5M-3

大規模な異種データ解析のための情報基盤

山田 拓人† 鈴木 一徳† 和良品 友大† 林 隆史†
会津大学†

1. はじめに

近年、Hadoop などのデータを分散処理するフレームワークが普及したことにより、蓄積された大量のデータを分析することが容易になりつつある。一方で、温度や湿度を測定するセンサーや、機器の監視ログなどを生成するデータソースも増え、様々な情報を取得・蓄積することも可能になっている。これらの様々な種類のデータを組み合わせることでイノベーションが生まれる可能性があり、それを支援する仕組みが必要である。しかしながらデータソースの多様性は、インターフェースやデータフォーマットの不統一を生み出し、データの利用や統合を困難にした。データの標準化によってある程度のデータの統合は可能ではあるが、標準は時代や地域、目的によって変わる場合もあるため、標準だけでなく別の統合方法も必要である。

そこで我々は、様々な種類のデータソースのデータの利用を容易にするために、様々な種類のデータをある一定の形式に変換することが可能な Network-Centric なデータ統合基盤を構築した。[図 1] 具体的には、ネットワークを通して収集した全てのデータを分析処理が容易な XML、または構造を持ったテキスト形式に変換し、データの利用者に提供する。

本稿では構築した Network-Centric なデータ統合基盤の概要と実装例を挙げるとともに、この基盤を用いて行ったデータ分析の実例を紹介する。

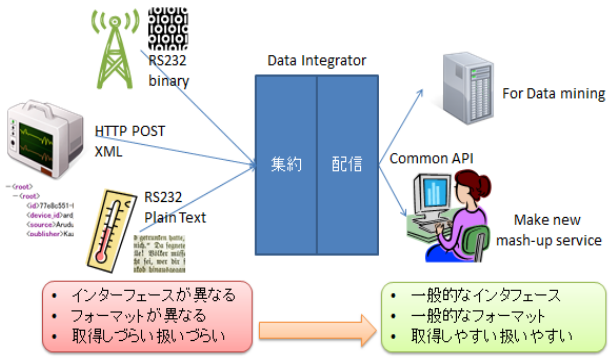

図 1 データ統合の概要

An Intelligent Infrastructure with Services for Heterogeneous Data Analysis
Takuto Yamada†, Kazunori Suzuki†, Tomohiro Warashina†, Takafumi Hayashi†
†The University of Aizu

2. データ統合のアプローチ

データソースには、各種センサデータ、気象情報、機器のログ、生体情報などがある。そこから生成されるデータを分析するアプリケーションやサービスを容易に実装するには、分析対象のデータが、

1) 標準的なインタフェースで取得可能である
2) 統一されたデータ形式で提供されている
3) 複数のユーザが同時に利用可能である

ことが必要である。

しかしながら、データソースはそれぞれ異なる通信インターフェースとデータフォーマットを持っているため、それらのデータはアプリケーションが利用しやすい形式に変換される必要がある。データソース自身がデータを分析者にとって適切な形式で提供する Publisher-Centric なシステムは、既存のデータソースに大きな変更を加える必要があることや、データソースの設置コストなどが大きくなってしまう問題がある。データの利用者がデータ変換などを行う Subscriber-Centric なデータ統合では、各々の分析アプリケーションが様々なインターフェースとデータフォーマットをサポートする必要があり、これは非効率かつ困難である。一方 Network-Centric な手法では、データソースとデータ利用者間の通信経路上でインターフェースやデータ変換を行う。必要な処理をネットワーク上で行うことにより、publisher(データソース)と subscriber(データ利用者)の両方の負担が減り、より容易なデータの提供とデータの分析が可能になる。

我々は、この Network-Centric なデータ統合によって上記 3 点を実現し、データ分析の実装を容易にする。

3. 提案システムの概要と実装例

上記 3 点を実現するための Network-Centric なデータ統合基盤を提案する。このシステムの概要を、我々が行った実装を例に交えながら述べる。データ統合基盤は、データソースからデータを収集する受信部、収集したデータを変換しフォーマットやデータ単位などを統一する変換部、変換されたデータを保存する保存部、利用者からの要求に応じてデータを提供する配信部で構成される。[図 2]

