AIOps Challenge

2023 CCF国际AIOps挑战赛决赛
暨“大模型时代的AIOps”研讨会

面向微服务架构系统中无标注、多模态运维数据的异常检测、根因定位与可解释性分析

国电南瑞科技股份有限公司 彭程

主办单位：中国计算机学会（CCF）、清华大学、中国建设银行股份有限公司、南开大学
承办单位：中国计算机学会互联网专委会、清华大学计算机科学与技术系、中国建设银行股份有限公司运营数据中心、南开大学软件学院、北京必示科技有限公司
赞助单位：华为技术有限公司、国网宁夏电力有限公司电力科学研究院、软通动力信息技术（集团）股份有限公司

目录
CONTENTS

第一章节 团队介绍
第二章节 选题分析
第三章节 解决方案
第四章节 总结展望

第一章节

团队介绍

团队介绍

AIOps Challenge

队伍名称：

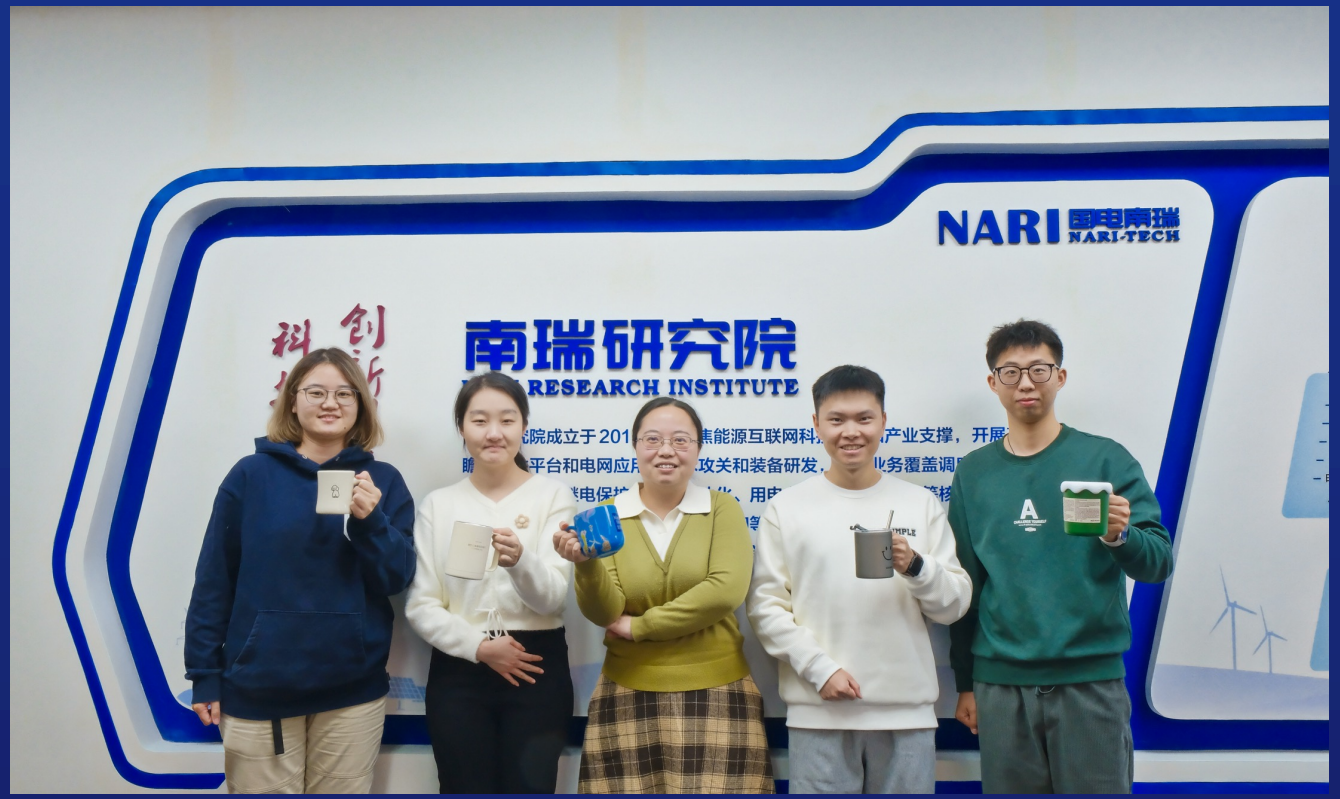
CheerX

参赛选手：

彭程、高尚、陈子韵、孔彦茹、田真龙（东南大学）

团队介绍：

CheerX团队来自于南瑞研究院系统平台研发中心，中心主要从事NUSP电力自动化通用软件平台的关键技术研究与软件研发。