



中华人民共和国国家标准

GB/T 38673—2020

信息技术 大数据 大数据系统基本要求

Information technology—Big data—Basic requirements for big data systems

2020-04-28 发布

2020-11-01 实施

国家市场监督管理总局 发布
国家标准化管理委员会

目 次

前言 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 1

5 大数据系统框架 2

6 功能要求 3

6.1 数据收集模块 3

6.2 数据预处理模块 3

6.3 数据存储模块 3

6.4 数据处理模块 4

6.5 数据分析模块 5

6.6 数据可视化模块 6

6.7 数据访问模块 6

6.8 资源管理模块 6

6.9 系统管理模块 6

7 非功能要求 6

7.1 可靠性要求 6

7.2 兼容性要求 7

7.3 安全性要求 7

7.4 可扩展性要求 8

7.5 维护性要求 8

7.6 易用性要求 8

前 言

本标准按照 GB/T 1.1—2009 给出的规则起草。

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。

本标准由全国信息技术标准化技术委员会(SAC/TC 28)提出并归口。

本标准起草单位：中国电子技术标准化研究院、华为技术有限公司、北京大学、中国人民大学、中兴通讯股份有限公司、浪潮电子信息产业股份有限公司、阿里云计算有限公司、天津南大通用数据技术有限公司、北京百分点信息科技有限公司、复旦大学、南京大学、东南大学、北京和仲宁信息技术有限公司、北京启迪区块链科技发展有限公司。

本标准主要起草人：梅宏、孙文龙、杜小勇、吴东亚、董建、张群、尹卓、许洁、李冰、李瑛、高琨、朱松、赵江、张展新、梁佳男、赵俊峰、符海芳、卫凤林、孙嘉阳、赵菁华、陈晋川、刘海军、孙伟、姜育刚、周志华、张敏灵。



信息技术 大数据

大数据系统基本要求

1 范围

本标准规定了大数据系统的功能要求和非功能要求。

本标准适用于各类大数据系统要求的设计、选型、验收和检测。

2 规范性引用文件

下列文件对于本文件的应用是必不可少的。凡是注日期的引用文件,仅注日期的版本适用于本文件。凡是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 35589—2017 信息技术 大数据 技术参考模型

3 术语和定义

GB/T 35295—2017 界定的以及下列术语和定义适用于本文件。为了便于使用,以下重复列出了GB/T 35295—2017 中的某些术语和定义。

3.1

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。

[GB/T 35295—2017,定义 2.1.14]

3.2

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式的计算模式。

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。

[GB/T 35295—2017,定义 2.1.22]

3.3

集群 cluster

一组相互独立的、通过高速网络互联的计算机或服务器。

3.4

租户 tenant

对一组物理和虚拟资源进行共享访问的一个或多个云服务用户。

4 缩略语

下列缩略语适用于本文件。

- API:应用程序接口(Application Programming Interface)
- CPU:中央处理器(Central Processing Unit)
- DAG:有向无环图(Directed Acyclic Graph)
- OLAP:联机分析处理(On-Line Analytical Processing)
- REST:表述性状态转移(Representational State Transfer)
- SQL:结构化查询语言(Structured Query Language)

5 大数据系统框架

GB/T 35589—2017 定义了大数据参考架构,如图 1 所示。大数据参考模型是一个通用的大数据系统概念模型,它表示了通用的、与技术无关的大数据系统的逻辑功能构件及构件之间的互操作接口,作为开发各种具体类型大数据应用系统架构的通用技术参考框架。

