采 购 文 件

项目编号：CA25B2E4A0008

项目名称： 长安汽车客户之声VOC聆听系统建设项目

采 购 人：重庆长安汽车股份有限公司

代理机构：东风国际招标有限公司

2025年5月

**编写要求**

1、本模板适用于系统实施类项目的技术要求编写。

2、要求逐项认真填写，文字表达明确严谨、用词精练、术语规范。

3、红色字体部分由项目组自行编写，黑色字体建议采用（可适当调整）。

4、\*号条款由项目组自行确定，但不宜过多，建议3-5条为宜。

5、字体一律用宋体，正文字号为小四号（封面标题为一号加粗，其他为三号加粗）。

目录

# 项目背景

# 项目目标

## 总体目标

## 分项目标

# 项目范围

# 项目假设

# 业务需求分析

## 总体流程分析与流程规划方案

## 业务流程及主要功能分析与设计方案

## 系统报表功能分析与设计方案

## 业务数据交互分析与设计方案

## 业务用户量规模与承载方案

## 业务数据量规模与承载方案

## 业务灾难恢复需求分析与方案

# 系统平台建设方案

## 总体要求

## 软件架构

### 本地化部署方案

### 独立训练框架方案

### 冗余架构设计方案

## 硬件架构

### 硬件架构部署方案

### 硬件资源容量估算清单

### 硬件资源切换方案

# 项目实施方案

## 时间/计划

## 质量保障方案

确保数据供给的可用性、稳定性、时效性和合规性是构建VOC体系的前提。

可用性

* 数据源广

数据源广：国内社媒数据全覆盖

|  |  |
| --- | --- |
| **采集站点类型** | **具体站点举例** |
| 问答 | 问答贴吧主流站点覆盖：知乎/百度贴吧......**知乎近6千万主帖，3.3亿+回答** |
| 知乎、百度贴吧、百度知道、天涯问答、悟空问答、宝宝树问答、马蜂窝问答、微头条、新浪门户\_新浪爱问、天涯问答 |
| 微博 | **实时获取，全面集成商业API，最高权限商业合作** |
| 微信 | **千万**账号，**日更，**支持**微信视频号**账号监测 |
| 小红书 | **亿级**账号，**小时级**监测， |
| 短视频 | 抖音**XXXXXX**、快手**亿级**账号**小时级**监测、火山、西瓜 |
| 视频 | **监测各大主流视频网站20+，B站上亿账号，小时级监测** |
| 优酷、bilibili、爱奇艺、腾讯视频、AcFun、乐视、芒果TV、搜狐视频、凤凰视频 |

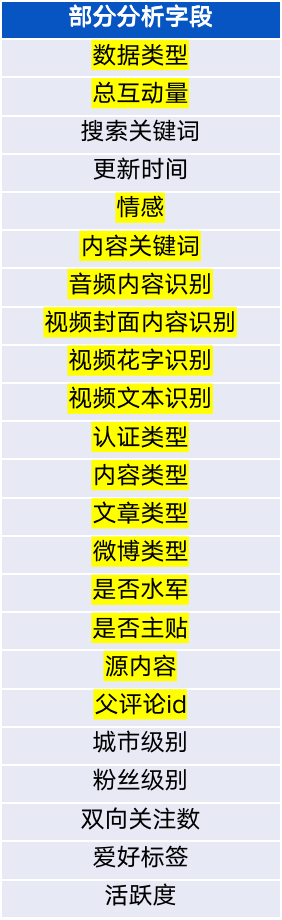
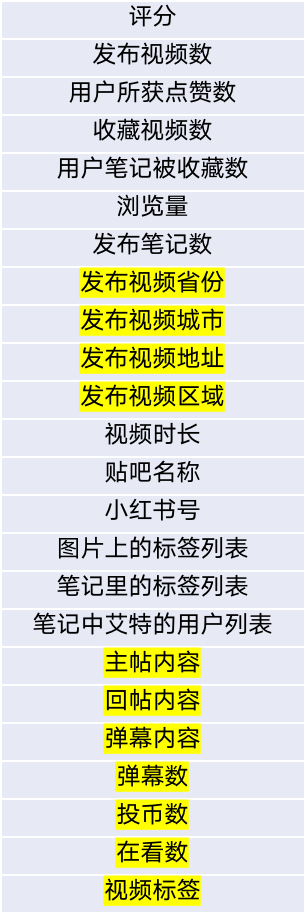
|  |  |  |
| --- | --- | --- |
| **采集站点类型** | **具体站点举例** | |
| 新闻/论坛 | **10万+站点，30万+板块，千万级自媒体账号，分钟级监测** | |
| 主流门户网站 | 今日头条、搜狐门户、腾讯门户、新浪、腾讯、凤凰网、百度天涯论坛、虎扑网、人民网、新华网、中国新闻网、环球网、中国经济网、糗事百科、MSN、光明网、TOM、中国网、中华网、中新网、中青在线、中广网、央视网、中青网、国际在线、千龙网、中国经济网、环球网、中国日报网等 |
| 科技类 | 36氪、虎嗅、果壳、极果、硅谷网、IT168、中国科技网、极客迷、中关村在线\_论坛、电子发烧友论坛、搜狐门户\_IT社区等 |
| 家电类 | 家电网、环球家电网、中国家电网、万维家电网等、各品牌官网 |
| 综合类 | 百度新闻网、网易网、ZAEKE、百度贴吧、豆瓣、什么值得买等 |
| 投诉类 | 黑猫投诉、消费保等 |
| 数码类 | IT时代周刊、机锋网、物联网在线、华为论坛、数字尾巴、华为云\_社区、安卓论坛\_bbs、PP助手论坛、智能电视网、西部数码论坛、中国电子DIY之家、三星手机论坛、电脑报官方论坛、开源电子网、电子电路网论坛、最数码等 |
| 汽车类 | 汽车之家、爱卡汽车网、懂车帝、车质网、太平洋汽车网、易车网、车主之家、中国客车网、极客汽车等 |
| 食品类 | 食品论坛网、食为天论坛、食品饮料招商网、国家食药监、每日食品、中国海洋食品网等 |
| 亲子母婴类 | 父母网、沪江育儿、99育儿、妈妈育儿网、家有宝宝、母婴公社、妈妈网、太平洋亲子网、妈妈帮论坛等 |
| 其他类 | 房产类、影音娱乐类、体育类、游戏类、军事类、商业财经类、旅游类 |

