ICS 35.020
L70

YD

中华人民共和国通信行业标准

YD/T 3773—2020

大数据 分布式批处理平台技术要求与测试方法

Big data - technical specification and test methods on distributed batch processing platform

2020-12-09 发布　　2021-01-01 实施

中华人民共和国工业和信息化部　发布

目　　次

前　言

本标准是大数据系列标准之一，该系列标准名称和结构如下。

——大数据 分布式批处理平台技术要求与测试方法

——大数据 分布式分析型数据库技术要求与测试方法

——大数据 分布式事务型数据库技术要求与测试方法

——大数据 分布式流处理平台技术要求与测试方法

——大数据 时序数据库技术要求与测试方法

——大数据 商务智能分析工具技术要求与测试方法

——大数据 数据管理平台技术要求与测试方法

——大数据 数据集成工具技术要求与测试方法

——大数据 数据挖掘平台技术要求与测试方法

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。

本标准由中国通信标准化协会提出并归口。

本标准起草单位：中国信息通信研究院、中国科学院计算技术研究所、华为技术有限公司、星环信息科技（上海）有限公司、中国移动通信集团有限公司、中兴通讯股份有限公司、新华三技术有限公司、腾讯云计算（北京）有限责任公司、阿里云计算有限公司、北京百度网讯科技有限公司、广州巨杉软件开发有限公司、中国电信集团有限公司、中国联合网络通信集团有限公司、北京国双科技有限公司、北京东方金信科技有限公司。

本标准主要起草人：魏凯、姜春宇、马鹏玮、王卓、詹剑锋、王磊、李经纬、朱松、陆扬、石在辉、高俊杰、柯正祥、林松涛、郑瑞、刘俊良、赵懿、石棋玲、侯海旭、堵俊平、张强、吴文峰、梁屾。

大数据 分布式批处理平台技术要求与测试方法

1 范围

本标准规定了用于对大数据进行分布式批处理的软件平台或服务应具有的技术要求及相关的测试方法。

本标准适用于大数据分布式批处理平台产品的设计、研发、测试、评估和验收等。

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 5271.18—2008 信息技术词汇第 18 部分：分布式数据处理

GB/T 32400—2015 信息技术 云计算 概览与词汇

GB/T 35295—2017 信息技术大数据术语

3 术语定义及缩略语

3.1 术语定义

下列术语和定义适用于本文件

3.1.1

分布式系统 distributed system

网络联通的多台计算机组成，不同计算机之间通过网络通信传递信息进行计算任务相关的交流与协同。

3.1.2

节点 node

连接至网络中的一个连接点，具体的定义根据所应用的网络和协议各有不同。在分布式系统中，一个节点指代系统中一台连接至网络的计算或存储设备。

3.1.3

批处理 batch processing

数据集合为单位，对已存储的数据进行自动批量处理计算的过程。

3.1.4

分布式计算 distributed computing

覆盖存储层和处理层的，用于实现多类型程序设计算法模型的计算模式。

[GB/T 35295—2017，定义 2.1.22]

3.1.5

分布式数据处理 distributed data processing

数据操作分散到计算机网络各节点进行计算的过程。

[改写 GB/T 5271.18—2008，定义 18.1.8]

3.1.6

分布式批处理 distributed batch processing

在分布式系统架构上根据分布式数据处理和计算模式进行的批处理计算。

3.1.7

分布式文件系统 distributed file system

能够管理分布在多个节点上的文件的文件管理系统，节点间通过分布式系统中的网络进行通信和数据传输。

3.1.8

非关系型数据库 NoSQL database

在非关系型数据库管理系统中数据不是以关系模型组织的，包括键值、列、文件和图数据库管理系统等。

3.1.9

水平扩展 scale out

集成的一群个体资源作为一个单系统使用的过程。

[GB/T 35295—2017，定义 2.1.17]

3.1.10

租户 tenant

一组物理和虚拟资源进行共享访问的一个或多个用户。

[改写 GB/T 32400—2015，定义 3.2.37]

3.1.11

多租户 multi-tenancy

物理或虚拟资源的分配实现多个租户以及他们的计算和数据彼此隔离和不可访问。

[GB/T 32400—2015，定义 2.1.24]

3.1.12

流数据 streaming data

经由接口传递，从连续运行的数据源产生的数据。

[GB/T 35295—2017，定义 2.1.24]

3.1.13

结构化数据 structured data

数据表示形式，按此种形式，由数据元素汇集而成的每个记录的结构都是一致的并且可以使用关系模型予以有效描述。

[GB/T 35295—2017，定义 2.2.13]

3.1.14

非结构化数据 unstructured data

不具有预定义模型或未以定义方式组织的数据。

[GB/T 35295—2017，定义 2.1.25]

3.1.15

测试环境 test environment

软件测试的基础，测试项目依托测试环境，运行于测试环境之上，测试环境是测试所需的所有软硬件的总称。

3.1.16

吞吐率 throughout rate

衡量业务系统单位时间内提供服务的能力指标，在分布式批处理平台中可以指单位时间内处理的数据量或作业量。

3.2 缩略语

下列缩略语适用于本标准。

CPU	中央处理器	Central Processing Unit
JDBC	Java 数据库连接	Java Database Connectivity
NoSQL	非关系型的数据库	Not Only SQL
ODBC	开放数据库连接	Open Database Connectivity
SQL	结构化查询语言	Structural Query Language

4 总体要求

分布式批处理平台的主要功能是存储海量结构化、非结构化的数据，提供批处理和流式计算的能力，支持多种数据分析和挖掘的手段，对数据进行全生命周期的处理。

4.1 参考架构

分布式批处理平台应包括如下功能模块：

a） 分布式存储模块；

b） 批处理计算模块；

c） SQL 引擎模块；

d） 资源调度模块；

e） 运维监控模块；

f） 安全管理模块。

分布式处理平台也可以包括如下功能模块。

a） NoSQL 数据库模块。

常见架构如图 1 所示。

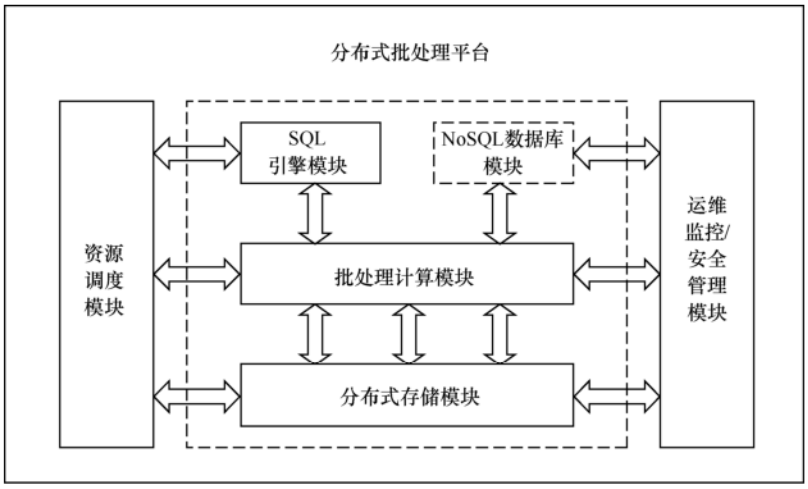

图 1 分布式批处理平台架构

4.2 功能要求

分布式批处理平台应具备数据处理的基本功能，包括数据导入、存储、计算和分析，同时还应该具备容错能力、安全能力、兼容性能力、水平扩展能力、多租户支持能力、运维管理能力、易用性能力等。

