SG88-4064-00
（英文原典：REDP-4704-00）

Wei-Dong Zhu
Markus Lorch

IBM Content Collector デプロイメントとパフォーマンス・チューニング

概要

この IBM® Redpaper™ では、IBM Content Collector のデプロイメントとパフォーマンス・チューニングを計画する場合に考慮する必要のある重要な点について説明します。

この資料は、2 つの大きな部分に分かれています。最初の部分では、適切なシステム設計とアーカイブ・ポリシーのために考慮する必要のある質問と、実装に役立つサポート情報が紹介されます。設計段階での選択が及ぼす影響について理解するために、さまざまな段階で生成されるワークロードと、フロントエンド・システム、Content Collector、およびバックエンド・システムでの各コンポーネントのアーカイブ・プロセスについて説明します。

この資料の 2 番目の部分では、Content Collector と連動して最適なパフォーマンスを得るために、アーカイブ・システムのモニターおよびチューニングなどの操作について説明します。コマンドとその説明も掲載されており、ログ出力を分析し、この情報を使用してアーカイブ・ポリシーとシステム設定を調整することができます。システムのスループットを測定する方法と、共通のボトルネックを診断する簡単な方法について説明します。

この資料では、特に以下のトピックについて取り上げます。

- コンプライアンス・アーカイブ・シナリオの計画
- モニタリングとヘルス・チェック
- パフォーマンスとスケーラビリティー
- 高可用性およびロード・バランシングの計画
- バックアップおよび災害復旧の計画

コンプライアンス・アーカイブ・シナリオの計画

ほとんどの場合、Content Collector をデプロイする主な目的は、E メール・サーバー、共有ファイル・システム、および Microsoft® SharePoint などのグループウェア・ソリューションのコンテンツをアーカイブすることです。Content Collector は、ソース・システム (E メール・サーバーなど) の代わりになることを目的として設計されてはいませんが、システム・オペレーターが、使用頻度がどの程度低いデータをアーカイブしてソース・システムから削除するかに関するポリシーを定義できるようにします。

ユーザーの要件を正しく引き出して理解することは重要です。Content Collector ソリューションが提供できるサービスに対する期待も現実的なレベルに設定する必要があります。ここで、注意すべき重要な点は、ソース・システムで現在使用されているコンテンツをそのシステムから削除 *してはならない* ということです。現在使用されているコンテンツを削除すると、元のシステムの機能が低下し、生産性が落ち、アーカイブ・システムに対するユーザーの許容度が下がる結果になります。理解しておくべき 2 番目の点は、コンテンツをアーカイブする際、ソース・システムでは、データの読み取り、マーク付け、ある場合はプレースホルダーとの置き換えなどが行われるため、追加の負荷が発生するということです。

アーカイブ・ソリューションは、多くの場合、データをもともと格納していたシステムのストレージ・コストを減らし、システム・パフォーマンスを高めることに重点を置いた技術的なプロジェクトです。技術的な面だけに重点を置くと、選択される特定の設計が、アーカイブ・データを使用する必要のあるユーザーに与える影響を忘れがちです。ユーザーは、データがアーカイブされるとデータに直接アクセスできなくなり、アクセス遅延が生じるようになるため、多くの場合、コンテンツがアーカイブされるのを好みません。そのため、技術的な目的に加えて、ユーザーの要件を満たすシステムにすることも重要です。ユーザーを忘れるべきでない分かりやすい例として、メールボックスの管理シナリオがあります。メールボックスの管理システムとアーカイブ・ポリシーを設計する場合、あまりに早い段階でコンテンツをアーカイブして、ユーザーの生産性を落とさないようにすることが重要です。さらに、エンタープライズ・アーカイブから切断された状態で頻繁に作業するユーザーのことも忘れないようにすべきです。そのためには、オフライン・リポジトリーのサポートを計画に含める必要があります。適切に設計されたアーカイブ・ポリシーでは、頻繁

に使用されるデータに直接アクセスできるようにしつつ、ストレージのスペースを大幅に減らすことができます。

シンプルに保つ

この一般的な設計原則は、アーカイブ・システムにおいて重要です。作成されるポリシーは、特にユーザーがアーカイブの効果を実感できる場合、理解しやすいものであるべきであり、システムの動作が許容できるものでなければなりません。アーカイブの選択とライフサイクルのルールを複雑にしないことによっても、アーカイブに伴う負荷の影響を抑えることができる場合があります。アーカイブ・ポリシーが少ないほど、システムの管理は容易です。

それで、基本的な設定から始めることをお勧めします。経験を積むにつれ、必要に応じて複雑なものにしていくことができます。そのお客様の環境で必要とされる機能のサブセットを選択してください。アーカイブ・ポリシーは、理解しやすく実行しやすいシンプルなものにします。ライフサイクルの管理ステージは少ない数に保ってください。

要件を集める

アーカイブ・システムのデプロイメントの計画を始める前に、アーカイブ・システムが対処する必要のある要件と運用される環境を理解しておくことは重要です。これらの要件を集め、アーカイブ・システムのアーキテクチャーと運用の影響を十分に理解するのを助けるため、一連の質問と説明を用意しました。それらを以下に示します。考え得るすべてのシナリオと要件の組み合わせを取り上げることはできませんが、以下のトピックによって、システムの設計に影響を与える最も重要な要件には対処できます。

- アーカイブ・ソリューションの主な目的は何か？
- アーカイブするデータにはどんな特徴があり、ボリュームはどれくらいか？
- アーカイブに費やせる時間はどれくらいか？

アーカイブ・ソリューションの主な目的は何か

まず最初に、アーカイブの主な目的を理解する必要があります。しばしば必要となる主な要件は以下のとおりです。

- 法的な要件を満たし、特定のユーザーにアーカイブされたコンテンツでの全文検索機能を提供するためのアーカイブ (通常、コンプライアンス・アーカイブ・シナリオおよび eDiscovery シナリオと呼ばれる)
- ファイル、E メール、またはグループウェア・サーバーで必要とされるストレージを減らし、ユーザーにアーカイブされたコンテンツでの全文検索機能を提供するためのアーカイブ (E メール・アーカイブ・シナリオ、またはメールボックス管理シナリオとも呼ばれる)

- コンテンツをビジネス・プロセスおよびレコード管理シナリオで使用できるようにするためのアーカイブ

一般に、コンプライアンス・アーカイブ・シナリオでは、ユーザーがコンテンツを使用して作業する方法には影響せず、システムのユーザーの数は限られたものになります。そのため、計画とデプロイはシンプルです。他のコンテンツ・アーカイブ・シナリオでは、ユーザーのユース・ケースおよび要件を考慮に入れる必要があります。「メールボックス管理の計画の考慮事項」(5 ページ)では、E メールのメールボックス管理シナリオを計画する場合に注意すべき点について説明します。同様の点は、E メール以外のコンテンツ・アーカイブ・シナリオに適用されます。

アーカイブするデータにはどんな特徴があり、ボリュームはどれくらいか

アーカイブするデータのボリュームと特徴を知っておくことは重要な要素です。特に、次の点を理解しておくべきです。

- 日単位のアーカイブ・ボリューム : 日単位でアーカイブする必要のあるアイテム (E メールまたはファイル) の数 (例えば、1 日の間にアーカイブされる E メール・ジャーナルのアイテム数およびユーザーのメールボックスのアイテム数)
- 初期アーカイブ・ボリューム : プロジェクトを開始する時点で、既存のデータ・バックログの一部としてアーカイブする必要のあるアイテム (E メールまたはファイル) の数
- 重複の割合 : 重複するアイテム (重複する E メール、ファイル、および E メールの添付ファイル) の予想される割合

  例えば、1 通の E メールに平均 5 人の受信者がいる場合、メールボックス管理シナリオでは、このメールが 6 通アーカイブされることを予測します (送信分が 1 通と受信分が 5 通)。E メール・ジャーナルもアーカイブされる場合、7 通目が処理されます。予想される重複数 (合計数 - 1) を知っておくと、正確なストレージ要件を予測するのに非常に役立ちます。
- アーカイブされるアイテムの平均サイズ、特にファイルまたは E メールの添付ファイルの予想サイズ。これらは、必要とされるストレージの大部分を使用します。

上記の詳細情報を収集する 1 つの方法は、実際のデータの代表的なサンプルを使用して、概念を実証するアーカイブを実行することです。Content Collector 監査ログを使用して、E メールの属性 (受信者、E メールのファイル・サイズ、および E メールの添付ファイルの数など) に関する情報を記録できます。E メール・アーカイブ・シナリオでは、Content Collector を実行する仮想マシン、および E メールの収集と、メタデータおよび添付ファイルの抽出のみを行い、データをアーカイブしないタスク経路を使用して、こうした情報のほとんどを収集できます。特性に関する情報に加えて、小規模なアーカイブ・テストを実行することによって (物理的なハードウェアで実行する場合)、既存の E メール・

サーバーまたはファイル・サーバー・システムから発生する可能性のあるスループットを判別することもできます。「Content Collector のログ」(21 ページ) では、Content Collector 監査ログを使用する方法が説明されています。

アーカイブに費やせる時間はどれくらいか

アーカイブに費やせる時間の量は、システム設計およびアーキテクチャーにおいて重要な要素になります。以下について知っておく必要があります。

- アーカイブに費やせる作業日ごとの時間
- 初期のバックログのアーカイブに計画される時間
- E メール・システムおよびエンタープライズ・コンテンツ管理アーカイブ・システムの他の使用のための時間 (バックアップ、再編成、および他のアクティビティー)

高ボリュームのアーカイブ (バックログのアーカイブまたはメールボックス管理) を実行する場合、このワークロードは作業が行われていない時間にスケジュールする必要があるかもしれません。これまでの経験から、作業が行われていない時間の大半をバックアップおよびデータベースの再編成のために費やすため、会社の E メール・システムには日常的に使用率が下がる時間がほとんどまたはまったくない可能性があります。ファイル・サーバーの場合には、概してアーカイブに費やせる多くの時間があります。

ヒント : 関係するすべてのシステムと、それらのシステムの時間の経過に伴う使用状況を記した時間表をまとめて、アーカイブに適した時間帯を計画するのが良い方法です。

これらのアクティビティーを行うために週末を利用することを考慮してください。定期的なメンテナンスのための時間を計画に含めることを忘れないでください。

メールボックス管理の計画の考慮事項

メールボックス管理シナリオのビジネス上の推進力となるのは、多くの場合ストレージ・コストの削減です。推進力となることがある別の点としては、ユーザーが管理するローカル・メール・アーカイブ (ローカルの PST または NSF ファイル) と E メール・コンテンツの高機能な検索メカニズムを統合することが挙げられます。このセクションでは、メールボックス管理シナリオを計画するときに考慮すべき推奨事項と実際の経験を紹介します。

ストレージのコスト削減がアーカイブを導入する主な目的のひとつである場合、高性能の E メール・サーバー・ストレージから安価なアーカイブ・ストレージにすべての E メールを移動すると逆効果となる場合があります。メールをアーカイブに格納することによって、必要なストレージ総量が増える可能性があり

ます。これは、非常に小さいメールの場合に特にそう言えます。アーカイブ・データベースのストレージ要件は、ある程度まではメール・サイズとは無関係であるため、小さなメールのスタブを保存する場合は E メール・サーバーのストレージは大きく減少しない可能性があります。

ただし、すべてのメールジがアーカイブされないと、アーカイブの検索で完全な結果が得られません。多くの場合、小さなメールもアーカイブすることによるストレージ量の増加よりも、アーカイブですべてのメールを検索できる利点の方が重要です。

