米国における大規模データ分析等の技術分野における研究開発動向等に関する調査

2011 年 3 月

NICT　ワシントン事務所

目 次

1 連邦政府におけるデータ集約型コンピューティングに関する取り組み

20 年以上にわたり、米国政府はデータ集約型コンピューティングを開発し、活用してきた。しかし、この間の大半の期間は、その活動のほとんどは特定分野において実施され、他の組織における活動にほとんど関わり無く、各々が独自にシステムとアプリケーションの開発を進められてきた。それらの分野には、国家インテリジェンス･データをはじめ、高エネルギー物理学、バイオインフォマティクス、天気・気候が含まれる。

しかし近年、政府の研究者と管理職の間において、データ集約型コンピューティングは、アプリケーション領域にまたがり多くの共通要素を持つ、コンピューティング研究領域と認識されるようになってきた。こうした認識は、一つには、グーグル（Google）やウォルマート（Walmart）といった企業に代表される、民間セクターにおけるデータ集約型コンピューティングの成功の影響を受けている。その結果、政府、特に全米科学財団（NSF：National Science Foundation）は、データ集約型コンピューティングを改善し、一般的に拡張するための研究に資金を投じている。

1.1 連邦政府によるデータ集約型コンピューティング助成の背景

1.1.1 データ集約型コンピューティングの位置づけ

データ集約型コンピューティングは研究と実践の両方において新興領域であり、その定義とスコープはまだ厳格に確立されていない。大雑把にいうと、データ集約型コンピューティングは大規模データセットの大規模分析をともない、多くの場合において高度に合成された結果をもたらす。データ集約型コンピューティングの目的は、大規模データセットのパターンやトレンド、相互関係などを分析することにより、ナレッジと見識を創造することである。

データ集約型コンピューティングは、低コストのセンサーや、遠隔センシング技術、インターネットによって可能になった迅速なコミュニケーション、コモディティ･コンピュータとストレージの出現、そして、結果として得られる「データの洪水」を格納、管理、分析、そして視覚化技術が統合され可能となってきた。データ集約型コンピューティングのクラスには、財務、トランザクション、供給、製造、そしてその他データを分析することによって、ビジネス･オペレーションをいかに改善するかについての見識を探るビジネス分析をはじめ、試験的または観測的データを分析することにより、生物学的構造、またはプロセスに関する見識を探るバイオインフォマティクス、そして、試験的、観測的、または計算データを分析することにより、新たな物理的プロセスや存在の発見を目指す科学的分析が含まれる。

データ集約型コンピューティングは、コンピューティング技術と科学の両方において新たなパラダイムを生んだ。ここでいう「パラダイム」とは、新たな科学的パラダイムは古いパラダイムに取って代わる、と唱えた、米国の物理学者兼哲学者であるトーマス・クーン氏のパラダイムとは異なることに留意されたい[1]。ここでは、より新しいパラダイムは、必ずしも古いものに取って代わるのではなく、古いものに追加されるからである。

コンピューティング技術における最初のパラダイムは、銀行口座への入金や引き出しのポスティングといったトランザクション・ベース（transaction-based）のコンピューティングであった。トランザクションに関して言えば、インプット・ストリーム（トランザクションと口座）とアウトプット・ストリーム（口座更新）の規模は大きいかもしれないが、数値計算の規模は小さい（トランザクションのポスティングと残高の更新）。トランザクション・コンピューティングでは通常、大金を扱うために、ハードウェアとソフトウェアの高い信頼性と精度が要求される。汎用コンピュータは、この種の計算に秀でるように設計された。

第 2 のパラダイムは、天気予報などのシミュレーション・ベースのコンピューティングである。シミュレーション・ベースのコンピューティングに関しては、インプット・ストリーム（物理的プロパティーや初期条件、境界条件など）は極めて小さい。アウトプット・ストリーム（研究対象のシステムの進化など）は中規模から大規模だが、数値計算の規模は膨大になる可能性がある（システムの状況の時間経過にともなう漸進的変化など）。スーパーコンピュータは、この種のコンピューティングを想定して設計された。

データ集約型コンピューティングは、歴史的気象記録を分析し、地球が温暖化しているかどうかを見極めるといったタスクのための、新たな第 3 のパラダイムといえる。この場合、インプット・ストリーム（気象記録）は莫大である可能性があり、また、コンピューティング（相関関係と傾向の検索）も計り知れない規模である可能性もあるが、アウトプット・ストリーム（温暖化：イエス／ノー）は少量という可能性がある。大抵の場合、汎用コンピュータとスーパーコンピュータのいずれも、そのような計算にはあまり適さない。しかし、並列コンピュータのクラスターは、並行してそれぞれ独立して検索が可能なデータの処理に理想的といえる。非常に規模の大きいデータセットでさえも、リレーショナル・データベースとして格納が可能であることが、これまでにわかっている。この場合、

[1] Thomas Kuhn (1970). *The Structure of Scientific Revolutions.* Chicago: University of Chicago Press.

分析の実行にあたり SQL（Structured Query Language）を使うことが可能であり、並列 SQL クエリーのための標準技術の利用が大抵の場合は成功する。

科学分野において自然現象を理解するために導入された最初のパラダイムは、アリストテレスが実践したような実験と観察だった。第 2 のパラダイムはそれに理論を追加し、ニュートンの法則とマクスウェルの方程式によって実証された。この第 2 のパラダイムは、試験的データを、新現象の予測に利用可能な方程式によって表現される理論へと体系立てた。第 3 のパラダイムであるコンピュテーショナル・シミュレーションは、複雑な予測方程式を分析的に解決する能力の不足に対処するために生まれた。代わりに、コンピュータはこれら方程式を数値的に解くために使われた。

軍事的問題を解決するために 1940 年代に最初に開発された第 3 のパラダイムは、極めて新しく、今もまだ進化を続けている。データ集約型コンピューティングは、科学の第 4 のパラダイムと称される[2]。このパラダイムでは、現象の理解と予測を目的に、大規模な試験的／観測的データと理論的モデルの融合が利用される。

科学的コミュニティにおいては、データ集約型分散コンピューティングは、e 科学（e-Science）とも呼ばれている。英国研究会議（UK Research Councils）は e 科学について、「超大型データ・コレクションや超大規模コンピューティング・リソース、そして高性能視覚化へのアクセスを必要とし、ネットワークによって可能になる分散型グローバル・コラボレーションを介して実行される大規模な科学」と定義している。最近の e 科学に関するワークショップでは、この分野の急速な成長が明らかにされている[3]。ワークショップでは、遺伝学をはじめ、システム生物学、医学、天文学、化学、環境科学、音楽学、データ標準、そしてデータのキュレーション・分析、視覚化技術を含む様々な領域が議論されている。

1.1.2 物理学におけるデータ集約型コンピューティング：素粒子物理学

素粒子物理学（または高エネルギー物理学。elementary particle physics）は、データ集約型コンピューティングを導入した最初の科学的領域の一つである。1970 年代後半から、米国のフェルミ研究所（Fermilab）やスタンフォード線形加速器センター（SLAC : Stanford Linear

[2] *The Fourth Paradigm: Data-Intensive Scientific Discovery*, Microsoft, 2009

[3] Microsoft eScience Workshop 2010, Berkeley, CA, 2010

Accelerator Center）、欧州原子核研究センター（CERN：European Centre for Nuclear Research）などの加速器研究所は、イベント再構成（event reconstruction）と呼ばれる新粒子と相互作用の探求において、実験データを分析するためにコンピュータの利用を開始した。

このタスクは、本質的に並行かつデータ集約的であり、これら組織はこの目的のために、当時のコモディティ・コンピュータを集めて大型「ファーム（farms）」システムを構築した。米国では、高エネルギー物理学ネットワーク（HEPNet：High Energy Physics Network）が構築され、加速器研究所と研究者の所属機関を結んだ。しかし、初期の HEPNet は、通信速度がわずかに毎秒 56kb と遅かったため、分散型コンピューティングはほぼ不可能だった。先駆的取り組みであったがゆえに、データ転送とストレージ、分析のための技術は今日の標準に比べて未熟だったが、それらは多くの発見と複数のノーベル賞受賞につながった。

最近では、国際的な物理学コミュニティが、オープン科学グリッド（Open Science Grid）に基づくデータ集約型コンピューティング開発の第一線にある。この活動の主要なけん引力となっているのは CERN の大型ハドロン衝突型加速器（LHC：Large Hadron Collider）におけるイベント再構成であり、世界の科学者によって利用されている。LHC では、年間ペタバイト級のデータの生成が始まったばかりである。ヒッグス粒子（Higgs Boson）のような新粒子を発見するため、また、素粒子のビヘイビアに対する理解を深めるために、データは保管・周知・分析される必要がある。オープン科学グリッドは、超広帯域コミュニケーションと、大規模データを移動、分析するためにカスタム化された専用コンピュータ･プロトコルに依存する、分散型コンピューティングの体系化されたモデルである。研究班が必要とするデータは、これらプロトコルを使って CERN から国家または地域リポジトリーまで分類され、分析のために大学やその他研究機関へと転送される。この活動は、クラウド･コンピューティングに幾分近いグリッド･コンピューティングの全分野の開発に役立てられている。

1.1.3 地球科学におけるデータ集約型コンピューティング：EOSDIS

米航空宇宙局（NASA：National Aeronautics and Space Administration）の地球観測システム・データ情報システム（EOSDIS：Earth Observing System Data and Information System）は 1994 年以来、参加科学者と一般のために、地球科学衛星とフィールド計測からのデータを管理、流布してきた。EOSDIS は、複数の地球観測衛星や、12 の全米データセンター、そして多くの調査サイトを含む広範囲に分散されたシステムである。データセンターでは、地球のグリッド上で、衛星が旋回するナローパスから通常収集される衛星データの処理が行われる。それには、計器や計測の科学的単位への変換の較正をはじめ、データの時空間ディストリビューションの合

理化、異常の解決、複数の衛星または計器のデータの融合と統合フォームへの変換、そして研究者がアクセス可能なデータベースの構築といった難しさがある。

データ同化（data assimilation）と呼ばれるこのプロセスでは、全体において大規模なコンピューティングが必要とされる。幸いなことに、それらの多くは、クラスター・マシンにおいて並列処理が可能である。EOSDIS に関与する研究者は、このデータを分析し、地球で起こる物理的および生物学的プロセスの研究と、観測結果を地球科学モデルの結果を比較するために役立てている。その他のアプリケーションの中でも、気候に関する人間の活動結果の研究と予測において、このシステムは非常に重視されている[4]。

1.1.4 生物学におけるデータ集約型コンピューティング：ヒトゲノム計画

1990 年代の間、科学者はヒトゲノム計画の一部としてヒトゲノムの配列を決定する技術を開発した。その中でも最も生産性が高かったのが、その成功をデータ集約型コンピューティングに依存するショットガン法（shotgun sequencing）と呼ばれる技術である。各染色体は、長さが短くて不揃いな、リード（read）と呼ばれる断片の重なりに、化学的に分断される。各リードは、それから配列が決定され、コンピュータ･プログラムが各リードの重なる部分を比較して、それらつなぎ合わせて連鎖を作る。不可避の配列エラーや配列の無作為さがあるために、コンピュータ･プログラムでは、有望な組み合わせを決定し、多数の反復的なリードの重なりに依存するためにも確率論を利用することで、個々のエラーによる影響の軽減を図っている。このプロセスでは大量の情報を分析するが、比較の大半は並行して実施することが可能であり、一般的な並列コンピュータを利用して行える。

1.1.5 インテリジェンス分野におけるデータ集約型コンピューティング：全情報認知

国防高等研究計画局（DARPA：Defense Advanced Research Projects Office）は 2002 年、全情報認知（TIA：total information awareness）という名の下に監視（surveillance）と情報技術を応用し、テロリストやその他の安全保障上の脅威をモニターするためのプロジェクトを発足した。捜査令状なしで既存のデータ･ベースを集約し、米国内の全ての人の個人情報を含む膨大なデータベースを構築しようとするものである。データマイニング技術を導入し、不審な関係や活動を当局に警告することを目指した本計画は 2003 年、一般からの抗議を受けて終了に追い込まれたが、TIA の一部活動は今も違う名前で継続されている。

---

[4] Earth Science Data and Information System Project, NASA

破壊分子的行動を特定するために、無作為な記録を確実に分析するという数学上の問題は大きく、誤検出が数多く発生する可能性がある。そのため、プロジェクトは本来の形で継続されてはいるものの、その実用性は不透明といわざるをえない。

1.1.6 商業におけるデータ集約型コンピューティング：ウォルマート

政府から資金援助を受けた取り組みではないが、ウォルマートはデータ集約型コンピューティングにおける先駆者の一組織である。1980 年代にウォルマートの幹部は、顧客のトランザクションデータをプロセスの最適化と売上げ拡大に利用できることに気がついた。そこで、全店舗のトランザクションデータを、アーカンソー州のコンピュータ･センターに毎晩アップロードするというシステムを構築した。アップロードには、当初は電話のダイヤルアップ接続を利用した。データは、テラデータ（Teradata）のコンピュータによって管理されるリレーショナル･データベースに格納された。テラデータ級コンピュータの利用は、当時としては非常にめずらしいことだった。それらは、ゆるく連結されたプロセッサによって構成され、プロセッサが一体となり超大型リレーショナルデータベースを「共有なし（shared nothing）」モードで管理するというものだった。（共有なしとは、データベースがパーティションによって複数のパーツに分けられて、それらパーツが各プロセッサのローカル･ディスクに保存される形態のこと。クエリーは各プロセッサに分配され、各プロセッサは、それぞれが制御する一部のデータベースにだけアクセスする。従って、それらは何も共有しておらず、それらが返信する部分的回答をマスター･プロセッサが組み立てる。）ウォルマートは、このシステムを非常に上手く活用し、在庫管理とサプライヤーへのデリバリータイム短縮に役立てた。これは逸話だが、ウォルマートは、ハリケーン警報によって、米国で人気のスナック菓子であるポップタルト（Poptart）の販売が大幅に増えることを発見し、天気予報によってポップタルトの仕入れを増大させ、在庫を一時的に増やしたことなどもある。

