IBM SPSS Analytic Server
バージョン 1

ユーザーズ・ガイド

IBM

お願い

本書および本書で紹介する製品をご使用になる前に、 19 ページの『特記事項』に記載されている情報をお読みください。

製品情報

本書は、IBM SPSS Analytic Server バージョン 1、リリース 0、モディフィケーション 0、および新しい版で明記されていない限り、以降のすべてのリリースおよびモディフィケーションに適用されます。

お客様の環境によっては、資料中の円記号がバックスラッシュと表示されたり、バックスラッシュが円記号と表示されたりする場合があります。

原典： IBM SPSS Analytic Server
Version 1
User's Guide

発行： 日本アイ・ビー・エム株式会社

担当： トランスレーション・サービス・センター

目次

第 1 章 概要

IBM® SPSS® Analytic Server はビッグ・データ分析のためのソリューションであり、IBM SPSS テクノロジーをビッグ・データ・システムと結合して、使い慣れた IBM SPSS のユーザー・インターフェースでの作業によって、以前は実現できなかったスケールで問題を解決することができます。

ビッグ・データ分析が重要な理由

組織によって収集されるデータの量は急激に増大しています。例えば、金融業界や小売業界では 1 年間 (または 2 年間、さらには 10 年間) のすべての顧客の取引を蓄積しており、電気通信事業者は Call Data Record (CDR) やデバイス・センサーの読み取りを蓄積しており、インターネット企業は Web クロールの結果を蓄積しています。

ビッグ・データ分析は、以下の場合に必要になります。

- 大量のデータ (テラバイト、ペタバイト、エクサバイトのレベル) が存在する。特に、構造化されたデータと構造化されていないデータが混在する場合
- データの変更や累積が急速である

ビッグ・データ分析は、以下の場合にも効果的です。

- 多数 (数千) のモデルを作成している
- モデルを頻繁に作成/更新する

課題

大量のデータを収集するこのような組織では、多くの場合、以下に示すさまざまな理由から実際のデータの活用に困難が生じています。

- 従来の分析製品のアーキテクチャーは分散計算に適合しません。さらに、
- 既存の統計アルゴリズムは、ビッグ・データを処理するように設計されていません (それらのアルゴリズムではデータをローカル環境に移動する必要がありますが、ビッグ・データの移動はコストが大きすぎます)。そのため、
- 最先端の分析をビッグ・データに適用するには、ビッグ・データ・システムに関する新しいスキルと詳細な知識が必要です。これらのスキルを持ち合わせたアナリストはほとんどいません。
- メモリー内で動作するソリューションは、中規模の問題であれば適用できますが、真のビッグ・データまでスケーリングすることはできません。

ソリューション

Analytic Server は以下の機能を提供します。

- ビッグ・データ・システムを活用するデータ中心のアーキテクチャー (HDFS 内のデータでの Hadoop Map/Reduce など)。
- データをリモート環境に置いたまま使用するように設計された新しい統計アルゴリズムを取り込んだ、定義済みのインターフェース。
- 使い慣れた IBM SPSS のユーザー・インターフェースを採用しているため、ビッグ・データ環境の詳細を知る必要がなく、アナリストがデータの分析に集中できます。
- あらゆる規模の問題に対応できる、スケーラブルなソリューション。

アーキテクチャー

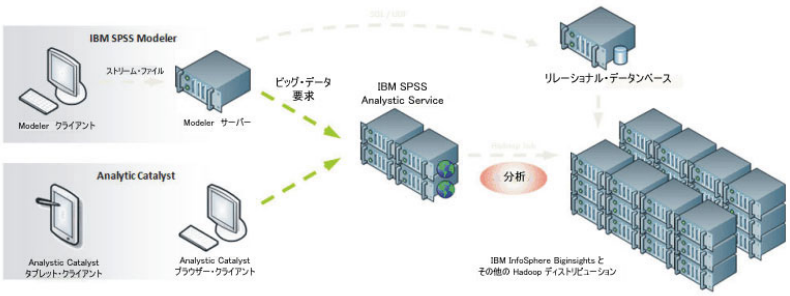

図 1. アーキテクチャー

Analytic Server は、クライアント・アプリケーションと Hadoop クラウドの間に位置します。データをクラウドに格納している場合、Analytic Server での処理の概略は以下のとおりです。

1. クラウド内のデータを対象に Analytic Server のデータ・ソースを定義します。
2. クライアント・アプリケーションで実行する分析を定義します。現行リリースでは、クライアント・アプリケーションは IBM SPSS Modeler および IBM SPSS Analytic Catalyst です。
3. 分析を実行すると、クライアント・アプリケーションが Analytic Server 実行要求を送信します。
4. Analytic Server が、ジョブを調整して Hadoop クラウドで実行し、結果をクライアント・アプリケーションに報告します。
5. その結果を使用して以後の分析を定義し、上記のサイクルを繰り返します。

第 2 章 SPSS Modeler の統合

SPSS Modeler は、分析に対する視覚的なアプローチを備えたデータ・マイニング・ワークベンチです。ジョブに含まれる個別のアクションは、データ・ソースへのアクセスから、レコードの結合、新規ファイルの記述、またはモデルの作成に至るまで、すべてキャンバス上のノードとして表されます。これらのアクションをリンクすることで、分析ストリームを形成します。

HDFS 内で実行できる SPSS Modeler ストリームを作成するには、『Analytic Server ソース』・ノードから開始します。SPSS Modeler は、できる限り多くのストリームを Analytic Server にプッシュし、その後、必要に応じて一部のレコードを HDFS からプルして、SPSS Modeler サーバー内でストリームの実行を「ローカルで」終了します。SPSS Modeler がダウンロードするレコードの最大数は、 4 ページの『Analytic Server ストリームのプロパティー』で設定できます。

HDFS へのレコードの書き込みで分析を終了する場合は、 4 ページの『Analytic Server エクスポート』・ノードでストリームを終了します。

Analytic Server ソース

Analytic Server ソースを使用すると、Hadoop Distributed File System (HDFS) 上でストリームを実行できます。Analytic Server データ・ソースの情報は、以下のようなさまざまな場所から取得されます。

- HDFS 上のテキスト・ファイル
- データベース
- HCatalog

通常、Analytic Server ソースを使用するストリームは、HDFS 上で実行されます。しかし、HDFS での実行がサポートされていないノードがストリームに含まれている場合は、できる限り多くのストリームが Analytic Server に「プッシュ・バック」され、その後、SPSS Modeler Server が、残りのストリームの処理を試みます。データ・セットの量が特に多い場合は、ストリーム内にサンプル・ノードを配置するなどして、データ・セットのサブサンプルを作成する必要があります。

