데이터베이스 워크로드에서의 자원 식별

오 정 석[†] · 이 상 호[††]

요 약

데이터베이스 응용분야에 따라 데이터베이스 워크로드는 서로 다른 자원 사용 형태를 보인다. 데이터베이스 관리자는 워크로드 특성을 반영하는 자원 관리를 통하여 시스템 성능을 향상시킬 수 있다. 본 논문은 성능지표와 자원간의 관계를 분석하여 데이터베이스 시스템 성능에 영향을 주는 자원을 선별하는 방법을 제시한다. 첫째, 본 방법은 피어슨 상관계수와 유의도 검정을 적용하여 데이터베이스 시스템 자원 확장에 대해 감소되거나 증가되는 성능지표를 선별한다. 둘째, 감소/증가 관계를 갖는 성능지표를 이용하여 데이터베이스 시스템에 성능에 영향을 주는 자원을 선별한다. 실험은 TPC-C 및 TPC-W 환경에서 본 방법을 수행하였으며, 제안된 자원 선별 방법에 대한 검증 시험을 수행하였다.

키워드 : 데이터베이스 워크로드, 성능 지표, 자원 사용

Resource Identification in Database Workloads

Jeong Seok Oh[†] · Sang Ho Lee[††]

ABSTRACT

Database workloads may show different resource usages for database applications. Database administrators can enhance the DBMS performances through resource management that reflects workload characteristics. We provide a method that can identify tunable resources from analyzing the relationship between performance indicators and resources. First, we select which performance indicators increase or decrease by expanding resources using a correlation coefficient and a significance level test. Next, we identify resources that can affect the DBMS performances by using increasing or decreasing performance indicators. We evaluated our method in the TPC-C and TPC-W environments.

Key Words : Database Workload, Performance Indicator, Resource Usage

1. 서 론

워크로드는 시스템에 부하를 가할 수 있는 요소의 집합이며, 데이터베이스 워크로드는 데이터베이스 질의 집합을 의미한다. 데이터베이스 시스템에서 수행되는 워크로드는 데이터베이스 응용 분야에 따라 다른 워크로드 특성을 보일 수 있다. 데이터베이스 응용 분야가 다양화되고 복잡해짐에 따라, 데이터베이스 응용분야의 독특한 워크로드를 고려한 데이터베이스 관리가 요구된다. 데이터베이스에서 자원의 변경은 워크로드에 따라 데이터베이스 시스템의 성능에 주는 영향이 다를 수 있으므로 데이터베이스 시스템의 자원 사용 형태가 측정되어야 한다. 자원 사용 형태의 측정은 일반적으로 운영체제나 데이터베이스 시스템에서 제공되는 성능지표들이 이용된다.

데이터베이스 워크로드를 분석하고 자원 사용을 고려한 연구는 문헌에서 찾아볼 수 있다. [3]은 수행되는 자원 영역에 따라 워크로드를 디스크 버퍼 관련 트랜잭션/질의 집합과 작업 메모리(working storage) 관련 트랜잭션/질의 집합으로 세분화하였고, 트랜잭션/질의별로 목표 응답시간을 설정하여 자원 할당 크기와 다중 메모리 단계(multiprogramming level)를 피드백 알고리즘에 의해 동적으로 조절하는 방법을 제안하였다. [8]은 전자 상거래 시스템에서 세 개의 응용분야에 대한 워크로드 특징을 분석하고 QoS(quality of service) 요구사항들을 정립하였고, QoS 요구사항을 동적으로 충족하는 DBMS 관리 기술인 Quartermaster 시스템을 개발하였다. [1]은 데이터베이스 시스템에서 성능을 하락시키는 자원을 진단하는 연구를 수행한다. 자원 진단은 상호 관련된 자원들의 영향에 관한 지식을 적용하는 진단 알고리즘을 제안하였다. [5]는 데이터 마이닝 기법인 의사결정 트리를 이용하여 워크로드 타입을 식별하는 연구를 수행하였다. 총 9개의 성능지표를 이용하여 워크로드 데이터를 축적하였으며 IBM DB2 Intelligent Miner에 의해 워크로드 모델을 생성하고 워크로드 타입 식별을 수행하였다.

기존의 연구 결과들은 워크로드에 따라 데이터베이스 시스템의 성능에 영향을 주는 자원이 틀릴 수 있음을 고려하

† 정 회 원 : 한국가스안전공사 가스안전연구개발원
†† 종신회원 : 숭실대학교 컴퓨터학부 교수
논문접수 : 2005년 11월 22일, 심사완료 : 2006년 2월 23일

지 않았다. 본 논문의 선행연구인 [21]에서 워크로드 종류에 따라 자원사용이 틀렸고, 데이터베이스 시스템에 영향을 주는 자원도 달랐다. 예를 들어, 데이터 버퍼 적중률은 워크로드 종류에 따라 다른 비율을 보였다. 이러한 결과에 의거하여 워크로드 종류에 따라 데이터베이스 시스템에 영향을 주는 자원을 선별하는 방법이 요구된다.

본 논문의 목적은 성능지표와 자원간의 관계를 분석하여 데이터베이스 시스템에 영향을 주는 자원을 선별하는 것이다. 성능지표와 자원간의 관계는 피어슨 상관계수와 유의도 검정에 의해 결정된다. 피어슨 상관계수는 자원 확장에 대한 자원 사용이 증가되거나 감소되는 경향을 조사하며 모든 상관이 실제로 증가되거나 감소되지 않기 때문에 유의도 검정을 적용한다. 유의도 검정은 실제로 감소되거나 증가되는 상관계수를 t-검정식에 의해 식별한다. 감소되거나 증가되는 성능지표가 존재하는 자원은 데이터베이스 시스템의 성능을 변화시키는 요인으로 간주되어 효과적인 데이터베이스 시스템 관리를 위해 선별된다.

