Vol.2014-OS-128 No.9
2014/3/7

Hadoop を用いた衛星画像データ解析処理の高速化の研究

塚本 勢児†1　布広永示†2

†1 東京情報大学大学院総合情報学研究科
†2 東京情報大学総合情報学部

東京情報大学では,学術フロンティアプロジェクトの一環として,NASAの衛星TerraとAquaに搭載されているセンサーで撮影された衛星画像データであるMODISデータを受信し,大学や研究機関に提供している.現在，MODISデータを蓄積，解析する衛星画像データ解析処理システムを開発している.本研究では,衛星画像データ解析システムのデータ解析性能を向上するために,オープンソースソフトウェアとして無償で公開されている分散並列処理基盤「Hadoop」を導入することを検討した.本報告では,衛星画像データ解析処理システムへの適用性の評価と処理性能の測定を実施し，その検証概要,検証結果について報告する.

Research to speed up the satellite image data processing using Hadoop

SEIJI TSUKAMOTO†1　EIJI NUNOHIRO†2

†1 Graduate School of Tokyo University of Information Sciences

†2 Tokyo University of Information Sciences

The Tokyo University of Information Sciences, as part of the Academic Frontier project, receives Moderate Resolution Imaging Spectroradiometer (MODIS) data captured by the sensor mounted on the Aqua satellite Terra of two aircraft of the United States NASA. As a system for the provision of satellite image data, we have developed a satellite image analysis data processing system. In this research, we introduce the distributed data processing based on "Hadoop" which is currently exhibited free as open source software into the satellite image data analysis system. Next, we evaluate the problem at the time of “Hadoop” system installation and affinity with satellite image data, and measured the data processing speed time. In this paper, we report system configuration, validation summary, and the results of the validation.

1. はじめに

東京情報大学では，学術フロンティアプロジェクトの一環として，アメリカ NASA の２機の衛星 Terra と Aqua に搭載されているセンサーで撮影された MODIS（Moderate Resolution Imaging Spectroradiometer)データを受信し，大学や研究機関に提供している（図１－1)．

本研究の対象となる衛星画像データ解析システム（Satellite Image Data Analysis System：SIDAS）は，複数の PC や高性能サーバで構成される並列分散処理が可能なシステムである．本システムの目的は，解析処理の実装・解析成果の一般公開，及び衛星画像データの一般公開・提供である．SIDAS が扱う衛星データは非常に大容量であり画像解析処理を高速化するためには，データ解析のプログラムを使用するためのデータ変換，蓄積などのデータ配信前処理を高速化する必要がある．

本研究では，SIDAS における衛星画像処理の高速化，ユーザへの衛星データ配信速度の短縮を目的として，近年ビッグデータの処理で多くの実績を残し，発展的な開発が行われ，オープンソースソフトウェアとして無償で公開されている Hadoop の導入を行った．

本報告では，最初に，SIDAS に対する Hadoop の実装方法，Hadoop に対する MODIS データの問題点や適応性について評価した．次に，Hadoop を適応することによる効果について実証し，その評価結果について考察する．

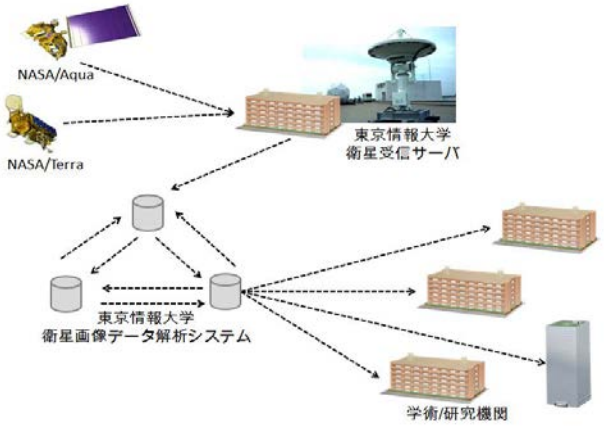

図 1－1 MODIS データ受信及び配信イメージ

2. 衛星画像データ解析システム(SIDAS)