各项能力的具体要求详见 5.1 至 5.8。

4.3 性能要求

分布式批处理平台能够满足特定应用场景对大数据批处理的性能要求，平台产品应披露性能衡量指标的具体情况，包括如下。

a） 有效存储空间：分布式批处理平台的有效存储空间不等同于实际的总存储空间，平台产品应相应的披露有效存储空间，以及该指标的具体计算方法；

b） 批处理任务执行时间：批处理任务的执行时间主要为从批处理任务提交开始至任务执行结束的时间总和。平台产品应挑选特定应用场景的典型批处理任务，并披露执行相应任务的执行时间和测试环境。

如果分布式批处理平台包含 NoSQL 数据库模块，还应满足特定应用场景对 NoSQL 数据库读写操作的性能要求，平台产品应披露相应性能指标的具体情况，包括：

a） NoSQL 读写操作的吞吐率：NoSQL 读写操作的吞吐率为在单位时间内执行的 NoSQL 读写操作数。平台产品应挑选特定应用场景的典型 NoSQL 数据库读写任务，并披露相应任务的读写操作吞吐率和测试环境。

5 技术要求

5.1 数据处理基本功能

分布式批处理平台应具备数据处理的基本功能，包括如下。

a） 数据导入：是将数据从外部数据源加载到分布式存储模块的过程。分布式批处理平台应支持结构化和非结构化数据的批量导入，导入任务可以定时执行也可以即时执行。

b） 数据存储：支持海量数据的分布式存储，具有多副本存储能力，能达到较高的数据存储持久性，支持多种数据格式和压缩算法。

c） 数据批处理计算：以数据集合为单位，对已存储的数据进行批量计算的过程。一般应支持。

1） 数据转换：包括数据清洗、数据内容和数据结构的转化。

2） 数据分析：基于 SQL 或其他编程语言的数据查询、统计、分析等能力。

分布式批处理平台宜具备数据处理的其他功能，包括如下。

a） 流式数据导入：分布式批处理平台可支持结构化和非结构化数据的流式导入，导入任务可定时执行也可即时执行。

b） 流式数据处理计算：对不断到达的无限数据集做实时逐条分析处理的过程。

c） 数据存储策略：支持按用户自定义的数据分级策略对数据的存储介质等进行管理。

5.2 容错要求

分布式批处理平台应支持各个模块在出现故障后系统进行恢复的能力，主要包括数据备份和恢复、各模块的主备节点切换等。主要的故障类型有插拔硬盘、网线、关机、重启等。

5.3 安全要求

分布式批处理平台应具有安全保护的技术，以防止恶意的访问和攻击，防止关键数据的泄露，支持完备的权限管理和审计日志功能。具体应包括如下：

a） 各个模块对用户进行身份认证的功能；

b） 至少按照管理员、审计员和操作员三种角色进行分级权限管理的功能，针对操作员按照数据库、表、行列等不同粒度进行权限控制；

c） 多种类型的数据加解密方式；

d） 对于操作进行审计的功能，包括用户操作、任务执行操作、运维操作、权限操作等。

5.4 兼容性要求

分布式批处理平台应对上层应用开放开发接口，支持 SQL 语法，支持数据库操作，具体包括如下：

a） JDBC 和 ODBC 接口的支持；

b） SQL 语法的支持，包括数据类型、运算符和函数等。

5.5 扩展性要求

扩展性要求具体包括：

a） 分布式批处理平台应具备在线水平扩展能力，系统性能随节点数量增加而增长；

b） 分布式批处理平台宜具备在线水平收缩能力。

5.6 多租户要求

分布式批处理平台应具备多租户能力，支持为多个不同租户分配和调度计算资源的能力，确保不同租户之间的计算资源相互隔离，具体包括：

a） 租户管理功能，包括租户建立、修改、删除和资源分配等；

b） 租户内的多种资源调度策略；

c） 租户间的资源相互隔离；

d） 租户资源的监控。

5.7 运维要求

分布式批处理平台应具备良好的运维管理能力，主要包括平台的部署能力、集群管理、用户管理、故障管理、资源监控、作业监控、配置管理、日志管理、平台升级等。具体包括如下：

a） 部署能力：支持以可视化向导的方式对所有模块进行自动化部署。

b） 集群管理：支持通过可视化界面对集群进行操作，例如服务启停或节点启停。

c） 用户管理：支持创建、修改和删除用户和用户组，分配不同的角色和权限，支持对用户密码和信息的修改。

d） 故障管理：支持对告警级别进行区分，对告警阈值进行配置，并通过可视化界面、电子邮件或短信等方式进行告警通知。

e） 资源监控：支持可视化资源监控，可以查看系统各模块软硬件配置、系统性能以及资源使用情况。

f） 作业监控：支持通过可视化方式查看各类作业的运行状态、资源使用情况以及日志信息。

g） 配置管理：支持对系统各模块进行配置管理的能力，包括配置的查看和修改。

h） 日志管理：支持对系统各模块日志进行查看、检索和下载。

i） 平台升级：支持在不宕机的情况下进行软件升级。

5.8 易用性要求

分布式批处理平台在易用性方面宜提供对于工作流的支持，即支持具有相互依赖、参数传递关系的复杂任务的可视化编排调度能力，主要包括如下：

a） 支持工作流的创建，包括作业配置、作业依赖关系设置、作业参数传递设置等；

b） 支持工作流的管理，包括对于工作流中作业的增加、删除、修改、查找，工作流的复制、导入导出、周期触发、事件触发等；

c） 支持工作流的监控和告警。

6 测试方法

本章定义了分布式批处理平台各类技术要求的测试方法，包括数据处理能力、可用性能力、安全能力、兼容性能力、扩展性能力、多租户支持能力、运维管理能力、易用性能力和性能。

6.1 数据处理能力测试方法

6.1.1 数据导入

测试编号：6.1.1
测试项目：数据导入
测试目的：验证分布式批处理平台能够支持结构化和非结构化数据的批量导入，导入任务可以定时执行也可以即时执行
预置条件： 1） 分布式批处理平台正常运行。 2） 结构化和非结构化数据源准备就绪
测试步骤： 1） 进行结构化（非结构化）数据导入配置，并记录数据源中数据内容； 2） 手动启动数据导入； 3） 查看导入后数据； 4） 删除导入的数据，并设置定时数据导入任务； 5） 到达定时任务启动时间时，观察任务执行情况； 6） 查看定时任务执行结果
预期结果： 1） 步骤 3）中，能够观察到导入后数据与步骤 1）中记录数据内容一致； 2） 步骤 5）中，能够观察到定时任务开始执行； 3） 步骤 6）中，能够观察到定时任务导入的数据与步骤 1）中记录数据内容一致

