交通大数据敏捷服务平台建设项目

建设方案

项目单位：上海公共交通卡股份有限公司

2022 年 01 月

第一章 项目概况

1.1 项目名称

交通大数据敏捷服务平台建设项目

1.2 项目单位

上海公共交通卡股份有限公司

1.3 建设内容和目标

1.3.1 建设目标

本项目针对交通行业目前普遍存在的数据采集固化与实时性差，数据供给服务种类固化和时效性差，数据共享数据安全和保护不完善的痛点，通过开发运营交通大数据敏捷服务平台，构建交通卡公司大数据管理能力模型，实现城市交通数据敏捷化归集和治理，敏捷化数据资产开发和服务，以及支持动态数据访问控制的数据脱敏、分级分层等隐私安全访问机制，同时，围绕上海交通卡公司乘客出行画像、渠道权益营销和久事客流宝等数据资产化运营开展应用示范，为用户、城市智慧交通服务运营企业、政府行业治理服务机构、互联网开放平台数字创新运营实体提供数据价值服务。

1.3.2 建设内容

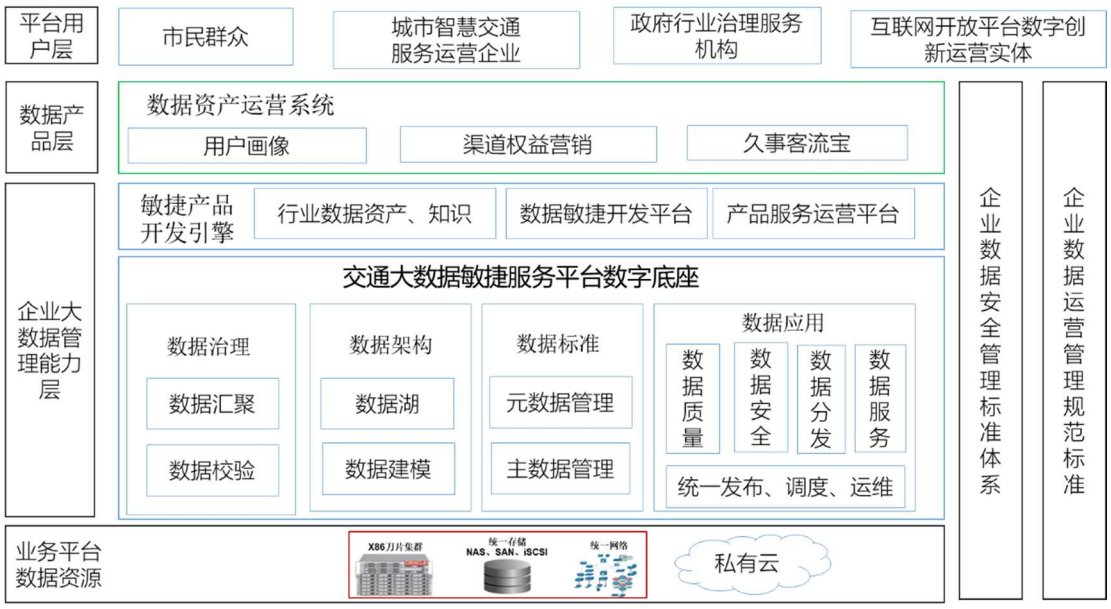

图1. 项目整体逻辑框架图

(1)敏捷化数据归集和数据底座

建设敏捷数据接入、采集和交换平台系统，提供广泛的多种数据接入方式和数据处理配置框架，实现在交通卡公司体系及行业体系的业务系统快速、敏捷的对接，实时、准实时的数据采集和在线实时的标准化处理与保存。达到敏捷化接入、敏捷化归集、实时化处理和标准化数据集合存储。

针对交通卡公司具体的数据结构和业务系统的实际情况，对卡公司的数据体系进行梳理和标准化规范管理，形成卡公司的元数据标准、主数据标准，并通过构建元数据和主数据管理平台辅助数据管理。完成企业数据资源梳理，形成企业数据规范。

建设企业数据底座，实现企业数据湖、核心模型建设和数据应用平台，数据应用系统实现统一调度、运维、发布，数据安全，数据服务、数据分发、数据治理等。

(2)敏捷实时数据服务引擎系统

实时数据服务引擎为敏捷产品开发和服务引擎，完善提升企业数据资产和知识规模，实现按需定制化数据需求快速开发和数据服务快速部署运营，打通和管理数据源、数据集和数据产品之间数据通道，实现实时性数据运营服务，满足需求快速研发实现、快速部署运营和数据高实时性的需求。

（3）研究数据脱敏机制和隐私保护系统

为推动数据共享和隐私安全，推进数据脱敏机制研究和实现，研发隐私安全的沙箱系统，实现数据脱敏和隐私保护，服务内外数据共享和公众化数据系列产品研发。

（4）客流宝系列数据资产化研发和示范运营

推进卡公司乘客出行画像、渠道权益营销和客流宝等数据资产化运营应用示范，

实现对内数字化转型推进和对外数据化产品输出和服务。实现数据共享和示范应用，成为交通行业数据中心和数字化场景应用创新平台。

本工程建成后，为久事集团和上海交通卡公司提供一个规模适度，容量可扩展、功能基本完备的交通大数据敏捷服务平台为卡公司提供在上海城市数字化生活中捕捉服务创新的业务机会。

本项目大数据赋能平台将汇聚上海久事集团交通板块数据资源，联动上海智慧城市丰富多元的数字生活场景，串联城市绿色出行的多维服务和治理需求，立足久事客流宝数据资产，开发出更加丰富的多元多维跨界的数据服务产品，构建城市交通数字服务产品的运营架构。

在数字化转型趋势中，本项目通过把握数字化转型“价值连接、生态共生、当下敏捷”本质特性的数字赋能运营实践，打造支持城市智慧交通数字化服务产品持续续创新、持续迭代的运营体系，助力上海智慧城市公共交通全域数据要素融合的服务平台建设，助力城市公共服务行业“数商”新业态、数字新生态的科技创新。

1.4 总投资估算及来源

本项目计划投资 1716 万元，主要用于上述系统建设所需软硬件采购，应用软件系统研发、运营服务支出等。项目投入的具体明细见后面内容。

1.5 经济及社会效益

1.5.1 社会效益：

（1）行业性的大数据中心汇聚行业数据，有助于支持城市数字化治理，减低社会运营成本。本项目建设了行业性的数据资源集合，汇聚了公交体系广泛的运行数据，形成数据齐全，规范标准，数据质量高，实效性强的城市行业数据集，有助于提高城市行业数字化精准管控，提高城市运营效率和降低城市运行成本。

（2）解决行业数据应用难问题。在交通大数据敏捷服务平台为用户提供数字产品开发和运营环境，实现平台企业客流宝系列数据产品示范性应用。平台提供的数据资产集合、知识集合，数据脱敏和隐私计算服务能力，产品开发和运营服务能力，有效解决数据开发和共享难点痛点，推进行业数据广泛互动融合和共享应用。

（4）支持国家数字化战略。交通大数据敏捷服务平台有效推进交通卡公司和交通行业的数字化转型，创新发展数字化应用场景切实支持和推动国家数字化转型战略和上海市数字化转型需求，推动了社会发展。

1.5.2 经济效益：

（1）降本增效赋能。通过交通大数据敏捷服务平台建设和示范应用，建立企业“可量化、可评估”的降本增效评价指标库，并形成降本增效的反馈机制，精准化实现企业降本增效。

（2）扩大企业经济收入。推进企业用户画像和营销平台应用，扩大用户规模和覆盖服务范围，实现企业市场份额扩大，通过市场规模和份额的增加，提升企业经济收入。通过企业创新发展的客流宝系列数据产品服务，增加新兴业务的收入。

（3）提高企业品牌价值。市场覆盖扩大，新兴技术应用，用户规模增加都有效提升企业品牌价值，在合作市场上提升企业竞争力，获得更广泛的市场影响力和经济收入。

1.6 结论与建议

通过交通大数据敏捷服务平台建设和运营，主动融合新兴技术和数字化转型洪流，转变成为用户和数据驱动性的企业经营格局，稳固和扩大公司行业地位。

项目的实施和运营将成为企业在数字化转型时代的基础性设施平台和创新动力之源，成为企业持续发展的关键设施。

第二章 现状

2.1 项目单位概况

1.单位职责、内设及下属机构、人员编制和业务情况

上海公共交通卡股份有限公司成立于 1999 年 5 月 25 日，是一家主营业务为本市交通卡、沪通卡、旅游卡系统建设、运营、结算和公共交通信息服务的公共服务类企业。上海公共交通卡系统于 1999 年 12 月 27 日开通试运行，在市委、市政府领导的关心支持下，在市交通委和久事集团的直接指导下，经过二十年的发展，公司承担的交通卡、沪通卡、旅游卡支付服务现已实现在本市交通领域应用的全覆盖，并向停车场、充电桩等交通相关领域以及公共交通信息服务等领域延伸。随着手机交通卡的陆续推出、“上海公共交通乘车码”的上线运行，公司的服务不断拓展新的应用场景，全面进入“互联网+”新时代，努力成为推进上海“智慧出行”的先行者。

公司是一家按照现代企业制度组建的股份制企业，截至 2019 年 12 月底，由上海久事（集团）有限公司（占比 52.63%）等 8 家单位共同出资组成，总资产 75.85 亿元，净资产 8.98 亿元，财务状况良好。

公司按照现代企业制度，建立了股东会、董事会、监事会和公司经营管理班子。内设党群工作部、组织人事部、行政办公室、计划财务部、销售服务部、技术开发部、研究发展部、清算管理部、运营服务部、ETC 客服部等 10 个部门。目前已投资组建三家下属企业，分别为销售服务公司、都市旅游卡公司和久誉软件公司。

2.近三年来职责和业务的调整情况及未来发展趋势

交通卡公司正在推进实施企业十四五规划和数字化转型规划，加速推动用户化、在线化、数据化的企业业务和系统升级换代，加快二维码应用范围，落实国家数字货币在公共交通领域的应用，推进久事集团的大数据中心建设和运营，推进数据产品研

发和运营销售。

通过本项目的实施，打造交通大数据敏捷服务平台，汇集行业数据，形成企业大数据模型和行业数据资产和知识建设，有效推动企业数字化转型，助力企业在交通体系的数字化服务产品持续续创新、持续迭代的发展，助力上海智慧城市公共交通全域数据要素融合的服务平台建设，助力城市公共服务行业“数商”新业态、数字新生态的科技创新，为上海的创新驱动、转型发展战略做出自己的贡献。

3.拟建项目与项目单位职责、业务的关系

拟建项目，既满足卡公司既有业务的数字化转型规划中数据的强烈需求，同时也是国家战略发展需求和企业自身推进新兴业务领域要求，与公司发展方向高度融合。公司经过多年的发展和建设，具备完整技术储备和相应的实施团队、拥有丰富的项目经验，完全能够实施本项目的建设任务。

2.2 数字化转型现状

1.单位数字化转型的整体框架规划或设想

交通卡公司在数字化转型战略上提出“一中心、三个数字化基础架构”数字化转型规划，助力卡公司的数字化转型。

1 个云混合计算中心：依托卡公司自有同城异地双机房，利用外部公有云构建卡公司的混合云计算中心。

3 个数字化：一是面向用户为中心，规划建设公司的数字化产品与服务体系；二是面向运营和决策，规划建设运营数字化运营管理和决策平台；三是数字化设施建设，推进数字化设施升级和数字化管理平台，以各类运营服务设备为管理对象，建立实时监控机制，确保交易环境、运营环境的长期安全稳定。

1 个大数据平台：构建企业的大数据平台，实现企业和行业性数据归集、数据治理、AI 平台、数据管理和数据应用拓展等能力，并形成行业性数据资产集，服务公司产品、管理的数字化战略和发展数据服务战略。

2.现有的信息化系统应用及运维管理制度

公司作为专注交通和支付结算的软件和信息服务业的企业，经历 20 多年的信息化系统建设和运营，在系统管理和项目管理上有多年的应用和运营经验，以及相应的完整完备的管理制度。企业信息化系统严格按照行业要求，严格遵循和达到支付企业信息化系统技术指引要求，全面实现国家安全等保三级规范要求，并通过审查。企业建有全面的运行维护系统和专业的运维团队，实现 7*24 运行制度。

