ZL UNIFIED ARCHIVE

グローバル シングルインスタンス ストレージ

ZL TECHNOLOGIES | ホワイトペーパー

ZLTECH

ZL UA のグローバル シングルインスタンス ストレージ (グローバル SIS) 技術は、マルチレイヤーの重複排除機能で、業界トップレベルのストレージ削減、およびメールとファイルの TCO (総所有コスト) の最適化を実現します。

企業が抱えるメールとファイルの容量は、一意のコンテンツ、重複コンテンツ、類似コンテンツを併せると膨大な量になるため、効果的なシングルインスタンス化を行うのは非常に困難です。電子メールは、通常組織の保有コンテンツの大半を占めており、ユーザー、メールサーバー、ファイルのアーカイブ形式 (PST と NSF)、部署、キャプチャ方式、日付範囲などのメールデータの違いに対応することが必要となるため、SIS (シングルインスタンスストレージ) 技術の導入において独特の課題を呈します。ファイルシステムファイルもまた、ファイルの保管と処理を目的とした添付ファイルと独立ファイルのバージョン管理や SIS など特有の問題を伴います。

これらの課題は解決が困難なため、大手ソフトウェアプロバイダーでさえ、シングルインスタンスストレージは諦めて、圧縮技術のみを使用した比較的効果の低いソリューションに甘んじているのが現状です。ZL UA は、高度な研究開発を通じて直接課題に取り組むことで、業界をリードするストレージ削減率を実現します。

このホワイトペーパーでは、真のストレージ最適化を実現するために ZL UA が業界で最も広範かつ粒度の細かいハッシングを実行する仕組みについて説明します。この手法により、メッセージと添付ファイルの両レベルでグローバル SIS を実行し、実稼働サーバーで最大 80% のデータを削減できます。

SIS の概要

さまざまな調査によると、大企業の場合、1 通のメールが企業ネットワーク全体で複製される平均回数は 8 回という結果が報告されています。その主な原因には、膨大な配信リスト、参考情報の送信や CC 宛のメール、社内の一括大量メール送信などがあります。メッセージ本文に加えて、添付ファイルもかなりのストレージ容量を占め、多くのメールやファイルシステム内で複製されることも少なくありません。

SIS の基本的な仕組みとしては、まずシングルインスタンス化の対象となるアイテム (メール、添付ファイル、ファイルシステムファイルなど) のコンテンツのダイジェスト (ハッシュ値) を算出し、そのハッシュ値を SIS ルックアップテーブルで検索します。一致データが検出された場合、そのアイテムは既にボルト内に存在しているため、既存コンテンツに追加ポインターが加えられます。一致データが見つからない場合は、アーカイブ内にアイテムが存在していないため、新しいコンテンツがボルトに書き込まれ、SIS テーブルに 行が追加されます。

各アイテムに対し、データベースに既存コンテンツのコピー数を表す現行参照数が記録され、参照数がゼロになるとアイテムが除去されます。

グローバル SIS とサイロ/ハイブリッド構成の SIS の比較

SIS ルックアップテーブルのサイズは、SIS のスケーラビリティを左右する主な要素となります。効率的なグローバル SIS では、スピードと一貫性を確保するため、1 つの大規模なルックアップテーブルが使用されます。ZL UA はこの手法を採用しているため、膨大数のアイテムに対応可能です。ZL UA は、柔軟なコンピューティンググリッドアーキテクチャと大規模なデータベースアーキテクチャを基盤とするため、全社レベルのグローバルなシングルインスタンスストレージに対応できます。

これに対して、サイロ化された複数の SIS ルックアップテーブルを使用する手法もあります。サイロ構成のルックアップテーブルは、効率的にコンテンツをパーティションに分割できる場合には有効ですが、パーティション分割は非常に煩雑なプロセスです。このモデルを導入した組織からは、大規模な展開の場合、SIS によるストレージ削減はほとんど、またはまったく達成されなかったことが報告されています。この手法で効果が得られないため、SIS を完全に排除して、有効性の劣る圧縮技術に依存した手法に切り換える組織も少なくありません。

