

工业互联网平台 工业大数据通用技术要求

1 范围

本文件规定了工业互联网平台的数据集成、存储、处理、分析、应用、安全能力，以及平台运维能力和数据安全能力。

本文件适用于指导工业互联网平台提供者的工业大数据服务能力建设，也适用于第三方机构对工业互联网平台提供者的工业大数据服务能力进行评估。

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 25069 信息安全技术 术语

GB/T 35295 信息技术 大数据 术语

GB/T 41778 信息技术 工业大数据 术语

3 术语和定义

GB/T 25069、GB/T 35295、GB/T 41778界定的以及下列术语和定义适用于本文件。

3.1

数据服务 data service

提供数据采集、数据传输、数据存储、数据处理（包括计算、分析、可视化等）、数据交换、数据销毁等数据生存形态演变的一种网络信息服务。

[来源：GB/T 41778—2022，3.8]

3.2

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 41778—2022，3.1]

3.3

数据 data

信息的可再解释的形式化表示，以适用于通信、解释或处理。

注：可以通过人工或自动手段处理数据。

[来源：GB/T 41778—2022，3.4]

3.4

实时计算 real-time computing

在规定时间内对数据进行处理及运算的操作。

注：在工业场景下的实时计算时间要求一般为秒级及以下，且主要包括数据的实时采集与存储、数据的实时处理与分析等两部分。

[来源：GB/T 41778—2022，3.43]

3.5

流式计算 stream computing

能够对流式数据进行实时处理的计算方式。

注：工业大数据相关的流式计算通常应用于工业生产活动中的实时监控、实时预警等场景。

[来源：GB/T 41778—2022，3.48]

3.6

结构化数据 structured data

一种数据表示形式，按此种形式，由数据元素汇集而成的每个记录的结构都是一致的并且可以使用关系模型予以有效描述。

[来源：GB/T 41778—2022，3.50]

3.7

非结构化数据 unstructured data

不具有预定义模型或未以预定义方式组织的数据。

[来源：GB/T 41778—2022，3.54]

3.8

垂直扩展 vertical scaling

为提高性能而提高处理速度、存储和内存等系统参数的过程。

[来源：GB/T 35295—2017，2.1.16]

3.9

水平扩展 horizontal scaling

将集成的一群个体资源作为一个单系统使用的过程。

[来源：GB/T 35295—2017，2.1.17]

3.10

分布式文件系统 distributed file system

多个结构化数据集分布在一个或多个服务器集群的各个计算节点的文件系统。

注：此类系统中，数据可能分布在文件和/或数据集层，更为普遍的是在数据块这个层级分布，同时支持集群中多个节点与大型文件和/或数据集的不同部分交互。

[来源：GB/T 35295—2017，2.1.21]

3.11

分布式计算 distributed computing

一种覆盖存储层和处理层的、用于实现多类型程序设计算法模型的计算模式。

注：分布式计算结果通常加载到分析环境。MapReduce是数据分布式计算中默认的处理构件。

[来源：GB/T 35295—2017，2.1.22]

3.12

分散-聚集 scatter-gather

大数据集的处理形式，其中所需的计算被划分并分布在集群的多个节点上，整体结果由每个节点的结果合并而成。

注：分散聚集通常要求对处理软件的算法进行改变。

示例：MapReduce(包含 Map 和 Reduce 两个计算过程的一种计算模型)就是采用分散-聚集的处理形式。

[来源：GB/T 35295—2017，2.1.23]

3.13

集群管理 cluster management

在以非关系模型方式驻留数据的集群资源之间提供通信的一种机制。

[来源：GB/T 35295—2017，2.1.59]

3.14

数据管理 data management

在数据处理系统中，提供对数据的访问，执行或监视数据的存储，以及控制输入输出操作等功能。

[来源：GB/T 35295—2017，2.2.3]

3.15

元数据 metadata

关于数据或数据元素的数据（可能包括其数据描述），以及关于数据拥有权、存取路径、访问权和数据易变性的数据。

[来源：GB/T 35295—2017，2.2.7]

3.16

数据挖掘 data mining

从大量的数据中通过算法搜索隐藏于其中信息的过程。

注：一般通过包括统计、在线分析处理、情报检索、机器学习、专家系统（依靠过去的经验法则）和模式识别等方法来实现。

[来源：GB/T 35295-2017, 2.2.9]

