YonghongTech
Talk with Data

智慧运营 数文未来

2016·永洪科技深圳大数据峰会

大数据在金融行业里应用案例分享

高伟达软件股份有限公司 | 主讲人：方长青

目录

中国金融业务发展面临的挑战

宏观环境

利率市场化 加剧市场竞争

- 2013年，银行收入增速下降11%，总利润下降2%
- 降低金融行业手续费用

传统金融业务 遭遇互联网金融冲击

- 余额宝用户数量突破2亿，规模逼近6000亿
- 如：2014年淘宝“双十一”的移动支付交易量由2013年的15.3%上升到47.6%

创新技术 逐步渗透业务

- 云计算、社交化、移动化、大数据，支撑金融转型和新业务开创，提升核心竞争力

监管力度加强 谨防科技风险

- 银监会发布39号文件，要求每年安全可控技术应用率提升15%
- 应于安全可控应用技术的预研投入不低于IT整体投入的5%
- 截止2019年，安全可控技术应用率达75%

工行董事长姜建清 提出建设“信息化银行”

- 互联网金融发展和大数据时代对银行传统经营模式带来重大挑战
- 互联网金融则是借助大数据、云计算、社交网络和搜索引擎等信息技术优势，从商品流掌握到企业的资金流、信息流，再延伸至银行支付、融资等核心业务领域，对商业银行经营模式甚至是中介功能的全面冲击
- 从银行信息化到信息化银行，是通过信息的集中、整合、共享、挖掘，使银行整个经营决策和战略制定从经验依赖向数据依据转化

建行董事长王洪章 提出建设 “大数据行”

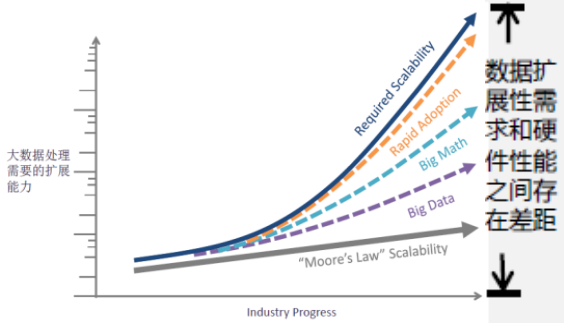
- 通过大数据实现大型银行战略转型
- 数据作为战略性资产，通过数据挖掘能力成为大型商业银行的核心竞争能力
- 银行整合完整的客户行为数据，充分了解客户消费及投融资偏好，能够据以及时为客户提供针对性服务
- 数据成为资产、不再追求精准、强调预测、关注相关性而非因果性成为大数据时代的关键思维特征

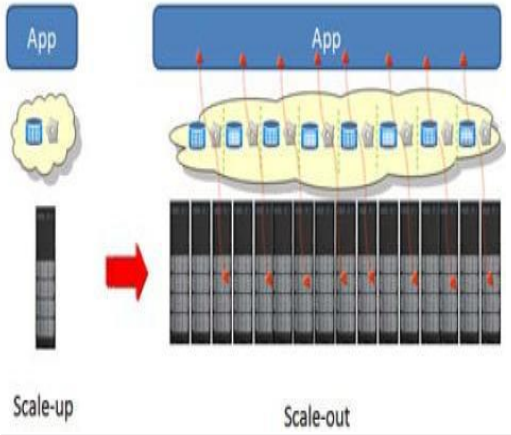
竞争、转型、创新演进的IT架构

大数据是银行解决面临问题的着力点

全球金融危机之后 ，各国银行都 在探索转型路径，寻找未来银行的发展方向。我们发现大部分银行的转型都有一个共同的特点，就是转型的设计方案都是建立在大量数据分析的基础上，数据已成为当前最突出的各种矛盾、潜力和机遇的一个集合点。

