# 大数据基准测试程序包构建方法研究

熊 文 喻之斌 须成忠

（中国科学院深圳先进技术研究院云计算技术研究中心 深圳 518055）

**摘 要** 基准测试程序是评估计算机系统的关键测试工具。然而，大数据时代的到来使得开发大数据系统基准测试程序面临着更加严峻的挑战，当前学术界和产业界还不存在得到广泛认可的大数据基准测试程序包。文章利用实际的交通大数据系统构建了一个基于 Hadoop 平台的交通大数据基准测试程序包 SIAT-Bench。通过选取多个层次属性量化了程序行为特征，采用聚类算法分析了不同程序-输入数据集对的相似性。根据聚类结果，为 SIAT-Bench 选取了有代表性的程序和输入数据集。实验结果表明，SIAT-Bench 在满足程序行为多样性的同时消除了基准测试集中的冗余。

**关键词** 大数据基准测试程序；输入数据集；程序相似性；城市交通系统；GPS 轨迹数据
**中图分类号** TP 39 **文献标志码** A

# An Approach to Build a Big Data Benchmark Suite

XIONG Wen YU Zhibin XU Chengzhong

( *Cloud Computing Technology Research Center*, *Shenzhen Institutes of Advanced Technology*, *Chinese Academy of Sciences*, *Shenzhen* 518055, *China* )

**Abstract** Benchmarks are important tools to evaluate the performance of a variety of computing systems. However, benchmarks for big data systems are lacking as big data is relatively new and researchers are interested in understanding how big data systems including hardware and software work but do not have data. In this paper, an approach to develop big data benchmarks was devised at first. Then a big data benchmark suite named SIAT-Bench, which contains five representative workloads from Shenzhen urban transportation system, was presented. To this end, the program behavior was characterized and the impact of input data sets was qualified by observing metrics from multiple levels such as micro-architecture, OS and application layer. Then statistical techniques such as Principal Component Analysis (PCA) and Clustering were employed to perform similarity analysis between different workload-input pairs. Finally, we built SIAT-Bench by selecting representative workloads and associated input sets according to the clustering results. Experimental results show that SIAT-Bench properly satisfies the requirements of a benchmark suite.

**Keywords** big data benchmark; workload-input pairs; similarity; urban traffic systems; GPS trajectory data

**收稿日期**：2014-4-18
**作者简介**：熊文，博士研究生，工程师，研究方向为大数据基准测试和并行计算；喻之斌（通讯作者），副研究员，研究方向为计算机体系结构和性能评估，E-mail：zb.yu@siat.ac.cn；须成忠，研究员，研究方向为并行与分布式系统、互联网与云计算、高性能计算和移动嵌入式系统。

# 1 引 言

## 1.1 大数据的特点

由于云计算、物联网和社交网络等新兴服务的出现，人类社会的数据种类和规模正以前所未有的速度增长和扩大，标示着大数据时代正式到来[1]。一份来自谷歌的报告表明：2011 年，全球互联网用户占全部人口的 32.77%。这意味着全世界 23 亿人每天都在产生新的数据。2012 年 3 月，IBM 公司报告全世界每天产生的数据量达到了 2.5 EB(1 EB＝1000000000 GB)[2]。

大数据区别于其他数据的特征主要体现在三点：Volume(数据量大)、Velocity(速度快)和 Varity(种类多)[2]。大数据的量大指单个数据集达到了 PB 以上；速度快指数据的增长速度非常快；种类多指数据格式繁多，包括结构化、半结构化和非结构化数据。非结构化数据包括视频、音频、日志文件和其他一切不能方便地存储到传统关系型数据表中的数据。除此以外，一些大数据研究组织和社区认为 Value 也是大数据的一个基本特征，指数据量大但价值稀缺。他们认为，大数据问题是真实的问题，一个好的大数据解决方案应该能够给商业组织和其客户创造价值。

为了更好地管理和分析如此大规模的数据集，工业界和学术界提供了一系列不同的大数据解决方案。然而，目前尚未有广泛认同的基准测试程序集去评估这些不同的大数据系统，并公平比较这些系统的性能差异。上述大数据的特征为大数据基准测试集研发带来了巨大的挑战[3]。

