ビッグデータ解析の現在と未来―Hadoop，NoSQL，深層学習からオープンデータまで―
原 隆浩著・喜連川 優コーディネーター
http://www.kyoritsu-pub.co.jp/bookdetail/9784320009202

AI時代を生き抜くにはその震源地ともいえるビッグデータを正しく理解することが必須

コーディネーター　喜連川　優

ビッグデータという名称はとりわけ2012年3月オバマ政権によるビッグデータイニシアティブ開始以降より広く利用されるようになった．本書の出版は5年半を経ているが米国並びに世界はこの方向感をさらに推進しており，短時間に使い捨てられるITバズワードではない．

ビッグデータは大きく二つのパートから構成される．データの収集とデータの解析である．データの収集において重要な役割を果たすのがIoT（Internet of Things）である．センサー技術ならびに半導体技術の進展により多くの事象が可観測となり容易にデータを収集することが可能となり，これが膨大なデータの新たな生成源となりつつある．データの解析部分は従来データアナリティクスと呼ばれていたが，最近はディープラーニングが生まれて以降AIブームとなり，解析部分がAIと呼ばれることが多い．3度目のブームを迎える今日のAIの最大の特徴は，膨大な学習データが利用可能になったという点にある．従来は，大量データはほとんどなかった．すなわち，今日のAIとビッグデータは不可分な関係にある．このように最近の3つのITキーワード，IoT，AI，ビッグデータは非常に近い領域を指しているといえる．あるいは，遠目には，おおむね一括りの技術といっても過言ではない．もちろん，IoTはセンターではなくアクチュエータとしての利用もあり，また，AIも必ずしも多量ではなく少量のデータの解析もある．しかし，大勢と

しては，上述のようにビッグデータという大きな幹が IT の中核として位置づけられるようになったことは確実である．このような流れを理解することは極めて重要であり，本書は全体の技術動向をうまくまとめている．

ビッグデータというブームの 8 年前に我が国では，情報爆発というキーワードで，文部科学省は大規模な研究プロジェクトを企画し，2005 年より，ビッグデータとほぼ同様の研究を本格的に開始していた．2007 年には経済産業省は情報大航海なるプロジェクトを立ち上げ，企業を主たるプレイヤとした国家プロジェクトが開始され，数多くの興味深い成果が生まれている．すなわち，日本は遅れているわけではない．本書から，そのような流れも汲み取れよう．

現時点ではさらに新しい研究の潮流が生まれつつある．社会には多くの IT をフルに活用した社会システムが稼働しているが（これらは algorithmic system と呼ばれる），当該システムが公平で倫理的に正しいサービスを社会に提供することをいかにして保障するかが活発に議論されている．アルゴリズムとデータの両方がフェアでなくてはならない．アルゴリズムが正しく動作するには，元となるデータが歪んでいないことを保証することが必須となる．すなわち，ビッグデータの質担保が極めて重要な要件となり，そのようなデータのデザインを研究する時代へと突入している．量が多ければよいというビッグデータの時代からさらに次のステップへ飛躍しようとしている．

制度面の整備も重要であるが，十分に進んでいるとはいえない．データは著作権の対象とはならない．データの保護は現時点では不正競争防止法によることとなる．一方で，データが大きく世界を変える中で，企業活動が適正になされるべく新しい制度についても知財戦略として議論が進められており，今後が期待される．

本書はビッグデータを中心に据えて，ビッグデータそのものの歴史や現状，さらにはその周辺の諸技術について俯瞰的に解説した希少な書籍といってよい．

以下では，本書の概要について，章構成に従って概要を示す．1章では，ビッグデータおよびその解析技術について概観する．とりわけ，ビッグデータの特徴や注目されるようになった背景，ビッグデータの解析技術の特徴（従来のデータ解析とは何が違うのか）などについて解説する．この章を読むことにより，生成されるデータ量の大量化・多様化だけではなく，それを解析するためのデータベースや機械学習，分散処理フレームワーク，クラウドサービスなどの諸技術や環境の成熟・発展が重要なきっかけになっていることが示される．

2章では，ビッグデータ解析の応用事例について，米国大統領選挙や都市部の人流解析，防災・災害対応，Yahoo! Japanのビッグデータレポートなど，代表的な事例を紹介し，私が領域代表者を務めて2005年度～2010年度に実施したビッグデータプロジェクトである「情報爆発プロジェクト」（文部科学省科学研究費特定領域研究「情報爆発時代に向けた新しいIT基盤技術の研究」）についても紹介する．著者の原氏は当該プロジェクトで活躍された．本章を読むことにより，ビッグデータ解析が実際にどのような応用に用いられるのか，さらには，ビッグデータが流行する遥かに前から我が国ではビッグデータの重要性が認識され，国を挙げて技術開発が行われていたことを知ることができるであろう．

3章では，ビッグデータ解析の典型的な流れとして，「データ収集」と「データ解析」の二つのフェーズに焦点を当て，各フェーズにおける処理内容や考慮すべき課題，用いられる技術などについて