본 논문에서는 제안된 방법을 적용하여 TPC-C와 TPC-W에서 수집된 워크로드 데이터를 이용하여 유의 수준 0.05에서 상관계수에 대한 유의도 검정을 수행함으로써 워크로드 종류에 따라 자원 확장에 대해 기록되는 성능지표가 감소되거나 증가되는 관계를 선출한다. 워크로드 데이터 구축을 위해 국제 표준 데이터베이스 성능평가로 인정받는 TPC-C와 TPC-W를 사용한다[16, 17]. TPC-C는 도매 업체의 재고 관리 시스템을 가상 운영할 수 있는 OLTP 환경의 워크로드를 제공한다. TPC-W는 인터넷 전자 서점의 전자 상거래 시스템을 모델링 하는 웹 기반 환경의 워크로드를 제공한다. 워크로드 데이터는 네 개의 자원(데이터 버퍼, 공유 메모리, 개인 메모리, I/O 프로세스)을 확장하면서 수행된 114회의 성능평가에서 14개의 성지지표 값을 워크로드 데이터로 수집하여 분석된다. 본 논문은 제안된 방식을 검증하기 위해 TPC-C와 TPC-W에서 각 자원 크기에서 최대 수행될 수 있는 웨어하우스와 EB수를 측정하여 본 논문의 자원 선별 방식의 결과와 비교한다. 본 연구의 결과는 데이터베이스 튜닝을 비롯하여 데이터베이스 시스템의 자동화된 DBMS 관리에 필요한 선행정보를 제공할 수 있다.

본 논문의 구성은 다음과 같다. 2장은 TPC-C와 TPC-W에서 워크로드 데이터를 수집하는 방법에 관해 설명한다. 3장은 변경된 자원과 성능지표간의 감소/증가 관계를 분석하고 검정하는 방법을 설명하고, 4장은 TPC-C와 TPC-W에서 수집된 워크로드 데이터를 분석하여 자원과 성능지표간의 감소/증가 관계를 추출한 결과를 기술한다. 5장은 TPC-C와 TPC-W 환경에서 적용된 본 방식의 결과를 검증한다. 6장은 수행된 내용에 대해 결론을 맺고 향후 계획을 제시한다.

2. 워크로드 데이터 수집

워크로드 데이터는 자원의 크기를 변경시킨 후 성능평가가 수행되는 동안 성능지표의 값을 수집하여 구축된다. 자원은 데이터베이스 시스템의 성능에 영향을 미치며 시스템 구동 중에 자원의 크기를 변경할 수 있는 것으로[1-3, 8-10, 18, 19]를 참조하여 데이터 버퍼, 공유 메모리, 개인 메모리, I/O 프로세스를 선별하였다. [3]은 데이터베이스 성능에 영향을 주는 메모리 자원을 데이터 버퍼와 작업 메모리로 구분하였다. 데이터 버퍼는 자주 사용되는 데이터를 오랫동안 보관한다. 적절한 버퍼의 크기는 디스크 접근 횟수를 감소시켜 질의 결과에 대한 응답시간을 감소시킨다. 작업 메모리는 사용자 질의에 대한 처리 정보를 보관하여 질의 구문 분석의 최소화 및 질의 데이터 처리의 최적화를 위해 사용된다. 본 논문에서는 [4]를 참조하여 사용 목적과 데이터 공유 여부에 따라 개인 메모리와 공유 메모리로 세분화하였다. 공유 메모리는 일반적으로 자주 사용되는 질의에 대한 구문 분석 및 실행 계획 등이 저장되고 데이터는 다른 사용자에게 공유된다. 개인 메모리는 조인, 정렬, 커서 등의 정보가 저장되고 데이터는 다른 사용자에게 공유되지 않는다.

데이터 버퍼, 공유 메모리, 개인 메모리, I/O 프로세스 자원에 대한 크기 조절은 사용된 데이터베이스 시스템에서 제공하는 db_cache_size, shared_pool_size, pga_aggregate_target, dbwr_io_size 파라미터를 사용하여 수행하였다. 파라미터의 변경은 데이터베이스 시스템의 성능 하락을 발생시킬 수 있으므로 주의 깊게 설정해야 한다. 파라미터의 초기값과 증가값은 데이터베이스 시스템에서 제공하는 파라미터 기본값으로 하였으며, 파라미터의 변경범위는 <표 1>과 같이 초기값을 기준으로 14번 변경시킨다. 파라미터 변경 방식은 변경 대상이 되는 파라미터만이 변경되며 다른 파라미터들은 초기값으로 설정된다. 예를 들어, 데이터 버퍼의 크기 변경에서 db_cache_size를 64MB로 변경할 때 다른 3개의 파라미터 값은 초기값으로 설정된다.

〈표 1〉 자원의 종류와 변경 범위

자원 (파라미터)	초기값	증가값	최대값
데이터 버퍼 (db_cache_size)	32MB	32MB	480MB
공유 메모리 (shared_pool_size)	32MB	32MB	480MB
개인 메모리(pga_aggregate_target)	20MB	20MB	300MB
I/O 프로세스 (I/O process)	1 개	1 개	15개

워크로드 데이터 속성은 [1, 3, 4, 5, 8, 12]를 참조하여 14개의 성능지표를 이용한다. 14개의 성능지표는 워크로드 특징을 구분할 수 있는 식별성(identities)과 데이터베이스 시스템 구동 중에 쉽게 접근할 수 있는 동적 접근성(dynamic accessibility)을 바탕으로 선별되었으며 종류는 <표 2>와 같다.

