

# 大规模分布式系统全链路监控探索实践

李明亮



# 李明亮

## 天眼平台项目总监

中国联通软件研究院天眼平台项目总监，天眼平台负责人，曾任高级业务架构师，负责腾讯王卡、蚂蚁宝卡等2I2C业务支撑，目前主要关注运维平台建设，DevOps文化践行与团队管理。

# 目录

contents

- 1 云原生下应用运维的挑战
- 2 分布式应用链路追踪实践
- 3 关于链路追踪演进的思考

# PART 1

## 云原生下应用运维的挑战

# 1.1 云原生下的挑战与需求

应用几何级数增长，中间件快速变化，部署动态切换



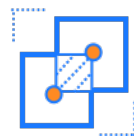
2019年起，软研院对**cBSS、新客服、公众中台、政企中台**等大型生产系统进行上云改造，主要使用**Kubernetes、Mesos、Edas**等多种云平台，涉及服务**5000+**，天眼平台亟需打造一款可以**跨数据中心、跨云平台、跨系统**的分布式应用性能管理产品：



自动拓扑：支持服务与服务，服务与组件，服务与主机等所有调用关系自动生成



故障发现：分钟级故障发现，实时告警运维人员，快速发现生产运行中的问题

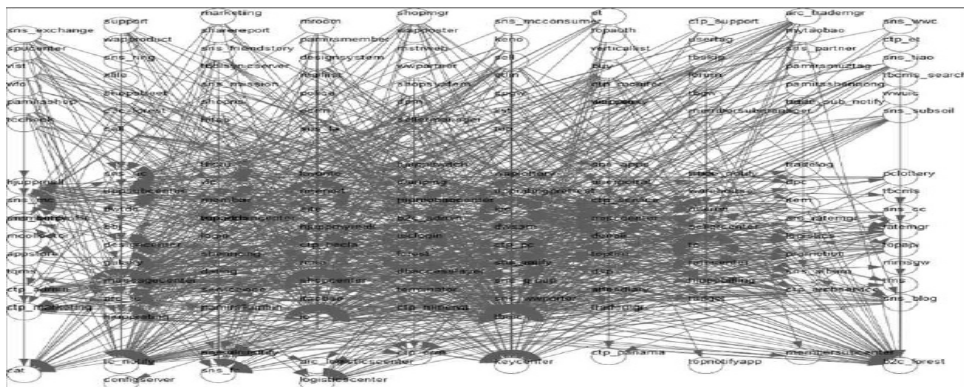


根因定位：多服务故障时，只发送根因服务信息，实现分钟级故障定位，支持根因服务、根因组件、根因主机等多种场景



一键诊断：一键诊断故障根因，傻瓜式操作提升故障定位速度，降低使用人员门槛

调用承载关系极其复杂，亟待引入运维工具



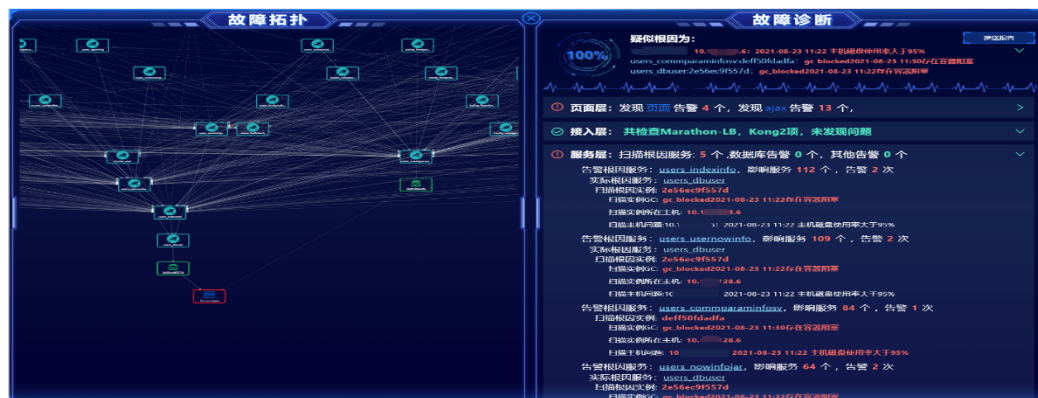
# 1.2 云原生下典型故障案例



故障根因在SaaS服务下的实例



故障根因在PaaS组件



故障根因在IaaS主机



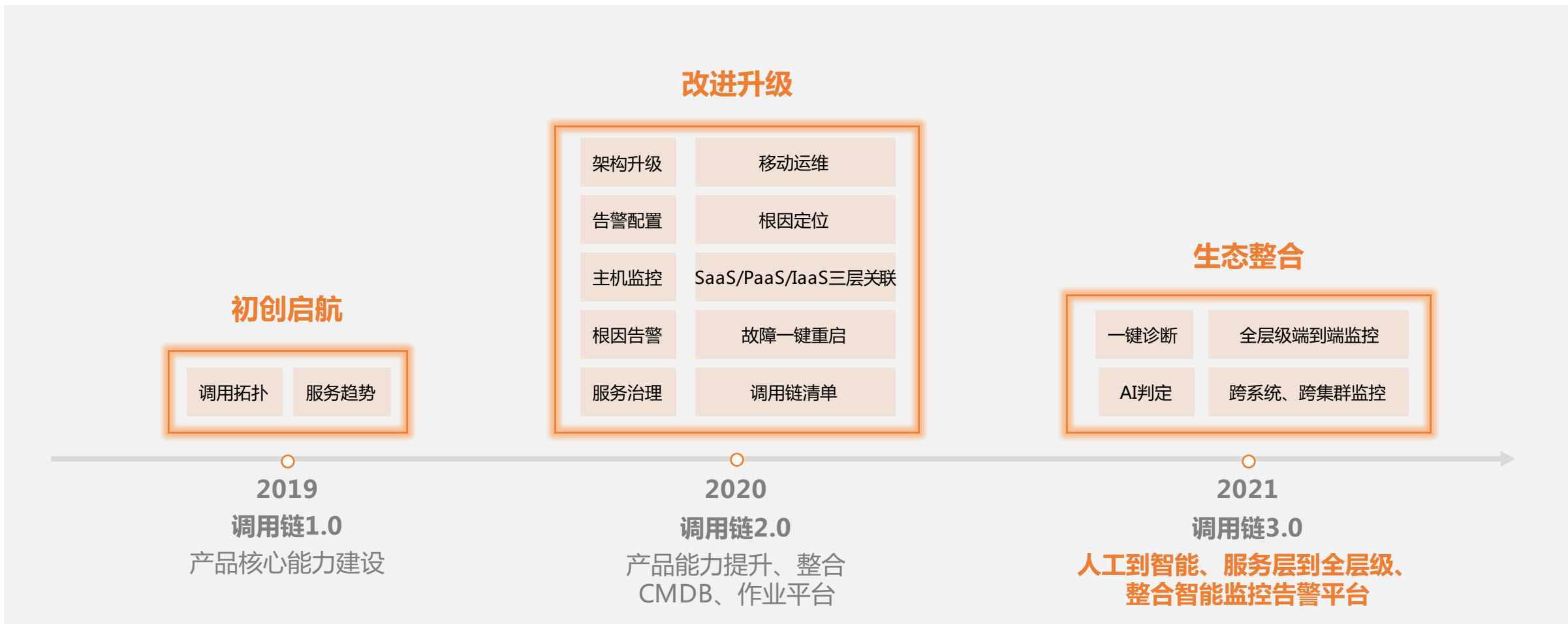
故障根因在外部接口

# PART 2

## 分布式应用链路追踪实践

# 2.1 天眼全流程调用链演进历程

天眼全流程调用链定位于解决云原生下监控问题，历经三个阶段

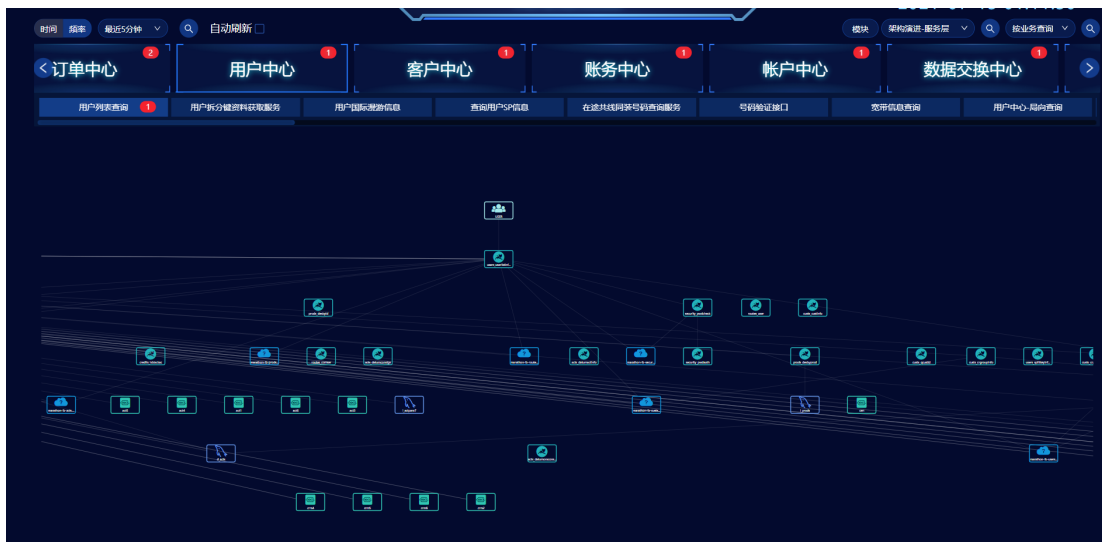




## 2.2 2019年初创起航-基本功能

2019年初创启航阶段，由使用开源产品逐步改造，形成自主分布式应用性能监控产品的尝试。