传统的数据处理系统面临的问题，呼唤新的技术



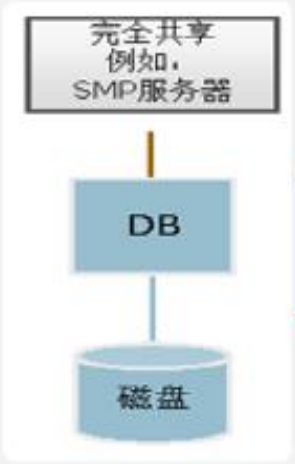


- 海量数据的高存储成本
- 大数据量下的数据处理性能不足
- 流式数据处理缺失
- 有限的扩展能力
- 单一数据源
- 数据资产对外增值

新的业务需求，需要新的大数据处理平台

数据处理技术分布式演进趋势：Hadoop成为开放的事实标准

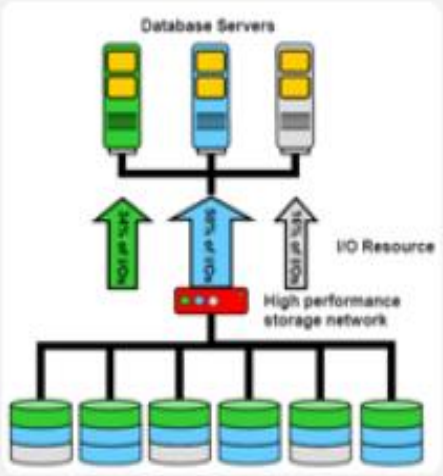
SMP



特点：单机、Scale up

- 性能存在瓶颈
- 扩展性差

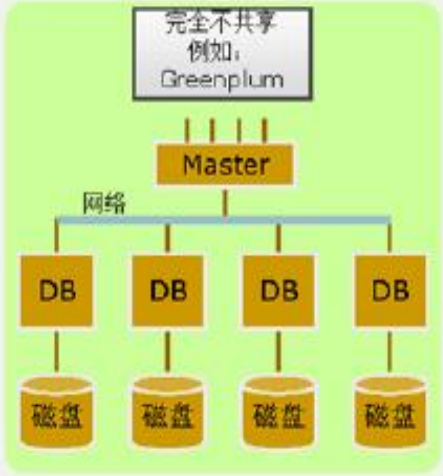
SMP+MPP混合



特点：集群、Share Everything

- 结构化、关系型
- Flash Cache+分布式块存储+IB

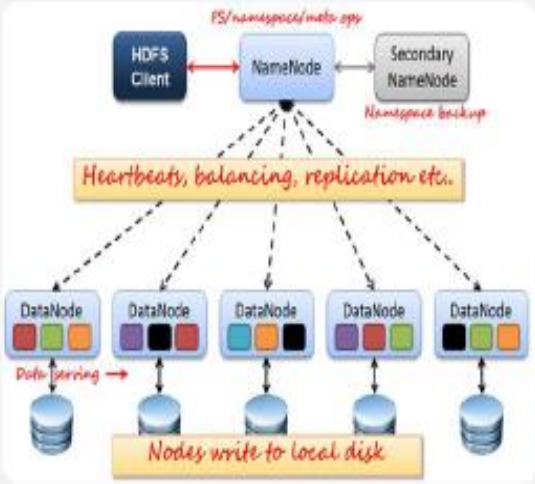
MPP



特点：集群、Share Nothing

- 结构化、关系型
- 通用的硬件

Hadoop



特点：集群、Share Nothing

- 开放、全球生态
- 结构化、半结构化、非结构化
- 高性能、实时

四大新技术发展逐步渗透金融业务

移动互联

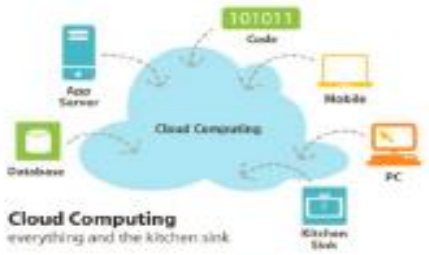
70+亿 用户
接近全球人口总数

78% CAGR 数据增长

社交

社交即业务

86% 企业在社交媒体上开展业务



云计算

云成为新一代IT基础设施

56%中小型企业购买云服务

大数据

数据即资产

未来5年，企业间的竞争在数据层面

未来银行：客户更加移动化、个性化、社交化，实时化

传统客户

固定时间地点获取服务
被动接受数据
寻求更多资源
信任市场信息
被动接受传播

随时随地获取服务
分析、创造数据
寻找有意义的体验
审视细节
互动参与内容、产品和体验的创建

新客户

标准化和产业化提供服务
关注过程和步骤
被动接受信息且信息来源单一
通过客户经理联系客户
固定渠道单一交互

传统银行

新银行

运营客户

全渠道

关注场景

互动

效率

分析

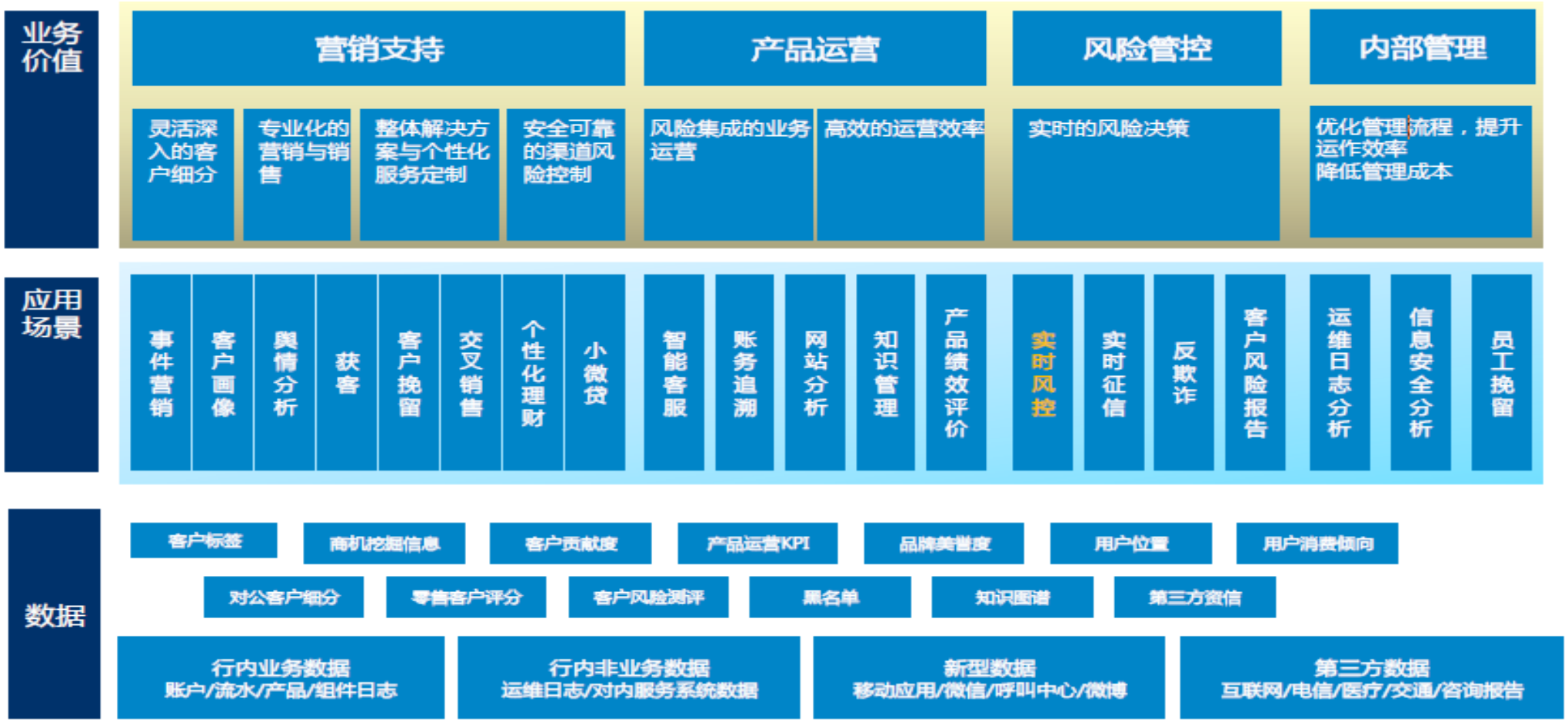
营销客户

推荐

个性化的灵活服务

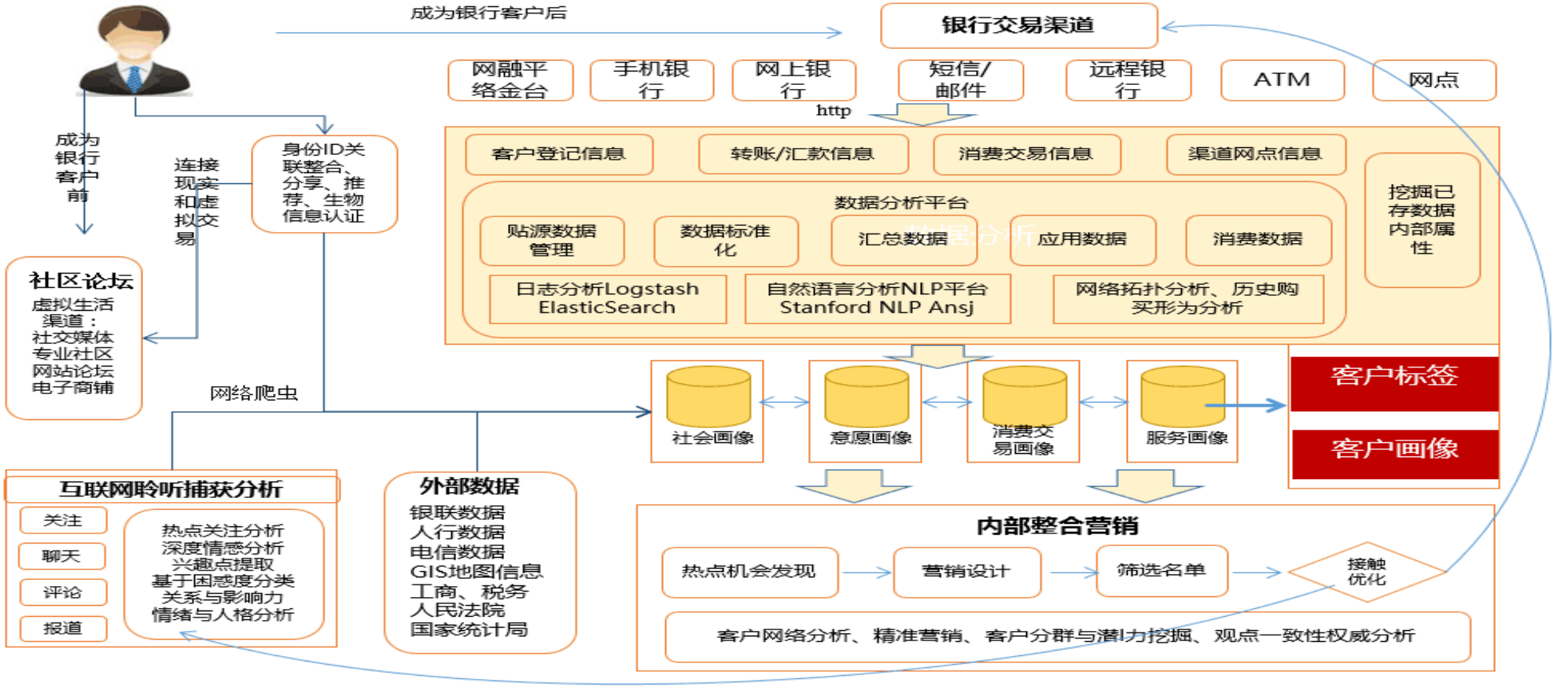
服务客户

金融行业大数据业务价值框架-全面的金融业务模型



目录

高伟达大数据客户分析平台

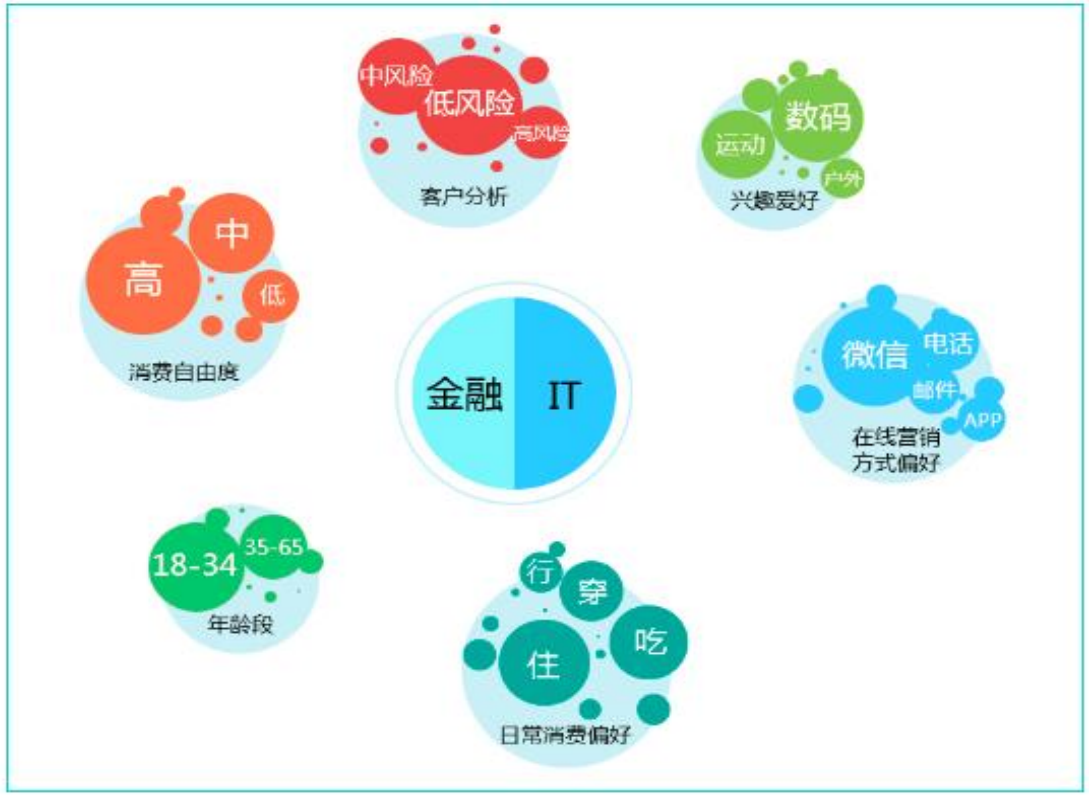


用户标签

- **人口统计学标签：**UID: 537821068，性别：男，年龄39，工作状态：在职；工作行业：IT，教育程度：硕士生，地域：湖北省，武汉市；活动区域：武汉_汉阳大道
- **信用属性标签：**职业；收入；资产；负债；学历；信用评分等。
- **金融特征标签:**资产信息特征; 收入贡献特征; 产品偏好特征; 消费行为特征; 渠道偏好特征; 生命周期特征；稳健投资客户；激进投资客户；财富管理客户。
- **资产标签：**房主：是；位置：市中心，是否拥有奢侈品：是， 是否股民：是，是否信托客户：是，是否保险客户
- **金融产品标签：**理财客户，房贷客户，车贷客户，保险客户，理财客户，差旅人群，境外游人群，旅游人群，餐饮用户，汽车用户，母婴用户。

- **兴趣偏好标签：**户外爱好者，奢侈品爱好者，科技产品发烧友，摄影爱好者，高端汽车需求者。
- **社交圈**
  - 社交圈属性： IT，篮球，红酒，军事
  - 粉丝数：1503
  - 关注数：423
  - 影响力排名：34%
  - 社交圈口碑：78
- **长期阅读喜好**
  - 喜好类别1：体育新闻；浏览比例：19
  - 喜好类别2：军事新闻；浏览比例：11
  - 喜好类别3：投资理财；浏览比例：32
- **短期阅读关注：**
  - 关注类别1：家庭装修；浏览比例：31
  - 关注类别2：出境旅游；浏览比例：26
  - 关注类别3：汽车报价；浏览比例：12

