ICS 35.240
L 67

中华人民共和国国家标准

GB/T 38676—2020

信息技术　大数据
存储与处理系统功能测试要求

Information technology—Big data—
Functional testing requirements for storage and processing systems

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、上海计算机软件技术开发中心、新华三技术有限公司、浪潮电子信息产业股份有限公司、深圳市金蝶天燕云计算股份有限公司。

本标准主要起草人:赵华、符海芳、卫凤林、张群、李瑛、陈敏刚、陈文捷、潘子健、李永平、赵江、林琳。

信息技术　大数据
存储与处理系统功能测试要求

1　范围

本标准规定了大数据存储与处理系统的基本功能、分布式文件存储、分布式结构化数据存储、分布式列式数据存储、分布式图数据存储、批处理框架、流处理框架、图计算框架、内存计算框架和批流融合计算框架的测试要求。

本标准适用于大数据存储与处理系统的测试。

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 37722—2019　信息技术　大数据存储与处理系统功能要求

3　术语和定义

GB/T 37722—2019 界定的术语和定义适用于本文件。

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)

GPU：图形处理器(Graphics Processing Unit)

MPI：消息传递接口(Message Passing Interface)

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

5　概述

本标准根据 GB/T 37722—2019 规定的大数据存储与处理系统的功能要求，给出了相应的测试要求。

6　大数据存储子系统的功能测试要求

6.1　基本功能的测试要求

大数据存储子系统基本功能的测试要求如下：

a) 应测试大数据存储子系统是否能对文件、图等数据进行上传和下载的操作；
b) 应测试大数据存储子系统是否能对目录进行创建、查看、权限修改、删除等操作；
c) 应测试大数据存储子系统能否通过 API 调用对文件、对象、图等数据进行查询、修改、删除、增加等操作；
d) 应测试大数据存储子系统能否通过开源或商业工具提供系统与传统关系型数据库之间交换数据和文件；
e) 应测试大数据存储子系统能否通过开源或商业工具提供系统与其他文件系统(如 ext2 等)之间交换数据和文件；
f) 应构造一个关键节点故障，验证大数据存储子系统中的数据读写是否正常；
g) 应测试大数据存储子系统能否通过界面/工具/命令行方式完成自动或手动备份。自动备份需配置相应的参数，如备份周期、备份数等；
h) 应对大数据存储子系统中存在的结构化数据、半结构化数据、非结构化数据执行批量更新、批量删除等操作，测试操作是否正常执行；
i) 应测试大数据存储子系统能否从消息队列读取数据，并将计算结果实时写入数据库；
j) 应测试大数据存储子系统能否将数据采集到实时检索平台，并根据索引主键进行实时查询。

注：本条的测试要求对应 GB/T 37722—2019 中 6.1 的要求。

6.2 分布式文件存储的功能测试要求

分布式文件存储的功能测试要求如下：

a) 应测试大数据存储子系统能否进行文件上传、下载、读写、复制、移动、删除、访问控制等操作；
b) 应测试大数据存储子系统能否对文件进行多副本备份，并能通过副本恢复出原始数据；
c) 应测试大数据存储子系统节点/软件发生故障时，如断电、数据节点失效等，不影响系统及业务的正常运行；
d) 应通过对副本文件进行写操作，然后查看块校验文件，验证副本文件所在节点的块校验文件相同；
e) 应测试删除节点前，单个数据节点先退出服务集群，集群数据重新分布，数据无损，业务不中断；
f) 应验证增加节点后，节点加入集群，系统数据重新分布，业务无中断；
g) 应测试按照已配置的压缩、加密算法，对数据进行压缩、加密和解密，原始数据无损坏和丢失；
h) 应测试大数据存储子系统能否对文件数据进行统一检索、编目、增加和删除操作；
i) 应测试大数据存储子系统能否通过命令或图形化界面对文件进行搜索、批量操作(创建/删除等)、文件删除后进入回收站、(创建/删除/查询等)快照操作；
j) 应测试大数据存储子系统能否根据配置的文件阈值，将存储系统中的小文件打包成大文件进行存储；
k) 应测试大数据存储子系统能否根据目录存储空间大小以及文件数量，对写文件设置最高上限。

注：本条的测试要求对应 GB/T 37722—2019 中 6.2 的要求。

6.3 分布式结构化数据存储的功能测试要求

分布式结构化数据存储的功能测试要求如下：

a) 应通过向大数据存储子系统中导入结构化数据，测试在数据节点上能否看到块数据分布在不同的节点上；
b) 应测试大数据存储子系统能否支持通过 API 调用对结构化数据进行查询操作，包括：元数据、业务数据等；

c) 应通过创建多张表，进行表之间的关联，测试大数据存储子系统能否通过规则过滤等方式查询到某张表中的数据；
d) 应通过导入数据，测试数据所在节点的块校验文件是否相同；
e) 应通过导入多行多列的数据，创建表进行映射关联，测试是否可以按行按列进行数据查询；
f) 应通过导入多行多列的数据，创建表进行映射关联，进行行列转换，测试行数据与列数据能否进行转换。

