ICS 35.240
L 67

中华人民共和国国家标准

GB/T 38643—2020

信息技术　大数据
分析系统功能测试要求

Information technology—Big data—Functional testing requirements for analytic system

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:浪潮电子信息产业股份有限公司、中国电子技术标准化研究院、中国人民大学、上海计算机软件技术开发中心、浪潮软件集团有限公司、勤智数码科技股份有限公司、深圳迅策科技有限公司、成都四方伟业软件股份有限公司、陕西省信息化工程研究院、中国铁道科学研究院集团有限公司、平安科技(深圳)有限公司、内蒙古大学、江苏中堃数据技术有限公司、重庆大数据研究院有限公司。

本标准主要起草人:赵江、苏志远、卫凤林、张群、杜小勇、陈敏刚、黄先芝、公维锋、陈文捷、蔡立志、王建华、李正、耿大为、赵志强、颜怀柏、顾美营、张勇、朱志祥、马小宁、吴艳华、赵正阳、韩梅、李华、魏清、张海静、王东强。

信息技术　大数据
分析系统功能测试要求

1　范围

本标准规定了大数据分析系统的数据准备模块、分析支撑模块、数据分析模块、流程编排模块的功能测试要求。

本标准适用于指导大数据分析系统的设计、开发和交付。

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 37721—2019　信息技术　大数据分析系统功能要求

3　术语和定义

GB/T 37721—2019 界定的以及下列术语和定义适用于本文件。

3.1

大数据分析系统　big data analysis system

在大数据存储和处理系统提供的原始数据和计算框架的基础上，集成了一系列数据分析生命周期过程中所用工具的系统。

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)

AUC：ROC 曲线下方的面积(Area under the ROC Curve)

CPU：中央处理器(Central Processing Unit)

GPU：图形处理器(Graphics Processing Unit)

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

JSON：JS 对象标记(JavaScript Object Notation)

OLAP：联机分析处理(On-Line Analytical Processing)

PCA：主成分分析(Principal Components Analysis)

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

ROC：接受者操作特性(Receiver Operating Characteristic)

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

SSD：固态硬盘(Solid State Drives)

XML：可扩展置标语言(Extensible Markup Language)

5 概述

本标准根据GB/T 37721—2019规定的大数据分析系统的功能要求，给出了相应的测试要求。测试示例参见附录A。

6 数据准备模块功能测试

6.1 数据抽取功能测试

数据抽取功能测试要求如下：

a) 应测试大数据分析系统数据准备模块是否支持按照需求抽取存放在存储系统中的数据；
b) 应测试大数据分析系统数据准备模块是否对结构化数据、非结构化数据提供不同抽取方法；
c) 应测试大数据分析系统数据准备模块是否提供全量抽取及增量抽取模式；
d) 应测试大数据分析系统数据准备模块是否支持主动抽取和被动追加；
e) 应测试大数据分析系统数据准备模块是否支持定时批量抽取；
f) 应测试大数据分析系统数据准备模块是否支持分布式数据抽取，并测试在数据抽取过程是否实现负载均衡。

6.2 数据清洗功能测试

数据清洗功能测试要求如下：

a) 应测试大数据分析系统数据准备模块是否支持数据一致性；
b) 应通过进行无效数据值删除、修正等操作测试大数据分析系统数据准备模块是否支持处理无效值；
c) 应通过填充缺失值或删除缺失值对应数据条目等操作测试大数据分析系统数据准备模块是否支持处理缺失值；
d) 应通过合并重复数据或者删除重复数据等操作测试大数据分析系统数据准备模块是否支持处理重复数据；
e) 应测试大数据分析系统数据准备模块是否提供清洗前后的数据比对功能；
f) 应测试大数据分析系统数据准备模块是否支持逻辑矛盾、关联性验证、不合理数据的清洗。

6.3 数据转换功能测试

数据转换功能测试要求如下：

a) 应通过对结构化数据进行列转换操作测试大数据分析系统数据准备模块是否支持结构化数据列转换；
b) 应通过对结构化数据进行行转换操作测试大数据分析系统数据准备模块是否支持结构化数据行转换；
c) 应通过对结构化数据进行表转换操作测试大数据分析系统数据准备模块是否支持结构化数据表转换；
d) 应测试大数据分析系统数据准备模块是否支持非结构化数据的结构化处理；
e) 应测试大数据分析系统数据准备模块是否支持对文本、网页类数据的规范化处理，是否支持将文档类数据转化成单一规范形式；
f) 应通过进行语音和音频输入，检测输入识别结果准确性，测试大数据分析系统数据准备模块是否支持对语音/音频数据的识别处理；

