Hadoopデータレイクにおけるデータの準備

Michael Lang, Teradata

10.14 EB8458 WHITE PAPER

TERADATA.

目次

Hadoop データレイクにようこそ

企業は、低コスト、スケーラビリティ、柔軟性などのさまざまな理由で、Apache™ Hadoop® に注目しています。特にスケーラビリティ、柔軟性は、企業のデータサイエンティストおよびその他のユーザーにとっての新たな可能性を提示しています。Hadoop 分散ファイルシステム (HDFS: Hadoop Distributed File System) は、あらゆる種類と形式のファイルを受け入れ、「データレイク」と呼ばれる画期的で新しい用途を満たしています。

データレイクにおいて、企業は、HDFS を使用して、以前は利用されていなかったデータの格納および処理、そして新しい方法による従来型のデータの結合を行なっています。データサイエンティストは、Hive、Pig、MapReduce などの Hadoop エコシステムのツールを使用し、メガバイト単位からペタバイト単位に至るあらゆるサイズのデータの中からパターンや傾向を探し出すことにより、データを探索し、関連性を調査します。この収集、準備、分析、レポートというプロセスが、データ・サイエンスのワークフローです。データレイクにおいては、アナリストは、ログファイルや位置情報のデータ、ソーシャルメディア・フィード、センサーデータを分析することができます。表形式に整えられたデータや、完全な非構造化テキスト、その中間にあるすべてのデータを、高速に処理することが可能です。このデータ準備フェーズは非常に反復的かつ探索的なものであり、その目的は、有意義な統計分析に適した形にデータを加工することです。準備作業のすべては、より秩序だった記述的分析、予測モデル、そして内部および外部の利用者のための可視化につながります。最終的には、データレイクおよびこのデータ・サイエンス・ワークフローが、全社規模のデータ主導型意思決定の基盤を形成します。

企業やアナリストにとっての課題は、どうすれば実際に流れ込んでくるファイルの意味を理解し、効果的にデータを管理することができるかです。データレイクを実現させる柔軟なファイル・システムが、ファイル種別や未知の起源を増殖させ、複雑に入り組んだディレクトリ群を作り出す恐れもあります。 Teradata Loom® および Teradata Loom Activescan サービスの拡張可能なレジストリは、Hadoop エコシステムにおいて他には見られないメタデータ管理機能を備えたソリューションの役割を提供します。ソース、データセット、加工、ジョブを包含する Teradata Loom のフレームワークは、このワークフローを総合的に見る視点をデータサイエンティストに与えます。

アナリストおよびデータサイエンティストのために、Teradata Loom は迅速な発見を可能にし、Teradata Loom Activescan が計算した統計情報は、さらなる分析のための確固とした出発点となります。アナリストがタスクに適したデータを入手した後、データ・サイエンス・ワークフローの残りの時間の多くはデータの準備に費やされます。アナリストが、適切な形でデータを取得する作業には、多くの場合70% から 80%、時には 90% もの時間を取られると証言しています。データの探索やアプローチの開発に加えて、ジョブに適した

