

[文章编号] 1003—4684(2020)05-0055-07

大数据软件测试技术研究

刘 珊 艳¹, 胡 秀¹, 严 武²

(1 荆楚理工学院计算机工程学院,湖北 荆门 448000; 2 荆门市优思信息服务有限公司,湖北 荆门 448000)

[摘 要] 在介绍了大数据测试的主要技术后,给出了面向大数据的测试过程,由于 ETL(Extract Transform and Load 抽取、转换和加载)测试是数据仓库测试中重要且复杂的阶段,给出了 ETL 测试的主要类型及 ETL 自动化测试的优势。通过对员工信息进行数据填充测试及测试结果分析,说明 ETL 测试是保证数据质量有效性的重要途径。研究表明合理的搭建测试环境,应用自动化测试技术,可以提高测试效率以降低大数据测试的难度。

[关键词] 大数据; 软件测试; 大数据测试; ETL 测试; 测试数据

[中图分类号] TP311 [文献标识码] A

大数据时代的到来,使得数据成为了重要的经济资产^[1]。大数据时代,也颠覆了以往对数据的惯性思考方式,要保证数据执行,软件质量、测试质量、数据使用场景等,都需要重新变换一个新的角度,对软件进行更全方位的思考。

软件测试是为了发现软件缺陷而运行测试的过程。对于常规的系统测试,测试人员根据需求规格说明的描述,判断系统的输出结果与需求描述的预期结果是否一致;若一致,系统的准确性就得到了保证;若不一致,系统就是有缺陷的。这个看似必然的测试准则在大数据系统测试中已经不成立了。大数据系统软件测试,在很多场景下系统预期的输出结果是无法直接确定的。

大数据时代对软件测试要求提升到了一个新高度。大数据系统架构设计的复杂性使得系统测试也非常复杂,软件测试的各个方面都会受到大数据的影响^[2],这使得大数据测试非常有挑战性。

1 大数据测试难点

1.1 测试理论

当前软件测试理论大部分是 20 多年前提出的基础测试理论。这些理论仍然可以设计出很完善的测试案例,前提是软件功能明确,且在需求规格说明中作了准确的描述。

假设功能明确的程序 P 是为了实现规格说明 S 中的所有需求: $I_1, I_2, \dots, I_n (n > 1)$ 是 S 的 n 个自变量;输入集合 $I = \{I_1, I_2, \dots, I_n (n > 1)\}$ 时,期望

的输出假设为集合 $O' = \{O'_1, O'_2, \dots, O'_m$; 对于按照规格说明 S 具体实现的程序 P ,在测试时,输入集合 $I = \{I_1, I_2, \dots, I_n (n > 1)\}$ 时,实际的输出为 $O = \{O_1, O_2, \dots, O_m$;比较 O 和 O' 的值,当 $O = O'$ 时,该测试用例 I 通过了测试。

在大数据场景下无论是趋势分析还是图论计算都变得极其困难。因此预期输出结果 O' 无法确定,在这种情况下,确定测试用例 I 是否能够通过也同样变得极其困难^[3]。

1.2 测试过程

大数据应用的鲜明特征之一就是数据的多样性,既有结构化的关系数据、图数据、轨迹数据,也有非结构化的文本数据、图片数据,甚至是视频数据等。大数据软件的一个基本要求就是能够支持结构化、半结构化、非结构化等多种数据类型的组织、存储和管理,形成以量质相融合的知识管理为中心、并以此提供面向知识服务的快速应用开发接口^[4]。故大数据应用程序的测试,除了要验证其功能、性能,还要验证数据处理。即数据能否正确地加载至系统,加载后的数据与源数据是否一致,数据映射、清洗过程是否正常,以及经过大数据处理框架后数据的准确性和完整性。

1.3 测试思维

大数据系统的应用,目的是为了得到数据和数据之间的关联关系。数据关联颠覆了以往对数据的惯性思考方式。保证数据质量、软件质量、数据利用率等,都需要站在更高的高度以全新的角度进行全

