2026年朝阳区人口大数据监测（移动手机信令数据）合同

本合同由以下双方根据平等自愿原则，协商一致在北京市朝阳区签订：

甲方：北京市朝阳区数据局

地址：北京市朝阳区日坛北街33号

授权代表人：冯峻

邮编：100020

联系电话：65099913

传真：65094287

乙方：中国移动通信集团北京有限公司

营业执照注册号：91110000722611700L

地址：北京市东城区东直门南大街 7 号

法定代表人：李强

联系电话：13810588472

账户名称：中国移动通信集团北京有限公司

开户银行：招商银行股份有限公司北京分行营业部

公司账号：8888015000006138

本《2026年朝阳区人口大数据监测（移动手机信令数据）合同》（以下简称**本合同**）经甲、乙双方友好协商签订，双方达成以下合同内容。

第一条　　服务内容与服务标准

本合同所涵盖的各项服务内容与服务标准详细内容见附件一《服务内容与服务标准》。

第二条　　服务期限

自【 2026 年 9 月 24 日至 2027 年 9 月 23 日】止。本项目成交结果可作为两次合同续签的依据，经采购人同意，且年度资金保障、乙方履约考核合格、服务内容与价格维持原成交标准不变的前提下，甲方与乙方可就新的自然年度续签合同（每次续签合同服务周期不超过 12 个月，累计续签上限为 2 次）。

第三条　　工作检查

甲方有权按本合同第一条约定的服务内容和标准对乙方工作进行工作检查，经甲方工作检查不合格的，甲方有权要求乙方在指定期限内采取弥补措施，乙方采取弥补措施之后再次提交甲方进行检查。经乙方【三】次弥补仍未能经甲方检查合格的，甲方有权单方解除合同，不予支付剩余合同价款。

第四条　　合同总价款及结算

1、本合同总价款为人民币【1476000】元（大写：【人民币壹佰肆拾柒万陆仟元整】），该费用已包含本合同项下全部费用，包括但不限于税费、人工费、交通费、服务费、保修费等。除此以外，乙方无权要求甲方因本合同支付其他任何费用。

本合同生效后【20】个工作日内甲方支付合同首款【885600】元(大写：【人民币捌拾捌万伍仟陆佰元整】)，支付中期款，即人民币

【295800】元（大写：【贰拾玖万伍仟捌佰元整】），剩余合同价款，即人民币【294600】元（大写：【人民币贰拾玖万肆仟陆佰元整】），经甲方验收合格后支付。甲方付款前，乙方应向甲方提供合法、有效的税务发票。

2、本项目如根据甲方要求提前终止部分服务内容的，应由甲乙双方协商一致后签署书面补充协议。甲方剩余价款支付金额以乙方提供的实际服务内容和服务时间综合核算结果为准。

第五条　乙方义务

1、乙方需安排专人负责与甲方对接本合同约定的相关工作，确保乙方按时保质完成工作，乙方安排人员未经甲方书面同意，不得随意更换。

2、乙方保证所有项目符合国家标准以及行业标准。

3、乙方严格遵守中华人民共和国法律、法规及合同对有关技术资料及技术的要求。

4、乙方确保其提供的本合同项下的所有服务所涉及的产品、技术、资料不会侵犯第三方的知识产权或所有权，否则由此造成的损失，由乙方承担全部责任。乙方安排人员在甲方服务期间，因故意或过失给甲方造成相应损失的，乙方应承担相应的赔偿责任。

第六条　知识产权

1、乙方保证乙方提供的服务不存在任何侵犯第三方知识产权的情形。如果第三方声称乙方向甲方提供的服务侵犯其知识产权，并已就此对甲方或乙方提起（包括威胁提起或很可能提起）法律诉讼程

序或知识产权行政执法程序（简称侵权诉讼），则知悉上述事项的一方应立即通知合同对方，甲方有权：（1）暂停履行对侵权诉讼所涉服务的采购或支付义务直至侵权诉讼完全解决，并要求乙方自担费用向甲方提供与该第三方协商、诉讼、和解所需的一切协助（包括但不限于向甲方提供证明侵权不存在的各类证据、安排人员参加协商、诉讼或会谈等）；（2）甲方有权选择与该第三方达成和解，并由乙方支付和解协议所约定的全部费用以及甲方因侵权诉讼而遭受的全部损失或费用（包括但不限于诉讼/仲裁费、律师费、交通费、通讯费、差旅费、对第三方的损害赔偿金、行政处罚罚款、获取该服务相应使用许可的费用、因停止使用或修改、替换侵权威胁所涉及的服务而遭受的损失等）。如果甲方选择继续参加侵权诉讼法律程序，乙方应当赔偿甲方因侵权诉讼及履行生效法律裁判而需支付的费用或遭受的损失，但生效法律裁判认定乙方服务不存在侵犯第三方知识产权情形的除外。

2、甲方在乙方提供的服务基础上开发、研制形成的技术成果（包括但不限于程序、文件、资料等）的知识产权归甲方所有。

3、不论本合同是否解除或终止，本合同项下第六条独立适用。

第七条　保密义务

1、乙方应对其知晓的甲方的商业、技术、市场、管理、人事、财务等任何方面的信息和资料予以保密，未经甲方事先书面同意，乙方不得披露、使用或以任何方式处置上述信息、资料，并应促使其员工、关联方承担相同保密义务，如果乙方员工、关联方违反上述保

密义务，视为乙方违反保密义务并适用本条第 2 点的约定。

2、乙方违反保密义务的，应当对甲方因此所遭受的损失承担赔偿责任。如果乙方在本合同有效期内违反保密义务，甲方有权单方提前终止本合同。

3、不论本合同是否解除或终止，本合同项下第七条独立适用。

第八条　不可抗力

1、由于发生不可抗力事件（如战争、暴动、严重火灾、水灾、台风、地震、政府行为和禁令等事件），致使合同任一方不能履行合同义务时，遭受不可抗力事件影响的一方负有在不可抗力事件发生之日起 15 日内尽快通知合同对方以减轻可能给对方造成的损失，并应当在合理期限内提供证明。

2、遭受不可抗力事件影响的一方在履行前述义务后免除违约责任，但其合同义务不因此免除。经合同双方协商同意，合同履行时间可合理延长，延长时间相当于因事件发生受到影响的时间。

第九条　违约责任

1、乙方违反本合同约定义务的，甲方有权要求乙方在指定期限内采取弥补措施，乙方未能弥补的，视为乙方违约，乙方应向甲方支付相当于合同总金额【百分之三】的违约金。

2、乙方延期履行的，每延期一日，按合同总金额【千分之三】向甲方支付延期履行违约金。

3、因乙方过错造成甲方或第三方损失的，乙方应赔偿甲方或第三方全部损失。

4、甲方有权直接在未支付的合同总金额中扣除本条约定的违约金及赔偿金，不足部分由乙方补齐。

第十条　争议管辖

本合同项下发生的争议，由双方当事人协商解决；协商不成的，任何一方均有权向甲方住所地人民法院提起诉讼。

第十一条　本合同自双方法定代表人或授权代表人签字并加盖公章之日起生效。本合同壹式肆份，双方各执贰份，具有同等法律效力。本合同的任何变更、补充或修改，应由双方协商一致并签署书面协议。

第十二条　通知与送达

任何一方根据本合同规定向另一方发出的通知应以书面形式作出，并以邮寄/快递、传真、专人送达方式发送。如以邮寄/快递方式发送，以邮寄/快递回执上注明的收件日期为送达日期。如以传真方式发送，收到传真机发出的确认信息后，视为送达。如专人送达，被送达人签署后，视为送达。各方联系信息以本合同首部所列为准，一方联系信息变化后，该方应在联系信息变化之前将变化情况书面通知其他方，否则该方应自行承担相应的风险、责任和后果。

第十三条　本合同中出现的“日”，除非明确写明为工作日，否则为自然日。

第十四条　本合同附件作为本合同内容的一部分，与本合同具有同等法律效力。

第十五条　如甲方因为机构调整变更主体，由变更后的主体继续履行本合同项下的内容。

附件一：《投标文件技术应答部分》

附件二：《验收标准》

附件三：《合作伙伴廉政协议》

（本页为签署页）

甲方：北京市朝阳区数据局

（公章）

授权代表人（签字）：

【2026】年【 9 】月【 11 】日

乙方：中国移动通信集团北京有限公司

（公章）

法定代表人或者授权代表人（签字）：

【2026】年【 9 】月【 11 】日

附件一：《服务内容与服务标准》

1.1. 项目背景

2017 年 9 月中央、国务院批复的《北京城市总体规划（2016 年 —2035 年）》相关要求，北京市以资源环境承载力为硬约束，严格实施人口规模减量管控，明确全市常住人口 2020 年控制在 2300 万人以内、2035 年长期稳定在 2300 万人左右的发展目标，同时建立“一年一体检、五年一评估”的规划实施监测机制，推行分区域差异化人口调控模式，重点推动中心城区人口减量、优化人口空间布局，严控人口无序集聚，引导人口随产业与城市功能有序疏解转移。

作为北京城六区核心区域和人口大区，朝阳区承接全市减量发展核心要求，在2019年12月编制完成的《朝阳分区规划（国土空间规划）(2017 年 —2035 年)》中进一步细化人口管控目标，明确 2035 年常住人口需控制在 333.4 万人以内的约束性指标，要求建立常态化人口动态跟踪机制，全方位监测全域及各街乡人口增减、流动变化情况，为人口疏解成效评估、分区规划落地考核提供量化支撑。

2026 年 1 月 16 日《北京市朝阳区国民经济和社会发展第十五个五年规划纲要》经北京市朝阳区第十七届人民代表大会第七次会议审议批准，是引领朝阳区 2026 至 2030 年经济社会发展的纲领性文件。纲要立足朝阳区超大城市中心城区功能定位，明确持续深化人口精细化治理，严格落实全市人口调控总体部署，压实人口规模管控任务，健全人口常态化动态监测、趋势分析和研判预警工作机制，精准掌握区域人口总量、结构分布、流动迁徙等动态变化特征，为人口调控施策、公共服务供给、空间资源优化配置提供科学依据。同时，纲要明确推进数

字赋能城市治理，加快智慧城市建设，完善城市运行数字感知体系，统筹各类数据资源汇聚共享与创新应用，运用大数据、数字化监测手段补齐城市治理动态感知短板，提升城市治理科学化、精准化、智能化水平。

当前朝阳区人口调控、疏解及精细化治理工作任务繁重，量化明确的人口管控任务对精准、高效、连续的人口数据支撑提出了极高要求，但现阶段区域人口监测工作仍存在明显短板，一方面传统统计人员入户抽样调查的数据获取方式存在时效性弱、精准度不足、覆盖范围有限等问题，无法适配新时代人口动态治理的工作需求；另一方面区域现有信息化系统主要包含“一网通办”区级政务服务平台、政务服务内部管理系统及大数据可视化展示系统等，整体功能较为基础，大数据深度分析能力薄弱，仅可实现基础数据汇总，缺乏实时分析、预测建模等核心功能，尚未充分利用移动信令数据等动态数据资源，且现有可视化平台仅支持静态图表展示，无法实现人口流动、人口密度变化等场景的动态监控和交互式分析，难以支撑复杂人口治理场景的决策、应急响应与资源调配工作。

在此背景下，本项目依托成熟的大数据技术，整合移动运营商手机信令数据开展用户位置数据连续大规模采样分析，可精准掌握朝阳区及北京市其他十五区人口总量变化趋势，细化监测朝阳区各街乡常住人口、工作人口动态变化情况，持续跟踪特定区域人口疏解调控成效，深度分析区域人口时空流向与变化规律，能够有效弥补传统统计方式与现有信息化系统的短板，精准对标朝阳区“十五五”时期人口常态化监测、数字赋能精细化治理的核心工作要求，同时项目在前期工作基础上持续开展常态化数据采样分析与运维服务，可精准把控朝阳区人口动态变化规律，全面提升区域人口现代化、智能化、精细化治理水平，为朝阳区人口

相关政策的科学制定、落地实施、成效评估及优化调整提供全方位、系统化、精准化的数据支撑和决策依据。

1.2. 需求理解

1.2.1. 服务内容需求

本项目是通过对移动运营商手机用户数据在朝阳区域内进行连续、大规模抽样采集并统计，通过对一段时间内的手机用户在不同时间维度的数据量变化、占比变化、各街乡、重点区域手机用户的流动情况等，从而掌握本区人口变化的主要趋势和特点，以此分析朝阳区人口流动、人口结构、职住通勤等特征，通过持续开展用户位置数据连续大规模采样分析，提供数据运维服务，把握朝阳区人口总量变化，提升人口治理能力，为政府人口相关政策提供更完善、更全面的数据依据。

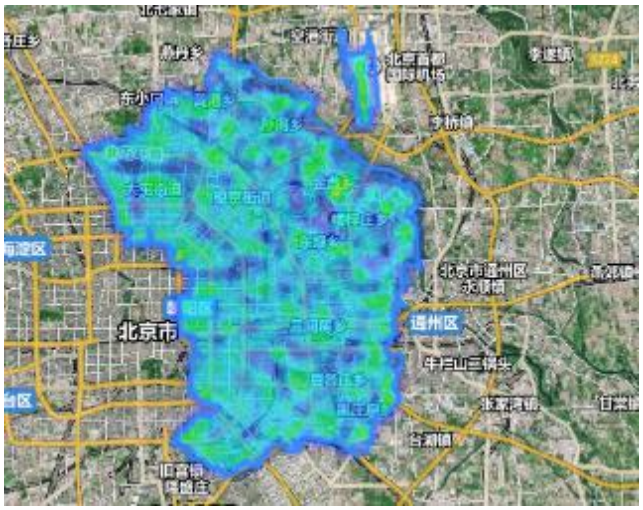
需求内容主要有：

1.提供涵盖朝阳区全境、43 个行政街乡、81 个重点区域、50 个重点村、12 个商圈和 24 个旅游景区及临时新增区域的 4G、5G 用户的网络信令数据结果，为满足人口监测数据同比、环比对比分析需要，数据可追溯周期至少覆盖服务期起始前一年，优先回溯至前两年。同时按照采购人规范要求，完成数据脱敏、清洗、加工、监控等全流程处理，全面保障数据使用安全、合规、有效。

2.上述采集到的人口数据标准化处理以后汇总到区指定平台，供相关单位分析，形成多维度的分析结果，可用于模型统计、可视化呈现和决策参考，以服务于区域人口调控中心工作。

3.采集到的数据对区指定平台热力图提供支撑，并实时提供数据展示。

参考图例如下：



1.按照三网融合平台建设项目的数据需求及技术规范，配合输出提供所需的统计级标准数据，对采集到的人口数据通过模型算法和结果汇总，形成多维度的分析结果。

2.针对客户临时需求，提供个性化业务分析服务，满足突发性的人口监测业务需要，例如重大节假日等特殊时期。

3.持续提升用户定位精准度并进行定位验证修正，同时根据项目服务需求构建用户标签体系。

4.根据客户需求提供数据质量、数据安全等服务保障工作，保障数据的准确性和安全性。

1.2.2. 服务性能需求

为满足项目对数据实时性和准确性的要求，对数据服务的性能需求如下：

1.实时数据接口传送频率不低于计划内时间，如每 30 分钟一次的接口数据，在本次 30 分钟内完成传送。

2. 月度人口统计数据在次月 10 日前完成统计。

3. 月度分析报告在次月 15 日内完成分析。

4. 统计数据交付格式为 csv，分析报告交付格式为 word。

1.2.3. 服务安全需求

为满足物理安全、网络安全、主机安全、应用安全、数据安全五个方面基本技术要求，要求系统平台在专网进行网络传输，只能传输统计级数据，并在数据传输过程中全程进行数据加密，以保证关键数据在传输过程中不被监听，避免敏感数据外泄，避免被未授权者操控。对数据服务的系统安全性能需求如下：

应用安全：应用系统应具备完善的用户身份认证、权限校验能力，建立规范的用户管理、权限管理体系，充分依托操作系统及数据库安全机制筑牢应用安全防线，软件运行全程留存完整操作日志，所有用户密码、系统密钥均采用加密存储方式，禁止明文存储。

数据安全：严格落实数据访问权限管控，数据库、业务文件仅对授权用户、专用终端开放访问权限，防范数据在本地存储、网络传输过程中被非法篡改、删除、窃取、破坏的风险，保障数据全生命周期安全。

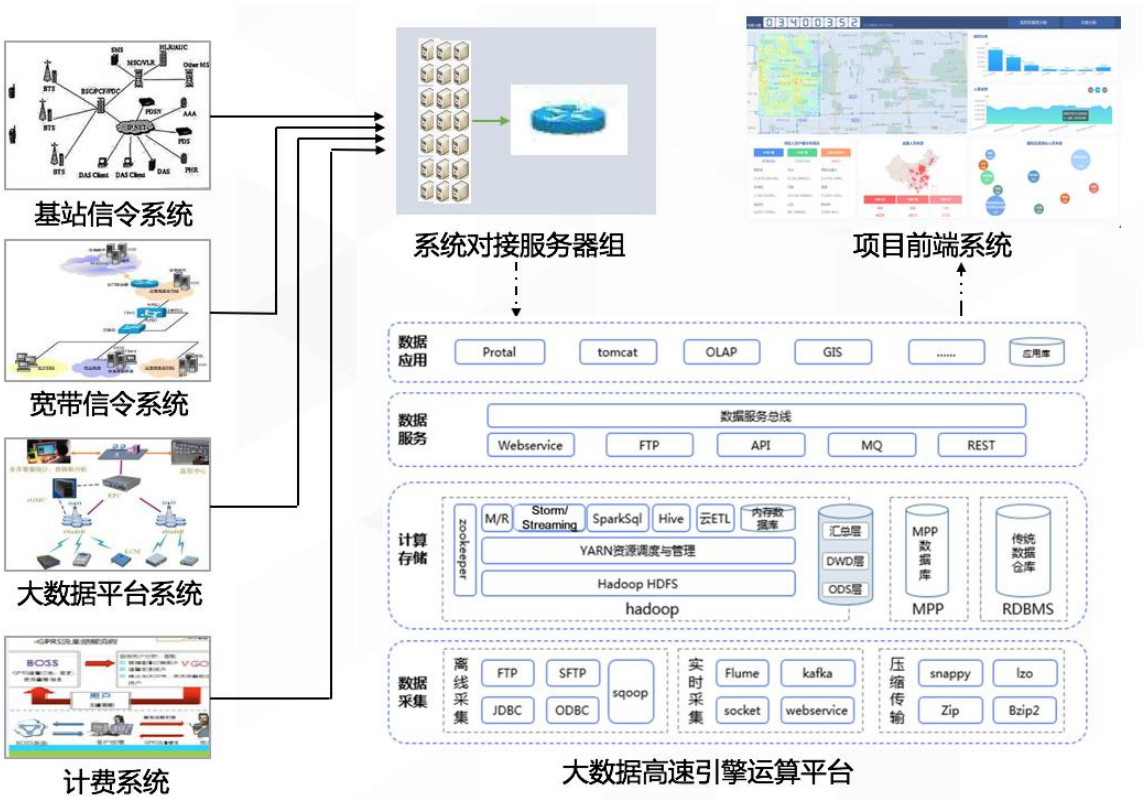
网络安全:系统应具备网络安全防护能力，支持访问控制、安全检测、攻击监控等一系列安全功能，应提供完整的网络安全监控、报警和故障处理功能，保障系统网络环境安全稳定。

1.3. 数据服务方案

1.3.1. 数据服务整体架构

在满足北京移动的数据采集和存储规范及输出的统计级数据成果的前提下，北京移动提供数据服务接入方案。

从通信网络各系统平台和基站等得到信令和业务数据，汇总至系统对接服务器组，通过大数据高速引擎运算平台进行去重、定位、纠偏、标识、模型过滤、分析、汇总等处理，将大数据分析结果提供给前台页面进行展现。前端系统负责展现数据并处理与使用者的交互功能。数据处理结果提供给前台界面，并做为数据接口、报告、报表的数据依据。

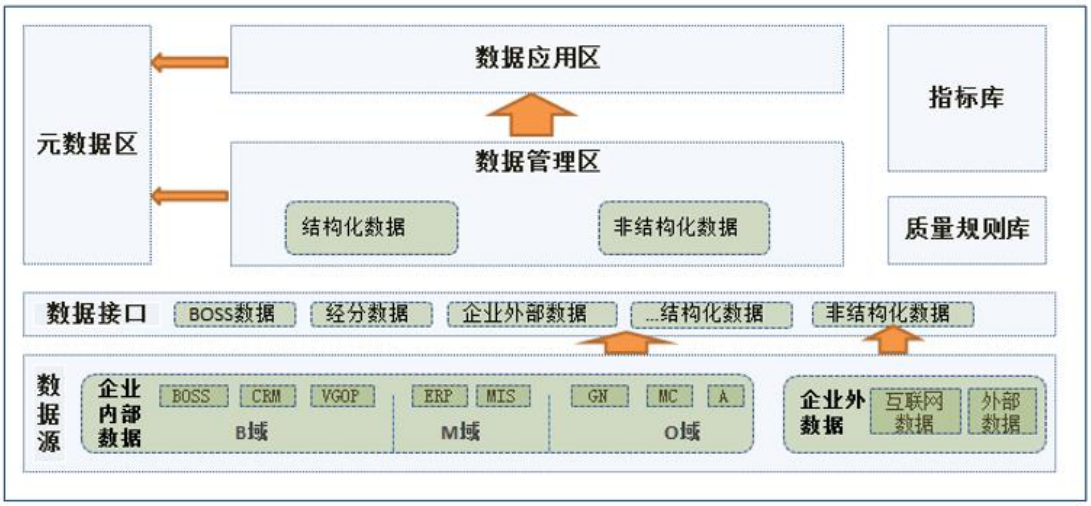


数据处理环节整体技术架构主要采用运营商 4 大系统数据，经数据采集、计算存储、数据服务、数据应用 4 层面处理，形成面向全端系统的结果数据。

- **数据采集层**：通过 Sqoop 或文件传输等工具分为两种方式传输，一部分是关系数据、文本数据采集，利用文件传输工具进行离线或压缩传输直接存储到 HDFS 上，还有一部分是流式数据的采集，利用 Flume 或 Kafka 等进行实时采集。
- **计算存储层**：计算存储分两种数据的处理，关系数据、文本数据（如日志数据），通过 Spark 或 Map Reduce 进行处理，流式数据通过 Spark Streaming 或 Storm 进行处理,处理过程及处理结果线落地到 HDFS 上，然后根据业务使用要求同步到 MPP、传统数据仓库等数据库平台中。
- **数据服务层**：主要是平台层提供数据服务的各种接口，如 Web Service、Rest、 API 等，通过安全验证提供数据应用服务。
- **数据应用层**：主要是通过 Tomcat、GIS、OLAP 等系统软件及分析型数据仓库为客户提供相应的数据应用服务。

1.3.2. 数据平台架构

数据架构主要分成六部分：数据源、数据接口区、数据管理区、元数据区、指标库和质量规则库，不同的数据区分别存储不同属性和功能的数据。具体的数据架构如下图所示：



数据源：主要指 B 域的经分、CRM、VGOP，M 域的 ERP、MIS，O 域的 GN、MC 等内部系统及互联网等外部系统中的结构化数据以及一些关键业务系统的日志或文本等非结构化业务数据。

数据接口区：数据接口区用于保存从数据源中采集的各种结构化数据或非结构化数据，如从经分系统中采集的经分数据，从 CRM 系统中采集的 CRM 数据等，这些数据的保存期限为 7+1 天，过期的数据平台定时清除。从 M 域采集的非结构化数据保留 1+1 天。

数据管理区：数据管理区主要有两个用途，一是非结构化数据的结构化处理，如关键业务的日志分析，可以将非结构化的数据转换成结构化的数据，二是对一些不能结构化的非结构化数据进行备份存储。

元数据区：元数据区用于保存在数据采集、清洗、转换和数据质量管理等各环节中产生的元数据信息，既包括业务元数据、也包括管理元数据和技术元数据。

指标库：指标库用于保存各业务分析的指标数据。

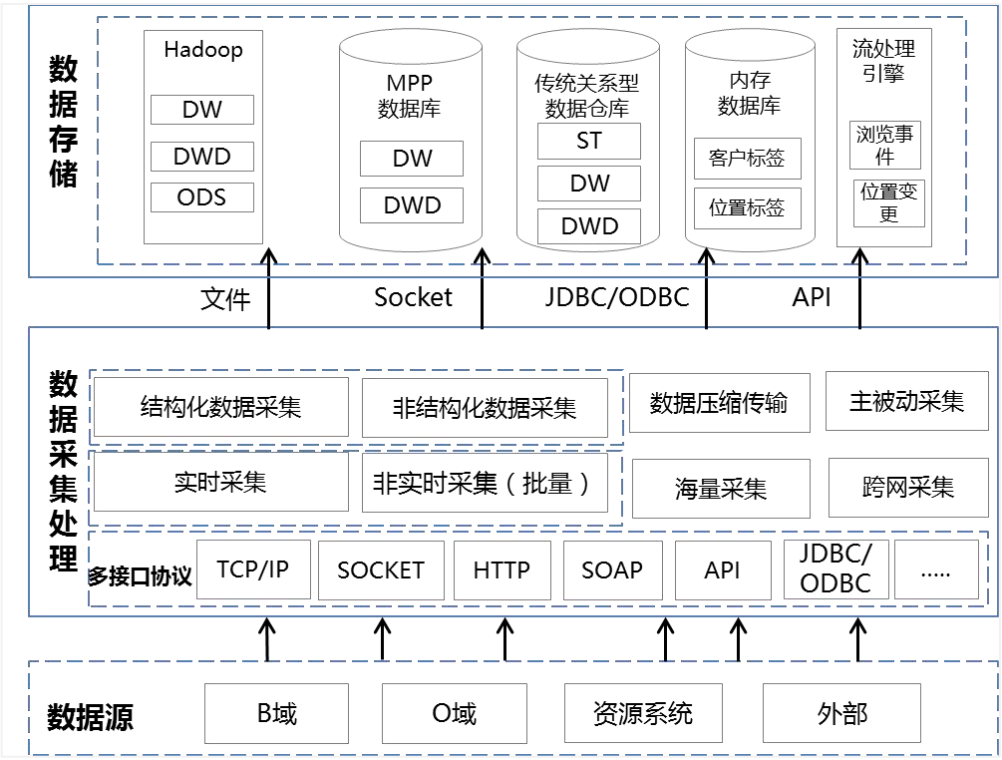
质量规则库：质量规则库主要用于存储数据质量管理的各种规则信息，以定量的方式对数据的质量进行评估和管理。

1.3.3. 数据采集

根据采购方的数据采集要求，完成相关数据的采集工作，基于获取到的信息，

完成号码关联工作，号码关联时前置条件为剔除三个月内无通话号码与剔除物联网卡。数据能满足采购方开展人口规模、人口结构、人口流动、职住通勤的数量及变化特点等业务以及人口定位、监测指标优化等工作。并按照三网融合平台建设项目的数据需求及技术规范，配合输出提供所需的统计级标准数据，对采集到的人口数据通过模型算法和结果汇总，形成多维度的分析结果。