项目团队潜心攻关，在系统异常检测、故障根因定位等方向发表多篇SCI论文，研发了全景监视、风险预警、故障诊断及辅助处置为一体的电力自动化系统运行监视平台，为应用故障发现、诊断和定位等环节提供全流程高效支撑，提升电力自动化系统在状态感知和故障诊断定位方面的智能化水平。



第二章节

选题分析

选题分析

AIOps Challenge

运维现状：多种问题造成运维系统可用性差

- 故障数据标签的获取成本高、精度低
- 很少使用多模态运维数据，主要还是指标异常检测
- 指标异常检测的海量告警问题
- 基于有监督学习的故障分类解释性差

当前目标：希望增强运维系统的可用性与可解释性

- 希望降低甚至避免故障标签的成本
- 希望利用多模态数据挖掘出更多故障信息
- 希望消除指标告警风暴，输出事件级故障根因
- 希望有完整清晰的故障诊断流程

选题方案：面向微服务架构系统中无标注、多模态运维数据的异常检测、根因定位与可解释性分析

方案特点：

- 采用无监督的解决方案
- 使用多模态数据挖掘故障信息
- 输出事件级别的告警，避免指标告警
- 输出故障事件的诊断链路

运维能力：

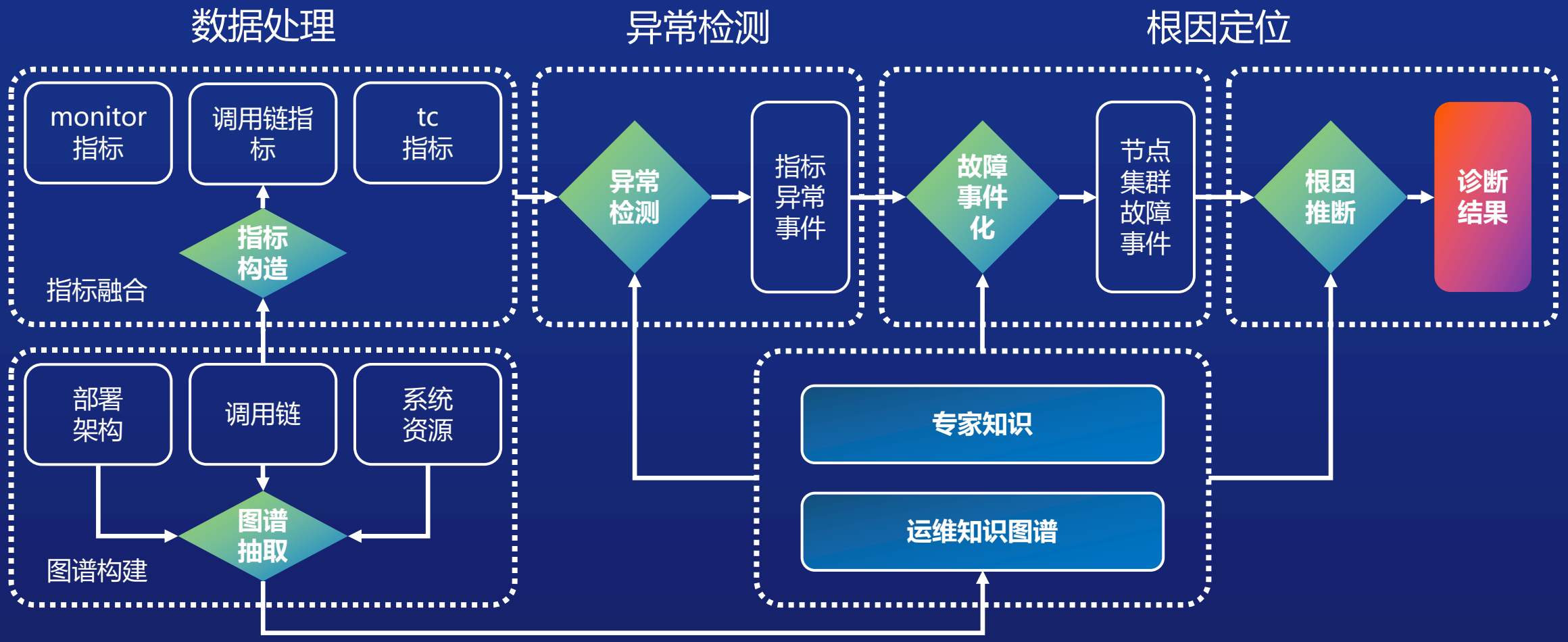
- 异常检测：快速准确定位异常时间和指标
- 根因定位：事件级别根因定位，形成故障链路
- 资源预警：系统资源使用率预警