数据源广：灵活性高、支持新增站点开发和自定义标签。已经为多位头部客户提供指定站点的数据采集服务，包含各种投诉网站、协会官方站点、政府机关网站、垂直类数据站点等。

|  |  |  |
| --- | --- | --- |
| **部分新增数据源类型** | **重点信源示例** | **重点模块** |
| **新闻站点及数字报刊** | 网信办媒体白名单全部媒体 | 全部子站、栏目及文章 |
| **政府机关网站** | 政府机关、监管机构，如网信办、工信部、银保监会、药品监管局、食品安全备案网等 |
| **行业协会** | 消协、全国12315平台、深圳手机行业协会等 |
| **垂直社区** | Coloros论坛、一加社区、realme社区、花粉俱乐部、小米社区、vivo官网社区、酷安开发者平台等 | 全部用户信息、主贴及跟帖 |
| **电商平台** | 垂直商城：欢太商城、vivo商城、华为商城、小米商城等  电商平台：丝芙兰、酒仙网、途虎养车等  二手电商：闲鱼、找靓机、转转、拍拍 | 全部用户信息、评论及问答 |
| **垂直站点** | 金融界、东方财富网、投资界、雪球网、虎嗅网等 | 全部文章、评论及用户信息 |
| **资讯类APP** | 股吧APP、雪球APP、财新APP等 |
| **投诉类网站** | 聚投诉、黑猫投诉 | 投诉信息 |
| **应用商店** | 三星、苹果、小米、googlePay、vivo、OPPO等应用商店 | 全部应用信息、主帖及跟帖 |
| **海外-主流的中文新闻媒体** | 联合早报、BBC中文网、星岛环球网、美国中文网、马来西亚环球日报、WSJ中文网等 | 全部子站、栏目及文章 |

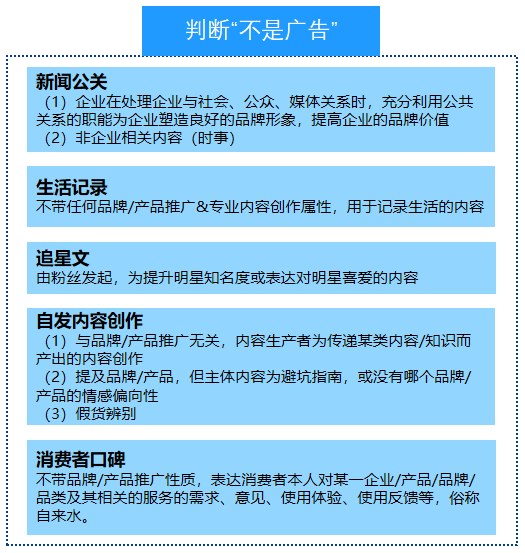
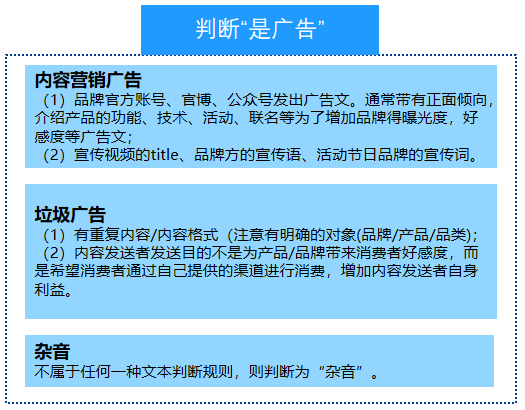
* 字段标签丰富