[收稿日期] 2019—11—19

[基金项目] 湖北省教育厅科研项目(B2018242, B2020193);荆门市科技局科研项目(2019YFZD010, 2018YDKY071)

[第一作者] 刘珊艳(1979—),女,湖北潜江人,荆楚理工学院讲师,研究方向为软件测试技术

方位的思考。这种思考方式称为大数据思维,从数据的应用场景切入,思考如何挖掘数据本身的价值,并将其转变为市场价值。所以大数据软件的思维方式和传统软件将完全不同。即需要全部数据样本而不是抽样、关注效率而不是精确度、关注相关性而不是因果关系^[5]。大数据测试除了关注系统功能和性能外,还要关注数据本身的价值。面对海量数据,思考如何应用这些数据,如何有效提升数据的价值是测试的关键所在。

2 测试的主要技术

2.1 数据处理流程

根据数据从数据源向商务决策报表转化的过程,总结整理大数据的处理流程如图 1 所示。

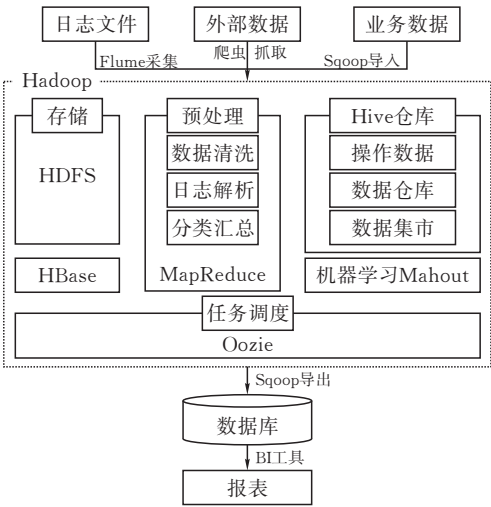


图 1 大数据处理流程

首先,进行数据采集、导入。这一步将根据具体的应用背景和业务需求,将各种数据源如网络日志、物联网、互联网文本和文件等存储到多个数据库中,在这一阶段可以使用 Flume 工具、爬虫工具和 Sqoop 工具。其次,进行数据清洗分析。数据加载到 Hdfs 后,MapReduce 开始对原始数据进行清洗和统计分析。把一堆杂乱无章的数据按照某种特例归纳起来,然后处理得到想要的结果。Hive 可以应用到数据分析的场景中来,它是一种 ETL(Extract-Transform-Load 抽取、转换和加载)工具,数据经过 ETL 产生中间表或产生最终的报表。第三,使用数据挖掘技术。对于前面阶段处理后的数据进行数据挖掘,使用不同的算法对数据进行计算,从而满足更进一步数据分析需求。数据挖掘作为一个术语,用于以下六类活动:分类、估算、预测、关联规则、聚类、描述^[6],主要使用的技术是机器学习以及能够从复杂数据中获取有价值的知识的深度学习等^[7]。最后,就是数据展示。当数据统计分析结束后,产生的

数据可以转移到数据仓库或有特定应用的数据集中,也可以使用 Sqoop 将产生的数据导出到传统的数据库如 Mysql 中。Hadoop 中任务的调度与协调由任务调度框架 Oozie 来完成。对于导出的数据可以使用 BI(Business Intelligence 商业智能)工具产生报表供用户作出准确的决策。

2.2 大数据测试策略

大数据应用程序测试更多的是验证其数据处理过程,而不仅仅是测试软件产品的单个特性。对于大数据测试工程师而言,如何高效正确的验证经过大数据工具/框架成功处理过的至少百万兆字节的数据的准确性将会是一个巨大的挑战。大数据处理的三个特性^[8]是大批量、实时性、可交互。

除此之外,在测试应用程序之前,有必要检查数据的质量,并将其视为数据库测试的一部分。它包括检查各种特征,如一致性、准确性、重复性、数据完整性等。

2.3 大数据测试传统数据库测试对比

大数据多源异构、规模巨大、快速多变等特性使得传统的计算不能有效支持大数据的处理、分析和计算^[9]。同样,在进行大数据的测试时,由于数据规模大,内在关联关系密切而复杂,价值密度分布极不均衡,这些特征都要求大数据测试与传统的数据库测试技术全然不同。表 1 给出了传统数据库测试与大数据测试的不同之处。