1.3.3.1. 数据采集架构



本项目中平台统一采集来自 B 域、O 域、资源系统以及移动外部的数据，并将清洗、转换处理后的数据按流程处理，为整体平台各项功能服务提供数据支撑。

采用基于分布式、高可靠、高可用的、面向流的数据采集方法，可处理海量结构化、非结构化数据的采集、聚合和传输。

支持实时性、非实时性数据采集，通过实时流实时采集信令数据，支撑基于

位置的实时营销以及对外应用。通过 JDBC 等批量数据采集，实现大批量非实时性数据采集

支撑数据压缩传输，对采集过来的原始接口数据，进行文件合并拆分、数据清洗、数据转换、稽核校验、数据分发等数据加工操作，支持将数据并行加载到目标数据库中。

支持常见的 TCP/IP 协议、HTTP/SOAP、SOCKET 协议等多种采集接口协议。

支持主动被动采集模式，满足不同采集场景的需要。

支持代理机制的跨网采集能力，满足远程数据采集以及通过代理实现采集的防封策略。

1.3.3.2. 数据采集范围

本项目人口数据覆盖范围为北京市朝阳区，覆盖范围涵盖朝阳区，43 个行政街乡、81 个重点区域、50 个重点村、12 个商圈和 24 个旅游景区及临时新增区域。其中，朝阳区地理坐标、43 个街乡、81 个重点区域、50 个重点村、12 个商圈和 24 个旅游景区具体如下表：

朝阳区全境		
区域地理坐标为北纬 39°49′至 40°5′；东经 116°21′至 116°38′，		
覆盖 470.8 平方公里		
朝阳区 43 个行政街乡		
垡头街道	潘家园街道	来广营地区
三里屯街道	大屯街道	王四营地区
十八里店地区	东湖街道	亚运村街道

朝外街道	首都机场街道	和平街街道
南磨房地区	豆各庄地区	小关街道
太阳宫地区	常营地区	小红门地区
八里庄街道	酒仙桥街道	三间房地区
六里屯街道	东坝地区	孙河地区
安贞街道	左家庄街道	团结湖街道
崔各庄地区	奥运村街道	劲松街道
东风地区	建外街道	呼家楼街道
管庄地区	香河园街道	望京街道
高碑店地区	将台地区	金盏地区
平房地区	双井街道	麦子店街道
黑庄户地区	/	/

81 个重点和个性化区域		
CBD 核心区	原管庄乡–郭家场村	十八里店乡–十八里店村
CBD 中心区	原管庄乡–咸宁侯村	原石各庄村
奥林匹克公园	原管庄乡–小寺村	首都儿科研究所附属儿童医院
奥林匹克森林公园–北园	广渠快速路	首都经济贸易大学
奥林匹克森林公园–南园	合生汇中心	四海官鑫日用品市场
奥林匹克森林公园–廉洁奥运主题文化园	原黑桥村	四合庄村
奥林匹克公园中心区	原黑庄户乡–定辛庄区域	松榆里市场
北京盛华宏林批发市场	弘善家园	孙河乡地区
北京第二外国语学院	原花车展览区域	天丰利生活用品市场
北京工业大学	原黄港村	原王四营京哈（京沈）高速南侧市场群
北京望京医院	金瀛农贸市场	望京电子城西区北扩

北京朝阳医院	来广营村	原萧太后河整治与周边
北郎东	来广营乡城锦苑小区	小红门乡 65 号院
长安街沿线二	原奶西村	亚星香园农副产品市场
长安街沿线一	南磨房乡–广华新城	原一廊文创实验区
大柳树市场	农光里农贸市场	育慧南路及周边
大望京	农展北路集贸市场	朝京金旭菜市场
大洋路农副产品市场	潘家园旧货市场	朝阳公园
道尔泰农贸市场	平房乡–平房村	原朝阳区–崔各庄乡–南皋村
原定辛庄东村	原前苇沟片区	原朝阳区–黑庄户乡–定辛庄西村
东坝乡欣和园小区	三里屯	原朝阳区–黑庄户乡–苏坟村
豆各庄乡青荷里南山园小区	原沙子营村	中国传媒大学
对外经济贸易大学	原上辛堡村	中国科技馆
垡头街道双合家园小区	芍药居北里	中国医学科学院肿瘤医院
高碑店乡泰和园小区	原十八里店十里河市场一条街市场群	中日友好医院
国家会议中心	国家体育馆	国家游泳中心
香江北岸	国家速滑馆	中国残疾人体育运动管理中心

50 个重点村		
四合庄村	万子营东村	皮村
何各庄村	大鲁店一村	西村
马泉营村	原定辛庄东村	小店村
草场地村	原定辛庄西村	来广营村
东辛店村	原黑庄户村	新生村
费家村	郎各庄村	原平房村
奶东村	郎辛庄村	原石各庄村
原南皋村	双树北村	三间房西村
豆各庄村	双树南村	原横街子村

六里屯村	原苏坟村	老君堂村
八里庄村	万子营西村	原吕家营村
司辛庄村	小鲁店村	原西直河村
原郭家场村	幺铺村	小武基村
原咸宁侯村	东村	原孛罗营村
大鲁店二村	东窑村	观音堂村
大鲁店三村	黎各庄村	南花园村
王四营村	原肖村	

12 个商圈		
CBD 商圈	燕莎商圈	朝外商圈
三里屯商圈	望京商圈	常营商圈
亚奥商圈	双井商圈	东坝商圈
朝青商圈	太阳宫商圈	蓝色港湾商圈

24 个旅游景区		
奥林匹克森林公园	兴隆公园	京城梨园景区
朝阳公园	将府公园	北京蓝调庄园
奥林匹克公园	大望京公园	国家体育场（鸟巢）
日坛公园	庆丰公园	北京欢乐谷
团结湖公园	北小河公园	蟹岛绿色生态农庄
红领巾公园	四得公园	中国紫檀博物馆
北京蓝色港湾	白鹿公园	中华民族园
古塔公园	斯普瑞斯奥莱旅游商业文化体验区	温榆河公园

北京移动将严格按照行政区进行划分，根据行政区域经纬度坐标实现边界测绘，并充分考虑测绘区域内覆盖的基站数量等情况，对测绘区域进行微调，以

保证人口统计的准确性。根据采购人实际要求提供不低于 81 个重点区域的数据监测，具体列表以采购人提供为准。

1.3.3.3. 数据采集技术

（1）多种源系统的采集

支持多种源系统的采集，包括且不仅限于 text 文件、消息、dfs（HDFS 文件）、数据库等；

本方案支持可扩展到多种采集接口技术：基于 API 接口数据采集、基于 FTP/SFTP 接口数据采集、基于 kafka/flume 接口数据采集。

（2）基于kafka/flume接口数据采集

针对本项目建设，通过引入开源技术 Flume、Kafka，以及结合 Storm 计算引擎，针对流式数据进行抽取和分析计算，并将计算结果进行展示。该采集接口扩展，在实时数据采集时引入该技术框架。

Flume

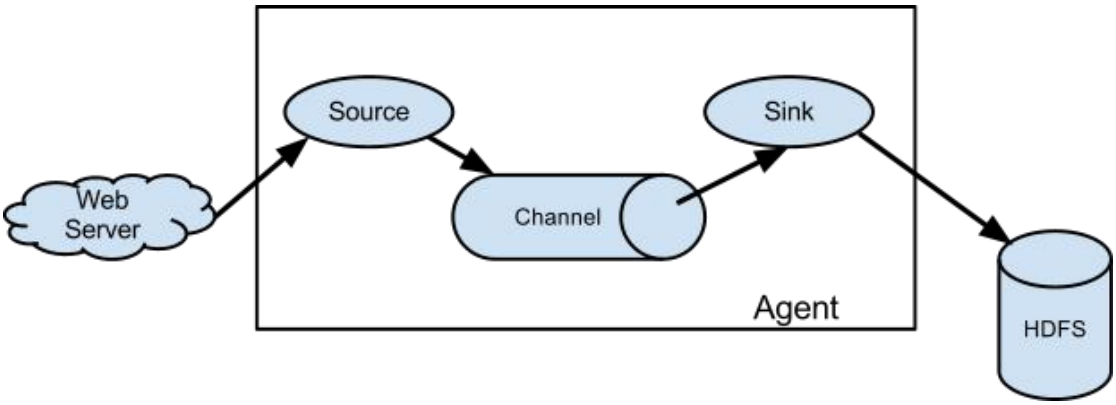
Flume是Cloudera提供的一个高可用、高可靠、分布式的海量日志采集、聚合和传输系统。Flume支持在日志系统中定制各类数据发送或采集源，用于收集数据，同时Flume提供对数据进行简单处理，并写入各类指定的数据接收方。

Flume具有以下优势或特点：

- 可靠的、可伸缩、可管理、可定制、高性能
- 声明式配置，可以动态更新配置
- 提供上下文路由功能
- 支持负载均衡和故障转移
- 功能丰富

◆ 完全的可扩展

Flume模型架构如下图所示：



Flume 模型架构

通过Flume实现对实时数据采集并存放到HDFS上，如网络日志类信息、网页和APP信息等。

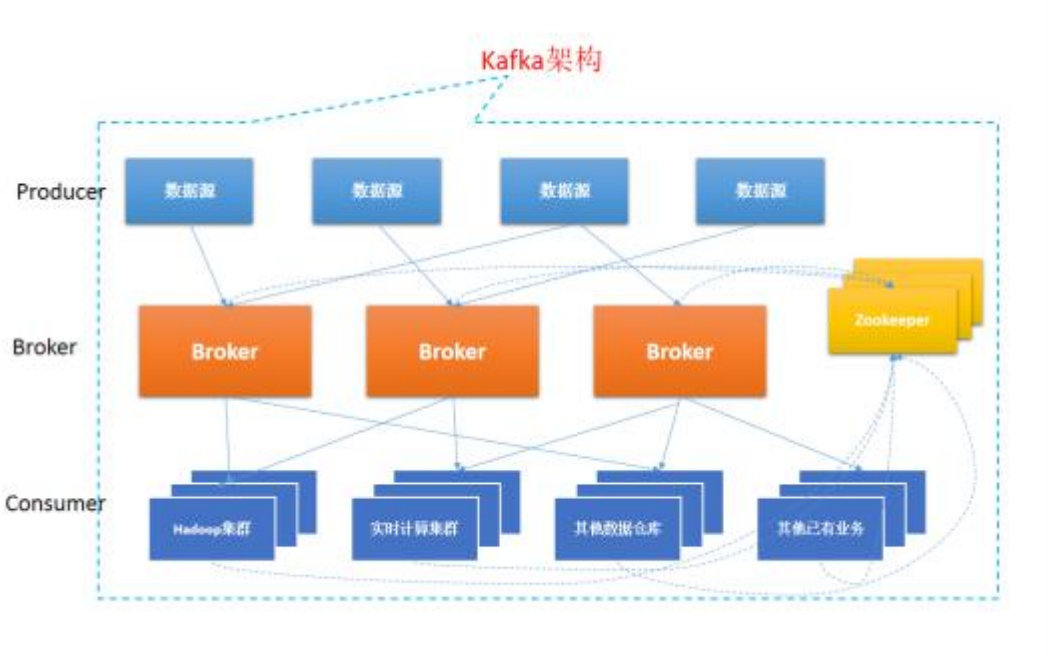
Flume是一个高可用、高可靠、分布式的海量日志采集、聚合和传输系统。Flume支持在日志系统中定制各类数据发送或采集源，用于收集数据，同时Flume提供对数据进行简单处理，并写入各类指定的数据接收方。

流式采集采用Flume日志处理数据流、Kafka等分布式消息系统。

Kafka

Kafka是一个低延迟高吞吐的分布式消息队列，适用于离线和在线消息消费，用于低延迟地收集和发送大量的事件和日志数据。

Kafka通过副本来实现消息的可靠存储，同时消息间通过Ack来确认消息的落地，避免单机故障造成服务中断。在Kafka中，Producer自动通过Zookeeper获取到Broker列表，通过Partition算法自动负载均衡将消息发送到Broker集群。Broker收到消息后自动分发到副本Broker上保证消息的可靠性。下游消费者通过Zookeeper获取Broker集群位置和Topic等信息，自动完成订阅消费等动作。如下图所示架构图：



Kafka 架构

Kafka 使用 Zookeeper 实现集群动态扩容，不需要更改 Producer 和 Consumer 的配置。集群新增的 Broker 自动向 Zookeeper 注册自身，客户端自动感知节点变化，并调整负载均衡策略。

Kafka 使用本地磁盘进行消息的持久化。磁盘的随机读写性能非常差，但是 Kafka 的模型是只有顺序读写，且充分利用了操作系统的 Pagecache，使用常见 SATA 盘可以达到高达上百兆的吞吐量，同时顺序读写的算法也是 O(1)稳定的，大大降低了数据延迟。

Kafka 多应用在前端和后端之前，完成架构解耦、一次发布多次订阅和消息的可靠存储。该技术目前已被广泛应用，可完成海量信息的输出，给系统架构带来多方面好处。

针对于持续不断的流式数据源，采用 Kafka 采集工具，通过调用 Kafka 提供的 API，将数据生产者将数据汇聚到 Kafka 集群，再由数据消费者 Storm 或其他流式数据处理平台进行计算处理或者存储。

（3）基于 API（webservice）接口数据采集

API 封装实现：

系统采用统一封装函数库的方工，通过类 sql 编写，或函数调度，来屏蔽底层差异性，实现跨平台的 API 调度。

通过可视化的流程界面，拖拽方式实现对函数的编排，对每个节点函数编写参数，实现数据加工功能，降低开发难度。

如：将某类数据文件加载到数据库中，开发功能只要指定数据文件路径和目标表。系统执行时调用 Hadoop 的命令。

API 函数封装示例

API 分类管理：

根据数据处理过程，将 API 分为：数据输入(Reader)、数据输出(Writer)、记录集处理(Process)、字段级处理、流程控制类、数据检查类、数据交换类，并提供了一套可扩展的机制。

API 调用方式：

通过 webservice 来调用

如，格式：http://dp 服务器地址/DP-Excuter/api/函数名?para=；支持同步或异步的调用，根据函数不同返回值不同。

通过 SDK 调用

DpClient =new DpClient(url,username,password)；

Result =DpClient.call(api 名，参数.....)；

或

Result=DpClient.execsql(sql 语句,dbname)；

通过 UDF 调用

有些函数 API 可以注册成 UDF 函数，在 sql 中就可以进行调用。通过 SQL 语句：Select dacp_map(dim_channel,dim_channel_map) from xxxx；可将表的 dim_channel 字段按照 dim_channel_map 的映射进行编码转换。

4）基于 FTP/SFTP 接口数据采集

基于 FTP/SFTP 接口，系统支持普通文件、压缩文件、压缩文件流等不同类

型文件的采集与上传，其支持的文件格式包括：XML 格式、Excel 格式、CSV 格式等。基于这些不同的文件类型、文件格式，通过采集、抽取过程可形成新的数据文件，以满足不同目标数据库的需要。FTP 接口适合于对实时性要求不高的场景下的大数据量信息交换。

接口信息配置

如下图所示，基于其 FTP 接口采集配置界面：

1.支持源服务器待采集目录、目标服务器存放路径的设定；

2.支持采集完成后，源服务器文件多种处理模式的选择，包括删除、保留、移动。删除，即会将源服务器上的文件进行删除处理；保留，即不改变源服务器文件；移动，即指将源服务器的文件移动到指定的转移目录中；

3.通过通配符的设置，支持文类型的筛选、过滤；如：指定 CSV 格式等；

4.支持采集后文件存放的备份设置，包括指本地备份，采集文件后在本地进行文件的备份；

5.压缩采集：实现传输过程的压缩处理，即，将源服务器文件传输过程实现压缩处理，目标服务器实现压缩文件格式存放，以提升文件后续传输、处理过程中的 IO 性能并节省宽带资源，系统支持不同类型的压缩格式。如远程某文件为 10G 的 txt 文件，把其压缩为 zip 文件后为 2G，文件采集组件会直接读取压缩流，不落地到服务器实现压缩流加载，使得 10G 的文件实际耗费的网络带宽为 2G，这样大大节省了带宽资源，并且有效的提高了采集速度。装载组件对流进行解压装载。

采集触发设置

通过采集任务 job 流程设置，并配置好触发时间、周期、优先级等相关信息，

以满足采集频率的设置，如：满足月、日、小时、分钟等多种周期粒度的数据。

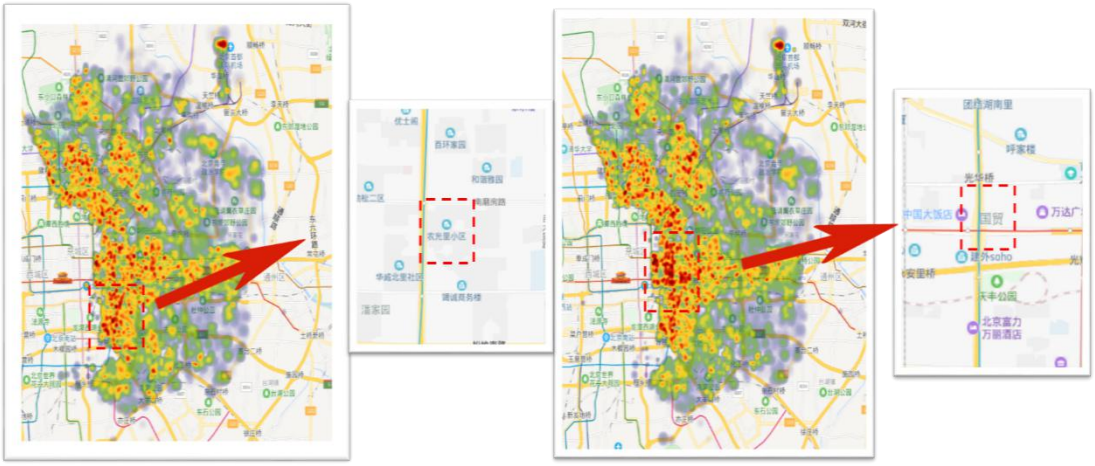
目标装载规则设置

采集过程中，通过设置的采集规则进行监控，对异常情况进行实时告警，并输出相应的日志信息。采集完成后依据数据采集校验规则，对数据进行校验，包括采集格式校验、采集数据一致性、完整性校验。

文件抽取支持定长和不定长两种格式。对于不定长抽取，允许功能在不违反管理的前提下指定分隔符，系统至少提供“|”，“,”，“TAB 符号”，三种缺省的分隔符方式工功能选择。通过约定，可实现源文件按约定的规则，将数据信息装载到目标库（或文件）中。

1.3.4. 热力图数据范围

采集到的数据对区指定平台热力图提供支撑，根据客户指定区域提供实时数据展示。



1.3.5. 三网数据融合

按照三网融合平台建设项目的数据需求及技术规范，配合输出提供所需的统

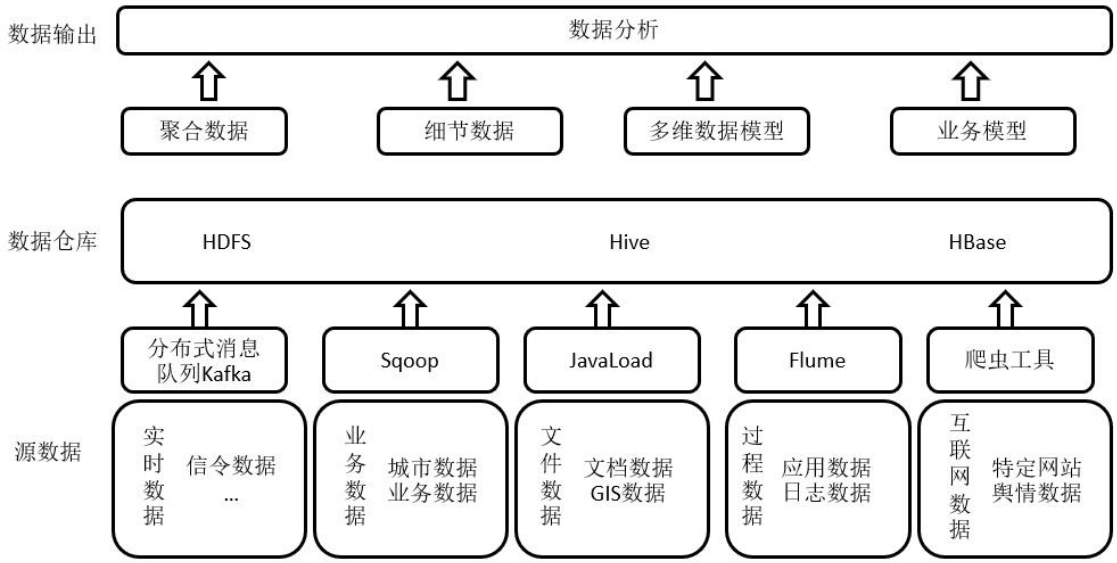
计级标准数据。

1.3.6. 数据标准化处理

1.3.6.1. 数据处理流程

1.3.6.1.1. 数据接入

由于数据时效性、数据存储介质、数据存储类型和数据传输方式的多样性，针对每一类数据，借助不同的导入工具，实现不同源数据、不同结构数据的接入，如下图所示：



实时批量数据以分布式消息队列的形式由kafka进行分发处理；结构化型文件数据由java程序分发处理；内部业务系统数据由Sqoop等工具导入；日志数据由Flume工具导入；对于互联网使用Python程序获取并导入；

1.3.6.1.2. 异构数据处理

大数据采集过程中通常有一个或多个数据源，这些数据源包括同构或异构的数据库、文件系统、服务接口等，易受到噪声数据、数据值缺失、数据冲突等影

响，因此需首先对收集到的大数据集合进行预处理，以保证大数据分析与预测结果的准确性与价值性。

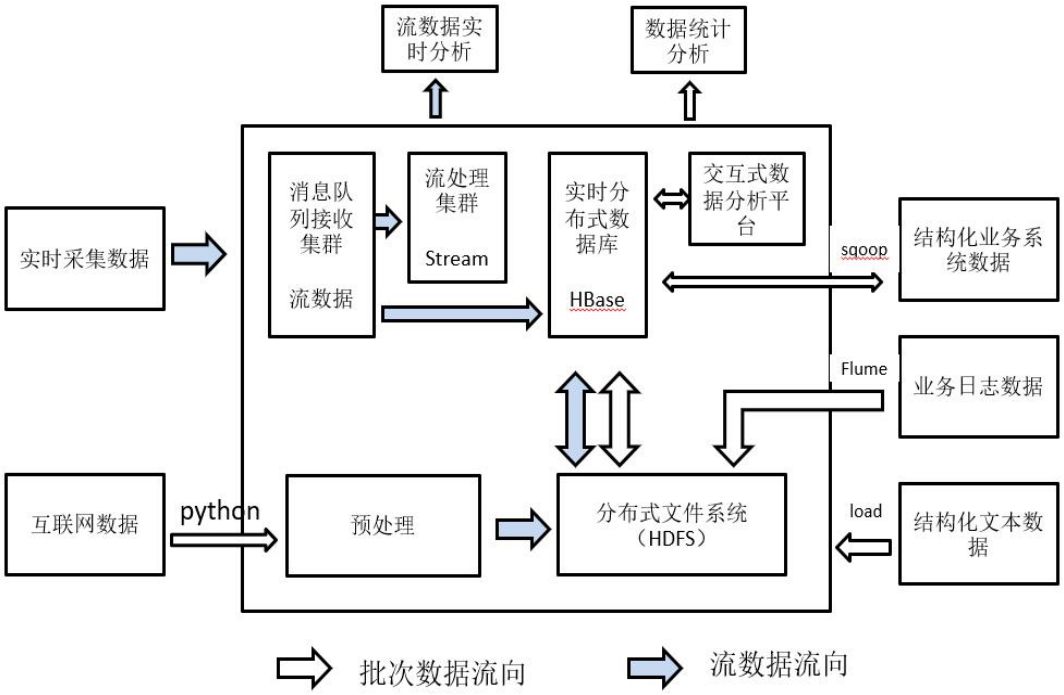
大数据的处理环节主要包括数据清理、数据集成、数据归约与数据转换等内容，可以大大提高大数据的总体质量，是大数据过程质量的体现。 数据清理技术包括对数据的不一致检测、噪声数据的识别、数据过滤与修正等方面，有利于提高大数据的一致性、准确性、真实性和可用性等方面的质量;

数据集成则是将多个数据源的数据进行集成,从而形成集中、统一的数据库、数据立方体等，这一过程有利于提高大数据的完整性、一致性、安全性和可用性等方面质量;

数据归约是在不损害分析结果准确性的前提下降低数据集规模，使之简化，包括维归约、数据归约、数据抽样等技术，这一过程有利于提高大数据的价值密度，即提高大数据存储的价值性。

数据转换处理包括基于规则或元数据的转换、基于模型与学习的转换等技术，可通过转换实现数据统一，这一过程有利于提高大数据的一致性和可用性。

总之，数据处理环节有利于提高大数据的一致性、准确性、真实性、可用性、完整性、安全性和价值性等方面质量，而大数据处理中的相关技术是影响大数据过程质量的关键因素。处理过程如下图所示：



1.3.6.1.3. 实时流处理

1) 实时流处理功能架构

流处理架构实现实时数据的统一采集、处理和存储等，平台包括集成开源组件模块以及自身的处理和管理模块，开源组件包括 Flume、Kafka、Storm、Redis 等，除此之外提供基于流数据的采集接入、信令处理引擎、数据共享输出、可视化配置和监控管理、一站式开发分析、系统模块化管理等功能。

采用业界主流的 Lamda 架构，支持使用开源组件作为计算引擎实现的实时流数据处理平台。具有如下特点：

- 实时性高，能够实时处理流数据
- 通过插件管理业务逻辑，简化新业务逻辑开发流程
- 易用性强，图形化部署和配置
- 扩展性强，可以动态扩充集群规模

使用 Lamda 架构设计，在并行处理中，对流入的数据不做改动，数据进入

流处理过程中进行分析,并将分析的结果流出系统。在整个流数据处理过程中(数据流入－>分析－>流出),避免使用落盘操作。极大地提升了数据的处理速度,保证了数据处理的实时性要求。支持多种数据流的输入方式,例如:Socket,文件,Kafka 等。可以满足多种场景不同数据源接入的需求。同时也支持将数据处理结果以多种方式输出,例如 Kafka,Codis,HDFS 等。此外还具有高实时性、高扩展性、高易用性等特点,能够方便用户快速搭建和使用流数据处理平台。

整体流处理平台功能包括:数据接入、Storm 处理引擎、数据输出、以及配置管理、监控管理、可视化开发工具、系统管理等。

数据采集:内置多种 adapter,支持流数据的接入、存储转发、数据处理和固化等,通过按照规则完成接口适配的配置,确保和数据源的连通性,支持 HTTP、socket、文件、HDFS 等。

处理引擎:对大吞吐量的数据实时性要求极高的前提下,处理来自不同的业务系统的多样数据,同时通过预处理很好的解决数据格式间的差异化,提供一个性能稳定且功能强大的处理引擎,来承载业务快速发展需求和业务系统对数据的灵活订阅能力,不仅能满足按需订阅,更加满足数据隔离的诉求,另一方面,实

