

**IBM SPSS Modeler Entity  
Analytics 17 ユーザー・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、61 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(r) SPSS(r) Modeler バージョン 17 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 Modeler Entity Analytics 17 User 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

前書き	v
<b>第 1 章 エンティティ分析</b>	<b>1</b>
エンティティ分析について	1
エンティティ分析と予測分析	2
<b>第 2 章 IBM SPSS Modeler でのエンティティ分析</b>	<b>5</b>
IBM SPSS Modeler でのエンティティ分析の使用	5
段階 1: ソース・データを SPSS Modeler に読み込む	6
段階 2: リポジトリを作成する	6
段階 3: SPSS Modeler をリポジトリに接続する	7
段階 4: 入力フィールドをリポジトリ・フィーチャーにマップする	7
段階 5: データをリポジトリにエクスポートして ID を解決する	7
段階 6: 解決済み ID を分析する	9
段階 7: リポジトリに対して新規ケースを解決する	10
段階 8: 警告を生成する	11
<b>第 3 章 エンティティ分析のタスク</b>	<b>13</b>
タスクについて	13
エンティティ・リポジトリの設定 (EA エクスポート・ノード)	13
エンティティ・リポジトリ	13
データ・ソースへの接続	14
リポジトリの作成	15
入力フィールドとフィーチャーのマッピング (EA エクスポート・ノード)	17
フィールド・マッピングの表示 (EA エクスポート・ノード)	19
エンティティ・リポジトリの構成	19
データ・ソース・マッピングの表示	20
リポジトリ・フィーチャーの保守	20
フィーチャーの追加または編集	22
リポジトリ・フィーチャーの匿名化	23
エンティティ・タイプの保守	24
エンティティの一致に対するしきい値の設定	26
リポジトリ構成の再利用	26
構成の変更の保存	27
構成ウィンドウを閉じる	27
解決済み ID の分析 (エンティティ分析 (EA) ソース・ノード)	27
データ・ソースの選択	27
データ・フィールド名の変更	28
データ・フィールドの型情報の設定	28
ストリームへのノードの追加	29

リポジトリに対する新規ケースの比較 (ストリーミング EA ノード)	29
入力フィールドとフィーチャーのマッピング (ストリーミング EA ノード)	30
フィールド・マッピングとデータ・ソースの表示 (ストリーミング EA ノード)	31
ストリーミング EA ノードからの出力	32
他の IBM SPSS 製品での IBM SPSS Modeler Entity Analytics の使用	33
管理タスク	33
ポート割り当ての構成	34
リポジトリ・データベースの管理者資格情報の管理	34
別のストレージ・ディレクトリへのリポジトリの移動	35
日時フィールドとタイム・スタンプ・フィールドに対するストリーム・プロパティの設定	36
タイムアウト設定の調整	36
SPSS Modeler Client と SPSS Modeler Server が同じ Windows システム上にある場合の IBM SPSS Modeler Entity Analytics の実行	36
エンティティ・リポジトリの消去	37
リポジトリからの未使用データ・ソースの削除	37
エンティティ・リポジトリの削除	38
リポジトリに接続できない場合のリポジトリの削除	38

<b>第 4 章 エンティティ分析の実行</b>	<b>41</b>
この例について	41
元のモデル	41
エンティティ分析の追加	45
リポジトリへのソース・データの取り込み	45
解決済み ID の読み取り	46
エンティティ分析出力と元のモデルとの比較	52
要約	56

<b>付録. IBM SPSS Modeler Entity Analytics のスクリプト記述プロパティ</b>	<b>57</b>
IBM SPSS Modeler Entity Analytics でのスクリプト記述	57
共通プロパティ	57
entityanalytics_exportnode プロパティ	57
entityanalytics_sourcenode プロパティ	58
entityanalytics_processnode プロパティ	59

<b>特記事項</b>	<b>61</b>
商標	62

<b>索引</b>	<b>63</b>
-----------	-----------



---

## 前書き

IBM® SPSS® Modeler は、IBM Corp. が提供する企業向けデータ・マイニング・ワークベンチです。SPSS Modeler は、データの詳細な分析を提供することにより、顧客との関係を深めることができるよう、企業を支援します。SPSS Modeler から得た知見を基に、既存顧客のつなぎとめ、クロスセル機会の特定、新規顧客へのアピール、不正行為の検出、リスクの低減、行政サービス提供の改善などを行うことができます。

SPSS Modeler のビジュアル・インターフェースには、ユーザー固有のビジネスの専門知識を適用することができます。これにより、予測モデルを強化し、解決までの所要時間を短縮することができます。SPSS Modeler には、予測、分類、セグメンテーション、関連検出の各アルゴリズムなど、さまざまなモデリング技法が組み込まれています。モデルを作成したら、IBM SPSS Modeler Solution Publisher を使用して、そのモデルを意思決定者やデータベースに対して全社的に配信することができます。

## IBM Business Analytics について

IBM Business Analytics ソフトウェアは、完全で正確な一貫性のある情報を提供します。意思決定者は、この情報に基づいて業績を改善することができます。ビジネス・インテリジェンス、予測分析、財務実績および戦略管理、分析アプリケーションから構成される包括的なポートフォリオにより、現在のパフォーマンスに関する明確で実践可能な知見が迅速に提供され、将来的な結果を予測することができます。さまざまな業種向けソリューション、これまでの実施例、専門的なサービスを組み合わせることにより、あらゆる規模の組織において、生産性を最大限に高め、意思決定を確実に自動化し、より適切な結果を得ることができます。

このポートフォリオの一部として組み込まれている IBM SPSS Predictive Analytics ソフトウェアを使用すると、将来的に何が起きるかを予測し、その知見に基づいて事前に対策を講じることにより、より良いビジネスの成果を得ることができます。IBM SPSS のテクノロジーは、不正行為やリスクを低減しながら、顧客へのアピール、顧客のつなぎとめ、顧客の拡大を行うための強力な手段として、企業、行政機関、教育機関など、世界中のユーザーから信頼されています。IBM SPSS ソフトウェアを日常業務に取り入れることにより、ビジネスの目標に適合した意思決定の指示と自動化を行い、測定可能な競争上の優位性を実現できる、予測型の企業になることができます。詳細情報や担当者へのお問い合わせについては、<http://www.ibm.com/spss> を参照してください。

## 技術サポート

お客様はテクニカル・サポートをご利用いただけます。IBM Corp. 製品の使用に関する支援や、サポート対象ハードウェア環境のインストールに関する支援については、技術サポートにお問い合わせください。技術サポートへのお問い合わせについては、IBM Corp. の Web サイト (<http://www.ibm.com/support>) を参照してください。サポートを要請する場合は、お客様のご本人確認、所属している組織、サポート契約について、事前に情報を準備しておいてください。



---

# 第 1 章 エンティティ分析

---

## エンティティ分析について

IBM SPSS Modeler Entity Analytics により、IBM SPSS Modeler の予測分析に新たな次元が追加されます。予測分析は過去のデータから将来の動きを予測しようとするのに対し、エンティティ分析は、レコード内での ID の競合を解決することにより、現在のデータの整合性と一貫性を高めることに重点を置いています。ID は、個人、組織、オブジェクトなどの任意のエンティティの ID にすることができますが、あいまいさが存在する場合があります。ID の解決は、カスタマー・リレーションシップ・マネジメント、不正行為の検出、マネー・ロンダリングの防止、国内外のセキュリティーなど、さまざまな分野で重要になります。

2 つの異なるソースに以下の顧客レコードがあり、それらが同一人物を参照しているのか、異なる人物を参照しているのかわからないとします。

### ソース 1

Record no.: 70001  
Name: Jon Smith  
Address: 123 Main Street  
Tax Reference: 555-00-1111  
Driv. License: 0001133107  
Cred. Card: 10229127

### ソース 2

Record no.: 9103  
Name: JOHNATHAN Smith  
Date of Birth: 06/17/1934  
Telephone: 555-1212  
Cred. Card: 10229128  
Email: jls@mail.com  
IP address: 9.50.18.77

2 つのレコード間にデータの完全一致は存在しません。ただし、3 番目のソースを導入すると、共通する属性がいくつか見つかります。

### ソース 3

Record no.: 6251  
Name: Jon Smith  
Telephone: 555-1212  
Driv. License: 0001133107  
Cred. Card: 10229132

運転免許証番号ではソース 1 とソース 3 のレコードがリンクしますが、電話番号ではソース 2 とソース 3 のレコードがリンクします。したがって、3 つのソースすべてが同じ人物を指していると考えられます。

一方、それほど簡単に区別できない場合はどうでしょうか。判断の基準となるデータが非常に少ない場合も考えられます。ここでは、以下の 2 つのレコードについて考えてみます。

### ソース 4

Record no.: S45286  
Name: John T Smith Jr  
Address: 456 Main Street  
Telephone: 703-555-2000  
Date of birth: 03/12/1984

Record no.: S45287  
Name: John T Smith  
Address: 456 Main Street  
Telephone: 703-555-2000  
Driv. License: 009900991

これは明らかに、前のレコードと同じ Smith 氏ではありません。同じではないことを示す十分な違いがあります。しかしながら、まだ問題があります。それは、同じデータ・ソースの 2 つの異なるレコードが同じ人物に関連しているように見えることです。これらは重複レコードでしょうか。別のソースなどから、追加の情報を提供する別の関連レコードが見つからない限り、重複レコードかどうかを判断することはできません。

ソース 5

Record no.: 769582-2  
Name: John T Smith Sr  
Address: 456 Main Street  
Telephone: 703-555-2000  
Driv. License: 009900991  
Date of birth: 06/25/1959

これで問題が解決します。ソース 4 の 2 つのレコードは重複レコードではなく、実際には同じ名前と同じ住所に居住し、同じ電話番号を使用している父親と息子でした。手動によるシステムならば、身元を解決する 1 件のレコードを見つけるまでに何週間もかかる可能性があります。自動化されたエンティティ分析システムを使用すれば、解決までの所要時間が大幅に短縮されます。

---

## エンティティ分析と予測分析

使用するすべてのデータが、完全に明確なレコードの単一ソースで構成されていれば、IBM SPSS Modeler で ID の競合を解決するのは比較的簡単です。予測分析だけを使用して、データを IBM SPSS Modeler に読み込んで処理を実行すれば、信頼性の高い結果が得られます。

しかしながら、実際の状況は大きく異なる場合があります。通常、データは完全な状態からはほど遠く、多くの場合は不明確で、さまざまなデータ・ソース上に存在し、記録されている属性もさまざまで、重複フィールドがほとんどないことも珍しくありません。エンティティ分析の価値の 1 つは、あらゆるソースからのデータをリポジトリと呼ばれる単一の中央ストレージ領域内に収集する点にあります。収集されたデータは、エンティティ分析システムによって詳細に検査され、競合が解決されます。同じ人物や組織から発生したレコードには、一意の識別子が付加されます。

以下の表に、2 種類の分析の違いを示します。

表 1. 予測分析とエンティティ分析の違い：

特徴	予測分析	エンティティ分析
学習データのタイプ	比較的小さなセットや数値範囲に基づく	名前やアドレスなどの大きなセット (データ型不明のフィールド) を活用できます



表 1. 予測分析とエンティティ分析の違い (続き) :

特徴	予測分析	エンティティ分析
学習データのサイズ	通常、大きなセット (データ型不明のフィールド) は無視されます	使用されるすべてのデータ
一般化	簡潔なモデルを作成するために、アルゴリズムによって学習データ全体が一般化されます。	エンティティの一致と関係の検出に適した構造でデータが永続化されず。
不正検出	レコード内に不正アプリケーションの典型的な特徴がある場合は、不正の可能性ありとしてフラグが立てられます。	レコードが既知の不正レコードに関連する場合や、同じ人物から発生しているにもかかわらず ID が異なる場合は、不正の可能性ありとしてフラグが立てられます。



---

## 第 2 章 IBM SPSS Modeler でのエンティティ分析

---

### IBM SPSS Modeler でのエンティティ分析の使用

データで ID の問題が発生した疑いが生じることがあります。例えば、個人が複数回表示されたり、異なる個人がマージされたり欠落しているように見える場合などです。IBM SPSS Modeler Entity Analytics を使用して、この問題に対処することができます。推奨手順を以下に示しますが、特定の要件に合わせてこの手順を変更してください。

- ソース・データを IBM SPSS Modeler に読み込む。
- データを格納するリポジトリを作成する。
- IBM SPSS Modeler をリポジトリに接続する。
- データ・フィールドをリポジトリ・フィーチャーにマップする。
- データをリポジトリにエクスポートして ID を解決する。
- 解決済み ID を分析する。
- リポジトリに対して新規ケースを解決する。
- 必要な警告 (バッチまたはリアルタイム) を生成する。

この時点で、IBM SPSS Modeler の仕組みをある程度知っておく必要があります。IBM SPSS Modeler は非常に使いやすいツールで、さまざまなノードを経由して流れるデータ・ストリームのグラフィカルな表現に基づいています。各ノードは、ワークフローの特定の段階を表します。

IBM SPSS Modeler 自体にはさまざまなノードが用意されており、標準的なデータ・マイニング機能をすべてカバーしています。IBM SPSS Modeler Entity Analytics は、特にエンティティ分析で使用されるノードを追加します。これらは、EA エクスポート・ノード、エンティティ分析 (EA) ソース・ノード、およびストリーミング EA プロセス・ノードです。

以下の図は、エンティティ分析プロセスを示したものです。

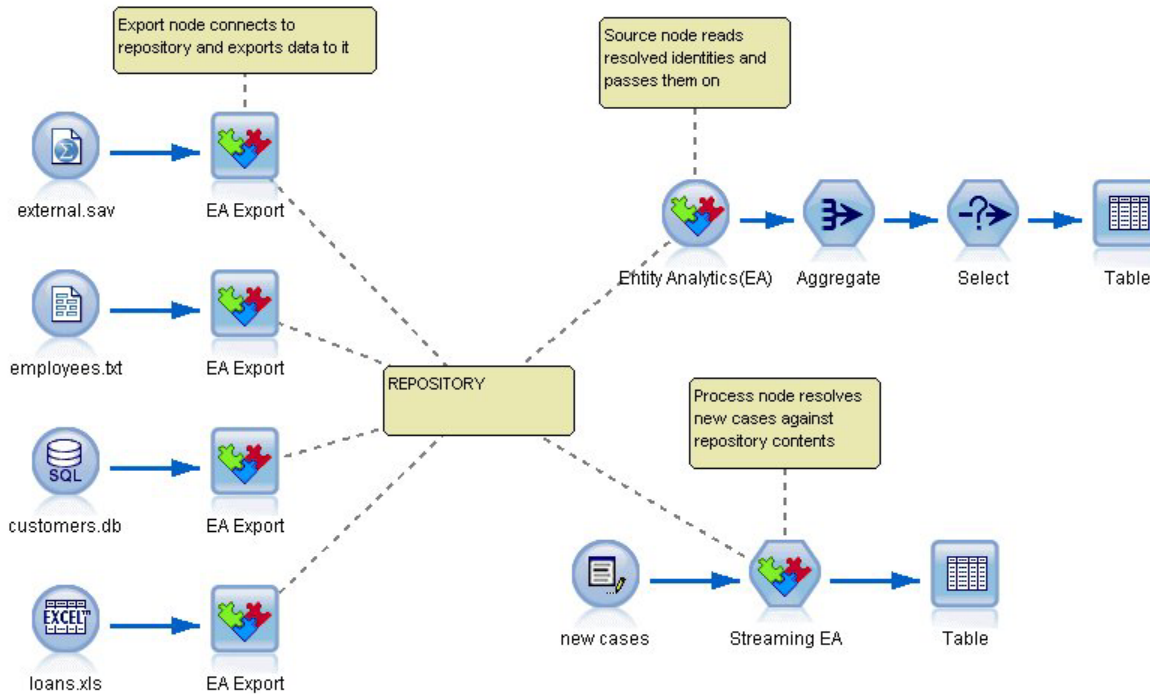


図 1. エンティティ分析プロセス

## 段階 1: ソース・データを SPSS Modeler に読み込む

まず、丸いアイコンによって SPSS Modeler で示されている 1 つ以上のソース・ノードを使用して、データを SPSS Modeler に読み込みます。

データの形式は、SPSS Modeler でサポートされる形式であれば何でも構いません (テキスト・ファイル、データベース表、スプレッドシート、XML ファイルなど)。ただし、各種の形式に対応する SPSS Modeler ソース・ノードが必要です。図では、データベース・ソース・ノードがこれに該当します。

各データ・ソース・ファイルには、各レコードを一意に識別するフィールドが 1 つ必要です。こうしたフィールドがデータ・ソース内に存在しない場合は、SPSS Modeler で簡単に追加することができます。詳しくは、14 ページの『一意のレコード識別子の追加』のトピックを参照してください。

詳しくは、14 ページの『データ・ソースへの接続』のトピックを参照してください。

注: 非ラテン文字データはサポートされていません。ラテン文字セット (例えば西ヨーロッパ) と非ラテン文字セットのレコードが混在しているデータでは、ラテン データのエントリーのみが解決されます。

## 段階 2 :リポジトリを作成する

エンティティ分析のすべての作業の中心に位置するのがリポジトリです。リポジトリは、すべてのデータ・レコードが収集される中央ストレージ領域です。

リポジトリを作成するには、最初に、正方形のアイコンで表される EA エクスポート・ノードにデータ・ソースを接続します。

このエクスポート・ノードで、新しいリポジトリを作成するか既存のリポジトリを選択します。これにより、エクスポートされたデータを受け取る準備が整います。

リポジトリ作成のプロセスについては、これ以降で詳しく説明します。詳しくは、13 ページの『エンティティ・リポジトリの設定 (EA エクスポート・ノード)』のトピックを参照してください。

**注:** リモート・サーバー・モードで実行する場合、Modeler Server マシン上にリポジトリを作成する必要があります (リポジトリを作成する場合は、Modeler Client から Modeler Server に接続する必要があります。これにより、EA リポジトリがサーバー・マシン上に作成されます)。

リポジトリを設定したら、さまざまな方法でリポジトリの内容を保守することができます。詳しくは、19 ページの『エンティティ・リポジトリの構成』のトピックを参照してください。

### 段階 3: SPSS Modeler をリポジトリに接続する

リポジトリを作成したら、そのリポジトリを SPSS Modeler ストリームに接続します。

詳しくは、16 ページの『エンティティ・リポジトリのオプション』のトピックを参照してください。

### 段階 4: 入力フィールドをリポジトリ・フィーチャーにマップする

データ・ソースには、さまざまなエンティティ情報を格納することができます。情報タイプには、大部分のエンティティ・データ・ソースに共通するものもあれば、特定のデータ・ソース固有のものもあります。エンティティ・リポジトリでは、これら各種の情報タイプを**フィーチャー**と呼びます。リポジトリにはさまざまなフィーチャーが標準で用意されていますが、独自のフィーチャーを作成することもできます。

