十年架构 成长之路

SACC 第十届中国系统架构师大会

SYSTEM ARCHITECT CONFERENCE CHINA 2018

2018年10月17-10月21日 北京海淀永泰福朋喜来登酒店

大型企业智能运维的探索和实践

孙 杰

目 录

构建新IT运维
管理体系

全景业务
服务管理

日志采集
监控告警

知识库
故障自治

01

构建新IT运维管理体系

传统运维软件逐渐不适应运维需求

事后

所有的运维软件大多是事后报警，此时损失已经造成，晚了！

零散

一种软件监控一类设备，无法提供整体的运维监控解决方案

“弱智”

智能化程度差，以监控和报表为主，不具备大数据关联分析和深度数据挖掘功能

单一

针对不同的用户提供的是相同的界面和视图，不能满足用户不同岗位、不同业务的运维要求

无用

由于无法发挥实质性的作用，且运行时间长之后性能影响显著，最终被弃用。

传统运维存在的突出问题

数据分散，不利于故障分析和问题跟踪

- 不同的数据存储在不同的运维系统中，无法关联
- 数据格式、时间戳等各不相同，不利于问题排查

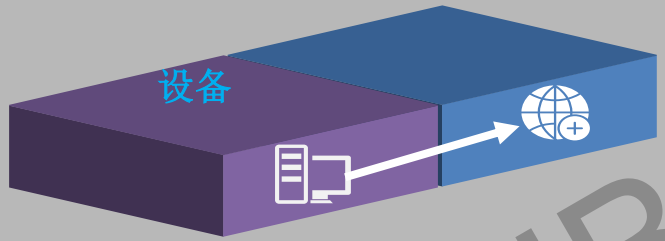
要的功能没有，没用的数据重复采集，影响正常业务

- 每个运维软件都有特长部分，但采集数据多有重复
- 有些甚至相互影响，干扰正常业务运行

投资浪费，增加运维压力

- 采购多种运维软件，在功能上、设备上存在投资浪费
- 没有减轻运维压力，还增加多种软件系统的维护工作

运维技术在持续升级



以设备为中心的维护

升级为

以数据为中心的运营

升级的3个原因：

技术进步

运维事故

运维压力

	人工	工具	自动化	智能
现状	目前大量的用户采用人工运维方式，包括自行运维、外包运维、原厂维保等	一些用户开始尝试自主开发工具、外购工具或者利用其他软件的附带工具进行运维	大部分互联网用户使用自动化运维；仅少量传统用户尝试自动化运维	很多客户开始探索使用大数据进行智能运维管理，并获得惊人收获
前景	人艰不拆	“弱智” “无用”	实施不易	未来趋势

运维的理想

无论云上云下，保障业务系统稳定运行都是最重要的工作。

- 通过部署智能运维系统，能够显著提升运维效率，大大增强运维团队的能力和价值；
- 通过部署智能运维系统，能够显著增加运维透明度，使管理和运维人员增加主动权和掌控力；
- 通过部署智能运维系统，能够显著降低故障频率，使运维更省心。

维护 ->运营

帮助用户将以设备为中心的维护升级为以数据为中心的运营。

“活着” -> 健康

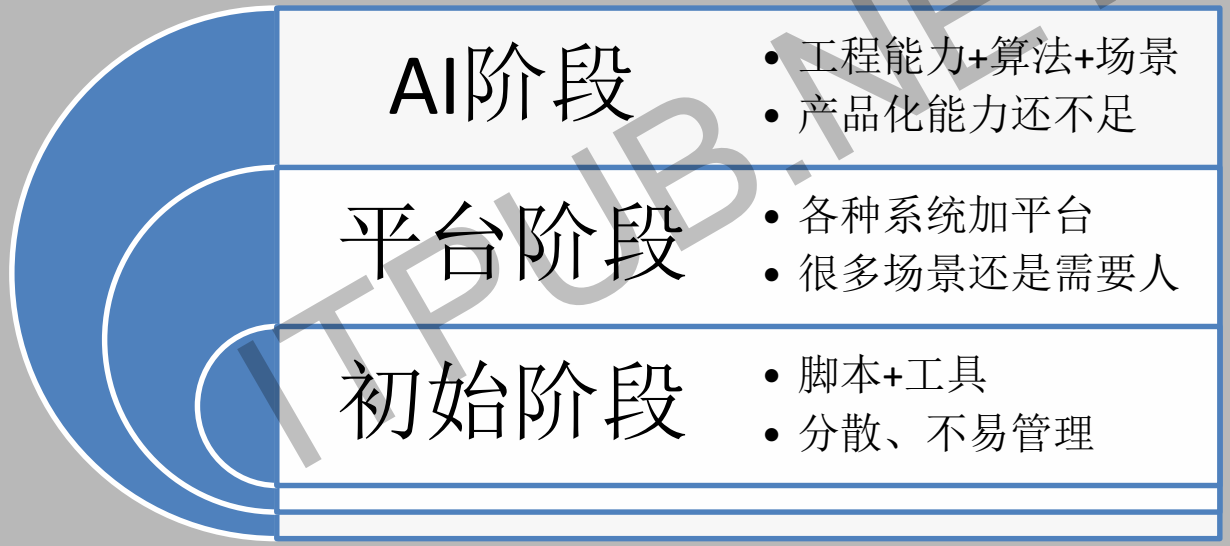
将运维质量的标准，从保证系统“活着”，升级为确保系统始终运行在最佳状态。

合规 -> 敏捷

将用户的运维管理，从满足流程要求的合规管理，升级为以事件响应为特点的敏捷管理。

AIOps运维阶段发展和演进

AIOps：即Algorithmic IT Operations，是由Gartner定义的新类别，基于算法的IT运维。通俗来说，就是将人工智能数据科学和算法用于传统运维领域，基于已有的运维数据（日志、监控信息、应用信息等），通过机器学习的方式来进一步解决自动化运维所未能解决的问题，提高系统的智能化、稳定性、降低IT成本，并提高企业的竞争力。



驱动的3个力量：

业务驱动

技术驱动

人才驱动

注：当基础设施固定下来后，运维模式最终也会固定下来。你所处的发展阶段，决定了你要做的事。

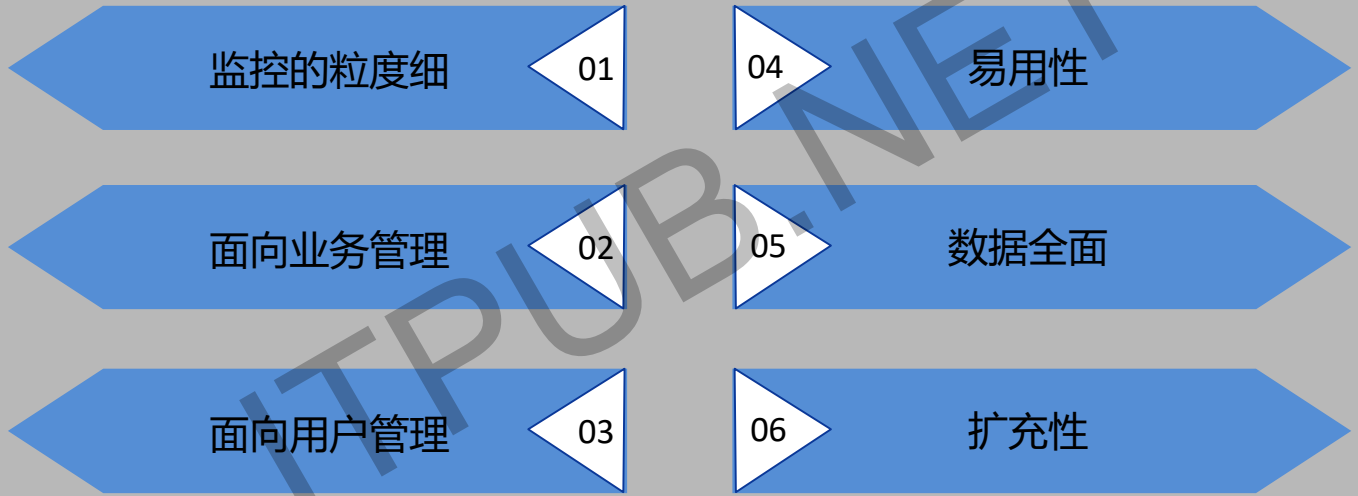
科学规划、分阶段实现

NHTSA	L0	L1	L2	L3	L4	
SAE	L0	L1	L2	L3	L4	L5
	无自动化	驾驶支持	部分自动化	有条件自动化	高度自动化	完全自动化
功能	夜视 行人检测 交通标志识别 盲点检测 并线辅助 后排路口交通警报 车道偏离警告	自适应巡航驾驶系统 自动紧急制动 停车辅助系统 前向碰撞预警系统 车身电子稳定系统	车道保持辅助系统	拥挤辅助驾驶	停车场自动泊车	
特征	传感探测和决策警报	单一功能（以上之一）	组合功能（L1/L2 组合）	特定条件部分任务	特定条件全部任务	全部条件全部任务