3.现有应用系统的情况

企业目前拥有交通卡综合业务系统平台，平台承担上海城市公共交通卡的业务运营，对接地铁票务系统、公交票务平台，其他公交行业，银行和第三方支付平台，交通部平台等等，平台处理能力为 2500 万笔/日。

ETC 发行结算平台，承担 ETC 的发行、充资服务、资金清算、客户服务和运营平台。对接交通部平台、公路收费平台、银行和第三方支付平台等等，平台交易能力 300 万笔/日。

上海旅游卡平台，平台承担上海旅游卡的发行、结算、渠道接入和终端系统、客户服务和运营平台系统。对接银行和第三方支付系统、各个接入点票务系统等，平台交易能力 200 万笔/日。

交通卡 APP 系统，作为公司融合移动互联时代需求，建设新一代的线上用户中心、业务和产品承载平台的信息化系统，初步实现了用户线上服务、新业务开展等目

标，并逐步成为公司后续核心业务平台。对接各个核心业务平台，线上支付平台等等。

数据仓库系统，建有企业数据仓库，聚集了数据进行了初步的挖掘，并为各相关政府部门和研究机构提供了信息。

其他信息化系统包括OA、ERP、财务系统、客服系统，运维平台等等。

4.拟建项目与已有系统的关系

项目建设的交通大数据敏捷服务平台相对独立于既有的信息化系统，同时也推动与服务既有信息化系统数字化转型。平台作为企业数据战略核心，实现公司内部数据集成和久事集团体系内数据接入服务，实现企业数据治理，形成企业数据中心，实现数据在线化、数据生产化、数据应用化、数据资产化和知识化，提升企业数据治理体系能级；推进企业数据治理和系统融合，为各个信息化系统提供数据和数据服务，加速企业数字化产品服务转型，企业信息化架构升级和服务升级，牵引公司业务的数字化升级。

5.现有网络、设备以及其他信息资源情况

公司拥有自有机房2个，信息化系统全面实现冗余设计开发和部署，企业核心业务系统实现同城异地应用级的灾备在线部署运行，支持企业信息化平台7X24的不间断业务运行。

公司机房拥有65个机柜，大中型交换机40台，小型机HP服务器20台，x86服务器200多台，存储系统12套。

第三章　项目的需求分析

3.1　项目建设的背景

在以数据、IT 技术的快速发展背景下，数字化已经成为全球性的发展趋势，中共中央多次提出加速发展数字经济等要求，上海市委全会确立了“全面推进城市数字化转型”的总体目标。

上海市委、市政府 2020 年年底公布《关于全面推进上海城市数字化转型的意见》，意见指出，要坚持整体性转变，推动“经济、生活、治理”全面数字化转型；坚持全方位赋能，构建数据驱动的数字城市基本框架；坚持革命性重塑，引导全社会共建共治共享数字城市。

《上海市全面推进城市数字化转型“十四五”规划》全面推进城市数字化转型规划。规划提出目前在数字化形势下，数字化推动科技创新进步和经济社会发展；**数据要素**对价值创造的乘数效应全面激发，不断催生新产业、新业态、新模式,带来了生活领域的革命性变革。

推进上海生活数字化转型，构建高品质数字生活行动方案（2021—2023 年）中提出**以“人”为根本，以“数据”为核心**，打造数字赋能生态和打造全流程场景体系。

上海交通卡公司作为一个城市公共交通保障型国资企业，在“人民城市人民建、人民城市为人民”的理念指引下，全面贯彻上海市城市数字化转型精神，规划自身企业的数字化转型规划，融入上海数字化转型体系，实现数据治理、数字归集、数据管理、数据服务和数据应用，支持企业自身数字化转型、行业数字转型对数据服务和互联互通的需求。

3.2　项目建设的依据

- 《中华人民共和国国民经济和社会发展第十四个五年规划和 2035 年远景目

标纲要》

- 《关于全面推进上海城市数字化转型的意见》
- 《上海市全面推进城市数字化转型“十四五”规划》
- 《推进上海经济数字化转型 赋能高质量发展行动方案（2021-2023 年）》
- 《推进上海生活数字化转型 构建高品质数字生活行动方案（2021—2023 年）》
- 《上海公共交通卡股份有限公司十四五发展规划报告》
- 《上海公共交通卡数据中台一期项目立项报告》

3.3 采用信息系统实现业务需要的需求分析

3.3.1.业务现状、存在的问题

(1)交通行业覆盖面广，数据规模大，数据利用迫切需要提升。交通卡公司和久事集团作为公共交通行业核心基础性的保障服务性公司，服务遍及全市范围的公交交通和全市市民，覆盖面广，运营过程中产生的数据全面和庞大，但是数据目前都存在于各个业务系统，数据标准不一，形成数据孤岛，不能有效的发挥数据价值。迫切需要将这些数据归集，融合加工，形成一个上海行业性的基础共享数据中心，助力交通卡公司和久事公数字化转型，服务公众、运营企业、城市管理和治理。

(2)建立数据仓库和初步数据服务能力，数据接入困难和数据服务固化。公司在 2008 年建设公司数据仓库系统，经过多年的运营，系统比较陈旧，接入数据内容少，数据模型一直不能同步提升，仅仅提供固化的一些指标，新数据需求难以快速满足，并提供不能有效运营能力。平台已经难以应对企业自身和外部对数据新需求。

同时体系内的信息化系统相对封闭，业务依赖性高，进行系统对接较为困难，导致数据归集接入困难，数据归集范围少，数据归集时效性差。迫切需要一套数据接入、采集和处理系统来实现既有系统和行业外数据和系统的敏捷接入、数据敏捷采集、实

时在线处理和数据标准化存储管理。

(3)数据化转型迫切，缺乏数据产品和服务能力。卡公司积极规划和布局企业数字化转型发展，布局智慧交通和城市生活领域，加快企业既有系统数字化升级和改进，深化数据在产品服务、管理决策和设施设备的应用，推进产品服务数字化、管理决策数字化和智能化设施装备升级，加速推进公司数据产品系统研发，创新数字化场景应用，以点带面推动企业数字化转型和服务升级。

(4)信息系统安全基本完备，尚需加强敏感数据和共享数据保护机制。交通卡公司完整安全防护能力体系已经涵盖基础网络安全、主机应用安全、WEB 深度安全、安全接入、负载均衡和安全运维审计等多个层面，包括传统安全产品防火墙、IPS、堡垒机等，并增加 VPN 安全接入、服务器负载均衡、DDOS 防护等，以及 WAF 应用防火墙、网页防篡改、网站安全监测等 WEB 应用安全服务能力。在大数据敏捷服务下，对数据安全提出进一步的需求，有效进行数据资源分类分级管控，敏感数据脱敏处理，隐私计算需求等保障数据有序流动和安全可控。

3.3.2.业务对信息系统的具体需求

数据在数字化转型发展中，成为核心的发展要素，如何有效获取数据，管理数据，应用数据成为数字化转型核心内容，以数据赋能推动企业流程改造和来创新发展新业务、新产品、新生态。交通大数据敏捷服务平台是响应数据作为数字化转型的核心要素基础核心平台，平台实现数据归集、标准建设、数据展现、数据产品研发、数据产品运营和数据安全管控需求。

项目实施中业务具体需求落在 4 个主要内容方面。

（1）数据敏捷归集和数据底座

构建敏捷的数据归集系统，平台能够支持对接各自架构的业务平台，提供ETL数据库、接口、消息队列、文件等等方式的数据对接模式，实现和数据源头的业务平台进行快速敏捷对接。

对采集的数据提供统一处理框架，便于各型数据进行标准化处理，数据采集效率和时效性实现准实时性，采集的数据实现在线实时处理，形成大数据ODS数据集合。

针对交通卡公司具体的数据结构和业务系统的实际情况，对卡公司的数据体系进行梳理和标准化规范管理，形成卡公司的元数据标准、主数据标准，并通过构建元数据和主数据管理平台辅助数据管理。完成企业数据资源梳理，形成企业数据规范。