워크로드 데이터는 총 114회(변경된 파라미터 수 × 파라미터의 종류 × 성능평가의 수-2)의 성능평가 수행을 통해 수집되었다. TPC-W 환경은 데이터베이스, 웹/응용프로그램, 이미지 서버를 별개의 시스템에 장착하였다. 서버에 사용되는 데이터베이스는 오라클 9i 버전을 이용하였으며 웹서버는 웹로직(WebLogic)과 아파치 웹서버를 이용하였다. TPC-C 환경은 클라이언트와 데이터베이스 시스템(오라클 9i)이 동일 기계에서 구동된다.

〈표 2〉 성능지표의 종류와 의미

성능지표	의미
데이터 버퍼 적중률	검색하는 데이터가 데이터 버퍼에 존재할 확률을 의미한다.
공유 메모리 적중률	검색하는 질의 정보가 공유 메모리에 존재할 확률을 의미한다.
시스템 카탈로그 적중률	검색하는 카탈로그 정보가 시스템 카탈로그에 존재할 확률을 의미한다.
래치 경합 비율	파싱을 위해 수행된 CPU 시간을 파싱 수행에 경과된 시간으로 나눈 비율을 의미한다.
메모리 정렬 비율	메모리에서 데이터가 정렬될 비율을 의미한다.
메모리 파싱 비율	파싱을 수행할 때 공유 메모리에 존재하는 데이터를 읽어 파싱을 수행할 비율을 의미한다.
데이터 변경률	디스크에서 메모리로 읽은 데이터가 변경될 비율을 의미한다.
데이터 버퍼 읽기량	디스크에서 데이터 버퍼로 읽는 데이터의 용량을 의미한다.
데이터 비버퍼 읽기량	디스크에서 데이터 버퍼가 아닌 메모리의 특정 부분으로 읽는 데이터의 용량을 의미한다.
데이터 버퍼 쓰기량	데이터 버퍼에서 디스크로 쓰이는 데이터 용량을 의미한다.
데이터 비버퍼 쓰기량	데이터 버퍼가 아닌 다른 부분에서 디스크로 쓰이는 데이터의 용량을 의미한다.
디스크 쓰기량(체크포인트)	체크 포인트가 발생했을 때 DBWR(database writer) 프로세스에 의해 디스크로 쓰이는 데이터의 용량을 의미한다.
디스크 쓰기량(비체크포인트)	체크 포인트가 아닌 다른 이유로 디스크로 쓰이는 데이터의 용량을 의미한다.
리두 로그량	생성되는 리두 로그 데이터 용량을 의미한다.

3. 자원 식별을 위한 자원과 성능지표의 관계 분석 방법

본 논문은 자원 확장에 따른 성능지표 변화의 판단을 위해 유의도 검정을 적용한 상관관계를 구한다. 유의도 검정을 적용한 상관계수법은 상관 계수에 의해 서로 다른 변수 간의 관계를 파악하며, 유의도 검정에 의해 데이터와 유의 수준에 따라 상대적으로 의미가 존재하는 관계를 구한다. 이용되는 상관계수 수식은 피어슨(pearson) 상관계수를 이용하며 (수식 1)처럼 공분산(covariance)을 두 변수에 대한 표준편차의 곱으로 나누어 구한다. $\overline{X}$와 $\overline{Y}$는 두 변수의 평균을 의미하며 n은 변수 집합의 개수를 의미한다. 피어슨 상관계수는 +1에서 -1사이의 값을 가진다. 상관계수의 값이 양의 수이면 한 변수가 증가할 때 다른 변수도 증가함을 의미하며 +1에 가까울수록 선형에 가깝다. 상관 계수의 값이 음의 수이면 한 변수가 증가할 때 다른 변수는 감소함을 의미하며 -1에 가까울수록 선형에 가깝다.

$$COE(X,Y)=\frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\overline{X})^2\sum_{i=1}^{n}(Y_i-\overline{Y})}} \quad \text{(수식 1)}$$

상관계수 수식을 자원과 성능지표간의 관계에 적용하면 $\overline{X}$와 $\overline{Y}$는 기록된 성능지표 집합과 변경된 자원 크기 집합의 평균을 의미하며 n은 기록된 성능지표 값의 수를 의미한다. 자원과 성능지표간의 관계가 양의 상관을 보이면 자원의 크기가 증가할수록 성능지표의 값이 증가하는 관계가 존재할 수 있다. 자원과 성능지표간의 관계가 음의 상관을 보이면 자원의 크기가 증가할수록 성능지표의 값이 감소하는 관계가 존재할 수 있다. 모든 상관계수의 값이 증가하거나 감소되는 관계가 아니므로 t-검정식을 이용하여 실제로 의미 있는 상관을 선별한다. t-검정식은 귀무가설(null hypothesis), 대립가설(alternative hypothesis), 유의 수준(significance level)을 설정하고, t-분포를 이용하여 기각영역(rejection area)과 채택영역(acceptance area)을 구분하는 임계값(critical value)을 결정하며, 통계량이 임계값에 따라 어느 영역에 속하는지 판별하여 결과를 해석하는 방법이다. 귀무가설이 상관이 없다고 가정될 때, 검정 통계량이 채택영역에 속하면 귀무가설이 채택되어 두 변수간의 상관이 없음을 의미하며 검정 통계량이 기각 영역에 속하면 귀무가설이 기각되어 두 변수간의 상관이 있음을 의미한다. 유의도 검정에 이용되는 t-검정식은 (수식 2)에서 보인다. r은 두변수간의 상관계수이며 n은 자료 집합의 개수를 의미한다.

$$t=\frac{r}{\sqrt{1-r^2}}\sqrt{n-2} \quad \text{(수식 2)}$$

자원 K와 성능지표 P와 Q에 대한 실제로 감소/증가 관계가 존재하는 상관 계수를 구하는 방법을 [예제 1]에서 보인다.