02

全景业务服务管理

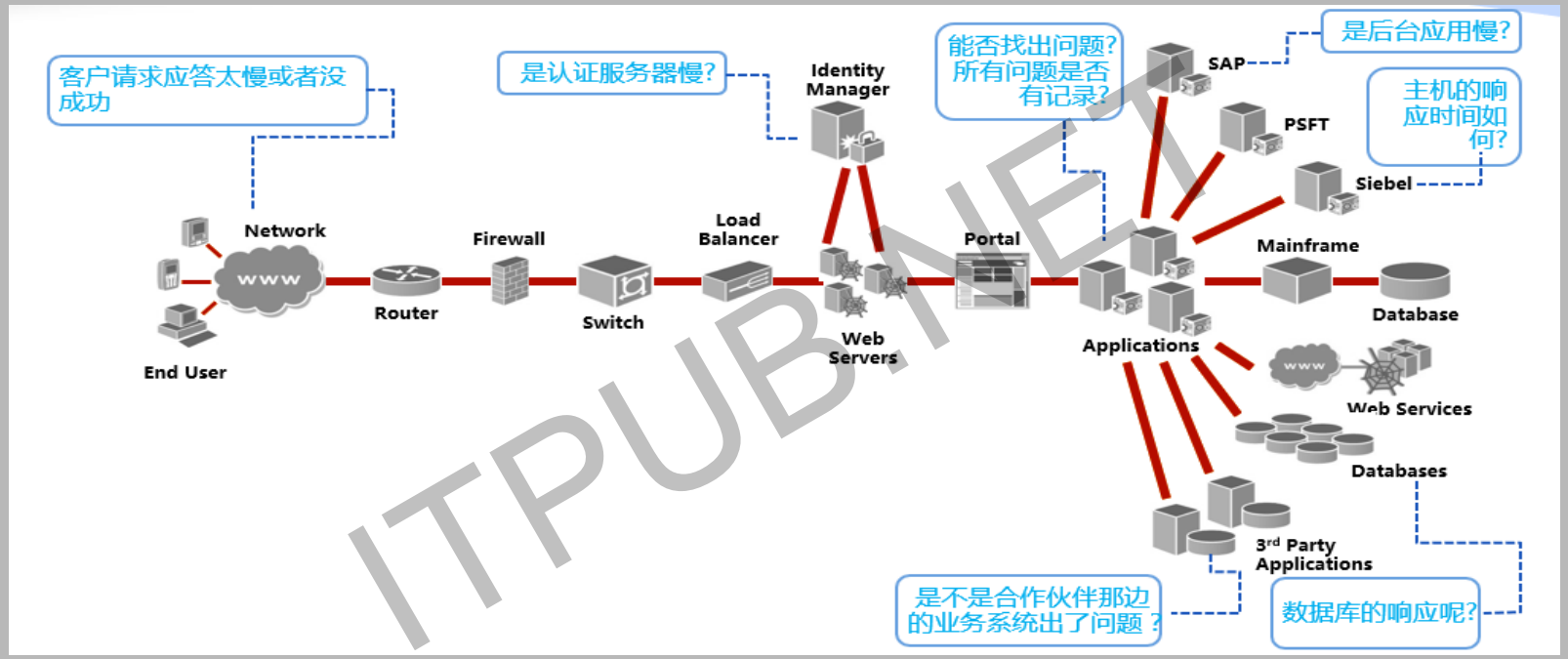
IT业务服务管理—特点



业务视角管理资源的视图

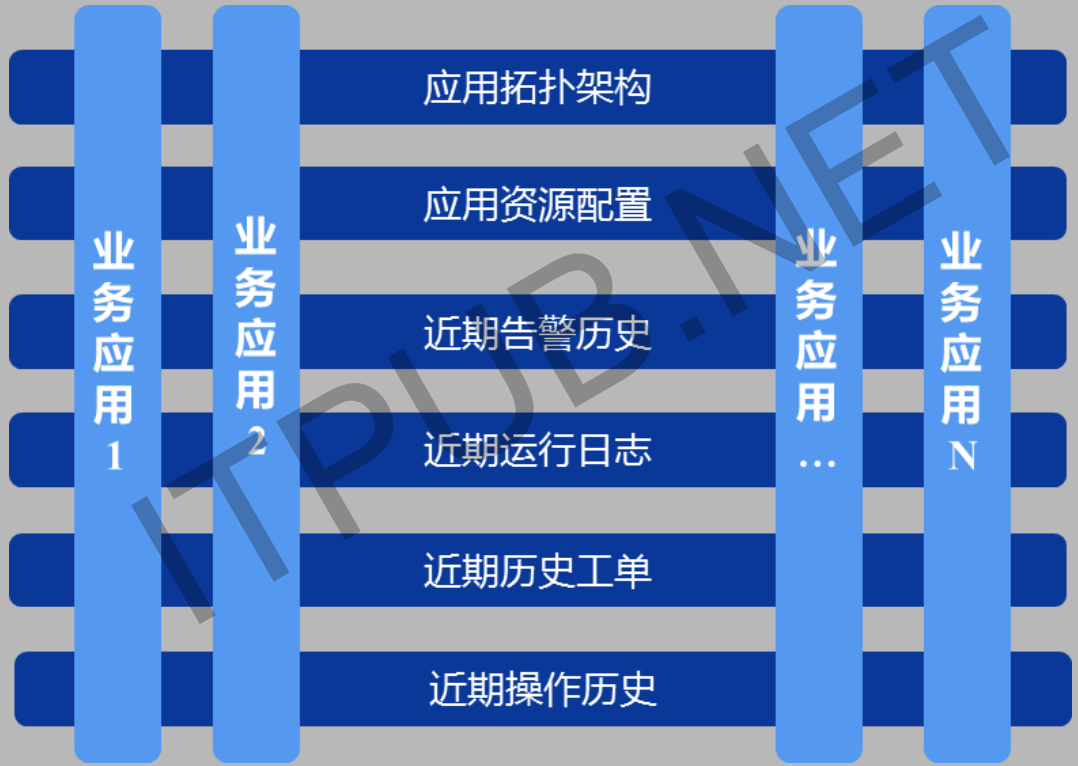
- 自动发现应用依赖关系
- 优先处理关键业务工作负载，然后再处理非生产工作负载
- 以与业务一致的方式管理基础架构