6.1.2 SQL 任务能力

测试编号：6.1.2
测试项目：SQL 任务能力
测试目的：验证分布式批处理平台能够将文件存储模块中的数据导入至 SQL 引擎模块，并能够支持 SQL 查询任务
预置条件： 1） 分布式批处理平台正常运行； 2） 文件存储模块中存有可进行 SQL 查询的数据

测试步骤： 1） 记录文件存储模块中已存有可进行 SQL 查询的数据内容； 2） 将步骤 1）中的数据从文件存储模块导入 SQL 引擎模块； 3） 查看导入后的数据； 4） 在 SQL 引擎模块中执行任意聚合操作； 5） 查看步骤 2）中 SQL 查询的执行情况； 6） 在 SQL 引擎模块中执行任意连接操作； 7） 查看步骤 4）中 SQL 查询的执行情况
预期结果： 1） 步骤 3）中，能够观察到导入后的数据内容与步骤 1）中记录的数据内容一致； 2） 步骤 5）中，能够观察到 SQL 查询成功执行并且执行结果正确； 3） 步骤 7）中，能够观察到 SQL 查询成功执行并且执行结果正确

6.1.3 NoSQL 数据库操作能力

测试编号：6.1.3
测试项目：NoSQL 数据库操作能力
测试目的：验证分布式批处理平台能够将文件存储模块中的数据导入至 NoSQL 数据库模块，并能够支持 NoSQL 数据库相关读写操作
预置条件： 1） 分布式批处理平台正常运行； 2） 文件存储模块中存有可导入 NoSQL 数据库的数据
测试步骤： 1） 记录文件存储模块中已存有可导入 NoSQL 数据库的数据内容； 2） 将步骤 1）中的数据从文件存储模块导入 NoSQL 数据库模块； 3） 查看导入后的数据； 4） 在 NoSQL 数据库模块中执行数据表建立、数据写入、数据读取等操作； 5） 查看步骤 4）中操作的执行情况
预期结果： 1） 步骤 3）中，能够观察到导入后的数据内容与步骤 1）中记录的数据内容一致； 2） 步骤 5）中，能够观察到操作均成功执行并且执行结果正确

6.1.4 机器学习能力

测试编号：6.1.4
测试项目：机器学习能力
测试目的：验证分布式批处理平台能够支持基本的机器学习算法，可以包括 Kmeans 算法、贝叶斯算法、PageRank 算法等
预置条件：

1） 分布式批处理平台正常运行
测试步骤： 1） 在分布式批处理平台的批处理计算模块中提交机器学习算法 1 的作业； 2） 查看步骤 1）中算法作业的执行情况； 3） 重复步骤 1)、步骤 1)，将机器学习算 1 更改为机器学习算法 2、机器学习算法 3 等
预期结果： 1） 步骤 2）中，能够观察到步骤 1）中机器学习算法 1 成功执行； 2） 在后续步骤中验证机器学习算法 2、机器学习算法 3 能够同机器学习算法 1 一样成功执行

6.1.5 流处理能力

测试编号：6.1.5
测试项目：流处理能力
测试目的：验证分布式批处理平台能够支持流式数据处理能力
预置条件： 1） 分布式批处理平台正常运行
测试步骤： 1） 选择任意流数据处理框架及组件接入分布式批处理平台； 2） 模拟流处理业务场景进行流式数据处理作业； 3） 查看流数据处理作业情况
预期结果： 1） 步骤 3）中，能够观察到流数据处理作业正确执行，并能够实时产生处理结果

6.2 高可用能力测试方法

6.2.1 文件存储模块主协调节点失效及恢复

测试编号：6.2.1
测试项目：文件存储系统主协调节点失效及恢复
测试目的：验证分布式批处理平台能够在文件存储系统的主协调节点故障时提供持续服务，可以继续使用备用协调节点进行同步
预置条件： 1） 分布式批处理平台正常运行； 2） 分布式批处理平台初始配置副本数为 3
测试步骤： 1） 准备较大存储量的文件（例如 100 个 100MB 的文件）； 2） 开始将步骤 1）中准备的文件上传至分布式批处理平台的文件存储模块； 3） 在文件传输过程中，将主协调节点所在进程终止，以模拟故障，并记录时间； 4） 观察进行文件传输作业的相关日志文件；

5） 文件传输作业结束后，将已传输至分布式批处理平台文件存储模块的文件下载至本地，并同步骤 1）中准备的文件进行对比
预期结果： 1） 步骤 4）中，能够在日志中发现错误记录，并且错误记录时间同终止进程模拟故障的时间相同； 2） 步骤 5）中，能够对比得出最终上传至分布式批处理平台文件管理模块的文件与步骤 1）中准备的文件的数量、内容、存储容量均相同

6.2.2 文件存储模块数据存储节点故障及恢复

测试编号：6.2.2
测试项目：文件存储模块数据存储节点故障及恢复
测试目的：验证分布式批处理平台能够在文件存储模块的数据存储节点故障时继续正常提供服务
预置条件： 1） 分布式批处理平台正常运行； 2） 分布式批处理平台初始配置副本数为 3； 3） 分布式批处理平台的文件存储模块有负数个数据存储节点
测试步骤： 1） 准备较大存储量的文件（例如 100 个 100MB 的文件）； 2） 开始将步骤 1）中准备的文件上传至分布式批处理平台的文件存储模块； 3） 在文件传输过程中，将分布式批处理平台文件存储模块的某个数据存储节点关机，以模拟故障，并记录时间； 4） 观察进行文件传输作业的相关日志文件； 5） 文件传输作业结束后，将已传输至分布式批处理平台文件存储模块的文件下载至本地，并同步骤 1）中准备的文件进行对比
预期结果： 1） 步骤 4）中，能够在日志中发现错误记录，并且错误记录时间同终止进程模拟故障的时间相同； 2） 步骤 5）中，能够对比得出最终上传至分布式批处理平台文件管理模块的文件与步骤 1）中准备的文件的数量、内容、存储容量均相同

6.2.3 资源调度模块资源管理功能失效及恢复

测试编号：6.2.3
测试项目：资源调度模块资源管理功能失效及恢复
测试目的：验证分布式批处理平台能够在资源调度模块的资源管理功能节点故障时继续正常提供服务
预置条件： 1） 分布式批处理平台正常运行； 2） 分布式批处理平台初始配置副本数为 3
测试步骤： 1） 对分布式批处理平台资源调度模块的高可用性功能进行配置； 2） 向分布式批处理计算模块提交一个运行持续时间较久的分布式批处理作业；

3） 在步骤 2）中作业运行的过程中，将分布式批处理平台资源调度模块的资源管理功能节点进程终止，以模拟故障； 4） 观察提交作业的运行情况
预期结果： 1） 步骤 4）中，能够观察到步骤 1）中的操作出现报错，随后能够继续正常执行