[예제 1] 워크로드 X에서 자원 K의 크기 확장에 대한 성능지표 P와 Q의 상관관계를 구하려 한다. 자원 K의 크기 변경(단위 : megabyte)과 기록된 성능지표 P와 Q의 값과 유의도 검정을 위한 가설이 [보기]와 같다고 가정하자.

[보기]
K = { 32, 64, 96, 128, 160, 192, 224, 256, 288, 320}
P = { 27.21, 27.49, 27.45, 27.22, 27.43, 26.5, 26.95, 27.1, 27.11, 27.03}
Q = { 74.32, 76.79, 78.25, 80.63, 81.69, 81.95, 84.3, 84.61, 87.7, 89.41}
귀무가설(H_0) : 변경된 자원과 기록된 성능지표 사이는 상관이 없다 ($\rho=0$).
대립가설(H_a) : 변경된 자원과 기록된 성능지표 사이는 상관이 있다 ($\rho\neq0$).
유의수준(α) : 0.01
H_0의 기각영역과 채택영역의 임계값 : $t_{\alpha/2}(n-2)$

[자원 K와 성능지표 P의 감소/증가 관계 판별 과정]

$\acute{K}$은 K의 각 원소를 K의 평균으로 감산한 집합이며, $\acute{P}$는 P의 각 원소를 P의 평균으로 뺀 집합이다. 공분산 $COV(K,P)$는 -124.32로 계산되고 K와 P의 표준편차의 곱 $STD(K) \times STD(P)$는 255.6256으로 계산된다. K와 P의 상관관계 $COE(K,P)$는 K와 P의 공분산을 K와 P의 표준편차의 곱으로 나누며 -0.48634로 계산되어 음의 상관을 보인다. 성능지표 P의 값은 자원 K의 크기 확장에 대해 감소되는 경향을 보인다. K와 P의 상관계수 대한 검정 통계량은 -1.5743으로 계산된다. 계산된 검정통계량은 t-분포에서 유의수준 0.01일 때 채택영역과 기각영역을 나누는 임계값을 찾아야 하며 t-분포표를 이용한 임계값은 $t_{0.005}(8)$이 되어 3.355로 설정된다. 즉, 채택 영역은 $-3.355 \le t \le 3.355$이며, 기각영역은 $t < -3.355$ 또는 $t > 3.355$가 된다. 계산된 t값은 유의수준 0.01에서 채택영역 안에 있으므로 귀무가설을 채택하여 자원과 성능지표 사이에 상관이 없음을 나타낸다. 유의도 검정결과에 의해 K의 크기 확장에 의해 P는 감소되는 관계로 인정될 수 없다.

$$\acute{K} = \{-144, -112, -80, -48, -16, 16, 48, 80, 112, 144\}$$

$$\acute{P} = \{0.061, 0.341, 0.301, 0.071, 0.281, -0.65, -0.2, -0.05, -0.039, -0.12\}$$

$$COV(K,P) = \sum_{i=1}^{n} (\acute{K}_i \times \acute{P}_i) = -124.32$$

$$\acute{K}^2 = \{20736, 12544, 6400, 2304, 256, 256, 2304, 6400, 12544, 20736\}$$

$$\acute{P}^2 = \{0.0037, 0.116, 0.091, 0.005, 0.079, 0.421, 0.04, 0.002, 0.0015, 0.014\}$$

$$STD(K) \times STD(P) = \sqrt{\sum_{i=1}^{n} \acute{K}_i^2 \times \sum_{i=1}^{n} \acute{P}^2} = 255.6256$$

$$COE(K,P) = \frac{COV(K,P)}{STD(K) \times STD(P)} = -0.48634$$

$$\text{검정통계량}(t) = \frac{-0.48634}{\sqrt{1-(-0.48634)^2}} \sqrt{10-2} = -1.5743$$

[자원 K와 성능지표 Q의 감소/증가 관계 판별 과정]

$\acute{K}$은 K의 각 원소를 K의 평균으로 감산한 집합이며, $\acute{Q}$는 Q의 각 원소를 Q의 평균으로 뺀 집합이다. 공분산 $COV(K,Q)$는 4084로 계산되고 K와 Q의 표준편차의 곱 $STD(K) \times STD(Q)$는 4127.42로 계산된다. K와 Q의 상관관계 $COE(K,Q)$는 K와 Q의 공분산을 K와 Q의 표준편차의 곱으로 나누어 0.98948로 계산되어 양의 상관을 보인다. 성능 지표 Q의 값은 자원 K의 크기 확장에 대해 양의 상관을 보이므로 증가되는 경향을 보인다. K와 Q의 상관계수 대한 검정 통계량은 19.34525로 계산된다. 채택 영역은 $-3.355 \le t \le 3.355$이며, 기각영역은 $t < -3.355$ 또는 $t > 3.355$이므로 계산된 t값은 유의수준 0.01에서 채택영역 안에 있지 않으므로 귀무가설을 기각한다. 귀무가설의 기각은 변경된 자원 크기 집합 K와 기록된 성능 지표 집합 Q 사이에 상관이 있음을 의미한다. 유의도 검정결과에 의해 K의 크기 확장에 의해 Q는 증가되는 관계로 인정된다.