「**データ・ソース**」。SPSS Modeler Server 管理者によって接続が確立されている場合は、使用するデータを含むデータ・ソースを選択します。データ・ソースには、そのソースに関連付けられたファイルおよびメタデータが含まれています。「**選択**」をクリックすると、選択可能なデータ・ソースのリストが表示されます。詳しくは、『Analytic Server の「データ・ソースの選択」』のトピックを参照してください。

新規データ・ソースを作成する必要がある場合や、既存のデータ・ソースを編集する必要がある場合は、「**データ・ソース・エディターの起動 (Launch Data Source Editor)...**」をクリックします。データ・ソースの作成および編集について詳しくは、 9 ページの『データ・ソース』を参照してください。

Analytic Server の「データ・ソースの選択」

「データ・ソース」表に、使用可能なデータ・ソースのリストが表示されます。使用するソースを選択して「**OK**」をクリックしてください。

データ・ソースの所有者を表示するには、「**所有者の表示 (Show Owner)**」をクリックします。

「**フィルター基準 (Filter by)**」を使用すると、リストに表示されるデータ・ソースを、「**キーワード**」を基準にフィルタリングできます。この機能では、データ・ソース名とデータ・ソースの説明、または「**所有者**」を対象に、フィルター基準が確認されます。フィルター基準として、ストリング、数字、またはワイルドカード文字 (%) の組み合わせを入力できます。検索文字列では、大/小文字が区別されます。「**更新 (Refresh)**」をクリックすると、「データ・ソース」表が更新されます。

Analytic Server エクスポート

Analytic Server エクスポートを使用すると、分析によって得られたデータを、既存の Analytic Server データ・ソースに書き込むことができます。例えば、Hadoop Distributed File System (HDFS) 上のテキスト・ファイルや、データベースに書き込むことができます。

Analytic Server エクスポート・ノードを使用するストリームも、通常は Analytic Server ソース・ノードから開始され、Analytic Server に送信されて、HDFS 上で実行されます。一方、「ローカル」のデータ・ソースを使用するストリームは、Analytic Server で使用する比較的小規模なデータ・セット (100,000 レコード以下) をアップロードするために、Analytic Server エクスポート・ノードで終了することができます。

「**データ・ソース**」。使用するデータを含むデータ・ソースを選択します。データ・ソースには、そのソースに関連付けられたファイルおよびメタデータが含まれています。「**選択**」をクリックすると、選択可能なデータ・ソースのリストが表示されます。詳しくは、3 ページの『Analytic Server の「データ・ソースの選択」』のトピックを参照してください。

新規データ・ソースを作成する必要がある場合や、既存のデータ・ソースを編集する必要がある場合は、「**データ・ソース・エディターの起動 (Launch Data Source Editor)...**」をクリックします。データ・ソースの作成および編集について詳しくは、9 ページの『データ・ソース』を参照してください。

「**モード (Mode)**」。 既存のデータ・ソースに追加する場合は「**追加 (Append)**」、データ・ソースの内容を置き換える場合は「**上書き**」を選択します。

「**このデータのインポート・ノードを生成 (Generate an Import node for this data)**」。選択すると、指定したデータ・ソースにエクスポートするデータのソース・ノードが生成されます。このノードは、ストリーム・キャンバスに追加されます。

Analytic Server ストリームのプロパティー

以下の設定により、Analytic Server を操作するためのいくつかのオプションを指定できます。

Analytic Server 外で処理するレコードの最大数 (Maximum number of records to process outside of Analytic Server)

Analytic Server データ・ソースから SPSS Modeler サーバーにインポートするレコードの最大数を指定します。

Analytic Server 内でノードを処理できない場合の通知 (Notification when a node can't be processed in Analytic Server)

この設定は、Analytic Server に送信されるストリームに、Analytic Server では処理できないノードが含まれていた場合の動作を決定します。警告を出してストリームの処理を続行するか、エラーをスローして処理を中止するかを指定します。

分割モデルのストレージ設定 (Split Model Storage Settings)

モデル・サイズ (MB) の超過時に Analytic Server への参照によって分割モデルを保管する (Store split models by reference on Analytic Server when model size (MB) exceeds)

モデル・ナゲットは、通常、ストリームの一部として保管されます。多数の分割を伴う分割モデルでは、大量のナゲットが発生することがあり、それらのナゲットをストリームと Analytic Server 間でやり取りすることにより、パフォーマンスに影響が及ぶ場合があります。その解決策として、分割モデルが指定されたサイズを超えた場合に、その分割モデルを Analytic Server に保管し、SPSS Modeler 内のナゲットには、そのモデルへの参照を格納します。

実行の完了時に Analytic Server への参照によるモデルを格納するデフォルト・フォルダー (Default folder to store models by reference on Analytic Server once execution is complete)

Analytic Server 上の分割モデルを保管する場所のデフォルト・パスを指定します。このパスの先頭は、有効な Analytic Server プロジェクト名にする必要があります。

プロモートされたモデルを格納するフォルダー (Folder to store promoted models)

「プロモートされた」モデルを保管する場所のデフォルト・パスを指定します。プロモートされたモデルとは、SPSS Modeler セッションの終了時にクリーンアップされないモデルのことです。

サポート対象ノード

多くの SPSS Modeler ノードでは、HDFS での実行がサポートされていますが、実行方法に相違点があるノードや、現時点でサポートされていないノードもあります。このトピックでは、現在のサポート・レベルについて詳しく説明します。

全般

- ここで示された Modeler のフィールド名では、通常どおりに受け入れ可能な文字の一部が、Analytic Server では受け入れられません。
- Modeler ストリームを Analytic Server で実行するには、ストリームを 1 つ以上の Analytic Server ソース・ノードで開始し、単一のモデル作成ノード、または Analytic Server エクスポート・ノードで終了する必要があります。結合は許可されていますが、fork は許可されていません。
- 連続型対象のストレージは、整数ではなく実数に設定することをお勧めします。スコアリング・モデルでは、連続型対象の出力データ・ファイルに必ず実数値が書き込まれるのに対し、スコアの出力データ・モデルでは、対象のストレージに従って処理が行われます。このため、連続型対象のストレージが整数である場合は、書き込み値とスコアのデータ・モデルに不一致が生じ、この不一致によって、スコアリングされたデータの読み取り時にエラーが発生します。