时数据常需要与用户历史业务数据相关联才能产生更大的价值，数据闪存能够将实时数据提供给传统的业务系统，实现实时数据与历史数据的高效访问关联。

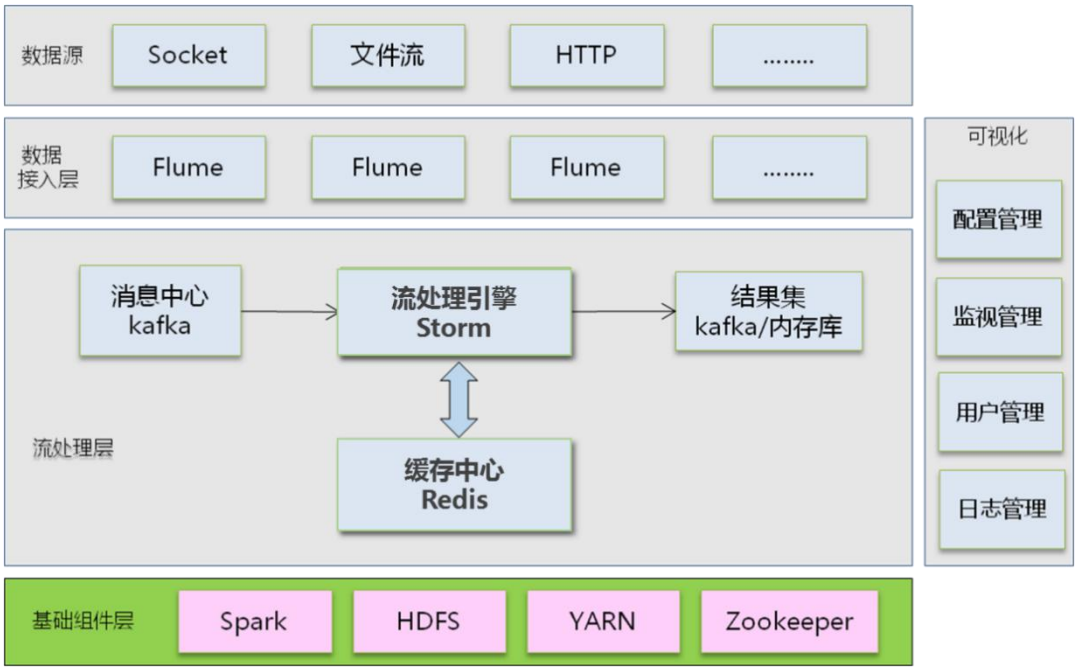
配置和监控：提供易用、直观的界面化的配置和监控功能，让系统管理和运维人员有效掌握系统配置和运行的信息，实现管理上的透明化与可视化。

可视化开发工具：流式数据实时获取、计算、分析、统计、分发的一站式开发工具。

数据传输：提供统一数据适配层，满足各种文件、消息的数据转换适配器，可支持多种异构数据源的统一采集与数据接入。对各种数据源进行接入适配，支持包括各种消息、文件等格式的数据接入。

系统管理：除了基础的用户管理、日志管理之外，还提供消息队列管理、复杂事件管理、内存访问管控等功能。

2) 实时流处理技术架构



数据接入层：

数据接入层的主要作用是将外部数据源的输入数据转换为流处理能够识别

和读取的消息队列 Kafka 数据。由 Flume 提供数据接入、转换功能（规则可通过可视化的配置管理进行配置），支持多种外部数据源的数据，默认支持 Socket 数据和文件流数据，可依据业务要求，进行插件式定制化开发接入其他类型的数据。

流处理层：

采用 Kafka 来接收和存储来自数据接入层 Flume 的数据流。

采用 Storm 来主动从 Kafka 中消费数据流，并且根据选定的标签进行数据加工，在处理过程中，使用 Key-Value 内存库保存迭代运算所需的关键业务数据以及中间数据，以提升整个流处理的效率。最终将按照订阅需求输出到 Kafka 或者内存库 Redis 中，提供外部系统使用。

采用 Redis 来存储相关业务码表数据和中间数据集。

可视化-配置管理：

流处理提供一站式的配置管理界面，支持系统参数配置和处理规则配置。

系统配置

流平台集群的系统相关配置。例如：Zookeeper 地址，Spark 运行模式，Redis 地址，Kafka 地址等。

处理规则配置

数据源接入方式：Socket、文件流等，Source 信息，Channel 信息，Sink 信息等。

消息队列参数配置：配置 Kafka 的重要参数，例如：接收线程数、I/O 线程数、处理任务线程数、消息大小、超时时间、topic 信息等。

流处理配置：预处理规则配置，标签配置，订阅规则配置，输出配置等。

3) 基于 kafka 构建的消息中心

流处理平台以 Kafka 分布式消息系统作为数据交换中心，具备高吞吐量、数据持久化、易于且不停机扩展、基于 Zookeeper 的无状态集群特性。

Kafka 是一个是分布式发布-订阅消息系统，最初由 LinkedIn 公司开发，用作 LinkedIn 的活动流（activitystream）和运营数据处理管道（pipeline）的基础，之后成为 Apache 项目的一部分。现在它已作为多种类型的数据管道（datapipeline）和消息系统应用于多家不同类型的公司。

Kafka 的主要特点：

1）同时为发布和订阅提供高吞吐量。据了解，Kafka 每秒可以生产约 25 万消息（50MB），每秒处理 55 万消息（110MB）。

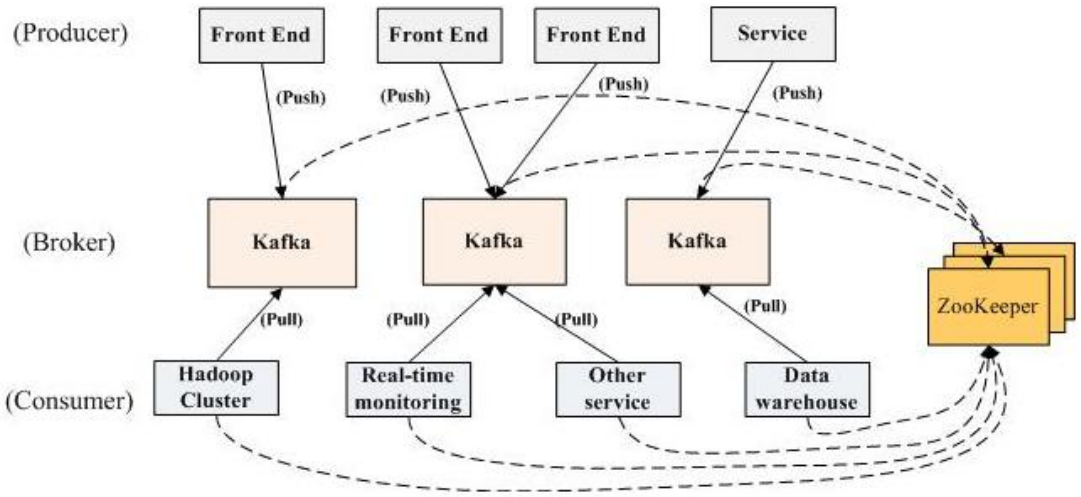
2）可进行持久化操作。将消息持久化到磁盘，因此可用于批量消费，例如 ETL，以及实时应用程序。通过将数据持久化到硬盘以及 replication 防止数据丢失。

3）分布式系统，易于向外扩展。所有的 producer、broker 和 consumer 都会有多个，均为分布式的。无需停机即可扩展机器。

4）消息被处理的状态是在 consumer 端维护，而不是由 server 端维护。当失败时能自动平衡。

5）支持 Online 和 Offline 的场景。

Kafka 的架构：



Kafka 的整体架构非常简单，是显式分布式架构，producer、broker（kafka）和 consumer 都可以有多个。Producer、consumer 实现 Kafka 注册的接口，数据从 producer 发送到 broker，broker 承担一个中间缓存和分发的作用。broker 分发注册到系统中的 consumer。broker 的作用类似于缓存，即活跃的数据和离线处理系统之间的缓存。客户端和服务器端的通信，是基于简单，高性能，且与编程语言无关的 TCP 协议。几个基本概念：

1. 主题（topic）指 Kafka 处理的消息源（feedsofmessages）的不同分类

2. 分区(Partition)是 Topic 物理上的分组，一个 topic 可分为多个 partition，每个 partition 是一个有序队列。它中的每条消息都会被分配一个有序 id（offset）。

3. 消息（Message）是通信的基本单位，每个 producer 可以向一个 topic（主题）发布一些消息。

4. 生产者（producer）是消息和数据生产者，向 Kafka 的一个 topic 发布消息的过程叫做 producer

5. 消费者（consumer）订阅 topics 并处理其发布的消息的过程叫做

consumers

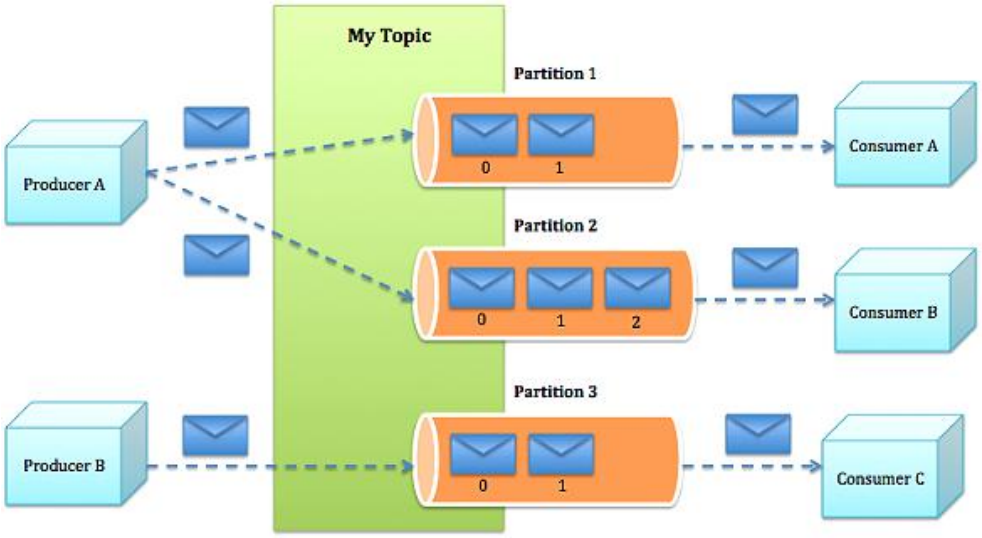
6. 缓存代理（broker）Kafa 集群中的一台或多台服务器统称为 broker

生产者（producer）通过网络发消息到 kafka 集群，而 kafka 集群则以下面这种方式对消费者进行服务：

1. Producer 根据指定的 partition 方法（round-robin、hash 等），将消息发布到指定 topic 的 partition 里面

2. Kafka 集群接收到 Producer 发过来的消息后，将其持久化到硬盘，并保留消息指定时长（可配置），而不关注消息是否被消费。

3. Consumer 从 Kafka 集群 pull 数据，并控制获取消息的 offset



基于 Kafka 的分布式消息队列具有高吞吐量。对于单次请求，可以实现请求时间低于 5 毫秒；对于高并发应用场景的请求，可以实现万次/秒、通过增加服务器数据可以实现更高的吞吐量。

流处理平台以 Kafka 分布式消息系统作为数据交换中心的底层服务，通过封装优化数据写入、读取的操作接口，达到数据的高效、高吞吐、高可用的消息流转。

1.3.6.1.4. 结构化文件数据处理

对于结构化的数据文件，通过 SFTP 等传输方式传至对应的文件接口服务器，由数据导入程序批量加载导入。数据处理流程如下：

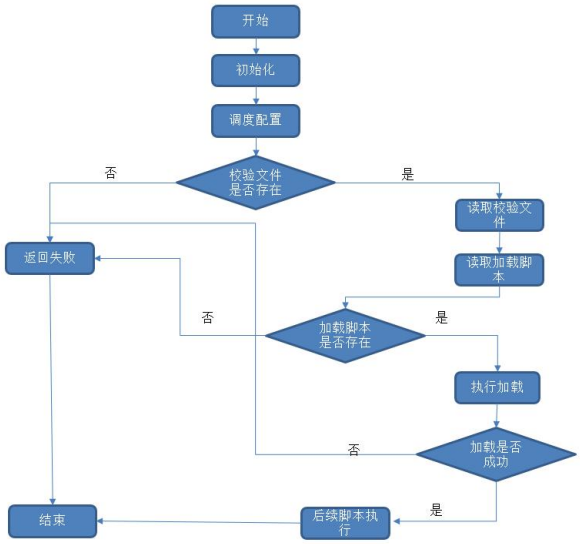


图-结构化文件数据加载

1）接口实现方式

在满足本接口规范要求的前提下，本接口系统的数据交换采用文本文件方式。

为了保证接口数据文件的一致性，避免数据类型和格式错误，在形成所传输的文件之前，数据提供方将数据转换成本接口规范所规定的数据类型和格式。

数据提供方与数据接收方的明细数据传送通过文件方式实现。一个接口的内容可以用一个或多个文件传输，但同一个文件不能传输两个或两个以上接口的内容。

2）传输方式

文件类接口传输方式为单向数据获取。该类接口的工作模式为：数据提供方周期性预先生成接口文件，由数据接收方获取。

数据提供方直接将接口文件存放到数据提供方接口机的指定目录，数据接收

方到数据提供方接口机的指定目录获取数据。

3) 传输过程

➢ 接口处理模式

数据提供方根据不同接口的要求，遵循统一的命名规则生成接口文件，传输到数据提供方接口机的指定目录下。数据接收方通过 FTP 或双方约定的应用层协议到该目录下获取接口文件。在接口文件接收后，数据接收方在本端接口机的指定目录下生成文件级校验报告和记录级校验报告，供数据提供方获取。

➢ 接口及时性要求

按小时接口：保证每个整点（最多延迟 40 分钟）在数据提供方接口机生成完上一个时段的数据；上一个时段无数据时需要传空文件。

按日接口：保证按本接口规范定义的时间点前，在数据提供方接口机生成完前一天的增量或全量数据；前一天无数据时需要传空文件。

按月接口：保证按本接口规范定义的时间点前，在数据提供方接口机生成完前一个月的增量或全量数据；前一个月无数据时需要传空文件。

➢ 接口文件保存时长

保证不同类型的接口文件在文件生成后，数据提供方按照一定的周期保存接口文件：

1) 按小时生成的接口文件，至少保留 48 小时；

2) 按日生成的接口文件，至少保留 7 天；

3) 按月生成的接口文件，至少保留 15 天。

➢ 接口文件格式

本接口规范中如无特殊要求，需遵循此章节定义的接口文件格式。

回车换行符（0x0D0A）；

ASCII 码 0x7C（‘|’）。

为了保证数据的准确性以及接口文件中的记录各值域在有效的取值范围内，数据中均不能包含 0x0D0A（回车换行符）和字段间分隔符。

➢ 接口数据要求

数据提供方在生成接口文件时，必须遵守如下数据转换规则。

➢ 编码格式

汉字：GBK 内码

西文：ASCII 码

➢ 数字格式

在接口文件中，数字的表示必须规范，小数点的前后必须有数字，如：0.01 或 34.0,不能用“.01”或“34.”表示；

符号处理：数字最高位的左边第一位为符号位。对于负数，符号位为“-”，正数不用加符号位

➢ 日期类型

日期类型统一采用 YYYYMMDD 格式，不允许出现空值，且 YYYYMMDD 必须为有意义的日期。

YYYY 为四位数字，必须是有效的年份

MM 为两位数字，必须是有效的月份（01-12），月份小于 10 的时候首位补 0 使得月份为 2 位。

DD 为两位数字，必须是有效的日期（01-31），日期小于 10 的时候首位补 0 使得日期为 2 位。

➢ 时间类型

统一采用 HHMMSS 格式：

HH 为两位数字，必须是有效的小时（00-23），24 小时制，小时小于 10 的时候首位补 0 使得小时位为 2 位；

MM 为两位数字，必须是有效的分钟（00-59），分钟小于 10 的时候首位补 0 使得分钟为 2 位；

SS 为两位数字，必须是有效的秒（00-59），秒小于 10 的时候首位补 0 使得秒位为 2 位。

- 接口字段要求

当字段备注列有“可为空”或“选填”时，字段无数据时置空。

- 补传重传说明

补传

针对累计数据

即保证累计所有周期数据时，一个营销活动对应的一个指标只有一条记录。

重传

适用于：所有接口

数据提供方告知接收方需重传的接口及周期，接收方清除旧数据周期后，再重新接入新周期数据。

4) 接口协议

文件类接口的传输协议应支持 SFTP。同时提供的 sftp 服务的账号应具备在数据存放目录上的增删改权限.

5) 文件命名规则

- <文件类别><接口编号>_<数据日期>_<4 位序列号>.AVL

说明：

该文件用于数据提供方向数据接收方提供本接口规范要求的数据；

- <文件类别>：为“I”表示文件为获取全量数据，为“A”表示文件为增量文件时；
- <接口编号>：填入 6 位数字编码,此编码参考【文件接口单元明细】章节；
- <数据日期>：填入 10 位数字，按不同的接口类型确定各位的值，小时取值为“YYYYMMDDHH24”，日接口则为“YYYYMMDD00”，若为月接口则为“YYYYMM0000”；
- <4 位序列号>默认为 4 位，为数字型，序列号是用于描述同一个接口单元在同一个抽取周期中的文件顺序号，若一个接口单元被分割成多个文件则其流水号必须不同，并且从 0001 依次递增(如： 0001、0002、0003)，即位数不够 4 位首位用 0 补足。

备注：一个接口若抽取文件大小超过 3GB 则需要拆分文件。

- 在接口文件生成过程中，定义接口文件扩展名为.tmp，待接口文件生成成功后，再改为.AVL，以保证数据文件的完整性。

6) 文件目录及维护

➢ 数据提供方接口机文件目录的划分

所有数据存放于同一个目录。

➢ 数据提供方接口机文件目录的维护

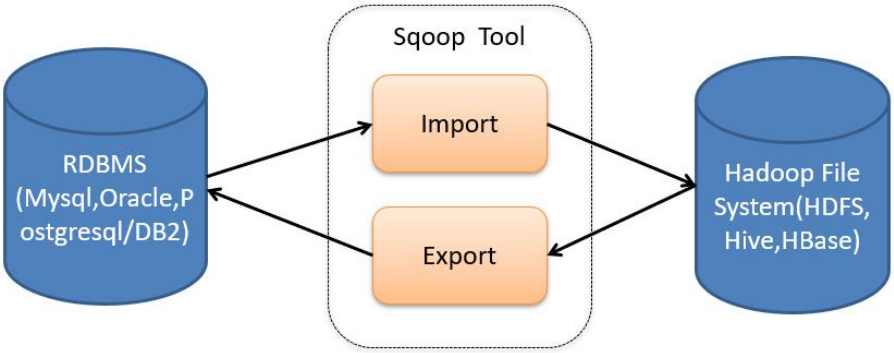
负责目录的创建和权限管理。

➢ 数据接收方接口机文件目录的维护

负责目录的创建和权限管理。

1.3.6.1.5. 业务数据处理

业务数据将通过 Sqoop 接入。Sqoop 是一个用来将 Hadoop 和关系型数据库中的数据相互转移的工具，可以将一个关系型数据库（例如 ：MySQL ,Oracle ,Postgres 等）中的数据导进到 Hadoop 的 HDFS 中，也可以将 HDFS 的数据导进到关系型数据库中。



对于某些 NoSQL 数据库它也提供了连接器。Sqoop，类似于其他 ETL 工具，使用元数据模型来判断数据类型并在数据从数据源转移到 Hadoop 时确保类型安全的数据处理。Sqoop 专为大数据批量传输设计，能够分割数据集并创建 Hadoop 任务来处理每个区块。

1.3.6.2. 数据清洗

数据清洗一般包括：

- 原始数据集的分割
- 缺失值的处理
- 变量的转化

a.原始数据集的分割

在数据挖掘中，一般会将原始数据集分成若干子数据集，这些子集互不相交（即任意两个子数据集都不包含原始数据集中的同一条观测），具体地说，原始数据集可以分割为以下几种子数据集：

训练数据集。所谓的训练数据集指的是原始数据集中用来创建模型的子数据集。

验证数据集。该数据也是原始数据集的一个子集，在模型创建后，可以用这部分数据进行验证，如果模型效果不佳，则需要对模型进行重新训练。

测试数据集。在建模过程中，不需要用到测试数据集。当有多个模型进行比较时，可以使用测试数据集。由于建模阶段没有用到测试数据集，因此在这个基础上进行模型优劣的比较会相对客观。

b.缺失值的处理

对缺失值的处理也是数据挖掘中的难点之一。如果一条观测在某一个变量上的取值为缺失值，那么称该观测为不完整观测。数据清洗加工中的多数过程步都不会考虑不完整观测。虽然这样处理比较方便、简单，但是也存在一定的问题，例如，不完整观测虽然在某个变量上的值缺失，但是在其余变量上的信息仍然是有用的。

对取值为连续型的变量，缺失值的处理方法一般有以下几种：

•指定默认值。

•平均值。

•中位数。

•（最大值+最小值）/2。

•最小值。

•最大值。

•不做处理。

对取值为离散的分类变量，缺失值的处理方法一般有以下几种：

•指定默认值。

•众数。(在变量的所有可能取值中，最经常出现的一个)

•最小值。

•最大值。

•不做处理。

c.变量的转化

在数据清洗加工中，变量的转化也是一个重要的环节。对于连续变量，常见的转换方法有：

简单变形（Simple Transformations）。常见的对输入变量进行简单变形的方法有：LOG、开方根、取逆、平方、取指数以及标准化等，通常使用变换后的新值来代替原来的输入变量进行分析。例如，在 LOGISTIC 回归中，理想情况下，LOGIT（因变量）与其他变量（自变量）之间呈线性关系。假设现有一个变量与 LOGIT 之间呈二次关系，在这种情况下，可以考虑对变量进行平方变形，将平方后的变量与原变量一起纳入建模分析的变量中。

BUCKET 转换法。该方法可以将连续变量转化为分类变量，具体做法是：将最小值与最大值之间的区间分成长度相等的 n 组,任何两组之间不包含同一条观测，且每一条观测恰好落入其中的某一组中。例如，将年龄分成 0~25、25~50、50~75 以及 75~100 四组，每一组所包含的观测数可能不一样。

数据清洗加工的基本目的是从大量的、可能是杂乱无章的、难以理解的数据中抽取并推导出对于某些特定的人们来说是有价值、有意义的数据。建立用于支撑快速查询和多维展现的指标体系及用户标签库,可以实现不同业务需求场景的快速数据支撑。

1.3.6.3. 数据关联

构建数据关联规则，对数据进行关联处理，保证数据关联关系的准确性。

1.3.6.4. 数据脱敏

根据国家数据隐私安全管理要求，对涉及到的个人隐私数据（包括用户手机号码、IMSI 号、IMEI 号、消费金额等）进行不可逆脱敏处理。

支持对数值、文本等多种类型数据进行脱敏，支持多种脱敏方式包括不可逆加密、区间随机和掩码替换等。

1.3.6.5. 数据提供和接口改造

经过加密脱敏等安全处理后的运营商数据需以安全可靠的方式提供给采购方，并在运营商机房内进行联合建模使用，对于实时数据提供，可采用数据库表或 Kafka 消息队列授权提供；对于非实时数据，采用数据库表授权提供。我方在服务期内根据采购方建模需求，协助配合采购方进行必要的接口改造工作。

1.4. 数据计算服务方案

1.4.1. 数据计算流程

1.4.1.1. 基础数据计算

人口大数据使用到的位置信令数据共两种，MME 口数据和 signal_5g 数据 MME 口数据。MME 口数据是 4G 手机位置数据，当用户在每进出一个基站都将产生数据信号，在手机完全静止的情况下，每 5 分钟产生一次数据。signal_5g 数据是 5G 手机位置数据,当用户处于 5G 网络时，进入基站离开基站以及在某

个基站停留,都会记录该用户基站位置信息，当用户在某个基站停留静止情况下,每 5 分钟产生一次数据。

使用计费账务系统登记的功能号码列表排除功能数据。通过语音通话话单排除连续 6 个月都没有语音通话的号码。排除物联网专用号段 106、144 开头的 13 位号码。

- 通过信令数据加工生成用户的位置信息，获取用户在不同时间出现的位置网格或经纬度；
- 通过信令数据加工生成用户的出行特征信息，获取用户在不同时间出行的 OD，轨迹，交通方式等信息；
- 通过 CRM 等数据，生成用户基础属性信息，包括性别、年龄、话费、籍贯地、归属地等

标签数据计算：通过原始数据和计算生成的基础数据，进一步计算输出用户的各类标签,包括用户职住类型的判定标签,用户是否商旅就医等到访类型标签,确认用户是否活跃的通话量标签、记录用户是否进出京信息的漫游特征标签、记录用户是否物联网卡等非人设备类型的终端属性标签、记录用户是否使用上网卡的上网属性标签等

地理数据计算：基于用户位置和出行数据，计算用户在北京市区街乡镇村的行政区划、重点监测区域、全城均匀网格划分区域等各类地理区域内的活动情况

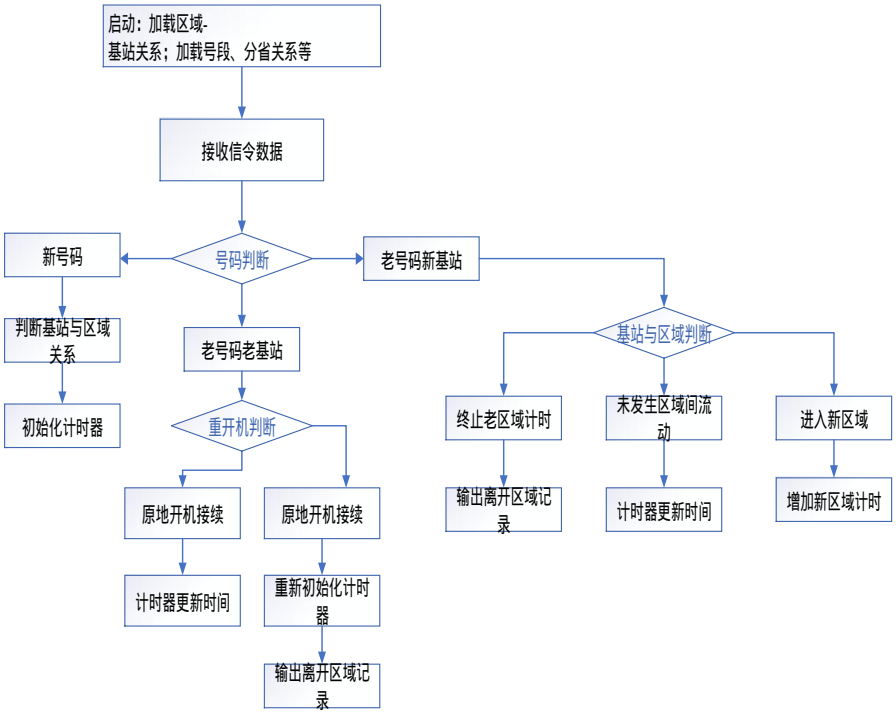
监测口径计算：根据业务指标统计需求，计算不同监测口径下的人口指标统计。

1.4.1.2. 用户信令流处理过程

流处理部分负责实时接收由信令采集系统采集到运营商基站扇区的信令数

据，数据实时更新，用流处理地计算按用户进行各区域的计时处理，并输出明细数据。

主要处理流程如下：



在计时过程中，需要判断开机关对统计结果的影响：

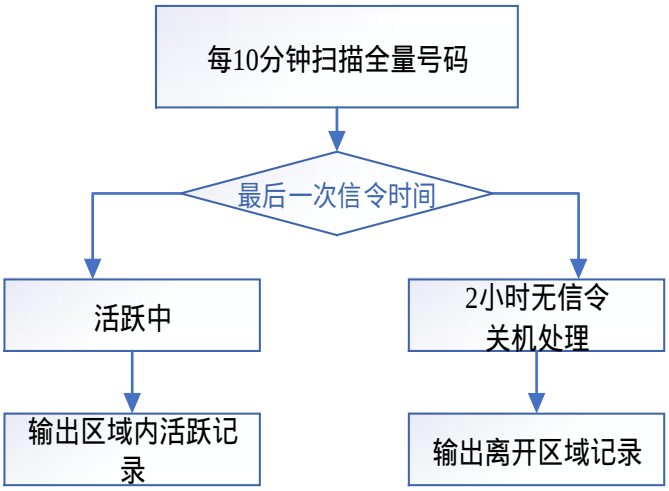
（1）用户关机的判断。用户手机关机后，信令数据不会再产生，需要对用户的开机关状态，区域计时器进行处理。目前的处理方式是当同一用户连续 2 小时没有新的信令数据时，设置为关机状态，用户的区域计时器停止计时。

