IBM DataWorks

クラウド上でシンプルかつ強力にデータの準備と移動を行う統合ソリューション

特長

- ユーザーが技術者である場合もそうでない場合も、迅速かつ簡単にデータから価値を創出できます。
- クラウド上でシンプルにデータの準備および移動を行うサービスにより、データの品質を向上できます。
- 先進的なクラウド・データ・サービスと連携することで、シームレスなデータ・マネジメント・プラットフォームを実現できます。

クラウド・コンピューティング、ビッグデータ、Internet of Things の時代に入り、企業においては情報の過多が発生しています。昨今においては、現状の時間やリソースではビジネス・インテリジェンスやデータ・サイエンスのチームが分析できないほどのデータが生成され、収集されています。実際に、Forrester の調査によると、IT 部門が対応している単純な BI リクエストのうち、68%のリクエストは完了までに数週間から数カ月以上もかかっています[1]。

このような時代におけるデータ・ニーズに対応し、競争力を維持するために、企業は、ビジネス部門の担当者をデータ処理能力の高いデータワーカーへと変貌させ、IT 部門の負担を減らす必要があります。しかし、このことは「データベース管理者やデータ・サイエンティスト並みの高い技術スキルをもたないビジネス・ユーザーでも、数多くのデータ・ソース (オンプレミスとクラウドのソースを含む) から迅速にデータを取り出し、加工し、分析できる必要がある」という困難な課題があることを示しています。

IBM® DataWorks のような最新のクラウド・サービスを活用することで、ローカルに保存した複数の Excel シートであれ、クラウド上にホストした大規模なデータベースであれ、ポイント・アンド・クリックによるアクセスだけで、技術者であってもそうでなくてもデータから有益な洞察を導き出すことができるようになります。

IBM DataWorks について

フル・マネージド形式でデータの準備と移動を行うサービスである DataWorks を導入すると、アナリスト、開発者、データ・サイエンティスト、およびデータ・エンジニアは、シンプルかつ強力なクラウド・ベースのインターフェースを通じてデータを活用することができます。IBM クラウド・データ・サービス・ポートフォリオの主要なコンポーネントである DataWorks を通じて、ビジネス・アナリストや Excel のパワー・ユーザーはデータの発見・標準化・加工・移動を行えるようになり、アプリケーション開発とアナリティクスの実行が可能になります。

DataWorks はクラウド・データウェアハウスの IBM dashDB™ をはじめ、NoSQL データベースの IBM Cloudant®や IBM Watson™ Analytics などのクラウド・データ・サービスと統合されています。これにより、DataWorks を活用することで、「オンプレミスおよびオフプレミスからのデータの取得」、「データの移動・加工」、「クラウド上の分析エコシステムで迅速な分析と可視化」をシームレスに行えるようになります。さらに、DataWorks ではお客様に継続的にサービスを提供しており、定期的に新規の機能や特性が追加されています。

本ソリューションの処理エンジンは Apache Spark™ に基づいています。Spark には、先進的なオープン・ソースのアナリティクス・プロジェクトである大規模な開発コミュニティーが存在しており、その規模は恒常的に成長を続けています。そのため、ビッグデータおよびクラウド・コンピューティングにおけるスピーディーなイノベーションと歩調を合わせた業界トップレベルのソリューションを実現できます。

