



GIAC

全球互联网架构大会

GLOBAL INTERNET ARCHITECTURE CONFERENCE

AIOps - 智能故障处理中的系统工程

王亚雷 云兴维智（北京）科技有限公司 CEO

100

第七届 全球软件案例研究峰会

运营增长

组织发展

爆款架构

人工智能

测试实践

运维体系

产品创新

研发效能

数据平台

AI/数据驱动

测试体系

AI/ops DevOps

体验设计

团队管理

工程实践

AI实践

大前端

区块链



扫描二维码
查看最值得学习的100+创新案例

2018年11月30日-12月3日
北京 | 国家会议中心

主办方 **msup**

上海精品公开课

管理3.0认证课程

时间：12月22-23日 | 地点：上海 | 讲师：林伟丹



大数据及AI挖掘技术

时间：12月22-23日 | 地点：上海 | 讲师：风清扬



高可用架构与设计

时间：01月12-13日 | 地点：上海 | 讲师：沈老师



K8S与service mesh

时间：01月12-13日 | 地点：上海 | 讲师：Jim



备注：扫码查看课程详情，两人以上报名有优惠，详情咨询：15802217295

msup[®]



AIOps 专业厂商面临的挑战

客户运维对象多样性、复杂性

客户运维数据的质量和不可控性

平台算法的可选性、可调性、可演进性 > 算法的先进性

完整的解决方案



AIOps建设三要素

AIOps是一个体系建设

应用
场景

数据
治理

算法和
平台

应用场景

深耕IT运维和
大数据的运维
团队

数据治理

多源、异构数
据的整合关联
和生命周期管
理

算法

多种优化算法及框架技
术、高性能计算平台



AIOps 建设路径

机器学习 / AI主导的
高效排障流程

算法驱动的成本
和性能优化

智能的资源预测和计划：
Capacity Forecasting
and Planning

全域运维数据服务：存、取、查询、更新