6.2.4 文件存储模块备份恢复能力

测试编号：6.2.4
测试项目：文件存储模块备份恢复能力
测试目的：验证分布式批处理平台能够实现文件存储模块的备份和恢复
预置条件： 1） 分布式批处理平台正常运行
测试步骤： 1） 在文件存储模块的/backup 目录中创建目录 a 和目录 b，分别向两个目录中上传 10G 数据，同时记录文件数、文件存储量和内容； 2） 对目录 a 执行全量备份； 3） 同时删除目录 a 和目录 b； 4） 使用步骤 2）中的备份进行恢复，观察目录/backup； 5） 向 a 目录中添加 1G 数据，记录文件数、文件存储量和内容，并自动执行增量备份； 6） 删除目录 a； 7） 使用步骤 5）中的增量备份进行恢复，观察目录/backup
预期结果： 1） 步骤 4）中，能够观察到目录 a 恢复，对比恢复后目录 a 中文件数、文件存储量和内容均与步骤 1）中记录相同； 2） 步骤 7）中，能够观察到目录 a 恢复，对比恢复后目录 a 中文件数、文件存储量和内容均与步骤 5）中记录相同，步骤 5）中追加的 1G 数据被正确恢复

6.2.5 双集群互备

测试编号：6.2.5
测试项目：双集群互备
测试目的：验证分布式批处理平台能够实现在两个集群之间进行互相备份
预置条件： 1） 分布式批处理平台集群 A 和集群 B 正常运行； 2） 集群 A 的文件存储模块、SQL 引擎模块、NoSQL 数据库模块中已存有数据
测试步骤： 1） 记录集群 A 的文件存储模块、SQL 引擎模块、NoSQL 数据库模块中已有的数据； 2） 将集群 A 的文件存储模块、SQL 引擎模块、NoSQL 数据库模块中的数据同步至集群 B
预期结果：

1） 步骤 2）中，能够成功完成集群 A 到集群 B 的数据同步，查看同步后的数据与步骤 1）中记录的数据相同

6.2.6 运维监控模块失效及恢复

测试编号：6.2.6
测试项目：运维监控模块失效及恢复
测试目的：验证分布式批处理平台能够在运维监控模块故障时继续正常提供服务
预置条件： 1） 分布式批处理平台正常运行； 2） 分布式批处理平台初始配置副本数为 3
测试步骤： 1） 打开运维监控模块可视化界面； 2） 将分布式批处理平台运维监控模块进程终止，以模拟故障，并记录时间； 3） 再次查看运维监控模块可视化界面； 4） 查看运维监控模块相关日志
预期结果： 1） 步骤 3）中，能够观察到运维监控模块可视化界面可以正常查看； 2） 步骤 4）中，能够在日志中发现错误记录，并且错误记录时间同终止进程模拟故障的时间相同

6.3 安全能力测试方法

6.3.1 身份认证

测试编号：6.3.1
测试项目：身份认证
测试目的：验证分布式批处理平台的各功能模块能够对用户身份进行认证，使合法用户能够进行操作，非法用户不能进行操作
预置条件： 1） 分布式批处理平台正常运行； 2） 分布式批处理平台的批处理计算模块、SQL 引擎模块、NoSQL 数据库模块均正常运行； 3） 身份认证功能未开启
测试步骤： 1） 准备两个用户 a 和 b，其中用户 a 为批处理计算模块、SQL 引擎模块、NoSQL 数据库模块的合法用户，用户 b 为非法用户； 2） 使用用户 a 和用户 b 分别向批处理计算模块提交批处理作业，向 SQL 引擎模块提交 SQL 查询任务，向 NoSQL 数据库模块提交 NoSQL 查询任务； 3） 开启身份认证功能，并记录启用方法； 4） 再次使用用户 a 和用户 b 分别向三种模块提交作业和任务
预期结果：

1） 步骤 2）中，用户 a 和用户 b 均能够成功提交批处理作业、SQL 和 NoSQL 查询任务； 2） 步骤 4）中，用户 a 能够成功提批处理交作业、SQL 和 NoSQL 查询任务，用户 b 无法提交

6.3.2 权限管理

测试编号：6.3.2
测试项目：权限管理
测试目的：验证分布式批处理平台能够对于用户和用户组关于系统内的目录、文件、数据库、数据表等的权限进行管理
预置条件： 1） 分布式批处理平台正常运行
测试步骤： 1） 创建用户组 A 和用户组 B，创建用户 a 和用户 b； 2） 通过运维监控模块向用户组 A 分配文件存储模块、SQL 引擎模块、NoSQL 数据库模块的访问和读写权限，向用户组 B 仅分配访问和读权限； 3） 使用用户 a 和用户 b 分别对三个模块进行访问和数据读写操作； 4） 重置运维监控模块中对于用户组 A 和用户组 B 的权限分配； 5） 通过文件存储模块、SQL 引擎模块、NoSQL 数据库模块直接向用户组 A 分配各模块的访问和读权限，向用户组 B 分配访问和读写权限； 6） 再次使用用户 a 和用户 b 分别对三个模块进行访问和数据读写操作
预期结果： 1） 步骤 3）中，用户 a 能够在文件存储模块、SQL 引擎模块、NoSQL 数据库模块中进行访问和数据读写操作，用户 b 仅能够进行访问和数据读操作； 2） 步骤 6）中，用户 a 仅能够在文件存储模块、SQL 引擎模块、NoSQL 数据库模块中进行访问和数据读操作，用户 b 能够进行访问和数据读写操作

6.3.3 存储加解密

测试编号：6.3.3
测试项目：存储加解密
测试目的：验证分布式批处理平台能够对于 SQL 引擎模块和 NoSQL 数据库模块的数据进行加解密的能力，对加密数据通过文件存储模块直接访问无法得到数据明文
预置条件： 1） 分布式批处理平台正常运行； 2） 分布式批处理平台的 SQL 引擎模块和 NoSQL 数据库模块正常运行； 3） SQL 引擎模块和 NoSQL 数据库模块中的数据处于未加密状态
测试步骤： 1） 通过文件存储模块查看 SQL 引擎模块和 NoSQL 数据库模块中的数据； 2） 对两模块中的数据启用加密； 3） 再次通过文件存储模块查看两模块中的数据；

4） 分别通过 SQL 引擎模块和 NoSQL 数据库模块访问各自模块中的数据
预期结果： 1） 步骤 1）中，能够查看到数据明文； 2） 步骤 3）中，查看到的数据为乱码，无法查看到数据明文； 3） 步骤 4）中，能够查看到数据明文

6.3.4 审计

测试编号：6.3.4
测试项目：审计
测试目的：验证分布式批处理平台具有审计功能，能够记录和查询用户的非法操作
预置条件： 1） 分布式批处理平台正常运行； 2） 系统存在非管理员用户 a，文件存储模块中存在数据目录 data
测试步骤： 1） 向用户 a 分配目录 data 的访问和数据读权限，不分配数据写权限； 2） 使用用户 a 在目录 data 下进行写入新文件操作； 3） 查看运维监控模块相关日志
预期结果： 1） 步骤 2）中，用户 a 操作时产生报错，无法进行新文件写入操作； 2） 步骤 3）中，能够在日志中查询到用户 a 对于目录 data 的非法操作记录

