2 ビッグデータ関連プログラム —米国とEUにおける動向—

山名 早人（早稲田大学）

政府による研究開発投資

今，世界でビッグデータ関連の研究への投資が加速している．ここでは，2012年から始まった米国のビッグデータイニシアティブと，2014年から始まったEUのHorizon 2020でのビッグデータ関連研究開発について，コンテンツ処理技術にかかわる基盤的なプログラムを中心に紹介する．

米国における動向

ビッグデータイニシアティブ

米国大統領府の科学技術政策局（OSTP）は，2012年3月に「ビッグデータ研究開発イニシアティブ」を発表し☆1，6政府機関（NSF，NIH，DOE，DOD，DARPA，USGS）に対して新規に2億ドル以上を投じ，大規模ディジタルデータへのアクセス，保存，そして発見をサポートするツールおよび技術を飛躍的に進歩させることを明らかにした．

この発表の中で，大統領科学技術諮問委員会（PCAST）のメンバでありOSTP所長でもあるJohn P. Holdren博士は「高性能計算やインターネットでの革新的な進歩が政府の情報技術研究開発により導かれたように，科学的発見，環境・生命科学研究，そして教育，安全保障の各分野において，本イニシアティブによってビッグデータの利活用が進むであろう」と語っている．同報告書☆2では，ビッグデータ処理基盤から目的指向のものまで多種多様なプログラム89件が記載されている．

以下では，コンテンツ処理技術に関連する基盤研究に資金を提供しているNSF，DARPA，DOEが推進するプログラムを紹介する．

■ NSF

米国国立科学財団（NSF）における本イニシアティブ関連予算の中で最大のプログラムは，情報科学工学局（CISE）が，2012年4月，カリフォルニア大学バークレー校（UCB）に総額1,000万ドル（5年間）を投じたAMP（Making Sense at Scale with Algorithms, Machines and People）である．

AMP以外にもビッグデータ科学・工学の発展に寄与するコア技術開発に関連し，2012年に2,500万ドル（NIHと共同実施），2014年に2,300万ドル，2015年に2,650万ドルの公募☆3が実施され，プログラム当たり年間20～50万ドルが3～4年間，全米の主要大学を中心に支出されている．2014年までの公募において，単一のプログラムとして最大のものは，ワシントン大学のMyria☆4であり，3年間で総額296万ドルの資金を受け，分散非共有ビッグデータ管理システムおよびクラウドサービスの研究開発が実施されている．

■ DARPA

国防高等研究計画局（DARPA）では，大規模データからの異常検知（Anomaly Detection at Multiple Scales, ADAMS），サイバー脅威検知（Cy-

☆1 https://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release.pdf
☆2 https://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_fact_sheet_final.pdf
☆3 2013年に公募がないのは2012年に開始された公募の採択が遅れたためだと考えられる．実際2012年中にスタートしたプログラム数は6件と少ない．
☆4 http://myria.cs.washington.edu/

ber-Insider Threat, CINDER）等，国防にかかわる内容を中心に研究開発が実施されている．一方でコンテンツ処理にかかわる技術として注目されるものに，PROCEED と XDATA がある．

PROCEED（Programming Computation on Encrypted Data）は，完全準同型暗号を用いた秘匿計算，すなわち「暗号化したままの状態で四則計算を進める手法」が実用化できないのは，通常計算に比較して 10 桁遅くなるためであることを問題として挙げ，これを 10 万倍高速化することを目指している．また，XDATA は，データ解析ツール開発を目指しており年間約 2,500 万ドルが 4 年間にわたり投じられている．

■ DOE

エネルギー省（DOE）は，総額 2,500 万ドルを投じて Scalable Data Management, Analysis and Visualization Institute（SDAV）と呼ばれる新たな研究機関を新設し，ローレンス・バークレー国立研究所を中心に他の 5 研究所，7 大学と連携し，超大規模データの「管理」「分析」「視覚化」を補助するツール構築を実施している．

「管理」では，計算機能力の向上にディスク性能が追いついていないことを挙げ，たとえば高エネルギー物理学データに代表される数百億を超える超大規模データから特定のパターンを高速に検索する手法で 100 倍以上の高速化を目指す研究が進められている．「分析」においては，スーパーコンピュータから得られる膨大なシミュレーション結果をそのままディスクに書き込むのではなく，不要な部分を事前に削除し必要な部分のみを保存するためのインシチュ処理を行う手法の開発が行われている．「視覚化」では，さまざまなパラメータにより得られた多様なシミュレーション結果を時空間上で分析者に分かるかたちで表示する手法について研究開発が進められている．