リポジトリ・フィーチャーは、エンティティ・データ・ソースで使用できる個別の情報タイプです。さまざまなデータ・ソースで使用できるフィーチャー (姓、名、生年月日など) もあれば、特定のデータ・ソース固有のフィーチャーもあります。通常、フィーチャーは、データ・レコード内のフィールドまたはデータベース表内の列に対応します。

リポジトリを作成し、そのリポジトリに接続したら、入力データのいずれかのフィールドを**固有キー**・フィールドとして指定します。このフィールドは、後の分析で使用されます。さらに、各入力データ・フィールドをリポジトリ内の対応するフィーチャーにマップします。定義済みのフィーチャーにマッピングすることにより、エンティティ・リポジトリに対して比較対象のフィールドとその比較方法 (重要) が示されます。EA エクスポート・ノードには、マッピングを作成するためのマッピング・テーブルが用意されています。

詳しくは、17 ページの『入力フィールドとフィーチャーのマッピング (EA エクスポート・ノード)』のトピックを参照してください。

### 段階 5: データをリポジトリにエクスポートして ID を解決する

各データ・ソース・ノードには個別の EA エクスポート・ノードが必要です。そのため、データが複数のソース上に存在していると、1 つのストリームで複数のデータ・ソースが使用され、各データ・ソースが別々の EA エクスポート・ノードに接続されることになります。詳しくは、5 ページの『IBM SPSS Modeler でのエンティティ分析の使用』のトピックを参照してください。

データ・ソースが複数存在する場合は、レコードの読み取り元となるデータ・ソースを 1 つにするか、複数にするか、すべてにするかを選択できます。エンティティ分析システムは、選択されたレコードを分析し、それぞれのレコードに \$EA\_ID という ID フィールドを追加します。これで、以前は不明確だった ID

に関連するレコードを 2 件以上解決できる場合、それらのレコードに追加された識別子は、リポジトリ全体を通して一意の識別子になります。また、レコードの発生元のデータ・ソースを示すフィールドも、システムによって追加されます。

各データ・ソース・ノードを個別の EA エクスポート・ノードに接続し、入力フィールドをリポジトリ・フィーチャーとマップしたら、ストリームを実行してデータを SPSS Modeler からリポジトリにエクスポートし、ID の競合を 1 回の操作ですべて解決することができます。この仕組みを説明するために、4 つの異なるデータ・ソースに以下のレコードがあると想定します。

## 外部データ

表 2. 外部データ

名前	電話番号	信用リスク
Mike	555-1234	560
Joe	555-4567	780

## 従業員

表 3. 従業員

名前	住所	電話番号
Michael	1234 5th Street	555-1234
Fred	543 1st Avenue	555-9876

## 顧客

表 4. 顧客

名前	住所	預金
Susan	1234 5th Street	\$1234
Joe	777 Oak Street	\$5

## ローン

表 5. ローン

名前	住所	電話番号	ローン
Sue	1234 5th Street	555-1234	\$10,000
Joseph	777 Oak Street	555-4567	\$50,000

これまでに説明したように、各データ・ソースをリポジトリにエクスポートします。このエクスポートにより、リポジトリ内で各レコードの解決が更新されます。リポジトリ内では、各レコードの先頭に ID フィールド (*\$EA-ID*) とソース・インジケター・フィールド (*\$EA-SRC*) が付加されます。ソース・インジケター・フィールドは、レコードの発生元のデータ・ソースを示します。したがって、この例の 4 つのデータ・ソースをすべてエクスポートすると、リポジトリの内容は次のようになります。

表 6. エクスポート後のリポジトリの内容の例：

<i>\$EA-ID</i>	<i>\$EA-SRC</i>	名前	電話番号	住所	信用リスク	預金	ローン
1	従業員	Michael	555-1234	1234 5th St			
1	外部	Mike	555-1234		560		

表 6. エクスポート後のリポジトリの内容の例 (続き):

\$EA-ID	\$EA-SRC	名前	電話番号	住所	信用リスク	預金	ローン
2	顧客	Joe		777 Oak St		\$5	
2	外部	Joe	555-4567		780		
2	ローン	Joseph	555-4567	777 Oak St			\$50,000
3	従業員	Fred	555-9876	543 1st Ave			
4	顧客	Susan		1234 5th St		\$1234	
4	ローン	Sue	555-1234	1234 5th St			\$10,000

エンティティ分析システムは、電話番号が共通していることと、割り当てられた ID がどちらも 1 であることを根拠として、外部 データ・セットの *Mike* と従業員 データ・セットの *Michael* は同一人物であると判断しました。

外部 データ・セットの *Joe* の場合は、少し複雑になります。この *Joe* は、顧客の *Joe* と同一人物でしょうか。これら 2 つのデータ・ソースだけでは判断できませんが、第 3 のソースのローンには *Joseph* が含まれています。ここで一致が見つかります。*Joseph* の電話番号は、外部 データ・セットの *Joe* の電話番号と同じです。これに基づき、システムはこれらの人物を同一人物であると判断し、ID 2 を割り当てます。

*Fred* についてはレコードが 1 件しかないため、ID 3 が割り当てられます。顧客の *Susan* とローンの *Sue* は住所が同じであるため同一人物と判断され、ID 4 が割り当てられます。

注: これは、説明をわかりやすくするための簡単なマッチングの例です。単純な名前や電話番号や住所だけでは完全一致にならないような、より複雑なルール・セットを選択し、両方のレコードに同じ識別子を割り当てることもできます。

## 段階 6: 解決済み ID を分析する

リポジトリ内で ID の競合を解決したら、さらに分析を行い、その結果を処理することができます。例えば、同じ ID について重複するレコードが存在し、不正な行為が疑われる場合は、重複をリストするレポートを作成することができます。

最初に、エンティティ分析 (EA) ソース・ノードを作成してリポジトリにリンクします。

このノードからの共通出力は、以下のフィールドから構成されます。

- システムによって追加された ID フィールド (段階 5 の例の *\$EA-ID*)
- システムによって追加されたソース・インジケター・フィールド (段階 5 の例の *\$EA-SRC*)
- 段階 4 で指定した固有キー・フィールド

また、関係性を調べる場合は、以下の出力が生成されます。詳しくは、27 ページの『データ・ソースの選択』のトピックを参照してください。

- エンティティ間の分離度 (*\$EA-DEGREE*)
- 親フィールド (*\$EA-PARENT*)
- 子フィールド (*\$EA-CHILD*)
- 関係性を識別するルール (*\$EA-RULE*)



SPSS Modeler で出力内容を表示するには、SPSS Modeler の出力ノード (テーブル・ノードやレポート・ノードなど) を接続して、ストリームのこの部分を実行します。出力のサイズは、非常に大きくなる場合があります。出力内容を要約する必要がある場合は、レコード操作ノード (集計ノードや条件抽出ノードなど) を組み込んでください。

エンティティ分析 (EA) ソース・ノードについては、これ以降で詳しく説明します。詳しくは、27 ページの『解決済み ID の分析 (エンティティ分析 (EA) ソース・ノード)』のトピックを参照してください。

## 段階 7: リポジトリに対して新規ケースを解決する

これまでの処理で、すべてのデータ・ソース内のすべてのレコードの ID が解決されました。しかし、スコアリングを改善するために、新しいレコードのセットを比較して、それらのレコードが既知の情報とどのように関係するかを確認したい場合はどうしたらよいでしょうか。このような場合は、ストリーミング EA ノードを使用します。

最初に、新規データをストリームに読み込むための、新しい SPSS Modeler データ・ソース・ノードを追加します。次に、このソース・ノードをストリーミング EA ノードに接続します。前述のとおり、出力内容を表示するにはテーブル・ノードを追加します。

ストリームのこの部分を実行すると、ストリーミング EA ノードが各新規レコードを読み取り、リポジトリの内容に対して比較を行います。ストリーミング EA ノードがリポジトリ内で一致レコードを検出した場合は、新規レコードとともに一致レコードがすべて出力され、新規レコードに ID フィールドとソース・インジケータ・フィールドが追加されます。一致が検出されなかった場合、プロセス・ノードは、ID フィールドとソース・インジケータ・フィールドが追加された新規レコードだけを出力します。

これを説明するために、リポジトリが現在、エンティティ分析 (EA) ソース・ノードによって出力された内容で構成されていると想定します。8 ページの表 6 を参照してください。

ここで、以下の新規レコードを受け取ったとします。これらのレコードは、既知の人物と関連があるでしょうか。

表 7. スコアリングされる新規レコード:

名前	住所	電話番号	ローン
Suzan	1234 5th Street	555-1234	\$100,000
Mark	888 9th Ave	555-9999	\$60,000

ストリーミング EA ノードは、新規データを既存のリポジトリ内容と比較して、既存のレコードで ID 4 が割り当てられている人物と最初の新規レコードをマッチングします。一方、2 番目の新規レコードでは一致が検出されないため、新しい固有 ID 5 が割り当てられます。

ストリーミング EA ノードは ID フィールドとソース・インジケータ・フィールドを追加し、新規レコードとともにそれらの一致レコードをすべて出力します。したがって、出力は以下のようになります。

表 8. ストリーミング EA ノードからの出力:

\$EA-ID	\$EA-SRC	名前	電話番号	住所	信用リスク	預金	ローン
4	顧客	Susan		1234 5th St		\$1234	
4	ローン	Sue	555-1234	1234 5th St			\$10,000
4	新規ローン	Suzan	555-1234	1234 5th Street			\$100,000



表 8. ストリーミング EA ノードからの出力 (続き):

\$EA-ID	\$EA-SRC	名前	電話番号	住所	信用リスク	預金	ローン
5	新規ローン	Mark	555-9999	888 9th Ave			\$60,000

次に、エンティティ分析 ID を集計キーとして使用してこの出力内容を集計し、他の下流ノードに渡してさらに処理を行うことができます。

ストリーミング EA ノードについては、これ以降で詳しく説明します。

## 段階 8: 警告を生成する

ここで再び、疑わしいアクティビティが判明したとします。今度は、既に 10,000 ドルのローンがある ID 4 の人物が、名前を少し変えて、10 倍の金額のローンを新たに申請しています。もちろん、不正な意図で行われたものではない可能性もあり、その場合はまったく問題ありません。ただし、ビジネス・ルールに従ってそのようなアクティビティが疑わしいと考えられる場合は、調査する必要があります。

例えば、SPSS Modeler のテーブル・ノードやレポート・ノードを接続して実行し、その出力ウィンドウの内容を印刷して誰かに読んでもらい、手動で警告を生成することができます。あるいは、ストリーミング EA ノードの出力内容を、IBM SPSS Modeler で以前に作成したリスク評価モデルに渡して、ビジネス・ルールをより正確に反映するスコアのセットを生成することもできます。また、出力内容をデータベースなどの媒体にエクスポートして、さらに処理を行うこともできます。IBM SPSS Modeler では、特定の要件に合わせてさまざまな選択肢からアクションを選択することができます。



---

## 第 3 章 エンティティ分析のタスク

---

### タスクについて

このセクションでは、以下のエンティティ分析タスクについて説明します。

- エンティティ・リポジトリの設定
- エンティティ・リポジトリの構成
- 解決済み ID の分析
- エンティティ・リポジトリに対する新規ケースの解決
- エンティティ・リポジトリの消去
- エンティティ・リポジトリの削除
- 他の IBM SPSS 製品でのエンティティ分析の使用
- エンティティ分析の管理

---

### エンティティ・リポジトリの設定 (EA エクスポート・ノード)

エンティティ・リポジトリの設定プロセスでは、以下のタスクを実行します。

1. データ・ソースに接続します。詳しくは、14 ページの『データ・ソースへの接続』のトピックを参照してください。
2. リポジトリを作成します。詳しくは、15 ページの『リポジトリの作成』のトピックを参照してください。
3. データ・ソース内の入力フィールドをリポジトリ内のフィーチャーにマップします。詳しくは、17 ページの『入力フィールドとフィーチャーのマッピング (EA エクスポート・ノード)』のトピックを参照してください。

マッピングを設定したら、現在のデータ・ソース、またはリポジトリで認識されているすべてのデータ・ソースについて、マッピングを表示できるようになります。詳しくは、19 ページの『フィールド・マッピングの表示 (EA エクスポート・ノード)』のトピックを参照してください。

**注:** バージョン 16 以降、SPSS Entity Analytics では、IBM DB2 製品上のリポジトリがサポートされるようになりました。リポジトリは SPSS Modeler のバージョンに固有で、以前のバージョンからインポートすることはできないため、既存のリポジトリを SPSS Entity Analytics バージョン 16 にアップグレードする場合は、新しい DB2 データベースで既存のリポジトリを再作成する必要があります。

### エンティティ・リポジトリ

リポジトリは、すべてのエンティティ情報のデータ・キャッシュとして機能する中央ストレージ領域を提供します。リポジトリは常にアクティブになっているため、リポジトリの状態は 1 つだけです。したがって、エンティティ・リポジトリにはバージョン管理という概念がありません。リポジトリにはすべての入力データの現在の状態が保持されるため、サイズが非常に大きくなる場合があります。

使いやすいグラフィカル・インターフェースにより、リポジトリの内容を保守することができます。詳しくは、19 ページの『エンティティ・リポジトリの構成』のトピックを参照してください。

**重要:** バージョン 16 以降、IBM SPSS Modeler Entity Analytics では、IBM DB2 製品上のリポジトリーがサポートされています。以前のバージョンの SPSS Entity Analytics では、IBM solidDB でホストされるリポジトリーがサポートされていました。既存の solidDB リポジトリーがある場合、SPSS Entity Analytics バージョン 16 以降にアップグレードするときは、新規 DB2 データベースにそのリポジトリーを再作成する必要があります。

**注:** IBM SPSS Modeler Premium に付属しているバージョンの IBM SPSS Modeler Entity Analytics では、SPSS Entity Analytics にバンドルされた IBM DB2 製品上でホストされる単一のリポジトリーのみをサポートしています。このバージョンを使用する場合は、新しいリポジトリーを作成する前に既存のリポジトリーを削除する必要があります。個別ライセンスによる SPSS Entity Analytics へのアップグレード (IBM SPSS Modeler Entity Analytics Unleashed と呼ばれます) を使用すると、複数のリポジトリーを同一システム上に共存させることができます。各リポジトリーは 1,000 万を超える行を格納でき、5 つ以上のプロセッサ・コアを使用することができます。詳しくは、IBM サービス担当員にお問い合わせください。

## データ・ソースへの接続

最初に、ソース・ノードを使用してソース・データを SPSS Modeler に読み込みます。

データ・ソースに接続するには、以下の手順を実行します。

1. SPSS Modeler のメイン・メニューの下部にあるノード・パレットの「ソース」タブから、ソース・データのタイプに対応するアイコンをダブルクリックします。このアイコンをダブルクリックすると、ソース・ノードがスクリーン領域に追加されます。
2. スクリーン領域でアイコンをダブルクリックして、ダイアログ・ボックスを開きます。
3. 「ファイル」フィールドで、ソース・データ・ファイルの場所と名前を入力します。
4. 必要に応じてダイアログ・ボックスの残りを入力して (詳細については「ヘルプ」ボタンをクリック)、「OK」をクリックします。
5. 各レコードを一意に識別するフィールドがソース・データ・ファイル内に存在しない場合は、フィールド作成ノードを使用して、各レコードを一意に識別するフィールドを追加します。詳しくは、『一意のレコード識別子の追加』のトピックを参照してください。

**注:** 非ラテン文字データはサポートされていません。ラテン文字セット (例えば西ヨーロッパ) と非ラテン文字セットのレコードが混在しているデータでは、ラテン データのエントリーのみが解決されます。

### 一意のレコード識別子の追加

エンティティ・リポジトリーに対する入力となる各データ・ソース・ファイルには、各レコードを一意に識別する 1 つのフィールドが必要です。データ・ソース・ファイル内にこのようなフィールドが存在しない場合は、SPSS Modeler のフィールド作成ノードを使用して、このフィールドを追加することができます。

一意のレコード識別子をデータ・ソース・ファイルに追加するには、以下の手順を実行します。

1. スクリーン領域で、前述のタスクで追加したソース・ノードをクリックします。
2. ノード・パレットの「フィールド設定」タブで「フィールド作成」アイコンをダブルクリックし、フィールド作成ノードをソース・ノードに接続します。
3. スクリーン領域でフィールド作成ノードをダブルクリックして、ダイアログ・ボックスを開きます。
4. 「フィールド作成」フィールドで、追加する ID フィールドのデフォルト名を分かりやすい名前 (「ID」など) に置き換えます。
5. 「データ型」フィールドが「CLEM 式」に設定されていることを確認します。

6. 「フィールドのデータ型」を「連続型」に設定します。
7. 「CLEM 式」テキスト・ボックスに「@INDEX」と入力して「OK」をクリックします。

## リポジトリの作成

すべての入力データを格納するリポジトリを作成する必要があります。

注: リモート・サーバー・モードで実行する場合、Modeler Server マシン上にリポジトリを作成する必要があります (リポジトリを作成する場合は、Modeler Client から Modeler Server に接続する必要があります。これにより、EA リポジトリがサーバー・マシン上に作成されます)。

リポジトリを作成するには、以下の手順を実行します。

1. SPSS Modeler のノード・パレットの「エクスポート」タブで、ストリーム領域に EA エクスポート・ノードを配置します。

注: リポジトリを初めて作成する場合は、EA エクスポート・ノードを使用し、そのリポジトリに入力したいデータを持つ SPSS Modeler ソース・ノード (または、固有の ID フィールドを取得するためにフィールド作成ノードを追加した場合は、そのフィールド作成ノード) に接続します。ノードに接続するには、以下の手順を実行します。

- a. SPSS Modeler ソース・ノードを右クリックします。
  - b. 「接続」を選択します。
  - c. EA エクスポート・ノードをクリックします。
2. EA エクスポート・ノードをダブルクリックしてダイアログ・ボックスを開きます。
  3. 「エンティティ リポジトリ」リストをクリックします。
  4. 「<参照...>」をクリックして「エンティティ リポジトリ」ダイアログ・ボックスを表示します。
  5. 「エンティティ リポジトリ」ダイアログ・ボックスで、「リポジトリ名」フィールドをクリックします。
  6. 「<新規リポジトリを作成...>」を選択して「リポジトリの作成」ウィザードを表示します。

### 「リポジトリの作成」ウィザード

#### ステップ 1

ここでは、IBM SPSS Modeler Entity Analytics とともにバンドルされている IBM DB2 製品を使用してローカル・リポジトリを作成するか、リポジトリ用に外部データベースを使用するかを選択します。

**ローカル・リポジトリの作成:** 作成対象のリポジトリをホストする IBM DB2 データベースの管理者ユーザー名とパスワードの詳細を指定します。パスワードを確認して「次へ」をクリックします。

注: ユーザー名にダッシュ記号やアンダースコア記号を使用することはできません。

IBM DB2 データベースに使用する必要がある資格情報は、ご使用のオペレーティング・システムによって異なります。UNIX ユーザーは、ユーザー名 g2user とパスワード G2password を使用する必要があります。

