ビックデータ

北村　一峰，中林　弘光，内村　祐之
Kazutaka KITAMURA，Hiromitsu NAKABAYASHI，Yushi UCHIMURA

1　はじめに

この数年間でスマートフォンおよびタブレットなどのモバイル端末を利用して日時・場所を問わずインターネットを利用できる時代となった．これらの機器の発展とネット上にソーシャルサービスなどが登場したことにより，個人または企業がリアルタイムに情報を発信するようになった．これにより，多量のデータが生み出され，データの総量が爆発的に増大し続けている．インターネット上に存在するデータの 90 %はこの 2 年以内に生成されたものである．蓄積されたデータの総量は 2 年ごとにほぼ倍増し，10 年後には 50 倍にも膨らむと言われている [1]．これらの膨大なデータは，コンピュータが管理を得意とする数値的な構造データだけではない．ウェブのアクセスログや画像，動画のようなコンピュータが管理を不得意とする非構造化データが 80 %に及ぶ．前述したような大容量で多様なデータがビックデータとして注目されている．企業は蓄積しているビックデータから新しい知識を探し出すことが課題となっている．本稿では，ビックデータの概要，処理方法，活用事例および今後の展望について述べる．

2　ビックデータの概要

ビックデータの共通定義は詳しく決まっていない．ここでは次のようにビックデータを定義する．ビックデータとはインターネットの普及および IT 技術の発展のため急増した大容量かつ多様なデータである．ビックデータの特性として下記が挙げられる [2]．

- 大容量なデータ
  ビックデータという言葉が表すように，データ容量の巨大さはビックデータの重要な要素である．テラバイトからペタバイト以上の膨大なデータ量のデータがビックデータと言われている．今までデータを管理する上で一般的に使用されてきた RDBMS（リレーショナルデータベース）で管理することが容易でない巨大なデータである．
- データの多様性
  これまで企業で管理・処理してきたデータは，構造化データと呼ばれる数値および文字列のような表形式の定型にはまるデータであった．しかし，近年扱うデータは，文書，画像および動画などのデータ，ウェブサーバのログデータなどが多い．これらのデータは，形式が一様ではなく非構造化データと呼ばれる．非構造化データはデータの種類が多様なため，処理が容易でない．
- データのリアルタイム性
  これまでのデータ管理・処理は，保存された過去のデータを対象の中心としてきた．ウェブサービスの発展やデジタルデバイスやセンサの高性能化のため，高頻度に情報が入力・収集されるようになった．

ビックデータは，前述した 3 つの特性を持っている．これらの特性のため，今までのデータ処理では扱うのが容易ではない．そこで，ビックデータを解析するために新しい技術が開発された．

3　ビックデータの処理方法

3.1　RDBMS と NoSQL

RDBMS　　行と列の表形式で管理される一般的なデータベースシステムである．構造化データを扱うために使用され，最も標準的なデータベースである．複数のテーブルを連結させることでデータを管理する（正規化）．正規化により，データの重複をなくすことができデータの整合性を保てる．また，格納するデータの総データサイズや更新データの書き込み量などを最小化することもできる．RDBMS にはトランザクション制御と呼ばれる重要な処理が存在する．トランザクションとは更新および計算などを行う処理の単位のことである．Fig. 1 にトランザクション制御の処理イメージを示す．

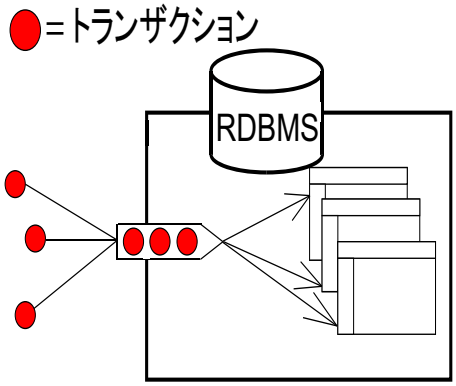

Fig.1　トランザクション制御のイメージ

図で示すように，トランザクション制御とは複数の互いに関係のあるトランザクションをまとめ，1 つずつ処理しすることである．トランザクションのすべてが完了またはキャンセルされるように制御する．これにより更新データに不整合が生じず，RDBMS ではデータの一貫性を保っている．また，RDBMS ではデータベース内でデータの定義，操作，および