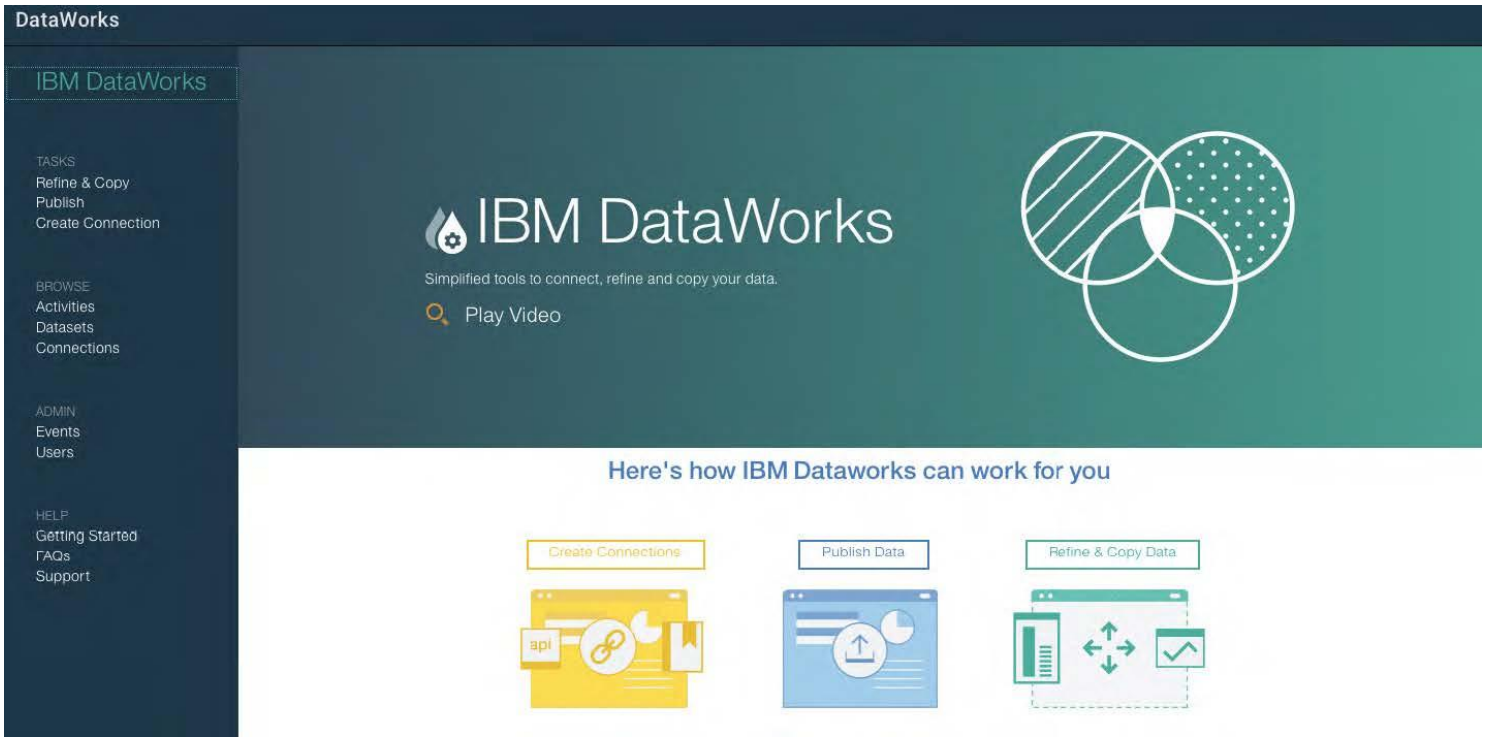

*図 1:*IBM DataWorks: ポイント・アンド・クリックに基づいてデータの準備を行うことができる、クラウド上のフル・マネージド・サービス

社内のすべてのユーザーにデータ・アクセスを提供する

誰もがデータ・サイエンティストになれる時代はまだ到来していないものの、DataWorks のようなツールが、すべてのユーザーに対してデータ・アクセスと先進アナリティクスを提供する方法を実現しつつあります。企業が DataWorks を活用する方法には様々なものがありますが、主なユース・ケースとしては以下が挙げられます。