共有なしデータ･クエリーにおけるウォルマートとテラデータの先駆的業績は、ウェブ検索の迅速な処理の中核を成す、グーグル（Google）のマップリデュース（MapReduce）ソフトウェア･フレームワークなどの最近の進歩の基盤となった。今日グーグルは、各購買客の全ての購入、つまり 1 日あたり約 2 億 6,700 万件のトランザクションを記録していると伝えられる。このデータは 4 ペタバイト・データ･ウェアハウスに格納され、ウォルマートの売上高と利益を改善するために常に分析作業が行われている[5]。

[5] Data-Intensive Scalable Computing, R.E. Bryant, Carnegie Mellon University, 2009

1.1.7 金融におけるデータ集約型コンピューティング：定量的モデル

1990 年代以降、先進的銀行やファンド、商社は博士号を持つ科学者（計量アナリスト、またはクオンツ《quants》と呼ばれる）を雇用し、証券の価格設計やトレードの時期決定のためのモデル開発に取り組んでいる。これらモデルは通常、市場の動きを予測し、企業に競争上の優位性を与える目的で、大量の歴史的データから方程式を作成することをベースとしている。モデルは、ほぼ経験によるものであり、合成された歴史的データによって、パラメータ化されている。つまり、非常にデータ集約的である。モデルの実際の稼働は、大抵の場合はリアルタイムで実施され、マイクロ秒ベースでトレーディング信号を出している。2000 年代初期の間、これらモデルは非常に上手く作動し、企業に大きなリターンをもたらしているかのようだった。しかし、2008～2009 年の金融危機では、モデルは無残にも機能しなくなり、経済的混乱を煽ることとなった。当時、モデルは大抵が線形であり、平衡近傍の状況においてのみ正確であることが露呈された。現在は、非均衡性データをより洗練されたモデルに組み込むか、あるいは少なくともモデルが失敗しそうなときを予測するための開発が実施されている[6]。

1.1.8 インターネットにおけるデータ集約型コンピューティング：グーグル

インターネットの検索エンジン提供に踏み切った最初の企業でこそないが、グーグルは群を抜いて最も成功した企業である。検索分野における同社の成功は、ページランク（page rank）と呼ばれる関連性の概念を導入したことから来ている。ページランクは当初、あるページが他のページによって言及された回数に基づいていた（この概念は、もともと NSF から資金提供を受けてスタンフォード大学で開発された）。グーグルは、グーグル・ファイル・システム（Google File System）上で稼働するデータベースのビッグテーブル（Bigtable）において、ペタバイト級のウェブページを保管している。グーグル・ファイル・システムとビッグテーブルはいずれも、グーグルが開発した[7]。マップリデュースは検索を管理するソフトウェアであり、同じくグーグルの手による[8]。マップリデュースでは、ユーザーは検索ディストリビューション・アルゴリズム（Map）と結果アセンブリー・アル

---

[6] The Minds Behind the Meltdown, Wall Street Journal, January 22, 2010. この記事は、Scott Patterson 氏の書籍“The Quants” の抜粋であり、このサブジェクトの良い入門編となっている。

[7] Bigtable: A Distributed Storage System for Structured Data, Seventh Symposium on Operating System Design and Implementation, November 2006

[8] MapReduce, op. cit.

ゴリズム（Reduce）をプログラムするだけでよいことから、新たな検索を極簡単に創出できる。フレームワークは、分散型サーバー上の全てのデータフローとコンピュテーションを管理する。マップリデュースは、この報告で後述する通り、多くの違う種類のデータ集約型コンピューティングにとって優れたプラットフォームであることが証明された。グーグルは、何千台という安いコモディティ・サーバーを使ってデータベースを保管し、クエリーを実行している。グーグルが導入したデータ集約型コンピューティング概念は、基本的に共有なしアプローチであり、その起源はウォルマートが採用した概念に見ることができる。

1.1.9 エネルギー分野におけるデータ集約型コンピューティング：石油産業のための地震探査データの視覚化

簡単に発見できるような油層はもう期待できないことから、石油会社は、アクセスがほぼ不可能な場所―地下深部や海底の下―にある石油を見つけるため、その方法の洗練を重ねてきた。石油を発見するための最も一般的な方法は、地震探査に加えて試掘井を掘ることである。油井検査のログデータを見れば、油井周辺の岩盤構造は正確に把握できるが、油井から離れた場所の地層についてはほとんど探知することができない。地震探査では、地表で作った雑音パルスが地下の岩盤に当たって跳ね返ってくる音をマイクで拾う。これらの記録は、処理を経て油井検査ログデータに融合され、地下の地層の 3 次元像が作成される。その後地質学者が 3 次元像を解釈し、潜在的石油貯留層を見つけ出す。地震記録は、キロメーターあたり 1000 データ・ポイント強を包含し、個々のデータセットは数テラバイトのデータを含んでいる可能性がある。

データ集約型コンピューティングは、音響特性と油井ログデータを反転させるために必要となる。というのも、データは不正確であり、本質的に不適正なデコンボリューション（逆重畳積分）プロセスが必要となるからである。確率論的、および決定論的数学的手法の両方が使用され、それらを繰り返しながら音響特性とログデータに対する 3 次元適合を発見する。

この分析の成功は、メキシコ湾の地下深くにある石油貯留層の発見によって証明された。これら石油貯留層は、通常は大きな塩類鉱床の下にある。塩類鉱床は音響信号のほとんどを反射する傾向があり、それらの下にある地層の視覚化は非常に難しい。深海の油井掘削は極めてコストが高く、それゆえに試掘井もほとんど掘られることがない。弱い音響シグナルから海の下の地層構造を得ることができるデータ集約型コンピューティングは、これら埋蔵物を経済的に回収可能にするとして高く評価されている[9]。

1.2 データ集約型コンピューティングにかかわる主要な技術的課題

1.2.1 大量データの整理と管理

データ集約型コンピューティングの原料は、現在一部アプリケーションの場合で最大ペタバイト（10^{15} バイト）級、数年以内にはエキサバイト（10 バイト）級に達するとされる大量の数値データである。このデータは、物理的、生物学的、そして環境的データのセンサーをはじめ、写真と動画、電話の会話などの音声、トランザクション・データ、政府および商業的データベース、ウェブサイトとインターネットのソーシャル・ネットワーク、電子メール、インスタント・メッセージ、ツイートなど、多くのソースに由来する。今日のデータ収集能力は、安価なセンサーや、正確な時間ベース、全地球測位システム（GPS：global positioning system）、コモディティ・コンピュータ、安価なストレージ・メディア、ビジネスおよび娯楽目的のインターネットの利用、書面情報のデジタル化による紙の代替、デジタル動画とスチルカメラの増殖、そしてデータ獲得と保管を安価にした多くの技術によって助成されている。例えば、コモディティ・コンピュータ・ディスク・ドライブの価格は、テラバイトあたり 50 ドル未満であり、ペタバイト級ストレージへの更新も最低 5 万ドルから可能である。調査会社 IDC では、創造されるデジタル・データの総量は、1 年あたり約 270 エキサバイトに達すると推定している。

一方、これらの原料を整理して管理するということは、大きな課題として残されている。

第一に、データ・コミュニケーションに掛かる費用が低減しているにもかかわらず、中央ストレージへのデータ転送費用は高額である。著名なコンピュータ・サイエンティストの Jim Gray 氏（マイクロソフトのリサーチ・サイエンティスト）をはじめとする専門家は、コンピューティングと通信の比較費用を理由に、コンピューティングにデータを持ち寄るのではなく、データにコンピューティングを持ち寄ることを助言している[10][11]。

第二に、関心のあるデータは、周知のスキーマを導入し高次に構造化されたデータベースから、説明のない写真などの完全な非構造化データまでさまざまである。後に分析することを念頭に、異種データをいかに保管してインデックスを作るか、またいかにアクセスするかを決定することは、難しい場合がある。

---

[10] *Distributed Computing Economics*, Jim Gray, Microsoft Research, 2003

[11] Towards 2020 Science, page 15, Microsoft Research, 2005

第三に、データ利用量に制限を設け、それを順守することにも問題が伴う。データは機密情報、または独占的所有物である可能性があり、他にもプライベートな個人データである可能性や、医療記録や著作権のある情報などのように、法的制約が設けられている可能性があるからである。データ利用にデータの権利と制約を強制することは、困難であるといえる。

最後に、データをいかに整理してデータベース化するかについても問題がある。例えば、データをリレーショナル･データベースに組み入れるのか、それともインデックス化されたファイル･システムに整理するのかなどである。幸いにも、コモディティ･コンピュータによって、ローカルにプロセシング機能とストレージを持つ安価な「データ･ブリック（data bricks）」が実現可能となった。データ･クエリーが並行処理されると仮定すれば、これらはデータ･ストレージと共同設置され、拡張して超大型データセットを処理させることもできる。グーグルは、ウェブ・クローラー（web crawler）が集めた情報の処理にこの技術を役立てており、共有なしモードでコモディティ･データ・ブリックのクラスターを管理するマップリデュース･ソフトウェア・システムを構築した。

1.2.2 データ・キュレーション

データ･キュレーションとは、データの選択、保存、維持、そしてアーカイブを行うことである。前掲の課題とも密接に関連しているが、データの質を重視する点で異なる。質には、精度の概念、意味、メタデータ経由などの記述、出所、利用制限、そして古い技術に関わらず再抽出可能かどうか（retrievability）などが含まれる。

キュレーションは、非常に高額になりかねないが、一方で、コストを制限し、精度を改善することができる、シア・キュレーション（sheer curation）というアプローチもある。シア･キュレーションに伴うキュレーション活動は、データ創造担当者のワークフローに統合される。例えば、写真は、ファイルに埋め込まれたフィールドにおいて、統制語彙（controlled vocabulary）を持つメタデータを使って最初にアーカイブされるときに、記述することが可能である。（デジタル写真の埋め込みタグ標準はすでに存在する。） これは、将来のファイル検索の単純化につながる。ドキュメント管理システムは、多くの場合において、ドキュメントを保存する前に、ドキュメント･ジェネレーターがメタデータ･フィールドを埋めることを要求する。マイクソソフト･ワード（Microsoft Word）のような一般的ツールでさえも、メニューのファイル、そしてプロパティー下に基本的なドキュメント管理ツールが盛り込まれている。

一方、テキストや画像などの非構造化データを含むデータの体系化については、コンピュータ･プログラムがそれらを理解して分類できるようにするため、複数の方法が提案されている。セマンティックス（semantics）とオントロジー（ontology）は、いずれも非構造化データを分類するための一

般的アプローチである。言語学的セマンティックス（意味論）は、テキストを分析し、それによってテキストを後に抽出できるように体系化することを目指している。また、オントロジーは、知識体系（Body of Knowldege）に応用され、将来的な抽出のために体系化に役立てることができる正式表現（formal representation）である。

- ***セマンティックウェブ***

ワールド･ワイド･ウェブ･コンソーシアム（W3C：World Wide Web Consortium）が提唱するセマンティックウェブ（Semantic Web）は、ウェブ上の情報について、それが何を意味するかをコンピュータに理解させるための方法である。最も単純なセマンティックウェブは、タグを付加したドキュメントによって構成される。タグは単語やフレーズに付加されて、それらの意味を表している。タグはメタデータの形をとり、テキスト自体に埋め込まれる。タグは、統制語彙（controlled vocabulary）から選ばれる言葉である。セマンティックウェブ用にドキュメントを作成するにあたり、W3C は、RDF（Resource Description Framework）、OWL（Web Ontology Language）、そして XML（Extensible Markup Language）を含むツールの使用を推奨している。セマンティックウェブは、セマンティクスとオントロジーの両方の要素を併せ持つ。その最も成功した導入例は、タグの統制語彙について賛同し、ドキュメントをマークアップするために QWL を使用する実践者などの小さなコミュニティによるものである。

コミュニティの中でも、生物学者と生化学者たちは、ドキュメント、特に研究論文へのセマンティックウェブの応用で最も成功している。その進歩の兆しとしてバイオケミカル･ジャーナル誌（Biochemical Journal）は最近、セマンティック検索を可能にするために、その記事や論文へ OWL を使ってタグ付けし、ユートピア･ドキュメンツ（Utopia Documents）を使って制作していることを明らかにしている[12]。

セマンティックウェブをめぐっては、多くのコンピュータ･サイエンティストとウェブユーザーたちの間で物議をかもし出している。彼らの主張は、ウェブはあまりにも膨大、かつ分散されて複雑であり、統一されたタグスキームに従うのは不可能というものである。