g) 应通过进行图像输入，检测输入识别结果准确性，测试大数据分析系统数据准备模块是否支持提取图像信息。

6.4 数据加载功能测试

数据加载功能测试要求如下：

a) 应测试大数据分析系统数据准备模块是否支持把经过清洗和转换之后的数据加载到大数据分析系统；

b) 应按照加载的目标结构将转换过的数据输入到目标结构中去，测试大数据分析系统数据准备模块是否支持全量加载；

c) 在目标结构中已经存在数据时，应通过在保存已有数据的基础上增加新的数据，测试大数据分析系统数据准备模块是否支持增量加载；

d) 应测试大数据分析系统数据准备模块是否支持实时加载或批量加载。

7 分析支撑模块功能测试

7.1 查询功能测试

7.1.1 查询接口测试

查询接口测试要求如下：

a) 应测试大数据分析系统分析支撑模块是否支持通过标准的数据库连接接口进行查询；

b) 应测试大数据分析系统分析支撑模块是否支持 REST API 查询接口进行查询。

7.1.2 查询优化测试

查询优化测试要求如下：

a) 应通过建立数据索引测试大数据分析系统分析支撑模块是否达到查询加速的效果；

b) 应测试大数据分析系统分析支撑模块是否支持精确查询和模糊查询；

c) 应测试大数据分析系统分析支撑模块是否支持基于规则或者基于成本的查询优化；

d) 应测试大数据分析系统分析支撑模块是否支持数据分片和多副本技术；

e) 应测试大数据分析系统分析支撑模块是否支持通过 SQL 进行复杂条件高并发查询；

f) 应测试大数据分析系统分析支撑模块是否支持二级索引。

7.2 机器学习功能测试

7.2.1 数据集管理功能测试

数据集管理功能测试要求如下：

a) 应测试大数据分析系统分析支撑模块是否能够将输入数据划分为训练集、验证集和测试集；

b) 应通过将训练、验证过的模型导入到大数据分析系统中，以及将大数据系统中训练所得的模型导出的操作，测试大数据分析系统分析支撑模块是否提供机器学习模型的导入和导出的功能。

7.2.2 支持算法测试

支持算法测试要求如下：

a) 应测试大数据分析系统分析支撑模块是否支持回归与分类算法；

b) 应测试大数据分析系统分析支撑模块是否支持聚类算法；

c) 应测试大数据分析系统分析支撑模块是否支持协同过滤算法；

d) 应测试大数据分析系统分析支撑模块是否支持降维算法；
e) 应测试大数据分析系统分析支撑模块是否支持频繁模式挖掘算法；
f) 应测试大数据分析系统分析支撑模块是否支持神经网络算法；
g) 应通过检查是否具有特征提取、特征转换、特征选择、模型选择、交叉验证、模型调优组件测试大数据分析系统分析支撑模块是否提供机器学习流程的其他组件；
h) 应测试大数据分析系统分析支撑模块是否支持 Java、Scala、Python、R 等一种或多种语言，并且是否支持二次开发增加新的算子。

7.2.3 模型评估功能测试

应通过检查机器学习模块中包含交叉验证、模型选择等核心评估组件测试大数据分析系统分析支撑模块是否能够支持算法模型的评估模块。

7.3 统计分析功能测试

统计分析功能测试要求如下：
a) 应通过计算最大值、最小值、求和、总数等统计量测试大数据分析系统分析支撑模块是否支持基本的数值统计；
b) 应通过计算平均数、中位数、众数等统计量测试大数据分析系统分析支撑模块是否支持分析数据集中趋势的统计；
c) 应通过计算极差、方差、标准差等统计量测试大数据分析系统分析支撑模块是否支持分析数据离散程度的统计；
d) 应通过计算协方差、相关系数等统计量测试大数据分析系统分析支撑模块是否支持分析多个随机变量的关系；
e) 应通过保存常用的统计分析方案测试大数据分析系统分析支撑模块是否支持统计分析的自定义模板能力。

7.4 可视化功能测试

可视化功能测试要求如下：
a) 应通过以 Excel、关系型数据库、JSON、XML 格式输入测试大数据分析系统分析支撑模块是否支持常见的数据源数据格式作为输入；
b) 应测试大数据分析系统分析支撑模块是否支持对高维数据的可视化展示；
c) 应通过检查是否可以以柱状图、饼图、折线图等方式展示测试大数据分析系统分析支撑模块是否支持可视化分析工具库；
d) 应测试大数据分析系统分析支撑模块是否支持算法模型的评估相关的可视化工具。

8 数据分析模块功能测试

8.1 分析模式测试

8.1.1 离线数据分析功能测试

离线数据分析功能测试要求如下：
a) 应测试大数据分析系统数据分析模块是否支持结构化查询语言；
b) 应测试大数据分析系统数据分析模块是否支持对离线数据的分布式分析；
c) 应测试大数据分析系统数据分析模块是否具有通过标准接口支持第三方应用的能力；

d) 应测试大数据分析系统数据分析模块是否支持分布式计算或并行计算等计算框架；
e) 应测试大数据分析系统数据分析模块是否支持对海量工作任务的切分和分布式调度；
f) 应测试大数据分析系统数据分析模块是否支持集成第三方的机器学习算法库；
g) 应测试大数据分析系统数据分析模块是否支持使用内存或 SSD 存储作为缓存；
h) 应测试大数据分析系统数据分析模块是否支持分布式执行计划层面的优化；
i) 应测试大数据分析系统数据分析模块是否支持对文本类、音视频类以及图像类数据的分析；
j) 应测试大数据分析系统数据分析模块是否支持对关系型数据库和大数据存储系统中的数据源进行交叉查询、聚合、关联操作的能力；
k) 应测试大数据分析系统数据分析模块是否支持使用 GPU 对特定算法加速分析。

8.1.2 流数据分析功能测试

流数据分析功能测试要求如下：

a) 应测试大数据分析系统数据分析模块是否支持按时间切片后进行批量处理；
b) 应测试大数据分析系统数据分析模块是否支持基于事件触发或者采样的流式处理；
c) 应测试大数据分析系统数据分析模块是否支持实时流上的数据统计；
d) 应测试大数据分析系统数据分析模块是否支持流式数据的排序；
e) 应测试大数据分析系统数据分析模块是否支持与静态表之间的关联；
f) 应测试大数据分析系统数据分析模块是否支持多个数据流的关联处理；
g) 应测试大数据分析系统数据分析模块是否支持采用滑动窗口方式的实时分析任务，并测试其时间窗口大小是否可调；
h) 应测试大数据分析系统数据分析模块是否支持实时数据的分组、优先级调度；
i) 应测试大数据分析系统数据分析模块是否支持对文本类、音视频类以及图像类数据的分析。

