

T/XJSIA

新疆维吾尔自治区软件行业协会团体标准

T/XJSIA 022—2024

碳达峰碳中和 大数据系统 第2部分：数据处理技术要求

Peak Carbon and Carbon Neutral Big Data Systems
Part1: Technical Requirements

2024 - 12 - 13 发布

2024 - 12 - 13 实施

新疆维吾尔自治区软件行业协会 发布

目 次

前言 II

引言 III

1 范围 1

2 规范性引用文件 1

3 术语和定义 1

4 缩略语 2

5 数据处理总体技术要求 2

6 数据接入技术要求 3

 6.1 数据源管理 3

 6.2 数据探查 3

 6.3 数据接入 3

 6.4 数据对账 4

7 数据处理技术要求 4

 7.1 数据提取与转换 4

 7.2 数据清洗与规范化 4

 7.3 数据关联 4

 7.4 数据比对 5

 7.5 数据分发 5

 7.6 数据处理技术 5

8 数据组织技术要求 6

 8.1 资源库 6

 8.2 主题库 7

 8.3 业务库 7

 8.4 索引库 8

 8.5 知识库 8

 8.6 数据存储技术 9

参考文献 9

前 言

本文件按照GB/T1.1—2020《标准化工作导则第1部分：标准化文件的结构和起草规则》的规定起草。

《碳达峰碳中和 大数据系统》系列标准分为四个部分：

- 第1部分：技术参考模型；
- 第2部分：数据处理技术要求；
- 第3部分：数据治理和服务技术要求；
- 第4部分：业务应用技术要求。

本文件为第2部分。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别这些专利的责任。

本文件由新疆维吾尔自治区软件行业协会提出并归口。

本文件起草单位：红有软件股份有限公司、新疆云计算应用重点实验室、新疆大学、克拉玛依众城石油装备研究院股份有限公司、新疆维吾尔自治区计量测试研究院、金风科技股份有限公司、克拉玛依碳和网络科技有限公司、广汇能源股份有限公司、新疆维吾尔自治区软件行业协会、新疆维吾尔自治区数字经济联合会、国家计算机网络与信息安全管理中心新疆分中心、新疆丝路六合科技股份有限公司、网新疆电力有限公司信息通信公司、乌鲁木齐大数据产业发展投资有限公司、新疆航天信息有限公司、新疆石文科技发展有限公司、新疆丝路智汇信息科技有限公司、新疆万维网络工程有限公司、新疆永兴同辉信息技术有限公司、新疆云天科技工程技术有限公司。

本文件主要起草人：何芳、孙素平、陈瑞军、唐虎强、刘永杰、张伟、刘明艳、陆勤华、汤杰、秦继伟、石飞、马君刚、柳青、王涛、谢远东、冯雪峰、于春伟、廖振良、刘雪松、周斌、杨晓、董伟伟、刘睿、葛磊、刘靓、吕莲花、曹永光、姚荣斌、张敏、郭江涛、李雅洁、杨柳、赵阳、陈勇、邱兰、魏利红、雷虎、赵欣善、李忠海。

引 言

为了促进大数据技术在碳达峰碳中和领域的广泛应用和创新发展，我们特制定碳达峰碳中和大数据系统系列标准，拟由以下四部分构成：

——第1部分：技术参考模型。提出了碳达峰碳中和大数据系统的技术参考模型。明确了系统的整体结构、数据流向、各部分的功能要求以及各组成部分之间的关系；

——第2部分：数据处理技术要求。规定了碳达峰碳中和大数据系统在数据接入阶段、数据处理阶段以及数据组织存储过程应具备的功能和技术要求；

——第3部分：数据治理和服务技术要求。规定了碳达峰碳中和大数据全生命周期数据治理活动和数据服务活动应具备的功能和技术要求；

——第4部分：业务应用技术要求。规定了碳达峰碳中和大数据系统应具备的业务应用功能。

本文件是拟建系列标准第2部分。本文件将指导碳达峰碳中和大数据系统在数据接入阶段实现多样化数据源的稳定、高效接入，确保数据的完整性与实时性；在数据处理阶段，通过智能算法与并行计算技术，实现数据清洗、去重、补全等预处理操作，提升数据质量，为后续分析与应用奠定坚实基础；在数据组织方面，构建合理的数据模型与存储结构，实现数据的有效组织与快速检索，同时采用分布式存储技术，确保数据的可扩展性与容灾能力。