$$\acute{K} = \{-144, -112, -80, -48, -16, 16, 48, 80, 112, 144\}$$

$$\acute{Q} = \{-7.645, -5.18, -3.72, -1.34, -0.28, -0.02, 2.335, 2.645, 5.735, 7.445\}$$

$$COV(K,Q) = \sum_{i=1}^{n} (\acute{K}_i \times \acute{Q}_i) = 4084$$

$$\acute{K}^2 = \{20736, 12544, 6400, 2304, 256, 256, 2304, 6400, 12544, 20736\}$$

$$\acute{Q}^2 = \{58.446, 26.78, 13.8, 1.782, 0.076, 0.0002, 5.452, 6.996, 32.89, 55.43\}$$

$$STD(K) \times STD(Q) = \sqrt{\sum_{i=1}^{n} \acute{K}_i^2 \times \sum_{i=1}^{n} \acute{Q}^2} = 4127.42$$

$$COE(K,Q) = \frac{COV(K,Q)}{STD(K) \times STD(Q)} = 0.98948$$

$$\text{검정통계량}(t) = \frac{0.98948}{\sqrt{1-(0.98948)^2}} \sqrt{10-2} = 19.34525$$

■

4. TPC-C와 TPC-W에서 자원과 성능지표간의 감소/증가 관계

본 연구는 TPC-C와 TPC-W에서 수집된 워크로드 데이터를 이용하여 자원과 성능지표 사이의 상관계수를 계산하고 유의도 검증을 수행한다. 유의도 검증을 위해 수립한 가설은 <표 3>과 같다. 자원과 성능지표 사이에 상관이 없다고 가설하고 상관계수의 검정 통계량을 계산한다. 유의수준 0.05 일 때 검정 통계량이 기각 영역에 해당되면 가설은 기각되고 자원과 성능지표 사이에 상관이 실제로 존재하여 본 논문에서는 자원 크기의 확장에 따른 성능지표가 감소/증가되는 것을 인정한다. 검정 통계량이 채택 영역에 해당되면 가설은 채택되어 자원과 성능지표 사이에 상관이 없는 것으로 간주하여 본 논문에서는 자원 크기의 확장에 따른 성능지표가 감소/증가 되는 것을 인정하지 않는다.

<표 3> 검증을 위한 가설과 가설의 기각/채택 영역

귀무가설(H_0) : 변경된 자원과 기록된 성능지표 사이는 상관이 없다 ($\rho=0$).
대립가설(H_a) : 변경된 자원과 기록된 성능지표 사이는 상관이 있다 ($\rho \neq 0$).
유의수준(α) : 0.05
t-분포에서 유의 수준 0.05와 데이터 자료에 의한 임계값 : |3.372|
H_0의 기각영역: t-검정통계량 < -3.372 또는 t-검정통계량 > 3.372
H_0의 채택영역: -3.372 < t-검정통계량 < 3.372

<표 4>는 TPC-C 성능평가에서 확장되는 db_cache_size 파라미터와 각 성능지표 사이의 상관계수와 유의도 검정에 의한 t-검정 통계량을 보인다. 유의 수준 0.05에서 데이터에

〈표 4〉 TPC-C 성능평가에서 db_cache_size 파라미터와 성능지표간의 상관계수와 검정 통계량

성능지표	상관계수	t-검정 통계량	유의
데이터 변경률	-0.29353	-1.10698	-
데이터 버퍼 적중률	0.996103	40.70532	○
공유 메모리 적중률	0.160614	0.58667	-
메모리 파싱 비율	0.144883	0.528017	-
시스템 카탈로그 적중률	-0.30929	-1.1727	
메모리 정렬 비율	0	0	-
래치 경합 비율	0.298839	1.128912	-
데이터 버퍼 읽기량	-0.9979	-55.5472	○
데이터 비버퍼 읽기량	-0.42433	-1.68945	-
데이터 버퍼 쓰기량	-0.91671	-8.27179	○
데이터 비버퍼 쓰기량	-0.42433	-1.68945	-
디스크 쓰기량(체크포인트)	0.870988	6.392314	○
디스크 쓰기량(비체크포인트)	-0.92264	-8.62323	○
로그 데이터량	-0.58489	-2.60002	-

〈표 5〉 TPC-C 성능평가에서 shared_pool_size 파라미터와 성능지표간의 상관계수와 검정 통계량

성능지표	상관계수	t-검정 통계량	유의
데이터 변경률	-0.86759	-6.29063	○
데이터 버퍼 적중률	-0.79973	-4.802867	○
공유 메모리 적중률	-0.08386	-0.30348	-
메모리 파싱 비율	-0.22212	-0.82137	-
시스템 카탈로그 적중률	0.434117	1.73749	
메모리 정렬 비율	0	0	-
래치 경합 비율	-0.68928	-3.43032	○
데이터 버퍼 읽기량	-0.55285	-2.39216	-
데이터 비버퍼 읽기량	-0.66143	-3.17971	-
데이터 버퍼 쓰기량	-0.9507	-11.0534	○
데이터 비버퍼 쓰기량	-0.69347	3.479	○
디스크 쓰기량(체크포인트)	-0.06651	-0.240342	-
디스크 쓰기량(비체크포인트)	-0.97353	15.3574	○
로그 데이터량	-0.95531	-11.652	○

대한 t-분포의 임계값이 ±3.372 이므로, 검정 통계량이 기각 영역에 존재하여 실제로 상관이 의미가 있는 성능지표는 다섯 개가 보인다. 확장되는 데이터 버퍼 자원에 대해 증가되는 성능지표는 데이터 버퍼 적중률, 디스크 쓰기량(체크포인트)이며, 감소되는 성능지표는 데이터 버퍼 읽기량, 데이터 버퍼 쓰기량, 디스크 쓰기량(비체크포인트)이다.