1. 複数のソースのデータを統合する: サポート対象のデータ・ソースであれば、どこからでもデータを取得し、それらのデータを任意に組み合わせることで、分析業務に必要なファイルやテーブルを作成できます。
   - 例: デジタル・ライツ関連の企業に勤めるデータ・サイエンティストが、顧客のメディア・アセットのポートフォリオに加え Nielsen、Rovi、Twitter、Rotten Tomatoes、EIDR などのサード・パーティーから得られるデータに基づいた広告配信アルゴリズムを開発したいと考えています。メディア・アセットのウェアハウスとして dashDB を使用し、さまざまな構造を持つコンテンツを保存するためには Cloudant を使用します。そして、これらのデータの統合や加工、クレンジングのために DataWorks を使用することで、すぐにレポーティングの準備が完了します。
2. ハイブリッド・クラウド環境でデータにアクセスする: 業界で汎用的なデータ・ソースへの接続法でデータのある場所にアクセスすることで、簡単かつセキュアにファイアウォールの内側にあるデータにもアクセスできます。
   - 例: あるユーザーが「クラウド上に存在している顧客センチメントのデータ」と「オンプレミスのデータベースにあるマーケティング・キャンペーンのデータ」の両方にアクセスして、マーケティング・キャンペーンの効果を評価する必要があるとします。DataWorks を活用すれば、ファイアウォールの内側に存在するデータを抽出するためのセキュアなトンネルを設定できます。
3. 分析のために生データを加工する: ソース・データの値と列をフィルタリングし、ソートし、重複データを削除することで、標準化されたスコアを通じてデータの品質を把握できます。
   - 例: ビジネス・アナリストが、昨年度の売上履歴のデータに基づいて売上予測を作成する必要があるとします。オンプレミスの売上データベースにはアクセスできるものの、レポートの作成前のデータの品質と関連性について確証が持てません。しかし、DataWorks を活用することで、ユーザーはデータ品質のスコアやデータのプレビュー機能を通じて、適切なデータが存在することをビジュアルに確認できます。また、DataWorks は不要な値を除外する機能も提供します。
4. 分析のためにデータをロードする: 必要なデータがどのような場所にあってもアクセスし、クラウド上のデータ・サービスにロードすることができます。
   - 例: データ・サイエンティストが、オンプレミスのデータウェアハウスから dashDB のクラウド・インスタンス上にいくつかのファイルをロードして、顧客維持プロジェクト用に統計モデルを構築する必要があるとします。DataWorks では、簡単なポイント・アンド・クリックによるアクセスが可能であるため、データ・ソースを指定して必要となるテーブルとファイルを選択するだけでロードを行えます。
5. Web アプリからデータのワークフローを制御する: DataWorks の API を使用すると、アプリケーションからワークフローのアクティビティーを作成し、制御できます。
   - 例: アプリケーション開発者は、DataWorks の API を用いることで、ビジネス・アナリスト、データ・サイエンティスト、IT 管理者らが作成するアクティビティーに基づいて、Internet of Things のセンサーや、モバイルやソーシャルなどの Systems of Engagement (協働のための情報活用システム) から発生するイベントをトリガーとしたデータの移動・クレンジング・加工が実行されるアプリケーションを開発できます。
6. リレーショナル・データや構造化データを半構造化データにマッピングする: 正規化されているテーブル・データを Cloudant の NoSQL ストアにロードできます。
   - 例: 開発者がリレーショナル・データを Cloudant にロードすることで Web アプリケーション上で使用可能とし、さらにその正規化されたデータを階層型の JSON ドキュメント構造に変換する必要があるとします。DataWorks では、リレーショナルなデータ・ソースから、移動先の NoSQL ベースの Cloudant にシームレスに接続し、リレーショナル・データを JSON ドキュメントに変換できます。