6.4 兼容性能力测试方法

6.4.1 ODBC 接口兼容性

测试编号：6.4.1
测试项目：ODBC 接口兼容性
测试目的：验证分布式批处理平台能够通过 ODBC 接口进行 SQL 引擎模块的数据插入、删除、修改、查看等 SQL 查询操作
预置条件： 1） 分布式批处理平台正常运行
测试步骤： 1） 进行通过 ODBC 接口对分布式批处理平台 SQL 引擎模块进行访问的配置； 2） 使用任意程序语言编写通过 ODBC 接口进行数据插入、删除、修改、查看等 SQL 查询操作的程序； 3） 运行步骤 2）中编写的程序； 4） 查看程序运行结果
预期结果： 1） 步骤 3）中，程序成功完成运行；

2） 步骤 4）中，能够观察到程序中的数据插入、删除、修改、查看等 SQL 查询操作均成功执行并得到正确返回结果

6.4.2 JDBC 接口兼容性

测试编号：6.4.2
测试项目：JDBC 接口兼容性
测试目的：验证分布式批处理平台能够通过 JDBC 接口进行 SQL 引擎模块的数据插入、删除、修改、查看等 SQL 查询操作
预置条件： 1） 分布式批处理平台正常运行。
测试步骤： 1） 进行通过 JDBC 接口对分布式批处理平台 SQL 引擎模块进行访问的配置； 2） 使用任意程序语言编写通过 JDBC 接口进行数据插入、删除、修改、查看等 SQL 查询操作的程序； 3） 运行步骤 2）中编写的程序； 4） 查看程序运行结果
预期结果： 1） 步骤 3）中，程序成功完成运行； 2） 步骤 4）中，能够观察到程序中的数据插入、删除、修改、查看等 SQL 查询操作均成功执行并得到正确返回结果

6.4.3 SQL 支持度

测试编号：6.4.3
测试项目：SQL 支持度
测试目的：验证分布式批处理平台对于 SQL 数据类型、操作符、函数的支持情况
预置条件： 1） 分布式批处理平台正常运行
测试步骤： 1） 在 SQL 引擎模块中执行包含各类常用 SQL 数据类型的 SQL 语句（建议数据类型包括数值、字符、时间日期、布尔等）； 2） 在 SQL 引擎模块中执行包含各类常用 SQL 操作符的 SQL 语句（建议操作符包括数值运算、比较运算、逻辑运算、字符串拼接、类型强制转换等）； 3） 在 SQL 引擎模块中执行包含各类常用 SQL 函数的 SQL 语句（建议函数包括数值函数、字符函数、时间日期函数、类型转换函数、条件表达式、正则表达式、聚合函数等）
预期结果： 1） 步骤 1）中，能够成功并正确的执行包含各类常用 SQL 数据类型的 SQL 语句； 2） 步骤 2）中，能够成功并正确的执行包含各类常用 SQL 操作符的 SQL 语句； 3） 步骤 3）中，能够成功并正确的执行包含各类常用 SQL 函数的 SQL 语句

6.4.4 跨数据库关联操作

测试编号：6.4.4
测试项目：跨数据库关联操作
测试目的：验证分布式批处理平台中存储的数据表能够同关系型数据库进行关联操作
预置条件： 1） 存有非空数据表 t1 的分布式批处理平台正常运行； 2） 存有非空数据表 t2 的任意关系型数据库正常运行
测试步骤： 1） 在分布式批处理平台的 SQL 引擎中查询表 t1 内容； 2） 在关系型数据库中查询表 t2 内容； 3） 在分布式批处理平台中执行表 t1 和 t2 的关联操作，查看结果并与表 t1 和表 t2 应有的关联结果进行对比
预期结果： 1） 步骤 3）中，能够观察到关联操作结果正确

6.5 扩展性能力测试方法

6.5.1 在线水平扩展能力

测试编号：6.5.1
测试项目：在线水平扩展能力
测试目的：验证分布式批处理平台能够进行在线水平扩展
预置条件： 1） 分布式批处理平台正常运行； 2） 可用于扩展的服务器准备就绪
测试步骤： 1） 在分布式批处理平台上运行一个耗时较久的作业（例如 1TB 数据量的 Terasort 作业），并记录完成时间； 2） 通过运维监控模块在线水平扩展复数个节点； 3） 通过运维监控模块查看集群状态和节点数量； 4） 再次执行步骤 1）中的作业，并记录完成时间
预期结果： 1） 步骤 3）中，能够观察到集群完成在线水平扩展； 2） 步骤 4）中，作业完成时间同步骤 1）中记录的完成时间相比较短

6.5.2 在线水平收缩能力

测试编号：6.5.2
测试项目：在线水平收缩能力
测试目的：验证分布式批处理平台能够进行在线收缩扩展

预置条件： 1） 分布式批处理平台正常运行
测试步骤： 1） 在分布式批处理平台上运行一个耗时较久的作业（例如 1TB 数据量的 Terasort 作业），并记录完成时间； 2） 通过运维监控模块在线水平收缩复数个节点； 3） 通过运维监控模块查看集群状态和节点数量； 4） 再次执行步骤 1）中的作业，并记录完成时间
预期结果： 1） 步骤 3）中，能够观察到集群完成在线水平收缩； 2） 步骤 4）中，作业完成时间同步骤 1）中记录的完成时间相比较长

6.6 多租户支持能力测试方法

6.6.1 多租户管理与资源分配

测试编号：6.6.1
测试项目：多租户管理与资源分配
测试目的：验证分布式批处理平台能过实现租户的建立、删除和资源数量修改，在建立租户时能够按照指定的方式向不同租户分配不同数量的计算资源
预置条件： 1） 分布式批处理平台正常运行
测试步骤： 1） 新建队列 A 和队列 B，为两队列分配不同的计算资源（CPU 和内存）； 2） 修改队列 A 的计算资源数量； 3） 删除队列 B
预期结果： 1） 步骤 1）中，能够成功建立队列 A 和队列 B，并正确分配不同计算资源； 2） 步骤 2）中，能够成功修改队列 A 的计算资源数量； 3） 步骤 3）中，能够成功删除队列 B

6.6.2 多租户资源管理

测试编号：6.6.2
测试项目：多租户资源管理
测试目的：验证分布式批处理平台能够在进行多租户资源管理时能够实现不同的调度策略，可以包括先进先出调度策略（FIFO schedule）、计算能力调度策略（Capacity schedule）、公平调度策略（Fair schedule）等
预置条件： 1） 分布式批处理平台正常运行； 2） 正确建立队列 A 和队列 B

测试步骤： 1） 选择调度策略 1，并完成相关配置； 2） 启动多个作业； 3） 查看不同作业的资源分配情况； 4） 重复执行步骤 1）到步骤 3），并将调度策略修改为策略 2、策略 3 等
预期结果： 1） 步骤 3）中，能够观察到不同作业按照步骤 1）中选择的调度策略进行计算资源分配； 2） 在后续步骤中验证策略 2、策略 3 能够同策略 1 一样正确实现