分析字段标签丰富：采集字段和分析字段构成丰富的字段列表

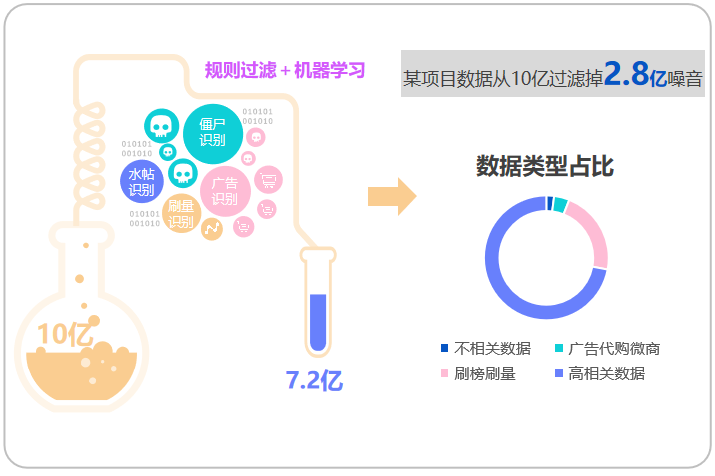


分析标签丰富：在现有字段基础上运用算法模型获得更多的分析标签

广告杂音算法（准确度85%+）



水军识别（准确度90%+）

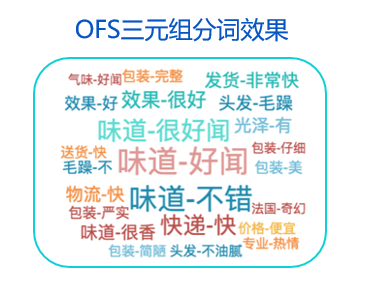


多层次情感判断模型（准确度75%～80%+）

通用情感算法

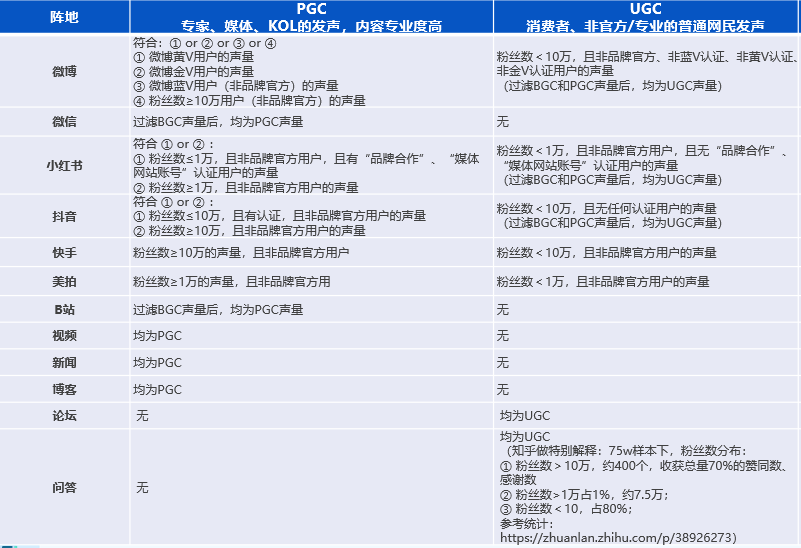