[12] The Semantic Biochemical Journal, Biochemical Journal

1.2.3 大規模データセットからのナレッジ抽出

大規模データベースに対してクエリーを実行する技術は、今でも研究課題に挙げられる。ある大学では、リレーショナル･データベースは何百ペタバイト規模への拡張が可能であり、大規模データセットの保存と分析のための望ましい方法であるという主張を貫いている。また別の大学は、データはますます非構造化しており、ファイルで保存し、適切な場合はインデックスを付けることが望ましいと主張する。最も成功したとされる分析ツールも、あるひとつの形式のデータ用に設計されていることから、今のところ「最善」の方法は存在しない。例えば、ウォルマートはリレーショナル･データベースを非常に上手く利用し、ペタバイト規模のトランザクション･データのクエリーを実行している。またグーグルは、同じくペタバイト規模のウェブページに対し、インデックス付きファイルを使ってクエリーを行っている。

また、最も成功したデータ分析ツールは、利用分野においても限定されている。今日に至るまで、真の汎用ツールはまだ登場していない。遺伝子同定ツールでは、BLAST が標準である。リレーショナル･データベースのビジネス分析に関しては SQL、それもクエリー生成と分析ツールによって増補された SQL が標準である。高エネルギー物理学の分野では、検出データを分析する専用ツールが標準的に用いられている。テキスト･データに関しては、グーグルのマップリデュースのような分析ツールが一般的である。

インテリジェンス分野では、テキストまたは口語データ中の単語、あるいは単語パターンを探すために専用ツールが使われる。また、データ間のコネクションやパターン検索にもツールは有用である。視覚化ツールは、人間のオペレーターがパターンを認識できるようにデータを提示するという点において、若干の成功を収めている。グラフ生成ツールは、工作員のネットワークを発見したり、誰が首謀者であるかを推測したりするために利用される。

1.2.4 “理解”創出のためのデータ削減

理解を生み出すためにデータを削減するというタスクは今日、大規模データの処理において最も標準化作業が進んでおらず、一般的に多くの人による介入を必要とする分野である。生物学の研究では、DNA 配列が分かっているという前提で、BLAST のようなツールを使えば高い精度で遺伝子を同定することが可能である。また、一部のケースでは、遺伝子がコードするタンパク質の種類を予測することもできる。気候研究においては、気温、アイスコア（ice cores）、木の年輪に関する大規模な歴史的記録とその他データソースを分析することにより、地球は最近の歴史の中では前例がないほど暖かくなっていることが、予想以上に確定的に示された。しかし、そこで使われた技術は、気候研究用に特化して設計されたものだった。高エネルギー物理学の分野では、

新粒子やその他異形を発見するために、粒子飛跡分析フォーム検出器を減らすための専用ツールが導入された。

分析した実験データを同じプロセスのシミュレーションと比較することは、理解に達する一助になる可能性がある。というのは、実験およびシミュレーション･データは、しばしば補完的関係にあるからである。実験は「現実（reality）」を生み出すが、そこで起きていること全てを、常に直接測定できるわけではない。シミュレーションは、その背後にあるモデルの出来具合によって評価も決まるが、小さすぎる／大きすぎる、あるいは速すぎる／遅すぎるなど、プロセスの詳細を測定不可能なスケールで示すことができる。また、シミュレーションでは、方程式やパラメーターを調整することで、何が重要なプロセスやパラメーターであるかを精査することも可能である。

理解に達するための異例の方法は、普通、インターネットを通じて、多くのボランティアに協力を求め、データを考察させることである。例えば、最近行われた何千人というボランティアによるテレスコープ･データの精査では、星雲の回転方向（地球から観測して時計回りか反時計回りか）が無作為であることが示された。小数の星雲を対象にした以前の分析では、優先的な回転方向のあることが示唆されていた。

しかしながら、ソースが何であるかにかかわりなく、無意識のエラー、あるいは悪意のある偽情報の流布のいずれかによって、多くのデータは不正確、または単に間違っている。ナレッジと理解を求めるに際し、不正確なデータの中から正確なデータを判別するのは困難である。

1.3 データ集約型コンピューティング達成手段としてのクラウド対グリッド

1.3.1 クラウドとグリッドの定義

クラウドとグリッドのいずれの単語も、本来の意味を失うことが危惧されるほど乱用されている。しかし、データ集約型コンピューティングにおけるそれぞれの役割を容易に理解できるような方法で、それらを定義してみたい。

クラウドという表現は、もともとネットワークの利用に由来しており、固定ポイント･ツー･ポイント通信経路（fixed point-to-point communications path）から、始点から目的地までの不確定経路の提供をインターネット･サービス事業者（ISP）に依存する方法への移行期に生まれた。一部の例では、競争上の理由から ISP が正確なルーティングを独占したこともあり、それを受けてネットワーク上、エンド･ポイント間で特定の経路が設定されない、エンド･ポイントのプラグイン先としてクラウドの利用が開始された。このコンテキストにおいては、クラウドは ISP による不確定ルーティングに言及している。

最近では、クラウドといえば、遠隔サーバーからインターネット経由でユーザーに提供されるコンピューティング･サービスである。これらサービスは、電子メールのように特定されていたり、データ･ストレージやコンピューティング･リソースのように一般的だったりする。ユーザーと遠隔サーバー間のデータ転送コストがそれほど高額でない限り、サービスを提供するリソースがどこにあるかは重要ではないことから、クラウドという表現は、そのようなサービスにとって相応しいといえる。一般にユーザーが所有する遠隔コンピューティング施設だけを利用するプライベート･クラウドに関しては、用語の意味は曖昧である。クラウド･コンピューティングの主な魅力は、ローカル･ユーザーによる費用負担とローカル･コンピューティング管理にかかわる問題を軽減し、ネットワーク上のどこからでもアクセスできることである[13]。

一方、グリッドという単語は、もともと科学的研究に由来し、試験的または観測的データを離れた場所にいる研究者に提供するために必要なネットワーキングのアレンジを指す。グリッドは通常、明確に定義され、かつ大抵の場合は専用の経路を利用して、これも明確に定義されたノードを接続する。

この点において、それらは旧式のポイント･ツー･ポイント型ネットワークと共通点がある。しかし、グリッドは多くの場合において非常に広帯域な光チャンネルを備えた特殊なネットワークであり、利用条件に応じて動的に設定されることもある。また、大量のデータを極迅速に転送するため、ノードではグリッド FTP（grid ftp）などの特殊なユーティリティーを利用する可能性もある。グリッド･コンピューティングの最大の価値は、ローカルな分析を必要とする遠隔地のユーザーに、データを迅速に届けられることである。グリッド上の遠隔コンピュータは、ユーザーに対し、大抵は視覚化のためにデータを分析して結果を返信する目的でも利用される。

1.3.2 データ集約型コンピューティングのためのクラウドとグリッドの相対的利点

前出の Jim Gray 氏は、データ転送コストがすぐにコンピューティング･コストを超えてしまう可能性を理由に、大量データの転送を最小限に留めることを強調している[14]。代わりに、コンピューティングは、データが保存されている場所で実行されるべきであるという。その原則をここに応用すると、仮にデータがクラウドに保存され、データ分析に使用するコンピューティング･リソースが共同設置されている場合、クラウド環境におけるデータ集約型コンピューティングの利用は理に叶っ

[13] NIST Definition of Cloud Computing, National Institute of Standards and Technology, 2009

[14] Distributed Computing Economics, Jim gray, Microsoft Research, 2003

ているといえる。クラウド･データ･ストレージが、（実験やセンサーなどからの）データ生成ポイントと共同設置されることはあまりない。そのため、仮に一つのある施設がデータの大半を生成している場合、データ転送とストレージ･コストを考えると、クラウド･コンピューティングは問題となる。

クラウド･ベースのデータ集約型コンピューティングは、複数の場所で生成され、追ってまとめて分析されることになる企業データへの展開も検討される。サービス･プロバイダーに転送され、リレーショナル･データベースに格納されたデータは、異なる場所からの遠隔 SQL クエリーを行うことが可能である。ユーザーへの結果転送には、通常は少量のデータを伴うだけである。しかし、遠隔ユーザーによるローカル・データ分析のための大量のデータ転送は、おそらくコスト効率がいいとはいえない。独自データ、あるいはその他の機密データをサービス事業者に委ねることもまた、問題の原因になり得る。米国における最近の訴訟では、ISP サーバーに保存された企業データの検索を、企業の資産に保存されたデータの検索よりも先に判事が認める可能性があることが示唆された。

一方、グリッド･ベースのデータ集約型コンピューティングは、生成データが、集約的あるいは反復的分析のためにデータを手元に保存しておきたいと考える、ある施設内の異なるユーザーによって生成された場合の利用に適しているかもしれない。また、遠隔分析の結果が、視覚化データなど、ローカルに表示することが求められる大型データセットである場合にも適切であると考えられる。

結論として、クラウドとグリッドはいずれも急速に進化していることから、データ集約型コンピューティングの利用者は、ローカル分析の代替技術としてどちらかを選ぶ前に、データ･ソース、分析の種類、ユーザーの場所といった課題を慎重に検討するべきである。

1.3.3 データ集約型コンピューティングと科学的コラボレーション

データ集約型コンピューティングは主に、一つの道具（天文台や高エネルギー粒子加速器など）によって生成された大量のデータを保管、管理する必要性と、これらデータを、遠隔地のユーザーが分析できるようにする必要性から生まれた。一方、データ集約型コンピューティングには、異なるタイプの科学的進歩に関連する第 2 のモードがあり、それがまた新たな研究課題を提示している。このモードでは、データは複数の機器によって生成され、個々の研究者や研究グループによってローカルに管理される。そして後にデータは集められ、一つの巨大なデータセットとして分析される。この種の科学は、科学的観測が何百人、または何千人という研究者の努力の結果であるかも知れないことから、「高度分散」と特徴付けられる。その最も極端なフォームでは、これら

の取り組みは「オープン・ソース科学」となり、個々の研究者ばかりか普通の人さえも、ローカルデータ収集というタスクを与えられ、分析と実験のための中央リポジトリへデータをアップロードする。

協調的データ集約型コンピューティングは、データ・アセンブリ、特にデータ・キュレーションにおいて混乱を引き起こす。仮にデータが複数の研究者グループによって生成されている場合、標準データ構造やターミノロジーを用いて体系化がされない限り、データは統合も分析もできないからである。これには、データ・アーキテクチャ設計に対する膨大な事前投資と、データ収集グループ間の高度なコーディネーションとコンプライアンスが必要となる。特にデータ・リポジトリーとして機能する組織は、収集されたデータの質を検証し、それがリポジトリに送信された後は、完全性を検証できなければならない。

データ・キュレーションにともなうこれらの難題は、データ集約型コンピューティングに対する大きな期待の一つ、つまり既存の科学的、技術的データセットを統合し、統合データに対して実験とシミュレーションを行なう能力への達成を妨げることになる。データ・フュージョンとも呼ばれるこの能力は、セマンティック・ウェブ技術を使うことにより（既存のデータベースがセマンティック・ウェブ標準を使っている限りにおいてだが）、ある程度は達成できるかもしれない。また、非構造化データ（ドキュメント、イメージ、サウンド、ビデオ・ファイル）の自動組織と分析に関する進歩は、データ・フュージョンにも寄与するものである。全米のインテリジェンスや本土防衛研究コミュニティは、特にこの分野において活発である。また、有望なアプローチのひとつとして、統合や分析が比較的簡単な既存データセットの少なくとも一部を取り上げて、統計学的モデリング技術を応用し、パターンを同定し結果を解釈することも考えられる。この種のアプローチは、不完全なデータを基に推論して新しいナレッジを得るわけだが、直接的観測を通じた科学的発見に比べて信頼性は低いものの、将来の研究の大発見につながるような有益な見識を得られる可能性がある。

2 データ集約型コンピューティング研究に対する連邦政府支援

米連邦政府は、データ集約型コンピューティングを強く支援している。しかし、個々のプログラムを特定することは、多くの場合困難である。「データ集約型コンピューティング」という言葉（より非公式に「ビッグ・データ」（Big Data）とも呼ばれる）は、比較的最近作られたものであり、多様な既存プログラムに適用されるが、これらプログラムは「データ集約型コンピューティング」としては分類されておらず、関連する研究も、一見して関係がないように見えるプログラムに組み込まれている場合がある。一般的には、データ集約型コンピューティング研究プロジェクトは、天文学、高エネルギー物理学、経済、そして気候学を含む、多くの科学的分野に見ることができる。本章の政府機関プロフィールでは、複数の研究領域において、将来的にデータ集約型コンピューティングの応用を可能する新技術やツールに主に集中したプログラムを紹介する。

2.1 全米科学財団

サイバー対応ディスカバリーとイノベーション・イニシアチブ（CDI：Cyber-enabled Discovery and Innovation initiative）は、データ集約型コンピューティング研究へのサポートを明確に示した、NSF の最初のプログラムの一つである。このプログラムは、NSF の分野横断型プログラム、つまり、このプログラム下で、NSF の複数の局が助成金を提供するものである。このプログラムは 2007 年 9 月に発足し、当時の全体予算は年間 2,000 万ドルだった。プログラムの目的は、当時の説明内容と同じく今も、「複雑かつデータリッチな相互作用的システムに対応するために、コンピュータに基づく次世代の発見概念とツールを開発することにより、国家のイノベーションを生む能力を拡大すること」である。プログラムでは、以下に示すテーマ領域を通じてそのミッションを追求している：

