ALEIYE

让 | 大 | 数 | 据 | 更 | 简 | 单

# 某运营商 XX 分公司
# 基于大数据的网络优化项目

北京数介科技有限公司

# 目录

# 1. 背景

## 1.1. 行业挑战

随着某运营商 XX 分公司（以下简称：XX 运营商）业务不断发展，其业务数据日益增长，数据量呈直线型增加，导致结果就是数据存储和处理的巨大压力。而传统的数据仓库由于本身特性问题无法线性扩容，由于各个系统独立运行，相互结构独立，呈现出竖井式结构，因此其管理难度很大，维护成本高。在 XX 运营商数据看来，已经具有上百种业务系统，IT 设备（包括交换机、防火墙、路由器、IPS 等网络、安全设备等）大概有几千台，每天数据量能达到 5、6 百 G。较好地存储、分析这些数据难度很大。

## 1.2. 网优现状及不足

移动互联网业务的快速兴起及用户对感知要求的提升、计算机大数据处理技术的广泛应用等诸多因素，以传统通信网络告警监控及性能数据统计分析为主要目的的支撑系统已无法满足运营需求，同时移动综合网管、无线网优支撑系统、室分监测系统、核心网网优支撑系统、一级架构感知系统、客服投诉处理系统等系统间松耦合的系统架构模式的导致了数据共享困难、无法端到端实施感知分析、网络故障及资源瓶颈无法准确定位等问题日益显现。

举个例子，在网优投诉工单统计分析过程中，将全 XX 省的地域划分为一个个的网格，结合各网格中产生的工单的分类、重要程度进行排名，为基站建设或者检修提供数据支撑。但目前没有系统可以很好的解决这个问题，人为处理又是一件很费时耗力的工作。

## 1.3. 国家大数据政策

2015 年 8 月 19 日国务院总理李克强主持召开国务院常务会议，讨论并通过了《关于促进大数据发展的行动纲要》，对消除信息孤岛、支持大数据产业发展、强化信息安全等提出了明确要求。要求各部门、各地方企业在发改委的牵头带动下，做好大数据技术和产业的创新和发展工作。此规定涵盖了工业、电信、金融、交通、医疗数据密集型行业。在电信行业，拥有着海量的高价值数据，例如掌握着用户的各类地理位置信息、商业活动、搜索历史和社交网络信息等大数据，具有维度丰富、群体性强、连续性好、网络行为全覆盖和关联性强等独特优势，某运营商如何更好的用好这些数据，实现数据价值，变得至关重要。