8.1.3 交互式联机分析功能测试

交互式联机分析功能测试要求如下：

a) 应测试大数据分析系统数据分析模块是否支持通过结构化查询语言对数据进行分布式的联机分析；
b) 应测试大数据分析系统数据分析模块是否支持通过结构化查询语言对数据进行即席查询；
c) 应测试大数据分析系统数据分析模块是否支持利用可视化中间件对数据分析结果进行显示；
d) 应测试大数据分析系统数据分析模块是否支持在交互式分析过程中定义计算公式和参数配置；
e) 应测试大数据分析系统数据分析模块是否支持交互式分析过程的自动保存和回退等操作；
f) 应测试大数据分析系统数据分析模块是否支持在交互式分析过程中对分析结果的保存和发布；
g) 应测试大数据分析系统数据分析模块是否支持基于在线联机分析的交互式数据分析；
h) 应测试大数据分析系统数据分析模块是否支持对非结构化数据的分析。

8.2 分析类型测试

8.2.1 预测型分析功能测试

预测型分析功能测试要求如下：

a) 应测试大数据分析系统数据分析模块是否支持趋势预测、回归分析等多种预测分析方法；
b) 应测试大数据分析系统数据分析模块是否支持准确率以百分比数值化形式呈现，并测试是否精确到小数点后至少 1 位；
c) 应测试大数据分析系统数据分析模块是否支持使用可视化方式进行显示分析结果；

d） 应测试大数据分析系统数据分析模块是否支持对训练好的模型的发布应用。

8.2.2 描述型分析功能测试

描述型分析功能测试要求如下：

a） 应测试大数据分析系统数据分析模块是否支持使用相关关系分析方法进行描述型分析；

b） 应测试大数据分析系统数据分析模块是否支持可视化展示样本数据的分析结果，是否支持展示模型训练效果，是否支持对训练好的模型可存储和发布；

c） 应测试大数据分析系统数据分析模块是否支持分析结果的良好直观呈现。

9 流程编排模块功能测试

9.1 工作流管理测试

工作流管理测试要求如下：

a） 应通过拖拉方式进行流程编排和修订等操作测试大数据分析系统流程编排模块是否支持可视化的流程编排操作界面；

b） 应通过配置工作流的触发时间的启动时间、执行周期测试大数据分析系统流程编排模块是否支持工作流的调度触发机制，并且是否支持配置触发时间或触发事件；

c） 应测试大数据分析系统流程编排模块是否支持通过管理界面对工作流进行启动、停止操作；

d） 应测试大数据分析系统流程编排模块是否支持并行执行多流程任务；

e） 应测试大数据分析系统流程编排模块是否支持通过数据管道实现工作流的串联；

f） 应测试大数据分析系统流程编排模块是否支持多人协同功能；

g） 应测试大数据分析系统流程编排模块是否支持流程编排结果的持久化保存。

9.2 告警和日志测试

告警和日志测试要求如下：

a） 应测试大数据分析系统流程编排模块是否支持跟踪计算或任务的执行状态，并测试是否对异常任务给出告警；

b） 应测试大数据分析系统流程编排模块是否支持任务执行状态的细节输出到日志。

附 录 A
（资料性附录）
测试示例

A.1 数据准备模块功能测试示例

A.1.1 数据抽取功能测试示例

测试示例见表 A.1～表 A.6。

表 A.1

功能要求	GB/T 37721—2019 6.1a)
测试项	6.1a)
测试示例	在全量/增量/负载均衡三种常见需求中选择抽取方法进行测试。分别执行表 A.3 或表 A.6 的测试示例

表 A.2

功能要求	GB/T 37721—2019 6.1b)
测试项	6.1b)
测试示例	a) 可选择以下至少一种结构化数据的抽取方法： 1) 数据库复制：从源数据库读取数据，写入目标数据库； 2) 数据库同步：在源数据库变化时，动态更新目标数据库中的数据，保持源数据库和目标数据库内容一致； 3) 数据抽取-转换：从源数据库中读取数据，经过转换处理，然后写入目标数据库。 b) 可选择以下至少一种非结构化数据的抽取方法： 1) 单文件复制：将单个文件从源存储地址复制到指定的目标存储地址； 2) 批量文件复制：将选取的多个文件从源存储地址复制到指定的目标存储地址； 3) 文件夹复制：将选取的一个或多个源文件夹中存储的所有文件复制到指定的目标存储地址； 4) 文件夹同步：采用同步更新机制实现源文件夹中存储的文件与目标存储的文件同步

表 A.3

功能要求	GB/T 37721—2019 6.1c)
测试项	6.1c)
测试示例	数据存储在源数据库或文件系统中，抽取到目标数据库或文件系统： a) 全量抽取操作：对待抽取的源数据库或源文件内容进行签名，全量抽取并存储到目标数据库或文件系统后，全量抽取后再进行签名，对比签名是否一致； b) 增量抽取操作：目标数据库或文件存储中已经包含全量抽取的内容，对待追加的数据记录或文件内容进行签名，向目标数据库或文件系统中追加新增的数据记录或文件，增量抽取后再签名，对比签名是否一致

