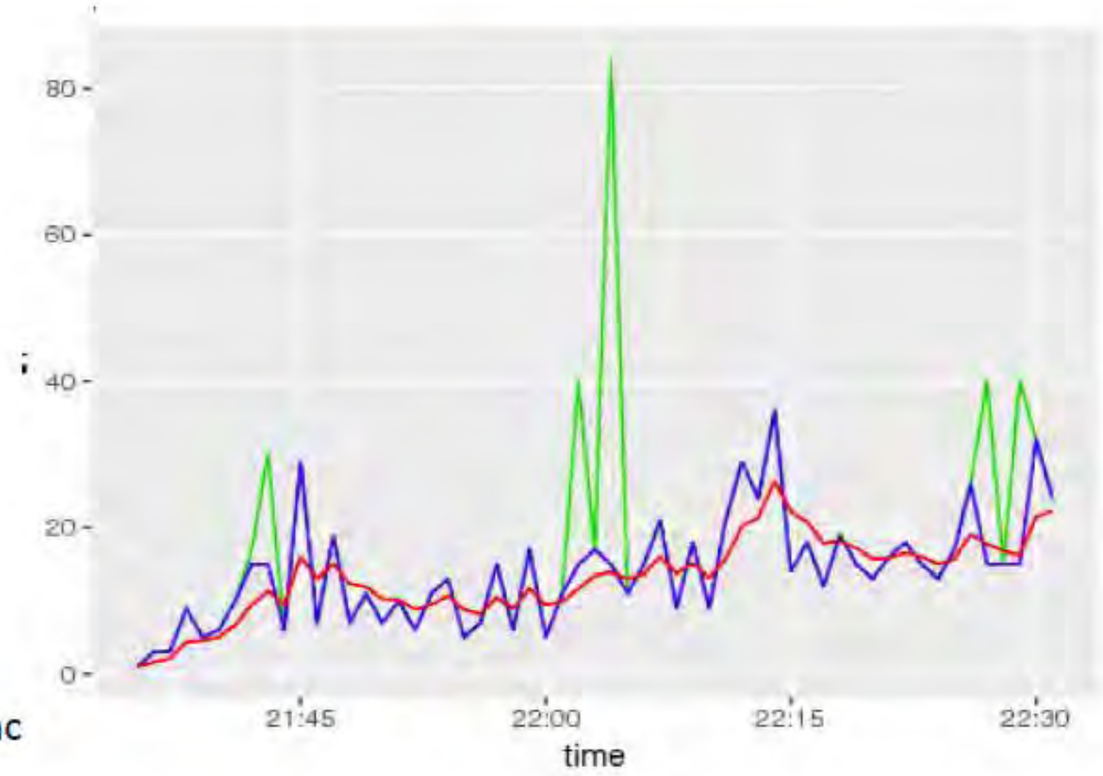
异常判定：

明确上下界: 预测值±δ

固定阈值

历史周期点的指数平滑

