O'REILLY®

Compliments of kinetica

基礎から学ぶデータ分析のためのGPU活用法

アクセラレーションコピューティングの進化と応用

Eric Mizell & Roger Biery

CPUに代わる GPUの役割

複雑な分析をCPUで対応することが困難になっています。この課題を解決するため、GPUの並列処理能力を活用するKineticaをゼロから設計しました。Kineticaは、最新の分析データベースであり、比類のない成果を実現します。CPUで構成されるインメモリデータベースと比較して、わずかなハードウェアを使用しながら100倍のパフォーマンスを実現します。Kineticaは、高度な分析処理、地理空間分析、機械学習などさまざまな分野で新しい可能性を切り開きます。

詳細は、kinetica.com をご覧ください。

kinetica

基礎から学ぶデータ分析のためのGPU活用法

アクセラレーションコンピューティングの進化と応用

*Eric Mizell*および*Roger Biery*

北京・ボストン・ファーンハム・セバストポル・東京 O'REILLY®

目次

まえがき

この数十年間にわたってCPUの価格は低下し、パフォーマンスは着実に向上してきましたが、CPUにおけるムーアの法則はついに終焉を迎えました。その理由は単純です。単一のチップに安価な方法で配置できるx86コアの数は実質的な限界に達しているのです。さらなる高密度化を目指して小型形状にしようとすると、非常に高価となり、ほぼすべての用途で使用できなくなってしまうのです。

このような制限から、プライベートおよびパブリッククラウドインフラストラクチャの両方を拡大するためにサーバーファームとクラスタが使用されるようになりましたが、このような力任せのスケーリングには多額の費用がかかり、データセンターで利用できる限られた設置面積、電力、および冷却リソースを使い果たしてしまう恐れもあります。

幸いなことに、データベース、ビッグデータ分析、機械学習アプリケーションに利用でき、コンピュート性能の拡張性を高め、優れた能力を発揮する、コスト効率の高いGPU（グラフィックスプロセッシングユニット）というソリューションを利用できます。GPUは、さまざまな用途で活用されてきた実績があります。GPUの設計は進歩を続けており、現在の企業が直面しているデータ量、多様性、スピードの絶え間ない増大や増加にも対応する理想的なソリューションとなります。

本書は、アクセラレーションコンピューティングテクノロジの進歩が、現在そして将来のデータベースとビッグデータ分析の課題にどのように対応するかを分かりやすく概説することを目的としています。本書の内容は、テクノロジ部門の幹部やプロフェッショナルを対象としていますが、ビジネスアナリストやデータサイエンティストにも適しています。

この電子ブックは、次の8つの章で構成されています。

- 1章、*「データ分析の進化」*は、データ分析におけるボトルネックがどのようにメモリI/Oからコンピュートへと変化し、現在の最大の課題となったのか、その経緯について説明します。
- 2章、*「GPU：革新的なテクノロジ」*では、グラフィックスプロセッシングユニットが、コンピュートに起因する制約をどのように克服し、継続的に価格を抑えながらパフォーマンスを向上できるようにするのかを説明します。
- 3章、*「新しい可能性」*では、GPUアクセラレーションの恩恵を受ける多くのデータベースやデータ分析アプリケーションについて説明します。
- 4章、*「機械学習とディープラーニング」*では、GPUデータベースでユーザー定義関数を使用すると、機械学習/ディープラーニングのパイプラインをどのように簡素化し加速できるかを説明します。
- 5章、*「モノのインターネットとリアルタイムデータ分析」*では、GPUアクセラレーションデータベースが、モノのインターネットやその他のソースのストリーミングデータをリアルタイムに処理する方法を説明します。
- 6章、*「インタラクティブなロケーションベースインテリジェンス」*では、要件の厳しい地理空間アプリケーションでも、GPUデータベースのパフォーマンスメリットを享受できる理由について詳述します。
- 7章、*「コグニティブコンピューティング：分析の未来」*では、膨大なコンピュートリソースが必要とされるコグニティブコンピューティングアプリケーションにも、GPUを活用できるようにするビジョンを説明します。
- 8章、*「新たな一歩を踏み出す」*では、GPUアクセラレーションソリューションをオンプレミスで、また、パブリック、プライベート、およびハイブリッドクラウドアーキテクチャに導入する方法を概説します。

1章
データ分析の進化

データ処理は、メインフレームコンピュータが登場して以来、継続的に進化し、その性能が強化されてきました。図1-1は、1990年以降のデータ分析の進化における4つの段階区分を示しています。

図1-1
ムーアの法則に基づいて、CPUはその価格を低下させながら、パフォーマンスを絶えず向上してきましたが、これはデータ分析アーキテクチャについても同じです。

1990年代には、データウェアハウスとリレーショナルデータベース管理システム（RDBMS）テクノロジにより、企業は安価にそして納得できるパフォーマンスでデータを保存および分析できました。ストレージエリアネットワーク（SAN）とネットワーク接続ストレージ（NAS）は、これらのアプリケーションで広く使用されてきました。しかし、データ量が増加し続けたため、このアーキテクチャでパフォーマンスを増強するには多額のコストがかかるようになりました。

2005年頃には、I/Oパフォーマンスを向上させるためにDAS（直接接続ストレージ）を活用した分散型サーバークラスタにより、データ分析アプリケーションを手頃な費用で拡張できるようになりました。HadoopとMapReduceは、サーバークラスタの特長である並列処理能力を活かすことができるように特別に設計されており、普及が進みました。このアーキテクチャは、バッチ処理型のデータ分析アプリケーションにおいて現在でも優れた費用対効果を発揮しますが、データストリームをリアルタイムに処理するためのパフォーマンスは欠落しています。

2010年までに、数テラバイトのランダムアクセスメモリ（RAM）を搭載したサーバーを低価格で構成できるようになり、インメモリデータベースを手頃な価格で利用できるようになりました。RAMへの読み取り/書き込みアクセスが劇的に向上したため（DASの場合100ナノ秒と10ミリ秒の差異）、パフォーマンスが劇的に向上しました。しかし、あらゆる部分におけるパフォーマンスが進化したため、アプリケーションの数が増大し続ける現在の環境の中で、今度はコンピュートがボトルネックになっていきました。