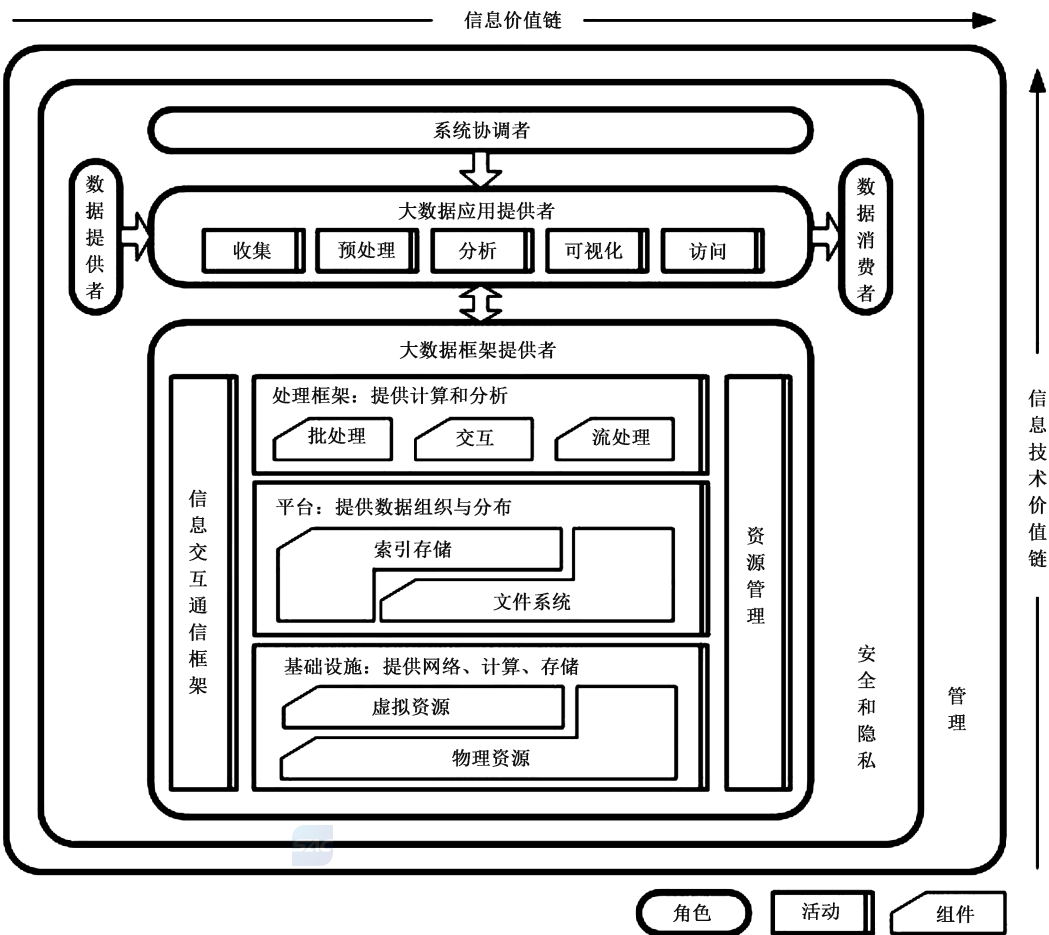


图 1 大数据参考架构

本标准参考大数据参考架构逻辑功能构件划分,将大数据系统划分为数据收集、数据预处理、数据存储、数据处理、数据分析、数据访问、数据可视化、资源管理、系统管理 9 个模块。大数据系统框架如图 2 所示。