第三章节

解决方案

解决方案-整体流程

AIOps Challenge

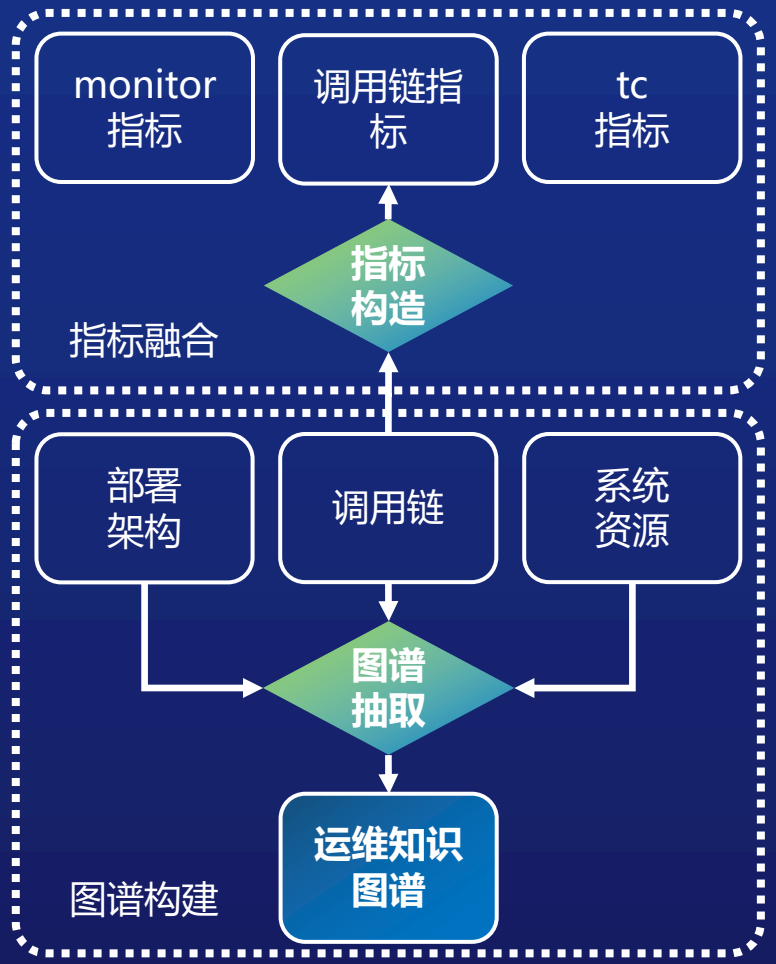


解决方案-数据处理

AIOps Challenge

亮点1：多模态指标构造与融合

- 从调用链数据构造出耗时，错误率等指标，与monitor指标、tc指标融合。
- 融合系统资源、部署架构和调用关系形成系统运维知识图谱。



monitor指标：

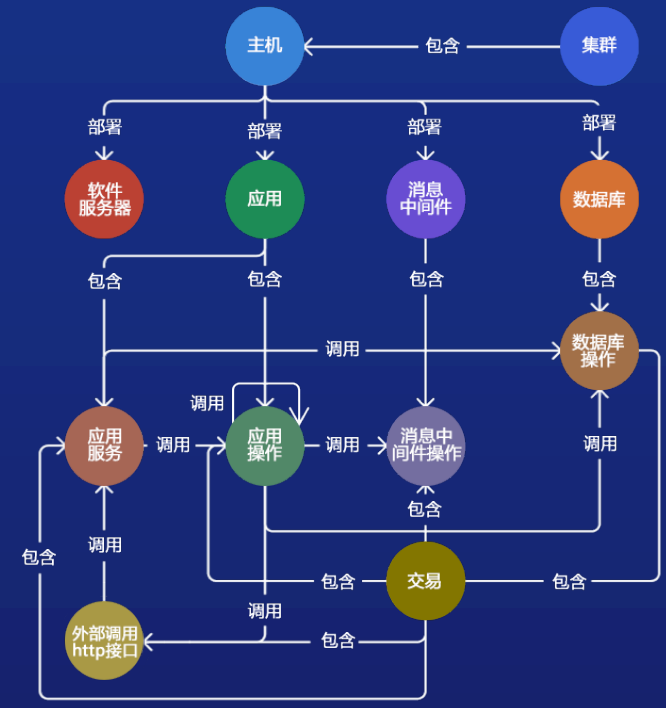
包含系统负载，cpu，磁盘，内存，tcp，网络等22种指标。

tc指标：

交易量，交易性能，平均响应时长，业务成功率。

抽取调用链数据构造指标：

- cost：调用的平均耗时
- selfCost：调用的平均自身耗时
- error_rate：调用错误率
- num：调用次数



运维知识图谱

包含12种概念、21种关系

解决方案-异常检测

AIOps Challenge

亮点2：通用可解释的指标异常检测算法