- 调用拓扑：开源采集服务调用指标，解析到neo4j图数据库，形成调用拓扑图
- 服务趋势：根据历史数据收集，进行调用趋势展现，呈现故障走势



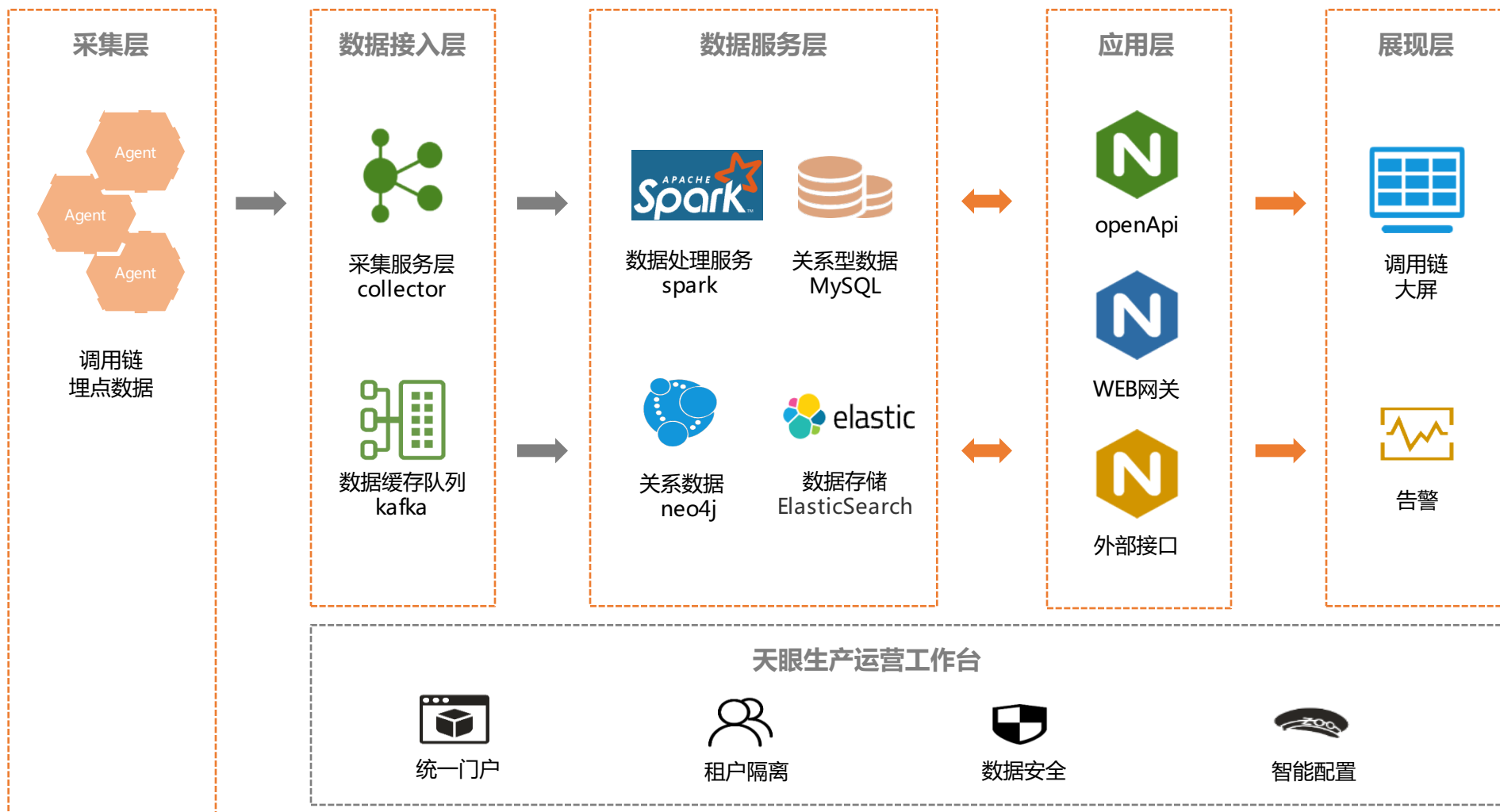
调用拓扑



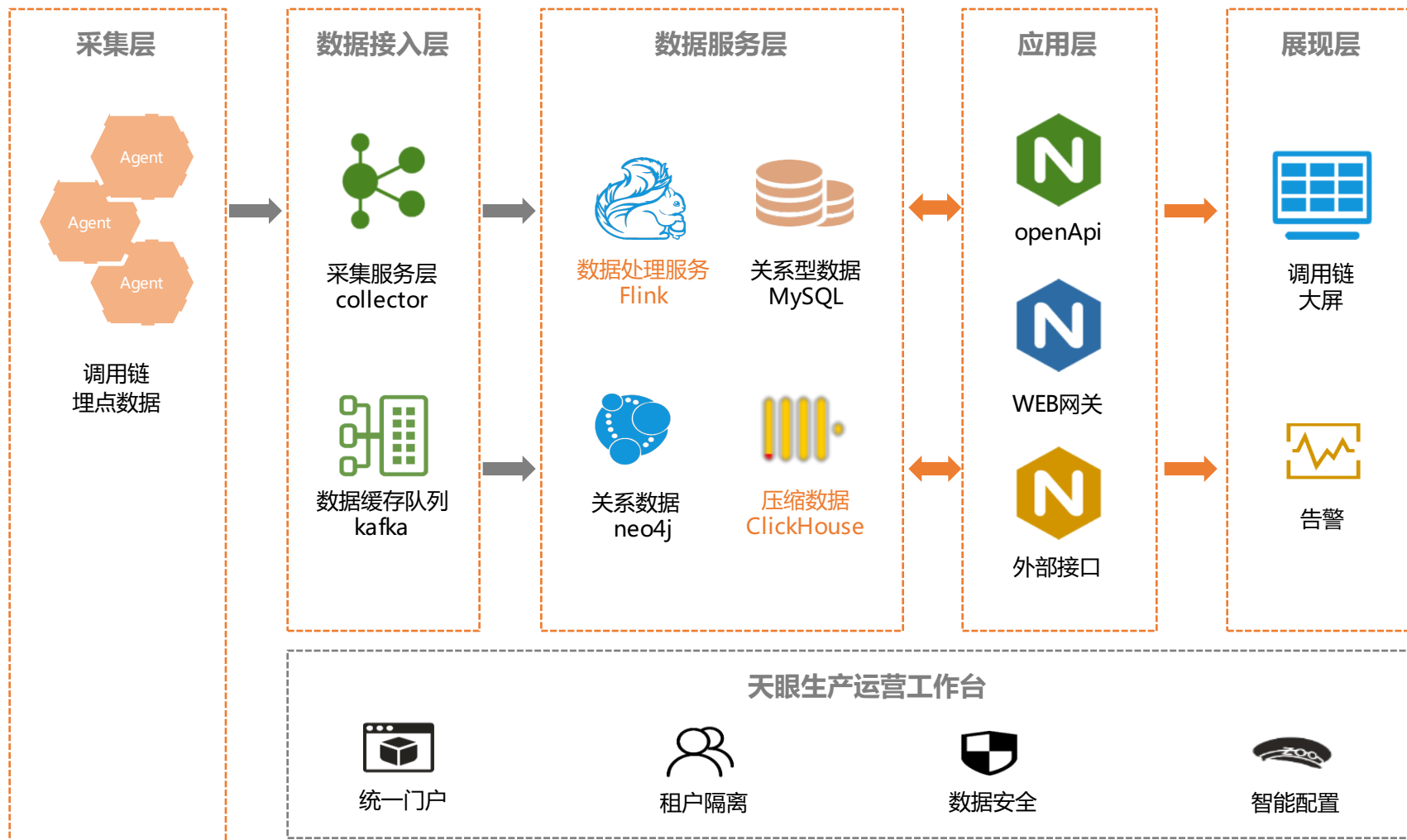
服务趋势

# 2.3 2019年初创起航-初始架构

2019年主要试点接入cBSS-2i项目。日均处理指标数据**30亿+**，架构中使用**ElasticSearch**存储+**Spark**计算。



# 2.4 2020年改进升级-第一次技术架构升级

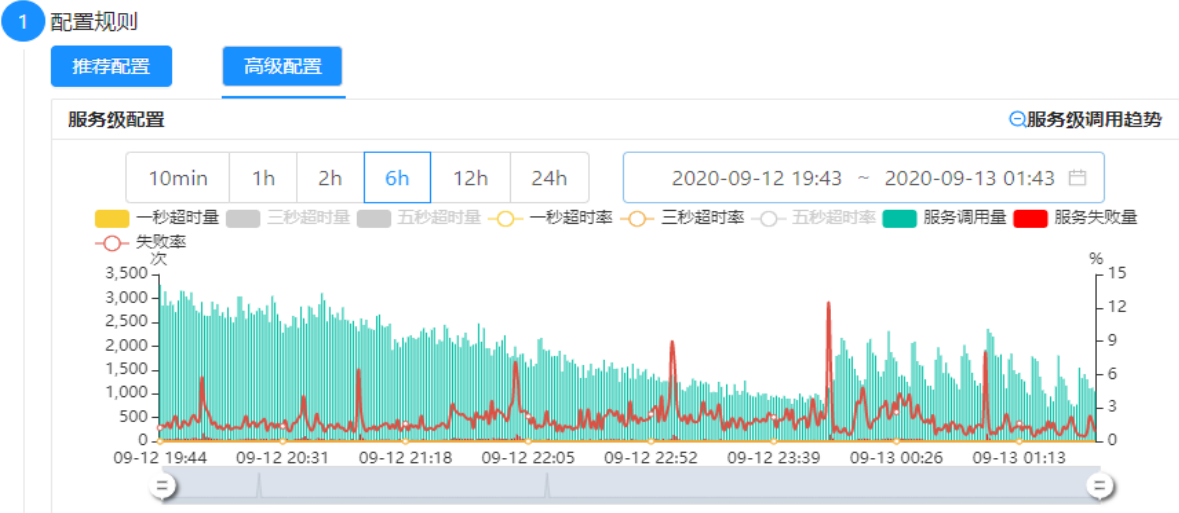


- **问题**：系统接入cBSS全量服务后数据量到每日**400亿**，服务**3000+**，经常发生数据积压，告警延迟**10分钟发送**

- **举措**：2020年1-2月，从**ElasticSearch**换成**ClickHouse**提升数据存储能力，从**Spark**换成**Flink**提升计算效率