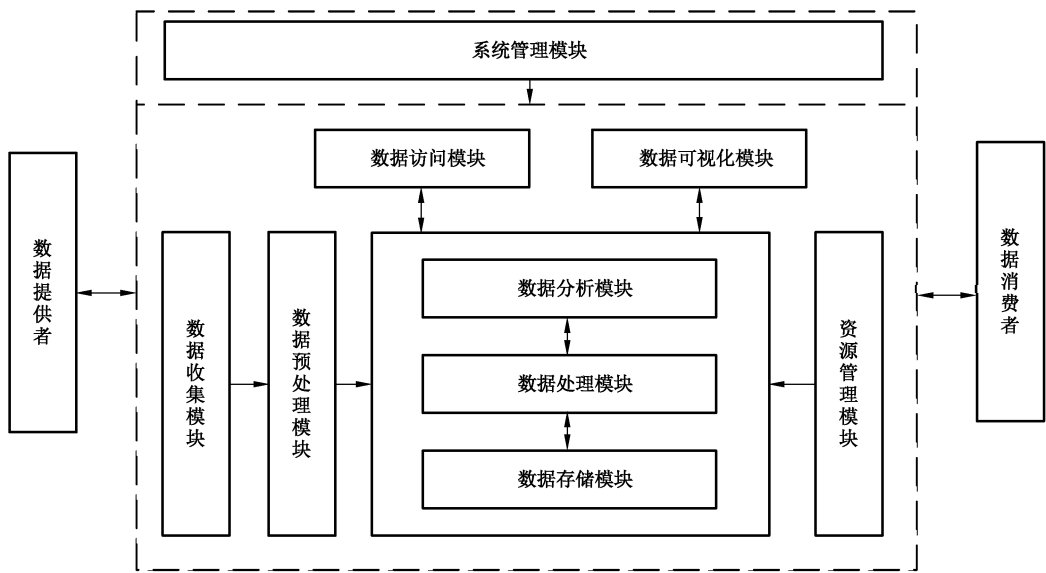


图 2 大数据系统框架

6 功能要求

6.1 数据收集模块

数据收集模块要求如下：

- a) 应提供数据导入功能，支持结构化数据、非结构化数据和半结构化数据导入；
- b) 应支持离线数据导入和实时数据导入；
- c) 应支持全量数据导入和增量数据导入；
- d) 应提供自动定时导入数据功能；
- e) 宜提供开放的数据导入 API；
- f) 宜提供图形界面实现数据导入功能。

6.2 数据预处理模块

数据预处理模块要求如下：

- a) 应提供数据抽取功能，支持对结构化数据、非结构化数据和半结构化数据进行抽取；
- b) 应提供数据清洗功能，支持对不一致数据、无效数据、缺失数据和重复数据的处理；
- c) 应提供结构化数据的列转换、行转换和表转换功能；
- d) 应提供数据加载功能，支持将经过清洗和转换的数据加载到数据分析模块；
- e) 宜提供清洗前后的数据比对功能；
- f) 宜支持非结构化数据的数据转换功能。

6.3 数据存储模块

数据存储模块要求如下：

- a) 应提供数据存储功能，支持结构化数据、非结构化数据和半结构化数据存储。
- b) 应提供与关系型数据库、其他文件系统之间交换数据或文件的功能。
- c) 支持分布式文件存储，实现以下功能：

- 1) 应支持文件系统基本操作,包括上传、下载、读写、复制、移动、删除、重命名、权限修改等;
- 2) 应提供数据块多副本存储、恢复功能;
- 3) 宜支持文件快速检索功能,支持数据资源的统一检索、编目、增加和删除操作;
- 4) 宜支持数据压缩存储功能。
- d) 支持分布式列式数据存储,实现以下功能:
 - 1) 应支持以键值形式存储数据的功能;
 - 2) 宜支持基于表、列族和列的用户权限管理功能,权限管理操作包括读、写、创建等。
- e) 支持分布式结构化数据存储,实现以下功能:
 - 1) 宜支持结构化数据的分布式存储,保证数据存储的可扩展性和一致性;
 - 2) 宜提供 API 实现数据的各类查询操作;
 - 3) 宜支持多表关联。
- f) 支持分布式图数据存储,实现以下功能:
 - 1) 宜支持由节点及边组成的数据模型;
 - 2) 宜支持图查询,支持单节点、多节点多层关系的扩线查询;
 - 3) 宜支持图遍历,支持最短路径、最优路径遍历查询;
 - 4) 宜支持图分析。

6.4 数据处理模块



数据处理模块要求如下:

- a) 支持批处理框架,实现以下功能:
 - 1) 应支持结构化数据、非结构化数据和半结构化数据的离线分析;
 - 2) 应支持多节点离线任务联动执行;
 - 3) 应支持分散-聚集的处理方式;
 - 4) 宜支持多种开发语言接口。
- b) 支持流处理框架,实现以下功能:
 - 1) 应提供实时计算功能,并将计算结果输出到消息队列或持久化;
 - 2) 应支持采用滑动窗口方式的实时分析任务,时间窗口大小可调;
 - 3) 应提供容错机制,出现故障时,可对故障进行处理;
 - 4) 宜提供用户级别的访问控制功能,支持对消息处理任务进行创建、浏览、中止、恢复等操作,并记录操作日志。
- c) 宜支持图计算框架,实现以下功能:
 - 1) 内置图数据查询类 API,支持同步或异步计算模型编写迭代算法;
 - 2) 在线图分析和查询功能;
 - 3) 基于属性图模型的图数据表达,包含节点/边上的标签和属性类型定义;
 - 4) 内置常用图指标计算功能,用以描述图的拓扑结构特征。
- d) 宜支持内存计算,实现以下功能:
 - 1) 通过分布式内存计算和 DAG 执行引擎提供数据处理能力;
 - 2) 支持多种数据类型,包括结构化数据、非结构化数据、半结构化数据的数据处理。
- e) 宜支持批流融合计算框架,实现以下功能:
 - 1) 批流融合统一查询 SQL 语言;
 - 2) 多场景下的流式 SQL,如位置信息分析等;
 - 3) 常用时间窗口,包括跳跃窗口、滑动窗口等。
- f) 宜支持按照任务间的依赖关系自动调度任务。

