IBM SPSS Statistics
パフォーマンス・ベスト・プラクティス

IBM

目次

概要

IBM SPSS Statistics は、データおよび統計分析のためのソフトウェアです。IBM SPSS Statistics により、データを素早く分析できるようになります。IBM SPSS Statistics には、さまざまな分析や検定が含まれており、ビジネスおよび研究上の複雑な課題を解決するのに役立ちます。本書を通じて SPSS Statistics ユーザーは、構成、データ準備、データ分析などの作業に関するベスト・プラクティスを確認できます。これらのベスト・プラクティスにより、SPSS Statistics の効率とパフォーマンスを向上し、より一層の最適化を図れます。

本書では、次の作業を効率化する方法について説明します。

- データの準備、特に自動データ準備 (ADP) の使用について。
- データ変換、コンパイル変換や最高のパフォーマンスを実現する変換のグループ化方法など変換
- データ分析、マルチスレッド化やキャッシュ圧縮など
- 大規模な出力からの有益な情報の抽出
- シンタックスの操作
- SPSS Statistics Server の使用

ADP を使用した自動的なデータの準備

IBM SPSS Statistics には、データ準備機能が含まれており、アクティブなデータ・セットでの異常または無効な状況、変数、およびデータ値を特定できます。また、モデリングのためのデータ準備が可能になります。分析のためにデータを準備することは、どのプロジェクトでも重要な手順ですが、最も時間のかかる手順でもありました。自動データ準備 (ADP) は、データの分析と修正箇所の特定、問題のあるフィールドや不要なフィールドの選別、必要に応じた新しい属性の取得、インテリジェントなスクリーニング手法を用いたパフォーマンスの改善などのタスクをユーザーに代わって処理します。

ADP を使用すれば、モデルを迅速かつ簡単に作成できるようにデータを準備でき、関連する統計の概念についての予備知識も必要ありません。モデルの作成とスコアリングが相対的に速くなり、さらに ADP によって、自動モデル作成プロセスの安定性が向上します。

本書では、ADP の使用に関する概要のみを説明します。詳細については、次の Web サイトからダウンロードできる、SPSS Statistics のドキュメントの「Data preparation」のセクションで確認できます。**ibm.com**/developerworks/spssdevcentral

SQL データベースからデータを取得する SQL プッシュバックの使用

SPSS Statistics Server は、SQL データベースへの並べ替えおよび集計のプッシュバックをサポートしています。並べ替えおよび集計の操作を SQL データベース内で実行するこの機能は、SQL プッシュバックと言われます。大規模なデータ・セットの情報源が SQL データベースの場合、SQL プッシュバックを使用すると、データベースで効率的に実行可能な操作はデータベースで実行されるようになります。

前提条件

SQL プッシュバック機能には、次の前提条件があります。

- SPSS Statistics Server
- SPSS Statistics Server への接続に使用する SPSS Statistics クライアント
- IBM DB2、Microsoft SQL Server、または Oracle Database などの SQL データベース

さらに、SQL 言語の知識がある方は、SQL 照会を修正して並べ替えや集計の処理をデータベース内で実行でき、それにより、SQL プッシュバックと同等のパフォーマンス改善を実現できます。

データ変換のグループ化

ほとんどの場合、未加工のデータが、実行する分析に完全に適していることはありません。事前分析によって、不都合なコード体系やコーディングの誤りが明らかになる可能性があり、さらに変数どうしを実用的に連携させるために、データ変換が必要になる可能性があります。データ変換では、分析用のカテゴリー削減のような単純なタスクから、新しい変数の作成などの高度なタスクを実行できます。代表的なユーザー・ジョブは、データ定義、変換、分析などです。

変換コマンドが分析プロシージャーに分散して組み込まれると、データ変換が反復的に実行されるため効率が低下することは明らかです。この場合、変換のグループ化が必要です。

