1D-3

大規模ログデータ分析エンジンの考察

佐藤　重雄　　東　辰輔
三菱電機株式会社　情報技術総合研究所

１．はじめに

近年、企業では、情報漏洩や不正アクセス防止などの情報セキュリティ対策、あるいは、内部統制に関する法整備への対応を目的として、様々な種類の大量のログの蓄積および分析処理が行われる傾向にある。大規模ログデータの分析システムでは、分析処理性能の観点から、目的や用途に応じて、蓄積されたログの一部を使用した分析専用のデータベースを構築することが行われている。

本稿では、ログ分析用データベースエンジンに必要な要件を定義し、要件の一つであるログデータとマスタデータの結合処理の評価結果について報告する。

２．ログ分析システム

大規模ログ分析システムの構成例を図 1 に示す。本システムは、以下の二つのデータベースから構成される。

(1) ログ蓄積用データベース

証拠保全などを目的として、ログをそのままの形式で蓄積保存するデータベースである。多様なログへの対応、高速検索・高速蓄積、ディスクの利用効率向上などが必要とされ、これらの要件を実現するログ専用データベース[1]が提案されている。

(2) ログ分析用データベース

分析の目的・用途に応じて、ログ蓄積用データベースから、抽出・変換処理を行って生成されるデータベースである。ログ蓄積用データベースには、発生したログがそのまま格納されるため、一般には非常に大規模なデータが格納される。ログ分析処理では、目的により、使用するログの範囲が異なり、通常は、一部のレコードや項目のみを使用した処理が行われるため、データマートの位置づけとなる分析専用データベースを構築することが、迅速な分析処理には効果的である。

このように、上記二つのデータベースは構築の目的が異なるため、それぞれの目的に応じた機能を有する必要がある。

A Study of Analysis Engine for Large Scale Log Systems
Shigeo Sato, Shinsuke Azuma
Mitsubishi Electric Corporation

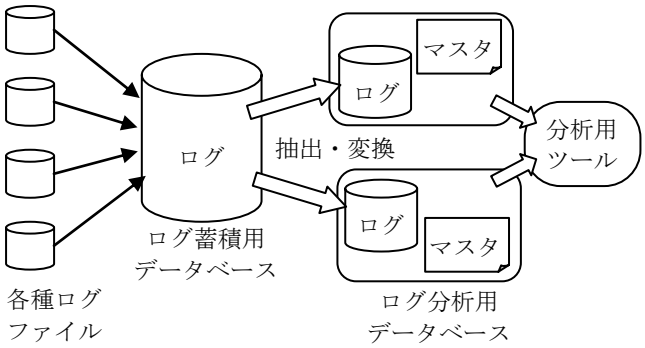

図 1. 大規模ログ分析システムの構成図

３．ログ分析用データベースの要件

ログ分析用データベースエンジンに必要とされる要件を以下に挙げる。

(1) 高速な構築処理

1 日に発生するログは、1 テラバイトを超える場合もあり、発生したログを短期間で分析用データベースに反映させるためには、高速なデータロード処理性能が要求される。

また、蓄積用データベースに比較して、データ規模は小さくなる傾向にあるが、データ圧縮などによるディスクの利用率向上が望まれる。

(2) 分析用ツールとのインタフェース

分析用ツールを使用した分析処理を行うため、標準 I/F（SQL、ODBC、JDBC など）のサポートが必要となる。

(3) ログデータの高速検索処理

ログデータに対する処理は、追加、および、検索・集計処理のみであり、更新などのトランザクション処理は発生しない。そのため、大規模データに対する検索・集計処理に特化した性能が要求される。

(4) ログデータとマスタデータの結合処理

分析処理では、単独のログデータに対する検索・集計処理に加えて、ログデータとは別に用意したマスタデータ（例：顧客マスタ）と結合した処理も必要となる。ログデータは、追加のみが発生するが、マスタデータは、必要に応じて追加・変更・削除が発生する。ログデータとマスタデータはその特性が異なるため、それぞれの特性に応じた処理方式が必要とされる。

４．ログ分析用データベースの評価

統合ログ収集・分析システム LogAuditor(*1)の LogAuditor/AQL をログ分析用データベースエンジンとしたシステムの評価を実施した。本システムは、SQL、ODBC、JDBC の汎用 I/F の提供、データ圧縮によるディスク利用率の向上を実現している。また、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）[2]の適用、集計処理の並列実行[3]により、CPU 数に比例したスケーラビリティを実現している。

本稿では、ログデータとマスタデータの結合処理の評価結果について報告する。

ログデータとマスタデータの結合処理は、以下のステップで実現する。

(Step1) ディスク上のマスタデータからデータを読み出し、メモリ上に展開
(Step2) メモリ上に結合処理用の索引を生成
(Step3) 索引を使用してログデータとの結合処理を実行