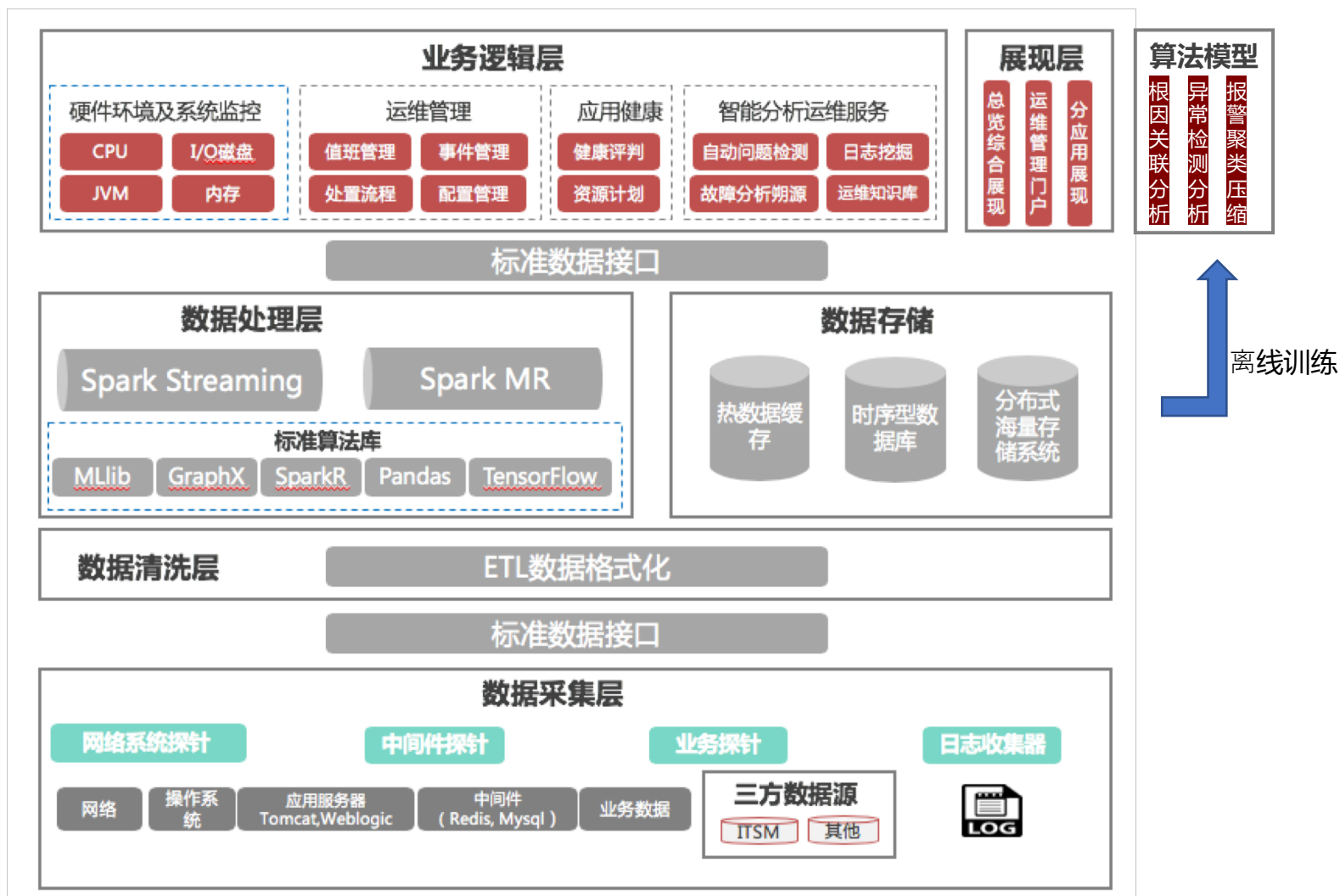
基于机器学习的运维知识图谱

统一的数据管理和治理：主题域、数据Schema

运维大数据平台：存储、处理、计算



AIOps 建设路径



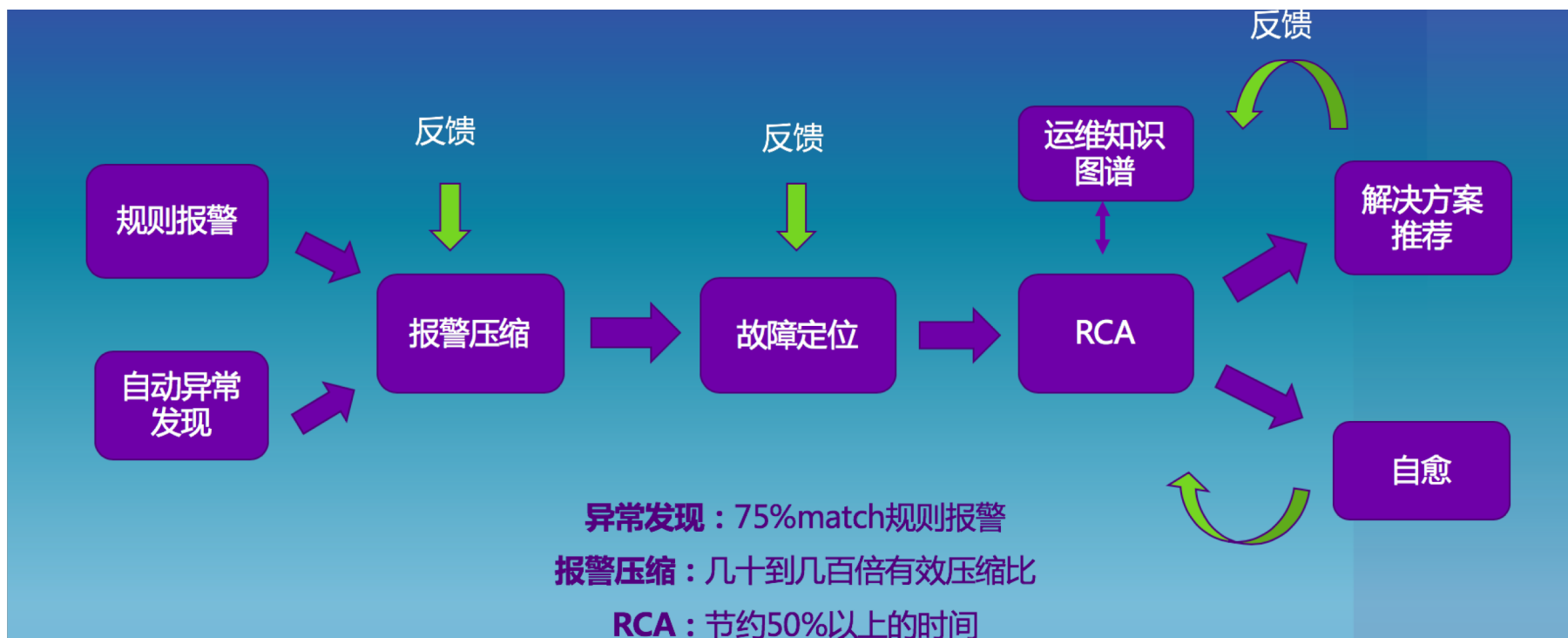


AIOps数据建设 – 运维数据主题模型





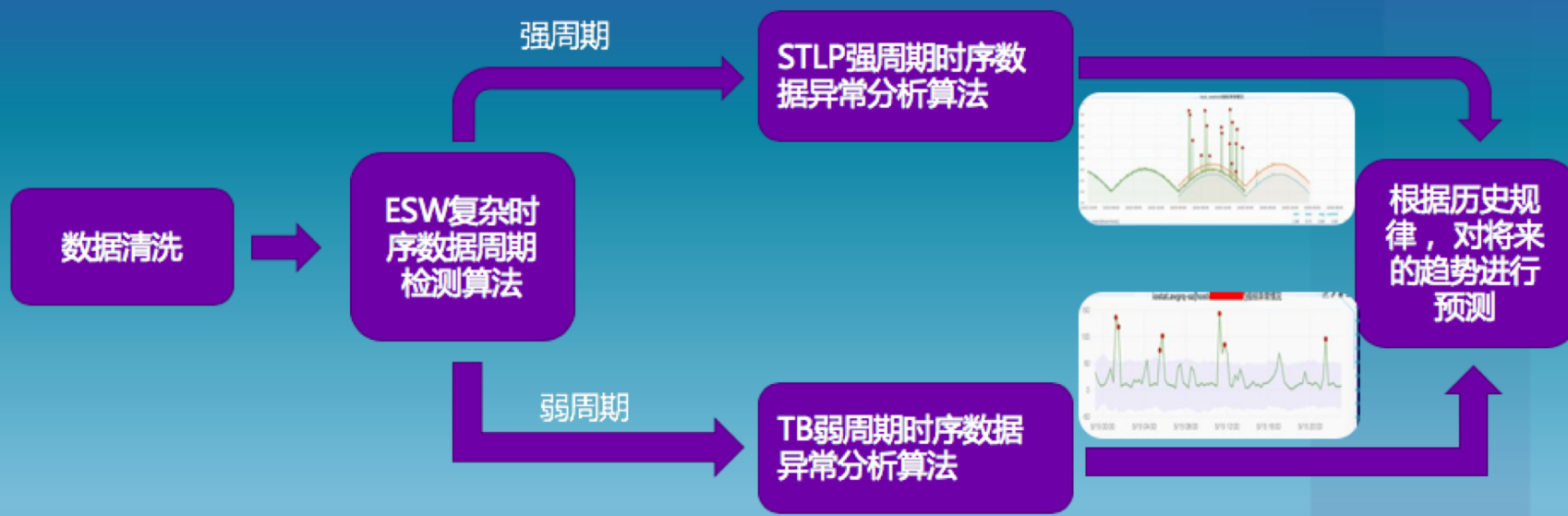
AIOps场景建设 - 高效排障流程





AIOps场景建设 - 异常检测

自动异常检测是通过查询历史数据计算未来时间的历史趋势并对比判断指标是否异常。
其中查询历史数据后依次需要进行数据清洗或补充, 使用机器学习算法分解趋势和季节性及噪音,
利用趋势曲线定义规律, 发现异常





AIOps场景建设 - 异常检测





AIOps场景建设 - 异常检测

	报警项	bppm(条)	Aiops(条)	备注
同时报警	OS:File system /xx/xx used is xxx%	20	18	
	OS:memory Page in over 293.19	4	3	
	OS:memory Page out over 1679.23	1	2	
	Sessionsused	0	6	人工反馈进行确认
	Phyreads	0	1	人工反馈进行确认
	Blkreads	0	2	人工反馈进行确认
仅bppm报警	OS:磁盘 hdisk219 平均响应时间为 35ms	3	0	Transit / 周期性



AIOps场景建设 - 异常检测

算法对周期性变化的处理





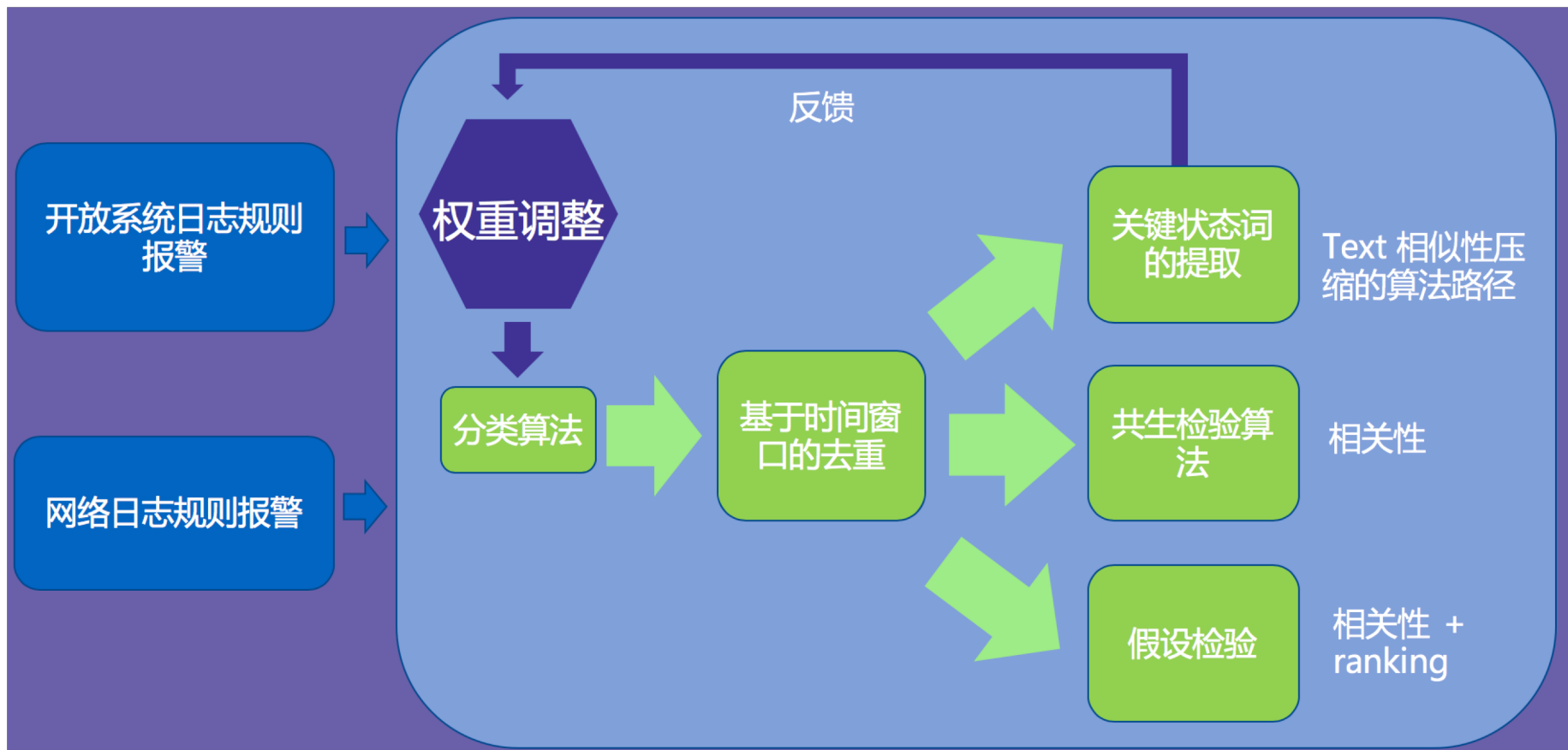
AIOps场景建设 - 报警压缩

某头部银行

1. 数据中心IT环境复杂度高
2. 日志种类多（网络告警日志，开放系统日志）
3. 网络告警日志20万条/天，开放系统日志3万条/天
4. 传统告警处置流程繁琐，依靠人力成本居多，效率低下
5. 无法发现关联告警