业务视角下全链路分析的必要

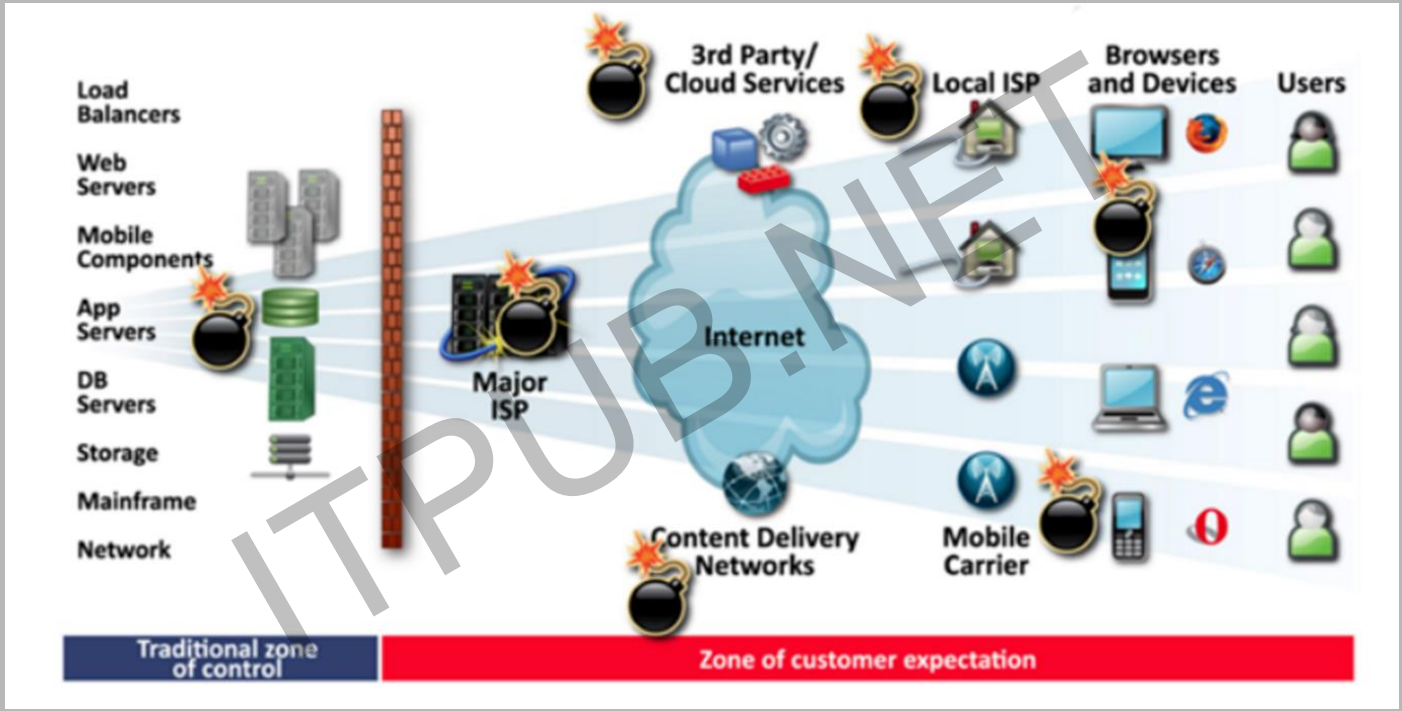


- 1、缺乏端到端的应用性能管理，无法快速准确定位故障原因，导致大量人员成本和时间成本的消耗
- 2、缺乏真实用户体验管理，导致IT部门的核心价值没有充分体现

业务视角的全方位分析



业务应用性能监控---发现瓶颈和故障



- 数据采集：
- 1、客户端：主动式探测和被动式监测　　2、服务端：旁路监听和应用探针

几种技术的对比

位置	方式	技术	侵入式	竞品对标	网络问题定位	全样本	代码级定位	后端服务监控
客户端	主动	基于自动化测试的拨测	--	O	O	--	--	--
	被动	浏览器嵌码	O	--	--	O	--	--
		App嵌码	O	--	O	O	O	--
服务端	被动	旁路监听	--	--	O	O	--	O
		应用探针	O	--	--	O	O	O

03

大数据日志采集与监控告警

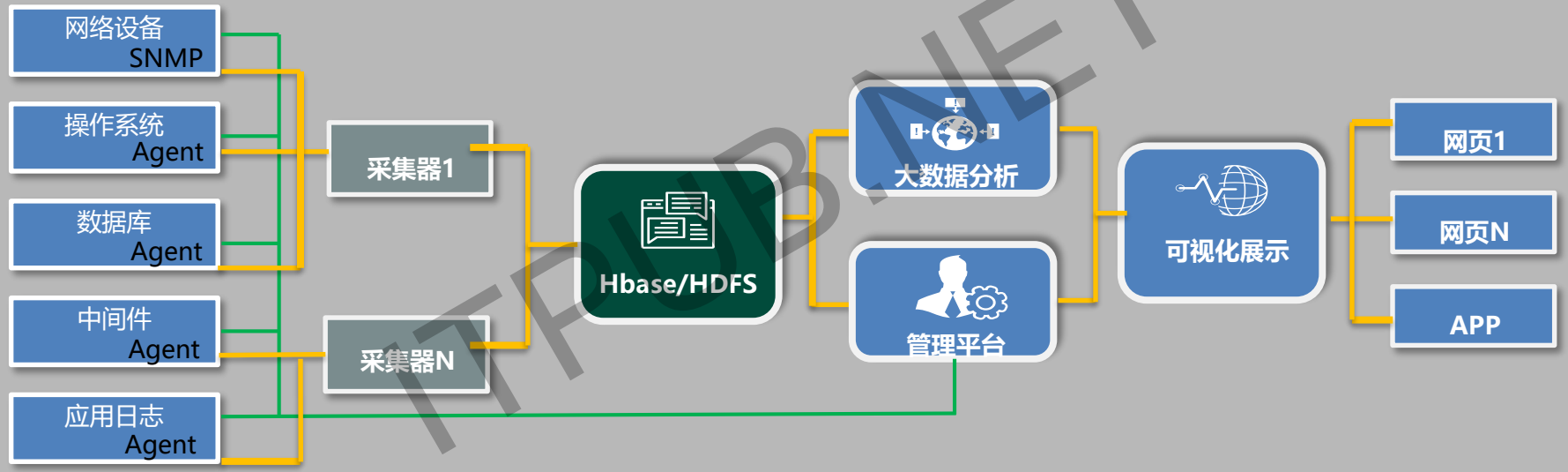
基于大数据平台的日志采集分析

基于大数据平台，提供日志采集和聚合处理

日志关联分析帮助准确全面定位，提升效能和满意度

智能预测与预警，为精细管理，科学决策提供量化依据

各种日志的采集分析



跨层采集与综合监控

T1 设备层

对机房内的各种设备进行监控，如：交换机、路由器、安全设备、服务器、UPS、精密空调等，实现物理层的实时监控和数据采集。

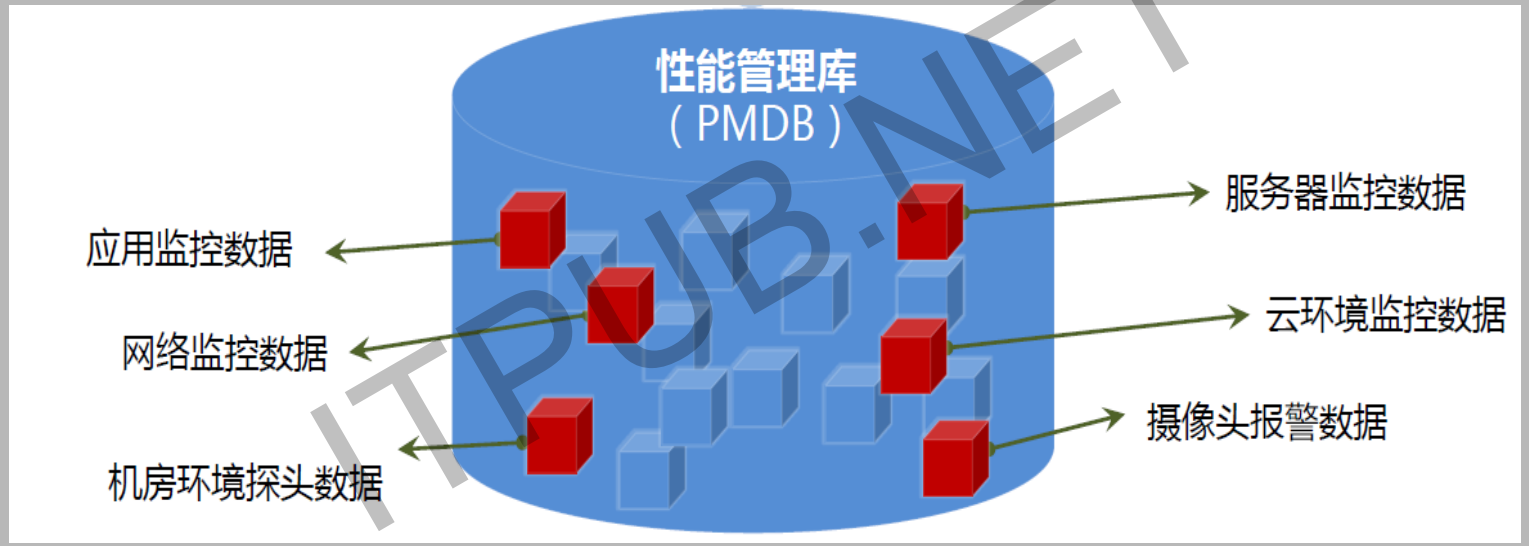
T2 系统层

以系统作为单位，对数据中心的主机(Linux主机和X86服务器)、操作系统(LINUX/Winwdos)、数据库（Oracle、Mysql等主流）、中间件、存储系统、应用软件API、HTTP端口、备份系统、容灾系统、数据同步系统，虚拟化系统，云平台进行实时监控、预警分析和故障定位。