智能三元组。从任意非结构化自然语言中，提取句子最精要的观点，细颗粒度情感分析：文本内容进行特征和情感的提取判断，情感判断更具针对性。



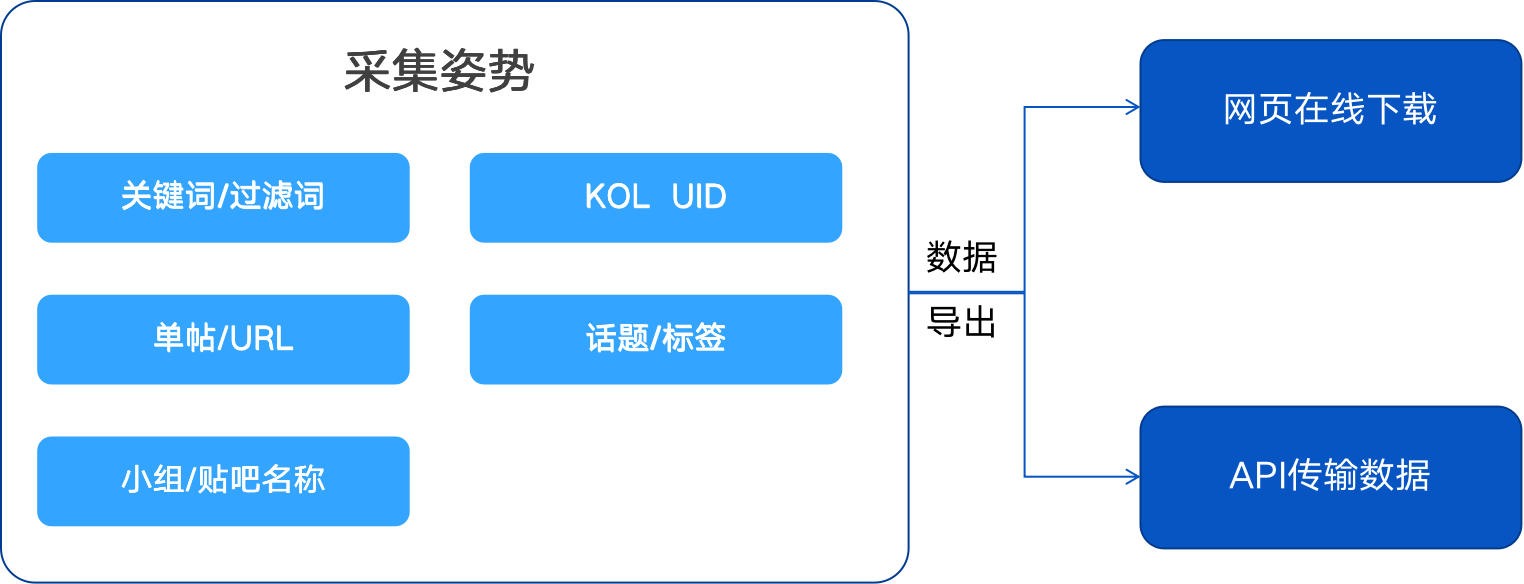
PGC/UGC判断

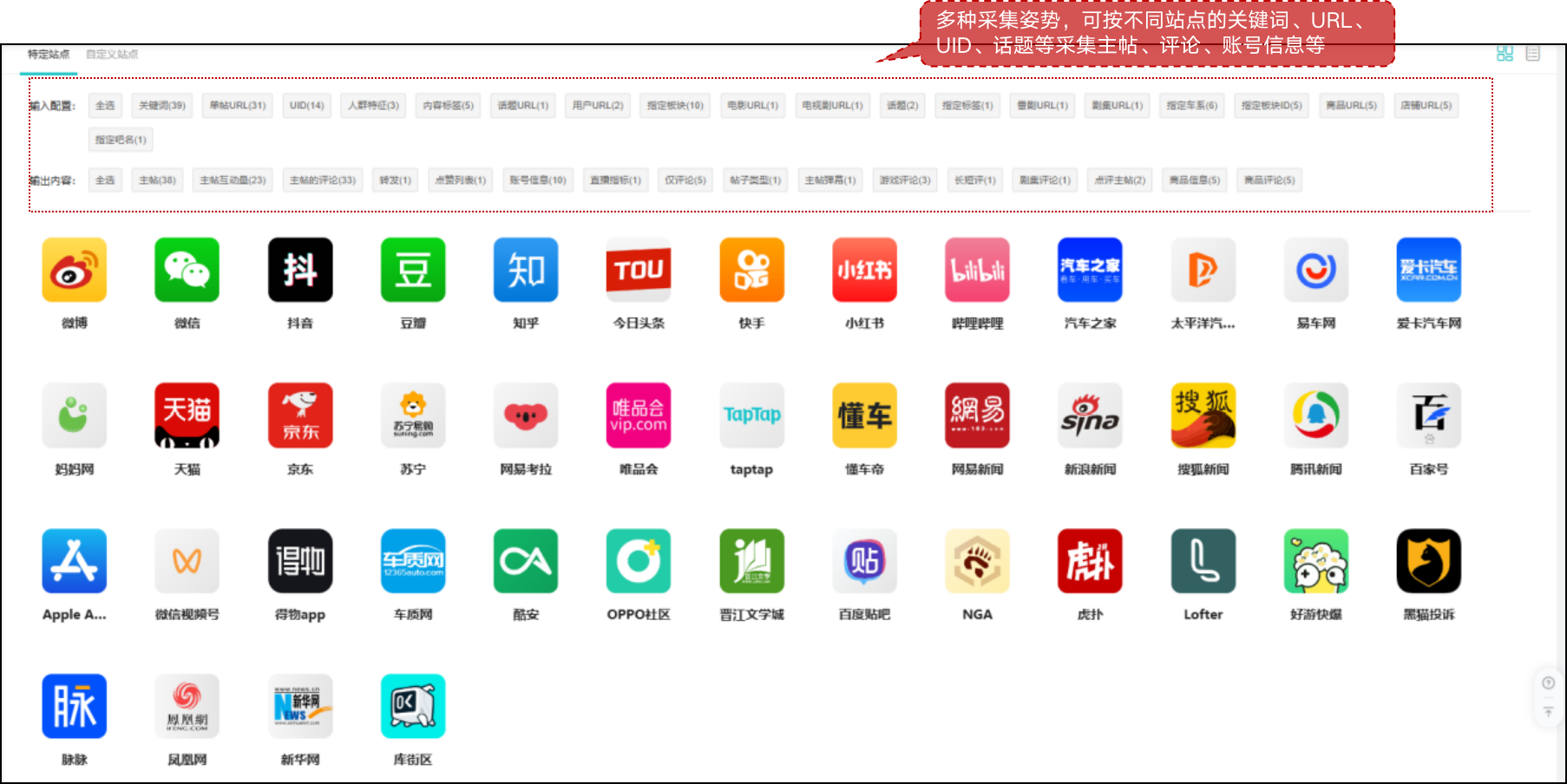
通过各个阵地对用户的官方认证等级、用户的粉丝量对声量类型划分为UGC、PGC。PGC：专家、媒体、KOL内容专业度更高（其中，微博、小红书、短视频、知乎、B站可获取认证类型）；UGC：消费者、普通网民发声



