ICS 35.020
L 72

中华人民共和国国家标准

GB/T 36345—2018

信息技术　通用数据导入接口

Information technology—General data import interface

2018-06-07 发布　　2019-01-01 实施

国家市场监督管理总局
中国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。

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：中兴通讯股份有限公司、华为技术有限公司、成都勤智数码科技股份有限公司、浪潮软件集团有限公司、北京软件和信息服务交易所有限公司、浪潮电子信息产业股份有限公司、上海天玑科技股份有限公司、天津南大通用数据技术股份有限公司、中国电子技术标准化研究院。

本标准主要起草人：黄峥、牛家浩、王源、张强、汪绍飞、刘宇峰、吴志刚、张安文、赵江、苏志远、王静。

信息技术　通用数据导入接口

1　范围

本标准规定了通用数据导入接口，包括数据源与大数据系统间应提供的主流通用的数据导入接口，及接口要求。

本标准适用于大数据系统的数据导入接口的研制和测试。

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

3　术语和定义

GB/T 35295—2017 界定的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　big data

具有数据量大、数据速度高、数据种类多样和(或)数据可变性高等主要特征，要求运用可扩展技术进行有效存储、操作、管理和分析的大规模数据集。

注 1：大数据通常以不同的方式被使用，例如，可作为处理大数据大规模数据集的可扩展技术的代名词。

注 2：大数据通常是一个或多个问题的集合：

a)　数据种类：不规则或异构数据的导航、查询和输入问题；

b)　数据量：处理大数据集时需要的并行计算、存储和管理问题；

c)　数据有效性/真实性：描述性数据和关于实时决策对象的自我查询问题；

d)　数据速度：数据的到达速率问题；

e)　数据可视化：数据集的呈现和聚集问题。

[GB/T 35295—2017，定义 2.1.1]

3.2

动态数据　data in motion

处于活动状态，其典型特征表现为大数据的速度和多变性特征的数据。

注：它们在网络上传输或暂时驻留于计算机内存中供读取或更新。对它们以实时或近实时方式进行处理和分析。

[GB/T 35295—2017，定义 2.1.36]

3.3

静态数据　data at rest

处于静止状态，其典型特征表现为大数据的体量和多样性特征的数据。

注：它们通常是存储于物理媒体中的数据。

[GB/T 35295—2017，定义 2.1.37]

4 接口要求

4.1 总则

根据数据的产生方式、存储状态、数据应用方法、实时性等，可以将数据源分为两大类数据：静态数据和动态数据。静态数据与动态数据都可以包含结构化、半结构化、非结构化类型的数据。

静态数据一般以文件方式存储；动态数据包括消息数据、流式数据等，由数据源以实时或准实时方式动态产生。动态数据通常通过消息中间件导入到大数据系统，消息中间件可以支持各种数据类型传输并满足实时性要求。