入力

- Analytic Server ソース・ノード以外のノードで開始されるストリームは、ローカルで実行されます。

レコード操作

すべてのレコード操作がサポートされています。以下では、これらのノードの機能について、特に注意が必要な点を示します。

条件抽出

- フィールド作成ノード (後述) でサポートされているものと同じ機能セットがサポートされています。

サンプル

- ブロック・レベルのサンプリングはサポートされていません。
- 複雑なサンプリング方法はサポートされていません。

レコード集計

- 連続キーはサポートされていません。
- 順序統計量 (中央値、第一四分位、第三四分位) はサポートされていません。

ソート

- 「最適化」タブはサポートされていません。

レコード結合

- 順序による結合はサポートされていません。
- 条件による結合はサポートされていません。
- 「最適化」タブはサポートされていません。
- サンプル・ノードまたはモデル・ナゲットの Analytic Server ソース・ノードおよびレコード結合ノード間への配置は、現在、サポートされていません。通常は、サンプル・ノードの機能の代わりに、条件抽出ノードを指定することができます。
- Analytic Server では、空の文字列キーに基づく結合は行われません。つまり、結合基準として指定したキーのいずれかに空の文字列が含まれる場合、空の文字列を含むレコードは、結合の出力から除去されます。
- 結合操作は比較的低速です。HDFS 内に使用可能なスペースがある場合は、各ストリームでデータ・ソースを結合するよりも、一度データ・ソースを結合してから、結合されたソースを以降のストリームで使用する方が、処理速度が大幅に向上する可能性があります。

フィールド操作

自動データ準備、データ型、フィルタリング、フィールド作成、アンサンブル、置換、データ分類、データ分割、RFM 分析、データ区分、フラグ設定、再構成、およびフィールド順序の各ノードがサポートされています。以下では、これらのノードの機能について、特に注意が必要な点を示します。

自動データ準備

- ノードの学習はサポートされていません。学習した自動データ準備ノードの変換の新規データへの適用はサポートされています。

データ型

- 「検査」列はサポートされていません。
- 「フォーマット」タブはサポートされていません。

フィールド作成

- 順序機能を除くすべてのフィールド作成機能がサポートされています。
- 分割フィールドは、分割として使用されている同じストリーム内では作成できません。分割フィールドを作成するストリームと、フィールドを分割として使用するストリームの 2 つのストリームを作成する必要があります。
- フラグ・フィールドを、比較内で単独で使用することはできません。つまり、if (flagField) then ... endif と指定すると、エラーが発生します。回避策として、if (flagField=trueValue) then ... endif を使用してください。
- ** 演算子を使用するときは、Modeler の結果に合わせて、指数を実数として指定することをお勧めします (x**2 ではなく x**2.0 など)。

置換

- フィールド作成ノード (前述) でサポートされているものと同じ機能セットがサポートされています。

RFM 分析

- タイの処理の「現在のまま保持」オプションはサポートされていません。RFM のリーセンシー、度数、およびマネタリーの各スコアは、同じデータから Modeler によって計算されたスコアと一致するとは限りません。スコアの範囲は同じですが、スコアの割り当て (ビン数) がそれぞれ異なっている場合があります。

グラフ作成

すべてのグラフ作成ノードがサポートされています。

モデル作成

モデル作成ノードでは、線形、ニューラル・ネットワーク、C&RT、CHAID、QUEST など、少数のノードがサポートされています。以下では、これらのノードの機能について、特に注意が必要な点を示します。

線形

- 既存の PSM モデルの継続学習はサポートされていません。
- 標準のモデル作成目的は、各分割のレコード数が多くなりすぎないように分割フィールドを定義する場合にのみお勧めします。ここで、「多すぎる」の定義は、ご使用の Hadoop クラスターの各ノードの処理能力によって異なります。一方、モデルを作成するためのレコードが少なくなりすぎることを防ぐため、分割をあまり細かく定義しないように注意する必要があります。
- 「ブースティング」目的はサポートされていません。
- 「バグ」目的はサポートされていません。
- 「特に大きいデータ・セット」目的は、レコードが少数の場合はお勧めしません。これは、モデルが作成されないか、品質の劣るモデルが作成されることが多いためです。また、使用されているアンサンブル・モデル作成アルゴリズムで仮定されているランダム性を損なうような組織的な方法で、入力レコードの順序が決定されている場合に、問題が発生することがあります。
- 自動データ準備はサポートされていません。このため、欠損値の多いデータに基づいてモデルを作成しようとしたときに、問題が発生する可能性があります。これらの値は、通常、自動データ準備の一環として代入されるからです。回避策として、ツリー・モデルまたはニューラル・ネットワークを拡張設定で使用して、選択した欠損値を代入することができます。
- 精度統計は、分割モデルについては計算されません。

ニューラル・ネットワーク

- 既存の標準または PSM モデルの継続学習はサポートされていません。
- 標準のモデル作成目的は、各分割のレコード数が多くなりすぎないように分割フィールドを定義する場合にのみお勧めします。ここで、「多すぎる」の定義は、ご使用の Hadoop クラスターの各ノードの処理能力によって異なります。一方、モデルを作成するためのレコードが少なくなりすぎることを防ぐため、分割をあまり細かく定義しないように注意する必要があります。
- 「ブースティング」目的はサポートされていません。
- 「バグ」目的はサポートされていません。

- 「特に大きいデータ・セット」目的は、レコードが少数の場合はお勧めしません。これは、モデルが作成されないか、品質の劣るモデルが作成されることが多いためです。また、使用されているアンサンブル・モデル作成アルゴリズムで仮定されているランダム性を損なうような組織的な方法で、入力レコードの順序が決定されている場合に、問題が発生することがあります。
- データ内に多くの欠損値がある場合は、拡張設定を使用して欠損値を代入してください。
- 精度統計は、分割モデルについては計算されません。

C&R ツリー、CHAID、および QUEST

- 既存の PSM モデルの継続学習はサポートされていません。
- 標準のモデル作成目的は、各分割のレコード数が多くなりすぎないように分割フィールドを定義する場合にのみお勧めします。ここで、「多すぎる」の定義は、ご使用の Hadoop クラスターの各ノードの処理能力によって異なります。一方、モデルを作成するためのレコードが少なくなりすぎることを防ぐため、分割をあまり細かく定義しないように注意する必要があります。
- 「ブースティング」目的はサポートされていません。
- 「バグ」目的はサポートされていません。
- 「特に大きいデータ・セット」目的は、レコードが少数の場合はお勧めしません。これは、モデルが作成されないか、品質の劣るモデルが作成されることが多いためです。また、使用されているアンサンブル・モデル作成アルゴリズムで仮定されているランダム性を損なうような組織的な方法で、入力レコードの順序が決定されている場合に、問題が発生することがあります。
- インタラクティブ・セッションはサポートされていません。
- 精度統計は、分割モデルについては計算されません。