表 A.4

功能要求	GB/T 37721—2019 6.1d)
测试项	6.1d)
测试示例	数据存储在源数据库或文件系统中，抽取到目标数据库或文件系统： a) 主动抽取操作：系统能够将待抽取的数据记录或文件从源数据库或文件系统，通过拉取(pull)方式进行全量抽取或增量抽取，测试示例与表 A.3 相同； b) 被动追加操作：外部系统通过数据准备模块的 API，将待抽取的数据以推送(push)方式追加到目标数据库或文件系统，追加前后分别对数据进行签名，对比签名是否一致

表 A.5

功能要求	GB/T 37721—2019 6.1e)
测试项	6.1e)
测试示例	源数据存放在数据库或文件系统中。对待抽取的数据进行签名，执行定时批量抽取操作到目标数据库或文件系统，然后对数据进行签名，对比抽取前后数据的签名是否一致： a) 设置分钟级定时任务，批量抽取过程中修改系统时钟； b) 设置小时级定时任务，批量抽取过程中修改系统时钟，并模拟抽取过程中跨天的情况； c) 设置天级定时任务，批量抽取过程修改系统时钟

表 A.6

功能要求	GB/T 37721—2019 6.1 f)
测试项	6.1 f)
测试示例	数据存放在数据库中，并能够继续追加数据。在数据库中存入足够多的文件内容足够大的数据，把监控探针分别部署到数据库每个节点(≥2)，然后进行为期 1 h 的数据抽取，分析监控探针传回的监测数据，得到每个节点的负载情况

A.1.2 数据清洗功能测试示例

测试示例见表 A.7～表 A.12。

表 A.7

功能要求	GB/T 37721—2019 6.2a)
测试项	6.2a)
测试示例	数据已经抽取到分析系统的结构化存储。对数据表中的数据进行检查，分析数据一致性。筛选出不一致的数据，对不一致的数据进行处理

表 A.8

功能要求	GB/T 37721—2019 6.2b)
测试项	6.2b)
测试示例	数据已经抽取到分析系统的结构化存储。对数据表中的数据项进行检查，删除或修改数据中的无效值

表 A.9

功能要求	GB/T 37721—2019 6.2c)
测试项	6.2c)
测试示例	数据已经抽取到分析系统的结构化存储。对数据表中的数据记录进行检查，删除存在缺失值的数据记录或将缺失值补全

表 A.10

功能要求	GB/T 37721—2019 6.2d)
测试项	6.2d)
测试示例	数据已经抽取到分析系统的结构化存储。对数据表中的数据记录进行检查，删除或合并重复数据记录

表 A.11

功能要求	GB/T 37721—2019 6.2e)
测试项	6.2e)
测试示例	数据已经抽取到分析系统的结构化存储并经过了数据清洗模块的处理。提供清洗前数据信息和清洗后数据信息的自动比对或人工比对功能，并输出数据清洗前后变化结果

表 A.12

功能要求	GB/T 37721—2019 6.2 f)
测试项	6.2 f)
测试示例	数据已经抽取到分析系统的结构化存储： a) 对数据表中的数据进行检查，分析数据逻辑，删除或修改存在逻辑矛盾的数据； b) 对数据表中的数据进行检查，分析数据关联性，删除或修改存在关联性错误的数据； c) 对数据表中的数据进行检查，分析数据合理性，删除或修改不合理的数据

A.1.3 数据转换功能测试示例

测试示例见表 A.13～表 A.19。

表 A.13

功能要求	GB/T 37721—2019 6.3a)
测试项	6.3a)
测试示例	数据已经抽取到分析系统的结构化存储，并经过了数据清洗模块的处理。对数据表的一个或多个字段的值进行转换或生成一个新的字段，包括但不限于以下操作： a) 分组或分级：如按照年龄段分组，按照用户消费额大小划分用户等级； b) 变换或替换：如字符与数值之间的变换，或用归一化的数值替换原来字段值； c) 拆分或组合：如将一个列拆分成多列，或将多列组合成一列

表 A.14

功能要求	GB/T 37721—2019 6.3b)
测试项	6.3b)
测试示例	数据已经抽取到分析系统的结构化存储，并经过了数据清洗模块的处理。对数据表按照行进行转换操作，包括但不限于以下操作： a) 行过滤：如按照某个标称值过滤掉不符合条件的行； b) 行变换：如把一行数据按照某种条件或规则分裂成多行数据，或把多行按组聚合成一行

表 A.15

功能要求	GB/T 37721—2019 6.3c)
测试项	6.3c)
测试示例	数据已经抽取到分析系统的结构化存储，并经过了数据清洗模块的处理。对数据表整体进行转换操作，包括但不限于以下操作： a) 从一个表抽取部分数据生成一个新的表； b) 抽取表的元数据信息，然后通过行列转置，生成一个新的表

表 A.16

功能要求	GB/T 37721—2019 6.3d)
测试项	6.3d)
测试示例	非结构化数据已经抽取到分析系统存储，并经过了数据清洗模块的处理。在文本、网页/文档/语音、音频/图片/图像五种常见需求中选择抽取非结构化数据类型功能要求的测试示例进行测试。分别执行表 A.17、表 A.18 或表 A.19 的测试示例

表 A.17

功能要求	GB/T 37721—2019 6.3e)
测试项	6.3e)
测试示例	非结构化数据已经抽取到分析系统存储，并经过了数据清洗模块的处理。 a) 文本、网页类数据的规范化处理操作：提取文本、网页类数据信息，将提取的信息生成结构化数据； b) 文档类数据的规范化处理操作：提取文档内容及文档属性信息，将提取的数据生成结构化数据