- **データからナレッジへ（From Data to Knowledge）：** 人間の認知を増進し、豊富な異種デジタル・データから新しいナレッジを生成する。
- **自然・人工・社会的システムの複雑性の理解（Understanding Complexity in Natural, Built, and Social Systems）：** 複数の相互作用する要素から構成されるシステムに関し、基本的洞察を導き出す。
- **仮想組織（Virtual Organizations）：** 組織的・地理的・文化的境界を越えて人とリソースを集めることにより、発見とイノベーションを強化する。

プログラムの提案要請文書の中で、NSF は以下のような期待を示している：

CDI プログラムは、

①大規模データセットにおけるパターンと構造を特定するためのトランスフォーマティブな発見（transformative discovery）を可能にする

②自然・社会科学とエンジニアリングの理解を深める手段としてコンピュテーションを開発する

③複雑な確率論的または混沌としたシステムを抽象化、モデル化、シミュレーション、そして予測する

④亜粒子（sub-particles）から銀河系まで、細胞下から生物圏（biosphere）まで、そして個人から社会に至るまで、自然の相互作業やコネクション、複雑な関係、そして相互依存性を調査してモデル化する； サイバー・リソースを強化、利用するために、将来の世代の科学者やエンジニアを訓練する

⑤科学とエンジニアリングの最前線を前進させ、さらに STEM 分野への関与を拡大させるため、創造的、かつサイバーに対応した境界横断型コラボレーションを、産業や国際的特徴を持つものも含めて促進する[15]。

CDI プログラムは、データ集約型コンピューティング・ツール以外にも多くのトピックを包含するが、データ集約型コンピューティングは、このプログラムを通じて推進される主要能力である。それら研究に関連する主要助成金提供先の一部を以下に示す：

- **ハーバード大学（Harvard University）：**「グラフィクス・プロセシング・ユニット（Graphics Processing Units）と半導体素子ストレージ（Solid-State Storage）を利用する、天文学、神経生物学、そして化学のための科学的コンピュテーション」関連プロジェクトを実施。内容は、「新しい市販のコンピューティング技術を活用し、実用的かつ透過的なスケーラブル異種コンピューティング（SHC：Scalable Heterogeneous Computing）を可能にする概念とツールを開発する（プログラミング戦略、汎用・ドメイン特定ライブラリ、など）。結果は、電波天文学、量子化学、そして神経科学の 3 つの最先端の課題に応用される」と説明されている。当初の助成額は 200 万ドル超だった。

[15] Cyber-enabled Discovery and Innovation program synopsis for FY2011, available at http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=503163.

- **カリフォルニア大学サンタバーバラ校（University of California, Santa Barbara）:** 「マルチモーダル・イメージングを通じた複雑な生物学的構造の発見と理解に関するコンピュテーショナルな課題（Computational Challenges in the Discovery and Understanding of Complex Biological Structures through Multimodal Imaging）」に関するプロジェクトを実施。バイオ医療イメージングの主要な問題—異なる種類の機器で生成された画像の統合が必要とされる、バイオ医療研究のための画像データの分析—に対処する。分析は、多くの場合に異なる組織サンプルの画像を使い、それら画像の融合を試みることがある。このプロジェクトでは、網膜の組織について、画像データを統合、解釈するための統計的プロセスを開発することにより、プロセスの自動化を目指す。本プロジェクトにはユタ大学も関与している。当初の助成額は 4 年間で 195 万ドルだった。

- **ミシガン大学アナーバー校（University of Michigan at Ann Arbor）:** 「宇宙天気予想を可能にする新サイバー技術（New Cyber Technologies to Enable Space Weather Forecasting）に関するプロジェクトでは、宇宙天気データの解析に新しいコンピュテーショナル・アプローチを活用することにより、宇宙の天候の向こう 18 時間の予報を可能にする。また、アルゴリズム、コンピュータ・コード、そして分析ツール開発のための NASA/NSF コミュニティ・コーディネーテッド・モデリング・センター（NASA/NSF Community Coordinated Modeling Center）とも協力する。当初の助成額は 3 年間で 170 万ドルだった。

また、NSF では、データ集約型コンピューティング分野における先駆的取り組み、およびデータ集約型コンピューティングの商業的利用の増加を評価し、2008 年度に CluE（Cluster Exploratory、クラスター探索）を発足させ、データ集約型コンピューティング分野の重点プログラムを開始した。CluE は、グーグル-IBM アカデミック・クラウド・コンピューティング・イニシアチブ（Google-IBM Academic Cloud Computing Initiative）との予算 500 万ドルの共同プログラムである。同イニシアチブでは、データ集約型コンピューティング分野の革新的研究案を探索する提案に資金を拠出し、研究者に対し、グーグル-IBM のコンピュータ・クラスターで稼働するソフトウ

ェアやサービスへのアクセスを提供してきた[16][17]。NSF は、このプログラム下で 14 の大学に助成金を拠出した。助成プロジェクトには、以下が含まれる[18]:

- **カーネギー･メロン大学（Carnegie-Mellon University）:** 第 1 のプロジェクトは、ウェブ検索をより効率的に実施する目的で、ウェブ･コンテントの時事性を特性化するものである。検索の局所的ルーティングは、従来の検索に比べて手間がかからないため、結果的に電算能力とコスト面の大幅な削減が可能になる。第 2 の研究プロジェクトは、機械翻訳のための統合クラスター･コンピューティング･アーキテクチャ（INCA: Integrated Cluster Computing Architecture）開発に焦点を当てている。

- **マサチューセッツ工科大学（Massachusetts Institute of Technology）、ウィスコンシン大学マジソン校（University of Wisconsin-Madison）、イエール大学（Yale University）:** クラスター･ベースの大規模データ分析に対するアプローチの比較研究。マップリデュースと並列データベース･システムはいずれも、何十万というノードを利用して拡張可能なデータ処理を提供する。しかし、研究者にとっては、新しいデータ集約型コンピューティング用アプリケーションを設計するにあたり、これら 2 つのアプローチのうちどちらがより適しているかを判断するために、2 アプローチの性能と拡張性の違いを知ることが重要である。

- **フロリダ国際大学（Florida International University）:** 災害と環境保護支援に役立てることを目的に、航空画像とオブジェクトを分析するためにクラウド･コンピューティングを活用する。

2009 年度と 2010 年度において、NSF のコンピュータ情報科学工学局（CISE：Directorate for Computer and Information Science and Engineering）は、データ集約型コンピューティングを、CluE プログラムを含む予算 1,000 万ドルの特別分野横断型プログラムと位置付けている[19][20]。

---

[16] Cluster Exploratory (CluE) Program Solicitation, National Science Foundation, 2008

[17] Cluster Exploratory (CluE) Program Solicitation, National Science Foundation, 2008

[18] Press Release, NSF, 2009

[19] CISE Cross-Cutting Programs: FY 2009 and FY2010, NSF, 2008

[20] CISE Cross-Cutting Programs: FY 2010, NSF, 2009

NSF は本プログラムの提案書の中で、以下のように述べている。「コンピューティングとストレージ技術の断続的な進化とコスト削減にもかかわらず、データ創出と収集は、我々のデータ処理と保存能力を越えている。そのため我々は、この大量のデータを管理—保存、抽出、探索、分析、通信—する方法を再考せざるを得ない。プログラムでは、我々のデータ集約型コンピューティング・システムとアプリケーションの構築および利用能力を向上させるとともに、それらの限界に対する理解を助け、さらにはそれらが我々の経済と社会においてますます大きな影響力を持つようになるに従い、これらシステムの運用と利用能力のある知識豊富な労働人口を育成するような、コンピュータと情報科学・工学の全領域のプロジェクトを対象に資金を提供する」。

NSF は、このプログラム下で 40 件を超える助成金を交付した[21]。NSF が支援する領域は、データ集約型コンピューティングのためのアーキテクチャや、データ集約型コンピューティング管理のためのミドルウェア、分散型データ分析技術、そして検索とパターン・マッチングを含むアプリケーションなどである。

また、NSF は 2009 年度には、コンピュテーショナルと物理的リソースが緊密に連携するシステムである、サイバー・物理システム（CPS：Cyber-Physical Systems）関連の研究プログラムに着手した[22]。CPS の例には、ロボット手術（robotic surgery）のようなローカル･システムや、スマート･エネルギー･グリッド（smart energy grids）などの高精度製造と分散システムが含まれる。このプログラムの下で実施された助成件数は、今日までにほぼ 150 件に達する[23]。

NSF は、複数分野に渡りデータ集約型コンピューティングへの投資の増加と拡張を継続するものとみられる。その一つのサインとして NSF は 2009 年、Dr. Myron Gutmann を社会・行動・経済科学（SBE：Social, Behavioral and Economic Sciences）局のアシスタント・ディレクターに任命した。Dr. Gutmann は、大規模コンピュテーショナル社会科学研究の著名な専門家であり、データ集約型コンピューティングの社会科学への応用を促進することが期待されている。

ネットワーキング・情報技術 R&D（Neworking & Information Technology R&D）プログラムに関する 2010 年 12 月のレビューでは、大統領科学技術諮問委員会（PCAST：President's Council

[21] CISE Awards in Data-Intensive Computing, NSF, 2010

[22] Cyber-Physical Systems, Program Solicitation, NSF, 2010

[23] Grant Awards: Cyber-Physical Systems, NSF, 2011

of Advisors on Science & Technology）は NSF に対し、特にデータ集約型コンピューティングに関して提案を行っている:

> NSF はデータ収集、保管、管理、そして分析に関する基礎研究へのサポートを拡大するべきである。プログラムでは、以下に示すようなトピックスに対応するべきである:
>
> - 物理的世界のセンサーから獲得したデータのコンテクチュアル・メタデータ（文脈上のメタデータ）
> - 相互相関に由来する情報
> - 多様なソース、異なるスケール、そして異なる種類とメタデータからのデータを一体化する情報融合アルゴリズム
> - 長期的保存
> - データの出所とインテグリティ
> - 不完全かつ不確かなデータに基づく推論
> - データに含まれる情報の深い分析
> - 複雑なデータと情報の抽象化、集約、視覚化

NSF では、2013 年度の助成金をカバーする次期予算要求の一部として、NITRD レビューの結果を考慮している。

2.2 国立衛生研究所

国立衛生研究所（NIH：National Institutes of Health）が支援するデータ集約型コンピューティングの大半は、分子生物学分野に対する統計学とコンピュータ･サイエンスの応用として定義される、バイオインフォマティクスの一般的カテゴリーに分類される。バイオインフォマティクスの主な用途は、大規模 DNA 配列解読に関してである。バイオ医療の進歩に与えたバイオインフォマティクスの価値は認識されており、そのことが、NIH 傘下の複数の機関で実施される多数のデータ集約型コンピューティング・プログラムに対し、NIH が支援を行うきっかけとなった。

バイオインフォマティクスに対する NIH の関心は、2004 年に開始された医療研究のための NIH ロードマップ（NIH Roadmap for Medical Research）によって弾みがついた。ロードマップでは NIH プログラム間のギャップを特定し、それらに対処するためのメカニズムを創設した[24]。その結果生まれたデータ集約型イニシアチブの大部分は、バイオ医療情報科学・技術イニシアチブ（BISTI：Biomedical Information Science and Technology Initiative）によって調整が図られている。

また、ロードマップの結果、7 つの全米バイオ医療コンピューティング・センター（National Centers for Biomedical Computing）が設立された。それらのうちの一つでミシガン大学（University of Michigan）にある全米統合バイオ医療インフォマティクス・センター（National Center for Integrative Biomedical Informatics）では、大量の生物学的データを統合し、それらの分析ツールを開発している。しかし、あるアナリストの経験に基づく推測によると、コンピューテーションとバイオインフォマティクスに対する助成金は、NIH 予算の 2～4%未満に過ぎないという[25]。

NIH プログラムの中で特筆されるのは、バイオ医療コンピューティング・コミュニティによって必要とされるソフトウェア開発・維持のための特別助成金である[26]。前掲のアナリストは、ソフトウェア開発が断念される傾向にある連邦政府研究機関の中にあって、NIH のこの判断は称賛に値する異例のことであると述べている。その他の関連プログラムとしては、バイオ医療コンピュテーショナル科学・技術における革新（Innovations in Biomedical Computational Science and Technology）[27]とバイオ医療インフォマティクスに対する研究助成（Research Grants in Biomedical Informatics）[28]がある。

---

[24] The NIH Roadmap, National Institutes of Health

[25] Bringing the Fruits of Computation to Bear on Human Health, Katherine Miller, Biomedical Computation Review, Spring 2009

[26] Continued Development and Maintenance of Software, NIH Program Announcement, 2010

[27] Innovations in Biomedical Computational Science and Technology, Funding Opportunity Announcement, NIH, 2010

[28] NLM Express Research Grants in Biomedical Informatics, NIH, 2010

NIH は多くの機関によって構成されており、NIH におけるデータ集約型コンピューティング関連活動を追うことは難しい可能性がある。最近、ある記事において、本研究に関与する NIH 傘下のさまざまな機関とプログラムが紹介されており[29]、それによると、BISTI では過去 6 年間に 4 件の包括的発表を実施しており、合計 297 件の研究助成、金額にして 3 億 5,500 万ドルを提供した。