建设企业数据架构，实现企业数据湖和核心模型建设，平台数据应用系统实现统一调度、运维、发布，数据安全，数据服务、数据分发、数据治理等。

（2）研发敏捷化数据产品开发和服务网关引擎

敏捷数据产品研发引擎上可以实现对数据资产和知识快速数据源配置、数据处理流程和服务输出配置，对接服务网关引擎实现数据产品的快速研发和部署交付。

数据服务网关提供数据产品的上线、发布和服务管理功能，支持插件化，方便开发人员自定义组件，支持横向扩展，高性能，可动态配置更新生效，不需要重启网关。

（3）数据脱敏机制和数据沙箱系统研发

研究数据脱敏的机制和方式，实现试点应用。在数据输出时，由脱敏系统对敏感信息进行脱敏处理，脱敏系统实现脱敏原则和规则定义与实现。

研究数据沙箱、孵化环境和部署算法算力资源库，实现数据隐私计算，避免

数据外泄。

（4）久事客流宝系列数据产品研发和运营

实现既有数据参考的数据服务功能集合。

研发上海交通客流宝系列数据产品，满足市场需求和城市运营对数据的需求。

3.4 业务流程分析

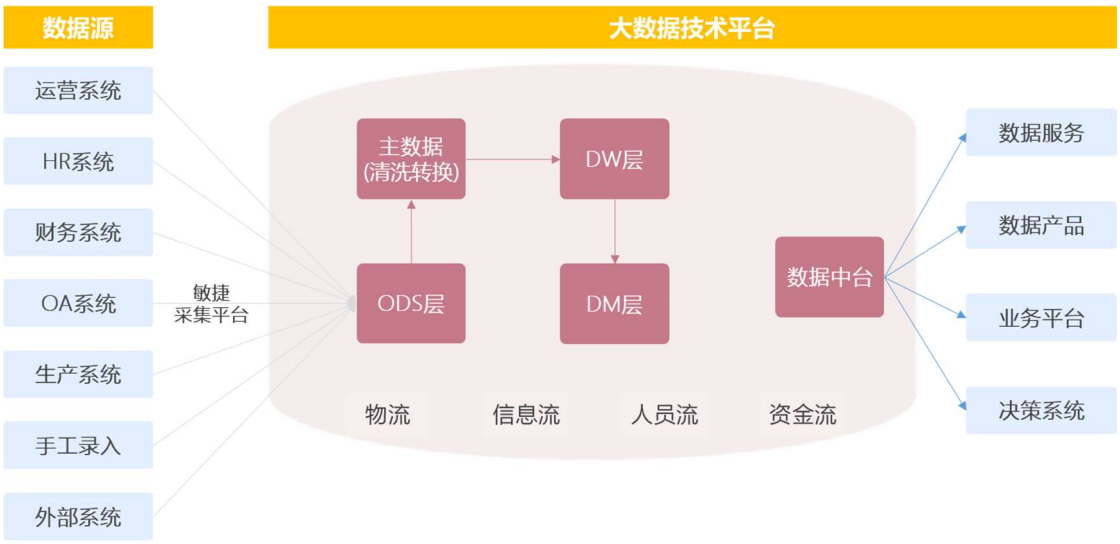

图2.　交通大数据敏捷服务平台数据流图

交通大数据敏捷服务平台数据流：

1．对接数据源，通过各自方式获取批量、实时、在线数据，进入大数据平台

2．大数据技术底座平台进行相应处理，并生产出数据集，形成数据资产

3、通过接口，数据底座等提供数据服务，支持数据产品化、支持业务数据化转型。

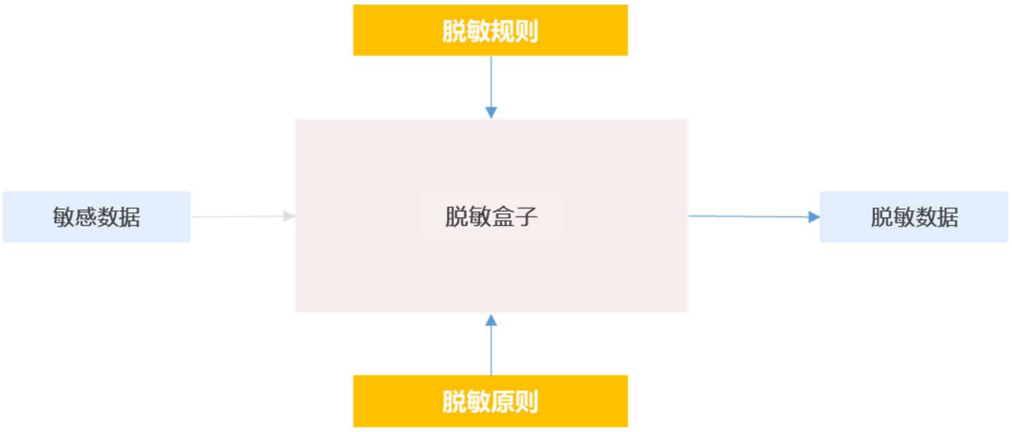

图3.　脱敏流程

数据脱敏流程：

1、定义脱敏规则和脱敏原则，输入脱敏盒子

2、敏感数据进入脱敏盒子，脱敏盒子完成数据脱敏

3、脱敏数据从盒子输出

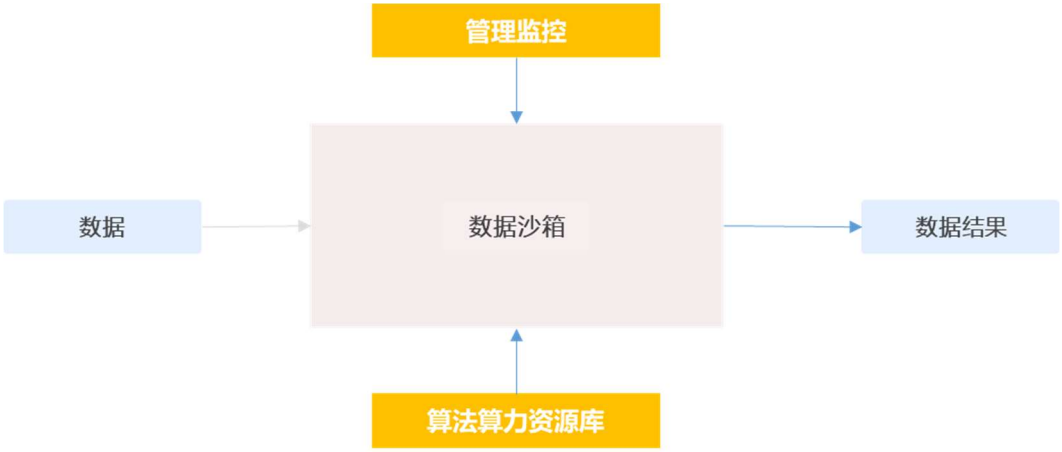

图4.　沙箱模型

数据沙箱流程：

1、数据沙箱创建

2、管理平台向沙箱配置处理系统

3、在数据平台按需配置数据，并导入沙箱

4、沙箱完成数据处理，并受控输出处理结果

5、数据沙箱销毁

3.5 功能需求分析

（1）交通大数据敏捷服务平台功能

1、完成既有系统业务梳理，编制主数据管理标准和元数据管理标准，并形成相应的管理平台，提出既有信息系统非标准化的改进方案。

2、一个高可靠、高稳定、高安全、高性能、高可扩展、高易用性的基础数据存储、计算、分析和管理平台。

3、敏捷数据采集平台，能够支持多业务明天的数据对接，实时接收行内各类系统所产生的数据，实现实时化数据处理并存入大数据平台存储层，完成数据清洗、数据转换、数据格式统一、数据入库等操作。

4、平台进行建模，以完成相关业务模型操作和计算，能够对数据进行定期批量处理，形成各个数据集、主题集、中间数据集等，为后期综合分析准备数据。

5、提供数据敏捷研发平台，能够支持在交通大数据敏捷服务平台上进行快速的数据应用开发。需要支持数据源管理、数据处理流程处理和数据输出功能。

6、数据服务网关，支持快速的数据服务发布、服务管理调度、服务供给和鉴权功能。

7、数据脱敏功能和数据沙箱功能。实现数据脱敏管控和数据沙箱功能管理。

（2）客流宝数据产品开发和数字化应用场景建设

1、研发上海交通客流宝系列数据产品，满足市场需求和城市运营对数据的需求

2、推进基于大数据敏捷平台数据权益服务应用数字化场景建设。

3.6 数据分析

3.6.1.数据流程和属性分析

交通大数据敏捷服务平台涉及的数据主要为业务过程中产生的交易数据，日志数据，中间埋点数据，加工中间数据和结果数据。这些数据分布于各个业务系统，数据规模较大。各个业务系统本身已经具备完整的数据存储、备份和应用系统容灾设计和部署，不在本项目考虑，本项目考虑主要对这些数据通过各自方式获取并存储到大数据技术平台，并进行大数据平台各个的存储、管理、加工和分析处理。

3.6.2.数据量分析

大数据敏捷服务平台数据需求：

大数据平台采集各个环节的业务过程数据、运行状态、中间数据、交易数据，日志数据，中间埋点数据，加工中间数据和结果数据，各个业务过程由多个功能流程构成，每个功能流程都产生相应的数据，系统日交易量1800万笔，按照每天3600万条过程数据，加上大数据存储和管理的倍乘数为3，每天需要1.08亿条数据，每条记录按照0.8k计算。

日数据存储量需求：0.8k*108,000,000/1024/1024=85G。

项目上线1年期间存储量：85G*365天=32T。

加上预留，大数据敏捷服务平台总计预计需要存储大约36T。

1、数据ETL及建模时间：3000万条数据1小时内。

2、大数据平台并发作业量：同一时间并发30个作业。

3、查询响应时间：1亿条数据3秒内。

4、采用日增量备份，备份数据存放磁带备份。

5、提供磁带备份功能，实施增量备份。

6、7×24小时高可用。

7、大数据敏捷服务平台正常备份/恢复：控制在 7 小时内。

3.7 安全需求分析

本项目安全需求达到国家三级安全等保规范。

1．数据安全管理

构建大数据安全综合防御体系，建立覆盖数据收集、传输、存储、处理、共享、销毁全生命周期的安全防护，推动数据脱敏、审计、防泄露、追踪溯源等技术手段在大数据环境下的增强应用，保障数据有序流动和安全可控。

2. 安全防护体系

安全防护体系涵盖基础网络安全、主机应用安全、WEB 深度安全、安全接入、负载均衡和安全运维审计等。

3．业务连续性

业务连续性需求是保障平台 7X24 小时运行。

3.8 设备需求分析

系统基于云服务的基础架构进行服务器和存储系统布署，云计算的特点决定了基于云计算的平台，可以满足超大规模的应用使用，数据多副本容错、计算节点同构可互换等措施来保障服务的高可靠性，可以动态伸缩、满足应用和用户规划增长的需要，具有极高的性价比。

计算节点：计算资源服务器是为云平台提供计算资源，云计算节点具有较强的计算和内存系统,

存储节点：存储资源节点为云平台提供存储资源，实现存储资源的动态管理和扩展能力。为云平台内系统提供存储服务。

网络设备：提供云平台网络和网络安全防护，形成网络拓扑。

3.9 软件需求分析

本次开发主要通过分布式计算框架,按照云原生架构进行设计部署。

主要软件包括开发工具、操作系统、关系型数据库,数据仓库公交、大数据引擎、标准 SQL 开发工具、数据 ETL 工具、元数据管理工具、工作流工具、可视化建模工具、数据挖掘建模工具、日志存储和分析工具等等。

第四章　项目建设方案

4.1　建设目标

本项目立足于构建集团型企业数据资源中心和数据底座的建设规划，建成一个高可靠、高稳定、高安全、高性能、高可扩展、高易用性的基础数据存储、计算、分析、管理和服务平台，成为一个企业数字化转型的数据能力平台，支持交通行业数字化转型。

1、完成既有系统业务数据梳理，编制大数据敏捷开发平台数据标准和数据管理制度。

2、建成一个高可靠、高稳定、高安全、高性能、高可扩展、高易用性的基础数据存储、计算、分析、管理和服务平台。

3、完成行业数据模型构建和数据资产集，研发敏捷数据采集系统实现接各个业务平台，实现数据采集、处理、存储和管理，形成数据集、数据资产化和知识化。

4、研发数据开发系统和数据服务网关。支持快速数据开发和发布，提供服务管理和运营能力。研究部署数据脱敏机制系统，提供统一化数据脱敏机制和服务。开发数据沙箱服务平台，提供数据沙箱隐私计算服务能力。

5、实现至少 1 个典型交通行业数据资产数字化场景应用。

6、实现至少 1 个客流宝系列数据产品研发、部署和上线运行。

4.2　总体架构

4.2.1 业务架构

平台的建设和运维是伴随着业务中台的规划建设及运营不断的完善和演进，同时数据底座为各业务单元提供与之对应的数据集市。

业务单元数据集市的构建时，根据其业务运营指标需求出发，分析该业务单元所产生的主数据以及需要由其他业务单元产生的主数据，形成主数据集。

围绕着这些业务单元的主数据，从业务流程维度出发，建立业务应用的主题模型，进而构建该业务单元的数据仓库。数据仓库面向主题，结合交通卡现有业务设计为当事人（用户、商户、客户）、产品、渠道、财务、营销等主题进行划分。

由以上构建的主数据集和数据仓库进而构建出业务单元相应的数据集市。

针对业务单元的数据集市、数据仓库，结合主数据以及相应的交易数据进行元数据管理。

综上，所形成的各个数据仓库、数据集市等共同构成了交通卡公司数据底座的数据湖，提供统一对内数据存储和对外服务。

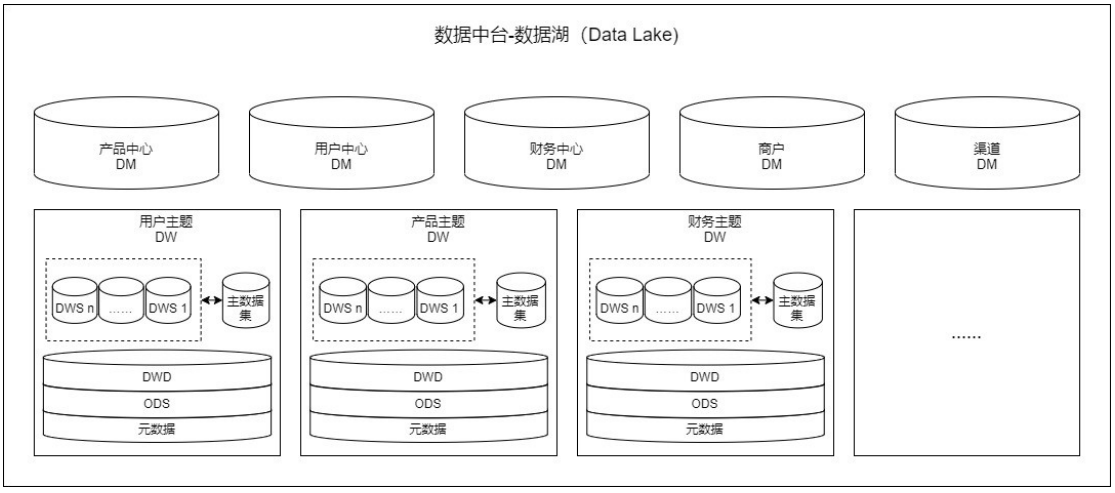

图5.　业务架构框图

数据模型设计

结合业务架构设计，在数据仓库层面向主题进行主题模型的设计。数据模型是关于城市通卡行业的完整信息的参考模型。其目标是设计出一个兼具效益与弹性、能适应未来业务及技术变化的数据模型以支撑行业各类业务系统数据模型的设计，是一

个逐步深化并持续改进的过程。

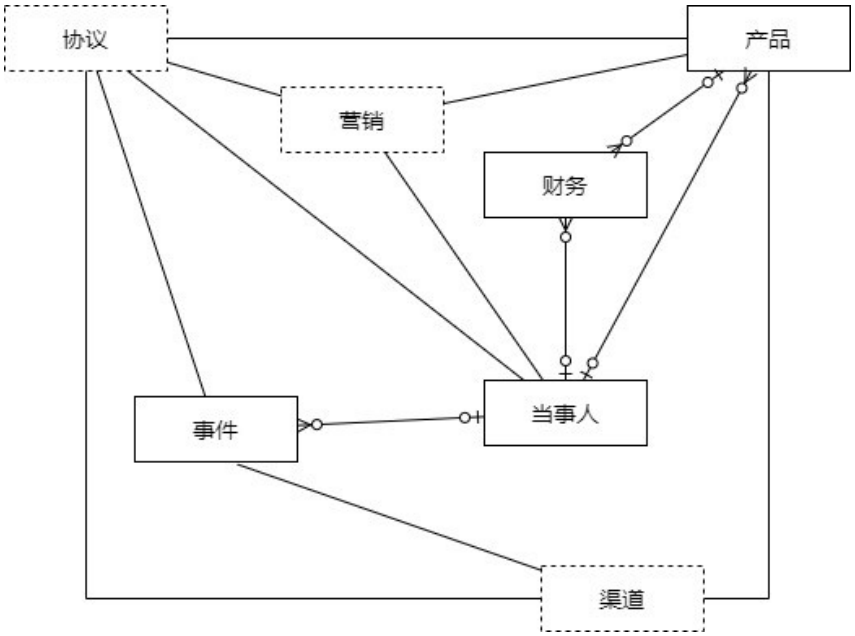

图6.　交通卡数据底座数据模型

交通卡公司数据底座的数据模型面向业务主题分为当事人、产品、财务、渠道、事件、协议、营销等主题域。

4.3.2 技术架构

在确定业务架构后，数据底座项目根据业务架构的规划结果进行技术选型，构建数据底座的工具平台，以支撑业务架构。数据底座的技术选型包括数据底座（包括数据存储、计算资源、网络资源等）、数据湖、数据总线（包括数据汇聚、消息队列、调度服务等）、大数据处理引擎、数据治理等并根据选型确定物理架构和网络架构。