本标准规定以下两类主流通用的接口，即静态数据的文件导入接口和动态数据的消息导入接口：

a) 静态数据的文件导入接口：实现将文件类的静态数据从数据源导入到大数据系统，简称文件接口。

b) 动态数据的消息导入接口：实现将消息数据、流式数据等动态数据从数据源导入到大数据系统，简称消息接口。

大数据系统的数据导入接口在大数据系统中所处的位置以及与其他部分的接口关系，如图 1 所示。

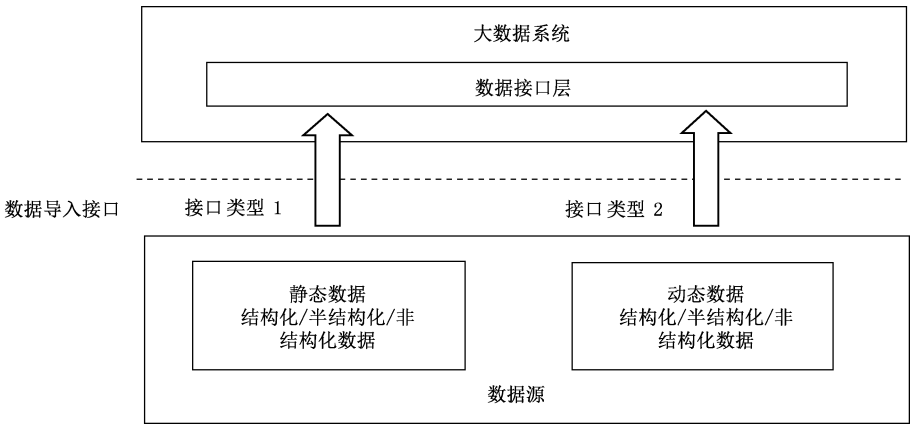

图 1 数据导入接口关系图

4.2 接口实现要求

4.2.1 概述

接口实现应遵循以下基本原则：

a) 接口应能够适配不同的大数据系统版本；

b) 接口应能够保证数据传输过程的安全性、可靠性、稳定性和完整性。

接口操作描述参见附录 A。

4.2.2 静态数据的文件导入接口

静态数据的文件导入接口，适用于客户端与服务器端进行批量文件传输，具有分布式、高吞吐等特性。

文件导入接口提供两种接口操作模式：

a) 操作模式一是数据源作为客户端，大数据系统作为服务器端，客户端与服务端之间采用 FTP 协议交互，客户端首先显式登录到服务器端，再进行文件上传和下载操作，如图 2 所示。

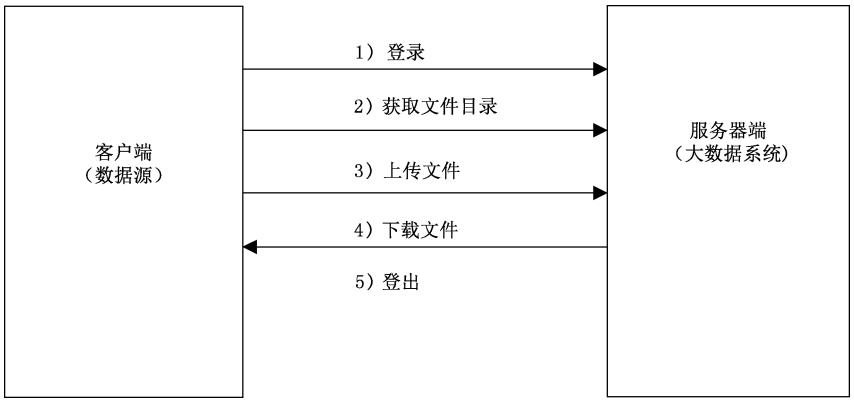

图 2　文件导入接口操作模式一

接口流程如下：

1）　客户端登录服务器端；

2）　客户端获取服务器端文件存储位置；

3）　客户端上传单个文件到服务器端指定文件存储位置；

4）　客户端从服务器端指定存储位置下载服务器单个文件；

5）　客户端登出服务器。

b）　操作模式二是客户端与服务器端之间通过大数据系统的数据传输协议进行文件传输操作，该模式不同于模式一，不需要显式登录服务端。该模式支持扫描满足规则的数据文件及并发传输多个文件。如图 3 所示。