- 加入**专家知识**对指标进行筛选和过滤，让指标检测算法检测出的结果有可解释性。

检测算法的结果体现的是数据形态上的离群，但这种离群值不一定是由业务系统**实际发生的故障**引起的。**只将检测算法的输出当作异常，会产生大量无意义的检测结果，造成告警风暴。**

举例：

- 资源使用率：应检测上升沿且应当大于一定值
- 业务成功率：应检测下降沿且应当小于一定值

……（其他专家知识）

解决方案-异常检测

AIOps Challenge

亮点2：通用可解释的指标异常检测算法

- 构建**通用**的指标检测流程，将形态各异的指标转换为**显著性序列**进行异常检测。

算法思路：

- 使用**卡尔曼滤波**计算趋势，保留异常特征
- 不管指标形状如何，异常一定是形态上显著的，因此使用**谱残差提取显著特征**
- 使用**箱型图检测**显著性序列异常

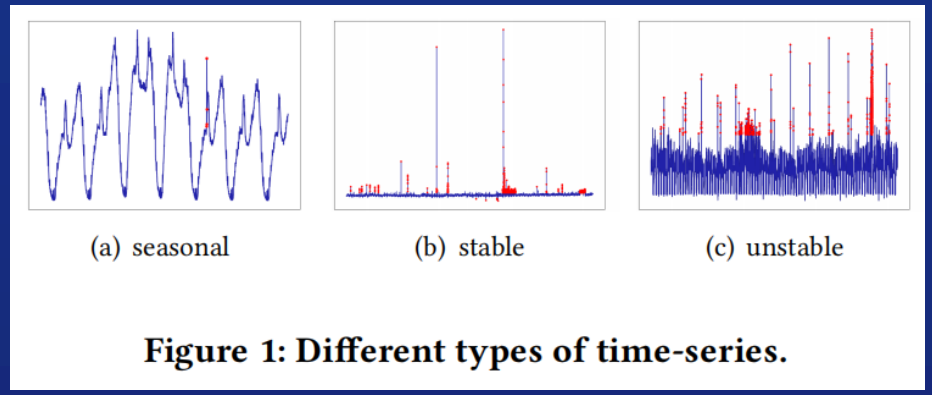


Figure 1: Different types of time-series.