6.6.3 多租户资源隔离

测试编号：6.6.3
测试项目：多租户资源隔离
测试目的：验证分布式批处理平台的多个租户间可以的实现计算资源相互隔离
预置条件： 1） 分布式批处理平台正常运行； 2） 正确建立队列 A 和队列 B
测试步骤： 1） 设置用户 a 可以使用对列 A 而不能使用队列 B； 2） 使用用户 a 向队列 A 提交资源需求超过队列 A 资源上限的作业； 3） 查看作业的资源使用情况和提交的队列； 4） 使用用户 a 向队列 B 提交作业
预期结果： 1） 步骤 2）中，能够成功向队列 A 提交作业； 2） 步骤 3）中，能够观察到作业提交在队列 A 上，仅使用了队列 A 的计算资源； 3） 步骤 4）中，不能向队列 B 提交作业

6.6.4 多租户资源使用情况监控

测试编号：6.6.4
测试项目：多租户资源使用情况监控
测试目的：验证分布式批处理平台能够实现对于不同租户资源使用情况的监控
预置条件： 1） 分布式批处理平台正常运行； 2） 正确建立队列 A 和队列 B
测试步骤： 1） 向队列 A 和队列 B 分配不同的计算资源（CPU 和内存）； 2） 分别向队列 A 和队列 B 提交作业；

3） 通过监控页面查看队列 A 和队列 B 中的资源使用情况
预期结果： 1） 步骤 3）中，能够成功查看队列 A 和队列 B 中的资源使用情况

6.7 运维管理能力测试方法

6.7.1 自动化部署

测试编号：6.7.1
测试项目：自动化部署
测试目的：验证分布式批处理平台所有模块支持以可视化向导的方式进行自动化部署
预置条件： 1） 硬件环境准备完毕； 2） 操作系统安装完毕； 3） 网络环境准备完毕
测试步骤： 1） 选择节点准备进行分布式批处理平台集群部署； 2） 以可视化向导的方式自动部署集群； 3） 部署完成后记录部署时间； 4） 查看集群系统状态
预期结果： 1） 在步骤 2）中，分布式批处理平台及各模块均可通过可视化向导的方式自动化部署； 2） 在步骤 2）中，可通过可视化形式提示部署进度； 3） 在步骤 4）中，可通过界面查看分布式存储模块、批处理计算模块均正常运行

6.7.2 资源监控

测试编号：6.7.2
测试项目：资源监控
测试目的：验证分布式批处理平台能够通过可视化界面的方式实现系统性能的监控管理功能
预置条件： 1） 分布式批处理平台正常运行
测试步骤： 1） 以网络页面或客户端形式启动资源监控功能； 2） 查看集群整体的资源使用情况； 3） 查看单个节点的资源使用情况； 4） 查看分布式存储模块基本信息以及资源使用情况； 5） 查看资源调度模块中当前作业以及历史作业的运行情况； 6） 查看系统性能历史数据

预期结果： 1） 在步骤 2）中，可通过可视化图表方式查看集群整体当前的 CPU、内存、存储、网络使用情况； 2） 在步骤 3）中，可通过可视化图表方式查看单个节点当前的 CPU、内存、存储、网络使用情况； 3） 在步骤 4）中，可查看分布式存储模块中块总数、总存储量、文件总数、剩余存储量、损坏块数量等信息； 4） 在步骤 5）中，可查看运行中和已完成的作业信息； 5） 在步骤 6）中，可通过可视化图表方式查看集群整体、单个节点的性能历史数据

6.7.3 作业监控

测试编号：6.7.3
测试项目：作业监控
测试目的：验证分布式批处理平台能够正常运行分布式批处理作业，同时能够通过可视化监控页面查看作业的作业状态、资源使用情况等
预置条件： 1） 分布式批处理平台正常运行
测试步骤： 1） 启动一个能够正常运行的分布式批处理作业； 2） 通过可视化界面查看作业的作业状态和资源使用情况； 3） 作业完成后查看作业状态和资源使用情况
预期结果： 1） 在步骤 2）中，可通过可视化界面查看当前作业的作业状态和资源使用情况； 2） 在步骤 3）中，可查看历史作业的作业状态和资源使用情况

6.7.4 集群操作

测试编号：6.7.4
测试项目：集群操作
测试目的：验证分布式批处理平台所有功能模块和各节点主机可通过可视化界面停止和启动
预置条件： 1） 分布式批处理平台正常运行
测试步骤： 1） 通过可视化界面一次性停止系统所有功能模块； 2） 通过可视化界面一次性启动系统所有功能模块； 3） 查看各个模块的运行状态； 4） 通过可视化界面依次停止、启动各功能模块； 5） 查看任一节点主机的运行状态； 6） 通过可视化界面启动、停止任一节点主机
预期结果：

1） 在步骤 1）中，通过一次操作成功停止系统所有功能模块；
2） 在步骤 2）中，通过一次操作成功启动系统所有功能模块；
3） 在步骤 3）中，可以正确查看各个功能模块的运行状态；
4） 在步骤 4）中，可以成功停止、启动各个功能模块；
5） 在步骤 5）中，可以正确查看任一节点主机的运行状态；
6） 在步骤 6）中，可以成功停止、启动任一节点主机

6.7.5 故障管理

测试编号：6.7.5
测试项目：故障管理
测试目的：验证分布式批处理平台在发生故障时，能够通过可视化界面、电子邮件、短信等方式进行告警，同时能够对故障进行分级、设置告警触发阈值、设置告警信息等操作
预置条件： 1） 分布式批处理平台正常运行
测试步骤： 1） 通过可视化界面对不同告警级别（假设为 A 级别和 B 级别）的阈值、告警方式、告警信息等分别进行不同的配置； 2） 人为制造应能够触发 A 级别告警而不触发 B 级别告警的故障； 3） 检查是否进行 A 级别告警，告警方式、告警信息是否与配置一致； 4） 检查是否未进行 B 级别告警； 5） 修改级别 B 告警阈值，使步骤 2）中故障应能够同时触发 A 级别和 B 级别告警； 6） 人为制造与步骤 2）中相同的故障； 7） 检查是否进行 B 级别告警，告警方式、告警信息是否与配置一致
预期结果： 1） 步骤 3）中，能够正确通过步骤 1）中配置的告警方式发送告警信息以进行 A 级别告警； 2） 步骤 4）中，并未进行 B 级别告警； 3） 步骤 7）中，能够正确通过步骤 1）中配置的告警方式发送告警信息以进行 B 级别告警

6.7.6 日志管理

测试编号：6.7.6
测试项目：日志管理
测试目的：验证分布式批处理平台能够通过可视化界面对系统各功能模块的日志进行查看、检索和下载
预置条件： 1） 分布式批处理平台正常运行
测试步骤： 1） 通过可视化界面查看系统任一功能模块的日志； 2） 检查后台存储的对应日志，将其内容与步骤 1）中查看的日志内容进行对比； 3） 通过可视化界面对步骤 1）中查看的日志内容进行检索；

4） 通过可视化界面对步骤 1）中查看的日志进行下载，并将下载后的日志内容与步骤 1）中查看的日志内容进行对比
预期结果： 1） 步骤 2）中，对比步骤 1）和步骤 2）中查看的日志内容发现两者一致； 2） 步骤 3）中，能够对日志内容进行正确的检索； 3） 步骤 4）中，能够对日志进行下载，对比步骤 1）和步骤 4）中查看的日志内容发现两者一致