■ プログラム事例

プログラム事例として，予算規模が大きい AMP と XDATA について紹介する．

AMP

AMP[☆5] は，UCB の Michael Franklin 教授を筆頭に 3 名の Director，教授，学生等を含め総勢 70 名体制で推進されている．AMP では，3 本柱として「アルゴリズム」「コンピュータ」「人」を掲げ，新しいデータ解析パラダイムの構築を目指している．

「アルゴリズム」の柱では，機械学習アルゴリズムのスケーラビリティ・正確性・効率性の向上を目指し，「コンピュータ」の柱では，倉庫大データセンタ（Warehouse Scale Computing, WSC）を実現するとともに，1 台のコンピュータのごとく利用できる仕組みを掲げている．そして，「人」の柱では，クラウドソーシング（本特集の「クラウドソーシング」を参照）に代表される人間とコンピュータの相互作用による協調を掲げている．

AMP では，**図 -1** に示す BADS（the Berkeley Data Analytics Stack）を提案し同モジュール開発を行っている．ポスト Hadoop とも言われる Apache Spark（AMPLab の前身である RADLab により開発）も組み込まれており，Hadoop で問題となる処理単位ごとでのディスクアクセスを RDD（Resilient Distributed Dataset）と呼ぶ独自のキャッシュ機構で最適化し高速化している．機械学習をターゲットとした評価では，10 ～ 100 倍の高速化を達成している．さらに，グラフ処理に特化したパイプライン処理による最適化を行うことでグラフ処理の高速化を達成する GraphX もある．現在構築中の MLBase は，分散環境下での機械学習を簡単，かつ効率よく行える仕組みの構築を目標としている．そして，Velox では，機械学習の一連のサイクルである「学習」「学習結果の利用」「直近データを用いた学習の更新」という実運用時のサイクルを効率化する仕組みを提案しており，実運用を目指した研究として興味深い．

XDATA

2015 年 4 月に開催された第 56 回 HPC User Forum での Wade Shen 教授（DARPA）の講演に

☆5 https://amplab.cs.berkeley.edu/

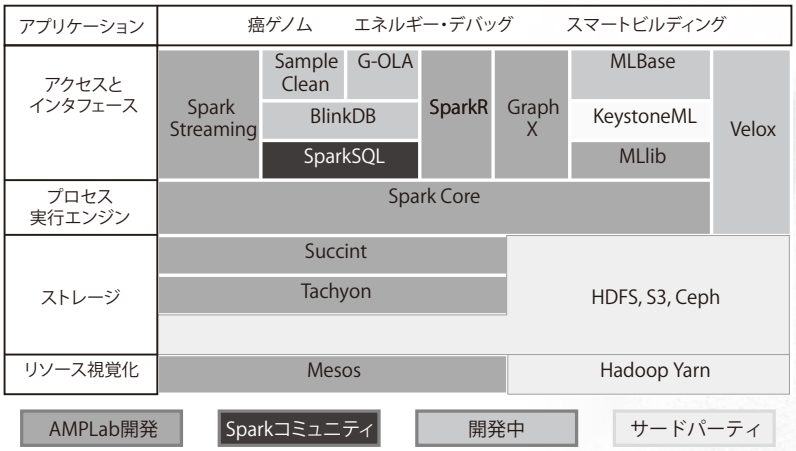


図-1 BADS（バークレー・データ分析スタック）☆6

よれば，XDATA は 28 の組織で構成されており，大学からは，UCB，CMU，ハーバード大学，スタンフォード大学等が参加している．XDATA は，「大規模なデータを対象に分析を行う場合，そのデータの性質や規模によって，十分な効率性を達成できない」ことを課題に挙げ，基盤技術の部分から新しいアプローチで並列分散処理やインタラクティブな可視化を実現するツールを構築し，オープンソースとして公開することを目指している．