- **效果**：告警处理时间缩短到**1分40秒**

# 2.5 2020年改进升级-告警配置与规则优化



- **问题**：早期告警都由**调用链研发自己后台配置**，随着服务数量及接入系统数量增加，研发人员已无法支撑

- **举措**：2020年2月增加**告警配置功能**，将告警配置开放给接入系统使用人员，由各接入系统自主配置

报错分类详情

系统异常 业务异常

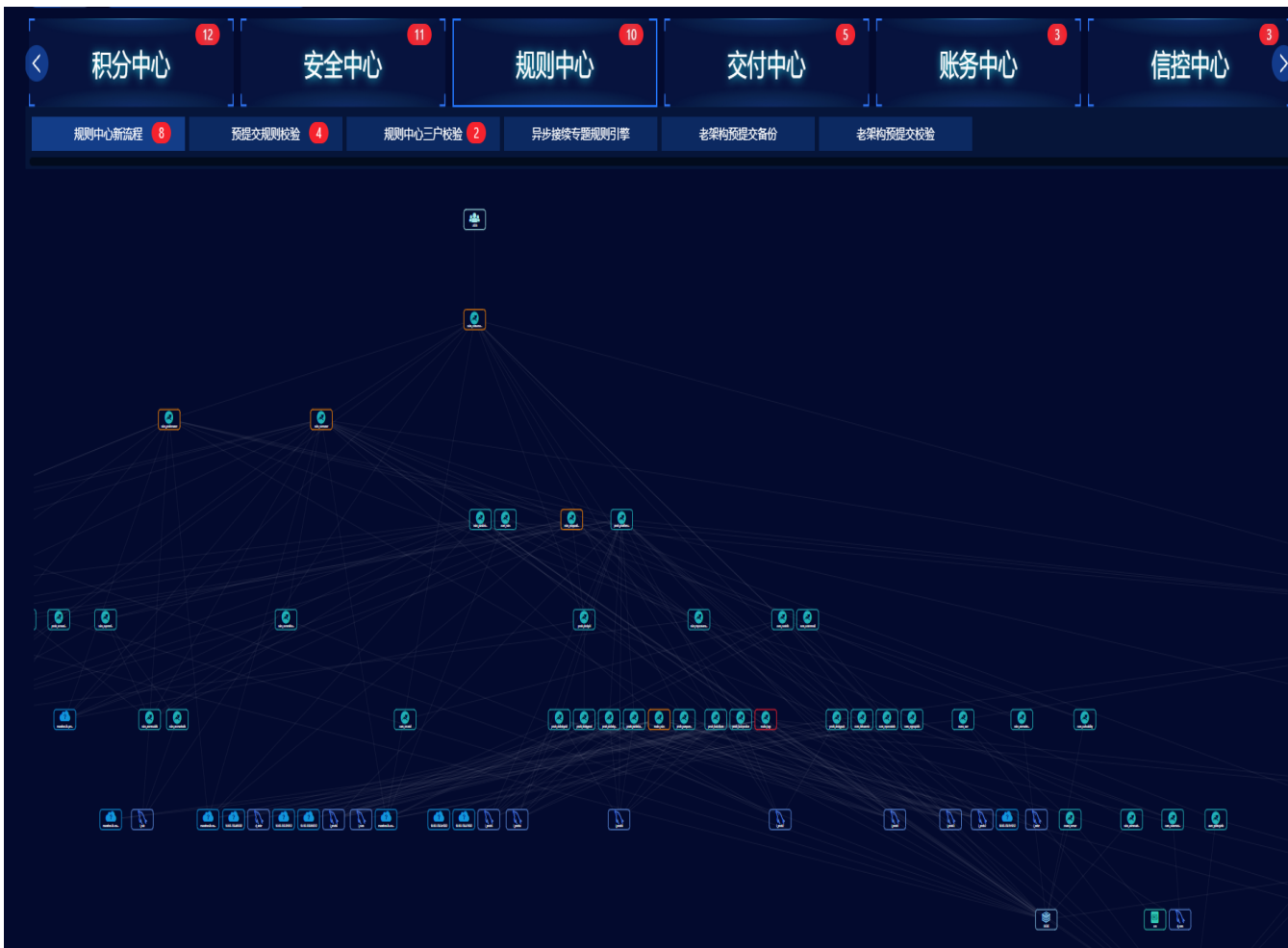
错误信息	错误类型	数量
fallbackCallTuxedo: null	方法异常	1403
call4GTuxSvcCmd3 failed and fallback fail...	方法异常	825

< 1 > 10条/页

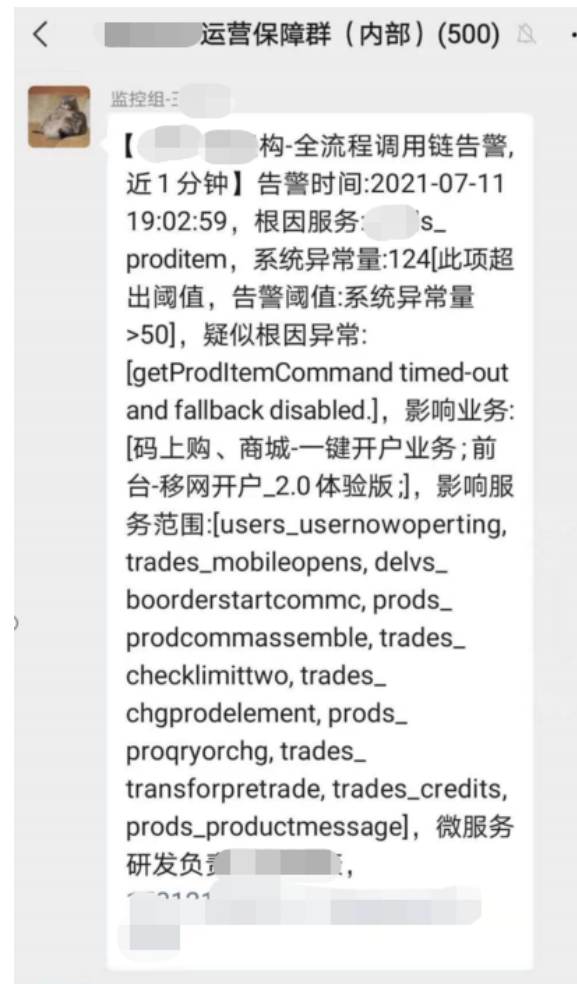
- **问题**：**服务失败、业务规则限制异常、系统级报错异常**，混杂一起告警不准确