（2）用户重新开机后，需要判断开机的位置和时间。当开机位置与关机位置处于同一个基站下，且间隔时间小于 6 小时的，表明同一用户位置没有发生变化，用户计时器接续到开机时间上。

（3）原地开机，但间隔时间超过 6 小时的，用户计时器切分为两个时间段，新驻留时间以开机时间重新计时。此逻辑的目的是避免在北京上班，在河北居住的人员通过同一出入京的道路往返北京河北之间，关机地是北京最后一个基站，

而第二天上班时返回北京，入网的也是同一个基站，这种情况需要将在河北的时间切出，不纳入统计。

关机与驻留区内未离开的判断如下：



在上述流程中，流处理计算引擎输出以下关键数据

1、用户离开区域数据：

用户 ID、区域 ID、区域进入时间、区域离开时间

2、用户在区域内的活跃数据：

当用户在一直处理区域内时，需要定时输出活跃信息，用户 ID、区域 ID、进入时间、活跃时间

3、用户关机数据：

用户 ID，区域 ID、区域进入时间、最后信令时间

1.4.1.3. 日数据处理

流处理引擎输出的数据为用户计时器的原始数据，需要按照不同区域的业务逻辑进行日汇总。流程如下：

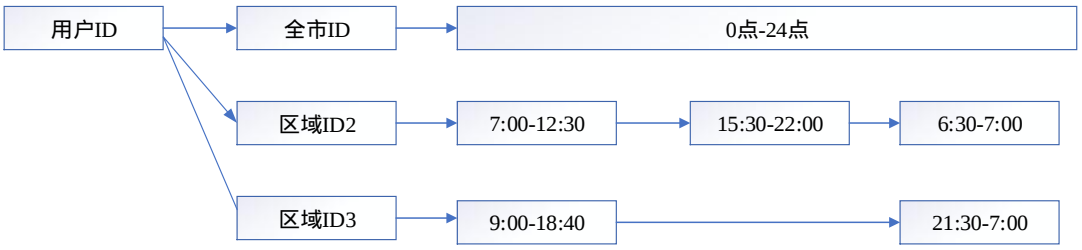
1、对同一用户的多条数据按时间点进行排序，按单个用户的位置变化生成

全天用户活动轨迹。

2、根据早 7 点-19 点 21 点-早 7 点对工作、居住时间进行秒数累计

3、分级输出区、街乡、重点区域用户级数据文件

处理逻辑如下图：



对每个用户当天所驻留过的区域 ID 按时间片段进行汇总，形成用户-区域关系图，并形成区域与时间片段的链表。

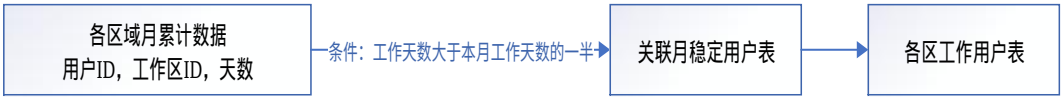
在时间片段链表上对全北京市做 0-24 点计时，其它地区按工作、居住时间按早 7 点-晚 19 点，21 点-第 2 天 7 点两段进行分段计时。

因重点区域之间可能出现重叠，同一用户可同时在多个重叠的重点区域内，其它地区数据不重叠。

1.4.1.4. 月数据处理

月处理是对全月每天累计出的数据按照业务规则生成全月用户数，要求如下：

(1)生成工作用户：



(2)生成居住用户：

1.4.2. 数据计算模型试算构建

1.4.2.1. 业务模型构建能力介绍

我司具备业务模型构建能力,能够构建一套全面覆盖不同业务场景的算法模型，模型具备高度的准确性、适应性和可扩展性；并持续优化和更新模型库，以应对业务发展和数据变化，并提供模型评估、验证、维护和支持服务，确保模型在实际应用中的稳定性和可靠性。

模型库构建：需要建立完善的业务算法模型库，包括各种适用于不同业务场景的算法模型，如人口流动性分析模型、瞬时人口统计、日度人口统计、月度人口统计模型等。

模型优化与更新：随着业务的发展和数据的积累，需要定期对模型进行优化和更新，提高模型的准确性和适应性。

模型集成与应用：能够将不同的模型进行集成，形成一套完整的解决方案，便于用户直接使用。提供模型的应用接口文档，方便用户将模型集成到自己的业务系统中。

模型评估与验证：需要对模型进行严格的评估和验证，确保模型的准确性和可靠性。包括使用真实数据进行测试、与其他模型进行对比、进行交叉验证等。

模型维护与支持：需要提供模型的维护和支持服务，包括解答用户疑问、解决模型使用过程中出现的问题、提供模型升级服务等。

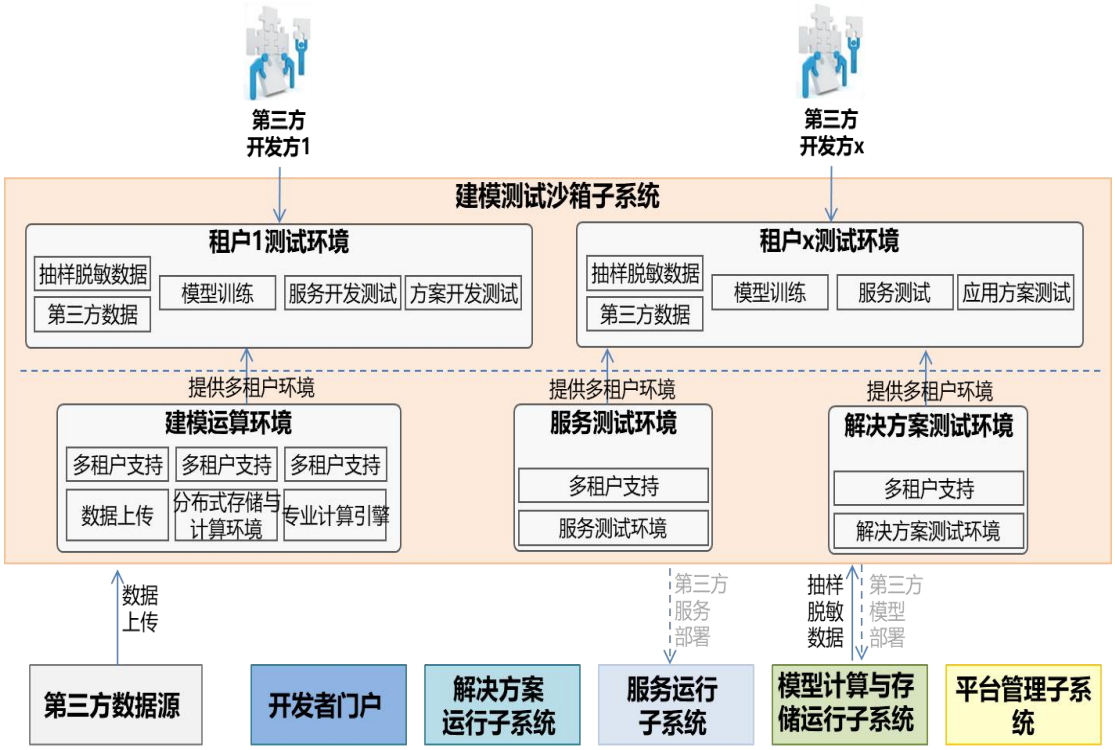
1.4.2.2. 模型服务环境

受限于国家相关个人信息、数据安全保护管理条例、运营商数据安全管理相关规定，运营商手机用户信令等相关数据无法离开运营商的数据环境直接输出。

因此，北京移动在其数据中心信令数据处理计算平台之上配备独立的沙箱资源环境，沙箱环境的硬件资源，存储空间需要满足服务期内所有数据的存储需求，计算资源需要满足实时加工人口统计结果的性能需求。

平台上需部署符合数据建模需求的软件环境和工具程序，开放各种简便易用的专业性分析界面和工具，提供 Spark、Hive 等多种大数据挖掘及建模环境，满足对各类对象数据、中间加工数据、统计结果数据的存储、智能分析及综合治理管控需求。

项目过程中北京移动提供沙箱服务系统配合客户方提供模型试算和环境，包括：模型试算环境、服务测试环境、解决方案测试环境。各环境都基于租户进行数据和资源的管理。其框架及与其他子系统间的关系见下图：

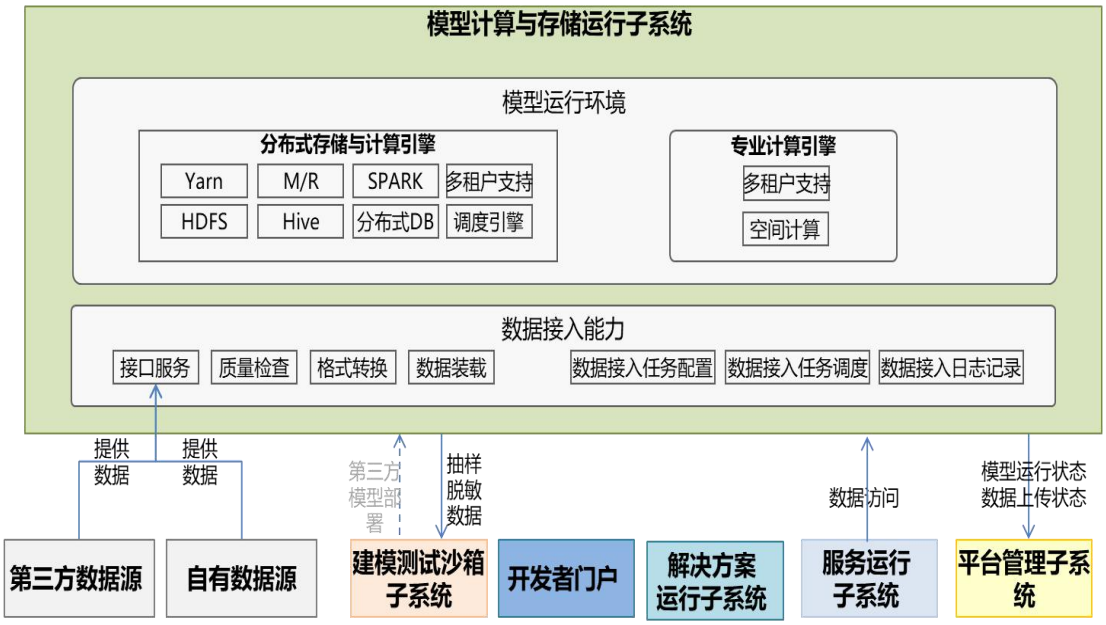


模型运行环境：提供分布式存储与计算引擎、专业计算引擎，并通过多租户支持能力隔离各个租户

数据接入和处理能力：提供数据接入通道、数据接入任务的管理与监控。完

成实时和批量数据的采集、数据清洗、数据脱敏、数据关联、数据入库等工作。

其框架及与其它子系统间的关系见下图：



1.4.2.3. 基础模型构建过程

构建有效的业务算法模型是一个系统化的过程，涉及多个阶段和步骤。

（1）业务理解与需求分析

明确目标：与业务团队紧密合作，明确项目的目标和预期成果。

问题定义：将业务问题转化为数据科学问题，确定是分类、回归、聚类还是其他类型的问题。

数据源识别：识别并获取所需的数据源，包括内部数据和外部数据。

（2）模型构建数据准备

数据收集：从不同来源收集数据，包括数据库、文件、API 等。

数据清洗：

1）处理缺失值：通过填充、删除或插值等方式处理缺失数据。

2）去除异常值：检测并处理异常值，以避免对模型的影响。

3）格式统一：确保数据格式的一致性，便于后续处理。

特征工程：

1）特征选择：选择对模型预测最有帮助的特征。

2）特征构造：基于现有特征生成新的特征，以提高模型性能。

3）特征缩放：对数值型特征进行标准化或归一化处理。

（3）模型选择与训练

选择合适的算法：根据问题类型和数据特性选择合适的机器学习或深度学习算法。

划分数据集：将数据划分为训练集、验证集和测试集，通常采用 70%（训练）、15%（验证）和 15%（测试）的比例。

模型训练：

1）初始训练：使用训练集对模型进行初步训练。

2）参数调优：通过交叉验证、网格搜索等方法调整超参数，优化模型性能。

3）验证与评估：在验证集上评估模型性能，确保模型不过拟合或欠拟合。

（4）模型评估与优化

性能评估：

1）选择合适的评估指标：根据问题类型选择合适的评估指标，如准确率、召回率、F1 分数、AUC-ROC 等。

2）多模型比较：如果使用了多种模型，对比它们的性能，选择最优模型。

误差分析：

1）分析错误案例：深入分析模型在验证集上的错误案例，找出改进方向。

2）特征重要性：分析特征的重要性，了解哪些特征对模型影响最大。

模型优化：

1）集成方法：考虑使用集成学习方法（如 Bagging、Boosting）进一步提升模型性能。

2）模型融合：结合多个模型的预测结果，以提高整体性能

（5）模型部署与监控

模型部署：

1）环境搭建：在生产环境中搭建运行模型所需的计算资源和软件环境。

2）API 封装：将模型封装成 API，方便与其他系统集成。

实时监控：

1）性能监控：定期监控模型的性能，确保其在生产环境中稳定运行。

2）异常检测：设置报警机制，当模型性能下降时及时通知相关人员。

持续迭代：

1）反馈循环：建立反馈机制，收集用户反馈和实际应用中的问题，用于模型的持续优化。

2）定期更新：定期重新训练模型，引入新数据，保持模型的时效性和准确性。

（6）模型评估与验证

对模型进行严格的评估和验证，确保模型的准确性和可靠性。包括使用真实数据进行测试、与其他模型进行对比、进行交叉验证等。

（7）模型维护与支持

提供模型的维护和支持服务，包括解答用户疑问、解决模型使用过程中出现的问题、提供模型升级服务等。

1.4.2.4. 数据模型建设

1.4.2.4.1. 不规则行政区域边界与基站对应算法

在实际的行政区域划分中，存在多种边界情况

➢ 飞地

如北京朝阳区，共43个街道组成，其中42个集中分布，仅首都机场1个街道内嵌到顺义区，与其它地区不接壤。

➢ 空洞地区

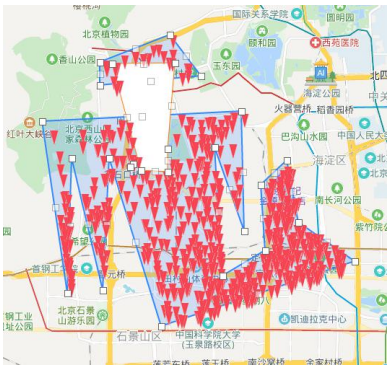
与北京朝阳区相邻的顺义区中对应的首都机场街道为空洞地区。

➢ 交错地区

大量行政区域边界为右图形态，成犬牙交错状态。

针对行政区域边界的实际情况，本方案建立了完善的基站与区域关算法。

行政区域边界采用多种多边形矢量描述，矢量边界分为外边界多边形和内边界多边形。



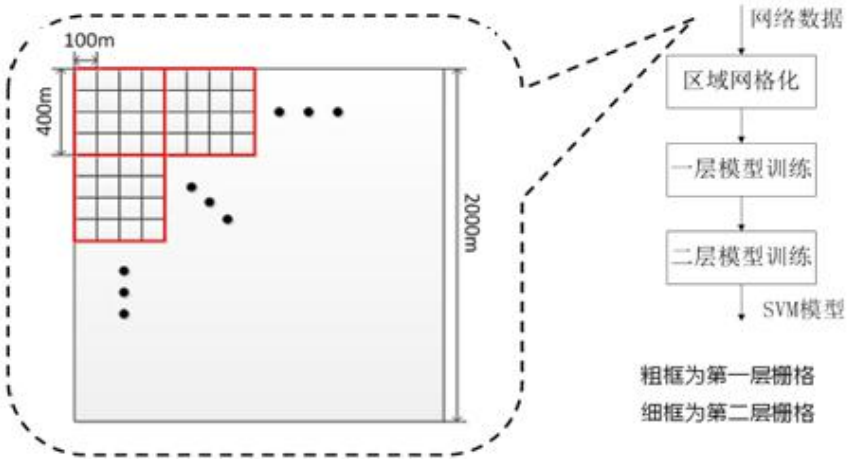
多个外边界描述该区域的轮廓，多个内边界描述区内的空洞地区

通过基站经纬度与朝向位置点与多个多形穿越边界射线算法（ray casting）计算该基站的归属关系。

右图为实际测试结果，蓝色边界为两个飞地边界，黄色为空洞地区，红色三角形为基站位置点。测试结果表明基站与不规则区域对应关系完全正确。通过算法，每日生成Channel_Cell表，用于后续运算。

1.4.2.4.2. 协同定位算法

移动通信网络中移动台的定位方法是近年来的新兴技术，随着移动通信用户数量的增加和用户需求的不断扩大，快速、精准、有效实用的定位方法已成为通信领域急需解决的关键技术之一。移动通信蜂窝网络中的移动终端的定位技术有着广阔的应用前景。移动通信网络通过扇区向移动台收发信息以及移动台在各个扇区间进行切换占用不同的扇区来实现移动通信，如何有效、准确的利用移动通信网络的相关数据进行可操作性强、精确度高的定位是移动通信网络定位技术的关键。一般场景下协同定位精度可控制在 60-80 米内。在经过基于主服务小区基站经纬度的初步定位选定出一个边长为 2 千米的正方形区域后，要先将该区域进行两级栅格划分，如下图所示：



之所以进行两级栅格划分是为了进行两级支持向量机定位，第一级支持向量机将定位范围从 2 千米缩小为 4 百米，第二级支持向量机将定位输出待定位数

据所属 100 米栅格并以 100 米栅格中心的经纬度作为定位结果输出。进行两级支持向量机定位相对于一级支持向量机定位计算复杂度要小，支持向量机分类阶段计算分为两部分，一部分为待分类样本点与所有类别支持向量的内积计算，一部分是决策函数的计算，假设支持向量机训练得到的分类机中每一类对应的支持向量数量都为 n（当训练数据在欧式空间均匀分布时会出现这种理想情形），且使用成对分类法处理多分类问题，则使用上述两级支持向量分类时第一级支持向量机求解的是一个 25 类分类问题，第二级支持向量机求解的是一个 16 类分类问题，在分类阶段总共要进行 $(25+16)\lceil n=41\rceil n$ 次待分类样本与支持向量的内积运算，$25\leftarrow(25-1)/2+16\leftarrow(16-1)/2=420$ 次决策函数计算。而使用一级支持向量机分类时求解的是一个 400 类分类问题，在分类阶段总共要进行 $400\lceil n$ 次待分类样本与支持向量的内积运算，$400\leftarrow(400-1)/2=79800$ 次决策函数计算，显然使用二级支持向量机进行定位相对于一级支持向量机将大幅降低分类阶段的计算复杂度。虽然使用二级支持向量机会使得训练阶段需要训练的分类机数量增多，并使分类模型的存储空间增大，但由于训练阶段是离线进行的，对时间复杂度的要求不高，而定位阶段也就是用支持向量机训练得到的分类模型对未知样本进行分类阶段的计算量直接影响定位方法的实时应用能力，因此用训练阶段计算和存储代价的增大换取分类阶段计算量的下降是可行的。

在实际应用中对第二级支持向量机定位时正方形栅格大小的选择决定了最终的定位精度，栅格划分越小定位精度越高，但由于分类问题总的类别数量会增大所以定位阶段的计算复杂度会相应提高，且随着划分栅格边长的减小，定位精度的增长会越来越小，所以具体第二级支持向量机定位是正方形栅格大小选择应综合实际应用对定位精度和速度的要求，第一级支持向量机定位正方形栅格大小

的选择则应在决定第二级栅格大小后依照使第一级分类问题和第二级分类问题的类别总数最小化的原则进行选择，以使定位阶段总的计算量最小。

在对进行支持向量机二次定位的 2 千米区域划分好二级栅格后，要分别对每个栅格进行编号作为分类问题的类别标识，为了之后方便对训练数据进行类别标注，可以采用经度和纬度方向优先，经度和纬度方向均从小到大的顺序进行编号，这样的编号可以使得在对训练数据进行类别标注时，只要将训练经纬度代入一个简单表达式中即可直接得出类别标识而不用建立栅格经纬度与编号的映射关系表进行查表操作。

（1）定位模型训练

基于支持向量机的定位方法训练阶段是通过将以移动终端对周围基站接收小区为特征，以划分好的栅格编号为类别标识的训练样本集代入到如下式所示最优化问题中，由于一般选择用成对分类法处理将移定位问题转化成的多分类问题，所以如果基于支持向量机定位的区域总共划分为 n 个栅格，则训练阶段要求解出 $n(n-1)/2$ 个决策函数。使用 LIBSVM 提供的支持向量机工具库实现基于支持向量机的二次定位训练阶段及定位阶段算法。

首先要将全部区域的训练数据集划分成 100 个子集，对于位于相邻 2 千米正方形区域的重叠区域中的训练数据要同时划分到两个子集中，这 100 个子集为进行第一级支持向量机定位的训练集，然后对于每个 2 千米正方形区域对应的训练数据子集，要进一步按 4 百米栅格划分成 25 个子集，总计 2500 个子集，这 2500 个子集为进行第二级支持向量机定位的训练集。

完成训练数据集的划分后要对处理所需格式的训练数据进行类别标注，将不同训练集中的每条数据的 GPS 经纬度代入到下面的表达式中：

$$d = 2\pi \times R_{earth} / 360$$

$$relativeX = d \times (lon - ref\_lon) \times \cos(lat + ref\_lat)$$

$$relativeY = d \times (lat - ref\_lat)$$

$$index = \left\lfloor relativeY / L_{grid} \right\rfloor \times (L_{area} / L_{grid}) + \left\lfloor relativeX / L_{grid} \right\rfloor$$

其中 d 为一度经度或纬度对应的实际物理距离，R_{earth} 为地球半径，$relativeX$ 与 $relativeY$ 为训练数据与正方形定位区域边界在经度和维度方向上的物理距离，lon 与 lat 为训练数据 GPS 经纬度，ref_lon 与 ref_lat 为正方形定位区域内的最小经度和纬度，L_{area} 与 L_{grid} 为正方形定位区域与划分栅格的边长，$index$ 为训练数据所在栅格编号。

在将训练集输入到 LIBSVM 的训练工具进行训练前，需要对支持向量机的一些参数进行设定，需要设定的参数包括选择的支持向量机类型，惩罚因子，核函数类型以及核函数中的参数值，其他参数可直接使用 LIBSVM 中的默认值，由于将定位问题转化为多分类问题，所以支持向量机类型应设定为求解多分类问题的支持向量分类机，在以往的移动终端定位相关工作中基于支持向量机的定位方法一般选择高斯核函数或多项式核函数，由于计算量更小，所以在本文中选择使用一阶多项式核函数，并使用网格寻优法对惩罚因子及核函数的参数进行配置，即先以指数步长测试当每个参数取集合 $\{-2^{10}, -2^{9}, ..., -2^{0}, 2^{0}, ..., 2^{9}, 2^{10}\}$ 中任意值构成的每种组合时对某训练集的交叉验证性能，然后取出性能最好的相邻两组参数组合，将这两组参数组合以步长 0.01 继续划分为新的取值集合，然后再次验证每个参数取这一取值集合中任意值时得到的每种组合对原训练集的交叉验证性能，选出性能最好的参数组合为最终的参数配置。

为 LIBSVM 配置好以上参数后，只要将划分好的 100 个 2 千米正方形区域

训练数据集和2500个4百米正方形区域训练数据集输入到LIBSVM的训练模块中进行训练，即可得到这些2千米正方形区域和4百米正方形区域的支持向量机定位模型，这些模型中存储了每个定位区域中的所有$n(n-1)/2$个决策函数(n为定位区域中划分的栅格数)，并以结构体的形式输出到内存中，将这些定位模型以二进制文件形式存储到硬盘中以供定位阶段调用。

由于在基于支持向量机的二次定位时选用的核函数为一阶多项式核函数，所以可以引入一种快速算法，该算法通过将支持向量机对未知样本分类时要进行的未知样本特征向量与支持向量的内积运算转移到训练阶段进行，从而大大降低分类阶段的计算复杂度。使用一阶多项式核时支持向量机的决策函数形式如下：

$$f(x)=\mathrm{sgn}[\sum_{i=1}^{l}\alpha_i y_i(\gamma \mathbf{x_i}\mathbf{x}+r)+b]$$

其中α_i，r，γ和b为参数，y_i为支持向量的类别标识，$\mathbf{x_i}$为支持向量，l为支持向量总个数，由此可见在没有引入快速算法的一阶多项式核支持向量机算法在分类阶段每次决策需进行l次内积运算，通过将上式中的小括号展开，可以转化为：

$$f(x)=\mathrm{sgn}[\sum_{i=1}^{l}(\alpha_i y_i\gamma \mathbf{x_i}\mathbf{x}+\alpha_i y_i r)+b]$$

上式中的参数α_i，r，γ，x_i和b都是在参数配置时确定或在训练阶段求解的，设：

$$\sum_{i=1}^{l}\alpha_i y_i\gamma \mathbf{x_i}=\boldsymbol{\beta}$$

$$\sum_{i=1}^{l}\alpha_i y_i r+b=b'$$

则转化为：

$$f(x)=\mathrm{sgn}[\boldsymbol{\beta}\mathbf{x}+b]$$

上式中$\boldsymbol{\beta}$和b都是可以在训练阶段计算得到的，在训练阶段计算出一阶多项式核支持向量机每个决策函数中的$\boldsymbol{\beta}$和b并将其保存在文件中，在分类阶段计算决策函数值时直接调用并代入到公式中，则每次决策只需进行 1 次内积计算即可。

在调用 LIBSVM 训练模块得到所有 2 千米正方形区域和 4 百米正方形区域的定位模型后，对模型中每个决策函数如式的参数进行了计算，并与 LIBSVM 输出的定位模型一起以二进制文件形式存储在硬盘中以供定位阶段调用。

（2）待分类样本定位

基于支持向量机的定位方法定位阶段是通过将由待定位移动终端对周围基站的接收信号组成的特征矢量代入到训练阶段得到的一组决策函数中，用成对分类法求解出移动终端的栅格，然后以栅格中心经纬度作为定位结果输出。