3.17

键值结构 key-value structure

一种存储范例，按此范例存储的记录由值中的关键词和一串数据组成。

注：在键值结构下，数据通过关键词检索，同时由非关系数据库软件处理对于值中的数据的访问。其作用相当于带单一索引字段和列的关系数据库表的子集和/或简化版。其变体之一是文档存储，在这种情况下，文档有多值字段，其中任何一个都可以用作索引/关键词。

[来源：GB/T 35295—2017, 2.2.15]

4 缩略语

下列缩略语适用于本文档

AI：人工智能（Artificial Intelligence）
API：应用程序接口（Application Program Interface）
CPU：中央处理器（Central Processing Unit）
CSV：逗号分隔值（Comma-Separated Values）
DDL：数据库模式定义语言（Data Definition Language）
DML：数据操纵语言（Data Manipulation Language）
JDBC：Java数据库连接（Java Database Connectivity）
HTTP：超文本传输协议（Hypertext Transfer Protocol）
IP：网际互连协议（Internet Protocol）
JSON：JS对象简谱（JavaScript Object Notation）
MFA：多重身份验证（Multifactor Authentication）
ODBC：开放数据库连接（Open Database Connectivity）
RPC：远程过程调用（Remote Procedure Call）
SDK：软件开发工具包（Software Development Kit）
SQL：结构化查询语言（Structured Query Language）
SSL：安全套接层（Secure Socket Layer）
TLS：安全传输层协议（Transport Layer Security）
TPC-DS：交易处理性能委员会的评估决策支持系统的基准测试（Transaction Processing Performance Council - Decision Support）
TTL：生存时间值（Time To Live）
UDF：用户自定义函数（User-Defined Function）
Web：全球广域网（World Wide Web）
XML：可扩展标记语言（Extensible Markup Language）

5 概述

工业大数据平台逻辑架构示意图见图1。本文件从数据集成、数据存储、数据处理、数据分析、数据应用、平台运维、数据安全7个方面规定了工业互联网平台提供者的工业大数据服务能力要求，其中：

- a) 数据集成能力：数据采集、数据清洗与预处理、数据映射与转换；
- b) 数据存储能力：分布式存储；
- c) 数据处理能力：实时流、大数据批处理、宽表数据库；
- d) 数据分析能力：交互式分析、图分析、数据搜索、AI 模型；
- e) 数据应用能力：数据应用；
- f) 平台运维能力：监控告警、健康检查、日志管理；
- g) 数据安全能力：数据集成、存储、处理、分析、应用安全等。

