新一代跨部门数据共享交换平台解决方案
智能一体化数据目录和数据交换 TurboDX SaaS 服务平台

一、解决方案的亮点

1.1 智能、科学，着眼于中心和部门“云边端”的实际需求

平台产品面向大数据中心和各部门提供开箱即用、操作简单、可自主服务的 AI 智能化数据目录管理和实时数据同步交换的功能服务，打通“最后一公里”，实现端到端的数据流和可视化管理。

与现有市场上需要大量定制接口开发的数据采集和数据交换、以及事后人工录入编目的项目产品不同，我们的创新产品采用自动化+AI 智能化+自下而上的数据梳理和编目方法，以及基于可视化数据目录、操作简单易用的 Web 界面“点击式”的数据实时交换任务配置，相比传统完全靠人工录入梳理和编目的工具和方法、以及通过开发接口及配置交换流程的陈旧数据接入方式，大大降低了项目人工成本和时间周期达 90%以上。更大的价值是通过机器算法丰富了实用的元数据及保证了数据的及时性和数据生命周期管理的可持续性运维，有效地为部门解决了“有什么、是什么”的数据资产可视化管理、以及大数据中心找数难、流动难、融合难、消费难的痛点问题，为数据资源有效地开发和利用，以及为促进传统政企信息中心发展到大数据管理中心、数据服务中心的数字化战略转型提供了科学的工具和方法。

1.2 融合 AI 智能元数据分析/分类技术和基于数据库日志 CDC 的实时复制同步技术

平台产品采用先进的基于元数据 AI 智能分析的数据标签，并通过非侵入式的数据库日志 CDC 技术以获取数据库的库表结构变化和数据实时增量，实现智能化的数据目录以及实时的数据复制同步、汇聚和分发。

平台产品将大数据智能分析技术应用于数据资源梳理和目录可视化管理领域，实现对数据源连接的自动化元数据采集、字段语义识别、数据关联性分析、业务活动分析、主数据和敏感信息等标签化梳理，通过机器计算产生有价值的元数据信息如字段的语义、字段的业务唯一性、字段的敏感性、活跃表、主数据表等有实际应用价值的元数据信息，并提供机器辅助分类/编目等实用性功能。通过元数据目录的功能，以及方便易用的勾选和"点击式"的 Web 界面操作，灵活地配置数据复制和同步任务，让用户可以实现"所见即可交换、结果即可验证"的"端到端"可视化管理和智能一体化的数据目录和数据交换应用体验，为政企数据资源管理、数据共享交换、数据归集融合和数据挖掘等大数据创新应用提供工具化和 SaaS 服务化的系统功能支撑。

1.3 平台采用新一代数据实时复制同步和云计算 SaaS 服务技术

支持批流一体化高性能的数据采集、数据复制同步以及加载大数据平台。

根据中国指挥与控制学会(CICC)的报道，美陆军 C5ISR 中心正在探索和利用数据复制和同步等云计算解决方案来实现多方无缝数据共享，实现下一代分布式远征指挥所，其技术突破的重点在于分散指挥所节点提高其生存性的同时，还必须以可靠的方式向各分散节点提供连续的数据同步服务。这种数据复制和同步

技术是实现美陆军战术云的基础。

TurboDX 系统平台采用先进的基于数据库日志的 CDC 数据实时复制和同步技术，以及国内首创基于 Web、内存流处理和云计算的技术架构。数据库日志扫描的非侵入式 CDC 数据增量获取技术，无需在应用系统数据库端部署任何代理(Agent)程序，对应用系统的运行影响降到几乎为零(<3%)，这个特性对于接入许多业务核心应用系统是至关重要的。

将数据实时复制和同步技术，通过云计算 SaaS 服务平台输出能力，为中心及各部门提供数据库、文件、WS/REST 接入大数据平台，以及数据库读写分离、上云数据迁移、数据汇聚整合、数据分发、数据发布/订阅服务等多样性的应用场景，并满足各种不同应用场景的统一使用和监控管理需求。

