



# 빅데이터 벤치마크 (Big Data Benchmarks) 현황



임태형 TTA 소프트웨어시험인증연구소 SW시험평가1팀 선임연구원

## 1. 머리말

빅데이터 기술의 확산으로 빅데이터 시스템의 객관적인 성능 비교를 위한 빅데이터 벤치마크의 필요성이 증가하고 있다. 그러나 빅데이터 플랫폼 및 솔루션이 계속 진화 중이고, 빅데이터 시스템이 다양한 독립적 구성요소로 구현되기 때문에 성능 기준의 정의가 쉽지 않은 점 등의 이유로 업계 표준 벤치마크가 통용되기까지는 조금 더 시간이 필요할 것으로 예상된다. 이른바 ‘빅데이터 킬러 앱’이 등장하여 시장을 이끌고, 모든 벤더가 동일 도메인 상에서 성능에 대한 경쟁을 시작할 때, 업계 표준 빅데이터 벤치마크도 윤곽을 드러낼 것이다. 본고에서는 TPCx-HS[1], BigBench[2], BigDataBench[3] 등을 중심으로 빅데이터 벤치마크 현황에 대해 살펴보고자 한다.

## 2. TPCx-HS

기존 DBMS(Database Management System) 분야의 업계 표준 벤치마크를 주도하고 있는 TPC(Transaction Performance Council)는 BI(Business Intelligence)를 위한 DSS(Decision-support System) 벤치마크표준인 TPC-H[4]와 TPC-DS[5]를 제정하는 등 빅데이터 시대를 위한 준비를 꾸준히 진행해 왔다. TPC-H는 기업의 비즈니스 의사결정을 위해 요구되는 복잡한 질의(Query) 처리 성능을 위한 벤치마크로, 1999년 7월 리비전 1.0.0을 시작으로 2014년 11월 리비전 2.17.1까지 개정되었으며, 2015년 9월까지 오라클, IBM, HP 등의 제품을 대상으로 254개의 벤치마크 결과가 공표되었다[6]. 그러나 TPC-H는 초기 DSS 환경을 시험하기에는 충분했지만 DSS 기술의 진화 속도를 따라가기에는 부족함이 있었다. TPC-H를 개선하기 위해 TPC는 차세대 DSS 벤치마크인 TPC-DS를

<표 1> TPCx-HS 측정 항목

측정 항목	측정 요소	측정 방법
<i>HSph@SF</i>	Performance (Sort Throughput)	$HSph@SF = SF / (T / 3600)$
$\$/HSph@SF$	Price/Performance	$\$/HSph@SF = P / HSph@SF$
Availability Date	시스템 가용성	벤치마크에 사용된 시스템을 고객이 구입할 수 있는 일자
Energy Metric (Optional)	Energy	$E / (T * HSph@SF)$

\* SF: 스케일 팩터 (Scale Factor), T: 경과 시간(초), P: 시험 대상 시스템(SUT) 구축 비용, E: 에너지 사용량(Watt)

<표 2> TPCx-HS 스케일 팩터 (Scale Factor)

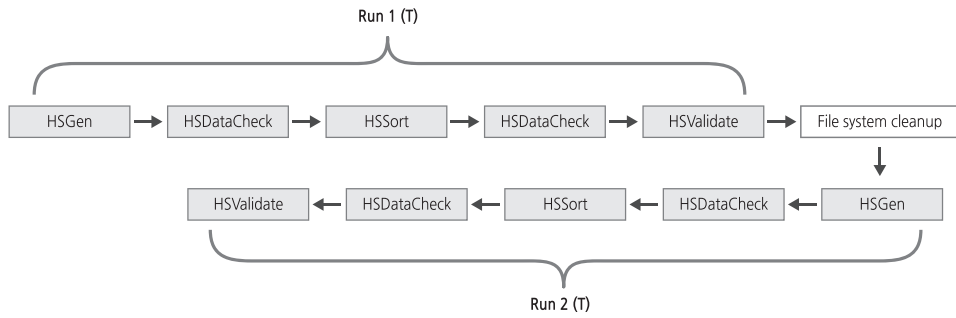
Scale Factor	1	10	30	100	300	1000	3000	10000	30000	100000
DB size (GB)	1	10	30	100	300	1,000	3,000	10,000	30,000	100,000

