4C-1

Hadoop上の空間データ問合せ処理フレームワーク

石川 佳治 †§　　杉山 武至 †　　鈴木 優 †
† 名古屋大学大学院情報科学研究科　†† 国立情報学研究所

1　はじめに

Hadoop は，クラウド環境における大規模な並列分散データ処理のためのフレームワークとして広く利用されている [3]．Map と Reduce という関数の組合せにより処理を記述する **MapReduce** のアプローチがその基礎であり，分散ファイルシステム上でのデータ処理を透過的に記述できる．データベースの研究分野においては，性能のチューニング方式の開発，フレームワーク自体の拡張，Hadoop に適したアルゴリズムの開発など，さまざまな研究がなされている [2]．

一方，**空間データ**（spatial data）は，電子化された地図データの普及や，モバイルコンピューティングや ITS の発展などを受け，その利用と重要性がますます増大している．また，近年，実世界の情報とオンライン情報を統合活用するサイバーフィジカルシステム（CPS）も重要となってきているが，そこでも空間データに関する技術は要素技術の一つとなる．

このような背景を受け，我々のグループでは，大規模空間データの利活用のためのクラウドコンピューティング技術に関する研究を進めている．特に Hadoop の活用に着目しており，たとえば [5] においては Hadoop 上で全 k 最近傍問合せを処理する効率的なアルゴリズムを開発している．

本稿ではこの延長線上として，現在研究を進めている，クラウド技術，特に Hadoop を活用した大規模空間データの分析のためのシステムフレームワークの概要について述べる．本研究ではこのような分析のことを**空間データアナリティクス**（spatial data analytics）と呼ぶことにする．その背景と構想，および技術的検討課題について述べる．

2　背景

ビッグデータの時代を迎えて，大規模なデータの処理技術が以前にも増して重要となっている．特に，分析に着目すると，大規模データの分析を意味するキーワードとして，近年**データアナリティクス**が注目されており，データベースの研究分野では分析機能の拡充や処理の効率化などのため，新たな技術開発が盛んに進められている．たとえば MADlib* では，*scalable in-database analytics* と銘打って，分析処理で頻繁に用いられる繰返しによる最適化や線形代数の操作をデータベース内で実現し，分析処理を効率的に行うことを実現している．このような流れは，確率的データベース（probabilistic database）に関する研究にも存在し，確率的知識表現の能力をデータベースに統合し，機械学習との連携機能を強化するような取り組みが見られる．しかし，これらの研究では一般的な学習機能の統合を目指しており，本研究が対象とする空間データの分析については考慮されていない．空間データを対象とした場合，道路網や地図データなどの表現・操作機能に加え，空間統計（spatial statistics）の機能や，移動状況等に関する統計処理機能など，空間データに関する分析要求に応じた機能とデータベースの大規模データ管理・処理機能を密に連携する必要がある．

3　システムの概要

想定するシステムの構成を図 1 に示す．メインとなるのは右側に位置する部分であり，データアナリティクスミドルウェアがその下位に位置する RDBMS，GIS，およびクラウド計算基盤（Hadoop と HDFS）を管理する．RDBMS には一般的なデータが管理され，GIS には地図データや移動軌跡データなどの空間データが管理される．ただし，RDBMS については，空間データの管理も一部担当し，空間データに対する高速な問合せ処理を支援する．GIS は，空間データに特有である詳細レベルの統計処理等を実現するために用いる．

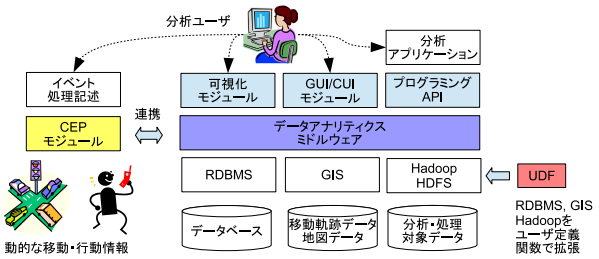

図 1: システムの構成

Hadoop および HDFS は，RDBMS や GIS により十分カバーできない処理を支援することが主な目的となる．一つには，バッチ的な処理による大規模空間データの一括処理が挙げられる．たとえば，[5] における全最近傍問合せはそのような例である．特にセンサ情報や履歴情報の処理などにおいては，RDBMS や GIS に格納するまでもなく即座に分析したいデータが存在するため，Hadoop の活用の意義が高い．また，Hadoop 利用の別の用途としては，RDBMS や GIS に入力するデータの前処理が考えられる．たとえば，

Spatial Data Query Processing Framework Based on Hadoop
Yoshiharu Ishikawa †§, Takeshi Sugiyama†, Yu Suzuki†
† Graduate School of Information Science, Nagoya University
§ National Institute of Informatics
*http://madlib.net/

データベース中の各点に対する全最近傍問合せの結果を RDBMS に事前に格納したい場合，RDBMS 上で求めるのではなく，前述の Hadoop 上の全最近傍問合せ処理をを適用し，その結果をデータベースにロードすることが考えられる．