- 数据底座

数据底座随着建设的演进，其对于数据存储、计算、网络等资源具有基础资源设施的可扩展性要求较高，因此目前主流的数据底座的数据底座均采用云平台架构。

对于交通卡公司的数据底座而言，考虑到目前集团已经建设了久事云平台，因此也具备了数据底座上云的条件，并且能够充分发挥基础资源的可扩展性及按需分配

的优势。所以数据底座的数据底座选型采用久事云平台。

- 数据湖

在确定了数据底座的选型后，需要进行数据湖的选型。交通卡数据底座未来不仅需要提供离线数据分析服务，还需要随着业务的发展提供实时、高速的数据流分析服务、数据挖掘服务等，并且存储的数据类型也包含了结构化数据、非结构化数据和半结构化数据等，因此需要建设一个支持原始格式存储的数据湖。

根据交通卡数据现状分析，交通卡核心业务的信息化数据存储均为关系型数据库，因此交通卡数据湖首先需要支持关系型数据的存储。

考虑到数据底座的数据湖不仅需要支持数据事务，更多需要提供大数据分析服务，因此在关系型数据存储选型采用的是 PostgreSQL。PostgreSQL 是目前最先进的开源关系型数据库，其提供丰富的数据类型更合适作为 OLTP/ OLAP 系统的数据库服务。同时 PostgreSQL 支持外部数据封装，能够更好地对接基于 Hadoop 技术栈的大数据平台。

由于交通卡数据底座未来的接入数据还包括非结构化数据和半结构化数据，为了更好地满足数据存储的扩展能力。因此数据湖的分布式数据存储采用 HDFS，通过分布式存储提供数据底座的数据存储水平扩展能力。考虑到久事云上已配有华为大数据平台产品，因此采用 Hadoop 平台，以 HIVE 作为数据底座的数仓存储。同时利用 Hadoop 技术通过 Spark 和 MapReduce 计算框架提高大数据量情况下的处理性能，利用列式存储在数据加工方面的优势，提供大数据分析、挖掘、处理能力。

- 数据总线

交通卡公司数据底座的数据总线模块主要为数据底座提供数据交互的服务能力，

主要分为离线数据交换和实时数据交换两部分。

离线数据交换主要采用 ETL 方式，通过与数据源的数据库系统进行对接，在不影响系统业务的前提下进行数据批量采集。交通卡公司数据底座离线数据交换模块选型采用 DataX 工具。DataX 是阿里巴巴集团内被广泛使用的开源离线数据同步工具，其支持目前市场上主流 RDBMS 数据库、NoSQL 数据库、时序数据库、无结构化存储等各类数据库系统及存储系统，可以高效地实现各种异构数据源之间高效的数据同步。

实时数据交换主要采用Rest API方式,接口服务采用SpringCloud微服务架构，实现接口服务的高并发能力和高可用性。同时通过消息队列系统提供数据服务的削峰和异步处理能力。消息队列系统采用 Apache Kafka，Kafka 为目前大数据处理主流消息队列系统，其具有高吞吐、高可用、实时性、可扩展等特点。

数据总线的离线数据交换由统一的调度系统进行管理，在本项目中调度系统采用 Apache Airflow，通过 Airflow 实现数据底座中的各种任务及任务之间的依赖和执行调度。

- 大数据处理引擎

大数据处理能力是数据底座的重要服务能力之一，其通过与数据湖相结合提供数据服务。交通卡公司数据底座的大数据处理主要采用 HQL(MapReduce)和 SparkSQL。根据 MapReduce 和 Spark 两种引擎的特点不同，在不同的数据处理场景下选择不同的引擎。

HQL 主要用于处理数据底座中复杂的批量大数据处理。例如数据从 ODS 层向 DWD 层进行 ETL 时，由于数据量较大，若使用 Spark 年前则需要大量内存支持，考虑到其对数据处理的时效性要求相对较低，因此选用 MapReduce 进行处理。

SparkSQL 主要用于处理数据底座中实时数据流的数据处理以及单次数据量有限的数据处理场景。例如数据从 DWD 层向 DWS、DM 层进行汇总计算时，以及未来实时接入数据底座并需进行实时计算的实时数仓服务均采用 Spark 作为计算引擎。

- 技术架构设计

基于以上技术选型，数据底座的技术架构示意图如下：

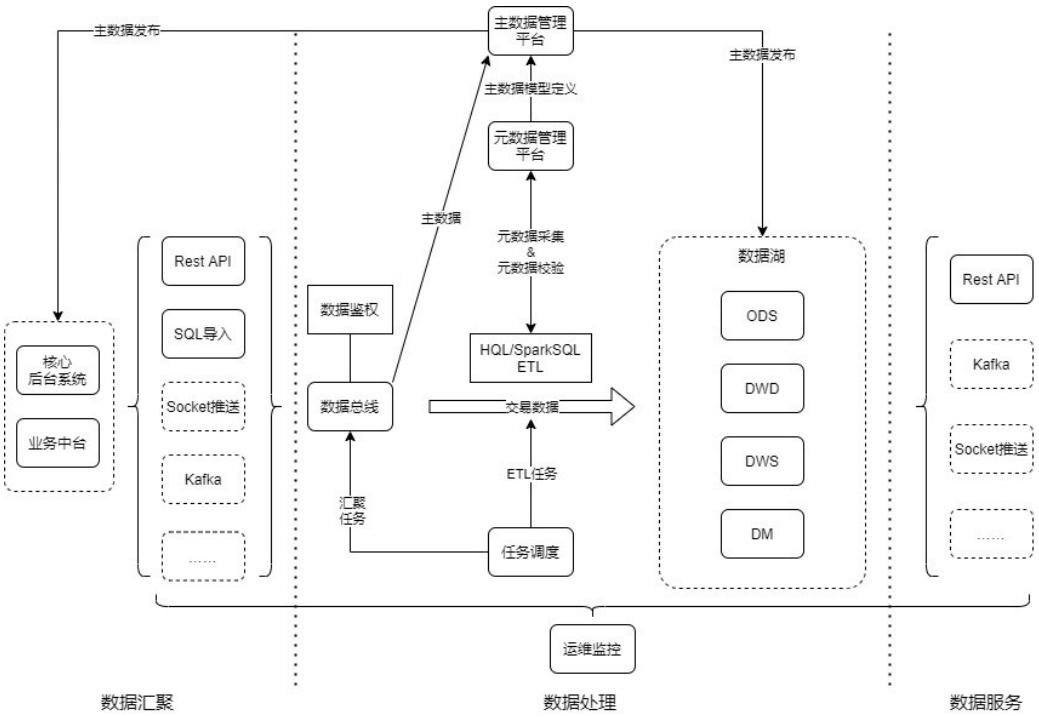

图7.　数据底座技术架构图

交通卡数据底座通过数据总线提供数据交换服务，具体交换方式包括 Rest API、SQL 导入，未来可根据实际需求提供 Socket 接口推送、kafka 消息队列系统等方式，以丰富的接口方式提供数据汇聚服务。

数据总线服务由数据底座任务调度系统统一进行任务管理，并在进行数据交换过程中对接入方进行身份鉴权。

主数据管理系统通过数据总线将主数据从核心后台系统/业务中台中进行抽取，并根据元数据管理平台中定义的主数据模型进行转换和存储。遵循主数据管理流程

向数据底座及核心后台系统/业务中台进行主数据版本发布。

数据湖采用分层架构进行数据的存储和服务，即数据操作层（ODS）、明细事实层（DWD）、数据汇总层（DWS）、数据集市层（DM），各层均采用单独的 schema 进行分层管理。

ODS 层作为数据操作层是数据汇聚存储的第一层。接入的数据根据类型分为主数据和伴随业务产生的交易数据。为确保数据接入的完整性，在 ODS 层只进行简单的数据校验清洗，尽可能使数据能够汇聚至数据底座中。ODS 层的数据根据数据的变化特点分别存入数据湖的关系型数据库或 HDFS 中。

DWD 层的数据由 ODS 层的数据进行清洗得到，数据清洗采用 SQL/HQL 实现，清洗的规则由元数据管理平台定义，数据存储采用 HDFS。在从 ODS 层向 DWD 层输出时，主要针对交易数据的标准化。考虑到主数据在主数据管理平台中已经根据主数据模型定义形成标准化，因此对于主数据 DWD 层不做处理，在做数据汇总时直接通过 ODS 层获取主数据信息进行数据处理。

DWS 层数据由 DWD 层数据汇总得到，该过程采用 SparkSQL 实现数据处理，并在处理过程中向元数据管理系统输出数据血缘信息。

DM 层为数据集市，该层以部门或产品为维度单独建立数据集市，DM 层数据由 DWS 层、DWD 层处理得到，采用 SparkSQL 作为数据处理引擎，并在处理过程中向元数据管理系统输出数据血缘信息。

DWS 和 DM 层的数据根据数据热度情况选择关系型数据库存储或 HDFS 存储。

数据底座的数据应用服务采用 SpringCloud 微服务框架提供数据接口服务或消息队列服务，提供数据服务的高性能和高可用性。

4.3.3 物理架构

基于上述技术架构设计，首先基于云平台的计算资源、网络资源和存储资源构建数据底座的数据底座。

整个平台云网络环境分为四大层，外联接入层、核心交换层、生产服务器层、大数据平台层。整体而言，网络环境做到了简介明了，利于后期维护管理。

外联接入层：主要分为外联网接入和内联网接入。由于后期必定存在互联网访问需求。在外联接入部分，配置高性能虚拟防火墙和虚拟负载均衡作为可靠支点，来应对外部数据高流量访问需求。同时从可靠性来讲，各节点间以冗余的方式存在。默认情况下所有外部访问请求为拒绝，内部服务器访问外网需开通访问策略。而目前内联接入部分，主要考虑的是与交通卡总部系统的访问需求，实际使用中会更多考虑安全访问因素，因此除与外联网访问需求雷同外，还需考虑交通卡公司数据接入层的访问控制策略。

核心交换层：主要配置高性能虚拟交换机，为系统中所有接入节点，做强力支撑，保障后台数据交换可靠、安全、高速。同时区分不同的生产网段，网段间的访问授权等。

生产服务器层：此层主要是所有生产服务器节点，全部虚拟云主机在此归类。实际生产环境下，我们区分为：数据交换组、应用安全组、数据库组、运维管理组。各组之间默认访问为拒绝，实际使用中，根据访问需求开通访问策略，达到安全访问的目的。所有新节点的增加，根据实际使用情况灵活申请，灵活增加。