E メール・サーバーのストレージを削減するためには、単一インスタンスの添付ファイル・ストレージのメカニズムを使用するすべてのメールボックスを、アーカイブの対象として考慮する必要があります。単一インスタンスのストレージにある特定の添付ファイルへの参照を保持するメールを部分的にアーカイブすると、添付ファイルは E メール・サーバー・ストレージに残ります。

Content Collector は常に E メール全体をアーカイブし (添付ファイルだけをアーカイブするオプションはありません)、オリジナルの E メールに対する様々な E メール・スタブ・オプションが用意されています。例えば、エンタープライズ・アーカイブにメール全体のアーカイブ・コピーを作成した後に、元の E メールから添付ファイルを削除するよう、Content Collector を構成することができます。アーカイブでは、E メールと添付ファイルが別々に格納されます。こうすることで、Content Collector はエンタープライズ・コンテンツ管理アーカイブと連動して、重複するメールを *1 回だけ* 格納することに加え、同じ添付ファイルが何通の E メールで使用されていようと、固有の添付ファイルそれぞれを 1 つだけ格納することが可能になります。

会社は、多くの場合、メールボックス・ストレージを直ちに削減することを望んでいるため、アーカイブされたコンテンツをメールボックスから削除したいと考えています。E メール・システムが機能低下してユーザーの作業効率が低下する結果にならないようにするためには、ユーザーのメールボックスから E メールを削除することは*慎重に計画する* 必要があります。ユーザーが頻繁に E メールにアクセスする間は、E メール・システムの E メールを常時使用できることが通常必要です。E メール・アーカイブは、E メール・システムに置き換わることを目的としていません。

ほとんどのメールボックス管理シナリオでは、ユーザーが頻繁にアクセスする間は、アーカイブされたメールのメール本文を完全な状態で保持しておくのが有益です。頻繁にアクセスが行われる期間は、業種によって異なります。一般には、3 カ月から 6 カ月の期間と考えられています。それでも、E メールの早期のアーカイブは実行可能です。しかし、それはスタブ処理を行って意図的に遅延を生じさせる結果となるにすぎません。

すぐにスペースを節約するには、E メールの添付ファイルを、アーカイブしたファイルへの参照に置き換えることができます。このスタブ・オプションは、E メールをそのまま完全な形で保つため、最初のアーカイブの際に直接適用で

きます。このオプションでは、E メールに対する 2 次的な変更、たとえば、30 日後にアーカイブを行い、90 日後にスタブ化を行うとすると、通常、Content Collector は、2 回に分けて、メールにアクセスして修正を行わなければならなくなりますが、そのような必要はなく、アーカイブ・タスク経路で直接適用できます。ユーザーのメールボックスには、元の形式で完全な E メールの本文が残されており、元のメールを復元することなく、それらの E メールに対して返信、転送、および他の操作を実行することができます。元の添付ファイルにはリンクを通じてアクセスできます。そのリンクには、ユーザーの組織に属するすべてのユーザーがアクセスできます。アーカイブされた添付ファイルには、そのメールの転送または返信を受け取ったユーザーがアクセスできます。ただし、エンタープライズ・アーカイブにネットワークでアクセスできる場合に限られます。E メールのスタブの URL には、一般に会社の外部のネットワークからはアクセスできないため、外部の受信者に転送される E メールに添付ファイルのデータを組み込むには E メールのスタブを復元する必要があります。

実際のデプロイメントでは、アーカイブされた E メールから添付ファイルを除去するだけで、ユーザーが E メールを使用して作業する方法にほとんど影響を与えることなく、E メール・サーバーの E メール・ストレージを最大で 80 % 削減できます。

オフライン・リポジトリー機能が必要な場合、これらのメールからコンテンツを除去する前に、アーカイブされたメールのローカル・コピーを作成するための時間をオフライン・リポジトリー機能が必要とするため、E メールをアーカイブの際に即時にスタブ化してはなりません。

E メール本文をそのまま残すと、ユーザーは E メール本文のテキストを検索するときに、E メール・クライアントのローカルの検索機能を使用できます。こうすると、E メール・アーカイブでの検索の負荷を大幅に減らすことができます。高頻度から中頻度のアクセスがある間に E メール本文をユーザーのメールボックスに保持しておく場合、E メール・アーカイブの検索機能のためのシステム要件を削減できます。

図 1 は、実際のデプロイメントから得られた、サンプルの 3400 通の E メールに関する、その平均的なサイズの分布とストレージ・サイズの分布を示したヒストグラムです。

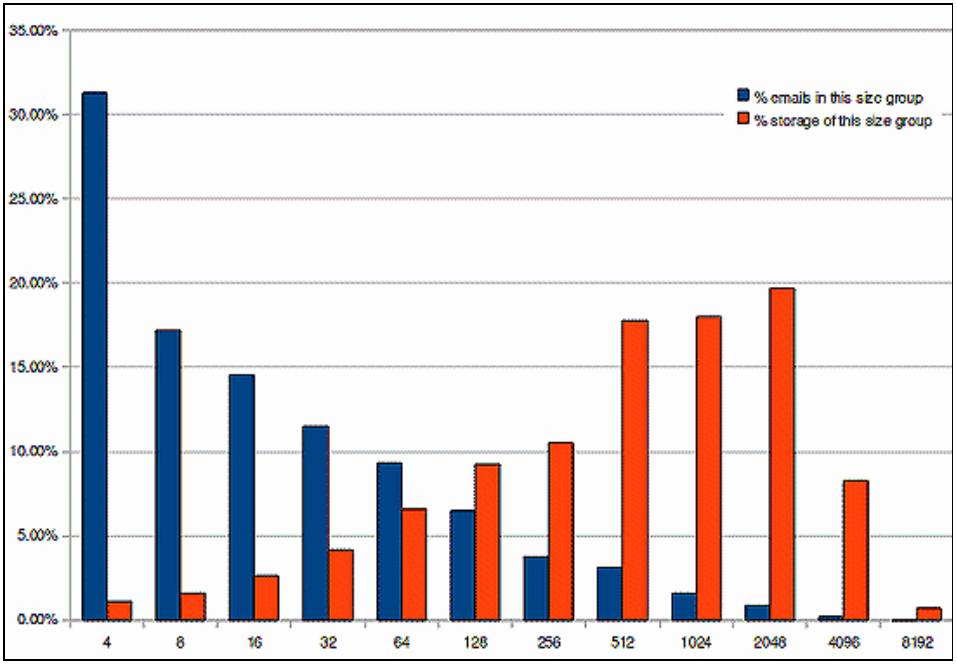

図 1　本文のサイズとサイズ・グループ (4 KB から 8 MB) ごとの E メールのヒストグラム

X 軸は、サンプル・データ・セットを異なるサイズ・グループ (4 KB から 8 MB) に分け、各グループに属する E メール数の割合と、そのグループで消費されている E メール・サーバーのストレージの割合を示しています。ここで、64 KB 未満の E メール数は全体の約 80% を占めますが、ストレージについては全体の約 15% に過ぎないということに注目してください。これは、大半のストレージは残りの 20% の E メールによって消費されていることを意味します。

この分析例は、数日間分のアーカイブに関する標準的な Content Collector 監査ログの E メール・サイズのデータに基づいています。添付ファイルの数は示されていません。他の経験上、実質的にサイズの大きな E メールには添付ファイルが含まれており、大半のストレージはこれらの添付ファイルによって使用されています (例えば、ほとんどの業界で添付ファイルを含まない E メールの本文が 64 KB を越えることは一般にありません)。E メール本文から添付ファイルを除去することによって、既存の E メール・システムが必要とするストレージの大半を解放することができるという結論を導き出すことができます。IBM Lotus® Domino® Attachment and Object Storage (DAOS) など、重複を検出する単一インスタンス・ストレージを採用している E メール・サーバー・システムでは、こうしたサイズの大きな E メールで消費される実際のストレージは大幅に減少します。

E メールを頻繁に使用する必要がなくなれば、E メールをスタブ化するか、完全に削除することができます。E メールがさまざまなアーカイブ・ステージ

(アーカイブ、アーカイブして添付ファイルを除去、アーカイブして本文を短縮、アーカイブして削除) を順に移行するプロセスを *ライフサイクル・スタブ化* と呼びます。E メールのライフサイクル・スタブ化を使用する場合、ライフサイクルのすべてのステージで、追加の処理による影響が出ることを覚えておいてください。Content Collector サーバーは、コレクターが実行するスケジュールされた各ライフサイクルの該当するすべてのメールボックスで、新しいライフサイクル状態に移行する必要のあるメールをスキャンして、ユーザーのメールボックスの対応する E メールを修正する必要があります。このアクションは、E メール・サーバーに大きな負荷をかけるため、注意深く計画する必要があります。E メールのスタブを削除する主な理由は、ユーザーの受信ボックス・フォルダー内のアイテム数が、ベンダー推奨の最大値を超えないようにするためです (特に、Microsoft Exchange の推奨事項を確認してください。これは、使用される Microsoft Exchange のバージョンによって異なります)。

多くの使用法シナリオでは、E メールの本文全体を保持することをお勧めします。E メールの本文の省略は、E メールのライフサイクルで、アーカイブして添付ファイルを除去した後にくる自然なステップのように思われます。しかし、E メール本文の省略、特に削除のメリットは、デメリットと注意深く比較する必要があります。例えば、本文を省略すると、E メールのストレージ要件は削減されますが、比較的小さな量に過ぎません (数 KB 程度のメールや、いくつかのシナリオでよくある非常に小さなサイズのメールについては、実質的に小さくなりません)。一方で、本文を省略すると、形式化された情報が失われ、ローカルの E メール・クライアントの検索機能は省略された E メールを確実には検出できなくなるため、E メールの検出率が大幅に下がります。

パフォーマンスの観点から言うと、セットアップ時には最初の高頻度アクセス期間が過ぎた E メールだけをアーカイブするため、E メール・システムとアーカイブ・システムで生じる影響は最小限になります。その後、ユーザーは不要になった E メールを削除することができ、処理されるメールの量を減らすことができます。さらにこのシナリオの最初のアーカイブ・ステップには、E メール・システムから添付ファイルを削除して、E メール本文の残りをそのまま残すという、E メールの即時スタブ化を実行するタスクが含まれます。残されたテキスト本文 (比較的小さい) は、E メールの一般的な存続期間ユーザーのメールボックスに保たれます。2 番目であり最終のステップ (*スタブ削除* と呼ばれる) では、この E メールが再び使用される可能性はないと思われるとき (例えば、6 カ月後) に、ユーザーの E メールボックスから完全に削除されます。このスキームでは、アーカイブの間に E メール・サーバーから E メールが取得され変更されるのは 1 回だけであるため、E メール・サーバーに与える影響は最小限に抑えられます。E メール・サーバーでの 2 番目であり最終の操作は、使用が予想される期間の最後に実行されるスタブ削除です。スタブが削除された後は、ユーザーは E メールを検出して参照するには、アーカイブ検索を使用する必要があります。

推奨事項

まとめとして、メールボックス・アーカイブでの推奨事項を以下に示します。

- メールのアーカイブは、早くても受信してから 30 日後にし、不要な E メールの削除を行えるようにしたり、高頻度で使われるようなデータをアーカイブしてアクセスされないようにしたりします。
- 初期アーカイブ・ステップの際に添付ファイルを削除してハイパーリンクで置き換えることによって、E メール・サーバーでのディスク使用量を減らし、E メール・サーバーでのアーカイブの負荷を最小限に保ちます。
- メール本文は、E メールが E メールボックスに保持されている間は常にそのまま残します。
- メール本文の削除は、通常、使用されなくなった後 (通常は受信して 3 から 12 カ月後) に行います。
- オフライン・リポジトリー機能が必要かどうかを検討し、オフライン・リポジトリー・ユーザーにはアーカイブとスタブ化を分けるための代替のアーカイブ・タスク経路を使用します。