## 1.2 大数据基准测试程序集研发的难点

构建大数据基准测试程序集主要面临五大挑战：(1)大数据系统的复杂性使得很难建立一个理想的基准测试模型；(2)大数据系统中应用领域的多样性使甄别典型的应用程序特征变得更加复杂；(3)大数据系统中的数据规模，为基准测试重现程序行为带来了巨大的挑战；(4)大数据系统的快速演化，要求基准测试包的更新能够跟得上数据系统的进化[2]；(5)没有真实的数据作为基准测试程序的输入。这些挑战使得目前还没有一个得到广泛认可的大数据基准测试程序集诞生。

## 1.3 已有的大数据基准测试程序包

由于大数据基准测试程序非常重要，许多机构和学者已经开始了相关工作，一些大型互联网公司和科研机构发布了相关领域的大数据基准测试程序包。如英特尔公司的 HiBench、雅虎公司的 YCSB 以及 YCSB＋＋和中国科学院计算技术研究所的 BigDataBench 等。然而，这些基准测试程序包都存在这样或那样的问题。

HiBench 是一个基于 Hadoop 平台的基准测试程序包，提供的基准测试程序既包括合成的基准测试也包括真实的应用程序。它以程序运行时间和系统吞吐率为基准测试的评价指标[4]；YCSB 是雅虎公司发布的一个关于云服务系统的基准测试程序包，它提供了一系列的核心基准测试程序和负载产生工具。这些负载可以有效对比 HBase、Cassandra、Yahoo!’s PNUTS 和 Sharded MySQL 等四种云服务平台的性能特征[5]，为基准测试受众者选择最优解决方案提供依据；BigDataBench 是中国科学院计算技术研究所发布的一个基于特定应用领域(网络搜索引擎)的大数据基准测试程序包[6]。

以上三个不同的基准测试程序包均较好地解决了各自领域的基准测试需求。但其相关介绍中并没有提及如何选择基准测试程序和为特定的基准测试程序选择输入数据集，且没有提供真实的输入数据集。

## 1.4 大数据基准测试程序集的要求

和传统基准测试程序包一样，大数据基准测试程序包也需要满足以下六个方面的要求：

(1)大数据基准测试应该覆盖多个应用领域

或同一领域的多个方面。当前，主要的大数据应用领域包括科学研究、健康护理、市场、金融、情报、社交媒体和零售等行业，这些不同应用领域对大数据系统提出了不同的要求[2]。

(2) 大数据基准测试应该覆盖多种数据类型，如结构化数据、半结构化数据和非结构化数据。具体来讲应该覆盖：图数据(如来源于社交网络或生物网络)、流式数据、地理信息数据和基因数据等。基准测试集的构建应该从应用程序级别开始，在这些不同应用和数据类型间甄别共有的关键数据处理程序，如排序等[2]。

(3) 大数据基准测试应该采用合成数据。在处理进行大数据基准测试时，从互联网上下载实际的大规模数据集代价非常昂贵，并且以当前的网络带宽来传输大数据集也不切合实际，因此，大数据基准测试包应该提供产生合成数据的算法和工具[2]。但对于这一点，学术界存在很多争议。有许多学者认为合成数据难以代表程序使用真实数据集时所表现出来的行为，本文亦赞同这一观点。

(4) 大数据基准测试应该考虑数据的隐私和安全。一些大数据集中包含了需要保密的信息，例如患者的医疗记录、保险公司的信息和军事数据等。因此，大数据基准测试使用者要求供应商提供保护隐私安全的大数据解决方案。

(5) 大数据基准测试应该考虑系统的可靠性。一些大数据系统往往需要批量处理任务和一些数据流信息，此时可靠性显得尤为重要，这些类型的应用对大数据解决方案提出了可靠性要求。

(6) 大数据基准测试标准应该学习已有的成功案例。我们在构建大数据基准测试应该学习传统计算机环境下已经被广泛认可的基准测试标准，如 TPC、SPEC 和 Top500 等。学习其构建模型的方法和性能评价指标等，甚至可以在其基础上直接扩展功能，添加大数据基准相关属性等方法来构建大数据基准测试集。

## 2　基准测试程序集的构建方法

大数据基准测试程序集的构建主要包括两方面的工作：(1) 选取有代表性的基准测试程序；(2) 为每个基准测试程序选取合适的输入数据集[7]。同时，大数据基准测试集还应该满足上述的几个要求。

