ビッグデータ

内村 祐之，西山 大貴, 平野 裕也
Uchimura Yushi，Nishiyama Daiki，Hirano Yuya

1　はじめに

昨今，スマートフォンやタブレット等のデジタルデバイス，SNS および動画配信サービスの普及によって，インターネット上に様々なデータが存在している．また，各種センサの小型化・低価格化により加速度センサやカメラ機器等の配備が容易になった．これにより，ログデータや画像データの情報収集が格段に増加した．

このようなデータの量は年々倍増しており，ネット上のデータのうち半分以上がここ数年で生まれたものと言われている [1]．Fig. 1 にデータの増加状況及び今後の予想を示す．

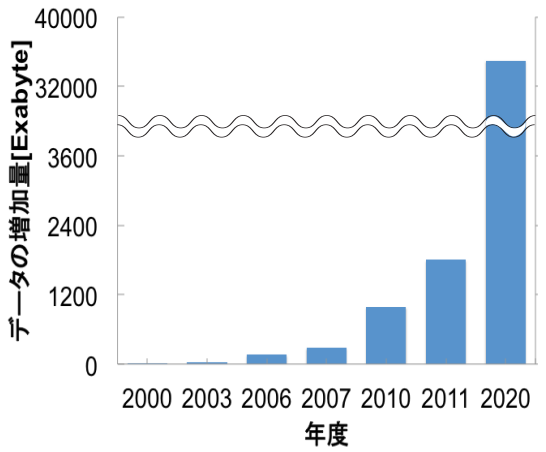

Fig.1　国際的なデータの増加状況及び予想

これら膨大なデジタルデータの総称がビッグデータである．

ビッグデータの処理のためには大量の記憶媒体とコンピュータが必要になる．クラウドが普及し大量の記憶媒体とコンピュータが利用可能になった．また，CPU の性能向上，記憶媒体の低価格も日々進んでいる．大量のデータが蓄積されるようになった現在，そのデータからいかにして価値を産み出すかが課題になっている．これがビッグデータに注目が集まるようになった背景である．本稿では，ビッグデータの概要，その処理技術および活用事例について述べる．

2　ビッグデータ

ビッグデータには明確な定義があるわけではない．大きなデータであればビッグデータと呼ぶこともあるが，以下の 3 つの要素が主な特徴として上げられる [2]．

- 多量性
- 非構造性
- 高頻度性

多量性とは，データの規模が大きいことを指す．ビッグデータの処理で扱うデータは，数 TB から数 ZB 規模である．

非構造性とは，データが構造化されていないということを指す．構造化されたデータとは従来の代表的なデータベースである RDB (Relational Database) に格納されて処理されるデータを指す．これに対して非構造化データは RDB で扱いにくい，文書，画像，音声およびセンサデータ等を指す．ビックデータはこれら両方のデータ構造からなる．

高頻度性とは，スマートフォンやタブレット等のデジタルデバイスや各種センサの普及によってデータが高頻度で入力・収集されることを指す．このようなデータを活用する場合，次々に収集されるデータをリアルタイムで処理しなければならない．

3　ビッグデータの処理技術

3.1　NoSQL/Not only SQL

表形式による RDB 管理システムとは異なる設計によって実装されたデータベースシステムである NoSQL は，センサやソーシャルメディア等から発生する非構造型データを含む多様なデータを大量にデータベース化するために利用される．

大量のデータを処理するために NoSQL では，トランザクション制御を犠牲にして同時並行処理を可能にした [3]．トランザクション制御とはデータの一貫性を保つための仕組みである．関連する一連の処理を 1 つの処理単位として実行することでデータから矛盾の発生することを防ぐことが可能になる．しかし，複数の更新要求を受け取った場合，トランザクション制御では一貫性を保つために 1 つずつ更新処理を行っていく．このため，RDB 管理システムでは，更新要求が増えればその分待たされる要求も増え，待ち時間も長くなる．

NoSQL では，一貫性を保持することをある程度諦めることで，待ち時間の解消を行っている．そのため，更新要求は即座に反映され，並列処理が可能となる．このような処理によってデータの不整合が発生する可能性があるが，NoSQL では更新処理方法を工夫することによって解決している．

データに対するアクセス方法は 4 種類ある．追加，読み取り，更新，削除である．このうち，更新処理にあたるのが，追加，更新，削除であるが，複数の更新処理が衝突するのは，更新と削除である．そこで，NoSQL では

更新と削除を極力行わず，追加を主とした更新処理で設計される．例をとると，100 という数字を誤って追加し，削除したい場合，NoSQL では 100 を削除するのではなく，-100 を追加する．このようにして NoSQL では最終的なデータの不整合を防いでいる [4]．