Content Collector サーバーのワークロード・コンポーネント

Content Collector デプロイメントを計画し、後からチューニングするには、E メール・アーカイブのデプロイメントの全体的なワークロードを構成する個々のワークロードを理解する必要があります。

Content Collector サーバーのワークロードは以下のようなカテゴリーに分けることができます。

- アーカイブ・ワークロード

  このワークロードには、ソースとなる E メール・サーバーで E メールのスキャンと読み取りを行い、Content Collector サーバーでこれらの E メールを処理し、E メールをエンタープライズ・アーカイブへ保存し、E メールを E メール・システムでアーカイブ済みとしてマークを付けてこのトランザクションを完了させることが含まれます。場合によっては、ソース・システムから添付ファイルを削除することや、メール全体を削除することも含まれます。

  このワークロードは、通常 Content Collector およびエンタープライズ・アーカイブ・システムの CPU とストレージに関する要件に最も影響するワークロードですが、E メール、アーカイブ、およびテキストの索引付けシステムにも大きな負荷が発生します。このワークロードで最も重要な数量は、定義された時間枠で Content Collector によって処理する必要のある E メールの数であり、次いでアクセスする必要のあるメールボックスの数です。

- メールボックスのスキャン・ワークロード

  このワークロードは、Content Collector が特定のメールボックス内のアーカイブ対象を確認する必要があるときに作成されます。アーカイブ・スケジュールを適切に設計すると、こうしたスキャンが最小限に抑えられます。メールボックスに多数のアイテムが含まれている場合、アーカイブされるメールのスキャンが頻繁に行われる (例えば、Content Collector が「常時」スケジュールに設定されている) 場合、および Content Collector に複雑なフィルター制約が設定されている場合は、ワークロードのこの部分で E メール・サーバーのディスク・サブシステムに大きな負荷が生じる場合があります。

- スタブ化ライフサイクル・ワークロード

  このワークロードには、以前にアーカイブされた E メールを E メール・システムでスキャンすること、および既にアーカイブ済みの E メールを定義されたスタブ化ライフサイクルに従って変更することが含まれます。例えば、E メールがアーカイブされてから 30 日後に添付ファイルが E メールから削除され、エンタープライズ・アーカイブ内の添付ファイルのアーカイブ・バージョンへの参照だけが残されます。このワークロードでは、残った E メール・スタブを再度 E メール・サーバーから取り出し、Content Collector によって変更し、E メール・サーバーに保存し直す必要が生じる場合があります。

  このワークロードは、E メール・サーバーと Content Collector サーバーに影響を与えます。エンタープライズ・アーカイブは、これらの操作に関係しません。規準となる重要な数値は、処理する必要のあるメールボックスの数、およびあるライフサイクル状況から次のライフサイクル状況に E メールが移動する頻度です。

- 対話式ワークロード (検索、表示、および復元)

  メールボックス管理シナリオでは、ユーザーは Content Collector サーバーに対してアーカイブ済みのコンテンツの検索、プレビュー、および復元の各要求を出します。このワークロードでは、Content Collector サーバーはエンタープライズ・コンテンツ管理バックエンドからコンテンツを取得し、解析および処理 (例えば、プレビュー操作) して、コンテンツをユーザーのコンピューターに転送する必要があります。待ち時間なくこれらの要求を満たすために十分なリソースが使用できるようにするには、作業が集中するピークの時間帯は他の重要でないアーカイブ・タスク (ユーザーのメールボックスからの古いメールの自動アーカイブなど) を減らす必要があります。

ワークロードの計画とアーカイブ・スケジュール

E メール・アーカイブ・シナリオでのアーカイブ・スケジュールを理解して計画するのに役立つ、2 つの重要な要素を以下に示します。

- アーカイブする E メールの選別と実際のアーカイブ操作
- Content Collector のスケジュールを「常時」に設定しない

アーカイブする E メールの選別と実際のアーカイブ操作

Email Server Connector は、アーカイブされる E メールを選択し、選択したメールの実際のアーカイブを個別の操作で実行します。システムは、コレクターの実行がスケジュールされている場合に、以下の 3 つのステップを使用して、E メール・サーバーの負荷を減らします。

1. Email Server Connector は、構成済みのグループ・メンバーを解決し、アーカイブする必要のあるメールボックスを特定します。
2. その後、メールボックスの処理が始まり、一度に 1 つのメールボックスでアーカイブ候補の照会が出されます。この照会の結果、処理されるアイテムのリストが生成されます (to-do リスト)。
3. この E メール (アーカイブされる) の to-do リストは、非同期的に並列で処理されます。

不要な Content Collector の実行は回避されます。以前の Content Collector の実行からアーカイブまたはスタブ化が保留中の E メールが依然として内部的な作業キューに入っている場合、アーカイブが保留中のメールのバックログがあるすべてのメールボックスでは、アーカイブされるメールのスキャンは行われません。

一般に、Content Collector の実行間隔は、次の実行がスケジュールされる前に既に識別されているすべての E メールをシステムがアーカイブするのに十分な時間を空けてスケジュールする必要があります。適切なスケジュール時間を導き出すには、E メールの初期バックログをアーカイブするために手動の Content Collector スケジュール (例えば、一度だけ実行) でシステムを実行する必要があります。システムが安定状態になった (バックログの処理が完了した) 後に、一般的なアーカイブの実行が完了するのに要する時間を計測し、この情報に基づいて適切なスケジュールを導き出すことができます。

時間の経過とともにすべてのメールボックスが同じ確率で処理されるようにするために、Email Server Connector は順番を変更して、ステップ 1 で識別されたメールボックスが各 Content Collector の実行で処理されるようにします。

Content Collector のスケジュールを「常時」に設定しない

Content Collector のスケジュールを「常時」に設定した状態で、実動タスク経路を実行しないようにします。このように設定すると、ソースの E メール・サーバーの構成されたすべてのメールボックスで、アーカイブされる E メールのスキャンが頻繁に行われます。この設定を考慮できる唯一の例外は、対話式アーカイブ・タスク経路の場合です。このユース・ケースでは、ユーザーのどのメールボックスからアーカイブ要求が出されるかを判別するために、Email Server Connector によってトリガー・メールボックスが頻繁にポーリングされます。トリガー・メールボックスはサイズが小さいため、このメールボックスで頻繁に行われるチェックによって、大きな影響が生じることはありません。対話式のアーカイブ要求が処理の前に数分待機することが可能な場合、トリ

ガー・メールボックスにアクセスする Content Collector での間隔を 15 分に設定することを考慮する必要があります。

コンプライアンス (ジャーナル) アーカイブ・シナリオの場合、ベスト・プラクティスはジャーナル・メールボックスのサイズを小さく保つことです。E メール・サーバーの資料を参照して、パフォーマンス低下問題を発生させずに、ジャーナルに保管できる E メールの数を確認してください。Content Collector 間隔 (ジャーナルにデータを入れるのに要する時間) を、推奨されている最大サイズの 25% 以下に設定することをお勧めします。こうすると、ジャーナル・メールボックスでパフォーマンス問題が発生するかなり前にアーカイブが開始されるとともに、ほぼ空の状態のジャーナル・メールボックスで不要なスキャンが行われるのを避けることもできます。これは、現在アーカイブ対象のすべてのメールが 1 回の Content Collector の実行で選択されることを Content Collector が保証しないのを理解するのに役立ちます。収集の間隔が長すぎると、アーカイブのバックログが発生し、ジャーナル・サイズが増加する可能性があります。

メールボックスの同じフォルダー内のアイテム数が低い状態に保たれる場合、E メール・システムには利点があります。通常、最初は頻繁にアーカイブするのが最適です。

高ボリュームのジャーナル・セットアップでは、通常、複数のジャーナル・メールボックスと E メール・データベースを使用して、ジャーナルを保持します。これらのセットアップには、1 つの大きなジャーナル・メールボックスを使用する場合と比べて以下のようないくつかの利点があります。

- ジャーナルから複数のディスク・サブシステムへ、ジャーナルおよびアーカイブの負荷を分散できます。
- 1 つのメールボックスが大きくなりすぎて、管理やアクセスが難しくなるのを防ぎます。
- ジャーナルおよびアーカイブの並列性を高めます。
- 個々のユーザー・グループに対して異なるサービス・レベルを設定できます。

Content Collector が、大量のメールを保持するメールボックスからアーカイブされるメールを E メール・システムで照会する場合、長い時間がかかる場合があることに注意してください。照会するメールボックス内のアイテム数、メールボックス内部の編成状態、キャッシュに入れられている可能性のある (最近使用された) 量、およびこのメールボックスを保持するディスク・サブシステムの速度が、照会時間に大きく影響します。

アーカイブの遅延は、アーカイブ対象の照会に時間がかかっている兆候です。Email Server Connector で通知ログを一時的に有効にすることによって、照会が長時間実行されることによるアーカイブの遅延が生じているかどうかを検出できます。このログで、アーカイブ対象の照会が送信されてから、最初のメール

がアーカイブされるまでの遅延を確認できます (「Content Collector のログ」(21 ページ) の「Email Server Connector の動作を理解する」(23 ページ) を参照してください)。遅延が数分を超える場合、解決する必要のあるボトルネックが存在している可能性があります。

照会されたメールボックス・データベースを含むディスクの E メール・サーバー・ディスクの使用率は、E メール・システムが照会を実行する間にピークとなります。ディスク・サブシステムの使用率が長時間高い状態にあり、このストレージが他のメールボックスと共有されている場合、同じディスク・ストレージでの他のすべての操作 (通常のユーザー E メール操作など) は、照会の間、非常に遅くなる可能性があります。

照会時間が改善されない場合、大きなメールボックスを専用のストレージ・ユニットに移動して他の操作への影響を防ぐことや、スケジュールの頻度 (アーカイブする E メールの照会) を最小限に抑えることを検討してください。

IBM Lotus Domino ベースのシステムの場合、ジャーナルのロールオーバー機能を使用して、現行のジャーナル以外のジャーナルをアーカイブすることをお勧めします。こうすると、Lotus Domino サーバーおよび Content Collector サーバーから現在アクティブなジャーナル・データベースに並行アクセスできなくなり、パフォーマンスが向上します。一般に、Lotus Domino メールボックスまたはジャーナルに格納できる E メールの数は、パフォーマンスに及ぼす影響が大きくなる前は、非常に大きな値にすることができます。しかし、メールボックスが非常に大きい (例えば、容量の最大値に近い) と、アーカイブ処理が遅くなる傾向があります。ロールオーバー・メカニズムを使用すると、ジャーナル・メールボックスのサイズを小さなファイル・サイズ (数ギガバイトを推奨) に保ちつつ、大規模なバッファーを提供して、アーカイブ間隔を低く抑えるのに役立ちます (例えば、1 時間ごとに E メールを収集)。