为了满足这两个属性，必须解决以下几个问题：

(1) 甄别有代表性的基准测试程序；

(2) 分析多个基准测试间程序行为特征的相似性，在保留程序行为多样性的同时去冗余；

(3) 为特定的基准测试程序选取适当的输入数据集；

(4) 为基准测试程序选取评价指标。

本节将简单介绍 SIAT-Bench 的构建方法和初步结论，其详细方法、流程和结果将在后续的研究成果中陆续发布。

### 2.1　选取典型的应用程序

如图 1 所示，一个典型的大数据系统类似于一个流水线，由多个不同的数据处理阶段组成。具体的大数据处理流水线可能各有不同，但基本组成一般都包括图 1 所示的五个阶段。

基准测试程序包一般包括系统级的基准测试程序和组件级的基准测试程序[2]。一个系统级的基准测试程序会围绕整个大数据系统流水线进行。这样的基准测试程序也被称之为端到端的基准测试。而一些基准测试研究者期待一个基准测试程序能够测试整个大数据系统，这是不切实际的。一个好的系统级的基准测试程序能够为受众者提供简单直接的方法来比较不同的大数据系统。参与基准测试的所有大数据系统均使用相同的测试程序并通过相同的标准进行对比。系统级基准测试程序的优势是为系统性能提供一个简单明了的视图，不需要区分组件级别基准测试程序在不同阶段或过程中的具体执行情况。

组件级的基准测试程序比系统级的基准测试程序更具灵活性，组件级的基准测试程序也相对较容易定义，并且只测试系统的某一个方面，容易部署并且只作用于系统的目标组件。

### 2.2 确定基准测试程序和输入数据集

在本节中，我们通过对程序行为特征进行相似性分析来确定基准测试程序和其对应的输入数据集。程序行为的相似性分析包括两步。首先，以一组属性量化程序行为特征；其次，利用统计相关技术如熵权法、主成分分析和聚类技术对程序行为相似性进行分析。

为构建 SIAT-Bench，我们使用不同层次的特征来对程序行为进行分析：

(1) 应用级的特征，如系统的 IO 吞吐率、map 输入输出数据量的比率和 map 阶段与 reduce 阶段的运行时间比率；

(2) 操作系统级的特征，如磁盘读写的数据量、网络传输的数据量等；

(3) 微体系结构级的特征，如 IPC(Instruction Per Cycle) 和缓存缺失率 (Cache Miss Ratio) 等；

(4) 分布式系统级的特征，如各计算节点间的不平衡性。

从以上的层次中，我们选取了 21 个属性来描述程序行为特征。

在实验过程中，我们以程序——输入数据集的组合为基本描述对象 (Program-Input Pair)，以一个 21 个属性构成的向量来代表一个描述对象的行为。因此，这些不同向量可以用来量化分析不同程序之间的相似程度，使基准测试程序集保持程序行为多样性的同时消除基准测试集中的冗余程序。也可以通过分析不同输入数据集对程序行为的影响来为基准测试程序选取典型输入数据集。

数据分析的主要流程包括：

(1) 对原始数据进行熵权运算，对每个属性进行权重排序，熵权体现评价对象的区分度。按一定的标准在原始属性中按权重选出一个子集；

(2) 对 (1) 中的输出数据进行正则化处理 (均值为 0，方差为 1)，消除不同属性间量纲的差异；

(3) 对 (2) 中的输出数据进行主成分分析 (PCA)，确定主成分个数；

(4) 用各主成分的新坐标表示程序和其输入数据集组合；

(5) 计算表示程序和输入数据集组合向量的欧式距离，进行层次聚类。

应用程序之间程序行为的差异以及输入数据集对程序行为的影响，可以很直观地通过散点图和层次聚类图表达。两个不同程序与输入数据集组合的程序行为越相似，与之对应的两个向量在程序行为空间内越接近；反之，如果两个点距离较大，说明其对应的程序行为差异较大。