注：本条的测试要求对应 GB/T 37722—2019 中 6.3 的要求。

6.4 分布式列式数据存储的功能测试要求

分布式列式数据存储的功能测试要求如下：

a) 应通过创建表，写入数据，测试数据能否以键值形式存储在大数据存储子系统中。
b) 应分别对表、列族和列设置用户权限，包括读、写、创建等，测试相应用户能否对表、列族和列进行创建、读、写等操作。
c) 应通过对指定列进行加密，创建表，写入数据。测试表的属性是否是列加密状态，数据是否是非明文存储。
d) 应测试大数据存储子系统能否对数据库对象包括：表、索引、函数、触发器等进行备份和恢复；测试数据备份和恢复任务的进展和历史记录。
e) 应通过批量导入数据，导入时创建二级索引，测试大数据存储子系统能否通过索引查询到导入的数据。
f) 应通过构造多张表，导入数据，测试大数据存储子系统根据关联规则/关系合并后的表内容与预期是否一致。

注：本条的测试要求对应 GB/T 37722—2019 中 6.4 的要求。

6.5 分布式图数据存储的功能测试要求

分布式图数据存储的功能测试要求如下：

a) 应通过定义图数据模型，上传数据文件和图规则映射文件，测试查询到的图数据与定义的数据模型是否一致；
b) 应通过写入/导入图数据，测试能否对图数据进行查询、遍历、分析操作；
c) 应测试大数据存储子系统能否通过图数据库开发接口进行元数据管理、图数据管理等操作；
d) 应通过写入/导入图数据，数据间存在多层关系，测试大数据存储子系统能否对数据进行单节点、多节点的扩线查询；
e) 应通过设置最短路径/最优路径模型，写入/导入图数据，测试搜索结果是否符合最短路径/最优路径算法；
f) 应测试大数据存储子系统能否对图数据顶点、属性的数据执行继承操作；
g) 应通过创建异步会话任务，测试任务状态是否为长任务。

注：本条的测试要求对应 GB/T 37722—2019 中 6.5 的要求。

7 大数据处理子系统的功能测试要求

7.1 基本功能的测试要求

大数据处理子系统基本功能的测试要求如下：

a) 应通过提交多个任务，测试任务是否可以在异构(包含 CPU、内存、GPU)的系统上部署，且资源均衡调度；

b) 应通过工具加压，测试计算框架部署的环境能否随任务的增加而增加节点个数；
c) 应通过创建多个任务，对其分别设置不同的优先级标签，测试高优先级任务能否优先执行；
d) 应测试大数据处理子系统能否通过图形化界面或命令行等查看系统资源情况，包括：所有资源、已分配资源、故障资源等信息；
e) 应测试任务的静态资源调度策略，设置每个任务使用固定的资源，查看任务运行过程中的资源占用是否达到预期；
f) 应测试任务的动态资源调度策略，查看任务运行过程中资源占用的变化情况；
g) 应通过创建一级租户，测试在其下是否能创建多个层级的子租户，并为其配置资源使用上限，每层子租户的资源总和不超过其父租户；
h) 应通过创建一级租户 A 和租户 B，分别分配一定量的资源，指定租户 A 提交任务，任务占用资源远超过其配置的资源容量，测试任务能否提交成功，租户 B 的资源能否减少。再指定租户 B 提交任务，任务占用资源为其满配，测试任务能否提交成功，租户 A 的资源能否减少；
i) 应通过同时提交各种分布式计算框架的任务，测试任务能否成功执行；
j) 应通过构造任务 A 和任务 B，A 的输出是 B 的输入，提交任务 A 和任务 B，测试 A 和 B 是否能够按照依赖关系自动调度；
k) 应通过提交任务，查看其对资源占用的变化情况；等任务运行完成，再次查看该任务对资源的占用情况，测试大数据处理子系统能否根据作业需求动态分配计算资源，自动管理回收资源；
l) 应通过提交多个任务，这些任务按照无回路有向图的方式进行拓扑排序，测试任务是否能够按照拓扑结果进行自动调度；
m) 应通过提交一个复杂任务，测试大数据处理子系统为不同子任务调度的 CPU、内存数量是否不同。