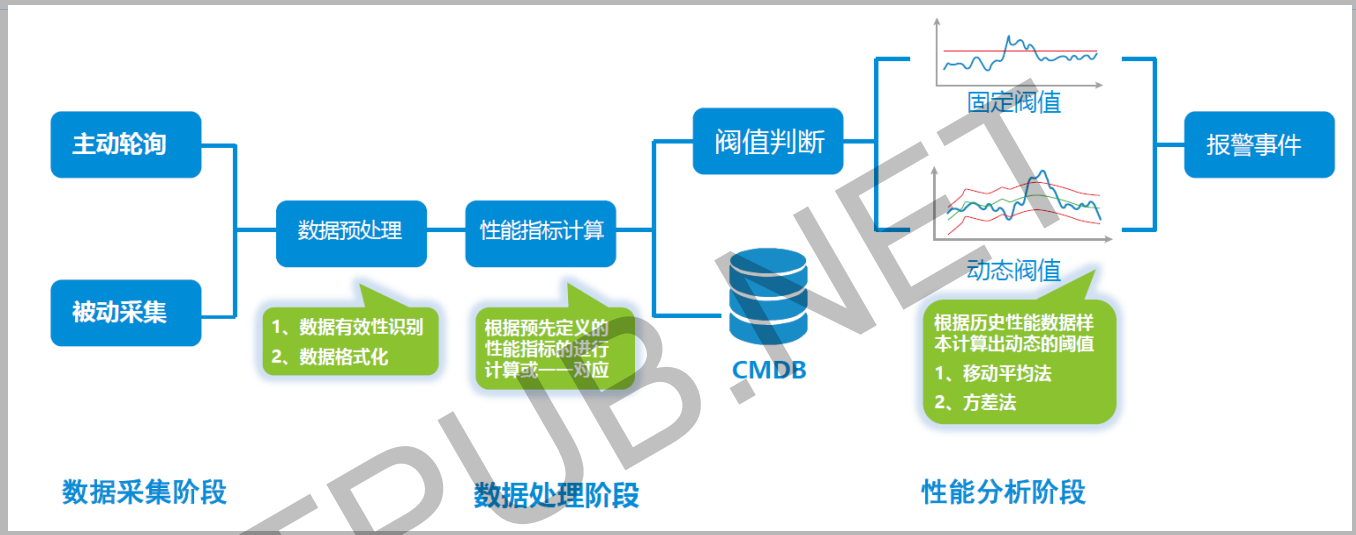
T3 业务层

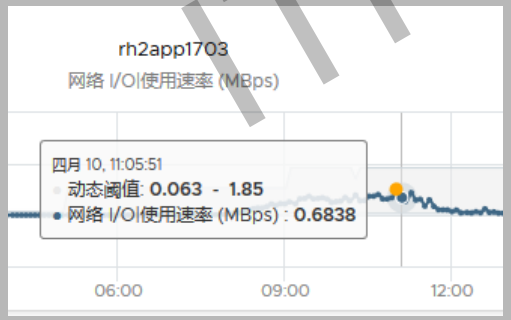
在条件许可的情况下，采集一定的业务数据，如用户数、连接数、业务并发量、日志量等等，通过多维关联和分析，对未来的业务运行进行分析和预测。