本文件适用于碳达峰碳中和领域大数据系统的规划、设计、开发、实施和维护阶段，旨在通过规范数据处理技术，提升大数据系统在碳达峰碳中和监测、报告与核查中的效能和准确性。同时，本文件也为相关领域的科研机构、企业、政府部门等提供了数据处理技术方面的指导和参考。本部分可与其他各部分联合使用。

碳达峰碳中和 大数据系统

第 2 部分：数据处理技术要求

1 范围

本文件规定了碳达峰碳中和大数据系统的数据处理功能架构，界定了数据接入、数据处理、数据组织的具体技术要求。

本文件适用于碳达峰碳中和大数据系统的设计、开发、管理和应用。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 35295-2017 信息技术 大数据 术语

GB/T 37722-2019 信息技术 大数据存储与处理系统功能要求

3 术语和定义

下列术语和定义适用于本文件。

3.1

大数据 big data

具有体量巨大、来源多样、生成极快、且多变等特征，并且难以用传统数据体系结构有效处理的包含大量数据集的数据。

[来源：GB/T35295—2017, 定义2.1.1]

3.2

大数据系统 big data system

实现大数据参考体系结构的全部或部分功能的系统。

[来源：GB/T35295—2017, 定义2.1.14]

3.3

批处理 batch processing

将一个大型作业分解成为多个任务交由多个节点分别处理，再将分解后多个任务处理的结果汇总起来，得出最终的分析结果的计算框架，具备高可用、高扩展、高并发等能力。

[来源：GB/T 37722-2019, 术语和定义3.2]

3.4

流处理 stream processing

能够对具有实时、高速、无边界、瞬时性等特性的流式数据进行实时处理的计算能力。

[来源：GB/T 37722-2019, 术语和定义3.3]

3.5

图计算 graph processing

以“图论”为基础的对数据的一种“图”结构的抽象表达，以及在这种数据结构上的计算模式。
[来源：GB/T 37722-2019，术语和定义3.4]

3.6

内存计算 in-memory processing

优先使用内存对数据进行计算、分析的一种数据处理技术。
[来源：GB/T 37722-2019，术语和定义3.5]

3.7

批流融合计算 integrated batch and stream processing

能够同时支持批处理和流处理的计算能力。
[来源：GB/T 37722-2019，术语和定义3.6]

4 缩略语

下列缩略语适用于本文件。
API：应用程序编程接口（Application Programming Interface）
SQL：结构化查询语言（Structured Query Language）
NLP：自然语言处理（Natural Language Processing）
JDBC：Java数据库连接（Java Database Connectivity）
CSV：逗号分隔值（Comma-Separated Values）
JSON：JavaScript对象表示法（JavaScript Object Notation）

5 数据处理总体技术要求

依据《碳达峰碳中和 大数据系统 第1部分：技术参考模型》提出的碳达峰碳中和大数据参考模型，碳达峰碳中和大数据提供者将新的数据或信息引入大数据系统。数据消费者使用大数据应用提供者提供的的应用。大数据计算框架提供者建立一种计算框架，在此框架中执行数据处理转换以及全生命周期大数据治理和数据服务，以满足系统协调者定义的需求。
本文件碳达峰碳中和大数据处理技术要求应至少符合GB/T35589-2017界定的要求，同时应具备数据接入、数据处理、数据组织三方面的技术功能要求。逻辑功能架构如图1所示。