メールボックス管理シナリオおよび他のアーカイブ・シナリオでは、ユーザーのメールボックスまたは他のソース・システム・コンテナー (例えば、ファイル・システム・フォルダー) で、アーカイブする必要のあるアイテムをチェックする頻度を定義する必要があります。アーカイブ対象のスキャンによって、ソース・システム (E メール・サーバーまたはファイル・サーバー) で負荷が生じるため、スキャンの頻度は予想されるアーカイブ対象の数とバランスを取る必要があります。例えば、ユーザーが 1 日に平均 20 通の E メールを受け取る環境では、メールの収集の間隔を 1 週間と定義すれば、1 回に収集されるメールは 100 通未満程度であるため、要件を満たせます。Content Collector をさらに頻繁に実行するよう構成されている場合、アーカイブ対象のスキャン完了にかかる時間が処理時間の大半を占めることになり、E メール・サーバーに不要な負荷がかかり、全体のスループットに悪影響が及びます。複雑なフィルター式も、アーカイブ対象のスキャンの実行にかかる時間に影響を与えます。

多数のユーザーが関係するデプロイメントの場合、メールボックスをグループ化して、アーカイブ・スケジュールとメンテナンス・スケジュール、および必要なアーカイブ頻度と連携する方法を考慮してください。例えば、10,000 の

ユーザーが関係し、各ユーザーに週に 1 回アクセスする必要があるデプロイメントの場合、約 2,000 ユーザーごとのグループを 5 つ作成し、各グループに週に 1 回アクセスする方法を考えるかもしれません。さらに、それよりも高い頻度または低い頻度でアーカイブする必要のあるアカウントのために特別なグループを作成することもできます。システムのリソースに、メンテナンスまたは再編成の間隔などが原因の制約がある場合、E メール・サーバーのクラスターに基づいてさらにグループ分けすることによって、あるクラスターのアーカイブをスキップする間に、メンテナンス、バックアップ、および再編成などを行えるようにして、別のクラスターのアーカイブのスループットを高めることができる場合があります。

E メールのスタブ化のワークロードは、通常週末にスケジュールできます (週に 1 回だけ実行)。ほとんどの場合、アーカイブされたメールの状態移行のスケジュールが数日遅れるとしても、このワークロードの次回の実行までに 1 週間の間隔が設定されているため、大きな問題とはなりません。

対話式アーカイブは、通常、作業が集中するピーク時間帯に行われます。これらの時間帯の Content Collector システムの負荷を低く抑えることによって、対話式アーカイブの要求が迅速に確実に実行されるようにし、作業が集中するピーク時間帯の E メール・サーバーとエンタープライズ・コンテンツ管理バックエンドの両方のアーカイブの負荷が減少するようにするのが賢明です。このように構成すると、対話式リトリーブ要求 (検索、表示、およびリストア) の応答を早くすることもできます。

E メールまたはファイルのバックログが、新規デプロイメントの初期フェーズでアーカイブされる必要がある場合、このアーカイブ・フェーズに使用できる時間を定義する必要があります。使用できる時間が多ければ多いほど、追加リソースの必要量は少なくて済み、対話式の毎日のアーカイブ・ワークロードが低いときにはバックログ・アーカイブによって生じる負荷を優先できます。例えば、この構成では最初、週末に実行されるようスケジュールできます。最初の年にバックログ・アーカイブで必要とされた追加の処理能力は、翌年以降の実動ではしばしばアーカイブ・ボリュームの増加分によって消費されます。

システムに高い負荷をかけるワークロードが重なるのを避けるためにアーカイブ・スケジュールを計画する場合、E メール、ファイル、またはグループウェア・サーバーの負荷とネットワークの負荷も考慮する必要があります (2 つのワークロードを並行で実行する場合にパフォーマンスが大幅に低下する例として、E メール・サーバー・データベースの再編成と E メール・アーカイブがあります)。

Content Collector のアーカイブでは、メールボックスと E メールは、一般のユーザーがこのデータに対話式にアクセスしたり変更したりするよりも非常に高い頻度でアーカイブが行われるため、E メール・サーバーには多くの作業が発生します。一般的にボトルネックとなるのは、E メール・サーバー・データベースをホスティングするディスク・サブシステム、および要求の合計数です。このシナリオには、作業時間にアーカイブの実行がスケジュールされている場合、

通常のユーザー要求と Content Collector サーバーからのアーカイブ要求が含まれます。WAN 接続が E メール・サーバーと Content Collector サーバーの間に位置している場合、ネットワークの帯域幅およびネットワークの待ち時間は制限的な要素となります。E メール・サーバーのアクセスに使用される RPC スタイルのプロトコルは、待ち時間の少ない直接接続で最適なパフォーマンスが得られます。Content Collector では、アーカイブの性質上、ローカル・メールのキャッシュなどの一部のパフォーマンス向上手段が役に立ちません。分散環境を使用しており、直接ギガビット・イーサネット接続を使用する場合、E メール・サーバーに隣接して Content Collector サーバーを配置して、待ち時間を低く保ち、スループットの制限が発生しないようにします。

システム計画に関する一般的な推奨事項

このセクションでは、アーカイブをセットアップする場合のベスト・プラクティスと思われる推奨事項をいくつか紹介します。デプロイメント後の変更は難しい、または不可能な場合さえあるため、これらの推奨事項の使用はデプロイメントを実行する前に検討してください。

ホスト名の別名

アーカイブ・ソリューションでホスト名の別名を使用して、その別名で Content Collector Web サービスがエンタープライズの *すべて* のユーザーに使用できるようにすることをお勧めします。このホスト名の別名は、すべてのスタブ・リンク (アーカイブされたコンテンツの代わりに置かれる URL) で使用される名前です。別のホスト名を指すようにスタブを変更することは多くの手間がかかり、不可能な場合もあるため、ホスト名は慎重に選択する必要があります。別名を使用すると、アーカイブ・サービスおよびリストア・サービスが実際に実行されるサーバーを割り当てたり、後から変更したりする場合の柔軟性が高まります。加えて、これは発行済みの HTTPS 証明書で使用されるホスト名です。

バックエンド・エンタープライズ・コンテンツ管理アーカイブへの直接リンクが使用される場合 (例えば、古いファイルのアーカイブ・シナリオの場合)、ホスト名の別名に関する同じ推奨事項がバックエンド・アーカイブにも適用されます。

最後に、Content Collector Web サービスによって複数のサービス・クラスが提供される場合 (アーカイブされたファイルのリトリーブとアーカイブされた E メールのリトリーブなど)、複数のホスト名の別名を使用することを考慮してください。この構成では、要求のクラスに基づいて特定の Content Collector インスタンスに経路指定するオプションが提供されます。これによって、異なるサービス・レベルをインプリメントするのに役立ちます。

ストレージの計画

Content Collector のアーカイブのインフラストラクチャーには、異なるサーバーに割り振られるいくつかのストレージ・ユニットが必要です。パフォーマンス・クリティカルなストレージを、標準的なロケーション (オペレーティング・システム・ディスク) に置かないようにします。パフォーマンス・クリティカルなストレージは、このストレージ域に想定されるワークロードとストレージ・デバイスのパフォーマンス特性に基づいて、特定のストレージ・デバイスに配置します。誤解が生じないために、マウント・ポイントおよびドライブ・ラベルには明示的で分かりやすい名前を指定するのが最善です。特定のドライブまたはマウント・ポイントの用途をはっきり示してください。すべての用途に同じタイプのストレージを使用する (例えば、コストの高いファイバー・チャネル (FC) ドライブによる RAID 10 構成) のは、コスト面で効果的ではありません。

以下のタイプのストレージを計画するための大まかなガイドラインが用意されています。

- バイナリー文書およびテキスト文書のアーカイブ・ストレージ
- 一時ストレージ
- テキスト索引ストレージ
- データベース・ストレージ
- ログ・ストレージ

バイナリー文書およびテキスト文書のアーカイブ・ストレージ

このストレージは、場合によっては固定ストレージ (WORM) デバイスになりますが、通常のディスク・ストレージになる場合もあります。最初、システムはアーカイブの際に低頻度の読み取り (リトリーブ) を使用して、このストレージに書き込みのみを行います。その後の段階で、ワークロードには、古いコンテンツの下層ストレージへの移行やコンテンツの削除 (有効期限の管理) が追加されます。ストレージの重複は、Content Collector システムが提供するメカニズムと、エンタープライズ・コンテンツ管理バックエンドのサポートを組み合わせたものによって解消されます。固定ストレージ・デバイスが使用される場合、これらのデバイスは通常、データの重複解除機能を備えています。

安価で大容量のドライブ (例えば、SATA ドライブ) を使用した RAID 5 アレイの採用をお勧めします。ご使用の環境およびセットアップ方法に応じて、SAN または NAS ソリューションを使用できます。事前構成されたコンテンツ・エンジン・セットアップ (FileNet® P8) では、分散ファイル・システムを使用する必要があります。ここでは、NAS ソリューションが推奨される一般的なメカニズムです。

一時ストレージ

さまざまなシステム・コンポーネントが、中規模サイズの一時ストレージ・ロケーションを必要とします (例えば、Email Server Connector または Text Extract Connector のための Content Collector 作業ディレクトリー、およびテキスト・

バージョンの E メールに索引付けする Content Collector インデクサーの作業ディレクトリー)。このストレージ・ロケーションに格納される情報は、失われても再生成が可能であるため、このストレージには、データ損失に対する保護の必要のない揮発性の高い文書が入れられます。この種類のストレージには、通常高速のローカル・ディスク (例えば、SAS ドライブによる RAID 0 アレイ) が使用されます。これは、Content Collector サーバーで 2 つの内部ディスクを使用する大きな理由となっています。そのため、一時ストレージ・ロケーションは、2 つのドライブ間でストライピングされます。

テキスト索引ストレージ

全文検索索引には、特別な使用パターンがあります。データの書き込みは通常、索引作成または再編成の間に大規模な順次書き込みによって行われますが、読み取りは小さなランダム・アクセスによって高頻度のトランザクションで行われる可能性があります。検索のパフォーマンスを高めるには、検索の負荷に見合った必要な IOPS (I/O operations per second) を提供できるように 1 つ以上の大規模なディスク・アレイが必要です。エンタープライズ・レベルの 15 K rpm の SAS または FC ドライブの RAID 10 アレイを使用することをお勧めします。ドライブの数は、必要な容量ではなく、想定される検索の負荷によって決定されます。テキスト索引の日付区分化機能の使用を強くお勧めします (「パフォーマンスとスケーラビリティー」(31 ページ) を参照)。この機能によって、必要なドライブ数を大きく減らすことができます。最新の Solid® State Disk (SSD) テクノロジーは、このユース・ケースに最適と思えますが、現在のところ SSD アレイの使用実績はわずかです。

データベース・ストレージ

システムの高いパフォーマンスの鍵となるのは、高速なデータベース・ストレージです。エンタープライズ・コンテンツ管理アーカイブ・データベースおよび Content Collector 構成データベースには、エンタープライズ・レベルの 15 K rpm の SAS または FC ドライブによる RAID 10 アレイを使用することをお勧めします。

ログ・ストレージ

ログは順次書き込まれ、通常は読み取られません。ログが専用ドライブに書き込まれる場合は、ディスクの負荷は比較的低く抑えられます。順次書き込みのみが行われる場合は、ドライブ・メカニズムは再配置なしで機能します。他の用途にも使用されるディスクにログが書き込まれる場合、定期的に再配置が行われて、全体として非常に高いディスクの負荷が生じ、他のディスク・ワークロードが遅くなります。システム・ログおよびトランザクション・ログには、専用のドライブを使用することをお勧めします。

FileNet P8 リポジトリーのアプリケーション・サーバー・ログには、細心の注意を払う必要があります。Content Collector の重複検出メカニズムによって、データベースの固有性制約違反のメッセージのために多数の重複が検出される場合、大きなロギング・アクティビティーが発生する可能性があります。