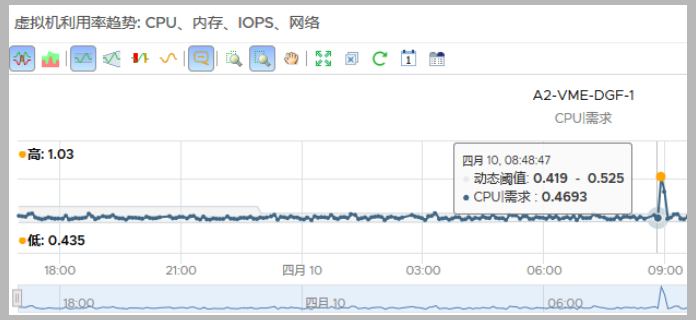
数据大集中--PMDB



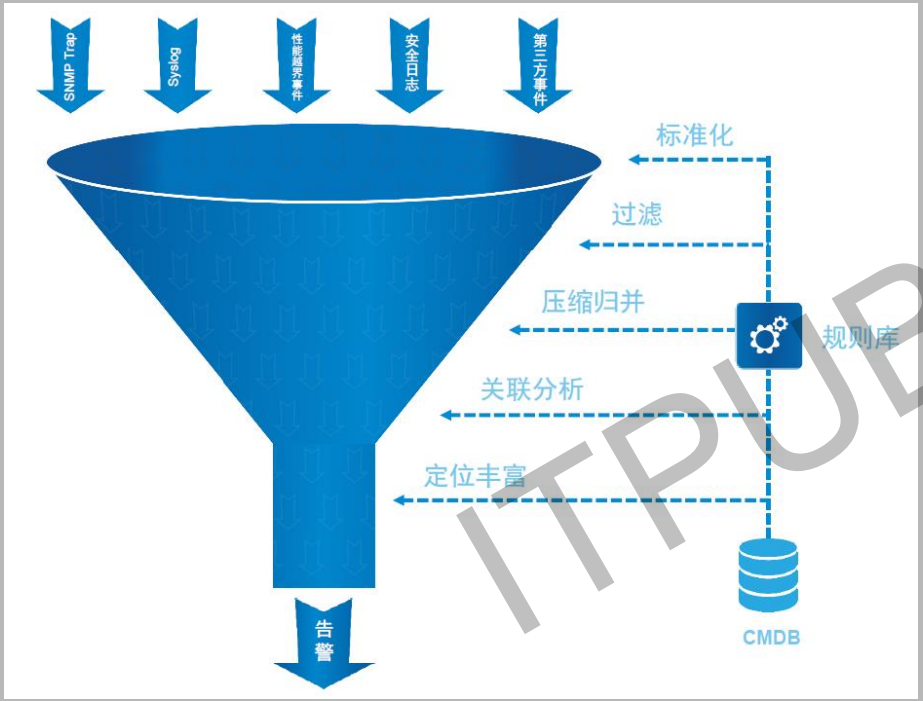
数据统一分析引擎和智能阈值—提前预警







数据的聚合

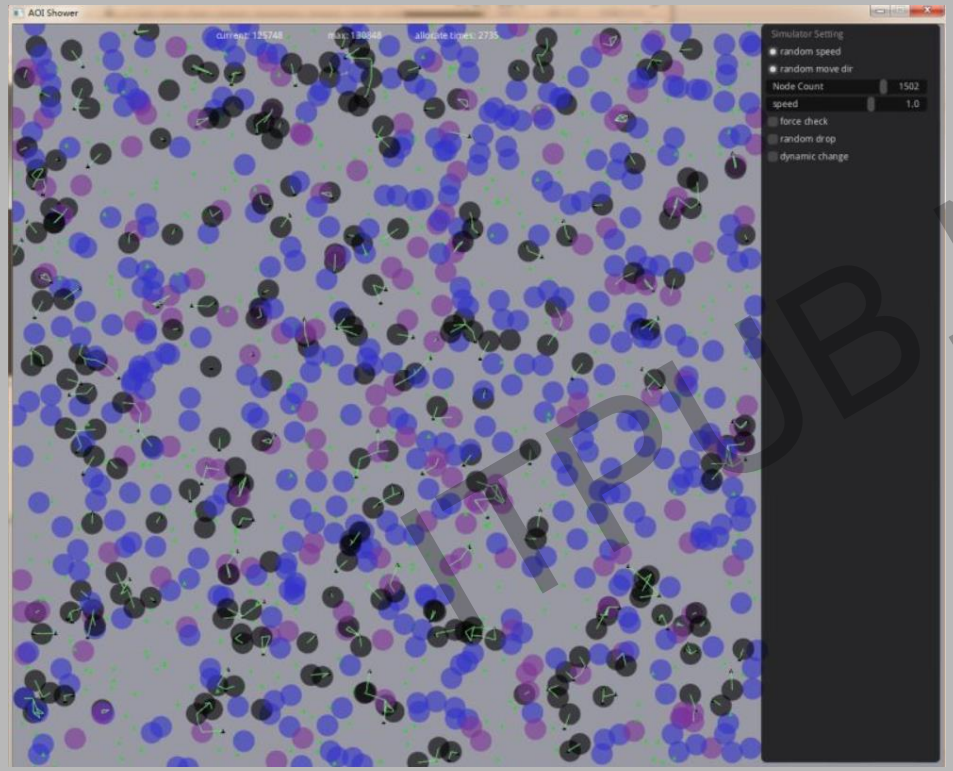


聚合的2个层面：

1、多维度的聚合（时间、位置、业务线、服务、事件、日志、接口等）

2、数据的聚合运算（计数、平均、抽样）

数据降维分析



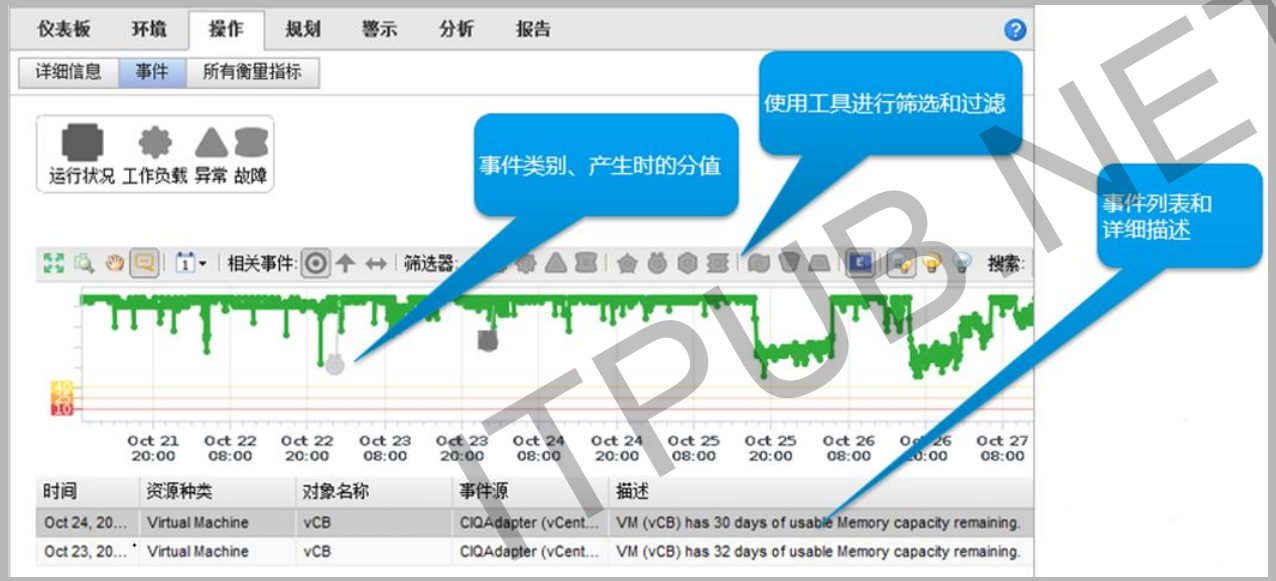
降维分析的几个方法：

1、将告警聚合成关联“事件”（AOI）

2、减少误报，告警分类

3、数据关联

事件和时序关联分析



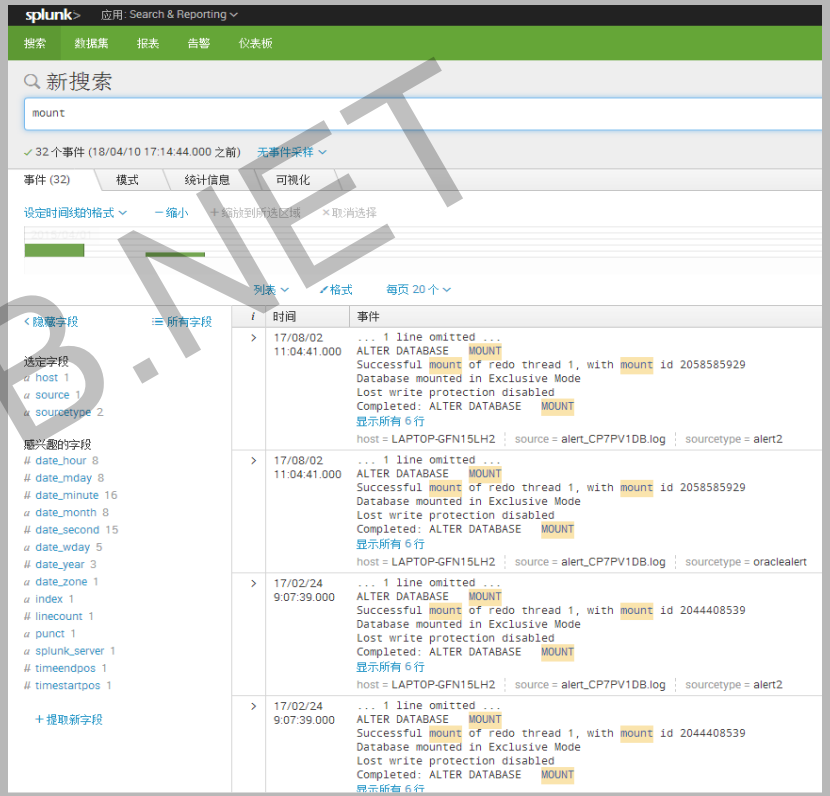
事件诊断一直是运维领域一个很重要的工作，事件和时序数据的相关性不仅可以为事件诊断提供很好的启发，而且在帮助进行根因分析等都能提供很好的线索。

主流时序数据库对比

时序数据库	技术栈	优点	缺点
Graphite	Python	提供丰富的函数支持，对Grafana的支持最好，维护简单，支持自动Downsample	Whisper存储引擎IOPS高，Carbon组件CPU使用率高，聚合分析功能弱
InfluxDB	Go	Metric+Tags，部署简单无依赖，实时数据Downsample，高效存储	开源版本没有集群功能，存在版本兼容问题，聚合分析功能弱
OpenTSDB	Java	Metric+Tags，集群方案成熟，写高效	查询函数有限，依赖Hbase，运维复杂，聚合分析功能弱
Prometheus	Go	Metric+Tags，适用于容器监控，具有丰富的查询语言，维护简单，集成监控和报警功能	没有集群解决方案，聚合分析功能弱
Druid	Java	支持嵌套数据的列式存储，具有强大的多维聚合分析能力，实时高性能数据摄取，具有分布式容错架构，支持类SQL查询	一般不能查询原始数据，不适合维度基数特别高的场景，时间窗口限制了数据完整性，运维较复杂
Elasticsearch	Java	具有强大的多维聚合分析能力，支持全文检索，支持查询原始数据，灵活性高，社区活跃，扩展丰富	不支持分析字段的列式存储，对硬件资源要求高，集群维护较复杂
ClickHouse	C++	具有强大的多维聚合分析能力，实时高性能数据读写，支持类SQL查询，丰富的函数支持，具有分布式容错架构，支持原始数据查询，适用于基数大的维度存储分析	比较年轻，扩展不够丰富，社区还不够活跃，不支持数据更新和删除，不支持事务，集群功能较弱，单存