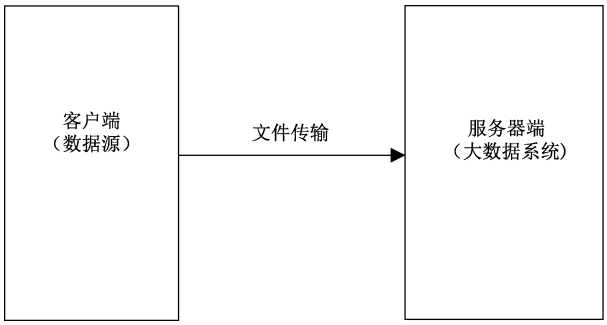

图 3　文件导入接口操作模式二

在操作模式二中，客户端先建立与服务器的连接，然后按规则扫描满足条件的本地文件，并通过大数据系统的数据传输协议将文件上传到大数据系统，传输完成后关闭连接。

接口流程如下：

1）　客户端隐式登录服务器，建立连接；

2）　客户端按照规则从源路径扫描本地文件；

3）　客户端通过大数据系统的数据传输协议，将扫描得到的文件数据批量上传到服务器的指定路径；

4）　数据传输完成后，客户端关闭与服务器的连接。

4.2.3　动态数据的消息导入接口

动态数据一般通过消息中间件，将结构化/半结构化/非结构化的数据实时或准实时地导入大数据

系统；可进行持久化操作，将消息持久化到磁盘，具有高吞吐、分布式、实时等特性，可用于批量消费以及实时应用。

消息中间件应用中包含生产者角色、代理角色和消费者角色。数据源是生产者，大数据系统是消费者，消息中间件是代理。

消息中间件接口关系，如图 4 所示。

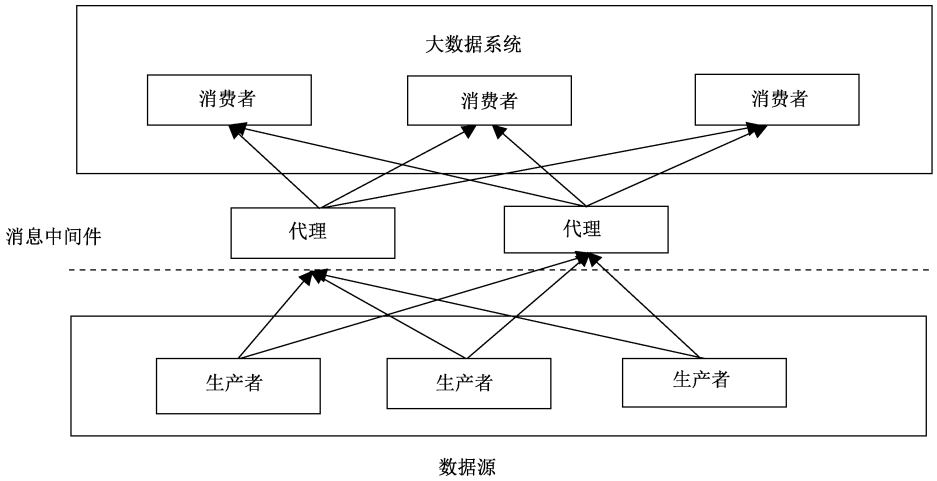

图 4　消息中间件接口关系图

接口处理方式：

建立生产者与消费者后，由生产者和代理建立消息并发送数据。消费者通过代理接收并消费生产者发送的消息数据。生产者、消费者使用完成后，需要关闭并释放资源。如图 5 所示。

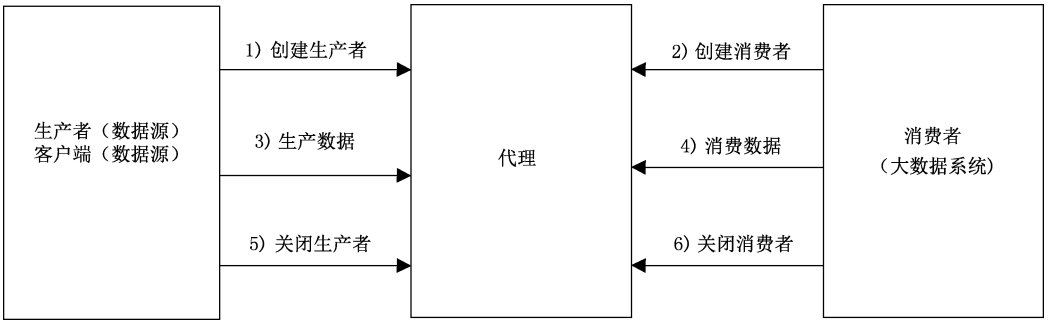

图 5　接口处理方式

接口流程如下：

1）　在发送消息前，数据源通过接口创建生产者，生产者向代理建立消息会话；

2）　在消费数据之前，大数据系统通过接口创建消费者，消费者建立向代理的消息会话；

3）　生产者向代理发送数据；

4）　消费者从代理接收数据；

5）　销毁生产者，生产者终止向代理的消息发送会话，并释放相关资源；

6）　销毁消费者，消费者关闭消息接收会话，并释放相关资源。

附 录 A
（资料性附录）
接口操作说明

A.1 静态数据的文件导入接口

A.1.1 操作模式一

A.1.1.1 登录

操作名：Login
描述：登录。
参数：见表 A.1。

表 A.1 Login

参数名	描述	类型	可选性
UserName	用户名	String	必选
Password	密码	String	必选
Url	目标地址	String	必选
Port	端口	String	必选
返回值：			
RetCode	返回结果	Boolean	必选

A.1.1.2 获取文件列表

操作名：GetFileList
描述：列举服务器端指定目录下的文件名列表。
参数：见表 A.2。

表 A.2 GetFileList

参数名	描述	类型	可选性
FilePath	路径	String	必选
Recursive	是否包含子目录	Boolean	可选
Comparator	排序规则	String	可选
FileNameFilter	文件名过滤器	String	可选
返回值：			
RetCode	返回结果	Boolean	必选
FileList	文件列表	List<String>	可选

A.1.1.3 下载文件

操作名：DownLoadFile
描述：下载单个文件。
参数：见表 A.3。

表 A.3 DownLoadFile

参数名	描述	类型	可选性
FilePathAndName	待下载文件路径及名称	String	必选
Callback	回调接口	String	可选
Timeout	超时时间单位秒(s)	Int	可选
返回值：			
RetCode	返回结果	Boolean	必选