レイヤーでの重複コンテンツの検出とストレージに与える影響

重複コンテンツを検出すると、エンタープライズ・コンテンツ管理アーカイブで必要とされるストレージ・スペースは大幅に削減されます。重複コンテンツの例として、複数のメールボックスに格納されていてアーカイブされる同一の E メール、異なる E メールで使用される同一の添付ファイル、または複数の場所に格納されている同一のファイルなどがあります。Content Collector は、ファイル、E メール、または添付ファイルが既にアーカイブされているもののコピーであるかどうかを検出でき、そのアーカイブに既に保存されているアイテムについては、新しいインスタンスに対する新しい参照だけを作成します。こうすることにより、膨大なストレージが削減されるだけでなく、コンテンツを一から再処理する必要がなくなるため、アーカイブ・サーバーおよび E メール・サーバーでの負荷も削減されます。

メールボックスの管理とコンプライアンス・ジャーナル・アーカイブの両方を使用するシナリオでは、Content Collector はメールボックスとジャーナルにそれぞれアーカイブされている E メールの重複も検出します。この機能により、コンプライアンス・アーカイブを実行する必要のある会社が、事実上追加のストレージ要件なしで、メールボックスの管理も実行できるようになります (メールボックスの管理の間に処理されるすべてのメールが、コンプライアンス・ジャーナル・アーカイブ・ワークロードの一環として既にアーカイブされている場合、またはその逆の場合)。

すべてのシナリオで、重複した E メールは、Content Collector によってハッシュ・アルゴリズムを使用して検出されます。他の重複コンテンツ (E メールの添付ファイルおよびファイルなど) の検出および管理方法は、使用されているバックエンド・エンタープライズ・コンテンツ管理アーカイブのバージョンに応じて、さらにストレージ・レイヤーに重複検出機能を備えたストレージ・サブシステムが使用されるかどうかに応じて異なります。

ストレージ・レイヤーで重複検出機能を提供するストレージ・サブシステムでは、Content Collector が既に重複検出を実行したコンテンツ (この例では E メール本文) について、重複をチェックする負荷を省略するよう構成をカスタマイズすることができます。一部のデバイスには、デバイスに格納されているファイルに重複検出を実行するかどうかに関して、ファイル・サイズのしきい値があります。しきい値よりも小さいファイルは、予想されるストレージ削減の効果が小さいため、チェックされません。この機能は、ストレージのサブシステムのパフォーマンスを向上させるために時折使用されます。

このメカニズムを使用して、大半の一般的な E メールの本文サイズよりも大きいサイズにしきい値を設定することによって、E メール本文の重複チェックを回避することができます。一般的な E メール本文の最大サイズの例は、IBM Lotus Notes® E メール・サーバーの場合は 32 KB であり、Microsoft Exchange E メール・サーバーが使用される場合は 64 KB です。

テキスト・インデクサーのワークロード

アーカイブ済みコンテンツのテキスト索引の作成は、ECM リポジトリーに付属の索引付けコンポーネントによって実行されます。データの具体的なフローは、使用されるエンタープライズ・コンテンツ管理アーカイブのバージョンによって大きく異なります。ここでは、使用されるアーカイブに応じて存在するデータ・フローを説明します。

IBM Content Manager がアーカイブ・リポジトリーとして使用される場合、データ (E メールとファイルの両方) はネイティブ・フォーマットでリポジトリーに取り入れられます。Content Collector インデクサー・コンポーネントは、Content Manager Library Server で非同期的に実行され、アーカイブされたコンテンツをバッチ処理できます。具体的には、リポジトリーから新たにアーカイブされたコンテンツのアイテム一式をリトリーブし、フィルターを使用してアーカイブされたアイテムに組み込まれているテキスト・データ (例えば、E メール本文と添付ファイルまたはアーカイブ・ファイル) をメタデータと共に中間的な XML 表現データに変換し、XML 文書一式を IBM Net Search Extender に送信して索引付けを行います。その後、Net Search Extender はこれらの一時的な XML 文書のテキスト・コンテンツを使用してテキスト索引を作成します。バイナリー・ファイルからテキスト・コンテンツを抽出するプロセスは、多数のプロセスが関係するため、Library Server の CPU リソースの大半を必要とします。デフォルトでは、Content Collector インデクサーによって多数の並列スレッドが発生します (スレッドの数はシステム上の CPU の数に合わせて自動的に調整されます)。インデクサーの実行中に CPU リソースが他のタスクに予約される場合、並列インデクサー・スレッドの数は低い数に構成されます。

Content Collector インデクサーをアーカイブ後から数時間、場合によっては数日遅らせるようにスケジュールすると、インデクサーのパフォーマンスが向上します。メールボックス管理 E メール・アーカイブ・シナリオでは、Content Collector がすべてのメールボックスにアクセスするのに要する時間を遅延として設定すると、文書が索引付けのために選択される前に、特定の文書のかなりの数の重複が取り込まれる可能性があります。特定のメールの重複するすべての E メール・インスタンスが、索引付けの必要な一連の E メールの一部としてアーカイブされている場合、XML 文書が 1 つだけ作成され、それがテキスト検索索引に追加される必要があります。こうすると、索引付けの必要なアイテムの数が大幅に減ります。

すべてのアーカイブ・シナリオで、1 回のインデクサーの実行で索引付けされる一連の E メールのサイズを大きくすると、索引の更新数が減り、索引は良く編成された状態が保たれ、再編成の実行が少なくて済みます。インデクサーに遅れをスケジュールする場合のマイナス面は、アーカイブされたコンテンツをすぐに検索できないことです。ただし、アーカイブ・ポリシーでソース・システムからすぐにコンテンツを削除しないようになっている場合、一般に検索機能はすぐには必要ではないため、テキスト検索用の索引付けを遅らせることは

可能です。1 回の索引付けの実行で、E メールの重複するインスタンスの大半を取り込める長さの遅延を設定することをお勧めします。

FileNet P8 Content Manager をエンタープライズ・アーカイブとして使用する場合、アーカイブされるコンテンツの種類に応じて、テキスト索引付けは異なります。標準的なファイル・アーカイブの場合、Content Collector がバイナリー・データをリポジトリーにアーカイブし、FileNet P8 Content Based Retrieval (CBR) 機能がこのコンテンツをテキスト変換と索引付けのために Content Search Engine に頻繁に送信します。

E メールの場合、プロセスは異なります。CPU を集中的に使用するテキスト変換は Text Extract タスクとして、一般の CPU リソースで動作する Content Collector サーバー上で実行されます。添付ファイルを含む、E メールのすべてのテキスト・コンテンツを保持する Content Collector サーバーで、XML 文書 (ICCEmailSearch 文書クラス) が作成されます。この XML 文書 (XML Instance Text (XIT) 文書と呼ばれる) はエンタープライズ・アーカイブに格納されます。こうして、Content Search Engine の CPU 集中型のテキスト抽出が軽減され、XML 文書のコンテンツをテキスト索引に追加するだけで良くなります。後からの再索引付けまたは更新のために XIT 文書もアーカイブに格納されます。E メールの重複インスタンスが検出される場合、Content Collector タスク経路は XIT 文書を更新して、重複するメールに関する情報を反映し、この情報を新しいバージョンの XIT 文書としてアーカイブに格納します。Content Search Engine は、新しいバージョンに索引を付け、索引から以前のバージョンを削除します。

モニタリングとヘルス・チェック

Content Collector のログは、システムのパフォーマンスと正常性を測るための有用な情報源です。定期的にログ・ファイルをモニターし、ヘルス・チェック (正常性検査) を実行してください。このセクションでは、Content Collector のさまざまなログを紹介し、パフォーマンス・カウンターを使用してシステムをモニターする方法を説明し、システムのパフォーマンスや正常性に関する情報を入手するために利用できる照会のサンプルを提供します。

Content Collector のログ

Content Collector を構成する際にログを 1 か所に格納することを、この資料の前半で推奨しました。Content Collector サーバーで 2 台の独立したディスク・ドライブを使用できる場合であれば、1 台のディスク・ドライブをオペレーティング・システムと Content Collector ログ・ファイルのために使用し、もう一方のディスク・ドライブを Content Collector の作業ディレクトリーと一時ディレクトリーのために使用することができます。最良の結果を得るには、2 台のエンタープライズ SAS ディスクで構成したローカル RAID 0 ディスク・アレイを作業ディレクトリー (Email Server Connector) と一時ディレクトリー (Text Extract

Connector) のために使用します。運用中、この 2 つのディスク・ドライブの使用率をモニターします。使用率が 60% を超えている場合は、そのディスク・ドライブが Content Collector のスループットを制限する要素になっている可能性があります。

Content Collector のロギング・レベルは、実動システムでは、通知、警告、またはエラー・ロギングのいずれかのレベルに設定してください。通知レベルのロギングは、システムの運用についてさらに洞察を得るために役立ちますが、システムに影響を与えるという問題もあります。システムを新しくセットアップしたときは、まず通知ロギングを使うとよいでしょう。システムが実動状態に入った後は、警告レベルまたはエラー・ログ・レベルが適切です。トレース・レベルのロギングは、通常のモニター作業には詳細すぎて不便ですが、特定の問題を解決するときに必要になります。

別の一般的な推奨事項として、ログの保存期間を数日に設定し、特に監査ログについては、ログをアーカイブするメカニズムを構成してください。このとき、現在アクティブなログ・ファイルをアーカイブしたり、アクセスをブロックしたりしないように注意してください。そうしないと、追加のログ項目を作成できなくなることがあります。

Content Collector Web サービスは、他のログ・ファイルに対して構成したのとは違う場所にログ・ファイルを出力することに注意してください (デフォルトの場所は `ContentCollector¥AFUWeb¥afu¥logs` ディレクトリー)。このディレクトリーを定期的に検査すれば、潜在的な問題を発見できます。

また、タスク経路について監査ロギング機能を有効にすることもできます (デフォルトでは有効になっている)。監査ログはカラム形式のログで、システムにおける成功したスループットと起こりうる障害を自動的にモニターするために役立ちます。この監査ログの詳しい調査 (例えば、アーカイブされた E メールの平均サイズの調査) は、スプレッドシート・プログラムで行うことができます。

もしまだであれば、タスク経路の監査ログ構成に以下のメタデータ項目を追加することをお勧めします。

- ファイル – ファイル拡張子
- ファイル – ファイル・サイズ
- E メール – 添付ファイル・フラグ
- E メール – 添付ファイル数

これらの値は、アーカイブしているコンテンツのサイズと添付ファイルの分布を理解するのに役立ち、ストレージ要件の予測と計画のために利用できます。また、システムのスループット数を理解する助けにもなります。

スタブ化タスク経路から意味のある監査ログ・データを取得するには、タスク経路に対して EC の「メタデータの抽出」タスク (Email Server Collector が提供

しているタスクの 1 つ) を追加する必要があります。このタスクは、スタブ化タスク経路のロギングに対してメタデータを利用可能にします。スタブ化タスク経路で特に意味のあるメタデータ項目の 1 つは、「E メール - 処理状態」です。この情報から、E メールがスタブ用に選択された時の状態 (例えば、ARCHIVED、STUB_NOTHING_ADD_TEXT、または STUB_ATTACHMENTS) を洞察できるため、E メールが 1 つのタスク状態から別のタスク状態にどのように移行するのかを分析するために役立ちます。