제정했다. TPC-DS는 2012년 1월 리비전 1.0.0을 시작으로 2015년 8월 리비전 1.4.0까지 개정되었으며, 현재 TPC-DS 버전 2를 위한 드래프트 버전의 리뷰를 진행하고 있다. TPC-DS 버전 2는 아파치 하둡(Apache Hadoop) 기반 시스템에서 실행 가능하도록 개정될 예정이며, 트랜잭션의 ACID(Automicity, Consistency, Isolation and Durability) 특성 중 지속성(Durability) 완화, 테이블 수정(Update) 금지 등의 사항이 변경될 예정이다. TPC-DS는 데이터 볼륨과 데이터 속도의 일부 측면에서 빅데이터 특성을 반영했으나 빅데이터의 핵심 요소인 비정형 데이터의 처리 등 데이터 다양성을 지원하지 않기 때문에 완전한 빅데이터 벤치마크로 사용되기에는 한계가 있다. 또한 아파치 하둡 등의 빅데이터 기술이 기업용 IT 생태 시스템에 중요한 부분으로 급부상함에 따라 아파치 하둡 맵리듀스(MapReduce) 및 HDFS(Hadoop Distribution File System)의 API(Application Programing Interface)와 호환되는 시스템에 대한 객관적인 성능 측정 방법이 요구되었다. 이에 부응하기 위해 TPC는 2013년 11월, 빅데이터 벤치마크

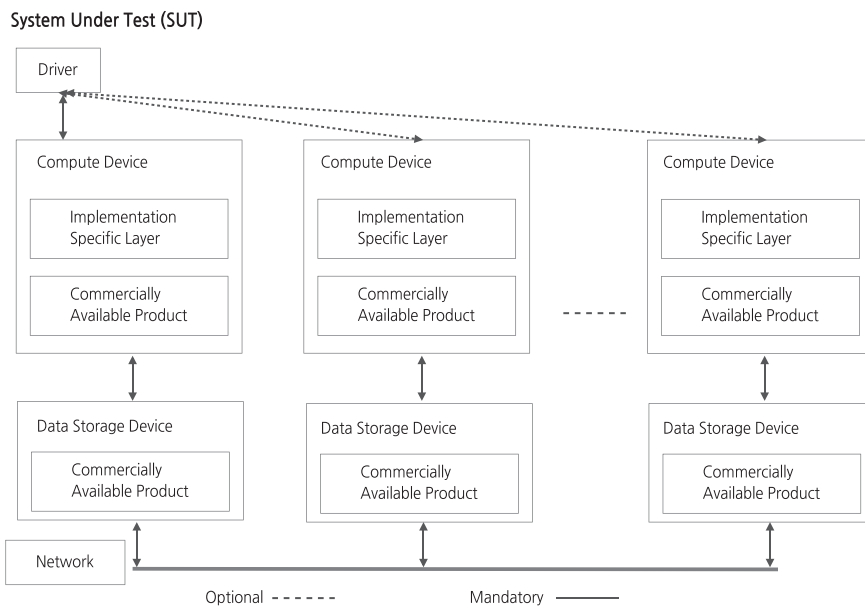
워킹 그룹인 TPC-BD(Big Data Working Group)를 만들고, 2014년 6월, 하둡 기반 빅데이터 시스템의 하드웨어 및 소프트웨어의 범용적인 성능을 측정할 수 있는 첫 번째 벤치마크인 TPCx-HS[1] 1.0.0을 제정하였다. TPCx-HS는 하둡 기반 시스템의 정렬(Sort) 성능을 위한 마이크로 벤치마크인 TeraSort[7]에 기반한 워크로드(Workload)를 사용하여 *HSph@SF*,  $\$/HSph@SF$  등의 성능을 측정한다.

TPCx-HS에서 정의한 스케일 팩터에 따라 데이터를 생성하는 HSGen은 TeraGen을 기반으로 하며, 정렬 프로그램인 HSSort는 TeraSort를 기반으로 한다. 또한, 출력 데이터를 검증하는 HSDataCheck는 TeraValidate를 기반으로 한다. TPCx-HS의 실행은 5개의 순차적 단계로 진행되며, 단계별 소요 시간이 기록된다.