変換コマンドをグループ化することで、すべての変換作業を一度に実行でき、複数の変換による余分な変換処理コストを削減できます。さらに、シンタックスが整理され、明確で整然となります。どの程度改善されるかは、構成、データ・サイズ、シンタックスなどによって異なりますが、明確な改善が見られます。

コンパイル済みデータ変換

コンパイル済み変換機能は、複雑な変換のパフォーマンスを改善するように設計されています。コンパイル済み変換を使用すると、変換コマンド (COMPUTE や RECODE など) が実行時にマシン・コードにコンパイルされ、パフォーマンスの向上につながります。この機能は、Windows Server で稼働している SPSS Statistics Server でのみ動作します。コンパイル済み変換機能には、次の前提条件があります。

- Windows で稼働する SPSS Statistics Server
- SPSS Statistics Server を構成するための SPSS Statistics管理コンソール
- GNU G++ コンパイラー

変換のコンパイルにはオーバーヘッドが存在するため、コンパイル済み変換は、多数のケースがあるときや複数の変換コマンドがあるときにのみ使用すべきです。

コンパイル済み変換を使用可能にするには、管理者がSPSS Statistics Server 設定を使用してコンパイル済み変換を有効にし、CMPTRANS を YES に設定する必要があります。管理者がコンパイル済み変換を有効にしていないと、警告メッセージが表示されてコマンドが無視されます。

データ分析のベスト・プラクティス

ここで紹介するベスト・プラクティスは、大規模なデータ・セットを効率的に分析し、CPU 負荷の高いプロシージャーの並列処理を強化するのに役立ちます。

大規模なデータ・セット向けのキャッシュ圧縮

大規模なデータ・セットで多くのプロシージャーを実行すると、データ取得のコストが明らかに増大します。アプリケーションは、プロシージャーごとに元のデータ・セットを読み取る必要があります。つまり、データ・テーブルはデータベース・ソースからデータを読み取るため、データの読み取りが必要なコマンドやプロシージャーごとに SQL 照会が再実行されなければならないことを意味します。このオーバーヘッドはキャッシュ圧縮によって回避できます。

データ・キャッシュを作成すると、データを複数回読み取る必要がなくなります。CACHE コマンドが、すべてのデータを一時ディスク・ファイルにコピーして、データを次に使用できるようにします。I/O コスト削減のため、一時データ・ファイルを圧縮することも可能です。CACHE と圧縮を組み合わせると、大規模なデータ・セットを処理する際に効率を改善することができます。

キャッシュ圧縮は、SPSS Statistics Server に接続している場合にのみ動作します。さらに、次の手順を完了します。

- 管理者が SPSS Statistics管理コンソールを使用してこの機能を使用可能にする。
- 分析プロシージャーの前に明示的に CACHE コマンドを発行する。
- シンタックス・ファイルで ZCOMPRESSION を YES に設定する。
- SPSS Statistics Server に接続している状態でシンタックスを実行するか、SPSS Statistics Batch Facility を使用してシンタックスを実行する。

マルチスレッド化

マルチスレッド化とは、1 つのタスクを、並行実行可能な複数のタスクに分ける際に使用される技術用語です。すべての分析プロシージャーに、マルチスレッド化のメリットがあるわけではありません。異なる CPU またはコアで同時に実行できるよう簡単に並列化でき、スケジューリングできるプロシージャーに最もメリットがあります。SPSS Statistics でマルチスレッド化されるプロシージャーを次の表に示します。

プロシージャー・ファミリー	プロシージャー名
相関	2 変量 偏相関
回帰	線形 序数 多項 ロジスティック
データの縮約	因子分析
生存分析	Cox 回帰 ロジスティック回帰
多重代入	欠損値代入

表 1: マルチスレッド化された分析プロシージャー

マルチスレッド化のメリットを得るためには、表 1 に示されたプロシージャーを複数のプロセッサーがある、または各プロセッサーが複数のコアを持つコンピューター上で実行する必要があります。

