5W-4

大規模な異種データ解析のためのスケーラブルな情報基盤

和良品 友大† 林 隆史†
†会津大学

1 はじめに

近年, 加速度センサーを用いた落下検知など, センサー情報から機械を制御する機構が普及した. 特に携帯電話などの電子機器は落下検知や消費電力を低減させるために, 多種多様なセンサーが搭載されている. そのセンサー情報を従来の用途と異なる検知に使用し, 所持者のライフログ [2] や認証 [3] に活用する動きがある.

しかしながら，複数の異種センサーを用いた場合，センサー自体の形式が異なるために統合が難しい。また，センサー情報は常に蓄積されるので,，規模が大きくなり処理機構が大規模になってしまう問題がある。そのため，大規模なデータを収集し，保持・管理を行う必要がある。また，それらはデータソースによって通信インターフェースやデータフォーマットが異なっている。そのためアプリケーションが利用したいデータを，利用しやすい形式に変換する必要もある。

これら問題解決のための要求を満たすためにデータソースが，分析者の利用したいデータを適切な形式で提供する，Publisher-Centric な手法 を用いると，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，設置コストが大きくなってしまう問題がある。一方，データの利用者がデータソースからすべての生データを受信しデータ変換を行う Subscriber-Centric なデータ統合では，各々の分析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るため困難である。

そこで，本論文ではデータソースとデータ利用者の間に情報基盤を構築して上記要求を満たすことのできる，Network-Centric な手法を提案する。[図 1] この情報基盤は，上記問題を解決するために必要となる処理をネットワーク上で行う。これにより，Publisher と Subscriber 両方の負担を減らし，より容易なデータ分析の実装を行うことができるようになる。また，この情報基盤の実装とこれを用いた応用例も併せて紹介する。

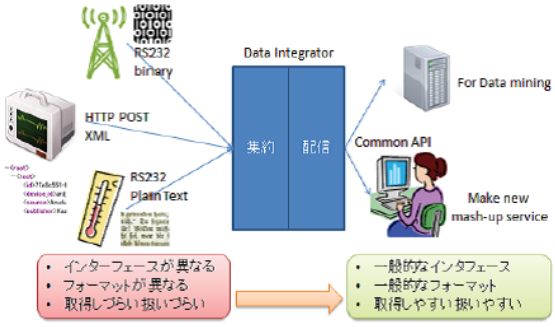

図 1: センサー情報活用のイメージ

2 提案と実装

上記大規模な異種データの問題を解決するための Network-Centric な情報基盤を，我々が行った実装を例に交えながら述べる。本論文では情報基盤を，アグリゲータ，統合基盤，パブリッシャーの 3 つで構築した。[図 2]

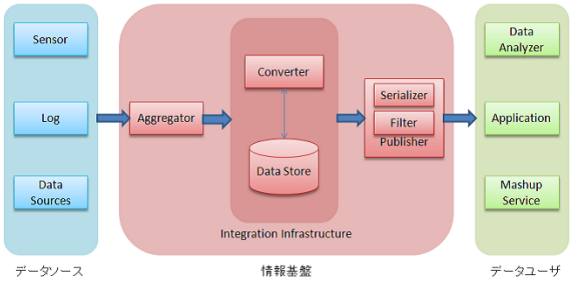

図 2: 情報基盤の構成

アグリゲータは，データソースから送られてくるデータを収集する。我々はこれを Flume [4] を用いて実装した。これはスケーラビリティと耐障害性に優れているデータ収集機である。これによって，大規模データを収集し統合基盤へと送ることができる。

統合基盤は，生データのフォーマット変換，およびデータの保持・管理を行う。生データのフォーマット変換は，収集した不統一な生データのフォーマットを一定の形式にシリアライズするために行う。具体的には，生データをスキーマレスの状態にする。この統合基盤は Hadoop [5] を用いて実装した。Hadoop は Hadoop 分散ファイルシステム (HDFS：Hadoop Distributed File

A Scalable Intelligent Infrastructure for Large-scale Heterogeneous Data Processing
†Tomohiro Warashina †Takafumi Hayashi
†The University of Aizu