表 1 传统数据库测试 VS 大数据测试

特性	传统数据库测试	大数据测试
数据	结构化数据有效的、明确定义的测试方法,测试人员可选择手动抽样测试,也可选择自动化穷举测试	结构化数据和非结构化数据 测试方法需要研发的支持 大数据抽样测试是一个挑战
基础架构	由于文件大小是有限的,不需要特殊的测试环境	由于数据和文件的巨大,需要搭建特殊的测试环境
验证工具	使用基于 excel 的宏命令或基于 UI 的自动化工具	范围很广,没有定义的工具
	拥有基本的操作知识和较少的培训就可以使用测试工具	需要掌握一套特定的技能和专业的培训来操作测试工具。此外,工具也是新生工具,随着时间的推移也会不断具有新的功能和特性

为了支持数据的读写删,测试需要对提供给用户的所有的基本功能(接口)进行测试,保证基本功能的正确。由于大数据系统往往由服务器集群组成,目前可达到成千上万个核的集群。测试需要在上百台甚至上千台 Blade 机器进行,以期覆盖几十

种操作系统。硬件支持,性能、压力、可用性、安全性、浏览器、数据库等都是大数据测试的要点^[10]。

大数据测试不同于常规的应用测试,为了应对数据爆炸性增长,应该具备以下一些基础环境:1)拥有足够的存储设备来存储和处理大数据;2)拥有集群来做分布式节点和数据处理;3)至少拥有足够的 CPU、内存来确保有高性能的处理基础。

2.4 大数据测试流程

大数据应用的测试过程见图 2,不同结构的数据首先被采集加载至 Hadoop 系统中,再通过 ETL 技术将业务系统的数据抽取、清洗转换之后加载到数据仓库,为 BI 商业智能的决策提供分析依据。

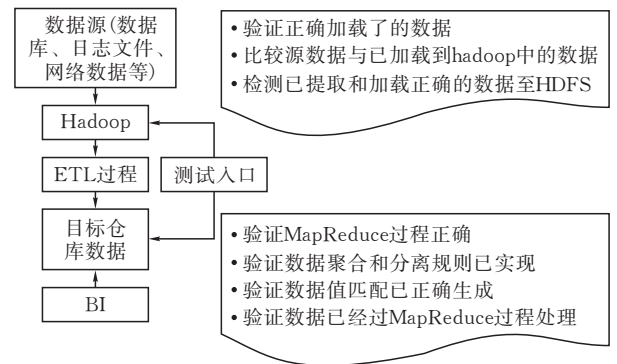


图 2 大数据测试流程

大数据测试大体可以分为三大步骤。

第一步,数据预处理验证。在进行大数据测试时,首先要在 Hadoop 前验证数据的准确性。数据来源可能是关系数据库、日志系统、社交网络等,应该确保正确的数据加载到系统中。要验证加载的数据和源数据是一致的。要验证数据被正确的提取和加载至 HDFS 中,且被分割、复制到不同的数据节点中。

第二步,MapReduce 数据输出验证。当数据加载进行 HDFS 后,MapReduce 开始对来自不同数据源的数据进行处理。在本阶段,主要验证每一个处理节点的业务逻辑是否正确。Map 过程和 Reduce 过程工作正常。验证数据聚合、合并是否正确。验证 MapReduce 处理过程的 key/value 对已正确生成。验证经过 MapReduce 后数据的准确性。

第三步,输出结果验证。当 MapReduce 过程结束后,产生的数据输出文件将被按需移至数据仓库或其它的事务型系统中。在此过程中,可能会由于不正确地应用转换规则,从 HDFS 中提取的数据不完全而带来问题,这阶段主要验证:验证数据转换规则是否正确应用;验证目标系统数据加载是否成功;验证数据的完整性;通过比较目标数据和 HDFS 文件数据来验证是否有数据损坏。