<표 5>는 TPC-C 성능평가에서 확장되는 shared_pool_size 파라미터와 각 성능지표 사이의 상관계수와 유의도 검정에 의한 t-검정 통계량을 보인다. 유의 수준 0.05에서 데이터에 대한 t-분포의 임계값이 ±3.372이므로, 검정 통계량이 기각 영역에 존재하여 실제로 상관이 의미 있는 성능지표는 일곱 개가 보인다. 확장되는 공유 메모리 자원에 대해 증가되는 성능지표는 없으며, 감소되는 성능지표는 데이터 버퍼 적중률, 데이터 변경률, 데이터 버퍼 쓰기량, 래치 경합 비율, 데이터 비버퍼 쓰기량, 디스크 쓰기량(비체크포인트), 로그 데이터량이다. TPC-C 성능평가에서 pga_aggregate_target와 dbwr_io_slave 파라미터의 확장은 유의도 검정에 의한 검정통계량이 모두 채택 영역에 존재하여 확장되는 자원에 따라 성능지표의 증가나 감소가 존재하지 않았다.

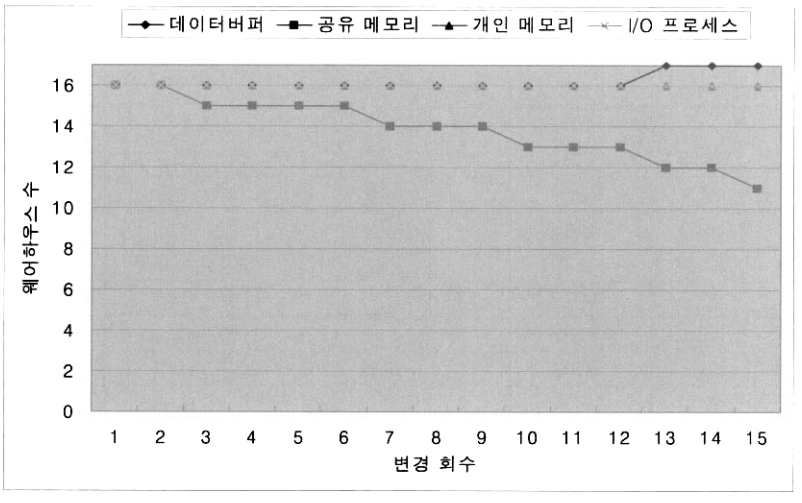
〈표 6〉 TPC-W 성능평가에서 db_cache_size 파라미터와 성능지표간의 상관계수와 검정 통계량

성능지표	상관계수	t-검정 통계량	유의
데이터 변경률	0.060751	0.219624	-
데이터 버퍼 적중률	0.780532	4.50149	○
공유 메모리 적중률	-0.24744	-0.92063	-
메모리 파싱 비율	-0.21744	-0.79221	-
시스템 카탈로그 적중률	0.045392	0.163861	
메모리 정렬 비율	0	0	-
래치 경합 비율	0.2117724	0.804217	-
데이터 버퍼 읽기량	-0.77491	-4.42021	○
데이터 비버퍼 읽기량	0.371154	1.441362	-
데이터 버퍼 쓰기량	-0.24389	-0.90678	-
데이터 비버퍼 쓰기량	0.216226	0.798403	-
디스크 쓰기량(체크포인트)	0.488516	2.018548	-
디스크 쓰기량(비체크포인트)	-0.2017	-0.77714	-
로그 데이터량	0.039129	0.141446	-

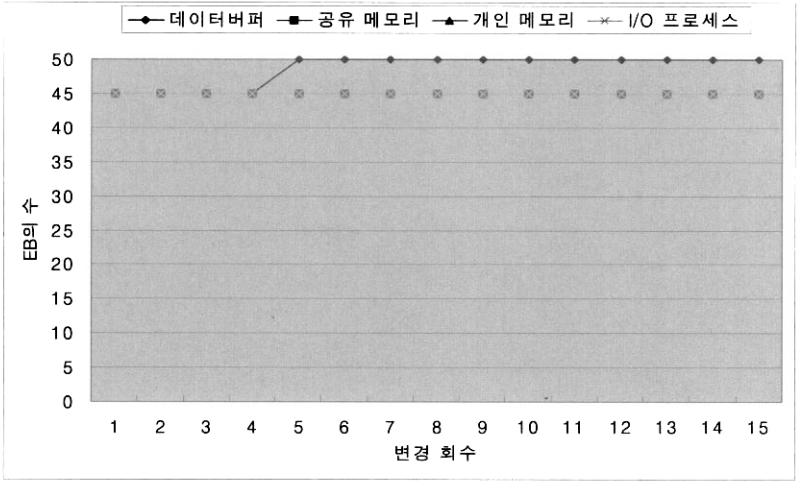
<표 6>은 TPC-W 환경에서 db_cache_size 파라미터와 각 성능지표간의 상관계수와 유의도 검정에 의한 t-검정 통계량을 보인다. 유의 수준 0.05에서 데이터에 대한 t-분포의 임계값이 ±3.372 이므로, 검정 통계량이 기각 영역에 존재하여 실제로 의미 있는 상관을 보이는 성능지표는 두 개가 보인다. 확장되는 데이터 버퍼 자원에 따라 증가되는 성능지표는 데이터 버퍼 적중률이며, 감소되는 성능지표는 데이터 버퍼 읽기량이다. shared_pool_size, pga_aggregate_target, dbwr_io_slave 파라미터의 확장은 유의도 검정에 의한 검정 통계량이 모두 채택 영역에 존재하여 확장되는 자원에 따라 성능지표의 증가나 감소가 존재하지 않았다.