首先对待定位数据集进行了一种合并处理，按表中所示时间戳字段以一定时间间隔将待定位数据集中的数据进行合并，具体合并方法是先统计参与合并的待定位数据所涉及到的所有基站编号，生成一个基站小区编号集合，然后对基站编号集合中的每个元素，将参与合并的待定位数据对应该基站编号的信息相加然后除以参与合并的待定位数据总数，得到以基站编号集合中的元素为矢量维度标识，参与合并的待定位数据对应该基站编号接收信息均值为矢量相应分量值的新特征矢量，定位将对该合并后的待定位数据进行，并认为参与合并的所有待定位数据的定位结果都是对合并后待定位数据的定位结果。对于基站编号集合中的某些元素，可能不是所有参与合并的定位数据都有对应该基站编号的信息，此时可以认为移动终端在路测到这条定位数据时接收到该基站的信号信息比较弱。

之所以可以对待定位数据进行合并处理是因为路测时间靠近的路测数据属于同一通话，且路测的地理位置也很相近，所以接收信号有很强的相关性，将它们合并能起到去噪以及增大特征矢量维度的作用，这将有助于提高最终定位精度，且由于待定位数据合并后将原来需要进行的多条待定位数据的定位减少为对合并后的一条待定位数据的定位，所以也将大幅减少定位阶段计算量，同时只要合并的时间间隔不要太大，使在此时间内路测距离控制在几十米的范围内，就不会因为用对合并后的路测数据定位结果代替所有参与合并的路测数据定位结果而引入太大的定位误差。

在实际应用中这种对待定位数据的合并也是可行的，只要将网络端接收到的移动终端测量报告按通话进行划分，在同一通话内进行合并，就可保证合并的移动终端测量报告的强相关性，达到上述提高定位精度和降低定位计算量的目的。

在进行完待定位数据的合并后，将合并的待定位数据以及训练阶段得到的定位模型一起输入到 LIBSVM 的分类模块进行分类判决，即可输出合并待定位数据的分类结果，即所处栅格的编号，然后将栅格编号代入以下表达式：

$$d = 2\pi \rceil R_{earth} / 360$$

$$lon = ref\_lon + [label \bmod (L_{area} / L_{grid}) + 0.5] \rceil L_{grid} / d$$

$$lat = ref\_lat + \mp M label / (L_{area} / L_{grid}) + 0.5'' \geq \rceil L_{grid} / d$$

其中 lon 与 lat 为二级支持向量机定位的最终结果。

在基于支持向量机的定位阶段同 LIBSVM 分类模块的源代码进行了修改，引入了一种优化的多分类问题求解方法，称为有向无环图(DAG)的多分类法。相对于使用普遍的成对分类法，可以是支持向量机在分类阶段进行的判决函数计算次数大幅减少，从而降低定位阶段计算复杂度。

成对分类法在分类阶段需要进行 $n(n-1)/2$ 次判决函数的计算(n 为分类问题的类别总数)，而 DAG 多分类法基于这样的假定：在分类阶段某个类只要在一次判决中被判决为不是待分类样本所属类，则这个类就不可能是待分类样本的最终分类结果，而不再进行这个类与其他类的判决函数计算，这样每个类就只用参与一次判决函数计算，在分类阶段总共进行 $n-1$ 次判决函数计算即可得到分类结果。由于分类误差的存在，DAG 多分类法的假定并不一定总是成立，所以相对于成对分类法一般会造成分类精度的下降，但根据仿真结果，引入 DAG 多分类法后对基于支持向量机的定位精度影响很小，几乎可以忽略不计，所以是一种可行的优化方法。

（3）基于 k-近邻法的三次定位

基于 k-近邻法的定位方法是通过将待定位移动终端以基站编号为维度标识，以接收小区信息的特征矢量与定位区域内历史路测的特征矢量进行某种相似度或距离度量的比较，选出与待定位移动终端特征矢量最相似或距离最小的路测特征矢量，可以认为它们在地理位置上应该也是最接近的，所以该路测特征矢量的经纬度即为对待定位移动终端的定位结果。之所以在基于支持向量机的二次定位后要引入基于 k-近邻法的三次定位，是因为根据的仿真结果，在移动网络户外移动终端定位上，基于 k-近邻法的定位方法定位精度比基于支持向量机的定位方法要高很多，但是在定位区域面积相同的情况下基于 k-近邻法的定位方法计算复杂度也比基于支持向量机的定位方法要大得多，所以为了利用 k-近邻法提升定位精度的同时尽可能减小定位方法整体的计算复杂度，采取了在基于支持向量机的二次定位后将进行 k-近邻法的定位区域相缩小到一个很小的范围，然后再进行基于 k-近邻法的三次定位的方法，使整体的定位方法同时具有了很高的

定位精度和定位速度。下面将详细介绍基于 k-近邻法的三次定位的步骤。

在进行基于支持向量机的二次定位后会输出一个经纬度，以该经纬度为中心选定一个边长为数百米的正方形区域，这个正方形区域即为进行基于 k-近邻法定位的区域。由于基于支持向量机的定位方法平均定位误差在 100 米以内，所以待定位数据的真实经纬度有很大的概率位于以基于支持向量机的定位方法得出的定位结果为中心的一个数百米正方形区域，当然正方形区域越大，待定位数据真实经纬度位于该区域的概率也会越大，可能有助于提升进行基于 k-近邻法的三次定位的定位精度，但区域越大，进行 k-近邻法的计算复杂度就越大，并且也可能为 k-近邻法的定位引入噪声(例如区域中距离待定位数据距离较远的那些训练数据)，所以在实际应用中应根据对不同基于 k-近邻法的三次定位定位区域大小选取的仿真结果，并结合对定位方法定位精度与速度的要求选择更适当的定位区域大小。根据仿真结果，在城市环境下选择边长为 100 米的正方形区域进行基于 k-近邻法的三次定位可以达到较好的定位精度和速度。

移动通信网络中移动台的定位方法是近年来的新兴技术，随着移动通信用户数量的增加和用户需求的不断扩大，快速、精准、有效实用的定位方法已成为通信领域急需解决的关键技术之一。移动通信蜂窝网络中的移动终端的定位技术有着广阔的应用前景。移动通信网络通过扇区向移动台收发信息以及移动台在各个扇区间进行切换占用不同的扇区来实现移动通信，如何有效、准确的利用移动通信网络的相关数据进行可操作性强、精确度高的定位是移动通信网络定位技术的关键。

1.4.2.4.3. 边界乒乓统计磁吸算法

随着移动通信大数据人口统计区域的不断细化下沉至街乡镇或更小的区域，

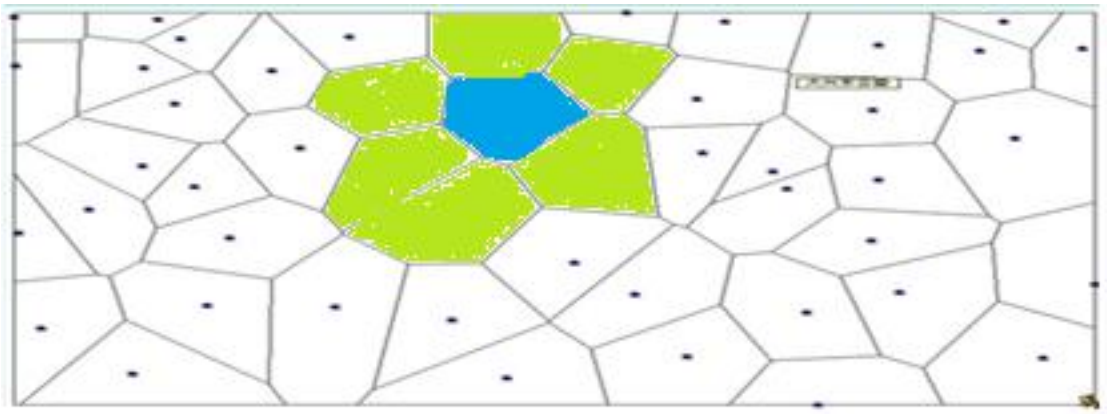
由于无线网络覆盖特性在统计区域边界容易出现人口乒乓统计的现象，统计区域边缘人口乒乓统计导致区域间人口流入流出指标异常，本方法旨在通过无线网络覆盖特点制定统计区域边缘人口乒乓统计抑制算法，以完成区域间流入流出指标的校验。

（1）乒乓缓冲区域划分

步骤一：根据泰森多边形做基站模拟覆盖模型，统计边缘区域基站一层邻接扇区邻区情况；

步骤二：以某个月份统计为人口统计技术，在后续月份统计中，对存在乒乓统计的人口中存在边缘区域基站一层邻区缓冲区内的人口进行补偿统计。示意图如下：

蓝色为统计扇区，绿色为一层统计邻区。

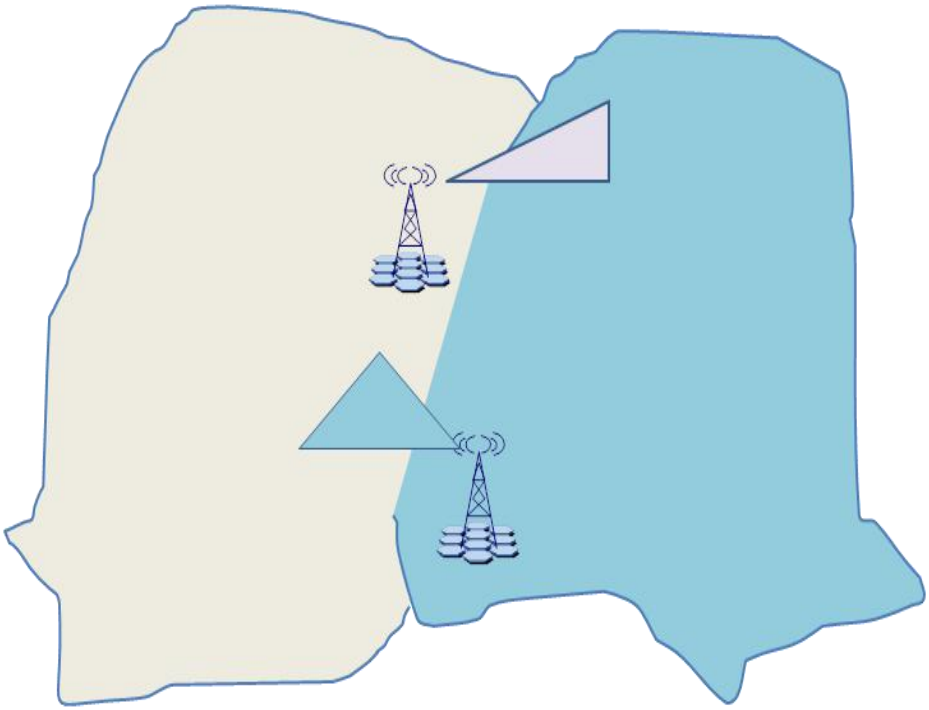


（2）磁吸算法

在乒乓缓冲区域基础之上，对存在乒乓统计的人口中存在边缘区域基站一层邻区缓冲区内的人口进行补偿统计。以某个月份统计为人口统计基础，在后续月份统计中，通过磁吸算法来抑制统计区域内相邻区域流入/流出人口。

方法：对居住于某区域的人口进行锁定，对锁定人群的居住地分布进行分析（即对流出人群分析），在某区域的居住人口，部分基站位于行政街道之间存在

人口乒乓统计基站扇区方位角变化所引起的居住地变化，流出区域多为周边流动：划定边界区域，9 月居住于某区域，10 月居住于某区域的缓冲区，则居住地不发生变化。在下个统计周期的居住地如下图：

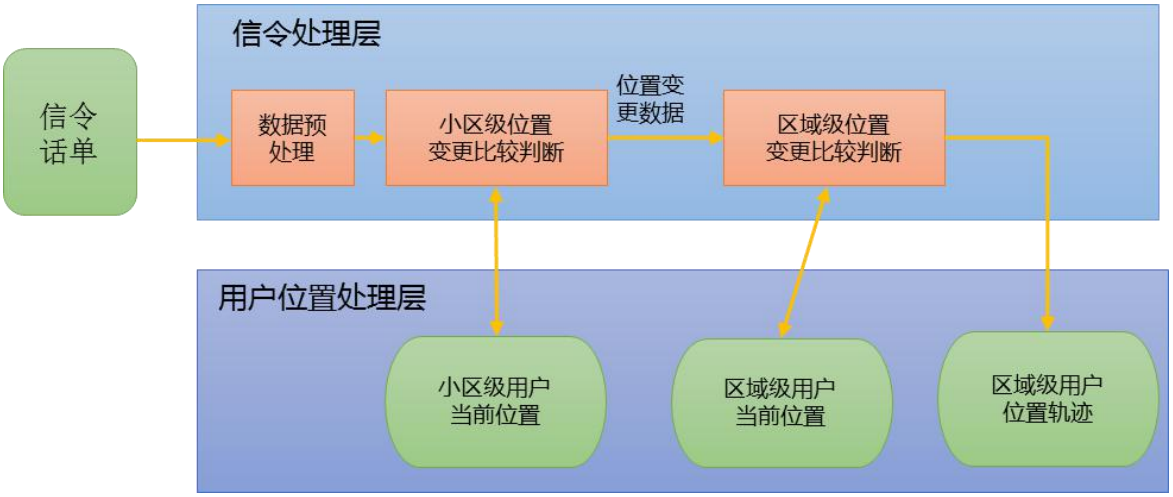


1.4.2.4.4. 位置识别模型

模型背景

核心模型，通过实时采集分析信令产生事件、小区基站，分析用户实时位置，并对位置变化实时记录和输出。

模型设计



图、位置识别模型设计

1、根据预处理之后的信令数据，得到信令中事件对于的基站信息；

2、按事件时序排序信令基站信息关联基站经纬度；

3、得出用户不同时刻的位置归属基站及基站经纬度；

3、实时输出用户更新后的位置信息到 MPP 数据库，形成用户位置轨迹。

分析周期：实时分析

实施方案

步骤 1：对信令数据进行预处理：增加唯一编码的内部 ID（UUID），用这个内部 ID 来作为来访总用户数的统计依据；针对本地用户，UUID 也生成，对应的是用户的 IMSI。

步骤 2：根据预处理之后的信令数据，计算小区级用户位置与历史轨迹。

步骤 3：根据小区级用户位置与历史轨迹，关联基站经纬度，得出用户经纬度。

步骤 4：根据小区和基站映射区域（如 XX 居民区、商圈、景区、广告牌位置等）。

输入输出

模型输入：

接口单元名称	信令信息				
接口单元编码					
接口单元说明	用户产生的信令数据				
数据源系统	信令口转发到 KAFKA				
抽取方式及周期	实时				
属性编码	属性名称	英文名称	属性描述	属性类型	备注
1	时间戳	data_time		timestamp	
2	IMEI	imei		varchar(50)	
3	IMSI	imsi		varchar(50)	
4	LAC	Lac_id		varchar(20)	
5	CI	ci	10 进制	varchar(20)	
6	网络类型	net_type	4G、5G	char(2)	01:4G 02:5G 需提供码表
7	数据类型	data_type	语音的主被叫、短信的发送接收等	char(2)	
8	事件 ID	event_id		varchar(50)	
9	上次 TMSI	LAST_TMSI		varchar(50)	
10	当前 TMSI	CURRENT_TMSI		varchar(50)	
11	上次用户所在位置区	LAST_Lac_id	lac+tmsi 在同一时间点,可以唯一标识一个用户	varchar(50)	

接口单元名称	基站清单同步			
接口单元编码	dim_lacci			
接口单元说明	4G、5G 基站清单			
抽取方式及周期	ETL 批量采集，一天一次			
属性编码	属性名称	英文名称	类型	备注
1	小区码 CI		VARCHAR(50)	

2	位置区 LAC 编码		VARCHAR(50)	
3	所属区县		VARCHAR(20)	
4	基站类型		VARCHAR(300)	
5	生命周期状态		VARCHAR(20)	
7	经度		VARCHAR(50)	
8	纬度		VARCHAR(50)	

接口单元名称	基站与特定区域关系表			
接口单元编码				
接口单元说明	基站与特定区域关系，包括 GIS 平台提供的校园区域+用户自定义区域			
数据源系统	GIS 平台			
抽取方式及周期	ETL 批量采集，一天一次			
属性编码	属性名称	英文名称	类型	备注
1	小区码 CI		VARCHAR(50)	
2	位置区 LAC 编码		VARCHAR(50)	
3	所属区域编码		VARCHAR(20)	
4	生失效标志		VARCHAR(2)	1-生效 0-失效

模型输出：

指标类型	指标名称	指标定义
基本信息	用户标识	用户编码
	用户号码	手机号码
	地市	地市
	区县	区县
	片区	片区
时间信息	时刻	2016101208:3003
	驻留时长	在该位置停留时间
位置信息	LAC	LAC
	CELL	CELL
	经纬度	23.45833,28.45833

1.4.2.4.5. 常驻用户识别

模型背景

区分出区域内的常驻人员、工作人员和途径人员，统计出每类人员的逗留时

长，是实现人群流量常态化监控的重要基础。本模型通过信令数据重点识别出区域内的工作人员。

模型设计

用户每天产生信令基站小区基本可以反应其一天的活动轨迹及行为信息。根据多数用户生活作息规律，设置时段，统计相应时段用户活动小区信息。因为每天的活动信息变化很大，所以需要累计一定天数的用户停留小区信息，然后根据相应规则准确统计出用户工作地信息，然后根据相应规则准确统计出用户工作地信息，根据工作地信息筛选出该区域的工作人员。

- **分析周期：**按月统计。
- **实施方案：**

通过连续分析用户在 A 接口的 TDR/CDR 记录，获取用户在一段日期内的连续生活轨迹信息，结合信令数据的特点，设计用户的工作地分析模型。

第一步：定义用户生活的特定时段范围：

T(D)上午工作时段：08:30—12:00;

T(E)下午工作时段：13:30—18:00。

每天均生成用户各个时段在各个小区的停留时长信息。

第二步：定义用户在各时段内的工作行为规则：

规则 1：上午工作时段计算用户在多个小区的累计停留信息，在该时间段内停留时间超过 90 分钟的小区作为其驻留小区。

规则 2：下午工作时段计算用户在多个小区的累计停留信息，在该时间段内停留时间超过 90 分钟的小区作为其驻留小区。

每天均按照此规则，根据第一步汇总的结果，统计生成每天用户各个时段唯

一符和条件的小区信息。

第三步：每天均按以上规则统计用户在各个时段停留小区信息；

第四步：以月为周期，剔除周末（week）和节假日因素，如当月无其他节假日，按照一个月通常包含 4 个周末日期共 8 天，剔除后，当月周期天数为 22 天。根据以下方法统计用户居住地信息。

1. 统计用户当月工作日（如 22 天）内早上工作时段在各个小区停留天数 D5。

2. 统计用户当月 22 天/当周 5 工作日（如 22 天）内下午工作时段在各个小区停留天数 D6。

3. 统计用户当月工作日（如 22 天）内在各小区以上时段停留天数之和 D7（D7=D5+D6）。

取 D7 为最大值的小区作为用户工作地。保留 D5、D6、D7 指标，这样可以灵活设置参数过滤出符合条件的用户工作地。

输入输出

- 模型输入：

指标类型	指标名称	指标定义
基本信息	用户标识	用户编码
	用户号码	手机号码
	地市	地市
	区县	区县
	片区	片区
时间信息	时段	（08:30-18:00）
	驻留时长	位置停留时间
位置信息	LAC	LAC
	CELL	CELL

- 模型输出：

指标类型	指标名称
基本信息	用户标识

	用户号码
	地市
	区县
	片区
位置信息	工作地 LAC
	工作地 CELL
	居住地 LAC
	居住地 CELL

1.4.2.4.6. 驻留时长分析

模型设计

基于位置识别模型输出的用户位置轨迹数据。轨迹数据中包含每个基站小区的进入时间、离开时间、小区下驻留时长（离开时间-进入时间）；区域驻留时长通过汇总区域映射的基站小区的可衔接的进入时间和离开时间的差值。

实施方案

在系统中输入特定的对象手机号码，系统分析在不同区域的驻留时长，绘制成驻留时长排序标识点。用户输入号码，选择时间段，通过 GIS 展现用户的驻留区域时长信息。查看该人员驻留位置（区域）信息、区域驻留时长。

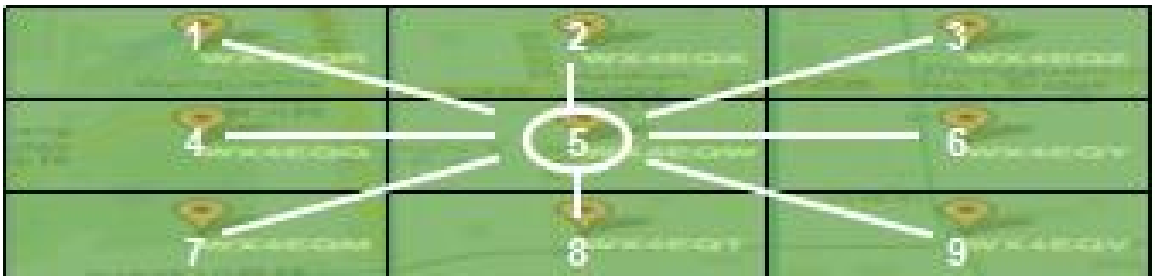
1.4.2.4.7. 一人多卡模型

基于获取到的信息，完成号码关联工作，号码关联时前置条件为剔除三个月内无通话号码与剔除物联网卡。

基于用户脱敏后网格位置信息，按驻留时间长短找出用户月常驻网格 Top 2（30 天内）。

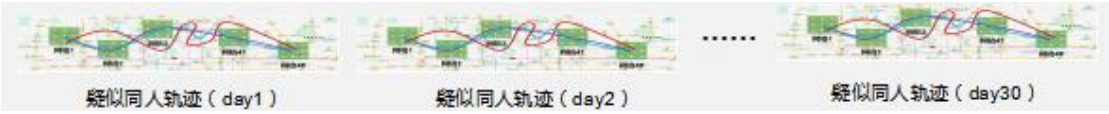
按近邻网格算法，通过对区域内所有用户两两配对，筛选相近用户，缩小用

户比对量。



采用 LCSS 最长公共子序列算法，按照时序顺序，比对轨迹链，找出疑似同人轨迹。

对月度疑似同人进行并集运算，并对日处理数据修订，避免日偶然事件带来的 LCSS 噪声数据，找出最终同人的轨迹链。



完成一人多卡去重分析。产品可按照用户定义的规则自动判断重复数据，并按照用户定义的规则处理重复的数据，同时可按照客户提供的方法进行加权汇总。

1.5. 数据服务内容方案

按照三网融合平台建设项目的数据需求及技术规范，配合客户侧输出提供所需的统计级标准数据，对采集到的人口数据通过模型算法和结果汇总，形成多维度的分析结果，同时根据客户需求适时开展专题人口监测的相关数据处理与计算工作。相关数据应能精准映射到网格，网格大小应不大于 250M*250M，具备提

供 100M*100M 网格的能力。同时，针对客户临时性的个性化专题需求，对重点节假日等个性需求提供人口监测分析支撑，为人口调控提供支撑。

1.5.1. 辖区及定制区域实时活跃用户数量

针对辖区及定制区域实时活跃用户流动情况进行监测分析，围绕北京市朝阳区全境， 43 个行政街乡、81 个重点区域及临时新增区域进行人口流动情况监测分析，包括区域内活跃用户人口规模，人口分布情况、人口来源及流向等情况。

实时活跃用户收集方式：每 30 分钟间隔步进进行统计，统计在每个区域下的人数。

输出形式：csv 格式平台输出。

1.5.2. 月度居住人口数量

分析整理规定区域内所有的居住人口的总体情况，围绕北京市朝阳区全境，43 个行政街乡、81 个重点区域及临时新增区域提供月度居住人口数量，包括居住人口总体规模变化情况、居住人口各街道分布情况、各街道居住人口占比情况等。

月度居住人口收集方式：用户 21 点-7 点（第二天）在某区域停留时长超过 360 分钟的区域为该用户的日居住地，取一个月内用户停留最多的日居住地为该用户的月居住地，统计月居住地区域的居住人数。

输出形式：csv 格式平台输出。

1.5.3. 月度工作人口数量

分析整理规定区域内所有的工作人口的总体情况，围绕北京市朝阳区全境，43 个行政街乡、81 个重点区域及临时新增区域提供月度工作人口数量，包括工作人口总体规模变化情况、工作人口各街道分布情况、各街道工作人口占比情况等。

月度工作人口收集方式：用户 7 点-19 点在某区域停留超过 480 分钟的区域为该用户的日工作地，取一个月内用户停留最多的日工作地为该用户的月工作地，统计月工作地区域的工作人数。

输出形式：csv 格式平台输出、数据分析报告，Word 格式电子版。

1.5.4. 区域及定制区域的 30 分钟流入流出人口数量

对监测区域内各个重点区域的 30 分钟内人口流动情况进行监测分析，围绕北京市朝阳区全境， 43 个行政街乡、81 个重点区域，及临时新增区域进行人口流动情况监测分析。

瞬时人口流入流出收集方式：用当前加工后的数据集与 30 分钟前的数据集取差集；当前数据集包含，30 分钟前的数据集不包含的为流入用户，当前数据集中流入用户的所在地为流入区域，统计流入区域的流入用户人数；当前数据集不包含，30 分钟前的数据集包含的为流出用户，30 分钟前的数据集中流出用户的所在地为流出区域，统计流出区域的流出用户人数。

输出形式：csv 格式平台输出。

1.5.5. 月度职住信息统计

1.5.5.1. 职住 OD

以月度为时间节点，围绕北京市朝阳区全境、43个行政街乡、81个重点区域及临时新增区域，收集监测区域内乡镇街道的职住OD分布情况。

职住OD数据收集方式:用户7点–19点在某区域停留超过480分钟的区域为该用户的日工作地，21点–7点（第二天）在某区域停留时长超过360分钟的区域为该用户的日居住地。取一个月内用户累计停留最多的日工作地居住地为该用户的月工作居住地，统计不同的月工作地居住地之间的职住人数。

输出形式：csv 格式平台输出、数据分析报告，Word 格式电子版。

1.5.5.2. 工作人口的居住地统计

以月度为时间节点，围绕北京市朝阳区全境、43个行政街乡、81个重点区域及临时新增区域，收集监测区域内乡镇街道的工作人口居住地分布情况。

工作人口居住地情况数据收集方式:用户 7 点-19 点在某区域停留超过 480 分钟的区域为该用户的日工作地，21 点-7 点（第二天）在某区域停留时长超过 360 分钟的区域为该用户的日居住地。取一个月内用户累计停留最多的日居住地为该用户的月居住地，统计月居住地的工作人数。

1.5.5.3. 居住人口的工作地统计

以月度为时间节点，围绕北京市朝阳区全境、43 个行政街乡、81 个重点区域及临时新增区域，分析监测区域所有居住人口的工作地情况分析，包括居住人口工作地分布情况。

居住人口工作地数据收集方式：用户 7 点-19 点在某区域停留超过 480 分

钟的区域为该用户的日工作地，21 点-7 点（第二天）在某区域停留时长超过 360 分钟的区域为该用户的日居住地。取一个月内用户累计停留最多的日工作地为该用户的月工作地，统计月工作地的居住人数。

1.5.6. 月度居住人口流向

以月度为时间节点,针对居住人口流向进行监测与分析,生成每日居住用户、每月人员，以及外迁居住人员在朝阳区全境、43 街乡、81 个重点区域的输出情况。

居住人口流向数据收集方式:用当前月居住用户明细的数据集与上个月的月居住用户明细数据集取差集；当月数据集包含，上月的数据集不包含的为流入居住用户，当月数据集中流入居住用户的所在地为流入区域，统计流入区域的流入居住用户人数；当月数据集不包含，上月的数据集包含的为流出居住用户，上月的数据集中流出用户的所在地为流出区域，统计流出区域的流出居住用户人数。