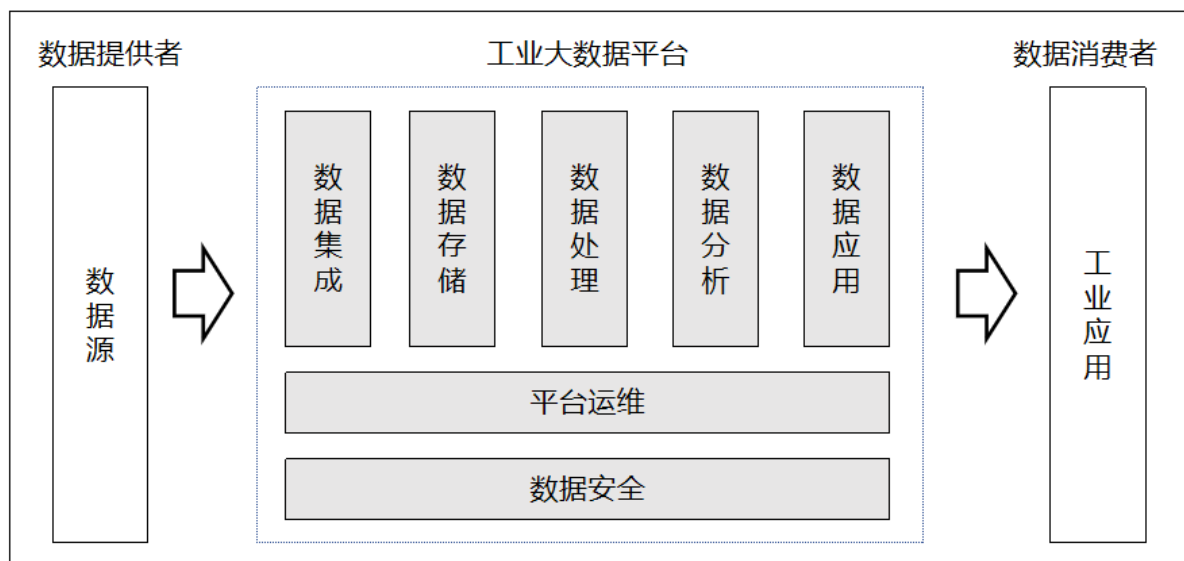


图1 工业大数据平台逻辑架构示意图

6 数据集成能力要求

6.1 数据采集

6.1.1 功能性

数据采集功能应符合下列要求：

- a) 支持数据采集的新增、删除和编辑功能；
- b) 支持常见的采集方法，包含且不限于文件、syslog、系统性能、数据库等数据源采集；
- c) 支持文件采集黑白名单；
- d) 支持文件路径或名称自适应能力；
- e) 支持数据的实时采集和批量采集；
- f) 支持断点续传；
- g) 支持自动生成目标表 DDL；
- h) 支持常见数据库作为源及目标端；
- i) 支持自定义采集列及自定义筛选条件；
- j) 支持增量采集及全量采集。

6.1.2 性能效率

数据采集性能效率应符合下列要求：

- a) 常见数据库采集速率应大于每秒 10000 条；
- b) 实时采集速率应大于每秒 1000 条；
- c) 数据实时采集延迟应小于 1 秒；
- d) 数据采集并发连接数应大于 10000 个；
- e) 批量采集每小时应大于 1TB；
- f) 在高负载情况下，CPU 利用率应保持在 70%-80%之间，避免过高负载导致性能瓶颈；
- g) 在高负载情况下，内存利用率应保持在 60%-70%之间，确保有足够的内存应对突发负载；

6.1.3 可靠性

数据采集可靠性应符合下列要求：

- a) 支持系统稳定运行；
- b) 数据采集过程中数据丢失率应低于 0.01%；

- c) 数据采集过程中数据准确率应达到 99.999%;
- d) 支持强制关机、断电、重启后,系统可快速恢复,故障恢复时间应小于 5 分钟。

6.1.4 易用性

数据采集易用性应符合下列要求:

- a) 提供友好、直观的用户界面,使用户能轻松配置和管理数据采集任务;
- b) 支持拖放操作来创建和配置数据采集 workflow,简化复杂操作;
- c) 具备配置向导,引导用户逐步完成数据源配置、数据采集规则设定等步骤;
- d) 具备常见数据源的预设模板,用户可以快速选择和应用,减少手动配置的工作量;
- e) 支持按时间间隔、特定时间点、事件驱动等多种调度策略,满足不同数据采集需求;

6.1.5 可扩展性

数据采集可扩展性应符合下列要求:

- a) 支持通过添加新的采集节点来增加整体的处理能力,扩展后性能应线性增加;
- b) 支持自定义数据源,自定义数据格式解析的插件化机制;
- c) 支持扩容、缩容不中断;
- d) 支持增加或减少节点过程服务不中断。

6.1.6 兼容性

数据采集兼容性应符合下列要求:

- a) 软件兼容:支持主流国产操作系统;
- b) 硬件兼容:支持主流国产服务器。

6.2 数据清洗与预处理

6.2.1 功能性

数据清洗与预处理功能应符合下列要求:

- a) 支持针对不同类型和等级的数据,制定相应的数据清洗加工策略;
- b) 支持采用支撑数据验证、清洗、标准化、格式化、存储和版本管理的方法或工具,开展数据预处理;
- c) 支持检查数据一致性,去除、补全或修改有缺失的、格式和内容错误的、逻辑错误的和不需要的无效值、缺失值或重复值;
- d) 支持生成数据预处理或数据清洗加工报告;
- e) 支持删除完全重复的记录或根据特定规则删除重复值;
- f) 自动将数据转换为统一格式,如日期格式、数值格式、字符串格式等;
- g) 支持不同数据类型之间的转换,如字符串到数值、数值到日期等。

6.2.2 性能效率

数据清洗与预处理性能效率应符合下列要求:

- a) 支持每秒处理大于 1,0000 条数据;
- b) 高效利用多核 CPU 资源,处理过程中 CPU 利用率应保持在 80%以上;
- c) 高效利用系统内存,利用率应在 70%-90%之间,确保处理大数据集时的稳定性;