しかし、最近はGPUアクセラレーションコンピュートを利用できるようになったことで、このパフォーマンスボトルネックは解消できるようになっています。2章で説明しますが、GPUは、スケールアウトとスケールアップの両方に対応する極めて強力な並列処理能力を備えており、ほぼすべてのデータベースやデータ分析アプリケーションで、比類のないレベルのパフォーマンスを達成し、価格を抑えながらパフォーマンスの大幅な向上を実現します。

データ分析の現在の課題

パフォーマンスの問題がビジネスユーザーに影響を与えている：

- インメモリデータベースのクエリー応答時間は、カーディナリティが高いデータセットで大幅に低下する。
- 取り込みとクエリーを同時に実行するのは難しく、ライブストリーミングデータで許容される応答時間を実現することが困難。

価格/パフォーマンスの向上が困難：

- 商用のRDBMSソリューションでは費用対効果に優れる方法でスケールアウトできない。
- x86ベースのコンピュートは、データ量と速度が激増すると、そのコストが許容できなくなる場合がある。

ソリューションが複雑であり、新しいアプリケーションの障害となっている：

- 期待されるパフォーマンスを達成するため、データ統合、データモデル/スキーマ、およびハードウェア/ソフトウェアの最適化において、頻繁な変更が必要となることが多い。
- 必要なスキルセットをすべて有するITスタッフを雇用し維持することがますます困難になっており、多額の人件費が必要。

2章

GPU：革新的なテクノロジ

CPU、メモリ、ストレージ、およびネットワーキングテクノロジは着実に進化を遂げており、手頃な価格で拡張性に優れるハイパフォーマンスなデータ分析のための基盤は既に確立されています。1章で説明しましたが、このような進化によって、パフォーマンスのボトルネックはメモリI/Oからコンピュートへとシフトしました。

高速で大規模な処理に対応できるように、現在、CPUには最大で32個のコアが搭載されていますが、大規模なサーバークラスタにマルチコアCPUを組み込んで使用しても、一握りの組織を除けば、高度な分析アプリケーションのコストを負担することは容易ではありません。

GPU（グラフィックスプロセッシングユニット）は、コンピュートパフォーマンスのボトルネックを解決する極めて優れた効果を発揮します。GPUを活用すると、CPUのみを実装する構成と比較して、最大で100倍高速にデータを処理できます。これだけ劇的に向上できる理由は、現在最もパワフルなCPUでもコア数は16〜32個ですが、一部のGPUはその200倍近い6,000個のコアを搭載しており、膨大な並列処理能力を発揮できるためです。例えば、Tesla V100は、最新のNVIDIA Volta GPUアーキテクチャを使用し、5,120個のNVIDIA CUDAコアと640個のNVIDIA Tensorコアを搭載しており、シングルGPUで最大100個のCPUパフォーマンスを実現します。

GPUの小型で効率的なコアは、類似性のある繰り返しの命令を並列処理するのに最適であり、今日のデータ解析アプリケーションでよく見られる処理負荷の高いワークロードを高速に処理する場合に理想的です。

パフォーマンスの向上をさらに低コストで

あるアプリケーションでは、シンプルな2ノードクラスタで、150億のツイートが含まれるGPUデータベースをクエリーし、1秒かからずにビジュアライゼーションを処理できました。各サーバーには、2.6GHzで動作する12コアのXeon E5プロセッサー2台とNVIDIA K80カード2枚が搭載され、合計で4個のCPUと4台GPUが搭載されました。

GPUの進化

GPUはその名前が示すように、当初はグラフィックスの処理に使用されていました。第1世代のGPUは、自身のメモリ（ビデオRAM、VRAM）を備えた別個のビデオインターフェイスカードに設置されていました。この構成は、高品質なリアルタイムグラフィック処理を追及するゲーマーに高い人気を博しました。時が経つにつれ、GPUの処理能力とプログラマビリティの両方が向上し、さまざまなアプリケーションに活用できるようになりました。

ハイパフォーマンスコンピューティングアプリケーション向けに設計されたGPUアーキテクチャは、当初、汎用GPU（GPGPU）として分類されましたが、グラフィックスとデータ分析アプリケーションで高速浮動小数点演算のための基本要件が共通していることがわかると、GPGPUという名前は若干おさまりが悪いこともあり、すぐに使用されなくなりました。

その後、フルプログラマブルGPU世代が登場し、ホストサーバーのCPUとメモリで、コア数の増加とI/Oの高速化という2つの方法でパフォーマンスを向上させました。例えば、NVIDIAのK80 GPUは4,992個のコアを搭載しています。また、ほぼすべてのGPUアクセラレータカードでは、現在16レーンのPCIeインターコネクトで双方向帯域幅が32GBpsのPCI Expressバスを使用しています。これだけのスループットであれば、大半のアプリケーションのニーズを満たすことができますが、CPUとGPU間やGPU間で5倍の帯域幅（160 GBps）を実現したNVIDIAのNVLinkテクノロジによってさらなる恩恵がもたらされます。

最新世代のGPUカードでは、図2-1に示すように、メモリ帯域幅が大幅に増加し、転送速度は最大732GBpsに達しています。この帯域幅を、Xeon E5 CPUの68GBps（PCIe x16バスの帯域幅の約2倍）と比較する

と、その違いは一目瞭然です。数千のコアに対応するこのような高速I/Oの組み合わせにより、16GBのVRAMを搭載したGPUカードは、9TFLOPS（1秒間の浮動小数点演算数）を超える単精度演算性能を実現します。

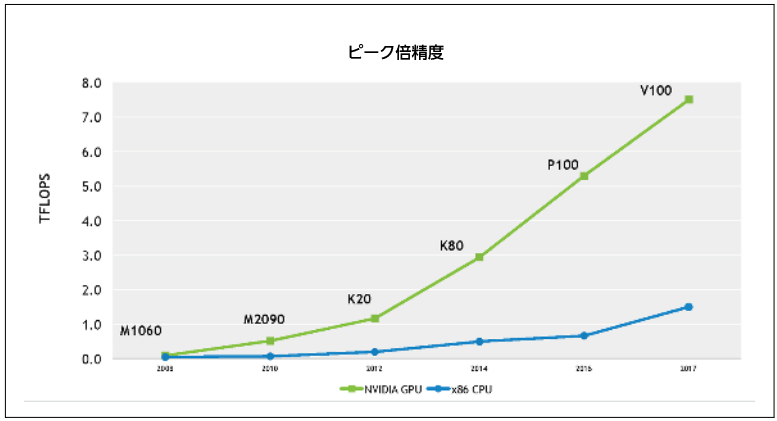

図2-1
マルチコアのx86 CPUは時間の経過とともにその性能が若干向上していますが、NVIDIAの最新世代のGPUには、約6,000個以上のコアが搭載されており、最高7.5TFLOPSの倍精度演算性能を実現します。（出典：NVIDIA）