最後の点として、補助コンポーネントによって作成されるログも、モニターの対象として考慮する必要があります。例えば、Content Collector のインデクサー・コンポーネント (IBM Content Manager アーカイブ・リポジトリーにある) などです。また、FileNet P8 アーカイブ・リポジトリーでのテキスト索引の作成をモニターするために CBR トレース・オプションを利用できます。

Email Server Connector の動作を理解する

新しく構成したシステムでは、Email Server Connector が処理のためにメールボックスをいつ「クロール」するか、そしてどのメールボックスをいつ処理するかについて詳しく理解しておくことが大切です。監査ログを調べると、この情報の一部が分かります。さらに概要レベルでの洞察を得るには、通知ロギング・レベルを設定した Email Server Connector の `sysout` ログ・ファイルを利用します。

特に注意するべき情報は、次の 2 種類の操作とその関連メッセージです。

- 処理のためのメールボックスのスキャン

  システムがメールボックスを検出するたびに「`crawling store` (ストアのクロール)」というストリングを含むメッセージがリストされます (例えば、処理対象のユーザー・グループなど、構成されたコレクターのデータ・ソースに基づいて)。このメッセージの時点で、メールボックスは、処理が必要なメールボックスの内部リストに追加されます。

- メールボックスのオープンと処理

  処理のためにメールボックスがオープンされ、メールをアーカイブするようにとの照会が E メール・システムに発行されるたびに「`Processing store` (ストアの処理)」というストリングを含むメッセージがリストされます。

  これらのメッセージを検索して分単位で集計すれば、有用な統計を入手できます。アーカイブ対象のメールボックスがどのくらいの数、いつ検出されたか (「`crawling store`」メッセージ)、そして検出されたメールボックスがいつ、どのくらいの速度で処理されたか (「`Processing store`」メッセージ) が判明することになります。

Email Server Connector を通知ロギング・レベルに構成した場合に頻繁に目にする別のメッセージは、「`pruning stores` (ストアのプルーニング)」メッセージです。これは無視してかまいません。このメッセージが出される理由を理解するには、Email Server Connector が効率化のためにどのように動作するかを知っ

ておく必要があります。Email Server Connector は、メールの処理が完了した後も E メール・ボックスすなわち E メール・ストアをクローズしません。同じストアで追加のメールを処理する必要が発生した場合に備えて、ストアをオープンしたままにするからです。多くの場合、未使用のストアが検出され、そのままクローズされます。このアクティビティーが「pruning stores」メッセージによって報告されるという仕組みです。

パフォーマンス・カウンターによるモニタリング

Microsoft Windows® ベースのサーバー・システムにおけるアーカイブ・ソリューションとしては、オペレーティング・システムのコンポーネントである Windows パフォーマンス・モニター (perfmon、「信頼性とパフォーマンス・モニタ」とも呼ばれる) を利用して、すべてのサーバーのパフォーマンス・データを共通の場所で収集して分析できます。Windows パフォーマンス・モニターの 1 つのインスタンスを使用して、Microsoft Windows オペレーティング・システム上で稼働しているすべての Content Collector サーバー、エンタープライズ・コンテンツ管理バックエンド・サーバー、および E メール・サーバーについて、システム運用の統計を収集できます。このようなセットアップを機能させるには、すべてのサーバーが同一の Windows ドメインに所属している必要があります。また、perfmon を実行するために使用されるドメイン・ユーザー・アカウントから、すべてのサーバーの必要なカウンターにアクセスできることも必要です。さらに、同一ドメイン内のすべてのマシンのシステム・クロックに大きな時間差があってはなりません。多くの場合、30 秒間隔であれば、データの細分性とディスク容量消費のバランスが取れます。24 時間を超えてモニターする場合は、1 日に 1 回、perfmon のカウンター・ログを停止して再始動することにより、ログ・ファイルのサイズを小さくすることをお勧めします (perfmon には、このようなスケジュール設定に利用できる組み込みスケジューラーが備わっています)。

32 ビット・バージョンのパフォーマンス・モニターの実行 : Content Collector のパフォーマンス・カウンターは、32 ビット・カウンターとして提供されます。64 ビットの Microsoft Windows 上で実行している場合は、パフォーマンス・モニターの始動時に次のようなコマンドを使用することにより、これらのカウンターをモニターする際に 32 ビット・バージョンのパフォーマンス・モニターを確実に使用するようにしてください。

```
mmc /32 perfmon.msc
```

そのようにすれば、カウンターの現在の値を perfmon で表示できます。32 ビットの Content Collector のカウンターを記録するには、カウンター値を収集してカウンター・ログ・ファイルに格納する役割を持つシステム・サービスも 32 ビット・バージョンを使用する必要があります。

- 64 ビットの Windows Server 2003 の場合は、「Performance Logs and Alerts」サービスを修正して再始動する必要があります。レジストリー・エディターを使用して、次のレジストリー・キーの ImagePath プロパティーを開きます。

  HKLM¥System¥CurrentControlSet¥Services¥Sysmonlog

  標準の値 (64 ビット・バージョンをポイントしている) を、次のように変更します。

  %SystemRoot%¥syswow64¥smlogsvc.exe

- 64 ビットの Windows Server 2008 の場合は、レジストリーを変更する必要はありません。新しい「Performance Counter DLL Host」サービスが導入されて、64 ビット・プロセスに対して 32 ビット・カウンターが使用可能になったからです。「Performance Logs and Alerts」サービスと「Performance Counter DLL Host」サービスの両方が実行されていることを確認するだけで済みます。

この後のセクションでは、個々のサーバーでモニターすることをお勧めするパフォーマンス・カウンターを紹介します。最初のカテゴリーは、その後のステップバイステップの解説でカウンター・ログに追加するものです。

すべてのサーバーで収集する必要のある汎用カウンター

次に挙げる汎用カウンターは、すべてのサーバーで収集する必要があります。

- Processor - % Processor Time (「すべてのインスタンス」を選択)
- Physical Disk - % Idle Time (「すべてのインスタンス」を選択)
- Physical Disk - Disk Transfers/sec (「すべてのインスタンス」を選択)
- Network Interface - Bytes Total/sec (「すべてのインスタンス」を選択するかどうか判断に迷う場合は、使用されているインスタンスを選択)
- Memory - Available MB
- Process - Percent of processor time (「すべてのインスタンス」を選択)

Content Collector サーバーの場合の追加カウンター

Content Collector サーバーでは、前述のカウンターに加えて、すべての CTMS カウンターをすべてのインスタンスについて収集してください。一部のインスタンスは、Content Collector の実行中でないと追加できないため、注意が必要です。また、タスク経路に固有のインスタンスのカウンターは、特定のタスク経路が最初にロードされた時点で生成されます。パフォーマンス・モニターのカウンター・ログを更新するには、次のようにします。

1. 少なくとも 1 つのタスク経路が存在する状態で、タスク経路サービスを始動します。

2. CTMS カウンターをカウンター・ログに追加し、すべてのカウンターのすべてのインスタンスが追加されていることを確認します。

こうしておけば、後ほどタスク経路のインスタンスが作成された時点で自動的にそのインスタンスがカウンター・ログに組み込まれるため、インスタンスが欠落して情報が失われるということがありません。

Microsoft Exchange E メール・サーバーのカウンター

Microsoft Exchange E メール・サーバーでは、CPU とディスク使用率に関する前述の汎用カウンターに加えて、Microsoft Exchange オブジェクトからのすべてのインスタンスについて次のカウンターを収集してください。

- MSExchangeIS オブジェクト :
  - Active connection count
  - Active user count
  - RPC operations/sec
  - RPC packets/sec
  - RPC requests
  - RPC averaged latency
  - Client:Latency >2、>5、および >10 (3 つのカウンター)
- MSExchageIS Mailbox オブジェクト :
  - Folder opens/sec
  - Message opens/sec

使用可能なカウンターや、どのカウンターが重要であるかについては、使用している Microsoft Exchange のバージョンによって異なります。カウンターの値を解釈するために、Microsoft の技術情報から Microsoft Exchange のパフォーマンスに関するトピックを参照することをお勧めします。

すべてのカウンターの構成を終えた後 (繰り返しになりますが、すべてのカウンターが 1 つの perfmon カウンター・ログに記録されていると便利です)、カウンター・ログを開始します。perfmon が作成するログの名前は `c:/perfmon/Content Collector_00001.blg` という形式になります。記録を保持する期間の 1 日ごとに、ログを生成するために約 250 MB が使用されることに注意してください。すべてのサーバーについてデータが記録されているかどうか確認するために、5 分ほど実行した後、ログ・ファイルを開き、CPU および物理ディスクのアイドル時間 % をすべてのマシンについてグラフに表示してみます。このグラフを見ると、システムの稼働状況を概観できます。

必要なグラフとカウンターを構成した後、クリップボードへのコピーのアイコンをクリックすれば、perfmon 構成のコピーを取得できます (表示される部分についてだけの情報)。この情報をメモ帳に貼り付けてファイルに保存しておきます。perfmon を再始動した後は、グラフおよびシステム・モニターの構成が消えているので、再度インポートする必要があります。保存したテキスト・

ファイルを開き、構成を perfmon にインポートします。インポートを実行するには「**プロパティー**」ボタンをクリックするか、Ctrl + Q キーを押します。

図 2 は、Lotus Domino E メールから IBM Content Manager アーカイブ・リポジトリーへのジャーナル・アーカイブ・アクティビティーのモニター結果です。

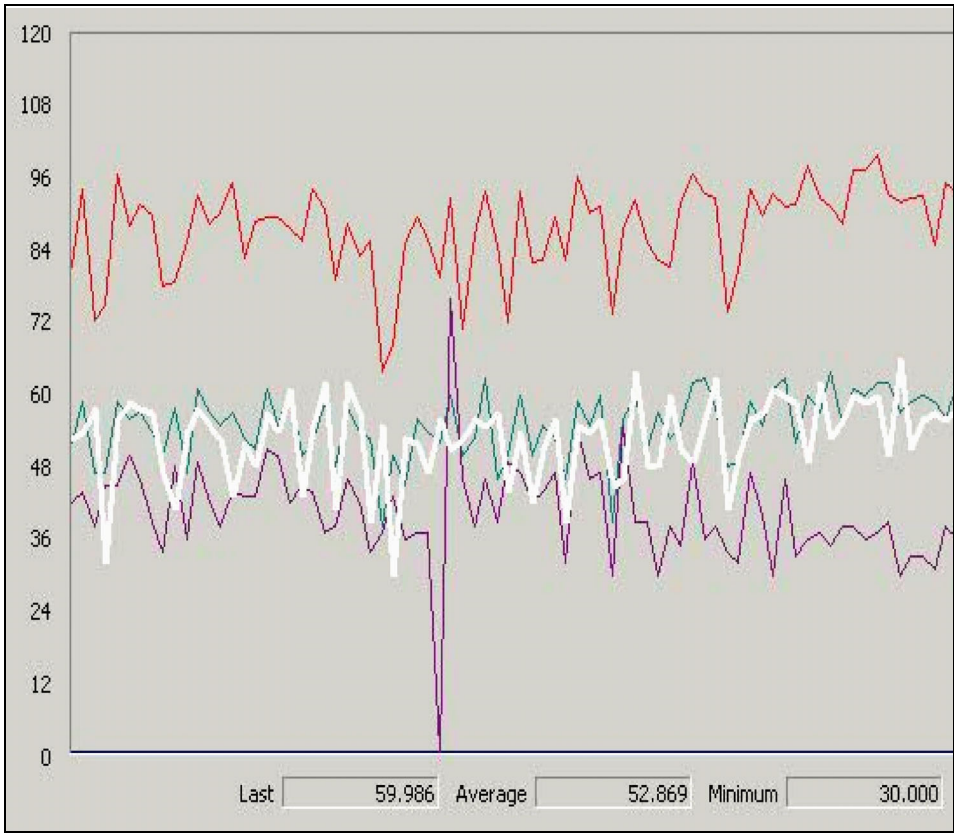

図 2 Lotus Domino から IBM Content Manager へのアーカイブのモニター

図 2 に含まれているのは、プロセッサー時間 %、1 秒あたりのエンティティー・アクセス数、アイテム・バックログ、そして 1 秒あたりの文書作成数です。このグラフを見ると、このマシンの 2-way AMD Opteron デュアル・コア・プロセッサーの合計利用率が約 90% であることが分かります。1 秒あたりのエンティティー・アクセス数と 1 秒あたりの文書作成数は、どちらも似通っていて、平均で 1 秒あたり 53 件の E メールです。タスク経路サービスの内部処理キューを表すアイテム・バックログは、常時 35 から 40 件の項目の処理が必要な状態となっています。