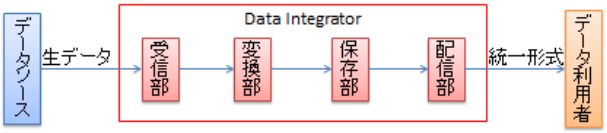

図 2 データ統合基盤の構成

我々はこのシステムを主にウェブサーバー(ASP.NET/IIS7.0)を用いて実装した。受信部は適切なデータ変換をするためにデータソースを識別する必要がある。実装では、HTTP Basic 認証を用いたデータソース識別を行った。変換部は、データソースの種類に対応したスクリプトによるデータ変換を実装した。スクリプトは、アップロードされた生データ(バイトストリーム)を attribute-value pair のリスト(連想配列)に変換する。スクリプトはデータソース提供者かボランティアによってシステムにアップロードされる。収集されるデータの attribute は多種かつ不統一であるため、保存部で利用するデータベースはスキーマレスである必要がある。今実装では、通常のファイルシステム上に構造を持ったテキスト形式で保存した。データ利用者は配信部に要求することにより、保存されたデータを適切な形式で取得することができる。実装では、配信技術に HTTP と Comet を使い、データ形式には XML, JSON, CSV 形式をサポートした。実装したシステム上でのデータ変換の流れを[図 3]に示す。

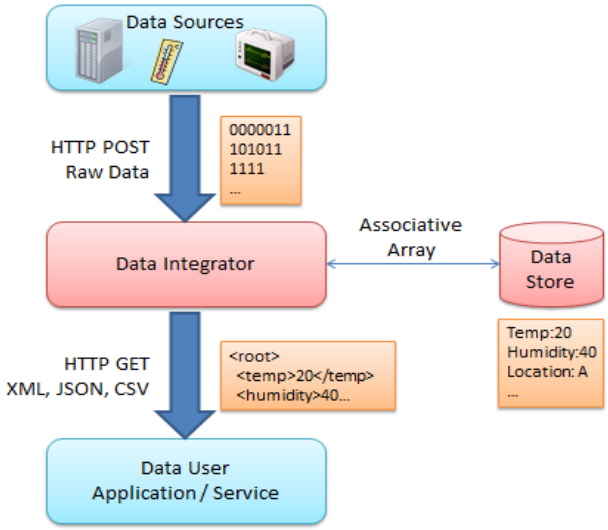

図 3 データ変換の流れ

4. システムの応用例

前節で実装したシステムを用いて行ったデータ分析の例を紹介する。

・複数センサデータのリアルタイム可視化

室内に設置した 3 種類・計 10 個の温度センサのデータを、HTML/javascript を用いて室温のリアルタイム可視化を行った。温度データをデータソースに依らない JSON 形式で取得することで、容易にデータを javascript で処理することができた。また、Comet によるデータ配信により、データソースからのデータをほぼリアルタイムに反映可能なことも確かめられた。この室温可視化プログラムは、小さい HTML と 20-30 行程度の javascript で実現できた。

・複数センサデータの Hadoop によるデータ解析

次に複数の異なる種類のデータソースが混在する環境におけるデータマイニングが可能であるかを確かめた。データソースとして、以下の環境データを測定するセンサを使用した。

- センサ A(1 機) 黒球温度、温度、湿度
- センサ B(3 機) 温度、湿度、気圧、CO2 濃度
- センサ C(5 機) 温度

これらのセンサから収集されたデータは全て CSV 形式で取得可能であり、解析プログラムは必要なデータ属性(温度や湿度など)を特別な処理を施すことなく利用できるため、Hadoop MapReduce を容易に実装することが出来た。試験的に行った温度、湿度、気圧、CO2 濃度の平均や最大値を求める MapReduce プログラムは 100 行程度で実装できた。

5. まとめ

本報告では、様々な種類のデータの解析を容易にするために、多種･多様なデータソースに対応可能な Network-Centric なデータ統合基盤を提案・構築した。このデータ統合基盤は一般的な Web サーバーと HTTP 技術で実現可能である。このデータ統合基盤を用いることで、必要なデータをデータソースに依らないインターフェースとデータフォーマットで取得できるため、データ分析アプリケーションを容易に実装できるようになった。今後は本格的な実装と運用のための、多様なデータ属性に対応可能なデータベースの選定、標準データモデルの定義を行う予定である。

参考文献

[1] T. Hayashi, J. Terazono, Y. Watanabe, and T. Sizuki, “Loosely coupled integration of sensor data and related services” IEICE Tech. Rep., vol. 110, no. 349, IA2010-65, pp. 43-47, Dec. 2010.

[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”, IEICE Tech. Rep., IA2008-5, pp. 23-26