デフォルトでは、SPSS Statistics は、内部のアルゴリズムを使用して、特定のコンピューターに対するスレッドの数を決定します。この設定は変更できますが、デフォルトは、多くの場合、最善のパフォーマンスをもたらします。コマンド SET THREADS=n を発行することで、デフォルト設定を上書きできます。ここでの n は、スレッドの数、すなわち、多くの場合は対応する CPU またはコアの数を示します。SET THREADSは、次のシナリオでデフォルト設定を上書きするのに適しています。

- デフォルトのスレッド数は、通常、処理単位の数と同じです。このスレッドが CPU リソースを消費すると、CPU 負荷の高い他のアプリケーションに必要な処理サイクルが減らされるおそれがあります。この場合、SET THREADS を使用してスレッド数を制限することができます。
- マルチスレッド化されたプロシージャーの場合、スレッド数が増加すると、データの分離、スレッドの管理、および結果のマージに関するオーバーヘッドも増加するため、パフォーマンスが改善されない場合があります。したがって、コマンド SET THREADS を使用することで、最適なスレッド数を見つけ、それを設定することが推奨されます。

SPSS Statistics クライアントでは最大スレッド数は 4 で、SPSS Statistics Server ではスレッド数の制限はありません。

大量の出力からの必要な情報の抽出

SPSS Statistics は、テーブル、図表、テキストなど、統計結果の表示にさまざまな方法を提供しています。デフォルトでは、結果は、SPSS Statistics Viewer ウィンドウで表示されます。ユーザーはその出力を操作して、必要な出力を的確に表示し、適切に調整と書式設定が行われた出力文書を作成できます。このセクションで紹介するベスト・プラクティスは、こうした目的を達成するのに役立ちます。

複数のプロシージャーを実行すると、SPSS Statistics は、多くの場合、テーブル、図表、ログ、テキストなどで構成される大量の結果を生成します。非常に多くの情報を検討して必要な情報を見つけることは、煩雑な場合があります。幸い、SPSS Statistics は、出力管理システム (OMS) と各種 OUTPUT コマンド (OUTPUT NEW、OUTPUT NAME、OUTPUT ACTIVATE、OUTPUT OPEN、OUTPUT SAVE、OUTPUT CLOSE) を備えており、ユーザーが出力を絞り込み、出力手順を指定するのに役立ちます。

OMS および OUTPUT コマンドは、次の操作に使用できます。

- 大規模な出力を別個の出力文書に区分する。
- 必要な情報を出力から選択して出力手順を指定する。
- 複数の公開出力文書を所定のセッションで使用する。
- OMS で出力を入力として使用する。

OMS および OUTPUT コマンドは、1 つ以上の出力文書をプログラムで管理する機能を提供します。これにより、出力をより簡単に処理できるようになります。詳細については、次の Web サイトにある、SPSS Statistics のドキュメントの「Working with Command Syntax」のセクションを参照してください。
ibm.com/developerworks/spssdevcentral

シンタックスからの不要な EXECUTE の削除

強力なコマンド・シンタックスは、数多くの一般的なタスクの保存および自動化を可能にします。また、メニューやダイアログ・ボックスに表示されていない機能を提供します。さらに、シンタックス・ファイルにジョブを保存できるので、後からその分析を繰り返すことができます。このセクションでは、コマンド・シンタックスを使用するためのベスト・プラクティスを紹介します。

EXECUTE コマンドは、ADD FILES、MATCH FILES、UPDATE、PRINT、WRITE などの変換コマンドおよび変換機能とともに使用するように設計されています。こうしたコマンドおよび機能は、その後にデータ読み取りプロシージャーがない限り、データを読み取らず、実行されることはありません。EXECUTE コマンドによってデータの読み取りが行われるため、不要な EXECUTE コマンドは、データの余分な受け渡しを招き、無駄な時間がかかる場合があります。

不要な EXECUTE コマンドを特定し、削除することで、シンタックスの配置を最適化し、データの読み取りに必要な時間を短縮することができます。この最適化は、特に I/O 負荷の高いプロシージャーに効果的です。