- **举措**：2020年2月通滤出**“系统异常”**，进行告警配置，大大提高告警准确性/可用度

# 2.6 2020年改进升级-根因服务告警



调用链展现

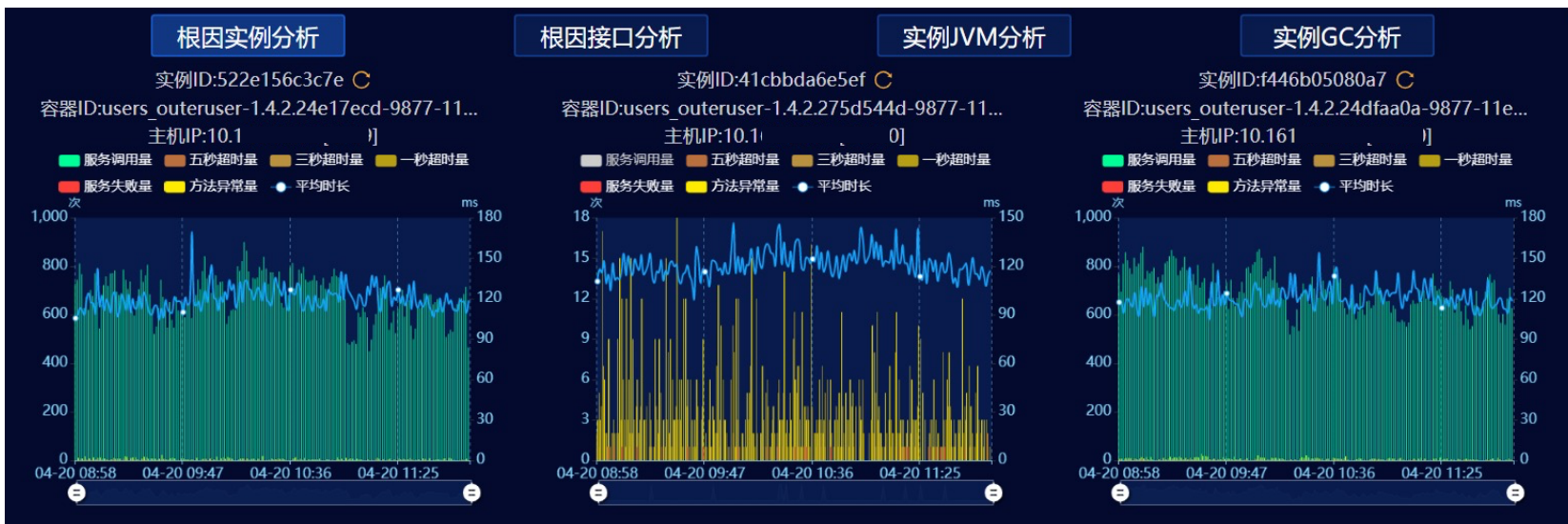


根因告警信息

- **问题**：故障发生时，如果是底层服务问题，会产生服务大范围的报错，容易形成告警风暴，使用人员定位难

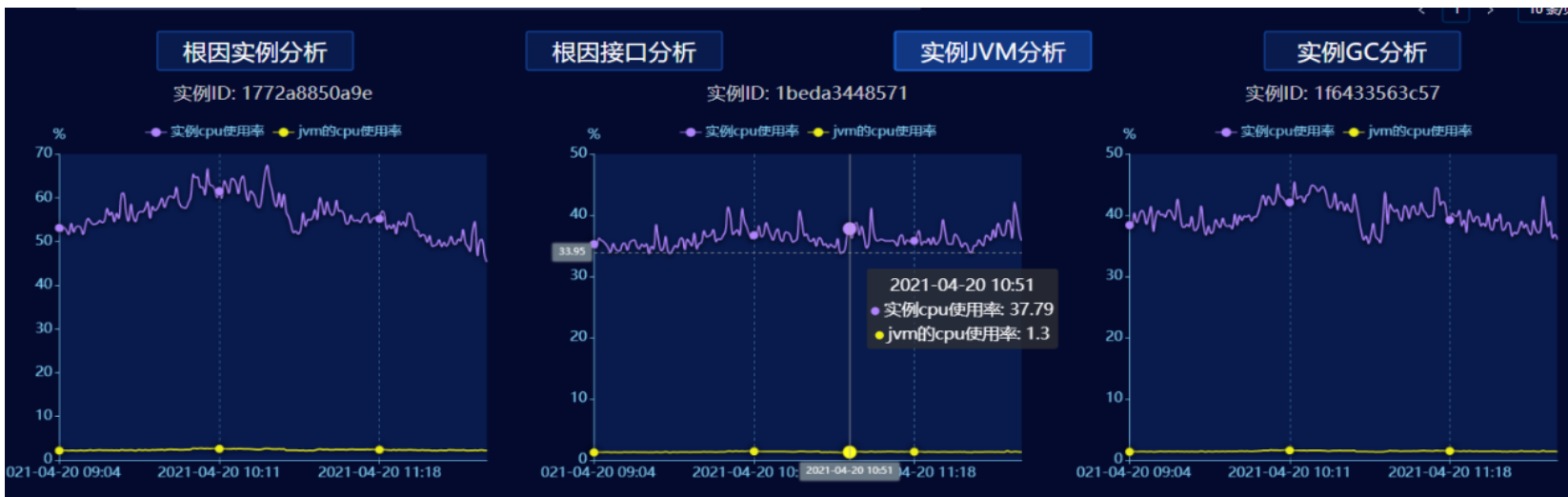
- **举措**：2020年3月通过根因判定后的规则收敛大大降低的告警数量，**当大量服务告警时只告警根因服务**