6.7.7 配置管理

测试编号：6.7.7
测试项目：配置管理
测试目的：验证分布式批处理平台能够通过可视化界面对各功能模块的配置信息进行查看和修改
预置条件： 1） 分布式批处理平台正常运行
测试步骤： 1） 通过可视化界面查看系统各功能模块的配置信息，并各选择一项可进行修改的配置信息进行记录； 2） 通过系统后台查看各功能模块的配置信息，检查步骤 1）中记录的配置项的配置信息； 3） 通过可视化界面对步骤 1）中记录的配置项的配置信息进行修改，并再次记录修改后的配置信息； 4） 再次执行步骤 2）
预期结果： 1） 步骤 2）中，所检查的配置信息与步骤 1）中记录的配置信息相同； 2） 步骤 4）中，所检查的配置信息与步骤 3）中记录的配置信息相同

6.7.8 用户管理

测试编号：6.7.8
测试项目：用户管理
测试目的：验证分布式批处理平台能够通过可视化界面对用户组进行增加、删除、权限修改，能够对用户进行增加、删除、所属用户组修改、密码初始化
预置条件： 1） 分布式批处理平台正常运行
测试步骤： 1） 通过可视化界面新建用户组 A 和用户组 B，向用户组 A 分配能够执行操作 X 但不能执行操作 Y 的权限，向用户组 B 分配能够执行操作 Y 但不能执行操作 X 的权限； 2） 通过可视化界面新建用户 a 和用户 b，将用户 a 分配至用户组 A，将用户 b 分配至用户组 B； 3） 分别使用用户 a 和用户 b 登录并进行 X 操作和 Y 操作； 4） 通过可视化界面将用户 a 所属的用户组修改为用户组 B，将用户组 B 的权限修改为能够执行操作 X 同时能够执行操作 Y； 5） 再次执行步骤 3）； 6） 通过可视化界面删除用户 a，删除用户组 A，同时对用户 b 的密码进行初始化；

7） 查看修改用户 b 所属用户组时可选择的用户组； 8） 分别通过用户 a 和用户 b 的账号和原密码进行登录； 9） 通过用户 b 的账号和初始化密码进行登录
预期结果： 1） 在步骤 3）中，使用用户 a 登录时可以执行 X 操作但不能执行 Y 操作，使用用户 b 登录时可以执行 Y 操作但不能执行 X 操作； 2） 在步骤 5）中，使用用户 a 和用户 b 登录时均可以执行 X 操作和 Y 操作； 3） 在步骤 7）中，发现修改用户 b 所属用户组时可选择的用户组不再包含用户组 A； 4） 在步骤 8）中，使用用户 a 和用户 b 的原密码均无法登录； 5） 在步骤 9）中，使用用户 b 的账号和初始化密码能够成功登录

6.7.9 集群无宕机升级

测试编号：6.7.9
测试项目：集群无宕机升级
测试目的：验证分布式批处理平台能够在不影响作业正常执行的情况下完成各功能模块的升级
预置条件： 1） 分布式批处理平台正常运行； 2） 分布式批处理平台某功能模块可进行升级
测试步骤： 1） 查看某个可进行升级的功能模块当前的版本信息，并进行记录； 2） 启动耗时较长的分布式批处理作业（可以是排序、SQL 查询等作业）； 3） 选择步骤 1）中可升级的功能模块进行升级； 4） 查看步骤 2）中启动的作业的运行情况； 5） 待升级完成后，查看进行升级的功能模块当前的版本信息
预期结果： 1） 步骤 4）中，升级开始后步骤 2）中启动的作业正常执行，能够成功执行完成； 2） 步骤 5）中，版本信息已与步骤 1）中记录的版本信息不同，已升级成更高级版本

6.8 易用性能力测试方法

6.8.1 工作流创建

测试编号：6.8.1
测试项目：工作流创建
测试目的：验证分布式批处理平台能够支持工作流的创建，包括进行作业配置、作业依赖关系设置、作业参数传递设置等
预置条件： 1） 分布式批处理平台正常运行

测试步骤： 1） 通过可视化界面将批处理作业、SQL 查询任务各类作业任务配置成工作流； 2） 通过可视化界面进行步骤 1）中工作流的配置为后续任务的开始执行需要依赖于前序任务的成功执行； 3） 通过可视化界面进行步骤 1）中工作流的配置为前序任务可以向后续任务进行参数传递； 4） 通过可视化界面对于步骤 1）中工作流中的各任务进行配置和执行； 5） 查看步骤 1）中工作流执行情况
预期结果： 1） 步骤 1）至步骤 4）中，均能够成功进行相应配置操作； 2） 步骤 5）中，能够观察到工作流成功按照步骤 1）至步骤 4）中的配置执行

6.8.2 工作流管理

测试编号：6.8.2
测试项目：工作流管理
测试目的：验证分布式批处理平台能够支持工作流的管理，包括对于工作流中作业的增加、删除、修改、查找，工作流的复制、导入导出、周期触发、事件触发等
预置条件： 1） 分布式批处理平台正常运行； 2） 已存在可正确执行的多任务工作流
测试步骤： 1） 对已有工作流进行作业的增加、删除、修改、查找等操作； 2） 对已有工作流进行复制； 3） 对已有工作流进行导入和导出操作； 4） 对于步骤 1）至步骤 3）中操作的执行情况和结果进行查看； 5） 对已有工作流进行触发条件配置，包括按照时间周期触发和根据事件触发； 6） 对于步骤 5）中的配置进行验证
预期结果： 1） 步骤 4）中，能够观察到步骤 1）至步骤 3）中的操作均成功执行，操作结果均正确； 2） 步骤 6）中，能够验证步骤 5）中的配置生效，能够成功根据配置触发工作流的执行

6.8.3 工作流监控

测试编号：6.8.3
测试项目：工作流监控
测试目的：验证分布式批处理平台能够支持对于工作流和工作流内各个作业的监控和告警
预置条件： 1） 分布式批处理平台正常运行
测试步骤：

1） 运行包含多任务的工作流； 2） 通过可视化界面查看工作流和其中各个作业的执行情况； 3） 通过可视化界面进行工作流告警配置； 4） 手动造成作业错误，模拟工作流中作业发生故障； 5） 检查是否进行告警
预期结果： 1） 步骤 2）中，能够通过可视化界面成功观察到工作流和工作流内各个作业的执行情况； 2） 步骤 5）中，能够正确的根据步骤 3）中的配置进行告警