RDB，GIS，およびクラウド基盤システムについては，システムの機能を実現するためのユーティリティ関数群を提供し，空間データアナリシスにおいて活用する基本機能群を実現する．また，ユーザ定義関数による拡張機能も提供し，アプリケーション固有の要求をシステムに取り込むことも可能とする．

図の左側に示しているのは，複合イベント処理（CEP）に関するモジュールである．センサ情報処理やサイバーフィジカルコンピューティングに関するアプリケーションでは，発生したイベントに応じて動的に処理を起動し，対応する分析処理やデータ管理を即座に実施する必要がある．このモジュールにより，柔軟なイベントベースの処理を実現する．

4 Hadoop を用いた空間データアナリティクス

4.1 基本的な考え方

Hadoop は，クラウド環境における並列分散コンピューティングのために広く普及している技術であるが，一般的にはその利用において Java 言語などを用いたプログラムを作成し実行する必要がある．しかし，探索的なデータアナリティクスの過程においては，プログラムを構築しコンパイル実行するようなステップは，手間と時間の面であまり受け入れられない．対話的なデータアナリティクスを支援するためには，比較的容易に記述できるインタプリタ型の処理言語の方が望ましい．このような点から，本研究では Pig [1] をベースにその拡張を図る．Pig は Hadoop 上で動作するデータフロー型のインタプリタ言語である Pig Latin を提供する．これは，Hadoop プログラムにコンパイルされて実行される．本研究では Pig に対して空間データの処理機能を追加し，これを Hadoop 上に展開することで，対話的な空間データアナリティクスの支援を図る．

4.2 空間データ処理機能による Pig の拡張

まず，図 2 のような二つのテキストファイルを考える．`usrs.txt` はユーザ情報を保持しており，ユーザ名や住所の座標値などの属性が含まれている．一方，`msgs.txt` はソーシャルネットワークサービス（SNS）等に書き込まれたメッセージの情報を保持する．ただし，座標が対応したメッセージのみが抽出されており，日付，位置などの情報が属性として得られているとする．ここで，SNS の各ユーザに対し，最近記述されたメッセージの中から，そのユーザの住所に近い（例：2km 以内）地理情報に触れているものを提示することを考える．

```
> cat usrs.txt       | > cat msgs.txt
1 John (25, 39) ...  | 1 1/10/2013 (93, 60) ...
2 Mary (80, 71) ...  | 2 1/10 2013 (40, 27) ...
3 Mike (75,  4) ...  | 3 1/11/2013 (30, 52) ...
...                  |
```

図 2: 空間情報を含むデータ

このような要求に応えるための Pig Latin の拡張構文による問合せを図 3 に示す．これは，「1 月 10 日のメッセージについて，ユーザの住所と近いものを取り出す」という問合せを想定している．具体的な拡張のポイントとしては以下の二つがある．

1. `LOAD` 演算でファイルをロードする際の型指定において，空間上の点に対応する `point` データ型が指定できる．
2. 空間結合を行う演算 `SPATIALJOIN` を使用できる．距離の閾値は `WHERE` 句内に記述される．

```
usrs = LOAD 'usrs.txt'
       AS (uid:int, name:chararray, uloc:point);
msgs = LOAD 'msgs.txt'
       AS (mid:int, date:chararray, mloc:point);
new _msgs = FILTER msgs By date == '1/10/2013';
R = SPATIALJOIN usrs BY uloc, new_msgs BY mloc
    WHERE within(20);
STORE R INTO 'output';
```

図 3: 拡張した Pig Latin による問合せ記述

このようなアプローチについて，初期的なアイデアを既に [4] で提案しているが，[5] で開発した最近傍結合処理アルゴリズムなども活用できるように開発を進めることが当面の目標である．また，道路ネットワークなどの地図に関連した空間データの処理を対話的に実行するための拡張も検討課題の一つである．

謝辞

本研究は文部科学省委託事業「地球環境情報統融合プログラム」および科研費（22300034）による．

参考文献

[1] A. F. Gates, et al. Building a high-level dataflow system on top of Map-Reduce: the Pig experience. *Proc. of VLDB Endowment (PVLDB)*, 2(2):1414–1425, 2009.

[2] K.-H. Lee, H. Choi, Y. D. Chung, Y.-J. Lee, and B. Moon. Parallel data processing with MapReduce: A survey. *SIGMOD Record*, 40(4):11–20, 2011.

[3] T. White. *Hadoop: The Definitive Guide*. O'Reilly, 3 edition, 2012.

[4] 横山, 石川. 大規模空間情報処理に対応するための分散処理フレームワーク Pig の拡張. 第 8 回情報学ワークショップ（WiNF2010）, pp. 221–224, 2010.

[5] 横山, 石川, 鈴木. Hadoop 環境における空間分割による並列全 k 近傍問合せ処理. 日本データベース学会論文誌, 11(1):25–30, 2012.