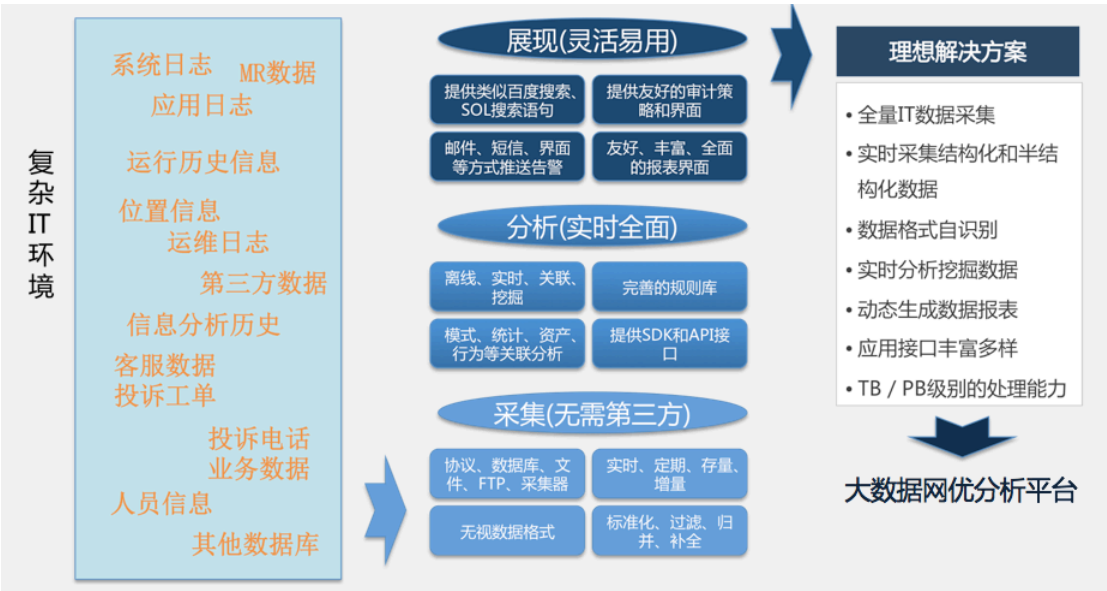
# 2. 解决方案探索

## 2.1. 理想解决方案的思路

在当前 XX 运营商复杂的 IT 环境里，各种不同的设备、业务系统时时刻刻在产生大量的数据，其中结构化数据主要为各业务数据，如 MR 数据，客户投诉故障数据、投诉工单数据等等；非结构数据包括操作系统运行日志、业务应用运行日志、安全、网络设备的运行日志、安全日志、地理化数据等等。

如此海量、不同类型的数据如何安全、可靠地进行存储、分析，高效地提取隐藏的数据价值呢？首先可以将 XX 运营商各个业务数据、IT 设备产生的非结构化数据全量采集、整合，然后还需要有完整的数据分析框架，可以完成统计、分析以及挖掘的工作，最后在展示层面有丰富多样的可视化图表很直观地进行体现。

下图是一种较为理想的解决方案示意图。




## 2.2. 大数据解决方案

### 2.2.1. 数据存储

XX 运营商有着繁杂的 IT 环境，包括了众多类型的 IT 数据，比如各业务系统中的业务数据，应用系统中的应用日志、IT 设备中的网络、安全设备的各种级别日志，将海量的、多个数据源、多种结构的数据做统一存储，统一管理是最佳的解决方案。

### 2.2.2. 数据收集

需要将数据从不同业务系统、不同设备中按相应收集方式进行采集，才能为后续的数据分析、展现做数据支撑。

- 多样数据采集

  根据数据不同的形式，需要使用不同的数据传输方式，有以下几种：

  - 数据库接入

    大部分业务数据都是由各个业务系统存储在自身独立的数据库的数据库表中，并且各数据库设计根据业务系统开发厂家不同而不同，甚至连数据库类型都不一样，因此，能够将如 Oracle、SqlServer、DB2、Mysql 等主流的数据库统一接入，是最佳的方案之一。

  - 协议传输

    网络、安全设备通常情况下会使用 SNMP 协议来获取流量、故障异常、设备性能数据，Syslog 协议会记录设备运行、警告等情况。因此，通过网络传输协议可以获取所有设备的运维信息。

  - 采集代理

    大量的数据在数据存储中以文件的形式存放着，并且有新数据产生时还会随之写入文件中。因此，如何实时动态的采集此类数据呢？通常情况下，会使用代理（Agent）方式进行采集。

- 数据预处理

数据源产生的数据为原始数据，无法直接进行数据分析。需要根据相应的数据类型进行提前做相应的处理，包括标准化、过滤、归并、信息补全步骤，将原始数据转化为格式统一、分类明确，并进行适当的过滤与归并，为后续关联分析、数据展现提供；最终输出已明确事件分类、各字段属性正确赋值，并补全了相关信息。

| 环节 | 功能要求 | 输入 | 输出 |
|---|---|---|---|
| 标准化 | 根据原始日志的来源、内容格式判断事件的类型，并按事件类型对应的属性提取相关字段内容，形成结构化数据，并最终对应为事件知识 | 原始数据（除数据库数据之个） | 数据的结构化数据 |
| 过滤 | 通过规则配置，将不影响后续分析的事件进行过滤，起到消除背景噪声、减少分析压力的效果 | 结构化数据 | 被丢弃数据只存储原始数据 |
| 归并 | 通过规则配置，在不影响分析效果的情况下将具有 | 结构化数据 | 经过归并的安全事件，只保留一条归并 |

| | | | |
|---|---|---|---|
| | 类似属性的多条数据合并为一条； | | 压缩后的信息，并更新发生次数、最后发生时间信息； |
| 信息补全 | 通过规则配置和分析，将事件与其它安全信息关联，如：发生事件的主机信息、产生操作的自然人等； | 安全事件的结构化数据 | 标准化事件 |