输出形式：csv 格式平台输出、数据分析报告，Word 格式电子版。

1.5.7. 月度工作人口流向

以月度为时间节点，收集监测区域工作的人口流向，分析监测区域各乡镇街道的工作人员新增和减少情况，以及新增人员及减少人员的去向分布情况等。

工作人口流动情况数据收集方式:用当前月工作用户明细的数据集与上个月的月工作用户明细数据集取差集；当月数据集包含，上月的数据集不包含的为流入工作用户，当月数据集中流入工作用户的所在地为流入区域，统计流入区域的流入工作用户人数；当月数据集不包含，上月的数据集包含的为流出工作用户，

上月的数据集中流出用户的所在地为流出区域，统计流出区域的流出工作用户人数。

输出形式：csv 格式平台输出、数据分析报告，Word 格式电子版。

1.5.8. 居住用户中外来人口比重

以月度为时间节点，围绕北京市朝阳区全境， 43 个行政街乡、81 个重点区域及临时新增区域居住人口中非京籍贯居住人口的分布情况。

居住用户中外来人口收集方式：用户 21 点-次日 7 点在某区域停留时长超过 360 分钟的区域为该用户的日居住地，取一个月内用户停留最多的日居住地为该用户的月居住地，统计月居住地区域的不同号段归属地的居住人数。

输出形式：csv 格式平台输出、数据分析报告，Word 格式电子版。

1.5.9. 月度居住用户流入流出

以月度为时间节点，对监测区域所有居住人口进行监测与分析，生成每日居住人员流动情况、每月人员流动情况，以及新增居住人口户籍分布和外迁居住人员流动去向等情况。

居住人口流动情况数据收集方式：用当前月居住用户明细的数据集与上个月的月居住用户明细数据集取差集；当月数据集包含，上月的数据集不包含的为流入居住用户，当月数据集中流入居住用户的所在地为流入区域，统计流入区域的流入居住用户人数；当月数据集不包含，上月的数据集包含的为流出居住用户，上月的数据集中流出用户的所在地为流出区域，统计流出区域的流出居住用户人数。

1.5.10. 月度工作用户流入流出

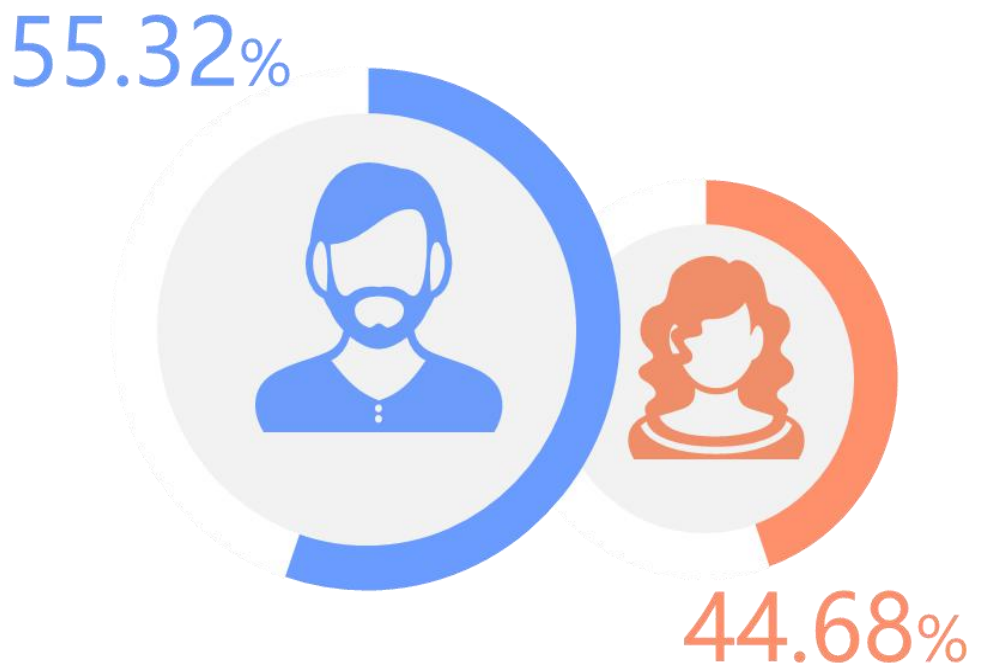
以月度为时间节点，收集监测区域内乡镇街道的工作人口分布情况，监测在监测区域工作的人口的流动情况，实时分析监测区域各乡镇街道的工作人员新增和减少情况，以及新增人员的居住地及减少人员的去向分布情况等。

工作人口流动情况数据收集方式：用当前月工作用户明细的数据集与上个月的月工作用户明细数据集取差集；当月数据集包含，上月的数据集不包含的为流入工作用户，当月数据集中流入工作用户的所在地为流入区域，统计流入区域的流入工作用户人数；当月数据集不包含，上月的数据集包含的为流出工作用户，上月的数据集中流出用户的所在地为流出区域，统计流出区域的流出工作用户人数。

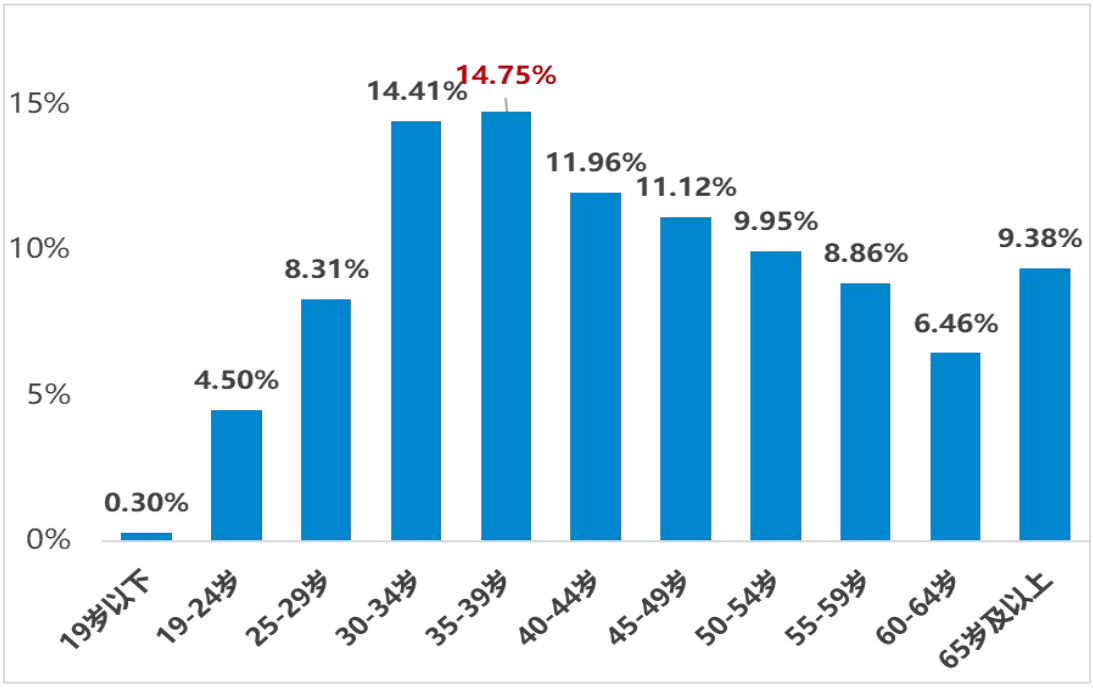
1.5.11. 用户结构画像

按月定期对客户需求监测区域人口结构（例如年龄、性别、归属地、籍贯地等）特征进行监测计算，并提供相关人口结构特征等结果数据。

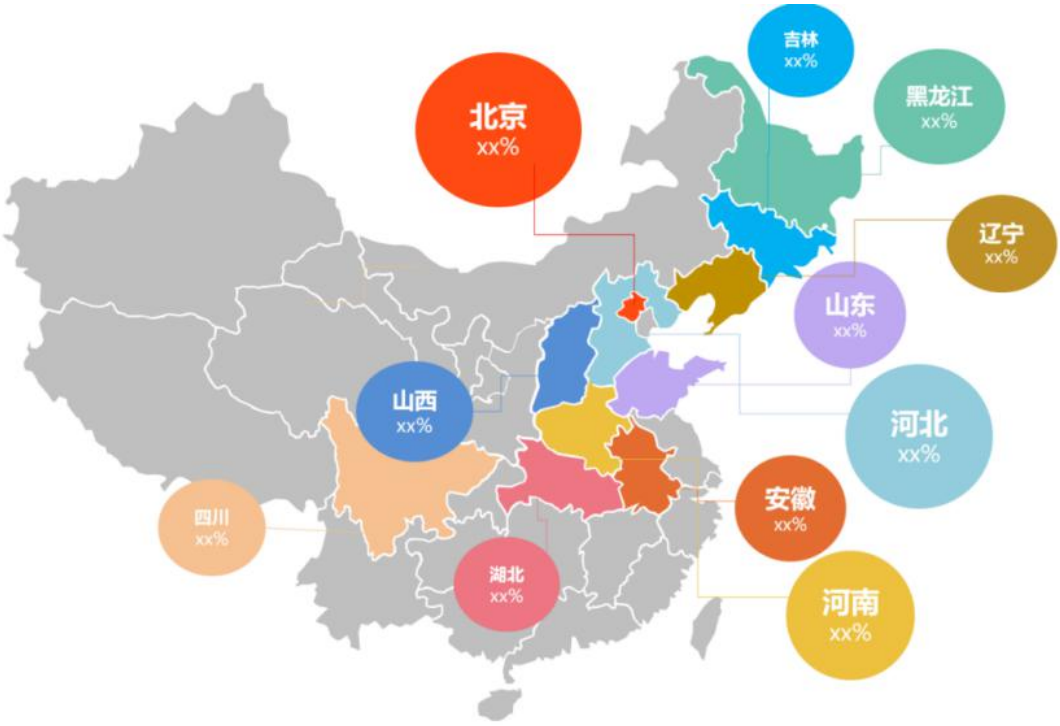
支持针对人口规模、人口流向、职住通勤等附加用户特征数据，对人口性别进行分析。



支持针对人口规模、人口流向、职住通勤等附加用户特征数据，对人口性别进行分析。



支持针对人口规模、人口流向、职住通勤等附加用户特征数据，对人口籍贯地进行分析。



支持针对人口规模、人口流向、职住通勤等附加用户特征数据，对人口号码归属地进行分析。

1.5.12. 商圈实时客流量

根据监测需求，针对 12 个商圈区域实时客流情况进行监测分析。

实时客流收集方式：每 30 分钟间隔步进进行统计，统计在每个商圈区域下的人数。

输出形式：csv 格式。

1.5.13. 商圈客流停留时长

根据监测需求，围绕 12 个商圈到访客流驻留情况进行监测分析。

商圈客流停留时长收集方式：统计每天在商圈不同停留时长的用户人数。时间主要划分为 0-30 分钟、30-120 分钟、120-240 分钟、240-480 分钟以及 480 分钟以上，5 个时间段。

输出形式：csv 格式。

1.5.14. 商圈客流日流量

根据监测需求，针对 12 个商圈每日客流累计到访情况进行监测分析。

商圈客流日流量收集方式：统计每天在各个商圈客流总人数，进行日累计去重。

输出形式：csv 格式。

1.5.15. 景区实时客流量

根据监测需求，针对 24 个旅游景区实时客流情况进行监测分析。

实时客流收集方式：每 30 分钟间隔步进进行统计，统计在每个景区区域下

的人数。

输出形式：csv 格式。

1.5.16. 景区游客停留时长

根据监测需求，针对 24 个旅游景区到访客流驻留情况监测分析。

景区游客停留时长收集方式：统计每天在旅游景区不同停留时长的用户人数。时间主要划分为 0-30 分钟、30-120 分钟、120-240 分钟、240-480 分钟以及 480 分钟以上，5 个时间段。

输出形式：csv 格式。

1.5.17. 景区游客日流量

根据监测需求，针对 24 个旅游景区每日到访客流累计情况进行监测分析。

实时客流收集方式:统计每天在各个旅游景区客流总人数,进行日累计去重。

输出形式：csv 格式。

1.5.18. 100×100 米网格标准化数据服务

我方可为采购人提供 100×100 米网格精度的标准化数据服务，覆盖北京市全部行政区域。该网格精度基于基站定位与信号强度加权算法实现，将信令数据中的 LAC/CI 基站标识映射为 WGS-84 坐标系下的经纬度点，再通过网格化引擎将连续空间离散化为 100 米×100 米的标准网格单元，每个网格单元具有唯一的空间编码。具体服务构建过程如下：

1.5.18.1. 数据接入与预处理

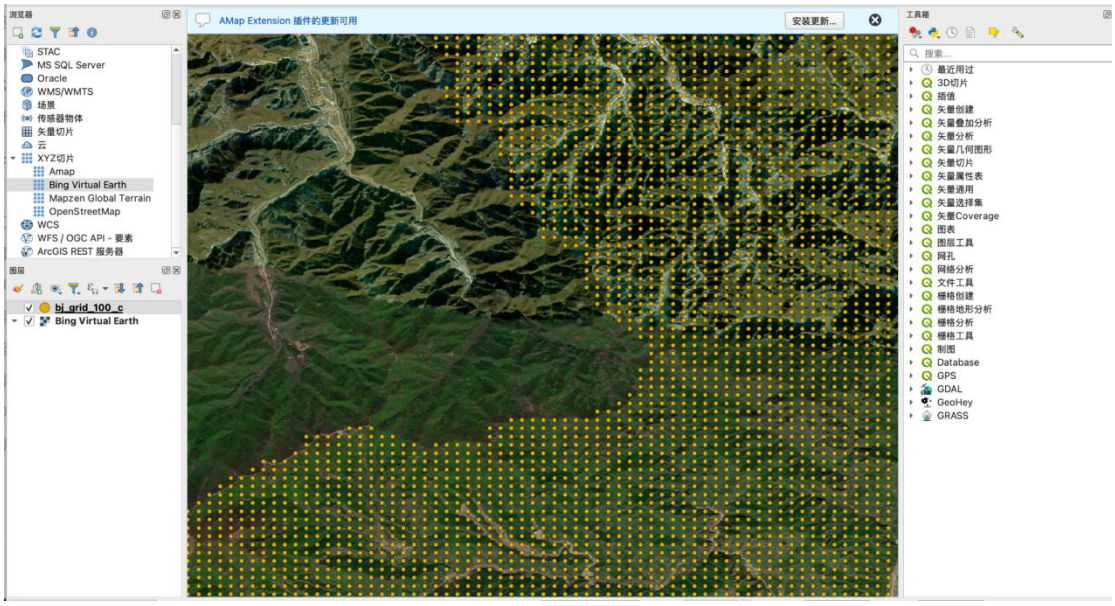
通过接入北京移动全市范围信令数据，对原始数据进行清洗，剔除异常基站标识、过滤乒乓切换噪声、去除重复及测试数据，确保数据真实有效。

1.5.18.2. 基站定位与坐标转换

建立北京市全域基站信息库，存储各基站的 LAC、CI、经纬度（WGS-84 坐标系）及工参属性。将信令记录中的 LAC/CI 标识匹配至基站信息库获取坐标；对于多基站重叠覆盖区域，采用信号强度加权算法计算加权中心坐标，使定位结果更贴近用户实际位置。

1.5.18.3. 网格化引擎构建

将 WGS-84 经纬度坐标转换为高斯-克吕格投影坐标（CGCS2000 坐标系），以北京市行政区域为范围建立规则网格矩阵。网格单元统一为 100 米×100 米，每个网格赋予唯一空间编码（格式：BJ_100M_XXXXX_YYYYY）。将投影坐标按网格尺寸取整映射，实现连续空间到离散网格的转换。



xmin	ymin	xmax	ymax	x	y
116.424447	39.442186	116.425618	39.443088	116.425033	39.442637
116.425618	39.442186	116.42679	39.443088	116.426204	39.442637
116.42679	39.442186	116.427962	39.443088	116.427376	39.442637
116.427962	39.442186	116.429133	39.443088	116.428547	39.442637
116.423275	39.443088	116.424447	39.443989	116.423861	39.443538
116.424447	39.443088	116.425618	39.443989	116.425033	39.443538
116.425618	39.443088	116.42679	39.443989	116.426204	39.443538
116.42679	39.443088	116.427962	39.443989	116.427376	39.443538
116.427962	39.443088	116.429133	39.443989	116.428547	39.443538
116.429133	39.443088	116.430305	39.443989	116.429719	39.443538
116.420932	39.443989	116.422103	39.44489	116.421518	39.44444
116.422103	39.443989	116.423275	39.44489	116.422689	39.44444

1.5.18.4. 属性赋值与数据聚合

为每个网格单元关联行政区划、地理特征及功能类型等空间属性。按网格编码聚合信令记录，生成人口数、人口流动量及人口画像等网格级指标。支持实时、日、月度等多时间粒度聚合，并保留全量历史数据供回溯查询。

1.5.18.5. 数据输出与服务

网格化结果存储于 HIVE 数据仓库，实时数据通过 Kafka 消息队列推送。提供标准 API 接口，支持按时间范围、行政区划等条件查询，支持空间范围检索及 CSV 等格式导出。100×100 米网格精度可满足街区级人口分布分析需求，支撑城市规划、交通流量分析、商业布局优化等常规业务场景的数据使用需要。

1.5.19. 个性化需求服务支撑

根据监测需求，针对重大节假日等特殊时期或突发性人口监测需求，提供个

性化专题监测服务报告，例如重点节假日洞察分析、重点商圈监测分析等。同时丰富用户标签体系提升人群定位的精准性。

1.5.19.1. 用户标签体系

基于统一的人口对象，项目构建多层级、多维度人口标签体系，实现人口属性的标准化表达和精细化管理。

标签体系主要包括基础属性标签、空间属性标签、行为属性标签、流动属性标签及专题属性标签五大类别。其中，基础属性标签包括人口类型、活跃程度、停留规律等；空间属性标签包括居住区域、工作区域、常驻区域、主要活动区域及跨区域活动范围等；行为属性标签包括通勤规律、出行频率、驻留偏好、昼夜活动特征、周末活动特征及节假日活动特征等；流动属性标签包括迁移方向、跨区域流动频率、活动半径、OD 流向特征等；专题属性标签则可结合具体业务需求构建商圈消费人群、景区游客、交通枢纽出行人群、重大活动参与人群等专题标签。

为保证标签体系的持续有效性，平台建立标签生命周期管理机制，根据人口行为变化情况定期更新标签内容，并支持新增标签、组合标签及业务专题标签的灵活扩展，形成可持续演进的人口标签工厂，为不同业务场景提供统一的人口标签服务。

1.5.19.1.1. 时空标签计算处理

时空标签计算是时间数据标签和空间数据标签间通过不同的规则组合形成时空融合数据，为时空数据分析计算提供能力支撑。

(1) 空间数据标签处理逻辑

根据基站信息、用户信令和时间信息进行合并计算，得到时间不同时间下的

用户、基站信息形成时间信令数据标签，如月度标签数、日度标签数据。

(2) 地理数据标签处理逻辑

以运营商数据为例,对地理数据进行标签计算,标签基础计算逻辑进行说明,根据地理位置信息,用户信令,基站信息三者合并进行计算得到用户的定位信息。然后根据用户定位信息，与地理位置信息计算相关的用户属性标签。

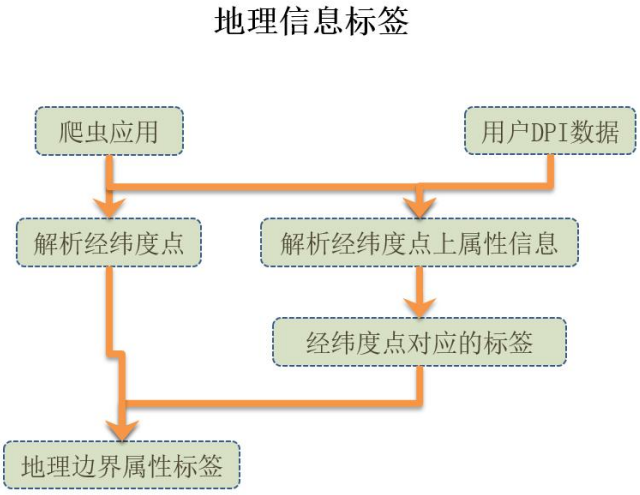


图 14-1 地理数据标签处理逻辑

(3) 时空数据关联逻辑

以运营商时空群体位置为例，对时空数据关联计算逻辑进行方案说明，通过对时序、基站信息、地理边界属性标签进行二次加工分析，形成时空数据关联计算。

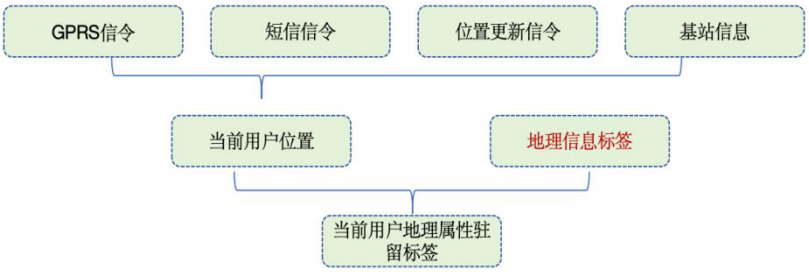


图 14-2 时空数据关联逻辑

通过对不同时间、空间标签关联计算分析，面向业务需求可提供形成不同时

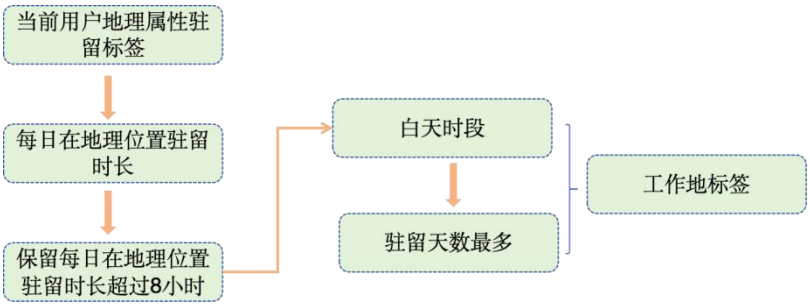
间下不同地理空间用户统计，例如医院日度用户统计、学校月度用户统计。

1.5.19.1.2. 其他业务数据标签处理

面向业务应用场景需求，对业务相关数据标签进行深入挖掘和组合分析，例如面向城市治理业务下的职住用户标签构建，通过对地理信息标签、空间地理信息标签、用户驻留特征标签进行融合分析，构建工作地标签、居住地标签后实现对职住标签的分析构建。

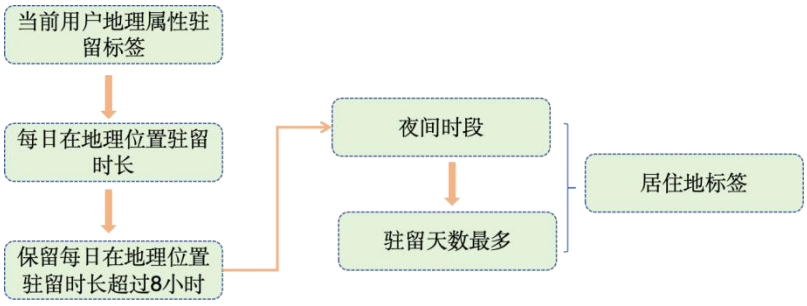
(1) 工作地标签构建

通过对用户地理属性驻留标签和时序标签数据进行逻辑计算，构建工作地属性标签，计算逻辑如下：



(2) 居住地标签构建

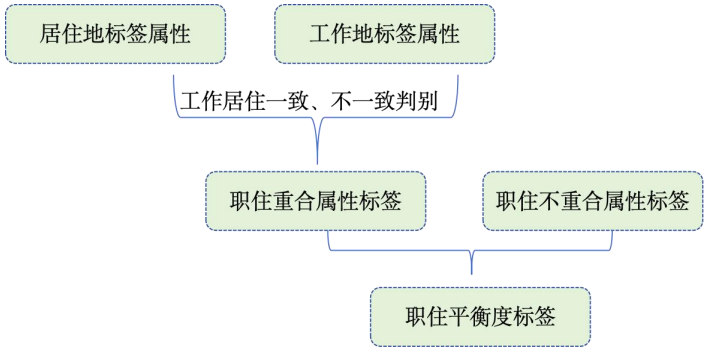
通过对用户地理属性驻留标签和时序标签数据进行逻辑计算，构建居住地属性标签，计算逻辑如下：



(3) 职住业务属性标签组合分析

通过对工作地标签数据和居住地标签数据整合，构建职住平衡属性标签，为

城市交通治理业务提供职住平衡、职住不平衡程度分析支撑。



1.5.19.2. 节假日人口流动监测分析

节假日期间人流流动变化性较大，面向节假日人口流出后朝阳区资源的调配需求，对人口规模的监测至关重要，项目服务期间，可针对客户需求对重点时段（春节、国庆等）用户规模、来源结构、去向分布等情况开展人口大数据监测分析服务。

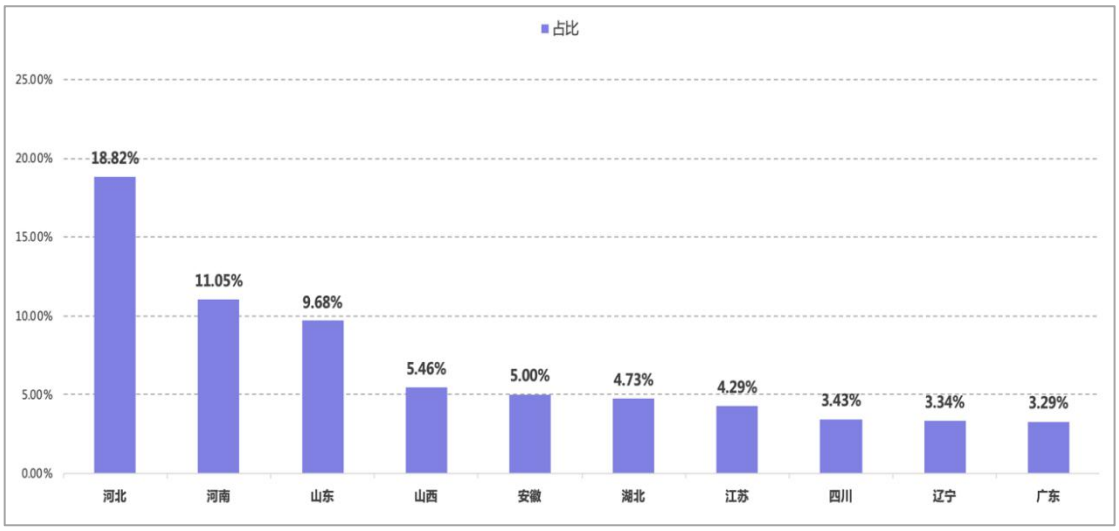
项目服务期间可针对重点节假日的离京用户、在京用户、重点景区客流量等用户规模分析，并提供数据表或分析报告的展示服务。