AIOps场景建设 - 报警压缩





AIOps场景建设 - 报警压缩

- 聚类算法对相似的报警进行压缩：编辑距离， 80% 压缩比
- 关键状态词的提取：编辑距离非常近，但不能进行压缩

Line protocol on the interface Serial x/x/x/xx:x is DOWN
Line protocol on the interface Serial x/x/x/xx:x is UP

Member[raw=Caused by: java.net.ConnectException: Connection timed out]
Member[raw=Caused by: java.net.ConnectException: Connection refused]

- Ngram + FT-Tree
 Normalize (归一化)



AIOps场景建设 - 报警压缩

提升准确率，依赖用户反馈

[CREATED, DELETED, OUT]

[DISCARDING, LEARNING]

[IN, OUT]

[存入, 消费]

[DOWN, UP, YQB_ECN_ACC_UNI]

[FLAP, UP]

[星期一, 星期二, 星期日]

[取款, 存入]

[INFO, WARN]

[CORRECT, WITHOUT]

[COMMAND-LOG, EXIT]

[32654, 39377, 49982]

[星期一, 星期二, 星期日]

[124, SERVER, 闪断]



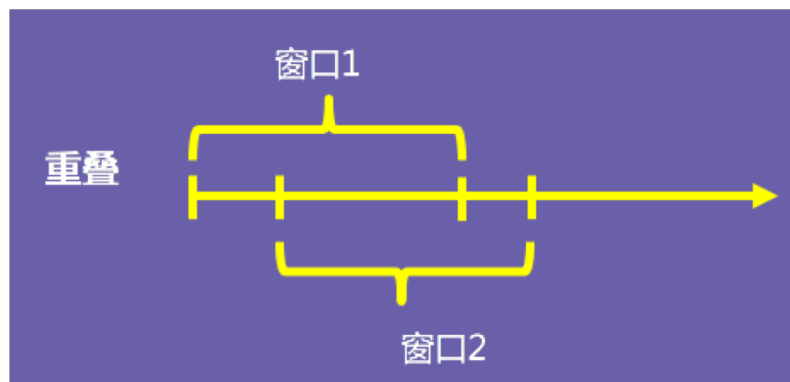
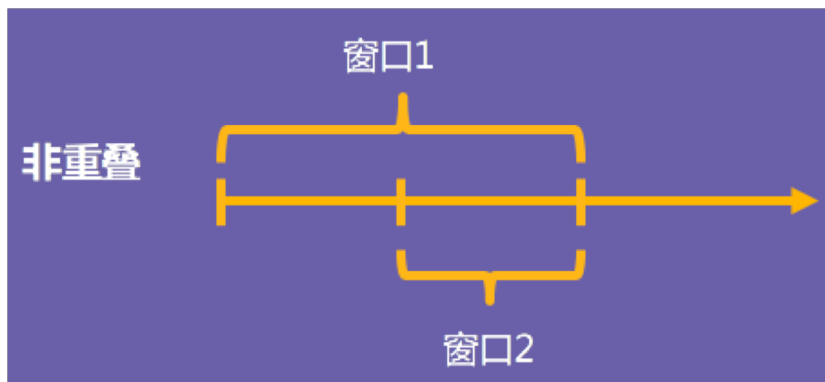
AIOps场景建设 - 报警压缩 - 相关性和根因

- 因为我们基于报警之间在同一时间窗口内的共生关系来进行报警压缩，所以需要时间轴划分为若干窗口。
- 我们根据经验和数据观察暂时将窗口长度定为十分钟。
- 划分的方式有两种重叠和非重叠两种形式。

非重叠形式为窗口长度和窗口移动步长相等。

重叠形式为移动步长小于窗口长度。

重叠形式相当于对于非重叠形式的抽样。





AIOps场景建设 - 报警压缩 - 相关性和根因

假设检验推断相关性

- 采用假设检验进行相关性筛选。先假定两个独立报警之间不相关，然后依据不相关的假定，计算出它们之间实际共生发生的概率，如果概率过低说明假设不成立，得出两者相关的推断。
- 设定概率阈值，筛选出相关的独立报警两两组合。通过报警之间的两两关联，我们可以构造出一个概率图。大部分独立报警都不在概率图中，也就是和任何其他报警都无时间上的相关关系。对于这样的报警我们不做压缩。在概率图中的独立报警种类数目虽然只占10%左右，但是在报警总数上占70%左右，所以在概率图中的报警有较大的压缩空间。
- 假定不相关的条件下的概率计算公式如下：

$$P_{N_c > n_c} = \sum_{k=n_c}^{k=\min(n_a, n_b)} \frac{\binom{N}{k} \binom{N-k}{n_a-k} \binom{N-n_a}{n_b-k}}{\binom{N}{n_a} \binom{N}{n_b}}$$

$$\binom{a}{b} = \frac{a!}{b! (a-b)!}$$



AIOps场景建设 - 报警压缩 - 相关性和根因

假设检验推断相关性

- 总共有1000个时间窗口的条件下，我们依照不同的逻辑假设计算了对应的不相关概率。
- 从下面的计算结果可以看出，虽然我们的假设检验是针对两两报警组合的检验，但是对于更复杂的报警组合关系，依然有很好的效果。

1. $A \perp B$

总窗口数 $N = 1000$

报警A发生次数 $N_A = 100$

报警B发生次数 $N_B = 100$

A和B共生次数的期望值 $E(N_C) = 10$

$P_{N_c > n_c} \approx 0.555 > 0.01$ 无法拒绝A和B
无关的假设

2. $A \rightarrow B$

总窗口数 $N = 1000$

报警A发生次数 $N_A = 100$

报警B发生次数 $N_B = 100$

A和B共生次数 $N_C = 100$

$P_{N_c > n_c} \approx 1.57 * 10^{-140} < 0.01$ 拒绝A和B无关
的假设





AIOps场景建设 - 报警压缩 - 相关性和根因

假设检验推断相关性

3. $A_1 \wedge A_2 \rightarrow B, A_1 \perp A_2$

总窗口数 $N = 1000$

报警 A_1 发生次数 $N_{A_1} = 100$

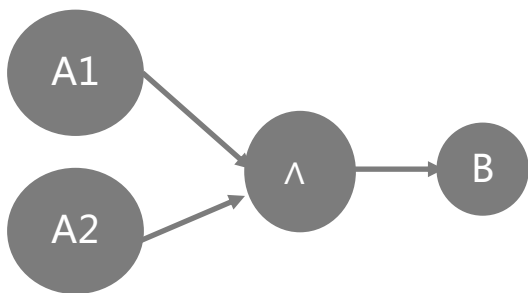
报警 A_2 发生次数 $N_{A_2} = 100$

A_1 和 A_2 共同发生的次数的期望 $E(N_{A_1 \wedge A_2}) = 10$

报警 B 的发生次数 $N_B = 10$

A_1 和 B 的共同发生次数 $E(N_C) = 10$

$P_{N_c > n_c} \approx 6.57 * 10^{-11} < 0.01$ 拒绝 A_1 和 B 无关的假设



4. $A_1 \wedge A_2 \rightarrow B, A_1 \perp A_2$

总窗口数 $N = 1000$

报警 A_1 发生次数 $N_{A_1} = 100$

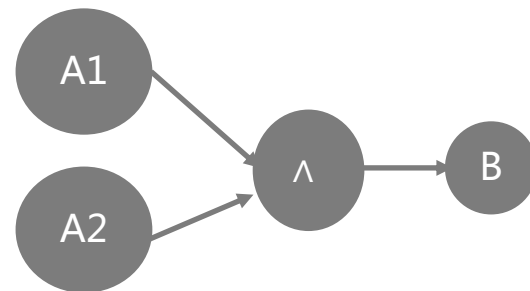
报警 A_2 发生次数 $N_{A_2} = 100$

A_1 和 A_2 共同发生的次数的期望 $E(N_{A_1 \wedge A_2}) = 10$

报警 B 的发生次数 $N_B = 190$

A_1 和 B 的共同发生次数 $E(N_C) = 100$

$P_{N_c > n_c} \approx 1.09 * 10^{-84} < 0.01$ 拒绝 A_1 和 B 无关的假设





AIOps场景建设 - 报警压缩 - 相关性和根因

报警ranking

- 在报警压缩的时候，如果同窗口出现的一组独立报警之间存在很大的相关性，我们需要决定压缩哪一个报警，留下另一个呈现给用户。
- 我们首先采用对于独立报警进行因果排序。在压缩报警的时候，我们会压缩排名较低的报警（偏向于果的一侧），保留排名较高的报警（偏向于因的一侧）。
- 我们采用构造马尔科夫矩阵的方法来对于独立报警进行因果排序，具体如下：