XDATA の成果（DARPA の OPEN CATALOG☆7 から参照可能）としては，UCB の AMP ラボが構築した，Hadoop のインメモリ版とも言える「Spark」（前述），コンティニュアム・アナリティクス社が構築したビッグデータ対応の Python（CPU/GPU を使った自動ベクトル化をサポート）「Anaconda/Numba」，キットウェア社が構築した，データを視覚化するための Web フレームワーク「Tangelo」，アンチャーティッド社が開発した，Hadoop や Spark から出力される大規模データをスケーラブルに視覚化できる「Tiles」（**図-2**）など，実用的なソフトウェアが数多く公開されている．

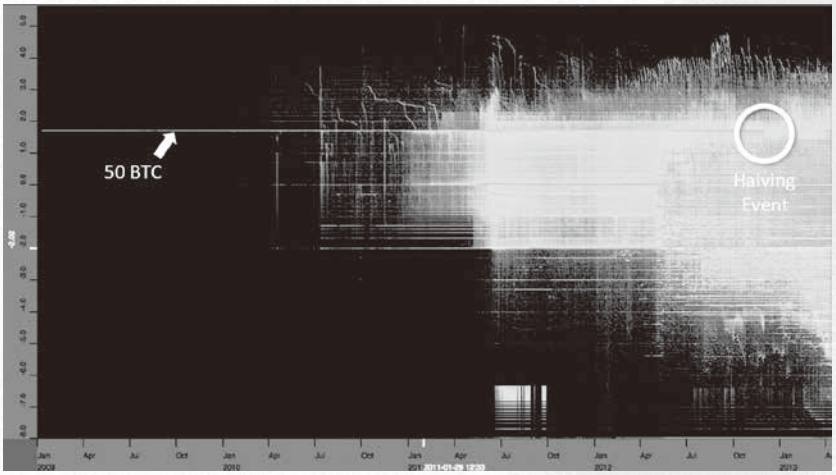
ビッグデータとプライバシー

2012 年以降に大統領科学技術諮問委員会がビッグデータに関連して発表した報告書に「ビッグデータとプライバシー：技術的観点から」（2014 年 5 月）がある．同じ 5 月には，大統領顧問である John Podesta 氏を中心としたワーキンググループによる報告書「ビッグデータ：機会を逃さず価値を保護する」も公開された．同ワーキンググループの報告は，ビッグデータの価値を最大化し，リスクを低減するための法制度を中心とした検討であるのに対し，PCAST からの報告は技術的側面からビッグデータがプライバシーに与える影響を分析している．特に，「ソーシャルメディア，画像，映像，医療などさまざまなデータを対象としたビッグデータ解析が新ビジネスを生み出すのはもちろんであるが，プライバシー保護のための研究が重要である」ことを指摘している．

PCAST からの報告書で述べられている提言は次の 5 つである：1）政策では実際の利用側面を考えるべきである，2）政策や規制は特定の技術に依存すべきではない，3）連邦政府のネットワーキングおよび情報技術研究開発（Networking and Information Technology Research and Development, NITRD）プログラム関連機関は，プライバシー関連技術とそれらの技術の成功を導く社会科学分野の研究を強化すべきである，4）プライバシー保護に関する教育・トレーニングの増強をすべきである，5）現在利用可能なプライバシー保護技術の利用を促進し国内外で主導的役割を果たすべきである．

☆6 https://amplab.cs.berkeley.edu/software/ に掲載されている構成図を基にトレース．
☆7 http://opencatalog.darpa.mil/

こうした提言を受け，2015年の予算教書では，NITRDプログラムにおいて，個人のプライバシーを適切に保護するとともにビッグデータから価値を引き出す研究に焦点が当てられた．また，2016年の予算教書に対するNITRDの補足資料では，2015年1月に開催されたビッグデータ戦略イニシアティブワークショップを踏まえ，2015年秋に「ビッグデータ・リサーチ・アジェンダ」が公開されることが述べられている．なお，ビッグデータとプライバシーについては，本特集の「ビッグデータ活用におけるガバナンス」を参照されたい．



約3,700万件の表示（x軸は時間，y軸は取引単価，輝度は取引量．白丸の部分は，マイナーの報酬が50BTCから半減した瞬間を表す．斜め右下の取引単価が小さく輝度が高い短形部分（取引量多）は，ビットコインへのDoS攻撃を表す）．