3.2 Hadoop

Hadoop とは，google が開発した分散処理ソフト「Google File System」と「Map Reduce」を模した Java ソフトウェア Framework である．専用のハードウェアを必要としないため，複数台の安価なサーバを連携させ数十 TB〜数 PB のデータを高速処理することができる．また，サーバを追加する際に基盤設計やアプリケーションに影響を及ぼすことなくスケールアウトすることができる．

Hadoop は分散ファイルシステムである「HDFS (Hadoop Distributed File System)」および分散処理を担う「Hadoop MapReduce」，データベース基盤となる「hBase」で構成される．

HDFS は 1 つのファイルを分散して保持する仕組みである．ファイルを分散する際に複製も同時に作成するため，障害が発生してもシステム全体に影響が発生しない．

MapReduce は 1 つの処理を分散して実行する仕組みである．MapReduce の処理は，入力されたデータから結果を得るのに必要な情報を抽出，加工および変換する「Map」および，抽出したデータを結合して並び替える「Shuffle」，並び替えたデータをまとめて結果を出力する「Reduce」という三つの手順で構成されている．MapReduce フレームワークには，分散処理に必要とされる共通機能があらかじめ実装されており，Hadoop 利用者は Map 関数と Reduce 関数の 2 種類を記述するだけで良い．そのため，複雑なプログラミングをすることなく大規模な分散処理を実現することができる．Fig. 2 に処理の概念図を示す．なお，矢印はデータの流れを意味する．

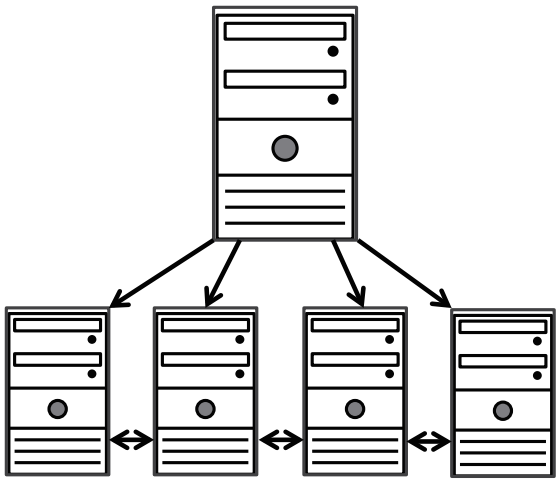

Fig.2　Hadoop による処理の仕組み

Fig. 2 のように，Hadoop は 1 台の親と多数の子によって構成されている．また，子同士でも通信するため，親に負担が集中しないようになっているのが Hadoop の特徴となっている．

「hBase」は 4.1 の節で述べた NoSQL の 1 つである．hBase ではデータを「キー」と「バリュー」の組み合わせで管理する．テーブル単位で管理する RDB に比べて非常にシンプルな構造になっている．このシンプルな構造によって，様々な形式のデータに対応できる柔軟性が得られる．

Hadoop は大量のデータを一括して変換するバッチ処理を得意する．そのため，大量に出力されたログファイルのような，巨大なファイルを集計する分析処理に適している．しかし，RDB のような特定の条件で絞り込むといった処理には適さない．また，円周率の計算のような計算結果に対して計算を積み重ねるような処理にも適していない．

4　ビッグデータの活用事例

新聞社である New York Times 社では，過去の掲載記事を PDF ファイル形式でユーザに提供するサービスを展開している．サービスを始めるに当たり，4 テラバイトにも上る 130 年分の過去記事をいかに効率良く PDF ファイルに変換するかが課題であった．

この課題に対して，同社では Hadoop を活用することで，4 テラバイトの画像データを，24 時間で約 1.5TB の PDF に変換した．これにかかった費用は 240 ドルであった [5]．

5　今後の展望

国際的なデータ量は，Fig. 1 で示したように，今後も増加する．このようなデータの洪水の中ではよりビッグデータの解析技術の向上及び解析する人財が求められる．

また，3.1 節では，分散処理のためにトランザクション制御を犠牲にした NoSQL について説明した．しかし，トランザクション制御無しでは処理できない分野も存在する．こういった分野のビッグデータを処理するために NoSQL と RDB を両立する NewSQL という概念も登場している．今後は NewSQL といった新しい技術を利用して，より広い分野でのビッグデータの解析が進むと考えられる．

参考文献

1)「ビッグデータ」とは？今知っておきたい旬キーワード！ http://smmlab.aainc.co.jp/?p=8630.
2) ビッグデータの研究開発推進の注目点. https://dspace.jaist.ac.jp/dspace/bitstream/0119/10980/1/kouen27_84.pdf.
3) よくわかる nosql 入門. http://itpro.nikkeibp.co.jp/article/COLUMN/20120828% slash418821/?ST=bigdata&P=3.
4) Rdb 開発者におくる nosql の常識（5）. http://www.atmarkit.co.jp/ait/aarticles/a1106/a23/anews117.html.
5) 第 1 回 hadoop で広がるビジネス領域. http://gihyo.jp/admin/serial/01/hadoop/0001.