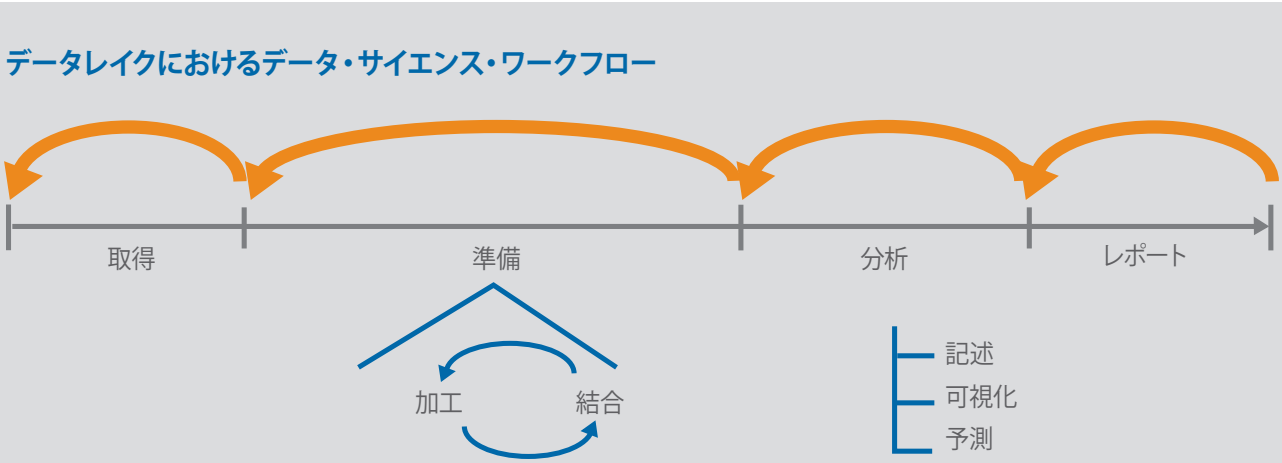

図 1.

ツールを探すだけでも時間がかかることがあります。データの準備に要する時間を短縮し、アナリストが「ビッグデータ」を使って一層効果的に作業できるようにすることは、Hadoopにとっての次のフロンティアです。そこで、Teradata Loomは新境地を開拓しているのです。

Hadoop データレイクにおけるデータの準備

データの取得とモデル化の間に行なわれる一連のアクティビティを表わす用語として定着したものはありません。私たちは、それらのアクティビティを「データの準備」と呼んでいます。データの準備は、新たに取得した「生」データを、意味ある方法での分析とモデル化が可能なクリーンなデータへと変えることを試みる作業です。データ・サイエンス・ワークフローのこのフェーズ、およびその一部には、データ・ラングリング、データ・マンジング、データ整理、データ・クレンジングなどの、さまざまな名称が付けられてきました。テラデータでは、業界用語を避けて「データの準備」という語を使用しています。

データの準備は、必ずしもビジネス・アナリストが注力してきた分野というわけではありません。従来型データウェアハウスでは、秩序があり、充分に定義されたデータがあらかじめ必要とされるため、ビジネス・アナリストが主に実行するのは、必要なデータの準備の大部分が済んだ後の記述的分析です。新たなデータを取得する作業と分析に適した形にデータを変える作業は、データ・エンジニアが担当します。この場合のデータ・エンジニアリングのスローガンは、「抽出、加工、そしてロード」です。このパラダイムにおいては、データ・エンジニアとビジネス・アナリストが、このデータ・サイエンス・ワークフローにおける責務を分け合います。このことが、データレイク環境においてはギャップを生み出します。データ・エンジニアには分析の基礎がなく、ビジネス・アナリストはデータの準備のためのツールやアプローチに精通していないためです。

データレイクの場合、スローガンは「抽出、ロード、そして加工」となり、データサイエンティストが、データ・エンジニアとビジネス・アナリストの間のギャップを埋める役割を果たします。データ・エンジニアは、データレイク内にて、多様なサイズ、起源、頻度のデータをインポートし、管理します。データサイエンティストは、データレイク内にてデータを準備し、データ・マイニングや予測モデリングなどの高度な分析を実施します。ビジネス・アナリストは、最終的な可視化を実現し、準備済みのデータからレポートを作成します。

当然ながら、これらの役割は一連の能力を表わしており、相互に排他的なものではありません。企業は、データの準備と分析の橋渡しに必要なスキルを持ったデータサイエンティストを雇うことにより、このデータ・サイエンス・ワークフローを完成させることができます。別の手段として、データレイクにおいて効果的な働きをできるような能力をデータ・エンジニアとビジネス・アナリストに与えることも可能です。いずれの場合も、これらのデータ・ワーカーには、このデータ・サイエンス・ワークフローを完遂して企業のために洞察を生み出すための、適切な種類のツールが必要となります。