表 A.18

功能要求	GB/T 37721—2019 6.3 f)
测试项	6.3 f)
测试示例	非结构化数据已经抽取到分析系统存储，并经过了数据清洗模块的处理。将语音/音频内容转换为计算机可读的输入，测试系统能否识别语音、音频中的词汇

表 A.19

功能要求	GB/T 37721—2019 6.3 g)
测试项	6.3 g)
测试示例	非结构化数据已经抽取到分析系统存储，并经过了数据清洗模块的处理。 a) 提取图片内容操作：将图片中的内容转换为字符文本； b) 提取图像信息操作：提取图像信息，将提取的信息生成结构化数据

A.1.4 数据加载功能测试示例

测试示例见表 A.20～表 A.23。

表 A.20

功能要求	GB/T 37721—2019 6.4a)
测试项	6.4a)
测试示例	数据已经抽取到分析系统的结构化存储，并经过了数据清洗模块的处理。在全量加载/增量加载/实时加载/批量加载四种常见需求中选择加载类型对应的测试示例进行测试。分别执行表 A.21、表 A.22 或表 A.23 的测试示例

表 A.21

功能要求	GB/T 37721—2019 6.4b)
测试项	6.4b)
测试示例	数据已经抽取到分析系统的结构化存储，并经过了数据清洗模块的处理。数据加载时，若目标结构中无数据，直接加载写入新数据；若目标结构中已有数据，删除原有数据并加载写入新数据

表 A.22

功能要求	GB/T 37721—2019 6.4c)
测试项	6.4c)
测试示例	数据已经抽取到分析系统的结构化存储，并经过了数据清洗模块的处理。数据加载时，若目标结构中的已有数据与加载的新数据不会产生重复记录，直接加载写入新数据记录；若目标结构会产生重复记录，丢弃加载的新数据记录或者以不同版本数据记录加载写入目标结构

表 A.23

功能要求	GB/T 37721—2019 6.4d)
测试项	6.4d)
测试示例	数据已经抽取到分析系统的结构化存储，并经过了数据清洗模块的处理。 a) 实时加载操作：实时通过流数据处理方式将转换过的数据输入目标结构中； b) 批量加载操作：采用批量导入方式，将转换过的数据输入目标结构中

A.2 分析支撑模块功能测试示例

A.2.1 查询功能测试示例

A.2.1.1 查询接口测试示例

测试示例见表 A.24～表 A.25。

表 A.24

功能要求	GB/T 37721—2019 7.1.1a)
测试项	7.1.1a)
测试示例	提供使用标准接口类型的验证程序样例源码、编译环境及预设目标。编译生成可执行的二进制程序连接数据库执行检测，检查结果是否完整以及是否符合预设目标

表 A.25

功能要求	GB/T 37721—2019 7.1.1b)
测试项	7.1.1b)
测试示例	建立匹配大数据分析系统 Rest API 接口相关的检测环境和预设目标。编译并执行测试样例程序，检查结果是否完整以及是否符合预设目标

A.2.1.2 查询优化测试示例

测试示例见表 A.26～表 A.31。

表 A.26

功能要求	GB/T 37721—2019 7.1.2a)
测试项	7.1.2a)
测试示例	审阅关于基于规则或基于成本优化的文件，看其是否规定了相应的优化途径。在同一套正常运行的大数据分析系统中构造检测数据环境，在该环境中通过两次执行相同的查询，一次带有数据索引的查询，一次不带数据索引，并观察结果

表 A.27

功能要求	GB/T 37721—2019 7.1.2b)
测试项	7.1.2b)
测试示例	建立完成检测所需数据及精准查询和模糊查询的查询语句，并设定检测目标。执行精准查询和模糊查询的查询语句，检测查询结果与预设目标的一致性

表 A.28

功能要求	GB/T 37721—2019 7.1.2c)
测试项	7.1.2c)
测试示例	系统具备基于规则或基于成本优化组件正常工作的证明手段，建立可体现查询优化功能的检测方法。执行构建的检测方法，并使用提供的证明手段进行结果检查

表 A.29

功能要求	GB/T 37721—2019 7.1.2d)
测试项	7.1.2d)
测试示例	构建检测方法及检测数据，具备可直观查看查询性能的手段，并设定检测目标。通过适当调整数据副本数，检查调整前后的查询耗时情况

表 A.30

功能要求	GB/T 37721—2019 7.1.2e)
测试项	7.1.2e)
测试示例	建立带有复杂条件的 SQL 语句及并发测试工具，并发过程中 SQL 的条件不应取固定值。执行并发测试工具并记录结果，通过工具输出物进行结果检查

表 A.31

功能要求	GB/T 37721—2019 7.1.2f)
测试项	7.1.2f)
测试示例	构建特定的能够建立二级索引的语句或程序，并设定检测目标。执行二级索引语句或程序，检测查询结果与预设目标的一致性

A.2.2 机器学习功能测试示例

A.2.2.1 数据集管理功能测试示例

测试示例见表 A.32 和表 A.33。

表 A.32

功能要求	GB/T 37721—2019 7.2.1a)
测试项	7.2.1a)
测试示例	加载数据集，并调用数据集划分 API，将数据按设定划分为训练集、验证集和测试集，检查训练集数据、验证集和测试集数据是否按预期正确划分