スモールデータだけでなくビッグデータ分析にも高い効果を発揮するGPU

サーバーで現在サポートされている数テラバイトのRAMと比較すると、GPUカードのVRAMは比較的小規模であるため、GPUアクセラレーションは「スモールデータ」を扱うアプリケーションにしか活用できないと考えるユーザーがいます。しかし、そのような考え方では、「ビッグデータ」アプリケーションで多く見られる2つの手法が無視されています。

まず、ビッグデータの分析では、期待される結果を得るためにデータセット全体を一度に処理することはほとんどありません。GPU VRAM、システムRAMとストレージ（直接接続ストレージ、ストレージエリアネットワーク、ネットワーク接続ストレージなど）を横断する階層的なデータ管理によって、ビッグデータワークロードに対して事実上無制限にスケーリングできるようになります。例えば機械学習では、訓練デー

タは必要に応じてメモリやストレージからストリーミングできます。モノのインターネット（IoT）および、KafkaやSparkなどの他のアプリケーションからのデータのライブストリームも、同じように「断片的に継続する」方式で処理できます。

2つ目の手法は、GPUアクセラレーション構成は垂直方向と水平方向の両方にスケーリングできることです。複数のGPUカードを1台のサーバーに配置し、クラスタ内に複数のサーバーを構成できます。このようなスケーリングによって、多くのコアとメモリを利用できるようになり並列で同時かつ高速に動作し、比類のないスピードでデータを処理します。つまり、GPUアクセラレーションの潜在的なプロセシングパワーにおける唯一の制限は予算です。

しかし、使用できる予算がどの程度があっても、CPUはGPUよりはるかに高価であるため、GPUアクセラレーション構成では1ドルあたりのFLOPSを向上することが可能です。つまり、単一のサーバーであってもクラスタであっても、GPUデータベースには価格/パフォーマンスにおける明確で大きな優位性があるのです。

3章
新たな可能性

GPUアクセラレーションがもたらすパフォーマンス向上のメリットは、さまざまなアプリケーションでさまざまな効果を発揮します。図3-1に示すように、通常、高い処理能力が求められるアプリケーションになるほど、メリットは大きくなります。

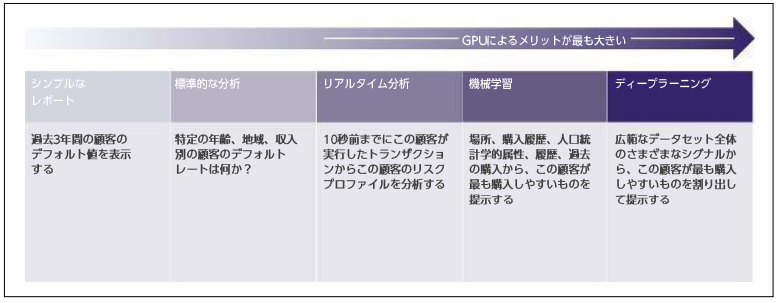

図3-1
ほぼすべてのデータ分析アプリケーションが、GPUの優れた価格/パフォーマンスのメリットを享受できますが、最も高い処理能力を必要とするアプリケーションでそのメリットは最大になります。

この章では、GPUアクセラレーションを使用して、さまざまなデータベース、データ分析、ビジネスインテリジェンス（BI）アプリケーションのパフォーマンスを向上し、コストを低減する方法について説明します。次の3つの章では、最も大きなメリットを受けることができる3つのアプリケーションまたはユースケースを中心に説明します。

- 機械学習とディープラーニング（4章）
- モノのインターネット（IoT）とリアルタイムデータ分析（5章）
- インタラクティブなロケーションベースインテリジェンス（6章）

高速/フルテキスト解析と自然言語処理

多くのデータ分析アプリケーションに共通する要件は、テキスト分析と自然言語処理（NLP）ですが、この要件こそが、GPUアクセラレーションがどのように補完的なメリットをもたらすのかを示す好例になっています。大規模な並列処理により、GPUは大規模なデータセットに対して以下の（これに限定されません）分析をリアルタイムで実行できます。

- 正確なフレーズ
- AND/OR
- ワイルドカード
- グループ化
- あいまい検索
- 近接検索
- 数値範囲

相互運用性と統合のための設計

さまざまなGPUベースのデータベースとデータ分析ソリューションが存在し、その機能は異なりますが、それらのすべてが既存のアプリケーションやプラットフォームを補完する、または統合されるように設計されています。一般的な手法のいくつかについてここで概説します。

ハードウェアの説明から始めましょう。ほぼすべてのGPUベースのソリューションが、x86 CPUを搭載した一般的な業界標準サーバーで動作するため、費用対効果の高い方法で構成をスケーリングでき、期待するパフォーマンスを達成できます。

GPUまたはVRAMを追加するか、さらに高速化して、通常はスケールアップできます。複数のGPUカードを搭載するサーバーのパフォーマンスは、NVLink（2章で説明）を使用することで、16レーンのPCIeバスで利用可能な5倍の帯域幅を使用でき、さらなるスケールアップを図ることができます。

スケールアウトする場合、クラスタに複数のサーバーを追加するだけで済み、分散型構成にして、信頼性を高めることもできます。

柔軟性を高めるために、GPUソリューションをオンプレミスまたはパブリッククラウドに展開できます。

ソフトウェア面では、ほとんどのGPUベースのソリューションはオープンアーキテクチャを採用しているため、事実上あらゆるアプリケーションに容易に統合でき、パフォーマンスの向上とコスト削減によるメリットがもたらされます（図3-2を参照）。従来型のリレーショナルデータベースや機械学習やディープラーニングなどの人工知能から、ストリーミングデータのリアルタイム分析や複雑なイベント処理を必要とするアプリケーションが、メリットを享受できる可能性があります。