平台产品完全适配和支持国产芯片、服务器和操作系统，通过了国产化鲲鹏认证。

1.4 多种应用服务模式

产品提供 SaaS 服务模式及本地化部署服务模式。

将基础的数据梳理、目录可视化管理和数据同步交换功能，以政府/企业云中心集约化 SaaS 服务化方式提供给大数据中心和各业务部门使用，“谁的资源谁管理谁负责”，首先满足和解决部门第一层次的数据目录管理可视化的基础需求。在此基础上，通过数据汇聚方式，实现“点-线-面”多层级的数据目录和数据交换可视化管理，达到“统一平台、部门共建、共享使用”的数据业务协同的应用效果，调动各部门、各应用系统开发商参与数据驱动型政企数字化转型发展的积极性和创造公平发展的机遇，这是解决目前跨部门数据共享交换、打通数据孤岛、解决

大数据"落地"归集的难点和痛点的正确有效路线。

对于没有或无法上云的部门应用系统数据源，部门也可以使用本地化部署产品，实现对本地、跨云和云上数据资源的梳理、数据目录可视化管理、以及数据复制同步和交换整合任务的配置和运行。

1.5 产品的技术和业务价值

(1) 提供简单易用、开箱即用、功能实用、可自主管理的智能一体化数据资源梳理、数据目录、和数据同步交换的工具化系统功能，以面向各政企部门的集约化 SaaS 服务平台，提升部门对数据管理和数据治理的能力，有效地解决找数难、流动难、融合难、消费难的痛点问题，相比传统完全靠人工录入梳理和编目的工具和方法、以及通过定制接口开发及配置交换流程的陈旧数据接入方式，大大降低项目人工成本和时间周期达 90%以上，大大提升数字化项目的效率和质量。

(2) 成功将大数据 AI 智能分析技术应用于数据资源梳理和目录可视化管理领域，把现在许多项目事后、被动、重复、繁琐、低价值的人工录入编目工作方式转化为事前、主动、高效、可持续、高应用价值的由部门内在需求驱动的自动化机器辅助的编目过程，有效地解决"有什么、是什么"的数据资源管理基础问题，为大数据归集融合和大型系统开发实施方案所涉及前期数据资源规划提供了前提条件，大大降低了项目的实施成本和开发周期。

(3) 采用新一代数据实时复制同步技术和云计算 SaaS 服务的解决方案，突破了过往老旧交换产品工具诸多的技术局限性。传统基于 ETL、中间件

或接口开发的交换产品工具在实时性、功能性、可靠性和性能、可管理性方面均达不到实际应用场景的多样性要求，造成丢数据、数据阻塞、同步延时拉长、以及开发工作量大、质量难于保证和长期运维困难等痛点问题。而采用新一代的复制同步技术和云计算 SaaS 服务平台工作方式，将以往交换产品单一的工具使用方式，转变为"统一平台、部门共建、共享使用"和"人人为我、我为人人"的数据业务协同服务模式，有效地调动各部门、各应用系统开发商参与数据驱动型政企数字化转型发展的积极性和创造公平发展的机遇，这是解决目前跨部门数据共享交换、打通数据孤岛、解决大数据"落地"归集难点和痛点的有效正确路线。

(4) 通过"看得见、摸得着"可自主操作和管理的成熟工具化产品的部署和使用，是帮助各政企部门培养自己的数据管理人才的有效方法和途径。数据就是业务，只有通过自身人才的培养，熟练掌握和使用可视化的数据管理工具，才能真正实现"业务数据化、数据业务化"为特点的数据驱动型政企的数字化转型。

(5) 通过使用成熟产品提供的系统化手段，让数据成为真正的资产，让业务更加敏捷和智能，为政企部门快速实现数据产品创新和应用服务模式创新赋能，创造了无限可能性。

二、行业问题和痛点

数据成为新生产资料，智能成为新生产力，要充分挖掘数据经济价值，政企都需要构建领先的数据基础设施，从而打通数据供应全流程，使能数据与业务

全连接，提升业务敏捷性。

对数据的加工处理通常包括"采-存-算-管-用"全生命周期管理，让数据存得下、流得动、算得快、用得好，帮助客户将数据资源转变为数据资产。虽然各行各业都已经公认数据中隐藏着巨大价值，但在实现过程中，面临多重挑战：