6.2.3 可靠性

数据清洗与预处理可靠性应符合下列要求:

- a) 支持系统稳定运行;
- b) 在数据清洗与预处理过程中,不应丢失任何原始数据或处理结果;
- c) 处理结果应确保一致性,不受系统状态或环境变化的影响;
- d) 支持记录和审计数据清洗与预处理过程中的所有操作和变更,以确保数据处理的透明性和可追溯性;

6.2.4 易用性

数据清洗与预处理易用性应符合下列要求：

- a) 数据清洗与预处理规则的版本控制，应确保规则修改后能回溯和恢复到先前的版本；
- b) 能自动检测和标记数据集中缺失值的位置，如提供多种处理缺失值的方法，如删除、插值、填充默认值、前向填充、后向填充等；
- c) 具备完善文档说明及操作引导；
- d) 具备内置的数据清洗及预处理规则。

6.2.5 可扩展性

数据清洗与预处理可扩展性应符合下列要求：

- a) 支持扩容、缩容不中断；
- b) 支持增加或减少节点过程中，服务不中断；
- c) 支持通过添加更多计算节点来处理更大规模的数据量；

6.2.6 兼容性

数据清洗与预处理兼容性应符合下列要求：

- a) 操作系统兼容：支持主流国产系统；
- b) CPU 兼容：支持主流国产芯片。

6.3 数据映射与转换

6.3.1 功能性

数据映射与转换功能应符合下列要求：

- a) 支持不同数据源数据结构的转换，包括数据类型转换、字段格式转换等，确保数据在目标系统中能被正确识别和处理；
- b) 支持将源系统中的字段映射到目标系统中的对应字段，确保数据能够正确地映射和识别，如将源系统中的“姓名”字段映射到目标系统中的“Name”字段；
- c) 能够将数据从一种格式转换成另一种格式，包括文本格式、JSON 格式、XML 格式等，以适应不同系统的需求；
- d) 支持将多个数据源的数据合并成一个数据集，或者将一个数据集拆分成多个部分，以便后续处理；
- e) 支持对数据进行转码和编码，以适应不同系统和环境的要求，确保数据被正确地解析和处理。

6.3.2 性能效率

数据映射与转换性能效率应符合下列要求：

- a) 高效利用多核 CPU 资源，处理过程中 CPU 利用率应保持在 80%以上；
- b) 高效利用系统内存，利用率应在 70%-90%之间，确保大数据集处理的稳定性；
- c) 单条记录的映射和转换延迟应控制在毫秒级别；
- d) 支持同时处理 10 个并发数据流，每个数据流每秒处理 10000 条记录。

6.3.3 可靠性

数据映射与转换可靠性应符合下列要求：

- a) 数据在映射和转换过程中应保持一致性，避免数据丢失、重复或不一致；
- b) 保证数据转换的准确性，确保数据转换结果符合预期；
- c) 具备自动故障检测和恢复机制，在计算节点故障时，自动切换任务到其他节点继续执行，并保证数据不丢失；
- d) 支持通过多节点冗余和负载均衡，确保在高负载和部分节点失效情况下仍能正常运行；
- e) 支持对数据转换操作进行完整性校验，确保数据没有被意外修改或丢失。

6.3.4 易用性

数据映射与转换易用性应符合下列要求：

- a) 支持通过拖放操作定义数据映射和转换规则，无需编写复杂代码；
- b) 具有详细的文档和视频教程，帮助用户快速上手；
- c) 具有常用的数据转换模板，供用户快速应用；
- d) 支持用户自定义数据转换规则和编写自定义函数；
- e) 支持用户实时预览数据转换效果，并通过调试工具修复错误；
- f) 支持多种编程语言进行数据处理脚本编写。

6.3.5 可扩展性

数据映射与转换可扩展性应符合下列要求：

- a) 支持在不改变系统架构的情况下，通过增加单个节点的硬件资源来提升性能；
- b) 支持通过添加更多计算节点来处理更大规模的数据量；
- c) 支持与其他大数据工具和平台无缝集成，支持多种数据格式和通信协议；
- d) 数据映射和转换模块独立设计，能单独扩展或升级。