离京用户：统计满足上月服务用户，在重点节假日期间，监测当日未在北京市出现信令的用户为当日离京用户。

返京用户：统计满足监测区域上月服务用户，在重点节假日后，监测当日在北京市出现信令且满足当日服务用户为返京用户。

针对离京返京用户可依托移动信令数据、实名制认证数据对用户离京去向地和返京来源地进行监测分析。

离京去向地：若用户为外地手机号，则以手机号归属地为离京用户去向地分布，若用户为北京市手机号，则以当日最后一条信令出现的区域作为离京用户去向地。



1.5.19.2.1. 节假日人口在京分析

支持对清明、五一、国庆人口在京状态情况分析，包括在京人口规模、人口结构分析，项目服务期间，可根据客户需求支持对在京用户状态提供数据监测服务。

(1) 重点时段下人口在京规模分析

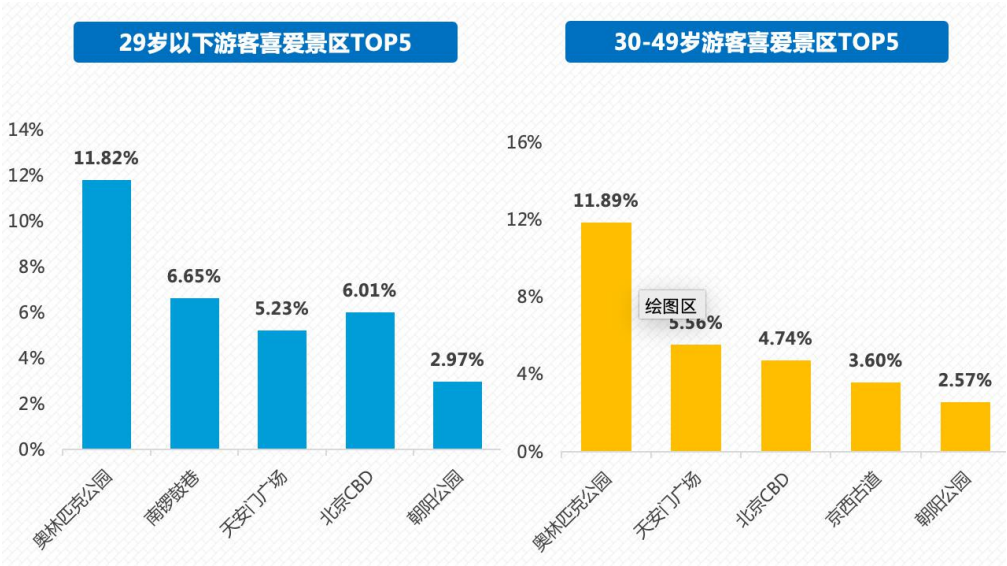
项目服务期间，可针对客户需求提供如清明、五一、国庆等重点时段下人口在京规模分析，并根据客户需求提供数据监测及分析报告服务。

在京用户：满足重点监测时段下在京服务用户则为在京用户。

客流用户：统计满足监测区域驻留超过 30 分钟剔除监测区域工作、居住用户。

(2) 重点时段下人口结构监测分析

项目服务期间，根据客户需求支持提供重点时段下人口结构监测服务，如五一、端午等重点节假日景区客流结构情况，助力客户了解不同景区对不同结构客流人群的吸引力，为区域相关公共资源的配置提供大数据服务支撑。

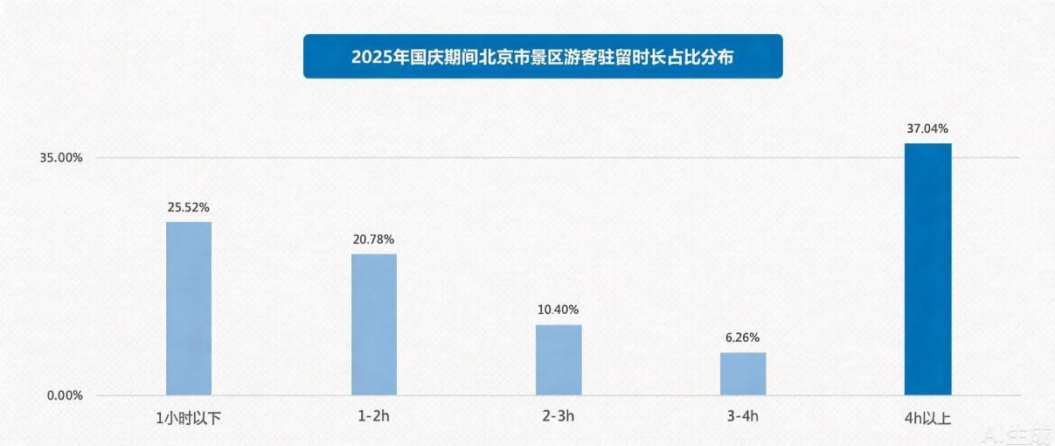


1.5.19.2.2. 节假日人口出行行为分析

项目服务期间，可针对客户需求提供人口出行行为情况分析，针对端午、清明等假期人口到访景区、墓园等区域情况、驻留时长、新进京客流来源地分布等情况进行监测分析，并提供数据表或分析报告的展示服务。

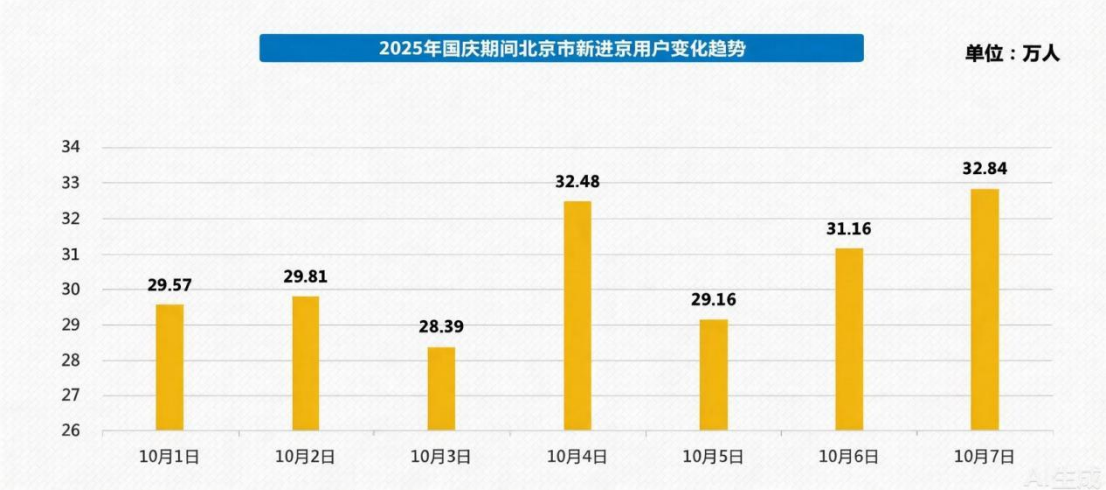
(1) 重点时段下人口驻留时长分析

项目服务期间，针对重点节假日等时段下人口外出行为进行监测，支持提供针对在重点景区或重点区域人口驻留时长监测分析，并提供数据表或分析报告的展示服务。



(2) 重点时段新进京用户规模统计

项目服务期间，针对重点节假日或重点时段期间，根据客户需求可提供新进京用户规模监测分析，并提供数据表或分析报告的展示服务。



1.6. 项目实施保障方案

1.6.1. 数据质量保障方案

1.6.1.1. 数据保障服务响应机制

北京移动提供专职保障服务团队，提供 7×24 小时的技术支持服务。

保障团队随时接受客户提问，接到技术支持要求时，2 小时内做出明确响应

和安排并为客户方提供维护服务和故障诊断报告。若需现场服务，工程师 4 小时内到达现场，24 小时内提供书面解决方案。

1.6.1.2. 数据保障思路

为保障数据服务过程中高质量服务，建立“三个特性、四个凡事、五大模块”的质量管理方法，并更具质量管理方法进行数据服务工作。

三个特性：数据的完整性、准确性、及时性

四个凡事：凡事有人负责、凡事有章可循、凡事有据可查、凡事有人监督

五大模块：完善质量管理体系、完整质量保障组织及职责、资源管理、数据采集计算实现、数据核验处理改进

以上述管理方法为基准，建立长效的数据质量监控机制，制定数据质量保障方案，逐步固化常规数据质量保障流程，组建数据质量保障小组，围绕原始数据质量监控、数据计算处理、数据传输处理、数据结果校核等方面，开展数据质量保障工作，具体如下：

定义：定义项目的数据保障工作目标，以指导整个项目运转过程中数据质量管理的工作。包括数据收集、汇总、计算、分析有关形式和信息环境。

度量：按照指定的数据质量维度对项目数据质量进行评估，

分析：使用数据分析技术分析、评估劣质数据对业务产生的影响，以及确定影响数据质量的客观原因，确定这些原因的影响的数据质量的级别。通过改进组织管理流程，最大限度控制由管理上的缺陷造成的数据质量问题。

改善：最终确定行动的建议，为数据质量改善制定方案，包括数据级和组织级的,建立数据错误预防方案，并改正当前数据问题，对数据和管理实施监控，维护已改善的效果。

1.6.1.3. 数据质量保障维度

1. 数据完整性

数据完整性是保障数据质量的一个重要标准，本项目中平台的建设以及对外应用服务离不开采集数据的完整性，尤其针对海量数据的采集，在采集过程中，会出现数据不一致、丢数据等诸类现象。因此，必须加强采集数据的完整性，保障采集过程中数据不丢失、数据一致。

系统通过提供丰富的数据校验手段来保障数据质量，在数据完整性方面，通过将数据校验依附在数据采集任务完成后，通过对数据源与目标数据库之间的数据进行对比分析，从而进一步来分析、发现与解决在数据采集过程可能产生的异常错误信息，数据检测达到以下要求：

（1）配置检测任务：通过配置生成检测任务，包括设置检测对象、检测周期、经验时间点等参数进行配置；

（2）执行检测任务：检测到数据完整性异常后，可采取智能补采（如定时补采等）和人工补采的方式进行。

数据校验从校验对象细粒度来分析，可以划分为文件级校验与记录级校验二大类，具体二者功能情况如下：

文件级校验：主要提供校验文件校验、引用完整性与唯一性检查等方面的检查内容。

记录级校验：主要包括提供字段类型、字段长度、数字精度、取值范围、可否为空、忽略字符、正则表达式等方式针对特定记录的校验处理。

数据校验模块还内置了部分的数据检查功能，主要包括：

数据唯一性检查：可以根据系统检查规则中设定的唯一字段进行校验，检测

出重复记录。

外键完整性检查节点：根据系统检查规则中设定的外键关系进行校验，检测外键引用是否完整。

数据处理节点进行数据采集过程中会根据数据的定义进行校验，校验内容有类型，长度，是否为空，精度，范围，格式等信息。如果数据不符合，会进行过滤，只有正确的数据才能继续使用。对于错误的数据，可以进行输出，包括错误原因和错误字段序号等信息。相关的错误类型和数量等统计信息也会绑定到流程变量中，以便后续节点进行判断使用。

2. 数据一致性

一致性是指数据是否遵循了统一的数据规范，数据集合是否保持了统一的数据格式。数据质量的一致性主要体现在逻辑规则与数据的组成规范。规范指的是，一项数据存在它特定的格式。一般的数据都有着标准的编码规则，对于数据记录的一致性检验是相对规范，符合标准编码规则即可。

3. 数据准确性

准确性是指数据记录的信息是否存在异常或错误。和一致性不一样，存在准确性问题的数据不仅仅只是规则上的不一致。更为常见的数据准确性错误就如乱码。其次，异常的大或者小的数据也是不符合条件的数据。

数据质量的准确性可能存在于个别记录，也可能存在于整个数据集。同时，一般数据都符合正态分布的规律，如果一些占比少的数据存在问题，则可以通过比较其他数量少的数据比例，来做出判断。将通过统计分析算法进行对比纠错，并借助专业的数据分析工具进行分析，保障数据质量。

4. 数据及时性

及时性是指业务数据从产生到可以利用、可以查看的时间间隔，也叫数据的延时时长。及时性对于数据分析本身要求并不高，但如果数据分析周期与数据采集的时间间隔过长，就可能导致分析得出的结论失去了借鉴意义。

1.6.1.4. 数据质量校验流程

数据质量校验流程涉及数据抽取、数据计算、数据统一汇总、数据内部自查以及数据问题处理五个阶段。

1.6.1.4.1. 数据抽取校验

大数据平台每天都有很多 ETL 任务定时执行加载数据，确保 ETL 加载数据的完整性、准确性是数据质量管理的基本要求。

1. 日常数据校验

数据质量管理人员每天要对数据清洗加载任务执行情况进行日常例行检查。

数据校验方法选择 ETL 数据质量校验采用以下三类方法来进行判断：记录数检查法；关键指标总量验证法；值域判断法。

数据校验周期考虑性能问题执行全部数据校验需要的时间过长，物理消耗过高，因此根据每个主题数据的评估等级确定校验频率。

评估等级与校验频率的对应关系如下：

一级：业务相关性较强数据，每次加载都执行数据校验

二级：业务相关性一般数据，每三次加载执行一次数据校验

三级：业务相关性较弱数据，每六次加载执行一次数据校验

对于业务相关性极强与特定的主题数据，额外增加经验审核法与细粒度的数据校验方案。

2. 定时数据抽取

数据清洗数据的校验确保每天加载的增量数据的完整性、准确性，在此基础上，数据质量管理小组每季度组织一次全量数据的定期抽查。

定期抽查的范围包括评估等级为一级的所有主题数据，评估等级为二级的二个主题的数据，评估等级为三级的一个主题的数据。

定期抽查采用数据质量评估方案中所涉及的全部定义方法。

1.6.1.4.2. 数据计算校验

数据计算校验包括日常计算校验与试算机制。

数据计算选择口径输入结果后，会去历史同时段数据做递归算法验证，以确保数据计算的正确性与有效性。

1.6.1.4.3. 数据统计校验

数据统计结果输出后，区级连续两日/两周/两月的同时段数据不会存在骤增或骤减，数据在每个时间周期的原始文件处理量检查结束后，自动启动解析结果统计校检，对比该时间周期的信令总条数、用户总数与前一日同时段数据是否存在骤增或骤减差异，若发现异常则触发预警告知维护人员检查处理。

1.6.1.4.4. 数据内部自查

1. 底层数据核验

所需原始数据为每 6 秒钟获取一个，每 10 分钟合计 100 个文件，其中正确解析处理的文件会重命名为.loaded 后缀，未能正确解析处理的文件会重命名为.error 后缀。平台每 10 分钟自动检查一次前 10 分钟内收到的文件数量、正确解析文件数量、未能正确解析处理的文件数量，若发现异常则触发预警告知维

护人员检查处理。

1）信令数据异常波动核查

接口 ETL 程序会记录周期时间内接收到的信令数据条数和信令数据正确条数，信令数据正确条数指的是过滤掉字段缺失、字段错误、数据错误、内容错误的信令数据后的条数统计，周期时间一般设定为 1-5 分钟，对这些周期性的数据进行同比和环比，可以检测出信令数据条数是否存在明显差异，当差异造成的波动幅度超过 10%则在系统内形成相应事件并进行报警，以便后续进行人工排查。

2）关机时间异常波动核查

每个用户的信令都是具有连贯性的，当单个用户连续长时间无新信令就会触发疑似关机事件，用户正常的关机则会触发关机事件。系统侦测到某一子节点下突然涌入大量关机和疑似关机事件时，形成相应事件并进行报警，以便后续进行人工排查。

2．阶段数据核验

在一般情况下，区级连续两日的同时段数据不会存在骤增或骤减，数据在每个 10 分钟的原始文件处理量检查结束后，自动启动解析结果统计校检，对比该 10 分钟的信令总条数、用户总数与前一日同时段数据是否存在骤增或骤减差异，若发现异常则触发预警告知维护人员检查处理。

另外，核验连续两周、两月同时段人口数据是否存在骤增或骤减情况。若存在骤增或者骤减情况，若发现异常则触发预警告知维护人员去核验数据提取日志，逐级追溯日、周、月的数据核验日志。

3．随机抽出式核验

数据平台每小时随机抽查十个客户加密串的信令最新更新时间，若发现超过一半用户加密串的信令最新更新时间与当前真实时间严重偏离，则触发预警告知维护人员检查处理。

4. 平台正确运行核验

数据平台所在服务器设定了定时检查任务，每小时自动向平台内部检查接口发送自检指令，检测各接口反馈值是否包含正确的内容，若发现异常则触发预警告知维护人员检查处理。由于平台采用了分布式架构，维护人员可单独重启某些模块来恢复平台正常而不影响整个平台的运作，重启过程一般不超过 1 分钟，重启过程中延迟处理的数据也会在启动完成后数秒内全部恢复。

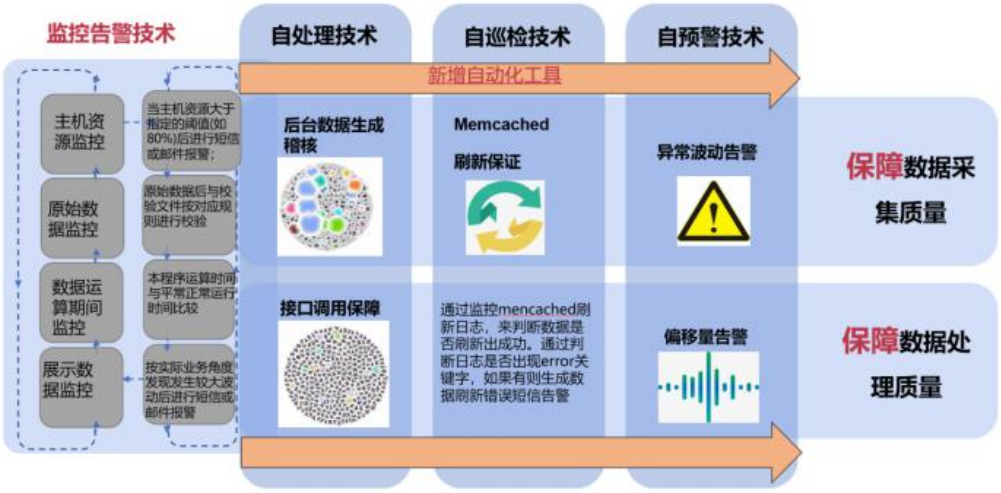
1.6.1.4.5. 数据问题处理

发现并确认源头数据存在缺失或故障数据包问题后，维护人员立即从备用数据源手工采集对应数据补充到本平台接口，可在 30 秒内完成数据补齐工作。若备用数据源也未采集到所需原始数据，则立即电话通知数据源维护人员检查原因并补齐原始数据，在 1 小时内完成数据补齐工作。

1.6.1.5. 数据质量保障措施

1.6.1.5.1. 自动化质量管理

运用自动化监控告警技术、自动处理技术、自动巡检技术、自动预警技术形成自动化质量管理机制，以保障数据采集及数据处理质量。



1.6.1.5.2. 全流程保障管理

建立数据采集、数据存储、数据处理全流程的数据质量保障措施，保障提供数据的准确性和稳定性。

1. 数据采集保障

采集的信令数据偶尔会存在不同程度的缺失和错误，相对信令数据，详单数据不够及时，但准确度较高，同一用户相同时间点理论上各个数据源应该连接相同的基站。详单数据不存在丢失的情况。基于详单数据来修正和补充信令数据有理可循。针对信令数据错漏问题，同一个用户详单数据相同时间点中包含了业务发生时的基站信息，可以使用话单数据来补充和替换信令中的错误数据。

2. 数据存储保障

针对不同类型数据采用不同存储方式，为了便于提升查询的高效性，标签数据存储在 Hive 中，明细标签数据存储在 HBASE。

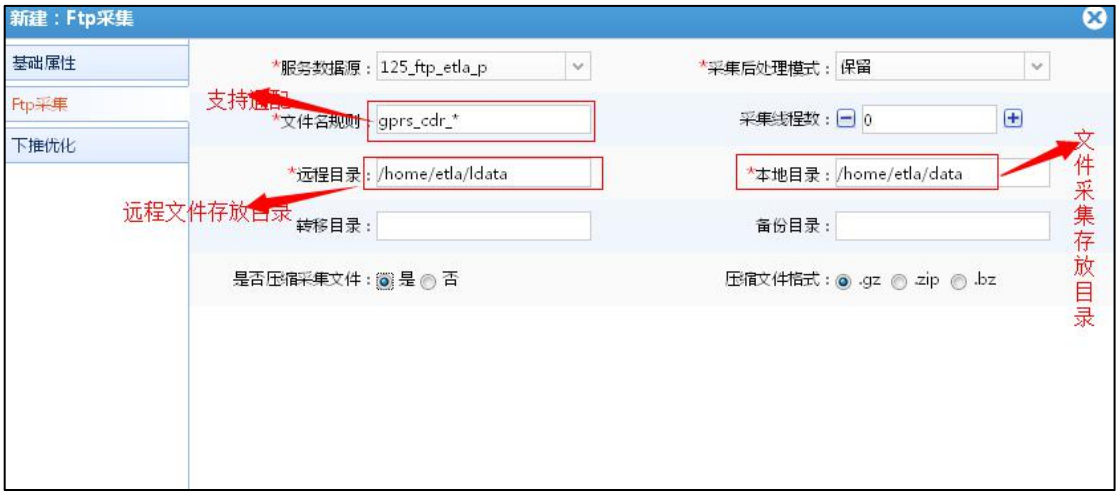
3. 数据处理保障

1) 大文件压缩传输

针对大数据压缩传输，数据的压缩技术是关键，在大数据应用领导，我司支

持主流的压缩方式，可以对大数据平台中的常规数据、索引数据、临时表空间数据等进行全面压缩，压缩能力可以从 1-10 级进行细粒度调整，在 IO 与 CPU 之间进行均衡配置。

系统提供相关的压缩采集可视化界面，让功能选择是否进采集压缩传输，界面效果如下图所示：



考虑异构环境的特殊性，不同的应用场景下，选择不一样的压缩方式，主要有以下场景：

（1）原始数据就是压缩的，那么最好按照原始数据的压缩格式直接上传 HDFS，在 HDFS 中并行处理过程中进行压缩流数据读取。

（2）MapReduce 或者 Hive 的中间结果的压缩。可以在 MapReduce 程序中设置 Map 后的中间结果用压缩模式，然后交给 Reduce。这种情况下，适合采用 snappy 与 lzo 这种压缩解压性能快的压缩算法。

2) 本地解压缩

（1）处理后数据或者冷数据的压缩，可以采用压缩率稍高的算法进行压缩以节约空间，比如 Zip 或 Bzip2，特别是在冷数据归档的情况下。但如果数据可能会频繁被即席查询的话，优先选择解压速度快一些的压缩算法。

（2）对于热数据传输落地后实现自动解压缩能力。

1.6.2. 数据安全保障方案

1.6.2.1. 安全需求分析

建立数据安全保障机制，对软硬件环境、数据传输的各个环节进行监控，确保实施流程的合规和安全；建立风险预警机制，加强网络安全应急保障，防范数据错误和数据泄露的风险。保证数据库和其它文件只能被授权用户和播控终端访问，防止在本地存储或者网络传输的数据受到非法篡改、删除和破坏。应用系统应具有认证功能，以鉴别用户身份并验证。系统应支持访问控制、安全检测、攻击监控等一系列安全功能，应提供完整的网络安全监控、报警和故障处理功能。

1.6.2.2. 安全保障机制

1.6.2.2.1. 遵守政策规范

遵守《中国个人信息保护法》、《中国移动大数据安全管理要求》管理规范要求，保证项目合法、合规运行。

安全管理贯穿信息系统的所有环节，是安全保障的重要环节。通过建立安全

管理制度，实施安全管理培训教育，使安全管理的科学化、系统化、法制化和规范化，确保平台的安全性。

关于安全管理制度，主要是指人员安全管理制度、设备安全管理制度、运行安全管理制度、安全操作管理制度、安全等级保护制度、有害数据防治管理制度、敏感数据保护制度、安全技术保障制度、安全计划管理制度等。

用户安全教育与培训也是安全管理设计的重要组成部分。对不同层次用户制定相应的教育培训计划及培训方式，提高用户安全意识、法制观念和技术防范水平，保障系统的合理使用和安全运行。

安全管理制度需制定信息安全工作的总体方针和安全策略，说明安全工作的总体目标、范围、原则和安全框架等，对安全管理活动中重要管理内容建立安全管理制度，并对安全管理人员或操作人员执行的重要管理操作建立操作规程。

技术上的安全性需要依赖制度来落实和保证，因此安全管理制度的制定和执行是系统安全性建设必不可少的工作。

1.6.2.2.2. 工作合规实施

明确安全管理机制及各项工作责任人。

确保网络安全管理、操作系统应用、数据库处理、应用软件配置、端口和账号口令管理、日志数据管理等方面加强管控，保障各项工作受控合规。

在系统的互操作性测试中，充分考虑对其他系统的影响，选择适当的时间、方法。

沙箱环境和生产环境完全隔离，保证数据安全。北京移动的生产环境进行封闭管理保证安全生产。

1.6.2.2.3. 数据合规审查

不断完善标准化数据提取制度及安全审查制度。

从数据提取确认，报告模板使用，接口的定义等一系列操作流程，保证数据生产的合规和效率。

1.6.2.2.4. 防范数据泄露

专人负责数据核算，利用自动化工具核查。

专项组负责交付物的确认，数据横向对标。

专人负责向客户解释数据口径，保证权威性。

1.6.2.2.5. 网络安全保障

保障数据安全，应用服务器与数据库分开部署。

关闭与业务无关的服务和端口，实现服务端口最小化。

系统补丁漏洞常态化更新，定期进行攻防演练。

1.6.2.2.6. 风险预警机制

节假日重点保障制度，7×24 小时保障体系。

操作日志和重要存储保存 90 天以上备查。

建立突发事件应急预案，健全日常检查机制，识别项目潜在风险。

1.6.2.3. 安全保障措施

1.6.2.3.1. 系统安全

服务器资源由 IT 资源池统一分配，系统资源（包括操作系统、数据库、中间件、WEB 服务器等）满足移动安全基线配置规范要求。

1.6.2.3.2. 应用系统安全

本系统对所有使用系统的人都要分配工号和经过授权鉴权才能登录系统，对于外部系统要调用本系统提供的服务需要通过本系统分配的用户和身份验证密

码才能与本系统的服务对接。系统的用户信息保存在数据中，用户的关键信息通过密文的方式保存起来，只能开发给应用程序或者 DBA 权限的操作员使用。运行时登录的和登出都登记了详细的日志。

应用程序的源代码严格按照安全规则存放在局方指定的介质上，不直接存放到应用服务器上。应用服务器上只存放编译后的程序文件。并严格控制系统应用程序文件访问的权限。

对应用提供严格的审查功能，对系统关键数据的每一次增加、修改和删除等操作都能记录相应的修改时间、操作人和修改前的数据记录。针对应用服务器上的日志文件、配置文件、程序文件、系统目录等都要设置及其访问权限。保证系统应用服务器安全运行。

1.6.2.3.3. 数据安全

对于后台、数据库统一通过 4A 进行登录访问。对于手工或 SQL 脚本直接在数据库中进行增删改，需要对数据库操作日志进行监控和审计，防止数据被恶意破坏。前台页面增加数据格式有效性校验，避免异常数据写入数据库。相关密码在数据库中通过密文存储。另外对数据安全性做到如下两点：1) 本项目数据服务内容要求的汇总统计及结果数据；2) 通过本地数字传输专线，统计结果以 FTP 格式输出。