データレイクにおけるデータの準備ツールが抱える主な課題は、対話的であるかということです。アナリストが単一のマシンまたはサーバー上でデータをインメモリで準備した場合、加工は多くの場合、ニア・リアルタイムで行なわれます。アナリストは、加工済みのデータセットをほぼ即時に取得し、さらなる探索および加工による反復を続けます。データレイクにおいて、一部のデータが驚異的な量であるために加工がリアルタイムで進まなくなる場合、データの準備には新たなアプローチが必要となります。そのアプローチの鍵となるのは、対話的サンプリングによるバッチ処理と、加工の反復的な見直しの必要性の最終的なバランスをとる、直感的なユーザー・インターフェースです。そのインターフェースは、データの準備用の既存のエコシステム・ツールに置き換わるものではなく、それらを補完するように設計されていなければなりません。

データ・サイエンス・ワークフローにおける役割

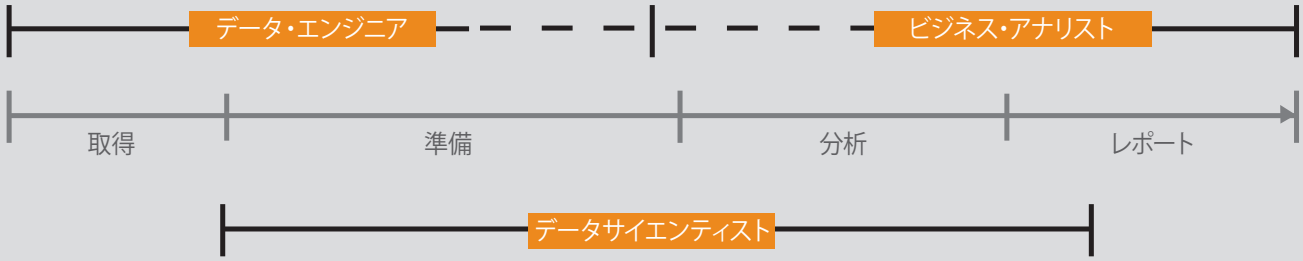

図2.

Hadoopデータセットのメタデータ管理とデータ加工の強固な基礎を確立したTeradata Loomは、Weaverという名称の機能によってデータの準備を行なう、新たなアプローチを提供するようになりました。データサイエンティストは、段階的かつ反復的にビッグデータを準備するための対話的メソッドという、データレイク向けのパワー・ツールをついに手にします。

Teradata LoomおよびWeaverによるデータの準備

データの準備には、構造化、探索、加工という3種類の能力が不可欠です。これらの能力、特に探索と加工は、反復的なものであり、重複する部分があります。Teradata Loom ActivescanおよびWeaverは、データの準備タスクの全領域をサポートすると同時に、他のIn-Hadoopツールやインメモリ・ツールを組み入れる柔軟性をユーザーに提供します。

構造化

特定のタスクに適したデータを見つけた後、データサイエンティストはそのデータを構造化しなければなりません。データの準備との関連において、データの構造化とは、通常、1つのフラット・ファイルまたはファイルの集合から表形式の構造を構築することを意味します。加工および分析用のツールでは、測定結果が行、フィールドが列に入った表形式またはマトリクス型の構造が要求される傾向がありますが、どのセルの内容も任意に複雑なものにすることができます。Teradata Loom Activescanは、データレイクにおいてデータを構造化するためのフレームワークとメカニズムを提供します。

目の前のタスクに応じて、データはアクセスしやすい形で利用できる場合とできない場合があります。多くのデータ・ソースは、区切り文字で区切られたテキストや固定幅のテキストなどのように、読み取りやすい形式になっています。表形式のデータ・ソースに加えて、XMLやJSONのように入れ子形式になっているデータが利用可能な場合もあります。また、データがバイナリ形式や独自の形式で圧縮または格納されている場合もあります。そして、データが「非構造化」テキストの中に存在している場合もあります。データレイクにおいては、これらのファイル形式すべてが共存可能です。データの構造化には、生データから特定の要素を抽出する作業が含まれることがあります。例えば、入れ子構造を平坦化することができますが、結果として生成されるテーブルの次元を単純化および削減する目的で、一部のデータを無視する必要が生じる可能性があります。