挑战 1：数据准备难

一个数据整合 BI 分析的项目，70%以上的时间都花在找到合适的数据，并判断这些数据是否具有可整合性和满足业务分析的需求。比如要花 24 小时采集数据、花 3 小时转换数据做 ETL 入库、花 1 小时准备训练数据、最后只花了半小时训练+推理，得到需要的决策数据。要解决数据准备难的痛点，需要有可视化数据梳理和目录产品工具的有力支撑，通过数据目录系统提供的导航、搜索和发现功能，快速找到、发现和定位所需的数据资源；通过对数据实体的元数据和相关性知识图谱，对数据需求的完整性、相关性和可整合性提供有力的分析工具。

挑战 2：数据流动难

解决数据流动难的问题，除了政企部门之间的业务协调外，技术平台要充分体现管理要素，要为交换中心端和部门接入端均提供方便易用的数据管理和接入操作的可视化工作平台；提供多层次的监控、报警和统计分析功能，确保交换业务运行稳定、可靠、可管、可控，数据可查可跟踪；并提供多种应用服务模式，满足各方不同管理需求，将接入管理工作在统一的平台上按职责合理分割和分解，各方职责边界清晰，可自主管理，在统一平台上实现各部门数据流动业务的协同。对于陈旧的基于 ETL 或中间件交换产品的单一化工具解决方案，我们常听到诸如"黑箱技术、易用性差、效率低、实时性差、本身又造成孤岛、缺乏服务输出能力、难于运维"等许多来自用户的差评。

根据 IDC 研究报告的统计，客户对于传统的 ETL 的解决方案在性能方面的满意度极低，满意的用户只占 17%左右。基于 20 多年前陈旧技术的 ETL 和中间件产品的解决方案面临着许多难于解决的挑战，主要有以下几个方面：

(1) 日益增加的异构数据源环境，包括各种关系型数据库、结构化及非结构化数据、以及 NoSQL 数据库和大数据平台(Hadoop、Kafka)的应用环境。

(2) 在政府私有云和混合云的计算环境下，传统产品的 C/S 架构难于满足构建"云边端"的数据"端到端"的采集和交换方案、以及在政务云中心的部署并提供多部门 SaaS 服务使用方式的要求，无法根本上解决部门端数据源连接及管理权限的隔离和安全等方面的需求，也就无法真正对接和打通部门端的业务系统和解决"最后一公里"的难题。在一个现有的政府环境中，往往不同项目要购买多套 ETL 或交换产品，各自成为孤岛，难于实现元数据集中的统一管理、共享和数据同步任务的监控和运维，造成元数据目录管理与数据交换割裂的"二张皮"。

(3) 需要编写（二次开发）脚本语言或所谓的"模板"组件，或者大量的定制化接口开发，产品易用性差、开发时间周期长、运维成本高昂，难于满足业务部门对数据的快速需求；而另一方面，越来越多的项目数据分析人员希望产品工具提供"傻瓜化"简单易用的功能，实现"自我服务"模式和数据"端到端"的可视化管理。

(4) 传统的 ETL 解决方案产品工具，往往采用批处理(batch)的数据采集/抽取方式，需要开发大量的任务，造成 ETL 任务服务器不堪重负以及丢数据、数据阻塞、实时性差，交换效率低下等痛点问题。

挑战 3：数据融合分析难

传统的烟囱式政企的 ICT 建设难以打通数据：技术众多、接口不统一、开发周期长；数据类型多，结构化/半结构化/非结构化；数据分析链路长，多系统集成难度大。数据归集融合难，一是缺乏可视化数据目录产品，对数据的相关性和可整合性提供有力的分析功能；二是市场上基于传统技术的 ETL 和中间件交换产品难于满足实时快速、简单易用、灵活可扩展性的要求，难于满足对多种异构数据源快速集成服务和实时性的使用需求，造成数据采集难、流动难、归集融合难、上线时间长、成本高昂和服务商绑架等一系列诟病。