表 A.33

功能要求	GB/T 37721—2019 7.2.1b)
测试项	7.2.1b)
测试示例	在文件系统中准备好训练好的机器学习模型和测试数据集，如 Logistic 回归算法及 Iris 数据集（可以选择其他机器学习模型与数据集），导入并加载已训练好的 Logistic 回归模型，使用该模型对 Iris 数据集中的一条或多条数据记录进行预测，并将 Logistic 回归模型导出到文件系统

A.2.2.2 支持算法测试示例

测试示例见表 A.34～表 A.41。

表 A.34

功能要求	GB/T 37721—2019 7.2.2a)
测试项	7.2.2a)
测试示例	调用系统的回归算法 API，对数据集进行回归分析，并检查回归分析的结果；调用系统的分类算法 API，对 Iris 数据集进行分类，并检查分类的结果

表 A.35

功能要求	GB/T 37721—2019 7.2.2b)
测试项	7.2.2b)
测试示例	调用系统的 K-均值聚类算法 API，对 Iris 数据集进行聚类，并检查聚类的结果

表 A.36

功能要求	GB/T 37721—2019 7.2.2c)
测试项	7.2.2c)
测试示例	调用系统的协同过滤算法 API，对数据集中的用户进行推荐，并检查推荐结果

表 A.37

功能要求	GB/T 37721—2019 7.2.2d)
测试项	7.2.2d)
测试示例	调用系统的降维 API,比如 PCA API,对 MNIST(手写体数字识别)数据集进行降维,并可视化降维结果

表 A.38

功能要求	GB/T 37721—2019 7.2.2e)
测试项	7.2.2e)
测试示例	调用系统的频繁模式挖掘 API,比如 Apriori 算法 API,对构造的数据集进行频繁模式计算,并查看关联规则

表 A.39

功能要求	GB/T 37721—2019 7.2.2f)
测试项	7.2.2f)
测试示例	调用系统的神经网络 API,对 MNIST 数据进行分类,并查看分类结果

表 A.40

功能要求	GB/T 37721—2019 7.2.2g)
测试项	7.2.2g)
测试示例	构造数据集,并调用系统的特征提取、特征转换、特征选择等 API,实现模型的训练、模型选择、交叉验证

表 A.41

功能要求	GB/T 37721—2019 7.2.2h)
测试项	7.2.2h)
测试示例	采用 Java、Scala、Python、R 等一种或多种语言编写一种或多种机器学习算法,并用数据集验证机器学习算法可在系统中正确运行

A.2.2.3 模型评估功能测试示例

测试示例见表 A.42。

表 A.42

功能要求	GB/T 37721—2019 7.2.3
测试项	7.2.3
测试示例	确认机器学习模块中包含交叉验证、模型选择等核心评估组件；确认机器学习模块中包含混淆矩阵、精度、召回率、ROC 曲线、AUC 等模型性能度量指标

A.2.3 统计分析功能测试示例

测试示例见表 A.43～表 A.47。

表 A.43

功能要求	GB/T 37721—2019 7.3a)
测试项	7.3a)
测试示例	选择相应的统计分析算子并连接需要统计的数据表，对数据表执行基本数值统计操作，包括但不限于： a) 求最大值； b) 求最小值； c) 求和； d) 求总数

表 A.44

功能要求	GB/T 37721—2019 7.3b)
测试项	7.3b)
测试示例	选择相应的统计分析算子并连接需要统计的数据表，对数据表执行数据集中趋势统计操作，包括但不限于： a) 求平均数； b) 求中位数； c) 求众数

表 A.45

功能要求	GB/T 37721—2019 7.3c)
测试项	7.3c)
测试示例	选择相应的统计分析算子并连接需要统计的数据表，对数据表执行离散程度统计操作，包括但不限于： a) 求极差； b) 求方差； c) 求标准差

表 A.46

功能要求	GB/T 37721—2019 7.3d)
测试项	7.3d)
测试示例	选择相应的统计分析算子并连接需要统计的数据表，对数据表执行分析多个随机变量的关系操作，包括但不限于： a) 求协方差； b) 求相关系数

表 A.47

功能要求	GB/T 37721—2019 7.3e)
测试项	7.3e)
测试示例	将多个统计分析算子按照分析需求进行自定义组合，形成统计分析 pipeline，并保存为自定义模板。后续统计分析可以使用该模板，作为常用的统计分析方案

A.2.4 可视化功能测试示例

测试示例见表 A.48～表 A.51。

表 A.48

功能要求	GB/T 37721—2019 7.4a)
测试项	7.4a)
测试示例	可视化功能连接的数据源可以支持多种常见的数据格式，包括但不限于： a) Excel 文件； b) 关系数据库； c) JSON 文件； d) XML 文件

表 A.49

功能要求	GB/T 37721—2019 7.4b)
测试项	7.4b)
测试示例	可视化功能能够支持 3 维以上的数据源数据的加载，并按照选择的维度和展现方式进行高维数据可视化展示

表 A.50

功能要求	GB/T 37721—2019 7.4c)
测试项	7.4c)
测试示例	可视化功能加载各种数据源数据，支持多种可视化展示方式，包括但不限于： a) 柱状图； b) 饼状图； c) 折线图； d) 表格； e) 散点图； f) 雷达图； g) 网状图； h) 时间线； i) 热力图； j) 地图

表 A.51

功能要求	GB/T 37721—2019 7.4d)
测试项	7.4d)
测试示例	在文件系统中准备好训练好的机器学习模型和测试数据集，导入并加载已训练好的模型，使用该模型对测试数据集做预测，检查系统能否将模型预测结果的混淆矩阵、精度、召回率、ROC 曲线等模型评估度量指标以可视化的形式显示