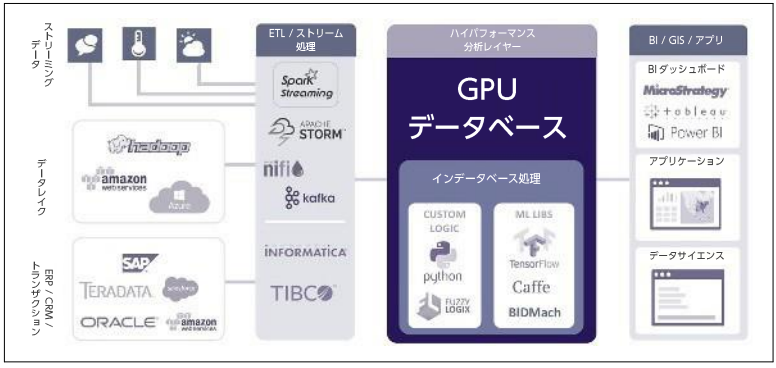

図3-2

GPUデータベースでオープンアーキテクチャが使用されており、さまざまな分析アプリケーションやBIアプリケーションに簡単に統合できます。

GPUデータベースは、例えば、Hadoopの高速クエリレイヤーとして使用できるなど、補完的な役割を担うこともできます。GPUアクセラレーションソリューションは、超低遅延のパフォーマンスを実現するため、大量のストリーミングや大規模で複雑なデータの同時処理と分析が必要となるアプリケーションにとって理想的です。

さまざまなビジネスに活用できる

大半のGPUアクセラレーションデータベースではオープンな設計が採用されており、さまざまなデータ分析アプリケーション、環境、ニーズに対応できます。オープンな設計の内容の例をいくつか示します。

- Accumulo、H2O、HBase、Kibana、Kafka、Hadoop、NiFi、Spark、およびStormなど、人気の高いオープンソースフレームワークとの統合を簡素化するコネクタ
- 既存のビジュアライゼーションやTableau、Power BI、SpotfireなどのBIツールとのシームレスな統合を可能にするOpen Database Connectivity（ODBC）およびJava Database Connectivity（JDBC）用のドライバ
- SQL、C ++、Java、JavaScript、Node.js、Pythonなど、一般的に使用されるプログラミング言語とのバインドを可能にするAPI
- 地理空間ビジュアライゼーションアプリケーションで使用されるジオリファレンスマップ画像を統合するためのWeb Map Service（WMS）プロトコルのサポート

ミッションクリティカルアプリケーションにおいてもGPUが利用されることを見越して、現在、多くのソリューションが高可用性と堅牢なセキュリティの両方を実現するように設計されています。高可用性機能として、2台以上のサーバークラスタ内で自動フェールオーバーを設定するデータレプリケーションと、データを個々のサーバーのハードディスクやソリッドステートストレージに保存することによってデータの整合性を確保する機能を利用できます。

セキュリティについては、ユーザー認証のサポートとロールベースおよびグループベースの承認がサポートされており、個人のプライバシーを保護する必要があるアプリケーションなど、政府の規制に遵守しなければならないアプリケーションにもGPUアクセラレーションを活用できます。これらの機能拡張により、パブリックおよびプライベートクラウドインフラストラクチャの両方で組織がGPUアクセラレーションを採用するあらゆるリスクが完全に排除されています。

一部のGPUベースのソリューションは、インメモリデータベースとして導入され、メモリ内で動作する他のデータベースと同様の機能を利用できます。GPUアクセラレーションデータベースの良し悪しは、大規

模な並列構成におけるピークパフォーマンスに合わせてどのようにデータを保存して処理するかによって決まります。

図3-3に示すように、GPUデータベースでは、利用可能なすべてのGPU全体での処理を最適化するために、データは通常、ベクトル化された列でシステムメモリに保存されます。次に、必要に応じてGPUのVRAMにデータが移動され、数学的および空間的なすべての演算が実行され、結果がシステムメモリに戻されます。小規模なデータセットやライブストリームの場合には、GPUのVRAMに直接データを保存して処理を高速化できます。システムメモリまたはVRAMのいずれに保存する場合でも、すべてのデータをハードディスクまたはソリッドステートドライブに保存でき、データは決して失われることはありません。

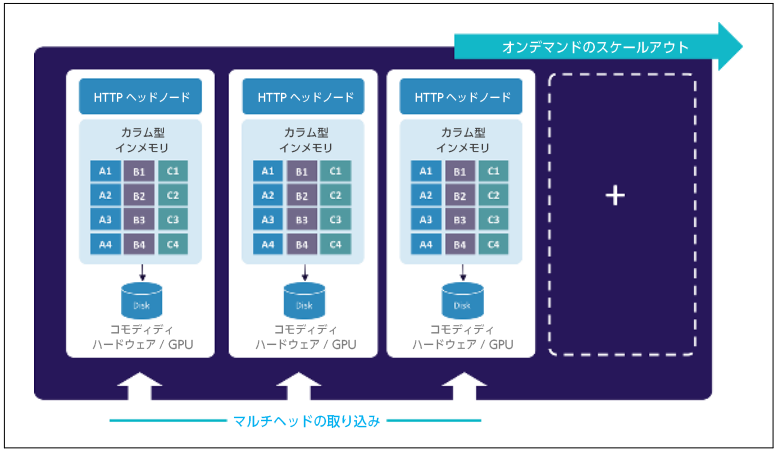

図3-3

GPUアクセラレーションインメモリデータベースは、多くのデータ分析およびビジネスインテリジェンスアプリケーションで優れたパフォーマンスを実現する「高速レイヤー」になります。

4章

機械学習とディープラーニング

機械学習（ML）とディープラーニング（DL）は、データに存在する有用なインサイト（知見）を発見することで、企業が詳細な分析から予測分析へと移行できるようにする実用的なテクノロジとして登場しました。ML/DLモデルは膨大なデータセットを処理し、そのパターン、異常、および関係性を自動的に発見し、データ連動型の意思決定を下し、大きな影響力をもたらすことがあります。

しかし、企業でMLを導入する場合、いくつかの課題もあります。これらの課題を克服し、MLのメリットを完全に活かすためには、ハードウェアアクセラレーションGPUデータベース、インメモリデータ管理、分散コンピューティング、TensorFlowなどの統合型でオープンソースのMLフレームワークのようなテクノロジを使用します。これによって、シンプルで使いやすく集約型のMLソリューションを利用できるようになります。