大数据平台层：此平台直接串联在数据底座网络环境中的子系统，主要的任务是做大数据分析。从独立子系统配置看，高性能虚拟防火墙、高性能虚拟交换机、高性

能虚拟主机根据实际需求情况申请购买。实际使用中，大数据平台核心交换网络划分单独网络接口与大数据平台直接相连，数据之间的访问通过虚拟防火墙做策略配置。

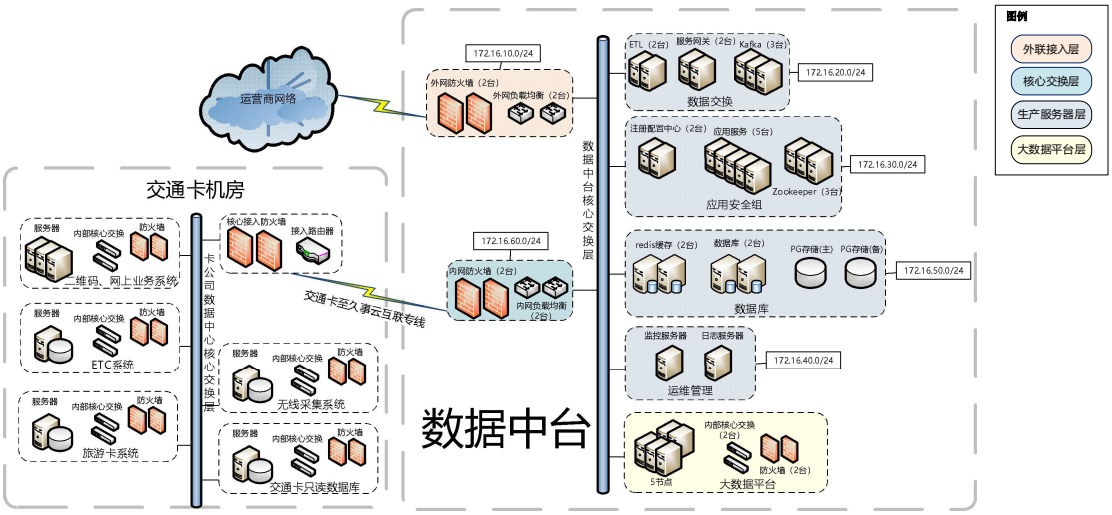

图8.　　平台物理架构图

为确保平台的可靠性，图中的集群内云服务器的宿主均为不同的物理服务器，并且物理服务器在不同的机架，以确保平台应用业务不受单台服务器或单个机架的故障影响。

平台内部根据业务情况定义 4 个安全组，依次为数据交换安全组、应用安全组、运维管理安全组、数据库安全组。根据不同的安全组定义不同，来引用不同的 IP 地址组，来实现可控的网络访问策略。默认情况下 4 组安全组相互拒绝访问，根据实际应用资源访问需求情况，逐步开通访问策略。

数据交换安全组由以下节点组成：

1）2 个 ETL 节点，用于部署 ETL 应用服务，以分布式并行形式提供从交通卡各业务系统抽取数据。

2）2 个服务网关节点，部署组成高可用服务网关节点，作为外部接口统一入口。

3）3 个对外 kafka 消息队列节点，部署组成高可用分布式消息队列系统，用于

提高外部数据的接入性能及流量削峰。

应用安全组由以下节点组成：

1）2 个注册配置中心节点，用于数据服务注册和配置管理。

2）5 组应用服务节点（应用服务及调度服务），用于部署主数据管理系统、元数据管理系统、调度系统及数据服务应用。

3）3 个 zookeeper 节点组成，用于分布式应用协调管理。

运维管理安全组由以下节点组成：

1）1 个监控服务节点，用于应用服务监控管理。

2）1 个日志服务节点，用于平台日志的统一管理。

数据库安全组由以下节点组成：

1）2 个 redis 缓存节点，用于部署平台缓存应用。

2）一主一备数据库节点，用于部署平台关系型数据库。

关于系统安全性设计，平台架构设计采用分阶段实施，初期部署和实施生产环境，预留灾备环境。后期灾备环境应实行同城异地灾备，灾备中心架构和主服务中心基本一致，两个中心通过专线实行互联互通。

4.3 应用系统

4.3.1 数据治理

1、 数据汇聚

数据汇聚交换系统的应用框架如下图所示：

图9.　数据汇聚交换系统的应用框架图

（1）端到端数据采集前置

代理前置系统是端到端数据采集部分，是汇集交换系统相连接的桥梁，是与部门内部业务系统及业务信息库相隔离的“堡垒”。其负责在业务系统中采集数据，并通过接入层接入数据汇聚交换系统，代理前置系统可以通过文件，接口，数据库拷贝，数据库查询等等各种适应业务环境的手段采集业务系统内数据，并通过安全规范对数据进行安全处理，完成数据采集，组包，通过接入层经数据传输至汇聚交换系统。

前置代理系统通过接入层接收数据交换系统转发的数据，进行安全处理，数据解析后将数据或者指令转发到业务应用系统。

接入层

接入层为各数据源提供接入服务。接入层提供接入点管理，接入数据源管理，提供通讯数据的安全认证、通讯协议内外格式转换、交易流量控制和负载均衡功能，完成数据上传和下送。接入层支持包括定 ftp、socket，http，rest 等方式接入，对报

文协议支持广泛，并能够通过插件扩展对其他报文协议支持的功能。

数据汇集采集中心

数据汇聚采集中心由数据汇集中心和数据交换中心以及管理功能组成，完成数据汇集和交换功能。

汇聚中心在调度引擎调度和数据字典支持下，对上送汇聚的数据进行数据接收、数据解析、数据转换、数据清洗，数据融合，并由存储系统存入大数据平台。在引擎控制下，通过接入层汇集来的数据按照业务进行划分，不同的业务由不同的处理模块处理，平台支持多个模块实例对数据进行并行处理。对需要交换下传的数据转发到交换系统。

交换中心在收到需要下传交换的数据，在交换引擎控制下，完成路由查询，格式转换，消息组装后由消息发送模块处理转发。路由识别根据数据信息确定下一步的节点。需要发往外部系统的交易根据路由信息分发到对应的接入层处理，交换平台支持互通数据校验功能，以及错误处理功能，可进行交易内容和身份、权限校验，并在身份验证失败、权限验证失败、通信超时、消息 ID 重复等异常情况下进行差错处理，确保差错处理过程中的数据完整性和一致性。

管理功能，对平台提供基础数据管理、权限管理服务，系统监控，运行维护，接入管理，路由管理，资源管理，业务配置和统计分析。

服务总线

服务总线作为连接各个业务处理模块的数据和信息交换通道。

2、 数据处理和存储

平台的数据在汇聚平台处理和存入大数据存储层，形成标准化的 ODS 层数据集

合。

大数据平台通过调度任务将 ODS 层数据进行清洗和校验，数据校验内容包括检查数据类型是否匹配，数据值域的正确性，数据一致性、完整性、合法性和可靠性。

4.3.2 数据架构

1、 数据湖建设

(1) ODS 层设计

通过对源数据进行ETL的抽取,对于满足抽取规则的数据进入ODS(数据操作层)。ODS 层用于存放在接口方提供的原始数据的基础上，经过清洗、转换后一致的、准确的、干净的数据，即对源系统数据进行了清洗（去除脏数据）后的数据，ODS 层的数据和源系统的数据采用同构、贴源原则。

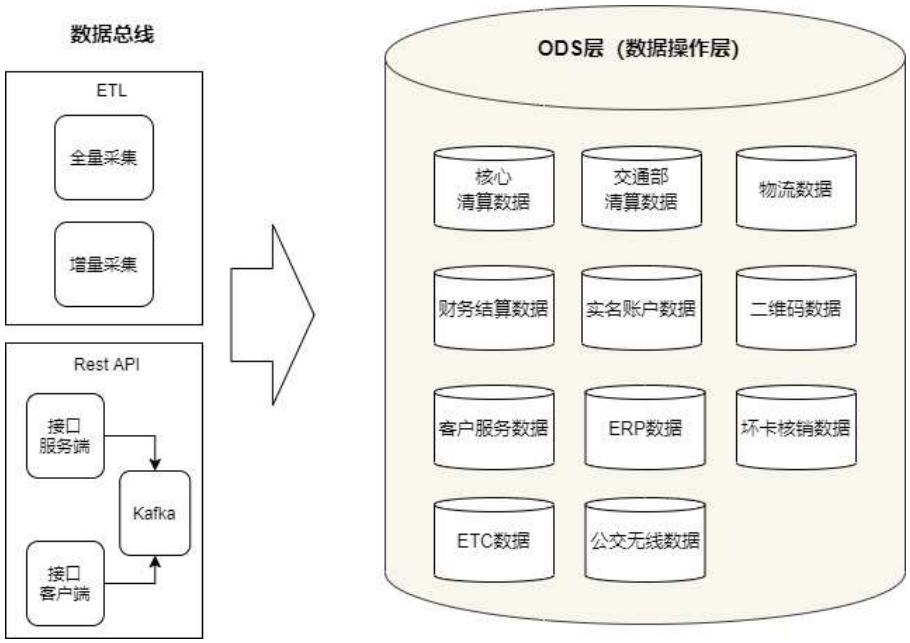

图10.　ODS 层设计示意图

(2) DW 层（DWD/DWS）设计

DW(数据仓库层) 是一个面向主题的，反映历史变化数据层，用于支撑管理决策，根据数据颗粒度又分成 DWD 层（明细粒度事实层）和 DWS 层（数据汇总层）。

进入 DW 层的数据需要进行数据清理、数据集成、数据变换、数据归约、数据离

散化等数据预处理操作，例如对噪声数据平滑处理，识别并删除孤立点，解决数据不一致性等。

进入 DWD 层的数据是一致的、准确的、干净的数据，即对源系统数据进行了清洗、集成、转换、归约和离散化后的数据，其数据粒度和 ODS 的粒度相同。

进入 DWS 层的数据是面向主题来组织的，在 DWD 层数据的基础上进行轻度汇总。从数据的时间跨度来说，主要的目的是为了满足用户分析的需求。从数据的广度来说，仍然覆盖了所有业务数据。

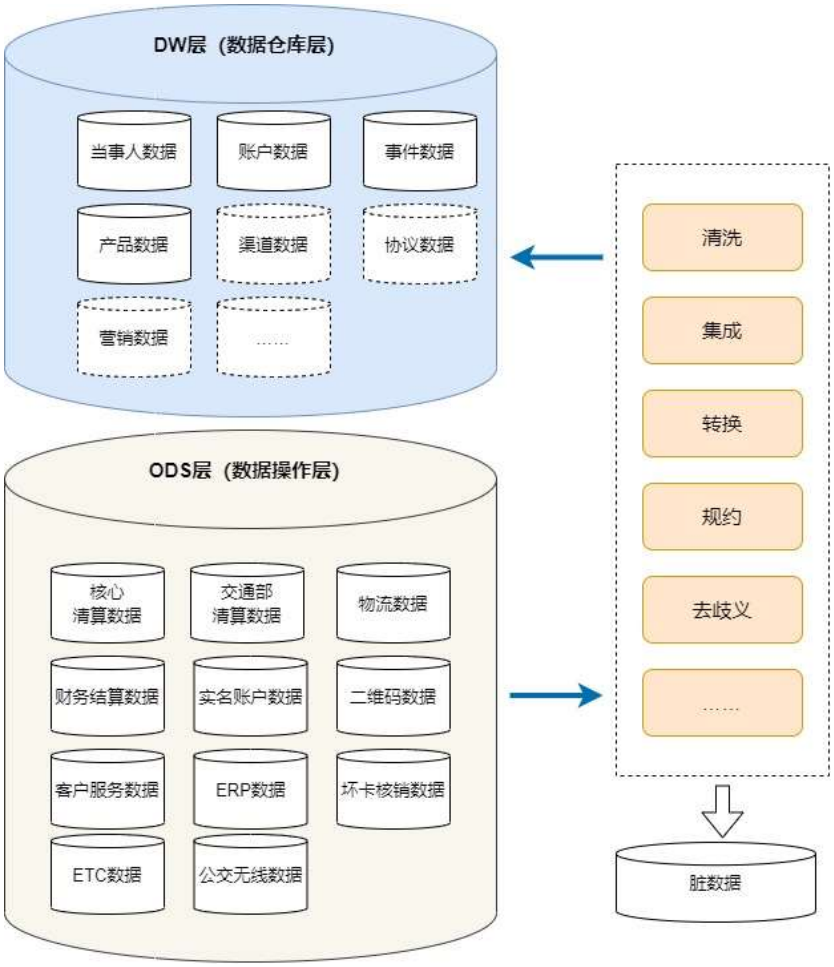

图11.　DW 层设计示意图

(3) DM 层设计

DM 层（数据集市）是数据底座数据分析和挖掘的承载。根据交通卡公司对于数

据决策分析和服务的需要，建立数据模型，数据模型基于业务主题，面向分析需求，提供各类决策支持。

DM 层数据是面向主题的，是一组特定的、针对某个主题域、部门或用户分类的数据集合。这些数据需要针对用户的快速访问和数据输出进行优化，优化的方式可以通过对数据结构进行汇总和索引。DM 层主要采用星形模型和雪花模型，满足行业分析、辅助决策等方面的要求。