1. HSGen을 통한 입력 데이터 생성
2. HSDataCheck를 통한 데이터 세트의 검증
3. HSSort를 이용하여 입력 데이터 정렬
4. HSDataCheck를 통한 데이터 세트의 검증
5. HSValidate를 통한 출력 데이터 유효성 확인



[그림 1] TPCx-HS Execution Phases

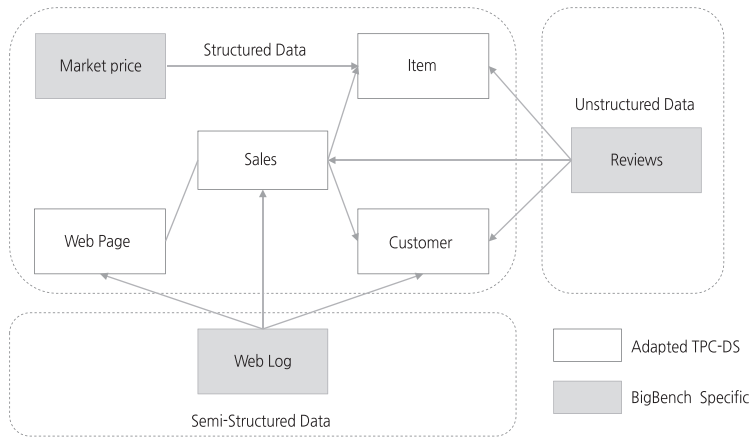


[그림 2] TPCx-HS System Under Test

TPCx-HS의 벤치마크는 상기 5단계를 순차적으로 진행하여 총 2회 시험을 수행하며 첫 번째 시험과 두 번째 시험 사이에는 파일 시스템 클린업(File System Cleanup) 이외에 시스템 튜닝이나 설정 변경 등의 어떠한 다른 작업도 허용되지 않는다. 2회 시험 결과 중 더 낮은 성능 측정 결과가 TPCx-HS의 성능 실행 결과(Performance Run)로 결과서에 기록된다.

TPCx-HS의 벤치마크를 위한 시험 대상 시스템

(SUT)은 드라이버(Driver), 연산장치(Compute Device), 데이터 저장 장치(Data Storage Device)로 구성되며, 각 구성요소에는 하드웨어와 소프트웨어가 모두 포함된다. TPCx-HS 벤치마크 구현 및 실행을 위해 생성된 드라이버 및 ISL(Implementation Specific Layer) 이외의 모든 하드웨어 및 소프트웨어 제품은 반드시 상용 제품을 사용해야 하며, 상용 제품을 제외한 드라이버 및 ISL의 모든 소스코드는 반드시 공개되는 것을 원칙



[그림 3] BigBench 데이터 모델

으로 한다. 현재 TPCx-HS의 표준 규격은 2015년 2월 리비전 1.2.1까지 완료되었고, 2015년 9월까지 시스코(Cisco)와 델(Dell) 제품을 대상으로 총 8건의 벤치마크 결과가 발표되었다[13].

### 3. CLDS Big Data Benchmarking Community

캘리포니아 대학 샌디에이고 캠퍼스(UC San Diego)의 CLDS(Center for Large-scale Data Systems research)[8]는 2012년 5월 미국 국립과학재단 (National Science Foundation)의 지원 및 Mallanox, Seagate, BROCADE, GREENPLUM 등 기업의 후원으로 'WBDB2012(Workshop on Big Data Benchmarking 2012)'를 개최하였다. WBDB2012에서는 학계와 산업계에서 초청된 60여명의 전문가들이 모여 빅데이터 벤치마크를 만들기 위한 기초 작업으로, 빅데이터 벤치마크 애플리케이션의 특징, 빅데이터 하드웨어/소프트웨어 성능 메트릭, 벤치마크 감사 규정 등 업계 표준에 관한 폭넓은 의견을 나누었다[9]. WBDB2012의 결과 빅데이터 분야의 산업 표준 벤치마크를

제정하기 위한 BDBC(Big Data Benchmarking Community)가 만들어 졌으며, 2015년 9월 현재까지 WBDB를 포함하여 총 8번의 워크숍이 개최되었다. WBDB에서 추구하는 빅데이터 벤치마크의 요건은 벤치마크의 구현 및 실행이 쉽고 (Simple), 비용이 합리적이며(Cost effective), 시장 및 기술의 흐름에 부합하고(Timely), 독립적인 방법을 통해 검증되어야(Verifiable) 한다는 것이다[10]. 'BigBench'[2]는 WBDB를 통해 제안된 주요 결과물 중 하나로, TPC-DS를 기반으로 정형 (Structured) 데이터뿐만 아니라, 반정형(Semi-Structured), 비정형(Un-structured) 데이터까지 지원하는 빅데이터 벤치마크이다. BigBench에서는 실제 현실을 반영하기 위해 5개의 주요 비즈니스 영역에서 빅데이터 분석을 위한 30개의 질의 (Query)를 지원하며, 질의 처리 시간(BBQpH)이 주요 측정 항목이다. WBDB의 활발한 활동은 TPC-BigBench의 소위원회 및 SPEC의 빅데이터 연구 그룹이 만들어지는데 크게 기여했으며, 2015년 12월 인도 뉴델리에서 열리는 7<sup>th</sup> WBDB까지 이어질 예정이다.