NIH 下の組織のひとつである国立癌研究所（National Cancer Institute）では、癌に関する研究と治療に関連する情報の収集、分析、統合、そして流布を目指す、癌バイオ医療インフォマティクス・グリッド（caBIG: Cancer Biomedical Informatics Grid）を支援している[30]。caBIG には、癌の遺伝要素に関する分子と臨床データのデータベースである癌ゲノム・アトラス（Cancer Genome Atlas）が含まれる。

2.3 諜報先端研究プロジェクト活動

諜報先端研究プロジェクト活動（IARPA：Intelligence Advanced Research Projects Activity）は、国家諜報コミュニティ（National Intelligence Community）の研究機関である。IARPA がそのプログラムを通じ、諜報機関からの情報を収集、分析、そして流布する方法を改善するための研究に資金を提供している。諜報データは、オープンソースと機密情報のいずれも量が爆発的に増えており、IARPA では、諜報データにデータ集約型コンピューティングを導入する目的で、2009 年にナレッジ・ディスカバリと流布（Knowledge Discovery and Dissemination）プログラムを開始した[31]。このプログラムでは、包括的分析を実現するために、異種データセットを統合するツール開発の重要性を強調している。またプログラムでは、新しいデータセットが追加されるに従い、ユーザーが新しいデータセットを無意識のうちに発見し、それを既存のデータ・リポジトリに統合、データ構造やアーキテクチャ、そしてモデルを動的に再構成することにより、データセットの横断的分析が実行できるようなシステムを想定している。

2.4 国防総省

1990 年代末以降、国防総省（Department of Defense）は、「ネットワーク中心戦争（network-centric warfare）」の新しいモードを追求するために、情報通信技術の能力を利用することを模索

[29] Update on Biomedical Computation at NIH, Peter Lyster, NIH, 2010

[30] Cancer Biomedical Informatics Grid (caBIG), National Cancer Institute, 2011

[31] Knowledge Discovery and Dissemination (KDD) Program, IARPA Program Solicitation, 2009

してきた。このアプローチの主要部分は、かつて「情報優勢（information superiority）」と呼ばれ、現在は「ナレッジ優勢（knowledge superiority）」として知られる分野である。国防総省とその内部機関が取り組むプロジェクトの一部は、データ・フュージョン（data fusion）に対する関心を反映している。データ・フュージョンとは、最新かつ関連性のある情報を戦場の指揮官に提供するために、センサーや観測衛星、無人偵察機（UAV：Unmanned Aerial Vehicles）による偵察、そして兵士の報告によって生成された膨大なデータセットをリアルタイムで統合、分析する技術である。

国防総省とインテリジェンス・コミュニティ（DOD/IC：Department of Defense and Intelligence Community）は 2008 年、米国の一部のエリート科学者とエンジニアによって構成される常任委員会の JASON に、センサー・システムによって生成される大量のデータを管理、統合、分析する際の、特に課題に関する報告書の作成を委託した。JASON は、国防総省を代理して MITRE（MITRE Corporation）が管理するグループで、その会員情報は機密扱いとされている。2008 年 12 月に発行された JASON による報告書は、大型データセット管理に関し、国防総省がいくつかの潜在的課題に直面しているとした上で、問題はそのスケールと本質において、高エネルギー物理学、天文学、そしてウォルマート（Wal-Mart）のような大企業において実施される同様の種類のデータ分析に匹敵すると結論した。しかし、報告書は研究の将来の方向性についても、重要な見解を示している。以下に一部を紹介する：

- データを収集、保管するハードウェアとは対照的に、データをより効率的に処理するためのソフトウェアへは、現在、十分な投資がなされていない。
- クラウド・コンピューティングのようなデータ構成や処理のアプローチは、将来的なデータ・フュージョンの促進と発見に、今のところ最も適していると思われる。
- イベント駆動型アーキテクチャとソフトウェアへの追加投資を伴って実施される、サービス指向アーキテクチャなどの技術への継続した投資は、DOD/IC エンタープライズ横断型データ・フュージョンの実現に有効と思われる。

また、JASON の報告は、国防総省内の複数研究機関の方向性に影響を与えることが多い。DARPA は、大型データセット分析のための新コンピュテーショナル・アプローチの開発でこれまで先行している。この領域の研究は、1960 年代と 1970 年代に、研究領域としての人工知能（artificial intelligence）の確立を果たした DARPA の先駆的役割に基づいている。DARPA は、大量のデータを分析し、戦場の司令官と兵士の意思決定をサポートするための、コンピューティング能力の利用に関する研究を長年サポートしてきた。2011 年度、この一連の研究は、以下に示す 2 つの主要な技術トレンドを活用する方向へとシフトしている：

(a) 大量の広範に分散したシステムを利用することにより、コンピュテーショナル処理能力を作り出すクラウド・コンピューティングの台頭

(b) 大人数の個人からのインプットと判断を協調的に管理することにより、人間の認知というユニークな能力とコンピュテーショナルな分析を統合するクラウド・ソーシング

そのミッションを理由にデータ集約型コンピューティングに最も直接的に関与している DARPA の 2 つのオフィスは、情報イノベーション室（Information Innovation Office）と戦略技術室（Strategic Technologies Office）である。以下は、この分野の DARPA プログラムの例である。

複数スケールにおける異常検出（ADAMS: Anomaly Detection at Multiple Scales）

ADAMS の目標は、異種ソースのデータへの長期的なアクセス、抽出、そして保管のための新しい分散型アーキテクチャを開発することである。これらのソースは、システム・ログ・ファイルや、センサー、新しい報告やテレビの番組などの非構造化データ、そして科学的データなどである。次に ADAMS では、個人、システム、グループ／組織、または国や州のレベルでデータを分析するにあたり、何日間、何ヶ月間、または何年間という特定期間にわたり、これらデータセットの連続的分析を行い、さらに「変則的（anomalous）」イベントや活動、ビヘイビアを検出できる技術を開発する。このプログラムの予算は、2011 年度に 450 万ドル、2012 年度に 1,800 万ドルとなっている。

ウェブ・スケール情報統合（WSII: Web-Scale Information Integration）

WSII プログラムでは、異種ソースから様々なメディア（テキスト、グラフィクス、オーディオ、ビデオ）の大量の情報とデータを取り込み、自動的にメタデータを生成、応用し、国防総省内のどこからでも情報を簡単に検察して、抽出できるようにするシステムを構築する計画である。新興のセマンティック・ウェブ技術の他に、視覚化、要約、そして機械ベースの分析など、情報をすぐに使用可能にする新技術を使うことを想定している。プログラム記述によると、このシステムには以下に挙げる機能も実装される予定である:

「重要な軍事的、科学的、経済的、そして社会文化的情報に関する豊富なナレッジを、人間が読解でき、かつ機械が処理できるフォーマットで自動的、かつリアルタイムに作成して維持する。セマンティック可能な検索と処理により、兵士の意思決定向上に資する情報発見と操作の自動化を図る一方で、そのように豊富なナレッジは、暴動鎮圧、世界的攻撃、そして、戦略、交戦規則、計画立案と実行を含む仲間同士の衝突（衝突直前の状態も含む）に対する根本的なコンテキストを提供する。」

このプログラムの 2011 年度予算は 1,080 万ドルであり、2012 年度要求は 1,550 万ドルである。

認知クラウド（Cognitive Cloud）

防衛科学研究室（Defense Sciences Office）が担当する認知クラウド・プログラムでは、クラウド・コンピューティングとクラウド・ソーシングの両方を利用し、大型異種データセットの分析処理能力の加速と合理化、拡張を図るための予備的研究を実施する。これら技術の初期アプリケーション領域は、オープン・システムにおけるサイバーセキュリティ脅威の特定、特性化、対応である。その他の潜在的アプリケーションとしては、広大かつオープンな地形の監視や、外国社会と政治システムの分析とモデリング、大型かつ複雑なソフトウェア・システムのデバッグなどが挙げられる。

2.5 国土安全保障省

国土安全保障省（DHS：Department of Homeland Security）はその設立以来、超大型データベースを収集、スキャン、解釈する能力が、同省にとってクリティカル・ミッションな能力であることを認識してきた。しかし、それゆえに、DHS が批判の的となることもあった。例えば、DHS が米国内の全ての金融トランザクション・データを収集して保管し、自動的方法を応用してテロリストへの金融支援が疑われるケースを発見すると提案したとき、米国政府による一般市民の不当な捜査に対する、憲法で定められた保護に違反しているという見方もされた。そのため DHS は、個人のプライバシー侵害を避けるとともに、当初想定されたよりは受身のやり方で、その能力を開発しなければならないとの結論に至った。

DHS 科学・技術局（Science and Technology Directorate）内部のコマンド・制御・相互運用性部（Command, Control and Interoperability Division）では、大型データセット統合と分析、解釈に関するほとんどの研究をサポートし実施する。この分野における DHS の最も先進的研究は、Dr. Joseph Kielman が率いる基礎／将来研究プログラム分野（Basic/Futures Research Program Area）である。DHS のデータ集約型コンピューティング・ツール開発努力は、そのプログラム分野の中の 2 つのプログラムに見て取ることができる。

全米視覚化・分析センター

DHS は、エネルギー省の国立研究所の一つであるパシフィック・ノースウェスト国立研究所（PNNL：Pacific Northwest National Laboratory）に、全米視覚化・分析センター（NVAC：National Visualization and Analytics Center）を設立した（なお、PNNL が実施するエネルギー省のためのデータ集約型コンピューティング研究は、次章で後述）。NVAC は研究を実施する一

方、DHS の本研究領域に対する助成のパイプ役として機能しているという点で、ユニークな存在である。NVAC は、DHS に様々な能力を提供することを目的としており、以下にその例を挙げた:

- 大量、多次元、複数ソース、時間依存性の情報ストリームを調査する。
- 情報ストリームを発見して発信する。
- 緊急を要する環境で意思決定を行う。.
- 脅威を防止、阻止、そしてその対応において、情報を最大限に利用し、適切と思われる人や部門と共有するために、人間の判断を応用する。

NVAC はまた、全米の研究施設に、5 つの地域視覚化・分析センター(Regional Visualization and Analytics Centers)を選んで開設している。さらに、ディスクリート科学研究所(Institute for Discrete Sciences)の 4 つの大学付属センター(University Affiliate Centers)に資金を提供している。これらセンターは、情報処理と分析、管理の先端手法に関する R&D を実施している。

ビジュアル分析と高精度情報環境

ビジュアル分析と高度情報環境(The Visual Analytics and Precision Information Environments)プログラムでは、複数のフォームとモード(テキスト、ビデオ、画像、オーディオ、データベース、センサー・データを含む)の情報を理解し操作するための画期的先端技術と手法に関する研究に出資し、実施する。開発中のシステムは、ビジュアル・インプットに基づき、認識的また直感的にデータを処理する人間の能力を利用するよう設計されている。また、これらの能力と、超大規模データベースの中で情報オブジェクトを統合し、評価、優先順位を決めるコンピュテーショナルな方法を統合する。モバイル・携帯型システム、デスクトップ分析ステーション、そして大規模な状況認識、または共通作戦図(common operating picture)設置を含む、幅広いカスタマイズされた利用のためのツールを開発する研究も進行中である。

3 研究実施機関

3.1 政府研究センター

3.1.1 パシフィック・ノースウエスト国立研究所

米エネルギー省（DOE）傘下のパシフィック・ノースウエスト国立研究所（PNNL: Pacific Northwest National Laboratory）は、データ集約型コンピューティング分野における相当な量の研究活動を維持している[32]。研究重点分野には、ソフトウェア・アーキテクチャ、ハイブリッド・ハードウェア・アーキテクチャ、そして分析アルゴリズムと視覚化が含まれる。そのうちソフトウェア・アーキテクチャに関しては、PNNL ではデータ集約型コンピューティング向けミドルウェアを開発した。ハイブリッド・ハードウェア・アーキテクチャ分野では、コンピュテーションを迅速化するための特別目的仕様ハードウェアの利用の探索を、分析アルゴリズムと視覚化分野では、超大型データを対象にしたリアルタイム分析と視覚化機能の開発に取り組んでいる。

データ集約型コンピューティング・イニチアチブ

2006 年に始まったデータ集約型コンピューティング（DCI：Data Intensive Computing Initiative Initiative）イニシアチブは、PNNL の主力研究領域（エネルギー、国家安全保障、環境科学）が、大規模データセットを管理し、大量のデータからナレッジを抽出し、人間が理解できるレベルまでデータを減らすという能力に、まずます依存するようになったという新しい認識に基づいて設置された。DIC は、ハードウェア、ソフトウェア、そしてアルゴリズム開発に同時に求められる進歩を調べることにより、この課題に着手した。DIC の責任者は、Deb Gracio 博士が務める。

PNNL は 2008 年、情報統合やデータ視覚化、そしてデータ分析フレームワークを網羅したデータ集約型コンピューティングのためのミドルウェア（MeDICi：Middleware for Data Intensive Computing）ソフトウェア統合フレームワークを開発した。このフレームワークは、ソフトウェアにおける分析手順の構築を推進する目的で、アプリケーションとソフトウェア・コンポーネント・ライブラリ間で大規模データセットを移動するための方法論とアーキテクチャを創造するものである。このフレームワークは、PNNL のチーフ・アーキテクトの Ian Gorton 氏率いる研究班が開発し、オープン・ソース・コードとしてリリースされた[33]。

---

[32] Data Intensive Computing at PNNL, Pacific Northwest National Laboratory

[33] See http://www.pnl.gov/main/publications/external/technical_reports/PNNL-18716.pdf.