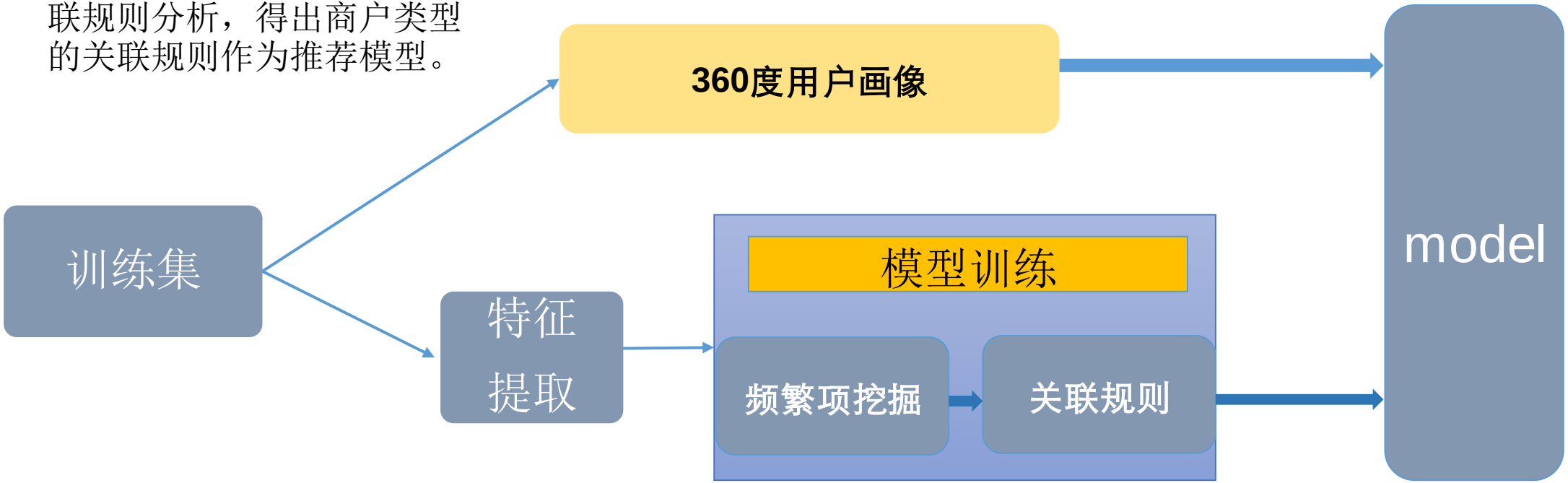
客户画像


中风险
低风险
高风险
客户分析
数码
运动
户外
兴趣爱好
高
中
低
消费自由度
金融
IT
微信
电话
邮件
APP
在线营销
方式偏好
18-34
35-65
年龄段
行
穿
住
吃
日常消费偏好


于靖童个人画像
理财
信贷资产
购物消费
投资理财
专家
生意达人
爱车一族
户外爱好者
高档酒店
吃货
男

数据分析与模型训练

从训练集中挖掘出用户频繁消费的商户类型，并进行关联规则分析，得出商户类型的关联规则作为推荐模型。



多维度的数据挖掘和精准营销

用户特征模型

获取用户业务访问特征值
构建用户个性化标签信息集合

用户消费模型

挖掘用户的消费属性并对相应的产品进行包装

预测现有客户购买相关产品的可能性

产品关联模型

获取客户其他业务的订购信息，寻找业务之间的关联性

进行产品的组合消费和引导

内容热度模型

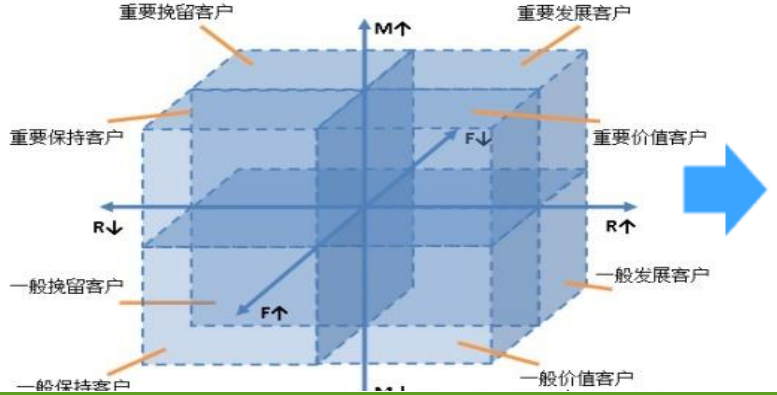
根据用户访问数据提炼热点信息、进行排行

根据热点内容进行有针对性的广告投放

价值核算模型

通过分析用户行为，评估某一内容的市场价值

预测未来可能火爆的内容



1. 用户为中心的面向主题的数据分析框架思想

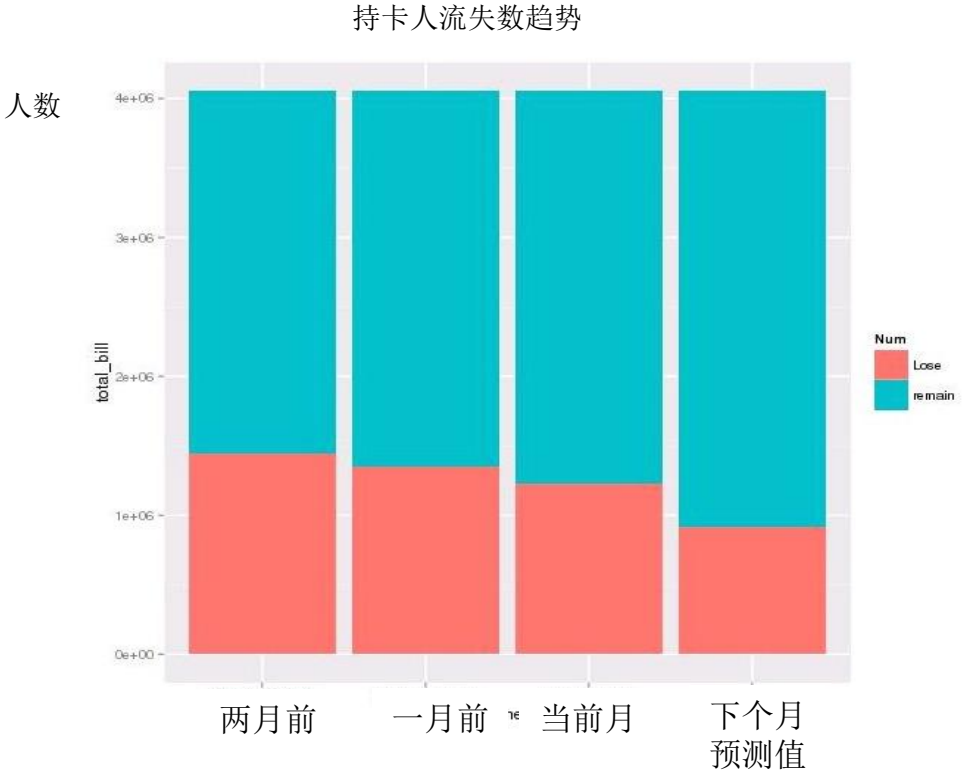
- 客户为中心的业务规划
- 面向主题的业务模型自定

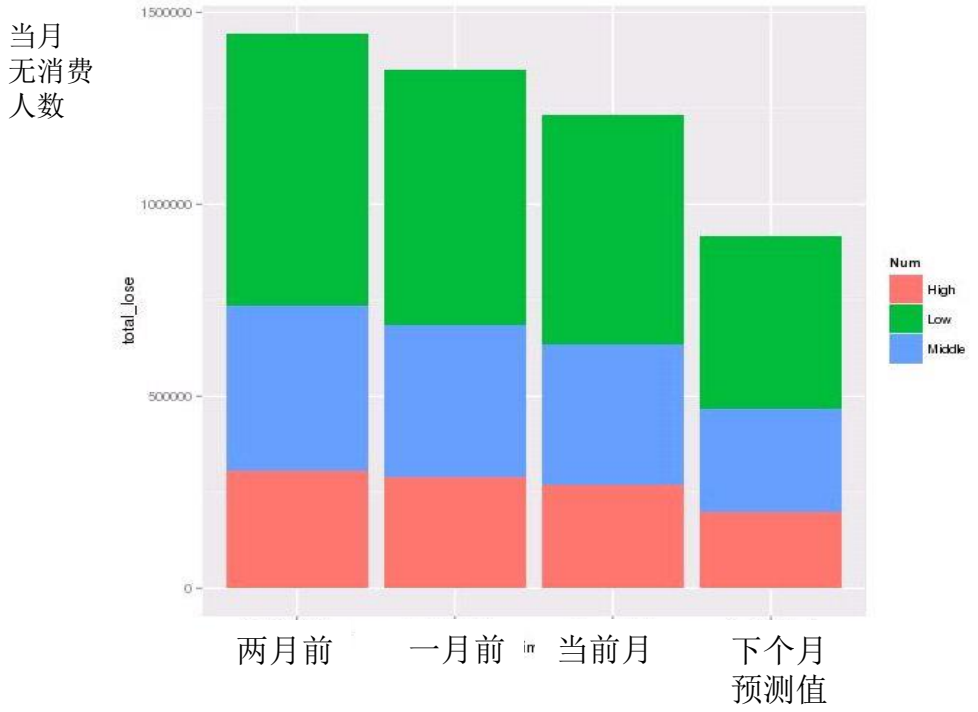
2.数据分析框架的主要事件

- 分类（Classification）
- 估计（Estimation）
- 预测（Prediction）
- 数据分组（Affinity Grouping）
- 聚类（Clustering）
- 描述（Description）
- 复杂数据挖掘