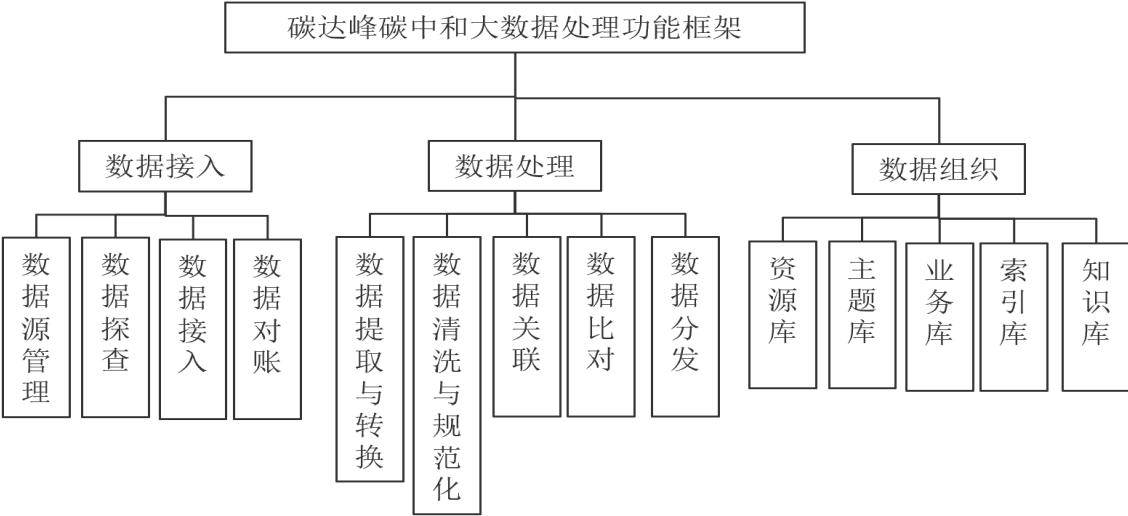


图1 碳达峰碳中和大数据处理逻辑功能框架

数据处理总体技术要求包括：

- 1) 数据接入：包括数据源管理、数据探查、接入实施与对账。明确数据源，深入探查，高效接入，确保数据一致性，建立对账机制，及时处理差异，为碳达峰碳中和业务决策提供准确数据支持。
- 2) 数据处理：包括数据提取与转换、清洗与规范化、关联与整合、比对与跟踪、分发与管理环节。应用分布式批处理、流处理、内存计算和批流融合计算等技术，确保双碳业务数据的高效、准确处理，推动业务的持续优化与发展。
- 3) 数据组织：包括资源库、主题库、业务库、知识库及索引库，结合分布式存储技术，确保双碳业务数据高效存储、规范管理，满足多样化数据处理和分析需求。

6 数据接入技术要求

6.1 数据源管理

6.1.1 数据源识别

1) 业务数据源识别：针对碳达峰碳中和管理业务，明确所需的数据源，包括但不限于企业能源管理系统、组织碳排放监测系统、政府发布的碳排放数据、行业报告、政策文件等。

系统协调者职责在于规范和集成各类所需的数据应用活动，以构建一个可运行的碳达峰碳中和垂直系统。

2) 分类管理：根据数据的来源、类型、更新频率和业务重要性，将数据源分为内部数据源（如企业内部的系统）和外部数据源（如政府、第三方机构提供的数据），并制定相应的管理策略。

6.1.2 接入策略

1) 接入方式：根据数据源的特性，选择适合的接入方式，如API接口对接、数据库直连、文件传输（如CSV、Excel、JSON等格式）等，确保数据能够高效、准确地接入系统。