滑动窗口的偏差标准差



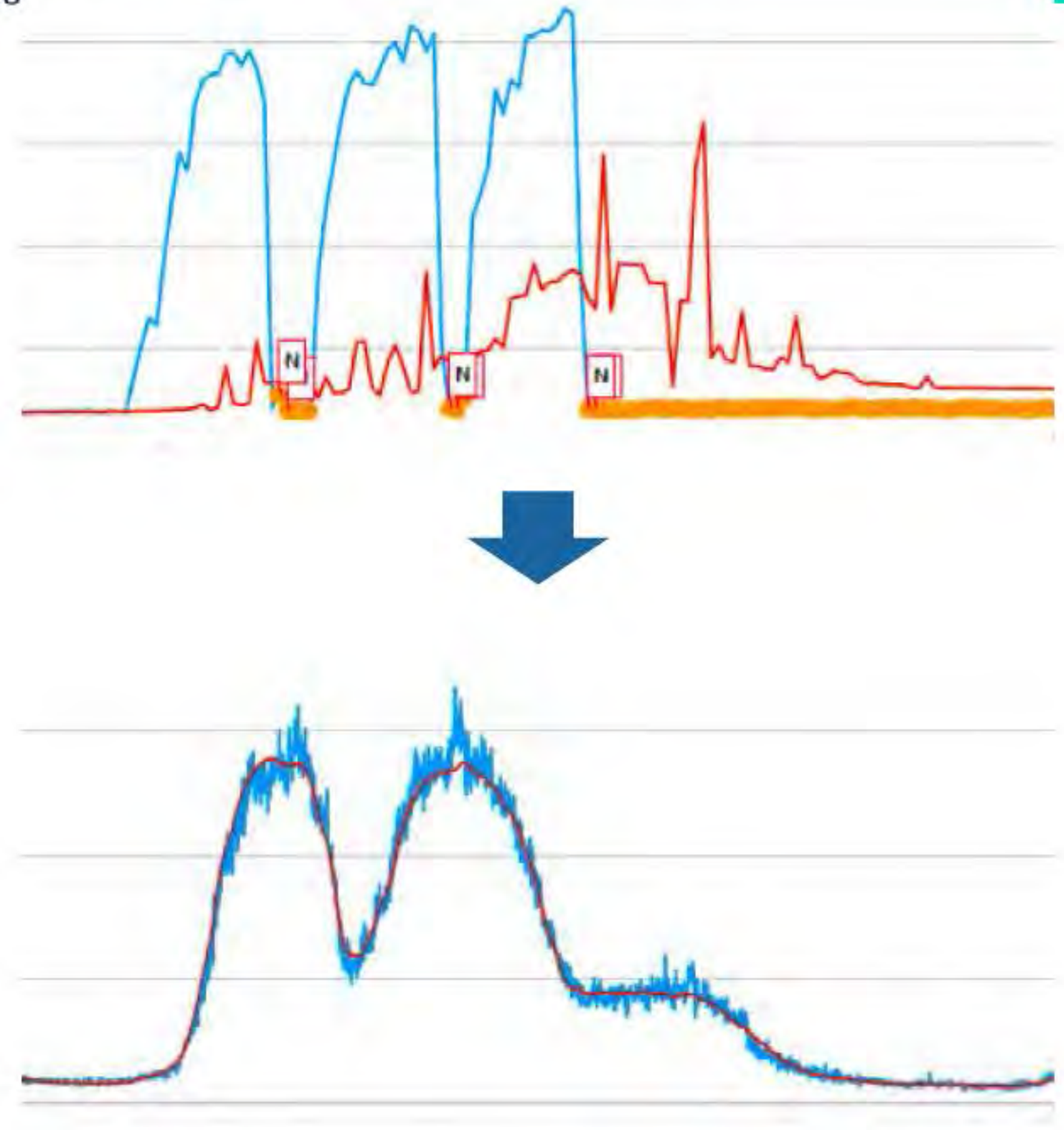
关键步骤

数据补全