Teradata Loom Activescanは、適切な種類のデータを見つけてそのデータを適切に構造化できるようにユーザーを支援します。Activescanは、ユーザーによる設定に基づいて、指定されたディレクトリにおいて指定された間隔で新規のファイルを特定します。区切られたテキストやHiveデータベースなどの標準の形式を補足するために、Activescanは、カスタム・プラグインを適用し、データの認識、解析、フォーマットを行ないます。例えば、正規表現として知られているテキスト・パターンを使用して、ログ・ファイルを認識し、それに応じてファイルの構文解析を行ないます。結果として作成されるテーブルは、後続の加工のために整然とフォーマットされます。同様に、ユーザーは2つのハイフンで区切られた10桁の数字のシーケンスをActivescanに認識させ、結果のテーブル内に電話番号の列を作成することも可能です。Activescanでは、複雑な構造化タスク用にHive SerDesを活用することもできます。

探索

データ・ソースからテーブル(の集合)を作成した後は、データサイエンティストは構造化されたばかりのデータについて知識を深めなければなりません。その目的は、データを統計分析に適したものにするためにはどのような加工をすればよいのかを理解することです。データセットに対するアナリストの理解は、記述統計、データ・サンプル、可視化の3つを基盤とします。Teradata Loom Activescanは、ユーザーがデータの重要な側面を前もって理解できるように支援します。一方　Teradata Loom Weaverは、サンプルを表示し、分析に必要な変更を計画するための直感的なインターフェースを提供します。

新規のテーブルが作成されると、Activescanはデータの品質を示してデータの探索をガイドする記述統計を自動的に計算します。数値型の列の場合、最小、最大、平均などの統計においてサニティ・チェックが行なわれます。例えば、人の年齢に関するデータが含まれる数値列に負の値が入っていてはならないはずです。文字列の列では、特異な値またはカテゴリの数と分布は、多くの場合、関心事の統計となります。列の種類がどのようなものであっても、Activescanは値の欠落やNULL値をユーザーに通知します。

記述統計以上に、ユーザーはデータに直接的に関与することを必要とします。テーブルの行と列が多い場合は、データを効果的にナビゲートする機能が不可欠となります。まずTeradata Loomには、多様な下位設定機能を備えた柔軟なデータ・プレビュー機能が用意されています。さらに、Weaverによってユーザーは組み込みのサンプルにアクセスすることができます。サンプルは、テーブルの最初もしくは最後の行、中間のいずれかの行、または無作為に抽出した行から引き出されます。フィルタを使用すれば、ユーザーは1つまたは複数の列またはフィールドの値に基づいてデータのサブセットを精査し、最終的には加工することができます。

最後に、データおよびメタデータ内にある有意義なパターンを明らかにするためには、可視化機能が欠かせません。Weaverは、Activescanが計算した統計を基盤に、数値型列や文字列型列の値の分布を表示するヒストグラムや棒グラフなどによるシンプルな可視化を実現します。

加工

データの探索を終えたデータサイエンティストは、列とテーブルを加工する作業に進み、最終的な統計分析用としてデータの準備が整うまで反復的に加工を続けます。Teradata Loom Weaver は、変換、すなわち加工のためのパワー・ツールであり、文字列型、数値型、日付/時刻型のオブジェクトを列ベースおよび行ベースで加工するためのビルトイン関数が含まれています。さらに、Weaverによってユーザーはテーブルの構造を変換することができます。ジョインや結合などの操作を通じて複数のテーブルから新規のテーブルを作成する場合、LoomはSQL/HiveQLを活用します。Teradata Loomは、そのような加工を自動的に追跡し、加工の経路を表示します。

以下に、加工の範囲の例を示します。文字列の加工により、新しいカテゴリ変数や一貫性のあるカテゴリ変数が作成されます。例えば、電話番号が入った列の最初の3桁を分離すれば、市外局番を示す新しい列を作成できます。その他の、大文字表記、置換、空白文字削除などの文字列加工により、一貫性のないデータをきれいに整えることができます。例えば、“usa”、“U.S.A.”、“USA” という文字列を、“USA” として標準化することが可能です。数値の加工