3 ETL 测试

3.1 ETL 测试过程

ETL 即数据抽取、转换、装载的过程。ETL 能够转换不同结构、类型的数据集为统一的结构,以便后续使用 BI 工具生成有意义的分析和报表。ETL 测试的目的是确保在业务转换之后从源加载到目标的数据是准确的。它还包括在源和目标之间使用的各种中间阶段验证数据。与其他测试过程相似,ETL 也经历了不同的阶段。ETL 过程和 ETL 测试过程的不同阶段如图 3 所示。ETL 测试要完成的任务分别是:确定数据源和需求,数据获取,实现业务逻辑和维度建模,建立和填充数据,生成报告。

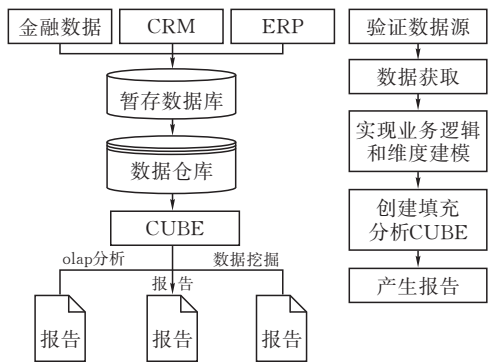


图 3 ETL 过程图和 ETL 测试过程

3.2 ETL 测试内容

随着业务的发展扩张,产生的数据越来越多,这些数据的收集方式、原始数据格式、数据量、存储要求、使用场景等方面有很大的差异。为了确保从源数据到目标数据经过不同阶段业务转换完成后是准确的,ETL 测试由此展开^[11]。ETL 测试的主要内容如下。

- 1)生产确认测试。该类型的 ETL 测试是在数据迁移至生产系统时进行的。为了提高商业决策的准确性,生产系统中的数据必须以正确的顺序进行排列。
- 2)源数据到目标数据测试。该类型的测试主要验证所转换的数据值是否是预期的数据值。
- 3)升级测试。该类型的 ETL 测试可以自动生成,能节省大量的测试开发时间。主要检查旧应用或存储库中提取的数据是否与新应用或新存储库中的数据完全相同。
- 4)元数据测试。元数据测试包括数据类型检查、数据长度检查和索引/约束检查。
- 5)数据完整性测试。为了验证所有预期的数据都从源加载到目标中,需要进行数据完整性测试。在数据完整性测试过程中,还可以进行一些简单的转换或无转换的源与目标之间的计数、聚合和实际

数据比较和验证的测试。

6)数据准确性测试。确保数据按预期准确加载和准确转换是数据准确性测试的目标。

7)数据转换测试。数据转换测试是一个复杂的过程,并不是简单的写一个源 SQL 查询并将输出与目标数据进行比较来实现的。可能需要为每一行运行多个 SQL 查询来验证转换规则。

8)数据质量测试。数据质量测试包含语法测试和参照测试。为了避免在业务过程中由于日期或唯一编号(例如订单号)引起的错误,需要进行数据质量测试。语法测试:根据无效字符、字符模式、不正确大小写等报告脏数据。参照测试:基于数据模型检查数据,例如客户 ID。数据质量测试包含:数字检查、日期检查、精度检查、数据检查、空值校验等。

9)增量 ETL 测试。增量测试验证插入操作和更新操作在增量 ETL 过程中是否按照预期被处理,并检查添加新数据时新旧数据的数据完整性。

10)GUI/导航测试。该类型测试主要检查大数据报告的 GUI/导航方面是否正常。

3.3 ETL 自动化测试

ETL 测试过程是测试数据仓库中一个至关重要的阶段,几乎也是最复杂的阶段,因为它直接影响数据的质量。每次 ETL 过程的失效都会导致在数据仓库中加载不正确的数据,从而导致提供错误的数据以供商务决策,最终导致不准确决策。ETL 测试的一般方法是使用 SQL 脚本或手工测试数据。这些 ETL 测试方法非常耗时,容易出错,并且很少提供完整的测试覆盖率。为了在生产和开发环境中加速、提高 ETL 测试的覆盖率、降低成本、提高缺陷检测率,自动化测试已经被证明是提高数据仓库系统质量的有效手段。但 ETL 自动化测试还面临一些挑战和限制^[12]:

1)人们普遍认为并不是所有的数据仓库测试都可以自动化,只是一些关键的和重复的测试可以使用自动化工具来实现。

2)数据仓库由许多表和记录组成,增加了测试的复杂性。

3)数据仓库中的数据来自多个源系统,在测试过程中,必须检查来自源系统和加载在数据仓库中的数据之间的数据。

4)自动化测试不能完全取代手工测试。仍然需要手工测试来处理自动化可能无法捕获所有内容的复杂情况。

5)业务对象报告测试仍然是自动化测试的一个挑战。

6)自动化工具的成本较高导致人们依赖于手工

测试。

图 4 给出了自动化测试的框架。建议的框架包括:三个存储库、映射文档、两个流程和结果报告。这三个存储库分别是:存储数据仓库元数据的元数据存储库、用于存储映射文档和元数据存储库所需信息的数据库模式存储库、存储质量参数定义及其相关测试用例的测试程序存储库。将使用两个流程根据前面存储库中存储的数据生成和执行测试用例,最后生成关于通过的测试用例的结果报告。

ETL 自动化测试的优势有^[13]:1)减少花费在测试阶段的时间;2)测试的可重用性;3)节省人力;4)利用工具生成测试报告并记录测试结果;5)每更改数据或业务规则后减少回归测试工作。

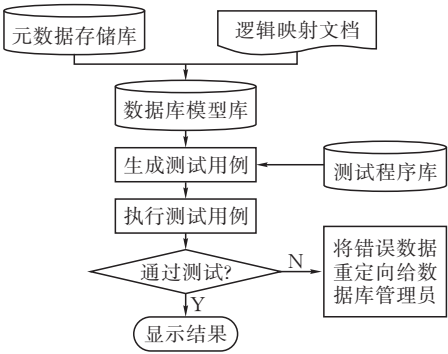


图 4 ETL 自动化进程测试框架

4 案例研究

4.1 案例描述

在本节中,使用员工信息数据库来检测 ETL 过程执行后,源数据到目标数据转换的准确性。员工信息表的数据由另一个数据库中的多表填充,图 5 显示用于填充员工维度表的源数据库。源数据库包括 Person(人员)表、Employee(员工)表、Employ-DepartmentHistory(员工部门历史)表、SalesPerson(销售人员)表、Department(部门)表。ETL 过程将对 DW(Data Warehouse 数据仓库)中员工信息表进行数据填充,从源数据库中抽取特定的列,抽取出的数据将与 DW 中表的数据进行比较,只有确定是新记录才能添加,其他记录被抛弃不使用。

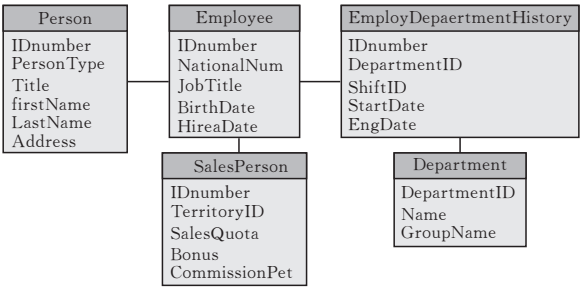


图 5 源数据库的联系

4.2 测试环境搭建

测试用例的产生和执行是使用图 4 所示的框架。第一步,定义引用涉及的数据库:源数据库、暂存数据库和 DW。第二步,涉及 ETL 逻辑映射文档和元数据存储库。逻辑映射文档包含从源数据库提供的每个字段到目标的过程。它包含以下字段:目标表名、目标表类型列名、SCD(Slowly changing Dimension 渐变维度)类型、源数据库、源表名、源列名和转换。最后,建立到元数据存储库的连接,目前市场上有许多工具可以在 DW 阶段实现元数据的交换,如:MDS 和 MIMS。