その他のソフトウェア製品としては、DNA 配列分析用 BLAST プログラムの並列版である ScalaBLAST や、アプリケーションを異種データ･ソースに連携させるバイオインフォマティクス、リソース・マネージャ（BRM：Bioinformatics Resource Manager）がある[34]。

また、バイオパイロット（BioPilot）は、PNNL とオークリッジ国立研究所（Oak Ridge National Laboratory）によるシステム生物学関連の共同研究プロジェクトである。科学者たちは、システム生物学に関連した複雑な生態系の予測モデリングを目的にコンピュテーショナル･ツールの構築に取り組んでいる。研究領域には、データ分析、モデル抽出、そしてモデリングが含まれる[35]。

PNNL の研究は、エネルギー、次世代スマート配電網・国家安全保障・サイバー･セキュリティのモニタリングと運用、異種データの体系化・組織化・視覚化、データ集約型科学研究などの分野に適用されている。また、PNNL は、基礎データのパターンと変化の発見と分析に資するための大型データセット向け視覚化ツールに関する研究で知られている。初期の視覚化研究の一部は国家安全保障を目的に実施され、諜報データの体系化と解釈のための視覚化ツールを提供した。最近では、これらツールは、CLIQUE モデリング･システム経由でサイバー･セキュリティとネットワーク･フロー分析に応用されている[36]。

3.1.2 ゴダード宇宙飛行センター

NASA 傘下組織であるゴダード宇宙飛行センター（GSFC：Goddard Space Flight Center）では、地球観測システム（Earth Observing System）宇宙探査機からの地球科学データの収集と流布を可能にする地球観測システム･データと情報システム（EOSDIS：Earth Observing System Data and Information System）を管理している[37]。これらの宇宙探査機には、合計 87 の計器が搭載されている。アーカイブの合計ボリュームは 4 ペタバイトを超え、エンド･ユーザーによるデータ･ディストリビューションは、1 日あたり平均で 6 テラバイトを上回る。

---

[34] Data Intensive Computing Downloads, PNNL

[35] Computational Biology & Bioinformatics, PNNL

[36] CLIQUE – Real-time Event Detection and Visualization, PNNL

[37] Earth Science Data and Information System Project, NASA

GSFC には、ゴダード地球科学データ・情報サービスセンター（GES DISC：Goddard Earth Sciences Data and Information Services Center）が設置されている。GED DISC は、研究科学者、アプリケーション科学者、アプリケーション・ユーザー、そして学生に地球科学データや情報、サービスを提供する分散型アクセス・アーカイブ・センター（DAAC：NASA Distributed Access Archive Center）のひとつである。GES DISC は、いずれもデータと情報の集合体である NASA 降水と水文学（NASA Precipitation and Hydrology）、および大気組成とダイナミクス（Atmospheric Composition and Dynamics）のアーカイブである。また、研究とアプリケーションのための近代レトロスペクティブ分析（MERRA：Modern Era Retrospective-Analysis for Research and Applications）データ同化データセットと、北米陸面データ同化システム（NLDAS：North American Land Data Assimilation System）とグローバル陸面データ同化システム（GLDAS：Global Land Data Assimilation System）も、GES DISC において格納されている。なお、MERRA は、GSFC のグローバル・モデリング同化室（Global Modeling and Assimilation Office が生成しており、NLDAS と GLDAS は、いずれも GSFC の水文科学部門（Hydrological Sciences Branch）が生成している。

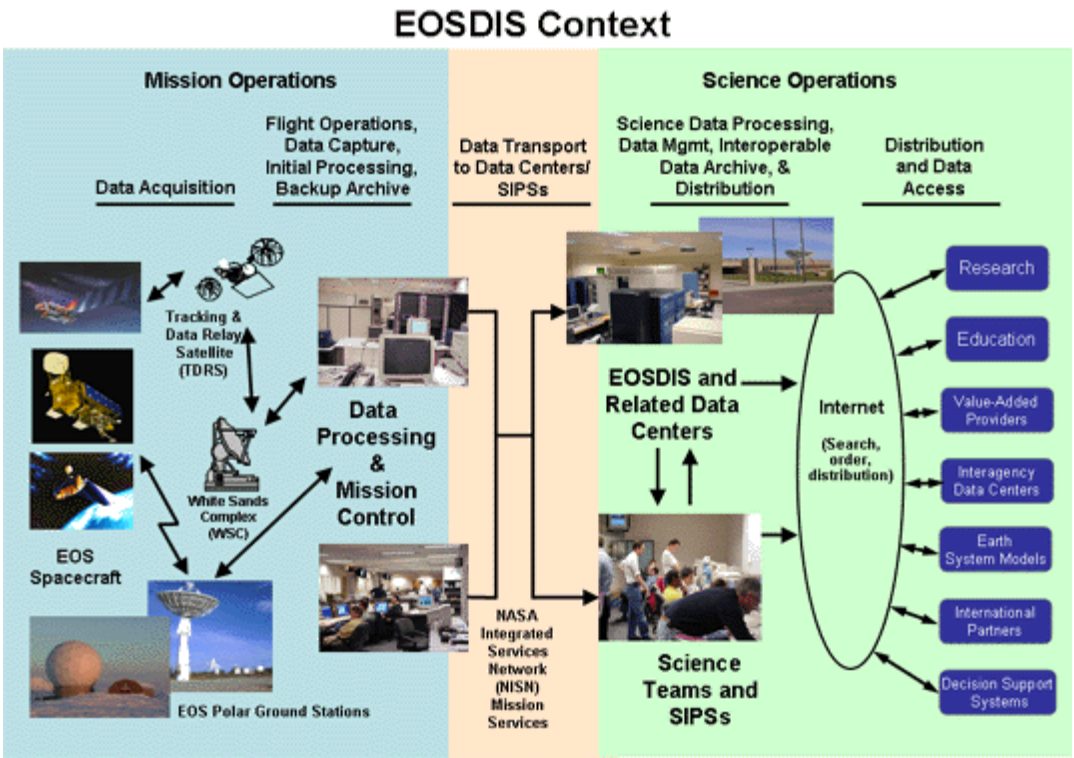
同センターでは、シンプル・スケーラブル・スクリプトベース科学処理アーカイブ（S4PA：Simple, Scalable, Script-based Science Processing Archive）と呼ばれる、データ統合とアクセスのためのアーキテクチャを開発した[38]。このアーキテクチャでは、ラジカル単純化（Radical Simplification）アプローチを採用し、超大型データセットを再組織、抽出時間の短縮と検索機能の簡便化を達成できる。また、同センターでは、ジョバンニ（Giovanni）というデータ分析と視覚化のためのウェブベースのツールも開発した。このツールを使えば、個々の研究者は、データをダウンロードすることなく、時系列プロット（time series plots）、緯度／経度グラフィクス（latitude/longitude graphics）、そして大気深度プロット（atmospheric depth plots）を作成することができる[39]。

[38] http://disc.sci.gsfc.nasa.gov/additional/techlab/s4pa.

[39] http://disc.sci.gsfc.nasa.gov/additional/techlab/s4pa.

図 1 に EOSDIS データのワークフローを示す。衛星が生成したデータは、NASA の専用宇宙ネットワークを利用して地上に送られる。その後、データは高速ネットワークを使って分布される。9 つの全米データ･センターがデータ処理を担っており、エンド･ユーザーにサービスを提供している。各センターでは、NASA が開発した専用ハードウェアとソフトウェアを使って EOS データのアーカイブと分析、分配が行われる。NASA では複数の技術を採用し、較正とデータ妥当性検証にデータ同化手法と統合地球システム･モデル（integrated earth system models）を組み合わせることにより、データの質を確保している[40]。また、テレメトリ･データの修正や正当性の実証、測定値の物理単位への変換、そして統一された 4 次元グリッドへの変数のマッピングを行うため、複数レベルの分析を応用している。

図 1: EOSDIS ワークフロー



出典：NASA[41]

[40] Earth Science Reference Handbook, NASA, 2006

[41] EOSDIS Context, NASA

3.1.3 サンディア国立研究所

サンディア国立研究所（SNL：Sandia National Laboratories）は、エネルギー省国立研究所の一つであり、国家核安全保障局（NNSA：National Nuclear Security Administration）内にある。SNL のミッションは、米国が保有する核兵器の原子の安定性をモデル化し、管理することであり、主にそのためにスーパーコンピューティング研究に長年携わってきた。SNL 内部では、データ分析・視覚化部門（Data Analysis and Visualization Department）が、大型データセットの管理と処理に関わる特有の問題に焦点を当てている。同部門では、先端アルゴリズムと革新的ハードウェア・アーキテクチャ、そして拡張性のある分析コンポーネントを組み入れることにより、データの複雑性や大きさ、不確かさに関わりなく、洞察を促進するツールを開発している。

SNL では、インフォマティクス・データの収集、処理、そして表示を目的に、柔軟性のあるコンポーネント・ベースのパイプライン・アーキテクチャを提供するタイタン・インフォマティクス・ツールキット（Titan Informatics Toolkit）を開発した[42]。タイタンは特に、科学的視覚化と情報視覚化の融合に言及し、分散型メモリー・プラットフォーム上の拡張性のある分析実行のためのフレームワークを提供する。タイタン・パッケージに欠かせないのが、非構造化テキストの分散型処理とモデリング、そして分析のためのソフトウェア・コンポーネントのセットである、パラテキスト（ParaText）である。タイタンと同様にパラテキストも、分散型メモリー・コンピュータ上での稼働を想定して設計された。パラテキストは、潜在的セマンティック分析に関する問題へ応用されている。潜在的セマンティック分析とは、ドキュメントのセットと、それらドキュメントに含まれる単語との関係を分析する手法である。タイタンとパラテキストのいずれも、オープン・ソース・ライセンスに基づきリリースされている。

3.1.4 ローレンス・バークレー国立研究所

エネルギー省国立研究所のひとつ、ローレンス・バークレー国立研究所（LBNL：Lawrence Berkeley National Laboratory）では、コンピュテーショナル研究部（Computational Research Division）内の 2 つのグループにおいて、データ集約型コンピューティングへの取り組みが行なわれている。二つのグループとは、Deb Agarwal 氏が率いるデータ集約型システムズ・グループ（Data Intensive Systems Group）[43]と、Arie Shoshani 氏率いる科学的データ管理センター

[42] The Titan Informatics Toolkit, Sandia National Laboratorie

[43] Advanced Computing for Science, Lawrence Berkeley National Laboratory

（Scientific Data Management Center）である[44]。Agarwal 氏のグループでは、センサー･ネットワークからの環境データの収集と分析を手掛ける世界的コンソーシアムの活動を調整している。Shoshani 氏のグループは、データ集約型コンピューティングのためのデータ管理の理論的基盤に取り組んでいる。

3.1.5 国立標準規格技術院

国立標準規格技術院（NIST：National Institute of Standards and Technology）によるデータ集約型コンピューティングへの関与は、限定的である。NIST では、独自のオープン機械翻訳評価（Open Machine Translation Evaluation）プログラム[45]を通じて機械言語翻訳に関する研究と試験を行っているほか、翻訳改善案の機能を評価するプラットフォームを提供している。また、サイバー･セキュリティに焦点を置いたクラウド･コンピューティングについて、他のサイトとの協調や研究も手掛けている[46]。

3.2 大学研究センター

データ集約型コンピューティングに関する学術研究は、コンピュータに関するハードウェア価格の急速な下落と、コモディティ・ベースの高性能クラスター・コンピューティングの開発を受けて進んでいる。現在では多くの大学が大規模データ・ストレージのためのインフラを持っており、グーグル（Google）や IBM、マイクロソフト（Microsoft）の新興ソフトウェア・ツールは、大規模データ研究に必要な能力を大学の研究者に提供している。

3.2.1 メリーランド大学

メリーランド大学（University of Maryland）では、情報研究学部（College of Information Studies）とメリーランド大学先端コンピュータ研究所（University of Maryland Institute for Advanced Computer Studies）において、自然言語処理、言語の機械翻訳、そしてウェブ･ベー

[44] Scientific Data Management Center, Lawrence Berkeley National Laboratory

[45] Machine Translation Program, National Institute of Standards and Technology

[46] Cloud Computing at NIST, National Institute of Standards and Technology

スのデータマイニングの各領域に主に焦点を置き、データ集約型コンピューティングに関する研究を行っている。コンピュテーショナル言語学と情報処理研究所（Computational Linguistics and Information Processing Laboratory）と、ヒューマン・コンピュータ・インタラクション研究所（Human Computer Interaction Laboratory）を含む複数の研究班がこれら問題に取り組んでいる。

クラウド・コンピューティング・センター（Cloud Computing Center）

2009 年秋に発足したメリーランド大学のクラウド・コンピューティング・センターには、研究、教育、そしてアウトリーチを通じ、共同でクラウド・コンピューティングの将来を構築することに関心のある人材が、複数の専門領域から集まっている。Ben Schneiderman 教授と Jimmy Lin 准教授が主導するプロジェクトであり、同センターでは、クラウド・コンピューティングのアプリケーションと、クラウド・アーキテクチャとインフラ、そして幅広い社会的問題に焦点を当てている。