MLアプリケーションは、ユーザー定義関数（UDF）が登場したことで、その実装がさらに容易になりました。UDFにより、シングルデータベースプラットフォームでリアルタイムにフィルタリングされたデータを受け取り、計算を実行し、別のテーブルに出力を保存できるようになります。データ生成、モデルトレーニング、モデルサービス機能の3つの主要プロセス（図4-1）を、GPUの大規模な並列処理能力を活用し、求められるパフォーマンスを提供する単一のソリューションに統合することで、MLのパイプライン全体を簡素化し加速できます。

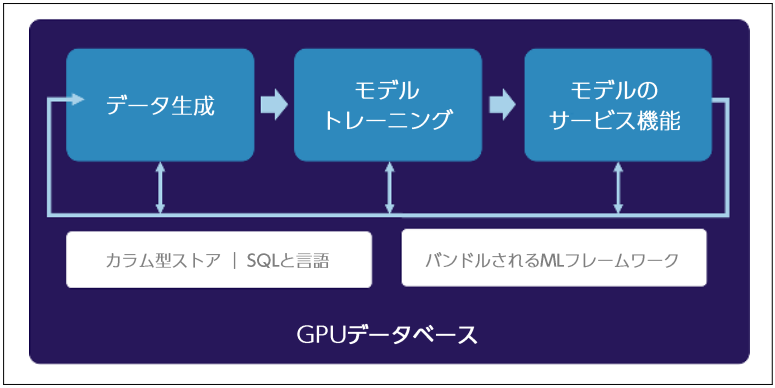

図4-1
GPUデータベースはMLのパイプラインを加速し、モデルの開発と展開を高速化します。

データ生成タスクには、機械学習モデルを訓練するためのデータセットの取得、保存、準備が含まれます。GPUデータベースは、これらの3つのすべてのデータ生成タスクに利点をもたらします。

- データ取得タスクでは、転送中および保存中のデータの両方を高速に取り込むことができるコネクタによって、分散しているシステム全体で数百万行のデータを数秒で簡単に取得できます。
- データ保管タスクでは、単一のGPUデータベースにさまざまな構造のデータ型を保存して管理する機能によって、ML/DLアプリケーションがすべてのテキスト、画像、空間および時系列データに簡単にアクセスできるようにします。
- データ準備タスクでは、SQL、C++、Java、Pythonなどの一般的な言語を使用してミリ秒の応答時間を実現でき、極めて大規模なデータセットであっても容易に探索できるようになります。

モデルトレーニングは、MLパイプラインで最も多くのリソースが必要となるステップであり、ボトルネックとなることが多くあります。GPUデータベースではUDFを利用でき、インラインモデルトレーニングのために、プラグインのカスタムコードやTensorFlow、Caffe、Torch、MXNetなどのオープンソースのMLライブラリをサポートするために必要なパフォーマンスが提供されます。GPUデータベースは、次の3つの方法でパフォーマンスを最大化します。

- アクセラレーション - GPUは大規模な並列処理に優れており、膨大なデータセットを集中的に処理するモデルトレーニングワークロー

ドに最適です。データサンプリングや多額の費用と多くのリソースが必要となるチューニングも不要になり、コモディティハードウェアを使用してパフォーマンスを100倍の向上できます。

- 分散型のスケールアウトアーキテクチャ - GPUデータベースはクラスタリングでき、複数のデータベースシャードにデータを分散できるため、モデルトレーニングを並列処理しパフォーマンスを向上できます。スケールアウトアーキテクチャにより、必要に応じてノードを追加してパフォーマンスとキャパシティを簡単に向上できます。
- ベクトル演算と行列演算 - GPUデータベースは、専用のインメモリデータ構造と処理最適化によって、最新のGPUで利用可能な並列処理を最大限に活用し、MLのワークロードで一般的なベクトル演算と行列演算のパフォーマンスを飛躍的に向上します。

図4-2に示すように、MLフレームワークをバンドルし、データ生成とモデルトレーニングに使用するのと同じGPUデータベースにモデルを展開することで、モデルサービス機能は、MLの操作性を高める能力を活用できます。このような統合により、モデルをインラインで評価でき、スコアリングを高速化でき、予測精度を向上できます。

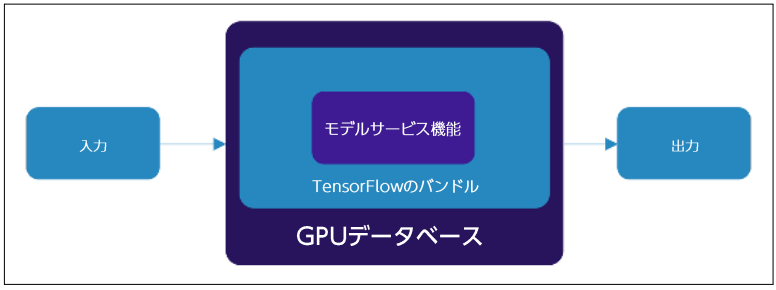

図4-2
MLフレームワークをバンドルし、GPUデータベースにモデルを展開することで、モデルサービス機能をインラインプロセスにすることができ、MLの操作性を高めることができます。

GPUデータベースは、コンピュートとモデル管理によって、データを統合し、あらゆる規模のデータの探索と準備を可能にします。GPUデータベースは、また、モデルトレーニングを加速し、本番環境へのモデルの展開を容易にし、モデルのライフサイクルを簡単に管理し、あらゆるMLプロジェクトの中核となるワークフローを簡素化します。

5章

モノのインターネットとリアルタイムデータ分析

ライブデータには膨大な価値があります。しかし、その価値が生まれるのはライブデータをリアルタイムに処理できる場合に限られます。これらのデータストリームをリアルタイムに処理・分析するために必要な処理能力がなければ、アプリケーションで扱うことができるデータ量と速度が制限されたり、処理は遅すぎてデータ本来の価値が失われたりして組織はビジネスチャンスを逸する恐れがあります。

このようなスピードに関するニーズは、モノのインターネット（IoT）にとって特に重要となります。IoTは、固定されているデバイスとモバイルデバイスの両方のコネクティッドデバイスから実用的なインサイトを引き出す絶好の機会を提供し、これらのデバイスをよりインテリジェントに、つまり効果的に動作させることを可能にします。