- g) 宜支持以有向无环图形式描述作业内多任务的依赖关系。
- h) 宜提供对复杂任务的调度能力。

6.5 数据分析模块

数据分析模块要求如下：

- a) 支持数据查询,实现以下功能：
 - 1) 应提供通过标准的数据库连接接口进行查询的功能；
 - 2) 应提供通过 REST API 查询接口进行查询的功能；
 - 3) 应提供建立数据索引的功能,达到查询加速的效果；
 - 4) 应支持精确查询和模糊查询功能。
- b) 支持机器学习,实现以下功能：
 - 1) 应提供数据集管理功能,可将数据划分为训练集、验证集和测试集；
 - 2) 应提供机器学习模型导入和导出功能；
 - 3) 应提供常用机器学习算法；
 - 4) 宜支持集成第三方机器学习算法。
- c) 支持统计分析,实现以下功能：
 - 1) 应提供基本数值统计,如最大值、最小值、求和、总数等统计量；
 - 2) 应提供数据集中趋势统计,如平均数、中位数、众数等统计量；
 - 3) 宜提供数据离散程度统计,如极差、方差、标准差等统计量；
 - 4) 宜提供随机变量关系的统计,如协方差、相关系统等统计量。
- d) 支持离线数据分析,实现以下功能：
 - 1) 应支持结构化查询语言；
 - 2) 应支持分布式计算或并行计算等计算框架；
 - 3) 宜支持对海量工作任务的切分和分布式调度。
- e) 支持流数据分析,实现以下功能：
 - 1) 应提供按时间切片进行批量处理的功能；
 - 2) 应支持基于事件触发或者采样的流式处理；
 - 3) 宜支持实时流上的数据统计；
 - 4) 宜支持流式数据的排序；
 - 5) 宜支持与静态表之间的关联；
 - 6) 宜支持多个数据流的关联处理。
- f) 宜支持交互式联机分析,实现以下功能：
 - 1) 通过结构化查询语言对数据进行分布式的联机分析,如 OLAP 等；
 - 2) 通过结构化查询语言对数据进行即席查询；
 - 3) 利用可视化中间件对数据分析结果进行显示；
 - 4) 在交互式分析过程中定义计算公式和参数配置；
 - 5) 在交互式分析过程中自动保存和回退；
 - 6) 在交互式分析过程中对分析结果的保存和发布；
 - 7) 基于在线联机分析的交互式数据分析。
- g) 宜支持可视化的流程编排操作,实现以下功能：
 - 1) 通过拖拽方式进行流程编排和修订；
 - 2) 支持工作流调度触发机制,可配置触发时间或触发事件；
 - 3) 支持流程编排结果的持久化保存。

6.6 数据可视化模块

可视化模块要求如下：

- a) 应支持使用常规图表展示数据,如表格、柱状图、饼图、折线图、热力图等;
- b) 宜支持第三方数据可视化工具的 API。

6.7 数据访问模块

数据访问模块应支持相应的访问接口,以便于第三方应用程序使用大数据系统的数据。

6.8 资源管理模块

资源管理模块要求如下：

- a) 应提供 CPU、内存等资源的调度和配置功能;
- b) 应提供对全局资源的集中管理功能;
- c) 应支持静态资源分配策略和动态资源分配策略;
- d) 应支持资源的弹性与抢占,即有空闲资源时,租户可使用超过其配置上限的资源,系统繁忙时,若租户使用的资源未达到其原始配置,则可抢占其他租户使用资源的超出部分;
- e) 宜提供设置任务优先级的功能,并按任务优先级对资源进行调度;
- f) 宜支持多层次的队列资源管理,队列资源实现隔离,即不为队列分配超过其资源上限的资源;
- g) 宜提供根据作业需求动态分配计算资源,自动管理回收资源功能。