図-2　ビットコイン運用開始から約3年半の取引☆8

EUにおける動向

✣ HORIZON2020

欧州では，2013年まで実施されていたFP7（第7次研究枠組み計画）に続くHorizon 2020（2014～2020）を通じて研究開発が進んでいる．Horizon 2020は，2010年に策定された「ヨーロッパ2020戦略」での革新的欧州連合を実現するEU最大の研究開発プログラムであり，2020年までの7年間で総額約800億ユーロが投資される．同プログラムの柱として，1）科学的卓越性，2）産業における先導性，3）社会的挑戦への取り組み，が掲げられており，基礎研究が「科学的卓越性」に，プロトタイピングが「産業における先導性」，出口を見据えた活動が「社会的挑戦への取り組み」に対応すると考えることができる．

Horizon 2020では，オープンデータポータルのさらなる整備とともに，複数国家間かつ産学官を含む複数のパートナー間の連携による研究開発に力を入れており，ビッグデータ関連研究への投資は，これら3本柱のいずれにも含まれている．一方で，公募名にビッグデータを冠しているものは，「産業における先導性」に含まれるICT技術分野「コンテンツ技術と情報管理」内の「ICT-15-2014：ビッグデータ革新と利用」（5,000万ユーロ）と「ICT-16-2015: ビッグデータ研究」☆9（3,900万ユーロ）の2つである．

ICT-15-2014では，中小企業への技術移転，オープンデータの利用・流通，多言語データサービス・製品を構築する上での技術移転を目指している．公募では106件の応募があり，評価点が基準以上であった54件の中から13件が採択された☆10．一方，ICT-16-2015☆11では，ビッグデータを対象とした解析，データ管理，予測，可視化を行うための革新的なデータ構造，ソフトウェアアーキテクチャ，最適化，言語理解の研究が公募対象としており，より基盤技術に近い．また，同公募では，マルチモーダルで多言語といった多様なデータを対象としたリアルタイムなクロスストリーム解析を特に明示している点に特徴がある．

☆8 2015年3月開催のSpark Summit EastにおけるUncharted Software社の講演資料を同社の許可のもと転載．
☆9 http://ec.europa.eu/research/participants/portal/desktop/en/opportunities/h2020/topics/9084-ict-16-2015.html
☆10 http://cordis.europa.eu/（2015年5月末時点）
☆11 2015年4月に公募が締め切られ，本稿執筆時点では選考中．

✣ プログラム事例

プログラム事例として，2015年から実施が始まったICT-15-2014の13件の中から予算規模が最も大きいODINEとその他のプログラム，そしてFP7からForgetItの事例を紹介する．

ODINE

ODINE[☆12]には，2015年2月からの3.5年間で総額822万ユーロが拠出される．ODINEの中心拠点は，英国サウスハンプトン大学である．同大学は，Tim Berners-Lee教授（WWWの仕組みを考案し現在のインターネットの基礎を築いた人物）とNigel Shadbolt教授が所属する大学であり，両名により2011年にODI（Open Data Institute）が設立され，英国におけるオープンデータの拠点となっている．ODINEは，中小企業によるオープンデータの利活用をサポートするプログラムであり，インキュベータとしての役割を担う．

他のICT-15-2014プログラム

その他のICT-15-2014での採択プログラムは，1）オープンデータの流通を加速することを掲げるBigDataEurope，EuDEco，2）多言語処理や意味理解のオープン基盤を目指すFREME，3）ソーシャルメディア等の感情分析を対象としたMixedEmotions，SSIX，4）産業への出口イメージを明確化したBISON（コンタクトセンタの電話音声をマイニング），proDataMarket（政府が管理する土地情報の市場化），AquaSmart（水産業界でのデータ流通），KConnect（医療関係のテキストデータ処理），AEGLE（医療関連各種データの高速処理によるパーソナライズされた医療提供），AutoMat（自動車から得られるデータの価値化），5）データサイエンティスト育成を目指すEDSA，の5つに分類することができる．

これらのプログラムに共通している点は，いずれも，公募が掲げる中小企業への技術移転やオープンデータの利用・流通の加速を狙っている点である．また，各プロジェクトの中心拠点は必ずしも教育研究機関ではなく，たとえば，AutoMatはフォルクスワーゲンが中心拠点となっている．AutoMatでは，これまで有効活用されてこなかった自動車の車載LANであるCAN-Busから得られる毎秒4,000個にもなる信号の自動車会社での活用，さらには自動車会社を超えたオープンデータとしての活用を目指している．