モデルのスコアリング

スコアリングがサポートされているモデル・ナゲットは、線形、ニューラル・ネットワーク、C&RT、CHAID、QUEST、ロジスティック、回帰、GenLin、TwoStep クラスター、C5、Bayesian Network、SVM、R、およびテキスト・マイニングです。

- 未加工または調整された傾向はスコアリングされません。回避策として、フィールド作成ノードで以下の式を使用して、未加工の傾向を手動で計算することによって、同じ効果を得られる場合があります: if 'predicted-value' == 'value-of-interest' then 'prob-of-that-value' else 1-'prob-of-that-value' endif
- Analytic Server は、モデルをスコアリングするときに、モデル内のすべてのフィールドがデータ・セット内に存在するかどうか確認しません。このため、Analytic Server で操作を実行する前に、このことを確認してください。

出力 クロス集計、精度分析、データ検査、変換、記述統計、および平均比較の各ノードがサポートされています。

エクスポート

ストリームを、Analytic Server ソース・ノードで開始して、Analytic Server エクスポート・ノード以外のエクスポート・ノードで終了することができますが、データは HDFS から SPSS Modeler Server に移動され、最終的にエクスポート場所に移動されます。

第 3 章 Analytic Server コンソール

Analytic Server には、データ・ソースおよびプロジェクトを管理するためのシン・クライアント・インターフェースが用意されています。このインターフェースには、以下の 2 とおりの方法でアクセスできます。

- Analytic Server ソースまたはエクスポートの各ノード。コンソールがブラウザーで開かれます
- ブラウザーを使用した直接アクセス
  1. ブラウザーのアドレス・バーに Analytic Server の URL を入力します。これはサーバー管理者から入手できます。
  2. サーバーへのログオンに使用するユーザー名を入力します。
  3. 指定するユーザー名に関連付けられているパスワードを入力します。

ログオン後のデフォルトの折りたたみ式リストは「データ・ソース」です。

コンソールのナビゲート

Analytic Server コンソールには、以下の 4 つのコンポーネントがあります。

- ヘッダーには、製品名と、現在ログインしているユーザーの名前が表示されます。
- 左側の列には、使用可能な折りたたみ式リストまたは機能グループが表示されます。いずれの折りたたみ式リストを選択するかによって、内容領域の表示内容が決定されます。
- 内容領域。現在選択している折りたたみ式リストに関連付けられたコントロールが表示されます。それぞれの折りたたみ式リストの内容について詳しくは、以降のセクションで説明します。
- フッターには、ログアウト用のリンクと、インストールされている Analytic Server のバージョンが表示されます。

データ・ソース

データ・ソースは、レコードのコレクションとデータ・モデルから成り、分析対象のデータ・セットを定義します。レコードのソースには、HDFS 上のファイル (区切り文字で区切られたテキスト・ファイル、固定幅テキスト・ファイル、Excel ファイル)、データベース、または HCatalog を指定できます。データ・モデルは、データの分析に必要なすべてのメタデータ (フィールド名、ストレージ、測定のレベルなど) を定義します。データ・ソースの所有者は、データ・ソースへのアクセスを認可または制限することができます。

左側の列

左側の列の、折りたたみ式リストの見出しの下に既存のデータ・ソースが表示されます。

- 内容領域に詳細を表示するデータ・ソースを選択して、そのプロパティーを編集します。検索領域に入力すると、リストがフィルタリングされ、名前にその検索文字列が含まれるデータ・ソースのみが表示されます。
- 「**新規データ・ソース**」ボタンをクリックして、「**新規データ・ソースの追加**」ダイアログで指定した名前および内容タイプで新しいデータ・ソースを作成します。
  - データ・ソース名では大/小文字を区別する必要があります。先頭と末尾の空白文字は無視されます。一部の名前は SQL インジェクションに対する保護のために拒否されます。
  - 選択可能な内容タイプは、ファイル、データベース、および HCatalog です。

注: 内容タイプは、一度選択すると編集できません。

- 「**データ・ソースの削除 (Delete data source)**」ボタンをクリックして、データ・ソースを削除します。この操作では、データ・ソースに関連付けられているファイルはすべてそのまま残ります。

内容領域

内容領域は、複数のセクションに分かれています。これらのセクションは、データ・ソースの内容タイプによって異なる場合があります。データ・ソースの設定を指定したら、「プレビューとメタデータ (Preview and Metadata)」をクリックして、データ・ソースをファイナライズします。

データ・ソースのプロパティー

すべての内容タイプに共通する設定。

名前 データ・ソースの名前を表示する編集可能なテキスト・フィールド。データ・ソース名では大/小文字を区別する必要があります。先頭と末尾の空白文字は無視されます。一部の名前は SQL インジェクションに対する保護のために拒否されることがあります。

説明 データ・ソースに関する説明テキストを指定するための編集可能なテキスト・フィールド。

公開 (Is public)

すべてのユーザーがプロジェクトを参照できるか (チェック・マークを付けた場合)、ユーザーおよびグループを所有者リストに明示的に追加する必要があるか (チェック・マークを外した場合) を示すチェック・ボックス。

共有 (Sharing)

ユーザーおよびグループを作成者として追加することで、データ・ソースの所有権を共有できます。

- テキスト・ボックスに入力すると、ユーザーおよびグループがフィルタリングされ、名前に検索文字列が含まれるものが表示されます。これらのユーザーおよびグループを作成者のリストに追加するには、「**参加者の追加 (Add participant)**」ボタンをクリックします。
- 作成者を削除するには、作成者リストでユーザーまたはグループを選択し、「**参加者の削除 (Remove participant)**」ボタンをクリックします。

注: 管理者は、作成者として明示的にリストされているかどうかにかかわらず、すべてのデータ・ソースに対する読み取りおよび書き込み権限を持ちます。

ファイル入力 (File Input)

ファイル内容タイプのデータ・ソースを定義する場合に固有の設定。

ファイル・ビューアー (File Viewer)