SIDASは，戦略的研究基盤形成に関する研究の一環として進められているMODISデータの解析処理を支援するデータ解析システムである．このSIDASを利用するユーザはWeb環境を利用し，解析処理を行うアプリケーションプログラムの実行や解析結果の確認を行う事が出来る．解析処理としては，林野火災探索，類似画像探索，土地被覆変化解析

や気象変動予測などがあり，現在，それらのアプリケーションを開発中である．

ユーザが本システムを利用して解析処理の要求や処理結果の確認，衛星画像データのダウンロードなどを行う場合は，Web ブラウザを利用する．また，衛星画像データという非常に膨大なデータを用いて実行される解析処理を実装し，ユーザのリクエストで動的に実行するためには，システムのリソースを大量に利用するため，本システムは複数の PC や高性能サーバで構成され,並列分散処理を可能にしている．データを格納するために大容量のストレージを搭載したデータベースを利用している．

現在，SIDAS は東日本大震災版である Version1．2 を経て，配信地域を拡大した Version2．0(図 2-1)を公開している．そして今後新たな SIDAS として，オープンデータやソーシャルデータといったビッグデータと連携された新たな SIDAS を開発する計画も立てられている．

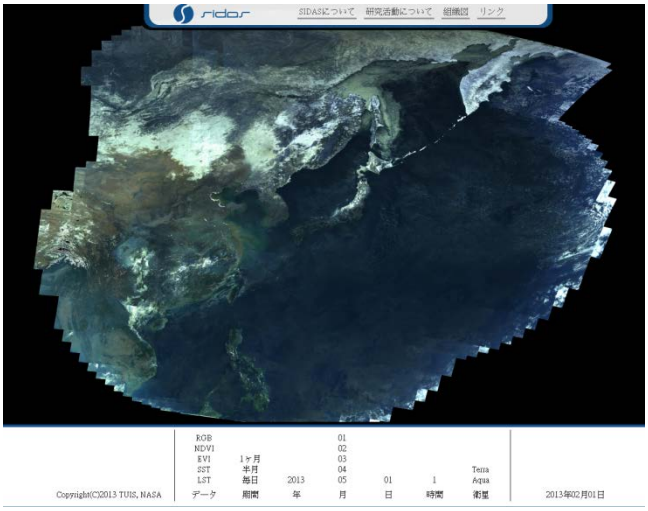

図 2－1　SIDAS Web ページ

2.1 SIDAS システム構成

Webサーバは，ユーザが利用しているWebブラウザとのやり取りを可能するため,Webブラウザに表示されるページの生成を行っている．Webサーバが行っている処理は，ユーザからの解析処理の要求受付や，解析処理結果の表示，衛星画像データの提供である．解析処理要求により，Webサーバは計算クラスタに解析処理の実行を命令し，衛星画像データを提供する際は，データベースと連携して動作する．

アプリケーションサーバは衛星画像データの事前処理や，スケジュールサーバ，PCクラスタとの連携を行う．

計算クラスタは，本システムにおける，高負荷な処理を実行するサーバである．これらが利用する衛星画像データは非常に容量が大きく，数が多いため，システムに大きな負荷を生む．そのため，計算クラスタは複数のPCと高性能サーバで構成され，並列分散処理が可能となっている．

データベースは，衛星画像データ，加工された衛生画像データを管理・格納する．(図 2－2)

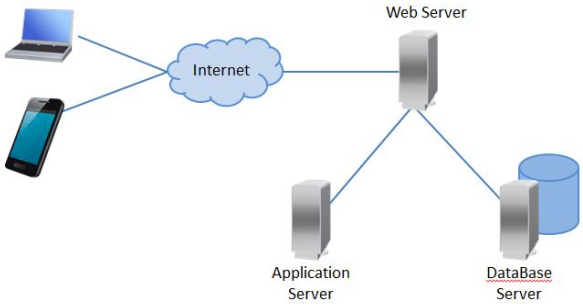

図 2－2 SIDAS システム構成図

2.2 MODIS データ

MODIS は，NASA によって打ち上げられた地球観測衛星 Terra（10 時 30 分に観測）と Aqua（13 時 30 分に観測） に搭載されている光学センサーの一つである中分解能撮像分光放射計である．MODIS の観測幅は 2330km で 36 のバンド（可視域から熱赤外域まで：0．405～14．385μm）と 3 つの空間分解能（250m，500m，1000m）を持っている．