# 2.7 2020年改进升级-多维根因定位

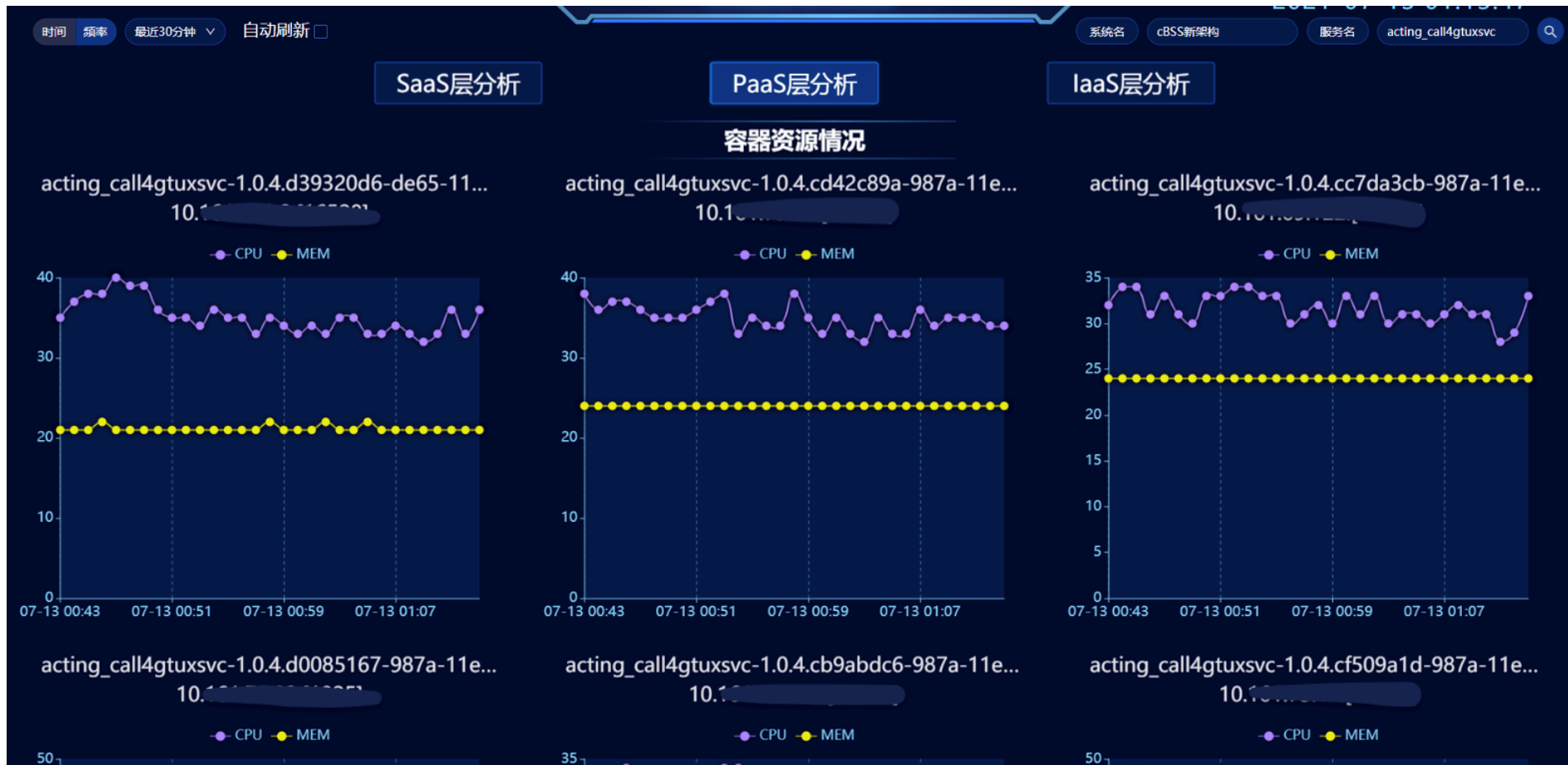


● **问题**：服务发生故障，经常由于**单个实例或者实例GC导致**，运维人员排查困难

● **举措**：2020年4月，天眼调用链提供了**根因定位**的能力，定位到根因服务后继续定位**根因实例、根因接口、实例JVM、实例GC**等能力



# 2.8 2020年改进升级- SaaS/PaaS/IaaS三层关联



● **问题**：服务发生故障时经常由于容器或者主机问题导致

● **举措**：2020年5月，天眼调用链创新地提供了跨**SaaS层->PaaS->IaaS**层，根因分析功能，发现对应**主机和容器指标**的异常

SaaS/PaaS/IaaS三层分析

# 2.9 2020年改进升级-SaaS/PaaS/IaaS三层关联



实例重启

● **问题：**服务问题经常由某个实例导致，研发运维需要在天眼定位问题后再去天宫平台重启实例

● **举措：**2020年7月，调用链实现了打穿**CMDB、作业平台**，可直接在调用链平台控制实例重启，**缩短了故障恢复时间**



# 2.10 2020年改进升级-移动端



告警墙



根因定位



服务趋势

● **问题**：运维人员无法24小时使用电脑进行故障定位与处理

● **举措**：2020年8月，调用链将核心监控能力研发到移动端，实现移动运维，支持使用人员**掌上运维**，一键重启实例更方便



# 2.12 2020年改进升级-调用链清单扩展异常清单

服务超时定位清单
服务异常定位清单

## 调用链清单

时间 频率 最近3小时
服务名称 dd\_c\_produce
txid 全部
调用情况 全部

服务名	业务时间	接口名	耗时(ms)	状态	实例id	txid
dd_c_produce	2021-07-12 22:33:07	cardDateSync	18385	成功	-bfd8cffd4-s9g5q	-bfd8cffd4-s9g5q^16...
dd_c_produce	2021-07-12 21:47:58	mixPreSubmit	12548	成功	-bfd8cffd4-vjncq	-bfd8cffd4-vjncq^16...
dd_c_produce	2021-07-12 22:38:52	mixPreSubmit	11794	成功	-bfd8cffd4-vjncq	-bfd8cffd4-vjncq^16...
dd_c_produce	2021-07-12 20:34:11	mixPreSubmit	10474	成功	-bfd8cffd4-s9g5q	-bfd8cffd4-s9g5q^16...
dd_c_produce	2021-07-12 20:35:50	mixPreSubmit	10231	成功	-bfd8cffd4-vjncq	-bfd8cffd4-vjncq^16...
dd_c_produce	2021-07-12 20:49:25	mixPreSubmit	10155	成功	-bfd8cffd4-s9g5q	-bfd8cffd4-s9g5q^16...
dd_c_produce	2021-07-12 20:30:16	mixPreSubmit	10121	成功	-bfd8cffd4-vjncq	-bfd8cffd4-vjncq^16...
dd_c_produce	2021-07-12 21:01:40	mobPreSubmit5G	10093	成功	-bfd8cffd4-vjncq	-bfd8cffd4-vjncq^16...
dd_c_produce	2021-07-12 21:44:11	mixPreSubmit	10087	成功	-bfd8cffd4-vjncq	-bfd8cffd4-vjncq^16...
dd_c_produce	2021-07-12 21:58:41	mixPreSubmit	10047	成功	-bfd8cffd4-vjncq	-bfd8cffd4-vjncq^16...

共 4747 条
< 1 2 3 4 5 >

序列	层级	服务名	实例id
0	1	dd_c_produce	-bfd8cffd4-vjncq com.chinaunicom.c...
1	2	dd_c_produce	-bfd8cffd4-vjncq ↓com.chinaunicom.c...
2	3	REDIS	-bfd8cffd4-vjncq ↓redis.clients.j...
3	2	dd_c_produce	-bfd8cffd4-vjncq ↓org.apache.cor...
4	2	prodmporder_mpo...	-bfd8cffd4-vjncq ↓com.mysql.jdb...
5	2	prodmporder_mpo...	-bfd8cffd4-vjncq ↓com.mysql.jdb...
6	2	dd_c_produce	-bfd8cffd4-vjncq ↓com.chinaunicom.c...
7	3	dd_c_produce	-bfd8cffd4-vjncq ↓com.chinauni...

### 调用链清单

服务名	业务时间	方法名	耗时(ms)	异常信息	状态	实例id	txid	错误信息	错误类型	数量
dd_c_produce	2021-07-12 21:28:54	com.chinaunicom.cb...	9010	调用提交失败...	失败	-bfd8cffd4-v...	-bfd8cffd4-v...	调用资源中心业务异常.W205:ICCL...	业务异常	42
dd_c_produce	2021-07-12 21:28:54	com.chinaunicom.cb...	9010	调用提交失败...	失败	-bfd8cffd4-v...	-bfd8cffd4-v...	调用资源中心业务异常.W010:库位...	业务异常	12
dd_c_produce	2021-07-12 21:28:54	com.chinaunicom.cb...	9010	调用提交失败...	失败	-bfd8cffd4-v...	-bfd8cffd4-v...	卡数据查询03:库位不对	业务异常	8
dd_c_produce	2021-07-12 20:35:37	com.chinaunicom.cb...	4393	调用提交失败...	失败	-bfd8cffd4-s...	-bfd8cffd4-s...	调用资源中心业务异常.W009:卡状...	业务异常	6
dd_c_produce	2021-07-12 21:57:28	com.chinaunicom.cb...	3727	sUniTrade接口...	失败	-bfd8cffd4-s...	-bfd8cffd4-s...	获取产品或者活动信息失效时间...	业务异常	3

共 104 条
< 1 2 3 4 5 ... 11 >

序列	层级	服务名	实例id	方法名	耗时(ms)	异常信息	状态
0	1	dd_c_produce	-bfd8cffd4-vjncq	com.chinaunicom.cbss2.ai.ordercenter.self.service.impl.OrderAopSubmitServiceImpl.mobPreSubmit(java...	9009	调用提交失败! s...	失败
1	2	dd_c_produce	-bfd8cffd4-vjncq	↓com.chinaunicom.cbss2.ai.ordercenter.query.com.service.impl.ParamCommQueryServiceImpl.com...	1		成功
2	3	REDIS	-bfd8cffd4-vjncq	↓redis.clients.jedis.Jedis.get(java.lang.String key)	1		成功
3	2	dd_c_produce	-bfd8cffd4-vjncq	↓org.apache.commons.dbcp.BasicDataSource.getConnection()	0		成功
4	2	prodmporder_mpo...	-bfd8cffd4-vjncq	↓com.mysql.jdbc.ConnectionImpl.prepareStatement(java.lang.String sql)	0		成功
5	2	prodmporder_mpo...	-bfd8cffd4-vjncq	↓com.mysql.jdbc.PreparedStatement.executeQuery()	1		成功
6	2	dd_c_produce	-bfd8cffd4-vjncq	↓com.chinaunicom.cbss2.ai.ordercenter.business.proxy.DefaultMicroServiceProxy.callService(com.c...	9007	调用提交失败! s...	失败
7	3	dd_c_produce	-bfd8cffd4-vjncq	↓com.chinaunicom.cbss2.ai.ordercenter.query.com.service.impl.ParamConfigServiceImpl.getParam...	1		成功

● **问题**：原生清单不支持按异常纬度查询，实际定位故障时研发运维需要通过异常直接查询故障链路

● **举措**：2020年11月，调用链支持每一笔请求的清单级展现，每个**方法执行**、**sql调用**一目了然，并提供专属**异常调用清单**分析，支持跨数据中心、云平台、系统拓扑