挑战 4：数据消费难

例如某企业 IT 系统，数据源 130+万张表，要从海量表中寻找目标数据，耗时 30 天左右，犹如大海捞针；然后将目标数据加工成业务可使用数据，烟囱多、步骤多，错综复杂，又耗时 7 天。导致找数难、取数难、数据消费难。很重要的原因之一是缺乏集中统一的可视化数据目录产品，提供有用的数据搜索/发现功能。数据搜索是否能提供有价值的信息，帮助用户准确理解原始数据产生的上下文语义环境及溯源，核心问题是基于元数据采集和元数据标注的数据目录系统能否提供丰富有价值的元数据信息，包括数据的相关性、可整合性、数据质量等一系列的信息描述；而对于政府存在大量的异构数据源来说，如果没有智能化机器辅助的元数据采集和智能分析、数据标签化梳理和智能分类、数据关系的分析、以及元数据的实时变更维护，完全靠人工采集梳理/编目的方法，这几乎是不可能的完成任务。

市场上基于传统人工数据采集、人工标注和分类的数据目录产品系统，有以下难于克服的主要问题：

问题 1：完全是靠人工录入标注梳理，有的产品甚至连物理库表和字段的元数据都无法实现连接采集，完全是人工填写，逻辑数据与真实的物理数据无法关联；形成数据目录与数据交换隔裂的"二张皮"；

问题 2：缺乏有价值和实用的元数据信息，例如该字段是否是可以用来关联和整合的业务主键，该字段的具体语义是什么，是否含有敏感信息；该库表（数据集）是否活跃高频变化、是否是实体主数据；库表之间有什么关系、是否能够整合等等。对于具有成千上万张表和动辄几十万甚至上千万的大数据来说，完全靠人工梳理是几乎不可能完成和不可持续的任务，利用机器算法分析进行数据梳理和机器辅助人工的智能化分类/编目是发展的必由之路。

总之，市场上现有隔裂的数据目录产品和交换产品实施难度大、开发和运维成本高昂、实用性差，无法满足政企各部门对数据资源管理以及对数据实时采集、交换、汇聚和融合的需求，难于有效地支撑对大数据资源的快速开发和利用。

市场上出现的一些误导及对策：

- 有些服务商将政企跨部门的数据共享交换平台或数据中台项目等同于(或转变为) 一个数仓或主数据或 BI 应用开发项目，有意无意地会将一款 ETL/或中间件交换产品/或 ESB 单一化的使用工具产品充当为政企统一的数据共享交换服务平台，阉割和弱化了数据共享交换服务平台作为公共基础设施平台所要求的：多功能一体化、灵活性、多部门 SaaS 服务模式、以及统一运维管理等作为基础设施服务平台的核心功能和服务

模式要求。事实上，数仓/主数据/或 BI 应用只是基于开放统一的数据共享交换服务平台所支撑的某个应用而已。

- 用户产品选型应考虑第三方专业公司提供的独立产品并由用户或第三方公司来运维公共服务平台，避免公共服务平台成为了某些服务商绑架用户和垄断的工具，失去了数据共享交换平台/数据中台的公共服务属性和业务灵活性。选用第三方独立厂商成熟的、功能强大的、灵活的、用户可自我服务和自主管理的 SaaS 化数据集成服务平台，才能最有效地服务于各部门的应用系统开发商/数据开发商，以及实现跨部门数据业务的协同配合性，达到"统一平台、部门共建、共享使用"的应用效果，从而调动各部门、各应用系统开发商参与数据驱动型政企数字化转型发展的积极性和创造公平发展的机遇，这是解决目前跨部门数据共享交换、打通数据孤岛、解决大数据"落地"归集难点和痛点的关键路线。