エンティティ分析ノード内のリポジトリ管理タスク (リポジトリの作成または破棄など) には、追加の権限が必要です。UNIX では、IBM SPSS Modeler Server にログインするユーザーは、root ユーザーまたは g2user のいずれかでなければならず、さらに db2iadm1 グループのメンバーでなければなりません。Windows では、IBM SPSS Modeler Server にログインするユーザーは、リポジトリ管理を実行するためには DB2ADMNS グループのメンバーでなければなりません。

後で管理者資格情報の変更が必要になった場合は、データベースのコマンド・ライン・エディターを使用して変更してください。詳しくは、34 ページの『リポジトリ・データベースの管理者資格情報の管理』のトピックを参照してください。

注: 使用できるユーザー名とパスワードの組み合わせは 1 つだけです。リポジトリにログインするすべてのユーザーが、同じユーザー名とパスワードを共有することになります。

**外部リポジトリの追加:** 外部データベースを使用してリポジトリをホストする場合は、このオプションを使用します。「リポジトリの .ini ファイルを選択」にデータベースの .ini ファイルの場所を入力して「次へ」をクリックします。

## ステップ 2

**新しいリポジトリ名:** 新しいリポジトリの固有名を入力します。

**構成のインポート元:** (ローカル・リポジトリのみ) 既存のリポジトリの構成をベースにしたい場合は、ここでそのリポジトリを選択します。それ以外の場合は、「デフォルト」を選択します。詳しくは、19 ページの『エンティティ・リポジトリの構成』のトピックを参照してください。

既存のリポジトリを選択し、接続の詳細が前の画面で入力したものと異なる場合は、接続の詳細を入力してください。

「OK」をクリックして新しいリポジトリを作成し、「エンティティ解決インスタンス」ダイアログ・ボックスを表示します。このダイアログ・ボックスを使用して、リポジトリに接続することができます。

## エンティティ・リポジトリのオプション

「エンティティ・リポジトリ」ダイアログボックスには、エンティティ・リポジトリの作成、接続、構成、保守に関するさまざまなオプションが表示されます。

**リポジトリに接続:** 新しいエンティティ・リポジトリを作成する場合や、既存のエンティティ・リポジトリに接続する場合は、以下のオプションを使用します。

- **リポジトリ名:** 現在のエンティティ・リポジトリが表示されます (存在する場合)。リポジトリが複数存在する場合に別のリポジトリを選択するには、必要なリポジトリをリストから選択します。

新規のリポジトリを作成する場合は、「<新しいリポジトリを作成...>」を選択します。この操作により、リポジトリの作成手順をガイドするウィザードが起動します。

- **ユーザー名:** 選択したリポジトリの有効なユーザー名を入力します。
- **パスワード:** このユーザー名のパスワードを入力します。
- **接続:** 現在のリポジトリに接続する場合にクリックします。

**リポジトリの管理.** この表には、現在のリポジトリ (接続先のリポジトリ) 内にロード済みのデータ・ソースがリストされ、各データ・ソース内のレコード数が表示されます。

- **リフレッシュ:** 新しいデータ・ソースを追加した場合や、既存のデータ・ソースのサイズを変更した場合などに、表内のデータ・ソース情報とサイズ情報を更新します。
- **すべて消去.** リポジトリからすべてのソース・データを削除します。ただし、構成の詳細はすべて保持されます。構成情報はまだ必要だが、リポジトリ内のデータ・レコードはすべて削除したいという場合は、このオプションを使用してください。詳しくは、37 ページの『エンティティ・リポジトリの消去』のトピックを参照してください。



- **未使用分を削除。**リポジトリから強調表示されたデータ・ソースを削除します。ただし、構成の詳細はすべて保持されます。詳しくは、37 ページの『リポジトリからの未使用データ・ソースの削除』のトピックを参照してください。
- **ソースの名前変更。**強調表示されたデータ・ソースの名前を変更できるダイアログ・ボックスが開きます。

注: これにより、リポジトリ内のデータ・ソースの名前が変更されます。参照元の既存のすべてのエクスポートまたはストリーミング・ノードで、この新規データ・ソース名を再選択する必要があります。

**リポジトリ全体を破棄。**現在のリポジトリの内容と構成の詳細を完全に破棄します。詳しくは、38 ページの『エンティティ・リポジトリの削除』のトピックを参照してください。

**リポジトリの設定:** 現在のリポジトリを構成するためのウィンドウを表示します。詳しくは、19 ページの『エンティティ・リポジトリの構成』のトピックを参照してください。

## 入力フィールドとフィーチャーのマッピング (EA エクスポート・ノード)

リポジトリには、さまざまな定義済みのフィーチャーが標準で用意されています。同じフィーチャーに対応する情報タイプの場合、異なるデータ・ソースで異なるフィールド名が使用される可能性があります (例えば、**Address1** や **Address Line 1** など)。重複を避けるには、入力データ・ソースのフィールドを特定のリポジトリ・フィーチャーにマップする必要があります。データ・セット内のすべてのフィールドをマップする必要はありません。他のデータ・セット内の同じフィーチャーに対応すると思われるフィールドだけをマップしてください。

データ・ソースで使用されているフィールドが、リポジトリで事前に定義されていない他の情報タイプに対応する場合は、「リポジトリ構成」ウィンドウから新しいフィーチャーを作成することができます。詳しくは、19 ページの『エンティティ・リポジトリの構成』のトピックを参照してください。

入力フィールドをフィーチャーにマップするには、以下の手順を実行します。

1. EA エクスポート・ノードをストリーム領域上のデータ・ソース・ノードに接続します。使用する各データ・ソース・ノードを、独自の EA エクスポート・ノードに接続する必要があります。
2. EA エクスポート・ノードを開いて「入力」タブを表示します。このタブには、入力フィールドをマップするためのオプションが表示されます。詳しくは、『マッピングに関するリポジトリ入力オプション』のトピックを参照してください。
3. EA エクスポート・ノードで「リポジトリ」タブを選択し、現在のデータ・ソースのマッピング割り当てを表示します。複数のデータ・ソースを使用している場合は、すべてのデータ・ソースのマッピング割り当てが表示されます。
4. 一連のマッピングの割り当てを保存するには (別のデータ・ソースのエクスポート・ノードで使用する場合など)、「**マッピングのエクスポート**」をクリックします。

最初のデータ・ソース・ノードのマッピングが終了したら、使用する他のすべてのデータ・ソース・ノードに対してこのプロセスを繰り返します。

## マッピングに関するリポジトリ入力オプション

「入力」タブには、リポジトリへのエクスポート準備が整っているリポジトリ・フィーチャーにデータ・ソース・フィールドをマップするためのオプションが表示されます。このタブでマッピング割り当てを設定し、オプションで「リポジトリ」タブをクリックして他のデータ・ソースのマッピングを確認したら、「**実行**」をクリックしてデータをリポジトリにエクスポートします。

マッピング・セットが既に XML ファイルに保管されている場合は、「マッピングのインポート」をクリックしてそれらのマッピングを使用することができます。

**モード:** ソース・ファイルのレコードをリポジトリの既存の内容に追加する場合は、デフォルトの選択値である「リポジトリに追加」のままにします。ソース・レコードを追加する前に、リポジトリの内容は消去し、構成情報は保持したい場合は、「エクスポートの前にリポジトリを消去」を選択します。

**エンティティ・リポジトリ:** 現在のエンティティ・リポジトリが表示されます (存在する場合)。リポジトリが複数存在する場合に別のリポジトリを選択するには、必要なリポジトリをリストから選択します。新しいリポジトリを作成するには、「<参照...>」を選択し、リポジトリを作成するためのダイアログ・ボックスを表示します。詳しくは、16 ページの『エンティティ・リポジトリのオプション』のトピックを参照してください。

**エンティティ・タイプにマップ:** リポジトリ内で定義されたエンティティ・タイプのリスト (つまりフィーチャーのセット)。リストからいずれかのタイプを選択するか、「<新しいエンティティ・タイプの追加...>」を選択してリポジトリ構成ウィンドウを表示します。このウィンドウで、新しいエンティティ・タイプを定義することができます。詳しくは、19 ページの『エンティティ・リポジトリの構成』のトピックを参照してください。

**ソース・タグ:** リポジトリが現在認識しているデータ・ソースを示すタグのリスト。リストからいずれかのタグを選択するか、「<新しいソース・タグの追加...>」を選択して新しいデータ・ソースのタグを作成します。

**固有キー:** (必須) データ・レコードの一意の識別子に使用する入力フィールド。

**マッピング・テーブル:** このテーブルでは、各入力フィールドを、リポジトリ内の対応するフィーチャーにマップすることができます。選択したエンティティ・タイプ内に適切なフィーチャーが存在しない場合は、ここで新しいフィーチャーを作成することができます。

- **フィールド:** 選択したデータ・ソース内の入力フィールドのセット。各フィールドには、フィールドの測定の尺度 (つまりデータ型) を示すアイコンがあります。
- **フィーチャーにマップ:** フィールドをフィーチャーにマップするには、フィールド行でこの列をダブルクリックして (またはスペース・バーを押して)、リストからフィーチャーを選択します。適切なフィーチャーがない場合は、「<新しいフィーチャーの追加...>」を選択してリポジトリ構成ウィンドウを表示します。このウィンドウで、このエンティティ・タイプ用の新しいフィーチャーを定義することができます。詳しくは、19 ページの『エンティティ・リポジトリの構成』のトピックを参照してください。
- **使用法:** 複数のコンテキスト (自宅の電話番号と会社の電話番号など) が考えられる特定のフィールドのコンテキストを示します。ADDRESS フィーチャーと PHONE フィーチャーには、複数の使用法タイプが事前に設定されています。また、すべてのフィーチャーについて、独自の使用法タイプを作成することができます。デフォルトの「(自動)」以外の使用法を設定するには、対象となる行でこの列をクリックし、いずれかの既存の使用法タイプを選択するか (存在する場合)、「<使用法の追加...>」をクリックして新しい使用法タイプを作成します。詳しくは、24 ページの『エンティティ・タイプの保守』のトピックを参照してください。

**マッピングのインポート:** 以前にエクスポートしたフィールド対フィーチャーのマッピング・セットを外部の XML ファイルからインポートします。このオプションを選択すると、異なるソースごとに同じマッピングを定義し直す必要がなくなるため、同じマッピング要件を持つデータ・ソースが複数存在する場合に便利です。



マッピングのエクスポート: マッピング・テーブルに表示されたフィールド対フィーチャーのマッピング・セットを外部の XML ファイルにエクスポートします。

## フィールド・マッピングの表示 (EA エクスポート・ノード)

「リポジトリ」タブで「リフレッシュ」ボタンをクリックすると、入力フィールドがマップされているリポジトリ・フィーチャーを表示することができます。これは、現在のデータ・ソース (このエクスポート・ノードに接続されたソース・ノードによって制御されるデータ・ソース) について表示することも、すべてのデータ・ソースについて表示することもできます。

**入力の表示対象:** 現在のデータ・ソースのマッピングを表示するか、またはリポジトリで認識されているすべてのデータ・ソースのマッピングを表示するかを選択します。

**リフレッシュ:** 選択された入力オプションの表示内容を更新します。

**フィーチャー:** 表示されたデータ・ソース内にマッピングが存在するフィーチャーをすべてリストします。マップされていないフィーチャーは表示されません。

**<データ・ソース>:** マッピングが定義されているフィーチャーごとに、特定のデータ・ソース内のマップ・フィールドを各列にリストします。

---

## エンティティ・リポジトリの構成

「リポジトリ構成」ウィンドウから、リポジトリの内容を保守することができます。このウィンドウには、リポジトリ全般に関する使いやすい視覚的なインターフェースが用意されています。

同じまたは類似した構成のリポジトリを複数使用する場合は、基本構成を設定してファイルにエクスポートし、他のリポジトリにインポートすることができます。詳しくは、26 ページの『リポジトリ構成の再利用』のトピックを参照してください。

**注:** バージョン 16 以降、SPSS Entity Analytics では、IBM DB2 製品上のリポジトリがサポートされるようになりました。リポジトリは SPSS Modeler のバージョンに固有で、以前のバージョンからインポートすることはできないため、既存のリポジトリを SPSS Entity Analytics バージョン 16 にアップグレードする場合は、新しい DB2 データベースで既存のリポジトリを再作成する必要があります。

### 注意:

既にデータが存在するリポジトリの構成を変更して保存する場合、そのリポジトリの内容を消去してデータを再ロードするよう求めるプロンプトが出されることがあります。これにより、リポジトリが不整合な状態になるのを防ぐことができます。

リポジトリ構成を設定するには、以下の手順を実行します。

1. 任意のエンティティ分析ノードを開きます。
2. 「エンティティ リポジトリ」リストをクリックします。
3. 「<参照...>」をクリックして「エンティティ解決インスタンス」ダイアログ・ボックスを表示します。
4. 「エンティティ解決インスタンス」ダイアログ・ボックスで、「リポジトリ名」リストをクリックします。
5. 構成を設定したいリポジトリを選択します。
6. まだ接続していない場合は、管理者ユーザー名とパスワードを入力して「接続」をクリックします。

7. 「リポジトリの設定」ボタンをクリックして「リポジトリ設定」ウィンドウを表示します (このボタンが使用可能になっている場合)。
8. 以下のセクションの説明に従い、構成の詳細を作成します。

「リポジトリ設定」ウィンドウの左側のナビゲーション・ペインには、ツリー構造が表示されます。このツリー構造を使用して、リポジトリのさまざまな特性を管理することができます。

表9. 「リポジトリ設定」ウィンドウの主な要素：

セクション	説明	
データ・ソース	すべてのデータ・ソースと各種リポジトリ・フィーチャーのマッピングを表示します。	詳しくは、『データ・ソース・マッピングの表示』のトピックを参照してください。
フィーチャー	新規フィーチャーの作成、または既存のフィーチャーの複製、編集、削除を行います。	詳しくは、『リポジトリ・フィーチャーの保守』のトピックを参照してください。
エンティティ・タイプ	新規エンティティ・タイプの作成、または既存のエンティティ・タイプの管理 (複製、名前変更、フィーチャーの関連付けまたは解除、削除) を行います。	詳しくは、24 ページの『エンティティ・タイプの保守』のトピックを参照してください。
解決ルール	エンティティの一致用のしきい値を設定します。	詳しくは、26 ページの『エンティティの一致に対するしきい値の設定』のトピックを参照してください。

## データ・ソース・マッピングの表示

「リポジトリ構成」ウィンドウの「データ・ソース」セクションの「すべてのソース」エントリには、すべてのデータ・ソースと各種リポジトリ・フィーチャーとのマッピングが読み取り専用で表示されます。

新しいデータ・ソースがリポジトリに追加された場合は、「リフレッシュ」をクリックしてリストを更新します。

注: ここでデータ・ソースをリポジトリに追加することはできません。データ・ソースを追加するには、SPSS Modeler のソース・ノードを作成してエンティティ分析エクスポート・ノードに接続する必要があります。詳しくは、14 ページの『データ・ソースへの接続』のトピックを参照してください。

## リポジトリ・フィーチャーの保守

リポジトリ・フィーチャーは、エンティティ・データ・ソースで使用できる個別の情報タイプです。さまざまなデータ・ソースで使用できるフィーチャー (姓、名、生年月日など) もあれば、特定のデータ・ソース固有のフィーチャーもあります。1 つのフィーチャーには、1 つ以上の要素を含めることができます。通常、各要素は、データ・レコード内のフィールドまたはデータベース表内の列に対応します。

「リポジトリ構成」ウィンドウの「フィーチャー」セクションの「すべてのフィーチャー」エントリを使用すると、すべてのリポジトリ・フィーチャーを保守することができます。以下の操作を実行することができます。

- 新しいフィーチャーを作成する。
- 既存のフィーチャーを複製する (既存のフィーチャーに基づいて新しいフィーチャーを作成する場合など)。

- 既存のフィーチャーを編集する。
- 既存のフィーチャーを削除する。

これらのタスクの実行手順は、このセクションで後述します。

フィーチャー・リストには、このリポジトリで定義されているすべてのフィーチャーが表示されます。リストの列には、フィーチャーに設定できる各種プロパティが表示されます。

**フィーチャー:** フィーチャーの名前。フィーチャー名の横に南京錠のシンボルが表示されている場合、そのフィーチャーはロックされています。ロックされているフィーチャーは、削除も複製もできません。それらに対する変更内容で保存できるのは、匿名化属性を変更した場合のみです。

**頻度:** このフィーチャーについて、同じ値を持つことができるエンティティの数を示します。有効な値は、「**1**」(例: パスポート番号)、「**少数**」(例: 住所)、「**多数**」(例: 誕生日)です。

**排他性:** 1 つのエンティティが当該タイプのフィーチャーを 1 つしか持たないことを指定します。例えば、生年月日や保険証番号の場合は、この値に「**はい**」を指定します。一方、住所やクレジット・カード番号の場合は、「**いいえ**」を指定します (1 つのエンティティが複数の住所やクレジット・カードを持つ場合もあるため)。

**安定性:** 当該フィーチャーの安定性の値 (つまり、エンティティの存続期間中に、当該フィーチャーが変わりそうにないかどうか) を示します。例えば、生年月日フィーチャーの場合は決して変わらないため、値に「**はい**」を指定します。一方、住所フィーチャーの場合は変わりやすい (したがって安定性が低い) ため、「**いいえ**」を指定します。注: 性別は、通常、存続期間を通して安定していますが、データ上の問題が原因で誤って指定されることがよくあるため、デフォルト構成では値が「**いいえ**」になっています。

**匿名化.** フィーチャーが匿名化されているかどうかを示します。入力項目は「**はい**」または「**いいえ**」のいずれかです。詳しくは、23 ページの『リポジトリ・フィーチャーの匿名化』のトピックを参照してください。

新しいフィーチャーを作成するには、以下の手順を実行します。

1. 以下のいずれかの操作を実行します。
  - 「新規フィーチャーの作成」ボタン (画面右側の一番上のボタン) をクリックします。
  - 画面左側のナビゲーション・ペインで「**すべてのフィーチャー**」を右クリックし、「**新規フィーチャー**」を選択します。
2. 「フィーチャーの追加/編集」ダイアログ・ボックスに入力します。詳しくは、22 ページの『フィーチャーの追加または編集』のトピックを参照してください。

既存のフィーチャーを複製するには、以下の手順を実行します。

1. 画面右側の表の「**フィーチャー**」列で、複製したいフィーチャーを選択します。
2. 「選択したフィーチャーの複製」ボタン (画面右側の 2 番目のボタン) をクリックします。
3. 「フィーチャーの追加/編集」ダイアログ・ボックスに入力します。詳しくは、22 ページの『フィーチャーの追加または編集』のトピックを参照してください。

既存のフィーチャーを編集するには、以下の手順を実行します。

**注意** リポジトリ内に既にデータが格納されている状態でフィーチャーまたはフィーチャー要素の編集、削除、匿名化を行う場合は、リポジトリの内容を消去してからデータを再ロードしてください。これにより、リポジトリが不整合な状態になるのを防ぐことができます。