因此根据聚类结果，可以很容易去除基准测试程序集中的冗余程序。对同一基准测试程序也很容易根据聚类结果选出有代表性的输入数据集。

### 2.3 基准测试评价指标

性能评价指标是基准测试和对比不同系

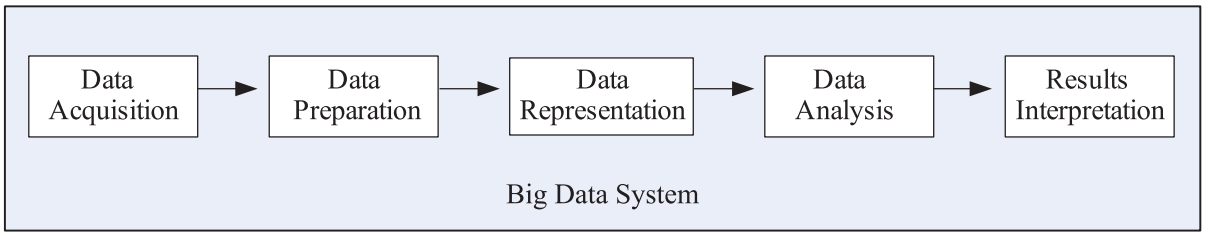


图 1 一个典型的大数据系统 Pipeline 模型

Fig. 1. Pipeline of a typical Big Data system

统的基础。一般情况下，除了系统的吞吐率(Throughput)，性能评价指标也包含性能(Performance)和成本(Cost)。基准测试的受众者会根据性能评价指标进行折中考虑，根据自身的需求选取性价比最高的大数据解决方案。

另外，基准测试结果的精确性(Correctness)和结果表现出的可预测性(Predictability)也是性能评价指标的重要方面。例如，如果测试结果表现出很好的可预测性，基准测试的受众者可以根据当前规模环境下的测试结果估算更大规模的基准测试结果。

## 3　实验和分析

本节以 Terasort 为例，使用 2.2 中描述的方法量化分析输入数据集对程序行为的影响，同时确定 Terasort 有代表性的输入数据集。

### 3.1　实验平台

在实验中，我们部署了一个包含 9 个节点的 Hadoop 集群，其中 8 个节点作为存储和计算节点，一个节点作为管理调度节点。全部节点均采用相同的软硬件配置并且通过一个千兆网卡相链接。具体配置如下：

每节点两个 Intel Xeon E5620 处理器，3 个 2 TB 的硬盘，16 GB 的 RAM；操作系统为 Ubuntu 12.04，内核版本是 3.2.0；Hadoop 的版本是 1.0.3，每个节点配置 8 个 map slot 和 8 个 reduce slot，每个 slot 分配 1 GB 的内存；JDK 的版本是 1.7.0；Terasort 来自 HiBench2.2；性能剖析工具为 oprofile-0.98。

我们使用 oprofile 获取微体系结构级别的信息，使用 Hadoop 平台自带的监控工具获取 job level 的信息，使用操作系统自带的命令 iostat 获取磁盘和网络资源使用信息，采用工具 ntp 进行集群时钟同步。

