D-018

データウエアハウスシステムの性能評価モデルの提案
The Benchmark Model for Datawarehouse Systems

高山　茂伸†　　佐藤　重雄†　　早川　孝之†　　白井　健治†
Shigenobu Takayama　Shigeo Sato　　Takayuki Hayakawa　Kenji Shirai

1．はじめに

データベースの性能を測定するベンチマークとしては、TPC[1]が一般的であるが、大規模なハードウェア構成が必要であり、またその複雑さゆえ様々なチューニング作業を要するなどいくつか問題がある。そのため、その他のベンチマークも研究されている[2]。また、TPC-H（TPC-R）はトランザクション的な処理をベースにした意思決定支援システムの性能評価であり、データウエアハウス(DWH)に特化したものではない。

本稿では、データウエアハウスシステムに特化した効果的な性能評価モデルを提案し、データウエアハウスシステムの処理特性を測定可能なベンチマークモデルを提案する。

2．既存のベンチマークの問題点

データベースのベンチマークの目的は、ある処理を行うアプリケーションプログラムに対して、幾つかのデータベースシステムを比較し、どのシステムが最も適切であるかを見極めることにある。しかし、意思決定支援システムのベンチマークとして最も一般的に用いられている TPC-H は、データベースベンダーや H/W ベンダーがハイエンドのスペックのマシンを用いて最高性能を競い合うという場になりつつあり、その結果から各ユーザがデータベースを選択するのは困難である。また、データベース技術者や SE が複数のデータベースシステムで実行しようと思うと、複雑なデータベースチューニングやシステムコンフィグレーションが必要なため容易ではないと言う問題がある。

3．DPA ベンチマークの提案

3.1　ベンチマーク作成の方針

データウエアハウスシステムにおいて、検索処理特性の評価を容易にするため、新しいベンチマークとして DPA（DWH Performance Analysis）ベンチマークの提案を行う。DWH を用いた様々な分析業務(流通業・小売業販売分析業務やサービス業・金融業顧客分析業務など)を調べた結果，以下の３つの処理が検索処理の特性をあらわしていることが判明した。本ベンチマークではこれらを評価することを作成方針とした。

（１）選択処理
選択列数、検索条件の内容、レコード選択率の性能への影響を調べるためにディスクからのデータの読み出し、および、検索条件の評価によるレコード選択処理の特性を評価する。

† 三菱電機（株）情報技術総合研究所
Shigenobu Takayama stakayam@isl.melco.co.jp
Shigeo Sato ssato@isl.melco.co.jp
Takayuki Hayakawa thaya@isl.melco.co.jp
Kenji Shirai shiraik@isl.melco.co.jp

表１　アクセスログの形式

項目	内容	タイプ
LOGDATE	アクセス時間	DATE
CIP	ユーザの IP アドレス	CHAR(16)
CNAME	ユーザ名	CHAR(16)
SNAME	Web サーバ名	CHAR(8)
SCOMP	Web サーバホスト名	CHAR(12)
SIP	Web サーバ IP アドレス	CHAR(16)
CMETHOD	HTTP メソッド	CHAR(8)
CREQ1	アクセスオブジェクト１	CHAR(80)
CREQ2	アクセスオブジェクト２	CHAR(100)
CQUERY	ASP で使用するデータ	CHAR(68)
HTTPSTAT	HTTP のステータス	CHAR(4)
WINSTAT	Win32 のステータス	CHAR(8)
SCBYTES	サーバが送信したバイト数	INTEGER(4)
CSBYTES	ユーザが送信したバイト数	INTEGER(4)
TIMTAKEN	データ送信時間	INTEGER(4)
PORT	ポート番号	INTEGER(4)
PROTVER	HTTP プロトコルバージョン	CHAR(8)
CAGENT	ユーザブラウザ情報	CHAR(140)
COOKIE1	COOKIE 情報１	CHAR(1)
COOKIE2	COOKIE 情報２	CHAR(24)
COOKIE3	COOKIE 情報３	CHAR(36)
REFERRER	ユーザ参照情報	CHAR(112)

（２）集計処理
レコード選択率、グループ化キーの内容の性能への影響を調べるために GROUP BY 句による集計処理、および、HAVING 句による条件設定処理の特性を評価する。

（３）ソート処理
ソート対象レコード数、ソートキーの内容の性能への影響を調べるために ORDER BY 句による結果のソート処理の特性を評価する。

3.2　ベンチマークの定義

近年データウエアハウスで取り上げられることが多く、上記の処理特性にも合致しどの分野の分析にも適用可能な、Web サーバに蓄積されたアクセスログに対する分析業務でベンチマークを構築する。アクセスログ（テーブル名：LOG）の形式を表１に示す。このアクセスログに対して実行されるビジネスモデルのいくつかの問合せを調べた結果、以下の４つの問合せを検索処理特性の評価として、ベンチマークで採用することとした。

これらの問合せ内容と検索処理特性との関係を表２に示す。問合せ 1 は全ての検索処理特性を含んだものである。

問合せ２は選択率が非常に低く、また比較処理に LIKE 述語を用いていることから、レコード選択処理の性能を測定することを想定したものである。問合せ３はキー数に着目して集計処理の性能を測定することを想定したものであり、問合せ４はキー長に着目して集計処理の性能を測定することを想定したものである。