サイロ構成の SIS を導入した組織では、効果の低下を補うために、サイロ化された複数のデータベースの連携を試みたケースもあります。この新たなモデルは、個々のデータベース間での複雑な調整作業が必要となるため、パフォーマンスやデータ損失などの問題を生じやすくなります。

ZL UA は、ストレージの効率性と信頼性の両観点から見て最も効果的な手法であるグローバル SIS をサポートします。

SIS マッチング技術

ZL UA が提供するマルチレベルの SIS を組み合わせて導入することで、業界でも比類ないストレージ削減レベルを達成できます。

完全重複データの排除

完全な重複データが組織内に多数存在していることがあります。これは、複数サーバー上に完全な重複メールのコピーが多数作成されるメールのジャーナルシステムなどで頻繁に見られる問題です。これらのファイルのアーカイブと処理には膨大なリソースが費やされるため、ZL はできる限り早期に重複データを処理して排除することで、運用コストとストレージコストの削減を図ります。

ファイルコンテンツの基本的なハッシュ値検証は、シンプルな完全重複データの識別とファイル整合性の検証に広く用いられています。しかし、アーカイブ目的では、ファイル外部のメタデータ (メール送信者や受信者から派生する部署データなど) における差異を、保管と処理ポリシー用に個別に追跡することが必要になることがあります。その場合、異なるポリシーの下に重複データがアーカイブされ、メッセージ SIS を使ってシングルインスタンス化されます。

ZL UA の完全重複データの排除機能は、完全重複データを破棄できる場合と、組織内外のポリシーに確実に準拠するために重複データをアーカイブしなければならない場合の両方に対応します。

類似データ用メッセージ SIS

メールサーバー内のメールでは一意であっても、メールフローにおいて各クライアントによりメールに変更が加えられます。その結果、1 通のメールに対し多数の類似コピーが存在することになります。類似コピーはほぼ同一ですが、メールサーバーによるメールの処理方法と変換形式により差異が生じます。企業内では膨大量のメールが生成されるため、これらの類似コピーを効果的に取り扱うことが、SIS においては極めて重要な要素となります。

ZL では、類似コピーの処理のため柔軟なマッチング技術を採用しており、ハッシュ値の計算を実行する前にまず比較値が正規化されます。値の正規化では、テキストを小文字に変換し、余分なスペースと HTML タグを削除するなどの作業が行われます。また、タイムゾーンの違いを調整するため、日付フィールドも GMT に正規化します。重要度の低いヘッダー (スパムチェックやウィルスチェックなど) は、組織のアーカイブの要件としてほとんど価値がないため、ハッシュ値計算では無視されます。

ZL はメールのさまざまなフィールドを対象に、多くのハッシュを実行します。

1. 本文ダイジェスト - メールの本文から取得されるハッシュ値
2. 添付ダイジェスト - メールのすべての添付ファイルから取得されるハッシュ値
3. 基本ヘッダーダイジェスト – さまざまなヘッダーフィールドから取得されるハッシュ値
4. エンベロープヘッダーダイジェスト – 送信者および受信者フィールドから取得されるハッシュ値

メッセージ SIS を実行するには、上記 4 つのハッシュ値すべてが同一であることが条件になります。

さらに柔軟なマッチングオプションとして、ZL はアグレッシブモードと非アグレッシブモードの 2 種類の SIS を提供します。非アグレッシブ SIS では、基本ヘッダーダイジェストに日付フィールドが含まれます。アグレッシブ SIS では、基本ヘッダーのハッシュ値計算に日付フィールドは含まれず、その後にシングルインスタンス化されたメッセージはすべてデータベース内で別々に保管されます。

SIS マッチングで一致データが検出されると、記録を必要とする非重複データは ZL UA のデルタフィールド技術を使って保管されます。これにより、特定の情報が ZL UA ボルト外部に保管され、該当アイテムの取得時にデータが再び付加されるようになります。