東京情報大学では北海道，東京，沖縄の 3 つの場所で受信をしており，各受信所からは 1 日昼 3～4 回，夜 3～4 回受信する．3 か所から受信する範囲は東経 100 度から 180 度まで，北緯 10 度から 70 度である．

本研究は，受信した MODIS データの中から MODIS 標準プロダクトである MOD02（大気上段反射率データ），MOD03（衛星姿勢データ），MOD11（地表面温度），MOD13（植生指数），MOD28（海面温度），MOD35（雲マスク）を利用する．MODIS データのバンド情報を表 2-1 に示す．

表 2-1　バンド情報

バンド	波長（nm）	空間解像度	利用分野
1	620~670	250m	土地被覆変化 植物葉緑素
2	841~876	250m	雲量 土地被覆変化
3	459~479	500m	土 / 植物識別 雲判断
4	545~565	500m	緑の植物
5	1230~1250	500m	葉 天蓋差
6	1628~1652	500m	スノー 雲識別
7	2105~2155	500m	雲プロパティ 土地プロパティ

・MOD02 は，衛星データの可視化と植生指数の計算をする

ためのデータであり，7 つのバンドで構成される．3 つの可視域バンド（バンド 1，4，3）は衛星データの可視化，近赤外バンド（バンド 2）は植生指数（光合成活動に相関が高い），3 つの短波長赤外バンド（バンド 5，6，7）は雪と雲の識別や異常高温地域検出に有効なデータである．

・MOD03（空間分解能 1KM）は，衛星の姿勢情報から，各画素の緯度経度情報とセンサーの角度や太陽の位置などが記録されている．MOD03 は MOD02，MOD09，MOD11，MOD28 と MOD35 を地図投影法に合わせて変換する際の地図情報データである．

・MOD11 （空間分解能 1KM） は土地被覆の放射率を用いて地表面温度を推定したデータである．都市域のヒートアイランド現象解析や土壌水分の状態解析や，ヒートフラックス解析に有効なデータである．

・MOD13 （空間分解能 250m）は大気補正を行った植生指数データである．

・MOD28（空間分解能 1KM）は海面温度データである．海面温度の変化や海流の流れの解析に多く用いるデータである．

・MOD35（空間分解能 1KM）は雲識別データである．このデータは，画素毎に晴れている確率を 0%，66%，95%と 99%の 4 段階に分類して保存してある．

3. 並列分散処理基盤「Hadoop」

Apache Hadoop(以下，Hadoop)とは，信頼性の高いスケーラブルな分散並列コンピューティングのオープンソフトウェア化されたフレームワークである．Hadoop は，元々 Google の Map&Reduce と，Google File System などの実装が進められ開発されている．また，Hadoop には，専用の対象データに対し高いスループットでアクセスを可能にした Hadoop Distributed File System(以下，HDFS)と，膨大なデータ群を Hadoop の各クラスタへ分散処理するためのソフトウェア Hadoop Map&Reduce で構成されている．

3.1 SIDAS への Hadoop 導入

Hadoop を導入する SIDAS は，衛星画像の配信地域拡大によって，利用するデータ量の拡大やリクエスト処理の増加が予想される．このリクエスト処理とは，システムの利用者がインターネットを介して希望する日付,データ形式,処理方法を選択し，その処理結果を取得する処理である．そのため，衛星画像データ，ネットワークトラフィックの増大が予想されため，Hadoop による分散並列処理を用いることで，データ処理速度の向上を考える．

3.2 Hadoop による画像加工処理

SIDAS には，衛星画像を配信する前段階の処理として，データ配信前処理というものがある．この処理は，受信直後の衛星の元データでは画像として表示することが出来ないため行われる画像加工処理である．

衛星の元データである HDF(Hierachical Data Format)データとは，衛星データをはじめとする各種データの共通フォーマットである．この HDF データを画像イメージデータの未加工データである RAW データへ変換する処理である．本研究では，この SIDAS で行っている画像加工処理に Hadoop を導入する．

