特集論文

高速集計検索エンジンとセンサデータベースへの応用

山岸義徳*
平井規郎*
西村達夫**

High Performance Aggregate Search Engine and Application for Sensor Database

Yoshinori Yamagishi, Norio Hirai, Tatsuo Nishimura

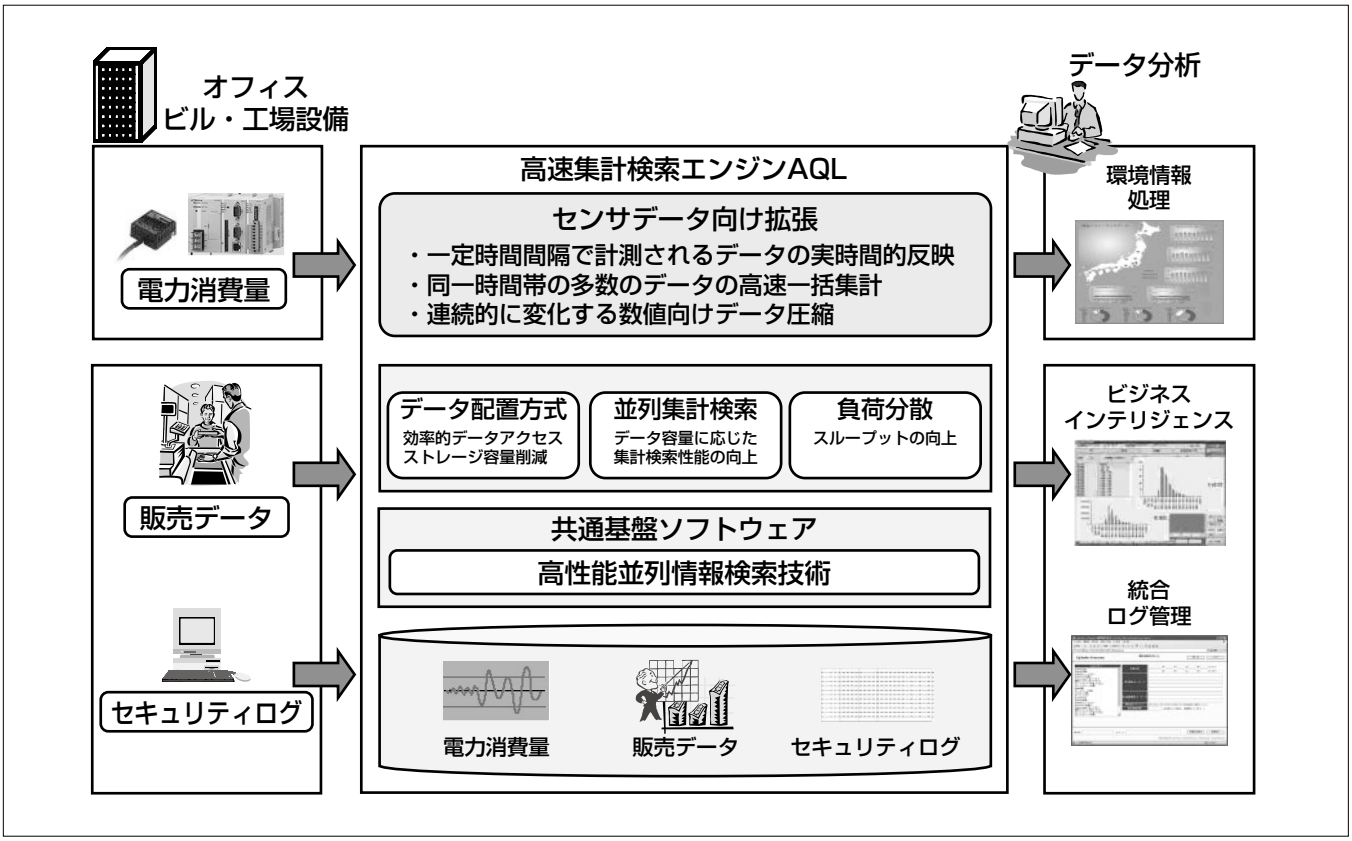
要　旨

情報技術の発展によって，オフィス，ビル・工場設備などに設置された情報機器が出力するデータは増加の一途をたどっている。環境情報処理，ビジネスインテリジェンス，統合ログ管理などの分野では，日々発生する大量のデータを活用した企業の経営判断や業務改善などのためのデータ分析が行われており，データ分析の基本処理である集計検索処理の高速化の必要性がますます高まっている。