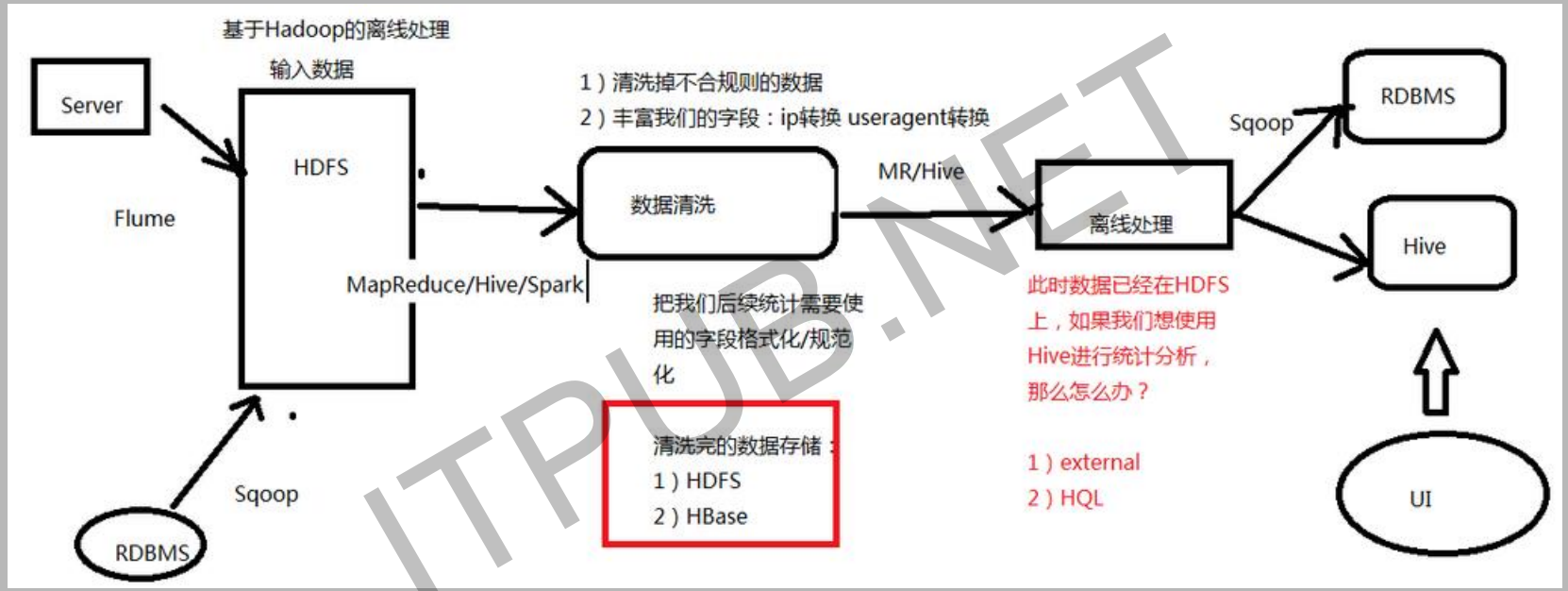
日志处理的几个问题

- 日志没有集中处理
  - 登陆每一台服务器，使用脚本命令或程序查看
- 日志被删除
  - 磁盘满了删日志
  - 黑客删除日志，抹除入侵痕迹
- 日志只做事后追查
  - 没有实时监控、分析
- 使用数据库存储日志
  - 无法适应TB级海量日志
  - 数据库的schema无法适应千变万化的日志格式
  - 无法提供全文检索