4.3 测试程序执行

从逻辑映射文档和 DW 元数据中提取所需数据后,这些数据将被加载到数据库模型中。对于数据模型中的质量参数,都分配了一些例行程序,通过研究影响每一个质量参数的所有质量问题,找到检测出质量问题的测试程序,手动的为每一个质量参数分配测试程序,指定的测试程序将在每一个维度表/实际表上执行。在这一步骤中测试用例中抽取出的每一个字段将被赋予来自数据库模型中的值(表名、列名等)。用于执行的算法如下:

开始
对每一个逻辑数据映射文档中的维度表/实际表执行以下操作:

开始
对于每一个测试程序库中的质量参数
开始
获取相应的测试用例;
对于每一个测试用例
将测试用例的字段映射为数据库模型中的对应值;
执行测试用例。
如果测试失败,重定向测试用例并提交测试详细信息;
否则 显示结果;
结束 如果;
结束// 结束每个测试用例
结束// 结束每个质量参数
结束// 结束每个表
结束

在对所选条目执行测试之后,成功的测试用例将直接通过到下一步,失败的测试用例可以使用各种策略来处理在需求分析阶段确定的 ETL 错误。例如:自动清除错误数据、抛出错误数据及常用处理方式——错误数据交给数据库管理员。

4.4 测试结果分析

通过对装载在维度表中的真实数据及代表无效 ETL 过程存在的各种错误类型的模拟数据这两个不同的数据库进行测试,得到测试结果见表 2。结果显示植入错误后,所有的测试用例都没有通过测试。

表 2 生成的测试用例及测试结果

DQ 维度	测试用例名称	期望输出	实际结果	测试结果
完整性	重复性检查	列值是唯一	目标表中数据有冗余	失败
	完整性约束检查	外键/主键可维护	违反外键/主键约束	失败
	超出边界值检查	目标表中的数据位于特定范围内	违反目标表中的范围规范	失败
	SCD 检查	SCD 列中的数据与指定类型兼容	SCD 属性的值与定义的类型不兼容	失败
	记录计数验证检查	目标表的记录数与源表中的记录数相同	源和目标之间属性数不匹配	失败
有效性	数据类型检查	源文件数据类型与目标数据类型相同	数据类型不匹配	失败
	字段长度检查	源表和目标表中字段长度一致	字段长度不匹配	失败
一致性	字段映射检查	字段映射依赖于逻辑映射文档规范	字段映射不匹配	失败
	度量值检查	度量值计算准确	度量函数结果错误	失败
准确性	超出边界值检查	目标表中的数据位于特定范围内	目标表中的值超出边界	失败
	截断值检查	目标表中的数据与源表相同	源表中的数据与目标表中的数据不匹配	失败

分析测试结果,就可以通过比较检测出的故障与植入错误数来评估系统的有效性了。系统生成的所有报告表明,所有植入的错误均被检测到了,验证了该方法在 ETL 阶段检测数据质量的有效性。通过在不同的数据集上进行测试(从 10000 到 50000 记录)证明了该方法在检测不同数据集上错误的有效性。

5 总结与展望

为了保证大数据系统的质量,需要对大数据应

用程序进行全面测试。大数据应用程序的测试重点之一就是验证数据处理,验证加载前数据的准确性,验证处理过程数据,验证输出结果数据。ETL 测试的目的就是确保在业务转换之后从源加载到目标的数据是准确的,它还包括在源和目标之间使用的各种中间阶段验证数据,以便后续使用 BI 工具生成有意义的分析表报。

未来大数据系统将面临许多重要挑战,这些挑战源于数据的本质:庞大、多样和不断发展,组织捕获和处理这些数据的能力受到了限制。当前的技

术、架构、管理和测试方法都需要进一步发展来面对数据的洪流,或者处理机构需要改变思维、计划、管理、处理和报告数据的方式,以真正发挥大数据的潜力。以下是研究人员和测试人员在未来几年必须面对的一些挑战。