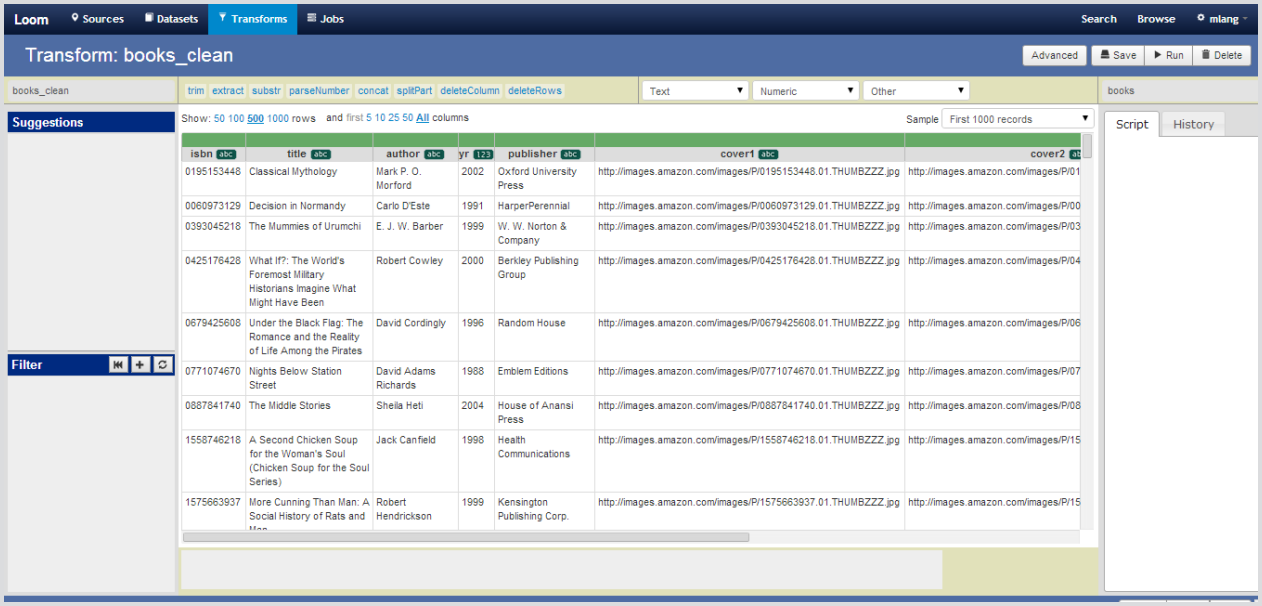

図 3.

は、数値型列の対数をとる計算などを行なう、数学関数または統計関数によって実行します。日付/時刻の操作は、文字列または数値を入力値として使用し、日付および時刻に関する固有の情報を持つオブジェクトを生成します。入力文字列には、“January 1, 2013 10:35:00”や“20130101103500”のようなものを使用できます。変換された日付/時刻オブジェクトにより、ユーザーは、元の文字列内には存在していなかった可能性のある要素(曜日など)を抽出できるようになります。テーブル・レベルの加工では、探索、クリーニング、分析を進めやすくするために、行および列のレイアウトが変更されます。列に対しては、並べ替え、削除、または名前の変更を行なうことができます。さらに集中的な操作としては、値の充填や行から列への転置などが挙げられます。

Weaverのセッションは、より規模の大きいテーブルのサンプルから開始されます。ユーザーは、必要な変更内容がすべてサンプルに反映されるまで、加工結果をサンプルに対して反復的に適用します。Weaverは、例えば文字列型の列を数値型または日時型の列に変えるために数値以外の値や適合しない値を排除することなど、次の加工に向けた提案を示すことによってユーザーを支援します。実行された加工はすべて、Weaverセッションの履歴に記録されます。加工の結果にユーザーが満足した時点で、Weaverは、HDFSにおいてMapReduceジョブを開始することにより、テーブル全体に対して同じ加工を実行します。ユーザーは、Activescanの統計およびWeaverの更新済みサンプルを参照して新規のテーブルを確認し、必要に応じてさらなる加工を続けます。この反復プロセスに関連するメタデータはTeradata Loomのレジストリ内に完全に捕捉され、データの加工は経路グラフに反映されます。