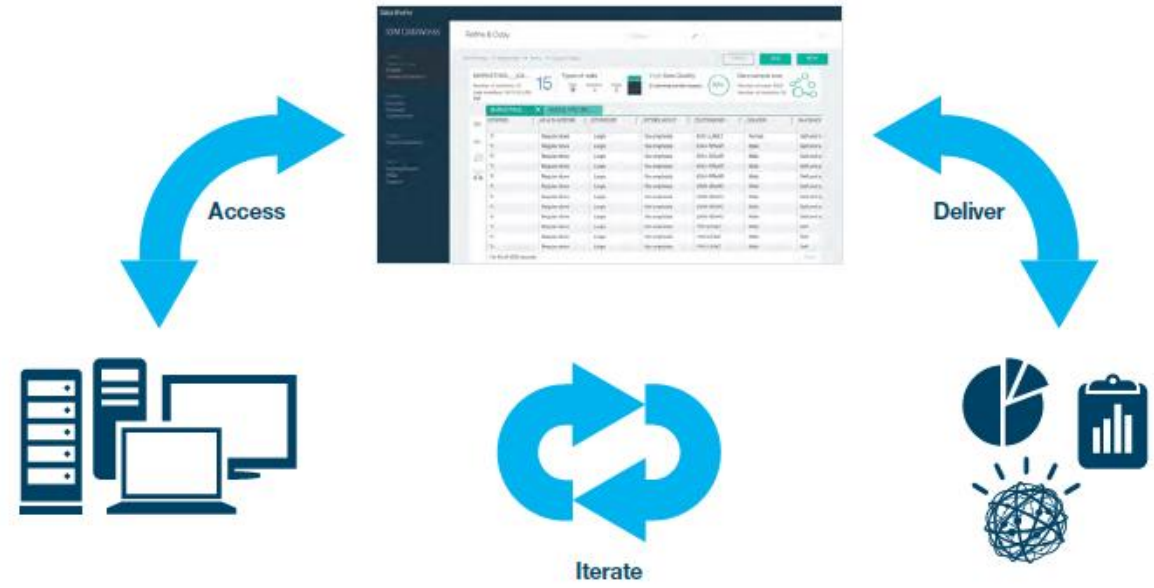

図 2: データにアクセスし、データを加工し、アナリティクスのクラウド・サービスにデータを提供した後、このプロセスを何度も繰り返すことができる。

ハイブリッド・クラウドの複雑な環境でシンプルなデータ・アクセスを実現する

現代の企業が抱えるデータ・アクセスとデータの移動に関する課題は、ハイブリッド型の IT 環境が複雑化することで生まれています。ハイブリッドという言葉の定義は人により異なり、「オンプレミスのインフラ」と「クラウド・サービス」の間でシームレスかつ完全なデータ同期を行うことを意味する場合もあれば、データのロケーションにかかわらずデータ・アクセスをサポートすることを指す場合もあります。このようにハイブリッドの定義は様々ありますが、ハイブリッド型の実行環境で共通して発生する明確なビジネス上の課題として、迅速かつセキュアなデータの移動とアクセスが挙げられます。

IBM DataWorks は、ハイブリッド環境で迅速かつセキュアにデータのアクセスと移動を管理するためのツールを提供します。DataWorks には、ハイブリッド環境を実現する以下の主要な 2 つのフィーチャーがあります。

1. セキュアなゲートウェイにより、お客様がクラウドからエンタープライズ・データへのアクセスを可能とするシンプルなソリューションを提供します。これにより、容易にインストール可能な SSL トンネルで、ファイアウォールの内側のデータにユーザーがアクセスできるようになります。このセキュアなゲートウェイは、汎用的な VPN アクセスに比べてはるかに単純であり、ユーザーはアウトバウンド・ポートを開き、オンプレミスでエージェントをインストールするだけで済みます。
2. データ準備に要する工程を分析し、できるだけ多くの処理をソース・データベース側にプッシュすることで、移行が必要なデータのサイズを削減します。これにより、データ・ソースのコンピューティング能力を活用してワークロードを分散し、クラウドに移動するデータ・セットのサイズを縮小することで、ターゲットで必要となるデータのみを転送できます。

ハイブリッド環境におけるこのようなプロセス全体を通じて、DataWorks はオンプレミス環境とクラウド環境の両方でセキュアな統合ポイントを提供することで、高レベルのセキュリティーを実現します。