为了保证结果的准确性，每次实验前我们都对系统进行了预热，并且每组实验都进行 3 次以上，实验结果为三次实验结果的平均值。

### 3.2　结果分析

如图 2 所示，熵权值最大的是 map 与 reduce 任务平均耗时的比值，说明 map 任务平均耗时的变化程度和 reduce 任务平均耗时的变化程度非常

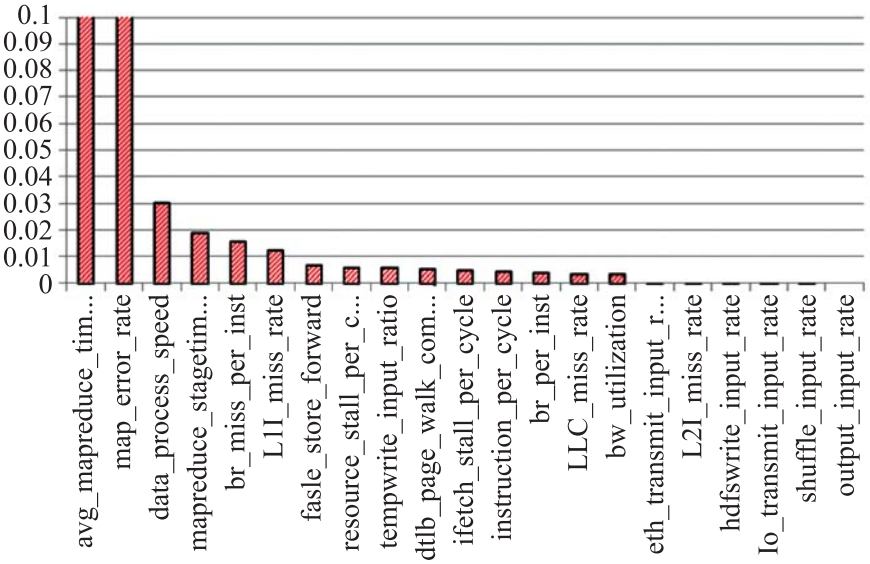


图 2　属性的熵权排序

Fig. 2.　Entropy weight for all metrics

不一致。原因如下：一方面，由于在运行过程中 reduce 任务的数量不变，输入数据集的增大导致了单个 reduce 任务平均运行时间的增加；另一方面，由于每个 map 任务处理的数据量是固定的，即使 map 任务需要处理的总数据量随输入数据集的增大而增加了，单个 map 任务的平均运行时间也基本不变。

熵权值次之的是失败的 map 任务数与总 map 任务数的比值。这是由于输入数据集的增加导致了 map 任务的总数增加，但发生错误的 map 任务个数基本固定(13～17)，并没有随着 map 任务总数的增加而增加。值得注意的是，目前的结论只针对 Terasort 程序。

图 2 中右侧几个量(L)的熵权值几乎为零。考虑到熵权是描述变量所起作用的权重，说明这些变量对描述程序行为基本不起作用。

主成份分析能去除原始变量中相互关联的变量，对原始数据进行降维，使数据的特征能够在平面图中更直观的展示。因此，我们对 Terasort 的特征数据进行了主成分分析。结果如图 3 所示，前四个主成份的贡献率分别是 57.44%、24.37%、8.96% 和 4.67%。前三个主成份其贡献率已累积达到 90.77%。

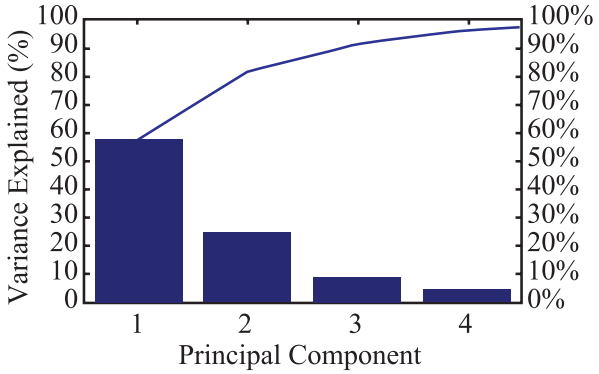


图 3 主成份的贡献率

Fig. 3. Rate of contribution of the principal component

图 4 是第一主成份和第二主成份的散点图。从图中可以看出有两个明显的点簇(即两个椭圆标示的区域)。图中的每个点表示一种尺寸的输入数据。输入数据小于 144 G(本文所使用的实验平台为 9 个节点的集群，共 144 GB 内存)的 Terasort 程序行为可以分为一组，而输入数据大于 144 G 的情况可以分为另一组。如果我们定义数据处理能力为 Terasort 在单位时间内能排序的数据量，则当输入数据集小于 144 GB 时，系统的处理能力保持在 0.09 GB/s 左右；但当输入数据集大于 144 GB 时，数据处理能力急剧降低到 0.03 GB/s 左右。这是因为输入数据集大于内存时，Terasort 处理程序受内存的限制需要将中间计算结果写到磁盘中，更多的 IO 导致了处理能力的下降。对于 Terasort 在输入数据集在 20 GB 的异常点，Terasort 全部操作都在内存中进行，磁盘 IO 相对较少，有着最高的 IPC 值。

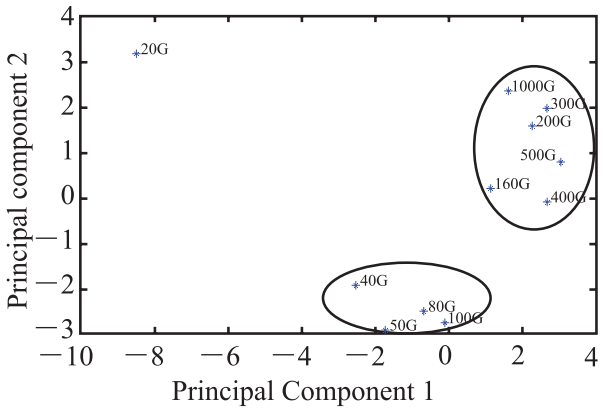


图 4 第一主成份和第二主成份散点图

Fig. 4. Scatter diagram of the first and second principal components

如图 5，该层次聚类图表达出除输入数据集在 20 GB 和 1000 GB 之外，数据集越接近，程序行为越相似，如 400 GB 和 500 GB、200 GB 和 300 GB、 80 GB 和 100 GB。按照一定的标准假设以距离 2.5 为划分，Terasort 的 11 个输入数据集可以划分为三类，分别是 20 GB、40 GB～500 GB 和 1000 GB。因此，我们推荐使用 20 G、40 G 和 1000 G 作为 Terasort 典型输入数集来进行基准测试。