### 2.2.3. 数据分析

当把多源异构的数据做了整合、统一、存储后，怎么才能进行数据分析呢？其中，主要包括实时分析、离线分析以及挖掘分析三种分析。

- **实时分析**

大数据的实时分析主要包括了数据的实时流式处理、数据实时检索以及异常数据的实时告警。

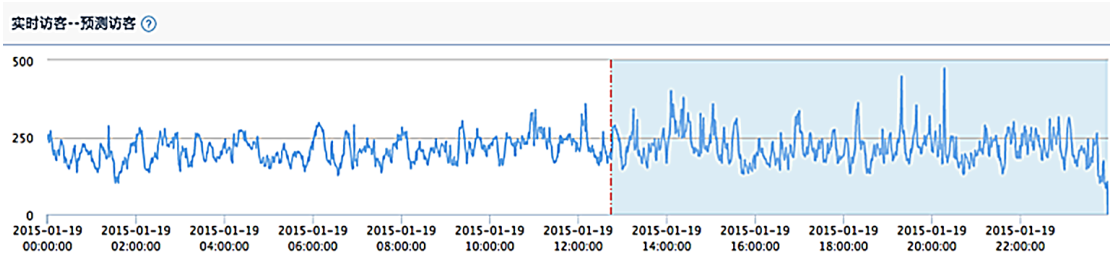
- **离线分析**

而离线分析主要是将海量的历史数据在不影响系统正常运行的情况下，通过导入的方式进行批量迁移，并且在此过程中给数据打标签，结合实时的数据进行业务扩展以及业务模型的自学习。

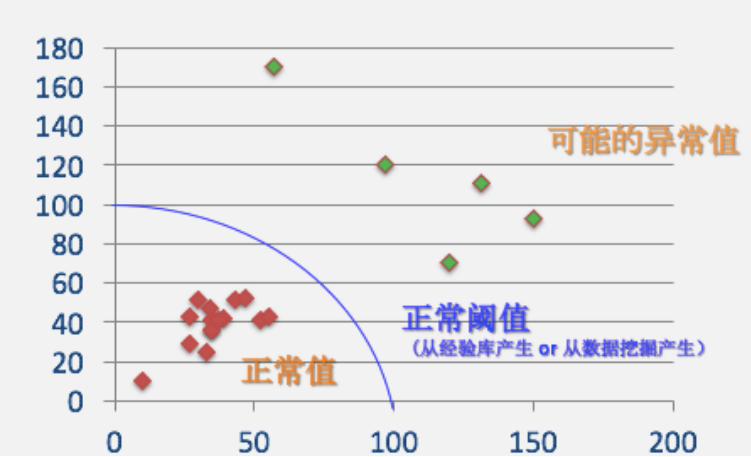
- **挖掘分析**

大数据最核心的功能即是将数据内部的价值通过挖掘技术给探索出来。其关键点包括两个，一个是趋势分析，第二个是关联分析。

趋势分析是指通过历史数据的规律性，结合相应的算法来预测未来的数据走向，并且可以数据阈值设置，为工作人员提前预防将会产生的异常或突发点。在 XX 运营商企业内部，包括 IT 运维中设备的性能预测、业务经营中的业绩预测都可以起到关键型作用。为企业决策层提供数据的客观支持。




