2C-1

センサデータ向けデータ圧縮ロード方式

立床　雅司†　　山岸　義徳†　　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

周期的に発生する大量のセンサデータに対してデータ圧縮機能を備えたデータベースに格納するデータ圧縮ロード方式について述べる．本方式では，センサデータの遅延時間およびデータベースのデータ圧縮処理の同時実行数の制約を考慮することにより，遅延時間短縮のための負荷分散を実現する．電力データを用いた評価により，本方式の有効性を検証した．

2　背景

製造現場や社会インフラにおいて日々発生するセンサデータを長期間蓄積し，異常検知や生産効率改善，省エネルギーに利用する試みがなされている．このようなセンサデータは，データ圧縮処理を備えたデータベースに格納することでストレージコストが削減できる．我々が開発しているセンサデータベース[1]上に実装したセンサデータロード方式[2]（以降，従来方式と呼ぶ）では，データロード時にセンサデータ圧縮方式[3]を適用してきた．

3　センサデータロードにおける制約

多数の収集周期が存在しデータ量が膨大となる場合，図 1に示すようなデータ圧縮処理の競合によるセンサデータの遅延時間超過が発生する．遅延時間超過の原因となるセンサデータロードの制約を収集周期，遅延時間，並列実行数の観点から述べる．

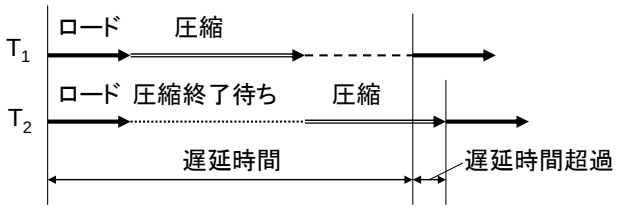

図 1 センサデータロードの制約による課題

3.1　センサデータの収集周期

センサデータは計測対象や重要度により複数の収集周期に分けられる．複数の収集周期を 1 つのテーブルにまとめる場合，周期の短いセンサデータの欠落がないように格納すると，周期の長いセンサデータに欠損値が格納され疎なテーブルとなる．周期の長いセンサデータに欠損値を格納しないようにすると，周期の短いセンサデータが欠落する．これらの理由から，周期が同一のセンサデータを 1 テーブルにまとめる．そのため，製造現場のような大量のセンサデータを扱う場合，複数のテーブルに分割して格納する必要がある．

3.2　遅延時間

遅延時間とは，データの登録からデータが参照可能となるまでの時間を指す．センサデータでは短い遅延時間が要求されるが，遅延時間を満たそうと細かい粒度で圧縮すると圧縮処理の効率が低下するため，まとまった粒度で圧縮できるよう圧縮処理を工夫する必要がある．

3.3　圧縮処理並列実行

ロード処理に含まれる圧縮処理は内部処理の並列度が高く，主記憶の消費量も大きいため，圧縮処理の並列実行可能数を制約する必要がある．圧縮処理が並列実行可能数を超えた場合は，圧縮処理の終了を待つロード処理が発生し，並列実行の効率が低下していた．

4　センサデータロード方式

センサデータロードにおける制約下で，遅延時間を平準化するセンサデータロード方式を実現する．本方式では，ロード処理において圧縮処理の実行判定を行い，圧縮処理が可能な場合のみ圧縮処理を実施する．本方式は(1)競合時の圧縮処理対象テーブルの決定，(2)圧縮対象以外の圧縮処理スキップ，(3)圧縮粒度の微細化を特徴とする．

4.1　圧縮処理対象テーブルの決定

複数のロード処理が競合した場合，競合したテーブルの優先度 p を求め，並列実行可能数に達するまで最も優先度高いテーブルの圧縮処理を実施する．これにより，テーブルの圧縮処理回数の平準化を図る．優先度 p は，圧縮処理スキップと残り遅延時間の観点から算出する．圧縮処理スキップ回数 f_{skip} に関連するパラメータを圧縮処理スキップ回数 c，蓄積データ量 r，テーブル ID i とし，残り遅延時間 f_{remain} に関するパラメータをテーブル毎に設定された遅延時間 t_d，

A Load Balancing Method with Data Compression for Sensor Data

†Masashi Tatedoko, Yoshinori Yamagishi, Mitsunori Kori
Information Technology R&D Center, Mitsubishi Electric Corporation

