

构建准确可信及便捷可消费的数据中台体系研究

陈吉平

杭州玳数科技有限公司, 中国·浙江 杭州 311121

摘要: 随着大数据时代的到来, 企业对数据的依赖日益增强, 构建一个准确、可信且易于消费的数据中台体系成为提升企业决策效率与业务创新能力的关键。论文旨在探讨如何构建这样一个数据中台体系, 以实现数据的高效治理、整合、分析与应用。通过分析数据中台的核心要素、技术架构、数据质量控制、安全保障措施以及用户友好界面的设计, 提出了一套系统性的解决方案, 旨在帮助企业构建数据驱动的决策支持系统, 促进数据资产的最大化利用。

关键词: 数据中台; 数据治理; 数据质量; 可信度; 数据消费

Research on Building an Accurate, Trustworthy, and Convenient Data Platform System for Consumption

Jiping Chen

Hangzhou Daishu Technology Co., Ltd., Hangzhou, Zhejiang, 311121, China

Abstract: With the advent of the big data era, enterprises' dependence on data is increasing day by day. Building an accurate, trustworthy, and easy to consume data middleware system has become the key to improving enterprise decision-making efficiency and business innovation capabilities. This paper aims to explore how to build such a data center system to achieve efficient governance, integration, analysis, and application of data. By analyzing the core elements, technical architecture, data quality control, security measures, and user-friendly interface design of the data center, a systematic solution is proposed to help enterprises build data-driven decision support systems and promote the maximum utilization of data assets.

Keywords: data center; data governance; data quality; credibility; data consumption

1 引言

在数字化转型的大背景下, 企业面临着海量数据的挑战与机遇。数据中台作为连接数据源与业务应用的桥梁, 承担着数据集成、处理、分析和服务的重任。论文聚焦于如何构建一个既准确可信又便于用户消费的数据中台体系, 以解决数据孤岛、数据质量问题 and 数据消费障碍, 促进数据价值的深度挖掘和高效转化。

关键的任务, 其中 ETL 工具、API 接口和数据湖技术是实现数据统一接入与格式转换的核心手段。

① ETL 工具应用: ETL 过程首先通过 Extract 阶段从各类数据源 (如数据库、文件系统、Web 服务等) 中抽取数据, 随后在 Transform 阶段对数据进行清洗、去重、格式转换、数据标准化等处理, 确保数据质量与一致性。在 Load 阶段将处理后的数据加载到目标数据仓库或数据湖中, 为后续分析做好准备。主流 ETL 工具如 Apache NiFi、Talend、Informatica 等, 提供图形化界面和强大的数据处理能力, 便于用户设计、执行和管理 ETL 流程 (见图 1)。

2 数据中台体系架构设计

2.1 数据集成层

在构建数据中台体系时, 整合多源数据是一项基础而

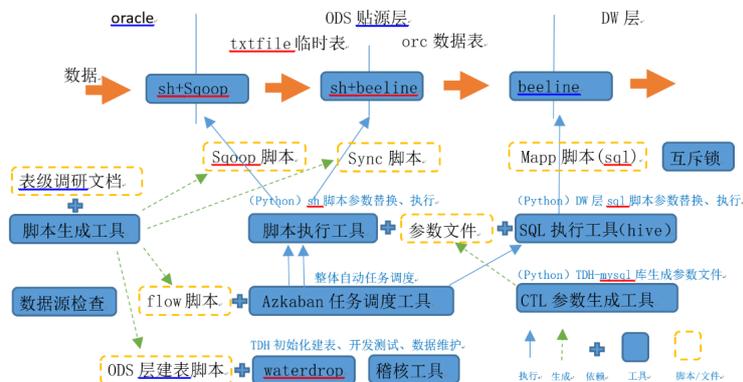


图 1 ETL 工具示意图

② API 接口集成：通过开放 API 接口，可以实现系统间的数据交换与集成。企业内部不同业务系统的数据可通过 RESTful API、SOAP 协议等方式被安全、高效地抽取，确保数据实时性与灵活性。API 集成不仅简化了数据接入过程，还支持数据的增量更新和按需调用，为数据中台提供了动态数据接入能力（见图 2）。