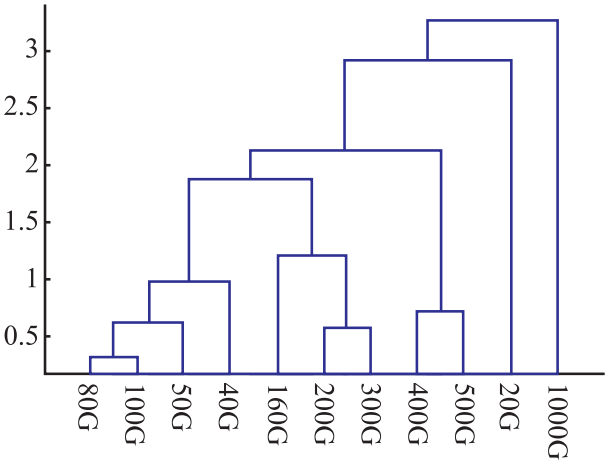


图 5　Terasort 多个数据集层次聚类

Fig. 5.　Hierarchical clusterings of Terasort datasets

## 4　SIAT-Bench

图 6 示意了中国科学院深圳先进技术研究院云计算技术研究中心为深圳市交通委员会开发的交通大数据处理系统。系统的数据来源于两部分：(1)深圳市出租车和公交车的实时 GPS 轨迹数据；(2)深圳市地铁公交智能卡实时交易数据。系统部署的应用分为两类，一类是面向交通委员会的非实时数据分析应用；另一类是面向公众的实时查询业务。

系统由三个子系统构成，分别是：(1)数据采集子系统。负责接收终端设备如 GPS 终端和智能卡读卡器的数据，验证数据的有效性，将数据存储到 HDFS 或 HBase 中；(2)数据存储子系统。这是一个多层次的混合式云存储系统，由 HDFS、HBase 和传统的关系型数据库 mysql 构成。数据在不同阶段将被存储在不同的存储子系统中；(3)数据应用子系统，在系统中部署的两类对外提供服务的应用。

系统规模包含 1500 多万张深圳通卡，30000 多辆出租车和公交车。平均每天产生 1200 万条深圳通卡交易记录和 9000 万条 GPS 轨迹数据。目前系统保存了近一年以来的全部数据，累计数