* 采集姿势多

按需提供多种采集姿势，适用不同的取数用途。





* 数据质量高

准确、及时、完整

**准确性指字段无错漏，若有任意字段错误则认为是无效数据。有效数据占比超过95%即为达标。**

监控&巡检手段：

（1）自动化

• 不能为空的字段：简单的正则校验；

• 可能为空的字段：设定一定比例，若超过比例则告警。

（2）人工巡查

• 巡查安排：

各站点对点负责到人，30条/每站点/每周，主要检查字段错漏问题；

• DPS巡查能力：

每人每天巡查约1200条，每周可安排2人用1天时间集中巡查重点站点，即每周巡查2400条左右。

**及时性-增量方案保证调用频次及任务时效性，新闻论坛20分钟入库，微博重点1-5分钟滚动更新。**

监控&巡检手段：

（1）自动化

• 由采集系统将相关资源信息输出到ELK上，到ELK上查看资源使用次数；例如抖音设备信息，统计所有设备的总调用次数、成功次数、失败次数，查看每天成功的调用次数是否达标。

（2）人工巡查

• 在做完整性巡查时，一起巡查及时性指标。

**完整性指采集回的数据量对比原站点数据量的完整度，主回贴数据量漏数率不超过5%，覆盖率在95%以上。**

监控&巡检手段：