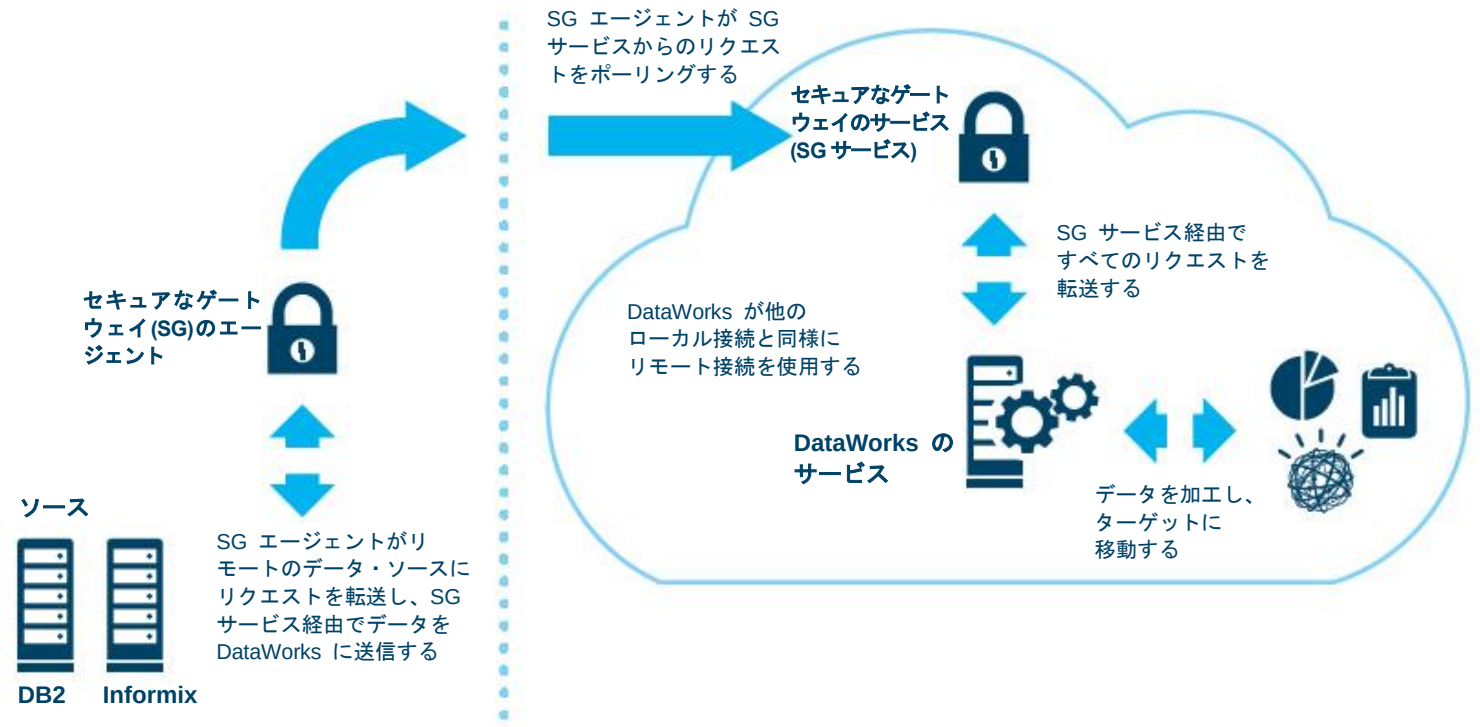

図 3: DataWorks のセキュアなゲートウェイを活用することで、ファイアウォールの内側のデータにアクセスする

データ品質のジレンマ: データ活用の前に

現在、不完全や不正確、無関係なデータが存在することで生じるデータ品質の問題が原因となり、非常に多くのアナリティクス・プロジェクトが停滞し、遅延し、未完成のままになっています。実際、Forrester によると、42% ものビジネス部門の専門家が、データ分析を行う前のデータの修正および検証作業に業務時間の 40% 以上を費やしています[2]。このようなデータ品質に関わる問題には、データの準備に関する新しいアプローチで対応することができます。

データの準備とは、広範囲な接続により、あらゆるロケーションに存在するデータにセルフサービス形式でセキュアにアクセスすることを指します。これは、ビジネス上の洞察を得るためのデータ分析を行う前に、データの品質と完全性を最適化するという従来の ETL (データの抽出・加工・ロード) のコンセプトに基づいたものです。文字列や整数といった内容は技術スキルを持つユーザーが専門とする領域であり、ビジネス・ユーザーはできるだけシンプルな方法でデータの意味を理解したいと考えています。従来のプロセスでは、ビジネス・ユーザーは IT 部門がデータ・セットを検証してくれるのを待つ必要がありましたが、現代の新たなデータ分析のニーズとして、このデータ検証の障壁を下げることが求められています。現代においては、ビジネス・ユーザーは自分でデータの準備を行う能力を持つ必要があるのです。