データ・ソースに追加できるファイルが表示されます。「**プロジェクト**」モードを選択すると、Analytic Server プロジェクト構造内のファイルが表示され、「**HDFS**」を選択すると、HDFS の残りの部分が表示されます。どちらのフォルダー構造も参照は可能ですが、HDFS はまったく編集できず、Analytic Server フォルダー構造は定義されたプロジェクト内でのみ編集可能です。つまり、「**プロジェクト**」モードのルート・レベルでは、ファイルの追加、フォルダーの作成、または項目の削除ができません。プロジェクトを作成、編集、または削除するには、「プロジェクト」折りたたみ式リストを使用します。

- 「**HDFS にファイルをアップロード (Upload file to HDFS)**」ボタンをクリックすると、現在のプロジェクト/サブフォルダーにファイルがアップロードされます。

- 「**新規フォルダーの作成**」ボタンをクリックすると、「新規フォルダー名」ダイアログで指定した名前の新しいフォルダーが、現在のフォルダーの下に作成されます。
- 「**ローカル・ファイル・システムへのファイルのダウンロード (Download file to the local filesystem)**」ボタンをクリックすると、選択したファイルがローカル・ファイル・システムにダウンロードされます。
- 「**選択したファイルの削除 (Delete the selected file(s))**」ボタンをクリックすると、選択したファイル/フォルダーが削除されます。

データ・ソース定義に含まれるファイル (Files included in data source definition)
移動ボタンを使用して、選択したファイルを追加したり、データ・ソースから削除したりします。データ・ソース内の選択された各ファイルについて、「設定」をクリックすると、ファイル読み取りの仕様を定義できます。

ファイル出力 (File Output)
ファイル内容タイプのデータ・ソースには、Analytic Server で実行したストリームからの出力を付加できます。「**書き込み可能にする (Make writeable)**」を選択すると、付加を有効にして、新規ファイルの書き込み先となる出力フォルダーを選択できます。

データベース選択 (Database Selections)

レコードの内容を含むデータベースに対する接続パラメーターを指定します。

データベース
接続するデータベースの種類を選択します。DB2、Oracle、SQL Server、TeraData、または Netezza から選択してください。

サーバー・アドレス (Server address)
データベースをホストするサーバーの URL を入力します。

サーバー・ポート
データベースが listen するポートの番号。

データベース名
接続するデータベースの名前。

ユーザー名
データベースがパスワードで保護されている場合は、ユーザー名を入力します。

パスワード
データベースがパスワードで保護されている場合は、パスワードを入力します。

テーブル名
使用するデータベースの表の名前を入力します。

最大同時読み取り数 (Maximum concurrent reads)

データベース出力 (Database Output)
データベース内容タイプのデータ・ソースには、Analytic Server で実行されたストリームからの出力を付加できます。「**書き込み可能にする (Make writeable)**」を選択すると、付加を有効にして、出力データの書き込み先となる出力データベース表を選択できます。

HCatalog の選択 (HCatalog Selections)
Apache HCatalog の下で管理されているデータにアクセスするためのパラメーターを指定します。

データベース
HCatalog データベースの名前。

テーブル名
使用するデータベースの表の名前を入力します。

フィルター
表がデータ区分された表として作成されている場合は、表のデータ区分フィルターです。

HCatalog スキーマ
指定した表の構造を表示します。HCatalog では、高度に構造化したデータ・セットをサポートできます。そのようなデータに対して Analytic Server データ・ソースを定義するには、構造をフラットにして単純な行と列にする必要があります。スキーマの要素を選択して移動ボタンをクリックすると、その要素を分析のためにフィールドにマップすることができます。ツリー・ノードのすべてをマップできるわけではありません。例えば、複合タイプの配列またはマップは「親」と見なされるため、マップできません。そのようなノードは、ツリー内でラベルの末尾が ...:array:struct または ...:map:struct であることから識別できます。

HCatalog のフィールドのマッピング (HCatalog Field Mappings)
データ・ソース内のフィールドへの、HCatalog 内の要素のマッピングを表示します。「**生データのプレビュー (Preview Raw Data)**」をクリックすると、HCatalog に保管された状態のレコードを表示できます。これは、HCatalog スキーマをフィールドにどのようにマップするかを決定するときに役立つことがあります。

HCatalog の要素 (HCatalog Element)
編集するにはセルをダブルクリックします。HCatalog の要素が配列またはマップである場合は、セルを編集する必要があります。配列の場合は、フィールドにマップする配列のメンバーに対応する整数を指定します。マップの場合は、フィールドにマップするキーに対応する文字列を引用符で囲んで指定します。マップのインデックスに対応する文字列を、生データのプレビューを使用して判別する方法の例については、13 ページの図 2 を参照してください。

マッピング・フィールド (Mapping Field)
Analytic Server データ・ソースに表示されるフィールド。編集するにはセルをダブルクリックします。「マッピング・フィールド (Mapping Field)」列では重複した値は許可されず、エラーとなります。

ストレージ (Storage)
フィールドのストレージ。ストレージは HCatalog から取得され、編集できません。

注: 「プレビューとメタデータ (Preview and Metadata)」をクリックして HCatalog データ・ソースをファイナライズする場合、編集オプションはありません。

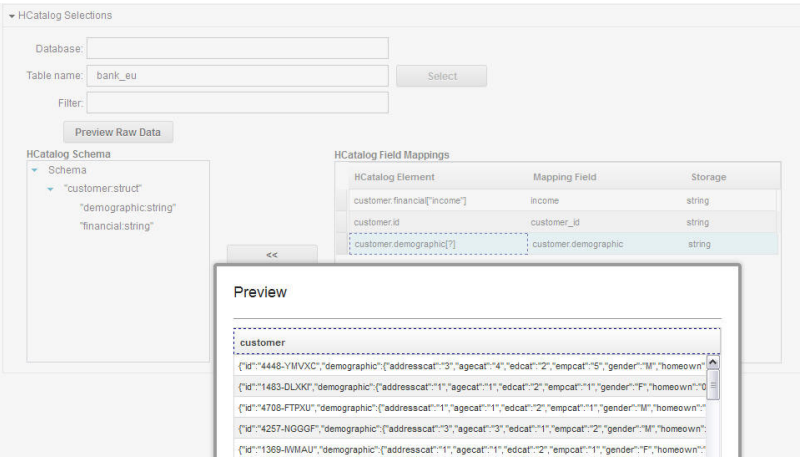

図 2. HCatalog データ・ソースを定義する「データ・ソース」折りたたみ式リスト

設定 (ファイル・データ・ソース)