1. 画面右側の表の「フィーチャー」列で、編集したいフィーチャーを選択します。注: 編集できるのは、自分で作成したフィーチャーだけです。システムが提供するフィーチャーを編集することはできません。
2. 「選択したフィーチャーの編集」ボタン (画面右側の 3 番目のボタン) をクリックします。
3. 「フィーチャーの追加/編集」ダイアログ・ボックスに入力します。詳しくは、『フィーチャーの追加または編集』のトピックを参照してください。

既存のフィーチャーを削除するには、以下の手順を実行します。

**注意** リポジトリ内に既にデータが格納されている状態でフィーチャーまたはフィーチャー要素の編集、削除、匿名化を行う場合は、リポジトリの内容を消去してからデータを再ロードしてください。これにより、リポジトリが不整合な状態になるのを防ぐことができます。

1. 画面右側の表の「フィーチャー」列で、削除したいフィーチャーを選択します。注: 削除できるのは、自分で作成したフィーチャーだけです。システムが提供するフィーチャーを削除することはできません。
2. 以下のいずれかの操作を実行します。
  - 「選択したフィーチャーの削除」ボタン (画面右側の一番下のボタン) をクリックします。
  - 画面左側のナビゲーション・ペインで「すべてのフィーチャー」を右クリックし、「削除」を選択します。
3. 「続行」をクリックして、フィーチャーの削除を確認します。

**注意:**

フィーチャーの削除を取り消すことはできません。

## フィーチャーの追加または編集

**注意** リポジトリ内に既にデータが格納されている状態でフィーチャーまたはフィーチャー要素の編集、削除、匿名化を行う場合は、リポジトリの内容を消去してからデータを再ロードしてください。これにより、リポジトリが不整合な状態になるのを防ぐことができます。

「フィーチャーの追加/編集」ダイアログ・ボックスで、新しいリポジトリ・フィーチャーを作成したり、既存のフィーチャーの複製や編集を行うことができます。

注: 既存のフィーチャーがロックされている場合、このダイアログ・ボックスでその詳細を編集することはできません。

**フィーチャー・タイプ:** このフィーチャーに関する情報のタイプを示すラベル。このラベルは、フィーチャー識別子の最初の部分となります。

**説明:** フィーチャー・タイプについて簡単に説明するテキスト (情報を提供する目的でのみ使用します)。

**頻度:** このフィーチャーについて、同じ値を持つことができるエンティティの数を示します。有効な値は、「1」 (例: パスポート番号)、「少数」 (例: 住所)、「多数」 (例: 誕生日) です。

**排他性:** 1 つのエンティティが当該タイプのフィーチャーを 1 つしか持たないことを指定します。例えば、生年月日や保険証番号の場合は、この値に「はい」を指定します。一方、住所やクレジット・カード番号の場合は、「いいえ」を指定します (1 つのエンティティが複数の住所やクレジット・カードを持つ場合もあるため)。

**安定性:** 当該フィーチャーの安定性の値 (つまり、エンティティの存続期間中に、当該フィーチャーが変わりそうにないかどうか) を示します。例えば、生年月日フィーチャーの場合は決して変わらない

ため、値に「はい」を指定します。一方、住所フィーチャーの場合は変わりやすい (したがって安定性が低い) ため、「いいえ」を指定します。注: 性別は、通常、存続期間を通して安定していますが、データ上の問題が原因で誤って指定されることがよくあるため、デフォルト構成では値が「いいえ」になっています。

**要素テーブル:** このフィーチャーを構成する要素のリスト。

- **要素:** 要素の名前。
- **説明:** 要素の内容についての簡単な説明。
- **データ型:** この要素で使用できるデータ型。使用可能な型は、String、Integer、Real、および Date です。

「**新しい要素の追加**」ボタン: 要素テーブルに新しい行を追加します。これにより、新しい要素を定義できるようになります。

「**要素の削除**」ボタン: 選択された行を要素テーブルから削除します。この操作を元に戻すことはできません。

**注意** リポジトリ内に既にデータが格納されている状態でフィーチャーまたはフィーチャー要素の編集、削除、匿名化を行う場合は、リポジトリの内容を消去してからデータを再ロードしてください。これにより、リポジトリが不整合な状態になるのを防ぐことができます。

**匿名化。** データ保護のために、リポジトリにデータを追加するときにデータを匿名化するように選択できます。フィーチャーに対してこれを有効にするには、「はい」を選択します。詳しくは、『リポジトリ・フィーチャーの匿名化』のトピックを参照してください。

## リポジトリ・フィーチャーの匿名化

データ・セキュリティの一環として、個人の識別情報が誤って開示されるリスクを軽減するために、リポジトリにデータを追加するときにデータの匿名化が必要になることがあります。

匿名化されたデータをリポジトリにエクスポートする場合は、匿名化されたデータによるエンティティーの解決を可能にする匿名化の方法が必要になります。例えば、ある個人のクレジット・カードの詳細に対する 2 つのデータ・レコードが「anon\_s21」と「anon\_s9271」として匿名化されている場合、それらの関係性は失われます。しかし、これらのレコード間の内部のバックグラウンド・リンクを使用すると、一方の名前が他方の簡易形式であることがシステムによって理解されます。

匿名化されたデータのリンクを可能にするバックグラウンド・リンクと識別子は、リポジトリの作成時に生成され、リポジトリに固有のものです。暗号化されたデータは内部で保管され、ストリームがリポジトリに接続するときに読み取られます。

リポジトリの構成時に、個々のフィーチャーをそれぞれ匿名化するかどうかを指定できます。フィーチャーを匿名化すると、その要素がすべて匿名化され、フィーチャーはその使用タイプと関係なく常に匿名化されます。詳しくは、22 ページの『フィーチャーの追加または編集』のトピックを参照してください。

**注:** SPSS Entity Analytics のフィールドをすべて匿名化することは避けてください。すべてを匿名化すると、どのデータが返されるかを識別できなくなります。後で、元のデータに再度マージできるように、少なくとも (単に行番号であろうとも) 1 つのフィールドは匿名化しないままにすることをお勧めします。

「リポジトリ構成」ウィンドウのフィーチャー・リスト内の列には、どのフィーチャーが匿名化するように設定されているかが表示されます。入力項目は「はい」または「いいえ」のいずれかです。



注: フィーチャーが匿名化される前のデータが既存のリポジトリに含まれている場合は、最初にすべてのデータを消去する必要があります。消去しない場合、匿名化されたフィーチャーと匿名化されていないフィーチャーの間での一致が発生しなくなります。

## エンティティ・タイプの保守

**エンティティ・タイプ**は、論理的に同じものとして分類されたりポジトリ・フィーチャーから構成される名前付きセットです。例えば、顧客データ・セットでの使用を目的としたエンティティ・タイプは、名前、生年月日、性別、住所、電話番号などのフィーチャーで構成されます。

IBM SPSS Modeler Entity Analytics リポジトリにはエンティティ・タイプの標準セットが用意されていますが、独自のセットを追加することもできます。

「リポジトリ設定」ウィンドウの「エンティティ タイプ」セクションには、作成済みの各種エンティティ・タイプがリストされます。以下の操作を実行することができます。

- 新しいエンティティ・タイプを作成する。
- 既存のエンティティ・タイプを複製する (既存のエンティティ・タイプに基づいて新しいエンティティ・タイプを作成する場合など)。
- エンティティ・タイプにフィーチャーを関連付ける。
- エンティティ・タイプからフィーチャーの関連付けを解除する。
- エンティティ・タイプの名前を変更する。
- エンティティ・タイプを削除する。

**エンティティ・タイプ:** 選択されたエンティティ・タイプの名前。

**フィーチャー:** このエンティティ・タイプを構成する有効なフィーチャーのリスト。

**使用タイプ:** (オプション) このフィーチャーが使用される可能性のあるさまざまなコンテキストを示します。使用タイプを追加または編集するには、この列をダブルクリックします。複数の使用タイプを指定する場合は、コンマとスペースで区切ってください。「入力」タブでフィーチャーの「使用法」列をクリックすると、ここで指定した値により、EA エクスポート・ノードまたはストリーミング EA ノードに表示される値が定義されます。詳しくは、17 ページの『マッピングに関するリポジトリ入力オプション』のトピックを参照してください。

使用タイプに関する一般情報

- • 使用タイプは任意のラベルです。
- • ほとんどすべてのテキスト・エントリーから使用タイプを作成できますが、スペースや無効な文字を入力することはできません。
- • 入力内容は自動的に大文字に変換されます。
- • 必要な数の使用タイプを作成することができます。
- • 使用タイプの名前をわかりやすいものにする必要はありませんが、自分や他のユーザーにとってわかりやすい命名規則を使用すると、後でマッピングを行う場合に便利です。
- • マッピングを行う際に、一部の要素がある使用タイプにマップされていて、他の要素が別の使用タイプにマップされている場合は、警告が赤のフォントで表示されます。

通常、2 つのフィールドを同じフィーチャー要素にマップしようとする、エラーが表示されます。使用タイプは、複数のフィールドを同じフィーチャー要素にマップし、それらのフィールド間での比較を可能にするための方法です。

例えば、*HOMEADDRESS* と *WORKADDRESS* という異なる 2 つのフィーチャーを定義した場合、これらのフィーチャー間での一致は存在しません。あるエンティティーが *HOMEADDRESS* というフィーチャーを持ち、このフィーチャーが別のエンティティーの *WORKADDRESS* というフィーチャーと同じであっても、これらは異なるフィーチャーであるため、これらの間に一致は存在しません。ただし、異なる複数の使用タイプを持つ単一のフィーチャーを再使用する場合、解決ルールは、*ADDRESS.WORK* と *ADDRESS.HOME* を同じものとして認識します。

異なるフィーチャーについて使用タイプを再使用することも、異なる使用タイプを使用することもできます。例えば、電話番号について *HM* と *WK* を使用し、住所について *HOME* と *WORK* を使用することができます。電話番号と住所を比較することはないため、これで問題ありません。ただし、使用タイプを統一しておく、後でフィールドの特定やグループ化を行う場合に便利です。

同じフィーチャーを使用している場合、複数のエンティティー・タイプを 1 つのリポジトリに格納しても、使用タイプの違いは問題にはなりません。例えば、エンティティー・タイプ *COMPANY* の *ADDRESS* の使用タイプとして *WK* と *HM* を定義した場合でも、これらの使用タイプは、*PERSON* の *ADDRESS* の使用タイプである *WORK* と *HOME* に対して比較されます。

新しいエンティティー・タイプを作成するには、以下の手順を実行します。

1. 画面左側のナビゲーション・ペインで、「**エンティティー タイプ**」を右クリックします。
2. 「**新規エンティティー タイプ**」を選択します。
3. エンティティー・タイプの固有名を入力して「**OK**」をクリックします。
4. エンティティー・タイプにフィーチャーを関連付けます (次のセクションを参照)。

エンティティー・タイプにフィーチャーを関連付けるには、以下の手順を実行します。

1. 画面左側のナビゲーション・ペインでエンティティー・タイプを選択します。
2. 「フィーチャーの添付」ボタン (画面右側の一番上のボタン) をクリックします。
3. 使用可能なフィーチャーのリストから、1 つ以上のフィーチャーを選択して「**OK**」をクリックします (複数のフィーチャーを選択する場合は、**Ctrl** キーを押しながらフィーチャーをクリックします)。

エンティティー・タイプからフィーチャーの関連付けを解除するには

1. 画面左側のナビゲーション・ペインでエンティティー・タイプを選択します。
2. 画面右側の関連付けられたフィーチャーの表から、1 つ以上のフィーチャーを選択します。複数のフィーチャーを選択する場合は、**Ctrl** キーを押しながらフィーチャーをクリックします。
3. 「フィーチャーの添付解除」ボタン (画面右側の一番下のボタン) をクリックします。

既存のエンティティー・タイプを複製するには、以下の手順を実行します。

1. 画面左側のナビゲーション・ペインで、複製したいエンティティー・タイプを右クリックします。
2. 「**重複エンティティー タイプ**」を選択します。
3. 新しいエンティティー・タイプの固有名を入力して「**OK**」をクリックします。
4. 必要に応じて、エンティティー・タイプにフィーチャーを関連付けたり、エンティティー・タイプからフィーチャーの関連付けを解除したりします (前述の手順を参照)。

エンティティー・タイプの名前を変更するには、以下の手順を実行します。

**注意** リポジトリ内に既にデータが格納されている状態でフィーチャーまたはフィーチャー要素の編集、削除、匿名化を行う場合は、リポジトリの内容を消去してからデータを再ロードしてください。これにより、リポジトリが不整合な状態になるのを防ぐことができます。

1. 画面左側のナビゲーション・ペインで、名前を変更したいエンティティ・タイプを右クリックします。
2. 「名前を変更」を選択します。
3. エンティティ・タイプの新しい名前を入力して「OK」をクリックします。

エンティティ・タイプを削除するには、以下の手順を実行します。

**注意** リポジトリ内に既にデータが格納されている状態でフィーチャーまたはフィーチャー要素の編集、削除、匿名化を行う場合は、リポジトリの内容を消去してからデータを再ロードしてください。これにより、リポジトリが不整合な状態になるのを防ぐことができます。

1. 画面左側のナビゲーション・ペインで、削除したいエンティティ・タイプを右クリックします。
2. 「削除」を選択します。
3. 「OK」をクリックして、エンティティ・タイプの削除を確認します。

**注意:**

エンティティ・タイプの削除を取り消すことはできません。

## エンティティの一致に対するしきい値の設定

「リポジトリ構成」ウィンドウの「解決ルール」セクションで、エンティティの一致が発生するしきい値を選択します。

リポジトリを作成すると、一致率がデフォルトのしきい値に事前設定されます。

エンティティを解決できるだけの十分な一致がレコード内に見つからない場合は、「**積極的解決に設定**」を選択します。

他のいずれかの設定からデフォルトのしきい値に戻るには、「**デフォルトの解決に設定**」を選択します。

検出された一致が多すぎる場合は、「**従来 of 解決に設定**」を選択します。

エンティティと関係性の両方のリポジトリを作成するには、「**関係を含める**」を選択します。このオプションが使用可能になるのは、IBM SPSS Modeler Entity Analytics Unleashed と呼ばれる個別ライセンスによるアップグレードがある場合に限られることに注意してください。

## リポジトリ構成の再利用

設定済みの構成を別のリポジトリで使用したい場合は、既存の構成を XML ファイルにエクスポートし、そのファイルを別の (ターゲットの) リポジトリにインポートします。これは、既存のセットアップ内でのみ可能です。例えば、IBM SPSS Modeler のあるバージョンから別のバージョンに、または、あるデータベース・タイプから別のデータベース・タイプに、リポジトリ構成を移行することはできません。

既存の構成を再利用するには、以下の手順を実行します。

1. 使用したい構成が設定されているリポジトリの「リポジトリ構成」ウィンドウを表示します。詳しくは、19 ページの『エンティティ・リポジトリの構成』のトピックを参照してください。
2. このウィンドウのメニューから、以下のように選択します。

「構成」 > 「構成のエクスポート」

3. 「名前を付けて保存」ダイアログ・ボックスで、エクスポート XML ファイルの名前と場所を選択します。
4. ターゲット・リポジトリの「リポジトリ構成」ウィンドウを表示します。



5. このウィンドウのメニューから、以下のように選択します。

「構成」 > 「構成のインポート」

6. 「開く」ダイアログ・ボックスで、上記の手順でエクスポートした XML ファイルの名前と場所を選択して「開く」をクリックします。

## 構成の変更の保存

構成の変更を保存するには、以下の手順を実行します。

「リポジトリ構成」ウィンドウのメニューから、以下のように選択します。

「ファイル」 > 「保存」

## 構成ウィンドウを閉じる

構成ウィンドウを終了するには、以下の手順を実行します。

「リポジトリ構成」ウィンドウのメニューから、以下のように選択します。

「ファイル」 > 「終了」

構成に対する変更内容が保存されていない場合、変更内容を保存するには「OK」をクリックし、保存せずに終了するには「キャンセル」をクリックします。

---

## 解決済み ID の分析 (エンティティ分析 (EA) ソース・ノード)

データをリポジトリにエクスポートした後で、エンティティ分析 (EA) ソース・ノードを使用して、解決済み ID を別の IBM SPSS Modeler ノードに渡し、さらに分析や処理 (解決済み ID をリストするレポートの作成など) を行うことができます。

解決済み ID を分析するには、以下の手順を実行します。

1. エンティティ分析 (EA) ソース・ノードをストリームに追加します。
2. エンティティ分析 (EA) ノードを開きます。
3. 「データ」タブで、エンティティ・リポジトリとその入力データ・ソースを 1 つ以上選択します (レコード・カウントを更新するには、「リフレッシュ」をクリックします)。詳しくは、『データ・ソースの選択』のトピックを参照してください。
4. さらにノードをストリームに追加し、必要な処理を実行します。詳しくは、29 ページの『ストリームへのノードの追加』のトピックを参照してください。

## データ・ソースの選択

「データ」タブで、さらに処理を実行するリポジトリ内のデータ・ソースを 1 つ以上選択します。リストされたデータ・ソースのレコード・カウントを更新するには、「リフレッシュ」をクリックします。

**エンティティ・リポジトリ:** 現在のエンティティ・リポジトリが表示されます (存在する場合)。リポジトリが複数存在する場合に別のリポジトリを選択するには、必要なリポジトリをリストから選択します。新しいリポジトリを作成するには、「<参照...>」を選択し、リポジトリを作成するためのダイアログ・ボックスを表示します。詳しくは、16 ページの『エンティティ・リポジトリのオプション』のトピックを参照してください。

データ・ソースのレコードを含める。この表には、リポジトリに入力された各種のデータ・ソースと、各ソース内のレコード数がリストされます。さらに分析や処理を行うために使用したいデータ・ソースの「含める」チェック・ボックスを選択します。すべてのデータ・ソースを選択または選択解除するには、「すべて含める」または「すべて除外」をそれぞれクリックします。

**関係性。**リポジトリに含める関係性のタイプを選択します。これが使用可能になるのは、IBM SPSS Modeler Entity Analytics Unleashed と呼ばれる個別ライセンスによるアップグレードがある場合、および関係性を含めるようにリポジトリが構成されている場合に限られることに注意してください。

- **関係なし。**関係性の詳細は使用されません。
- **近い関係性。**密接に関連したエンティティのみを選択します。関係性の近さは、マップされたフィーチャーのプロパティなどの多くの変数、共有されるフィーチャー、および解決が従来または積極的のいずれかに設定されているかによって異なります。
- **すべての関係性。**すべての関連したエンティティを選択します。

**最大分離度。**「近い関係性」または「すべての関係性」のいずれかが選択されている場合にのみ使用可能です。関係性の識別に使用する分離度を選択します。例えば、Ann と Bob は知り合いではないが、John は Ann と Bob の両方を知っている場合、Ann と Bob は分離度 2 で関係しています。

