ICS 35.240
L 67

中华人民共和国国家标准

GB/T 38672—2020

信息技术　大数据　接口基本要求

Information technology—Big data—Interface basic requirements

2020-04-28 发布　　2020-11-01 实施

国家市场监督管理总局
国家标准化管理委员会　发布

目　次

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:华为技术有限公司、中国电子技术标准化研究院、浪潮电子信息产业股份有限公司、浪潮软件集团有限公司、智慧神州(北京)科技有限公司、美林数据技术股份有限公司、深圳讯策科技有限公司、北京软件和信息服务交易所有限公司、内蒙古大学、中电长城网际系统应用有限公司、西藏国路安科技股份有限公司。

本标准主要起草人:光亮、符海芳、杨彦林、王为中、尹卓、赵江、王功明、黄先芝、张慧敏、赵志强、刘雪、董艳、李华、闵京华、龙祥、孙嘉阳、李冰。

信息技术　大数据　接口基本要求

1　范围

本标准给出了基于大数据参考架构的接口框架，规定了接口的基本要求。

本标准适用于指导大数据系统的设计、开发和应用部署。

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、GB/T 35589—2017 界定的以及下列术语和定义适用于本文件。

3.1

图数据　graph data

基于图解模型的数据。

注：图解模型是一种可以呈现数据元素之间关系的大数据记录存储类型。

3.2

地理数据　geographic data

与地球某个地点直接或间接相关的数据。

[GB/T 17694—2009，定义 B.206]

3.3

日志数据　log data

记录在系统运行中发生的事件的数据。

3.4

时间窗口　time window

一段连续的时间区间。

注：用于对数据进行基于时间的分析。

3.5

内存计算　in-memory processing

优先使用计算机内存对数据进行计算、分析的一种数据处理技术。

3.6

数据同步　data synchronization

建立源和目标数据存储之间的一致性的过程。

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)

CLI:命令行界面(Command-Line Interface)

DDL:数据定义语言(Data Definition Language)

DML:数据操纵语言(Data Manipulation Language)

HDFS:分布式文件系统(Hadoop Distributed File System)

OLAP:在线分析处理(On-Line Analytical Processing)

REST:表述性状态转移(Representational State Transfer)

SNMP:简单网络管理协议(Simple Network Management Protocol)

SQL:结构化查询语言(Structured Query Language)

5 接口框架

5.1 概述

依据 GB/T 35589—2017 描述的大数据参考架构,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应用。大数据应用提供者执行数据生命周期操作,以满足系统协调者定义的需求以及安全和隐私保护需求。大数据框架提供者建立一种计算框架,在此框架中执行转换应用,同时保护数据完整性和隐私。在数据提供者、数据消费者、大数据应用提供者、大数据框架提供者、安全和隐私、管理模块之间存在丰富的接口,支持模块间的信息传递和互操作,对大数据系统集成、兼容性、互操作性有着重要影响。模块内的组件间也有接口支持组件间的互操作。按照所连接的模块可以对接口进行分类。各类别的接口提供相应的信息传递和交互功能,满足接口使用对象的需求。

本标准基于 GB/T 35589—2017 描述的大数据参考架构,给出接口框架(见图 1),包含数据提供者、数据消费者、大数据应用提供者、大数据框架提供者、安全和隐私、管理模块之间的接口:

接口 1 是数据提供者与大数据应用提供者之间的接口;

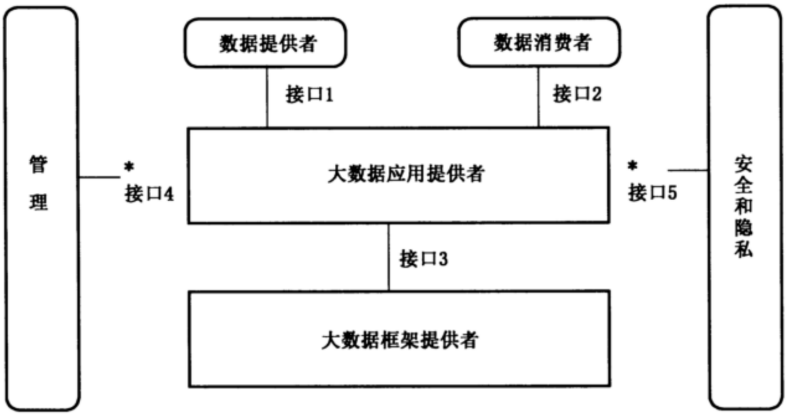
接口 2 是数据消费者与大数据应用提供者之间的接口;

接口 3 是大数据应用提供者与大数据框架提供者之间的接口;

接口 4 是管理模块与其他模块(数据提供者、数据消费者、大数据应用提供者、大数据框架提供者、安全和隐私)间的接口;