収集周期t_c，前回圧縮処理時間t_lとすると優先度pは，次式にて求める．

$$p = f_{skip}(c,r,i) + f_{remain}(t_d,t_c,t_l) \quad (1)$$

4.2 圧縮処理スキップ

従来方式では，圧縮処理を実施するタイミングで圧縮処理並列実行数を超過した場合，圧縮処理が実行可能となるまで待機していた．本方式では，圧縮処理対象テーブルとならなかった場合，圧縮処理を行わずにロード処理を終了する．スキップされた圧縮処理は，次回以降のロード処理にて実施する．

4.3 圧縮粒度の微細化

圧縮処理にかかる時間は，圧縮後のデータサイズである圧縮粒度に比例する．従来方式では，圧縮効率を高めるため圧縮粒度を大きくしていた．本方式では，遅延時間の制約を考慮して既定のサイズを従来手法より微細化し，圧縮処理時間の縮小を図る．

5 評価

本方式と従来方式との比較により有効性を評価した．ロード処理の遅延時間と圧縮率および検索処理時間を評価対象とした．圧縮率は次式にて求めた．

圧縮率 = 1 − 圧縮後サイズ/圧縮前サイズ (2)

5.1 評価データ

表 1は，評価データの収集周期，テーブル数，テーブルあたりのロード件数である．評価データは電力データを用いた．1 テーブルあたり 960 個のセンサを割り当て，合計 15 個のテーブルを用意した．全体のセンサ数は 14,400 点である．収集周期は 4 種類とした．各テーブルの遅延時間は 30 秒とした．各テーブルは 1 秒周期で 2,000 秒間繰り返しロードした．

表 1 評価データ

収集周期	テーブル数	ロード件数/テーブル
4ms	3	480,000,000 件
10ms	3	192,000,000 件
40ms	3	48,000,000 件
100ms	6	19,200,000 件

5.2 評価環境

表 2に示すPCサーバを用いて評価した．データをロードするクライアントとデータベースサーバとは同一のPCサーバ上で動作させた．

表 2 評価用 PC サーバ

OS	Windows Server 2008 R2 Standard SP1
CPU	Xeon E5620 x2
Memory	16 GB
Storage	SAS2.0, 450GB, 10000rpm x4, RAID5

5.3 評価結果と考察

本方式および従来方式を用いて評価データのロードを行った．ロード件数の最も多いテーブルにおける各手法による遅延時間の比較を図 2 に示す．

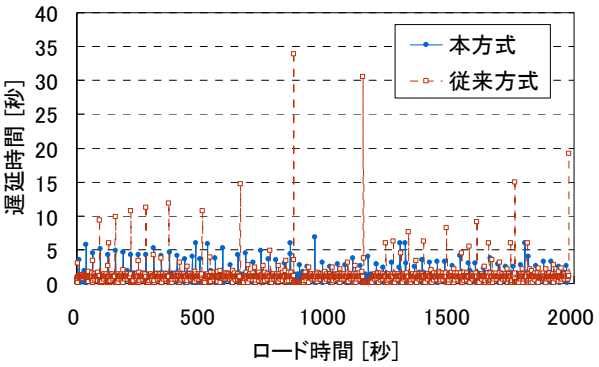

図 2 遅延時間比較

従来方式では，ロード処理の競合により遅延時間が拡大しており，870 秒程度経過した段階で遅延時間である 30 秒を超過している．本方式では，遅延時間が平準化され，最大でも 10 秒以下に圧縮処理が完了している．圧縮率と検索処理時間の比較を表 3に示す．検索時間は，5 件のセンサの合計値を求める問合せを 10 回繰り返し，処理時間の平均値により得た．圧縮率は，本方式が 94.3%であり従来方式の 93.3%より向上した．また，検索時間は，本方式と従来方式ともに 0.02 秒であり，違いは見られなかった．以上より，本方式の有効性を確認した．

表 3 圧縮率，検索時間比較

	本方式	従来方式
圧縮率	94.3 %	93.3 %
検索時間	0.02 秒	0.02 秒

6 まとめ

センサデータ向けのデータ圧縮ロード方式をセンサデータベース上に実装し，その有効性を検証した．今後は，より大規模なデータで評価し，圧縮処理対象の判定精度の向上を検討する．

参考文献

[1] 山岸 義徳ほか：高速集計検索エンジンとセンサデータベースへの応用，三菱電機技報，Vol.83，No.12，pp.11-14(2009)．

[2] 竹田 義聡ほか：環境情報データベース向けリアルタイムセンサデータロード方式，情報処理学会第 73 回全国大会講演論文集，pp1-555- 1-557(2011)．

[3] 加藤 守ほか：環境情報データベース向け高性能センサデータ圧縮方式，情報処理学会第 73 回全国大会講演論文集，pp1-559- 1-561(2011)．