6.9 系统管理模块

系统管理模块要求如下：

- a) 应提供配置管理功能,包括对大数据集群软硬件资源的配置管理,支持配置管理的分角色、分组管理及自动化;
- b) 应提供租户管理功能,包括租户的角色、权限、资源等功能;
- c) 应提供监控告警管理功能,包括多维度、可视化的大数据系统的监控、告警等;
- d) 应提供服务管理功能,包括对大数据系统组件服务的管理;
- e) 宜提供健康检查管理功能,支持以图形界面方式实现集群健康检查。

7 非功能要求

7.1 可靠性要求

7.1.1 高可用

高可用要求如下：

- a) 应提供系统自动故障探测及管理功能;
- b) 应确保系统组件不存在单点故障风险;
- c) 集群任意节点发生故障时,不应出现服务中断、数据丢失或数据不一致的情况;
- d) 集群任意单元发生故障时,系统操作应不受影响;
- e) 应保证系统长期无故障不间断运行。

7.1.2 数据冗余存储与分布

数据冗余存储与分布要求如下：

- a) 应提供元数据多副本存储功能,任意节点发生故障时不影响系统继续提供服务的能力;
- b) 应提供基于分区容错的主副本规划功能,具有提前规划各副本数据物理分布的能力。

7.1.3 数据备份和恢复

数据备份和恢复要求如下:

- a) 应提供分布式文件存储备份和恢复功能;
- b) 应提供分布式结构化数据存储备份和恢复功能;
- c) 应提供分布式列式存储备份和恢复功能;
- d) 宜支持数据全量备份和增量备份;
- e) 宜支持数据自动备份和手动备份。

7.1.4 故障恢复与迁移

故障恢复与迁移要求如下:

- a) 任意节点发生故障后,系统应提供将修复后的节点接回系统的能力;
- b) 故障恢复与迁移过程不应影响系统用户数据的完整性与一致性;
- c) 故障恢复与迁移过程不应影响系统整体服务能力。

7.2 兼容性要求

大数据系统应兼容不同品牌的操作系统。

7.3 安全性要求

7.3.1 用户管理

用户管理要求如下:

- a) 应对登录用户进行身份标识和鉴别,保证用户身份标识唯一性;
- b) 用户身份鉴别信息应满足一定的复杂度要求,并定期更换;
- c) 应提供登录失败处理措施,如结束会话、限制非法登录次数、登录连接超时自动退出等措施。

7.3.2 权限管理

权限管理要求如下:

- a) 应以系统组件为单位配置角色和用户;
- b) 应按照权限最小化的原则为用户配置权限;
- c) 应支持按照数据表级、数据列级的粒度为用户分配权限;
- d) 应支持按照不同操作类型(如增、删、改、查、执行等)为用户分配权限。

7.3.3 日志管理

日志管理要求如下:

- a) 应提供记录系统操作日志功能,记录用户的重要操作;
- b) 应保证系统操作日志无法删除、修改或被覆盖;
- c) 操作日志应包括日期、时间、操作者信息、操作类型、操作描述和操作结果等;
- d) 应提供对系统操作日志进行统计、查询、分析及生成报表的功能。



7.3.4 数据安全

数据安全要求如下:

- a) 应提供数据存储加解密功能,支持数据库级数据加密;
- b) 应提供系统敏感数据加密传输功能,并且加密密钥可被替换;
- c) 宜支持数据列级的数据加密。

7.4 可扩展性要求

系统可扩展性要求如下:

- a) 应提供集群在线扩容和减容功能;
- b) 应提供集群离线扩容和减容功能。

7.5 维护性要求

系统可维护性要求如下:

- a) 应提供安装部署管理功能,对大数据集群中管理节点和数据节点软件进行安装部署;
- b) 应提供查看系统版本信息的功能;
- c) 应提供系统在线升级功能,支持单组件升级、升级过程中回滚等;
- d) 应提供错误诊断功能,发生错误时可提供准确的诊断信息以便于定位错误;
- e) 应提供各类计算任务运行进度、状态的实时跟踪及上报功能;
- f) 宜提供系统降级功能,支持单组件降级、降级过程中回退等。

7.6 易用性要求

系统易用性要求如下:

- a) 应提供图形界面的系统安装配置工具,以便于系统部署;
- b) 应提供完整的产品文档,包括但不限于安装部署手册、管理员使用手册、应用开发指南、用户操作手册等。