**エンティティ・タイプの実出力。**デフォルトでは、リポジトリに詳細が含まれている場合、ここにはリポジトリにリストされている最初のエンティティ・タイプが表示されます。リポジトリにエンティティ・タイプが複数含まれている場合に、あるエンティティ・タイプをここで選択すると、「フィルター」タブに表示されるフィーチャーが変更され、そのタイプのフィーチャーがリストされます。リポジトリで使用される任意のエンティティ・タイプから選択できます。

## データ・フィールド名の変更

「フィルター」タブを使用すると、さらなる処理のために下流に渡される解決済み ID フィールドの名前を変更することができます。例えば、下流で別のデータ・セットと結合する際にフィールド名の互換性を維持するなどの目的で、解決済み ID フィールドの名前を変更する場合があります。

元の名前を持つフィールドは次のとおりです。

表 10. 解決済み ID フィールド

フィールド	説明
\$EA-ID	エンティティ識別子
\$EA-SRC	レコードの作成元のデータ・ソースを識別するソース・タグ
\$EA-KEY	データ・ソース・ファイル内の固有キーとして指定されたフィールド

注: 「フィルター」タブを使用してフィールドをフィルタリングすることもできますが、解決済み ID フィールドはエンティティ分析処理に必要な最低限のフィールドであるため、ここではフィルタリングしないでください。

## データ・フィールドの型情報の設定

「タイプ」タブでは、さらなる処理のために下流に渡される解決済み ID フィールドの各種プロパティを表示したり変更したりすることができます。

変更できるプロパティは、通常の SPSS Modeler データ型ノードの「タイプ」タブで変更できるプロパティと同じです。

表 11. フィールドのデータ型プロパティ

プロパティ	説明
尺度	フィールド内のデータの特性を記述するために使用される測定の尺度 (つまりデータ型)。
値	データ・セットからデータ値を読み取るためのオプションを指定します。
欠損値	フィールドの欠損値の処理方法を指定します。
検査	フィールドの値が指定された値または範囲に対応しているかを確認するための検証オプション。
役割	モデル作成ノードまたはモデル・ナゲットにデータが渡された場合のフィールドの使用方法を指定します。

## ストリームへのノードの追加

SPSS Modeler のさまざまなノードをストリームに追加して、エンティティ分析 (EA) ソース・ノードからの出力に対して分析操作や処理操作を実行することができます。例えば、以下に示す 1 つ以上のノードを追加することができます。

- レコード集計ノードまたは重複レコード・ノード: 非常に大きくなる可能性がある出力を要約します。
- 条件抽出ノード: 出力のサブセットを選択します。
- テーブル・ノード: エンティティ分析 (EA) ソース・ノードからの出力を表示します。
- レポート・ノード: レポートの出力内容を印刷します。
- SPSS Modeler エクスポート・ノード: 出力を別の形式 (スプレッドシートやデータベースなど) でエクスポートします。

詳しくは、「*IBM SPSS Modeler Source* 入力ノード、プロセスノード、出力ノード」でレコード設定ノード、出力ノードおよびエクスポート・ノードに関するセクションを参照してください。

---

## リポジトリに対する新規ケースの比較 (ストリーミング EA ノード)

リポジトリ内で何らかの ID 解決が既に実行されている場合は、ストリーミング EA ノードを使用して、その後に検出された新しいケースとリポジトリの内容を比較することができます。このノードは、新規データ・ソースのレコードを処理し、リポジトリ内で解決済みのエンティティとレコードを比較して、さらなる処理のために一致するレコードを渡します。一致の程度は、完全一致に設定することも、既存エンティティとの関連で緩く設定することもできます。

EA エクスポート・ノードと同様に、ストリーミング EA ノードも単一の SPSS Modeler ソース・ノードを入力として使用します。ただし、ストリーミング EA ノードは次の点で異なります。エクスポート・ノードは、その入力レコードに関連するすべてのエンティティについてレコードを出力しますが、ストリーミング EA ノードは、リポジトリ内で解決済みのエンティティに関連するエンティティについてのみレコードを出力します。詳しくは、32 ページの『ストリーミング EA ノードからの出力』のトピックを参照してください。

リポジトリに対して新規ケースを比較するには、以下の手順を実行します。

1. 既存のエンティティと比較する、新規レコードが存在するデータソースに接続します。詳しくは、14 ページの『データ・ソースへの接続』のトピックを参照してください。
2. 「レコード操作」タブで、ストリーミング EA ノードをデータ・ソース・ノードに接続します。
3. エンティティ分析エクスポート・ノードをダブルクリックして、ダイアログ・ボックスを開きます。
4. 「エンティティ リポジトリ」リストをクリックします。

5. 「<参照...>」をクリックして「エンティティ リポジトリ」ダイアログ・ボックスを表示します。
6. 「エンティティ リポジトリ」ダイアログ・ボックスで、「リポジトリ名」フィールドをクリックします。
7. 使用したいリポジトリの名前をクリックします。
8. このリポジトリのユーザー名とパスワードを入力し、「接続」をクリックします。リポジトリに接続したら、「OK」をクリックします。
9. 「ストリーミング EA」ダイアログ・ボックスで、マップ先とするエンティティ・タイプを選択します。詳しくは、24 ページの『エンティティ・タイプの保守』のトピックを参照してください。
10. データ・ソース内の入力フィールドをリポジトリ内のフィーチャーにマップします。詳しくは、『入力フィールドとフィーチャーのマッピング (ストリーミング EA ノード)』のトピックを参照してください。
11. オプションで、データをスコアリングするときに、リポジトリ内のレコードをリアルタイムで更新できます。詳しくは、『入力フィールドとフィーチャーのマッピング (ストリーミング EA ノード)』のトピックを参照してください。
12. 「出力」タブをクリックして、リポジトリに入力したさまざまなデータ・ソースの詳細を表示し、既存のエンティティを取得するための選択基準を設定します。詳しくは、31 ページの『フィールド・マッピングとデータ・ソースの表示 (ストリーミング EA ノード)』のトピックを参照してください。
13. 「フィルター」タブをクリックして、リポジトリに格納された入力フィールドとフィーチャーの詳細を表示します。ノードにおいてマップされていないフィーチャーはデフォルトでフィルターで除外されますが、この設定は必要に応じて変更できます。
14. ノードが正しく設定されたら、「OK」をクリックします。
15. テーブル・ノードをストリーミング EA ノードに接続し、ストリームを実行します。

テーブル・ノードの出力ウィンドウには、データ・ソース内の新規レコードと一致する取得済みエンティティがすべてリストされます。出力フィールドには \$EA- という接頭辞が付加されます。詳しくは、32 ページの『ストリーミング EA ノードからの出力』のトピックを参照してください。

注: ストリーミング EA ノードを実行すると、「サーバー データ モデルで不正な数のフィールドが検出されました」というエラーが返される場合があります。ストリーミング EA ノードの作成後にリポジトリ構成を編集すると、このエラーが発生することあります。このような状況で構成を編集すると、ノードから出力されるフィールドの数や名前が変更される可能性があります。この問題を解決するには、ストリーミング EA ノードを開き、「リフレッシュ」ボタンをクリックします。この操作により、出力フィールドの数と名前が再計算されます。

## 入力フィールドとフィーチャーのマッピング (ストリーミング EA ノード)

「入力」タブには、このノードへの入力に含まれるフィールドをリポジトリ・フィーチャーにマップするためのオプションが表示されます。このタブでマッピング割り当てを設定するか、「表示」タブを選択してリポジトリ内のすべてのデータ・ソースの詳細を確認してから、「OK」をクリックします。

マッピング・セットが既に XML ファイルに保管されている場合は、「マッピングのインポート」をクリックしてそれらのマッピングを使用することができます。

**エンティティ・リポジトリ:** 現在のエンティティ・リポジトリが表示されます (存在する場合)。リポジトリが複数存在する場合に別のリポジトリを選択するには、必要なリポジトリをリストから選択します。新しいリポジトリを作成するには、「<参照...>」を選択し、リポジトリを作成するためのダイアログ・ボックスを表示します。詳しくは、16 ページの『エンティティ・リポジトリのオプション』のトピックを参照してください。

**エンティティ・タイプにマップ。**リポジトリ内で定義されたエンティティ・タイプのリスト (つまりフィーチャーのセット)。リストからいずれかのタイプを選択するか、「<新しいエンティティ・タイプの追加...>」を選択してリポジトリ構成ウィンドウを表示します。このウィンドウで、新しいエンティティ・タイプを定義することができます。詳しくは、19 ページの『エンティティ・リポジトリの構成』のトピックを参照してください。

**永続検索。**データをスコアリングするときにリポジトリ内のレコードをリアルタイムで更新する場合は、このオプションを選択します。

**ソース・タグ:**「永続検索」を選択した場合にのみ使用できます。リポジトリが現在認識しているデータ・ソースを示すタグのリスト。リストからいずれかのタグを選択するか、「<新しいソース・タグの追加...>」を選択して新しいデータ・ソースのタグを作成します。

**固有キー:**「永続検索」を選択した場合にのみ使用できます。データ・レコードの一意的識別子に使用する入力フィールド。

**マッピング・テーブル:**このテーブルでは、各入力フィールドを、リポジトリ内の対応するフィーチャーにマップすることができます。選択したエンティティ・タイプ内に適切なフィーチャーが存在しない場合は、ここで新しいフィーチャーを作成することができます。

- **フィールド:** 選択したデータ・ソース内の入力フィールドのセット。各フィールドには、フィールドの測定の尺度 (つまりデータ型) を示すアイコンがあります。
- **フィーチャーにマップ:** フィールドをフィーチャーにマップするには、フィールド行でこの列をダブルクリックして (またはスペース・バーを押して)、リストからフィーチャーを選択します。適切なフィーチャーがない場合は、「<新しいフィーチャーの追加...>」を選択してリポジトリ構成ウィンドウを表示します。このウィンドウで、このエンティティ・タイプ用の新しいフィーチャーを定義することができます。詳しくは、19 ページの『エンティティ・リポジトリの構成』のトピックを参照してください。
- **使用法:** 複数のコンテキスト (自宅の電話番号と会社の電話番号など) が考えられる特定のフィールドのコンテキストを示します。詳しくは、24 ページの『エンティティ・タイプの保守』のトピックを参照してください。

**マッピングのインポート:** 以前にエクスポートしたフィールド対フィーチャーのマッピング・セットを外部の XML ファイルからインポートします。このオプションを選択すると、異なるソースごとに同じマッピングを定義し直す必要がなくなるため、同じマッピング要件を持つデータ・ソースが複数存在する場合に便利です。

**マッピングのエクスポート:** マッピング・テーブルに表示されたフィールド対フィーチャーのマッピング・セットを外部の XML ファイルにエクスポートします。

## フィールド・マッピングとデータ・ソースの表示 (ストリーミング EA ノード)

「出力」タブでは、リポジトリに入力されたさまざまなデータ・ソースの詳細を表示することができます。これらのデータ・ソースは、一致するエンティティを検索して取得するために、このノードへの入力が処理されるデータ・ソースです。「リフレッシュ」をクリックすると、レコード・カウントが更新されます。

**データ・ソースからの一致を含める (Include matches from data sources).** この表には、リポジトリ内の使用可能な各種のデータ・ソースと、各ソース内のレコード数がリストされます。



一致: 以下のオプションを使用して、「入力」タブで指定したフィールド対フィーチャーのマッピング情報を、候補レコード (つまりリポジトリ全体の内容) に対してどの程度厳密に一致させるかを指定します。一致基準を厳密にするほど、取得されるエンティティが少なくなります。

注: 一致の検出数が 20 を超える場合は、最初に検出された 20 件だけが返されます。

- **完全一致のみを含める:** これは最も厳密な一致基準で、選択されるレコード数が最も少なくなります。完全一致とみなされるエンティティだけが返されるようにするには、このオプションを使用します。
- **考えられる一致を含める。** 一致するエンティティと、同じ識別子を共有するエンティティ (頻度値 One で構成されたフィーチャー (一致するクレジット・カード番号や納税 ID 番号など) を持つエンティティ) の両方が返されるようにするには、この設定を使用します。
- **すべての一致を含める。** リポジトリ内で共有フィーチャーを持つエンティティができるだけ多く返されるようにするには、このオプションを使用します。これは最も緩い一致基準で、選択されるレコード数が最も多くなります。このオプションを選択すると、完全一致のほか、ほぼすべてのフィーチャーを共有するエンティティ (通常は、頻度値が One または Few のエンティティ) が返されます。例えば、同じ納税 ID 番号を持つエンティティと、類似する住所を持つエンティティの両方が返されます。

**関係性。** 関係性を含めるようにリポジトリが構成されている場合にのみ使用できます。関係性を含めるようにリポジトリを構成するには、IBM SPSS Modeler Entity Analytics Unleashed と呼ばれる個別ライセンスによるアップグレードが必要です。出力に含める関係性のタイプを選択します。

- **関係なし。** 関係性の詳細は使用されません。
- **近い関係性。** 密接に関連したエンティティのみを選択します。関係性の近さは、マップされたフィーチャーのプロパティなどの多くの変数、共有されるフィーチャー、および解決が従来または積極的のいずれかに設定されているかによって異なります。
- **すべての関係性。** すべての関連したエンティティを選択します。

**最大分離度。** 「近い関係性」または「すべての関係性」のいずれかが選択されている場合にのみ使用可能です。関係性の識別に使用する分離度を選択します。例えば、Ann と Bob は知り合いではないが、John は Ann と Bob の両方を知っている場合、Ann と Bob は分離度 2 で関係しています。

**エンティティ・タイプの出力。** デフォルトでは、リポジトリに詳細が含まれている場合、ここにはリポジトリにリストされている最初のエンティティ・タイプが表示されます。リポジトリにエンティティ・タイプが複数含まれている場合に、あるエンティティ・タイプをここで選択すると、「フィルター」タブに表示されるフィーチャーが変更され、そのタイプのフィーチャーがリストされます。リポジトリで使用される任意のエンティティ・タイプから選択できます。

## ストリーミング EA ノードからの出力

ストリーミング EA ノードからの出力は、取得された各レコードに関する以下のフィールドから構成されます。

フィールド	説明
<i>Field1</i> [, <i>Field2</i> [, ... <i>FieldN</i> ]]	新しいレコードが含まれたデータ・ソースのフィールド。
\$EA-ID	リポジトリ内での当該レコードのエンティティ ID。
\$EA-SRC	当該レコードの発生元のデータ・ソースを識別するソース・タグ。
\$EA-KEY	データ・ソース・ファイル内における当該レコードの固有キーの値。

フィールド	説明
\$EA-SC	当該レコードと、リポジトリ内で観測されたエンティティとの一致度を示すフィールド。値の範囲は 1.0 (一致度が低い) から 10.0 (一致度が高い) です。
\$EA-Feature1[, \$EA-Feature2[, ... \$EA-FeatureN]]	リポジトリ内で当該レコード用にマップされたフィーチャーの値。

リポジトリ内で関係性のフィールドが有効になっていて、「出力」タブの分離度がゼロより大きい場合、ストリーミング EA ノードからの出力には、取得された各レコードに関する以下のフィールドも含まれます。

フィールド	説明
\$EA-DEGREE	分離度。
\$EA-PARENT	分離の計算元のレコードの識別子。
\$EA-CHILD	分離の計算先のレコードの識別子。
\$EA-RULE	

## 他の IBM SPSS 製品での IBM SPSS Modeler Entity Analytics の使用

以下の製品で IBM SPSS Modeler Entity Analytics を使用できるようにするためのインストーラーが用意されています。

- IBM SPSS Collaboration and Deployment Services
- IBM SPSS Modeler Batch for Windows
- IBM SPSS Modeler Solution Publisher

IBM SPSS Modeler Entity Analytics のフィーチャーをこれらの製品で使用するには、該当するインストーラーを実行する必要があります。詳しくは、「*IBM SPSS Modeler Premium* のインストール」ガイドを参照してください。

インストール後は、IBM SPSS Collaboration and Deployment Services Deployment Manager クライアントを使用して、Entity Analytics リポジトリ サーバー定義を作成する必要があります。Entity Analytics ノードを使用する IBM SPSS Modeler ストリームを IBM SPSS Collaboration and Deployment Services ジョブで使用する (Entity Analytics ストリームを IBM SPSS Collaboration and Deployment Services で使用する) ために必要です。サーバー定義は、ストリーム内のリポジトリ名と一致している必要があります。この定義を使用して、ストリームに対して、リポジトリの場所の指示および必須の接続情報の指定が行われます。

## 管理タスク

Entity Analytics 内に作成されたリポジトリの場合、新しいデータベース・サービスは、IBM DB2 製品を使用して作成されます。DB2 に関連する管理タスクがいくつかあります。これらのタスクは通常、データベース管理者またはシステム管理者によって実行されます。該当するタスクは以下のとおりです。

- ポート割り当ての構成
- リポジトリ・データベースの管理者資格情報の管理

すべてのリポジトリに適用される、その他の管理タスクは以下のとおりです。これらのタスクは必要に応じて実行してください。

- 別のストレージ・ディレクトリへのリポジトリの移動
- 日時フィールドとタイム・スタンプ・フィールドに対するストリーム・プロパティの設定
- タイムアウト設定の調整
- SPSS Modeler Client と SPSS Modeler Server が同じ Windows システム上にある場合の IBM SPSS Modeler Entity Analytics の実行
- エンティティ・リポジトリの消去
- エンティティ・リポジトリの削除
- リポジトリに接続できない場合のリポジトリの削除

## ポート割り当ての構成

各 DB2 データベース・サービスには、マシン上で稼働する他のサービスに割り当てることができないポートを割り当てる必要があります。データベース・サービスは、IBM SPSS Modeler Server と同じマシン上に常駐します (あるいは、IBM SPSS Modeler を IBM SPSS Modeler Server と接続せずに使用する場合は、IBM SPSS Modeler が稼働するマシン上に常駐します)。

Entity Analytics は、デフォルトで 1320 から 1520 の範囲でポートを割り当て、最初に作成されたリポジトリにポート 1320 を割り当てます。競合が発生した場合は、<Modeler Server のインストール・パス>/ext/bin/pasw.entityanalytics/ea.cfg ファイルを編集し、*min\_port* 設定と *max\_port* 設定に適切な値を指定して、ポートの割り当てを構成してください。このファイルのデフォルトの内容を以下に示します。

```
# port range configuration for entity analytics
#
#   this port range controls which ports DB2 databases
#   (created to store Entity Analytics Repositories in)
#   may use. Configure this if the default port range will
#   introduce a conflict on your system.
#
# default min_port = 1320
# default max_port = 1520
min_port, 1320
max_port, 1520
```

## リポジトリ・データベースの管理者資格情報の管理

リポジトリが作成されると、エンティティ・リポジトリをホストしている DB2 データベースの管理者ユーザー名とパスワードが定義されます。現在の資格情報が分かっている場合は、DB2 SQL エディターを使用してこれらの詳細情報を変更することができます。

DB2 SQL エディターを開始するには、以下の手順を実行します



1. クライアント・マシンで、コマンド・プロンプト・ウィンドウを開きます。
2. 以下のように入力します。

```
cd modeler_install_dir\ext\bin\pasw.entityanalytics\DB2\bin
```