制御を行う SQL 言語が存在する．この SQL 言語を使用することで，条件を指定しデータの検索を行うことができる．

NoSQL　RDBMS の特徴である汎用性やデータの一貫性を犠牲にすることで，扱うことが容易でなかった大量の非構造化データを処理できるように設計されたデータベースシステムである．複数のサーバにデータを分散して管理・処理することで，保存領域の拡大および処理能力の向上を実現した．一般的な NoSQL では行と列を使用せずに，データにキーを付与することで管理する．データ処理を行う際は，付与されたキーが入力されると，データを読み出すようになっている．これにより，データの形式が一様ではない非構造化データを一様に扱うことができる．しかし，RDBMS には存在した SQL 言語を使用できないため，RDBMS のように細かな条件を指定し検索することができない．高速処理を実現するために，前述したトランザクション制御を行わない．Fig. 2 に NoSQL におけるデータ処理のイメージを示す．

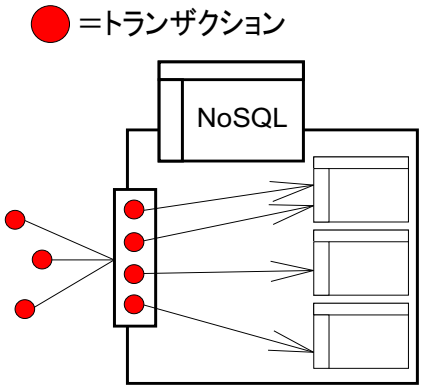

Fig.2　並列処理のイメージ

図で示すように複数のトランザクションを並行して処理する．高速処理を行うことができるが，処理順序に応じてデータに不整合が生じる恐れがある．このため，データの一貫性が低下している．

3.2　ビックデータと NoSQL

NoSQL はビックデータを扱うために生まれたデータベースである．しかし，ビックデータの管理が NoSQL が適しているというわけではない．数値や文字列などの構造化データの処理は正規化が行われる RDBMS の方が適している．しかし，はじめにで述べたように現在のデータの 80 %は非構造化データと言われており，非構造化データを扱うのはキーで管理できる NoSQL が適している．また大容量なデータを処理するのにも，分散処理を行うことができる NoSQL が適している．つまり，構造化データの処理は RDBMS で行い，非構造化データの処理は NoSQL で行う．こういったようにデータベースを使い分けるのがビックデータの最も適切な処理である．

4　ビックデータの利点と活用

ビックデータを活用する利点は，蓄積してきた膨大なデータから今まで見い出せなかった規則性を見い出すことができることである．様々な分野の企業がビックデータを持っており，企業にあった活用方法を考えることが重要である．具体的な例として，金融・保険業界の場合，取引傾向および不正取引などの解析が存在し，製造・販売業界の場合，品質および需要の解析が存在する [3]．Fig. 3 に様々なビックデータの活用例を示す．

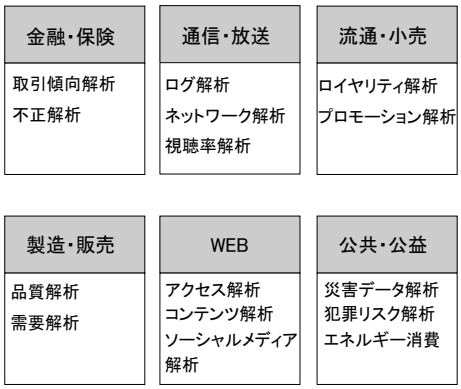

Fig.3　ビックデータの活用例

これらの活用例のように様々な用途に活かすために，ビックデータの解析が行われている．しかし，これらの活用例はビックデータが注目される前から統計的な解析として行われてきた．今までの解析とビックデータの解析の大きな違いは，解析に使用するデータの量，種類およびリアルタイム性にあるといえるだろう．今までは処理が容易でなかった非構造化データをリアルタイムに処理できることが即時に反応が求められる場面で重要になる．

5　今後の展望

大容量で多様な蓄積データから，活用方法を見つけることが今後も重要である．現在は蓄積したデータを処理できず，ビックデータを活用できていない企業が多い．3.1 節に示した処理方法である RDBMS と NoSQL の両方の特長をもつ NewSQL という技術が存在する．この新しい処理技術の発展により，様々な分野で蓄積されているビックデータを分析する企業が増えると考えられる．ビックデータの解析により今まで見い出せなかった有益な価値を見い出し，サービスの質の向上に活かしていくと考えられる．

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) Application-delivery-strategies. http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.
3) ビッグデータへの道 第 2 回「ビッグデータの活用範囲」．http://www.hitachi.co.jp/products/it/bigdata/column/column02.html.