1)实时性:大数据生成速度很快,很多数据是在线的,并且需要实时处理。大数据的实时分析是电子商务提供在线服务的关键。且所产生的数据中还存在大量的噪声对象、不完整对象、不准确对象、不精确对象和冗余对象^[14],大数据的规模在飞速增长,到 2020 年将达到 35 ZB。如何判断是否从大数据中发现并提取有价值的信息,实现快速响应和实时决策,是大数据测试研究的重点。

2)隐私安全:大数据技术的发展也极大威胁着个人隐私,政府机构在利用数据分析作出决策的同时还要致力于保护本国公民的隐私权,澳大利亚政府通过了《2012 年隐私修正案(加强隐私保护)法案》,加强了对个人信息的保护,并为个人信息的使用设定了更清晰的界限。欧洲法院(Europen Court of Justice)制定“被遗忘权”,规定欧洲公民有权要求搜索引擎删除被认为不准确、不相关或过度开发条目^[15]。美国、英国都相继成立数据保护的立法,我国于近几年颁布《网络安全法》、《个人信息保护法》等法律法规^[16]。政府机构在收集或管理公民数据时,受到一系列立法控制,必须遵守一系列行为和法规。这些立法工具旨在维持公众对政府作为有效和安全的公民信息存储库和管理者的信心。政府机构使用大数据不会改变这一点,相反,它可能会在管理信息安全风险方面增加一层额外的复杂性。大数据源、机构内部和跨机构的传输系统,以及这些数据的端点,都将成为本地和国际黑客感兴趣的目标,需要得到保护。研究开发的大数据系统应该具有相当高的安全性能,以达到保护公民隐私权的目的。

3)隐藏的大数据:大量有用的数据正在丢失,因为新数据大部分是基于无标记文件和非结构化数据的。国际数据公司(IDC)大数据研究表明,大数据技术趋于成熟,但数据壁垒仍然存在,数据价值没有充分发挥出来。

4)当前可以为大数据系统测试的规模,在很大程度上取决于 CPU 集群和 GPU 等高性能计算架构的增长。不幸的是,计算性能的提高远远落后于大数据的增长速度。大数据的最大问题之一就是基础设施的高成本,即使有了云计算技术,硬件设备也是非常昂贵的^[17]。

迎接大数据带来的挑战将是困难的,数据量已经非常巨大,而且每天都在增加。越来越多的设备

连入互联网,导致数据产生和增长速度正在加快,且生成的数据种类也在不断增加,大数据采集、分析、处理、测试的需求正在所有的科学和工程领域兴起。借助大数据技术,有望提供相关性更大、更准确的社会感知反馈,更好地实时了解社会,也进一步鼓励公众参与和社会和经济活动的数据制作圈来。

[参 考 文 献]

[1] 李昊,张敏,冯登国,等. 大数据访问控制研究[J]. 计算机学报,2017,40(1):72-91.

[2] Chandrashekar A M, Bhavana S. Extending search based software testing techniques to big data applications[J]. International Journal of Research and Scientific Innovation ,2017,4(6):142-146.

[3] 蔡立志,阎婷. 大数据背景下软件测试的挑战与展望[J]. 计算机应用与软件,2014,2(31):5-8.

[4] 杜小勇,卢卫,张峰. 大数据管理系统的历史、现状与未来[J]. 软件学报,2019,30(1):127-141.

[5] Mayer-Schönberger V, Cukier K. Big data: a revolution that will transform how we live, work, and think. Houghton Mifflin Harcourt [M]. London: Houghton Mifflin Harcourt, ,2015.

[6] Rajitha M, Sravanthi P. Data mining with big data[J]. International Journal of Research in Advanced Computer Science Engineering,2015,1(3):55-60.

[7] Zhang qingchen, Laurence T. Yang, Zhikui Chen, et al. A Survey on deep learning for big data[J]. Information Fusion,2018(42):146-157.

