5D-2

集計処理並列化に関する評価

佐藤　重雄　　早川　孝之
三菱電機株式会社　情報技術総合研究所

１．はじめに

企業では、経営戦略立案のために日々の業務活動で収集したデータを DWH（Data Warehouse）に蓄積することが行われている。蓄積されるデータは、流通小売業における売上データや、コンピュータに対するアクセスログのように多種多様で、かつ、膨大な規模のデータである。このようなデータを有効に活用するためには、複雑な非定型処理を高速に実現し、迅速な分析／判断を可能とするシステムが求められている。

本稿では、スケーラブルインテリジェントストレージアーキテクチャ（SISA）[1]を、三菱電機のビジネスインテリジェンス向けシステム DIAPRISM に適用し、集計処理並列化による高速処理システムを実現した実装方式およびその効果について報告する。

２．集計処理並列化

今回、実装を行った集計処理並列システムの概要を図 1 に示す。本システムでは、データファイルからの読み出し、レコードの選択／射影処理、および、読み出したデータに対する部分的な集計処理を、異なる CPU で並列に実行する。集計処理は、各 CPU で実行する部分集計処理と、部分集計の結果をまとめる全体集計処理の二段階に分けて処理を行う。

実装方式の概略は以下の通りである。

- データの受け渡しはバッファ単位で行う。
- データファイルから Read したデータは入力バッファに読み込まれる。
- 選択／射影処理では、条件に合致するデータを入力バッファから処理バッファに、バッファを満たすまで転送する。
- 部分集計処理では、処理バッファ内のデータに対して集計処理を実施する。部分集計処理の結果、処理バッファ内に空きが生じた場合は、再び、入力バッファから条件に合致するデータを処理バッファに転送する。部分集計結果が処理バッファを満たすと、バッファを全体集計処理に転送する。
- 全体集計処理では、各 CPU での処理の結果得られたバッファから全体集計を行う。

本システムでは、集計処理並列化により、複数 CPU の有効活用、全体集計処理での処理対象データ削減を実現し、大規模データに対する集計処理性能の高速化を図ること目的としている。

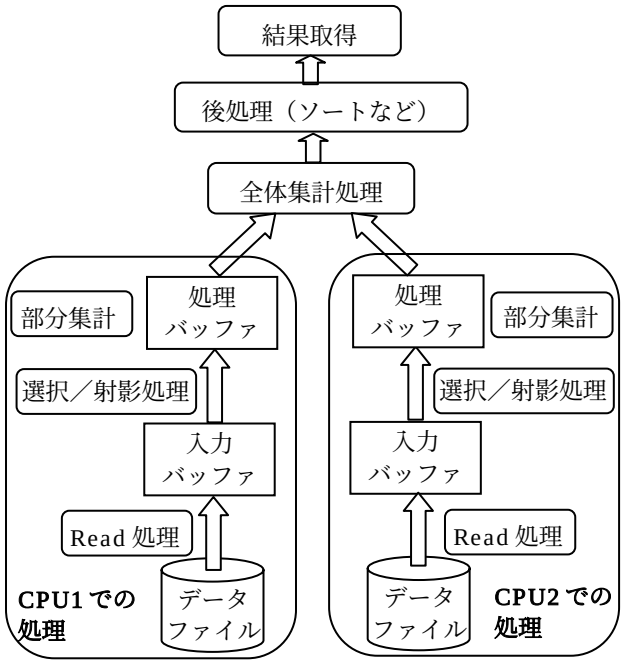

図 1．集計処理並列化の概要図

３．評価結果

3.1 評価内容

集計処理並列化の効果を、表 1 に示す評価環境で、表 2 に示すモデルを使用して評価した。

モデル 1 では、問合せの特性に応じた効果を評価し、モデル 2 ではレコード選択率を変化させた場合の特性についての評価を行った。

表 1．評価環境

CPU	Intel XeonMP(2.20GHz)×4
Memory	4GB
Disk	73.2GB(Ultra320SCSI, 15000rpm)×5
OS	WindowsServer2003 Standard

表 2．評価モデル

	レコード長	件数	問合せ
モデル 1	708Bytes	10 億件	8 種類
モデル 2	1068Bytes	2 億件	選択率変化

Evaluation of Parallel Processing for Aggregation
Shigeo Sato, Takayuki Hayakawa
Mitsubishi Electric Corporation

3.2 評価結果と考察

図 2、図 3 に、集計処理並列化の効果を示す。効果の指標には、問合せ処理の実行時間性能を用い、集計処理並列化を行わない場合の性能に対する性能比で表す。

全体集計処理が問合せ処理全体に占める割合が多くなると、集計処理並列化の効果が見込みにくい。特にレコード選択率が高い場合は、大量のデータに対する集計処理が必要であり、部分集計により、全体集計で処理対象とするデータがどれだけ削減されるかが性能向上のポイントとなる。図 4、図 5 に、全体集計処理対象バッファ数が、集計処理並列化を行うことにより、どの程度削減されたかを示す。ここでは、全体集計処理の対象となるバッファ数を、集計処理並列化を行わない場合に対する相対比で示す。