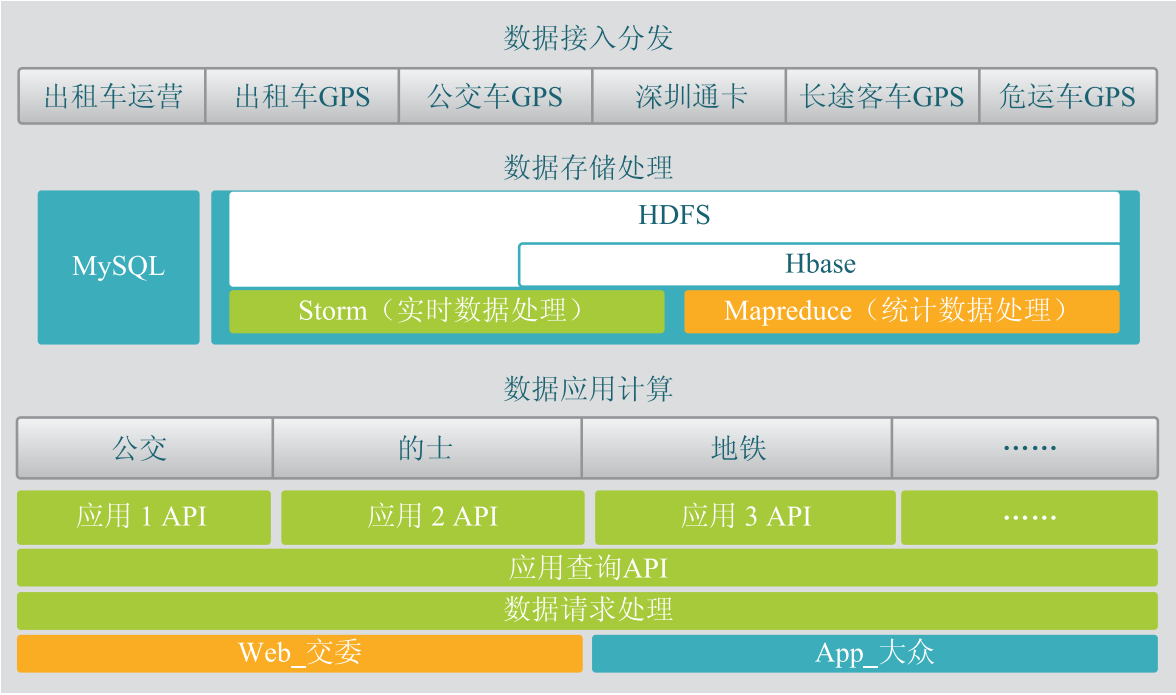


图 6　深圳市交通大数据系统架构

Fig. 6.　Architecture of a Big Data system for transportation system in Shenzhen

据总量达到 7 TB 以上。

SIAT-Bench 以该实际系统为基础，目标是实现一套交通大数据基准测试程序包，使其能够准确代表交通领域的典型应用如数据采集、存储和索引、分析和挖掘等。利用 SIAT-Bench 能够准确评估交通大数据系统的性能和特点。

### 4.1 SIAT-Bench 特点

SIAT-Bench 中程序的功能特点为：(1)建立交通领域数据采集，分析和挖掘的应用模型；(2)能准确刻画交通数据处理过程中的典型应用场景，准确评估同类交通大数据系统的性能；(3)客观重现了典型交通数据处理程序的行为特征；(4)支持大规模数据集，每天 9000 万条 GPS 轨迹数据和 1500 万条智能卡刷卡数据。

### 4.2 基准测试程序介绍

SIAT-Bench 目前包含 5 个应用程序，均基于 Hadoop 平台。其中有 2 个程序使用 Apache pig 平台实现，3 个程序由 java 语言实现。程序具体描述如下：

(1) Mapmatching(GPS 轨迹数据地图匹配)。由于实际采集的 GPS 轨迹数据(经度和纬度)与数字地图中的道路存在偏差，用 Mapmatching 来将出租车和公交车的 GPS 数据轨迹和数字地图进行准确匹配。它使用 Java 实现，是其他应用如交通流量分析的基础。

(2) Secondarysort(二次排序)。该程序以时间戳和出租车车牌号为主键对交通数据进行排序，它是数据预处理的主要步骤，基于 pig 实现。

(3) Traffic hotregion(出租车时空分布分析)。该程序统计分析深圳市全部出租车的时空分布，基于 java 实现。图 7 示意了该程序的运行结果，从图中可以明确看出在某时刻机场东站区域有空的士 85 辆，载客的士 310 辆。

