5X-6

大規模かつ多様なデータをリアルタイム解析のための情報基盤

伊藤 悠哉† 戸倉 一† 山崎 治郎† 阿部 泰裕†
福原 英之‡ 宮崎 敏明† 岩瀬 次郎† 林 隆史†
†会津大学 ‡ネットワンシステムズ

1. はじめに

近年、様々なデータを用いたデータマイニングやマッシュアップにより、新しい知見やサービスが産み出されている。しかし、データソースが増加・多様化することで、データが大規模化し、フォーマットが不統一になり、データの管理や処理が困難になっている。また、スマートグリッドをはじめとする分野では、生成されたデータをリアルタイムに解析することが求められてきている

Pub/Sub モデルにおいて、多対多かつ動的に Publisher と Subscriber が変わる点にこの問題の難しさがある。このような問題を解決する方法として、我々は Network-Centric な手法を提案してきた[1-2]。Network-Centric なシステムを実際に用いるためには、低遅延リアルタイムシステムを実現する具体的な検証が必要である。

本稿では、リアルタイム分散処理フレームワークを用いることで、リアルタイムにデータを統合するシステムを提案する。また、統合されたデータを用いたリアルタイム解析やマッシュアップの例を挙げる。

2. 提案システムの概要と実装例

種々のデータソースから生成される大規模なデータを容易に、かつリアルタイムにデータ利用者が利用するためには、利用対象のデータが、
・標準的なインターフェースで取得可能である
・統一されたデータ形式で提供される
・複数のユーザーが同時に利用可能である
ことが必要である[1]。また、データ統合基盤が、
・スケーラブルである
・リアルタイムシステムである
ことが必要である。インターフェースやデータ形式は、厳密に標準化されていなくてもネットワーク上で整形することで標準インターフェースや統一フォーマットに変換可能であれば構わない。

An Intelligent Infrastructure for Real-Time Analysis of Large Multi-Format Dara Sets
Yuya Ito†, Hajime Tokura†, Jiro Yamazaki†, Yasuhiro Abe†, Hideyuki Fukuhara‡, Toshiaki Miyazaki†, Jiro Iwase†, Takafumi Hayashi†
†The University of Aizu
‡Net One Systems

ここで本稿でのリアルタイムシステムというのは、データが生成されてから利用されるまでの間、全ての処理がストリームデータ処理であること、と定義する。

我々は、上記を実現するデータ統合のためのシステムを提案する。提案システムでは、データが生成されてから、メッセージング・ネットワークとデータ統合基盤を用い、利用者が統合されたデータを利用する[図 1]。

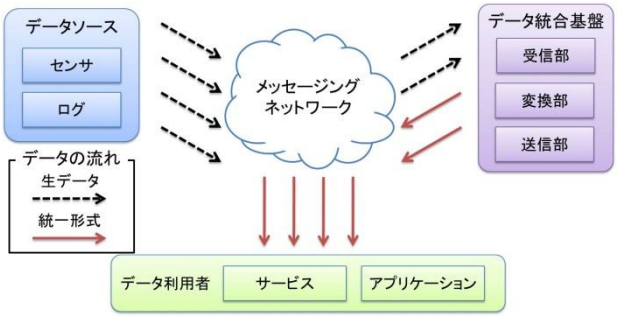

図 1 システムの概要とデータの流れ

2.1. メッセージング・ネットワーク

メッセージングとは、ネットワーク上の特定の複数間におけるメッセージのやりとりをいう。この場合のメッセージとは、ネットワーク上の 2 者ないし複数の者の間でやりとりされる情報の総称である。このメッセージングを行うネットワークをメッセージング・ネットワークとよぶ[2]。メッセージングはルータを用いて行うことができる。ルータは、クライアントが分かっているため、データソースやデータ利用者を認証するのに適している。また、経路制御・フィルタリングの手法としてトピック・ルーティングを用いる。トピック・ルーティングとは、Publisher はルータの構造化されたトピックへメッセージを送ることで、そのトピックの全ての Subscriber はメッセージを受け取ることができる。このような機能を行うことが出来るルータをメッセージルータと呼び、Solace[3]が利用可能である。Solace は、超高速・低遅延なメッセージングを可能にし、JMS, Java, .Net, そして javascript などの API で利用可能である。

データソースから生成されるデータ(生データ)は、そのままではメッセージングに利用出来ない。そのため、生データをメッセージに変換

するためのアダプタを設ける。このとき、メッセージには ID などの関連する情報も含ませる。また、トピックには、データソースの名前やバージョンなどの情報を含ませる。