A.1.1.4 上传文件

操作名：UpLoadFile
描述：上传单个文件。
参数：见表 A.4。

表 A.4 UpLoadFile

参数名	描述	类型	可选性
SourcefilePathAndName	待上传源文件路径及名称	String	必选
TargetfilePath	文件上传目的路径	String	必选
FileRegion	是否以文件创建时间建立目标目录	Boolean	可选
CompressionEnable	是否压缩	Boolean	可选
CompressionType	上传压缩方式	String	可选
EncodeType	上传文件的编码方式	String	可选
BackupEnable	上传后是否备份	Boolean	可选
CheckEnable	是否开启校验	Boolean	可选
CheckClass	校验模式	String	可选
Callback	回调接口	String	可选
Timeout	超时时间单位秒(s)	Int	可选
返回值：			
RetCode	返回结果	Boolean	必选

A.1.1.5 登出

操作名：Logout
描述：登出。

参数：无。

返回值：无返回值。

A.1.2 操作模式二

A.1.2.1 文件传送

操作名：SendFile

描述：向服务器传输单个/多个文件。

参数：见表 A.5 和表 A.6。

表 A.5 SendFile

参数名	描述	类型	可选性
UserName	用户名	String	必选
Password	密码	String	必选
Url	目标地址	String	必选
Port	端口	String	必选
FilePathAndFilterRuler	文件路径及传输规则	List＜FilePathAndFilterRuler＞	必选
返回值：			
RetCode	返回结果	Boolean	必选

其中，文件传输规则 FilePathAndFilterRuler 支持的类型和参数见表 A.6。

表 A.6 FilePathAndFilterRuler

参数名	描述	类型	可选性
FilePath	文件传输路径(包括源路径、目的路径)	String	必选
FileName	文件名(含过滤规则)	String	必选
FileRegion	是否以文件创建时间建立目标目录	Boolean	可选
CompressionEnable	是否压缩	Boolean	可选
CompressionType	上传压缩方式	String	可选
EncodeType	上传编码方式	String	可选
BackupEnable	上传后是否备份	Boolean	可选
CheckEnable	是否开启校验	Boolean	可选
CheckClass	校验模式	String	可选
ScantimeEnable	是否设置扫描时间	Boolean	可选
ScanTime	扫描时间间隔单位秒(s)	Int	可选

A.2 动态数据的消息导入接口

A.2.1 创建生产者

操作名：CreateProducer

描述：在发送消息前，需要首先创建生产者和消费者。生产者向代理建立消息发送会话。

参数：见表 A.7。

表 A.7 CreateProducer

参数名	描述	类型	可选性
Brokerlist	Broker 连接配置	List< String >	必选
Producerconfig	配置文件	String	可选
Property	配置项列表	List <prop>	可选
Topic	消息的 Topic 名称	String	必选
Compressioncodec	消息编解码方法	Int	可选
返回值：			
RetCode	返回结果	Boolean	必选
Producer	producer 实例	Producer	必选

A.2.2 消息发送

操作名：SendMessage

描述：生产者向代理（数据接口层）发送数据。

参数：见表 A.8。

表 A.8 SendMessage

参数名	描述	类型	可选性
Topic	消息的 Topic 名称	String	必选
Message	要发送的消息流	Byte[]	必选
返回值：			
RetCode	返回结果	Boolean	必选

A.2.3 销毁生产者

操作名：DestoryProducer

描述：销毁生产者。生产者终止向代理的消息发送会话，并释放相关资源。

参数：无。

返回值：无。

A.2.4 创建消费者

操作名：CreateConsumer

描述：创建消费者。消费者建立向代理的消息会话。

参数：见表 A.9。

表 A.9 CreateConsumer

参数名	描述	类型	可选性
Bootstrapserver	代理的连接信息	List<String>	必选
ConsumerConfig	配置属性文件	String	可选
Property	配置项	List <prop>	可选
Topic	消息的 Topic 名称	String	必选
返回值：			
RetCode	返回结果	Boolean	必选
Consumer	Consumer 实例	Consumer	可选

A.2.5 消息接收

操作名：ReceiveMessage

描述：消费者（数据平台接口层）接收数据。

参数：见表 A.10。

表 A.10 ReceiveMessage

参数名	描述	类型	可选性
Topic	消费的 Topic 名称	String	必选
返回值：			
RetCode	返回结果	Boolean	必选
Message	接收的消息流	Byte[]	必选

A.2.6 销毁消费者

操作名：DestoryConsumer

描述：销毁消费者。消费者关闭消息接收会话，并释放相关资源。

参数：无。

返回值：无。