4. 適用・検証・評価

4.1 評価環境構成

Hadoop を SIDAS で実行されている処理に近い環境で検証するための計算機器並びに使用しているソフトウェアの情報を表 4－1，システム構成図を図 4－1 に示す．

表 4-1 Hadoop 使用機器一覧

Name Node(Master Server)1 台	
製品名	BTO 製品
OS	CentOS6．0(安定板)
CPU	Corei7-3770 3．4GHz
HDD 容量	2TB SATA 7200rpm
メモリ容量	16GB
SSD	128GB(OS，その他ソフトウェア)
オプション	NIC 増設 通信速度 1000MBbps
ソフトウェア	OS ：CentOS6．0 JAVA ：JDK 10．6．2-1 (1．6．0-24) Hadoop：Ver．2．0
Data Node(Slave Server) 4 台	
製品名	BTO 製品
OS	CentOS6．0(安定板)
CPU	Corei7-3770 3．4GHz
HDD 容量	2TB SATA 7200rpm
SSD	128GB(OS，その他ソフトウェア)
メモリ容量	16GB
オプション	NIC 増設 通信速度 1000MBbps
ソフトウェア	OS ：CentOS6．0 JAVA ：JDK 10．6．2-1 (1．6．0-24) Hadoop：Ver．2．0

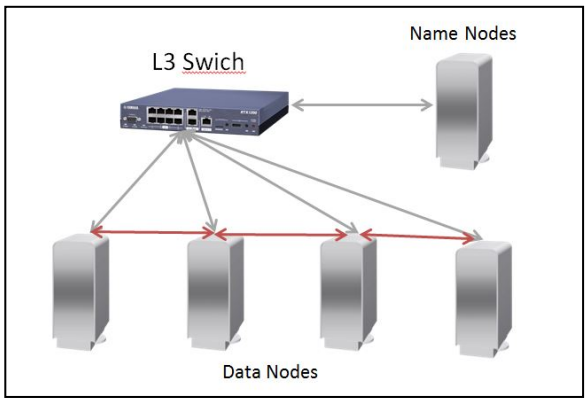

図 4－1 検証環境構成図

4.2 評価方法

本研究では,Hadoop を HDF データのような一つ一つのデータ単体の容量が大きく，大量に存在する場合の画像加工処理に対する Hadoop のデータ処理時間の計測を行った.測定対象は従来の SIDAS システムで使用している環境と同じ単体処理と本研究の Hadoop システムの処理時間を計測する。そして，Hadoop の挙動から HDF データのような特性をもつデータを処理した場合について適用性を評価する．Hadoop 導入による処理時間の短縮した場合のイメージを図 4-2，検証方法，測定データを表 4-2 に示す.

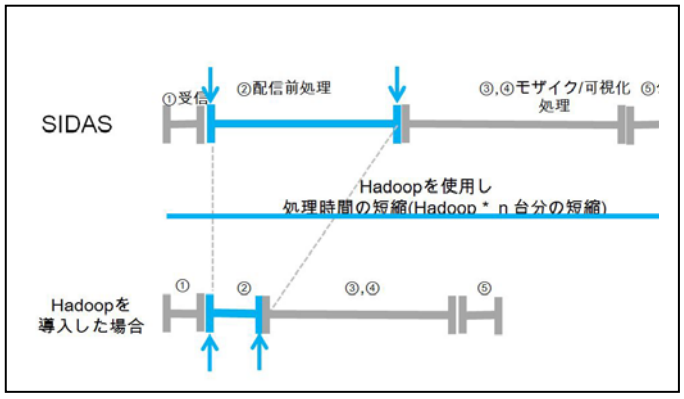

図 4-2　Hadoop 導入による処理時間の短縮
した場合のイメージ

測定方法は，プログラムの実行から処理結果を Master Server へ出力するまでを計測するターンアラウンドタイムという測定方法を用いる．対象データは SIDAS で利用する容量のデータの最大量を想定し，1 日に受信する HDF データ量の最大容量である 30GB，3 日相当の 90GB，1 週間相当の 210GB，2 週間相当の 420GB，その他にも中間地点として 10GB，50GB，100GB のパターンでも処理を行った．容量のパターン毎にファイル件数が定まっていないのは，人工衛星で撮影が成功した範囲，撮影回数が地球の周回回数に応じて 1 HDF ファイル内の容量が様々であるため，その再現を行うためにテスト用 HDF ファイルをランダムに選択したものを選んでいるためである．また，処理を失敗した場合に限り再処理をかけて再計測を行っている．