6.3.6 兼容性

数据映射与转换兼容性应符合下列要求：

- a) 操作系统兼容：支持主流国产系统；
- b) CPU 兼容：支持主流国产芯片。

7 数据存储能力要求

7.1 概述

数据存储能力在工业互联网平台主要是分布式存储能力，包括功能性、效力性能、可靠性、易用性、维护性、可扩展性、兼容性等要求。

7.2 分布式存储功能性

分布式存储功能应符合下列要求：

- a) 支持结构化、半结构化、非结构化数据类型的存储和读取；
- b) 支持数据生命周期配置，指定生效时间、对象保留时间及过滤条件等规则来实现定时自动删除数据对象；
- c) 支持在不同地域的存储空间之间自动、异步复制文件，将源存储空间中文件的改动（新建、覆盖、删除操作）同步到目标存储空间中；
- d) 支持热、温、冷等 3 种及以上存储类别；
- e) 支持集中和分布式管理元数据能力；
- f) 支持 JDBC、接口等模式访问元数据；
- g) 支持数据存储副本数配置的能力；
- h) 支持请求不同服务器的数据一致性和准确性；
- i) 支持数据存储内容进行版本管理，上传同名对象时不覆盖原有对象，而是产生一个新的版本对象作为当前版本独享；
- j) 分布式存储服务软件具备友好的部署能力，线上平台支持友好的服务开通能力；
- k) 支持分布式对象存储服务配置进行动态修改的能力；
- l) 支持管理员对分布式存储进行审计、权限管理；
- m) 支持添加新用户、创建用户账户、设置用户密码、查看当前存在的用户、删除用户；
- n) 支持通过用户名和密码对用户身份进行认证；
- o) 支持在分布式节点、关键部件，如 CPU、内存、硬盘、网卡等运行状态异常下进行告警；
- p) 支持查看在线或离线的用户操作日志，用户操作日志记录每个用户具体访问时间和操作内容等详细信息；
- q) 支持查看在线或离线的系统日志，系统日志记录系统运行情况；

- r) 日志支持导出或备份操作；
- s) 支持对包括但不限于存储空间信息，文件对象数量、大小，服务流量变化，服务可用性，延迟等；
- t) 具备自定义数据源接入的能力和自定义数据格式解析的插件化机制，方便用户根据需要扩展。

7.3 分布式存储性能效率

分布式存储性能效率应符合下列要求：

- a) 支持至少 1000 个存储节点，且增加节点后系统性能线性提升；
- b) 至少支持 PB 级别的数据存储容量（如 10 PB 以上）；
- c) 数据压缩率应达到 50%以上，即存储数据量减少至少 50%；
- d) 在高负载情况下，CPU 利用率应保持在 70%-80%之间，避免过高负载导致性能瓶颈；
- e) 内存利用率应保持在 60%-70%之间，确保有足够的内存应对突发负载；
- f) 具备元数据秒级读取能力。

7.4 分布式存储可靠性

分布式存储可靠性应符合下列要求：

- a) 数据跨机房存储：支持写入的数据在切块后，散落在不同的服务器中存储；
- b) 数据跨地域容灾：支持数据对象自动化跨地域进行复制；
- c) 元数据信息同步：支持元数据自动备份，实现元数据冷热访问模式；
- d) 硬件故障：支持电源插拔，硬盘插拔，网线插拔，交换机，机架，机房故障下的服务能力（故障监测、自动切换、副本自动补齐）；
- e) 操作系统故障：支持 CPU 资源占用，输入/输出（I/O）资源占用，内存资源占用，磁盘空间占用下数据存储服务能力（故障监测、自动切换）；
- f) 单节点故障：支持单节点故障后，数据请求自动指向其他副本数据进行访问。

7.5 分布式存储易用性

分布式存储易用性应符合下列要求：

- a) 支持可视化安装部署分布式存储系统；
- b) 支持可视化进行数据访问；
- c) 支持对分布式存储的相关参数进行可视化配置。

7.6 分布式存储维护性

分布式存储维护性应符合下列要求：

- a) 错误检测和提示：支持对用户错误操作、命令或非法数据输入的检验能力和提示情况，要求出错信息中提供差错产生的原因；
- b) 日志管理：支持对所有操作行为进行日志记录，支持关键词查看日志错误问题定位。

