Journal of Computer Applications
计算机应用,2015,35(4):1137 -1142
ISSN 1001-9081
CODEN JYIIDU

文章编号:1001-9081(2015)04-1137-06
doi:10.11772/j.issn.1001

大数据评测基准的研发现状与趋势

周晓云[1],覃雄派[2],王秋月[2*]
(1.江苏师范大学 计算机科学与技术学院,江苏 徐州 221116;2.中国人民大学 信息学院,北京 100872)
(*通信作者电子邮箱 qiuyuew@ruc.edu.cn)

摘　要: 工业界、学术界,以及最终用户都急切需要一个大数据的评测基准,用以评估现有的大数据系统,改进现有技术以及开发新的技术。回顾了近几年来大数据评测基准研发方面的主要工作。对它们的特点和缺点进行了比较分析。在此基础上,对研发新的大数据评测基准提出了一系列考虑因素:1)为了对整个大数据平台的不同子工具进行评测,以及把大数据平台作为一个整体进行评测,需要研发面向组件的评测基准和面向大数据平台整体的评测基准,后者是前者的有机组合;2)工作负载除了SQL查询之外,必须包含大数据分析任务所需要的各种复杂分析功能,涵盖各类应用需求;3)在评测指标方面,除了性能指标(响应时间和吞吐量)之外,还需要考虑其他指标的评测,包括系统的可扩展性、容错性、节能性和安全性等。

关键词: 大数据;评测基准;性能;可扩展性;容错性;节能性;安全性
中图分类号: TP311.13　　**文献标志码:** A

Big data benchmarks: state-of-art and trends

ZHOU Xiaoyun[1], QIN Xiongpai[2], WANG Qiuyue[2*]
(1. *School of Computer Science and Technology, Jiangsu Normal University, Xuzhou Jiangsu* 221116, *China*;
2. *School of Information, Renmin University of China, Beijing* 100872, *China*)

Abstract: A big data benchmark is needed eagerly by customers, industry and academia, to evaluate big data systems, improve current techniques and develop new techniques. A number of prominent works in last several years were reviewed. Their characteristics were introduced and the shortcomings were analyzed. Based on that, some suggestions on building a new big data benchmark are provided, including: 1) component based benchmarks as well as end-to-end benchmarks should be used in combination to test different tools inside the system and test the system as a whole, while component benchmarks are ingredients of the whole big data benchmark suite; 2) workloads should be enriched with complex analytics to encompass different application requirements, besides SQL queries; 3) other than performance metrics (response time and throughput), some other metrics should also be considered, including scalability, fault tolerance, energy saving and security.

Key words: big data; benchmark; performance; scalability; fault tolerance; energy saving; security

0　引言

随着存储器价格的下降,电子商务和互联网应用(如社交网络等)的迅速发展,以及科学研究的需要等,人们收集了一些大规模的数据集,其数据量可以达到上百拍字节(Peta Byte,PB),甚至更多。这些数据具有不同的形式(包括结构化和非结构化数据),产生的速度也不同(有些数据是由传感器自动产生的,必须尽快地进行处理)。

经过近40年的发展,关系数据库技术已经成为数据管理和数据分析的标准选择,关系数据库管理系统(Relational Database Management System,RDBMS)成为一项成熟的技术,被广泛应用在在线事务处理(Online Transaction Processing,OLTP)以及在线分析处理(Online Analytic Processing,OLAP)中,可以有效地处理结构化数据。虽然大多数RDBMS系统通过扩展插件能够处理非结构化数据,但是因为其内在的扩展性、容错性等方面的局限,RDBMS不能够完全地解决大数据管理和分析的挑战。近几年来,人们提出并且实现了一系列NoSQL数据库,对大数据进行有效的管理和分析处理。

面对众多的大数据管理和分析系统,用户需要通过评测,选择符合其需求的产品;而厂商和学术界的研究人员,同样需要一个公认的评测基准。厂商通过评测,找出产品的性能瓶颈,进而可以持续地改进其产品;研究人员则可以利用评测基准分析评测现有技术,并不断改进,从而加速创新的步伐。为了对大数据工具与平台进行评价和选择,用户、厂商和研究人员都需要一个大数据评测基准[1-3]。

工业界和学术界已经为关系数据库产品的评测研发了一系列的评测基准。其中最重要的是被工业界和学术界广泛接受的由事务处理性能委员会(Transaction Processing Performance Council,TPC)为决策支持系统中的数据库管理系统研发的评测基准TPC-C和TPC-H,它们分别针对OLTP和OLAP应用场景。以TPC-H为例,它包含8张表(数据集)和22个业务查询(负载),建模对象是货物的分发与销售,其