接口 5 是安全和隐私模块与其他模块(数据提供者、数据消费者、大数据应用提供者、大数据框架提供者、管理)间的接口。

注:本标准不涉及大数据参考架构中的系统协调者与大数据应用提供者间的接口。



* 表示一个模块与多个模块之间的接口。

图 1　基于大数据参考架构的接口框架

5.2　接口 1

一方面，数据提供者通过接口 1 向大数据应用提供者传送需要分析或处理的数据，包括结构化数据、半结构化数据和非结构化数据等各种数据格式。另一方面，大数据应用提供者可以通过接口 1 向数据提供者传递数据请求。

接口 1 的常见类别包括但不限于：

a) 数据访问接口：用于实时访问数据提供者的数据信息，包括业务系统访问接口、数据库访问接口、文件访问接口等各种数据来源接口，可由数据提供者开放二次开发接口、从数据提供者应用层面接口重构生成、或通过数据库生成数据库表访问接口等方式提供。

b) 数据采集接口：用于将数据提供者的数据信息按照一定的原则和规则进行获取，并按照指定的方式存储。

c) 数据核验接口：用于按照数据共享开放准则、信息保护和访问服务约束，进行跨层级、跨部门、跨系统数据在线核验，返回核验信息。

注：例如来源于某一信息系统的单一信息核验、多项信息联合核验，及来源于多信息系统的某一信息联查对比、多项信息联合比对核验等。

5.3　接口 2

一方面，数据消费者通过接口 2 向大数据应用提供者下发所需的各类数据处理指令，例如查询、检索、数据挖掘、报表生成、可视化等。另一方面，大数据应用提供者通过接口 2 返回数据消费者所需的数据处理结果。

接口 2 的常见类型包括但不限于：

a) 流数据分析接口：用于实时流数据的查询，基于 SQL 扩展提供高级分析功能（如复杂事件分析、在线机器学习、图计算等）。这类接口在物联网、车联网、智慧城市、金融、在线推荐等应用场景中广泛使用。

b) 图数据分析接口：用于对图数据进行查询、修改、子图和全图遍历及各类图分析（例如关系分析、路径规划等）。这类接口广泛用于社交关系分析、反欺诈、推荐、精准营销等场景。

c) 日志数据分析接口：用于查询和分析日志数据，提供日志的全文检索、统计分析、关联分析、监控预警等操作。这类接口在大数据系统的运维、企业信息技术系统的管理和维护、业务及用户行为分析等场景有广泛应用。

d) 数据同步接口：用于不同业务系统之间的数据同步和共享，支持一次采集多系统共享，跨系统

共享数据同步更新，不需要重复录入。这类接口在跨层级、跨系统业务数据共享共用、系统互联互通中广泛使用。

5.4 接口 3

一方面，大数据应用提供者通过接口 3 向大数据框架提供者下发数据计算、存储或访问指令，利用大数据框架提供者的各类计算、存储和网络资源。另一方面，大数据框架提供者通过接口 3 向大数据应用提供者返回数据计算的结果或需要访问的数据。大数据框架提供者包括各类计算、存储组件，一般基于开源版本进行增强。

根据访问的大数据框架提供者组件，接口 3 的常见类型包括但不限于：

a) 离线计算接口：用于对数据进行离线计算，支持数据读取、分发、聚集、输出等操作。同时对计算任务进行编排和调度；

b) 内存计算接口：用于使用内存对数据进行计算、分析，支持数据聚集、数据集转换等操作；

c) 分布式文件存储接口：用于对分布式文件数据进行交互，支持文件系统连接、文件访问、文件流及存储空间管理等操作；

d) 分布式列式存储接口：用于对分布式列式数据进行交互，支持实时查询、分析等操作；

e) 关系型数据库接口：用于对关系型数据进行交互，支持数据库连接、数据库管理、数据表管理、数据访问等操作；

f) 多维分析数据库接口：用于对多维数据进行交互，支持数据查询，和数据表的动态修改等操作；

g) 分布式内存数据库接口：用于对分布式内存数据进行交互，支持数据库连接、数据访问、数据管理等操作；

h) 海量全文检索接口：用于对海量文本数据进行检索和查询，支持索引库连接、数据表管理、数据访问等操作。

5.5 接口 4

一方面，管理模块通过接口 4 向大数据系统的其他模块发送监控、配置指令，监管大数据系统其他模块的资源和运行状态。另一方面，大数据系统的其他模块通过接口 4 向管理模块传送自身状态、配置请求、出错或告警信息。

5.6 接口 5

一方面，安全和隐私模块通过接口 5 向大数据系统的其他模块传送安全和隐私相关配置和指令，支持身份管理、访问授权、安全审计等操作。另一方面，大数据系统的其他模块通过接口 5 将数据安全、系统安全、用户隐私相关的状态、操作、验证请求等发送给安全和隐私模块。