<표 3> BigBench 측정 항목 및 파라미터

측정 항목 / 파라미터	설명
BBQpH	BigBench Queries per Hour = $(30 * 3 * 3600) / (T_L + T_p + (T_{TT1}/S) + T_{DM} + (T_{TT2}/S))$
$T_L$	Execution time of the loading process
$T_p$	Execution time of the power test
$T_{TT1}$	Execution time of the first throughput test
$T_{DM}$	Execution time of the data maintenance task
$T_{TT2}$	Execution time of the second throughput test

<표 4> BigDataBench Data Sets

Data Sets		Raw data size
1	Wikipedia Entries	4,300,000 English articles(unstructured text)
2	Amazon Movie Reviews	7,911,684 reviews(semi-structured text)
3	Google Web Graph	875713 nodes, 5105039 edges(unstructured graph)
4	Facebook Social Network	4039 nodes, 88234 edges(unstructured graph)
5	E-commerce Transaction Data	Table 1: 4 columns, 38658 rows. Table 2: 6 columns, 242735 rows (structured tables)
6	ProfSearch Person Resumes	278956 resumes(semistructured table)

#### 4. BigDataBench

BigDataBench는 중국의 중국과학원(Chinese Academy of Sciences)이 만든 오픈 소스 빅데이터 벤치마크 프로젝트로, 2013년 7월에 검색 엔진을 위한 6개 워크로드로 구성된 BigDataBench 1.0, 데이터 분석을 위한 11개 워크로드로 구성된 DCBench 1.0, 혼합 데이터 분석을 위한 워크로드로 구성된 CloudRank 1.0의 총 3가지 형태로 구분된 벤치마크로 시작하였다. 2013년 12월 기존 3가지 벤치마크를 BigDataBench 2.0으로 통합하여, 현실성이 반영된 6개의 데이터 세트, 19개의 워크로드를 사용하여 마이크로 벤치마크(Micro Benchmark), 클라우드 온라인 트랜잭션 처리(Cloud OLTP), 관계형 질의(Relational Query), 검색 엔진, 소셜 네트워크, 전자 상거래의

6개 빅데이터 애플리케이션 시나리오를 지원하였다. 2014년 4월 32개의 빅데이터 워크로드를 지원하는 BigDataBench 3.0이 발표되었으며, 2015년 9월 현재 최신 버전은 2014년 12월에 발표된 BigDataBench 3.1이다. BigDataBench 3.1은 실제 서비스 중인 14개의 데이터 세트와 33개 워크로드를 지원하며, 텍스트, 그래프, 이미지, 오디오, 비디오, 테이블 데이터 등 여러 가지 형태의 비정형, 반정형, 정형 데이터를 모두 지원한다. 또한, BigDataBench 3.1에서는 기존 애플리케이션 이외에 바이오인포매틱스(Bioinformatics)와 멀티미디어가 애플리케이션 도메인에 추가되었다. 중국의 학계 및 산업계가 참여한 중국의 첫 번째 빅데이터 업계 표준 벤치마크인 BigDataBench-DCA는 BigDataBench의 하위 프로젝트로 중국의 정보기술산업부(China's Ministry of Industry and


Information Technology)에 표준 제정을 위해 제출되어 검토 중이다.