A.3 数据分析模块功能测试示例

A.3.1 分析模式测试示例

A.3.1.1 离线数据分析功能测试示例

测试示例见表 A.52～表 A.62。

表 A.52

功能要求	GB/T 37721—2019 8.1.1a)
测试项	8.1.1a)
测试示例	数据存放在分布式文件系统或数据库中，调用结构化查询语言，验证数据查询结果

表 A.53

功能要求	GB/T 37721—2019 8.1.1b)
测试项	8.1.1b)
测试示例	集群正常运行，在 HDFS 上准备测试数据集，上传分布式离线数据分析测试程序，运行测试程序

表 A.54

功能要求	GB/T 37721—2019 8.1.1c)
测试项	8.1.1c)
测试示例	第三方应用可通过标准接口，获得离线分析的结果

表 A.55

功能要求	GB/T 37721—2019 8.1.1d)
测试项	8.1.1d)
测试示例	可对存储在分布式文件系统或数据库中的数据，利用数据分区的机制，实现数据在多台计算节点中分布式计算和结果汇总

表 A.56

功能要求	GB/T 37721—2019 8.1.1e)
测试项	8.1.1e)
测试示例	可对存储在分布式文件系统或数据库中的数据，按计算任务进行切分，并实现任务在多台计算机节点的调度

表 A.57

功能要求	GB/T 37721—2019 8.1.1f)
测试项	8.1.1f)
测试示例	分析系统可正确安装、配置第三方机器学习算法库，如 scikit-learn。运行第三方机器学习算法库的自带案例检查是否运行正确

表 A.58

功能要求	GB/T 37721—2019 8.1.1g)
测试项	8.1.1g)
测试示例	将分析系统常用数据缓存到内存或 SSD 中，对缓存中的数据进行分布式计算。数据分布式计算的时间应小于没有缓存加速计算时间

表 A.59

功能要求	GB/T 37721—2019 8.1.1h)
测试项	8.1.1h)
测试示例	通过比对数据的分布式执行计划的优化配置与非优化配置，比对数据处理的时间性能。分布式执行计划层面优化后的数据计算时间应小于非优化配置的计算时间

表 A.60

功能要求	GB/T 37721—2019 8.1.1i)
测试项	8.1.1i)
测试示例	在系统中存储文本、图像及音视频类的数据，利用系统提供的机器学习 API 对文本、图像与音视频数据实现预期的数据分析，如对文本实现自然语言处理、对图像实现分类等

表 A.61

功能要求	GB/T 37721—2019 8.1.1j)
测试项	8.1.1j)
测试示例	在关系型数据库和大数据存储系统中分别存储数据表，利用 SQL 语句实现不同数据源的交叉查询、聚合和关联操作

表 A.62

功能要求	GB/T 37721—2019 8.1.1k)
测试项	8.1.1k)
测试示例	使用神经网络算法对 MNIST 数据集进行分类，比对使用 GPU 与使用 CPU 分类算法的时间，是否使用 GPU 后分类算法的时间应小于使用 CPU 的时间

A.3.1.2 流数据分析功能测试示例

测试示例见表 A.63～表 A.71。

表 A.63

功能要求	GB/T 37721—2019 8.1.2a)
测试项	8.1.2a)
测试示例	将流数据按时间比如按 30 s 切后，输入流数据分析系统，系统对切片时间周期内容数据进行处理

表 A.64

功能要求	GB/T 37721—2019 8.1.2b)
测试项	8.1.2b)
测试示例	在流数据构造事件或采样模式，如日志中的 Error 事件或按 1 s 采样一次，系统对事件和采样得到的数据进行处理

表 A.65

功能要求	GB/T 37721—2019 8.1.2c)
测试项	8.1.2c)
测试示例	将流数据输入流数据分析系统，系统可实现数据的统计功能，如总数、均值等

表 A.66

功能要求	GB/T 37721—2019 8.1.2d)
测试项	8.1.2d)
测试示例	将一个时间周期，如 1 min 的流数据输入流数据分析系统，系统可实现对时间周期内的数据按需求进行排序

表 A.67

功能要求	GB/T 37721—2019 8.1.2e)
测试项	8.1.2e)
测试示例	将流数据输入流数据分析系统，系统可实现实时数据与静态表中数据的关联查询

表 A.68

功能要求	GB/T 37721—2019 8.1.2f)
测试项	8.1.2f)
测试示例	将两个流数据输入流数据分析系统，系统可实现对两个流数据进行关联查询

表 A.69

功能要求	GB/T 37721—2019 8.1.2g)
测试项	8.1.2g)
测试示例	将流数据输入流数据分析系统，系统可采用滑动窗口的方式实现 Top-K 计算，其中时间窗口可调节，并且在不同时间窗口下的结果都应该符合预期结果

表 A.70

功能要求	GB/T 37721—2019 8.1.2h)
测试项	8.1.2h)
测试示例	将流数据输入流数据分析系统，系统可按字段对数据进行分组，并根据优先级对分组数据进行处理

表 A.71

功能要求	GB/T 37721—2019 8.1.2i)
测试项	8.1.2i)
测试示例	将文本类、图像类及音视频类流数据输入流分析系统，系统可按预期对各类实时数据进行数据分析

A.3.1.3 交互式联机分析功能测试示例

测试示例见表 A.72～表 A.79。

表 A.72

功能要求	GB/T 37721—2019 8.1.3a)
测试项	8.1.3a)
测试示例	将数据表存入分布式数据存储系统，调用结构化查询语言，对数据进行分布式的联机分析，如对 1 个事实表和多个维表进行 OLAP 操作