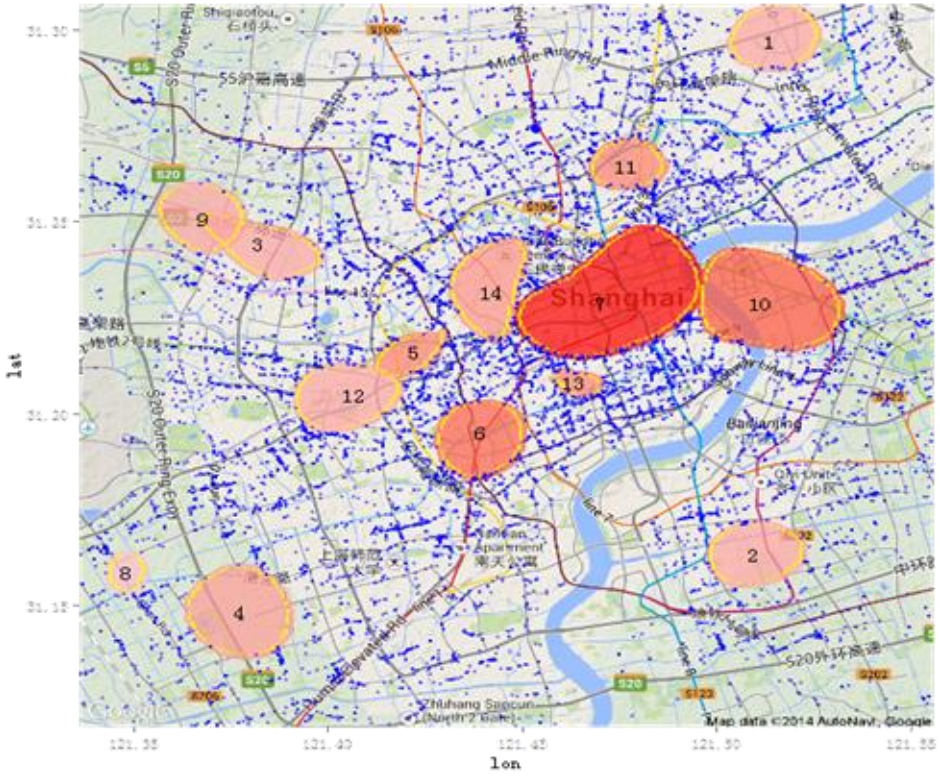
用户流失分析

- 流失预测SVM迭代1000次，在15分钟内训练和预测出所有持卡人的流失情况。





商圈分析

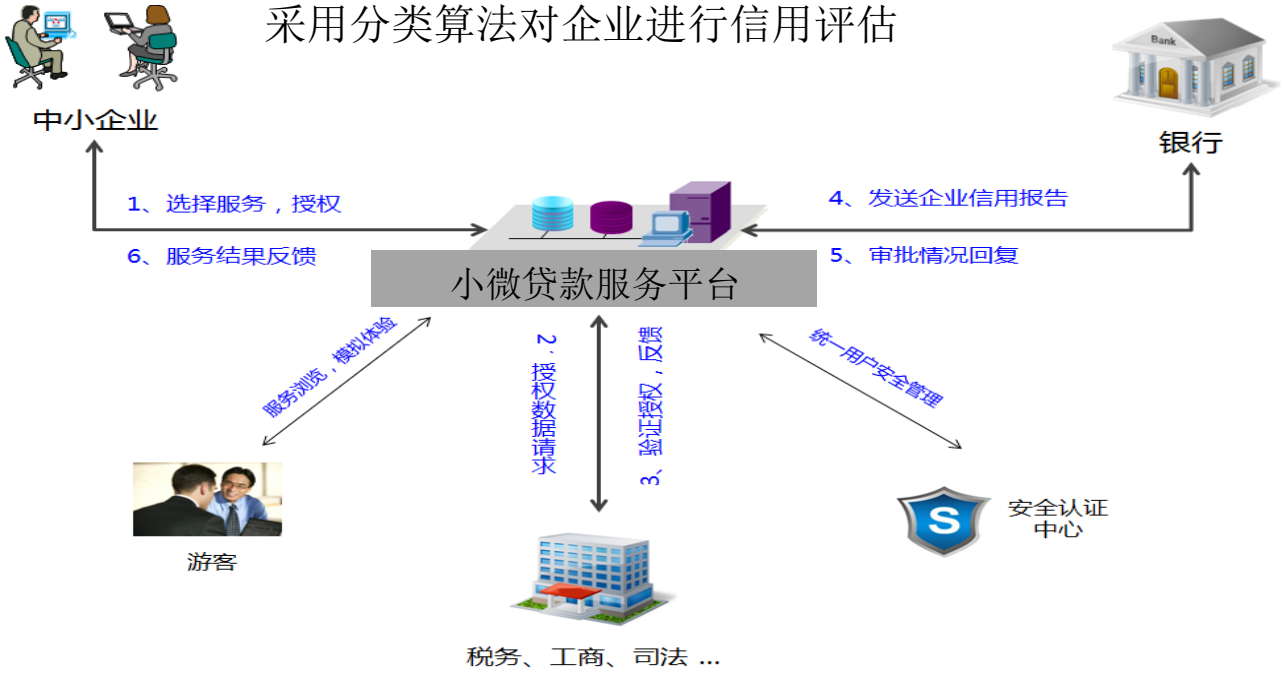


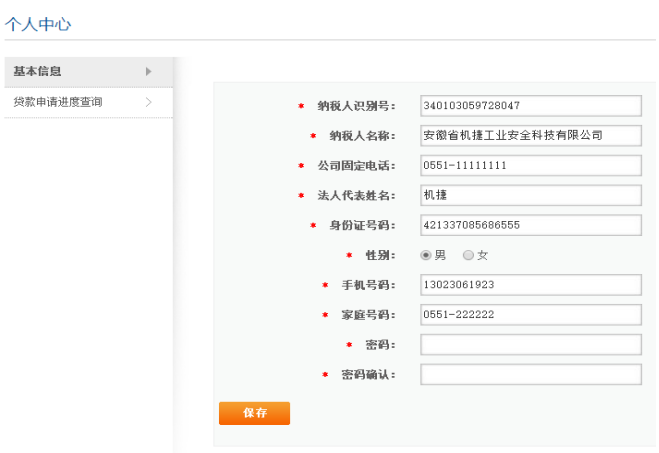
ID	名称	ID	名称	ID	名称
1	五角场	6	徐家汇	11	大柏树
2	浦东建材市场	7	静安寺-南京路-人民广场	12	娄山关路
3	金沙江路中环路口	8	虹莘路	13	新世界
4	漕河泾	9	金沙江路祁连山路	14	长寿路
5	中山公园	10	陆家嘴		

- 实时刷卡信息（来自银联）
- 定义商圈
- 商圈聚类模型分析与选择
- 模型拟合
- 动态商圈区域即时呈现，收缩变化一目了然
- 二级商圈的挖掘
- 人群密度趋势研判

大数据企业征信

选取企业的各种财务指标
采用分类算法对企业进行信用评估





目录

IT运维面对的挑战

结构复杂化

- 业务发展迅速-涉及多个业务领域、各种业务流程域、众多的业务系统相互关联，内部逻辑复杂多变
- 技术更新换代较快-从标准化的软硬件体系到Iass资源池实现虚拟化以及PasS和应用资源池化实现集群级弹性伸缩，导致技术的复杂度在快速增加造成日常操作不便和学习成本增高

数据碎片化

- 各个业务流程域、应用系统间、运维工具间的数据孤岛，在跨领域协作时，由于信息不对称，导致大量的理解偏差和额外的沟通成本
- 缺乏从业务至应用、服务器、网络的端到端分析的全景视图，导致对系统整体的理解存在一定偏差，不利日常的故障处理与分析