この課題を解決するために、DataWorks はクラウド上で容易にデータの準備と移動を行えるサービスを提供します。このサービスには、技術者であってもそうでなくても容易にアクセスできます。データ準備に関する技術は非常に複雑であり、専門のスキルなしでデータ・サイエンティストのように効果を発揮するのは非常に困難です。一方で、DataWorks の開発の根底には、データ・サイエンティストではないユーザーでもこの製品を活用することで先進的なデータ洞察を得ることを可能にする、という考え方があります。Excel のパワー・ユーザーであっても、データ・セットやデータ構造に関する深い知識を必要とせずに、データ・セットを制御でき、よりスピーディーかつ高精度にレポートを作成できるのです。

DataWorks は、ビジネス・アナリストや Excel のパワー・ユーザーにとって非常にアクセスしやすいスプレッドシート形式のインターフェースを提供しており、そこでデータのクレンジングや加工、データ操作を通じたデータの簡単な可視化ができます。ユーザーはインタラクティブなガイドに従うことで、迅速にアクティビティーを構築することが可能であり、サイズの小さなスプレッドシートから数テラバイトのデータベースまで、あらゆるサイズのデータ・セットに対してそのアクティビティーを実行できます。このオンデマンドでデータの作成・移動・加工を一括処理できる機能を活用することで、技術スキルがそれほどないビジネス・ユーザーでも、IT 管理者やデータベース管理者からのサポートを待つことなく、先進アナリティクスのプロジェクトを進行できます。データ管理者が事前にガバナンス・ポリシーを設定し、接続を確立しておけば、あらゆるビジネス・ユーザーがセルフサービスでデータ準備と加工を行えるようになり、活用されていない未加工のデータも活用できるようになります。

Apache Spark: パフォーマンスの向上を実現するための秘策

DataWorks は様々なデータ・ソース (dashDB、Salesforce.com、IBM DB2®、Cloudera Impala、Apache Hive、Sybase など) に対するコネクターを提供しており、データ移行における強力なソリューションとなります。しかし、これほど多くのデータ・ソースと連携しながらパフォーマンスとスケールアップの機能を維持するためには、強力なツールが必要です。そのため、DataWorks は、ビッグデータ処理を行うオープン・ソース・エンジンの中でトップレベルである Apache Spark を活用します。

Spark は使いやすく効率性の高いツールでありながら、無償で利用できます。本ツールには、大規模かつ成長中であるオープン・ソースの開発コミュニティーが存在し、そのコミュニティーの成長に伴い Spark は日々進化を続け、より多くのデータ処理機能と機械学習機能が追加されています。Spark は、クラスターによるコンピューティング・モデルを採用する Apache Hadoop のデータ処理モデルをさらに強化しています。また、使いやすいプログラミング・インターフェースによって、最近の Web やモバイル・アプリでよく見られるストリーミング・データや継続的なクエリーのワークロードに最適なツールとなります。Spark は、そのパフォーマンスや、柔軟性、使いやすさにより、大量のデータ・セットから迅速に答えを得るための最適なテクノロジーといえます。

DataWorks の内部では、Spark エンジンがリアルタイムでデータを迅速かつ大規模に操作するためのバックエンドの処理を実施します。ユーザーはログインし、接続を設定し、セキュアなゲートウェイを指定するだけで、オンプレミスやクラウド上のデータにアクセスできます。ユーザーからは見えないバックエンドで 、DataWorks が Spark クラスターに接続することで、ソースからデータを迅速にロードし、ソート、再配置、列の操作などの処理を実施します。そして、DataWorks は、この Spark が実施するプロセスを任意のスケジュールで繰り返し実施できるアクティビティーとして保存します。 このようにして、ユーザーは、手作業でのデータ操作に気をとらわれることなく、データに基づく新たな洞察からビジネス上の成果をより迅速に実現することに注力できるのです。DataWorks と Spark を活用することで、Excel レベルのユーザーであっても、シンプルかつ簡単に、そしてセキュアに大量のオンプレミスのデータとクラウド・ベースのデータを管理できるようになります。