## 5. SPEC RG Big Data Working Group

TPC와 함께 컴퓨팅 시스템 성능에 관한 표준 벤치마크를 주도하고 있는 SPEC은 다른 기관이나 단체에 비해 조금 늦은 2014년 10월 빅데이터 벤치마크를 위한 연구 그룹을 만들어 활동을 시작했다. SPEC RG BDWG(Standard Performance Evaluation Corporation Research Group Big Data Working Group)는 산학이 협력하여 빅데이터 시스템 벤치마크를 위한 성능 측정 방법론에 관한 연구를 수행하는 것을 목표로 한다. SPEC RG BDWG는 빅데이터 시스템에서 측정할 주요 성능 목표를 정의하고, 성능 측정을 위한 기준 및 규정을 개발하여 기존에 제안된 벤치마크와의 협력을 조성하는 것을 세부 목표로 한다. 이를 위해 SPEC RG BDWG는 2015년에 열리는 6<sup>th</sup> WBDB와 7<sup>th</sup> WBDB의 개최를 돕고 빅데이터 벤치마크에 관한 학회 참석 등 연구 활동에 집중할 계획이다[12].

## 6. 맺음말

WBDB2012에서부터 시작된 빅데이터 벤치마크에 대한 논의는 빅데이터 시스템의 진화와 더불어 계속 진행될 전망이다. TPC는 하둡 기반 빅데이터 시스템의 하드웨어 및 소프트웨어의 범용적인 성능을 측정할 수 있는 첫 번째 벤치마크인 TPCx-HS 1.0.0을 제정하였고, CLDS BDBC는 WBDB를 통해 빅데이터 벤치마크를 주도적으로 이끌어 나가고 있다. 중국은 정부 주도하에 산학연이 참여하여 BigDataBench를 만들어 국가 표준으로 진행하

고 있다. 반면, 국내의 경우 정부가 공공데이터 개방 및 규제 완화 등의 빅데이터 정책을 통해 빅데이터 산업을 육성하고자 노력하고 있으나, 빅데이터 벤치마크에 대한 준비나 관심은 크지 않은 실정이다. 빅데이터 벤치마크를 통해 국내 벤더들의 기술력을 향상시키고, 고객들이 자신들이 원하는 시스템을 쉽게 비교 선택할 수 있는 환경을 조성하는 것도 빅데이터 산업 육성의 좋은 정책적 선택이 될 수 있다. 빅데이터 시대에 대비하기 위해서 빅데이터 벤치마크에 대한 국내 참여가 더욱 활발해지기를 기대해 본다. 

### [참고문헌]

- [1] Transaction Processing Council (TPC), 'TPC EXPRESS BENCHMARK™ HS Standard Specification version 1.3.0,' Feb. 19<sup>th</sup> 2015
- [2] A. Ghazal, M. Hu, T. Rabl, F. Raab, M. Poess, A. Crolotte, and H.-A. Jacobsen, 'Bigbench: Towards an industry standard benchmark for big data analytics,' In SIGMOD 2013, 2013년
- [3] <http://prof.ict.ac.cn/BigDataBench/industry-standard-benchmarks/> [Accessed on: 2015-09-15]
- [4] Transaction Processing Council (TPC), 'TPC BENCHMARK™ H Standard Specification Version 2.17.1,' Nov. 13<sup>th</sup> 2015
- [5] Transaction Processing Council (TPC), 'TPC BENCHMARK™ DS Standard Specification Version 1.4.0,' Aug. 27<sup>th</sup> 2015
- [6] [http://www.tpc.org/tpch/results/tpch\\_results.asp](http://www.tpc.org/tpch/results/tpch_results.asp) [Accessed on: 2015-09-15]
- [7] <http://sortbenchmark.org>. [Accessed on: 2015-09-15]
- [8] <http://clds.ucsd.edu/bdbc> [Accessed on: 2015-09-15]
- [9] Baru, Chaitanya, 'Setting the Direction for Big Data Benchmark Standards,' Lecture Notes in Computer Science, Vol. 7755, 2013, Pages 197-208
- [10] <http://clds.sdsc.edu/wbdb2014.de> [Accessed on: 2015-09-15]
- [11] <https://research.spec.org/working-groups/big-data-working-group.html> [Accessed on: 2015-09-15]
- [12] [http://www-conf.slac.stanford.edu/xldb2015/Talks2015/LT\\_9\\_Tues\\_Poelman\\_XLDB-Presentation-RGBD\\_V2.pdf](http://www-conf.slac.stanford.edu/xldb2015/Talks2015/LT_9_Tues_Poelman_XLDB-Presentation-RGBD_V2.pdf) [Accessed on: 2015-09-15]
- [13] [http://www.tpc.org/tpcx-hs/results/tpcxhs\\_results.asp](http://www.tpc.org/tpcx-hs/results/tpcxhs_results.asp) [Accessed on: 2015-09-15]