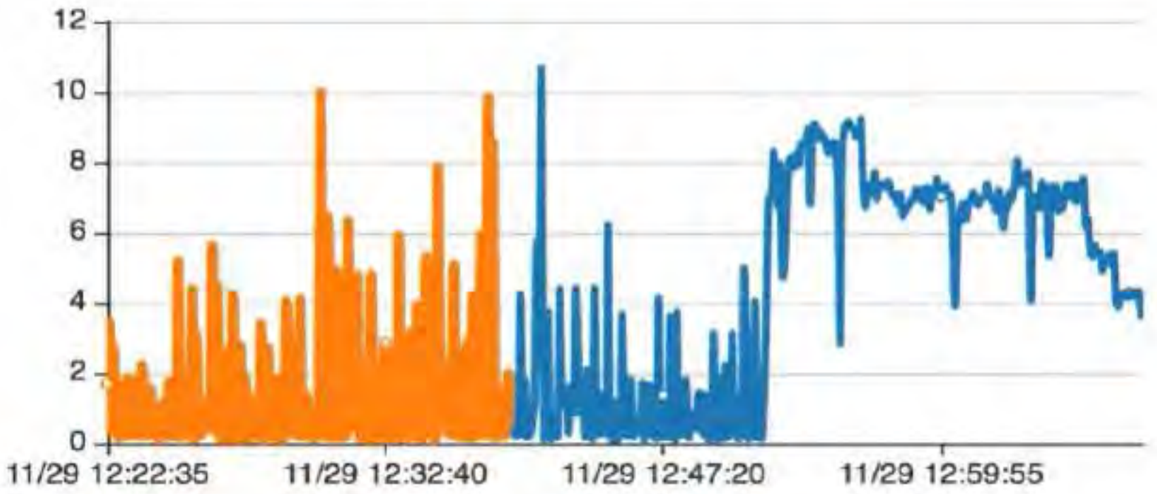
日期类型划分

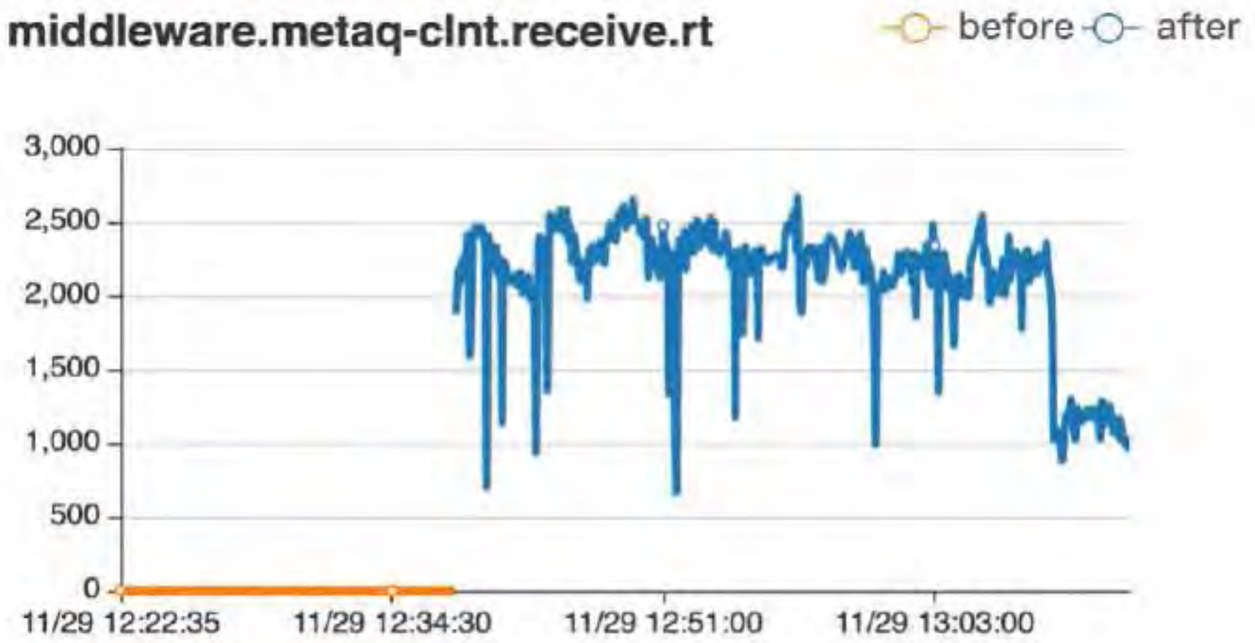
预测“趋势”