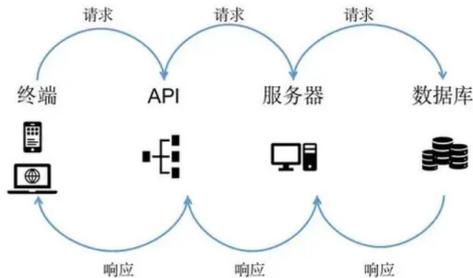


图 2 API 接口集成示意图

③数据湖技术部署：数据湖作为一种集中存储企业原始数据的体系架构，支持结构化、半结构化及非结构化数据的统一存储。在数据湖中，数据无需经过预定义模式即可被存储，便于后续的灵活分析。借助 Hadoop、Spark 等大数据处理框架，数据湖技术能够高效处理海量数据，实现数据的快速接入与格式转换。同时，配合元数据管理工具，可有效追踪数据来源、格式及变换过程，提升数据的可追溯性和可信度。ETL 工具、API 接口与数据湖技术相结合，不仅实现了多源数据的高效整合与格式标准化，还为数据中台构建了一套灵活、可扩展的数据接入与处理体系，是支撑企业数据战略、驱动业务洞察与决策的重要基石。

2.2 数据处理层

数据清洗是提升数据质量的首要步骤，涉及剔除重复项、修正错误值、处理缺失数据，以及去除异常值等操作，确保数据纯洁度。转换过程则关注数据格式与结构的适配，如将文本数据编码、日期格式标准化，或进行数据类型转换，以满足分析需求。标准化与规范化处理则是为了消除量纲影响，通过最大 - 最小缩放、Z-score 标准化等方法，使得不同变量间可比，提升模型训练的效率与准确性（见图 3）。

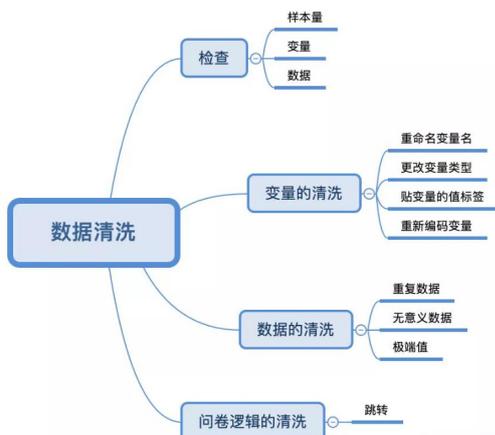


图 3 数据清洗流程示意图

模型构建前的数据准备工作还包括特征选择与工程，即识别影响预测或分类的关键变量，通过衍生新特征、聚合统计指标或降维技术，提炼数据中的有用信息，同时减少噪声与冗余。这一系列精心设计的步骤共同确保了数据分析的基础坚实可靠，为后续模型训练提供了高质量的数据输入，是构建精确预测模型与洞察数据背后隐藏规律的基石。

2.3 数据服务层

设计高效的数据 API 是实现数据流通与访问灵活性的关键。通过 RESTful 或 GraphQL 等现代 API 设计原则，封装数据访问逻辑，提供安全、标准化的数据访问接口，使得应用程序能够实时获取所需数据，促进数据的跨系统共享与集成。

数据仓库的构建专注于长期历史数据的整合存储，采用分层架构（如 ODS、DW、DM 层）来组织数据，支持复杂的分析查询。通过 ETL 过程，将清洗转换后的数据加载至数据仓库，为报表生成、趋势分析等批量处理任务提供高性能支持。数据集市作为数据仓库的一个分支，针对特定业务部门或分析需求定制化设计，存储预聚合的数据视图。它简化了数据访问路径，使得业务用户能更快地进行即时查询和敏捷分析，加速决策过程。

3 数据治理与质量控制

3.1 数据治理框架

构建全面的数据治理体系，核心涵盖数据资产管理、元数据管理和数据质量管理三大支柱。数据资产管理侧重于数据生命周期的监控与优化，确保数据资源的可见性、可控性及最大化利用。元数据管理则致力于创建和维护一个统一的元数据 repository，详细记录数据的来源、格式、含义及变化，为数据的发现、理解和交互提供坚实基础。数据质量管理强调通过建立数据质量规则、执行监控与评估，及时发现并纠正数据问题，保证数据的准确性、完整性和一致性，从而提升决策的有效性和数据驱动业务的能力。这一综合体系为组织构建了坚实的数据基础架构，赋能数字化转型与智慧决策。

3.2 数据质量监控

在数据质量管理实践中，集成规则引擎技术以自动执行预定义的质量检查规定，对数据集进行严格筛查，即时捕获偏离标准的异常情况。同时，采用先进的异常检测算法，通过机器学习模型识别数据中的非典型模式和离群点，进一步增强异常监测的灵敏度与精确度。此外，实施质量评分卡机制，为数据集的完整性、一致性及准确性打分评级，便于量化评估并追踪数据质量的改善进度。这些策略协同工作，不仅实现了数据质量的持续自动化监控，还促进了问题的快速定位与闭环改进，为组织构建了稳固的数据信任基石。