5. 자원 선별에 대한 검증

자원과 성능지표 사이의 감소/증가 관계는 자원에 대한 성능지표의 변화를 탐지하므로 자원 선별을 위해 사용된다. 상관계수와 유의도 검정을 적용한 결과에 의하면 TPC-C 환경에서는 데이터 버퍼와 공유 메모리에서 성능지표들이 감소/증가되므로 데이터 버퍼와 공유 메모리의 변경은 데이터베이스 시스템의 성능에 영향을 줄 수 있다. TPC-W 환경에서는 데이터 버퍼에서만 성능지표들이 감소/증가되므로 데이터 버퍼만이 데이터베이스 시스템의 성능에 영향을 줄 수 있다. 자원 선별 방식을 검증하기 위해, 본 논문은 TPC-C와 TPC-W 환경에서 각 자원 크기에서 최대 수행될 수 있는 웨어하우스 수와 EB 수를 각각 측정하였다.



(그림 1) TPC-C에서 자원크기별 최대 수행된 웨어하우스 수



(그림 2) TPC-W에서 자원크기별 최대 수행된 EB의 수

(그림 1)은 네 개의 자원의 변경 크기에서 TPC-C 성능평가의 수행조건에 부합되면서 최대 수행될 수 있는 웨어하우스의 수를 측정하였다. 데이터 버퍼는 416MB(13단계)이상의 확장에서 최대 수행될 수 있는 웨어하우스의 수가 16개에서 17개로 증가되었다. 데이터 버퍼 자원의 확장은 디스크 입/출력 횟수를 줄여 데이터베이스 시스템의 성능 향상을 유도할 수 있다. 하지만 TPC-C에서는 빈번하게 갱신 질의가 발생되므로 본 시험에서는 416MB 이상의 데이터 버퍼 확장에서 웨어하우스 수가 증가한다. 공유 메모리는 최대 수행될 수 있는 웨어하우스의 수가 16개에서 96MB(3단계), 224MB(7단계), 320MB(10단계), 416MB(13단계), 480MB(15단계)로 확장할 때 한 개씩 감소되었다. TPC-C에서 공유 메모리 파라미터의 확장은 수행된 TPC-C 성능평가의 3분의 2가 TPC-C의 제약조건(웨어하우스 당 9 tpmC이상 기록되는 제약조건)을 만족하지 않아 발생한다. 오라클에서 지나치게 큰 공유 메모리는 TPC-C와 같이 빈번하게 질의객체 할당과 해제가 빈번하게 일어는 환경에서 자유공간(free space) 리스트의 검색과 할당에 소요되는 시간을 지연시킴으로써 시스템 성능을 하락시킨다. 반면에, 개인 메모리와 I/O 프로세스의 확장은 최대 수행될 수 있는 웨어하우스의 수가 변화되지 않았다.

(그림 2)는 네 개의 자원의 변경 크기에서 TPC-W 성능평가의 수행조건에 부합되면서 최대 수행될 수 있는 EB의 수를 측정하였다. 데이터 버퍼는 160MB(5단계) 이상의 확장에서 최대 수행될 수 있는 EB 수가 45개에서 50개로 증가되었다. 다른 자원들은 모든 확장된 자원 크기에서 최대 수행될 수 있는 EB 수를 변화되지 않았다. 최대 부하의 측정 결과는 본 논문의 자원 선별 방식의 결과와 일치하여 본 논문의 자원 선별 방법이 옳음을 입증하고 있다. TPC-W 성능평가에서 데이터 버퍼의 확장은 갱신질의가 많지 않기 때문에 본 시험에서 160MB 이상의 확장에서 EB수를 증가시켰다. 반면에, TPC-W 성능평가 데이터베이스 환경에서 공유 메모리의 확장은 자유공간에 질의객체를 빈번하게 할당/해제하지 않아 객체 할당 및 검색 시간을 지연시키지 않기

때문에 데이터베이스 시스템의 성능에 영향을 주지 않았다.

6. 결론 및 향후계획

본 논문에서는 성능지표와 자원간의 관계를 분석하여 데이터베이스 시스템 성능에 영향을 주는 자원을 선별하는 방법을 제시하였다. 본 방법은 상관계수와 유의도 검정에 의한 자원과 성능지표간의 감소/증가 관계를 이용하여 데이터베이스 시스템에 성능에 영향을 주는 자원을 선별한다.

본 논문은 TPC-C와 TPC-W 환경에 자원 선별 방식을 적용하여 데이터베이스 시스템의 성능에 영향을 주는 자원을 선별하였다. TPC-C에서는 데이터 버퍼 적중률, 데이터 버퍼 읽기량, 데이터 버퍼 쓰기량, 디스크 쓰기량(체크포인트), 디스크 쓰기량(비체크포인트)에서 데이터 버퍼 자원과 성능지표간의 감소/증가 관계가 존재하였고 데이터 변경률, 데이터 버퍼 적중률, 래치 경합 비율, 데이터 버퍼 쓰기량, 데이터 비버퍼 쓰기량., 디스크 쓰기량(비체크포인트), 로그 데이터량에서 공유 메모리 자원과 성능지표간의 감소/증가 관계가 존재하였다. TPC-W에서는 데이터 버퍼 적중률과 데이터 버퍼 읽기량에서 데이터 버퍼 자원과 성능지표간의 감소/증가 관계가 존재하였다. 즉, TPC-C에서는 데이터 버퍼와 공유 메모리가 데이터베이스 시스템의 성능에 영향을 주는 자원으로 선별되었으며, TPC-W에서는 데이터 버퍼만이 선별됨을 의미한다.