三菱電機が開発した高速集計検索エンジンAQL（Analytical Query Language）は，三菱電機インフォメーションテクノロジー㈱（MDIT）のデータ分析プラットフォーム“DIAPRISM”で，販売データの分析などに活用されている。また，MDITの統合ログ管理ソリューション“LogAuditor（注1）”では各種セキュリティログの分析などに活用されている。

近年では多数のセンサによって計測されたセンサデータの応用が広がっている。例えば，MDITの環境経営推進ソリューション“MELGREEN”は，AQLを利用して電力消費量などのセンサデータを集計し，省エネルギー対策に役立てている。センサデータには，一定時間間隔で計測されるデータの実時間的反映，同一時間帯の多数のデータの一括集計，連続的に変化する数値などビジネスインテリジェンスや統合ログ管理のデータとは異なる特徴がある。このため当社では，AQLをセンサデータ向けに拡張し，センサデータベースとして活用するための研究開発に取り組んでいる。

（注1）　LogAuditorは，三菱電機インフォメーションテクノロジー㈱の登録商標である。



高速集計検索エンジンAQLの概念図

大規模・高速なデータ分析処理を可能とする高速集計検索エンジンAQLは，ビジネスインテリジェンスや統合ログ管理で広く活用されてきた。近年では，環境情報処理などに代表されるセンサデータの応用が広がっており，当社ではAQLをセンサデータ向けに拡張するための研究開発に取り組んでいる。

*情報技術総合研究所　**本社

1. まえがき

情報技術の発展によって，様々な情報機器が日々出力するデータは増加の一途をたどっている。1日当たり数億件のログや，数万に及ぶセンサによって計測されたデータなどの膨大なデータの分析を実現するために，集計検索処理の高速化の必要性が高まっている。

本稿では，当社の開発した高速集計検索エンジンAQLと，AQLを応用したセンサデータベースへの取り組みについて述べる。

2. 高速集計検索エンジンAQLの応用分野

高速集計検索エンジンAQLは，追記型データベースの高速処理技術“高性能並列情報検索技術”[1]に基づいて，多次元集計処理をはじめとする集計検索処理を高速に実行することを目標に開発された。AQLは，標準的なデータベース問い合わせ言語であるSQL(Structured Query Language)[2]に準じる問い合わせインタフェースを提供するデータベース管理システムとして動作し，集計検索処理だけでなく，対象データを絞り込む選択／射影などデータ操作の基本機能を備えている。

AQLは，MDITのデータ分析プラットフォーム“DIAPRISM”，統合ログ管理ソリューション“LogAuditor”，環境経営推進ソリューション“MELGREEN”で活用されている。

このAQLが活用される分野と，そこで使われるデータの特徴について述べる。

2.1 ビジネスインテリジェンス／統合ログ管理

近年，販売データなどをデータウェアハウスと呼ぶ分析専用のデータベースに取り込んで，企業の経営判断や業務改善に活用するビジネスインテリジェンスと呼ばれるデータ分析が広く行われている。例えば，商品を販売するごとに生成される販売データをデータベースに格納し，日付，地域，商品など複数の視点から集計する，いわゆる多次元集計によってデータの傾向を把握する。

また，情報漏洩(ろうえい)事故の多発や日本版SOX法(金融商品取引法)の施行を契機に，様々な情報システムの生成するログを保存する統合ログ管理システムが構築されている。例えば，入退室やパソコンの操作などを行うごとに生成されるログをデータベースに格納し，日付，場所，人などの視点による多次元集計によって異常の有無を確認する。

これらの応用で発生するデータは，イベントごとに発生し，時間とともに蓄積される追記型のデータである。このような販売データやログなどのデータを総称して“ログデータ”と呼ぶことにする。AQLは追記型のデータの高速集計処理に適した構造を備えており，ログデータの高速集計検索に広く利用されてきた。

2.2 環境情報処理

ネットワーク化の進展によって，全国のビル・工場からネットワークを介してセンサデータを収集できる環境が整いつつある。こうした中，改正省エネ法(エネルギーの使用の合理化に関する法律)の施行などを背景に，企業の環境対策として，電力消費量や温度などのセンサデータが収集されるようになってきた。多数のセンサから一定の時間間隔で電力や温度などを計測し，時間帯，拠点，用途などの視点による多次元集計を行うことによって，電力消費量の傾向把握を行い，省エネルギー対策に役立てる。