## 异常清单

GOPS 全球运维大会2021·上海站

# 2.13 2020年改进升级-主机实例监控

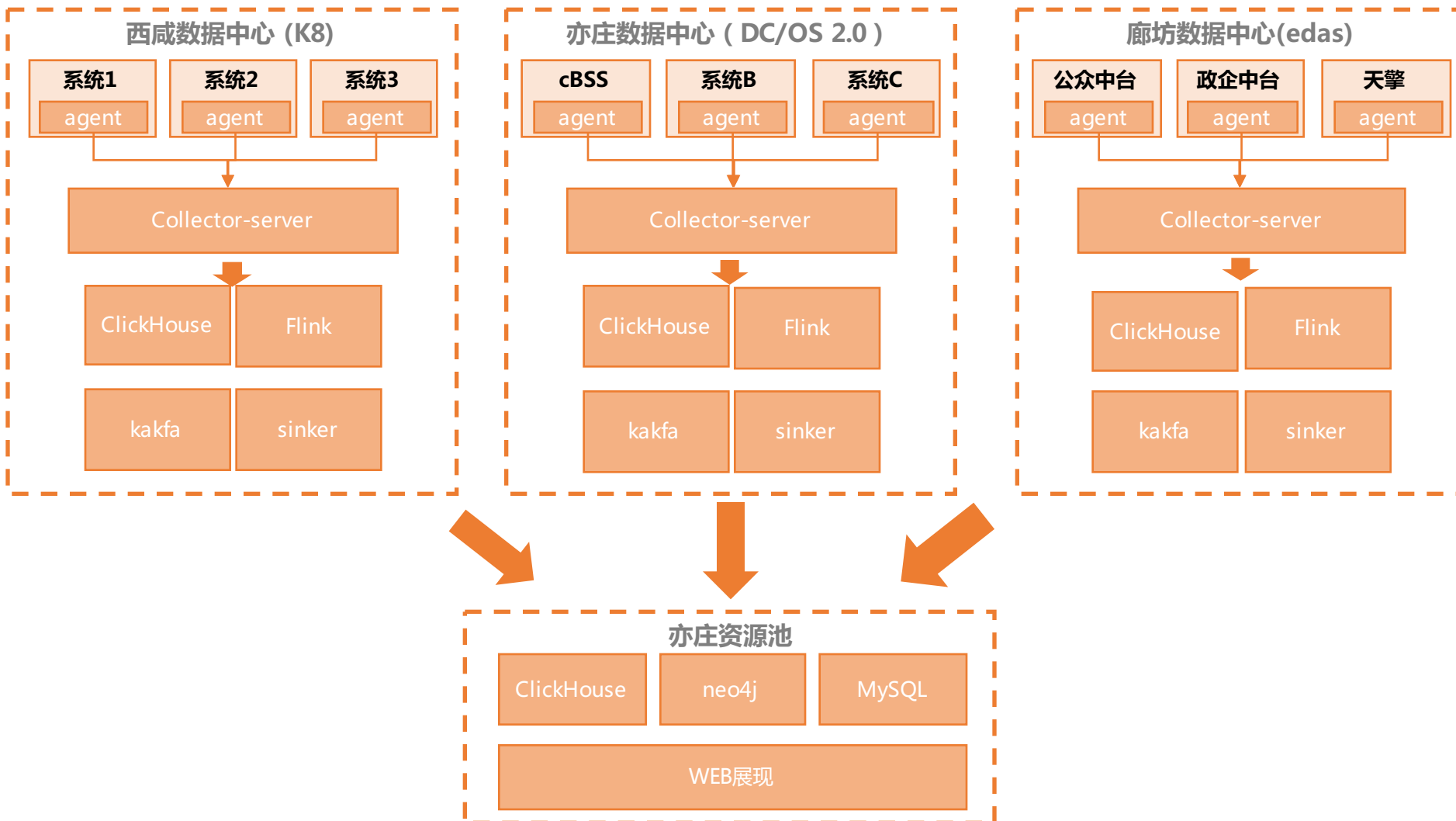


主机实例监控

● **问题**：当从服务实例定位到主机问题时，缺少主机影响实例范围，常见故障是主机CPU高，但是不知道哪个高

● **举措**：2020年12月，调用链通过整合CMDB，实现了**主机实例监控**，快速发现主机上是否存在占用资源过高容器，以及**影响的服务实例范围**

# 2.14 2021年生态整合-第二次技术架构升级



● **问题**：2021年2-6月，随着**公众中台、政企中台、cBSS、新客服系统**接入，多个系统之间的调用无法串联，数据存储在不同**数据中心**，不同系统使用k8s、mesos、edas**不同云平台**，存在数据采集不到情况