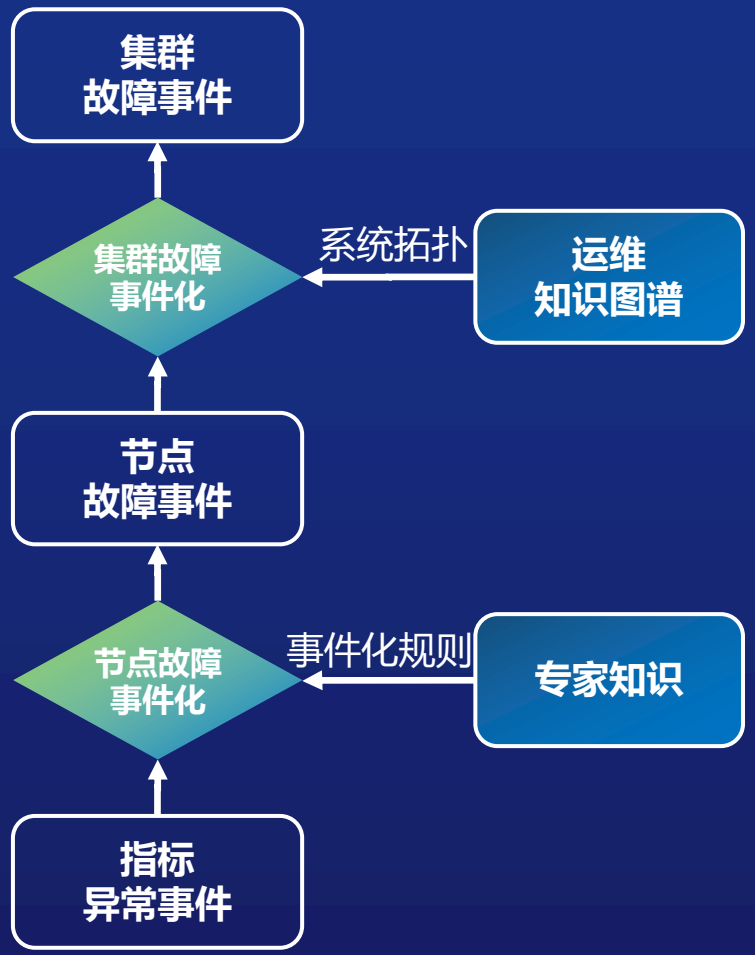
参考论文：Time-Series Anomaly Detection Service at Microsoft

解决方案-根因定位

AIOps Challenge

亮点3：基于故障事件化和知识图谱的根因定位

- 将指标检测结果进行**故障事件化**，将指标级别的告警合并成事件级别的告警，减少告警量。



根据运维知识图谱和专家经验构建故障事件化规则，得到事件级别告警

指标异常事件	CPU类	LINUX_CPU_USAGE_ANOMALY	CPU 使用率升高
		LINUX_CPU_IO_USEAGE_ANOMALY	CPU IO使用率升高
		LINUX_CPU_USER_USEAGE_ANOMALY	CPU 用户使用率升高
	负载类	LINUX_LOAD_ANOMALY	系统负载 升高
	磁盘类	LINUX_DISK_USAGE_ANOMALY	磁盘分区 使用率升高
		LINUX_FS_INODE_USAGE_ANOMALY	磁盘分区 inode使用率升高
		LINUX_IO_READ_AWAIT_ANOMALY	磁盘 读响应时间增加
		LINUX_IO_WRITE_AWAIT_ANOMALY	磁盘 写响应时间增加
		LINUX_IO_UTIL_ANOMALY	磁盘 IO使用增加
	内存类	LINUX_MEM_USAGE_ANOMALY	内存使用率（应用角度)升高
		LINUX_MEM_REAL_USAGE_ANOMALY	内存使用率（系统角度)升高
	网络类	LINUX_TCP_RETRANS_PCT_ANOMALY	TCP 重传率 升高
		LINUX_TCP_SPECIAL_STATUS_NUM_ANOMALY	TCP 特殊状态数 增加
		LINUX_NET_BYTES_RCVD_ANOMALY	网卡入流量 下降
		LINUX_NET_BYTES_SEND_ANOMALY	网卡出流量 下降
		LINUX_NET_PACKAGES_IN_COUNT_ANOMALY	接收数据包数 下降
		LINUX_NET_PACKAGES_OUT_COUNT_ANOMALY	发送数据包数 下降
		LINUX_NET_PACKAGES_IN_COUNT_ANOMALY	接收数据包数 下降
		LINUX_NET_PACKAGES_OUT_COUNT_ANOMALY	发送数据包数 下降
	TC类	TC_AMOUNT_ANOMALY	交易量下降
		TC_APDEX_ANOMALY	交易时间性能下降
		TC_BUS_SUCCESS_RATE_ANOMALY	交易业务成功率下降
		TC_AVG_RES_TIME_ANOMALY	交易平均响应时间上升
	Trace类	TRACE_NUM_ANOMALY	调用量下降
		TRACE_ERROR_RATE_ANOMALY	调用错误率上升
		TRACE_COST_ANOMALY	调用耗时增加
		TRACE_SELFCOST_ANOMALY	调用自身耗时增加