2.2. データ統合基盤

提案するデータ統合基盤は、おおまかに受信部、変換部、そして送信部に分かれる[図 2]。

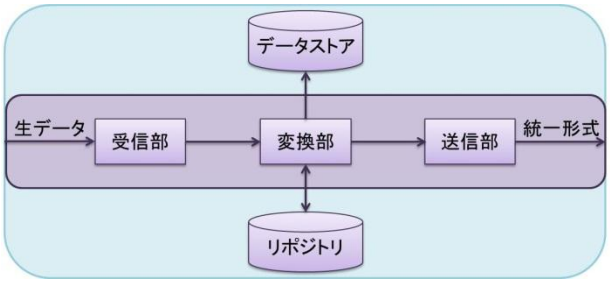

図 2 データ統合基盤

まず受信部は、トピックの情報をもとに生データを含むメッセージを受け取り、適切な変換プログラムが用意されている変換部に生データや ID などを送る。変換部は、生データなどの情報を key-value pair のリストに変換する。このとき、ID を関連付けリポジトリから位置情報などのデータソースに関する情報をこのリストへ追加し、送信部へこのリストを送る。また、データのロールバックやバッチ処理のために変換前後のデータや ID をデータストアへ保存しておく。送信部は、このリストを適切な形式に変換し、この統一形式のデータを含むメッセージを作成しメッセージルータへ送る。データ形式には XML と JSON をサポートした。このときのトピックの情報は、データのタイプや形式、位置情報などを含む。

この受信部、変換部、そして送信部を、リアルタイム分散処理フレームワークである Storm[4]によって実現した。これにより、スケーラビリティを持ち、ストリームデータ処理が可能なデータ統合基盤を構築できた。

3. システムの応用例

このシステムのデータ利用者の応用例を挙げる。データソースとして

・温度センサ(7 個) 温度

・環境センサ(3 個) 温度、湿度、気圧、照度、CO2

の 2 種類・計 10 個のセンサを室内に設置した。

● 複数のセンサデータのリアルタイム可視化

室内のセンサデータのリアルタイム可視化を行った。データソースからデータが生成されてからリアルタイムに反映させることを可能にした。これによって JSON 形式で取得したものを、javascript で容易に扱うことが出来るようになった。また、トピックを変えることにより、温度、湿度、気圧、照度、CO2 それぞれの値を反映させることができる。この実装には、HTML と javascript を用いた。

● 複数のセンサデータのリアルタイム解析

日立 uCosminexus Stream Data Platform (uCSDP)[5]を用いたリアルタイム解析を行った。uCSDP では、SQL に似た CQL という言語でストリームデータ処理をしていく。ここでは、データソースの位置ごとに一分間の温度、湿度、気圧、照度、CO2 の平均値を求めた。uCSDP へデータを送るためのプログラムを 1 つの Java プログラムだけで実装した。XML 形式でデータを取得し、特別な処理をすることなく実装できた。高度なデータマイニングのアルゴリズムを導入することにより、複合イベント処理にも応用可能である。

4. まとめ

本報告では、大規模かつ多様なデータをリアルタイムに利用することを容易にするための、Network-Centric なデータ統合基盤を提案した。このデータ統合基盤を用いることで、種々のデータソースから生成される大規模なデータを、統一の形式、そしてインターフェースでリアルタイムに利用することができる。今後は、より本格的な実装を行い、その性能を評価することで、様々な応用システムへの実現方法の検討をする予定である。

参考文献

[1] 山田 拓人, 鈴木 一徳, 和良品 友大, 林 隆史, “大規模な異種データ解析のための情報基盤,” *全国大会講演論文集,* pp. 635-637, 2011.

[2] 川内 見作, 高橋 友一, 福原 英之, 古瀬田 勇, 藤田 龍太郎, 衣川 昌宏, 宮崎 敏明, 斎藤 梅朗, 林 隆史, “メッセージング・ネットワークを用いた環境情報統合,” *電子情報通信学会技術研究報告. IA,* pp. 23-26, 2008.

[3] Solace Systems, ″Solace's Unified Messaging Platform,″ [Online]. Available: http://solacesystems.com/library/ump-paper.php.

[4] N. Marz, ″nathanmarz/storm · GitHub,″ [Online]. Available: https://github.com/nathanmarz/storm.

[5] 日立製作所, “取扱説明書 uCosminexus Stream Data Platform 01-02 ストリームデータ処理基盤,” 2012.