本評価結果から以下のことが言える。

- モデル 1 では、集計処理並列化の効果には、ばらつきが見られる。これは、集計処理並列化の実装方式が、問合せの内容とデータ特性に依存するためである。

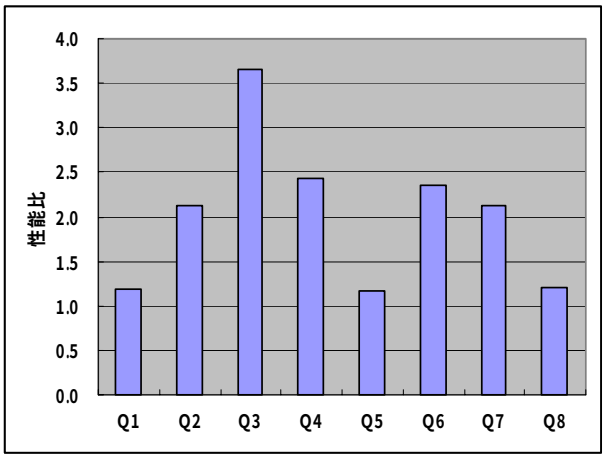

図 2． 集計処理並列化の効果（モデル 1）

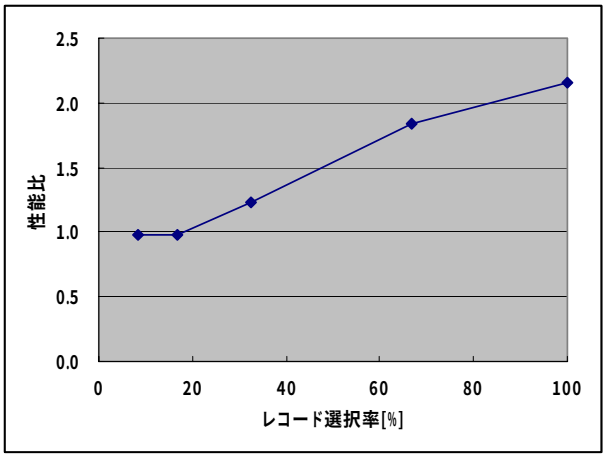

図 3．集計処理並列化の効果（モデル 2）

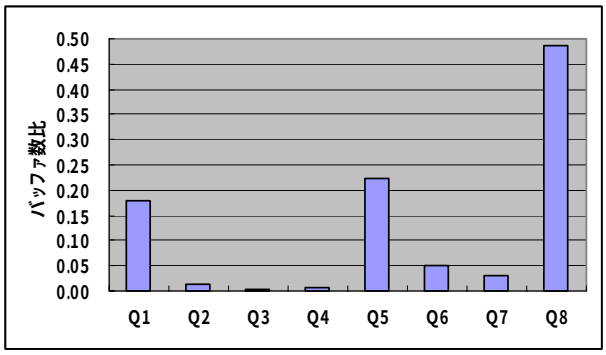

図 4．全体集計処理のバッファ数比(モデル 1)

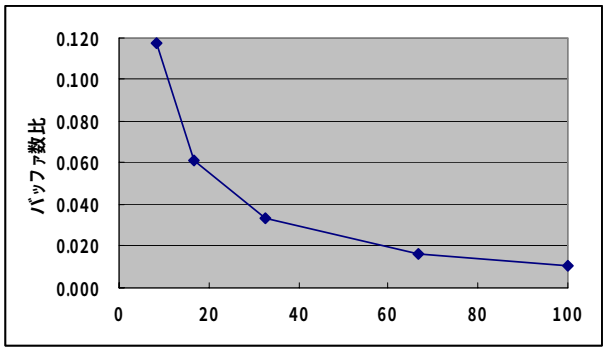

図 5．全体集計処理のバッファ数比(モデル 2)

- 効果の少ない問合せ（Q1、Q5、Q8）は、副問合せの集計処理を含み、最初の集計処理結果のサイズが大きい（集計キー数が多い）という特性がある。この場合は、全体集計処理の対象となるバッファ数の削減が小さく（図 4）、部分集計の効果は小さい。
- 集計キーの種類が少なく、部分集計により全体集計処理対象バッファ数の大幅な削減が実現できる場合は、2 倍以上の効果が得られた（Q2、Q3、Q4、Q6、Q7）。
- レコード選択率だけが異なる問合せでは、レコード選択率が高いほど、部分集計の効果は大きい（図 3）．これは、選択率が高いほど、全体集計対象バッファ数の削減程度が大きくなることからも分かる（図 5）。

4．おわりに

集計処理並列化による大規模データの検索性能の効果について評価を行った。集計キーの種類が多く、全体集計処理がネックとなる場合の性能向上が、今後の課題として挙げられる。

参考文献

[1] 郡，他、”検索機能を備えたストレージシステムによる大規模並列全文検索”，信学技報、CPSY-2002-47、Aug 2002