收稿日期:2014-10-28;修回日期:2014-12-21。
基金项目:国家自然科学基金资助项目(61170013,61202331);江苏省自然科学基金资助项目(BK2012578)。
作者简介:周晓云(1971-),女,辽宁岫岩人,副教授,博士,主要研究方向:高性能数据库、大数据分析;　覃雄派(1971-),男,广西百色人,讲师,博士,CCF会员,主要研究方向:高性能数据库、大数据分析、信息检索;　王秋月(1974-),女,山西定襄人,讲师,博士,CCF会员,主要研究方向:数据库、信息系统、信息检索、知识库、自然语言问答。

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

性能指标包括 QphH@ Size(Query-per-hour at Different Data Sizes)和 Price/QphH。QphH@ Size 指标指测量目标系统在不同的数据规模下每小时完成的查询数量,对来自不同厂家的数据库系统,可以在相同的数据规模下进行性能评测和比较;Price/QphH 指标则用来度量系统的性价比。除了 TPC-C 和 TPC-H 评测基准外,还有几个评测基准得到了广泛应用,包括 TPC-W、TPC-DS 和 SSB(Star Schema Benchmark)等。TPC-W 是面向 OLTP 类 Web 应用的评测基准,而 TPC-DS 和 SSB 则是面向 OLAP 类应用以及决策支持系统等分析型应用的评测基准。

TPC 委员会的一系列评测基准,在研发出来以后的几十年间,没有大的改动,有利于人们的理解和使用,进行不同系统的纵向比较(不同时代的硬件软件系统的性能比较)和横向比较(同一时代的不同厂家的系统之间的比较)。但是,TPC 系列评测基准对于评测近几年出现的各类大数据系统是不够用的。因为大数据管理和分析系统(包括各种 NoSQL 数据库)是近几年才研发出来的,在面向的应用、采用的数据模型、一致性约束等方面,和传统的数据库不同。大数据本身以及大数据之上的应用,随着时间不断演化,需要新的大数据评测基准来对不同系统进行客观深入的评测。

下面将对近几年来研究人员提出和研发的各种大数据评测基准进行回顾,并且对它们进行比较和分析。在此基础上,本文对研发新的大数据评测基准提出了一系列考虑因素,并介绍我们对主要大数据处理系统的容错性的评测工作。

1 大数据评测基准的回顾与比较

1.1 大数据评测基准的回顾

1.1.1 MRBench

MapReduce(Hadoop 是对 MapReduce 技术的开源实现)已经成为大数据处理的标准工具[4]。Kim 等设计了 MRBench 评测系统[5],使用 TPC-H 负载对 MapReduce 平台进行评测。他们首先在 MapReduce 平台上实现了一些基本的操作,包括选择(Selection)、投影(Projection)、叉乘(Cross Product)、连接(Join)、排序(Sort)、分组(Grouping)、聚集(Aggregation)等,然后在上面实现 TPC-H 的各个查询功能。他们在不同的数据规模上,通过改变节点数量和 Map 任务(Map Tasks)数量等参数,进行了一系列的实验,并给出了实验结果。

对于利用 MapReduce 平台运行数据仓库查询的应用,MRBench 可以给人们一些评测的结果。但是,最近不同厂家在 SQL on Hadoop 系统(包括 EMC 的 HAWQ,HortonWorks 的 Stinger,Cloudera 的 Impala,Apache 的 Drill 等)[6]研发中,已经不再采用把 SQL 查询转换成 MapReduce 作业(Job)的技术路线;而且一些大数据应用,已经超越 SQL 查询,向更复杂的深度分析发展,所以 MRBench 评测是具有局限性的。

1.1.2 天文物理仿真数据分析的评测

文献[7]提出了包含5个代表性查询(构成一个应用场景),面向大规模天文物理数据分析的评测方法。他们在分布式 RDBMS 系统和 Pig/Hadoop 系统上实现了该应用场景,并且进行了性能评测,对两者的评测结果进行了比较。分布式 RDBMS 系统和 Pig/Hadoop 系统都获得了与他们原有的接口定义语言(Interface Definition Language,IDL)方法媲美的性能,同时具有更高的扩展能力。然而,这5个查询不能构成一个完整的评测基准,因为这个应用场景仅仅考虑了天文物理数据分析应用的需求,不能覆盖其他应用领域工作负载的特点。

1.1.3 PUMA

PUMA[8]是针对 MapReduce 平台设计的一个评测基准。该评测基准的负载包含了大量的数据分析任务,包括:词项-向量操作(Term-vector)、倒排索引(Inverted-index)、自连接(Self-join)、邻接列表(Adjacency-list)、聚类(*K*-means)、分类(Classification)、直方图操作(Histogram-movies,Histogram-ratings)、序列计数(Sequence-count)、排序的倒排索引(Ranked Inverted-index)、TB(Terabyte)级数据的排序(Tera-sort)、字符串匹配(GREP),以及单词计数(Word-count),等等。