1 定义报警两两之间的影响因子

在固定时间窗口下，两个报警之间相互影响因子相等。时间窗口总数为 N ，报警A发生在 n_a 个窗口中，报警B发生在 n_b 个窗口中，两者共同发生的窗口个数为 n_c 。

2 构造影响因子矩阵

影响因子矩阵为一个 $N*N$ 的对称正定矩阵， N 为在前一部分的概率图中的独立报警数目。若报警A和B之间存在相关关系，则A行B列的元素为 F_{ab} ，B行A列的元素为 F_{ba} 。如果A和B之间无相关关系，则对应行列元素为0。

$$AB \text{ 影响因子 } F_{AB} \stackrel{\text{def}}{=} \frac{A \text{ 和 } B \text{ 共同发生的概率}}{A \text{ 发生的边缘概率} \times B \text{ 发生的边缘概率}} = \frac{P(AB)}{P(A)P(B)}$$

$$\begin{bmatrix} F_{11} & F_{12} & F_{13} & \cdots \\ F_{21} & F_{22} & F_{23} & \cdots \\ F_{31} & F_{32} & F_{33} & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix}$$



AIOps场景建设 - 报警压缩 - 相关性和根因

报警ranking

3 从影响因子矩阵到马尔科夫矩阵

我们将第二部的影响因子矩阵的每一列除上每一列的和，是矩阵的每一列的和都为1，就得到了一个马尔科夫矩阵。

5 马尔科夫矩阵的稳定状态

设上一步得到的马尔科夫矩阵为A。马尔科夫矩阵存在一个性质，对于任意一个非负初始向量 u_0 。马尔科夫矩阵不断与该向量相乘，会收敛于一个稳定状态向量 u_s 。且 u_s 和 u_0 无关。也就是存在 u_s ，对于任意 u_0 ，都有

$$\lim_{k \rightarrow +\infty} A^k u_0 = u_s$$

4 利用稳定状态对于警报排序

因为我们构造的马尔科夫矩阵代表了警报之间的相互影响，所以稳定状态向量上的每一个元素代表了其对应警报在整个概率图中的因果影响力。在稳定状态数值越大的警报，我们认为其越接近果。。

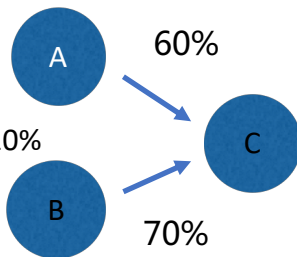


AIOps场景建设 - 报警压缩 - 相关性和根因

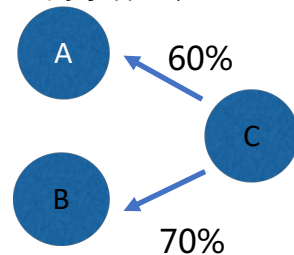
报警ranking

- 我们认为在我们的马尔科夫矩阵排序中，分数比较高的警报更可能是因果关系中的果，而分数比较低的警报更可能是因。
- 这里列出了两个迷你例子，在第一个例子中我们设定A和B是C发生的原因，在第二个例子中我们设定C是A和B的发生原因。
- 在第一个例子中，最为结果的C的分数最高，而在第二个结果中作为共因的C的分数最低，一定程度上支持了我们的推断。

- 假设A和B是C发生的原因
- A发生的时候在60%的情况下C发生
- B发生的时候在70%的情况下C发生
- A和B之间相互独立，各自发生的概率为10%
- 其对应的马尔科夫矩阵和稳态向量如下



- 假设C是A和B发生的原因，在给定C发生的条件下，A和B的发生独立。
- C发生的时候在60%的情况下A发生
- C发生的时候在70%的情况下B发生
- C发生的概率为10%
- 其对应的马尔科夫矩阵和稳态向量如下



马尔科夫矩阵

	A	B	C
A	0.634213	0.0603785	0.26087
B	0.0634213	0.603785	0.304348
C	0.302366	0.335836	0.434783

A	0.311629
B	0.327234
C	0.361138

马尔科夫矩阵

	A	B	C
A	0.454545	0.291667	0.333333
B	0.272727	0.416667	0.333333
C	0.272727	0.291667	0.333333

A	0.363216
B	0.339616
C	0.297168

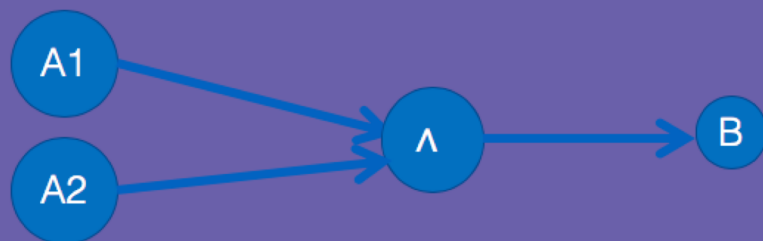


AIOps场景建设 - 报警压缩 - 相关性和根因

模型训练和预测

- 在报警组合筛选以及报警排序中，我们发掘的都是根据两两报警之间的共生关系。但是仅仅依靠两两共生关系无法满足我们的压缩的功能。
- 在下面的例子中，报警 A_1 和报警 A_2 共同发生时会导致报警B的发生，这也反应在列出的迷你数据集中， $P(B = 1|A_1 = 1, A_2 = 1) = 1$ 。
- 但是如果只考虑两两的共生关系，那么 $P(B = 1|A_1 = 1) = P(B = 1|A_2 = 1) = 0.4 < 0.5$ 。两个条件概率都不足以让我们预测出报警B的发生进而进行压缩。
- 因此我们需要引入预测模型来对多个报警之间的共生关系进行学习。

时间窗口index	A_1	A_2	B
0	1	0	0
1	1	0	0
2	1	0	0
3	1	1	1
4	1	1	1
5	0	1	0
6	0	1	0
7	0	1	0
test	1	1	?