同センターのプロジェクトのひとつは、アイボリー（Ivory）と呼ばれるソフトウェア・パッケージの構築である。アイボリーは、マルコフ・ランダム・フィールズ（Markov Random Fields）に基づく抽出エンジン、その名も「MRF を利用する検索」（SMRF：Searching with Markov Random Fields）を利用する、ウェブ・スケールの情報抽出研究のためのハドゥープ（Hadoop）・ツールキットである。このオープン・ソース・プロジェクトは同センターが開設される前の 2009 年春にすでに開始されており、メリーランド大学とヤフー・リサーチ（Yahoo! Research）のコラボレーションによるものである。アイボリーは、インデックス作成と抽出のために、ハドゥープ分散環境（マップリデュース・プログラミング・モデルとその根本にある分散ファイル・システム）を最大限に利用しており、クルーウェブ 09（ClueWeb09）コレクション上で、ハドゥープと相互動作するように特に留意して設計された[47]。クルーウェブ 09 は、カーネギー・メロン大学が作成した、情報抽出や言語技術研究を支援するためのデータセットで、2009 年 1 月、2 月に収集された、10 言語、10 億ウェブページ（25TB）を含んでいる。

[47] https://wiki.umiacs.umd.edu/ccc/index.php/Main_Page

3.2.2 イリノイ大学

イリノイ大学（University of Illinois）シカゴ校では、Robert Grossman 氏のグループが、マップリデュース、またはハドゥープに似たデータ集約型環境を構築した[48]。この環境は、セクター（Sector）という分散ファイル･システムと、スフィア（Sphere）という関連プログラミング･フレームワークによって構成される。研究者によると、セクター／スフィア・システムの処理性能は、ハドゥープよりもつねに約 2～4 倍早いとのことである。

3.2.3 ジョンズ・ホプキンス大学

データ集約型科学とエンジニアリング研究所（Institute for Data Intensive Science and Engineering）は、大型データ研究分野の主要な大学の研究センターである。同機関のリーダーであり天文学教授の Alex Szalay 博士は、大型データ集合体と、それが科学における影響を含むさまざまなプロジェクトにおいて、マイクロソフトの Jim Gray 氏と協力してきた[49][50]。同博士のグループは、スローン・デジタル・スカイ・サーベイ（Sloan Digital Sky Survey）のためにテラバイト級のアーカイブを構築したほか、博士自身は天文学とデータ集約型コンピューティングに関する数多くの論文を出版した。その他のプロジェクトには以下が含まれる：

- ライフ・アンダー・ユア・フィート（Life Under Your Feet）： 温度や水圧といった土壌の変数を測定する、世界最大級のセンサー・ネットワークの一つからのデータ。
- 公共乱気流データ・クラスター（Public Turbulence Data Cluster）： 研究者が流体力学的乱流の大型シミュレーションのクエリーを実行できるオンライン・リソース。クラスターは、大型データセットと相互作用するための抜本的に新しい方法である。

IDIES の研究者は、他のデータ集約型エンジニアリングと科学プロジェクトにも参加している。以下に例を挙げる：

---

[48] Towards Efficient and Simplified Distributed Data Intensive Computing, IEEE Transactions on Parallel and Distributed Systems, 2010

[49] Research Activities of Alex Szalay

[50] Pooling Data in Astronomy and Particle Physics, National Science Foundation

- 最先端レベルの革新的地震発生シミュレーション
- オンコスペース（OncoSpace）に関する JHU 医学部とのコラボレーション。オンコスペースは、放射線腫瘍学分野で使われる大型分析データベースの概念である。
- データ集約型コンピューティング関連の画期的プロジェクトに取り組むポスドク研究員 6 名に、ムーア・ケック財団（Moore and Keck Foundations）が助成するポスドク・フェローシップ・プログラムを提供
- 天文学コミュニティ向けデータに関するオーバーレイ・ジャーナルに関して JHU シェリダン図書館（Sheridan Libraries）と協力

また、Szalay 博士は、データ・スコープ（Data-Scope）と呼ばれるシステムを構築する、210 万ドルの新規 NSF 助成研究の研究責任者でもある。データ・スコープは巨大情報セットを処理できる最新鋭コンピュータのクラスターで、一旦構築された後は 5 ペタバイトのデータ処理が可能になる。NSF による助成は 2 年に渡って実施され、ほかにジョンズ・ホプキンス大学が約 100 万ドルを出資する。プログラムは 2011 年 5 月の開始が予定されており、初年度に新しい機器の設計と構築を行う予定である。

3.2.4 インディアナ大学

インディアナ大学（Indiana University）では、インフォマティクスとコンピューティング学部の Geoffrey Fox 氏、Randall Bramley 氏、Judy Qi 氏、そして関連研究員が、クラウドとマルチコア・コンピュータに関するデータ集約型コンピューティング研究に取り組んでいる。研究は生命科学アプリケーションに重点を置き、マップリデュースと、従来の並列および分散コンピューティング・アプローチを利用している。研究班は、サービス集約型リンクド・シーケンシャル・アクティビティーズ（SALSA：Service Aggregated Linked Sequential Activities）の下に、その活動の一部を組織している。SALSA とは、分散ハードウェアをデータ集約型アプリケーションに拡張する、コンピュテーショナル・アーキテクチャである。ランタイム・ミドルウェアとして、ハドゥープ／マップリデュース++／MPI（Hadoop/MapReduce++/MPI）または DryadLINQ/MPI のいずれかを使い[51]、

[51] Performance of Cloud and Cloud Technologies, SALSA Group, Indiana University, 2009

クラスター管理は、エクストリーム・クラウド・アドミニストレーション・ツールキット（xCAT: Extreme Cloud Administration Toolkit）を利用している[52]。

3.2.5 カリフォルニア大学デイビス校

カリフォルニア大学デイビス校は、Kwan-Liu Ma 教授先導のもと、エネルギー省のプログラムである最新鋭コンピューティングによる科学的発見（SciDAC : Scientific Discovery through Advanced Computing）の一部である、SciDAC ウルトラスケール視覚化研究所（Institute for ultra-scale visualization）を主導している。

この SciDAC は、期間 5 年のプログラムとして 2001 年度に始まった研究プロジェクトで、物理学者、数学者、コンピュータ科学者、そしてコンピュテーショナル科学者のチームが協力し、基礎エネルギー科学（Basic Energy Sciences）、高エネルギー物理学（High Energy Physics）、核物理学（Nuclear Physics）、先端科学的コンピューティング研究（Advanced Scientific Computing Research）、フュージョン・エネルギー物理学（Fusion Energy Sciences）、そして生物学的環境研究（Biological and Environmental Research）における問題を解決するための、主要なソフトウェアとアルゴリズムの開発に取り組んでいる。また、このプログラムでは、データ集約型コンピューティングの研究を手掛ける 4 つの研究所に資金を提供している。

SciDAC ウルトラスケール視覚化研究所は、DOE SciDAC プログラムが助成する期間 5 年の研究およびアウトリーチ活動として、2006 年 9 月 15 日に設立された。同研究所を主導する UD デイビスは、シカゴのアルゴンヌ国立研究所（Argonne National Laboratory）の支援を受けている。同研究所のミッションは、コンピュテーショナル科学とエンジニアリングが直面する、やがて訪れるペタスケール視覚化に関わる課題に対応することである。研究プログラムには、プラットフォームを横断して移植できる包括的並列視覚化スウィートの構築プロジェクトが含まれる。このスィートは、このスケールでの科学的発見を可能にするとともに、アプリケーション科学者に、これらツールの最善の使用法を示すものである。

この研究所では、視覚化、高性能コンピューティング、そして科学的アプリケーション領域の著名な専門家を招集し、SciDAC 科学者および幅広いコミュニティにとって、並列視覚化技術をコモディティ化することを目指している。また、産業界に対しては、大規模視覚化のための計算をサポートするため、ハードウェアとソフトウェアのアーキテクチャ、そしてプロトコルの見直しを提言するこ

[52] xCAT Extreme Cloud Administration Toolkit, Sourceforge

とになる。教育の面では、チュートリアルの組織、ソフトウェアとハードウェアのベンチマーキング、学生向け夏季訓練の計画などの責任を担う。

3.2.6 パデュー大学

パデュー大学（Purdue University）は、コマンド・コントロール・相互運用性環境のためのビジュアル分析（VACCINE：Visual Analytics for Command, Control and Interoperability Environments）として知られる、国土安全保障省が助成するマルチセンター・イニシアチブの中心的組織である。David Ebert 教授率いる VACCINE は、情報のコミュニケーションと流布、そして大量の国土安全関連データの中から洞察を引き出すための双方向ビジュアル分析環境に関する、教育、研究、開発、展開に焦点を当てている。また、国土安全担当者が抱える大量のデータに対する、リアルタイムで拡張性のある双方向ビジュアル・コンピュテーショナル分析、調査、計画立案、仮説検定、管理、そして意思決定のための新手法と技術を、研究、評価、移行することを念頭に設計された。具体的な研究対象は、ビジュアル分析、メディア特定分析、地理情報科学、情報合成、データとナレッジの管理、拡張性のあるデータ集約型コンピューティング、ヒューマン・コンピュータ・インタラクション、認知科学、ユーザビリティ・エンジニアリング、コミュニケーションなどであり、いずれもビジュアル分析科学を発展させるために実施される。VACCINE には、以下を含む合計 19 大学が参加する。

- ジュージアエ科大学（Georgia Institute of Technology）
- ペンシルベニア州立大学（Pennsylvania State University）
- スタンフォード大学（Stanford University）
- ノースカロライナ大学シャーロット校（University of North Carolina at Charlotte）
- ワシントン大学（University of Washington）
- フロリダ国際大学（Florida International University）
- インディアナ大学（Indiana University）
- ジャクソン州立大学（Jackson State University）
- ノースカロライナ A&T 州立大学（North Carolina A & T State University）

- テキサス大学オースティン校（University of Texas at Austin）
- バージニア工科大学（Virginia Tech）
- ヒューストン大学ダウンタウン（University of Houston, Downtown）
- 加サイモン・フレーザー大学（Simon Fraser University）
- 加ブリティッシュ・コロンビア大学（University of British Columbia）
- 独シュトゥットガルト大学（University of Stuttgart）

主要プロジェクトには以下が含まれる：

- 調査とインテリジェンス分析のための視覚化（JigsaW：Visualization For Investigative And Intelligence Analysis）： JigsaW は、ドキュメント・コレクションのビジュアル・インデックスであり、分析者が閲覧したい特定のドキュメントを見つけることを支援する。
- 流行病視覚化（PavViz：Pandemic Visualization）： 公衆衛生担当者や当局を対象に、汎発性インフルエンザ拡大を分析するための一連のビジュアル分析ツールを提供する。当局による、さまざまな決定（学校閉鎖、メディアリポート、戦略的国家備蓄の利用）と、それがインフルエンザ流行に及ぼす影響の分析を可能にする。また、PanViz ツールは、当局によるインフルエンザ流行の追跡と、流行期間中のあらゆる時点でのさまざまな意思決定手段の実践を可能にする。
- DHS エクセレンス・センターのための視覚化（Visualization for the DHS Centers of Excellence）： DHS エクセレンス・センターは、数多くの研究や教育材料、プロジェクトを生んできた。VACCINE 研究班は、双方向ビジュアル・フォームのこの大規模ナレッジ・リソースを操作、調査するために、ビジュアル分析手法を利用する。このプロジェクトでは、簡単にアクセスできるフォーマットで、DHS エクセレンス・センターのプログラムを、テーブル・ベース、グラフ・ベース、そしてタイムライン・ベースで視覚化する一方で、データに関連する複数の面での局面的ブラウジングやクエリーもサポートする。これらの視覚化は、将来的に再設計や再構築をしなくても、簡単にデータを追加して拡張できる。

3.3 企業研究センター

3.3.1 マイクロソフト

マイクロソフトリサーチ（Microsoft Research）は、故 Jim Gray 氏の下、長年にわたりデータ集約型分散型コンピューティングの研究に携わってきた。研究班は天文学者や生物学者と協力し、データ集約型アプリケーションの研究に取り組んでいる。マイクロソフトは最近、大型 PC クラスター向けに大規模データ並列アプリケーションを書くためのプログラミング環境、DryadLINQ をリリースした[53]。DryadLINQ は、分散型コンピューティングを管理するコンピューティング・エンジンの Dryad と、マイクロソフトの.NET フレームワークの一部として開発された、データ・クエリーと更新のための機能セットである LINQ を合体したものである。マイクロソフトは、学究目的での使用に限り、DryadLINQ のソース･コードと Dryad のバイナリー･コードのダウンロードを無料で提供している[54]。このシステムでは、クラスターを、それが逐次プログラムを稼働する一台のコンピュータであるかのように扱うプログラミング・モデルを提示することにより、クラスターでのプログラミングと稼働タスクを単純化することを目指している。2006 年に製品化が開始された Dryad は、ビング（Bing）分析の実行エンジンであり、複数ペタバイト級データを日常的に処理している。マイクロソフトは、データマイニング、グラフ分析、画像処理、シミュレーション、そして機械学習に Dryad/DryadLINQ を応用する研究プロジェクトも展開している[5556]。