局部趋势反馈



系统指标

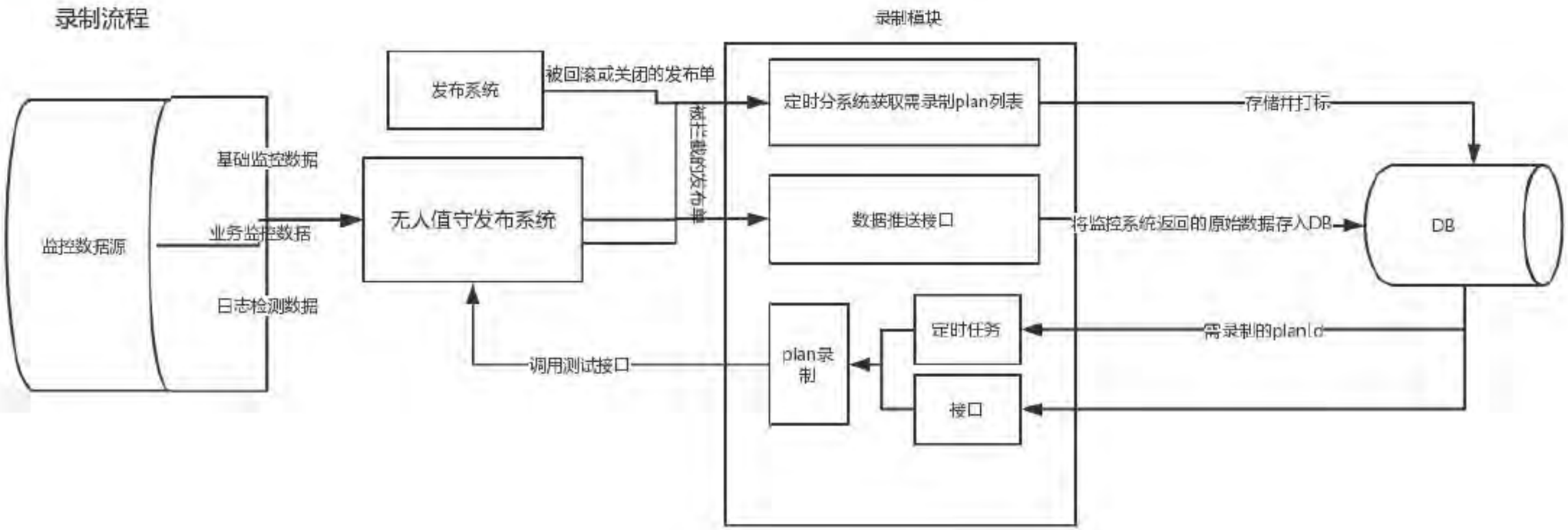

middleware.metaq-clnt.receive.rt
before
after
12
10
8
6
4
2
0
11/29 12:22:35
11/29 12:32:40
11/29 12:47:20
11/29 12:59:55


middleware.metaq-clnt.receive.rt
before
after
3,000
2,500
2,000
1,500
1,000
500
0
11/29 12:22:35
11/29 12:34:30
11/29 12:51:00
11/29 13:03:00