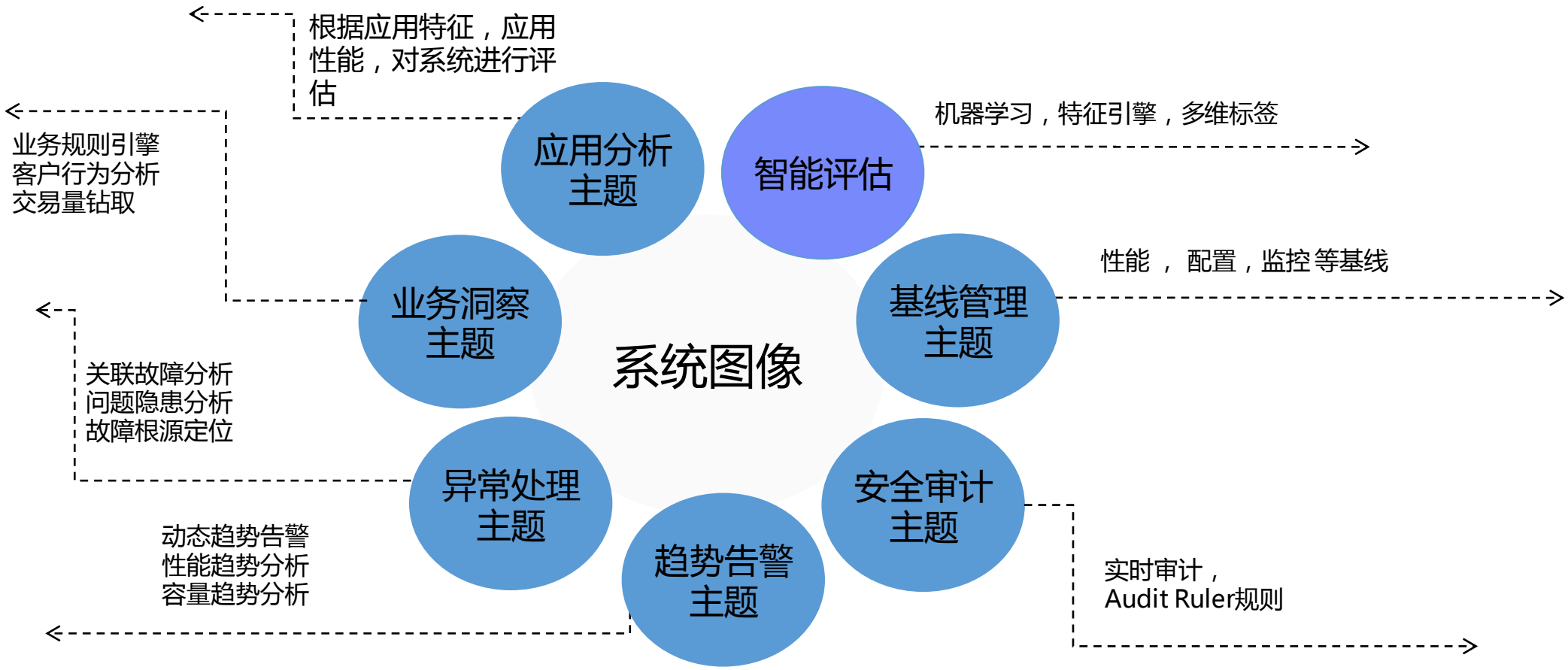
变化常态化

- 业务和新技术的迅速发展和诞生，导致了系统版本需要频繁变更管理成本增大

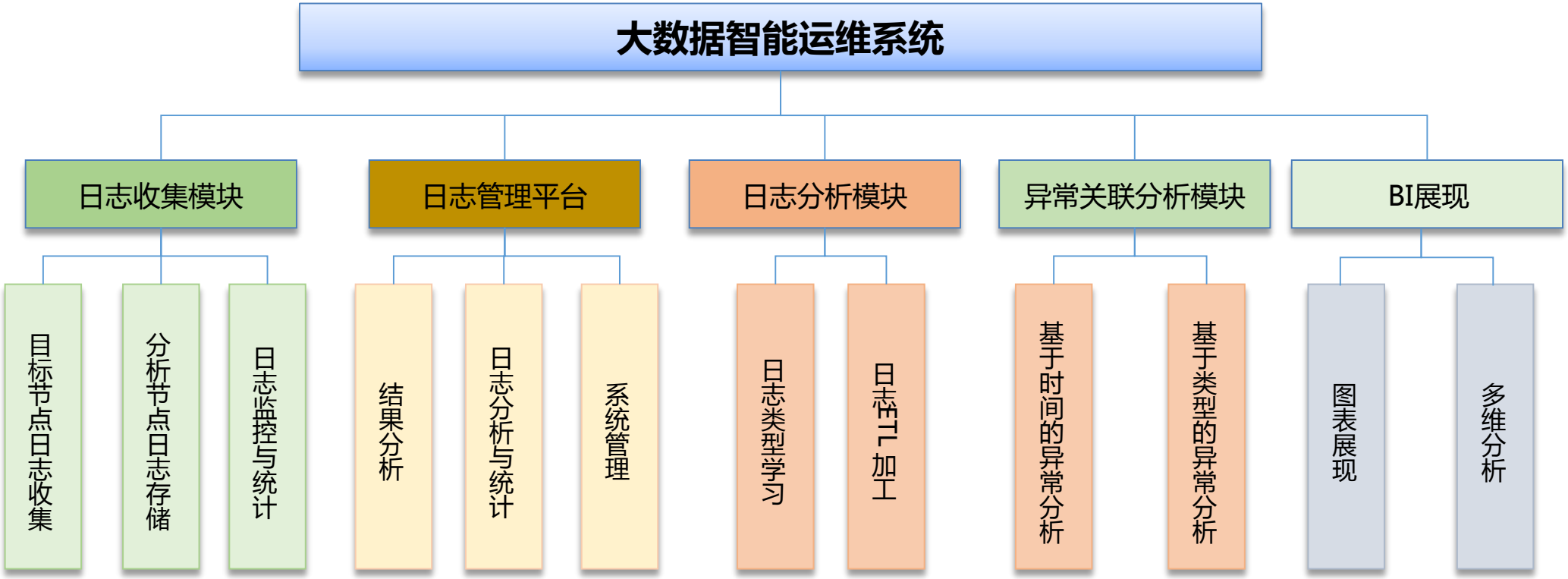
机制板结化

- IT运维工具的升级，需要经历漫长的需求调用、概要设计、详细设计、开发、测试、实施、试运行的过程，导致上线周期长，资源开销大，市场响应速度慢， IT运维产品失去活性，难以发展和适变
- 由于IT运维标准、规则一刀切，IT运维人员的个体运维经验难以融合为组织知识资产，失去活性，专业能力难以发展和适变

大数据IT运维体系–系统画像



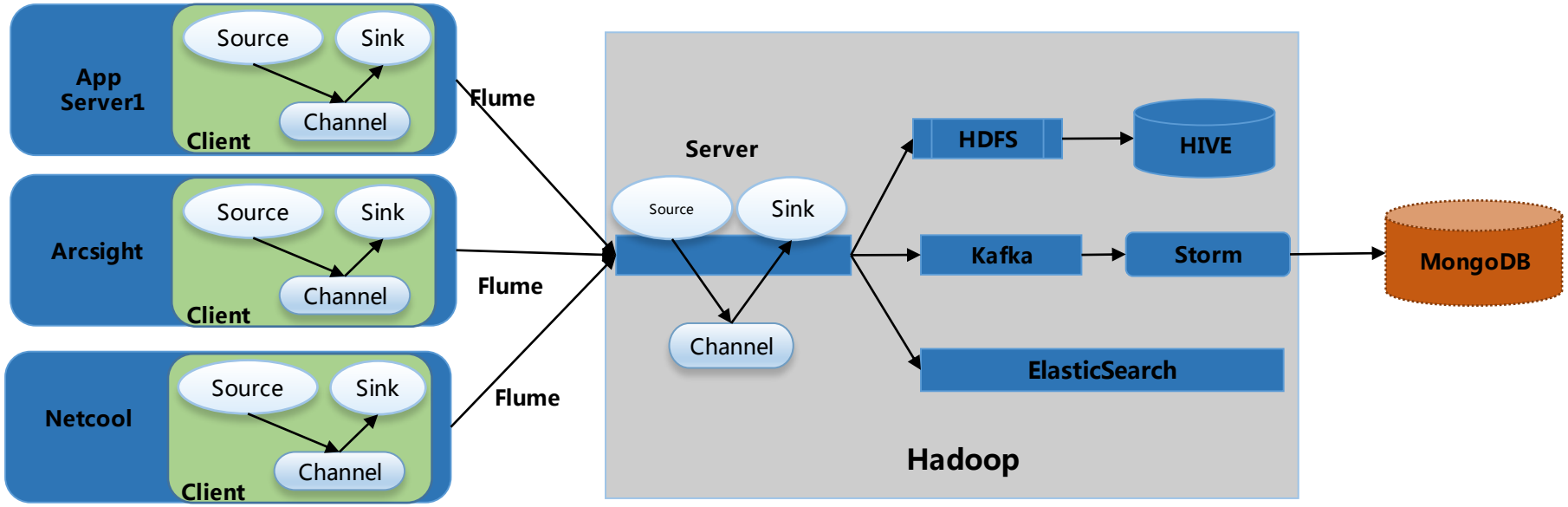
大数据智能运维系统功能



原始日志信息采集示意图

原始日志信息采集方案

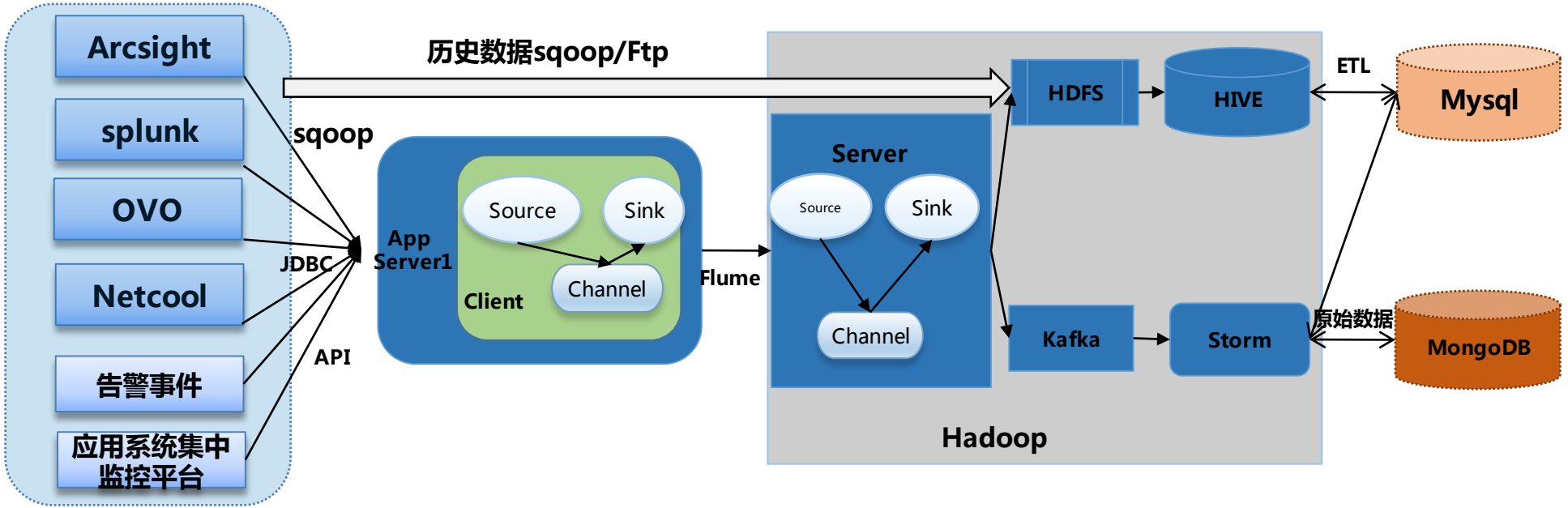
通过在每台主机上部署采集点（如主机性能数据）或从监控系统（如Netcool）直接抽取的方式，可以将系统日志发送到统一的Syslog收集器中，在Syslog收集器中配置Syslog转发到Flume Agent中，通过使用Flume的Syslog Source以及HDFS Sink,可将采集到的Syslog信息直接存入HDFS中。同时存入MongoDB和ElasticSearch中，MongoDB和ElasticSearch主要用于短期日志信息的查询，一般存储1-3个月，定期进行清理。