*modeler\_install\_dir* は、SPSS Modeler がインストールされているディレクトリーです。

3. 以下のように入力します。

```
solsql -c "C:\Documents and Settings\All Users\Application Data\IBM\SPSS\Modeler\version\EA\repositories\repos_name
```

*version* は、SPSS Modeler のインストールのバージョン番号、*repos\_name* は、リポジトリーの名前です。

4. プロンプトで、現在のデータベース管理者のユーザー名とパスワードを入力し、solsql> プロンプトを表示します。

データベース管理者のパスワードを変更するには、以下の手順を実行します。

1. solsql> プロンプトで、以下のように入力します。

```
alter user username identified by password;  
commit work;
```

*username* は、データベース管理者の現在のユーザー名、*password* は、新しいパスワードです。

2. 「exit;」と入力してエディターを閉じます。
3. SPSS Modeler Client を再起動します。

DB2 データベースで実行するその他の管理タスクについては、<http://publib.boulder.ibm.com/> で、該当するバージョンの IBM DB2 の資料を参照してください。

## 別のストレージ・ディレクトリーへのリポジトリーの移動

リポジトリー・ファイルは、以下の場所に存在する *EA* というディレクトリーにデフォルトで保管されます。

- C:\Documents and Settings\All Users\ApplicationData\IBM\SPSS\Modeler\version\EA (Windows システム)
- *modeler\_install\_directory*\ext\bin\pasw.entityanalytics\EA (UNIX システム)

リポジトリーの保管に使用されるファイルのサイズは非常に大きくなる場合があるため、必要に応じてそれらのファイルを別のディスクやデータ区分に移動して、空き領域を増やしてください。

リポジトリーを別のディレクトリーに移動するには、以下の手順を実行します。

1. SPSS Modeler を終了します。
2. 元の場所 (上記を参照) から新しい場所に *EA* ディレクトリーを移動します。例えば Windows の場合、*F:\data\EA* などの新しい場所に移動します。
3. <Modeler Server のインストール・パス>\ext\bin\pasw.entityanalytics\ea.cfg ファイルを編集し、以下のオプションを追加します。

```
repository_data_directory, new_location
```

*new\_location* は、*EA* ディレクトリーの移動先ディレクトリー (*F:\data\EA* など) です。

## 日時フィールドとタイム・スタンプ・フィールドに対するストリーム・プロパティの設定

日時データまたはタイム・スタンプ・データを格納するフィールドがソース・データに含まれている場合は、それに対応するストリーム・プロパティが IBM SPSS Modeler Entity Analytics で認識可能な形式に設定されていることを確認してください。

ストリーム・プロパティの形式を設定するには、以下の手順を実行します。

1. SPSS Modeler のメインメニューで、以下のように選択します。

「ツール」 > 「ストリームのプロパティ」 > 「オプション」

2. 「日付/時刻」を選択します。
3. 「日付のフォーマット」を「YYYY-MM-DD」に設定します。
4. 「時間のフォーマット」を「HH:MM:SS」に設定します。
5. 「OK」をクリックします。

## タイムアウト設定の調整

処理速度の遅いシステムや負荷の高いシステムで、リポジトリの作成時やアクセス時にエラーが発生した場合は、必要に応じて、エンティティ分析エンジンまたはエンティティ分析データベース・サーバーを開始/停止するためのタイムアウト設定値を大きくしてください。

エンティティ分析エンジンのタイムアウトを調整するには、以下の手順を実行します。

1. SPSS Modeler を終了します。
2. <Modeler Server のインストール・パス>/ext/bin/pasw.entityanalytics/ea.cfg ファイルを編集し、以下のオプションの値を増やします。

`timeout, value`

`value` は、エンティティ分析エンジンのタイムアウト値 (秒単位) です (デフォルトは 60)。

エンティティ分析データベース・サーバーのタイムアウトを調整するには、以下の手順を実行します (DB2 のみ)

1. SPSS Modeler を終了します。
2. <Modeler Server のインストール・パス>/ext/bin/pasw.entityanalytics/ea.cfg ファイルを編集し、以下のオプションの値を増やします。

`timeout, value`

`value` は、エンティティ分析 DB2 データベース・サーバーのタイムアウト値 (秒単位) です (デフォルトは 100)。

## SPSS Modeler Client と SPSS Modeler Server が同じ Windows システム上にある場合の IBM SPSS Modeler Entity Analytics の実行

同じ Windows システム上にある SPSS Modeler Client と SPSS Modeler Server の両方に IBM SPSS Modeler Entity Analytics がインストールされている場合は、デフォルトで、クライアントとサーバーの両方が同じリポジトリを共有することになります。クライアントとサーバーで別々のリポジトリを使用するには、一方のシステム上で構成ファイル `ea.cfg` を編集し、別のポート範囲とリポジトリ・フォルダーを使用するように構成する必要があります。

注: 特に、32 ビットの SPSS Modeler Client と 64 ビットの SPSS Modeler Server (またはその逆) を使用している場合は、この手順を実行する必要があります。

1. 編集するファイル <Modeler [Server] のインストール・パス>/ext/bin/pasw.entityanalytics/ea.cfg を開きます。
2. 他方のシステムとは異なるポートを使用するように、min\_port 設定と max\_port 設定を変更します。詳しくは、34 ページの『ポート割り当ての構成』のトピックを参照してください。
3. 他方のシステムとは異なるディレクトリーを使用するように、repository\_data\_directory 設定を変更します。
4. ea.cfg ファイルを保存して閉じます。

## エンティティー・リポジトリーの消去

エンティティー・リポジトリーからデータ・レコードを消去したいが、構成情報は維持したいという場合は、リポジトリーからデータを消去することができます。

リポジトリーからすべてのデータを消去するには、以下の手順を実行します。

1. エンティティー分析ノードを開きます。
2. 「エンティティー リポジトリー」リストをクリックします。
3. 「<参照...>」をクリックして「エンティティー解決インスタンス」ダイアログ・ボックスを表示します。
4. 「エンティティー解決インスタンス」ダイアログ・ボックスで、「リポジトリー名」リストをクリックします。
5. 消去するリポジトリーを選択します。
6. まだ接続していない場合は、管理者ユーザー名とパスワードを入力して「接続」をクリックします。
7. 「すべて消去」ボタンが有効な場合、クリックします。
8. 「すべてのデータ・ソースの消去 (Purge All Data Sources)」ダイアログ・ボックスで「消去」をクリックしてリポジトリーの消去を確認します。

## リポジトリーからの未使用データ・ソースの削除

エンティティー・リポジトリー内に使用しなくなった、または不要になったデータ・ソースがある場合、リポジトリーからソースを削除できます。削除するデータ・ソースを 1 つ以上選択できます。

選択したデータ・ソースをリポジトリーから削除するには、以下の手順を実行します。

1. エンティティー分析ノードを開きます。
2. 「エンティティー リポジトリー」リストをクリックします。
3. 「<参照...>」をクリックして「エンティティー解決インスタンス」ダイアログ・ボックスを表示します。
4. 「エンティティー解決インスタンス」ダイアログ・ボックスで、「リポジトリー名」リストをクリックします。
5. データ・ソースの削除元のリポジトリーを選択します。
6. まだ接続していない場合は、管理者ユーザー名とパスワードを入力して「接続」をクリックします。
7. 「リポジトリーの管理」リストで、削除するデータ・ソースを選択します。必要に応じて、Ctrl キーを押しながらクリックして、追加のデータ・ソースを選択します。
8. 「未使用分を削除」ボタンが有効な場合、クリックします。

9. 「未使用データ・ソースの削除 (Delete Unused Data Sources)」ダイアログ・ボックスで、「削除」をクリックしてリポジトリの消去を確認します。

## エンティティ・リポジトリの削除

不要になったリポジトリは、完全に削除することができます。

注意：この操作を実行すると、リポジトリが完全に削除されます。この操作を元に戻すことはできません。リポジトリを削除してよいかどうか確信がない場合は、「消去」ボタンを使用してソース・データをすべて削除してください。この場合、リポジトリ構成は削除されません。詳しくは、37 ページの『エンティティ・リポジトリの消去』のトピックを参照してください。

注：以下の手順では、SPSS Modeler からリポジトリに接続できること、リポジトリをホストするデータベースの管理者ユーザー名とパスワードが分かっていることを前提としています。そうでない場合は、リポジトリに接続できないときにリポジトリを削除する手順を実行してください。詳しくは、『リポジトリに接続できない場合のリポジトリの削除』のトピックを参照してください。

リポジトリを削除するには、以下の手順を実行します。

1. エンティティ分析ノードを開きます。
2. 「エンティティ リポジトリ」リストをクリックします。
3. 「<参照...>」をクリックして「エンティティ解決インスタンス」ダイアログ・ボックスを表示します。
4. 「エンティティ解決インスタンス」ダイアログ・ボックスで、「リポジトリ名」リストをクリックします。
5. 削除するリポジトリを選択します。
6. まだ接続していない場合は、管理者ユーザー名とパスワードを入力して「接続」をクリックします。
7. 「リポジトリ全体の削除」ボタンをクリックします (このボタンが使用可能になっている場合)。
8. 「削除」をクリックして、リポジトリの削除を確認します。
9. 「OK」をクリックして、正しく削除されたことを確認します。

## リポジトリに接続できない場合のリポジトリの削除

SPSS Modeler の接続の問題、またはユーザー名やパスワードを忘れたことにより、エンティティ リポジトリを削除するときにそのエンティティ リポジトリに接続できない場合は、以下の手順を実行してください。

リポジトリ データベースをホストしているマシンで、以下の手順を実行します。

### Windows システム

1. 「コマンド・プロンプト」ウィンドウを開きます。
2. 以下のように入力します。

```
cd modeler_install_dir
cd ext%bin%pasw.entityanalytics
delete_repository.bat repos_name
```

*modeler\_install\_dir* は SPSS Modeler がインストールされているディレクトリ、*repos\_name* はリポジトリの名前です。

注：リポジトリ名では、大文字と小文字が区別されます。

3. このセクションで後述する「手順の完了」に進みます。

## UNIX システム

1. シェルを開きます。
2. 以下のように入力します。

```
cd modeler_server_install_dir  
cd ext/bin/pasw.entityanalytics  
./delete_repository.sh repos_name
```

*modeler\_server\_install\_dir* は SPSS Modeler Server がインストールされているディレクトリー、*repos\_name* はリポジトリーの名前です。

注: リポジトリー名では、大文字と小文字が区別されます。

## 手順の完了 (すべてのシステム)

1. プロンプトで「Y」と入力して、リポジトリーの削除を確認します。
2. 削除したリポジトリーと同じ名前のディレクトリーを削除します。このディレクトリーを削除できない場合は、マシンを再始動してもう一度実行してください。





## 第 4 章 エンティティ分析の実行

### この例について

IBM SPSS Modeler を使用することによって優れた結果が得られることは既に説明しましたが、エンティティ分析を追加することにより、さらに結果を改善することができます。この例では、その点について説明します。

この例では、ストリーム *loan\_entity\_analytics.str* を使用します。このストリームは、データ・ファイル *loan\_applications.csv* を参照します。このファイルは、IBM SPSS Modeler Entity Analytics もインストールされている IBM SPSS Modeler のインストール済み環境の *Demos* ディレクトリーに格納されています。*Demos* ディレクトリーには、Windows スタート・メニュー上の IBM SPSS Modeler プログラム・グループからアクセスすることができます。*loan\_entity\_analytics.str* ファイルは *Entity\_Analytics* ディレクトリーに格納されています。

注: このサンプル・ストリームを実行する前に、システム上にリポジトリーを作成する必要があります。リポジトリーを作成してから、この例を実行してください。詳しくは、15 ページの『リポジトリーの作成』のトピックを参照してください。

ここでは、銀行の経営陣が、保留になっているローン申請について、債務不履行になる可能性があるかどうかを懸念している状況を想定します。この銀行の IT 部門では、SPSS Modeler を長期にわたって利用しています。そのため、銀行のスタッフによって既にストリームが作成されており、この銀行が過去に扱った 700 件のローンに関する既存データから予測モデルが構築されています。これらのローンは、返済されているものもあれば、顧客が返済を怠っているものもあります。

### 元のモデル

ここでは、銀行のスタッフがどのようにモデルを構築し、そのモデルからどのような情報を得たのかを確認します。



図 2. モデル作成ノードを使用した初期のストリーム

*loan\_applications.csv* データ・セットには、過去のローンの詳細のほか、ローン申請がまだ保留になっている 150 名の顧客の詳細も格納されており、レコードの合計は 850 件になります。

このデータ・セットのすべてのフィールドが、予測の作成に使用されるわけではありません。例えば、*name* フィールドは無視してもかまいません。こうしたフィールドの役割を「なし」に設定すると、データ型ノードでそれらのフィールドが無視されます。予測の作成に使用されるフィールドの役割は「入力」に設定され、モデルが値を予測しようとするフィールドの役割は「対象」に設定されています。



図3. データ型ノードで設定されたフィールドの役割

このモデルでは、過去のデータだけに基いて予測を作成する必要があるため、ストリーム内の条件抽出ノードには、保留中のマークが付いていないローンだけが含まれます。そのため、保留中の150件のローンは破棄されます。

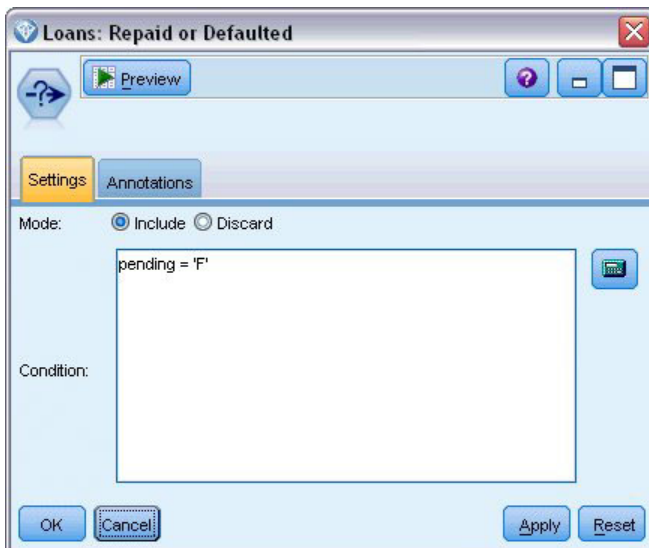


図4. 保留中のローン申請を破棄

保留中のローンが破棄されたため、残り700件のローン（返済済みまたは債務不履行）の詳細だけがモデル作成ノードに渡されます。銀行は、SPSS Modelerの各種アルゴリズムのいずれかを使用して、適切なモ

デルを作成することができます。この例では、C&R ツリー・ノードを使用しています。このノードを使用して、銀行の顧客の過去の履行状況に基づき、債務不履行になる可能性がある顧客を予測するためのモデルが構築されます。

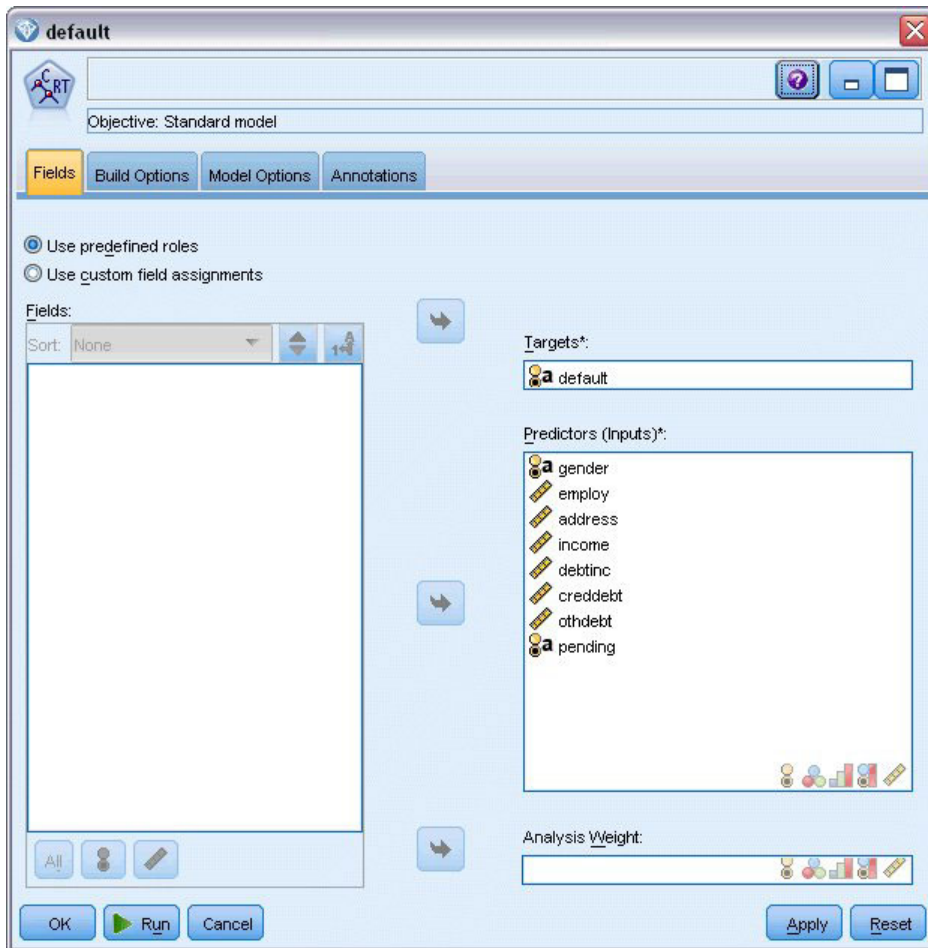


図5. 予測フィールドと対象フィールドの割り当て

データ型ノードで定義されているとおり、予測の作成で使用されるフィールドは予測フィールドとして指定され、モデルが値を予測しようとするフィールド（この場合は **default**）は対象フィールドとして設定されます。

このストリームを実行すると、予測フィールドから構築されたモデルを持つモデル・ナゲットが生成されます。

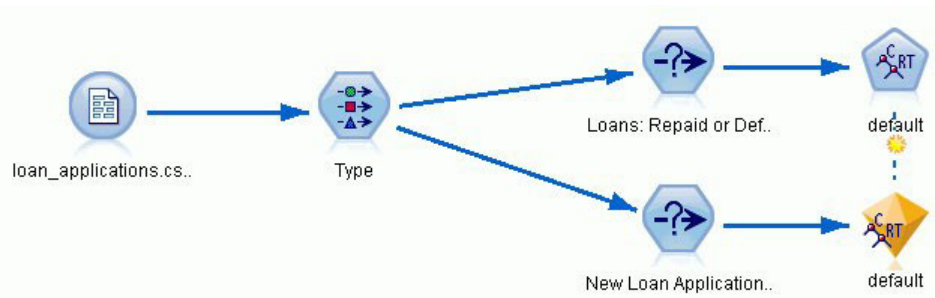


図6. モデル・ナゲットが追加されたストリーム

これで、銀行のアナリストはこのモデルを使用して、返済が保留になっている顧客が債務不履行になる可能性があるかどうかを予測することができます。アナリストは元のデータ・セットを使用して、条件抽出ノードを挿入します。この場合、保留中としてマークされた 150 件のローン・レコードだけがこのノードに組み込まれ、破棄されることはありません。アナリストはこれらのレコードをモデルに直接渡し、モデルの予測を視覚的に表示するための分布ノードを追加します。