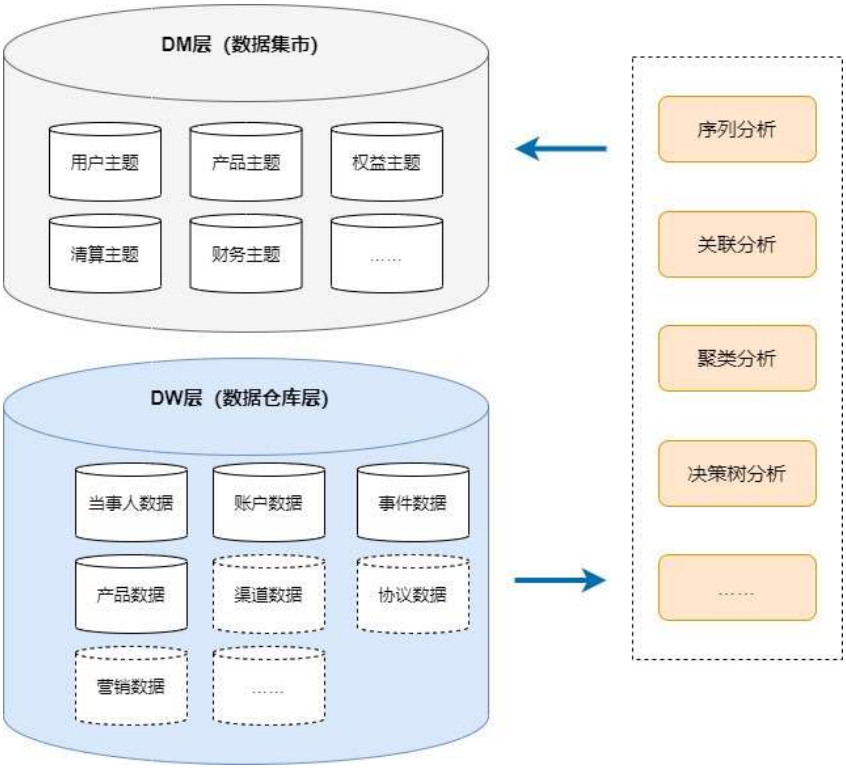

图12.　DM 层设计示意图

2、 数据模型建设

(1) 产品主题

产品是交通卡公司赖以生存的生命线，其重要性不言而喻。本文结合交通卡公司的实际业务情况，将产品分为线上产品、线下产品和其他类产品。

(2) 模型设计

通过分析得到无论线上、线下产品还是其他类产品，其具有共同的基础属性的特征，比如上线和下线时间、销售形式、销售金额等等，因此将不同的产品抽象形成产品模型，以产品代码作为唯一标识，通过产品类型编码、产品类型描述属性对产品进行分类和说明。

考虑到不同产品具有独特属性，因此将分别将不同产品类型的产品特有属性作为扩展信息，并通过产品代码与产品基本信息进行关联。比如线下产品的制作批次、制作押金、制作储值属性以及线上产品的产品有效期等属性。

另外，产品主题域还包括产品技术规格、产品组合（关系）、产品变化历史、产品适用用户、线下产品库存以及线下产品的供应商信息。

其中产品技术规格用于记录产品的规则和型号信息；产品组合（关系）主要维护组合产品，例如定位册作为组合产品包含的纪念卡信息和纪念册信息；产品变化历史用于记录产品的历史变更情况；产品适用用户用于记录哪类用户可以使用该产品，未来可与协议主题形成关联；线下产品库存分为两个模型表示，一个是本部的库存模型，另一个是销售代理商库存模型。

4.3.3 数据标准

平台项目建设是一个迭代优化的过程，由于交通卡公司经过近 20 年的信息化发展，在信息化系统建设过程中对元数据、主数据的管理变得愈发重要，因此在本项目需结合交通卡公司业务的复杂性，针对交通卡公司具体的数据结构和业务系统的实际情况，对卡公司的数据标准体系进行梳理和标准化规范管理，形成卡公司的元数据标准、主数据标准，并通过构建元数据和主数据管理平台辅助数据管理。

1、 主数据标准

主数据是具有共享性的基础数据，可以在企业内跨越各个业务部门被重复使用

的，因此通常长期存在且应用于多个系统。由于主数据是企业主数据，数据来源单一、准确、权威，具有较高的业务价值，因此是企业执行业务操作和决策分析的数据标准。

2、 元数据管理

元数据常见的定义是：“关于数据的数据”。更准确一点说：元数据是描述流程、信息和对象的数据。这些描述涉及像技术属性（例如，结构和行为）这样的特征、业务定义（包括字典和分类法）以及操作特征（如活动指标和使用历史）。将原始数据转换为知识时需要元数据，元数据使数据产生意义。例如元数据能将数字表格识别为收入数据，此外还提供如何翻译数据的业务观点，以及关于数据表格的组织与架构的技术观点。

元数据为业务分析人员提供数据的相关数据，例如源数据、以及如何转换至数据底座；元数据为数据底座提供关于数据组织的信息。这份信息包括表格的组织与链接、使用者的存取与更新信息、非关连式数据与关连式格式如何互相对应等；元数据提供数据架构与组织的相关信息，包括所在位置，以及每一表如何放入数据底座中。

元数据（Metadata）管理是定义、描述系统数据的重要技术手段。元数据管理贯穿数据底座建设的整个生命周期，为用户提供统一的数据描述和定义。通过元数据对业务术语、逻辑模型、物理数据结构和数据转换过程等基础数据信息进行统一分类管理，建立“概念层—逻辑层—物理层”的端到端全局数据视图。

3、 数据标准化

交通卡公司数据的标准化建设通过元数据管理和主数据管理来实现。

数据标准化的实现思路：

1）通过元数据管理平台来定义数据底座的元数据，初始的元数据信息通过元数据采集功能完成，后续通过元数据维护功能补充完整元数据信息。

2）数据底座的数据主要由两部分组成，主数据和交易数据。两者的数据标准化过程为：

主数据：主数据模型由元数据管理平台进行定义，主数据管理平台根据主数据模型，通过主数据采集功能进行主数据的采集。采集过程中根据元数据管理系统定义的元数据标准和清洗规则进行数据校验和清洗，完成主数据的标准化。最后经主数据管理流程发布后由数据底座和其他业务系统进行生效。

交易数据：数据底座在从源系统进行数据抽取加载或数据接口服务的过程中根据贴源原则进行数据存储，在此过程中为避免数据丢失不做强校验，但会将元数据关系回送至元数据管理系统中。在进行 ODS 层向 DWD 层清洗和数据加工时根据元数据管理平台所定义的规则对所抽取的数据进行转换、清洗，以此完成数据标准化，并根据数据模型进行数据的合并和加工。同时记录元数据间关系并回送至元数据管理平台，便于其后续进行数据血缘分析。

数据底座在进行 DWS 层、DM 层数据汇总时，记录数据变换并回送给元数据管理平台，供数据血缘分析。

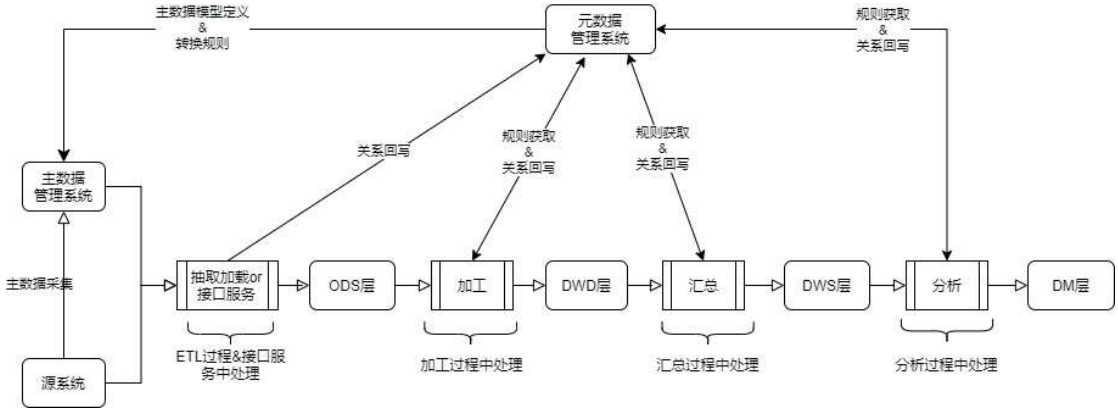

图13.　数据底座标准化流程图

4.3.4 数据应用

1、 数据安全

(1) 传输安全

ETL 方式考虑到大数据平台部署在久事云，中台与数据源间采用专线方式传输，因此在接入安全上通过配置防火墙策略进行采集方的合法性控制，同时DataX json 配置数据源信息通过密文方式存储，在任务启动时解密执行，以此最大化保障 ETL 方式的数据接入安全，并有效保护数据源端信息安全。

RestfulAPI 接口方式，首先通过 https 协议保证数据传输过程的安全性。同时大数据平台的接口设计中规范 http head 包含 app-id、sign 和 token 等信息，通过对头部信息验签确保请求的合法性。（拟通过 oauth2.0 进行资源授权控制）。

对外 kafka 消息队列接口方式通过启用 jks 加密，基于 jks 生成公私钥，以此保证数据传输安全性。

由于大数据平台的内部网络为受控环境，且加密影响系统性能，因此不做加密。

(2) 存储安全

大数据平台提供数据存储脱敏，用于关键信息进行加密存储和解密展示。大数据平台的存储加解密有两种模式。

模式一：通过硬件加密设备进行密钥管理及加解密运算，应用在获取数据后将关键信息通过加密设备运算后存储密文，当需要读取显示时再通过加密设备进行解密运算。

模式二：通过数据安全系统进行加解密管理，应用在获取数据后将关键信息送入数据安全系统加密后存储，数据读取时同样送入数据安全系统进行解密后显

示。

模式三：通过大数据平台实现软件加密，数据在存取过程中调用加解密服务处理，加解密密钥通过合规方式存取使用。

大数据平台将根据交通卡公司整体数据安全架构规划选择以上三种模式中的一种进行实施。

(3) 可视化脱敏

数据可视化脱敏是指大数据平台在进行数据展示时，由后台应用在读取数据后即对敏感信息进行脱敏处理，脱敏方式主要采用部分信息遮蔽。

脱敏原则：

- 敏感信息定义：客户个人隐私信息，包括客户姓名、电话号码、手机号码、证件号码、通信地址、邮箱、银行卡号、开户行账号、退款联系电话、联系人姓名、联系人电话、开票人姓名、开票人手机号码、开票人证件号码等等。

脱敏规则：

- 信息查询及展示：由于业务需要查询展示客户信息的，根据用户级别和权限，展示信息时进行截断显示，对于低权限用户系统展示遮蔽后信息，对于高权限用户则可以看见完整信息不做限制。
- 日志记录：日志中不明文存储客户敏感信息，如系统业务或维护需要，则对敏感信息进行截断处理。

2、 数据质量

大数据平台的数据质量服务包括两部分：一部分是针对数据的接入质量进行

管理和监控；另一部分是对大数据平台的元数据质量的监控和管理。

数据接入质量管理包括，包括对 ETL、API 接口、消息队列接口等方式的接入质量监控管理。

ETL 方式接入质量通过调度系统收集每个工作任务的执行日志，并将其记录到大数据平台数据库中供前端界面查询。ETL 数据质量监控内容包括启动日期及时间、任务消耗时间、接入数据总量、失败数、写入速度等。

API 接口方式质量监控通过 API 被调用时记录调用方信息，定时对被调用记录进行汇总生成监控项。

消息队列方式在本项目中作为 API 接口方式的异步处理方案，不直接暴露给外部系统使用，因此数据质量监控同 API 接口。

元数据质量管理包括数据命名管理和数据使用反馈管理。

数据命名管理是指大数据平台根据元数据管理系统中所定义的元数据命名规则对已接入的数据进行校验，根据元数据标准判断并列出不符合标准的元数据命名。

数据使用反馈管理是指对大数据平台中所存储数据的元数据进行分析，判断数据的数据类型、精度以及值域是否符合元数据管理的定义。

3、 文件/数据分发

大数据平台文件/数据分发模块为大数据平台数据服务提供的服务方式之一，包括文件分发和数据分发。文件分发主要指提供制定文件格式的数据导出功能，数据分发主要指提供大数据平台作为消息生产者通过消息队列方式对外进行数据分发或以 rest api 接口方式对外提供数据服务。