（1）自动化

• 所有爬虫：直接看ELK对应爬虫的失败率，超出阈值时告警。

（2）人工巡查

• 增量：

a. 巡查安排：每周每个站点抽取15个板块，每个板块抽取3条数据到全量库查询是否有采集到，做覆盖率统计和及时性统计；

b. DPS巡查能力：每人每天巡查约1200条，每周可安排2人用1天时间集中巡查重点站点，即每周巡查2400条左右；

• 异步搜索巡查安排：

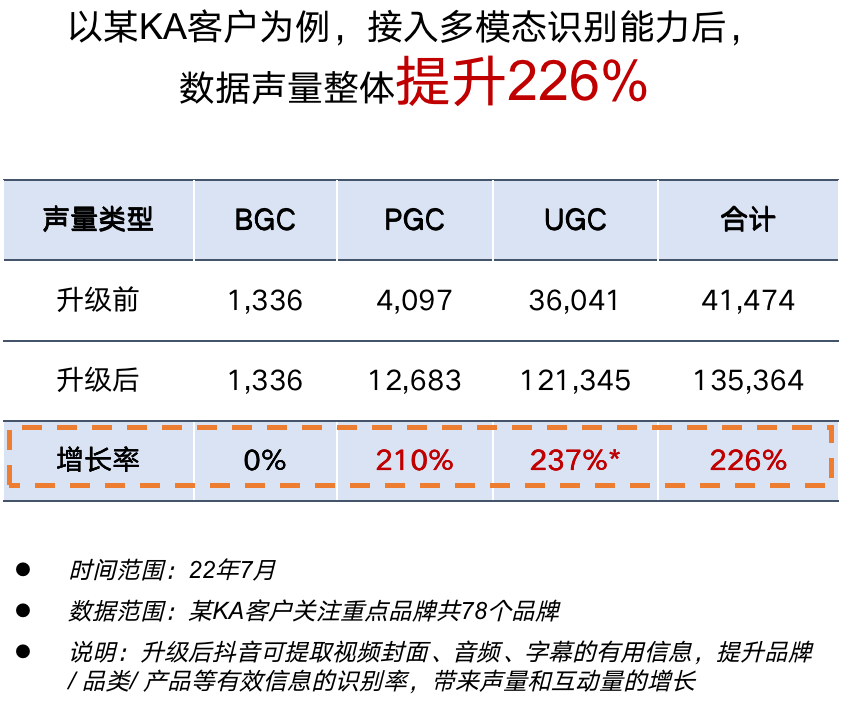
从入全量库数据的业务中，找到在使用的重点站点抽取部分关键词巡查覆盖率和及时性。

* 音视频解析

视频语音转文本。转化长/短视频的语音为可分析的文本数据，扩充长/短视频的数据库。

图片识别。通过解析电商图片，获取文本数据，有效进行相应的数据分析，增强数据匹配能力。





* 系统可用性强

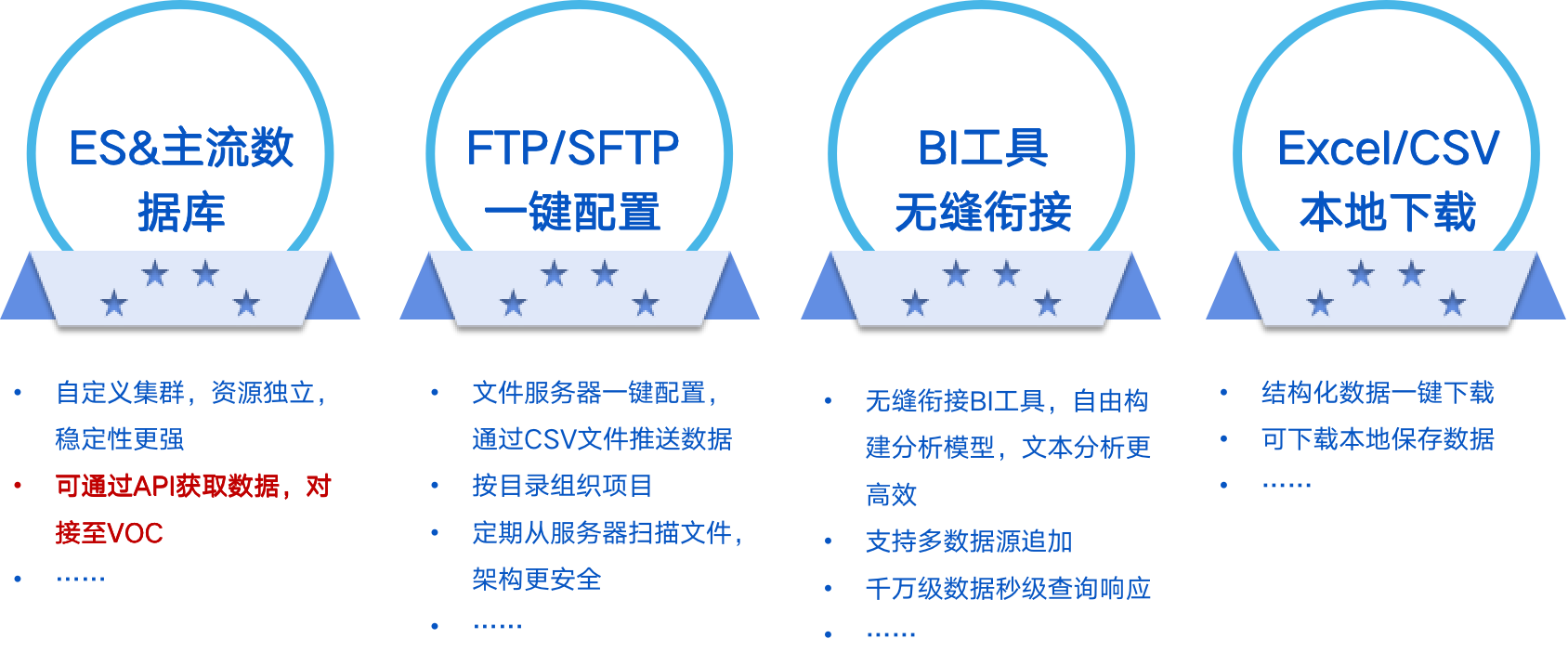
关键词设置/取数姿势灵活设置，对于广告、文本识别可能引起的杂音，可调整取数方式，保障数据全面性和有效性