针对 MapReduce 平台研发的大数据评测基准有一个共同的局限性,即它们的负载并未涵盖大数据应用的所有需求,仅仅评测了数据管理和分析系统的一些行为。实际应用系统一般不会同时只运行一个作业(Job),或者只运行一些特定的 Job。文献[9-10]提出的评测基准也是面向 MapReduce 平台进行研发的,同样具有这些局限性。

1.1.4 加州大学伯克利分校的工作

加州大学伯克利分校(UC Berkeley)的一些研究人员深入讨论了要研发一个大数据评测基准[11]所需要考虑的因素。他们认为,必须基于一个统一的模型来研发大数据评测基准。这个模型描述了目标应用工作负载的各个方面,包括数据模型,主要的数据存取模式,负载随着时间变化的特点,系统应该具备的分析功能,为了达到分析功能的一些基本数据操作方法(包括 SQL 的基本操作,如选择、投影、连接、排序、分组、聚集等)和复杂的 SQL 查询(比如相关子查询),以及各种复杂分析算法。他们列出了大数据评测基准研发的一些准则:

1)代表性(Representative):评测基准应该面向实际应用进行建模,使用实际应用领域关注的评测指标,并且整个评测过程必须和实际的计算场景一致。

2)可移植性(Portable):评测基准应该易于移植到不同的数据管理和分析系统上(包括各类 NoSQL 系统和 RDBMS 系统)。

3)可扩展性(Scalable):评测基准能够对不同大小规模的系统进行评测,包括不同规模的集群系统和不同规模的数据集。

4)简单性(Simple):评测基准必须足够简单,易于被用户、厂商和研发人员所理解。

5)系统多样性(System Diversity):大数据系统需要对不同类型的数据进行管理和分析。有时多样性意味着在设计评测系统时,在一些关键目标上是互斥的。

6)快速的数据演化能力(Rapid Data Evolution):评测系统能够在不同层面上进行扩展,在数据类型、数据格式等方面,数据集能够代表典型的大数据应用。

把大数据应用的典型功能需求分析出来,仅仅是完成研发一个全面的大数据评测基准的第一步。需要注意的是,大数据涵盖了很多的应用领域。OLTP 是传统数据库系统擅长的领域,除此之外,还需要包括如下典型的应用:1)高延迟批量数据处理。这是 MapReduce 技术所擅长的,MapReduce 技术原先即为这样的应用场景而设计。2)低延迟交互式的分析应用。OLAP 以及类似的决策支持应用,即归于此类。3)高速流式数据的分析。对流数据的持续、不间断的监控和分析。

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

1.1.5　YCSB

YCSB(Yahoo! Cloud Serving Benchmark) [12] 是 Yahoo 公司提出的一个评测基准,用于评测基于云平台的数据服务系统(Cloud Data Serving Systems) 的性能。所谓基于云平台的数据服务系统，包括 BigTable、PNUTS、Cassandra、HBase、Azure、CouchDB、SimpleDB、Voldemort,以及其他各类 NoSQL 数据管理系统。由于 YCSB 是面向云平台的评测基准，该评测基准考虑了云平台的横向扩展特性、系统弹性(Elasticity) 和高可用性(High Availability) 等特性。设计者还考虑了评测过程中的一些折中(Trade off) 方案，包括读/写性能的折中(Read vs. Write Performance) 、查询延迟和持久性的折中(Query Latency vs. Durability) 、同步复制和异步复制的折中(Synchronous vs. Asynchronous Replication) ,以及不同数据分片方式 (Data Partitioning) 的折中等。这是和传统的 TPC-C 评测基准具有明显区别的地方。传统的 OLTP 应用必须符合 ACID (原子性 Atomicity、一致性 Consistency、隔离性 Isolation、持久性 Durability) 事务特性的约束，事务提交以后，数据必须持久化。而 YCSB 则对“基于云平台的类 OLTP 系统(Cloud Based OLTP Systems) ”进行评测，一般不要求支持严格的 ACID 事务，所以能够对这些因素进行折中。

YCSB 从三个层面对目标系统进行评测，包括性能层面(Performance Tier) 、扩展性层面 (Scaling Tier) 和可用性层面(Availability Tier) 。其设计者在四个广泛使用的数据库系统上运行了这个评测基准，包括 Cassandra、HBase、Yahoo 的 PNUTS,以及一个分割的 (Sharded) MySQL 数据库，具体结果可以参见文献 [12] 。

1.1.6　非结构化数据分析系统的评测

Smullen 等 [13] 认为，在未来的大数据系统中，系统存储和处理的数据，大部分是非结构化的数据，包括文本、图像、音频、视频等。他们提出了一个非结构化数据处理的评测软件包。在负载方面，他们抽象出四个主要的操作，包括边缘检测(Edge Detection) 、邻近搜索(Proximity Search) 、数据扫描(Data Scanning) 和数据融合(Data Fusion) 等，并且对每个操作进行了详细的定义。这些操作并未囊括非结构化数据处理应用所有的负载特点。即便这个评测软件包仅仅针对非结构化数据的处理，负载的代表性(Representative) 有限，但是这个工作给我们一个启示: 非结构化数据的管理和分析是大数据评测基准需要考虑的关键问题之一。