$$P(B = 1|A_1 = 1) = 0.4 \quad P(B = 1|A_2 = 1) = 0.4$$

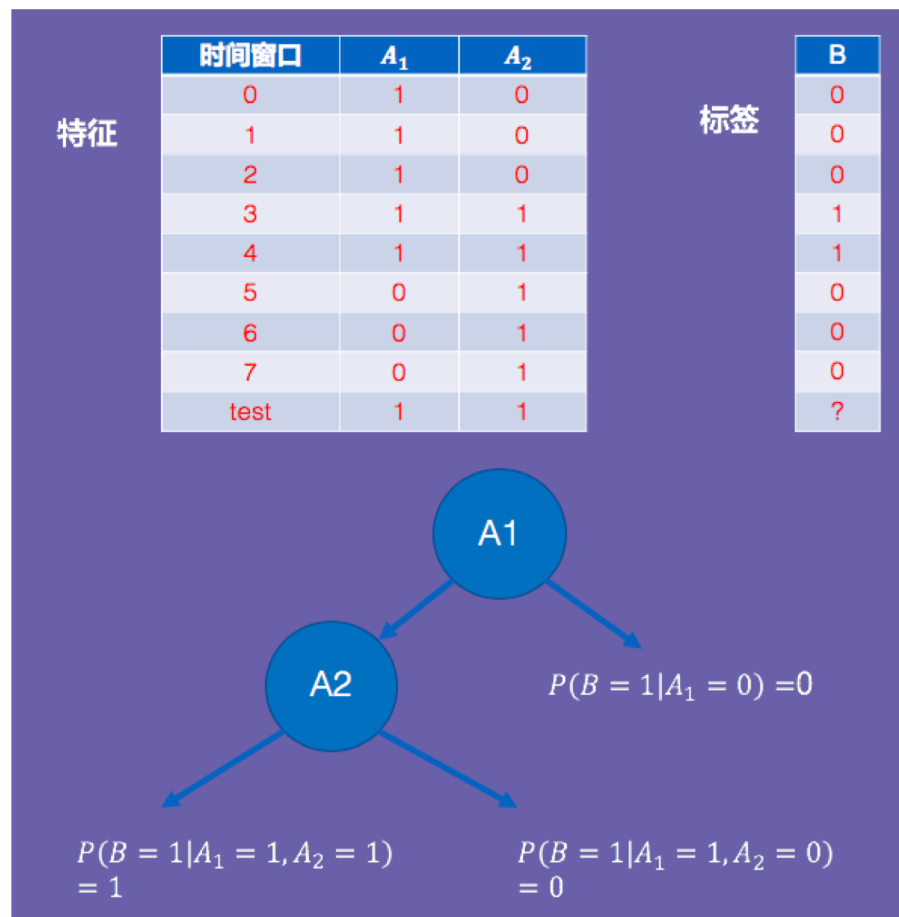
$$P(B = 1|A_1 = 1, A_2 = 1) = 1$$



AIOps场景建设 - 报警压缩 - 相关性和根因

模型训练和预测

- 我们从因果排序中排名最低的警报开始，注意训练判断该警报在某一时间窗口内会否发生的模型。模型的因变量为该警报的发生情况（1 为发生，0 为没发生）。其他报警的发生情况作为特征。只有符合以下两个条件的警报才有资格成为特征，一是和因变量警报两两相关（在概率图中是因变量警报的邻居），二是在因果排序中排名比因变量警报高。
- 因为所有项目以及标签都是01变量，很适合决策边界具有阶梯型（锯齿形）的学习器，比如决策树。但是由于数据集噪音较大，而决策树又很容易产生过拟合。我们选择以决策树为自学习器的随机森林模型来做训练预测。
- 从因果排名中的低到高，我们对于每一个警报都会训练一个预测模型。
- 在预测时，我们在当前窗口中从排名最低的警报开始预测，如果预测结果为不会发生，则不进行压缩，如果预测结果为会发生，则进行压缩，并将该警报归入其当前发生的邻居中排名最高的警报类下。





AIOps场景建设 - 报警压缩 - 相关性和根因

模型测试样例

报警总数: 63499 提取到有效报警总数: 7951 (根因报警数: 7337, 被压缩报警数: 614)



报警时间	报警内容
- 2018-08-02 03:03:36	"Interface GigabitEthernet1/0/23, changed state to down" ==> INTERFACE GIGABITETHERNET### CHANGED STATE TO DOWN
2018-08-02 03:03:53	"Outbound Speech Call Cleared or Failed, Disconnect Cause: 11 (Peer Address: 018518365075)" ==> OUTBOUND SPEECH CALL CLEARED OR FAILED DISCONNECT CAUSE PEER ADDRESS
2018-08-02 03:04:03	OSPF Link State Request Packet Retransmitted to 10.3.128.254 (10.197.132.229) ==> OSPF LINK STATE REQUEST PACKET RETRANSMITTED TO \$IP\$ \$IP\$
2018-08-02 03:04:54	hh3cfanfailure: Fan failure ==> HH#CFANFAILURE FAN FAILURE