FP7でのプログラム事例

2013年まで公募が行われていたFP7から，特徴的なプログラムとして，ForgetIT[☆13]を紹介する．ForgetITは，2013年2月から3年間，総額908万ユーロの予算により，ライプチッヒ大学を拠点とした合計11組織から，情報，データ解析，個人情報保護，クラウドコンピューティング，法律，経済などの専門家が加わり研究が進められている．目指すゴールは，柔軟な「保存・忘却フレームワーク」（**図-3**）の構築である．

ビッグデータ時代，マルチメディアデータを含めあらゆるデータが保存されている．一方，将来に渡って利用されないデータや，ある時期を過ぎると急速に価値がなくなるデータがある．人間の記憶は時間とともに薄れていくが，何かのきっかけで思い出すこともできるし，逆に完全に忘れることもある．こうした人間の記憶のメカニズムを参考に，将来のデータ保管のあり方について研究が進んでおり，IBMを中心に「PoFミドルウェア」の開発が行われている．なお，直接の説明はないものの，EU個人データ保護規則案第17条「忘れられる権利」に関連しているとも言える．

ビッグデータ研究の今後

これまで見てきたように，各国においてビッグデータ研究が進んでいる．ソフトウェアについてはオープンソース化，データについてはオープンデータ化が主流である．DARPAのOPEN CATALOGにおけるソフトウェア公開をはじめとした各種ツール公開は，本分野の研究開発に大きく貢献するもの

☆12 http://opendataincubator.eu/
☆13 http://www.forgetit-project.eu/

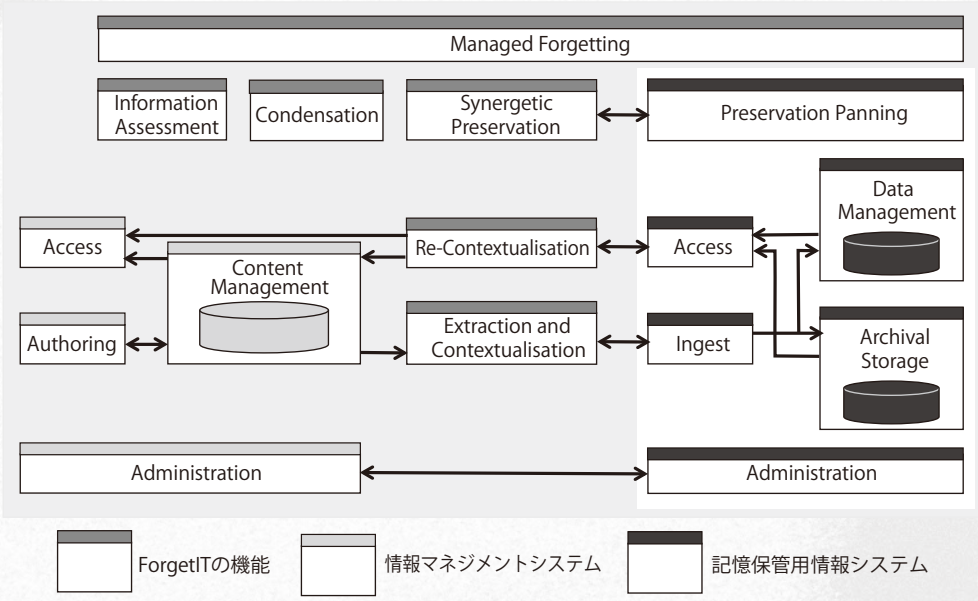


図 -3　ForgetIT のアーキテクチャ☆14

と考えられる．また，将来のビッグデータのあり方を考えさせてくれる ForgetIT や，DARPA での高速秘匿計算技術開発，PCAST 報告書でのプライバシー関連技術の重要性指摘などから，今後，プライバシー関連技術の開発が今以上に加速される可能性がある．我が国においても，昨今，プライバシーが社会問題化しており，こうした分野への取り組みがますます重要となってくると考えられる．

(2015 年 5 月 30 日受付)

山名 早人（正会員）yamana@waseda.jp
早稲田大学・理工学術院・教授．ビッグデータ解析およびユーザインタフェース研究に携わる．

☆14 http://www.forgetit-project.eu/fileadmin/fm-dam/downloads/2013-05-24_forgetit_brochure.pdf に掲載されている ForgetIT Architectural Framework を基にトレース．