文字セットのエンコード (Character set encoding)
ファイルの文字エンコード。「UTF-8」、「ISO-8859-2」、「GB18030」などの Java 文字セット名を選択または指定します。デフォルトは「UTF-8」です。

ロケール
ロケールを定義します。オプションです。デフォルトはサーバー・ロケールです。ロケール文字列は、<language>[_country[_variant]] の形式で指定する必要があります。ここで、

language
ISO-639 で定義された、小文字 2 文字の有効なコードです。必須です。

country
ISO-3166 で定義された、大文字 2 文字の有効なコードです。オプションです。

variant
ベンダーまたはブラウザー固有のコード。オプションです。

空白の切り取り (Trim white space)
文字列フィールドの先頭および/または末尾から空白文字を削除します。デフォルトは「なし」です。以下の値がサポートされています。

なし 空白文字を除去しません。

左 文字列の先頭から空白文字を除去します

右 文字列の末尾から空白文字を除去します

両方 文字列の先頭と末尾から空白文字を除去します。

グループ化記号
桁区切り文字に使用されるロケール固有の文字を使用するかどうかを設定します。

フィールド区切り文字
フィールドの境界を示す 1 つ以上の文字。それぞれの文字は、個別の区切り文字と解釈されます。例えば「,¥t」の場合は、コンマまたはタブのいずれかがフィールド境界を示すと指定されま

す。制御文字によってフィールドが区切られる場合は、制御文字に加えて、ここで指定された文字が、区切り文字として扱われます。制御文字によってフィールドが区切られない場合のデフォルトは「,」、それ以外の場合は空の文字列です。

制御文字によるフィールドの区切り (Control characters delimit fields)
LF と CR を除くすべての ASCII 制御文字をフィールド区切り文字として扱うかどうかを設定します。デフォルトは「**いいえ**」です。

最初の行がフィールド名を含む (First row contains field names)
最初の行をフィールド名の決定に使用するかどうかを設定します。デフォルトは「**いいえ**」です。

スキップする先頭文字数 (Number of initial characters to skip)
ファイルの先頭でスキップする文字の数。この値は、負でない整数です。デフォルトはゼロです。

空白の結合 (Merge white space)
複数のスペースやタブが連続しているときに、それらを単一のフィールド区切り文字と見なすかどうかを設定します。スペースもタブもフィールド区切り文字でない場合は無視されます。デフォルトは true です。

行末コメント文字 (End-of-line comment characters)
行末コメントを示す 1 つ以上の文字。レコード上で、この文字以降のものはすべて無視されます。それぞれの文字は、個別のコメント・マーカーと解釈されます。例えば「/*」の場合は、スラッシュまたはアスタリスクのいずれかでコメントが開始すると指定されます。複数の文字からなるコメント・マーカー (「//」など) を定義することはできません。空文字列の場合は、コメント文字を定義しないことを示します。定義した場合は、引用符を処理する前、またはスキップする先頭文字をスキップする前に、コメント文字が検査されます。デフォルトは空文字列です。

不正な文字
無効な文字 (エンコードの文字に対応しないバイト・シーケンス) の処理方法を指定します。空の文字列を指定すると、無効な文字が破棄されることを示します。空でない文字列 (通常は単一の文字) を指定すると、無効な文字が文字列の内容に置き換えられることを示します。デフォルトは空文字列です。

単一引用符
単一引用符 (アポストロフィ) の処理を指定します。デフォルトは「**保存**」です。

保存 単一引用符が特別な意味を持たず、他の文字と同じように処理されます。

除去 (Drop)
引用符の付いていない単一引用符は削除されます。

ペア (Pair)
単一引用符が引用符文字として処理され、単一引用符のペアの間にある文字が特別な意味を持ちません (引用符で囲まれていると見なされます)。単一引用符で囲まれた文字列の中に単一引用符自体を含めることができるかどうかは、設定「**二重化によって引用符を引用符で囲むことができる (Quotes can be quoted by doubling)**」によって決定されます。

二重引用符
二重引用符の処理を指定します。デフォルトは「**ペア (Pair)**」です。

保存 二重引用符が特別な意味を持たず、他の文字と同じように処理されます。

除去 (Drop)
引用符の付いていない二重引用符は削除されます。

ペア (Pair)
二重引用符が引用符文字として処理され、二重引用符のペアの間にある文字が特別な意味

を持ちません (引用符で囲まれていると見なされます)。二重引用符で囲まれた文字列の中に二重引用符自体を含めることができるかどうかは、設定「**二重化によって引用符を引用符で囲むことができる (Quotes can be quoted by doubling)**」によって決定されます。

二重化によって引用符を引用符で囲むことができる (Quotes can be quoted by doubling)
「**ペア (Pair)**」に設定したときに、二重引用符で囲んだ文字列内で二重引用符を表現できるかどうか、および単一引用符で囲んだ文字列内で単一引用符を表現できるかどうかを示します。「**はい**」の場合は、二重引用符で囲んだ文字列の内側で二重引用符を 2 個連続して記述するとエスケープされ、単一引用符で囲んだ文字列の内側で単一引用符を 2 個連続して記述するとエスケープされます。「**いいえ**」の場合は、二重引用符で囲んだ文字列の内側で二重引用符を使用することはできず、単一引用符で囲んだ文字列の内側で単一引用符を使用することもできません。デフォルトは「**はい**」です。

プレビューとメタデータ (Preview and Metadata) (データ・ソース)

「**プレビューとメタデータ (Preview and Metadata)**」をクリックすると、レコードのサンプルに基づいて、データ・ソースのデータ・モデルが生成されます。ここでは、基本的なメタデータ情報を確認および編集できます。

プレビュー
「プレビュー」タブには、レコードの小規模なサンプルと、それらのフィールド値が表示されます。

編集

「編集」タブでは、ユーザーが基本的なフィールドのメタデータを修正できます。

フィールド
フィールド名をダブルクリックして編集します。

測定 (Measurement)
指定されたフィールド内のデータの特性を示す測定の尺度です。

役割 (Role)
フィールドが、マシン学習プロセスの入力 (予測フィールド) と対象 (予測済みフィールド) のどちらであるかを、モデル作成ノードに示すために使用されます。使用可能な役割には、「データ区分 (Partition)」の他に、「両方」および「なし」もあります。「データ区分 (Partition)」は、レコードを、学習、テスト、および検証用の個別のサンプルに区分けするために使用されるフィールドを示します。値「分割 (Split)」は、フィールドに入力される可能性のある値ごとに、個別のモデルが作成されるように指定します。