2) 接入频率与时机：根据碳达峰碳中和业务的实时性和周期性需求，设定合理的接入频率，如实时接入、每日/每周/每月定时接入等，确保数据在关键业务节点前完成接入。

6.2 数据探查

6.2.1 数据结构与格式分析

1) 字段分析：详细分析数据源中的字段，包括字段名称、数据类型、数据长度等，确保数据能够正确解析和存储。

2) 数据格式验证：检查数据的格式是否符合系统要求，如日期格式、数字格式等，对于不符合要求的数据进行格式转换或预处理。

6.2.2 数据质量与业务逻辑验证

1) 数据质量评估：通过数据抽样、数据校验等方式，对数据源进行初步的质量评估，识别可能存在的数据异常、缺失、重复等问题。

2) 业务逻辑验证：根据碳达峰碳中和业务的逻辑和规则，对数据进行业务逻辑上的验证，确保数据符合业务需求。

6.3 数据接入

6.3.1 数据接入实施

1) 数据读取：根据数据源的类型和格式，选择合适的读取工具和策略，如使用Kafka、Spark等工具进行批量读取或流式读取。

2) 数据存储：将读取的数据按照业务需求和系统设计存储到数据组织活动指定的数据库或数据仓库中，确保数据的准确性和可访问性。

6.3.2 接入验证与监控

1) 数据一致性验证：在数据接入过程中，对比数据源与系统内已有数据，验证数据的一致性，确保数据在接入过程中没有发生错误或遗漏。

2) 接入监控：建立数据接入监控机制，实时或定期监控数据的接入情况，包括接入成功率、接入速度等指标，及时发现并处理接入过程中出现的问题。

6.4 数据对账

6.4.1 对账机制与流程

1) 对账策略：根据碳达峰碳中和业务的特性和需求，针对关键数据制定对账策略，如每日对账、每周对账等，确保数据的准确性和一致性。

2) 对账流程：建立清晰的对账流程，包括数据提取、数据对比、差异分析、问题处理等步骤，确保对账工作的有序进行。

6.4.2 差异处理与反馈

1) 差异分析：对于对账过程中发现的数据差异，进行详细的分析和调查，查明原因并采取相应的处理措施。

2) 反馈机制：建立数据对账反馈机制，及时将对账结果和问题反馈给相关人员或部门，确保问题得到及时解决和处理。

7 数据处理技术要求

7.1 数据提取与转换

7.1.1 数据提取

1) 提取规则：根据知识库中预设的数据提取规则和识别算法，进行结构化数据提取、文本信息提取、视频信息提取、音频信息提取以及图像信息提取。

2) 提取技术：运用SQL查询和ETL工具，遵循知识库中的规则，精确提取结构化数据；采用NLP文本语义识别技术，依据预设算法，从文本中抽取文本关键内容；通过图像识别技术，配合特定规则，从视频中捕获所需关键信息；利用语音识别和声纹识别技术，结合算法，从音频中提取数据；应用图像识别算法，依据规则，从图像中识别并提取关键信息。