IBM Watson Analytics: DataWorks の活用事例

DataWorks のデータ準備機能とクラウド・サービスとの連携機能の活用事例として、業界トップレベルのデータの分析と可視化のツールである IBM Watson Analytics 内部での活用が挙げられます。DataWorks は Watson Analytics に組み込まれており、分析とレポーティングの前にデータの品質を改善したいと考えるビジネス・アナリストに対して、単一の統合エクスペリエンスを提供することができます。Watson Analytics は DataWorks を取り込むことで、以下のような新しい機能を実現しています。

- 様々なエンタープライズ・データ・ソースにアクセスする: オンプレミスとクラウドを問わず、多くのデータ・ソース (Amazon Redshift、Apache Hive、Cloudera Impala、IBM DB2、IBM Informix®、IBM Netezza®、IBM SQL Database、IBM dashDB、Microsoft Azure、Microsoft SQL Server、MySQL、Oracle、Pivotal Greenplum、PostgresSQL、Salesforce.com、Sybase、Sybase IQ など) にアクセスすることで、Watson Analytics を通じてより深い分析と BI レポーティングを実行できるようになりました。
- ロード前にデータを加工する: ユーザーはデータ・ソースからデータをそのまま Watson Analytics にロードするだけでなく、ロードの前にデータを加工することもできるようになりました。このデータ加工の機能により、ユーザーはデータ品質の評価や、プレビュー表示、列の値に基づくフィルタリング、不要な列の削除、複数ソースのデータの統合ができます。
- ファイアウォールの内側のデータにセキュアにアクセスする: ユーザーは DataWorks のセキュアなゲートウェイを介して、ファイアウォールの内側でのみ利用可能なデータにアクセスできます。これにより、管理者は制御されたアクセス環境で SSH トンネルを確立し、オンプレミスのデータ・ソースやその他の安全性の高いデータ・ソースに接続することが可能になります。