7.7 分布式存储可扩展性

分布式存储可扩展性应符合下列要求：

- a) 节点动态扩展：支持分布式存储节点动态增加；
- b) 扩容、缩容不中断：支持增加或减少节点过程中，服务不中断。

7.8 分布式存储兼容性

分布式存储兼容性应符合下列要求：

- a) 操作系统兼容：支持主流国产系统；
- b) CPU 兼容：支持主流国产芯片；
- c) 协议兼容：HTTP、超文本传输安全协议（HTTPS）等协议兼容；
- d) 开发语言兼容：支持 Java、C/C++、Python 等主流开发语言的访问兼容。

8 数据处理能力要求

8.1 实时流

8.1.1 功能性

数据处理的实时流功能性应符合下列要求：

- a) 流数据源：支持常见流数据源，如传输控制协议（TCP）、用户数据报协议（UDP）、HTTP、消息系统(Kafka, RocketMQ 等)；
- b) 表数据源：支持常见表数据源，例如通用 JDBC、ODBC，专用特定数据库接口访问等；
- c) 事件数据格式：支持常见数据格式(CSV、JSON、AVRO、XML 等)和自定义格式解析；
- d) 简单事件处理：支持逐一对数据流中单个事件进行提取、清洗、转换等处理；
- e) 复杂事件处理：支持持续地从事件流中查询出符合要求的事件序列，进行过滤、关联、聚合等处理；
- f) 事件窗口和水位：支持滑动和滚动等多种窗口划分，支持事件水位设置；
- g) 支持通过 SQL 对实时流数据进行操作，完成常用实时数据处理功能；
- h) 缓冲和背压：支持对事件进行缓冲和背压设置；
- i) 支持多数据源及多输出目标；
- j) 事件优先级：支持事件优先级和按优先处理的能力。

8.1.2 性能效率

数据处理的实时流性能效率应符合下列要求：

- a) 简单事件处理性能：对于简单事件处理，单机性能不应低于每秒 50000 条；
- b) 复杂事件处理性能：对于复杂事件处理，单机性能不应低于每秒 5000 条；
- c) 简单事件处理延迟：在事件流以每秒 50000 条到达，水位为 500 毫秒情况下，延迟时间中值不应高于 1s，最大值不应高 2s；
- d) 复杂事件处理延迟：在事件流以每秒 5000 条到达，水位为 500 毫秒情况下，延迟时间中值不应高于 2 s，最大值不应高于 5 s。

8.1.3 可靠性

数据处理的实时流可靠性应符合下列要求：

- a) 长时间负载测试：多个实时流并发处理、运行时间 7×24h，事件处理成功率不应低于 95%；
- b) 单点故障测试：支持集群部署，单点故障下系统功能正常；
- c) 处理一致性：支持一次和至少一次语义；
- d) 状态存储和容错：支持定期状态存储，集群或单机全部停机，重启后系统能从存储的状态中恢复。

8.1.4 易用性

数据处理的实时流易用性应符合下列要求。

- a) 手册的一致性和准确性。人机交互界面操作与其相应的文档一致、准确。
- b) 界面的规范性。人机交互界面规范、合理和一致。
- c) 界面元素的易理解性
 - 1) 人机交互界面应使用适当的术语、图形表示背景信息和帮助；
 - 2) 人机交互界面上窗口、菜单、图标、文字、鼠标等界面元素能准确理解。
- d) 易浏览性
 - 1) 具有必要的信息，指导用户使用程序；
 - 2) 输出结果应简洁、直观、方便阅读、易懂和使用；
 - 3) 人机界面风格相对一致；
 - 4) 界面、人机交互、输出中的用语应与业务用语一致；
 - 5) 菜单深度不超过 3 级。
- e) 易操作性
 - 1) 提供辅助输入手段（如选择输入、默认值等），数据检索方便、灵活；
 - 2) 安装参数应给出默认值或提示。

f) 辅助工具。提供图形化的辅助工具。

8.1.5 维护性

数据处理的实时流维护性应符合下列要求：

- a) 错误检测和提示：具备对用户错误操作/命令或非法数据输入的检验能力和提示能力，出错信息中应提供差错产生的原因；
- b) 日志信息：支持输出实时日志，对实时记录人为操作及错误信息记录等；
- c) 配置管理：支持各类配置参数的修改；
- d) 导入与导出：支持配置导入与导出。