解説する．本章により，ビッグデータ解析の全体像を把握でき，本書が対象とする範囲が把握できるであろう．

4 章～7 章では，ビッグデータを支える諸技術として，分散処理フレームワーク，ストリーム処理エンジン，データベース，機械学習について解説する．これらの章は，本書の技術書としての根幹の部分といえ，まず 4 章では，大量のデータを超並列分散処理によって高速に処理するための分散処理フレームワークについて，代表的なシステムを紹介する．ビッグデータがこれだけ注目されることになったのは，これまで不可能と考えられてきた大規模なデータ解析が，Hadoop などの分散処理のオープンフレームワークの登場によって，従来高価であったライセンスフィーがなくなり，広く利用が進展したことがわかる．本章では，Hadoop 分散ファイルシステムや MapReduce，YARN などの Hadoop 関連技術に加えて，最近特に注目されている Spark，Storm，Flink，Mahout などの Apache ソフトウェア財団によるオープンソースフレームワークや Jubatus，Spatial Hadoop など，多様な広がりについて紹介し，その特徴を紹介する．本章により，多数存在する分散処理フレームワークの全体像を理解することができよう．

5 章では，M2M や IoT，各種センサーや監視機器などから連続的に発生するストリームデータを対象として，ストリームデータを効率的に処理するための技術基盤（エンジン）について，学術レベルから商用レベルまで代表的なシステムを紹介する．ストリーム処理の歴史は実は古く，2000 年前後からセンサーネットワークの分野とデータベースの分野を中心に研究が進められてきた．本章では，そのような歴史のあるストリーム処理の技術が，なぜ最近のビッグデータの流行時に再び脚光を浴びているのかを含めて，ストリーム処理エンジンのアーキテクチャや関連技術を俯瞰的に解説する．特

に，ストリーム処理のための代表的な問合せ言語である CQL（連続問合せ言語）について，リレーションのデータとストリームのデータを相互変換するための演算（例えば，ストリームをリレーションに変換するウィンドウ演算）と問合せの例を紹介する．また，集中型と分散型のエンジンについて，それぞれのアーキテクチャや代表的な例について説明する．本章により，Spark や Storm などでその存在が知られていたストリーム処理について，その歴史や全体像を把握することができる．

6 章では，ビッグデータ解析のためのデータベース技術として，NoSQL と呼ばれるデータベース群について解説する．リレーショナルデータベースの長所であるスキーマ定義や正規化・結合処理，ACID 特性の保障がビッグデータ解析では必ずしも必要でない場合があることを指摘し，それを解消・緩和した NoSQL データベースについて，いくつかの分類と各分類での代表的なものを紹介する．一方，最近の動向として，多くの NoSQL データベースにおいて，従来のリレーショナルデータベースで用いられている技術（上記の長所となる機能）が拡張機能として実装され始めていることを紹介している．これは，従来のデータベース技術者と，そのコミュニティ外からの NoSQL の技術者（ビッグデータ解析者）の歩み寄りを表しているともいえ，旧来のデータベース要件がかなり本質的であることから，避けられない様相を呈しているともいえる．本章を読むことにより，NoSQL の現状と今後の方向性について，理解を深めることができよう．

7 章では，ビッグデータの解析技術として重要な機械学習と，その中でも特に最近注目されている深層学習について概説する．4 章～6 章で紹介した技術がビッグデータのデータ基盤のための技術であるのに対し，機械学習はデータ解析を行うための技術である．こ

の章において，AI の隆盛・衰退を繰り返す歴史にも触れつつ，代表的な機械学習手法のクラス（教師あり，教師なし）とその中での代表的な手法について，SVM，ニューラルネットワーク，決定木，クラスタリング，トピックモデルを概説する．さらに，深層学習のための代表的な技術に関して，畳込みニューラルネットワークや，自己符号化，再帰型ニューラルネットワーク，LSTM などについて説明する．ビッグデータの流行と AI の流行について関係性にも触れる．

8 章では，関連する動向として，最近新たなブームになっているオープンデータに関し，国内外の動向と，オープンデータ化がそれほど進まない現状の問題点について解説する．さらに，オープンデータをビッグデータ解析に用いた際に生じる課題について，プライバシーや信頼性，オーナーシップ・トレーサビリティなど多角的に紹介している．ここでの議論は，データベースやデータ解析の研究開発に携わっている研究者・技術者にとっても，最先端の領域といえる．

最後に 9 章では，本書のまとめとして，ビッグデータに関する将来展望について議論する．オープンデータの効果的な利用を促進するプラットフォームや，人に関わる，人を介したデータ処理について，諸課題とその解決に必要となる技術を解説する．

このように，本書はビッグデータに関する広範囲にわたる多様な技術をすっきりとまとめており，関連する技術を概括的に理解することが可能となる．ビッグデータ解析に関しても，最近の AI について全体を俯瞰しており，容易に理解が可能となる．「データが世界を変える」という今世紀の大きな潮流を感じることのできる書物であり，当該分野へ挑戦する方々が増えることを希望する次第である．