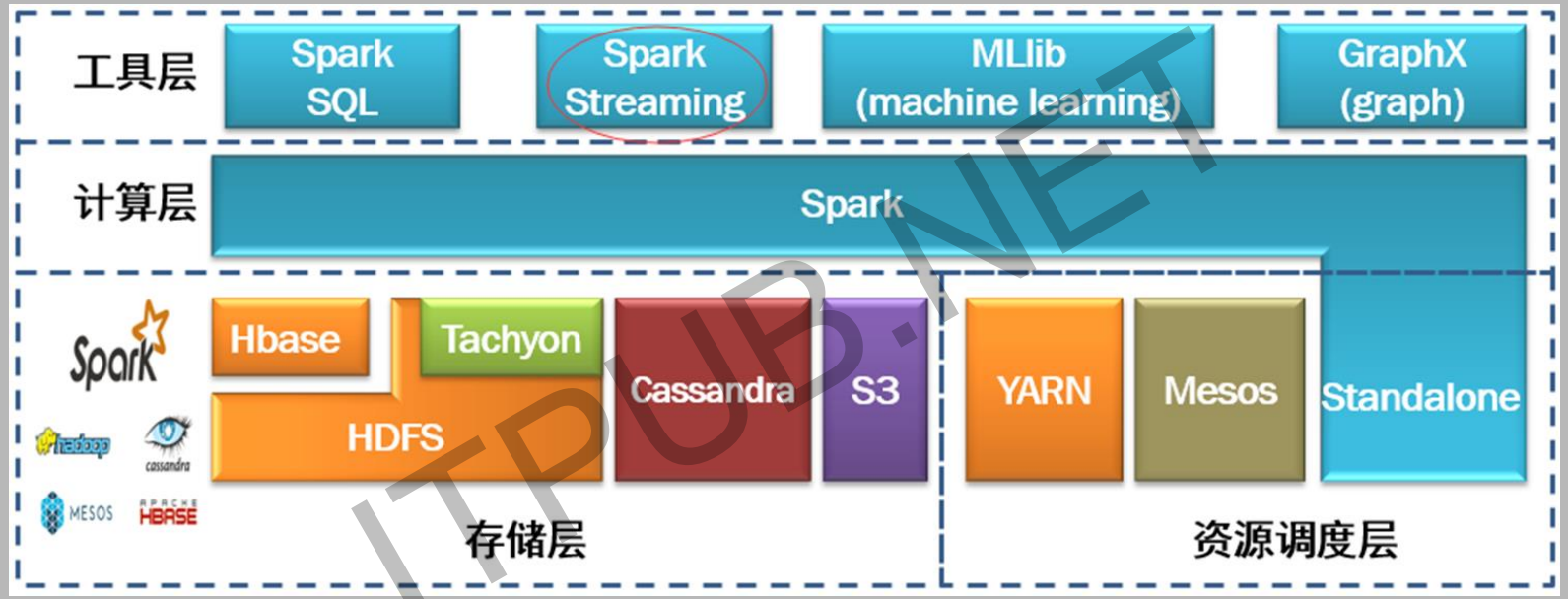
日志标准化：1、日志内容可理解 2、格式相对统一 3、能自我标识

离线日志处理的几个问题



重点：1、离线计算管理 2、离线计算的原子控制 3、离线计算的数据质量

实时在线日志处理的几个问题



优化的几个关键点：

1、设置合理的批处理时间 2、增加Job并行度 3、使用Kryo系列化类

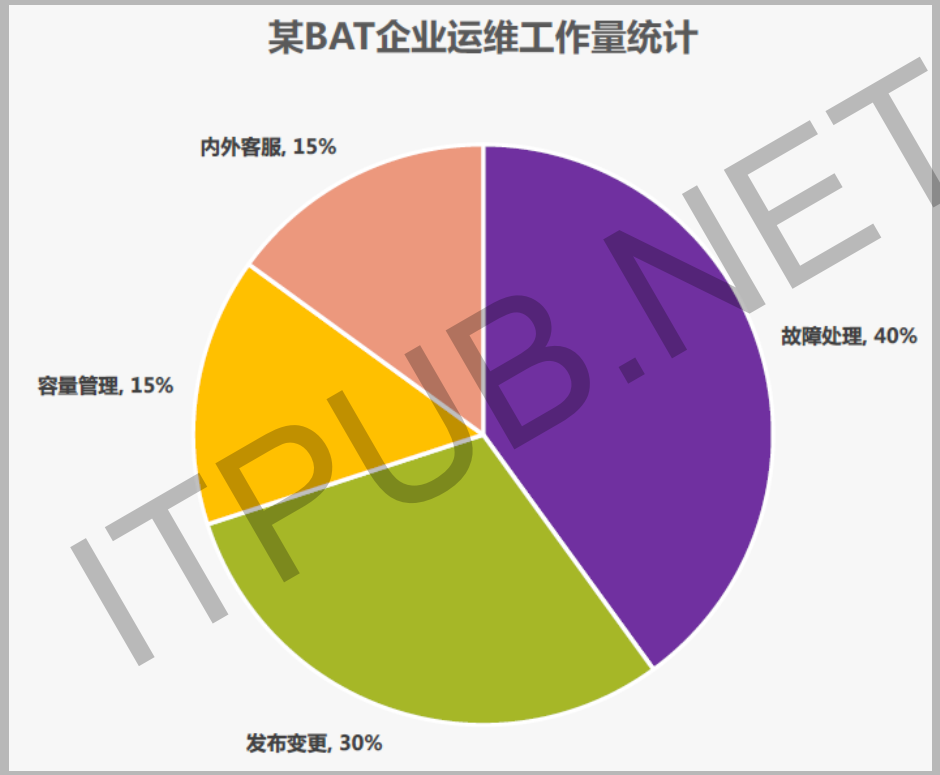
4、减少数据重复计算 5、设置合理的GC 6、设置合理的CPU数量

04

知识库与故障自治管理

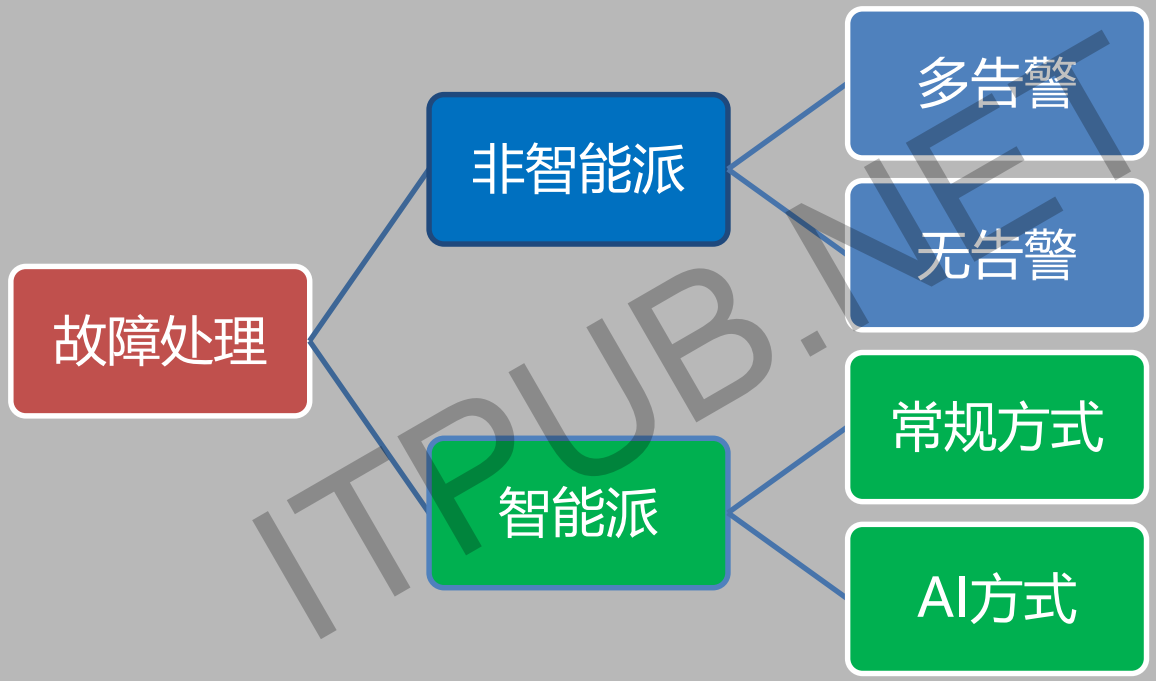
如何从错综复杂的运维数据中形成知识库

运维工作量统计



故障处理是运维人员耗时最多的第一场景

故障处理的演变



问题：
1、非常依赖经验
2、定位成本高
3、实时性不够
4、不够精确

逐步摆脱对专家知识结构化的依赖，降低使用门槛
实现知识的机器自学习，提高智能化

策略知识库的构建

基于架构
基于经验
基于概率

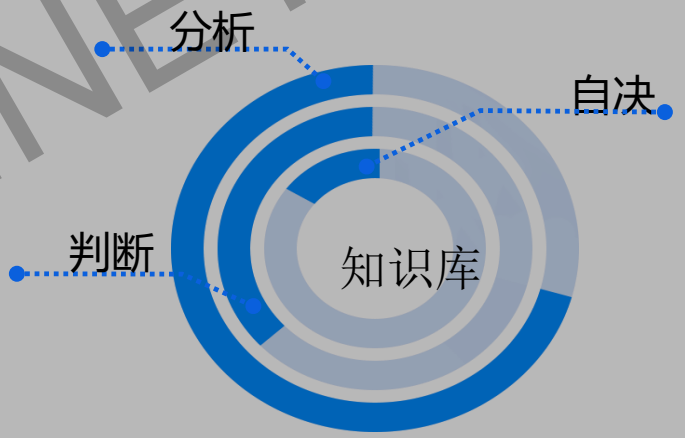
基于规范
基于分工

基于数据
基于模型

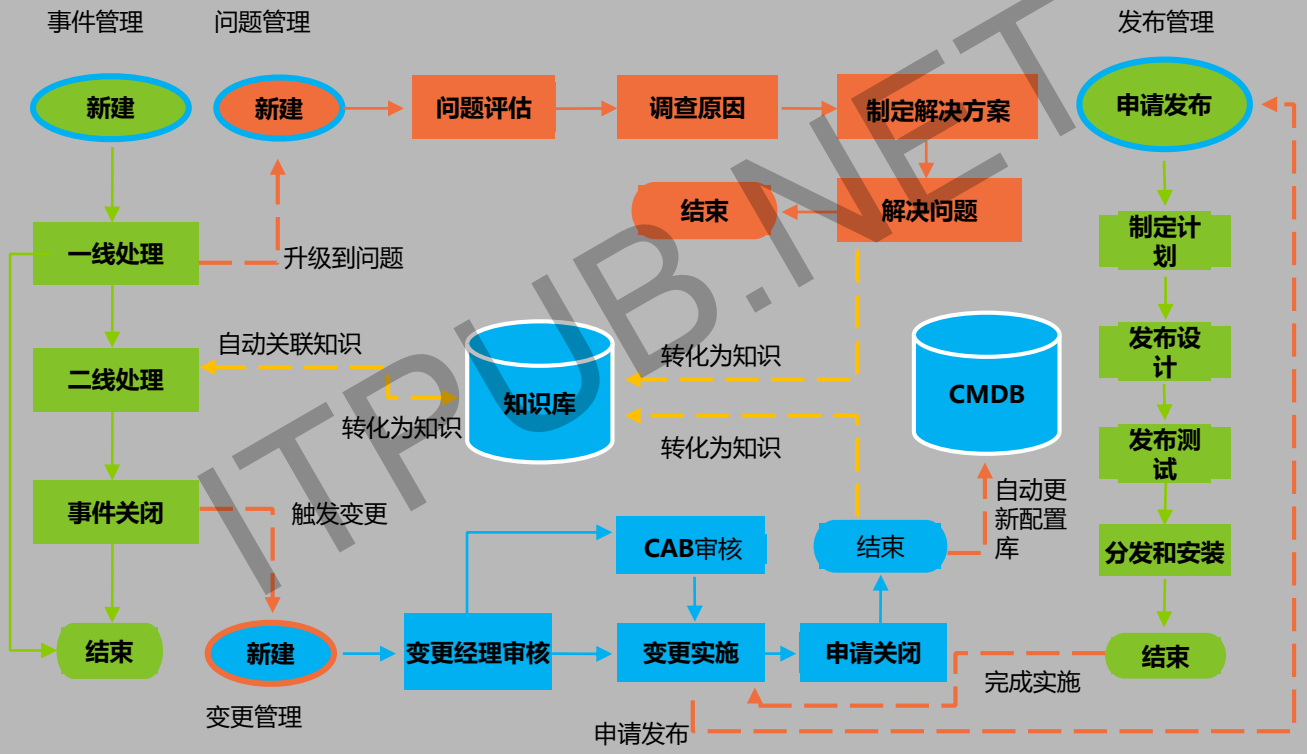
收敛告警事件

产生告警事件

提高事件处理能力



企业内部知识库构建



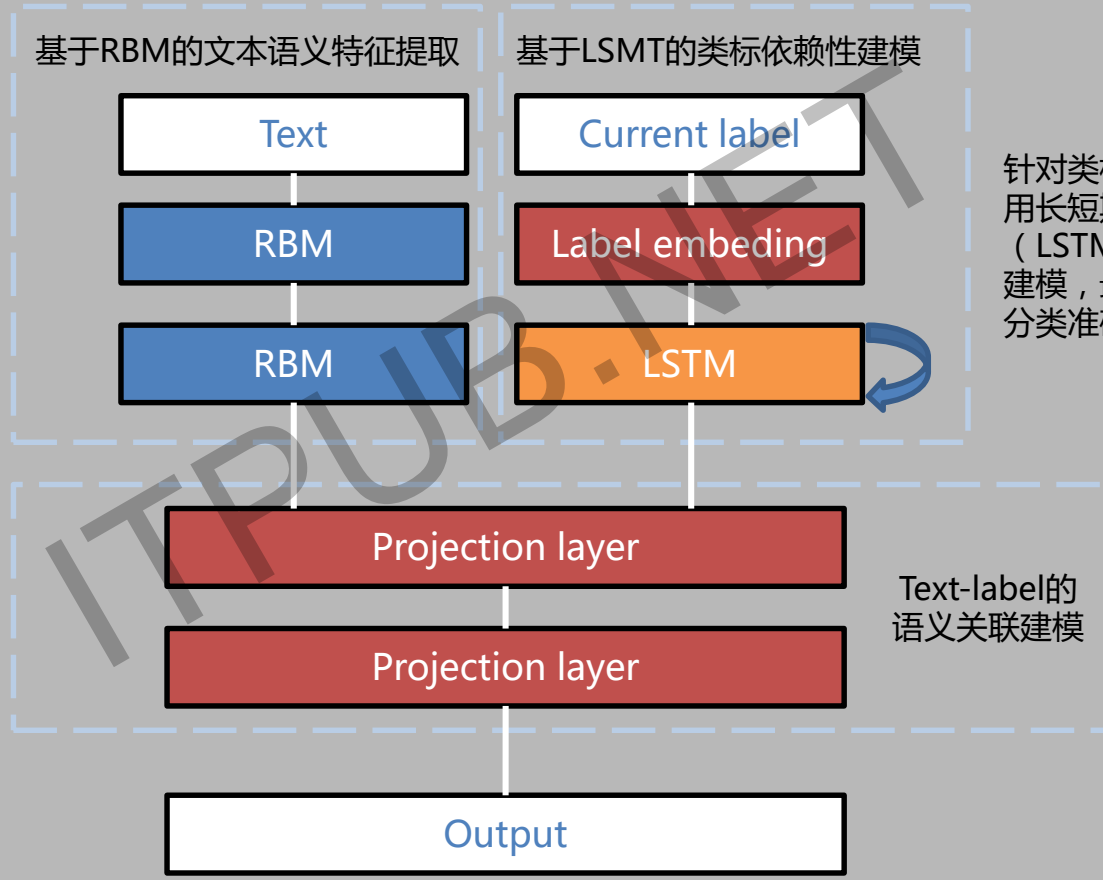
突破与成果--自动分类

- **互联网数据**：对数据信源进行标注，并爬取文章自带的标签
- **本地上传的文件**：在上传过程中可为该文件进行标签的标注

自动分类模型：针对性的采用基于RBM+LSTM的深度学习的方法，面向海量信息进行机器学习自动分类训练

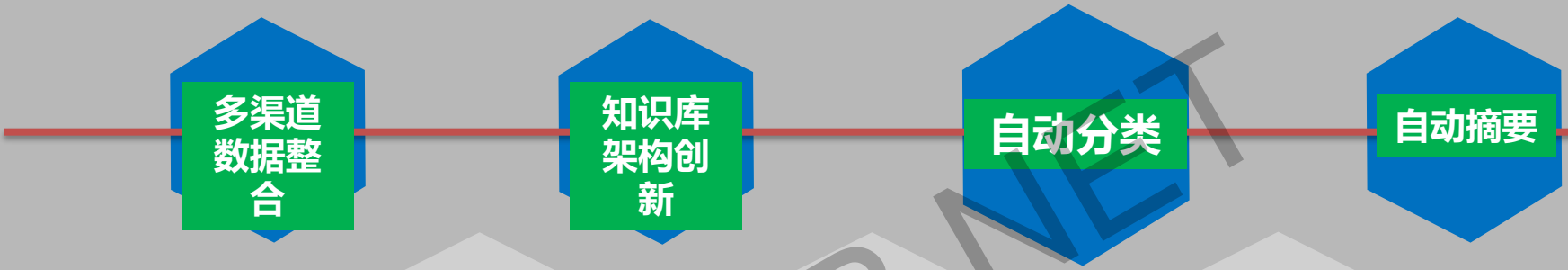