* 数据提供格式多样

数据提供格式多样：满足不同的分析途径和存储格式，支持对接海尔HVOC系统。





稳定性

* 系统的稳定性

**从数据巡检到核心站点全面保障数据的稳定性**

核心站点「抖音/知乎/B站/天猫/京东/微博/微信/小红书」稳定性95%+

数据底层能力的稳定性

系统化的去重方案将kafka数量与入库量、入库和首次入库的差异性小于5%

推进聚合全部异步采集任务迁移至新调度框架，实现异步任务框架层面99.5%+稳定性

全量库ES集群完成由单集群向多集群拓展

全量库API的查询场景在1.5秒以内，全面提速取数场景

算法框架的优化提升算法离线平台小任务运行效率

全数据服务巡检制度

健全的数据服务巡检制度，定时巡检，统一汇报，及时处理

* 采集的稳定性

多种采集能力和反爬技术构成完善的采集稳定性闭环

全量采集

普通公司基于单次任务采集，受限于采集网站无法回溯，采集时间长，相关性差；

数说持续不间断无差别的采集所有新闻、论坛、微博、微信数据，可回溯，实时性强。

集群和代理IP资源

上百个节点组成的分布式采集集群， 配备千兆级的带宽；

自建多地域的VPS代理集群，每天能切换数十万的动态IP， 并覆盖了海外节点；

每日并发采集任务数上百万， 请求数近千万，日新增数据量几千万。

专业团队

专人开发： 高效采集规则配置，字段解析处理成结构化数据；

专人维护：持续监控网站的改版和变更， 专人及时响应网站改版和维护；

专业运维：管理集群机器，保证99.99%的可用率。

特色采集的能力

分布式的采集调度平台，管理上百台节点和任务分发；

智能提取算法： 新闻类、论坛类；

持续采集： 每日的海量微博、微信、新闻、论坛等全网的持续采集，日新增数据三千万；

图片、视频、音频解析下载。

反爬公关能力

图片的文本识别、验证码OCR识别；

海量代理IP防封锁；

自有账号资源池： 微信小号、微博小号、淘宝账号等，突破站点采集限制。

新型数据源定制

特定APP内数据、直播视频弹幕等数据采集。

* 存储的稳定性

数据存储备份策略

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **存储引擎** | **分类** | **高可用策略** | **备份策略** | **恢复策略** | **备注** |
| Elasticsearch | 分布式全文索引 | 集群支持高可用，内置Shard备份，某节点故障不影响使用 | 分布式存储架构，内置多shard备份（**热备**）  可选支持每天增量备份（**冷备**） | shard可自行恢复  增量的冷备数据可按需全量恢复 | 用于存放全量采集信息与标签信息 |
| Hadoop | 大数据 | 集群支持高可用 | 作为计算节点无需备份 | 作为计算节点无需备份 |  |
| MySQL | 关系数据库 | 主备虚拟IP漂移方式支持高可用 | 主备方式支持**数据热备**  支持定期的**自动备份**  支持**手动备份** | 主机故障可以自动切换到备机  可使用备份数据全量恢复 | 存放系统配置数据 |
| Redis | KeyValue数据库 | 多机集群支持高可用 | **持久化机制**，缓存一定时间内状态。 无需冷备机制 | 重启时自动从文件恢复 |  |
| CEPH/对象存储 | 分布式对象存储系统 | 集群支持高可用 | 批量**数据复制**  跨地域复制的**容灾备份**方案 | 通过备份数据进行恢复 |  |

* API 的稳定性

完善、标准化的API接口



* 基础配套运维服务的稳定性

采集和平台SLA是运维服务的保障

采集SLA

|  |  |  |  |
| --- | --- | --- | --- |
| **指标** | **保障对象** | **保障形式** | |
| **自动化** | **人工巡查** |
| 准确性 | 主帖 | 不能为空的字段：简单的正则校验，至少可以用于判空，数值之类可以用\d+方式判错  可能为空的字段：设定一定比例，若超过比例则告警 | 每周花2人天的时间去巡查对应站点的数据，确认是否有错漏情况（2人天巡查的数据量约2400条） |
| 回帖 |
| 及时性 | 接口调用形式 | ELK统计各资源的调用次数、成功率、失败率，当失败率超过阈值时告警 | 在巡查完整性的同时巡查及时性 |
| KOL/板块监控形式 | 调度频次保证：调度器输出调度每个种子的日志到ELK上  任务时效性保证：将执行时间超过调度间隔的任务输出到ELK上 |
| 完整性 | 主帖 | ELK统计爬虫失败率  使用模拟脚本和人工巡查做相互验证 | 每周每个站点抽取15个板块，抽查靠前的数据是否有采集到，做覆盖率统计 |
| 回帖 |