ストレージ (Storage)
ストレージは、データをフィールドに保管する方法を示します。例えば、値 1 および 0 を含むフィールドには、整数データが保管されます。これは、測定の尺度とは異なります。測定の尺度は、データの使用法を示すもので、ストレージには影響しません。例えば、値 1 および 0 を含む整数フィールドの測定の尺度は、フラグ (Flag) に設定できます。これは、通常、1 が True、0 が False であることを示します。

プロジェクト

プロジェクトは、入力を保管し、ジョブの出力にアクセスするためのワークスペースです。これは、ファイルおよびフォルダーを追加するための最上位の編成構造です。プロジェクトは、個々のユーザーおよびグループと共有することができます。

左側の列

左側の列の、折りたたみ式リストの見出しの下に既存のプロジェクトが表示されます。

- 内容領域に詳細を表示するプロジェクトを選択して、そのプロパティーを編集します。検索領域に入力すると、リストがフィルタリングされ、名前にその検索文字列が含まれるプロジェクトのみが表示されます。
- 「**新規プロジェクト**」をクリックすると、「新規プロジェクトの追加 (Add New Project)」ダイアログで指定した名前で新しいプロジェクトが作成されます。名前は大/小文字が区別され、先頭および末尾の空白文字が無視され、SQL インジェクションから保護されます。
- プロジェクトを削除するには「**プロジェクトの削除 (Delete Project)**」をクリックします。この操作では、データ・ソースに関連付けられているファイルはすべてそのまま残ります。

内容領域

内容領域は、「**設定**」、「**所有者**」、および「**バージョン**」の各タブに分かれています。

設定

プロジェクトの説明 (Project description)
プロジェクトに関する説明テキストを指定するための編集可能なテキスト・フィールド。

公開 (Is public)
すべてのユーザーがプロジェクトを参照できるか (チェック・マークを付けた場合)、ユーザーおよびグループを所有者に明示的に追加する必要があるか (チェック・マークを外した場合) を示すチェック・ボックス。

「**保存**」をクリックすると、現在の設定の状態が保存されます。

プロジェクトのデータ・ソース (Project data sources)
プロジェクトに関連付けられたすべてのデータ・ソースをリストする編集不可能な領域。

プロジェクト構造ペイン

右側のペインには、現在選択しているプロジェクトのプロジェクト/フォルダー構造が表示されます。フォルダー構造を参照することができますが、ボタン以外の方法で編集することはできません。

- 「**ローカル・ファイル・システムへのファイルのダウンロード (Download file to the local filesystem)**」をクリックすると、選択したファイルがローカル・ファイル・システムにダウンロードされます。
- 「**選択したファイルの削除 (Delete the selected file(s))**」をクリックすると、選択したファイル/フォルダーが削除されます。

所有者

所有者はプロジェクトのすべての権限を持つメンバーであり、プロジェクトの他に、その中のフォルダーおよびファイルを変更することができます。

「使用可能なユーザー/グループ」リストには、現在このプロジェクトに関連付けられていないアクティブなテナントのユーザーおよびグループが表示されます。

- 検索ペインに入力すると、ユーザーおよびグループがフィルターに掛けられ、名前に検索文字列が含まれるものが表示されます。
- リストの上にあるユーザー・アイコンを選択すると、選択可能なユーザーが表示されます。アイコンを選択解除すると、ユーザーが非表示になります。このアイコンは、デフォルトで選択されています。

- リストの上にあるグループ・アイコンを選択すると、選択可能なグループが表示されます。アイコンを選択解除すると、グループが非表示になります。このアイコンは、デフォルトで選択されています。

移動ボタンを使用すると、ユーザーおよびグループを「プロジェクト・ユーザーおよびグループ (Project Users and Groups)」に移動することができます。これらのユーザーおよびこれらのグループのメンバーは、IBM SPSS Modeler から Analytic Server に接続しているときに、このプロジェクトに対する読み取り権限 (Analytic Server ソース・ノード) および書き込み権限 (Analytic Server エクスポート・ノード) を持ちます。

注: 「所有者」タブで行われた変更は、すぐに自動的に適用されます。

注: 管理者は、所有者として明示的にリストされているかどうかにかかわらず、すべてのプロジェクトに対する読み取りおよび書き込み権限を持ちます。

バージョン

プロジェクトは、ファイルおよびフォルダーの内容に加えられた変更に基づいてバージョン管理されます。プロジェクトの属性 (説明など) を変更する場合は、公開されているかどうか、およびどのユーザーと共有しているかにかかわらず、新しいバージョンは不要です。ファイルまたはフォルダーの追加、変更、または削除を行う場合は、新しいバージョンが必要です。

プロジェクトのバージョン管理表

表には、既存のプロジェクト・バージョン、作成日、確定日、各バージョンの担当ユーザー、および親バージョンが表示されます。親バージョンとは、選択したバージョンの基となるバージョンのことです。

- 「**ロック**」をクリックすると、選択したプロジェクト・バージョンの内容を変更することができます。
- 「**確定 (Commit)**」をクリックすると、プロジェクトに対するすべての変更が保存され、そのバージョンが現行の参照可能状態のプロジェクトになります。
- 「**破棄**」をクリックすると、ロックしたプロジェクトに対する変更がすべて破棄され、参照可能状態のプロジェクトが最新の確定バージョンに戻ります。
- 「**削除**」をクリックすると、選択したバージョンが削除されます。

バージョン数が超過した場合に自動的にクリーンアップ (Automatically clean up when number of versions exceeds)

バージョンの数が指定の数を超えた場合に、最も古い確定プロジェクト・バージョンを自動的に削除します。

特記事項

本書は米国 IBM が提供する製品およびサービスについて作成したものです。

本書に記載の製品、サービス、または機能が日本においては提供されていない場合があります。日本で利用可能な製品、サービス、および機能については、日本 IBM の営業担当員にお尋ねください。本書で IBM 製品、プログラム、またはサービスに言及していても、その IBM 製品、プログラム、またはサービスのみが使用可能であることを意味するものではありません。これらに代えて、IBM の知的所有権を侵害することのない、機能的に同等の製品、プログラム、またはサービスを使用することができます。ただし、IBM 以外の製品とプログラムの操作またはサービスの評価および検証は、お客様の責任で行っていただきます。

IBM は、本書に記載されている内容に関して特許権 (特許出願中のものを含む) を保有している場合があります。本書の提供は、お客様にこれらの特許権について実施権を許諾することを意味するものではありません。実施権についてのお問い合わせは、書面にて下記宛先にお送りください。