4 可信度与安全性保障

4.1 数据安全策略

在保障数据安全的综合策略中，加密技术扮演着核心

角色，通过使用如 AES（如图 4 所示）、RSA 等加密算法，对敏感数据进行加密处理，确保数据在传输过程中不被非法截取和解读，同时在静止状态下于存储介质上以密文形式保存，增加数据泄露难度。访问控制机制按照最小权限原则，严格限制用户和系统对数据的访问权限，采用角色 -based 访问控制（RBAC）或属性 -based 访问控制（ABAC）等策略，确保只有授权实体才能访问指定资源。数据脱敏技术则针对非授权环境或测试用途，对敏感信息进行变形或遮盖处理，如替换、混淆、泛化等，既保护了个人隐私与商业机密，又不影响数据分析的有效性。这些多层次安全手段的综合应用，为数据生命周期的每个阶段构建了严密的安全防线。



图 4 AES 加密算法示意图

4.2 数据审计与合规

为了确保数据处理活动符合国际及地区法律法规要求，组织需建立严谨的数据审计机制。该机制需深入贯彻《通用数据保护条例》（GDPR）、《加州消费者隐私法》（CCPA）等法规精神，通过定期执行数据保护影响评估（DPIA），监控数据收集、存储、处理、传输和销毁的全链条，验证合规性。审计过程涉及检查数据主体权利的尊重情况、数据最小化原则的执行、跨境数据转移的合法性，以及安全措施的有效性。同时，建立透明的数据处理记录，详细跟踪数据流动轨迹，确保可追溯性。通过这些措施，不仅强化了内部数据治理结构，还提升了对外部监管审查的应对能力，有效降低了违规风险。

5 用户友好与数据消费便利性

5.1 用户界面设计

为了实现数据的高效利用与普及化应用，设计一套直观易用的仪表板和自助式分析工具显得尤为关键。仪表板应以用户为中心，采用清晰的布局与图表设计，直观展示关键业务指标（KPIs）、趋势分析、异常报警等核心信息。通过交互式元素，如可点击的图表、下钻分析功能，用户能轻松探索数据背后的故事，而不被复杂的技术操作所困扰。

自助式分析工具则需集成拖拽式操作界面，允许用户自由选择数据维度、指标和可视化组件，无需编程知识即可快速构建个性化报告。内置的智能推荐算法可依用户行为和数据分析结果，自动提供数据洞察和故事线建议，进一步降低数据分析的门槛。同时，提供多样化的数据展示模板与

预置分析模型，满足不同业务场景和分析需求，确保用户能够迅速获取有价值的信息，促进数据驱动的决策制定。

5.2 数据产品化

将数据转化为可消费的产品是实现数据价值最大化的重要途径。通过精心设计的数据 API，数据如同流动的血脉，被高效、安全地输送到各个业务系统和应用中，支持实时数据集成与交互，为用户提供无缝的数据服务体验。数据 API 的标准化和文档化，确保了开发者能够轻松调用，促进创新应用的快速开发与迭代。

数据包则是预处理和打包的数据集合，根据特定业务场景或分析需求定制，用户无需从海量原始数据中筛选，即可直接获取到针对性强、格式规范的数据集，降低了数据消费的复杂度，加速了数据分析与洞察的进程。

数据应用的开发，如数据分析工具、业务智能仪表盘、预测模型等，直接面向最终用户，将复杂的数据处理逻辑封装于后台，前端展示直观、操作简便的界面，使得非技术背景的业务人员也能轻松利用数据进行决策。这些应用不仅提升了工作效率，还促进了跨部门间的协作与知识共享。

6 结语

构建准确可信且便捷可消费的数据中台体系，核心在于高效的数据集成、严格的治理机制、智能的数据处理与安全共享能力。面临的挑战包括数据源多样性导致的整合难题、数据质量参差不齐、用户数据消费习惯的培养，以及持续的技术更新与合规性要求。AI 驱动的数据治理将成为常态，利用机器学习自动优化数据流程、提升治理效率与智能化水平。边缘计算技术的融入，将实现数据的近源处理与即时分析，减少延迟，提高数据处理的灵活性与效率，进一步推动数据中台向更敏捷、智能的服务模式演进。

参考文献：

- [1] 贺前程,曹炳尧.面向用户画像的轻量化数据中台研究[J].工业控制计算机,2023(11):137-139.
- [2] 车品觉.数据中台:被忽视的创新枢纽[J].中欧商业评论,2018(12):34-37.
- [3] 邹健.构建中台能力助力企业数字化转型[J].消费电子商讯,2023(1):44-46.
- [4] 卫梦瑶.数据价值再认知:“数据中台”是伪命题吗?——一个基于营销传播视角下的讨论[J].现代广告,2020(20):36-41.
- [5] 王菊梓,郭树行,陈诺.基于数据中台构建数字化商业营销能力[J].中国高科技,2020(20):122-123.

作者简介: 陈吉平(1977-),男,中国河南郑州人,本科,高级工程师,从事人工智能研究。