また最近、マイクロソフトはウィンドウズ・アジュール・クラウド（Windows Azure Cloud）サービス･プラットフォームを投入した。これは、クラウド･ベースのアプリケーション開発を簡便化するアプリケーション開発環境とともに、世界中にあるマイクロソフトのデータセンターにおいてホステッド･クラウド･サービスを提供するものである[57]。アジュール・マーケットプレース・データマーケット（Azure Marketplace Datamarket）で、民間および公共機関からの膨大な第 3 者データセットを

---

[53] DryadLINQ, Microsoft Research

[54] Dryad and DryadLINQ Academic Release, Microsoft Research

[55] Dryad and DryadLINQ, Yuan Yu, Microsoft Research, 2009

[56] DryadLINQ: A System for General-Purpose Distributed Data-Parallel Computing Using a High-Level Language, Yuan Yu and others, Microsoft Research

[57] Windows Azure, Microsoft

提供するという点は、アジュール・クラウドの興味深い機能の一つである。データセットは、自社データとの統合を希望する顧客向けに貸し出し（または、一部の例では無償で使用することもできる）も行っている。

3.3.2 IBM

IBM リサーチ（IBM Research）は、ハードウェアとソフトウェア両方を含む先端コンピューティング技術分野のリーダーとして長年君臨を続けている。例えば、データ集約型コンピューティングの実現に必要なハード・ディスク・ストレージ容量（現在、ドライブあたり 3TB で、さらに増加中）の大きな進歩は、IBM アルマデン（IBM Almaden）における、巨大磁気抵抗効果に基づく読み取り・書き込みヘッドの発明によって可能になった[58]。最近では 2007 年、IBM は研究と教育用にクラウド・ベース・サーバーを提供するため、グーグルと提携した[5960]。

NSF は 2008 年、CluE イニシアチブのもと、この提携に参加した。2.1 章で示したとおり、CluE イニシアチブは、教育機関の研究活動に大型クラスター上で稼働するソフトウェアとサービスを提供する取組みで、データ集約型コンピューティング分野の革新的研究および教育案を探索する目的で実施されている[61]。また、ニューヨーク州にあるワトソン研究所（Watson Research Center）では、Liang-Jie Zhang と Rong Chang の両氏を含む研究者が、グリッドとクラウド・アーキテクチャの概念の開発に取り組んでいる。

3.3.3 グーグル

データ集約型コンピューティングの本番使用に加えて、グーグル・リサーチ（Google Research）ではデータ集約型コンピューティングとクラウド・コンピュータ分野に活発な研究グループを維持しており、これら領域の学究研究に資金を提供している[62]。グーグルのオペレーションをサポートす

---

[58] The Giant Magnetoresistive Head: A Giant Leap for IBM Research, IBM

[59] Google and I.B.M. Join in “Cloud Computing” Research, New York Times, 2007

[60] Google and IBM Announce University Initiative to Address Internet-Scale Computing Challenges, 2007

[61] A CluE in the Search for Data-Intensive Computing, NSF, 2008

[62] Data-Centric, Data-Intensive Computing, Google Research

る主要アーキテクチャとソフトウェアは、グーグル･リサーチによって開発された。コンポーネントには、何十というデータセンターに分散された複数ペタバイト級のデータを管理する分散型グーグル･ファイル･システム（GFS：Google File System）や、全データにわたって分散される検索をサポートするマップリデュース言語、データを体系化するデータベース構造のビッグテーブル（Bigtable）、そしてグーグルのオペレーションの中核にあるフォールト・トレラント・ハードウェアとハードウェア管理システムが含まれる。グーグル･リサーチはまた、統計的言語翻訳のためのグーグル・トランスレート（Google Translate）も開発した。グーグル・トランスレートは、NIST による 2005 年と 2006 年の中国語-英語、そしてアラビア語-英語翻訳の評価において、良い成績を収めている[63]。

3.4 コンソーシアム

3.4.1 オープン・サイエンス・グリッド

オープン･サイエンス･グリッド（OSG: Open Science Grid）は、約 90 の大学や国立研究所によって構成される主に米国ベースのコンソーシアムである。当初は、European Organization for Nuclear Research （CERN）の大型ハドロン衝突型加速器から得られる年間あたり複数テラバイト級データのアーカイブと分析を行うためのグリッド･コンピューティング技術を開発、応用するために設立された。しかし、現在ではその技術は追加の科学的研究にも応用されている[64]。主要な資金源はエネルギー省と NSF であり、他に会員組織が資金を拠出している。

OSG に付属するものとして、ローレンス・バークレー国立研究所（Lawrence Berkeley National Laboratory）が管理する ESNet の一部である先端ネットワーキング・イニシアチブ（Advanced Networking Intitiative）では、大型データセットを転送するために 2 つの超高速ネットワークを備えている。それらは、毎秒 10Bb のエンド・ツー・エンドのサーキットをオンデマンドで提供する高性能かつ再構成可能なテストベッドと、それよりも小さな毎秒 100Gb のプロトタイプ･ネットワークである[65]。

---

[63] Machine Translation, Information Technology Laboratory, NIST

[64] Open Science Grid, A National, Distributed Computing Grid for Data-Intensive Research.

[65] Advanced Networking Initiative (ANI), ESNet, Lawrence Berkeley National Laboratory

ウィスコンシン大学率いる研究班によって開発された OSG1.0 は、コンピュータ･スケジューリングとデータ転送を司る中核ソフトウェアである。OSG は、大型ファイルの超高速転送のための gridftp を含む、グリッド･コンピューティング・ソフトウェアのグローバス（Globus）パッケージソフトウェアを含有している[66]。また、異種コンピュータを横断する分散型コンピューティングのためのコンドル（Condor）システムも含まれる[67]。

3.4.2 オープン･クラウド･コンソーシアム

オープン･クラウド･コンソーシアム（OCC: Open Cloud Consortium）は、約 15 の会員が所属する非営利団体であり、クラウド･コンピューティングのための参照実装（reference implementation）やベンチマーク、標準の開発を手掛けている[68]。OCC は、分散型コンピューティング･プラットフォームのオープン・クラウド・テストベッド（Open Cloud Testbed）を運用している。このテストベッドは、全米ラムダ・レイル（National Lambda Rail）が提供する毎秒 10Gb のネットワーク（現在、毎秒 100Gb へのアップグレードが実施されている）によって接続される、全米 4 箇所（ボルチモア、シカゴ、ラホヤを含む）の 9 ラックにある 1,000 超のコアを包含する。

また、OCC は、科学者による中・大型科学的データセットの管理、分析、統合、そして共有を可能にするクラウド・ベース・インフラ、オープン・サイエンス・データ・クラウド（Open Science Data Cloud）を管理しており、NSF から一部資金提供を受けている。オープン・サイエンス・データ・クラウドは、ゲノミクスをはじめ、天文学、地球科学の各領域の研究をホストしており、分散型データ集約型コンピューティング向けマスルトーン（MalStone）ベンチマークを開発、維持することでも知られる。

---

66 Globus Grid Software

67 Condor High Throughput Computing

68 Open Cloud Consortium

4 データ集約型コンピューティングにかかわる将来的な課題

本稿 1.2 章では、データ集約型コンピューティングへの移行に関わる技術的課題を挙げた。しかし、現在の研究結果を見る限り、それらは解決からはほど遠く、解決可能であることすら示されていない。そして、それら技術的課題に加えて、情報・コンピューティング技術分野の研究によるデータ処理や分析へのアプローチ方法において、より根本的な変化が必要であることが最近の経験から示されている。以下に示すこのような新しい課題は、技術的トピックから派生している一方で、データ集約型コンピューティングに関連する戦略、組織、経済における変化の必要性を指摘している。

4.1 データ集約型コンピューティング科学の創造

データ集約分野における最新の研究は、非常に探索的か、あるいはマップリデュースや BLAST など数件の並外れた成功に派生しているかの、どちらかのようである。データ集約型コンピューティングの真の科学を達成する方法、あるいはそれをより確立された科学と連携させる方法のための土台は、ほんとできていない。この意見は、批判のではない。というのも、新興科学のほとんどは、このようにして始まるからである。40 年強という開発期間を経た今も、コンピュテーショナル・モデリングとコンピュータ科学はその基盤を作っているところである。しかし、データ集約型コンピューティングの支持者は、将来の成果は、この科学が創造されるまで実現しないとの予測を示している。例えば、大変思慮深い研究者である Dan Reed 氏は、Vannever Bush 氏が提唱したメメックス（Memex）概念である「記憶を拡張させ、創出的な思考を支援すること」は、データ集約型コンピューティングを通じて創造が可能であることを示唆している[69]。しかし、今日の我々は、その実現からはまだ遠い。

スーパーコンピュータ用の最適なアーキテクチャの決定がコンピュテーショナル科学にとってそうであったように、データ集約型コンピューティングのための最適なコンピュータ・アーキテクチャの決定は、科学創造の重要な一部である。そのための案は出されているが、今日まで最も一般的なアーキテクチャの成功であるコモディティ・プロセッサのクラスターは、おそらく偶然の産物であった[70]。この課題には、フォルト・トレランスに関する問題、具体的には、収集された、または保存

[69] Clouds and Manycore: The Revolution, Dan Reed, Microsoft, 2008

[70] Data-Intensive Supercomputing: The Case for DISC, Randal Bryant, Carnegie Mellon University, 2007

されたデータのエラーやギャップ、ハードウェアまたはソフトウェアの脆弱性または誤動作、データ転送またはアグリゲーションにおける問題、そしてデータ･ストレージに対する陳腐化したフォーマットが含まれる。

科学とエンジニアリングにおいては、現行の①試験ツールまたは②観測ツール、③理論ツール、そして④コンピュテーショナル科学ツールを活用したデータ集約型コンピューティングの合成は、成功すれば非常に強力なものとなる。しかし今日では、後者 3 つのツール間の合成でさえ不完全であることから、データ集約型コンピューティングを追加するのは簡単ではない。

4.2 異種・非構造化データセットの統合

データ集約型コンピューティングの最大の成功は、データが小規模コミュニティの統制化にあるような、非常に限定的な制御された環境においてみることができる。前掲したように、これらの成功例としては、ウォルマートによって完全に制御されているトランザクション･データの管理や、わずか 4 つの別々の化合物の操作と BLAST のような検索プログラムの利用を行った遺伝子情報の解読が挙げられる。

非構造化データのパターン検索に際し、マップリデュースのようなツールを利用した部分的成功も達成されている。しかし、マップリデュースの開発会社であるグーグルは、マップリデュースによって戻されたパターン・マッチから理解、あるいはナレッジを得ることさえまだできていない。つまり、マップリデュースは、データから直接パターンを検知しており、なぜそれらのパターンが存在するのか、または、それらのパターンがどのように人による意思決定に影響を及ぼすべきか、といった見識機能のようなものは一切ない。

他のアプローチの中でも、おそらくセマンティックウェブは、この統合を達成する案としては最も歴史が古いアプローチだが、その実現可能性には賛否両論がある。人間の脳、ただし非常に記憶力のいい脳によって実行される統合を達成することは、データ集約型コンピューティングの中心的な課題である。

マップリデュースのアプローチを超えた次のステップの検索技術として、ウルフラム・アルファ（Wolfram Alpha）[71]と呼ばれる試みがある。このウルフラム・アルファは、定量的なデータ分析のための Mathematica（マセマティカ）ソフトウェアパッケージを開発したウルフラム・リサーチ

[71] WolframAlpha Computational Knowledge Engine

（Wolfram Research）社によって開発された技術である。ウルフラム・アルファの開発者、そして、ウルフラム・リサーチ社の設立者であるスティーブン・ウルフラム氏は、ウルフラム・アルファという新しい取り組みを、自然言語処理、セマンティックウェブ技術、構造型・非構造型データの計算操作の統合として特徴つけている。単に、パターンを特定するのではなく、ウルフラム・アルファ検索エンジンは、平文の言語質問を受入れ、その質問トピックに関連した複数のデータセットを特定し、これらのデータセットから特定のデータを抽出し、回答を合成することができる。この技術は、人間によって理解しやすいアウトプットを生成できることから、データ集約型コンピューティング分野の現在のプロジェクトにおいて最も洗練されたもののひとつといえる。しかしながら、ウルフラム・アルファの技術は、ユーザの質問を構成要素に分析し、それらの質問に関連した方法で多異種データ構造を解釈することにおいてより多くの作業が必要であるため、まだ部分的な成功でしかない。

4.3 大規模データの政策および法令

データ集約型コンピューティングの登場は、TIA プロジェクトをはじめ、ウェブ･ユーザー追跡、個人の医療および財務データ、ソーシャル・ネットワーキング・データ、そして知的財産の利用に関する論議によって例示されているように、何が収集可能であり、誰がそれを所有していて、適切な使用は何かなど、データをめぐる進行中の議論に拍車をかけた。

一般大衆は、ユニバーサル・データが手に入る便利さを求める一方で、政府であろうと大企業であろうと、「ビッグ・ブラザー（big brother）」すなわち独裁者に追われることを嫌っており、両極面を激しく提示している。この課題の解決は簡単ではない。それは一つには、公共の議論と新法によって解決可能であろうが、それにはいつも面倒なプロセスを伴うからである。この課題に対処するための技術的要因には、おそらく新たなデータ所有形態や（現在の米国の法律は、データ集合体（data collections）の著作権を認めていない）、データファイルに付加されるメタデータ、そして標準データ利用クエリーが含まれる。