図7. 分布ノードが追加された、新規ローン申請を選択するストリーム

分布ノードには、モデル内の  $\$R-default$  フィールドの値の分布が表示されます。C&R ツリー・ノードを実行すると、このフィールドがデータ・モデルに追加されます。このフィールドには、各新規申請者が返済するか債務不履行になるかの予測が格納されます。このフィールドは、エンティティ分析を追加した場合の効果を比較するために、後で使用します。

アナリストがストリームのこの部分を実行したところ、分布ノードの出力により、150 名の新規申請者のうち 137 名がローンを返済するという予測になることがわかりました。残りの 13 名については債務不履行になることが予測されるため、アナリストはこれらの申請を拒否するよう銀行に提案することになります。

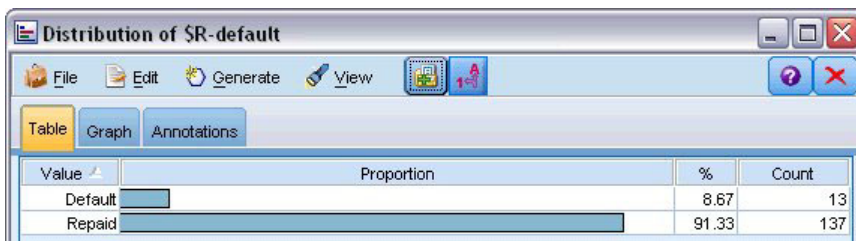


図8. エンティティ分析のない分布ノードからの出力

## エンティティ分析の追加

ここでは、式にエンティティ分析を追加することにより、状況を改善できるかどうかを考えてみます。例えば、ソース・データ内の顧客レコードに不正なエントリが含まれている疑いがあり、エンティティ分析の専門家がその調査を銀行から依頼されたとします。データの入力ミスによってレコードが重複している可能性もありますが、ローン申請者が自分の身元を偽ろうとしている可能性も考えられます。いずれにせよ、銀行は実態を知る必要があります。

この例では、エンティティ・リポジトリーはすでに作成されているものとします。詳しくは、15ページの『リポジトリーの作成』のトピックを参照してください。

## リポジトリーへのソース・データの取り込み

最初に、EA エクスポート・ノードをデータ・ソース・ノードに追加して、ソース・データをエンティティ・リポジトリーにエクスポートできるようにします。

データをエクスポートする前に、データ・ソース内のフィールドをエンティティ・リポジトリー内のフィーチャーにマップする必要があります。さまざまなデータ・ソースが同じ種類の情報に対して異なるフィールド名を使用する場合があるため、この処理が必要になります。エンティティ・リポジトリーには、重複を避けるための情報タイプ（「フィーチャー」と呼ばれます）の標準セットが用意されています。

EA エクスポート・ノードで、リポジトリーに関する詳細として、接続詳細、ソース・タグ（データ・ソースを識別するため。この場合は **TEST**）、エンティティ・タイプ（使用するフィーチャーのセット。この場合は **PERSON**）、固有キー・フィールド（各レコードを一意に識別するため）を設定します。ここでは、固有キーとして **key** フィールドを使用します。

ここで、マッピングを設定することができます。使用するフィーチャー・セットには、*fname*、*mname*、*lname*、*generation*、*dob*、*gender*、*addr1*、*city*、*country*、*postcode*、*phone*、*email*、*ssn*、*drlic*、*passport* の各フィールドに対応するフィーチャーがあります。

最初に、*fname* のマッピングを設定します。表の *fname* 行で「フィーチャーにマップ」列をダブルクリックし、**NAME.GIVEN\_NAME** エントリまでスクロールして、このエントリーをクリックしてマッピングを作成します。

対応するフィーチャーを持つ残りのフィールドをマップします。すべて完了すると、以下のようなマッピング・セットになります。

表 12. リポジトリーのフィーチャーにマップされるフィールド：

フィールド	フィーチャーにマップ
<i>fname</i>	NAME.GIVEN_NAME
<i>mname</i>	NAME.MIDDLE_NAME
<i>lname</i>	NAME.SUR_NAME
<i>generation</i>	NAME.NAME_GEN
<i>dob</i>	DOB.DOB
<i>gender</i>	GENDER.GENDER
<i>addr1</i>	ADDRESS.ADDR1
<i>city</i>	ADDRESS.CITY
<i>country</i>	ADDRESS.COUNTRY
<i>postcode</i>	ADDRESS.POSTAL_CODE



表 12. リポジトリのフィーチャーにマップされるフィールド (続き):

フィールド	フィーチャーにマップ
phone	PHONE.PHONE_NUM
email	EMAIL_ADDR.ADDR
ssn	SSN.ID_NUM
drlic	DRLIC.ID_NUM
passport	PASSPORT.ID_NUM

「実行」をクリックし、データをリポジトリにエクスポートします。エクスポートには多少時間がかかります。「実行のフィードバック」ダイアログ・ボックスが閉じると、エクスポートが完了します。

## 解決済み ID の読み取り

データをリポジトリにエクスポートすると、考えられる ID 競合の解決処理がエンティティ分析システムによって開始され、固有のエンティティ識別子が割り当てられます。この識別子は、後で *\$EA-ID* フィールドとして表示されます (注: これは、EA エクスポート・ノードの「固有キー」フィールドとは異なります。このフィールドは、データ・ソース・レコードを一意に識別するために使用されます)。

解決済み ID を読み取るには、最初に、エンティティ分析 (EA) ソース・ノードをストリームに追加します。この段階では、このソース・ノードをどこにも接続しないでください。

エンティティ分析 (EA) ソース・ノードを開き、エンティティ・リポジトリの詳細を設定します。リポジトリにエクスポートされているデータ・ソースのリストが表示されます。この場合、データ・ソースは 1 つだけです。

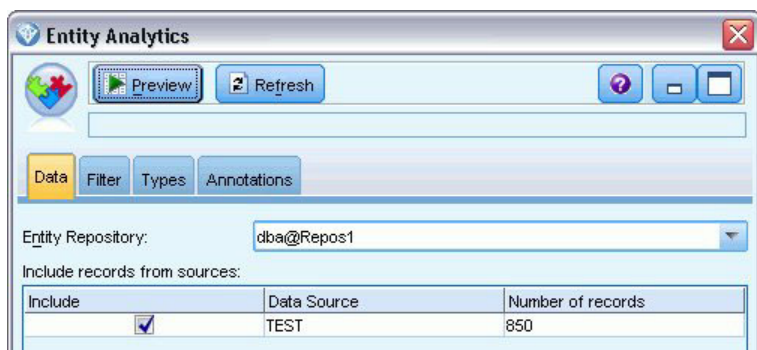


図 9. リポジトリ内のデータ・ソースを選択

「TEST」データ・ソースのチェック・ボックスを選択して「OK」をクリックします。

ここで、エンティティ分析システムがデータに対して実行した内容を確認してみます。テーブル・ノードをエンティティ分析 (EA) ソース・ノードに接続し、テーブル・ノードを開き、「実行」をクリックしてテーブル・ノードの出力ウィンドウを表示します。



	\$EA-ID	\$EA-SRC	\$EA-KEY
1	1	TEST	REF00000001
2	2	TEST	REF00000002
3	3	TEST	REF00000003
4	4	TEST	REF00000004
5	5	TEST	REF00000005
6	6	TEST	REF00000006
7	7	TEST	REF00000007
8	8	TEST	REF00000008
9	9	TEST	REF00000009
10	10	TEST	REF00000010
11	11	TEST	REF00000011
12	12	TEST	REF00000012
13	13	TEST	REF00000013
14	14	TEST	REF00000014
15	15	TEST	REF00000015
16	16	TEST	REF00000016
17	17	TEST	REF00000017
18	18	TEST	REF00000018
19	19	TEST	REF00000019
20	20	TEST	REF00000020

図 10. テーブル・ノードからの出力

これまでに見たことがあるフィールドは、*\$EA-KEY* フィールドだけです。これは、実際にはソース・データの *key* フィールドです。このフィールドは、EA エクスポート・ノードの「固有キー」フィールドとして選択されているため、ここに表示されます。

システムによって追加されたフィールドは、他にも 2 つあります。*\$EA-ID* フィールドは、ソース・レコードではなく解決済み ID の一意の識別子です。その違いについては、すぐに説明します。*\$EA-SRC* フィールドは、データの提供元を識別します。EA エクスポート・ノードでこのフィールドに割り当てたソース・タグは **TEST** であるため、ここでは **TEST** が表示されています。

ソース・データ内の他のフィールドはどうなっているのでしょうか。これらのフィールドは、まだリポジトリ内に存在しています。エンティティ分析 (EA) ソース・ノードは、パフォーマンス上の理由から、必要最小限のフィールド・セットだけを、さらなる処理のために下流に渡したにすぎません。

次に、テーブル・ノードの出力を 385 行目までスクロールします。

	\$EA-ID	\$EA-SRC	\$EA-KEY
376	376	TEST	REF00000376
377	377	TEST	REF00000377
378	378	TEST	REF00000378
379	379	TEST	REF00000379
380	380	TEST	REF00000380
381	381	TEST	REF00000381
382	382	TEST	REF00000382
383	383	TEST	REF00000383
384	384	TEST	REF00000384
385	45	TEST	REF00000385
386	386	TEST	REF00000386
387	387	TEST	REF00000387
388	388	TEST	REF00000388
389	389	TEST	REF00000389
390	390	TEST	REF00000390
391	391	TEST	REF00000861
392	392	TEST	REF00000392
393	393	TEST	REF00000393
394	394	TEST	REF00000394
395	395	TEST	REF00000395
396	396	TEST	REF00000396

図 11. テーブル出力の行と \$EA-ID 番号との違い

ここで、\$EA-ID 番号が順序どおりになっていないことに注意してください。エンティティ分析システムは、レコード REF00000385 がエンティティ 45 として識別される人物を参照し、この人物がレコード REF00000045 も持っていることを検出しました。この出力をさらに下にスクロールすると、485 行目、517 行目、520 行目など、順序どおりになっていない番号がほかにもあることがわかります。注意して確認する必要があります。

最初に、\$EA-KEY フィールドの名前を *key* に変更することにより、ソース・データ内の *key* フィールドのデータが \$EA-KEY フィールドに格納されることを確認します。フィルター・ノードをエンティティ分析 (EA) ソース・ノードに接続し、フィルター・ノードを開きます。2 番目の「フィールド」列で文字列 \$EA-KEY をダブルクリックし、「key」と入力します。

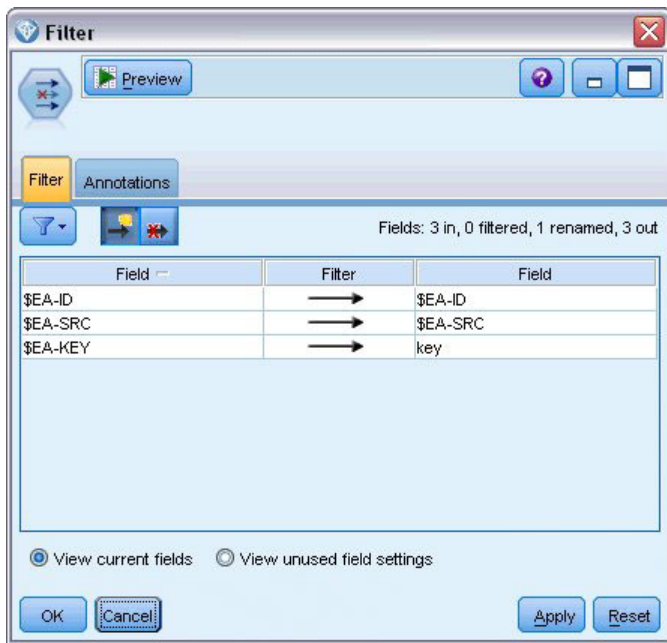


図 12. \$EA-KEY フィールドの名前の変更

「OK」をクリックしてフィルター・ノードを閉じます。

ここで、\$EA-ID エンティティの ID を昇順にソートする必要があります。最初に、ソート・ノードをフィルター・ノードに接続します。次に、ソート・ノードを開き、「ソート項目」表の横にある 1 番上のボタンをクリックして「\$EA-ID」を選択し、「OK」をクリックします。



図 13. エンティティ ID を昇順にソート

ソート順を「昇順」のままにして「OK」をクリックします。

ここで、レコードが一意か重複しているかを示す追加のフィールドを作成する必要があります。最初に、フィールド作成ノードをこのソート・ノードに接続します。次に、フィールド作成ノードを開き、「派生フィ

ールド」の名前を IsDuplicate に設定します。「データ型」リストから「フラグ型」を選択すると、「フィールドのデータ型」も「フラグ」に設定されます。「真の値」フィールドを「重複」に設定し、「偽の値」フィールドを「一意」に設定します。

重複するレコードを検出するには、@OFFSET という特殊なシーケンス関数を使用します。この関数は SPSS Modeler に組み込まれています。

「If」フィールドに、以下のように入力します。

```
'$EA-ID' = @OFFSET('$EA-ID',1) or '$EA-ID' = @OFFSET('$EA-ID',-1))
```



図 14. フィールド作成ノードでの条件設定

@OFFSET 関数は、エンティティ ID が昇順にソートされている状態で、隣接するエンティティ ID が同じかどうかを確認します。ID が同じであれば、レコードが重複していることになります。ID が重複している場合は、その ID の IsDuplicate 値が Duplicate に設定され、そうでない場合は Unique に設定されます。

「OK」をクリックしてノードを閉じます。

フィールド作成ノードの効果を確認するには、テーブル・ノードをフィールド作成ノードに接続し、テーブル・ノードを開いて「実行」をクリックします。次に、テーブル・ノードの出力ウィンドウを 45 行目までスクロールします。

	\$EA-ID	\$EA-SRC	key	IsDuplicate
39	39	TEST	REF00000039	Unique
40	40	TEST	REF00000040	Unique
41	41	TEST	REF00000041	Unique
42	42	TEST	REF00000042	Unique
43	43	TEST	REF00000043	Unique
44	44	TEST	REF00000044	Unique
45	45	TEST	REF00000045	Duplicate
46	45	TEST	REF00000385	Duplicate
47	46	TEST	REF00000046	Unique
48	47	TEST	REF00000047	Unique
49	48	TEST	REF00000048	Unique
50	49	TEST	REF00000049	Unique
51	50	TEST	REF00000050	Unique
52	51	TEST	REF00000051	Unique
53	52	TEST	REF00000052	Unique
54	53	TEST	REF00000053	Unique
55	54	TEST	REF00000054	Unique
56	55	TEST	REF00000055	Unique
57	56	TEST	REF00000056	Unique
58	57	TEST	REF00000057	Unique

図 15. フィールド作成ノードからの出力

エンティティ分析 (EA) ソース・ノードから出力を直接確認したときのことを思い出してください。レコード REF00000385 がエンティティ 45 と同じ人物を参照していることが、システムによって既に検出されています。これは既に確認済みで、レコード REF00000045 と REF00000385 がどちらもエンティティ 45 を参照しているため、これらのレコードが重複していることも確認しています。

出力ウィンドウをさらに下にスクロールすると、重複フラグが付いているレコードがほかにもあることがわかります。

重複レコードをリストするレポートを作成するには、(ノード・パレットの「出力」タブから) レポート・ノードを *IsDuplicate* フィールド作成ノードに接続します。次にレポート・ノードを開き、「テンプレート」タブの入力フィールドに以下のテキストをコピーして「実行」をクリックします。

```
<html>
<h1>List of duplicate customer records.

<h2>This report was generated: [@TODAY]

<h2>Duplicate records
<table>
  <tr>
    <td>Entity ID</td>
    <td>Key</td>
  </tr>

#WHERE IsDuplicate = "Duplicate"

  <tr>
    <td>['$EA-ID']</td>
    <td>[key]</td>
  <tr>
```

```
#  
</table>  
</html>
```

これにより、以下の出力が作成されます。

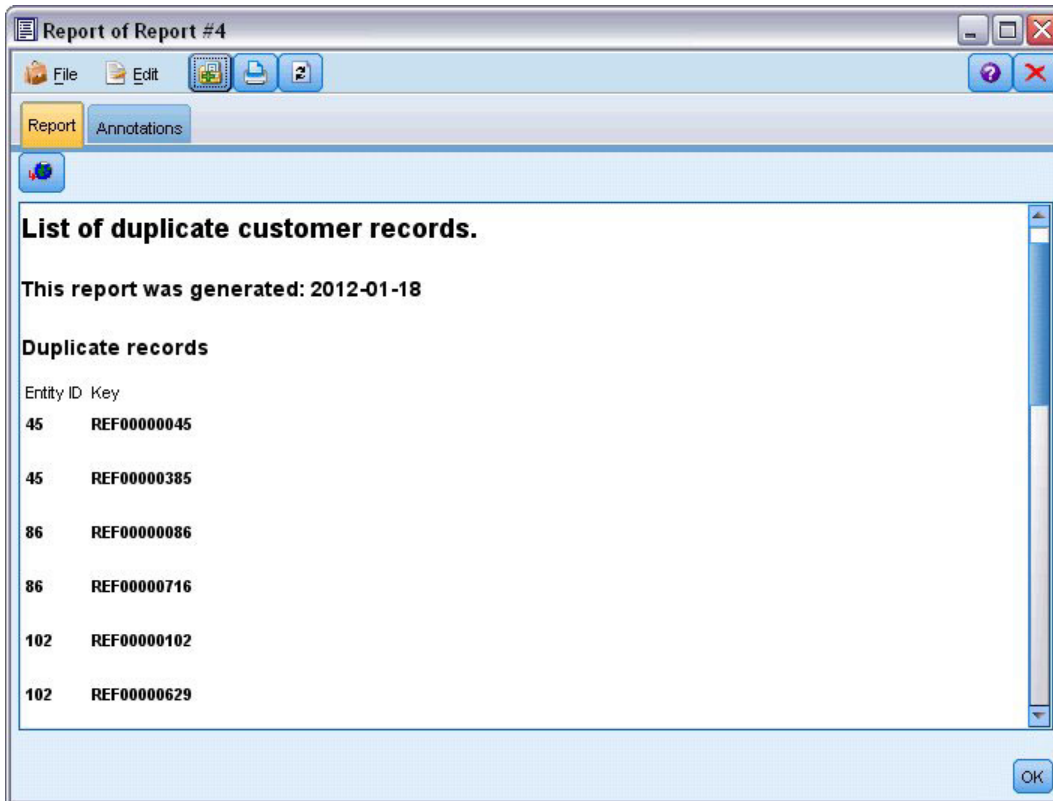


図 16. レポート・ノードからの出力

このレポートでは HTML 形式が使用されていますが、XML 形式や ASCII 形式を使用することもできます。

## エンティティ分析出力と元のモデルとの比較

この例の最終段階では、エンティティ分析を追加した場合に、銀行の元の予測と比較して違いが出るかどうかを確認します。元のモデルでは、保留になっている 150 件の申請のうち 13 件が債務不履行になると予測されました。ここでは、結合ノードを使用して、このモデルからの出力と、エンティティ分析からの重複レコードに関する情報を結合し、これによって予測に変化が出るかどうかを確認します。