突破与成果--自动分类

采用受限玻尔兹曼机（RBM）和池化（pooling）的操作对文本的语义特征进行恰当的提取，降低文本的特征的稀疏性和冗余性



针对类标之间的依赖性，采用长短期记忆人工神经网络（LSTM）对类标进行有效地建模，最后进行分类，提高分类准确率。

突破与成果--自动摘要



01 网络类

02 操作系统类

03 数据库类

04 安全优化类

05 存储类

突破与成果--自动摘要

技术文章

01

论文格式文章

针对有提示性词语“摘要、内容、综述”等，将摘要字段后面的一段话提取为摘要

非论文格式技术文章

- 没有提示摘要的词语时，有一级标题和二级标题的，选择第一个一级标题前面的内容，加上文章的一级标题和二级标题为摘要
- 没有提示摘要的词语时，有一级标题无二级标题的，选择第一个一级标题前面的内容，加上文章的一级标题和一级标题下每段首句为摘要

突破与成果--自动摘要

02

技术类或解决方法类

含有提示性词语的文章

在后台提供属性配置功能，建立属性规则，可以自主添加关键词属性，针对涉及关键词属性部分进行摘要

无提示性词语的文章

- 选择文章段首内容加上文章的一级标题，加尾段，作为摘要
- 选择文章段首和文章一级标题，以及一级标题下每段首句内容，再加尾段，作为摘要

问答结构类文章

一般选择问题和问题答案的每段首句作为摘要，如果问题有多个回答则将多个回答的每段首句罗列成为摘要

kB类的文章

选择KB标题、现象和解决方案或方法的每段首句内容作为摘要

AIOps的应用场景分析

效率提升方向

智能变更
智能问答
智能决策
容量预测

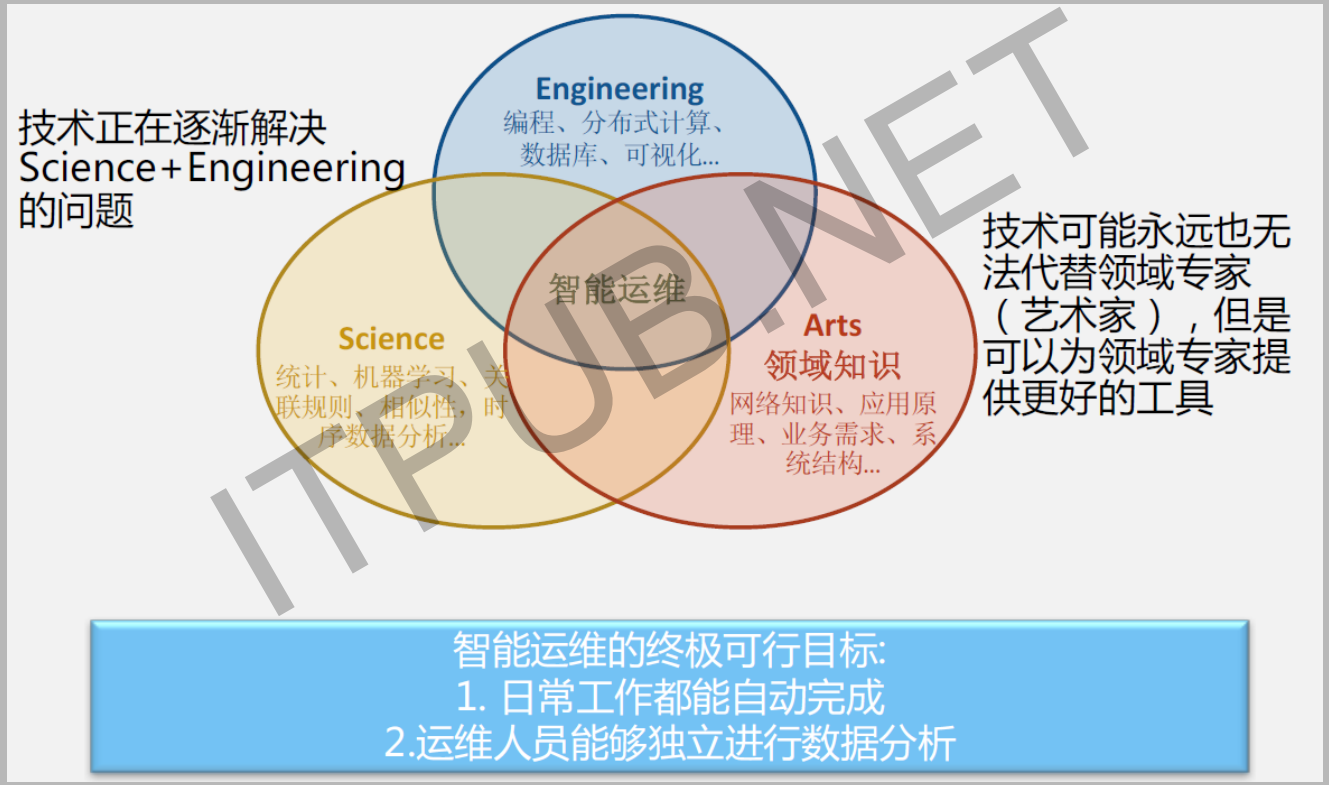
质量保障方向

异常检测
故障诊断
故障预测
故障自愈

成本管理方向

成本优化
资源优化
容量规划
性能优化

减少对人的依赖，信任机器，实现自判自断自决



THANKS