メールメッセージとスタンドアロンファイルを対象とした添付 SIS

企業がアーカイブに保管してデータ保持管理ポリシーの対象として扱わなければならないコンテンツは増大しつつあり、その大半を占めているのはメールメッセージです。一方、平均的なメールサイズも膨大化しており、その主な原因となっているのは添付ファイルです。PowerPoint プレゼンテーションや PDF ファイルをはじめとし、サイズの大きな各種ファイルが頻繁にメールで送信されるようになり、1 通のメールがメガバイト単位のサイズになることも珍しくありません。さらに、送信者側または受信者側で、添付ファイルをメール外部で表示して編集し、各種のファイルシステム上で独立したドキュメントとして保管されることも多くあります。これらの要素を考慮すると、添付ファイルのシングルインスタンス化により、メールとスタンドアロンファイルの両方を対象に劇的なストレージ最適化機能を提供できます。

ZL にメッセージが取り込まれると、添付ファイルの解析が実行されます。ZL では、シングルインスタンスストレージの最適化のため、 添付 SIS の実行対象となる添付ファイルの最小サイズを設定できます。添付ファイルのサイズが非常に小さい場合は、メールと一緒に保管するよりもシングルインスタンス化する方が多くのストレージが必要になることもあるため、この設定により一定サイズ以下の添付ファイルを SIS の対象外とすることができます。指定サイズを超える添付ファイルはすべてメッセージから抽出されて、個別に保管されます。

抽出された添付ファイルがシングルインスタンス化の対象となります。添付ファイルは Raw バイナリ形式に抽出され、各ファイルのハッシュ値が算出されます。メッセージ SIS のプロセスと同様に、添付 SIS でも ZL がハッシュ値を検索します。ハッシュ値がテーブルに含まれている場合は、添付ファイルが既にアーカイブに保管されているため、ポインターが追加されます。ハッシュ値が見つからない場合は、添付ファイルが Raw バイナリ形式で保管され、添付 SIS テーブルにレコードが追加されます。

ファイルシステム SIS の注意点

データの保持、処理、検出、訴訟ホールドなどの目的で、アーカイブを利用してファイルシステムファイルを管理することを検討する場合、ファイルシステム特有の要件を考慮する必要があります。これには、ファイル名、保存場所、変更日、アクセス制御などの要件が含まれます。例えば、同じファイルが異なるユーザーにより違う名前で保存されている場合があります。ZL UA のファイルシステムアーカイビングでは、SIS による効果的なストレージ削減を図るため、このメタデータが無視されて透過的に処理される一方で、情報が失われることがないようにアーカイブに保管されます。

ZL UA File System Archiving は、ファイルのバージョン管理もサポートするため、コンテンツ管理のため、同一ファイルの複数のバージョンをキャプチャ、検索、参照できます。

結論

増え続ける電子メールやその他の非構造化コンテンツを管理する上で、さまざまなレベルと形式で存在する重複コンテンツへの取り組みは不可欠です。ZL UA は、細かい粒度でシングルインスタンス化を効果的かつ透過的に管理します。全社レベルの完全重複データ、類似データ、添付ファイル、バージョン管理に対応し、各ファイルを正確、安全かつ効率的にシングルインスタンス化し、最大 80% の SIS 削減を実現します。ZL UA のデルタフィールド機能などの構成オプションを利用すると、データ保持の要件に合わせて、シングルインスタンス処理の詳細を透過的に調整して、最適なレベルでストレージを削減できます。ZL のユニファイドされた完全統合型のスケーラブルなアーキテクチャにより、データベースストレージ、処理能力、信頼性を低下することなく、グローバル SIS を実行できるようになります。ZL はスピード、スケーラビリティ、TCO といったあらゆる面で市場トップに位置付けられていますが、その重要な鍵となっているのがグローバル SIS です。