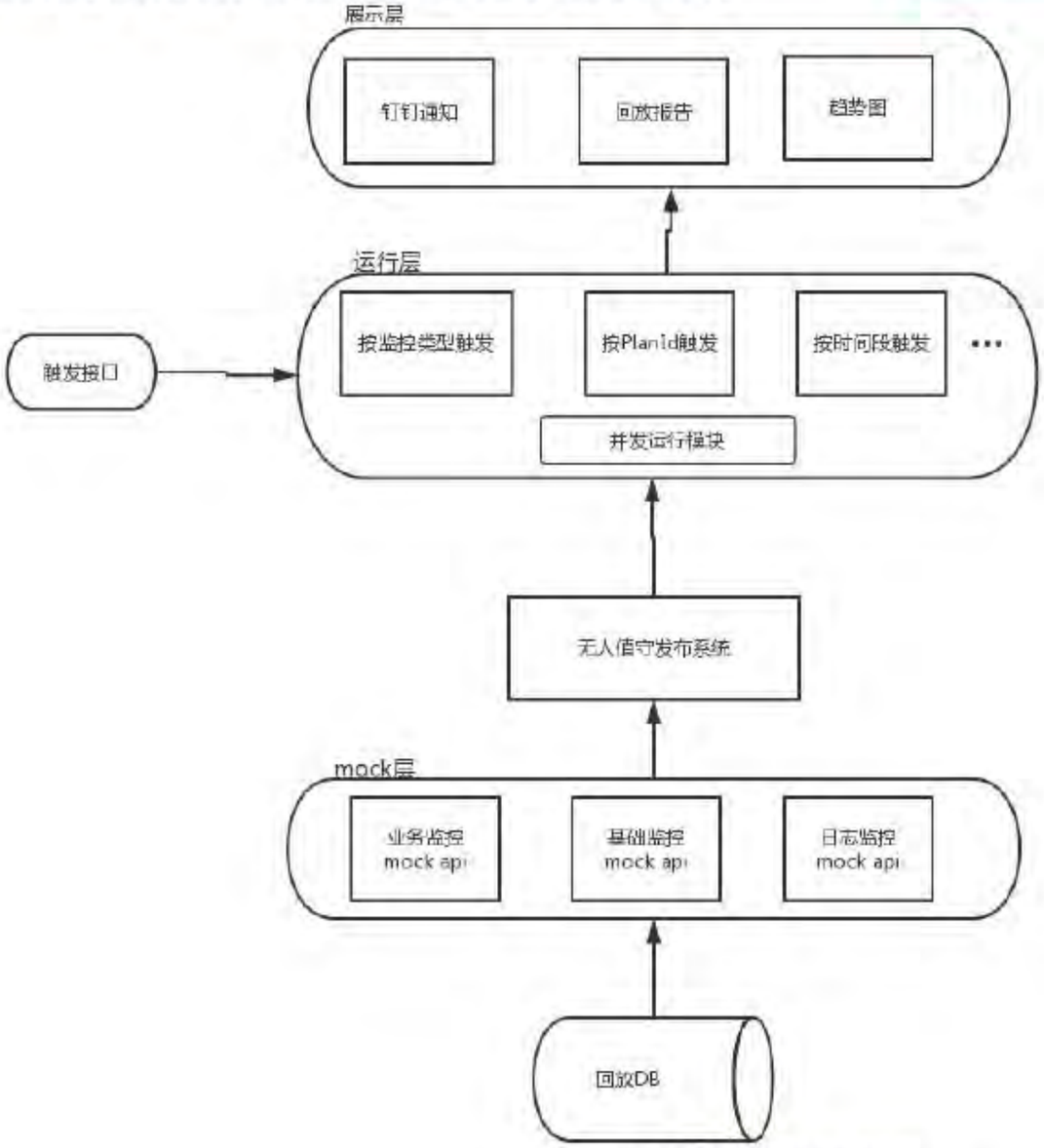
效果

Method	Precision	Recall	F1	False Positives	False Negatives	Total Positive
AUC	0.33	0.90	0.49	18	1	9
DTW	0.57	0.82	0.67	7	2	11
AUC+DTW	0.80	0.80	0.80	2	2	10
Risk-free	0.33	1.00	0.40	24	0	12

反馈 – 录制


录制流程
录制模块
发布系统
被回滚或关闭的发布单
定时分系统获取需录制plan列表
存储并打标
基础监控数据
被拦截的发布单
监控数据源
业务监控数据
无人值守发布系统
数据推送接口
将监控系统返回的原始数据存入DB
DB
日志检测数据
定时任务
需录制的planId
plan录制
调用测试接口
接口

反馈 – 回放


展示层
钉钉通知
回放报告
趋势图
运行层
触发接口
按监控类型触发
按PlanId触发
按时间段触发
...
并发运行模块
无人值守发布系统
mock层
业务监控
mock api
基础监控
mock api
日志监控
mock api
回放DB

问题和挑战

应用画像

日志异常/error库

接入成本

日志路径很难自动探测到

业务指标缺失，需要配置，甚至要改代码（500个应用 1500个指标）；机器维度的聚合

数据质量

数据打标少、准确度低

监控数据准确性，断图

GOC故障分类、变更关联

数据数量

应用发布时多数业务量较少甚至没有

判定标准

固定阈值