System) と MapReduce から構成される，分散フレームワークである。HDFS は大きなファイルを複数のコンピュータ (ノード) にまたがり格納することができる。デフォルトのレプリケーション数が 3 のとき，データは 3 つのノードに保存される。その 3 つのノードのいずれかに障害が起こった場合，残りのノードからその故障したノードの保持していたデータを他のノードへ複製することにより，自動的にバックアップを行う。これにより，大規模データの保持・管理を実現できる。

MapReduce は多数のノードを用いて，巨大なデータセットに対して並列処理を行うためのフレームワークである。Map 処理と Reduce 処理からなり，それぞれの処理は並列的に行われる。そのため，複数あるノードの一部に障害が起こり，処理が実行できなくなった場合でも，処理を再スケジュールして実行させることが可能となる。これを用いれば，大規模な異種データをスキーマレスな中間フォーマットに変換することができる。

パブリッシャーは必要なデータを抽出し，それをユーザが望むフォーマットにシリアライズして配布する。これを ASP.NET/IIS7.0 と HTTP を用いて実装した。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。

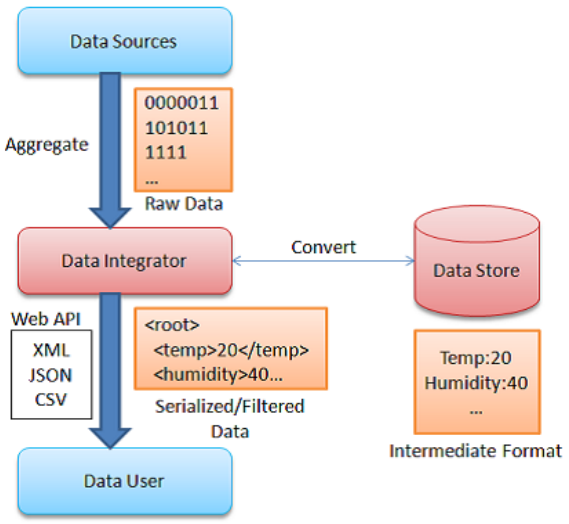

図 3: データ変換の流れ

3 システムの応用例

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，以下の環境を測定するセンサーを使用した。

- センサー A (1 機) 黒球温度，温度，湿度
- センサー B (3 機) 温度，湿度，気圧，CO_2 濃度
- センサー C (5 機) 温度

これらセンサーから収集されたデータはすべて 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，プログラムを容易に実装することができた。試験的に行った温度，湿度，気圧，CO_2 濃度の平均や最大値・最小値を求める Java プログラムは，解析アルゴリズムのみの 100 行未満の行数で実装できた。

4 まとめ

本論文では，大規模な異種データの解析を容易にするために Network-Centric な情報基盤を提案・構築した。情報基盤にクラウド基盤技術を用いて実装することにより，大規模なデータの収集および，保持・管理を容易に行うことができるようになった。また，Web サーバと HTTP 技術で，ユーザが利用したいデータを，利用しやすい形式にシリアライズする機能も実現した。これにより，ユーザが利用したいデータをセンサーの形式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，データ分析者は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を行う予定である。

参考文献

[1] Takuto Yamada, An Intelligent Infrastructure with Services for Heterogeneous Data Mining and Mashup.

[2] 伊藤 達明，石原 達也，中村 幸博，武藤 伸洋，阿部 匡伸，携帯電話を用いた歩行モニタリング (情報セキュリティ, ライフログ活用技術, ライフインテリジェンス, オフィス情報システム, 一般)，IEICE technical report 110(282), 9-14, 2010-11-10

[3] 石原 進，太田 雅敏，行方 エリキ，水野 忠則，端末自体の動きを用いた携帯端末向け個人認証，2005 年 12 月, 情報処理学会論文誌 Vol. 46 No. 12

[4] Apache, ”Apache Flume” Available: https://cwiki.apache.org/FLUME/

[5] Apache, ”Apache Hadoop” Available: http://hadoop.apache.org/