表 4－2　研究測定手法・詳細表

検証方法	HDF データをテキストへ変換したデータをワードカウントし，処理速度を測定した．
測定方法	ターンアラウンドタイム方式による測定
使用する処理	HDF データ内を読み込み，単語，数値を収集し単語，数字の個数をカウントする処理
測定パターン	・10GB-データ件数:56 件 ・30GB-データ件数:172 件(HDF1 日相当) ・50GB-データ件数:252 件 ・90GB-データ件数: 411 件(3 日相当) ・100GB-データ件数:470 件 ・210GB-データ件数:1002 件(1 週間相当) ・420GB-データ件数: 2003 件(2 週間相当)
単体処理マシン	Hadoop Master Server

4.3 評価結果

今回の結果では Hadoop で HDF データをテキスト化したものを一日の受信容量相当である 10GB から 30GB， 50GB，90GB ，100GB 210GB 420GB のパターンで処理をかけて測定，容量に応じたパターンの時間測定を行った．測定結果を表 4－3，図 4－2 に示す.

表 4－3 Hadoop 処理と単体処理の処理時間対応表

	単体処理	Hadoop	処理時間倍率
10GB	7:43	2:28	3．12
30GB	8:21	8:21	2．72
50GB	9:09	9:09	2．74
90GB	9:51	9:51	2．6
100GB	10:11	10:11	2．33
210GB	13:39	13:39	2．4
420GB	19:48	19:48	3．1

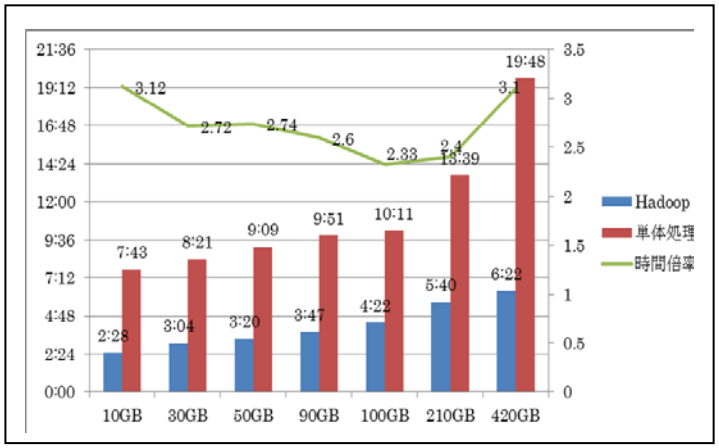

図 4-2　測定結果

4.3.1 評価結果-性能について

今回実施した検証では，測定した処理速度が Hadoop の処理時間が，単体処理に比べて約 2 分の 1 から 3 分の 1 の処理時間を，処理容量が追加増量された場合でも維持をしている．そして，420GB(2 週間相当)の処理結果では 3 分の 1 以上の処理時間の短縮を達成し，性能の効果が確認することが出来た．このことから Hadoop 内で行われるデータの分散処理時間の増加の割合が単体処理をし続けた場合の処理時間を上回り始めていることが分かる．

今後，SIDAS では HDF データの一括処理が今後行われる予定もあるため，今回試験的に行ったデータ件数，容量よりも多くの処理をすることが予定されている．そのことからも十分効果を期待出来ると思われる．しかし，今回使用した台数は全 5 台に対し，処理時間の短縮量は約 3 分の 1 であるため導入した台数比例の効果を Hadoop から得られていない．さらに，処理時間，処理の成功率に安定性があるとはあまり言えない．