7.1.2 数据转换

1) 转换规则：根据业务需求，制定数据转换规则，如将不同单位的数据转换为统一单位、将不同格式的数据转换为统一格式等。

2) 转换工具：选择合适的转换工具，如数据转换软件、脚本等。

7.2 数据清洗与规范化

7.2.1 数据清洗

1) 清洗方法：采用知识库中预设的清洗方法，包括数据过滤、数据去重、空值检验、范围校验、规则校验、格式转换。

2) 清洗工具：选择合适的清洗工具，如数据清洗软件、脚本等。

7.2.2 数据规范化

1) 规范化标准：制定数据规范化的标准，如数据格式标准、数据命名标准等。

2) 规范化工具：选择合适的规范化工具，如数据标准化软件、脚本等。

7.3 数据关联

7.3.1 数据关联

1) 关联规则：根据业务需求，制定数据关联规则，如根据时间、地点、企业名称等字段进行关联回填、关联提取、关联分析等

2) 关联工具：选择合适的关联工具，如数据关联软件、脚本等。

7.3.2 数据整合

1) 整合方法：采用知识库中预设的整合方法，如数据合并、数据汇总等。

2) 整合工具：选择合适的整合工具，如数据整合软件、脚本等。

7.4 数据比对

7.4.1 数据比对策略

1) 数据比对：将清洗和关联后的温室气体数据与基准数据（如历史数据、行业标准等）进行比对，发现数据差异。

2) 预警机制：设定预警阈值，当数据差异超过阈值时，触发预警，提醒相关人员进行处理。

7.4.2 标识与跟踪

1) 数据差异标识：对比对过程中发现的数据差异进行标识，如标记为异常、错误或待处理。

2) 数据差异跟踪：跟踪数据差异的处理情况，确保问题得到及时解决。

7.5 数据分发

7.5.1 数据分发策略

1) 目标库选择：根据数据的特性和业务需求，选择数据组织目标库进行分发。标准库用于存储符合统一标准和规范的数据；资源库用于存储原始数据、处理后的中间数据和最终结果数据；主题库根据特定业务主题组织数据；业务库则直接服务于具体业务应用。

2) 分发方式：采用内存数据库到持久化数据库的导出机制，如使用数据库自带的导出工具、ETL工具或自定义脚本等，将处理后的数据高效、准确地分发至目标库。

3) 分发时机：根据业务需求和数据更新频率，设定合理的分发时机，如实时分发、定时分发或按需分发。

7.5.2 数据分发管理

1) 分发权限控制：建立严格的数据分发权限管理机制，确保只有经过授权的用户或系统才能访问和分发数据。

2) 分发日志记录：记录每次数据分发的详细信息，包括分发时间、分发源（内存数据库）、分发目标（持久化数据库）、分发数据量等，以便后续跟踪和审计。

3) 分发质量监控：对数据分发过程进行质量监控，确保分发数据的准确性、完整性和一致性。如果发现数据分发错误或异常，应及时进行处理和纠正。

7.5.3 数据分发后处理

1) 数据校验：在数据分发至目标库后，进行数据校验，确保分发数据的准确性和完整性。

2) 数据更新与维护：根据业务需求和数据更新策略，对目标库中的数据进行定期更新和维护，确保数据的时效性和准确性。

7.6 数据处理技术

7.6.1 分布式批处理技术

1) 批处理规则：应利用分布式批处理框架，根据业务需求制定批处理规则。这些规则可能涉及按时间周期（如日、周、月）批量处理碳排放数据、能源消耗数据等，进行数据的汇总、统计和分析。同时，根据地点、企业名称等字段，对批量数据进行关联处理，以形成全面的双碳业务数据视图。

2) 批处理工具：应选用适合双碳业务需求的分布式批处理框架，如Apache Hadoop、Apache Flink的批处理模式等。这些框架提供了强大的数据处理能力，能够处理PB级的数据量，并支持复杂的数据处理逻辑。通过编写MapReduce任务、Spark作业等，实现对双碳业务数据的批量处理。

7.6.2 流处理技术

1) 流处理规则：在双碳业务中，利用流处理框架处理实时数据监测和数据分析。根据业务需求制定流处理规则，涉及对碳排放数据、能源消耗数据等进行实时采集、处理和分析，以及对异常数据进行实时预警和报警。同时，通过流处理框架，实现对实时数据的关联处理，以形成实时的双碳业务数据视图。

2) 流处理工具：选用适合双碳业务需求的流处理框架，如Apache Flink、Apache Storm等。具备高效的实时数据处理能力，能够处理高速的数据流，并支持复杂的实时数据处理逻辑。通过编写流处理任务，实现对双碳业务数据的实时采集、处理和分析。

7.6.3 图计算技术

1) 图计算规则：在双碳业务中，企业间的关联关系、碳排放的传递路径、产品碳足迹等可以通过图数据结构来表示。利用图计算框架，根据业务需求制定图计算规则。这些规则可能涉及对图数据进行遍历、搜索和分析，以发现企业间的关联关系、碳排放的传递路径等。