本システムで用いた結合処理用の索引は、図 2 に示すように、ハッシュと 2 分木で構成される。

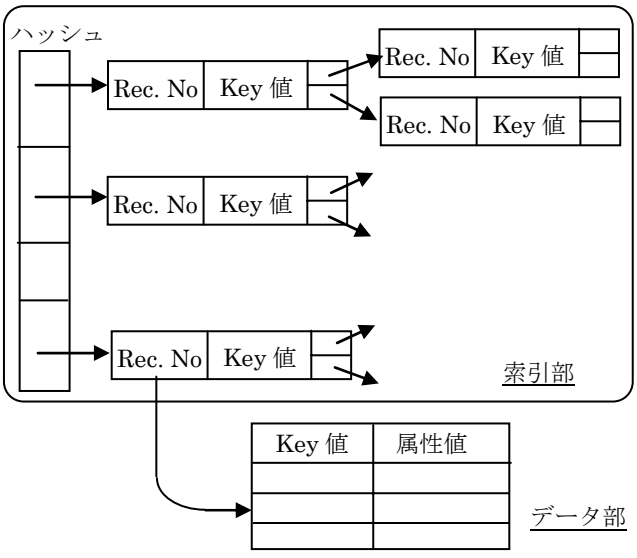

図 2. 結合処理用索引の構造

上記索引を用いた結合処理時間（結合処理の Step3）の特性を図 3 に示す。図 3 は、マスタデータ件数が 1 万件～1000 万件の 6 パターンに対して、ログデータ件数を変化させた場合の結合処理時間を測定した結果である。図 3 より、本システムの索引を用いた結合処理時間は、ログデータ件数にほぼ比例することが確認された。

また、結合前処理（Step1、Step2）の時間が全体処理に占める割合と、マスタデータ件数の関係を図 4 に示す。結合前処理時間は、マスタデータサイズに依存して増加するため、マスタデータ件数が多く（100 万件以上）、ログデータ件数が比較的少なく（1 億件以下）結合処理負荷が小さい場合には、結合前処理が全体処理に占める割合が大きくなり、全体の処理性能に影響を及ぼす結果となる。

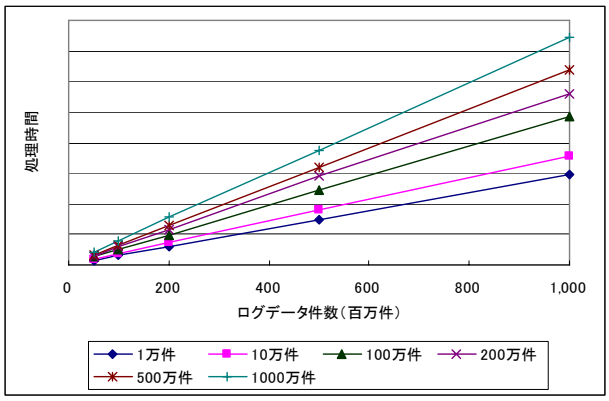

図 3. 結合処理時間の特性

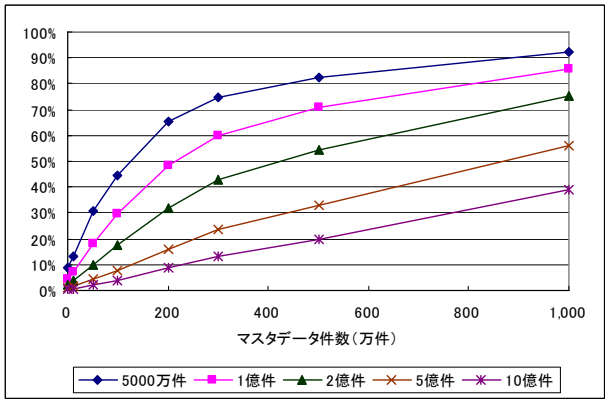

図 4. 結合前処理の全体処理に占める割合

５．おわりに

大規模ログデータ分析エンジンに必要な要件を定義し、要件の一つであるログデータとマスタデータの結合処理の評価を行った。結合処理性能向上のためには、結合前処理を検索処理の事前に行う方式などが有効と考えられる。

参考文献

[1] 中村、他、大規模ログデータベースの実現、第 68 回情報処理学会全国大会 講演論文集 (3) 29-30、2006
[2] 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
[3] 佐藤、他、集計処理並列化に関する評価、第 68 回情報処理学会全国大会 講演論文集 (3) 59-60、2006

(*1) LogAuditor：様々な形式のログデータの収集・蓄積・分析を可能とする製品。開発・販売は三菱電機インフォメーションテクノロジー株式会社。