1.1.7　HiBench

HiBench [14] 也是面向 MapReduce 平台的评测基准。HiBench 的设计者对先前的若干评测基准进行了全面的分析。

首先，面向排序的评测基准，比如 TeraSort 等，已经被众多的厂商包括 Google 和 Yahoo [15] 用来测试和优化 MapReduce 技术。虽然对数据进行排序是非常重要的数据处理任务，但是排序仅仅是数据处理的一个方面。其次，GridMix [16] 是包含在 Hadoop 软件包中的一个人工构造的评测软件。GridMix 通过混合若干不同类型的 Hadoop 作业(Job) ，来对数据存取的模式进行建模。这些 Job 包括 3 阶段链式 MapReduce Job、大数据排序、选择、文本排序等。GridMix 不能看作是面向大数据平台的完整的评测基准，因为其负载并未涵盖实际应用的完整特征。此外，Hive 性能评测基准(Hive performance benchmark) [17] 用于对 Hadoop 系统和并行数据库系统的性能进行比较,它包含 5 个重要的分析任务，包括 GREP 任务，还有文献 [18] 里面使用的面向结构化数据处理的另外四个分析任务，即选择、聚集、连接和基于用户自定义函数(User Defined Function，UDF) 的聚集。最后，DFSIO [19] 是文件系统一级的评测基准，用于对 HDFS(Hadoop Distributed File System) 性能进行评测，其关注点是在大量并发任务同时进行读/写操作时的文件系统性能。

基于对上述工作的分析，HiBench 的设计者在评测基准中包含了几个主要的组件，使之成为一个超越上述工作的新的评测基准。主要的组件包括: 1) 微观层面的评测基准(Micro Benchmarks) ，包含了排序、单词计数和大数据排序(TeraSort) ; 2) Web 搜索，包含 Nutch 索引技术和 Page Rank 计算; 3) 机器学习算法，包含 Bayesian 分类和 *K*-means 聚类; 4) HDFS 评测基准，这是一个文件系统层面的子评测基准。这些组件使得 HiBench 成为一个非常全面的评测基准。

HiBench 评测基准的突出特点是，它把机器学习等复杂分析功能集成到了整个评测基准的框架中，和大数据的复杂分析大趋势相吻合。这也是其他综合的大数据评测基准，比如 CloudRank-D 和 Big Bench 等所具备的特点。

1.1.8　CloudRank-D

中国科学院计算技术研究所的詹剑锋团队基于他们的并行计算背景，提出了面向云平台上大数据处理的评测基准 CloudRank-D [20] 。在工作负载方面，这个评测基准不仅包含了数据分析的基本操作(主要是简单 SQL 查询和数据仓库查询) ，同时加入了关键的数据挖掘和机器学习算法，比如分类 (Classification) 、聚类 (Clustering) 、推荐算法 (Recommendation) 、序列模式学习(Sequence Learning) 、关联规则分析(Association Rule Mining) 等。这个评测基准还集成了一个代表性的应用——ProfSearch (http://prof. ict. ac. cn/) 。在评测指标方面，针对基于云平台的大数据处理系统，他们提出了两个关键指标，分别是每秒钟处理的数据量(Data Processed per Second) 和每焦耳处理的数据量(Data Processed per Joule) 。前者定义为所有 Job 的输入数据量除以这些 Job 处理的总时间，这个时间是从第一个 Job 提交到最后一个 Job 完成的时间。另外,把节能概念引入评测基准是这个工作的一个鲜明特点。

1.1.9　Big Bench

近两年来，来自学术界和工业界的具有并行计算背景的一些专家聚集在一起，召开了一系列大数据评测讨论会(Workshop on Big Data Benchmarking) [21] ，讨论如何构建一个全面的大数据评测基准。

这些专家讨论了大数据评测基准的各方面细节，包括代表性的负载、数据集的构成、新的指标、实现方面的一些选择、工具集与数据生成工具、评测基准的运行规范、扩展(Scale) 因子的定义等。

文献 [22] 是他们工作的结晶。该文献提出了一个端到端的(即以应用场景驱动的评测，关注业务需求的满足，而非基本的数据操作和分析算法) 大数据评测基准建议稿 Big Bench。该建议稿涵盖了结构化、非结构化和半结构化等数据模型，并且针对大数据的多样性(Variety) 、高速度(Velocity) 和大数据量(Volume) 等特点，给出了评测策略。

在结构化数据处理的评测方面，Big Bench 从 TPC-DS 评测基准借鉴了主要的数据模型，然后在这个模型之上，加上了非结构化和半结构化数据模块。其中半结构化数据处理和评测模块对注册用户和普通用户(Guest) 在零售商网站上的点击流进行了建模; 而非结构化数据处理和评测模块则对用