关联分析是指按业务数据按时间或者业务特性，将分散在不同系统、设备的数据，有机的按事务发展结合起来，很多情况是人的肉眼无法观察的情况通过合理的数据算法可以洞察到。

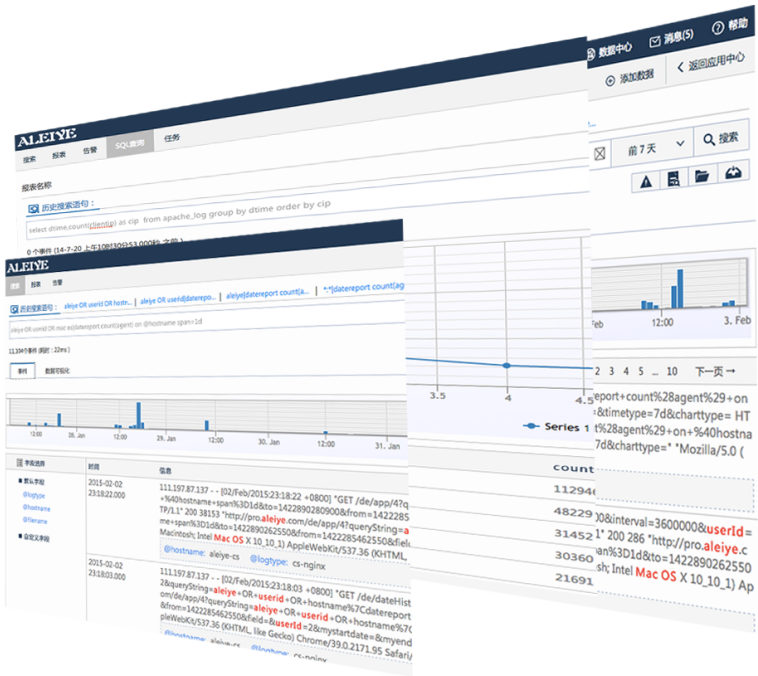



## 2.2.4. 数据可视化

数据可视化是客户与数据之间最直接的联系。在前端页面上，可以自由进行数据的设置，并对数据进行筛选、条数限制、数据公式等操作，最终完成生成可视化图表的过程。

- **检索命令快速实现图表展现**

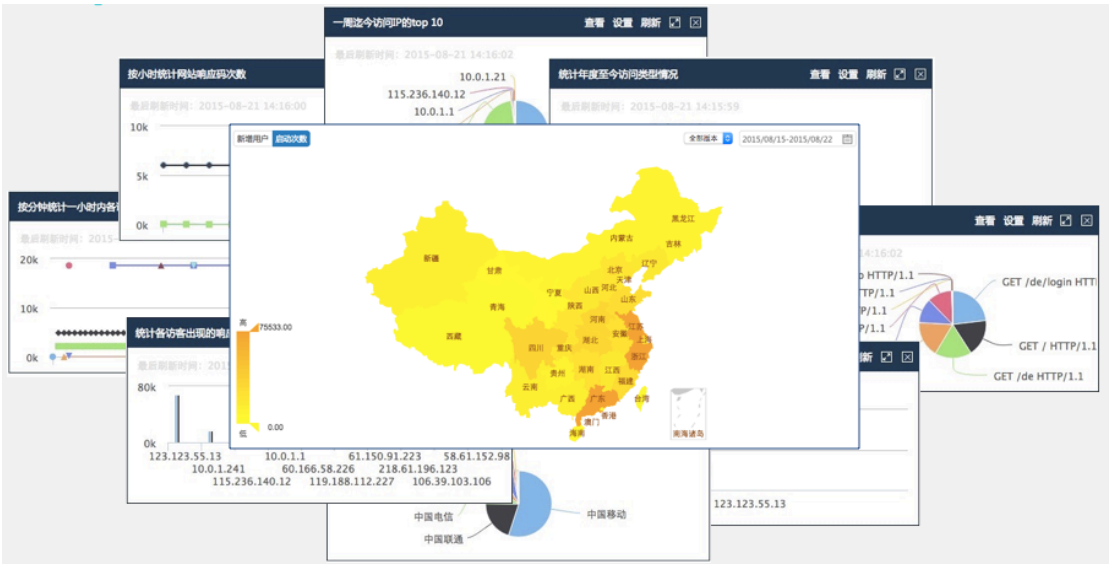
通过写检索命令来完成数据分析和数据挖掘，方便进行自主式探索分析。



- **多图表类型支持**

包括饼状、折线图、堆栈型条形图、堆栈型柱状图、分组柱状图、分组条形图、比例柱状图、比例条形图、区域图、比例区域图、散点图、

地图、气泡图、漏斗图等图表类型支持。




## 2.3. 常见解决方案

随着数据量日益增加，传统的数据解决方案已经无法支撑目前的企业数据，目前涌现出新的企业数据解决方案，例如非结构化数据解决方案和大数据解决方案。以下是传统数据分析产品与大数据分析产品的对比。