TPC-C와 TPC-W에서 본 방식의 결과를 검증하기 위해 본 논문은 각 자원의 변경 크기에서 최대 수행될 수 있는 부하를 측정하였다. 본 논문의 결과는 효과적인 데이터베이스 튜닝을 비롯하여 자동화된 DBMS 관리를 위한 선행 정보를 제공하며, 워크로드에 따라 데이터베이스 시스템에 영향을 주거나 활용 방식의 개선이 필요한 자원을 식별함으로써 데이터베이스 시스템의 튜닝 범위를 명확하게 해준다. 향후 계획으로는 유의수준이 상관관계에 많은 영향을 미치므로 보다 객관적인 유의수준을 선별하는 방법과 선별된 자원에 따라 수행될 수 있는 데이터베이스 시스템의 성능향상 방법에 대하여 연구한다.

참 고 문 헌

[1] D. G. Benoit, "Automated Diagnosis and Control of DBMS Resources", Ph.D Workshop on EDBT Conference, 2000.

[2] K. P. Brown, M. J. Carey, and M. Livny, "Goal-Oriented Buffer Management Revisited", Proceedings of ACM SIGMOD Conference, pp.353-364, Montreal, 1996.

[3] K. P. Brown, M. Mehta, M. J. Carey, and M. Livny, "Towards Automated Performance Tuning For Complex Workloads", Proceedings of 20th VLDB Conference, pp.72-84, Santiago, 1994.

[4] M. Cyran, "Oracle 9i : Database Performance Guide and Reference, Release 2(9.2)", Oracle Corporation, 2001.

[5] S. Elnaffar, P. Martin, and R. Horman, "Automatically Classifying Database Workloads", Proceedings of 11th CKIM Conference, pp.622-624, McLean, 2002.

[6] D. M. Lane, "Hyerstat Online: An Introductory Satistics Textbook and Online Tutorial for Help in Statistic", http://davidmlane.com/hyperstat/index.html

[7] J. Neter, M. J. Kunter, C. J. Nachtsheim, W. Wasserman, "Applied Linear Regression Models", IRWIN Books, 1996.

[8] P. Martin, H. Y. Li, M. Zheng, K. Romanufa, and W. Powley, "Dynamic Reconfiguration Algorithm : Dynamically Tuning Multiple Buffer Pools", Proceedings of 11th DEXA Conference, pp.92-101, London, 2000.

[9] P. Martin, W. Powley, H. Y. Li, and K. Romanufa, "Managing Database Server Performance to Meet QoS Requirements in Electronic Commerce Systems", International Journal on Digital Libraries, Vol.3, No.4, pp.316-324, 2002.

[10] D. Menasce, D. Barbara, and R. Dodge, "Reserving QoS of E-Commerce Sites through Self-Tuning : A Performance Model Approach", Proceedings of 3rd ACM-EC Conference, Florida, 2001.

[11] D. S. Moore, "Statistics Concepts and Controversies (the fifth edition)", W.H.Freeman and Company, 2001.

[12] T. Morals and D. Lorentz, "Oracle 9i : Database Reference, Release 2(9.2)", Oracle Corporation, 2001.

[13] J. S. Oh and S. H. Lee, "Resource Selection for Autonomic Database Tuning", Proceedings of IEEE International Workshop on Self-Managing Database Systems, pp.66-73, Tokyo, 2005.

[14] V. Signhal and A. J. Smith, "Analysis of Locking Behavior in Three Real Database Systems", The VLDB Journal, Vol.6, No.1, pp.40-52, 1997.

[15] D. E. Shasha, "Database Tuning : A Principled Approach", Prentice Hall PTR, 1992.

[16] TPC Benchmark C Specification (Revision 5.0), 2001, http://www.tpc.org/tpcc/default.asp

[17] TPC Benchmark W (Web Commerce) Specification (version 1.8), 2002. http://www.tpc.org/tpcw/default.asp

[18] G. Weikum, C. Hasse, A. Moenkeberg, and P. Zabback, "The COMFORT Automatic Tuning Project", Information Systems, Vol.19, No.5, pp.381-432, 1994.

[19] G. Weikum, A. C. Konig, A. Krasis, and M. Sinnewell, "Towards Self Tuning Memory Management of Data Servers", Bulletin of the Technical Committee on Data Engineering, IEEE Computer Society, Vol.22, No.2, pp.3-11, 1999.

[20] 박정식, 윤영석, "현대 통계학(제 4판), 다산 출판사, 2005.

[21] 오정석, 이상호, "데이터베이스 워크로드 분석: 실험적 연구", 정보처리학회 논문지, 제 11-D권, 4호, pp.747-754, 2004.

오 정 석

e-mail : dbstar@nate.com
1996년 서경대학교 정보처리학과(학사)
1998년 숭실대학교 대학원 컴퓨터학과 (석사)
2006년 숭실대학교 대학원 컴퓨터학과 (박사)
관심분야 : 메타 검색, 데이터베이스 시스템 튜닝 및 성능평가, 데이터베이스 워크로드

이 상 호

e-mail : shlee@compu.ssu.ac.kr
1984년 서울대학교 컴퓨터공학과(학사)
1986년 미국 노스웨스턴 대학교 전산학과 (석사)
1989년 미국 노스웨스턴 대학교 전산학과 (박사)
1990년~1992년 한국전자통신연구원 선임연구원
1992년~현재 숭실대학교 컴퓨터학부 교수
1999년~2000년 미국 George mason 대학교 교환 교수
관심분야 : 인터넷 데이타베이스, 데이타베이스 튜닝 및 성능평가, 웹 기술