8.1.6 可扩展性

数据处理的实时流可扩展性应符合下列要求：

- a) 垂直扩展：支持垂直扩展，性能随硬件提高而提高；
- b) 水平扩展：支持集群部署，支持水平扩容。

8.1.7 兼容性

数据处理的实时流兼容性应符合下列要求：

- a) 软件兼容：支持主流国产操作系统；
- b) 硬件兼容：支持主流国产服务器。

8.2 大数据批处理

8.2.1 功能性

大数据批处理的功能性应符合下列要求。

- a) 数据抽取的数据源支持多源异构
 - 1) 支持关系型数据库、数据仓库等结构化数据；
 - 2) 支持文本格式数据（TXT、CSV、JSON 等）半结构化数据；
 - 3) 支持多媒体数据等非结构化数据。
- b) 数据抽取过程支持多种导入方式
 - 1) 支持通过 SQL 语法导入数据；
 - 2) 支持通过表述性状态转移（REST）接口规则导入数据；
 - 3) 支持通过文本传输协议（FTP）等协议导入数据。
- c) 数据转换过程支持多元操作
 - 1) 支持将相同类型数据的整合，进行一致性转换；
 - 2) 支持数据不同粒度级别的转换；
 - 3) 支持根据相应业务规则进行数据聚合、排序等操作。
- d) 数据转换过程支持数据清洗策略
 - 1) 支持对不完整数据的补全或者剔除；
 - 2) 支持对数据进行校验，剔除错误数据；
 - 3) 支持对数据进行去重清洗。
- e) 具备对多源异构数据的多元存储策略
 - 1) 支持关系型数据库、非关系型数据库、内存数据库、分布式文件系统、数据仓库等多种存储策略；
 - 2) 批处理过程支持多种方式 支持通过 SQL 脚本进行批处理操作；
 - 3) 支持通过计算机壳层（shell）脚本支持批处理操作；
 - 4) 支持通过 java 等开发批处理服务执行批处理操作。
- f) 批处理过程支持图形化界面设计
 - 1) 支持采用图形化界面设计批处理流程；
 - 2) 批处理流程可拆分为多个子流程，支持以图形化的方式设计子流程间的调度关系。
- g) 批处理过程支持任务调度

- 1) 支持采用定时任务启动、停止批处理任务;
- 2) 支持以分、时、日、周、月、年等单位的时间进行调度。
- h) 数据分析集成相应的机器学习算法。支持常见的机器学习模型训练(分类、聚类、回归、异常点检测、关联规则等)。
- i) 数据分析支持图表等可视化功能。支持以饼图、柱状图、条形图、气泡图、热力图、瀑布图、突出表、折线图、散点图、交叉表等进行数据可视化呈现。
- j) 数据分析支持对分析结果的上报及存储。支持数据分析结果存储至数据库或者以 REST 接口方式提交给上层应用。

8.2.2 性能效率

大数据批处理的性能效率满足下列要求:

- a) 数据导入性能: 记录入库 1T 数据的时间, 不应高于 10h;
- b) 数据导出性能: 记录出库 1T 数据的时间, 不应高于 10h;
- c) 读写性能: 1T 数据装载后, 小表(1w)、中表(100w)、大表(1000w)的精确查询或插入单条记录时间应分别小于 200ms、1s、5s;
- d) 数据吞吐量: 在 10s 内处理的数据量, 吞吐速率应低于 10s;
- e) 存储性能: 集群在单实例模式下数据入库的数据容量上限, 不应低于 500TB;
- f) 支持 5000 以上并行任务;
- g) 系统应达到 99.9%的年度可用性, 即每年最多允许 8.76 小时的非计划停机时间;
- h) 集群节点数可扩展到 1000 以上。

8.2.3 可靠性

大数据批处理的可靠性应符合下列要求:

- a) 长时间负载测试: 支持在 7×24 小时内进行批处理任务, 且保证全时段执行正常;
- b) 完全备份/还原: 支持完全备份、还原, 实现全库所有对象的备份和还原;
- c) 误操作恢复: 支持对索引、表的删除进行还原;
- d) 集群故障处理: 支持磁盘损坏后, 集群节点可以正常读写操作不影响集群业务;
- e) 高可用能力: 支持集群节点挂起时, 集群可以正常提供服务;
- f) 副本能力: 支持多副本存储。