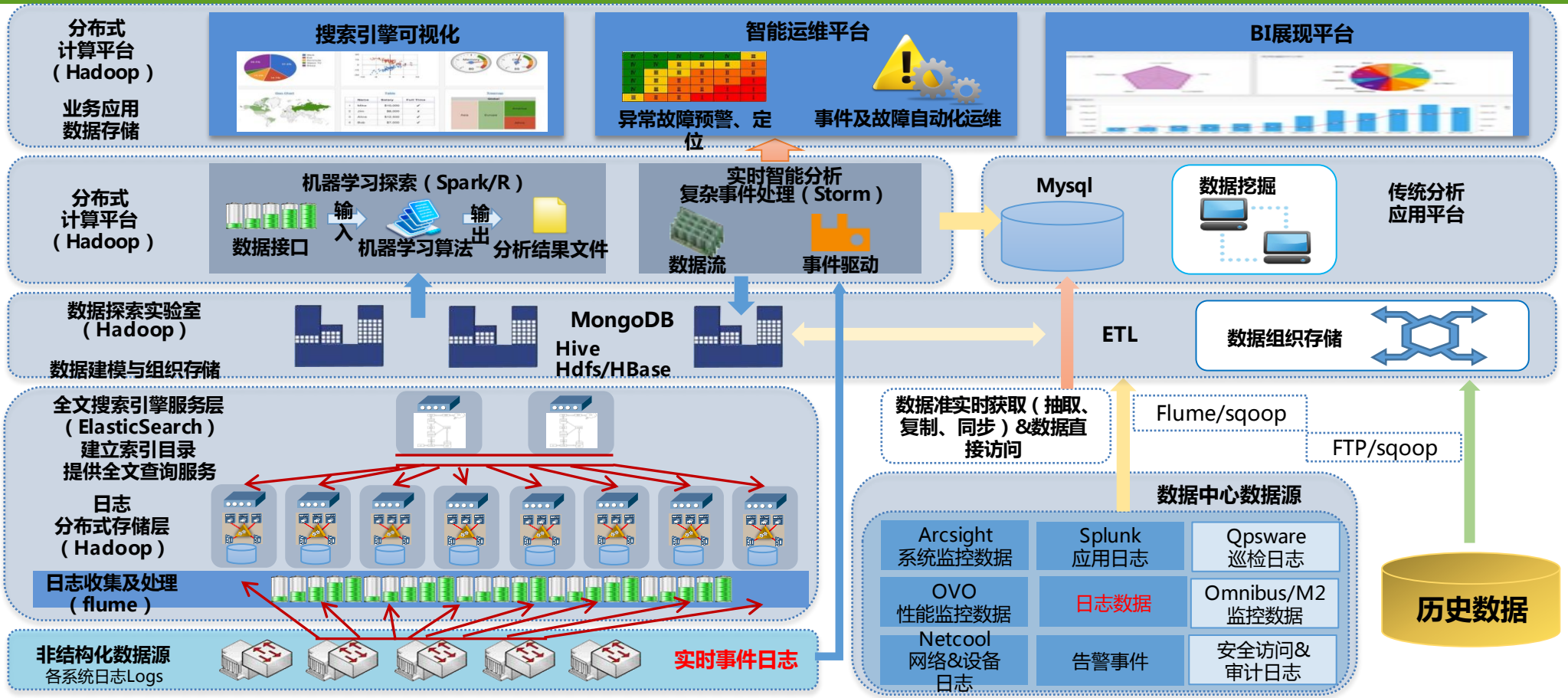
监控数据采集存储示意图

监控系统数据采集存储方案

监控系统数据是通过监控系统（如： **Arcsight 、 splunk 、OVO、 Netcool 等**）的数据采用Sqoop、JDBC接口等方式抽取到Flume的客户端的某一固定目录下，再通过Flume Agent监控其动态变化，将其变化的数据传输到Flume Sever端，最后通过Flume的Source以及Sink,可将采集到的数据分别存入HDFS和MongoDB中，如果需要实时加工，则通过Kafka在Storm中加工后存入Mysql中。日终将HTFS的数据导入HIVE，并进行ETL加工后将数据存入Mysql中。以供分析加工或报表展现使用。同时Mysql中还存储一些配置管理、参数等数据。



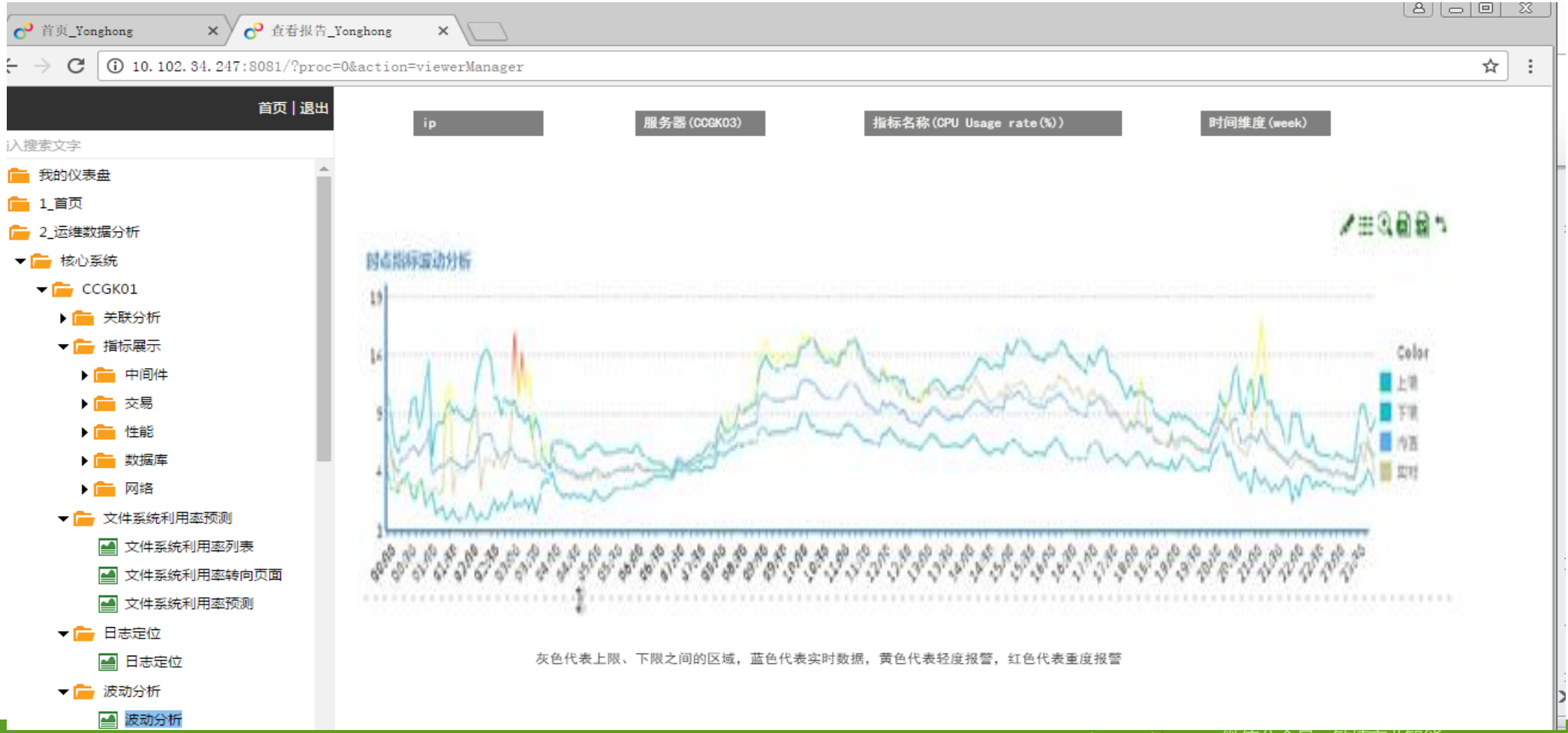
大数据智能运维系统整体架构图



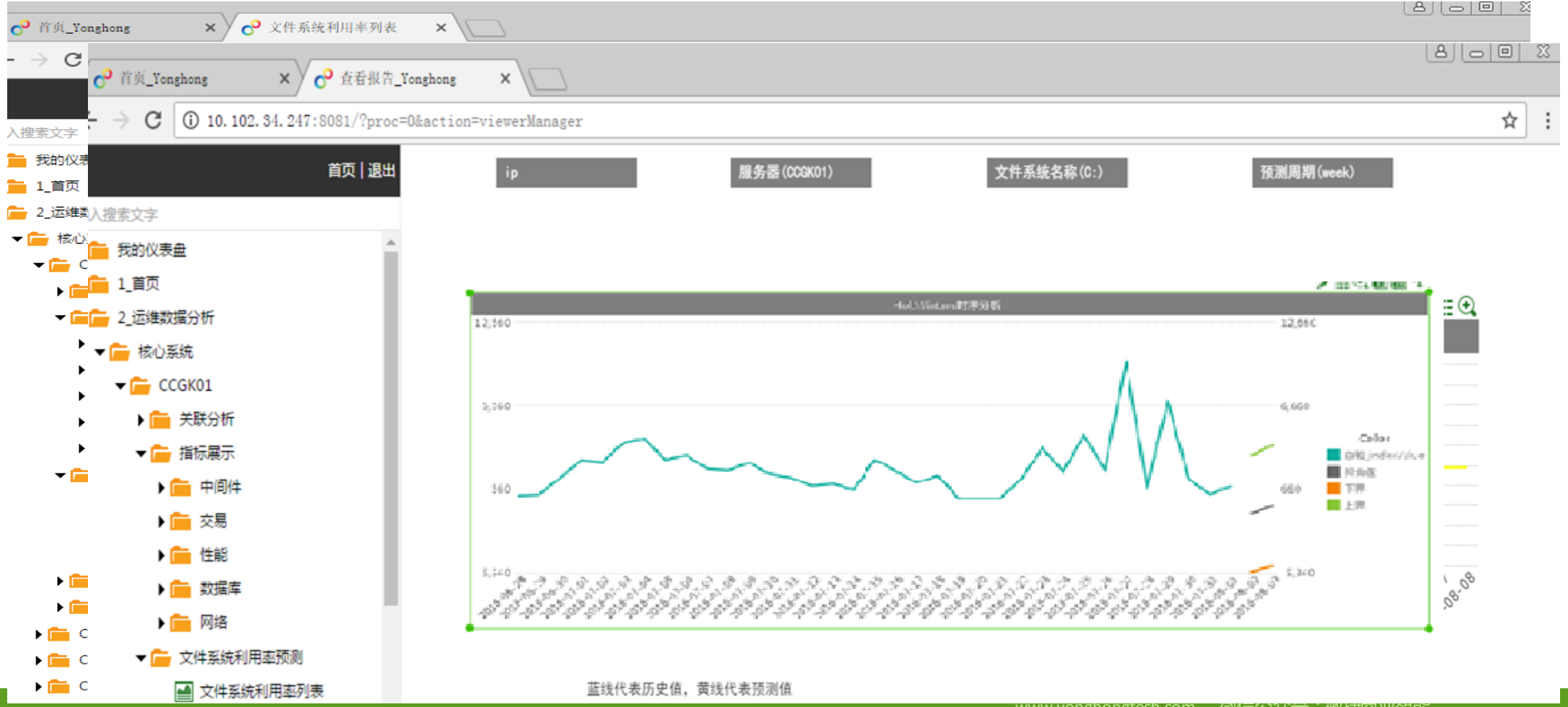
指标推荐综合展现


首页_Yonghong
查看报告_Yonghong
10.102.34.247:8081/?proc=0&action=viewerManager
首页 | 退出
输入搜索文字
我的仪表盘
1_首页
四宫格
2_运维数据分析
核心系统
CCGK01
CCGK02
CCGK03
CCGK04_PER
CCGK05_PER
CXGZ-MQ
DEP-DEV1
ESSCDB01
ESSCDB02
IVMOVOU9
redhat01
SOP05
VMIPP002
VMOVOCS
网络
3_综合查询
HIVE_查询
Mongodb_查询
中间件
实时
上限
下限
主机名 类型 指标
CCGK03 I/O /dev/hdisk0_I/O Usage rate(%)
主机名 类型 指标
CCGK03 CPU CPU Usage rate(%)