2) 图计算工具：选用适合双碳业务需求的图计算框架，如Apache Giraph、GraphX等，能够提供了高效的图计算能力，处理大规模的图数据，并支持复杂的图计算算法。通过编写图计算任务，实现对双碳业务中的图数据进行处理和分析。

7.6.4 内存计算技术

1) 内存计算规则：需要频繁访问和计算的数据，如碳排放的实时监测数据、能源消耗数据等。利用内存计算框架，根据业务需求制定内存计算规则，将数据加载到内存中，以提高数据访问速度和计算效率。

2) 内存计算工具：选用适合双碳业务需求的内存计算框架，如Apache Spark的内存计算模式、Apache Flink的内存状态管理等。这些框架提供了高效的内存计算能力，能够充分利用内存资源，提高数据访问速度和计算效率。通过编写内存计算任务，实现对双碳业务中的数据进行快速处理和分析。

7.6.5 批流融合计算技术

1) 批流融合计算规则：双碳业务中，既需要处理历史数据（如历史碳排放数据、能源消耗数据等），又需要处理实时数据（如实时监测数据）。利用批流融合计算框架，根据业务需求制定批流融合计算规则，将历史数据和实时数据进行融合处理，以形成全面的双碳业务数据视图，并支持对历史数据和实时数据进行联合分析和挖掘。

2) 批流融合计算工具

选用适合双碳业务需求的批流融合计算框架，如Apache Flink的批流一体模式等。能够提供高效的批流融合计算能力，同时处理历史数据和实时数据，并支持复杂的批流融合计算逻辑。通过编写批流融合计算任务，实现对双碳业务中的历史数据和实时数据进行联合处理和分析。

8 数据组织技术要求

8.1 资源库

8.1.1 定义与功能

资源库专注于存储、管理和维护数据的原始形态及其经过标准化处理后的形态。它确保数据的可访问性、完整性、一致性和规范性，为数据处理、分析和决策支持提供坚实的基础。

8.1.2 数据类型与内容

1) 原始数据：直接来源于各类监测设备、传感器、统计报表等未经处理的数据，保持其最原始、最真实的形态。

2) 标准化数据：经过清洗、转换、关联等标准化处理后的数据，符合特定的数据格式、结构和质量标准，便于后续的数据分析和应用。

8.1.3 数据存储与管理

1) 分布式存储：采用分布式存储技术，确保大规模数据的存储和高效访问。

2) 数据格式支持：支持多种数据格式的存储，如CSV、JSON、数据库表等，以适应不同类型的数据需求。

3) 元数据管理：对数据的描述、来源、格式、质量等元信息进行详细记录和管理，便于数据的查找、理解和使用。

8.1.4 数据组织与分类

1) 数据分层：根据数据的处理阶段和用途，将数据分为原始数据层和标准化数据层，确保数据的层次清晰、易于管理。

2) 数据分类：根据数据的性质和用途，对数据进行细致的分类管理，如能源数据、温室气体排放数据、碳减排数据、政策数据等，便于数据的快速定位和应用。

8.2 主题库

8.2.1 定义与功能

主题库根据特定的业务需求或分析主题，整合、归纳和提炼资源库中的相关数据，形成具有特定主题的数据集合。它为用户提供针对特定业务场景或分析需求的数据视图，提高数据分析和决策的效率。

8.2.2 主题分类与构建

1) 主题分类：根据碳达峰碳中和领域的业务需求和分析目标，划分为多个子主题，如能源消费主题、温室气体排放主题、减排措施主题等。每个子主题下又可进一步细分为更具体的主题，如能源消费主题下的电力消费、煤炭消费等子主题。