AIOps场景建设 - 故障根因分析和RCA

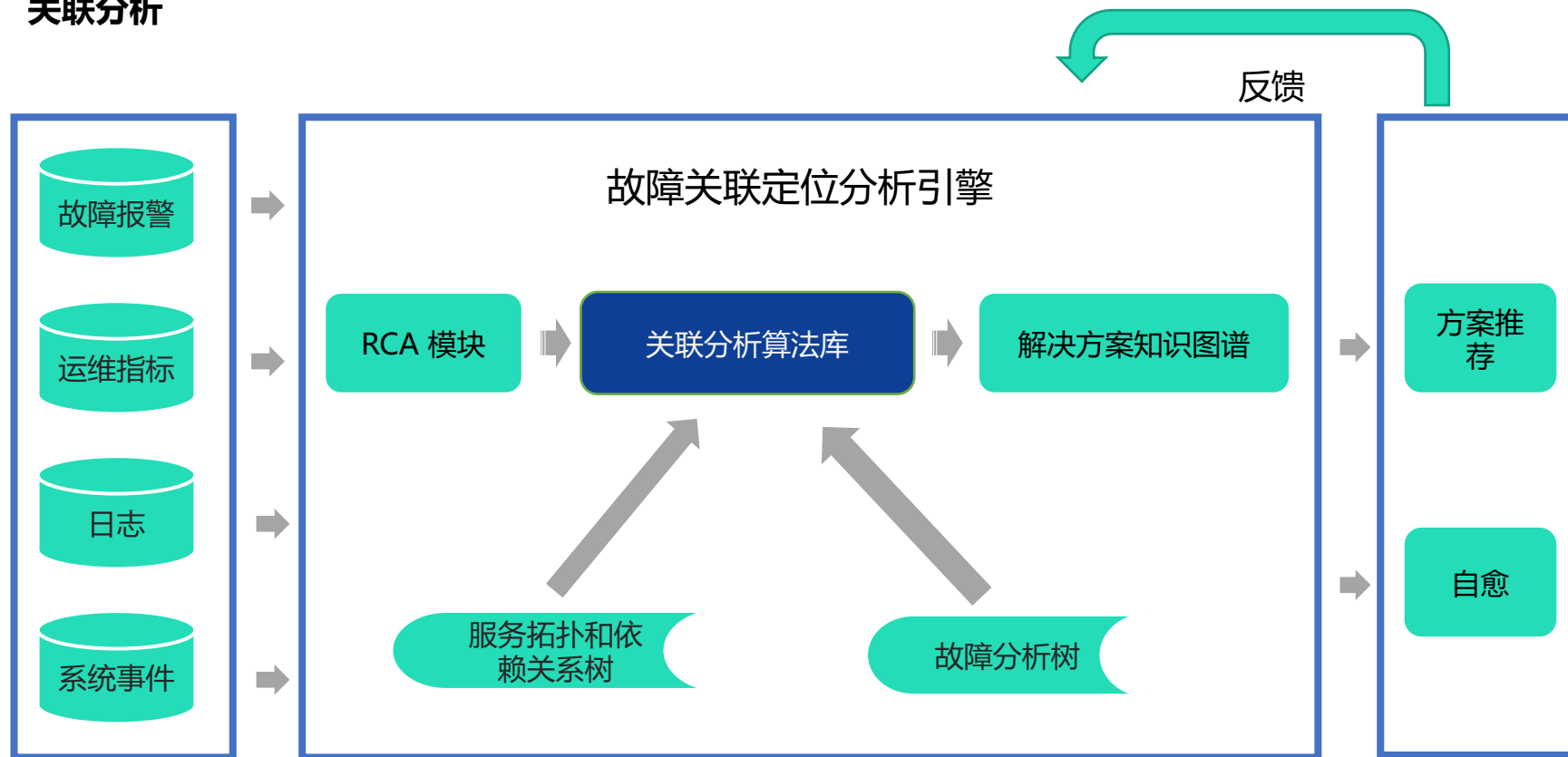
基于关联分析的根因定位





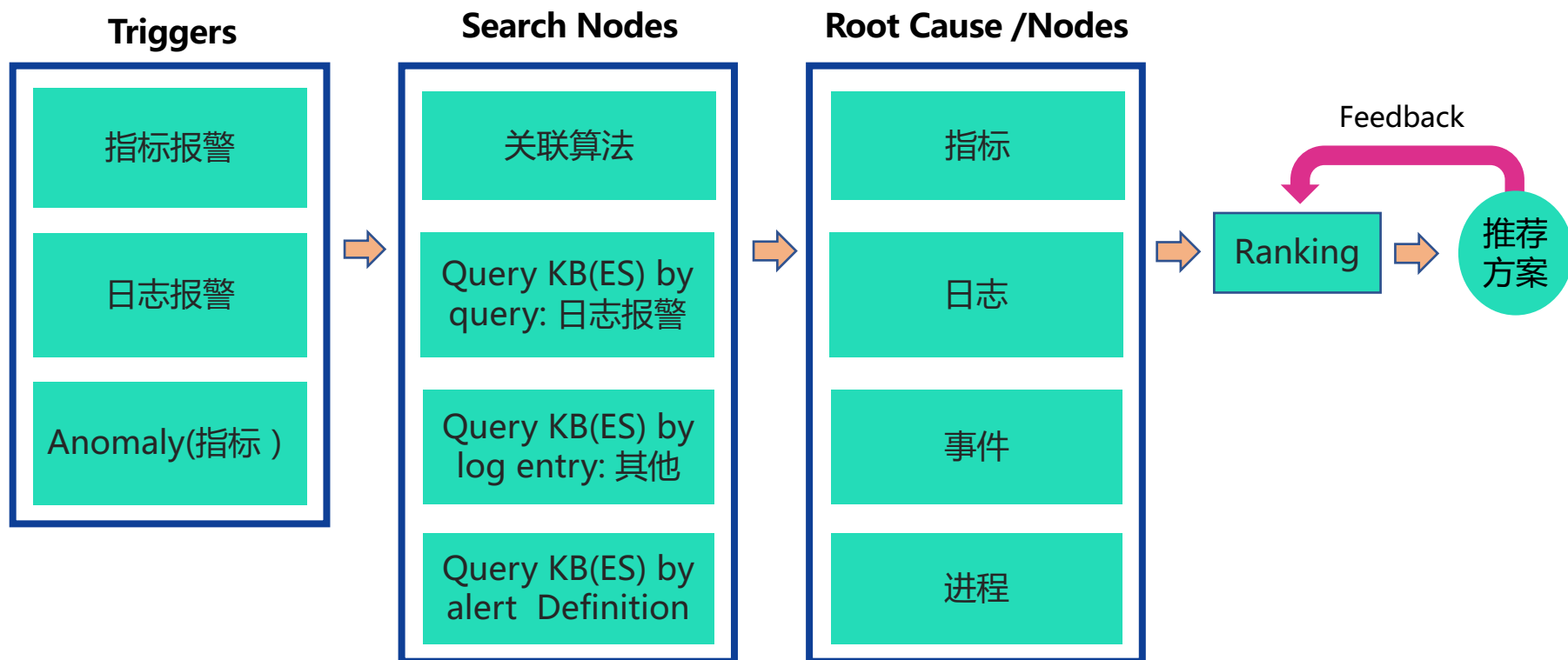
AIOps场景建设 - 故障根因分析和RCA

关联分析



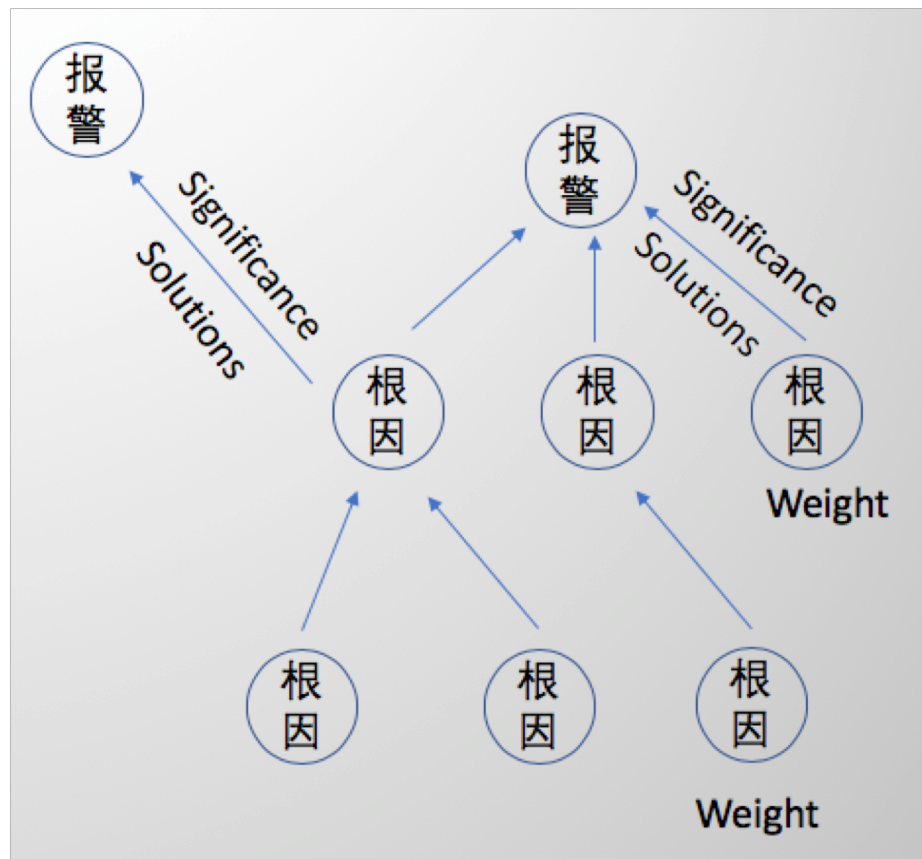


AIOps场景建设 - 故障根因分析和RCA





AIOps场景建设 - 故障根因分析和RCA





AIOps场景建设 - 故障根因分析和RCA

The screenshot shows a web browser window with the URL 'lerts/status'. The page is titled 'Cloudwiz' and has a sidebar with navigation options: '实时监控' (Real-time Monitoring), '指标' (Metrics), '日志' (Logs), '机器' (Machines), and '服务' (Services). The main content area is titled '报警指标状态' (Alert Metrics Status) and contains a search bar and a table of alerts.

报警名称	报警内容	主机名	报警值/阈值	报警级别	报警时间	处理人员	处理方法	操作
磁盘使用程度严重, 有大量的io	iostat.percentag_util	bigdata65	97.60 / 60	严重	2018-08-13 01:20:03	自动关闭		诊断分析
磁盘使用程度严重, 有大量的io	iostat.percentag_util	bigdata53	52.24 / 50	警告	2018-08-13 01:15:00	自动关闭		诊断分析
磁盘使用程度严重, 有大量的io	iostat.percentag_util	bigdata53	61.79 / 50	警告	2018-08-13 01:38:19	自动关闭		诊断分析
磁盘使用程度严重, 有大量的io	iostat.percentag_util	bigdata53	89.09 / 50	警告	2018-08-13 01:53:19	自动关闭		诊断分析
磁盘使用程度严重, 有大量的io	iostat.percentag_util	bigdata53	74.75 / 60	严重	2018-08-13 02:06:31	自动关闭		诊断分析
磁盘使用程度严重, 有大量的io	iostat.percentag_util	bigdata53	99.95 / 60	严重	2018-08-13 02:15:19	自动关闭		诊断分析
磁盘使用程度严重, 有大量的io	iostat.percentag_util	bigdata54	86.34 / 60	严重	2018-08-13 01:08:25	自动关闭		诊断分析
磁盘使用程度严重, 有大量的io	iostat.percentag_util	bigdata54	82.10 / 50	警告	2018-08-13 02:04:20	自动关闭		诊断分析
磁盘使用程度严重, 有大量的io	iostat.percentag_util	bigdata55	80.65 / 60	严重	2018-08-13 01:11:08	自动关闭		诊断分析
磁盘使用程度严重, 有大量的io	iostat.percentag_util	bigdata55	62.10 / 60	严重	2018-08-13 03:00:19	自动关闭		诊断分析
磁盘使用程度严重, 有大量的io	iostat.percentag_util	bigdata56	87.35 / 60	严重	2018-08-13 01:22:32	自动关闭		诊断分析
磁盘使用程度严重, 有大量的io	iostat.percentag_util	bigdata57	100.0 / 60	严重	2018-08-13 01:13:06	自动关闭		诊断分析
磁盘使用程度严重, 有大量的io	iostat.percentag_util	bigdata57	99.04 / 60	严重	2018-08-13 03:26:31	自动关闭		诊断分析
磁盘使用程度严重, 有大量的io	iostat.percentag_util	bigdata58	74.25 / 50	警告	2018-08-13 02:39:19	自动关闭		诊断分析
磁盘使用程度严重, 有大量的io	iostat.percentag_util	bigdata58	70.45 / 60	严重	2018-08-13 02:49:19	自动关闭		诊断分析
磁盘使用程度严重, 有大量的io	iostat.percentag_util	bigdata60	100 / 50	警告	2018-08-13 01:05:55	自动关闭		诊断分析
磁盘使用程度严重, 有大量的io	iostat.percentag_util	bigdata60	54.39 / 50	警告	2018-08-13 03:25:31	自动关闭		诊断分析
磁盘使用程度严重, 有大量的io	iostat.percentag_util	bigdata61	92.84 / 60	严重	2018-08-13 01:11:45	自动关闭		诊断分析
磁盘使用程度严重, 有大量的io	iostat.percentag_util	bigdata62	56.54 / 50	警告	2018-08-13 03:44:19	自动关闭		诊断分析



AIOps场景建设 - 故障根因分析和RCA

The screenshot displays a web-based AIOps dashboard for monitoring and analyzing system anomalies. The interface is in Chinese and includes a sidebar with navigation options like '实时监控' (Real-time Monitoring), '实时报警' (Real-time Alerts), '分析诊断' (Analysis and Diagnosis), '系统配置' (System Configuration), and '运维知识库' (Operations Knowledge Base).