(4) Sztod(出租车或人群流动统计)。该程序

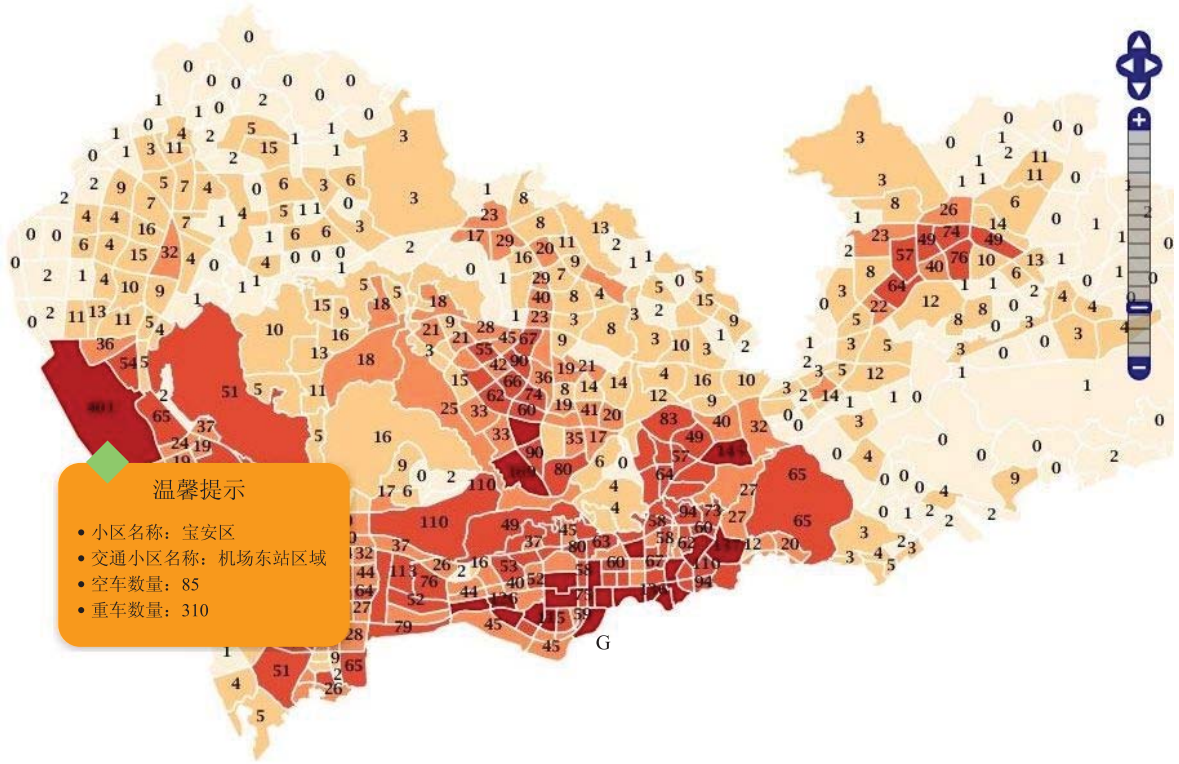


图 7 深圳市某一时刻出租车的分布情况

Fig. 7. Taxicab distribution at a certain time in Shenzhen

基于 java 来统计指定时间段内从区域 A 到区域 B 的出租车或人的数量。

（5）Traffic hotspot（交通热点分析）。该程序通过 pig 来统计市内交通热点如火车站、购物中心和机场等地点交通流量并进行分析。

## 5　结论和工作展望

从实验分析过程可以看出，在构建大数据基准测试程序包时，量化大数据典型程序行为的方法并对程序行为的相似性进行分析可以有效满足开发基准测试程序包的两个要求，在保持程序行为多样性的同时消除冗余性。

对于 Terasort，当输入数据集在 200 G～1000 G 变化时，程序行为相似。因此针对 Terasort 只需选定 200 G 为其代表性的输入数据集，既可以准确评估 Terasort 在输入数据为 1000 G 时的行为特征，又能准确推算程序的运行时间，这个方法节约了 80% 的系统评估时间。

我们将进一步完善 SIAT-Bench 的功能，建立准确的数据更新模型，构建更加准确的基准测试程序。

## 参 考 文 献

[1] Meng XF, Ci X. Big data management: concepts, techniques and challenges [J]. Journal of Computer Research and Development, 2013, 50(1): 146-169.

[2] [EB/OL]. http://www.cse.wusl.edu/~jian/cse567-13/ftp/bigdata/index.html.

[3] Chen YP. We don’t know enough to make a big data benchmark suite-an academia-industry view [C] // Workshop on Big Data Benchmarking, 2012.

[4] Huang SS, Huang J, Dai J, et al. The HiBench benchmark suite: characterization of the MapReduce-based data analysis [C] // 2010 IEEE 26th International Conference on Data Engineering Workshops, 2010: 41-51.

[5] Gao WL, Zhu YQ, Jia Z, et al. Bigdatabench: a Big Data Benchmark Suite from Web Search Engines [Z]. arXiv preprint arXiv:1307.0320, 2013.

[6] Cooper BF, Silberstein A, Tam E, et al. Benchmarking cloud serving systems with YCSB [C] // Proceedings of the 1st ACM Symposium on Cloud Computing, 2010: 143-154.

[7] Eeckhout L, Vandierendonck H, De Bosschere K. Quantifying the impact of input data sets on program behavior and its applications [J]. Journal of Instruction-Level Parallelism, 2003, 5(1): 1-33.

[8] Phansalkar A, Joshi A, John LK. Analysis of redundancy and application balance in the spec cpu2006 benchmark suite [C] // Proceedings of the 34th Annual International Symposium on Computer Architecture, 2007: 412-423.