### 2.3.1. SPSS Clementine 简介

SPSS Clementine 是一款专门用于数据挖掘的产品，由 SPSS 公司收购。它集合了多个数据挖掘算法，形成了完整的、丰富的数据挖掘功能。Clementine 结合商业技术可以快速建立预测性模型，进而应用到商业活动中，帮助人们改进决策过程。强大的数据挖掘功能和显著的投资回报率使得 Clementine 在业界久负盛誉。

### 2.3.2. 实时大数据分析引擎

实时大数据分析引擎需具有强大的数据采集能力和强大的数据分析能力，需满足大数据存储、分布式计算的能力。并在此基础上，针对全行业不同业务，尤其是运营商、银行、证券、政府行业拥有大数据策略的业务解决方案。为了便于对比分析，这里采用了 Aleiye 实时大数据分析引擎。

### 2.3.3. 大数据分析引擎 VS SPSS 对比分析

| | SPSS Clementine | 大数据分析引擎 |
|---|---|---|

| | | |
|---|---|---|
| 数据类型 | 结构化<br>（数据库如 SQL、Oracle、CSV 等） | 结构化、非结构化<br>（数据库、设备、业务等全量 IT 数据） |
| 采集方式 | 手动导入、通过节点连接数据库 | 支持全自动采集、手动导入、FTP 自动获取、数据库自动对接 |
| 分析方式 | 关联分析、挖掘分析<br>（需人为触发，不支持实时、定时任务） | 实时分析、离线分析、关联分析、数据挖掘 |
| 策略中心 | 架构化数据的统计模型 | 关联规则：行为、资产、统计、模型<br>机器学习：开源大数据机器学习库 Mlib、Mahout 等 |
| 数据中心 | 结构化数据库：SQL Server、MySQL | 快速查询：Index<br>分布式非结构化数据库：MongoDB、HDFS<br>结构化数据库：Mysql |
| 交互界面 | SPSS Client 应用界面 | 灵活的、可定制化的 web 应用，并支持 API 调用，二次前端开发 |

通过以上报表，可以比较清楚地看出来，SPSS 下的 Clementine 虽然是一款具有丰富、强大数据挖掘能力的专业软件，但做为企业数据分析所需的软件应该具备如大数据分析解决方案里所具有的能力和功能。包括非结构化数据分析、可进行分布式部署能力、数据处理中可进行并行计算，最重要的是可以支撑 TB/PB 级数据量地处理。

# 3. 大数据分析引擎测试

XX 运营商网络优化部为了更好地统计分析客服投诉情况，需要自前台客服系统、网优派单系统、核心华为 V600 系统中相应提取出投诉工单以及派单数据来，进行按工单类型、状态、完成情况等不同维度进行统计分析。

因此在 2015 年 3 月份开始，对 Aleiye 实时大数据分析引擎做了系统功能、性能的测试工作。在 8 月份时基本已完成了相关的测试工作。

## 3.1. 测试环境

### 3.1.1. 硬件环境

将 Aleiye 部署到一个 IBM 的服务器上，其硬件配置如下：

| 指标 | 配置参数 |
|---|---|
| CPU | X3650M4-2*Xeon 2.9GHz |

| | |
|---|---|
| 内存 | 64G（8*8G） |
| 硬盘 | 2*300G(2.5in SAS)-12*1T(2.5 SATA) |

### 3.1.2. 操作系统

CentOS 6.3.1 64位

## 3.2. 数据来源

由投诉系统的厂家将数据从数据库以Excel形式导出，目前提供了2014年全年的数据。原始的投诉工单数据都由此而来，而投诉用户的个人、使用套餐等信息可通过信息化数据文件中获取。其中投诉区域等信息需要利用工单中的经、纬度在地图数据运算获取。