The main content area shows an alert titled 'hbase master 服务异常' (hbase master service anomaly). The alert details include the time (2018-11-13 00:02:53), the content ('hbase.master.alive'), and the status ('严重' - Severe). A '暂缓报警' (Pause Alert) button is visible.

Below the alert, there is a horizontal navigation bar with several options, each with a corresponding icon and a status indicator (green smiley face for good, orange sad face for bad). The '上下游关联' (Upstream and Downstream Association) option is highlighted with a blue box. Other options include '引发报警的高消耗进程' (High consumption processes causing alerts), '可能引起报警的事件' (Events that may cause alerts), '报警相关关联指标 (部分)' (Alert related indicators (partial)), '报警相关关联日志 (部分)' (Alert related logs (partial)), and '采纳方案' (Adopted solution).

The '上下游关联' section shows a diagram illustrating the relationship between services. It includes a checklist of items:

- 已扫描 hbase.master 服务的上下游 (Scanned upstream and downstream of hbase.master service)
- 上游 Zookeeper 服务正常工作, 未扫描出任何异常指标 (Upstream Zookeeper service is working normally, no abnormal indicators were scanned)
- 下游 Hbase.regionserver 服务正常工作, 未扫描出任何异常指标 (Downstream Hbase.regionserver service is working normally, no abnormal indicators were scanned)

The diagram shows a flow from the upstream service (Zookeeper) to the downstream service (Hbase.regionserver), with the hbase.master service in the middle. The status of each service is indicated by a green box with a green smiley face.



AIOps场景建设 - 故障根因分析和RCA

The screenshot displays the AIOps system interface. The top navigation bar includes tabs for '大数据智能运维管理系统' and a browser address bar showing '10.85.62.96/diagnose?type=now&source=alert&alertId=2.4.0df0e311-152e-4242-b6c7-5c5b0af0c389&host=bigdata53&ts=1542038573000'. The left sidebar contains navigation items: '实时监控', '实时报警', '分析诊断', '系统配置', and '运维知识库'. The main content area shows a warning titled 'hbase master 服务异常' with details: '报警时间: 2018-11-13 00:02:53', '报警内容: hbase.master.alive', '再次报警时间: 2018-11-13 17:38', '报警触发值: 1.00', and '报警描述: hbase master 服务可能不可用, 请立刻重启'. The right side shows '主机名: bigdata53' and '状态: 严重'. Below the warning, there are six cards: '上下游关联', '引发报警的高消耗进程' (highlighted), '可能引起报警的事件', '报警相关指标 (部分)', '报警相关日志 (部分)', and '采纳方案'. The '引发报警的高消耗进程' card shows a list of processes with columns: '进程 ID', 'CPU 使用率', '内存 使用率', '读 硬盘速度', and '命令'. The table lists several processes, including '/sbin/init', '/usr/java/jdk1.7.0_55/bin/java', 'crond', and 'hald-addon-input: Listening on /dev/input/event4 /dev/input/event1 /dev/input/event0'.

进程 ID	CPU 使用率	内存 使用率	读 硬盘速度	命令
1	0	0.01%	25 GBps	/sbin/init
25347	0	8.02%	3 GBps	/usr/java/jdk1.7.0_55/bin/java -cp /home/yarn/spark-1.2.0-bin-hadoop2.3/lib/mysql-connector-java-5.1.17.jar/home/yarn/...
1982	0	0.00%	2 GBps	crond
1677	0	0.00%	33 kBps	hald-addon-input: Listening on /dev/input/event4 /dev/input/event1 /dev/input/event0
1935	0	0.01%	2 GBps	/usr/libexec/postfix/master
116369	0	0.04%	1 GBps	/opt/cloudwiz-agent/altenv/bin/python /opt/cloudwiz-agent/altenv/bin/supervisord -c /opt/cloudwiz-agent/altenv/etc/super...
116371	0	0.01%	-	/bin/bash /opt/cloudwiz-agent/uagent/run-uagent.sh
116372	0	0.01%	-	/bin/bash /opt/cloudwiz-agent/agent/run



AIOps场景建设 - 故障根因分析和RCA

The screenshot displays the AIOps dashboard interface. The top navigation bar includes tabs for '大数据智能运维管理系统' and a specific URL: '10.85.62.96/10.85.62.96:5001/rca/'. The main content area shows an alert titled 'hbase master 服务异常' (hbase master service anomaly) with details: '报警时间: 2018-11-13 00:02:53', '报警内容: hbase.master.alive', '再次报警时间: 2018-11-13 17:38', and '报警触发值: 1.00'. The alert status is '严重' (Severe) and the host is 'bigdata53'. Below the alert, a diagnostic flow is shown with steps: '上下游关联' (Upstream/Downstream Association), '引发报警的高消耗进程' (High consumption process causing the alarm), '可能引起报警的事件' (Events that may cause the alarm), '报警相关指标 (部分)' (Some alarm-related metrics), '报警相关日志 (部分)' (Some alarm-related logs), and '采纳方案' (Adopted solution). The '可能引起报警的事件' step is highlighted with a blue border. Below this step, a list of events is shown, all marked as '已扫描' (Scanned) and '没有检测到任何可能导致报警的' (No events detected that may cause the alarm):

- 已扫描机器上所有事件
- 没有检测到任何可能导致报警的 服务启动/关闭 事件
- 没有检测到任何可能导致报警的 机器启动/关闭 事件
- 没有检测到任何可能导致报警的 内存溢出/保护 事件
- 没有任何可疑 编辑 命令行操作
- 没有任何可疑 终止进程 命令行操作
- 没有任何可疑 涉及 start / stop 命令行操作
- 没有任何可疑 删除 命令行操作

The bottom of the screen shows the Windows taskbar with the date '2018/11/13' and time '16:44'.



AIOps场景建设 - 故障根因分析和RCA

The screenshot displays a web-based monitoring dashboard for AIOps. The browser address bar shows the URL: `10.85.62.96/diagnose?type=now&source=alert&alertId=2.4.0df0e311-152e-4242-b6c7-5c5b0af0c389&host=bigdata53&ts=1542038573000`. The dashboard is titled "hbase master 服务异常" (hbase master service abnormal). It shows the alert time as 2018-11-13 00:02:53, the alert content as "hbase.master.alive", and the alert trigger value as 1.00. The host name is "bigdata53" and the status is "严重" (Severe). The alert description states: "hbase master 服务可能不可用，请立刻重启" (hbase master service may be unavailable, please restart immediately).

The dashboard includes a navigation menu on the left with options: 实时监控 (Real-time Monitoring), 实时报警 (Real-time Alerting), 分析诊断 (Analysis and Diagnosis), 系统配置 (System Configuration), and 运维知识库 (Operation and Maintenance Knowledge Base). The main content area features a series of cards for analysis: 上下游关联 (Upstream and Downstream Association), 引发报警的高消耗进程 (High consumption processes that triggered the alert), 可能引起报警的事件 (Events that may cause the alert), 报警相关指标 (部分) (Alert related metrics (part)), 报警相关日志 (部分) (Alert related logs (part)), and 采纳方案 (部分) (Adopted solutions (part)). The "报警相关指标 (部分)" card is highlighted with a blue border.