1.6.2.3.4. 4A 要求

系统资源（包括操作系统、数据库、中间件等）均纳入业务支撑系统 4A 统一管理，实现账号统一管理、单点登录、集中认证、授权与审计。

所有 WEB 应用接入 4A，应用系统不应存在独立的登录页面，若因集团业务探测等特殊原因必须保留应用系统自身登录页面的必须从应用层面实现源 IP

地址和帐号口令鉴权等访问控制机制。

1.6.2.3.5. 日志和审计要求

针对系统所发生的所有事件都必须有详细的日志记录，并进行定期审计。这些事件包括系统用户登陆和注销，应用程序用户的登陆和注销、数据库系统用户的登陆和注销、关键数据的读取、修改和删除、系统维护记录、日志管理等。

1.6.2.3.6. 数据库高可用

数据库之间采用双节点模式构建，单点故障可以自动切换。

1.6.2.3.7. 应用高可用

后台应用采用主备模式部署，当主进程死掉后，备用进程（部署在其它机器）自动接管主进程的任务处理。

1.6.2.3.8. 服务高可用

WEB 服务在多服务器上部署。HTTP 请求通过负载均衡设备分发给 WEB 服务。当其中一个服务器宕机或 WEB 服务僵死后，负载均衡设备会把任务请求派给其它服务，从而实现 WEB 服务高可用。

1.6.3. 项目进度保障方案

1.6.3.1. 项目实施进度控制

(1) 项目负责人实时掌握整个项目周期的各节点工作进程，任务的划分以天为单位，项目经理组织人员进行系统的实施，并作好对应的记录，对已解决的问题做好详细记录。

(2) 沟通与交流，项目经理定期与项目人员进行交流，调动项目人员积极性，了解实际的进展以及数据的准确程度。

(3) 做好项目的过程管理，对项目实施中的各类突发情况进行及时的记录、上报、查找原因，并第一时间进行处理解决，且有详细的说明和记录。

(4) 所有项目工作，做到日清日结，不堆积项目工作，不拖延项目进度。

1.6.3.2. 项目实施进度保障

项目进度管理是项目管理者围绕项目要求编制计划，付诸实施且在此过程中经常检查计划的实际执行情况，分析进度偏差原因并在此基础上，不断调整，修改计划直至项目交付使用，通过对进度影响因素实施控制及各种关系协调，综合运用各种可行方法、措施，将项目计划控制在事先确定的目标范围之内，在兼顾

成本，质量控制目标的同时，努力缩短时间。

本次项目进度管理主要通过以下方案进行保障：

1、制定责任拆分机制

将项目进度目标进行责任拆分，责任落实到每个进度控制人员，同时在项目进行过程中每周定时进行两次及以上的项目沟通例会，项目组所有人员进行进度汇报，同时将问题进行汇总沟通，及时推进项目运行。

（1）识别进度计划所有者

识别所有者或负责开发所有或部分项目进度计划的个人，通过采用 WBS(作业分解结构)或组织分解结构作为进度开发的基础，通过功能范围划分落实所有者进度推进工作。

（2）划定项目任务和里程碑

对每个最低级别的 WBS 元素，识别任务和里程碑对应交付的元素。可交付物通常设置为里程碑，产生可交付物的活动被称为任务。里程碑是一个时间点，被用于管理检查点来测量成果。

（3）排序工作活动

在确定了项目任务和里程碑之后，进行项目逻辑排序，来反映将被执行的工作方式。排序建立了任务和里程碑之间的依赖，并被用于计算交付产品的进度。

（4）核实项目进度

通过确认个人项目任务及个人工作的排序工作，通过项目任务和时间节点的不断沟通和检验，来推动整个项目的进度，由项目负责人统一把控项目任务进度的核实，并在项目期间结合实际进行调整。

2、制定项目激励机制

项目过程制定完善的奖惩机制，对进度控制人员的工作进行协调和考核，利用相关激励手段督促工作人员的进度控制。

项目周期由项目负责人或项目经理进行整体把控，项目任务划分以天为单位，并由各组的负责人对项目组人员的任务进度进行汇总记录，对于及时完成项目任务项目人员进行相应的奖励。

3、制定项目组沟通机制

项目团队制定了顺畅的沟通机制，保证小组间的信息沟通渠道顺畅，减少了因沟通不畅导致项目的延后情况。

（1）建立明确沟通机制

项目采用周例会的形式作为项目实施过程中的常态沟通机制，坚持“常规工作定时报，重点工作随时报”的原则，向小组领导及项目负责人进行汇报，及时对工作中的问题进行解决。

（2）树立统一的目标观念

项目过程中做到项目信息共享，项目负责人对项目小组人员定期举办思想交流会议，提升团队凝聚力，激发员工工作热情。

（3）缩短团队组建与磨合时间

项目负责人定期组建项目小组活动，公开讨论团队成员选择原因，成员间的互补能力，通过信息激励，激发团队意识，推动小组成员快速融入团队，建立团队自我认知。

项目负责人通过引导成员参与项目计划制定及任务分配，同时激励成员积极表达问题，提升成员团队参与意识，进一步增进团队组建凝聚力，缩短成员间磨合时间。

1.6.4. 项目交付成果

1.6.4.1. 月度监测数据

按照三网融合平台建设项目的数据需求及技术规范，在次月 10 日前配合输出提供所需的统计级标准数据，对采集到的人口数据通过模型算法和结果汇总，并通过专线传输至指定平台，实时数据接口传送频率不低于计划内时间。月度监测数据具体内容如下：

1） 实时活跃用户数量

2） 月度居住人口、工作人口数量

3） 区域及定制区域的 30 分钟流入、流出人口数量

4） 月度职住信息统计：包括工作人口的居住地、居住人口的工作地统计

5） 月度居住人口流向情况、月度工作人口流向情况

6） 居住用户中外来人口比重：按照身份证户籍统计分析

7） 月度居住用户、工作用户流入流出情况，流入人口来源地及流出人口去向地

8） 实时客流量（商圈、旅游景区）

9） 客流停留时长（商圈、旅游景区）

10） 游客日流量（商圈、旅游景区）

1.6.4.2. 人口洞察报告

本项目每月 15 日前提供上月人口分析报告，分析区域内不同时间维度的数据量变化、占比变化等，掌握区域内人口流动的主要趋势和特征，分析报告以

Word 形式呈现，报告具体包括但不限于以下维度：

1） 统一口径及指标说明

2） 北京市各区县本月用户分析

3） 朝阳区本月用户分析

4） 朝阳区本月街乡居住用户、工作用户分析

5） 街乡居住用户、工作用户增幅前十名

6） 朝阳区本月重点区域居住用户、工作用户分析

7） 重点区域居住用户、工作用户增幅前十名

8） 分街乡居住时段（02 点）用户实时总量数据统计

9） 分街乡工作时段（14 点）用户实时总量数据统计

10） 全区居住时段（02 点）和工作时段（14 点）用户实时总量数据统计

11） 本月进入街乡工作的人，来源地分布(分区县)

12） 本月进入街乡居住的人，来源地分布(分区县)

13） 朝阳区本月重点村居住用户分析；

14） 朝阳区本月重点村流出用户分析；

15） 朝阳区本月重点村流入用户分析；

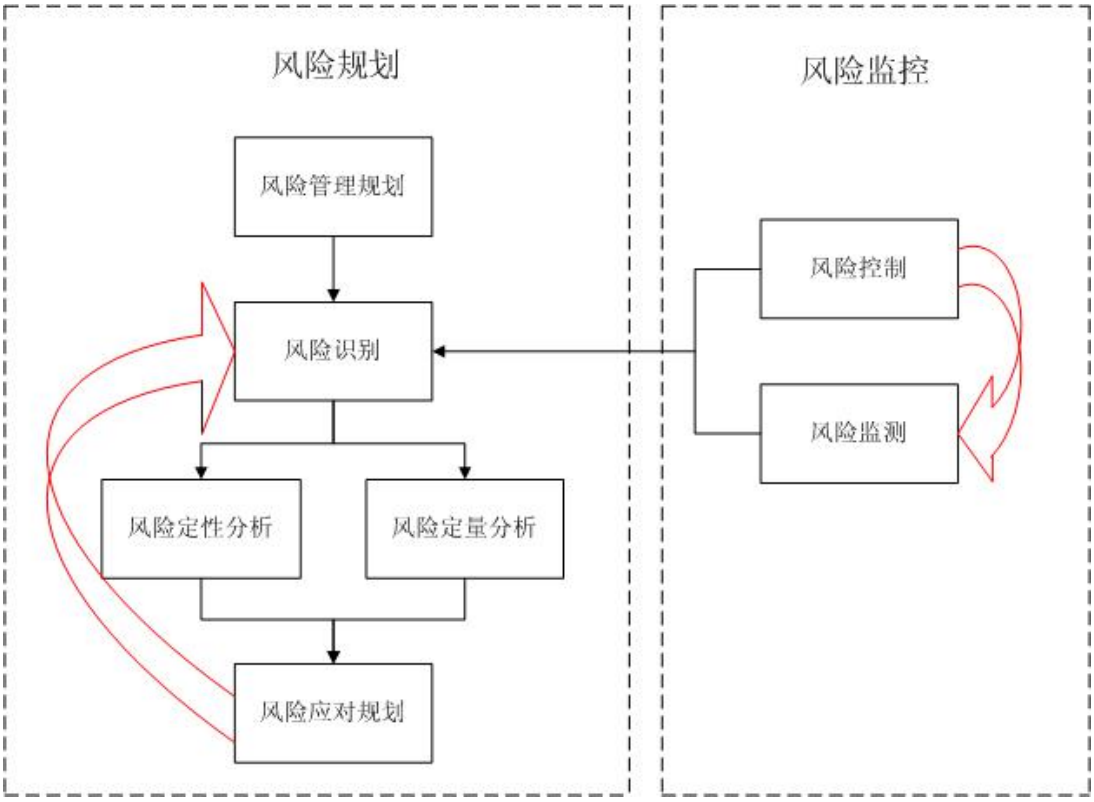
1.6.5. 项目风险管理

1.6.5.1. 项目风险识别

项目风险是项目管理过程中潜在的问题。项目风险可能引起项目不能按时交付，或者达不到预期的质量，或需要增加项目成本。

风险管理是一种对项目风险进行识别、分析、应对的系统管理过程。它包括鼓励对项目目标有正面影响的风险发生并加强其影响、减小对项目目标有负面影响的风险发生并减弱其影响。

针对风险管理策略如下图所示：



1) 风险管理规划：决定如何进行与规划项目的风险管理活动。

2) 风险识别：判断哪些风险会影响项目，并做正式记录。

3) 风险分析：

- 风险定性分析：对风险及其条件进行定性分析，并依其对项目目标影响进行排序。
- 风险定量分析：量度风险的概率与后果，估计其对项目目标造成的影响。

4) 风险应对规划：制订为项目目标增加机会、减轻风险的程序与技术。

5) 风险监测与追踪：在项目整个生命周期内监测残余的风险、识别新的风险，执行减轻风险计划，并对这些计划的有效性进行评估。

风险是由项目团队成员进行识别和分析的。风险必须上升到项目级对待。

风险管理是一个连续的前瞻性的过程，它是业务和技术管理过程的重要组成部分。风险管理需要处理可能危及关键目标的问题。应用持续风险管理的方法来确保有效地抵御和缓解项目生存周期中具有关键影响的风险。

有效的风险管理包括，按照项目策划过程中所拟订的共利益者介入计划，与共利益者合作，早期识别风险。

1.6.5.2. 项目风险类别

1) 需求风险

本项目的建设是一个涉及相关部门较多、数据量大的复杂国程，在项目实施过程中与多与业务部门进行沟通，才能逐步明晰项目需求。同时，明确项目建设需求和需要的技术，定期召开进度会议，规避项目需求的频繁变更，使得项目进展严重滞后，最后造成项目的失败。

2) 协调与沟通风险

在项目实施过程中公司需要协调多个部门，与这些部门的沟通与协调可能直接影响到本项目的质量与进度。因此，建立高效的协调与沟通机制，减少相互之间的误解与拖延，是保障本项目成功实施的关键点之一。这需要各相关单位充分理解项目沟通管理的重要性，严格遵守项目管理的各项规章制度，提高协调沟通的效率，降低项目协调与沟通的风险。

3) 项目人员风险

项目人员压力会随着项目的进展逐渐加大，工作效率也可能会随着项目的进

展逐渐降低，造成工作效率低下，甚至会造成项目成员的不稳定。这就需要招标方与公司相互理解，明确共同的目标，发挥团队精神，同时要合理规划项目进度，作到劳逸结合，提高项目人员的积极性，降低项目人员的风险。

在本项目中，针对以上涉及的风险，是需要仔细考虑并加以规避：

项目人员风险规避：由于该项目实施周期有限，所以必须保证在项目期间人力资源的稳定性，否则将影响工期。北京移动将调动稳定且具备专业性的人员进行组建团队，使项目工作能够按时保质的完成。

数据安全问题：在数据管理过程中需要严格把控数据处理期间人员的权限管理，减少不必要的数据与不相关人员的接触问题，规避数据安全隐患。

项目过程中，项目组要紧密监控项目过程，跟踪项目风险，及时发现、识别项目中的风险变化及新的风险点，更新风险跟踪表，快速做出调整，消除、控制不利的风险因素，将风险控制在可接受的范围内。

1.6.5.3. 项目风险管控

1）需求风险管控

在建设过程中多与业务部门进行沟通，才能逐步明晰系统的需求。为了减少该项目需求不清和需求频繁变更导致项目进度延后，北京移动建立一套专业需求风险管理方法：

- 明确需求：明确需求责任人和需求提出流程，规定需求提出的截止时间和封版时间。
- 需求评审：完成工作任务后，召集相关的需求提出人、项目组成员召开需求评审会，再次确认需求完成情况，并做好需求评审记录。
- 需求变更管理：原始需求提出变更后，进入需求变更流程，召开需求变更评审会，分析变更对项目范围、进度、资源的影响，做出变更决策。

2） 质量风险管控

项目团队在项目实施过程中主要通过以下手段管控项目实施质量风险。

➢ 组建专业、稳定项目团队

通过派驻经验丰富的项目经理和业务能力专业、稳定的项目成员，保障项目过程中的质量；

➢ 制定项目组内评审机制

对项目实施过程项目方案、项目实施计划、项目实施管理、售后管理等记录文档进行项目组内评审，通过项目组各负责人审核，方可进行后续实施。

➢ 制定过程监督机制

组建项目实施监督小组，对项目实施过程进行监督管理，制定项目实施过程监控关键节点、检查机制缺陷弥补机制，保障项目建设过程不偏离。

1.6.6. 培训服务方案

1.6.6.1. 培训目的

为了保证我司提供运维的系统能稳定运行，提高系统使用效率，需要对用户进行培训。本着“一个目标、一个团队、一个标准”的态度，对各级相关人员进行系统培训，使其对系统有充分了解，熟悉系统的设计原理和工作方式，掌握系统的工作流程和操作方法。

1.6.6.2. 培训人数

培训人数根据采购方要求确定，具体人数可根据项目需求进行协商。

1.6.6.3. 培训对象

相关培训主要面向操作人员开展，根据其业务需要和自身素质，安排合适的

培训内容，才能保证不同层次人员都学有所成，胜任各自所承担岗位职责，也为以后系统的移交打下良好的基础。相关人员经过培训将能熟练地使用系统操作界面，培训内容包括平台及业务系统简介、业务管理详细讲解、数据分析服务详细解等。

1.6.6.4. 培训方式

为了对用户实施专业化，有针对性地培训，采用多样化，多形式的培训方法，以达到最佳培训效果，使受训人员能够在最短的时间内熟悉掌握培训内容，尽快投入实际应用。根据培训队伍的建设方案，我公司将制定出一系列的培训方式，具体培训方式如下：

1、网络远程培训

利用系统设计的培训系统，提供在线仿真培训教程，便于数量最大的基本业务人员快速准确地掌握系统的使用方法和技巧。

2、分班培训

按照不同的培训对象、培训内容，制定出不同的培训班进行强化培训，如：一般操作人员班；系统管理员班；系统维护人员班。

3、授课培训

授课是培训最常用的方式，组织分班培训也是为了授课方式的集中进行，授课时，教员可以按照系统的逻辑结构将学员一步步地引导到正常的操作方法上来，容易给学员一个完整的系统工作模式，使学员在最短的时间内接受最正规的教导。

4、实例示范与练习

这种方式给学员的印象最直观，它通过具体的示范可以将内容具体化、形象

化，具有很强的说服力，易于理解，能够将复杂的操作环节清晰化。

5、辅助学习工具

具体培训中，可以采用教学视频、教学录音等方式，方便学员经常性的学习，更有利于掌握所学内容。

6、综合培训方法

对于项目培训，不能拘泥于一种培训方法，将在具体培训中根据情况，按照“分班集中授课，辅助实例练习”的综合方法，力争在最短的时间内，使相关工作人员熟练使用系统。

1.6.7. 项目售后服务方案

1.6.7.1. 基本保障承诺

1.6.7.1.1. 服务期限

（1）服务期限为一年，自合同签订之日起开始计算。我公司保障在此时间之前完成项目成果的提交。

（2）服务保证期为一年，我公司对服务内容提供一年的保证期，自验收合格之日起计算。在该期间，如我公司的服务质量存在问题，我公司承诺采取相应的补救措施。

1.6.7.1.2. 服务人员管理承诺

（1）人员管理：我公司保证，派驻稳定的专业团队，且应答文件中罗列的骨干成员与实际中标后的执行人员保持一致。我公司保证项目人员稳定，在未经过项目单位同意的情况下，不随意更换人员，项目人员更换将提前 1 个月申请，经项目单位同意并办理交接手续后方可更换。项目组将严格把控项目人员进出场

控制，做好项目人员管理。项目单位有权因团队服务人员达不到工作要求提出换人，我公司在 15 天内配备符合能力要求人员。

（2）团队成员：我公司承诺项目成员不少于 8 人专业技术人员负责本项目的实施和服务工作。

1.6.7.1.3. 项目全过程管控

我公司承诺建立项目集中管控模式，从项目沟通、项目进度、项目质量、项目风险等多角度进行管控，并从项目准备、数据分析、项目实现、专业交流活动等各阶段进行全过程管控，保证项目顺利开展。

1.6.7.1.4. 规范管理制度建立考核评价机制

我公司承诺建立项目管理制度，明确各职责工作分工，为了实现项目最终顺利完成，对项目进行有效地计划、组织、领导和控制。同时建立合理全面的项目考核评价机制，每月对项目进行考核评价，从考勤、服务、规范等维度考核，进行通报，并做好后勤跟项目宣传管理，全面支撑项目顺利进展。

1.6.7.2. 运维服务承诺

（1）公司保证项目执行期间团队成员的稳定性，保障不出现任何岗位的空缺状态。

（2）我公司中选后，核心团队人员原则上不变更；如若变更至少提前一月向项目单位提出申请，获得同意才进行变更，并保障变更交接期。

（3）项目过程中，我公司及时将项目进度和存在的问题以周报形式书面反馈给项目单位项目负责人，并完善下一阶段工作计划；在项目过程中，以双周报和月报方式，反馈专题分析报告；在项目结束时以结项汇报的形式向项目负责人汇报，并交付相关交付物。

1.6.7.2.1. 运维服务方式

1）服务措施

我公司重点采取以下措施，为客户的系统使用扫除后顾之忧：

我公司在总部成立提供 7 x 24 小时响应服务的开发和技术支持中心，在现场成立专门的技术支持小组，充分保证对系统的长期技术支持，提供及时高效且方式全面的服务，包括但不限于邮件、电话、远程维护、现场服务等方式；

合同履行过程中，我公司保证及时对采购方提出的关于项目的任何问题，按照采购方要求进行当面解答或作出书面解答等。

2）维护方式

我公司客户服务中心将以以下方式提供技术支持及服务：

维护方式	维护内容	服务时间
热线服务	我公司提供 7 x 24 小时/周热线服务，可快速响应用户请求。所有用户问题及解决结果均存档备查，客服中心对问题解决过程进行监督、跟踪并输出客服工作报告。	合同期内 7 x 24 小时/周
邮件服务	我公司中标后，会设立专门的邮箱，来响应客户的需求。来解决电话里说不清的问题。	合同期内 7 x 24 小时/周
远程支持	对部分故障采取远程技术支持进行高效的故障诊断和排除。对于电话、	合同期内 5x 8 小时/周

	传真或 E-MAIL 中无法讨论解决的复杂问题，技术工程师将利用远端测试手段，对系统故障进行远程诊断，进行远程故障排除或向局方提供详尽解决方案。 历史经验表明，远程维护方式可以高效诊断 90%以上的故障原因，并且修复 80%以上的故障。	
	（1） 诊断并查明系统的错误或故障；	
	（2） 修正错误或故障，并将系统恢复至最佳状态；	
	（3） 提供最佳临时性解决方案；	
	（4） 向采购方提供修正或改进后的结果并说明修正或改进后的差异。	

1.6.7.2.2. 运维服务流程

用户报障的途径有很多种，无论是工程师、销售还是其他人员接到用户的保障，都有义务交接到客服中心，进行记录，并及时交接到技术支持部门。为了保证用户的所有问题都依规范的程序得到响应和解决，我公司将与采购方共同建立按故障的严重程度排列优先顺序的标准，这一标准将使服务资源得到合理分配，

有效提高服务效率。

故障级别定义：

编号	故障级别	故障现象
P1	一级故障	现有的系统停机，或对最终用户的业务运作有严重影响
P2	二级故障	现有系统的操作性能严重下降，或由于系统性能明显下降，对最终用户的业务运作重要影响
P3	三级故障	系统的操作性能受损，但最终用户大部分业务运作仍可正常工作
P4	四级故障	在产品功能、安装或配置方面需要信息或支援，很显然对最终用户的业务运作几乎无影响，或根本没有影响

故障处理 SLA：

1. 临时解决时限指对终端用户来说恢复了系统服务，问题有可能被采用临时方案解决，也可能是最终解决方案。如果我方提供临时方案以后，故障级别可以降级。

2. 最终解决时限指采取最终方案解决故障。

故障级别	一级故障	二级故障	三级故障	四级故障
响应时间	15 分钟	15 分钟	30 分钟	30 分钟
临时解决时限	90 分钟	4 个小时	1 个星期	3 个星期
最终解决时限	2 个星期	30 天	60 天	90 天

1.7. 项目实施团队

项目经理：负责实施计划制定、实施推进，确保各项工作任务按时间节点保质保量完成，确保项目顺利交付和验收；协调处理项目实施中出现的相关问题，确保项目稳定实施。

业务分析组：负责行业发展态势研究、行业数据分析，负责行业咨询、节假日分析报告编写等。

数据分析组：支持运营商大数据的数据清洗、开发、计算处理并针对数据问题进行解答与需求对接。

业务咨询组：支撑行业应用前瞻性研究和行业分析，支撑深度咨询报告。

数据运维组：支撑数据质量、数据安全的管理与保障。

服务人员配置方案

姓名	拟在本项目担任职务	从业年限	服务内容
刘亚东	项目经理	9	项目实施计划制定
胡馨月	项目经理	6	项目统筹
唐艳如	项目经理	3	项目流程推动
翟师若	产品与解决方案经理	3	项目方案设计
左元钧	业务分析	8	需求理解及方案设计
张娟	业务咨询	7	业务咨询及产品设计
王堃	业务咨询	11	业务咨询及产品设计

朴冬临	数据分析	17	大数据平台数据开发与数据加工计算
郑洪超	数据分析	7	大数据平台数据开发与数据加工计算
邓智群	数据分析	18	大数据平台数据开发与数据加工计算
刘伟	数据挖掘	15	数据模型及算法制定
姚金言	数据开发工程师	6	数据开发
倪默语	数据开发工程师	4	数据开发
陈安琪	运维工程师	6	系统及产品运维
李琪鸣	运维工程师	6	系统及产品运维
刘岩	测试工程师	5	产品测试
张菀桐	测试工程师	4	产品测试

附件二：《验收标准》

一、数据监测范围

本项目人口数据覆盖范围涵盖朝阳区全境、43 个行政街乡、81 个重点区域（项目服务过程中根据实际情况进行调整，重点和个性化区域监测数量不少于 81 个）、50 个重点村、12 个商圈和 24 个旅游景区及临时新增区域。

二、数据验收要求

数据监测指标要求

针对客户实施过程要求，提供监测统计数据服务，数据监测指标包括但不限于以下内容：

1) 实时活跃用户数量

2) 月度居住用户、工作用户数量及流入流出情况，流入人口来源地及流出人口去向地

3) 实时客流量、客流停留时长（商圈、旅游景区）

数据报告指标要求

针对客户实施过程要求，按月提供数据分析报告编写支撑，监测分析指标包括不限于以下内容：

1) 统计口径及指标说明

2) 朝阳区本月街乡及重点区域居住用户、工作用户分析

3) 全区及分街乡居住时段（02 点）和工作时段（14 点）用户数据统计

数据交付时间要求

1) 月度人口统计数据按月提供，次月 10 日前提供当月的月度监测相关数据。

2) 月度人口分析报告按月提供，次月 15 日前提供当月的月度人口洞察报告。

3) 统计数据交付格式为 csv，分析报告交付格式为 word。

数据安全性要求

1) 本项目数据服务内容要求的汇总统计及结果数据。

2) 通过本地数字传输专线，统计结果以 FTP 格式输出。

附件三：

合作伙伴廉政协议

为共同维护商业活动公平竞争秩序，保证双方在业务交往活动中做到诚信、廉洁、高效和共赢，特订本协议如下：

1、双方在合作过程中，自觉遵守国家法律、法规，按照《中华人民共和国反不正当竞争法》、《中华人民共和国招标投标法》、《关于禁止商业贿赂行为的暂行规定》以及其他相关法律规定开展商业交易活动。

2、在供应商资格预审、投标、开标与评标等过程中，做到公平、公正、公开，禁止暗箱操作。甲方监察部门或其授权人员按照规定对招投标项目实施监督，对于乙方有关招投标的投诉和举报，监察部门认真调查，秉公处理，及时回复。

3、双方相关工作人员及其亲属，不得收受对方馈送的现金、贵重物品、有价证券等，不得要求或接受对方为其住房建修、婚丧嫁娶、出国等提供资助，不得介绍亲友从事与双方合作有关的业务活动，不得收受回扣，不得参加影响公正执行公务的高档娱乐、宴请、健身或旅游活动，不得由对方报销应由本人支付的费用。

4、双方相关工作人员不得为谋取私利私下就材料供应、数量变化、材料质量问题处理等进行私下商谈或者达成默契。

5、不做违反商业道德、扰乱正常竞争秩序、有损双方形象的事情，不围标、串标，不泄露双方机密，不排挤其他经营者的公平竞争，不在预决算、招投标和商务报价中弄虚作假或恶意高估冒算。

6、双方人员发现对方人员有受贿或索贿行为以及徇私舞弊、滥用职权、严重渎职等情况时，有义务向对方监察或相应部门举报。

7、乙方若违反廉政协议将被列入廉政黑名单，甲方将视情节轻重，停止该项目或类似项目的合作，一至三年内不再邀请其投标或比选，不再开展新的业务合作。

8、甲方人员若违反廉政协议，造成公司重大损失或恶劣影响的部门和人员，将按照相关规定进行处理，对于涉嫌犯罪的，按照有关程序移送司法机关。

甲方举报渠道　　乙方举报渠道

电话：65099922　　电话：52186699