EXECUTE コマンドを削除する際には、EXECUTE の前後のコマンドが完全に独立している必要があります。そうでない場合、結果が変わってしまうおそれがあります。

SPSS Statistics Server を使用した、データの引き渡しコストの削減

SPSS Statistics Server は堅牢で強力な分析ソリューションであり、一部門から、全社的な数百名のユーザーにわたる分析のニーズを処理するシームレスな拡張性を持ちます。SPSS Statistics のすべて機能を提供するばかりでなく、高速なパフォーマンス、大規模なデータ・セットの効率的な処理、エンタープライズの実装における拡張されたセキュリティーを提供します。

拠点が遠隔地に分散した企業の場合、拠点間で大規模なデータ・ファイルにアクセスするため、非常に時間がかかります。大規模なデータのネットワーク上での受け渡しは、帯域幅を飽和させかねず、他のアプリケーションの通常の使用を妨げるおそれがあります。

SPSS Statistics Server を使用すると、データはサーバー・マシンから読み取られ、大規模なデータ・セットがエンド・ユーザーのデスクトップに転送されるのを防ぐことができます。ネットワークを介して転送されるデータが最小限にとどめられ、パフォーマンスが改善されます。これにより、帯域幅の飽和を回避でき、SPSS Statistics のパフォーマンスのほか、電子メール、エンタープライズ・リソース・プランニング (ERP)、カスタマー・リレーションシップ・マネジメント (CRM) といった他の基幹業務のアプリケーションのパフォーマンスが改善されます。

以下の表では、次の場合のデータ・アクセスに要する時間を比較しています。

- SPSS Statistics クライアントはローカル・モードで稼働し、データ・センターのファイルに広域ネットワーク (WAN) を介して直接アクセスします。
- SPSS Statistics クライアントは分散モードで稼働し、データ・センターに実装された SPSS Statistics Server に接続されています。

ファイル・サイズ	SPSS Statistics クライアントがデータに WAN (T1 3.0 Mbps) を介して直接接続	SPSS Statistics クライアントがデータ・センターの SPSS Statistics Server に WAN (T1 3.0 Mbps) を介して接続	秒単位で見た、SPSS Statistics Server によって短縮された時間
50 MB	2 分 10 秒	4 秒	2 分 6 秒
250 MB	10 分 50 秒	40 秒	10 分 10 秒
1 GB	43 分 17 秒	80 分	41 分 57 秒

表 2. 秒単位で見た、データ・ファイルにアクセスするまでの時間

表 2 で示すように、分散拠点でファイルにアクセスするときに SPSS Statistics Server を使用すると、時間を大幅に短縮することができます。例えば、25 MB ファイルの場合で 2 分、250 MB ファイルの場合で 10 分、1 GB ファイルの場合で 42 分の短縮です[1]。

SPSS Statistics Server による 64 ビット・コンピューティング

RAM からデータへアクセスする速度はディスクからデータへアクセスする速度よりもかなり速いので、物理 RAM の容量はパフォーマンスにとって非常に重要です。高速パフォーマンスのためには、データ・セット全体が RAM 内にあることが最も望まれます。しかし、サポートされている RAM の全体の容量はプロセッサーに依存します。理論上では、32 ビット・プロセッサーは、4 GB の RAM へのアクセスに制限されます。64 ビット・マシンに移行すると、RAM の容量を 32 ビット・マシンの数倍に拡大することができます。64 ビット・マシン上の大規模なデータ・セットで分析プロシージャーを実行する場合、速度はさらに速くなります。

SPSS Statistics Server は、Windows Server、IBM AIX、Sun Solaris、Red Hat Enterprise Linux、および SUSE Linux Enterprise Server を含む、さまざまなサーバー・オペレーティング・システムにおける 64 ビット・コンピューティングを強力にサポートしています。ほとんどの分析プロシージャーは、32 ビット SPSS Statistics クライアントよりも 64 ビット SPSS Statistics Server でより高速に実行されます。