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

户在线提交的产品评论进行建模。

在工作负载方面，设计者围绕上述数据模型设计了一系列查询，涵盖了不同类型的数据分析功能，评测了目标系统的关键维度包括数据类型、查询处理和分析算法等。该文献最后报告了他们在 Aster Data 数据库上的评测基准的实现以及实验结果。在评测基准的实现中，他们使用了 Aster Data 的 SQL-MR 技术。这个评测基准很容易被移植到其他大数据管理和分析平台。

1.1.10 GRAPH 500 和 Link Bench

GRAPH 500[23] 和 Link Bench[24] 是专门针对图数据库进行评测的评测基准。在 GRAPH500 评测基准中，仅仅评测了两个核心功能，其中一个核心功能是从输入数据建立一张图；另外一个功能则评测图上的广度优先搜索(Breadth-first Search)能力。由此可以看出，GRAPH500 不能算是一个全面的图数据库评测基准。2013 年，Facebook 发布了 Link Bench 作为社交网络图数据库的评测基准。从某种意义上看，Link Bench 更多的是作为一个图数据服务(Graph-serving)的评测基准，而不是一个图数据处理和分析的评测基准。两者的区别在于，前者对交互式社交网络应用中的事务型负载(简单的增加、删除、修改和查询)进行建模，而后者则主要关注分析型负载，比如路径计算、关键节点发现、社区发现等。

本文认为把图数据的处理和分析纳入整个大数据评测基准是一个重要的趋势，图数据的处理和分析应该作为评测基准框架的一个评测模块。

1.2 大数据评测基准的比较

在对近几年大数据评测基准工作的介绍和分析的基础上，本文把所有的工作放在一个表格里(见表1)，并且列出了其主要特点和主要应用场合，方便读者进行对比和把握。

表1 大数据评测基准已有工作的比较

基准名称	评测对象	评测内容	备注
TPC-C*	RDBMS	面向事务处理 负载包含简单查询和数据更新	OLTP 应用
TPC-H*	RDBMS Hadoop Hive	面向报表和决策支持系统	OLAP 应用
TPC-W*	RDBMS NoSQL	面向 Web 应用	Web OLTP 应用
SSB*	RDBMS Hadoop Hive	面向报表和决策支持系统	OLAP 应用
TPC-DS*	RDBMS Hadoop Hive	面向报表和决策支持系统 包含四类查询：报表类查询、即席查询、迭代式查询和数据挖掘查询	OLAP 应用
TeraSort	RDBMS Hadoop	大规模数据的排序	只考虑数据排序
YCSB	NoSQL database	基于云平台的数据服务系统，主要针对 NoSQL 数据库	Web OLTP 应用
非结构化数据分析的评测基准[14]	unstructured data management system	面向非结构化(主要是文本)数据分析 包含四类分析：边缘检测、临近搜索、数据扫描、数据融合	工作负载不能全面代表各类应用需求
GRAPH 500	graph database	仅仅对图数据处理进行评测	工作负载不能全面代表各类应用需求
Link Bench	RDBMS graph database	对 Facebook 的实际应用进行建模 仅仅对图数据处理进行评测	工作负载不能全面代表各类应用需求
DFSIO	Hadoop	文件系统级别的评测基准	工作负载不能全面代表各类应用需求
Hive performance benchmark	Hadoop Hive	负载包含 GREP、选择、聚集、连接和基于 UDF 的聚集	工作负载不能全面代表各类应用需求
GridMix	Hadoop	负载是一系列 Hadoop Job 的混合	工作负载不能全面代表各类应用需求
MRBench	MapReduce	TPC-H 查询	OLAP 应用
PUMA	MapReduce	工作负载相当全面，包括：词项-向量操作、倒排索引、自连接、邻接列表、聚类(*K*-means)、分类、直方图操作、序列计数、排序的倒排索引、TB 级数据排序、字符串匹配(GREP)，以及单词计数等	全面的、代表各类应用需求的工作负载
HiBench	MapReduce	1) 微观层面的评测基准，包含了排序、单词计数和 TB 级数据排序；2) Web 搜索，包含 Nutch 索引技术和 page rank 计算；3) 机器学习算法，包含 Bayesian 分类和 *K*-means 聚类；4) HDFS 评测基准，文件系统层面的子评测基准	全面的、代表各类应用需求的工作负载
CloudRank-D	RDBMS Hadoop	不仅包含了数据分析的基本操作(主要是简单 SQL 查询和数据仓库查询)，同时加入了关键的数据挖掘和机器学习算法(比如分类、聚类、推荐算法，以及序列模式学习、关联规则分析等) 对能耗状况进行评测	全面的、代表各类应用需求的工作负载
Big Bench	RDBMS Hadoop	1) 涵盖结构化数据、非结构化数据、半结构化数据的处理；2) 设计了一系列针对这些数据的查询分析任务；3) 给出了针对大数据系统的数据多样性、高速度和大数据量特点的评测方案	全面的、代表各类应用需求的工作负载