このように一定の時間間隔で計測され，時間とともに蓄積されるデータを“センサデータ”と呼ぶことにする。センサデータは，ログデータと同様に追記型のデータであるが，次のような相違点がある(**表1**，**図1**)。

(1) 発生間隔

ログデータはイベント発生時に生成されるが，センサデータはほぼ一定の計測間隔で発生する。また，センサデータは，到着したデータを比較的短い遅延で反映する必要がある場合が多い。

(2) 項目の種類

ログデータでは，レコード内に発生したイベントに関する時刻，場所，人，操作内容など，様々なデータ項目が含まれることが多い。一方センサデータでは，データ項目の種類は少ないが，同一時間帯に対応する多数(数百～数万)のデータ項目が発生する。

表1．ログデータとセンサデータの比較

	ログデータ	センサデータ
応用例	ビジネスインテリジェンス 統合ログ管理	環境情報処理
発生間隔	イベントごとに発生するため，間隔は一定しない	一定の間隔で発生
項目の種類	1レコード中に様々な種類のデータ項目が含まれる	数百～数万の同種のデータ項目が発生
値の範囲	商品コードやユーザーIDなど限定された種類の値	時間経過と共に連続的に変化する数値

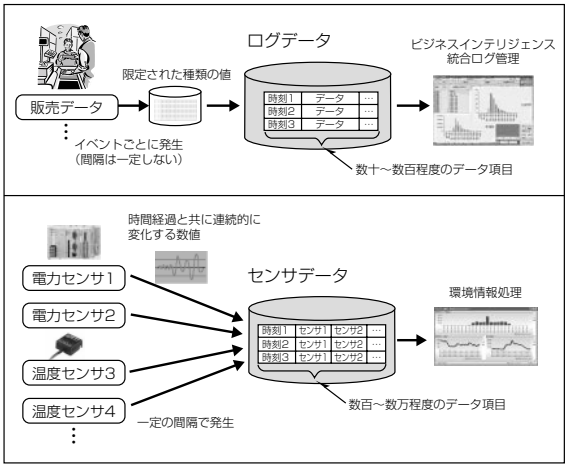


図1．ログデータとセンサデータ

(3) 値の範囲

ログデータに含まれるデータ項目には，商品コードやユーザーIDなど，限られた種類の値を取る場合が多い。これに対し，センサデータは時間とともに変化する連続的な数値である場合が多い。

3. 高速集計検索エンジンAQLの高速化技術

次に，AQLの主な高速化技術であるデータ配置方式，並列集計検索，負荷分散について述べる。

3.1 データ配置方式

大量データの一括集計検索処理では，メモリ上へのデータのキャッシュや，インデックスの効果が小さく，ストレージアクセスが性能のボトルネックとなりやすい。

AQLでは，ブロック化トランスポーズドデータ配置方式とマルチストリームデータ圧縮によって，集計検索処理におけるストレージアクセスの高速化とストレージ容量の削減を実現した。データ圧縮の符号化方式として，ログデータに適したRID(Run-length，Index and Difference)符号化方式[3]を採用し，高圧縮率と高速処理を両立した。データ圧縮によるストレージ容量と集計検索時間の削減効果を**図2**，**図3**に示す。典型的な3種類の販売データを対象に6種類の集計検索の問い合わせを実行し，ストレージ容量が約2～14%，集計検索の処理時間が約5～23%に削減されている。

3.2 並列集計検索

データ容量の増加に対応した高速処理を実現するためには，多数のプロセッサ，ディスク等のハードウェア資源を効率的に利用する必要がある。

AQLでは，集計検索処理の並列化によって，プロセッサの処理能力に応じた集計検索速度を実現する。**図4**に，ログ10億件を対象とした年月をキーとする全件集計における，プロセッサ数と処理速度の関係を示す。プロセッサ数にほぼ比例した速度性能が得られている。

3.3 負荷分散

データベースへのロードと集計検索を同時に行うと，高速な応答速度が要求される集計検索の速度が低下する。従来はロード処理を夜間バッチで実行し，昼間実行される集計検索と時間帯を分けることによってこの問題を回避することが多かった。しかしこのような運用では，データが反映されるまでに最大1日の遅延が生じ，タイムリーなデータ分析を行う上で課題があった。