## 3.3. 需求描述

由于网络投诉涉及多个部门，并且跨了多个系统，因此投诉数据整合、投诉统计、积分计算方面便遇到了极大的问题。总结如下：

| 序号 | 一级需求 | 二级需求 |
|---|---|---|
| 1 | 投诉工单自定义指标统计分析 | 由于投诉经纬度、类型等信息存储在“服务内容”同一字段中，无法结构化数据 |
| 2 | | 从系统数据库中直接读取原始数据 |
| 3 | | 投诉统计数据无法快速形成电子文件 |
| 4 | 投诉工单地理化统计 | 全省已划分出60多万个区域网格，无法将投诉工单对应至相应区域网格内 |
| 5 | 投诉统计（A类）、网格属性（B类）、用户价值（C类）积分统计分析 | 全XX省网格通过三个维度计算积分，以季度为周期进行排名 |
| 6 | | 投诉积分数据无法快速形成电子文件 |
| 7 | 投诉引用热点分析 | 无法快速分析出XX省各地市信息引用率情况 |
| | | 无法统计故障基站、规划基站、网优数据排名情况 |
| 8 | 对规划基站提供参与意见 | |

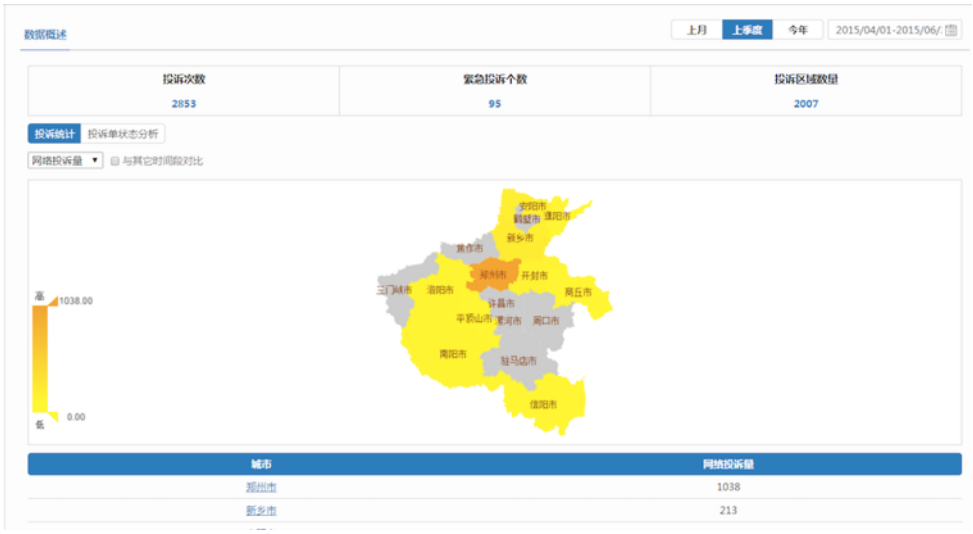
## 3.4. 已实现功能

| 序号 | 模块 | 功能 |
|---|---|---|
| 1 | 数据标准化 | 自动结构化“服务内容”数据，形成对应字段 |
| 2 | | 自 Excel 文件读取静态数据 |
| 3 | 概述 | 投诉、热点区域综合统计 |
| 4 | | 地图化显示各地市投诉情况 |
| 4 | 投诉工统计分析 | 全省已划分出 60 多万个区域网格，无法将投诉工单对应至相应区域网格内 |
| | | 投诉工单多维度分析、查询分析 |
| | | 投诉统计数据导出电子文件 |
| 5 | 投诉积分统计分析 | 全 XX 省网格通过投诉统计（A 类）、网格属性（B 类）、用户价值（C 类）三个维度计算积分，以季度为周期进行排名 |
| 6 | | 投诉积分数据无法快速形成电子文件 |
| 7 | 投诉引用热点分析 | 统计分析 XX 省各地市信息引用率、关闭引用率 |
| | | 按故障基站、规划基站、网优数据统计排名情况 |
| 8 | 对规划基站提供参与意见 | 各地市故障基站、规划基站排名 |