4、 数据服务

大数据平台数据服务主要以 HTTP REST API 方式提供服务。采用 swagger 框架规范 API 的设计，系统发布时自动获取 API 列表及 API 的版本信息和参数信息，通过控制台提供 API 的查询界面及设置 API 访问权限功能。

大数据平台的数据服务采用 Springboot + Nacos + Gateway 等 Spring Cloud 微服务生态圈。大数据平台通过 Nacos 组件实现注册、发布、检索、发现能力。

外部系统在进行服务请求时，由 Gateway 实现服务调用路由控制，并根据服务策略实现服务降级和服务限流。

5、 数据应用管理

(1) 统一发布

大数据平台统一发布模块主要实现大数据平台对外数据服务的管理功能，主要通过搭建 jenkins 平台实现应用软件的自动化构建、测试、发布。具体包括查看提交的作业、资源、函数，支持对发布包进行审核和发布，支持发布版本管理等。

(2) 统一调度

大数据平台基于 airflow 调度系统实现管理任务流程，其支持多种交互接口（如 Hive、Presto、MySQL、HDFS、Postgresql），并具有任务依赖管理、进度监控、任务触发等功能。

大数据平台基于 airflow 提供统一任务调度能力，调度周期支持年、月、周、日、时、分钟等，并通过 DAG（有向无环图）实现任务的上下游依赖关系控制。

大数据平台通过配置 airflow pools 提供任务调度的资源配置，并结合

priority_weight 参数实现作业优先级配置。

大数据平台通过 airflow 提供任务自依赖调度及作业出错自动重试。

(3) 统一运维

大数据平台基于 airflow 调度系统的维护功能提供统一运维，包括任务作业数、运行中作业数、失败作业数、已完成作业数、待运行作业数等信息的运行情况总览。

统一运维提供任务工作流的上下游关系、任务情况的监控，并提供任务的重跑、终止、跳过、下游重跑等功能。

大数据平台自身应用系统监控及运维采用 Prometheus + Grafana 实现。Prometheus 是一套开源监控报警系统，并自带时序列数据库(TSDB)。Prometheus 通过 HTTP 协议周期性抓取被监控组件的状态，不需要任何 SDK 或者其他的集成过程，适用于虚拟化环境的系统监控。

4.3.5 实时数字服务引擎系统

为满足数据服务产品的研发和运营，配套开发实时数字服务引擎系统，提供快捷的数据产品开发和服务系统。后续在此平台上，为公司和行业研发客流宝系列的数据产品提供能力服务平台。

1、 数据感知工作流

通过数据感知工作流管控，实现数据产品的快速开发，通过数据源配置、数据处理规则配置和数据输出配置，支持快速敏捷数据产品开发。

2、 实时数据服务网关

数据服务网关支持插件化，方便开发人员自定义组件，支持横向扩展，高性能，可动态配置更新生效，不需要重启网关。网关支持数据产品发布、更新、服务

的生命周期管理功能。系统能够满足按需配置数据服务产品，并实现产品的有效上架、设置、下架、推出等管理，产品销售统计等功能。

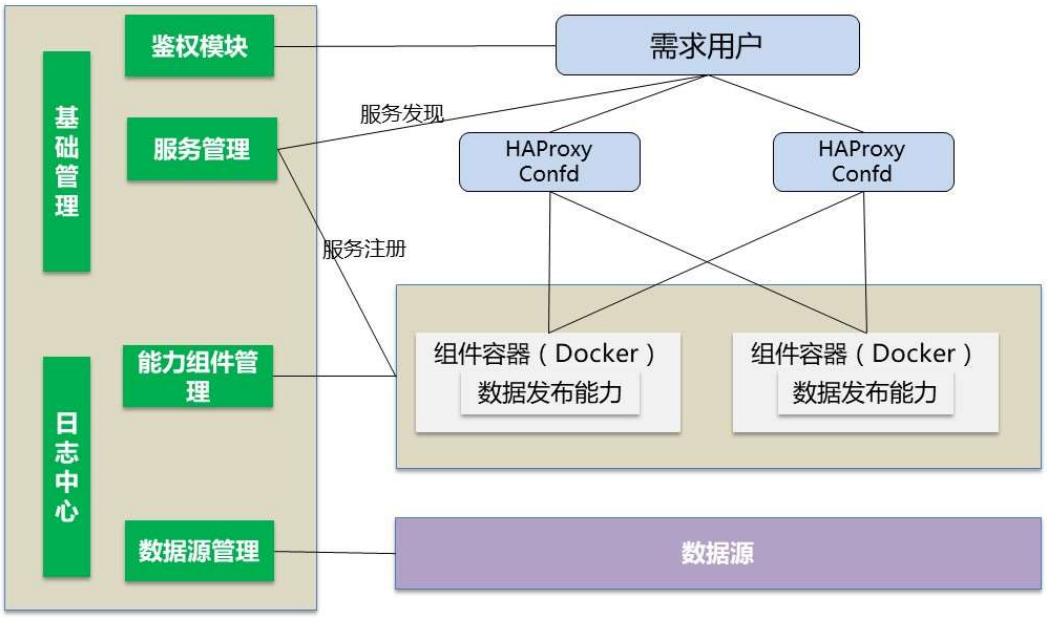

图14.　数据服务网关

4.3.6 客流宝系列数据服务产品

开展数据资产运营和场景应用示范。

1、 用户画像数据产品

用户画像数据产品利用交通卡用户中心、清算系统的数据，设计用户画像指标和标签，设计数据处理规则，形成数据集合输出。为企业开展市场产品设计服务。

2、 渠道权益营销产品

针对企业互联网移动业务拓展和服务营销需求，在交通大数据敏捷服务平台上开发一套高吞吐，实时的营销系统。系统通过配置营销规则，配置引入实时数据流源，实现基于数据的精准营销。

3、 久事客流宝数据产品

面向外部行业性对交通行业的数据产品，在数据脱敏基础上，研发公共性久

事客流宝系列数据产品。

4.4 网络系统

本项目系统拓扑图如下：

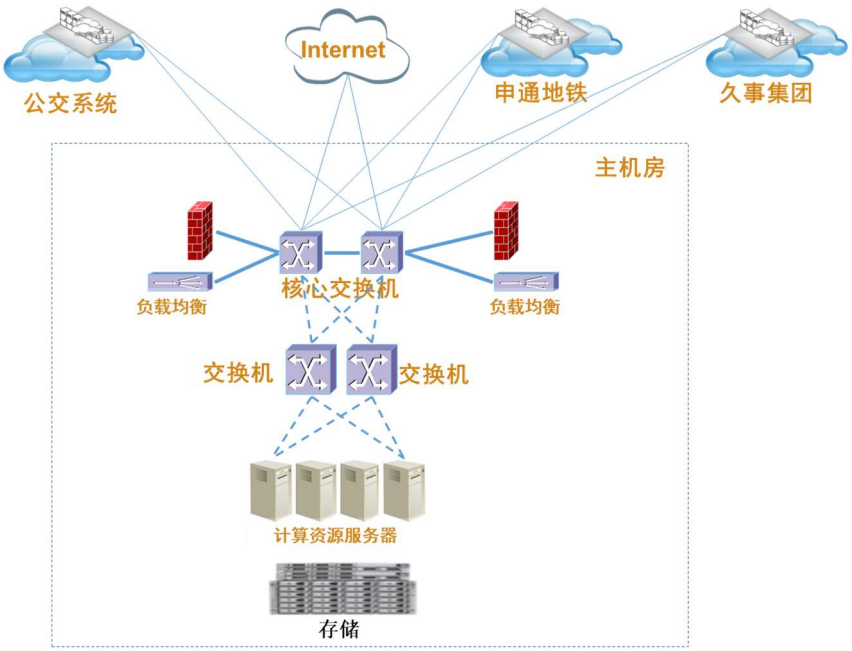

图15. 网络物理拓扑图

本次项目在新购设备上部署，系统部署于交通卡公司自有机房。

4.5 服务器和存储系统

服务器和网络设备包括服务器、核心交换机、防火墙等设施。

数据存储的硬件设备，拟采用服务器磁盘阵列的方式为基础构建云计算及存储服务。

序列	类型	描述	数量
1	应用服务器	应用服务器：2288H V5、2 个 2.3GHz CPU、6*32G 内存、2*480GB SATA 固态硬盘、2*16GB HBA 卡、1 块 1GB 以太网卡、2*10GB SFP	9

2	数据服务器	数据服务器：5885H V5、4 个 2.5GHz CPU、24*64G 内存、2*480GB SATA 固态硬盘、2*16GB HBA 卡、1 块 1GB 以太网卡、2*10GB SFP	2
3	存储	磁盘阵列 双控直连 SAS\FC 磁盘阵列 12*12T	2
4	外防火墙	外网防火墙，6 电口 4 光口，三层吞吐量 8Gbps，应用吞吐量 2.5Gbps	2
5	内防火墙	内网防火墙，，6 电口 2 光口，三层吞吐量 12Gbps，应用吞吐量 1.5Gbps	2
6	存储交换机	FC 交换机	2
7	网络交换机	三层交换机，万兆，48 口	2

4.6 软件

基础的软件需求包括操作系统、数据库、大数据技术、开发软件等等。

具体选型如下：

序号	软件需求	选择理由
1	操作系统： Linux	公司云平台基于X86体系的统一云化服务器，选择linux,具有安全性高，运行效率高，使用范围广等。
2	关系型数据库： PostgreSQL	PostgreSQL是目前最先进的开源关系型数据库，其提供丰富的数据类型更合适作为

		OLTP/ OLAP系统的数据库服务。同时PostgreSQL支持外部数据封装，能够更好地对接基于Hadoop技术栈的大数据平台。
3	数仓存储:Hive	hive是基于Hadoop的一个数据仓库工具，用来进行数据提取、转化、加载，这是一种可以存储、查询和分析存储在Hadoop中的大规模数据的机制。hive数据仓库工具能将结构化的数据文件映射为一张数据库表，并提供SQL查询功能，能将SQL语句转变成MapReduce任务来执行。Hive的优点是学习成本低，可以通过类似SQL语句实现快速MapReduce统计，使MapReduce变得更加简单，而不必开发专门的MapReduce应用程序。hive十分适合对数据仓库进行统计分析。
4	消息队列软件：Kafka	Kafka为目前大数据处理主流消息队列系统，其具有高吞吐、高可用、实时性、可扩展等特点。
5	数据抽取及转换：DataX工具	DataX是被广泛使用的开源离线数据同步工具，其支持目前市场上主流RDBMS数据库、NoSQL数据库、时序数据库、无结构化存储

		等各类数据库系统及存储系统，可以高效地实现各种异构数据源之间高效的数据同步。
6	开发系统：SpringCloud	采用SpringCloud微服务架构，支持云原生开发，提供高并发能力和高可用性。
7	大数据处理引擎：HQL（MapReduce）和SparkSQL	HQL主要用于处理数据底座中复杂的批量大数据处理。例如数据从ODS层向DWD层进行ETL时，由于数据量较大，若使用Spark年前则需要大量内存支持，考虑到其对数据处理的时效性要求相对较低，因此选用MapReduce进行处理。 SparkSQL主要用于处理数据底座中实时数据流的数据处理以及单次数据量有限的数据处理场景。例如数据从DWD层向DWS、DM层进行汇总计算时，以及未来实时接入数据底座并需进行实时计算的实时数仓服务均采用Spark作为计算引擎。