AQLでは，ロード処理と集計検索処理をそれぞれロード専用サーバと集計検索専用サーバに割り付けて処理する負荷分散によって，データロード中の高速な集計検索を実現した。追記型データベースの特長を生かし，ロード専用サーバに追加された差分データをそのままの形式でデータベースから取り出し，集計検索専用サーバのデータベースに配布，追加することによって，ロード専用サーバと集計検索専用サーバの高速なデータ同期を実現している(**図5**)。

4. センサデータベースへの取り組み

当社では，ここまでに述べたような技術によって追記型データの高速集計を実現したAQLの特長を生かしつつ，センサデータの処理向けの拡張を行ったセンサデータベースの研究開発に取り組んでいる。次に，センサデータベースにおける技術課題とそれに対する取り組みについて示す。

4.1 ストレージアクセス効率を高めるブロック最適化

AQLのような追記型のデータベースでは，短い間隔で到着するセンサデータを実時間的に反映するためには，到

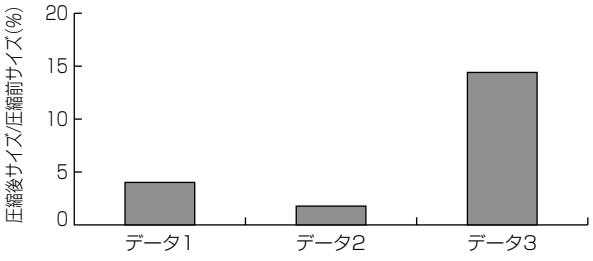


図2．データ圧縮によるストレージ容量の削減効果

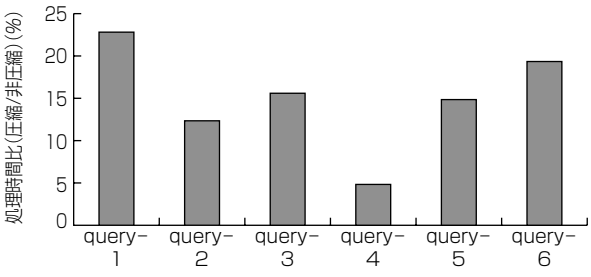


図3．データ圧縮による集計検索時間の削減効果

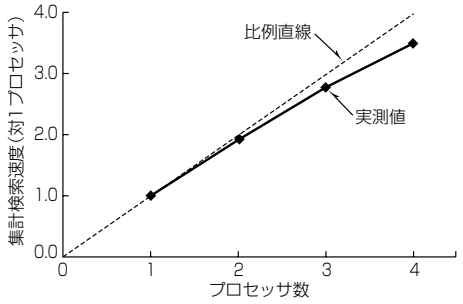


図4．集計検索処理のスケーラビリティ

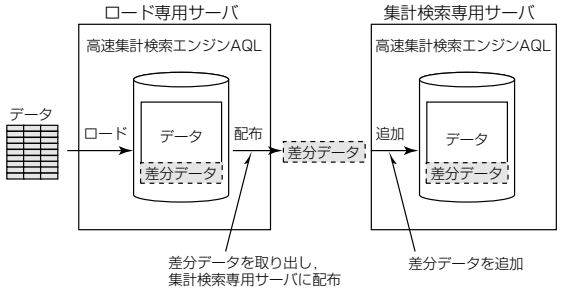


図5．AQLのデータ同期方式

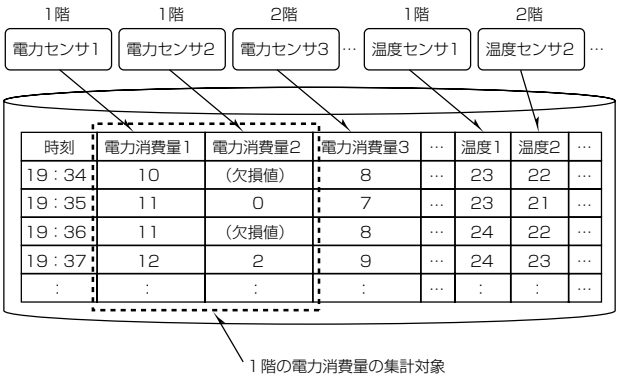


図6．センサデータベースの論理構造

着したデータを細かい単位（ブロック）でストレージに書き込む必要がある。しかし，細かいブロックによるデータの入出力はストレージのアクセス効率を低下させ，データのロード及び集計検索の速度を低下させる要因になる。