6 基本要求

6.1 总体要求

各类接口应满足如下总体要求：

a) 开放性：符合产业习惯，兼容主流开源接口，减小接口定制化带来的重新设计、适配成本；

b) 易用性：尽可能设计成抽象程度高、屏蔽底层实现、语法易理解的接口；

c) 扩展性：同一接口可通过增加函数、操作符、语句等形式支持新的功能。

6.2 接口 1

接口 1 的要求包括：

a) 应支持多种数据来源(业务系统、数据库、文件等)、多种数据类型(例如业务相关数据、监控数据)、多种数据格式(结构化、半结构化、非结构化等)的数据访问;
b) 应支持按照大数据采集留存规则进行数据采集;
c) 应支持接口运行情况(如主要函数调用时延)的监控,能及时发现错误、产生告警信号;
d) 宜支持多种数据访问接口实例的定时(例如按天、小时、分钟等)调用启动;
e) 宜支持对接常用数据库采集工具、主流日志采集工具(如 Filebeat);
f) 宜支持对接基于系统应用服务接口(数据提供者提供或第三方重构的)实现数据采集的工具;
g) 接口生成宜不受业务系统的开发语言、所处网络环境、系统形态等限制。

6.3 接口 2

接口 2 的常见类型接口有如下要求:

a) 流数据分析接口的要求包括:
   1) 应基于 SQL 扩展支持流数据的查询和分析;
   2) 应支持流和表、流和流的连接;
   3) 应提供流的聚合查询,支持常见的聚合函数(如汇总、均值、最大、最小);
   4) 应支持基于时间窗口的聚合查询;
   5) 应支持基于时间窗口的模式识别、复杂事件分析;
   6) 应支持基于时间窗口的地理数据分析;
   7) 应支持多种时间窗口,包括但不限于跳跃窗口、滑动窗口、会话窗口;
   8) 宜支持基于时间窗口进行流的连接;
   9) 宜支持机器学习函数(例如聚类、分类、回归等)。
b) 图数据分析接口的要求包括:
   1) 应兼容主流开源接口(例如 Gremlin),支持标签属性图数据模型;
   2) 应支持基本的图操作,包括但不限于增加或删除顶点、增加或删除边,增加、删除或修改顶点或边的属性,支持关联删除,定义属性类型等;
   3) 应支持顶点查询、路径查询、子图查询和全图查询;
   4) 应支持查询图拓扑结构的基本指标,包括但不限于中介中心度、紧密中心度等;
   5) 应支持基本图分析,包括但不限于社团发现、三角计数、k 核算法等;
   6) 宜提供主流开发语言(例如 Java、Scala、Python)接口。
c) 日志分析接口的要求包括:
   1) 应兼容主流开源组件(例如 Elasticsearch)接口;
   2) 应提供人工智能分析能力,包括但不限于关联分析、异常问题识别等;
   3) 应支持复杂报表分析,包括但不限于来源分析、热点页面、平均响应时间等;
   4) 应支持对相关指标(如平均相应时间)进行预警设置。
d) 数据同步接口的要求包括:
   1) 宜支持一个系统一次录入,其他共用信息系统的相同数据同步写入;
   2) 宜支持跨系统在业务层面信息的互通互联操作;
   3) 宜支持跨系统共享共用数据的同步更新。

6.4 接口 3

接口 3 的要求包括:

a) 应支持业界常用接口,兼容主流开源接口,支持系统集成;
b) 应支持海量数据的分布式离线计算,支持 MapReduce 计算模型,支持数据的读取、分发、聚集、

输出等处理功能；

c) 应提供内存计算的多种语言开发接口(例如 Scala，Java)，提供高度抽象算子构建分布式数据处理应用；

d) 应支持分布式文件存储的文件操作(包括但不限于文件创建、读取、写入、删除、文件状态信息查询等)及文件夹操作(包括但不限于文件夹创建、删除、状态信息查询、内容统计信息查询等)；

e) 应支持分布式文件系统设置，包括但不限于设置访问权限、设置文件所有者、设置访问时间或者修改时间等；

f) 应支持关系型数据库的 DDL、DML 操作，支持标准 SQL；

g) 应支持多维分析数据库以标准 SQL 查询和分析；

h) 应支持海量全文检索，提供结构化、非结构化文本的多条件检索、统计和报表生成；

i) 应支持自定义函数的定义、加载、使用机制；

j) 应支持海量结构化数据的交互式 OLAP 分析；

k) 宜支持分布式内存计算框架接入多种数据源(例如 HDFS、HBase、Hive)，支持离线计算程序平滑转接；

l) 宜支持分布式列式存储不同类型索引(例如主键索引、组合索引、全文索引)的创建、查询、重建；

m) 宜支持关系型数据库的事务操作，包括但不限于事务开启、提交及回滚等。

6.5 接口 4

接口 4 的要求包括：

a) 应支持大数据组件的安装部署、支持升级配置(包括但不限于查询可升级的版本、需配置的参数等)；

b) 应支持用户管理，包括但不限于增加、删除、修改用户，增加、删除、修改角色、用户权限控制等；

c) 应支持监控告警，对资源使用情况、资源运行状态等进行监控，并提供多种展示方式，支持健康检查；

d) 应支持各类日志的收集和存储，包括但不限于运行日志、操作日志；

e) 应支持标准管理协议(例如 SNMP)，提供 REST API、CLI 等交互方式。

6.6 接口 5

接口 5 宜提供 REST API。