6.9 性能测试方法

6.9.1 概述

分布式批处理平台应该能够满足特定应用场景对大数据批处理的性能要求，性能测试主要考虑如下方面：

a） 测试对象包括批处理框架、SQL 引擎和机器学习框架的性能；

b） 测试数据生成应关注数据的格式、类型和规模；

c） 测试任务应该根据实际的常用应用场景进行选择。

6.9.2 测试前审核

测试前应该进行如下审核：

a） 测试数据审核：应审核测试数据的格式、类型和数据量。

b） 测试环境审核：应审核分布式存储模块的副本数，系统缓存清理情况。

c） SQL 查询性能审核：应审核测试数据索引、聚合情况，待执行的 SQL 语句及执行顺序。

d） NoSQL 数据读写操作性能审核：如果进行 NoSQL 数据读写操作性能测试，应审核读写操作命令，NoSQL 数据库模块的服务节点数量。

e） 机器学习算法性能审核：如果进行机器学习算法性能测试，应审核测试数据的生成脚本，测试算法的执行脚本。

6.9.3 测试执行中

测试中应按照测试步骤记录测试执行的操作过程。

6.9.4 测试后

a） 测试环境记录：应记录整体的测试环境，包括 CPU、内存、硬盘、网络设备等硬件环境信息，操作系统、平台各模块软件、其他中间件和应用程序等软件环境信息，分布式系统总节点数、各功能模块节点数、网络拓扑结构等其他环境信息。

b） 测试完成后，在测试报告中应完整真实的记录性能测试结果和相应测试环境信息。

6.9.5 SQL 查询性能测试

测试编号：6.9.5
测试项目：SQL 查询性能测试
测试目的：验证分布式批处理平台执行 SQL 查询任务的性能，参考指标主要为 SQL 查询语句的执行时间
预置条件： 1） 分布式批处理平台正常运行； 2） SQL 引擎模块正常运行
测试步骤： 1） 将本地数据上传至分布式存储模块； 2） 在 SQL 引擎模块中创建外表和执行表，导入数据； 3） 执行清理缓存操作； 4） 配置调优参数； 5） 连续执行给定的若干条 SQL 查询语句，SQL 查询语句可以基于标准基准测试集，也可以根据用户具体需求确定； 6） 查看步骤 5）中所有 SQL 查询语句的执行结果； 7） 记录步骤 5）中所有 SQL 查询语句的执行时间
预期结果： 1） 步骤 6）中，步骤 5）中所有 SQL 查询语句的结果均正确； 2） 步骤 7）中，记录的步骤 5）中所有语句的执行时间为该项性能测试结果

6.9.6 NoSQL 数据库写操作性能测试

测试编号：6.9.6
测试项目：NoSQL 数据库写操作性能测试
测试目的：验证分布式批处理平台执行 NoSQL 数据库写入操作的性能，参考指标主要为 NoSQL 数据库写入操作执行时的吞吐率
预置条件： 1） 分布式批处理平台正常运行； 2） NoSQL 数据库模块正常运行
测试步骤： 1） 执行清理缓存操作； 2） 通过给定数量的客户端（如 10 个），分别执行数据写入操作，规定每个客户端写入若干条数据（如 2 亿条），并设定每条数据大小约为 1KB（例如使用 YCSB 时设定参数为 fieldcount=10、fieldlength=100 即每条数据大小为 10×100/1024，约为 1KB）； 3） 写入操作执行完成后，记录每个客户端的吞吐率，吞吐率=数据条数/完成数据写入的时间，系统总吞吐率为各客户端吞吐率的总和； 4） 查看已写入的数据量
预期结果：

1） 步骤 3）中，记录的所有客户端的吞吐率及吞吐率总和为该项性能测试结果； 2） 步骤 4）中，能观察到已写入的数据量与步骤 2）中预期的数据量近似

6.9.7 NoSQL 数据库读操作性能测试

测试编号：6.9.7
测试项目：NoSQL 数据库读操作性能测试
测试目的：验证分布式批处理平台执行 NoSQL 数据库读操作的性能，参考指标主要为 NoSQL 数据库执行 95%读操作和 5%更新操作时的吞吐率
预置条件： 1） 分布式批处理平台正常运行； 2） NoSQL 数据库模块正常运行
测试步骤： 1） 执行清理缓存操作； 2） 通过给定数量的客户端（如 10 个），分别执行 95%的数据读操作，和 5%的更新操作； 3） 步骤 2）中操作执行完成后，记录每个客户端的吞吐率，吞吐率=数据条数/完成数据写入的时间，系统总吞吐率为各客户端吞吐率的总和
预期结果： 1） 步骤 3）中，记录的所有客户端的吞吐率及吞吐率总和为该项性能测试结果

6.9.8 NoSQL 数据库读写混合操作性能测试

测试编号：6.9.8
测试项目：NoSQL 数据库读写混合操作性能测试
测试目的：验证分布式批处理平台执行 NoSQL 数据库读写混合操作的性能，参考指标主要为 NoSQL 数据库执行 50%读操作和 50%写操作时的吞吐率
预置条件： 1） 分布式批处理平台正常运行； 2） NoSQL 数据库模块正常运行
测试步骤： 1） 执行清理缓存操作； 2） 通过给定数量的客户端（如 10 个），分别执行 50%的读操作，和 50%的写操作； 3） 步骤 2）中操作执行完成后，记录每个客户端的吞吐率，吞吐率=数据条数/完成数据写入的时间，系统总吞吐率为各客户端吞吐率的总和
预期结果： 1） 步骤 3）中，记录的所有客户端的吞吐率及吞吐率总和为该项性能测试结果

6.9.9 Kmeans 算法执行性能测试

测试编号：6.9.9
测试项目：Kmeans 算法执行性能测试
测试目的：验证分布式批处理平台批处理计算模块执行 Kmeans 机器学习算法的性能，参考指标主要为 Kmeans 算法执行完成的时间
预置条件： 1） 分布式批处理平台正常运行； 2） 批处理计算模块正常运行； 3） 下载 Spark Bench 工具，利用数据生成工具生成指定规模的数据（如 400GB）
测试步骤： 1） 执行清理缓存操作； 2） 设置算法参数： NUM_OF_POINTS=400000000 NUM_OF_CLUSTERS=500 DIMENSIONS=60 SCALING=0.6 MAX_ITERATION=10 NUM_RUN=1 3） 执行按照步骤 2）中参数，对已生成好的数据执行 Kmeans 算法； 4） 步骤 3）执行完成后，记录算法执行时间
预期结果： 1） 步骤 4）中，记录的 Kmeans 算法执行时间为该项性能测试结果

6.9.10 SVM 算法执行性能测试

测试编号：6.9.10
测试项目：SVM 算法执行性能测试
测试目的：验证分布式批处理平台批处理计算模块执行 SVM 机器学习算法的性能，参考指标主要为 SVM 算法执行完成的时间
预置条件： 1） 分布式批处理平台正常运行； 2） 批处理计算模块正常运行； 3） 下载 Spark Bench 工具，利用数据生成工具生成指定规模数据（如 1.8TB）
测试步骤： 1） 执行清理缓存操作； 2） 设置算法参数： NUM_OF_EXAMPLES=500000000 NUM_OF_FEATURES=200

MAX_ITERATION=3 3） 执行按照步骤 2）中参数，对已生成好的数据执行 SVM 算法； 4） 步骤 3）执行完成后，记录算法执行时间
预期结果： 1） 步骤 4）中，记录的 SVM 算法执行时间为该项性能测试结果