〒103-8510
東京都中央区日本橋箱崎町19番21号
日本アイ・ビー・エム株式会社
法務・知的財産
知的財産権ライセンス渉外

以下の保証は、国または地域の法律に沿わない場合は、適用されません。 IBM およびその直接または間接の子会社は、本書を特定物として現存するままの状態で提供し、商品性の保証、特定目的適合性の保証および法律上の瑕疵担保責任を含むすべての明示もしくは黙示の保証責任を負わないものとします。国または地域によっては、法律の強行規定により、保証責任の制限が禁じられる場合、強行規定の制限を受けるものとします。

この情報には、技術的に不適切な記述や誤植を含む場合があります。本書は定期的に見直され、必要な変更は本書の次版に組み込まれます。IBM は予告なしに、随時、この文書に記載されている製品またはプログラムに対して、改良または変更を行うことがあります。

本書において IBM 以外の Web サイトに言及している場合がありますが、便宜のため記載しただけであり、決してそれらの Web サイトを推奨するものではありません。それらの Web サイトにある資料は、この IBM 製品の資料の一部ではありません。それらの Web サイトは、お客様の責任でご使用ください。

IBM は、お客様が提供するいかなる情報も、お客様に対してなんら義務も負うことのない、自ら適切と信ずる方法で、使用もしくは配布することができるものとします。

本プログラムのライセンス保持者で、(i) 独自に作成したプログラムとその他のプログラム (本プログラムを含む) との間での情報交換、および (ii) 交換された情報の相互利用を可能にすることを目的として、本プログラムに関する情報を必要とする方は、下記に連絡してください。

IBM Software Group
ATTN: Licensing
200 W. Madison St.
Chicago, IL; 60606
U.S.A.

本プログラムに関する上記の情報は、適切な使用条件の下で使用することができますが、有償の場合もあります。

本書で説明されているライセンス・プログラムまたはその他のライセンス資料は、IBM 所定のプログラム契約の契約条項、IBM プログラムのご使用条件、またはそれと同等の条項に基づいて、IBM より提供されます。

この文書に含まれるいかなるパフォーマンス・データも、管理環境下で決定されたものです。そのため、他の操作環境で得られた結果は、異なる可能性があります。一部の測定が、開発レベルのシステムで行われた可能性がありますが、その測定値が、一般に利用可能なシステムのものと同じである保証はありません。さらに、一部の測定値が、推定値である可能性があります。実際の結果は、異なる可能性があります。お客様は、お客様の特定の環境に適したデータを確かめる必要があります。

IBM 以外の製品に関する情報は、その製品の供給者、出版物、もしくはその他の公に利用可能なソースから入手したものです。 IBM は、それらの製品のテストは行っておりません。したがって、他社製品に関する実行性、互換性、またはその他の要求については確証できません。 IBM 以外の製品の性能に関する質問は、それらの製品の供給者にお願いします。

IBM の将来の方向または意向に関する記述については、予告なしに変更または撤回される場合があり、単に目標を示しているものです。

表示されている IBM の価格は IBM が小売り価格として提示しているもので、現行価格であり、通知なしに変更されるものです。卸価格は、異なる場合があります。

本書はプランニング目的としてのみ記述されています。記述内容は製品が使用可能になる前に変更になる場合があります。

本書には、日常の業務処理で用いられるデータや報告書の例が含まれています。より具体性を与えるために、それらの例には、個人、企業、ブランド、あるいは製品などの名前が含まれている場合があります。これらの名称はすべて架空のものであり、名称や住所が類似する企業が実在しているとしても、それは偶然にすぎません。

それぞれの複製物、サンプル・プログラムのいかなる部分、またはすべての派生的創作物にも、次のように、著作権表示を入れていただく必要があります。

本書には、日常の業務処理で用いられるデータや報告書の例が含まれています。より具体性を与えるために、それらの例には、個人、企業、ブランド、あるいは製品などの名前が含まれている場合があります。これらの名称はすべて架空のものであり、名称や住所が類似する企業が実在しているとしても、それは偶然にすぎません。

それぞれの複製物、サンプル・プログラムのいかなる部分、またはすべての派生的創作物にも、次のように、著作権表示を入れていただく必要があります。

© (お客様の会社名) (年). このコードの一部は、IBM Corp. のサンプル・プログラムから取られています。

この情報をソフトコピーでご覧になっている場合は、写真やカラーの図表は表示されない場合があります。

商標

IBM、IBM ロゴおよび ibm.com は、世界の多くの国で登録された International Business Machines Corporation の商標です。他の製品名およびサービス名等は、それぞれ IBM または各社の商標である場合があります。現時点での IBM の商標リストについては、http://www.ibm.com/legal/copytrade.shtml をご覧ください。

Adobe、Adobe ロゴ、PostScript、PostScript ロゴは、Adobe Systems Incorporated の米国およびその他の国における登録商標または商標です。

IT Infrastructure Library は英国 Office of Government Commerce の一部である the Central Computer and Telecommunications Agency の登録商標です。

インテル、Intel、Intel ロゴ、Intel Inside、Intel Inside ロゴ、Centrino、Intel Centrino ロゴ、Celeron、Xeon、Intel SpeedStep、Itanium、および Pentium は、Intel Corporation または子会社の米国およびその他の国における商標または登録商標です。

Linux は、Linus Torvalds の米国およびその他の国における商標です。

Microsoft、Windows、Windows NT および Windows ロゴは、Microsoft Corporation の米国およびその他の国における商標です。

ITIL は英国 The Minister for the Cabinet Office の登録商標および共同体登録商標であって、米国特許商標庁にて登録されています。

UNIX は The Open Group の米国およびその他の国における登録商標です。

Java およびすべての Java 関連の商標およびロゴは Oracle やその関連会社の米国およびその他の国における商標または登録商標です。

Cell Broadband Engine は、Sony Computer Entertainment, Inc. の米国およびその他の国における商標であり、同社の許諾を受けて使用しています。

Linear Tape-Open、LTO、LTO ロゴ、Ultrium および Ultrium ロゴは、HP、IBM Corp. および Quantum の米国およびその他の国における商標です。

索引

日本語, 数字, 英字, 特殊文字の順に配列されています。なお, 濁音と半濁音は清音と同等に扱われています。

［ア行］

［サ行］

［タ行］

［ハ行］

A

Printed in Japan

日本アイ・ビー・エム株式会社
〒103-8510 東京都中央区日本橋箱崎町19-21