データレイクの能力を最大限に引き出すために、データサイエンティストはテーブルの結合も行なう必要があります。ジョインや結合を生成および実行するために、Teradata LoomにはSQL/HiveQLへの直接的なインターフェースが用意されています。これらのクエリー言語は、関係型の加工のために馴染みの深い抽出をMapReduce上で実行する機能を提供します。ユーザーは、説明、キーワード、他のメタデータなどを、必要に応じて加工結果に追加できます。Weaverでの加工と同様に、入力と出力はTeradata Loomの経路グラフにおいて自動的に追跡されます。

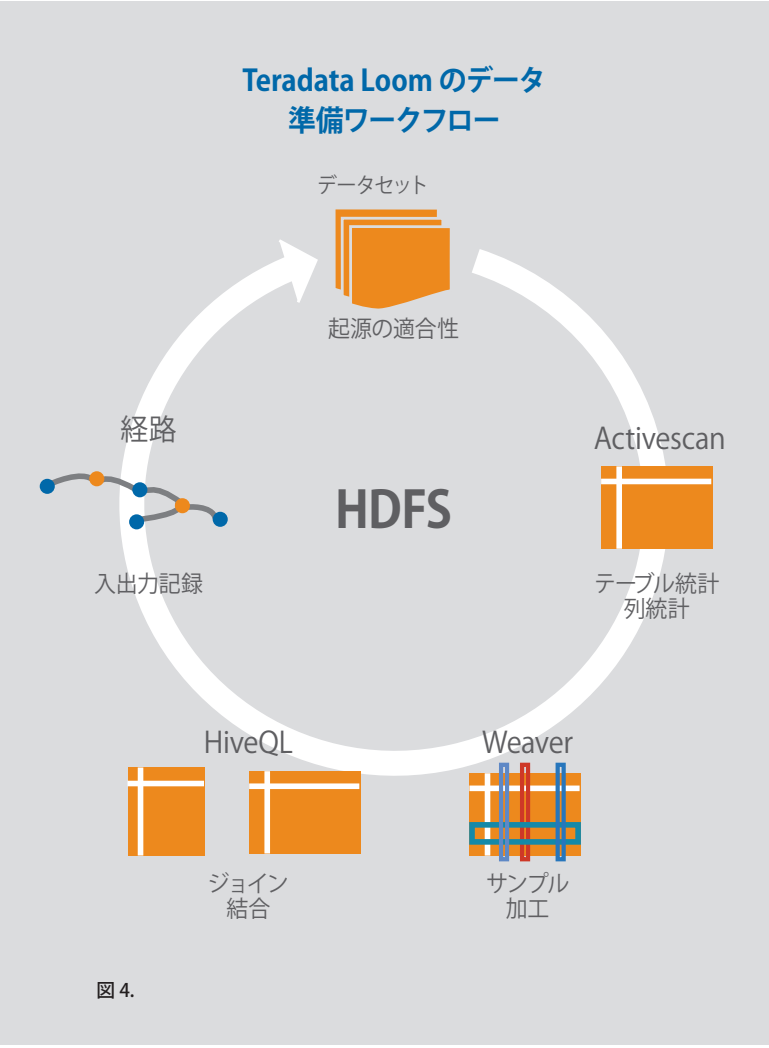

図 4.

結論

Teradata Loomは、Hadoop向けの完全なデータ管理を実現する初のソリューションを提供します。データ・エンジニア、ビジネス・アナリスト、そしてデータサイエンティストは、データレイクにおける効果的かつ効率的な作業に適したツールを手にします。Teradata Loomにより、データ・ワーカーは、データの発見、構造化、探索、加工をより迅速に行なえるようになると同時に、起源、経路、他のメタデータの明確な記録を保持できるようになります。結果として、企業は、連続的なデータ・サイエンス・ワークフローから、より良好で迅速な洞察を得られるという利点を享受します。Hadoopがこれ程までにエンタープライズ用途で利用できる状態になったことは、いまだかつてありません。

詳細情報

Hadoop および Teradata Loom でのデータおよびメタデータ管理の詳細や、Hadoop からさらに価値を引き出すためにテラデータがどのように支援できるかの詳細情報については、テラデータの担当者にお問い合わせください。

10000 Innovation Drive, Dayton, OH 45342 **Teradata.com**

10.14 EB8458 TDMK-5048(1503)