合并后的节点故障事件：

节点故障事件	Linux类	LINUX_DISK_BUSY	磁盘IO繁忙
		LINUX_DISK_CAPACITY_CONSUMPTION	磁盘空间消耗
		LINUX_MEM_CAPACITY_CONSUMPTION	内存消耗
		LINUX_NETWORK_FAILURE	网络故障
		LINUX_NETWORK_TRAFFIC_REDUCE	网络流量下降
		LINUX_CPU_CAPACITY_CONSUMPTION	CPU消耗
	TC类	TC_LOW_EFFICIENCY	交易低效
		TC_FAILURE	交易失败
	Trace类	TRACE_FATAL	调用故障
		TRACE_FAILURE	调用失败
		TRACE_LOW_EFFICIENCY	调用低效

合并后的集群故障事件：

集群故障事件	Linux类	LINUX_DISK_BUSY	集群磁盘IO繁忙
		LINUX_DISK_CAPACITY_CONSUMPTION	集群磁盘空间消耗
		LINUX_MEM_CAPACITY_CONSUMPTION	集群内存消耗
		LINUX_NETWORK_FAILURE	集群网络故障
		LINUX_NETWORK_TRAFFIC_REDUCE	集群网络流量下降
		LINUX_CPU_CAPACITY_CONSUMPTION	集群CPU消耗
	Trace类	TRACE_CLUSTER_FATAL	集群调用故障
		TRACE_CLUSTER_FAILURE	集群调用失败
		TRACE_CLUSTER_WAIT	集群调用慢

