1D-1

大規模ログデータベースの評価

竹内 丈志　　山岸 義徳　　中村　隆顕　　郡 光則
三菱電機株式会社 情報技術総合研究所

1. はじめに

近年、情報セキュリティ分野を中心に、問題発生時の証拠保全などを目的として、ログを保存・分析する動きが進んでいる。しかし、従来のデータベース管理システムは、ログの効率的な管理に必ずしも適していなかった。そこで、ログの管理における従来の課題を解決し、効率的なログの蓄積・管理を可能とする、ログ専用のデータベース管理システム”ログ DB”を開発した[1]。

ログ DB は、多様なログへの対応、高速蓄積・高速検索、ディスクの利用効率向上を目的として、設計されている。本稿では、このログ DB の設計の妥当性を検証し、その有効性を確認したので、報告する。

2. 評価方針

2.1. 評価対象ログ

多様なログへの対応について確認するため、2 種類の形式の異なるログを用意した。表 1に評価対象ログのデータ内容を示す。

1 つは、セキュリティログである。これは、PC の操作履歴を記録したもので、CSV 形式となっている。

もう 1 つは、メールである。これは、ヘッダ情報や本文を含む、メール本体のファイル全体を指す。

表 1:評価対象ログ

	セキュリティログ	メール
データサイズ (MB)	4,105.150	2,111.850
件数 (件)	6,075,161	907,700
平均レコード長 (Bytes/件)	708.551	2,439.606

なお、検索性能では、検索結果がセキュリティログの場合は全件の 0.012%(7,240 件/6,075,161 件)、メール本文の場合は全件の 0.2%(1,900 件/907,700 件)、となる検索条件を用意した。これは、数百万～数千万件の評価対象ログから、数千件まで検索結果が十分に絞り込まれる状況を想定して、設定したものである。

2.2. スケーラビリティ

ログ DB が動作基盤として用いている高速処理技術(SISA)[2]に含まれる並列処理技術に着目し、それの適用による効果を検証するため、スケーラビリティ評価を行った。具体的には、1 台のサーバ上での SMP による並列処理において、その CPU 数を変化させたときの、蓄積・検索性能の変化を調査した。ここでいう蓄積・検索性能とは、速度性能のことを指す。

2.3. 圧縮率

ディスクの利用効率について調査するため、蓄積時のログデータの圧縮率を評価した。圧縮率は以下に示す式で表される。

$$\text{圧縮率} = \frac{(\text{元のデータサイズ} - \text{圧縮後のデータサイズ})}{\text{元のデータサイズ}}$$

3. 評価システム構成

評価システムの構成は表 2に示す通りである。

なお、ログデータは 8 個のディスクに分けて蓄積した。

また、CPU は Hyper-Threading 対応であり、1CPU 当たり 2 個の圧縮処理スレッドまたは伸張・照合処理スレッドにより、並列処理を行う。

表 2:評価システム構成

OS	Windows 2003 Server x64 Enterprise Edition
CPU	Intel Xeon MP 3.66GHz x 4
Memory	15.9 GB
HDD	個数:10 個, 回転速度:15000rpm, キャッシュ:8MB, 接続形態:Ultra SCSI 320

4. 測定結果

図 1および図 2に、蓄積性能と検索性能を示す。

The Evaluation of the Large-Scale Log Database
Takeshi Takeuchi, Yoshinori Yamagishi, Takaaki Nakamura, Mitsunori Kori
Information Technology R&D Center, Mitsubishi Electric Corporation.