（問合せ１）ページのアクセス数順を調べる
```
SELECT CREQ1, COUNT(*) AS FREQ  FROM LOG
WHERE CREQ1 <> '-'  GROUP BY CREQ1
ORDER BY FREQ DESC
```
（問合せ２）リンク元 URL 別訪問者数を調べる
```
SELECT REFERRER, COUNT(*) AS FREQ
FROM LOG WHERE REFERRER NOT LIKE '%melco%' AND
REFERRER NOT LIKE '%MELCO%' AND REFERRER NOT
LIKE '%Melco%' AND REFERRER <> '-' AND
REFERRER NOT LIKE '%ftp/%' AND REFERRER NOT
LIKE '%www/%'
GROUP BY REFERRER ORDER BY FREQ DESC
```
（問合せ３）時間毎のアクセス成功数を調べる
```
SELECT  TO_CHAR(T1.“HOUR”,
 ’YYYYMonDD HH24:MI’)
  “HOUR”,T1.HTTPSTAT, COUNT(*) FROM
 (SELECT TRUNC(LOGDATE,’HH24’) “HOUR”
,HTTPSTAT FROM LOG) T1
GROUP BY ROLLUP(T1.”HOUR”,T1.HTTPSTAT)
```
（問合せ４）リンク元ページ別訪問者数を調べる
```
SELECT REFERRER, CREQ1, COUNT(*) AS FREQ
FROM LOG WHERE CREQ1 <> '-'
GROUP BY REFERRER, CREQ1  ORDER BY FREQ DESC
```

表２　ベンチマークの問合せ内容と処理特性

	問合せ	1	2	3	4
選択	選択率	33%	0.39%	100%	33%
	検索条件	比較述語	LIKE 述語	—	比較述語
集計	キー長	80	112	20	192
	キー数	678	208	250 万	1593
ソート	キー項目	集合関数	集合関数	—	集合関数

４．ベンチマークの実装と評価

提案手法を用いたベンチマークの方法を以下に説明する。対象となるデータベースとしては、２つの商用データベース（商用 DB-A、商用 DB-B）を選択した。それぞれのデータベースに対しては、特別なチューニングは実施していない。本ベンチマークは、表１に示す LOG テーブルに対する、データロードと検索の２つのフェーズから構成される。

データロードは、300 万件のロード処理を２つの商用データベースに対して順次行った。データ量は約 2G バイトである。非定型な検索を想定して、インデックスの生成は実施しない。検索はそれぞれのデータベースにおいて、シングルユーザモードで LOG テーブルに対して、上で用意した SQL 文を１つづつ実行した。

２つの商用データベースに対して実行した検索結果を表３を用いて説明する。表３では、それぞれの商用データベースにおいて、問合せ 1 の処理に要した時間を１とした場合に、他の問合せの処理に要した時間の比率を記載している。

表３　ベンチマークの結果

問合せ	1	2	3	4
商用 DB-A	1.0	2.2	1.5	1.7
商用 DB-B	1.0	4.0	1.5	1.0

問合せ 1 は全ての検索処理特性が含まているので、それぞれのデータベースの性能を測定する指標とした。問合せ１に対する相対比で他の問合せの性能を測定し、処理特性を評価することとした。問合せ２の結果を見ると、DB-A は 2.2 であるのに対して、DB-B は 4.0 となっており、DB-A はレコード選択処理が得意であることがわかる。また、問合せ４の結果を見ると、DB-B が 1.0 であるのに対して DB-A は 1.7 であり、DB- B はキー長が長い場合の集計処理がより得意であることがわかる。

データベース選定時には絶対性能による評価も重要であるが、それとは別に、ここで提案した問合せによる検索処理特性を評価することで、データベースの拡張性やシステム構成を決めるための指標として使用することが可能である。例えば商用 DB-A であれば、キー長が長いためにオンメモリで集計処理ができていない可能性がありメモリの増設を検討する、商用 DB-B であればディスクからのデータの読み出しを高速化するためにディスクの増設を検討するなどが考えられる。

これらのことから、DPA ベンチマークが、DWH の検索処理特性の評価に有効であることがわかった。

５．おわりに

提案した DPA ベンチマークは、評価環境を容易に構築でき、DWH の検索処理特性を評価できること示した。

今後はさらに問合せの数を増やして、ソート性能を含めて、より詳細に検索処理特性を評価できるように充実させるとともに、データロードについても評価を実施していく。また、それぞれの処理における CPU 使用率、メモリ使用率やディスク転送量などの調査も合わせて行うことにより、より正確な性能特性を調査できるモデルの構築を行っていく。なお、本ベンチマークで用いたデータは公開しており、連絡いただければ提供可能である。

参考文献

[1] Transaction Processing Performance Council. http://www.tpc.org/

[2] K. Keeton and D. Patterson. "Towards a Simplified Database Workload for Computer Architecture Evaluations," presented at the Workshop on Workload Characterization, Austin, Texas, October 1999.