解决方案-根因定位

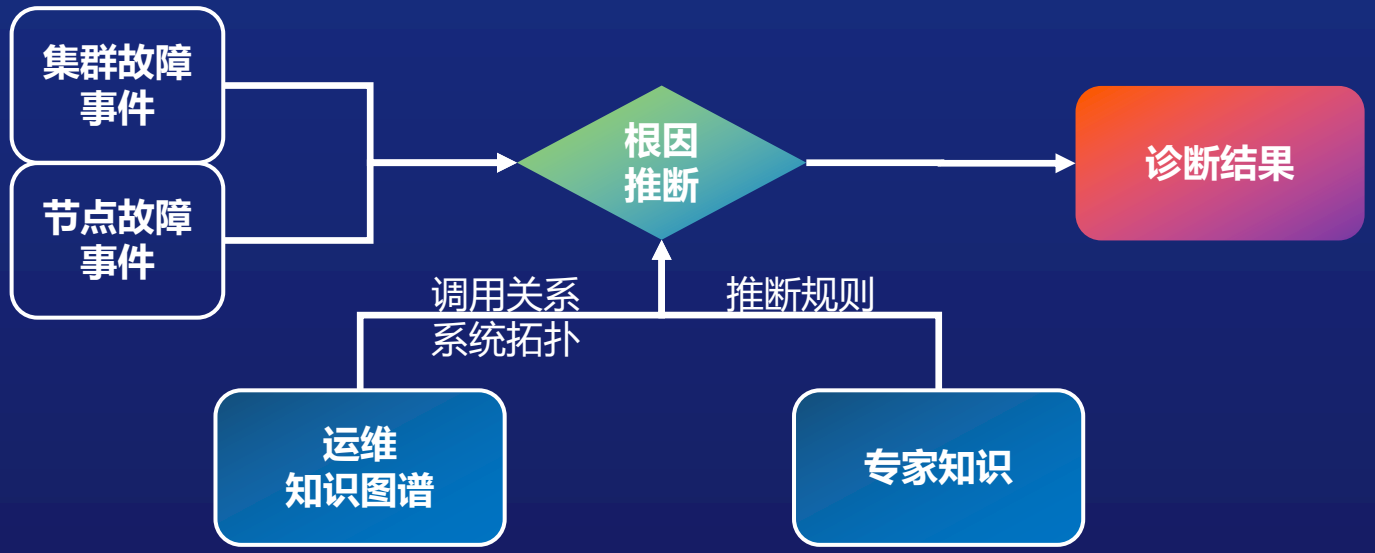
AIOps Challenge

亮点3：基于故障事件化和知识图谱的根因定位

- 通过运维知识图谱和专家知识推断故障发生事件，根据调用关系和系统拓扑合并故障事件形成**诊断链路**。

定位思路：

- 根据异常的交易推断公共的故障调用链
- 根据调用链的调用关系推断实际发生调用故障的位置
- 根据系统拓扑推断故障调用所在主机和集群，查看是否存在硬件故障
- 根据调用或者拓扑关系进一步合并故障，输出合并之后的诊断结果



结果输出：

故障时间，故障事件和故障诊断树。

诊断树上包含了指标、节点和集群的故障事件，输出了完整的诊断过程。

```
======= 报告时间 ：1694636700 =======
发生故障事件1 ： 网络故障
--事件： 网络故障 | cluster: 外网WEB集群
 --事件： 网络故障 | cmdb_id: nginx_10
   --事件： TCP 重传率 升高 | cmdb_id: nginx_10 | device: null
   --事件： TCP 特殊状态数 增加 | cmdb_id: nginx_10 | device: null
 --事件： 网络故障 | cmdb_id: nginx_11
   --事件： TCP 重传率 升高 | cmdb_id: nginx_11 | device: null
   --事件： TCP 特殊状态数 增加 | cmdb_id: nginx_11 | device: null
 --事件： 网络故障 | cmdb_id: nginx_12
   --事件： TCP 重传率 升高 | cmdb_id: nginx_12 | device: null
   --事件： TCP 特殊状态数 增加 | cmdb_id: nginx_12 | device: null
 --事件： 网络故障 | cmdb_id: nginx_13
   --事件： TCP 重传率 升高 | cmdb_id: nginx_13 | device: null
   --事件： TCP 特殊状态数 增加 | cmdb_id: nginx_13 | device: null
 --事件： 网络故障 | cmdb_id: nginx_14
   --事件： TCP 重传率 升高 | cmdb_id: nginx_14 | device: null
   --事件： TCP 特殊状态数 增加 | cmdb_id: nginx_14 | device: null
 --事件： 调用故障 | inst_name: http://LLMN_18:18088/p5mcppl4 | inst_type: http_client
   --事件： 调用故障 | inst_name: AppId008:TxController.doTx | inst_type: app_op
     --事件： 调用故障 | cmdb_id: Weblogic_19 | inst_name: AppId008:TxController.doTx | inst_type: app_op
       --事件： 调用失败 | cmdb_id: Weblogic_19 | inst_name: AppId008:TxController.doTx | inst_type: app_op
         --事件： 调用低效 | cmdb_id: Weblogic_19 | inst_name: AppId008:TxController.doTx | inst_type: app_op
           --事件： 调用耗时增加 | cmdb_id: Weblogic_19 | inst_name: AppId008:TxController.doTx | inst_type: app_op
           --事件： 调用量下降 | cmdb_id: Weblogic_19 | inst_name: AppId008:TxController.doTx | inst_type: app_op
         --事件： 调用错误率上升 | cmdb_id: Weblogic_19 | inst_name: AppId008:TxController.doTx | inst_type: app_op
       --事件： 调用自身耗时增加 | cmdb_id: Weblogic_19 | inst_name: AppId008:TxController.doTx | inst_type: app_op
     --事件： 调用失败 | cmdb_id: Weblogic_20 | inst_name: AppId008:TxController.doTx | inst_type: app_op
       --事件： 调用低效 | cmdb_id: Weblogic_20 | inst_name: AppId008:TxController.doTx | inst_type: app_op
         --事件： 调用耗时增加 | cmdb_id: Weblogic_20 | inst_name: AppId008:TxController.doTx | inst_type: app_op
         --事件： 调用量下降 | cmdb_id: Weblogic_20 | inst_name: AppId008:TxController.doTx | inst_type: app_op
       --事件： 调用错误率上升 | cmdb_id: Weblogic_20 | inst_name: AppId008:TxController.doTx | inst_type: app_op
     --事件： 调用故障 | cmdb_id: Weblogic_21 | inst_name: AppId008:TxController.doTx | inst_type: app_op
:
```