注：* 这些测试基准不是针对大数据处理的评测基准，罗列在此用于比较不同测试基准的特点和所面向的应用的不同。

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

2　测试基准改进思考

2.1　代表性的应用——组件式评测与端到端评测

使一个大数据评测基准成为一个有用的工具，不仅能够帮助用户评估各类目标系统，而且能够帮助厂商和研究人员改进现有技术，其数据模型和工作负载必须来自真实的应用需求。

对实际应用进行建模的一种方法是，以 Facebook 和 Google 等大型互联网公司的数据管理系统和上层应用为蓝本进行建模。但是，由于商业秘密保护要求，不能确定这些大型互联网公司有多大意愿共享他们的数据集和工作负载; 此外，单一模型可能无法代表千差万别的实际应用。

本文认为，评测基准的建模可以采用自底向上和自顶向下相结合的策略。自底向上策略想办法抽象出数据处理的基本操作和分析算法; 而自顶向下则考虑各种实际应用的特点，把它们的特点纳入整个评测基准的框架里。

人们相信，一个工具不能解决所有问题。在一个大数据处理平台里，应该包含各类数据处理工具模块。人们需要从不同类型的数据中，攫取有用的信息用于决策。这些数据包括各个阶段的数据，也包括各种类型的数据，数据本质上是多结构的(Multi-structured) 。一般来讲，大数据平台包含 3 个不同的组件(不同的应用场合，这些组件可以随意组合) : 1) 流数据处理引擎，负责流数据的分析; 2) 结构化数据处理引擎，可以建立在 RDBMS 之上，一般是一个面向 OLAP 处理的列存储数据库(OLTP 类应用由 RDBMS 完成，不纳入大数据平台，也不由大数据评测基准进行评测，而是由诸如 TPC-C、TPC-W 等基准进行评测) ; 3) 非结构化数据处理引擎，一般建立在 NoSQL 数据库或者 Hadoop 之上。特别需要指出的是，在一些特定应用比如社交网络分析应用中，图数据分析是必不可少的一个组件。

面向这样一个大数据平台，用户需要组件式的评测，也需要端到端的评测。用户可以从评测基准框架中，选择其中一个子评测基准，对某个组件进行评测，比如单独评测图数据库，或单独评测流数据处理引擎; 也可以通过一个应用场景，把若干子评测基准组合起来，以工作流的方式，完成若干组件的整体评测，测试它们满足业务目标的能力。

文献 [25] 认为，大数据分析一般是由一系列分析任务构成的工作流，而不是单一的分析任务。大数据分析是一个过程，跨越大数据平台里的多个数据处理工具。比如，在一个客户关系分析系统里，应用程序可以从社交媒体数据(主要是文本，一般由 Hadoop 进行保存和处理) 中提取社交网络信息和情感信息。这些情感打分信息可以和数据仓库(一般由 RDBMS 进行存储和处理) 中的客户信息进行关联。基于上述分析结果，可以进行后续的数据挖掘，比如找出由于没有得到满意的服务而选择要离开的客户，以便留住他们。从社交媒体中提取的社交网络关系信息，可以注入到图数据库中，进行后续分析，找出一些重要的网络关系，以便进行交叉销售(Cross Selling) 。这是一个生动的端到端的评测实例。

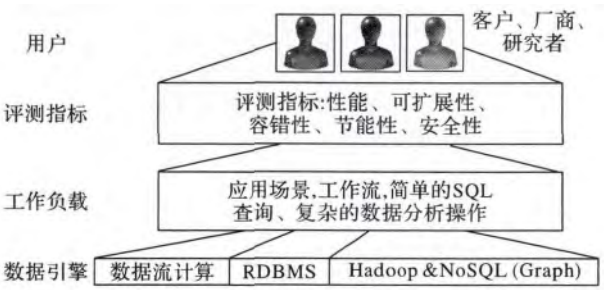
2.2　工作负载——超越简单 SQL 查询，集成更复杂的分析

简而言之，评测基准可以看作是规定了数据集合、工作负载和评测指标的一个软件规范。

通过对现有大数据评测基准的分析，本文认为工作负载应该包含简单分析和复杂分析，即除了简单 SQL 查询和汇总之外，关键的数据挖掘、机器学习、信息检索算法也应该纳入到负载里面来。

近年来，随着社交网络和各种信息网络(如语义网、知识库等) 的发展，图分析算法变得越来越重要，这些算法应该作为整个负载不可或缺的成分。在某些特定的应用场合中，系统需要对多媒体数据以及科学数据进行处理和分析(主要是矩阵和向量运算) ，也需要给予建模。