### 3.4.1. 投诉统计

根据网优投诉系统的工单数据，可进行以下维度数据分析：

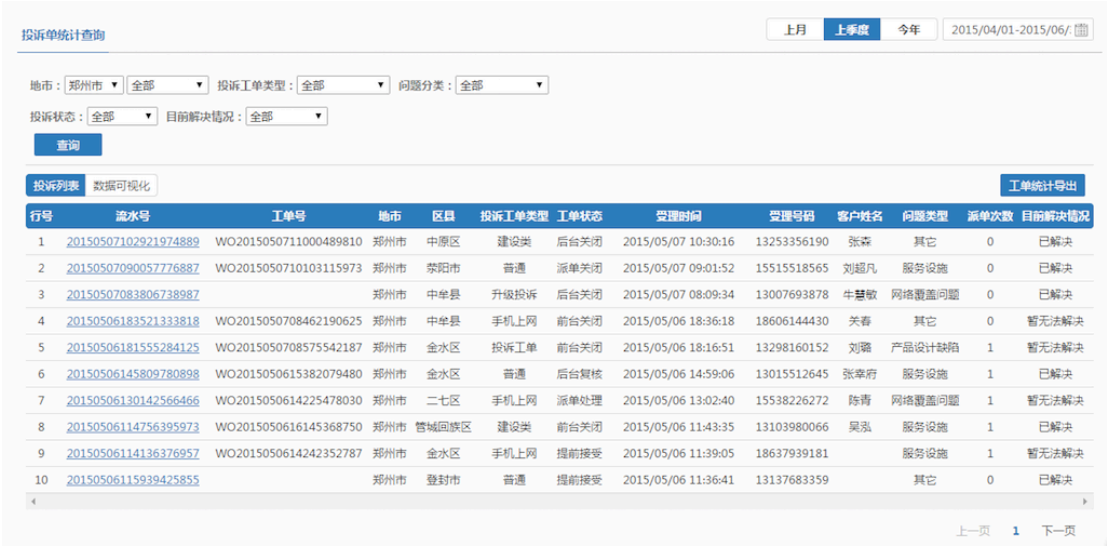
- 全省周期性投诉次数统计
- 紧急投诉次数统计
- 周期内投诉区域分布
- 各地市投诉量分布统计