解决方案-资源预警

AIOps Challenge

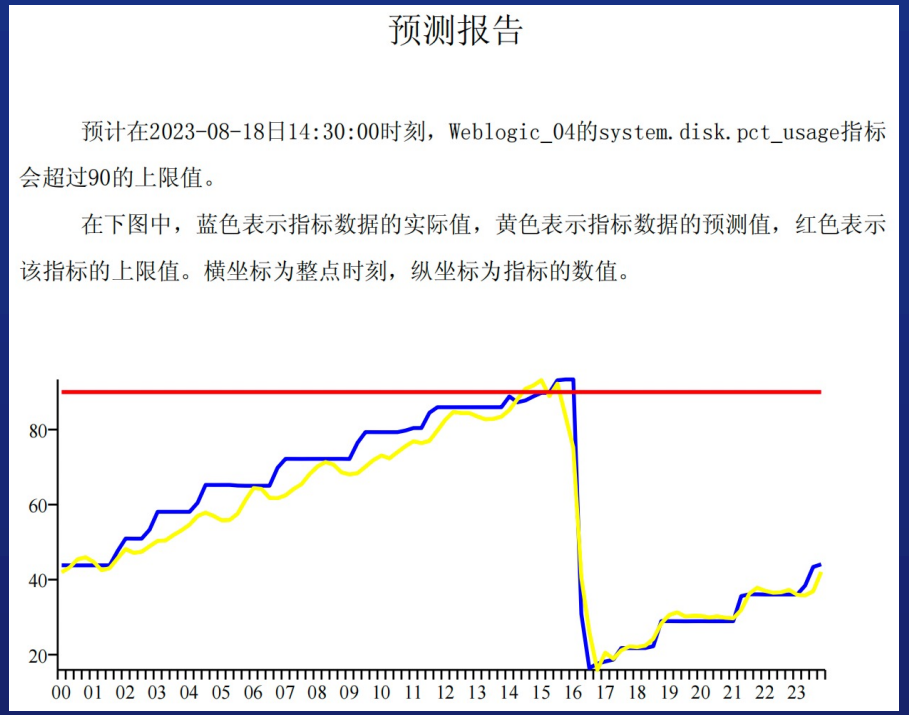
运维能力：基于STL分解与线性回归的资源预警

- 对cpu、磁盘、内存等资源使用率指标进行预测，生成资源预警报告。

预测思路：

- 使用STL分解算法分解资源指标序列
- 对趋势序列，使用线性回归算法进行预测
- 对周期序列，使用历史同时段平均进行预测
- 对两种结果求和，得到预测值

预测结果：



第四章节

总结展望

解决方案-总结展望

AIOps Challenge

面向微服务架构中系统复杂度高、数据多源异构的挑战，本团队提出了一种基于无标注、多模态数据的异常检测、根因定位与可解释性分析智能运维方案。

方案优越性

- 运维人员友好
- 实用性强
- 检测结果准确高效
- 数据与知识双驱

方案创新性

- 多模态数据融合
- 运维知识图谱构建
- 通用的异常检测流程
- 故障事件化与根因定位

方案扩展性

- 结合大模型扩充完善运维知识库
- 结合大模型分析日志数据

2023 CCF国际AIOps挑战赛决赛暨“大模型时代的AIOps”研讨会

THANKS

2023 CCF国际AIOps挑战赛决赛暨“大模型时代的AIOps”研讨会

THANKS