IoTが登場する前であっても、ライブデータをリアルタイムで（多くの場合、保存されているデータと一緒に）分析する必要性は広く高まっていました。組織によって業界固有のストリーミングデータソースを扱うことがありますが、ほぼすべての組織には、データネットワーク、Webサイト、インバウンドおよびアウトバウンドコール、暖房および照明制御、稼動記録、ビルセキュリティシステムやその他のインフラストラクチャがあり、これらのすべてが潜在的な価値のある（しかし、その価値は失われやすい）データを常に生成しています。

現在、IoT（一部の有識者はIoE（Internet of Everything）と呼んでいます）環境において、データをストリーミングするデバイス数は、さまざまな情報源をもとに、2020年までに300億台以上に拡大すると予測されています。

IoTを最大限に活用するために求められる処理能力やその他の能力があるのは、GPUデータベースだけです。特に、GPUには膨大な数の小型の効率的なコア全体で繰り返し実行される同じような命令を並列処理する高い能力があり、IoTアプリケーションにとって理想的です。多くの「モノ」が時間と場所の両方が重要となるデータを生成します。そのため、GPUの地理空間機能を活用することができ、最も要求の厳しいIoTアプリケーションへの対応も可能になります。

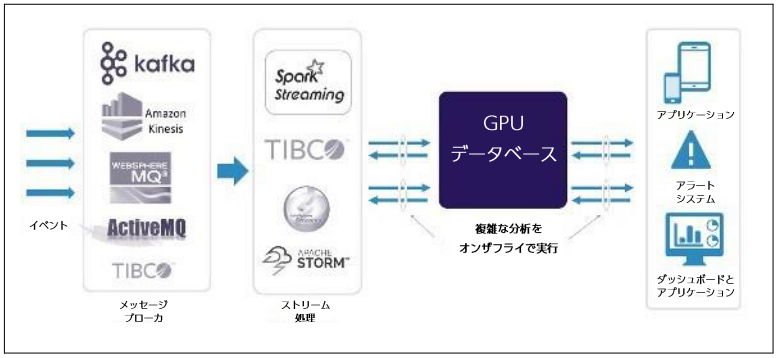

図5-1
GPUデータベースはストリーミングデータをリアルタイムで取り込み、解析し、処理することができるため、IoTアプリケーションにとって理想的です。

このような理由から、Ovum社は、IoTのユースケースにおいて、リアルタイムストリーミングの画期的な処理を可能にしたGPUの能力を評価し、GPUデータベースを同社の「*2017 Trends to Watch*」で革新的な成功事例にしています。

GPUデータベースにはIoTのストリーミングデータをリアルタイムで取り込み、分析および処理する能力があるため、あらゆるIoTのユースケースへの活用が可能です。IoTのユースケースは、業界および組織によって大きく異なりますが、GPUの優れた能力と可能性を示す3つの例を紹介します。

- カスタマーエクスペリエンス - GPUデータベースは、顧客情報やオンラインアカウントなどのさまざまなソースから、購買行動をリアルタイムで監視・分析するための情報を取り込むことができます。これは、POSシステム、ソーシャルメディアストリーム、天気予報、および他の情報ソースのデータを相関させ、全方位的に顧客をとら

える「Customer 360」アプリケーションを展開する小売業者にとって特に有用です。

- サプライチェーンの最適化 - GPUデータベースを使用して、サプライヤ、ディストリビュータ、物流、運輸、倉庫、小売店舗の場所など、サプライチェーン全体でリアルタイムの位置ベースのインサイトを提供できるため、企業は需要を正確に理解し、供給を適切に管理できるようになります。
- フリートマネージメント — 車両を所有し運営する公共部門の機関や企業は、GPUデータベースを使用してライブデータをフリートマネジメントシステムに統合できます。リアルタイムに位置を追跡するIoTアプリケーションは、GPUの地理空間処理能力を有効に活用できます。

IoTは今なお成長を続けており、組織に多くの可能性をもたらしています。GPUデータベースを利用することで、IoTを最大限に活用できるようになります。IoTのインサイトを活用できるオンライン分析処理やその他のビジネスインテリジェンス（BI）アプリケーションでは、SQL-92やBIツールなどの標準をサポートするGPUアクセラレーションデータベースを利用できます。また、このようなアプリケーションで求められることも多い高可用性と堅牢なセキュリティをサポートするGPUデータベースも利用できます。

6章
インタラクティブなロケーションベースインテリジェンス

現在、ほぼすべての組織で、少なくとも一部のデータをリアルタイムに処理することが求められており、データ分析アプリケーションに位置情報を統合することも重要な課題になっています。

車両やスマートフォンなどのモバイルソースから利用可能なデータが増加しており、これらのデータの地理空間的な意味を分析して視覚化することで、さらに多くのビジネスチャンス機会が得られます。しかし、主に静的なマップを作成するために設計されている従来型の地理空間マッピングツールでは、現在のニーズにはほとんど対応ができません。

あらゆるインタラクティブな操作に対応しながら、大規模なデータセットを分析し、ニアリアルタイムに大規模な地理空間を解析するためには、現在最も強力なCPUでも十分な計算能力を提供できないこと、また、収集した地理空間情報から点、線、ポリゴンを極めてシンプルなビジュアルでしかブラウザでレンダリングできないという、2つの課題を克服する必要があります。

GPUは、グラフィックス処理がそのルーツであり、大規模なデータセットの地理空間アルゴリズムをリアルタイムで処理し、その結果をレンダリングしてマップベースのグラフィックスとして一般的なブラウザに瞬時に表示するのに特に適しています（図6-1を参照）。GPUアクセラレーションデータベースでは、また、単一のプラットフォームで結果を取り込み、分析し、レンダリングできるため、求める結果を得るために異なるレイヤーやテクノロジ間でデータを移動する必要もありません。