4.7 信息安全保障方案

作为一个涉及社会民生安全的数字化基础设施，交通卡公司已经具备了完善的安全体系，拥有完整的安全防护能力和安全管理规范和运维管理制度。

本项目建设的系统部署和运行于交通卡公司自有机房的云平台，涉及的网络、存

储和服务器都在交通卡安全体系保护范围，不需要额外增加安全系统和软硬件设备。

1、完整安全防护能力体系

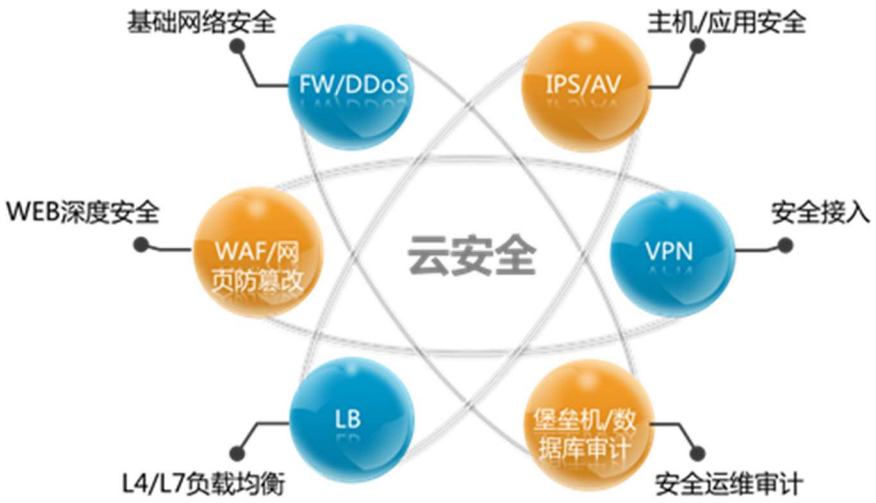

图16.　安全防护体系

安全防护体系涵盖基础网络安全、主机应用安全、WEB 深度安全、安全接入、负载均衡和安全运维审计等多个层面，包括传统安全产品防火墙、IPS、堡垒机等，并增加 VPN 安全接入、服务器负载均衡、DDOS 防护等，以及 WAF 应用防火墙、网页防篡改、网站安全监测等 WEB 应用安全服务能力。

防火墙具有高速网络流量处理能力并适用于多种网络环境，为网络提供不同层次及深度的安全控制以及接入管理，例如访问控制、IPSec/SSL VPN、应用带宽管理、病毒过滤、内容安全等。

入侵防御系统（IPS）用于实现专业的入侵攻击检测和防御的安全系统。主要部署在服务器前端、互联网出口以及内网防护等场景。

堡垒机是运维统一安全管理平台，平台内置各类法令法规（如 SOX、PCI、企业内控管理、等级保护等）对运维审计的要求，实现统一账户管理与单点登录，支持监控与历史查询，全方位风险控制。

负载均衡是实现对网络流量分流扩展网络设备和服务器的带宽、增加吞吐量、加强网络数据处理能力、提高网络的灵活性和可用性，支持业务并发和冗余。

2、数据安全管理

构建大数据安全综合防御体系，建立覆盖数据收集、传输、存储、处理、共享、销毁全生命周期的安全防护，推动数据脱敏、审计、防泄露、追踪溯源等技术手段在大数据环境下的增强应用，保障数据有序流动和安全可控。

数据完整性。针对数据完整性策略，在数据传输阶段涉及跨越公司之外的外部系统传输，采取专线通道和 VPN 加密通信两种方式进行数据传输安全，专线方式提供了标准的终端接入设备接口，以安全、透明、高速的方式进行数据传输，VPN 加密通信采用了加密整个通信链路的方式对数据传输前的身份认证过程和传输中的数据机密性、完整性提供了安全保障；在数据存储阶段基于数据分类存储的安全原则，对数据进行分类、敏感数据加密，静态数据加密存储等方式进行存储。对所有敏感数据进行数据库层面字段级加密，并严格控制密钥管理，确保敏感数据不会泄漏；数据使用阶段：基于最小特权原则，权限管控和沙箱等对数据的使用进行有效的隔离和监控，并对所有操作记录了详细的操作日志供事后审计、追溯等，保证了对数据在使用过程中的可问责性。

数据可用性，数据的可用性从数据的一致性、准确性、完整性、时效性等几方面实施。数据一致性，确保数据信息系统中各相关数据信息之间相容、不产生矛盾；数据的准确性确保对数据进行操作的各个环节都可能影响数据准确性；数据的完整性完全满足对数据进行各项操作的要求；数据的时效性：是指在不同需求场景下数据的及时性和有效性。

3、业务连续性

业务连续性在应用层面通过系统冗余和异地灾备方式实现，所有的业务系统设计按照云原生设计，支持冗余和动态扩展，关键核心业务平台在灾备机房部署灾备系统，实现在线灾备。

4、安全规程和标准

本项目的安全需求严格遵循公司的安全规范和管理制度，安全需求达到国家三级安全等保规范。

严格遵循交通卡公司的安全管理制度、信息安全规范和标准体系，满足合规性审查，保障系统等级保护的有效性。以合规和等保实施为引导，通过安全运维、安全防护、审计分析、安全制度四部分的部署加固，实现事前检测、事中防护、事后追溯的安全要求，结合安全服务对管理制度的完善，提供对重要信息系统起到一体化安全防护，保证了核心应用和重要数据的安全。满足三级等级保护对相关检测项目的要求。

4.8 采用的标准

- 《计算机软件需求规格说明规范》（GB/T 9385-2008）
- 《信息技术软件生存周期过程》（GB/T 8566-2007）
- 《 计算机软件质量保证管理计划规范》（ GB/T 12504-1990）
- 《计算机软件文档编制规范》（GB/T 8567-2006）
- 《计算机软件测试文档编制规范》（GB/T 9386-2006）
- 《计算机信息系统安全保护等级划分准则》(GB 17859-1999)
- 《GM/T 0054-2018 信息系统密码应用基本要求》
- 《计算机信息系统安全》（GA 216.1－1999）

- 《中华人民共和国计算机信息系统安全保护条例》（中华 人民共和国国务院令 147 号发布）
- 《Web 页面设计规范》

4.9 数据管理方案

数据管理方案是一系列改变数据使用行为的政策过程和制度体系，在一定程度上可以定义为：在组织内外所涵盖的管理资源的基础上，制定数据相关标准规范、管理规章及人员权责，通过长期执行既定制度，协调，整合现有技术资源，不断完善组织内数据资产管理的长期过程。

大数据技术平台的数据管理是为了规范业务数据规划、数据标准、数据质量、数据认责中的各类管理任务和活动而建立的组织、流程与工具；通过一个常态化的数据管理组织，建立数据集中管理长效机制，规范数据管理流程，提升数据质量，促进数据标准一致，保障数据共享与使用安全。

1)管控对象

管控的对象为数据模型、元数据、主数据、事务数据、共享交换数据。

根据信息系统数据对象的粒度划分，可以将数据对象划分为以下五类，具体包括：业务术语、指标定义、数据模型（逻辑/物理）、数据元素和基础编码。轨道数据模型管控的对象应覆盖以上五类数据对象。

2)管控支撑工具

实现有效的数据管控，需要从管控组织、管控流程、管控工具和评价考核等方面进行建设，这四项工作内容相互作用、相互支撑。

3)管控任务

管控的任务包括：数据标准管理、数据质量管理、数据安全管理、数据审计管理等。

数据管控是一项多元化的协同工作，需要考虑管理 定位、责任分工、工作流程、系统支持、考核体系等多方面因素。在数据管控工作过程中，通过借鉴行业经验，建立规划、组织、制度、技术工具和专项考核等因素相结合的综合型数据管控机制。

（1）规划层面：完成数据管理制度与流程体系的整体规划。

（2）组织层面：成立了集团领导亲自挂帅，信息化办公会领导下的跨部门的数据管控组织。

（3）制度层面：制定了数据标准管理、数据录入 维护管理、数据质量管理、元数据管理、数据模型管 理、数据平台数据交换管理、报表需求管理、手工数据 采集管理、数据仓库管理等 9 项数据管理办法，需要覆盖了数据定义、产生、存储、加工、交换和应用等数据 全生命周期管理，为数据管理工作提供制度保障。

（4）技术层面：通过建设了大数据管理系统，依据数据管控的内在关系将各项管控任务有机联系，包括元数据管理、数据标准管理、数据质量管理，参与数据管控工作的重 要技术手段，并持续开展系统升级工作，建立大数据资产地图，实现元数据管理。

（5）考核层面：通过设立数据管控专项考核指标，保障数据管控制度的落地执行。

第五章　项目实施进度和组织安排

5.1　项目建设周期

本项目的设计、建设、试运行和验收，预计约 24 个月。

2022 年 1 月 1 日 - 2023 年 12 月 31 日。

5.2　实施进度计划

2022 年 1 月-2022 年 3 月，完成项目方案确认、总体设计、详细设计、原型设计等。

2022 年 4 月-2023 年 9 月，完成平台数据汇聚系统和数据存储平台部署，具备数据归集和存储能力。

2022 年 10 月-2022 年 12 月，完成其他系统开发部署，进入试运行。

2022 年 10 月-2023 年 12 月，开展数据接入和数据资产开发，发展数据产品，推进运营。

5.3　责任人和组织保障

项目总负责人：张弛

项目技术总负责人：费晔

项目总协调：王联凤

技术研发负责人：彭春露

项目总架构师：刘凯

项目经理：沈盛

数据负责人：王冕

协调组：周琪、徐振宇、应鹏、张文杰、陈铭亮

开发运维组：顾春辉、崔怡婕、顾隽、王君青、施锦、朱欣荣、薛惠然、马志春

产品设计组：罗维维、李柏丰、倪慧、付宇航、姜亚萍、黄皓如

第六章　项目风险及控制措施

6.1 项目实施的内部风险及控制措施

1、质量风险

应对措施：加强系统开发中从设计、开发到测试、部署等质量风险分析，严格企业内部的项目管理制度，设定专门岗位负责质量监管，做到风险可控。

2、进度风险

进度延期风险是信息化未按项目计划进度推进，导致项目延期和费用增加。

应对措施：加强项目需求管控，加强项目组织管理工作，提高项目效率，加强人员培训，提升项目协同和沟通，严格控制项目进度及考核，控制相应风险。

6.2 项目长期运行风险及控制措施

加强规划，建立长期运营策略和演进线路图，形成可持续发展规划，推进平台长期演进和运营。

建立长期有效的管理制度及运维团队，实现持续运维。

第七章　总投资及所申请专项资金的详细估算和资金来源

7.1　总投资

本项目总投资 1716 万，其中自筹资金 1216 万，申请专项资金 500 万。该项目的费用计划包括以下各项：

单位：万元

资金来源	合计	1716
	申请专项资金	500
	自筹资金	1216
	其中：银行贷款	0
总投资	合计	1716
	设备购置费	250
	硬件资源与服务租赁费	0
	软件购置费	0
	软件开发费	1360
	数据服务费	45
	系统集成费	20
	审计费	5
	测评费	8
	其他	28
申请专项资金	合计	500
	设备购置费	100

	硬件资源与服务租赁费	0
	软件购置费	0
	软件开发费	400
	数据服务费	0
	系统集成费	0
	审计费	0
	测评费	0

7.2 设备概算表

序号	设备名称及型号	单价（万元）	数量	总价(万元）
1	应用服务器	9	9	81
2	数据服务器	32	2	64
3	存储	9.5	2	19
4	外防火墙	12	2	24
5	内防火墙	12	2	24
6	存储交换机	13	2	26
7	网络交换机	6	2	12
	合计	250		

7.3 应用开发概算表

序号	软件名称及型号	单价(万元）	数量	总价(万元)

1	交通卡数据底座	580	1	580
2	实时数字服务引擎	340	1	340
3	数据资产运营系统	440	1	440
	合计	1360		

第八章　经济和社会效益

8.1　项目经济效益

（1）降本增效赋能。研究建立集团“可量化、可评估”的降本增效评价指标库，借助可视化技术和大屏，实现集团业务精准业务过程和成本分析，实现降本。为各企业提供数据支撑，如利用交通卡等数据，定期为公交运营提供客流数据支持，增加效益服务。

（2）数字化营销赋能。通过数据挖掘反哺业务，形成客户跟踪评估和细分市场客户画像机制，为各板块服务优化提供支持。

（3）提供企业数据服务，吸引企业合作运营，预计每年代理营销合作运营增加400-800 万收入。

8.2　项目社会效益

（1）行业性的大数据平台汇聚行业数据，有助于支持城市数字化治理，减低社会运营成本。项目归集公交体系广泛的出行与运行数据，形成规范标准、种类丰富，覆盖面广、实效性强的交通行业高质量数据集合，成为行业性数据中心，为行业提供丰富数据资源，有助于提高城市行业数字化精准管控，提高城市运营效率和降低城市运行成本。

（2）解决行业数据应用难问题。在交通大数据敏捷服务平台为用户提供数字产品开发和运营环境，实现平台企业客流宝系列数据产品示范性应用。平台提供的数据资产集合、知识集合，数据脱敏和隐私计算服务能力，产品开发和运营服务能力，有效解决数据开发和共享难点痛点，推进行业数据广泛互动融合和共享应用。

（3）支持国家数字化战略。交通大数据敏捷服务平台有效推进交通卡公司和交通行业的数字化转型，创新发展数字化应用场景切实支持和推动国家数字化转型战

略和上海市数字化转型需求，推动了社会发展。