アーカイブ・エンジンと全文索引作成エンジンの照会

エンタープライズ・コンテンツ管理アーカイブ・リポジトリーには、基礎となるデータベースにアーカイブされたコンテンツのメタデータが格納されます。このセクションでは、アーカイブされた E メールの数と重複する E メールの数、および添付ファイルの数と重複する添付ファイルの数を調べるために使用する、これらのデータベースのサンプル照会を説明します。E メール以外のコンテンツがアーカイブされている場合の照会は、添付ファイル数を調べるために使用する E メール・アーカイブの照会に似ています。ここで説明するサンプル照会は IBM DB2® データベース・システムに対するもので、他のデータベース・システムで利用するには手直しが必要です。

現在アーカイブに入っている項目数を調べるには、次に挙げる照会が役立ちます。

IBM Content Manager ベースのアーカイブ・リポジトリーに対する照会

Content Collector 2.1.1 のデータ・モデルでは、次に挙げる照会を実行できます。

- 個々の E メール・インスタンス数 (重複を含まない) を調べるには、次の照会を実行します。

  ```
  select count(*) from ICCEmailCmpLDO01 with ur;
  ```

- E メール・インスタンスの合計数 (重複を含む) を調べるには、次の照会を実行します。

  ```
  select count(*) from AFUEChildO01 with ur;
  ```

- 個々の添付ファイル数 (重複を含まない) を調べるには、次の照会を実行します。

  ```
  select count(*) from ICCAttachmentsO01 with ur;
  ```

- 添付ファイル・インスタンスの合計数 (重複を含む) を調べるには、次の照会を実行します。

  ```
  select count(*) from AFUAChildO01 with ur;
  ```

Content Collector 2.1.0 のデータ・モデルでは、次に挙げる照会を実行できます。

- 個々の E メール・インスタンス数 (重複を含まない) を調べるには、次の照会を実行します。

  ```
  select count(*) from ICCEmailLDO01 with ur;
  ```

- E メール・インスタンスの合計数 (重複を含む) を調べるには、次の照会を実行します。

  ```
  select count(*) from AFUEChildLDO01 with ur;
  ```

基本表名より使いやすいビュー (例えば、ICCEmailLD001) を検出するには、次の照会を使用します。

```
select COMPONENTVIEWID, COMPONENTTYPEID, ITEMTYPEID, COMPONENTVIEWNAME,
CONCAT (CONCAT('ICMUTO',CAST(COMPONENTTYPEID as char(4))),'001') TABLE
from icmstcompv iewdefs where itemtypeid in (select keywordcode from
icmstnlskeywords where keywordclass=2 and keywordname in
('ICCEmailCmpLD', 'ICCAttachments')) with ur
```

特定の日にどれほどの量の項目が取り込まれ、変更されたかについて調べるには、次の照会を実行して、項目表をシステム表の名前に合わせる必要があります。

```
select date(changed) date, count(*) count from ICMSTITEMS001001 group
by date(changed) with ur;
```

オープン・タスクのデータベース表の名前は、AFU1FTIOPEN<*item type ID*> という組み合わせで生成されます。ここで、<*item type ID*> は、項目 ID の数値です。必要なら先行ゼロが付加されて 5 桁の ID になります。

Content Collector インデクサーが処理する必要のある項目のバックログをモニターするには、次の例のような照会を実行します (末尾の数字を、モニター対象の項目タイプの内部番号に合わせます)。

```
db2 "select count (*) from ICMADMIN.AFU1FTIOPEN01007"
```

Content Collector インデクサー・コンポーネントには、コンソール出力として表示できるパフォーマンス統計の機能が組み込まれています。詳細については、製品資料を参照してください。

パフォーマンス統計のコンソール出力を取り込んで保存するには、インデクサーをスクリプトから実行して、コンソール出力をファイルにリダイレクトします。AIX® では、インデクサー・コンポーネントを手動で実行するときに **nohup** コマンドを使用します。このコマンドを使用すると、操作がコンソールから切り離して実行されるため、誤操作や接続の切断などでコンソールが閉じても、プロセスの中断が回避されます。

また、インデクサーの構成を拡張することにより、DB2 Net Search Extender を使用してテキスト索引を実際に更新する前や後に、追加のスクリプトを実行できます。この機能を利用して索引の更新前と更新後の索引ディスク利用状況をモニターすれば、索引に文書を追加するバッチごとの索引サイズの標準的な増加量について情報が得られます。

テキスト索引が現在保持している項目数を調べるには、db2ext.textindexes 表の number_docs 列を照会します。

```
db2 "select number_docs from db2ext.textindexes"
```

複数のテキスト索引がある場合は、追加の列を表示して、どのテキスト索引がどの項目タイプに属しているかを識別する必要があります。

索引の更新操作中に DB2 Net Search Extender が現在処理しているアイテム数を調べるには (Content Collector インデクサー・コンポーネントの提供するモニター機能がアクティブでない場合)、次のコマンドを実行して索引更新プロセスから現在の統計を取得します。

```
db2text control list all locks for database icmnlsdb
```

FileNet P8 アーカイブ・リポジトリーに対する照会

シンボル名からオブジェクト ID を取得するには、次の照会を使用します。

```
db2 => SELECT object_id,symbolic_name FROM ClassDefinition
x'9B06EAE70C99F944B9CA279435EA42E2' ICCMailInstance
x'974F613023895F45A804E28A65E97FBA' ICCMailJournal
x'F890E9972A0123469EA9C3ED141DBE72' ICCMail
x'BE22769A46A02041AE1B197790B9B5F4' ICCMailSearch
```

ここで、

- *ICCMailInstance* (E メール・インスタンス (EI) ともいう) は、カスタム・オブジェクトです。アーカイブされる E メール 1 通ごとに、このインスタンス・オブジェクトが 1 つ作成されます。これにはメタデータだけが入ります。
- *ICCMail* (個別 E メール・インスタンス (DEI) ともいう) は、Document クラスから派生される標準オブジェクトです。主として、E メールの本文と添付ファイルのコンテンツ要素を保持します。
- *ICCMailSearch* (XML インスタンス・テキスト (XIT) ともいう) は、Document クラスから派生される標準オブジェクトです。バージョン管理機能とコンテンツ・ベース・リトリーブ (Content Based Retrieval、CBR) 機能を持ち、検索結果を表示するためのメタデータを保持します。E メールのテキスト・コンテンツと添付ファイルを XML 形式で表現した、索引作成のためのコンテンツ要素です。

object_id の値は、後続の照会で利用できます。

項目数 ICCMail (個別の E メール・インスタンス) と ICCMailSearch (CBR のためのテキスト・オンリー・バージョン) を DocVersion 表を使用して判別するには、次のコマンドを実行します。

```
db2 => SELECT count(*) FROM OSFMD.DocVersion WHERE
object_class_id=x'BE22769A46A02041AE1B197790B9B5F4' with ur;
```

アイテム数 ICCMail (個別の E メール・インスタンス) と ICCMailSearch を、日付範囲を指定して判別するには、次のコマンドを実行します。

```
db2 => SELECT count(*) FROM OSFMD.DocVersion WHERE create_date between
'2009-10-21-15.10.00.000000' and '2009-10-21-16.12.00.000000' and
object_class_id=x'F890E9972A0123469EA9C3ED141DBE72' with ur;
```

ICCMail 内の個別の E メール・インスタンスで作成日の異なる重複したメールがある場合、ICCMailSearch 内のテキスト・オンリー・バージョンが重複したメールによって更新された最終日付によっては、特定の日付範囲を指定したときに報告される項目数が異なることがあります。

ICCMailInstance オブジェクト (重複した項目を含む、アーカイブされた E メールの合計数) も同じような方法で照会できますが、このオブジェクトはカスタム・オブジェクトであり、DocVersion 表ではなく、generics 表に含まれています。このオブジェクトを取得するには、次の照会を実行します。

```
select count(*) from generic where
object_class_id='E7EA069B-990C-44F9-B9CA-279435EA42E2' with ur;
```

パフォーマンスとスケーラビリティー

新しく構築したシステムで最大限のパフォーマンスを得るには、実稼働状態に入る前に、初期チューニングが必要です。パフォーマンスのチューニングは、複雑で繰り返しの多い作業になります。定量的な目標の設定と、一定条件下でのシステムのモニタリングが必要で、時間の経過とともに目標に到達するまで、注意深く選択的にチューニングを行うことが求められます。

ユーザーとパフォーマンス・チームは、モニタリングの結果に応じてシステムのリソースとアプリケーションの構成を調整することにより、システム・パフォーマンスを最大にし、ダウン時間を削減し、潜在的な問題や危機的な状況を回避することを目指します。

システム・パフォーマンスの改善作業を行う時には、次のようなパフォーマンス・チューニングの一般的なガイドラインに従うことをお勧めします。

- 定量的で測定可能な、現実的な目標を立てます。
- システム全体を理解して検討します。
- 変更するパラメーターは一度に 1 つだけにします。
- 測定し、レベルに応じて再構成します。
- 設計を検討し、再設計します。
- 収穫逓減の法則を念頭に置きます。
- パフォーマンス・チューニングの限界をわきまえます。
- 構成の選択肢とトレードオフを理解します。
- チューニングのためだけのチューニングを行わないようにします。
- ハードウェアやソフトウェアの問題も点検します。
- チューニングを開始する前に、フォールバック手順を確立しておきます。

目標のスループットに向けて Content Collector をチューニングする

Content Collector のシステム・スループットを決める主要な要素は、次の 2 つです。

- スレッド数
- キュー・サイズ

この 2 つのパラメーターはどちらも、Configuration Manager の「ツール」メニューから**「タスク経路サービス構成」**を選択して設定できます。

スレッド数

スレッド数パラメーターによって、並行して処理される項目の最大数が決まります。経験法則として、CPU の物理的なコア 1 つに対して 2 から 4 のスレッドを設定すると、スループットの高いシナリオを構成できます。Content Collector の出荷時のデフォルト設定は 16 スレッドです。ほとんどのデプロイメント・シナリオでは、この設定が適切です。使用できる CPU コア数が多い場合は、スレッド数を増やすことができます。Content Collector サーバーのスループットを絞る (スロットルする) 必要がある場合は、数を減らすことができます。典型的なスロットルのシナリオとしては、対話式の検索、表示、復元のワークロードのために CPU リソースを確保する場合や、E メール・サーバーまたはファイル・サーバーの過負荷を避ける場合があります。スレッド数を 2 以下に設定することは避けてください。タスク経路サービスには少なくとも 2 つのスレッドが必要です。