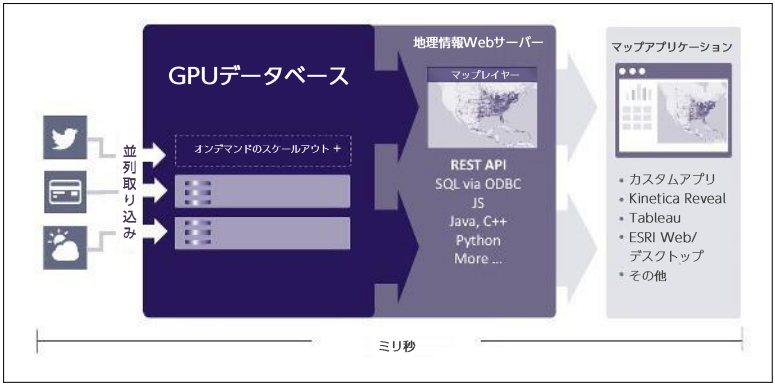

図6-1
GPUアクセラレーションデータベースは、ニーズが高まっているインタラクティブな位置情報ベースの分析に理想的です。

GPUの大規模な並列処理能力は、地理空間オブジェクトとネイティブフォーマットでの操作の両方をサポートできます。領域、軌跡、独自形状、ジオメトリ、またはその他の変数によるフィルタリングなどの地理空間処理をデータベース上で直接実行する機能があるため、最高のパフォーマンスを実現します。点、線、ポリゴン、軌跡、ベクトル、ラベルなどの地理空間オブジェクトを標準形式でサポートしており、未加工のデータの取り込みや他のシステムへの結果のエクスポートも簡単に実行できます。

ヒートマップ、ヒストグラム、および散布図を含むさまざまなビジュアライゼーションとしてブラウザで結果をレンダリングする場合、高品質なユーザーエクスペリエンスを確実に提供するうえで、標準が非常に重要になります。ほとんどのグラフィカル情報システム（GIS）データベースは、Open Geospatial Consortiumが推進している標準をサポートしており、これらの標準をサポートするGPUデータベースの数も増えています。OGC標準は、GISイメージを一般的なグラフィックス形式に変換する方法と、GPUデータベースに直接組み込むことができる標準のWebサービスソフトウェアを使用してグラフィックをどのように転送するかを指定しています。

このアプローチにより、Google、Bing、ESRI、MapBoxなどの主要なマッピングプロバイダのデータを簡単に統合でき、ユーザーは、これらのビジュアライゼーションをさらに活用したり、結果の表示方法を変更したりできるようになります。一部のソリューションでは、分析アプレット、データテーブル、およびその他の「ウィジェット」をドラッグアンドドロップするだけで、完全にカスタマイズされたダッシュボードを作成できるようになっています。

ユーザー定義関数（UDF）を使用すると、カスタムコードをGPUデータベースで直接実行でき、地理空間の分析をさらに拡張できます。このような分析手法をデータに適用することで、データを別のシステムに抽出する必要がなくなります。

このようなカスタマイズ形態によって、高度な地理空間予測のために、TensorFlowなどの機械学習ライブラリを使用するなど、さまざまな可能性が生まれます。機械学習は、たとえば、交通状況に基づいて時間通りに到着する可能性の低い配達にフラグを設定したり、運転に基づいて事故に遭う可能性が最も高いドライバーを予測したり、天気モデルを基準として資産の保険リスクを計算したりすることを可能にします。

地理空間データとリアルタイムでやりとりできるため、ビジネスアナリストは優れた意思決定を迅速に下すことができます。GPUデータベースは低コストで優れたパフォーマンスを実現するため、現在、その能力をほぼすべての組織が利用できます。

地理空間データの多くの次元

GPUアクセラレーションデータベースは、宇宙のような4次元の時空に存在する地理空間データをリアルタイムで処理するのに理想的です。3つの空間次元は、ベクトル（点、線、ポリゴン/形状）またはラスターイメージデータに基づくネイティブオブジェクトタイプを利用できます。ラスターイメージデータは、通常、インタラクティブな位置ベースのアプリケーションで使用されるマップオーバーレイイメージを生成するためにBaseMapプロバイダによって利用されます。

地理空間データの操作に使用される各種の機能（その多くは4次元で動作します）によって処理負荷が増大しますが、GPUアクセラレーションソリューションはこれらの負荷に対しても理想的です。これらの機能の例を以下に示します。

- 領域、属性、系列、ジオメトリなどによるフィルタリング
- ヒストグラムなどでの集約
- トリガーに基づくジオフェンシング
- イベントのビデオの生成
- ヒートマップの作成

実環境におけるユースケース

ここでは、さまざまな業界組織におけるGPUアクセラレーションソリューションの活用事例をいくつか紹介します。

大手製薬会社は、医薬品開発プロセスでGPUデータベースを利用し、化学反応シミュレーションを高速化できることを発見しました。化学反応データを複数ノードに分散することにより、シミュレーションを迅速に実行でき、新薬開発までの期間を大幅に短縮できます。研究者は、SQLなどの言語を使用して、最初に従来のRDBMS環境で分析を実行した後に、必要に応じてGPUアクセラレーションデータベースで同じ分析を実行できます。

大手ヘルスケアプロバイダは、リアルタイムのGPUアクセラレーションデータウェアハウスを使用して処方薬における不正を低減し、動的な地理空間分析とヘルスケアのモノのインターネット（IoT）データを使用し、Patient 360（全方位的な患者データ管理アプリケーション）を強化しています。

ある大規模な公益事業組織は、予測インフラストラクチャ管理（PIM）にGPUアクセラレーションデータベースを使用しています。GPUデータベースは、インフラストラクチャのヘルスを監視、管理、予測するための俊敏な処理を実現するレイヤーとして機能します。この公益事業組織は、GPUアクセラレーションを利用して、各地に展開されている資産の位置データなどの複数のデータフィードを一元的なデータストアに同時に取り込み、分析して、モデル化しています。

国際的な大手銀行は、GPUアクセラレーションデータベースを使用して取引先顧客のリスクを分析しています。このリスク分析はこれまでは一晩かかっていた処理でしたが、トレーダー、監査人、経営幹部がリアルタイムに利用できるアプリケーションになりました。特定の取引が処理されるときに銀行にトレーディング勘定の公正価値を決定するように求めた新しい規制が施行されたために、この変更が取り入れられました。今後数年にわたる評価調整を予測する必要があるため、そのリスクアルゴリズムは処理が複雑すぎて膨大な演算処理が必要となり、CPUのみの構成では対応が困難になっていました。

世界最大手の小売業者の1つは、サプライチェーンと在庫管理を最適化するためにGPUアクセラレーションデータベースを使用しています。小売業者のアナリストは、GPUデータベースを利用することで、ソーシャルメディアの感情分析、購買行動、オンラインおよび実店舗の購

買などの、顧客情報を統合でき、これまで数時間を要していたクエリーを1秒以内に実施できるようになりました。このアプリケーションはさらに拡張され、天候やウェアラブルデバイスに関連するデータを追加して、顧客行動をより正確に表示できるようになりました。