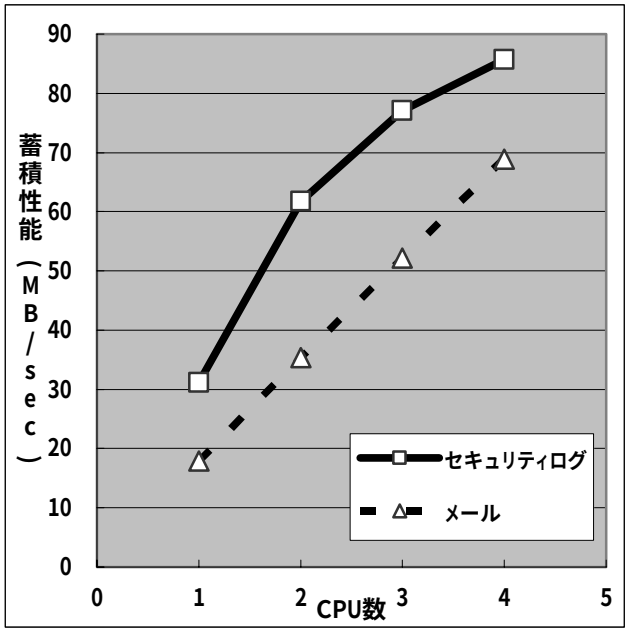

図 1:蓄積性能

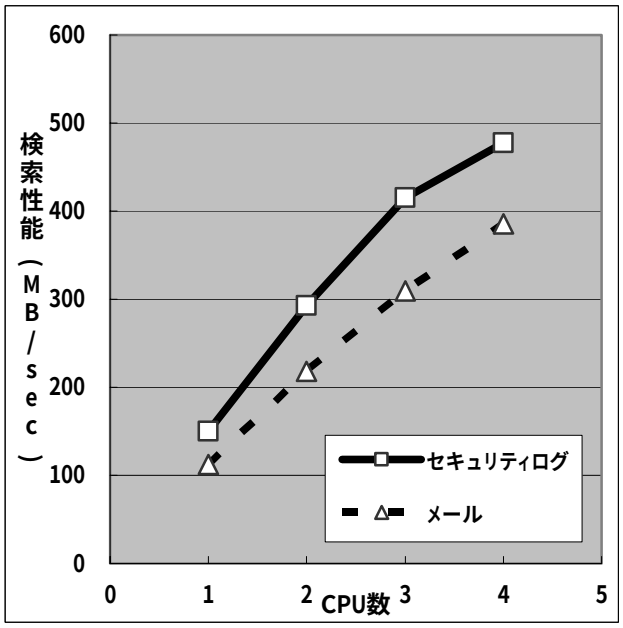

図 2:検索性能

また、表 3は圧縮率を示したものである。

表 3:圧縮率

	圧縮率
セキュリティログ	94.74 %
メール本文	74.35 %

5. 考察

5.1. ログ別の性能に対する評価

5.1.1. 蓄積性能についての評価

図 1より、セキュリティログに対する性能の方が、メールに対する性能よりも、高い性能を示していることがわかる。いずれの評価対象ログについても、CPU 数の増加に伴って性能を上げていることから、この性能の差は、ログ DB で蓄積時に並列化されている部分の性能差と考えられる。

つまり、評価対象ログの圧縮率の差が影響し、圧縮率の高いデータほど圧縮処理単位が高い性能を示すことが、評価対象ログによる性能差の原因と推定される。

5.1.2. 検索性能についての評価

図 2からセキュリティログに対する性能の方が、メールに対する性能より高い性能を示していることがわかる。いずれの評価対象ログについても、CPU 数の増加に伴って性能を上げていることから、この性能の差は、蓄積性能と同じように、ログ DB で検索時に並列化されている部分の性能差と考えられる。

つまり、評価対象ログの圧縮率の差が影響し、圧縮率の高いデータほど伸張・照合処理単位が高い性能を示すことが、評価対象ログによる性能差の原因と推定される。

5.2. スケーラビリティの評価

5.2.1. 蓄積性能について

図 1から、2 種類の評価対象ログのいずれも CPU 数増加に伴った性能向上を示しており、並列処理の効果を確認できた。

5.2.2. 検索性能について

図 2から、2 種類の評価対象ログのいずれも CPU 数増加に伴った性能向上を示しており、蓄積性能と同じく、並列処理の効果を確認できた。

5.3. 圧縮率の評価

ログの形式により圧縮率は変化するものの、圧縮後のデータサイズは、セキュリティログで約 1/20、メール本文で約 1/4 程度になり、ログの蓄積に必要なディスク使用量が抑えられることを確認できた。

6. まとめ

本稿では、ログ専用データベース管理システムとして開発したログ DB に対し、その特長である多様なログへの対応、高速蓄積・高速検索、ディスクの利用効率向上について評価を行い、その有効性を確認した。

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，1D-2，2006.

[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，Vol.102，No.276，pp.41-46，2002.