2) 数据整合：通过数据接入分发或从资源库中提取与特定主题相关的数据，形成主题数据集。

3) 数据提炼：对整合后的数据进行提炼，提取出对主题分析有用的关键信息和特征，如碳排放量的变化趋势、能源消费的结构分布等。

8.2.3 数据组织与存储

1) 数据模型：为主题库构建合适的数据模型，以优化数据的存储和查询性能。

2) 数据索引：为主题库中的数据建立索引，以提高数据的查询效率。包括为关键字段建立索引、为常用查询建立复合索引等。

8.2.4 数据更新与维护

1) 数据更新：定期或实时地从资源库中获取最新的数据，更新主题库中的数据集合。包括数据的增量更新和全量更新两种方式。

2) 数据校验：对更新后的数据进行校验，确保数据的准确性和一致性。包括对数据的完整性、唯一性、正确性等方面进行校验。

8.3 业务库

8.3.1 定义与功能

业务库支撑碳监测、碳汇、碳减排、碳资产、碳交易、碳足迹等各个专业领域内的业务活动，记录业务过程，并为这些业务活动提供必要的的数据支撑。业务库由各专业领域的业务部门主导建立，确保数据能够紧密贴合业务实际，满足业务需求。

8.3.2 数据内容与组织

1) 专业领域数据：业务库存储的数据紧密围绕各专业领域的业务需求，包括但不限于碳排放数据、碳汇量数据、减排措施数据、碳资产数据、碳交易数据以及碳足迹数据等。这些数据通过数据接入分发活动获得，或从资源库和主题库中提取、整合、转换而来，确保与业务需求的精准匹配。

2) 数据组织方式：业务库采用符合业务逻辑的数据模型和数据结构，以优化数据的存储、检索和处理效率。业务库应支持数据视图、索引等机制，以进一步提升数据访问性能。数据按照专业领域进行划分，形成独立的数据库或数据表，便于管理和维护。

8.3.3 数据处理与更新

1) 业务逻辑处理：根据具体业务场景的需求，对预处理后的数据进行进一步的业务逻辑处理，如数据聚合、关联分析、趋势预测等，以生成对业务决策具有指导意义的洞察。这些处理策略由各专业领域的业务部门根据业务需求进行定制。

2) 数据更新机制：业务库支持数据的实时或定期更新，确保业务数据与实际业务场景的同步。包括从资源库和主题库中自动获取最新数据，并进行必要的处理和整合，以保持业务库的时效性和准确性。同时，业务库还支持数据的手工录入和修改功能，以满足业务过程中的特殊需求。

8.4 索引库

8.4.1 定义与功能

索引库存储和管理用于数据检索和查询的索引信息。索引库通过构建和优化索引，提高数据查询和分析的速度和准确性，为业务的数据处理提供强有力的支持。

8.4.2 索引库的构建与管理

1) 索引设计：根据业务的数据特点和查询需求，设计合理的索引结构和算法。包括选择合适的索引类型（如全文索引、结构索引等）、确定索引的粒度（如字段级、记录级等）以及优化索引的存储和访问方式。

2) 索引技术：在索引构建过程中，采用先进的索引技术，如倒排索引、B树/B+树索引、哈希索引等，以提高索引的查询效率。这些技术能够根据不同的查询需求和数据特点，提供快速、准确的索引服务。

3) 索引创建与更新：在数据接入或更新时，根据索引设计的要求，创建或更新相应的索引信息。包括对新数据的索引构建、对已有索引的更新和维护等操作。同时建立索引库的更新机制，确保索引信息与数据保持一致性和时效性。

4) 索引优化与监控：定期对索引库进行优化和监控，以提高索引的性能和稳定性。包括调整索引的存储结构、优化索引的访问路径、监控索引的负载情况等操作。同时，应建立索引库的故障预警和恢复机制，确保索引库在出现异常时能够及时恢复和正常运行。