米国郵政公社（USPS）は、米国最大の物流公社であり、年間を通じてみると、UPS、FedEx、DHLよりも4時間早く個人向け荷物を配送しており、何十万台もの車両を使用して1億5400万以上の住所に毎日配送しています。運送・輸送状況をより明確に可視化するために、すべての郵便集配人は荷物をスキャンするデバイスを利用しています。このデバイスは、すべての配送ルート効率を最大化するなど、大規模な操作のさまざまな要素を改善するために毎分正確な地理的な位置情報を発信しています。このGPUデータベースは、合計200,000を超えるスキャンデバイスからのデータストリーミングを分析し、15,000の同時セッションをサポートしています。

7章

コグニティブコンピューティング：分析の未来

人間の思考や推論をリアルタイムでシミュレーションするコグニティブコンピューティングは、ビジネスインテリジェンス（BI）の究極の目標と言われることもあります。IBMのスーパーコンピューターであるWatson（ワトソン）は、既存のテクノロジを利用してこの目標を達成できることを実証しました。

しかし、本当の課題は今後にあります。コグニティブコンピューティングはいつ実用的になり、多くの組織が利用できるほど手頃な価格になるのでしょうか？

GPUの登場により、コグニティブコンピューティングを実際に活用できる時代がいよいよ現実になりつつあります。さまざまな方法で、人工知能（AI）や他の分析プロセスとストリーミング分析を統合することで、リアルタイムに人のような認知・認識を実現できる可能性があります。このような「思考のスピード」レベルの分析は、GPUの大規模な並列処理がもたらす前例のない価格とパフォーマンスがなければ、実用的ではなく、実現は不可能です。

コグニティブコンピューティングにおけるGPUの役割

リアルタイムでなければ、それは本当のコグニティブコンピューティングではありません。結局のところ、ジェパディで相手がクイズに回答する前に（回答が完全に読まれる前に）自分が回答する能力がないかぎり、ワトソンは1ポイントも獲得できず、勝利することはできないのです。今日のコグニティブコンピューティングをリアルタイムで実現する最も

費用対効果の高い方法は、GPUアクセラレーションを使用することです。

コグニティブコンピューティングアプリケーションでは、ビジネスインテリジェンス、AI、機械学習、ディープラーニング、自然言語処理、テキスト検索と分析、パターン認識など、あらゆる分析プロセスを利用する必要があります。これらのすべてのプロセスを、GPUを使用して高速化できます。実際、GPUに搭載される数千もの小型で効率的なコアは、膨大な処理能力を必要とするこれらのすべてのワークロードで見られる繰り返しの命令を並列処理するのに特に適しています。

コグニティブコンピューティングサーバーとクラスタは、必要に応じてスケールアップまたはスケールアウトでき、一秒未満から数分以内で、求められるパフォーマンスをリアルタイムに提供できます。GPU向けに最適化されたアルゴリズムとライブラリを使用することで、パフォーマンスをさらに向上できます。

GPUアクセラレーションソリューションは、Watsonスーパーコンピュータレベルのパフォーマンスを達成するためのコストやその他の障壁を打破し、コグニティブコンピューティング時代の牽引役となるでしょう。

8章
新たな一歩を踏み出す

GPUアクセラレーションは、多くのデータベースおよびデータ分析アプリケーションにおいて、CPUのみを搭載する構成と比較して、パフォーマンスと価格の両方で利点をもたらします。

パフォーマンス面では、GPUアクセラレーションにより、大規模で複雑なストリーミングデータをリアルタイムで取り込んで、分析し、視覚化することが可能になります。ベンチマークテストと実際に運用されているアプリケーションの両方で、GPUアクセラレーションソリューションは、1分間に数十億のストリーミングレコードを取り込み、数ミリ秒で複雑な演算とビジュアライゼーションを実行できることが証明されています。このような比類のないパフォーマンスレベルは、コグニティブコンピューティングなど、最も高度なアプリケーションでさえ、実用化する可能性があります。スケールアップまたはスケールアウトの両方に対応できるため、必要に応じて、段階的に予測可能な方法で、そして手ごろな価格でパフォーマンスを向上できます。

経済的な面でも、GPUアクセラレーションは同様に優れています。GPUの大規模な並列処理を、CPUのみで構成した場合のパフォーマンスと比較すると、ハードウェアコストは10分の1、電力および冷却コストは1/20となります。例えば、アメリカ陸軍情報保全コマンド（INSCOM）部隊は、1日あたりの1000億以上の記録を生成する200以上のストリーミングデータソースを処理するアプリケーションを稼働していた42台のサーバーから構成されるクラスタを、1つのGPUアクセラレーションサーバーに置き換えることができました。

しかし、同じように重要なのは、GPUのパフォーマンスと価格上のメリットは、今やすべての組織が活用できるようになっていることです。

オープンな設計が採用されているため、GPUベースのソリューションを既存の事実上すべてのデータアーキテクチャに簡単に組み込むことができ、オープンソースと商用の両方のデータ分析フレームワークと統合できます。

パブリッククラウドにおけるGPU

パブリッククラウドでGPUを利用できるため、GPUベースのソリューションをさらに手頃な価格で簡単に活用できるようになっています。Amazon Web Services、Microsoft Azure、Googleなど、主要なすべてのクラウドサービスプロバイダが、現在GPUインスタンスを提供しています。パブリッククラウドでGPUアクセラレーションを利用できるようになってきたことは、ハードウェアに投資することなく、GPUアクセラレーションを利用することを検討している組織にとって特に歓迎すべきニュースでしょう。

専用設計のGPUソリューションを使用すると、期待するパフォーマンスを得るためにこれまで使用していたテクノロジに付帯していた痛みを伴うことなく、利点のみを享受できる可能性があります。組織のデータ分析要件が今後どのように変わったとしても、リアルタイムにデータを取り込み分析できるようにするためにインデックス作成やスキーマの調整、アルゴリズムの調整/微調整、さらにはクエリーをあらかじめ決定する必要はなくなります。

もちろん、新しい取り組みを進める場合には必ずそうですが、さまざまな選択肢を調査し、分析ニーズをすべて満たし、必要に応じてスケーリングでき、GPUを最大限に活用できる専用のソリューションを選択してください。GPUアクセラレーションデータベースの能力と可能性は百聞は一見に如かずです。このテクノロジを確認できるパイロットプロジェクトからその一歩を始めてください。