注：本条的测试要求对应 GB/T 37722—2019 中 7.1 的要求。

7.2 批处理框架的功能测试要求

批处理框架的功能测试要求如下：

a) 应通过创建结构化、非结构化类型数据，对数据进行建表，测试批处理框架能否对创建的表进行离线分析；
b) 应通过提交一个离线任务，测试图形化界面能否查看离线计算任务进度与状态；
c) 应在分布式计算环境上，测试大数据处理子系统能否执行批处理任务；
d) 应测试大数据处理子系统能否使用多种语言开发接口进行分析任务，例如 Python、Java 等；
e) 应通过创建多个任务，设置任务之间依赖关系，测试任务能否按照依赖关系顺序执行；
f) 应通过创建分布式任务，测试任务能否在多节点分布、并行执行，任务执行完成后，结果能否进行聚集；
g) 应通过提交分布式任务，测试任务申请资源是否正常，任务执行是否成功。

注：本条的测试要求对应 GB/T 37722—2019 中 7.2 的要求。

7.3 流处理框架的功能测试要求

流处理框架的功能测试要求如下：

a) 应测试大数据处理子系统从消息队列获取实时消息，对数据进行高吞吐、低延迟的实时计算后，再将结果数据写入消息队列操作；
b) 应通过创建消息处理任务，测试用户能否对任务进行浏览、中止、激活、去激活等操作；所有操作记录是否在审计日志中；
c) 应通过创建滑动窗口方式的实时分析任务，测试窗口大小、滑动步长是否可调节；

d) 应通过构造流处理框架故障(如服务进程故障等),测试流处理服务是否正常、无中断;

e) 应通过创建消息处理任务,在消息处理过程中构造节点、进程等异常,测试异常节点上的消息是否会重分布到其他正常节点、进程进行处理。

注:本条的测试要求对应 GB/T 37722—2019 中 7.3 的要求。

7.4 图计算框架的功能测试要求

图计算框架的功能测试要求如下:

a) 应通过 API 读取图数据,寻找某条边/某个点的属性,测试结果与预期是否一致;

b) 应测试大数据处理子系统是否提供工具/命令行/图形化界面进行数据的全量导入、增量导入以及自定义导入;

c) 应测试大数据处理子系统是否支持对实时图数据进行分析和查询;

d) 应通过定义图数据模型,上传数据文件和图规则映射文件(节点/边上的标签和属性),测试文件上传是否成功;

e) 应测试大数据处理子系统是否支持内置常用图指标计算功能,如统计点边数量等;

f) 应通过工具加压,测试大数据处理子系统能否对增加的图数据计算和查询业务自动分配到不同节点;

g) 应通过模拟多个客户端发起图数据库查询请求,测试能否查询到相应的数据结果。

注:本条的测试要求对应 GB/T 37722—2019 中 7.4 的要求。

7.5 内存计算框架的功能测试要求

内存计算框架的功能测试要求如下:

a) 应在分布式计算环境上,测试将任务转化为 DAG 图的功能,并测试能否正确执行分布式计算任务;

b) 应通过工具加压,测试系统对增加的数据处理业务能否自动分配到不同处理节点,自动负载均衡,处理性能没有下降;

c) 应测试内存计算引擎能否对结构化、半结构化、非结构化不同类型的数据进行处理;

d) 应通过运行贝叶斯计算任务,测试任务能否成功执行;

e) 应测试大数据处理子系统能否直接读取非关系型数据库中的数据,不需要对数据做迁移。

注:本条的测试要求对应 GB/T 37722—2019 中 7.5 的要求。

7.6 批流融合计算框架的功能测试要求

批流融合计算框架的功能测试要求如下:

a) 应通过流式读取数据,对读取到的数据执行 SQL 查询语句,测试查询到的数据是否与预期的一致;

b) 应测试大数据处理子系统是否支持位置信息分析、广告浏览统计等场景下的流式 SQL 处理能力;

c) 应测试批流融合计算/处理的时间窗口是否支持多种类型,包括跳跃窗口、滑动窗口等;

d) 应测试大数据处理子系统能否执行对样本数据进行清洗、去重等批、流处理操作,结合模式识别算法,得出识别结果;

e) 应测试大数据处理子系统能否从消息队列中实时读取并批量处理数据,并将统计分析结果实时更新到存储系统中;

f) 应测试大数据处理子系统能否执行事件驱动的流处理,构造事件触发的 action,读取能触发 action 的消息队列数据,测试 action 是否被触发;

g） 应通过读取消息队列数据，测试大数据处理子系统是否支持流数据添加“事件读取”的时间和水印，并对时间窗口内的数据做处理；测试构造多个由简单事件构成的事件流，识别简单事件之间的内在联系，输出多个符合一定规则的简单事件构成复杂事件；

h） 应测试大数据处理子系统能否进行深度学习训练、MPI 等任务的调度。

注：本条的测试要求对应 GB/T 37722—2019 中 7.6 的要求。