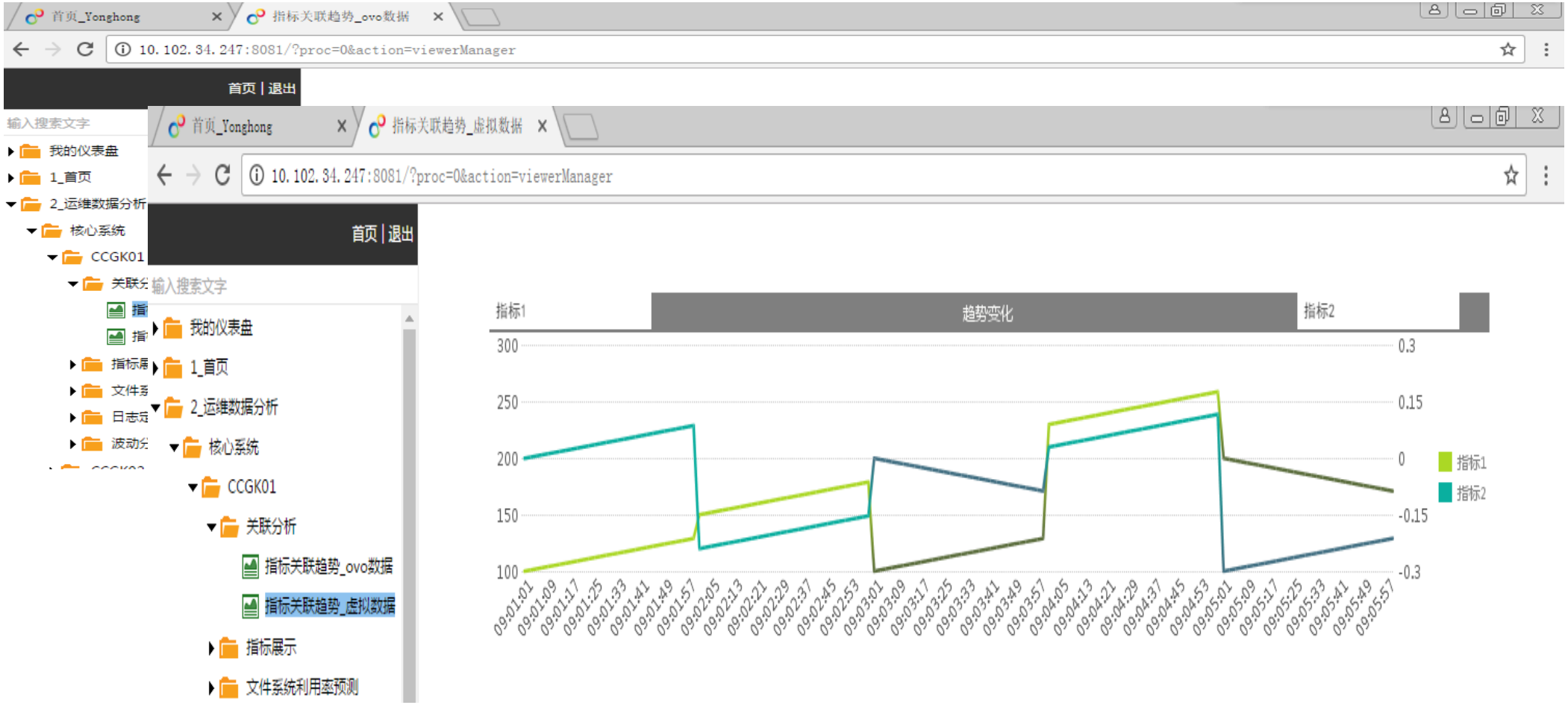
指标波动分析模型


首页_Yonghong
查看报告_Yonghong
10.102.34.247:8081/?proc=0&action=viewerManager
首页 | 退出
ip
服务器(CCGK03)
指标名称(CPU Usage rate(%))
时间维度(week)
我的仪表盘
1_首页
2_运维数据分析
核心系统
CCGK01
关联分析
指标展示
中间件
交易
性能
数据库
网络
文件系统利用率预测
文件系统利用率列表
文件系统利用率转向页面
文件系统利用率预测
日志定位
日志定位
波动分析
波动分析
Color
灰色代表上限、下限之间的区域，蓝色代表实时数据，黄色代表轻度报警，红色代表重度报警

磁盘容量/交易量预测模型



指标关联对比分析

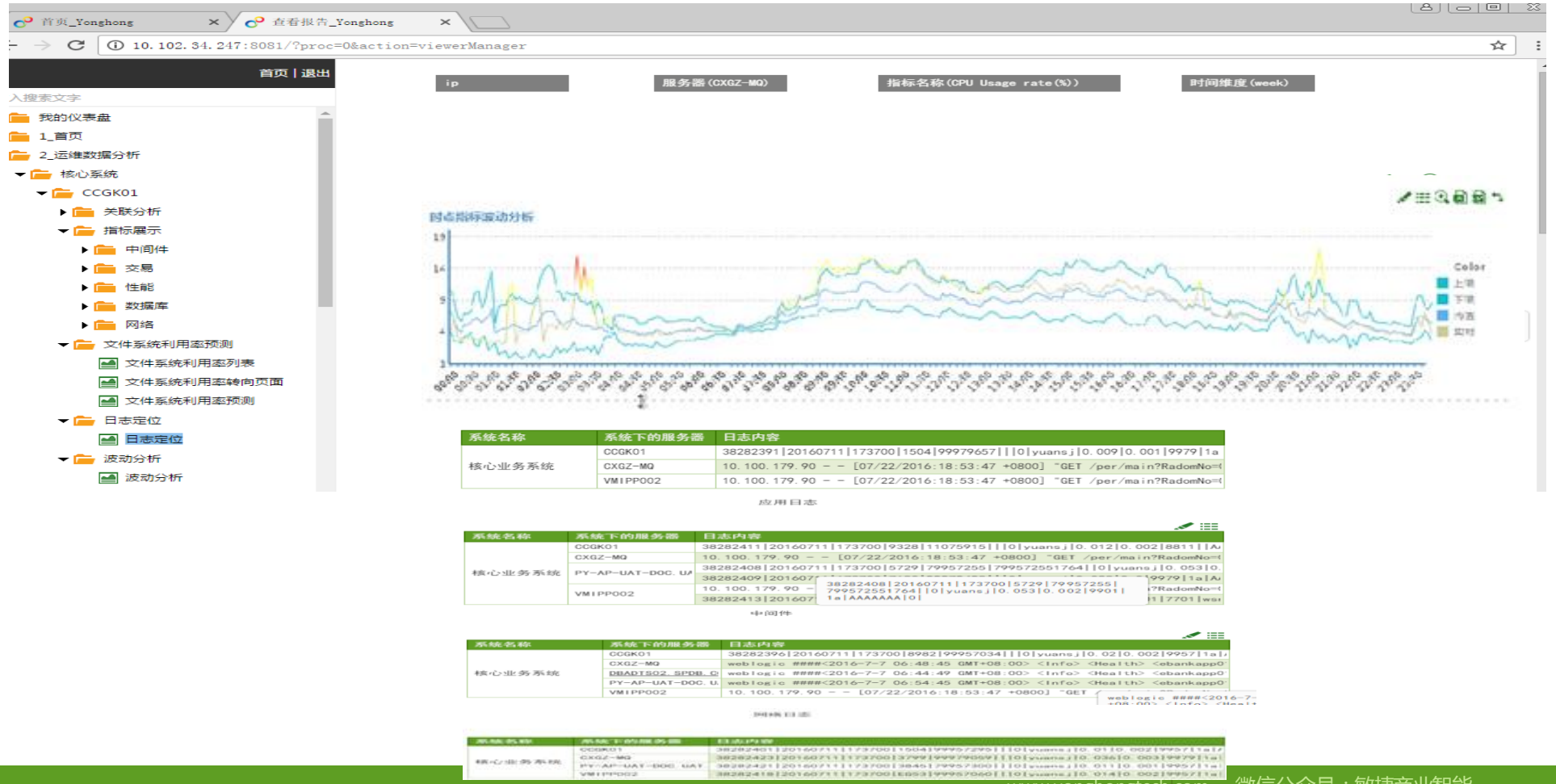

首页_Yonghong
指标关联趋势_ovo数据
10.102.34.247:8081/?proc=0&action=viewerManager
首页 | 退出
输入搜索文字
我的仪表盘
1_首页
2_运维数据分析
核心系统
CCGK01
关联分析
指标关联趋势_ovo数据
指标关联趋势_虚拟数据
指标展示
文件系统利用率预测
首页_Yonghong
指标关联趋势_虚拟数据
10.102.34.247:8081/?proc=0&action=viewerManager
首页 | 退出
指标1
趋势变化
指标2
300
250
200
150
100
0.3
0.15
0
-0.15
-0.3
指标1
指标2

故障分析与快速定位

实现应用、系统、网络全量日志数据、性能数据、监控告警归集，从日志关键字、性能容量异常、监控报警等海量信息中，快速统计、汇聚出可能造成异常的报错信息，帮助运维人员快速、便捷的聚焦关键报错，同时也为日常巡检提供便利。