## 3.4.2. 投诉工单分析

根据网优投诉系统的工单数据，可进行以下维度数据分析：

- 投诉工单搜索查看
- 工单各状态对比分析
- 投诉工单下载归档

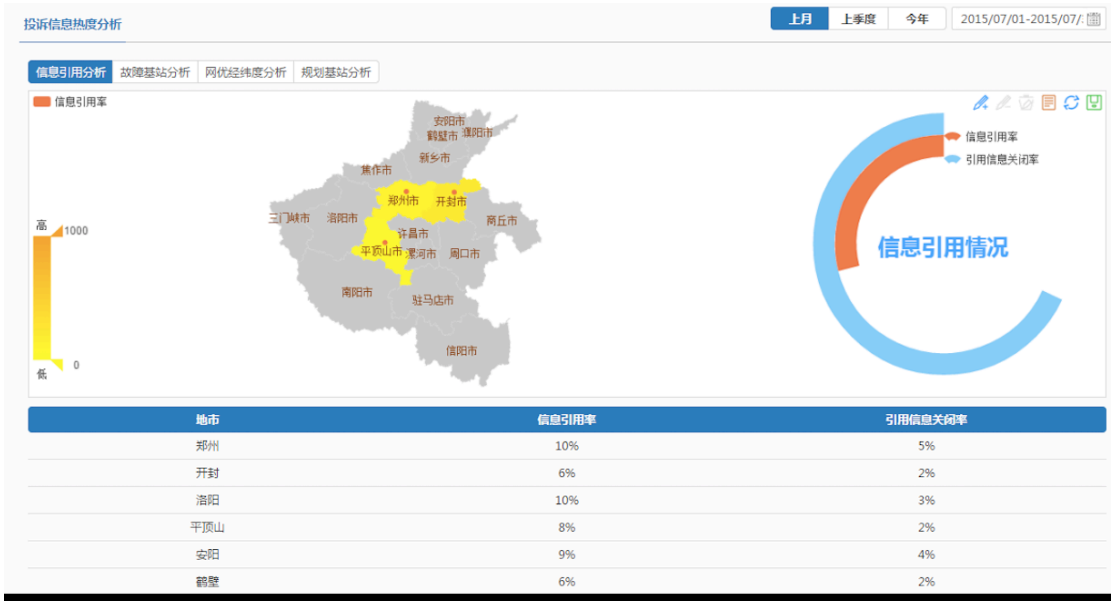



## 3.4.3. 投诉信息热度分析

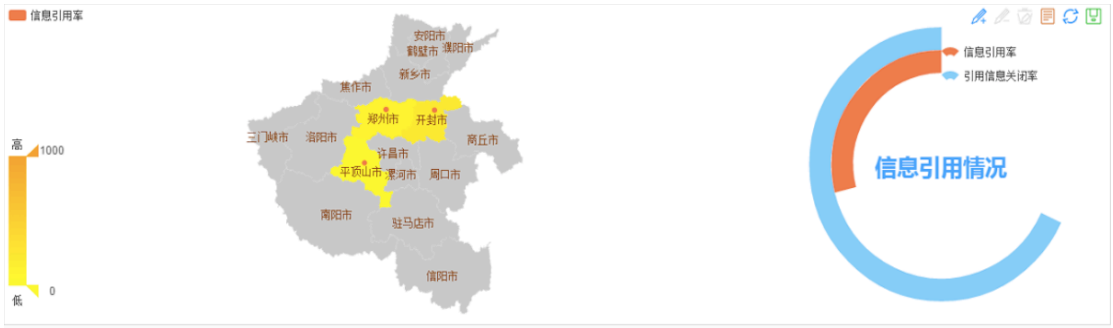
按投诉与公告的关联关系，将投诉工单与XX运营商发布的OA公告进行分析，指标如下：

- 各地市信息引用率、关闭率对比分析
- 故障基站排名统计

- 规则基站排名统计
- 网优经纬度排名统计




| 地市 | 信息引用率 | 引用信息关闭率 |
|---|---|---|
| 郑州 | 10% | 5% |
| 开封 | 6% | 2% |
| 洛阳 | 10% | 3% |
| 平顶山 | 8% | 2% |
| 安阳 | 9% | 4% |
| 鹤壁 | 6% | 2% |




| 序号 | 基站名称 | 地市 | 入线量 |
|---|---|---|---|
| 1 | 郑州火车北站W | 郑州 | 236 |
| 2 | 郑州经贸职业学院13至16号宿舍楼MW | 郑州 | 123 |
| 3 | 郑州大学老校区宿舍楼6MD | 郑州 | 76 |

## 3.5. 性能测试

Aleiye 实时大数据分析引擎测试，在每天（24 小时）产生约 1 亿条事件，共 100G 数据量的情况下，的处理性能结果如下：

| 指标 | 一级需求 | 时长 |
|---|---|---|
| 数据采集 | 10W 事件/秒 | |
| 数据处理 | 1W 事件/秒 | |
| 数据关联 | 5000 事件/秒 | |
| 数据搜索 | 15 亿事件/秒 | |

| | | |
|---|---|---|
| 数据图表展现 | 1 亿事件/秒 | |

### 3.6. 测试总结

Aleiye 实时大数据分析引擎已完成了对投诉统计分析、投诉积分统计、地理化数据展现、投诉信息热度分析等功能，基本满足了网优部门对网优投诉业务的数据分析工作。在性能方面，也基本满足现有数据量的支撑工作。

# 4. 后续实施方案

目前已通过静态数据的方式完成网络客服支撑系统的数据采集及分析。下阶段目标：通过地理化与无线侧基站关联，逐步完成多系统的数据汇聚和分析，形成多维度网络质量展现与深度挖掘的网络分析平台。

ALEIYE

让 | 大 | 数 | 据 | 更 | 简 | 单

公司地址：北京市西城区新街口外大街28号普天德胜大厦A座405

邮　　编：100088

联系电话：010-82053991

电子邮箱：service@ALEIYE.cn