Below the cards, the "报警相关指标 (部分)" section shows a table of metrics. The table has three columns: 异常指标 (Abnormal metrics), 描述 (Description), and 健康值 (Health value). The table contains one row of data:

异常指标	描述	健康值
yarn.nodes.usedMemoryMB	The total amount of memory currently used on the node (in MB)	25

The table also shows a status of "发现 1 个当前机器的 智能告警" (Found 1 intelligent alert on the current machine). A note at the bottom states: "没有关联任何指标，请点击进入 分析模式 查看详情" (No associated metrics, please click to enter analysis mode to view details).



AIOps场景建设 - 故障根因分析和RCA

hbase master 服务异常

报警时间: 2018-11-13 00:02:53
再次报警时间: 2018-11-13 17:38
报警内容: hbase.master.alive
报警触发值: 1.00
报警描述: hbase master 服务可能不可用, 请立刻重启

主机名: bigdata53
状态: 严重

上下游关联
引发报警的高消耗进程
可能引起报警的事件
报警关联指标 (部分)
报警关联日志 (部分)
采纳方案²

报警关联日志 (部分)

已扫描所有相关日志

发现 4 类与报警相关的 关联类日志

- 1 (2) \$Y-M-D_H:M:S\$ ERROR [master:bigdata#/#] LL master:#x#e#d# quorum=bigdata#/# bigdata#/# bigdata#/# bigdata#/# baseZNode= hbase ***** LL KeeperErrorCode = Session expired for hbase maste...
- 2 (1) \$Y-M-D_H:M:S\$ INFO LL Opening socket connection to server bigdata# \$IP\$#. Will not attempt to authenticate using SASL (unknown error)
- 3 (1) \$Y-M-D_H:M:S\$ ERROR [master:bigdata#/#] LL ZooKeeper getData failed after # attempts
- 4 (1) \$Y-M-D_H:M:S\$ ERROR [main] LL Master exiting LL HMaster Aborted at LL at LL at LL at LL at LL

根据快捷查询名称 hbase 连接 zookeeper 超时 发现 1 类 可能导致报警的 关联类日志

- 1 (1) \$Y-M-D_H:M:S\$ INFO LL LL Client session timed out have not heard from server in #ms for sessionid #x#e#d# closing socket connection and attempting reconnect 解决方案



AIOps场景建设 - 故障根因分析和RCA

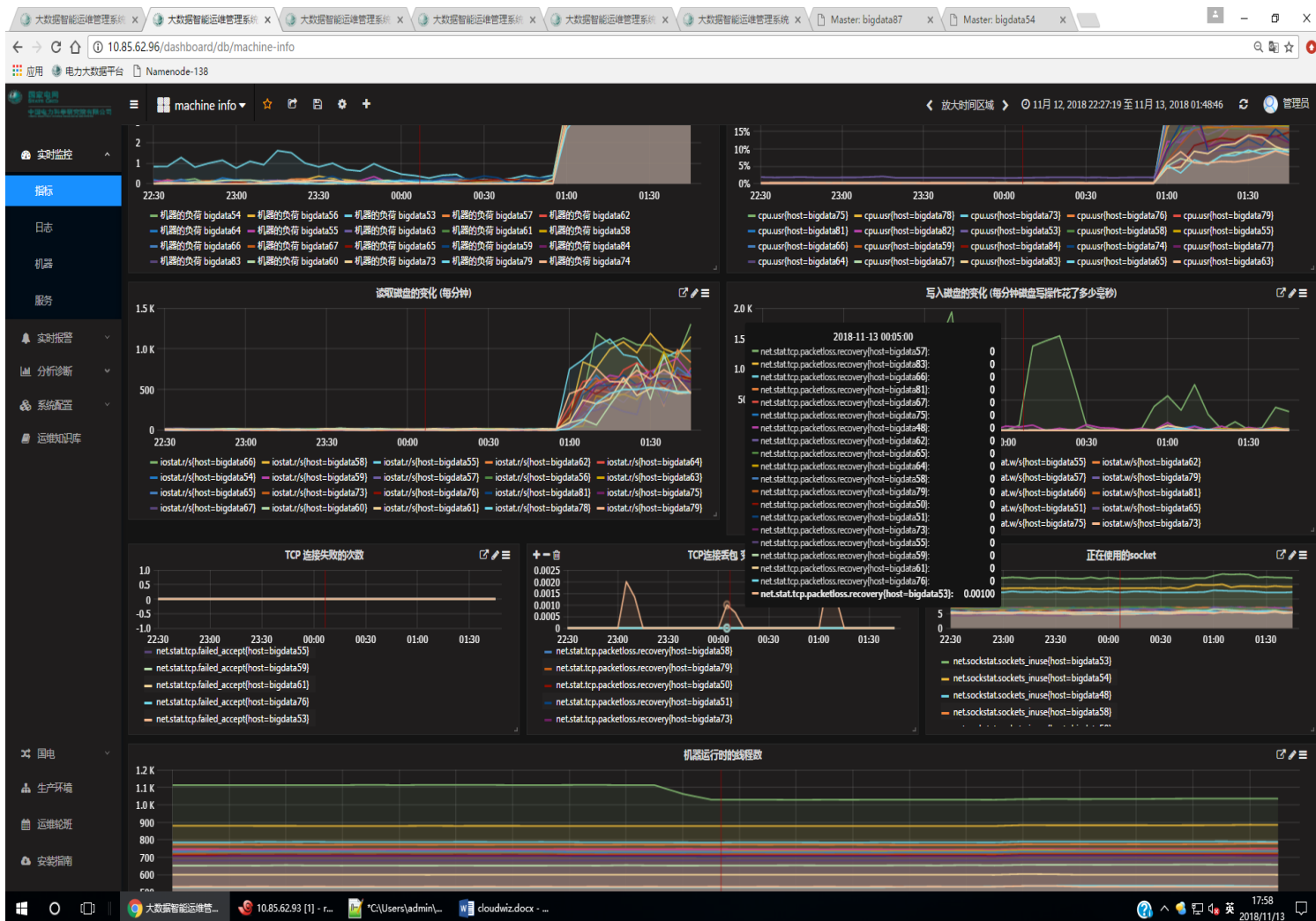
The screenshot displays the AIOps dashboard interface, which is used for monitoring and managing system health. The main panel shows a list of anomalies, with the top entry being "hbase master 服务异常" (hbase master service anomaly). This entry includes details such as the report time (2018-11-13 00:02:53), the last report time (2018-11-13 17:38), and the report description (hbase master 服务可能不可用, 请立刻重启). Below this, there are two cards showing the status of the service: "上下游关联" (Upstream/Downstream Association) and "引发报警的高清处理" (High-resolution processing triggered by the alarm). The "上下游关联" card shows a 100% recommendation rate and 0 adoption rate. The "引发报警的高清处理" card shows a 67% recommendation rate and 0 adoption rate.

The right panel shows a detailed view of the "hbase master 服务异常" entry. It includes a "问题描述" (Problem Description) section with the following text: "hbase master 服务可能不可用, 请立刻重启" (hbase master service may be unavailable, please restart immediately). Below this, there is a "解决方案" (Solution) section with the text "重启服务" (Restart service). The "错误日志" (Error Log) section shows a log entry from 2018-11-13 00:01:33, indicating a zookeeper connection timeout and a fatal error in the hbase master service.

The bottom panel shows a list of anomalies, with the top entry being "zookeeper连接超时导致 HMaster 服务异常" (zookeeper connection timeout causing HMaster service anomaly). This entry includes details such as the report time (2018-11-13 00:02:53), the last report time (2018-11-13 17:38), and the report description (hbase master 服务可能不可用, 请立刻重启). Below this, there are two cards showing the status of the service: "上下游关联" (Upstream/Downstream Association) and "引发报警的高清处理" (High-resolution processing triggered by the alarm). The "上下游关联" card shows a 100% recommendation rate and 0 adoption rate. The "引发报警的高清处理" card shows a 67% recommendation rate and 0 adoption rate.



AIOps场景建设 - 故障根因分析和RCA



GIAC

全球互联网架构大会

GLOBAL INTERNET ARCHITECTURE CONFERENCE



关注公众号获得
更多案例实践