8.5 知识库

8.5.1 定义与功能

知识库结合知识图谱、大模型等技术，构建一个全面、准确、智能的知识体系，为业务决策、技术研发、数据分析等提供科学依据和智力支持。

8.5.2 知识内容与组织

1) 知识体系：知识库整合业务领域政策法规、技术标准、案例库、专业知识、专家资源等多方面的知识和信息，形成完整的知识体系。

2) 知识图谱构建：采用自然语言处理、实体识别、关系抽取等技术，从海量数据中自动构建业务知识图谱，实现知识的互联互通和高效检索。

3) 大模型应用：引入深度学习等先进技术，构建业务大模型，能够理解和生成领域相关文本，提供智能问答、知识推理等高级功能。

8.5.3 技术实现与更新

1) 知识图谱技术: 利用知识图谱技术, 将领域的知识进行结构化处理, 形成知识节点和关系网络, 支持知识的智能检索和推理。

2) 大模型训练与优化: 基于大规模的领域数据, 训练和优化业务大模型, 提升模型的准确性和泛化能力。同时, 建立模型更新机制, 确保模型能够紧跟领域的发展动态。

3) 数据整合与更新: 定期收集、整理和审核领域的新知识, 确保知识库的时效性和准确性。同时, 建立知识库与业务库的联动机制, 实现知识的实时更新和共享。

8.5.4 知识应用与服务

1) 智能问答: 基于业务大模型, 提供智能问答服务, 快速响应业务人员的知识需求。

2) 知识推荐: 根据业务人员的兴趣和需求, 利用知识图谱和推荐算法, 提供个性化的知识推荐服务。

3) 知识培训与教育: 依托知识库中的专业知识和专家资源, 开展知识培训和教育活动, 提升业务人员的专业素养和综合能力。

8.6 数据存储技术

8.6.1 分布式文件存储技术

宜采用分布式文件存储技术。数据分散存储在多个物理节点上, 通过网络连接共同构成一个分布式的存储系统。它支持大数据量存储、高效的数据访问以及数据共享和管理, 非常适用于双碳业务中的环境监测图像、设备日志文件等非结构化数据的存储。

8.6.2 分布式结构化数据存储技术

宜采用分布式结构化数据存储技术。采用分布式数据库系统, 将数据分散存储在多个节点上, 并通过分布式数据库管理系统实现数据的访问和管理。适用于双碳业务中的碳排放量、能源消耗量、企业信息等结构化数据的存储, 具有高性能、高可用性和可扩展性等特点。

8.6.3 分布式列式数据存储技术

宜采用分布式列式数据存储技术。将数据按列存储, 适用于双碳业务中的大数据分析场景, 如海量数据的实时查询和分析。通过列式存储, 可以显著提高数据查询和分析的效率, 同时降低存储成本。

8.6.4 分布式图数据存储技术

宜采用分布式图数据存储技术。专门用于存储和管理图数据, 如图结构数据等。在双碳业务中, 利用图数据存储技术来构建能源网络、碳排放网络等图结构, 实现数据的高效查询和分析。同时, 图数据存储技术还支持复杂的数据关系和模式挖掘, 为双碳业务提供更加深入的数据洞察和分析能力。

参 考 文 献

- [1]GB/T5271.17-2010 信息技术词汇第17部分: 数据库
- [2]GB/T5271.18-2008 信息技术词汇第18部分: 分布式数据处理
- [3]GB/T32150-2015 工业企业温室气体排放核算和报告通则
- [4]GB/T32151(所有部分) 温室气体排放核算与报告要求
- [5]GB/T33760 基于项目的温室气体减排量评估技术规范通用要求
- [6]科技部等九部门: 科技部等九部门关于印发《科技支撑碳达峰碳中和实施方案(2022—2030年)》的通知: 国科发社(2022)157号
- [7]国务院. 关于印发2030年前碳达峰行动方案的通知: 国发(2015)50号
- [8]国务院. 国务院关于印发促进大数据发展行动纲要的通知: 国发(2015)4号

[9]工业和信息化部. 工业和信息化部关于印发“十四五”大数据产业发展规划的通知:工信部规(2021) 179 号

[10]国家发展改革委等部门. 国家发展改革委等部门关于印发《完善碳排放统计核算体系工作方案》的通知: 发改环资〔2024〕1479 号