この課題を解決するため，ブロック最適化によってブロックの細分化を防止し，大量のセンサデータの高速なロードと集計検索を実現する。

4.2 高速な一括集計を可能とするデータの格納方式

同一時間帯に発生した数万点に及ぶセンサデータを拠点や用途別に集計するには，同一時間帯のセンサデータをまとめて一つのレコードに配置すると効率的である（**図6**）。AQLのブロック化トランスポーズドデータ配置方式は，必要なデータ項目だけを選択して高速に読み出し可能なため，このような形式のデータの格納や集計検索に適している。しかし，ネットワークを経由して収集されるセンサデータの到着のタイミングはセンサごとに異なり，またデータの到着が遅れる場合や欠損する場合もある。

この課題を解決するため，データ到着の遅延や欠損を考慮したデータの格納を行い，効率的なデータロードと高速な一括集計を実現する。

4.3 連続的に変化する数値データ向けのデータ圧縮

AQLのデータ圧縮におけるRID符号化方式は，データ項目ごとに限られた種類のデータが出現することが多いというログデータの特性を利用している。一方，センサデータは，時間経過とともにデータの値が連続的に変化するという，ログデータとは異なる特性を持つ。

この課題を解決するため，センサデータ向けのデータの変化量を考慮した符号化方式を併用し，高圧縮率と高速処理を実現する。

5. 環境情報処理への適用事例

AQLの適用事例である環境経営ソリューション“MELGREEN”[4]について述べる。

図7に示すように，ビル・工場の電力，空調，照明設備などが逐次出力する電力消費量，温湿度，照度などの環境情報データをネットワーク経由で収集し，データセンタにある環境統合データベースで一元管理する。環境・省エネルギーのPDCA（Plan Do Check Action）サイクルを回す持続的な取り組みを行うためには，膨大なデータを長期間にわたり保存する必要がある。例えば，1万個のセンサを対象に1分間隔で収集したデータを5年間保存する場合，そのデータ件数は約263億件にも及び，高速な集計・分析と効率的な蓄積が必要となる。

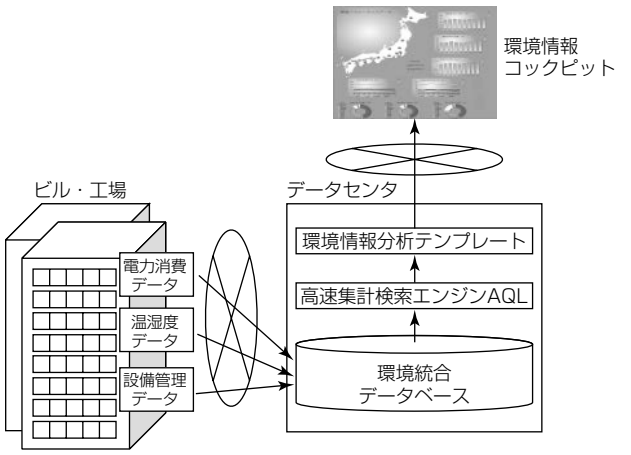


図7．環境経営推進ソリューション“MELGREEN”

環境統合データベースにAQLを適用することによって，テラバイト級の大規模データに対して，事業所別，ビル別，フロア別の電力消費量や温度変化など，場所ごとのエネルギー消費量の詳細把握や分析，削減に向けた対策立案などを可能とする。さらに，環境情報コックピットによっていろいろな角度から必要な情報をチャートやグラフによって表示し，効率的に意思決定を支援する。

6. むすび

大規模・高速なデータ分析処理を可能とする高速集計検索エンジンAQLと，そのセンサデータベースへの応用について述べた。データ量の増大とデータ分析の即時性に対する要求は当分続くものと予想されており，高速集計検索エンジンの活用される分野は今後ますます広がるものと考えている。

参考文献

(1) 郡　光則，ほか：高性能並列情報検索技術，三菱電機技報，**83**，No.12，705～708（2009）
(2) データベース言語SQL（JIS X3005-1995），㈶日本規格協会（1995）
(3) 郡　光則：データウェアハウス向け高性能データ圧縮方式，情報処理学会論文誌，**47**，No.SIG13，TOD31（2006）
(4) 松井陽子，ほか：省エネルギーのPDCAの管理基盤環境経営推進ソリューション“MELGREEN”，三菱電機技報，**83**，No.7，413～416（2009）