● **举措**：天眼对开源agent进行了改造支持了edas平台，对跨数据中心、跨系统调用情况进行了数据串联，打通了跨系统拓扑与清单

# 2.15 2021年生态整合-跨数据中心、跨平台、跨系统串联

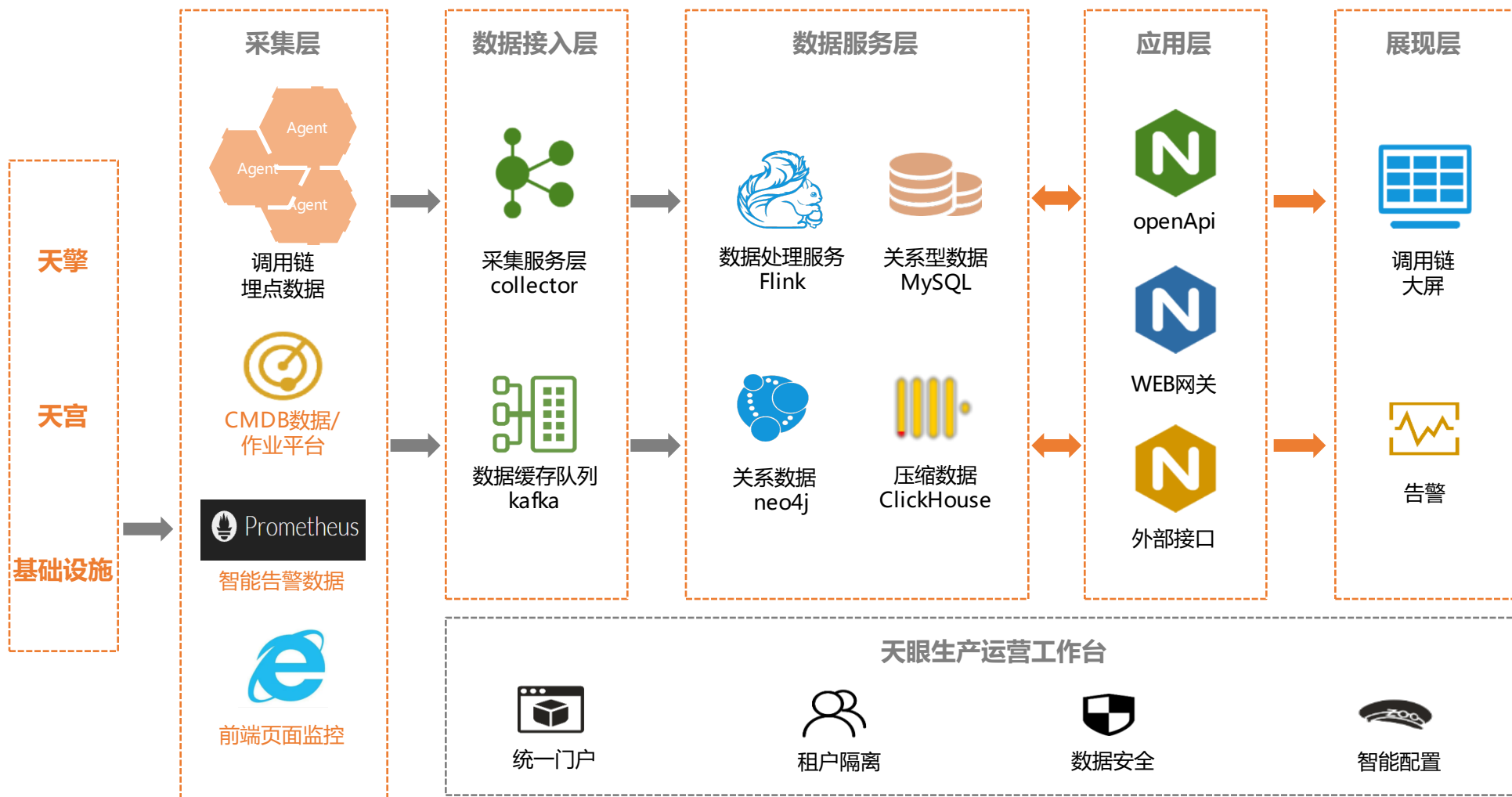


- **问题**：原有采集器不支持政企、公众平台、不支持跨系统、不支持跨数据中心的串联，导致采集不到，数据串联不上多种问题

- **举措**：2021年4-5月，调用链通过技术改造实现了**跨系统**的调用串联、**跨数据中心**的调用串联及**跨平台**的采集适配

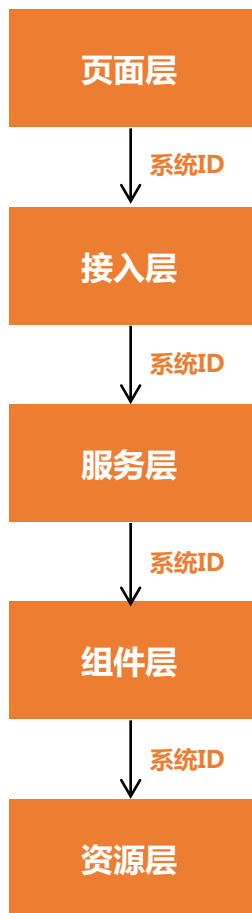
跨系统拓扑

# 2.16 2021年生态整合-第三次技术架构升级



- **问题:** 各个层次指标数据分散, 各管一块, 不能发挥更大价值
- **举措:** 2021年1月, 数据入湖。把所有层次监控数据进行整合, 将Prometheus (共享、天宫、天擎) 和 ClickHouse、CMDB数据等资源拉通, 含 (页面层、接入层、服务层、组件层、资源层)

# 2.17 2021年生态整合-全层级端到端监控



全层级根因定位

● **问题**：一个系统有多少个告警说不清，监控分散，故障影响范围说不清

● **举措**：2021年5月，调用链通过**系统ID**实现页面层、接入层、服务层、组件层、资源层关联，实现**端到端监控**，一屏看清当前系统所有问题



# 2.18 2021年生态整合-从人工定位到一键故障诊断



● **问题**：故障定位时主要通过使用人员一步一步定位，因每个人的能力不一样定位问题的效率不一致

● **举措**：2021年6，7月，调用链通过对**五层数据关联**，进行**系统自动诊断**，实现傻瓜式故障定位，大大**降低定位时间**

故障拓扑

故障诊断

PART 3

# 关于链路追踪演进的思考

# 3.1 2022生态演进-第四次技术架构升级

针对OpenTracing体系下的分布式链路追踪与日志进行整合，在系统监控的基础上强化业务监控能力



调用链、日志、性能管理

链路监控

生产压测

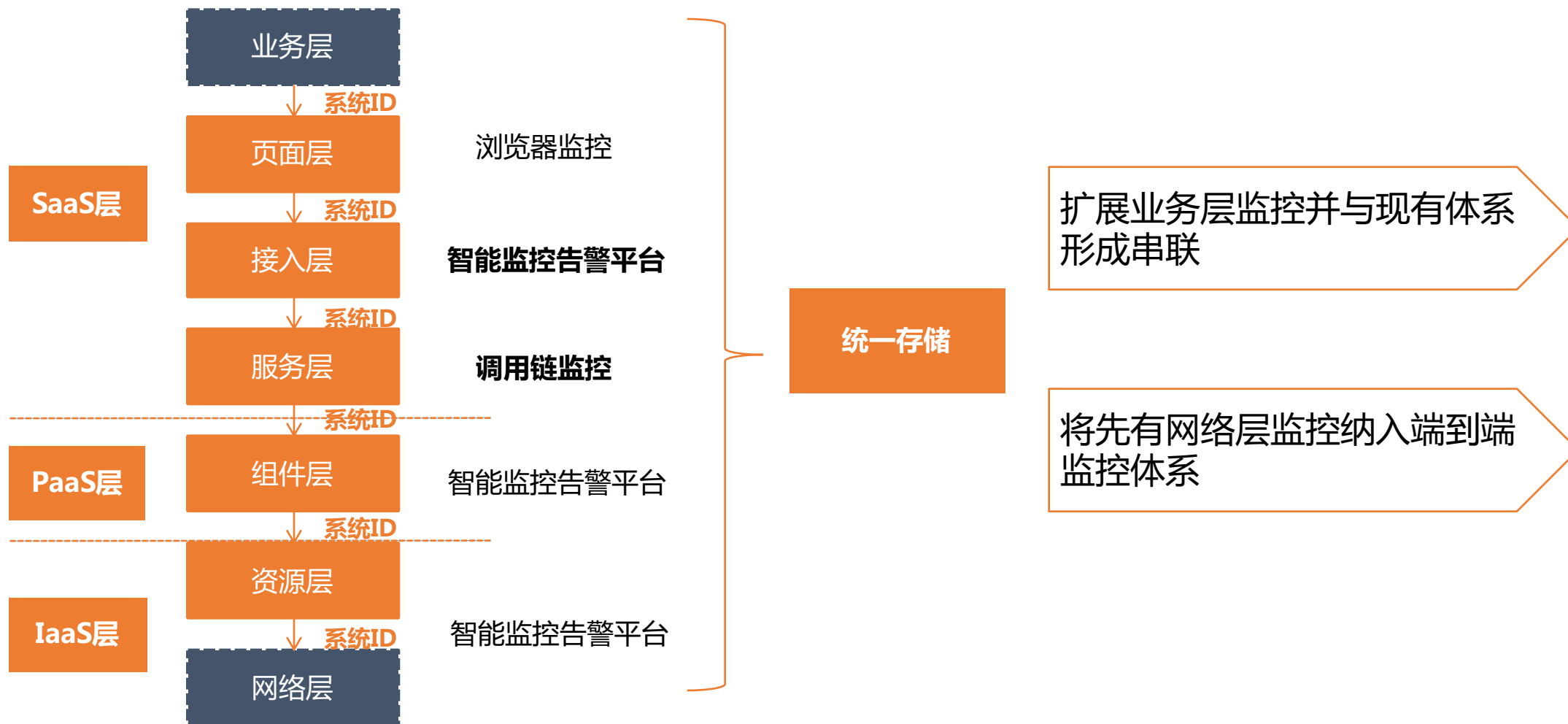
消息日志

流量回放

One Agent

# 3.2 2022生态演进-全层级数据再整合

强化**一键诊断**能力，向上串联业务评估故障影响，向下关联网络进行综合诊断





# Thanks

高效运维社区  
开放运维联盟

荣誉出品