利用故障关联性分析，把由同一个故障引起的所有故障日志都聚到一起，并按照发生的时间先后顺序排 序，为管理员判断故障的根本原因提供帮助。

故障/异常日志定位

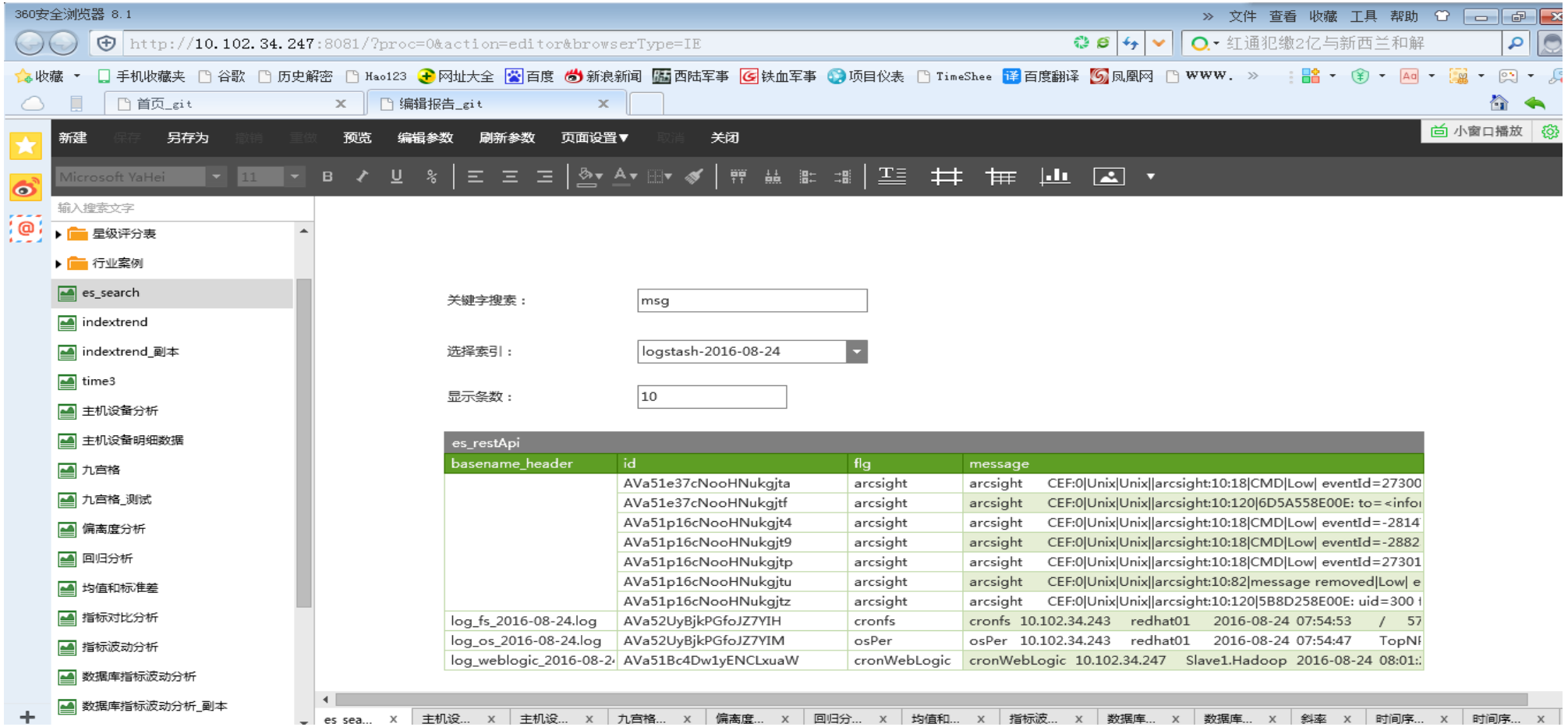

首页_Yonghong
查看报告_Yonghong
10.102.34.247:8081/?proc=0&action=viewerManager
首页 | 退出
入搜索文字
我的仪表盘
1_首页
2_运维数据分析
核心系统
CCGK01
关联分析
指标展示
中间件
交易
性能
数据库
网络
文件系统利用率预测
文件系统利用率列表
文件系统利用率转向页面
文件系统利用率预测
日志定位
日志定位
波动分析
波动分析
ip
服务器(CXGZ-MQ)
指标名称(CPU Usage rate(%))
时间维度(week)
时点指标波动分析
Color
系统名称
系统下的服务器
日志内容
核心业务系统
CCGK01
CXGZ-MQ
VMIPP002
应用日志
中间件
网络日志

系统临界值分析

分类	指标大类	指标细类	说明	数据来源
业务情况(20%)	业务发展情况(80%)	月均值环比（一天）(25%)	月均值一天的交易量比上一个月增加量30%～+∞ 25、20%～30% 20、10%～20% 15、5%～10% 10、0%～5% 5、-∞～0% 0	splunk
		月均值同比（一天）(25%)	月均值一天的交易量比去年同一个月增加量：：50%～+∞ 25、30%～50% 20、20%～30% 15、10%～20% 10、0%～10% 5、-∞～0% 0	
		峰值环比（一天）(25%)	同月均值环比	
		峰值同比（一天）(25%)	同月均值同比	
	交易异动现象(10%)	交易出现陡增陡降(50%)	每陡增一次加1，最高50	
		交易出现陡（50%）	每陡增一次加1，最高50	
	交易处理能力(10%)	交易耗时，每个月的日均值（50%）	10～ 50、5～10 30、3～5 10、0～3 0	
		业务处理能力，TPS峰值与业务处理能力上限（需要维护基础表）的比值（50%）	90%～ 50、80%～90% 40、60%～80% 20	
设备软硬件（25%）	设备老旧(10%)	设备老旧程度（运行系统）（70%）	设备使用年限：、7～+∞ 70、5～7 20、0～5 0	硬件EOS日期和软件清单来源CMDB，软件EOS清单来源人工导入
		设备老旧程度（灾备系统）（30%）	设备使用年限：、7～+∞ 30、5～7 10、0～5 0	
	硬件EOS(10%)	硬件EOS时间（运行系统）(70%)	硬件设备距离EOS的年限：已经EOS 70、0～1 60、1～2 50	
		硬件EOS时间（灾备系统）(30%)	已经EOS 30、0～1 20、1～2 10	
	软件EOS(5%)	软件EOS时间（运行系统）(70%)	操作系统EOS的年限：已经EOS 70、0～1 60、1～2 50	
		软件EOS时间（灾备系统）(30%)	已经EOS 30、0～1 20、1～2 10	
资源使用情况（20%）	CPU资源(50%)	月均值（一天）（50%）	CPU月均值使用率：50%以上 50、20%～50% 30、0%～20% 0	OVO
		月峰值（一天）（50%）	CPU月峰值使用率：80%以上 50、50%～80% 30、0%～50% 0	
	内存资源(50%)	月均值（一天）（50%）	内存月均值使用率：80%～100% 50、60%～80% 30 50%～60% 10、0%～50% 0	
		月峰值（一天）（50%）	内存月峰值使用率：90%～100% 50、80%～90% 40、50%～80% 20、0%～50% 0	
流程数据（10%）	生产事件(80%)	事件量（100%）	一个月内的事件量：20～+∞ 100、10～20 80、5～10 50、0～5 20	101系统
	变更（10%）	变更次数(100%)	一个月内的变更次数：20～+∞ 100、10～20 80、5～10 50、0～5 20	Butterfly
	维护（10%）	维护次数(100%)	一个月内的维护次数：20～+∞ 100、10～20 80、5～10 50、0～5 20	手工维护
其它（25%）	系统级别(25%)	级别大小（100%）	A(重要系统) 100、A(其它系统) 80、B 50、C 20	系统自行维护

- 对系统的各项指标进行多维的加工，再根据各系统设定的权重进行统计分类。当系统到达某一级别时提示运维人员提前进行设备或系统的健康巡检。

日志信息快速查询

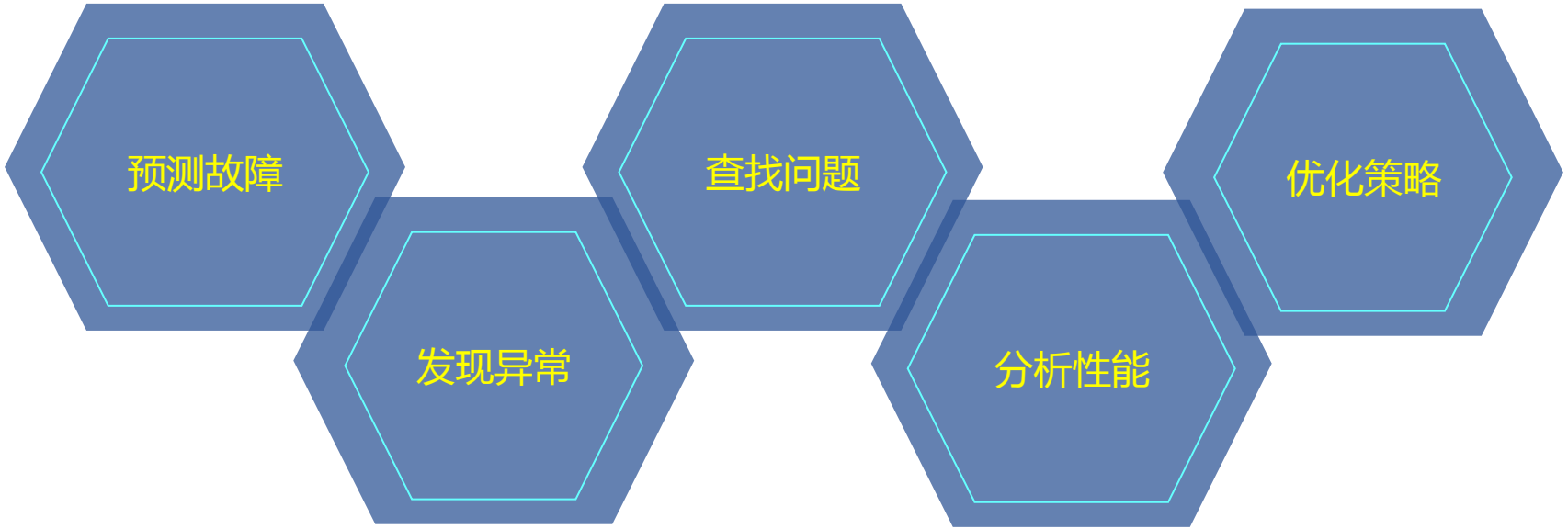

360安全浏览器 8.1
http://10.102.34.247:8081/?proc=0&action=editor&browserType=IE
首页_git
编辑报告_git
新建
另存为
预览
编辑参数
刷新参数
页面设置▼
关闭
小窗口播放
输入搜索文字
星级评分表
行业案例
es_search
indextrend
indextrend_副本
time3
主机设备分析
主机设备明细数据
九宫格
九宫格_测试
偏离度分析
回归分析
均值和标准差
指标对比分析
指标波动分析
数据库指标波动分析
数据库指标波动分析_副本
关键字搜索：
msg
选择索引：
logstash-2016-08-24
显示条数：
10
es_restApi
basename_header
id
flg
message
AVa51e37cNooHNukgjta
AVa51e37cNooHNukgjtf
AVa51p16cNooHNukgjt4
AVa51p16cNooHNukgjt9
AVa51p16cNooHNukgjtp
AVa51p16cNooHNukgjtu
AVa51p16cNooHNukgjtz
log_fs_2016-08-24.log
AVa52UyBjkPGfoJZ7YIH
log_os_2016-08-24.log
AVa52UyBjkPGfoJZ7YIM
log_weblogic_2016-08-2
AVa51Bc4Dw1yENCLxuaW

大数据对IT运维带来的实际价值

- 故障预警及预处理
- 性能指标异常提前感知

- 海量数据中集中索引多方面信息
- 跨域故障定位和分析

- 根据大数据分析结果对IT系统进行优化
- IT运维管理流程优化



- 系统行为规律发现与总结
- 历史异常规律挖掘

- 基于性能基线实现阀值优化
- 系统变更后性能影响分析

YonghongTech
Talk with Data

智慧运营 数造未来

永洪一站式大数据分析平台