工作负载可以用两层关系进行组织，如图 1 所示。首先，在简单 SQL 查询和基本分析算法之上，创建工作流; 其次，基于典型工作流，针对若干应用场合，把工作流组合起来，对应用场景进行建模。



注: 图数据管理和分析是大数据平台不可或缺的组成部分。

图 1　大数据评测基准的模块

当用户对目标系统进行评测时，他们可以选择部分或者全部应用场景(即工作流的组合) 来测试目标系统，以了解目标系统是否满足他们的业务目标。

2.3　性能指标与其他关键指标

传统的数据库系统评测基准在指标的设计上主要关注目标系统的性能指标，包括响应时间和吞吐量。但对于大数据评测基准来讲，这些指标是不够的。设计者应该考虑更多的指标，包括系统的可扩展性(Scalability) 、容错性(Fault Tolerance) 、节能性(Energy Saving) 和安全性(Security Guarantee) 等。

为了对大数据进行有效的处理，大数据处理系统一般建立在大型的集群之上，一部分集群系统则部署到云平台的虚拟节点上。在这种情况下，必须测试大数据系统的扩展能力。同时，因为云平台具有弹性(Elastic) 计算的特点，即云平台不仅可以动态扩展规模(Scale Out) ，而且可以动态缩小规模(Scale In) 。这种动态的可扩展性，要求用户能够在底层集群环境动态变化时(可能是由于性能原因或者成本原因，比如当需要更高性能时，增加节点规模; 当需要降低成本时，缩减节点规模) ，对目标大数据系统进行有效的评测。

在大型集群系统里，节点的故障是司空见惯的事情，容错保证是大型集群需要提供的一个重要功能。在大数据系统评测中，应该允许用户有计划地往目标系统注入一些故障，以便观察目标系统是否能够有效处理故障，达到一定的服务水平。

此外，大量的节点并行处理大数据，消耗大量的能源，节能成为一个严峻的问题。用户、厂商和研究人员都非常重视这个问题，能量消耗应该作为目标系统评测的一个重要指标。

最后，大数据评测基准的设计者应该把安全测试纳入整个评测基准的框架之中。

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

3 我们的测试工作

本文基于“人大行云”云计算平台开展了 SQL on Hadoop 系统的容错性实验。本文的实现策略是,在查询开始以后,以随机分布的时间点,在查询执行过程中,通过 Open Stack 虚拟机管理系统发送虚拟节点关闭指令,模拟节点发生当机的状况。本文比较了 Hive 系统和 Cloudera 公司的 Impala 系统。经过初步实验发现:当查询的数据集比较大,但是中间结果比较小,并且查询可以在一个 MapReduce Job 里面完成时,当机对 Hive 系统上的查询响应时间的影响(平均延长 10% 到 35% 不等),比对 Impala 系统上的查询响应时间的影响要小(平均延长 50% 到 100% 不等)。经过分析,其原因是 Hive 在执行查询时,在每个 Map/Reduce 任务结束后,都进行结果集的存盘操作,以达到容错的目的,即有节点当机,查询可以从已有的中间结果开始;而 Impala 则使用流水线(Pipeline)方式,在查询的各个子操作间进行中间结果集的传递,减少了存盘开销,提高了查询速度,但是一旦有节点当机,那么查询必须从头开始。对于复杂的多表连接查询,虽然 Hive 在当机情况下增加不多的查询时间,但是其查询总时间仍然比 Impala 重新执行整个查询时间要长。Hive 和 Impala 处在查询容错的两个极端,Hive 每一步都进行中间结果的持久化,出现当机无需重启整个查询;而 Impala 则通过流水线直接进行各个子操作的数据交换,出现当机需要完全重启整个查询。如何根据集群节点当机概率、中间结果集大小等因素,在流水线数据交换和中间结果持久化之间作出折中,在查询的子操作或者操作步中插入中间结果集持久化操作,保证已有的查询工作不至于白费,同时保证查询的响应时间,是一个值得研究的问题。

4 结语

本文对大数据评测基准近几年的工作作了一个回顾,大数据评测基准和传统的面向事务处理的评测基准,以及和面向分析处理的评测基准在若干方面存在差异,需要关注数据规模、数据多样性、数据处理速度等重要因素。

本文提出了针对大数据评测基准的若干设计原则:组件式的评测和应用导向的端到端的评测同等重要;除了 SQL 查询之外,典型的数据挖掘和机器学习算法应该纳入工作负载,以测试目标系统深度分析的能力和性能;除了性能指标(包括响应时间和吞吐量等)之外,其他指标比如能量消耗和安全级别等指标,需要定义和予以测量。

参考文献:

[1] VENTANA RESEARCH. Hadoop and information management: Benchmarking the challenge of enormous volumes of data [EB/OL]. [2014-10-10]. http://www.ventanaresearch.com/research/benchmarkDetail.aspx?id=1663.