一時ファイルに複数の場所を使用

SPSS Statistics Server がデータを処理するとき、多くの場合、そのデータの一時的なコピーがディスクに保持されます。さらに、一部のプロシージャー (CACHE、SORT、AGGREGATE、変換など) は、実行中に一時ファイルを作成する場合があります。一時ファイルのサイズは、データ・ファイルのサイズからデータ・ファイルのサイズの 3 倍まで、さまざまです。一時ファイルは書き込み可能であり、非常に大きくなる可能性があるので、入出力操作を管理することが難しく、I/O 負荷の高いユーザーが同時に複数存在する場合は特に顕著になります。この場合、一時ファイルの場所を複数設定することが必要です。

利点

一時ファイルの場所を複数使用すると、次のことが可能になります。

- アクセス対象のディレクトリーを操作するユーザーに制限を設ける。
- 各ユーザーに割り当てられた一時ファイル・スペースを、パーティション化されたドライブを指定することで制御する。
- 該当の場所が異なるスピンドル上にある場合にパフォーマンスを改善する。このオプションでは、サーバー・ワークステーションが複数の物理ディスクを備えていることが必要です。

結論

本書では、IBM SPSS Statistics の効率、パフォーマンス、および最適化を改善するためのいくつかのベスト・プラクティスを提供しています。これらのベスト・プラクティスには、データ準備、データ変換、データ分析、出力、コマンド・シンタックス、および Statistics Server が含まれます。SPSS Statistics ユーザーは、これらの事例から学ぶことで、自らの作業を最適化し、全体的なパフォーマンスを改善することができます。

SPSS Statistics ソフトウェアのパフォーマンスを改善する方法についての詳細は、次の SPSS コミュニティーから参照することができます。 **ibm.com**/developerworks/spssdevcentral

IBM ビジネス・アナリティクス について

IBM ビジネス・アナリティクス ソフトウェアは、企業のよりスマートな活動と競合他社をしのぐ業績実現を支援する、データに基づいた洞察を提供します。この包括的なポートフォリオには、ビジネス・インテリジェンス、予測分析と意思決定管理、パフォーマンス管理、およびリスク管理のソリューションが含まれています。

ビジネス・アナリティクス ソリューションは、ビジネス・パフォーマンスに重大な影響を及ぼし得る、顧客分析などの分野の傾向やパターンを企業が特定し、視覚化できるようにします。また、シナリオの比較、潜在的な脅威や機会の予測、経営資源に関するより優れた計画、予算作成、および予測、リスクと期待される利益とのバランスの維持、ならびに規制要件を満たすための対応を実現します。企業は、分析を広く利用できるようにすることで、戦術的および戦略的な意思決定をビジネス目標の実現に向けて調整することができます。詳しい情報とお問い合わせ先は、次の Web サイトをご覧ください。 **ibm.com**/software/jp/analytics

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

IBM のホーム・ページはこちらからご覧になれます。
ibm.com

本書の情報は最初の発行日の時点で得られるものであり、予告なしに変更される場合があります。すべての製品が、IBM が営業するすべての国で入手できるわけではありません。

本書に含まれるパフォーマンス・データは、特定の動作および環境条件下で得られたものです。実際の結果は、異なる可能性があります。本書の情報は、現存するままの状態で提供され、商品性の保証、特定目的適合性の保証、および第三者の権利の不侵害の保証も含むすべての明示もしくは暗示の保証責任を負わないものとします。IBM 製品は、IBM 所定の契約書の条項に基づき保証されます。

[1] 結果は、使用可能な帯域幅が 3.0 Mbps であるという仮定に基づきます。電子メール、ネットワーク・バックアップ、他のネットワーク・リソースなどの、ほかのアプリケーションによって帯域幅が取られているため、実際には、短縮される時間はより大きくなります。ここで示すデータは、例として示す目的でのみ提供されています。実際の結果は、WAN の構成、帯域幅、および待ち時間によって異なります。したがって、同様の試験を行う企業は、同じ結果にならない可能性があります。


Please Recycle

YTW03248-JPJA-01