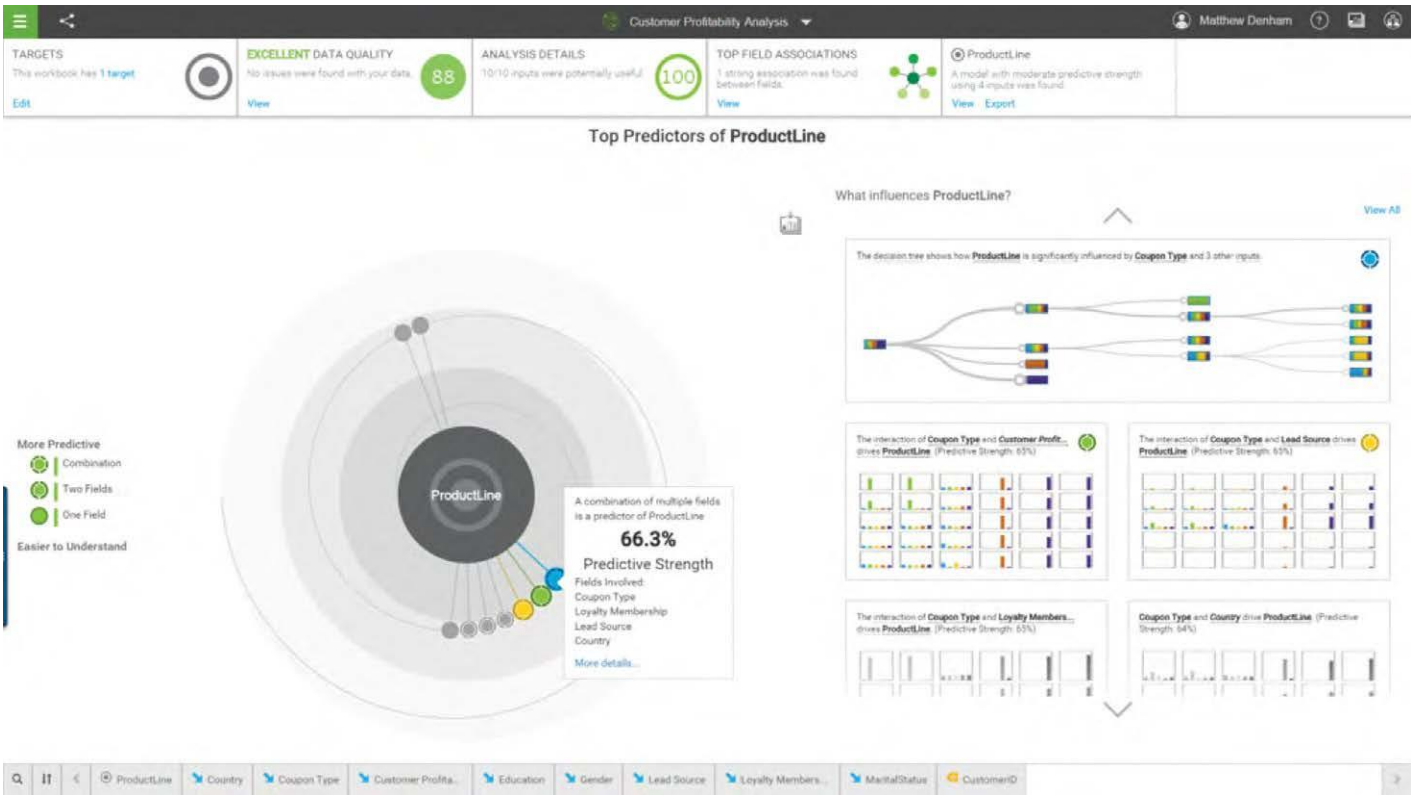

図 4: Watson Analytics を通じて、顧客の行動の説明変数を把握する

ソリューションの使用を開始する: データをビジネスに活用する

DataWorks は無償ですぐに使用を開始できます。具体的には、Bluemix.net にアクセスし、アカウントを作成するだけ使用を開始できます。Bluemix は IBM による Platform-as-a-Service (PaaS) オファリングであり、DataWorks と連携する広範な種類のクラウド・データ・サービス (クラウド・データウェアハウスの dashDB や NoSQL データベース・サービスの Cloudant を含む) を使用するための入り口としての役割を果たします。データの行数が 1,000 行未満の場合は DataWorks は無償で利用できるため、金銭的なリスクなしにデータのロードと加工を開始し、後からシステムを拡張していくということもできます。より大量のデータ・セットを使用する場合も、DataWorks は完全な従量課金モデルで提供されるため、お客様が使用しないインフラに対してまで料金を支払うということはありません。

詳細情報を確認し、本ソリューションを使用する際には、ibm.biz/DataWorksJP にアクセスしてください。

IBM クラウド・データ・サービスについて

IBM クラウド・データ・サービスは、開発者とデータ担当者にコンテンツ、データ、アナリティクスをサポートする豊富な機能を包括的に統合データ・サービスとして提供します。クラウド・データ・サービスのオファリングは、製品の開発スピードを迅速化し、アップタイムを改善することで、Web アプリケーションとモバイル・アプリケーションの開発者により大きな価値を提供することができます。IBM クラウド・データ・サービスが、開発者のために、どのように画期的なサービスを実現し、どのようにそのサービスを提供するのかをさらに詳しく確認する際は、Twitter で IBM をフォローし (アカウント: @IBMdashDB および @IBMcloudant)、ibm.com/analytics/jp/ja/technology/cloud-data-services にアクセスしてください。

日本アイ・ビー・エム株式会社
IBM Cloud

〒103-8510
東京都中央区日本橋箱崎町 19 番 21 号

Produced in Japan
2015 年 12 月

1 *Accelerate BI Initiatives With Self-Service Data Discovery And Integration* – Forrester, June 2015.

2 *Data Preparation Tools Accelerate Analytics* – Forrester, February 2015. (https://www.forrester.com/Brief+Data+Preparation+Tools+Accelerate+Analytics/fulltext/-/E-res119975)

Please Recycle

CDW12352-JPJA-00