最初に、エンティティ分析によって追加された新規フィールドのデータ型 (SPSS Modeler では測定の尺度と呼ばれます) が正しいかどうかを確認する必要があります。データ型ノードを **IsDuplicate** フィールド作成ノードに接続し、このデータ型ノードを開いて「値の読み込み」ボタンをクリックします。



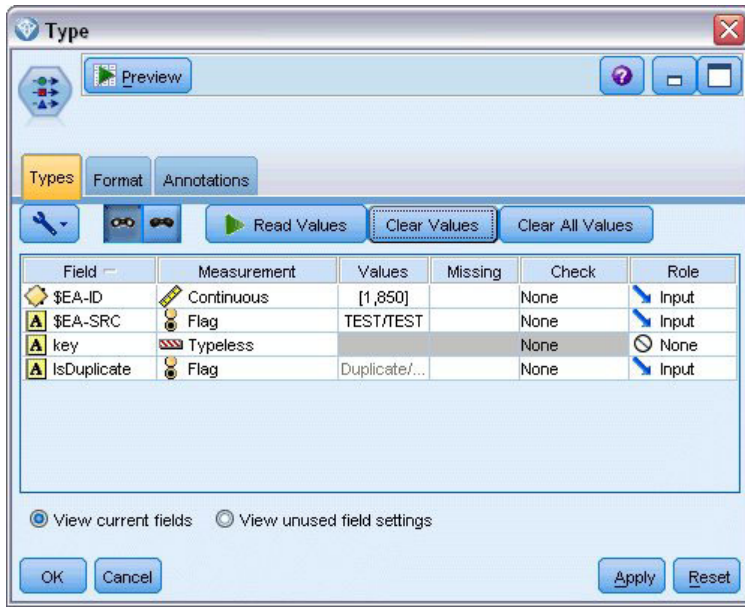


図 17. データ型ノードの設定

これで、結合ノードを追加できるようになりました。このノードをデータ型ノードに接続し、さらに元のモデルを持つ金色のナゲットにも接続します。これを行うには、金色のナゲットを右クリックして「接続」を選択し、結合ノードをクリックします。この操作により、結合ノードに 2 つの入力矢印が設定されます。

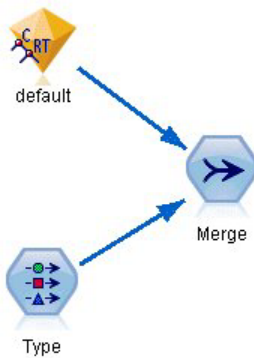


図 18. 結合ノードへの入力

結合ノードを開いて「レコード結合方法」を「キー」に設定し、右矢印ボタンをクリックして「key」フィールドを「キーの候補」から「結合キー」に移動して、「OK」をクリックします。

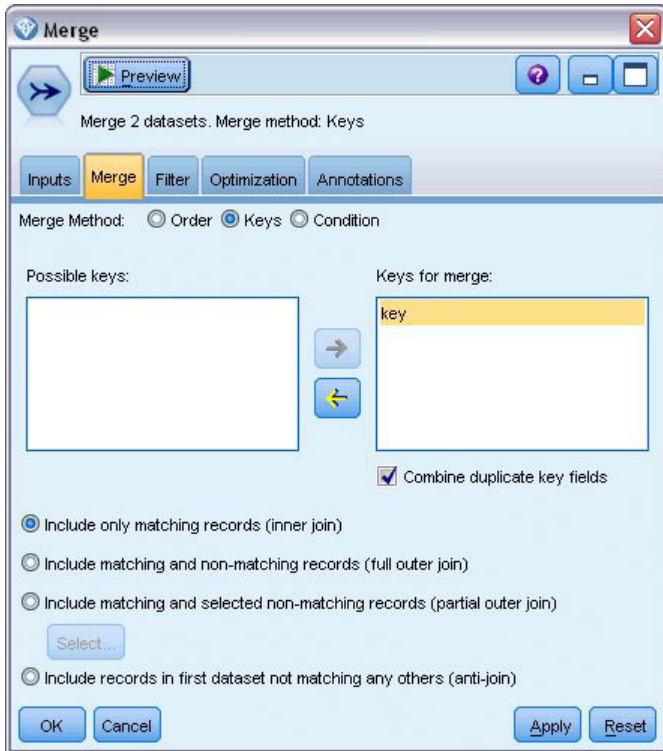


図 19. 結合操作のキー・フィールドを指定

これで、比較を行うための準備はほぼ完了しました。ただし、分布ノードを接続してこの時点で実行しても、元の予測との違いは表示されません。これは、元のモデル・ナゲットからの出力と、エンティティ分析によって作成された新規フィールドがストリームによって結合されていますが、データ・モデル内の予測フィールド自体 (*\$R-default*) は、新しい情報で更新されていないためです。

この更新を行うには、フィールド値を置き換えることができる置換ノードを使用します。置換ノードを結合ノードに接続して、置換ノードを開きます。

「対象フィールド」の右にある上部のボタンをクリックし、リストの下部までスクロールします。次に、「*\$R-default*」を選択して「OK」をクリックします。このダイアログ・ボックスの残りの部分で指定した条件が満たされると、このフィールドの値が変更されます。

条件を指定するには、「置換」が「条件を指定」に設定されていることを確認し、「条件」フィールドに以下のように入力します。

```
default /="default" and IsDuplicate = "Duplicate"
```

「置換値」フィールドに、以下のように入力します。

```
"default"
```

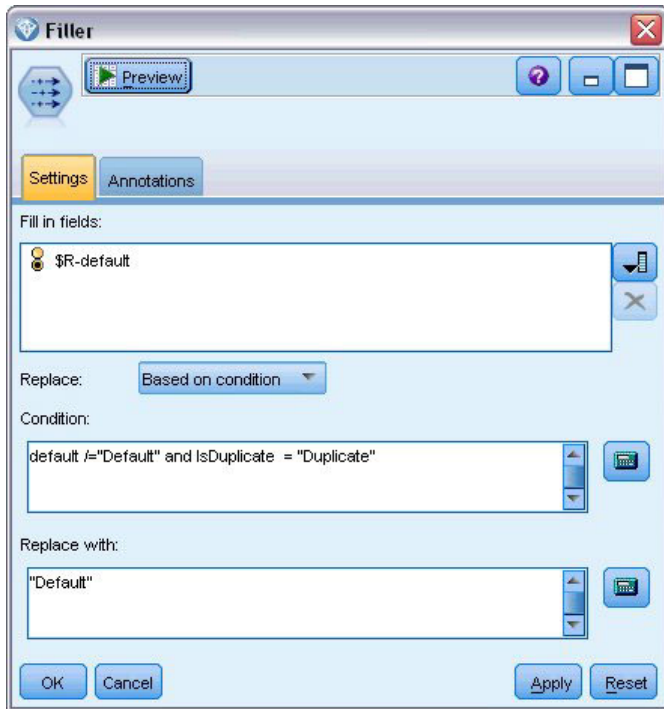


図 20. フィールド値を置き換える条件を指定

これらの設定について、簡単に説明します。この条件は、元のデータ・セットの *default* フィールドの値が **Default** ではない、重複フラグが付いている各レコードについて、モデル内の *\$R-default* フィールドの値を **Default** に設定することを示しています。

*\$R-default* フィールドはモデル内のフィールドで、顧客のローンが債務不履行になる可能性があるかどうかの予測がこのフィールドに格納されます。このようにして、重複レコードを持つ顧客は、債務不履行になる可能性がある顧客としてモデルに追加されます。

「**OK**」をクリックして置換ノードを閉じます。

これで、エンティティ分析による違いを確認する準備が完了しました。「グラフ」パレットから、分布ノードを置換ノードに接続し、分布ノードを開きます。次に、「フィールド」リストをクリックして「**\$R-default**」を選択します。

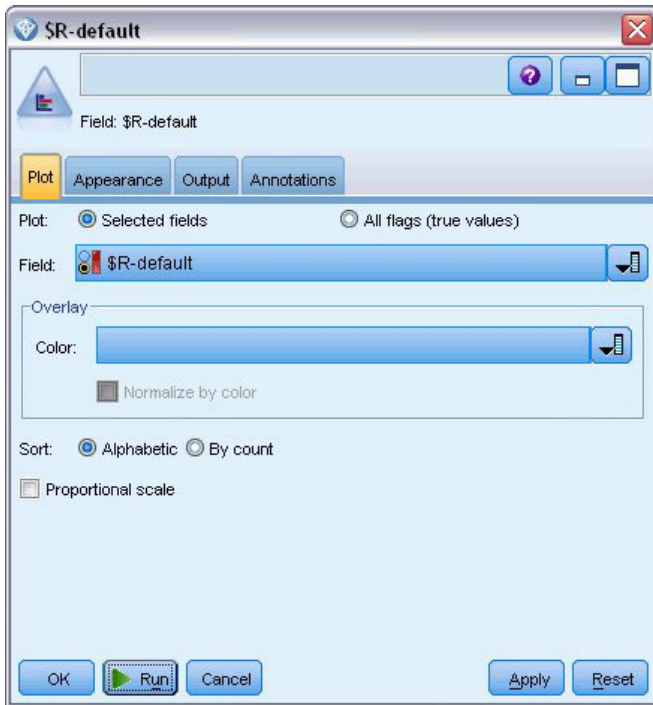


図 21. 分布ノードの設定

「実行」をクリックして新しい予測のグラフを作成します。

Value	Proportion	%	Count
Default		10.67	16
Repaid		89.33	134

図 22. エンティティ分析後の分布ノードからの出力

債務不履行になる可能性がある申請は、13 件から 16 件に増えました。増えた分の申請が実際に債務不履行になれば、かなり大きな損害になる可能性があります。このように、銀行のリスク評価業務にエンティティ分析を追加することの利点をグラフィカルに実証することができます。

## 要約

この例では、エンティティ分析を使用して、人物や組織に関するデータの重複レコードを削除し、予測の品質を改善する方法について説明しました。

注: 重複するレコードは、他のプロセスを実行する前に削除することをお勧めします。その後、自動データ準備 (ADP) ノードを使用して、データの分析と修正の特定、問題があるフィールドまたは有用ではないフィールドの除外、必要に応じた新規属性の派生、高度なスクリーニング技術によるパフォーマンスの改善を行うことができます。

エンティティ分析と自動データ準備を組み合わせると、可能な限りクリーンなデータでの作業が可能になります。

## 付録. IBM SPSS Modeler Entity Analytics のスクリプト記述プロパティ

### IBM SPSS Modeler Entity Analytics でのスクリプト記述

IBM SPSS Modeler Entity Analytics のスクリプト記述機能は、ユーザー・インターフェースでのプロセスを自動化する強力なツールです。スクリプトにより、マウスやキーボードを使用した場合と同じタイプのアクションを実行することができます。また、頻繁に繰り返すタスクや手動での実行に時間がかかるタスクを自動化することができます。スクリプト記述の使用については、IBM SPSS Modeler に付属している *ScriptingAutomation.pdf* ガイドを参照してください。

### 共通プロパティ

以下の表に、IBM SPSS Modeler Entity Analytics のノードの共通プロパティを示します。特定のノードの情報については、これ以降のセクションで説明します。

表 13. 共通プロパティ

プロパティ名	データ型	プロパティの説明
entity_repository	[ <i>field</i> ',' <i>field</i> ', ... , <i>field</i> ']	リポジトリ接続文字列。形式: [ ' <i>reposname</i> ', ' <i>username</i> ', ' <i>password</i> ' ] 例 : entity_repository = [ 'repos1', 'dba', 'psw1' ]
entity_type	<i>string</i>	使用するエンティティ・タイプ (フィーチャー・セット)。 例 : entity_type = 'PERSON'

### entityanalytics\_exportnode プロパティ



EA エクスポート・ノードは、データ・ソースからエンティティ・データを読み取り、エンティティを解決するためにそのデータをリポジトリにエクスポートするターミナル・ノードです。

表 14. entityanalytics\_exportnode プロパティ

entityanalytics_exportnode プロパティ	データ型	プロパティの説明
mode	Add PurgeFirst	エクスポート・モード。Add は、ソース・ファイル・レコードをリポジトリの既存の内容に追加します。PurgeFirst は、エクスポートの前に既存の内容を削除します。

表 14. *entityanalytics\_exportnode* プロパティ (続き)

entityanalytics_exportnode プロパティ	データ型	プロパティの説明
source_tag	string	データ・ソース ID。 例： source_tag = 'CUST'
unique_key_field	string	データ・レコードの一意の識別子で使用する入力フィールド。 例： unique_key_field = 'ID'
field_mapping	[ [ 'field_name' 'feature.element' 'usage_type' ]... ]	入力フィールドをリポジトリ内の対応するフィーチャーにマップします。 例： field_mapping = [['fname' 'NAME.GIVEN_NAME' ''] ['addr1' 'ADDRESS.ADDR1' 'PRIMARY']] 注: usage_type を "(Auto)" と同等に設定するには、上記の最初の例のように ' ' を使用します。

## entityanalytics\_sourcenode プロパティ



エンティティ分析 (EA) ソース・ノードは、リポジトリから解決済みエンティティを読み取り、さらに処理を行うために (レポートの書式設定など)、このデータをストリームに渡します。

表 15. *entityanalytics\_sourcenode* プロパティ

entityanalytics_sourcenode プロパティ	データ型	プロパティの説明
source_tags	list	リポジトリから引き出されるデータ・ソースのタグのリスト。 例： source_tags=[ 'LOANS', 'CUSTOMERS' ]
relationships	None Close All	リポジトリから関係性の詳細を取得するための一致基準。 None は、関係性を返しません。 Close は、分離度などの詳細に応じて、近似一致を返します。 All は、考えられるすべての関係性を返します。
max_degree_separation	integer	最小値 0、最大値 3。
output_entity_type	文字列	リポジトリで使われるエンティティ・タイプのリスト。



## entityanalytics\_processnode プロパティ



ストリーミング EA ノードは、新しいケースをリポジトリ内のエンティティ・データに対して比較します。

表 16. entityanalytics\_processnode プロパティ

entityanalytics_processnode プロパティ	データ型	プロパティの説明
match	Exact ByIdentifier All	リポジトリからエンティティを取得するための一致基準。Exact は、完全一致だけを返します。 ByIdentifier は、完全一致のほかに、同じ ID を共有するエンティティを返します。 All は、考えられる一致をすべて返します。
save_search_records	boolean	
relationships	None Close All	リポジトリから関係性の詳細を取得するための一致基準。None は、関係性を返しません。 Close は、分離度などの詳細に応じて、近似一致を返します。 All は、考えられるすべての関係性を返します。
max_degree_separation	integer	最小値 0、最大値 3。
output_entity_type	文字列	リポジトリで使用されるエンティティ・タイプのリスト。



---

## 特記事項

本情報は全世界で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を提供し、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。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認できません。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をお願いします。

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ます、単に目標を示しているものです。

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。

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](http://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。

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登録商標です。

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。

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。

## 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]

- エクスポート
  - データをエンティティ・リポジトリに 7
- エクスポート・ノード
  - エンティティ分析 7, 13
- エンティティの一致, しきい値の設定 26
- エンティティの一致に対するしきい値, エンティティ分析 26
- エンティティ分析
  - 他の IBM SPSS 製品での使用 33
  - 定義 1
  - 予測分析との比較 2
  - IBM SPSS Modeler での使用 5
- エンティティ分析 (EA) ソース・ノード 27
- エンティティ・タイプ
  - エンティティ分析 24
  - エンティティ・リポジトリ 17
- エンティティ・リポジトリ 13
  - オプション 16
  - 管理者資格情報の管理 34
  - 管理タスク 33
  - 構成 19, 26, 27
  - 削除 38
  - 作成 6, 15
  - 消去 37
  - 新規ケースの比較 29
  - ストリーム・プロパティの設定 36
  - 設定 13
  - 匿名化 23
  - フィーチャー 22
  - 別のストレージ・ディレクトリへの移動 35
  - ポート割り当ての構成 34
  - 保守 20
  - 未使用データの削除 37
  - IBM SPSS Modeler への接続 7

### [カ行]

- 解決済み ID, エンティティ分析による分析 27

- 解決ルール, エンティティ分析 26
- 型情報, エンティティ分析用の設定 28
- 管理者資格情報
  - エンティティ分析用の管理 34
- 構成
  - エンティティ・リポジトリ 19, 26, 27
- 固有キー
  - エンティティ分析 7
  - エンティティ・リポジトリ 17

### [サ行]

- 削除
  - エンティティ・リポジトリ 38
- 消去
  - エンティティ・リポジトリ 37
- 使用タイプ, エンティティ分析 24
- 新規ケース, エンティティ分析リポジトリに対する比較 29
- スクリプト記述
  - プロパティ 57
- ストリーミング EA ノード, エンティティ分析 29
- ストリーム・プロパティ
  - エンティティ分析用の設定 36
- ソース・タグ
  - エンティティ・リポジトリ 17
- ソース・ノード
  - エンティティ分析 9, 27

### [タ行]

- データ・ソース
  - エンティティ分析での接続 6, 14
  - エンティティ分析用の表示 16, 31
- データ・ソース, エンティティ分析用の選択 27

### [ナ行]

- 名前の変更
  - エンティティ分析用のデータ・ワールド 28
- ノード
  - エンティティ分析ストリームへの追加 29

### [ハ行]

- フィーチャー
  - エンティティ・リポジトリ 7, 17, 19, 20, 22, 30, 31
- フィーチャーの匿名化
  - エンティティ・リポジトリ 23
- フィールドのマッピング
  - エンティティ・リポジトリ・フィーチャーとの 7, 17, 19, 20, 30, 31
- プロセス・ノード
  - エンティティ分析 10, 29
- プロパティ
  - スクリプト記述 57
- ポート割り当て
  - エンティティ分析用の構成 34

### [マ行]

- 未使用データの削除
  - エンティティ・リポジトリ 37

### [ラ行]

- リポジトリ
  - エンティティ分析 6, 7, 13, 15, 16, 17, 19, 20, 22, 23, 24, 26, 27, 29, 30, 31, 37, 38
  - エンティティ分析の管理 33
  - エンティティ分析用のストレージ・ディレクトリ, 変更 35

## E

- EA エクスポート・ノード, エンティティ分析 13

## I

- ID の解決, エンティティ分析 7

## O

- output
  - エンティティ分析からの 32









Printed in Japan