[2] BIG DATA TOP 100. An open, community-based effort for benchmarking big data systems [EB/OL]. [2014-10-15]. http://bigdatatop100.org/benchmarks.

[3] HEMSOTH N. A new benchmark for big data [EB/OL]. [2014-11-01]. http://www.datanami.com/datanami/2013-03-06/a_new_benchmark_for_big_data.html.

[4] SAKR S, LIU A, FAYOUMI A. The family of MapReduce and large scale data processing systems [EB/OL]. [2014-03-01]. http://arxiv.org/abs/1302.2966.

[5] KIM K, JEON K, HAN H, *et al*. MRBench: a benchmark for MapReduce framework [C]// Proceedings of the 14th IEEE International Conference on Parallel and Distributed Systems. Piscataway: IEEE Press, 2008: 11-18.

[6] HARRIS D. SQL is what's next for Hadoop: here's who's doing it [EB/OL]. [2014-09-15]. http://gigaom.com/2013/02/21/sql-is-whats-next-for-hadoop-heres-whos-doing-it/.

[7] LEOBMAN S, NUNLEY D, KWON Y C, *et al*. Analyzing massive astrophysical datasets: can pig/Hadoop or a relational DBMS help? [EB/OL]. [2014-03-10]. http://nuage.cs.washington.edu/pubs/iasds09.pdf.

[8] AHMAD F, LEE S, THOTTETHODI M, *et al*. PUMA: purdue MapReduce benchmarks suite [R]. West Lafayette: Purdue University, 2012.

[9] MOUSSA R. TPC-H benchmark analytics scenarios and performances on Hadoop data clouds [C]// Proceedings of the 4th International Conference Networked Digital Technologies, CCIS 293. Berlin: Springer-Verlag, 2012: 220-234.

[10] CHEN Y P, ALSPAUGH S, GANAPATHI A, *et al*. SWIM statistical workload injector for MapReduce [EB/OL]. [2013-04-30]. https://github.com/SWIMProjectUCB/SWIM/wiki.

[11] CHEN Y P, RAAB F, KATZ R H. From TPC-C to big data benchmarks: a functional workload model [R]. Berkeley: University of California, 2012.

[12] COOPER B F, SILBERSTEIN A, TAM E, *et al*. Benchmarking cloud serving systems with YCSB [C]// Proceedings of the 1st ACM Symposium on Cloud Computing. New York: ACM Press, 2010: 143-154.

[13] SMULLEN C W, TARAPORE S R. GURUMURTHI S. A benchmark suite for unstructured data processing [C]// Proceedings of the 2007 International Workshop on Storage Network Architecture and Parallel I/Os. Piscataway: IEEE Press, 2007: 79-83.

[14] HUANG S S, HUANG J, DAI J Q, *et al*. The HiBench benchmark suite: Characterization of the MapReduce-based data analysis [C]// Proceedings of the 26th IEEE International Conference on Data Engineering Workshops. Piscataway: IEEE Press, 2010: 41-51.

[15] MALLEY O O, MURTHY A C. Winning a 60 second dash with a yellow elephant [EB/OL]. [2009-11-01]. http://sortbenchmark.org/Yahoo2009.pdf.

[16] GRIDMIX. GridMix benchmark [EB/OL]. [2012-12-15]. http://hadoop.apache.org/docs/r1.1.1/gridmix.html.

[17] JIA Y, SHAO Z. A benchmark for Hive, PIG and Hadoop [EB/OL]. [2012-06-15]. http://issues.apache.org/jira/browse/HIVE-396.

[18] PAVLO A, PAULSON E, RASIN A, *et al*. A comparison of approaches to large-scale data analysis [C]// SIGMOD 2009: Proceedings of the 2009 ACM SIGMOD International Conference on Management of data. New York: ACM Press, 2009: 165-178.

[19] DFSIO PROGRAM. DFSIO of Hadoop source distribution [EB/OL]. [2012-12-10]. file://src/test/org/apache/hadoop/fs/TestDFSIO.

[20] LUO C, ZHAN J, JIA Z, *et al*. CloudRank-D: benchmarking and ranking cloud computing systems for data processing applications [J]. Frontier of Computer Science, 2012, 6(4): 347-362.

[21] GHAZAL A, RABL T, HU M, *et al*. BigBench: towards an industry standard benchmark for big data analytics [C]// SIGMOD 2013: Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data. New York: ACM Press, 2013: 1197-1208.

[22] GRAPH 500. Graph 500 benchmark [EB/OL]. [2013-01-01]. http://www.graph500.org/specifications.

[23] KING R. Facebook releasing new social graph database benchmark: LinkBench [EB/OL]. [2013-12-25]. http://www.zdnet.com/facebook-releasing-new-social-graph-database-benchmark-linkbench-7000013356/.

[24] FERGUSON M. Architecting a big data platform for analytics [R]. Riverton: IBM Corp, 2012.

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net