三、创新产品解决方案

产品的解决方案可总结为一句话：基于元数据和 AI 智能化的数据梳理、数据目录、数据复制同步和交换整合功能为一体的云计算 SaaS 服务平台。

北京数贝软件科技有限公司自主研发的 TurboDX SaaS 服务系统平台，利用成熟的元数据 AI 算法技术、实时复制同步技术以及云计算 SaaS 服务技术，提供政府部门数据资源的元数据目录可视化管理，利用字段级语义识别、数据关联性分析、业务活动分析等并行计算的算法，提供实用有价值的数据标签化梳理、机器辅助分类/编目，以及为不同角色的数据管理人员提供不同视图的数据目录可视化应用服务功能。通过元数据目录的功能，以及方便易用的勾选和"点击式"的

Web 界面操作，灵活地配置数据实时复制同步及交换整合任务，让部门用户可以实现"所见即可交换、结果即可验证"的"端到端"可视化管理和智能一体化的数据目录和数据交换应用体验。在以下几个方面具有国内领先的技术创新：

1. 采用数据源连接的元数据采集和数据库实时增量 CDC 探针技术，实现各种类型数据库的日志增量数据，以及库表级和字段级技术元数据信息的自动化采集。
2. 采用大数据 AI 识别技术，自动分析关键字段的语义、业务主键字段、敏感字段等，提供丰富有价值的数据标签元信息，可应用于数据整合及数据脱敏等数据治理工作。
3. 采用大数据 AI 智能分析技术，分析数据库中业务数据的相关性，建立数据关系的知识图谱，并梳理出主数据，为主数据的管理和共享服务提供有力的工具支持。
4. 采用机器辅助自下而上的智能化分类/编目方法，实现物理数据与逻辑分类数据的关联及数据交换统一应用的"一张皮"。
5. 采用新一代基于数据库日志 CDC 的数据实时复制同步技术的解决方案，突破了过去交换产品工具诸多的技术局限性。传统交换产品工具在实时性、功能性、可靠性和性能、可管理性方面均达不到实际应用场景的多样性要求，造成丢数据、数据阻塞、同步延时拉长、以及开发工作量大、质量难于保证和长期运维困难等痛点问题。
6. 应用云计算 SaaS 服务平台技术，支持不同部门或不同应用系统开发商以 SaaS 服务方式使用工具化的平台功能。对于没有或无法上云的部门应用系统数据源，部门也可以使用本地化部署的产品，实现"云边端"的

"端到端"的数据流可视化管理。

四、平台优势和特色

相比陈旧的前置机(库)架构的数据共享交换平台，基于 TurboDXSaaS 数据共享交换平台的优势体现如下：

一是平台部署完毕即可实现跨部门数据端到端可达，各个接入部门不需再单独解决业务库到前置库的"最后一公里"联通问题。

二是支持接入配置工作合理安排到各接入部门，满足接入部门数据安全访问和管理需求，同时减少中心集中的接入配置工作量和责任。

三是为接入部门提供统一接入数据网关，避免一个部门需要 N 个前置机的情况，也不再需要前置库，减少前置机和前置库的购置、维护成本，避免前置库数据维护职责不清的情况发生。

四是为接入部门提供了统一的工作界面和工作平台，支持接入部门实现接入配置、运行监控和分析。

基于 TurboDX SaaS 的数据共享交换平台的特色主要体现在如下三方面：

一是简单易用。平台提供简单易用的 Web 用户界面，屏蔽了数据交换复杂的操作过程，通过点击几步界面操作，就可实现数据交换共享，因而支持部门快速接入，降低接入成本；此外，平台提供直观图形化的数据接入管理、监控管理和运行管理。

二是充分体现管理要素。为交换中心端和部门接入端管理方均提供统一工作平台；提供多层次的监控、报警和统计分析功能，确保交换业务运行稳定、可靠、可管、可控，数据可跟踪；提供多种应用服务模式，满足不同管理需求，将

接入管理工作按职责合理分割，各方职责边界清晰。

三是高效安全。采用先进技术架构和数据传输技术，数据交换效率可达 3 万条记录/秒或 10M/秒；非结构化文件交换可达 40M/秒。通过引入管道服务，能够有效分离源和目标数据库的访问权限，提高数据库访问安全。