平台SLA

|  |  |  |  |
| --- | --- | --- | --- |
| **等级** | **程度** | **描述** | **处理及时性要求** |
| P4 | 轻微 | 未正式对外发布功能、暂无业务方、客户使用的功能 | 问题上报后7~14天内(一个产品版本) 完成问题修复 |
| P3 | 一般 | 终端客户可能感知，功能存在偶发性问题或体验问题，未影响系统正常运行，但可同客户沟通协商处理解决 | 问题上报后7天内（一个产品版本）修复解决问题 |
| P2 | 中等 | 影响终端客户正常使用、系统功能存在问题，影响正常业务流程等情况 | 问题上报后24小时内恢复环境正常使用  问题出现3天内发布紧急hotfix彻底修复问题 |
| P1 | 严重 | 系统不可用（系统崩溃或完全无法提供服务），或直接引起终端客户投诉、等问题 | 问题上报后立即响应，以恢复生产环境为优先目标，同项目组组建联合的应急响应团队原则上2小时内恢复生产环境状态，若超过2小时未恢复， 每半小时通报一次恢复进展。  24小时内给出彻底修复方案，在保障生产环境不再次出现故障的前提下，原则上7天内按修复方案发版修复。 |

时效性

* 数据更新的时效性

主流阵地最高分钟级监测



* API 接口的时效性

标准接口快速取数，传输数据无延迟。



* 运营服务的时效性

及时响应项目故障/问题管理流程

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **优先级** | **描述** | **服务需求特点** | **响应时间** | **应答时间** | **替代方案或永久方案**  **预计解决时间** |
| 1 | 关键的 | 关键系统处理停止了，而且用户不能进行操作。没有规避、绕开和其他的方法可用。  主要的经济影响条件存在。(关键系统被定义为主要系统运转停止并对服务操作有关键影响)。 | **工作时间20min内响应** | **2小时内** | 立即着手解决问题，80%的问题均可在8小时内得到解决 |
| 2 | 重要的 | 一个主要的部件、应用、关键系统停止了。处理受到严重影响。  次要的经济影响条件存在。潜在的对服务操作有关键影响。 | **4小时内** | 80%的问题在1天内得到解决 |
| 3 | 次要的 | 一个部件、次要应用或过程停止了或难以使用。有些操作受影响，但不是立即影响到服务操作。一个可接受的规避、绕开和其他的方法可用。  那些严重等级1或2的有规避、绕开和其他的方法可用的问题将被认为是严重等级3。 | **1个工作日内** | 80%的问题在3个工作日内得到解决 |
| 4 | 普通的 | 一个部件、过程或个人应用（对客户来说是非关键的）不可用了。 对业务没影响，单一的出错事件，而且一个可用的规避、绕开和其他的方法存在。推迟的维护可被接受。 | **2个工作日内** | 根据双方协商达成的共识 |

## 信息安全方案

法律上的合规性：现行法规下合法采集数据

遵守法律法规

我们遵守《网络安全法》、《数据安全法》、《著作权法》、《反不正当竞争法》等国内外法律法规，收集的全部数据及信息，均来自互联网公开可获取的内容，信息收集、处理等技术合法合规，不损害任何第三方的合法权益

隐私信息保护

我们遵守《网络安全法》、《个人信息保护法》关于网络用户个人信息保护的相关规定，不收集、使用或传输任何个人隐私信息

官方平台合作

我们通过互联网平台官方接口或其他平台许可的方式，取得平台关于其数据处理的权限许可，在授权范围内开展数据处理活动，拥有微博等平台的数据接口API最高权限

规范处理数据

我们作为中立的技术服务提供者，在法律法规规定的范围内收集数据，并对数据内容进行审查及清洗，如发现所收集的内容中涉及法律法规规定的违法数据类别，会及时停止收集并删除相关信息，不用于二次传播，所获取的数据不会用做法律法规所不允许的非法目的和用途

行业内的合规性：遵循行业规则采集公开网络数据

遵守技术规则

1) 我们所采用的爬虫技术属于合法获取互联网信息的方式，信息收集、处理等方式均合法合规，与用户浏览行为本质相同，不损害任何第三方的合法权益

2) 我们遵守爬虫技术通用的技术规则，规范爬虫的访问频率，每日收集的数据量不超过源网站日均流量的三分之一，原则上遵循平台规则、Robots协议，不会对源网站造成不当干扰、破坏，不影响源网站的正常运行

多项权威认证

作为国内领先的一站式大数据及AI智能应用提供商，我们致力于为企业提供安全可靠的大数据服务，持续构建个人信息保护和数据安全保障管理体系，目前已完成多项权威认证：

1) 质量管理体系认证 ISO 9001

2) 信息技术服务管理体系认证 ISO 20000

3) 信息安全管理体系认证 ISO 27001

4) 信息系统安全等级保护备案证明（二级）

## 培训方案

## 测试测试

## 项目交付物清单

## 项目组织架构

## 项目实施方法

## 项目变更方案

# 系统服务方案

## 迁移方案

## 运行维护服务方案

## 3年技术支持方案

## 许可证方案

# 验收方案

## 验收条件

### 交付物清单

### 稳定运行保障方案

### 运行错误处理方案

## 验收标准

# 安全保密方案

# 知识产权转移方案