8.2.4 易用性

大数据批处理的易用性应符合下列要求:

- a) 各组件界面应风格一致, 简洁、美观;
- b) 主界面到各组件界面跳转分级明确, 操作逻辑一致;
- c) 控件使用一致, 排列整齐、合理有序;
- d) 菜单和按钮命名合理, 具有相应的图标或文本提示信息;
- e) 必填项未填时应有提示信息, 用户提交表单时若存在必填项未填应有明确、醒目的提示;
- f) 页面提示信息与内容一致;
- g) 页面应对显示内容有筛选;
- h) 用户提交的内容应具备可更改功能;
- i) 重要、常用功能放在列表、菜单的前位;
- j) 不常用功能不应干扰用户正常使用、占用较少的页面面积或页面具备不常用功能隐藏、显示功能;
- k) 页面进行高权限、重大操作时对用户应进行提示;
- l) 页面报警或提示时, 应采用醒目的界面对用户进行提示;
- m) 对用户输入的内容应进行自动纠正, 如日期格式错误、文本框限制输入内容、数字超出范围自动限制等;
- n) 系统后台应具备存储操作日志功能;
- o) 系统页面应具备帮助、手册、向导等内容;
- p) 系统应支持补丁方式升级、升级时保证原有数据的完整性和继承性;

- q) 系统应具备友好的安装部署交互方式;
- r) 系统安装后应自动启动用户图形化界面服务, 并且图形化界面可用。

8.2.5 维护性

大数据批处理的维护性应符合下列要求:

- a) 具备维护手册、相关文档;
- b) 具备维护手册、文档, 用于指导运维人员进行维护;
- c) 具备平台状态监控和故障提示功能;
- d) 具备状态监控和故障提示功能;
- e) 具备故障日志, 系统应存储故障日志, 帮助运维人员分析故障所在;
- f) 具备组件管理、替换和重置能力;
- g) 具备备份与误操作还原功能;
- h) 具备还原功能, 在出现故障或误操作时可对平台进行回溯;
- i) 具备升级功能。

8.2.6 可扩展性

大数据批处理的可扩展性应符合下列要求:

- a) 在线扩容: 支持存储节点和计算节点各自扩容的能力, 以满足突发的业务请求;
- b) 分布式存储测试: 验证测试图分析软件数据是否均衡分布存储在各个节点;
- c) 分布式计算测试: 验证测试图分析软件计算任务是否平均分布到各个节点。

8.2.7 兼容性

大数据批处理的兼容性应符合下列要求:

- a) 浏览器兼容: 支持在不同浏览器上可以正常运行, 功能正常;
- b) 屏幕尺寸和分辨率兼容: 支持在不同分辨率下正常显示;
- c) 操作系统兼容: 支持在主流国产操作系统上正常运行, 功能正常, 显示正确等;
- d) 硬件兼容: 支持主流国产服务器。

8.3 宽表数据库要求

8.3.1 功能性

宽表数据库的功能性应符合下列要求:

- a) 支持 DDL 操作, 如 create、alter、describe、disable、drop、list 等;
- b) 支持 DML 操作, 如 put、get、scan、delete、append 等;
- c) 支持命名空间操, 支持命令空间操作, 如 alter_namespace, create_namespace, describe_namespace, drop_namespace, list_namespace 等;
- d) 支持列簇级别生命周期配置, 根据生命周期定期清理数据。支持表的列簇设置 TTL, 可根据 TTL 定期清理过期数据;
- e) 支持表格压缩, 如 PREFIX、DIFF、FAST_DIFF、ROW_INDEX_V1;
- f) 支持表格压缩, 如 GZIP、snappy、zstd 等;
- g) 支持分区服务分组管理;
- h) 支持二级索引, 具备按照某些列的值进行索引的能力。

8.3.2 性能效率

宽表数据库应支持数据批量入库, 数据量200GB, 加载速率不应低于30MB/Node/Sec。

8.3.3 可靠性

宽表数据库的可靠性符合下列要求:

- a) 长时间负载测试: 应支持在 7×24h 长时间负载运行;
- b) 备份/恢复: 应支持元数据备份/恢复; 支持业务数据备份/恢复;

以上内容仅为本文档的试下载部分，为可阅读页数的一半内容。如要下载或阅读全文，请访问：<https://d.book118.com/056024154005010144>