表 A.73

功能要求	GB/T 37721—2019 8.1.3b)
测试项	8.1.3b)
测试示例	将数据表存入分布式数据存储系统，调用结构化查询语言，通过设置查询条件，对数据进行即席查询

表 A.74

功能要求	GB/T 37721—2019 8.1.3c)
测试项	8.1.3c)
测试示例	将数据表存入分布式数据存储系统，调用结构化查询语言，通过设置查询条件，对数据进行查询，并通过可视化中间件展现查询结果

表 A.75

功能要求	GB/T 37721—2019 8.1.3d)
测试项	8.1.3d)
测试示例	将数据表存入分布式数据存储系统，自定义用户函数，对数据进行自定义查询

表 A.76

功能要求	GB/T 37721—2019 8.1.3e)
测试项	8.1.3e)
测试示例	将数据表存入分布式数据存储系统，并通过结构化查询语言对数据进行多个查询操作，查询的过程可自动保存到文件系统，并可调用回退操作，撤销上一步的查询操作

表 A.77

功能要求	GB/T 37721—2019 8.1.3f)
测试项	8.1.3f)
测试示例	将数据表存入分布式数据存储系统，并通过结构化查询语言对数据进行查询操作，并将查询结果保存到文件系统，并可将查询结果发布到可视化系统

表 A.78

功能要求	GB/T 37721—2019 8.1.3g)
测试项	8.1.3g)
测试示例	将数据表存入分布式数据存储系统，调用结构化查询语言，对数据进行交互式联机分析

表 A.79

功能要求	GB/T 37721—2019 8.1.3h)
测试项	8.1.3h)
测试示例	将非结构化数据存入分布式数据存储系统，通过调用自定义查询语言，对数据进行交互式查询

A.3.2 分析类型测试示例

A.3.2.1 预测型分析功能测试示例

测试示例见表 A.80～表 A.83。

表 A.80

功能要求	GB/T 37721—2019 8.2.1a)
测试项	8.2.1a)
测试示例	选择相应的预测分析方法进行分析，包括但不限于： a) 趋势预测； b) 回归分析

表 A.81

功能要求	GB/T 37721—2019 8.2.1b)
测试项	8.2.1b)
测试示例	检查预测结果的准确率呈现形式，是否实现准确率数值化，是否实现百分比形式呈现并精确到小数点后至少 1 位

表 A.82

功能要求	GB/T 37721—2019 8.2.1c)
测试项	8.2.1c)
测试示例	检查分析结果的呈现方式是否能符合 7.4 的要求。分别执行表 A.48～表 A.51 的测试示例

表 A.83

功能要求	GB/T 37721—2019 8.2.1d)
测试项	8.2.1d)
测试示例	将表 A.80 中训练好的模型发布成应用，如通过服务的方式发布成应用

A.3.2.2 描述型分析功能测试示例

测试示例见表 A.84～表 A.86。

表 A.84

功能要求	GB/T 37721—2019 8.2.2a)
测试项	8.2.2a)
测试示例	检查是否能用相关关系分析方法进行分析

表 A.85

功能要求	GB/T 37721—2019 8.2.2b)
测试项	8.2.2b)
测试示例	检查分析结果的呈现方式是否能符合 7.4 的要求。分别执行表 A.48～表 A.51 的测试示例

表 A.86

功能要求	GB/T 37721—2019 8.2.2c)
测试项	8.2.2c)
测试示例	用可视化组件呈现分析结果

A.4 流程编排模块功能测试示例

A.4.1 工作流管理测试示例

测试示例见表 A.87～表 A.93。

表 A.87

功能要求	GB/T 37721—2019 9.1a)
测试项	9.1a)
测试示例	检查流程编排模块是否提供了可视化流程编排操作界面，测试是否可以通过拖拉方式进行流程编排和修订

表 A.88

功能要求	GB/T 37721—2019 9.1b)
测试项	9.1b)
测试示例	检查流程编排模块是否具有流程运行的调度功能，测试能否支持工作流的调度触发，以及是否可配置触发时间或触发事件，工作流的触发时间的启动时间、执行周期是否可配置

表 A.89

功能要求	GB/T 37721—2019 9.1c)
测试项	9.1c)
测试示例	检查是否提供工作流运行的管理界面，测试是否支持对工作流进行启动、停止操作

表 A.90

功能要求	GB/T 37721—2019 9.1d)
测试项	9.1d)
测试示例	测试是否支持多流程任务的并行执行功能

表 A.91

功能要求	GB/T 37721—2019 9.1e)
测试项	9.1e)
测试示例	测试是否支持能够通过数据管道实现机器学习任务的串联

表 A.92

功能要求	GB/T 37721—2019 9.1f)
测试项	9.1f)
测试示例	测试是否支持多人协同创建机器学习流程的功能

表 A.93

功能要求	GB/T 37721—2019 9.1g)
测试项	9.1g)
测试示例	创建新的机器学习流程编排,测试是否支持保存流程功能

A.4.2 告警和日志测试示例

测试示例见表 A.94 和表 A.95。

表 A.94

功能要求	GB/T 37721—2019 9.2a)
测试项	9.2a)
测试示例	运行选定的机器学习流程任务,测试是否能够支持跟踪计算或任务的执行状态,对异常任务是否能够给出告警

表 A.95

功能要求	GB/T 37721—2019 9.2b)
测试项	9.2b)
测试示例	检查流程编排模块的运行日志,测试任务执行状态的细节是否已经输出到日志