[8] Big data testing tutorial. what is, strategy, how to test Hadoop[EB/OL]. [2019-11-10]. <https://www.guru99.com/big-data-testing-functional-performance.html>.

[9] 程学旗,靳小龙,王元卓,郭嘉丰,张铁赢,李国杰. 大数据系统和分析技术综述[J]. 软件学报, 2014, 25(9): 1889-1908.

[10] 代亮,陈婷,许宏科,等. 大数据测试技术研究[J]. 计算机应用研究,2014,31(6): 1606- 1611.

[11] Krishna. ETL testing or data warehouse testing tutorial [EB/OL]. [2019-11-10]. <https://www.guru99.com/ultimate-guide-etl-datawarehouse-testing.html>.

[12] Vucevic D, Yadow W. Testing the data warehouse practicum- assuring data content, data structures and quality, trafford[R]. Trafford,2012.

[13] Tarek M Mahmoud, Sara B. Dakrory, Abdelmgeid A. Ali. Automated ETL testing on the data quality of a data warehouse [J]. International Journal of Computer Applications,2015,131(16):9-15.

[14] Saha B, Srivastava D. Data Quality: The other face of big data[C]. Proceedings of IEEE International Conference on Data Engineering, 2014:1294 - 1297.

[15] Yu, Shui. Big Privacy: Challenges and opportunities of

privacy study in the age of big data[J]. IEEE Access, 2016(4): 2751-2763.

[16] 余佳,刘逸帆,葛云.加强个人信息保护促进社会和谐进步-访中国电子商务协会政策法律委员会副主任阿拉木斯[J].社会治理,2017(5):37-41.

[17] Sivarajah U, Kamal MM, Irani Z, et al. Critical analysis of big data challenges and analytical methods[J]. Journal of Business Research,2017(70): 263-286.

Research on Software Testing Technology of Big Data

LIU Shanyan¹, HU Xiu¹, YAN Wu²

(1 *Department of Computer Engin., Jingchu Univ. Sci. and Engin., Jingmen 448000, China* ;
2 *Jingmen Yousi Information Technology co. LTD, Jingmen 448000, China* ;)

Abstract: With the development of information technology, big data has become a new stage in the information age. For software testing, what should be tested, how to test and how to measure product quality for big data system are all problems that need to be solved urgently. Firstly, the paper analyzed several challenges of big data software testing, including testing basic theory, testing process, testing thinking. After introducing the main technology of big data testing, the paper analyzed the process of big data testing. Since ETL (Extract Transform and Load) testing is an important and complex stage in data warehouse testing, the main contents of ETL testing and the advantages of automated ETL testing were discussed. Through data filling test and test result analysis on employee table, it is shown that ETL test can improve the validity of data quality. The research shows that reasonable test environment construction and application of automated test technology can improve test efficiency and reduce the difficulty of big data test.

Keywords: big data; software testing; big data testing; ETL testing; test data

[责任编辑: 张岩芳]

(上接第 16 页)

Research on Key Technologies of Panoramic View Parking System

QIAN Feng¹, LI Yong¹, WEN Shurong²

(1 *Hubei Key Lab of Manufacture Quality Engin., School of Mechanical Engin., Hubei Univ.of Tech., Wuhan 430068, China* ;

2 *Hubei Product Quality Supervision & Inspection Research Institute, Ezhou 436070, China*)

Abstract: A new panoramic parking aid system is designed to solve the problem of collision when the passenger car stops and starts because of the lack of space and the narrow vision of the driver. Through the simulation system of panoramic parking aid system, the technology of fish eye camera calibration, fish eye image distortion correction and image stitching in the panoramic parking aid system were studied. The simulation system imitated the panoramic parking assistant system and installed four fish eye cameras in the front, back, left and right of the simulation platform to collect and splice the images in four different directions around the platform into a panorama in real time. The experimental results show that the panoramic image of the simulation system had no splicing gap, and the splicing time was reduced by more than 35% and strong robustness.

Keywords: panoramic view parking system,; camera calibration; image splicing; top view picture

[责任编辑: 张 众]