予想される懸念点として，今回の検証でチューニングパラメータ設定部分を限定している点，全ての処理パターンにおいても動作を保障する最低限の変更と修正以外行っていない点である．また，検証データは HDF から RAW データ，つまり容量の大きいデータ(HDF)から容量の大きいデータ(RAW)への出力結果ではなく，容量の大きいデータ(HDF)から容量の軽い(テキストデータ)へ変換しているためメモリ内に格納する際の負荷率，スループットに変化が予想される点がある．より精度の高い設定と検証パターン，サーバの増設，サーバの導入台数以上の効果を得られるデータの件数と容量の検証を行っていくことでより正確なデータを求めていく必要がある．

4.3.2 評価結果 -適用性について

今回の検証において，Hadoop の処理で数回の処理失敗，処理の停止が見受けられた．表 4-4 に処理結果詳細履歴をまとめる。この原因としては，ヒープメモリの限界であるという警告などメモリ部分に対するエラー，もう 1 点は処理の失敗により Hadoop 内に発生してしまった不正な中間データ(メタデータ)が存在していたために起きているエラーである。チューニングの再設定，HDFS 内のデータを一度消去することで解決出来ているが，逐次的なチューニングや修正がかかせない点に不安定さと，SIDAS のような様々な画像処理やデータへの格納，変換が行われるシステムへの適用が非常に複雑で困難であることが予想される．

Hadoop にはその逐次的なチューニングや補正を動的に行い続け安定性と処理性能を引き上げる研究[1]も行われているため，実際に SIDAS で多様な処理を行う場合は，様々な現象を想定した実装が求められると思われる．

表 4-4 処理結果詳細履歴

処理パターン	処理結果
10GB	成功：2 分 28 秒
30GB -1 回目	失敗：ヒープメモリの不足
30GB -2 回目	成功：3 分 04 秒
50GB	成功：3 分 20 秒
90GB	成功：3 分 47 秒
100GB	成功：4 分 22 秒
210GB -1 回目	失敗：ヒープメモリの不足
210GB -2 回目	失敗：不正データの検出
210GB -3 回目	成功：5 分 40 秒
420GB -1 回目	成功：6 分 22 秒

また，Hadoop は全自動である分散ファイルシステム，サーバへのタスクの振り分け機能とは別に Map&Reduce のノウハウを持っていないと開発が容易ではないこと，MODIS データなどの特殊なデータを使用する場合は，Hadoop 内で使用するためのデータ構造のチューニングが必要であることが分かった．

5. まとめ

本報告における検証では，SIDAS で実際に行われる処理を切り出して試験的に評価した内容であるので，今後は厳密な効果の検証を実施する必要がある。しかし，今回の検証から単純なメモリ，HDD といったストレージの増設，サーバの台数を増やし，さらに検証を行うことでより Hadoop の性能を引出し，効果が得られる可能性は十分あると分かった．

一方，評価時に発生した処理の失敗やデータ毎の再調整

や修正から，様々なビッグデータを使用した処理においてHadoop導入の適用性については，様々なチューニングパターンが必要と思われる．

衛星データを使用した解析処理には一つ一つの容量が大きいデータを大量に使用する．さらに今後SIDASは，ソーシャルデータやオープンデータと言ったビッグデータの使用した新たな解析システムの開発をする案が考案されている．今後は，Hadoopを衛星画像解析やビッグデータ解析など，適応範囲を拡大するための研究を継続して行っていく予定である．

参考文献

1) Akihiro Nakamura, Jong Geol Park, Kenneth J. Mackin, Eiji Nunohiro, et al, ″Development and Evaluation of Satellite Image Data Analysis Infrastructure″, The Sixteenth International Symposium on Artificial Life and Robotics, Proceeding Index OS18-3, (AROB 16th’2011)
2) Hayao MORI 修士請求論文 衛星画像データ解析，システムの効率的利用に関する研究(2013)
3) 日立PCサーバーマガジン2013年6月号，http://www.hitachi.co.jp/products/it/server/portal/pcserver/magazine/itmedia/6th/
4) Apache Hadoop Project http://hadoop.apache.org/
5) Google, Inc. Map&Reduce http://staticgoogleusercontencom/media/research.google.com/ja//archive/mapreduce-osdi04.pdf
6) Google, Inc. Google File System http://staticgoogleusercontentcom/media/research.google.com/ja//archive/gfs-sosp2003.pdf
7) Hadoop徹底入門 第2版