キュー・サイズ

キュー・サイズ・パラメーターは、コレクターがサブミットできる項目の件数を定義します。サブミット・コネクターとタスク処理の間のバッファーのような働きをします。経験上、64 件 (ファイル・アーカイブ) から 128 件 (高スループットの E メール・アーカイブ・デプロイメント) の設定が良い結果を得ています。8 コアのサーバーに導入した Lotus Domino ベースのシステムに対して、最大で 256 件という値を使用できます (スケールアウト構成時)。

E メール・システムからの生のスループットの測定

33 ページの図 3 は、E メール・サーバーから E メールを取り込むために使用できるサンプル・タスク経路です。これを利用して、コンテンツのアーカイブや変更を行わずに E メール・サーバーから読み込めるデータの最大スループットを測定します。実際のアーカイブ・スループットは、多くの場合、この測定値より大幅に低くなります。メールをエンタープライズ・コンテンツ管理アーカイブにも書き込む必要があり、E メール・システムで「アーカイブ済み」とマークすることも必要だからです。このような不確実性があるとはいえ、このようなタスク経路を使用するサンプル処理により、多くの情報を入手できます (ファイル・アーカイブのシナリオについても、同じ原則が当てはまります)。

- 現在の E メール・サーバー・システムから読み取れる E メール数の最大スループットはどの程度か？

  この数が、必要なスループットより低いか、わずかに高い程度の場合は、この制限の原因 (この例では、E メール・サーバーのディスク入出力の限界や、WAN ネットワークの待ち時間の問題など) を解決するか、システム・アーキテクチャーを調整する (例えば、複数の E メール・サーバーを並行して稼働させる) 必要があります。

- アーカイブ対象のデータにどのような特性があるか？(例えばサイズなど)

  この情報を得るには、監査ログをスプレッドシート・ソフトウェアにインポートして分析します。

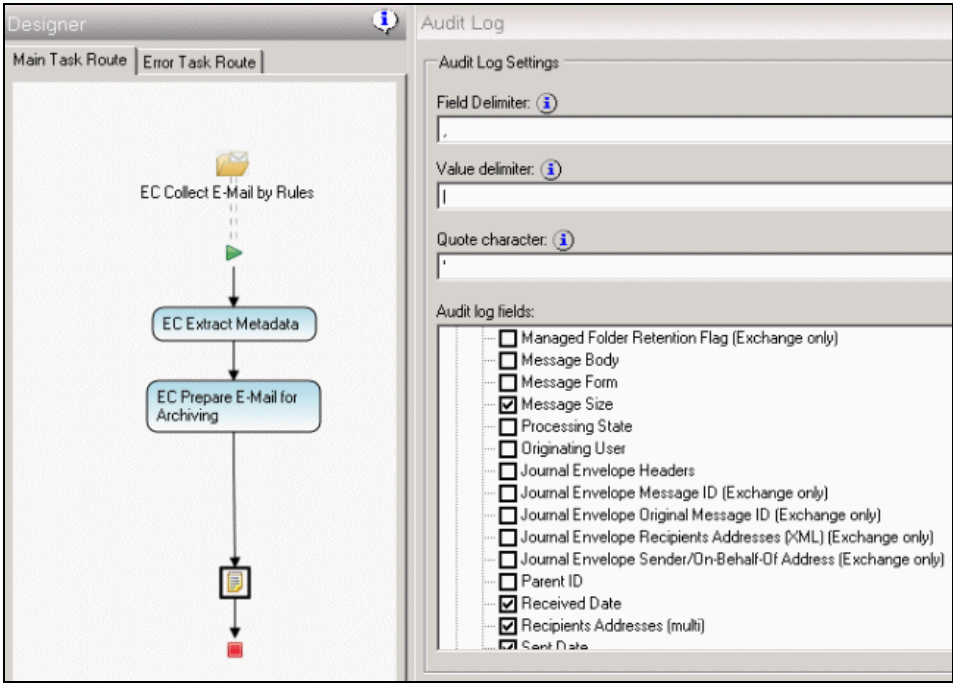

図 3　規則による E メールの収集、メタデータの抽出、およびアーカイブ用に準備のタスクで構成されるタスク経路

33 ページの図 3 に示したのは、次に挙げるコレクターとタスクで構成されたタスク経路です。

- EC の「規則による E メールの収集」コレクター
- EC の「メタデータの抽出」タスク : このタスクでは、宛先やメール・サイズなどのメタデータを抽出します。この情報は後ほど、監査タスクで利用できます。

- EC の「E メールをアーカイブ用に準備」タスク : このタスクでは、一時ファイルを Content Collector の作業ディレクトリーに書き込みます。
- 「監査ログ」タスク : 33 ページの図 3 の右側のペインは、監査ログに記録するデータの構成を示しています。この例では、メールの平均サイズ、平均の宛先数、平均の添付ファイル数を調べる目的で、「メール・サイズ (Message Size)」、「受信者アドレス (Recipients Addresses)」、および「添付ファイル数 (Attachment Count)」というメタデータ・フィールドを選択してあります。

現在デプロイされている E メール・サーバー・システムで達成できる潜在的なアーカイブ・スループットを調べるためにも、これと同様の方法を利用できます。システムが E メール・サーバーから E メールの添付ファイル・データも取得するように設定するために、EC の「添付ファイルの抽出」タスクを EC の「E メールをアーカイブ用に準備」タスクの後に追加する必要があります。添付ファイルの抽出を含むタスク経路と含まないタスク経路で達成されるスループットに大きな違いがある場合は、E メール・サーバーのディスクの制限か、ネットワーク帯域幅の制限がボトルネックになっている可能性が高いと考えられます。

Content Collector デプロイメントのスケーリング

Content Collector は、垂直方向へのスケーリング (複数の CPU コアを持つ性能の高いコンピューターを使用) と、水平方向へのスケーリング (Content Collector の追加のサーバー・インスタンスを使用) が可能です。どのようなスケーリング方法を選択するかは、デプロイされている E メール・サーバー・システムに大きく依存しています。IBM Lotus Domino ベースの E メール・システムの場合は、現行のマルチ・コア CPU のマシンによる 2-way または 4-way のシステムを使用した垂直スケーリングで良い結果が得られます。少し性能の低いマシンを複数使用するスケールアウト・アプローチも可能です。ハードウェアのコストという面でメリットが大きいのは複数のサーバーを使用する方法ですが、このアプローチを検討する際は、複数のシステムを管理するための管理コストを考慮に入れる必要があります。

Microsoft Exchange ベースのシステムでは通常、Content Collector デプロイメントのスケーリングに複数のサーバーを使用する必要があります。サーバー 1 台では、スループットの面で E メール・インターフェースが限定的になるからです。この環境では、多くの場合、少し性能の低いサーバーを多数使用する方法が効果を発揮します。

スケールアウト・セットアップでは、これまでの経験上、すべての Content Collector サーバーを同じハードウェア構成のマシンにデプロイする方法が最良の結果を得ています。スケールアウト・セットアップの 1 次サーバーとして機能するサーバー (ワークロードの分散を管理する) の CPU 要件が、拡張ノードのサーバーより若干高くなります。しかし、高可用性メカニズムが働くため、

どのサーバーが 1 次サーバーの役割を担うかは予測が難しく、運用中にこのワークロードが別のサーバーにシフトすることもあります。

Content Collector は、デフォルトで、大規模なメールボックス (例えば、ジャーナル・メールボックス) のアーカイブに対して最高のスループットを達成できるように最適化されています。そのため、スケールアウト・セットアップでは、現在選択されているメールボックスの作業にすべての Content Collector サーバーが処理能力を結集することになります。サーバー群は一度に 1 つのメールボックスを処理します。このメールボックスで現在選択されている項目の処理が終わると、次のメールボックスの処理に移ります。したがって、標準的な構成では、Content Collector が生成する E メール・サーバー負荷が一度に 1 台の E メール・サーバーにいつも集中することになります。E メール・サーバーが 1 台の Content Collector サーバーの生成する負荷を処理するだけで限界に達している場合、または Content Collector サーバーと E メール・サーバーの間のネットワーク接続がボトルネックになっている場合は、スケールアウト・セットアップで Content Collector サーバーの台数が増えると、既存のボトルネックにさらに負荷が掛かります。この状況が順繰りに発生して過負荷状態になり、セットアップ全体の合計処理能力が落ちることになります。このような状況の場合、複数のサーバーからのコンテンツを並行してアーカイブできるように構成を変更する必要があります。

Content Collector が複数の E メール・ソースを並行して処理できるようにチューニングするには、特定の Content Collector インスタンスに、異なるサーバー上にある特定の E メール・ユーザー・グループを割り当てます。IBM Lotus Domino を使用している場合は、複数の Content Collector インスタンスを構成しないでも、E メール・サーバー・コネクターの複数のクローラー・スレッドを構成することにより、同じ効果を得ることができます。クローラー・スレッドを調整する方法は、使用している Content Collector のバージョンによって異なるため、IBM サポート担当員に連絡してください。

テキスト検索のスケーラビリティーの確保

大規模なテキスト索引からコンテンツを検索する操作は、負荷の高い操作です。全文検索の機能をユーザーに提供し、公的機関からの大量の検索要請に対応するための負荷を減らすには、1 回の検索要求を実行するために検索コンポーネントが処理する作業量を最小限に抑えることが重要です。検索負荷を大きく低減するのに役立つ方法は、全文検索ではいつも日付範囲を指定するようにユーザーに習慣づけてもらうことです。

ヒント： メールボックス管理シナリオでは、ユーザー検索に対してデフォルトの日付範囲をいつも指定するようにします。

Content Collector の検索ユーザー・インターフェースでデフォルトの日付範囲をユーザーに推奨するには、環境変数 AFU_DATE_RANGE_IN_MONTHS にゼロ

より大きい値を設定します。この設定を行うと、指定した月数プラス今月の数値がデフォルトとして検索フォームの日付フィールドに表示されます。例えば、ユーザーが検索フィールドに特定の語句を指定して単純な照会を実行する場合、日付範囲を変更せずにその照会をサブミットすると、アーカイブ全体ではなく、指定した最近の月数以内だけが検索対象になります。デフォルトの動作では (Content Collector Web アプリケーションの始動時にこの環境変数が存在しない場合)、日付フィールドに値がまったく自動入力されないため、負荷の低減に役立つ日付範囲をユーザーが意識的に指定しなければなりません。

日付範囲を限定した検索を、日付で区分されたテキスト検索索引と併用することには、次の 2 つの利点があります。

- 処理してユーザーのクライアント・マシンに送り返す必要のある検索結果が少なくなります。
- 日付範囲を基準にしたデータ分割を IBM Content Manager (日付範囲で分割した複数の項目タイプを使用する) および FileNet P8 (索引範囲ごとにデータ分割してアーカイブする) で構成すれば、個々の照会で日付範囲を指定することにより、検索サーバーの負荷を軽減できます。照会の対象が、それぞれ指定の日付範囲に対応するテキスト検索索引または索引コレクションに限定されるからです。例えば、過去 4 年間の E メールが 6 カ月ごとに 1 つの索引という日付範囲で分割アーカイブされている場合であれば、過去 1 年間の E メールを対象にした照会は、1 年分 (合計テキスト索引サイズの 4 分の 1) に対応する 2 つの全文索引を検索すればよいことになります。これにより、検索サーバーのディスク入出力負荷が大幅に軽減されます。古いテキスト検索索引 (一般に使用頻度が低い) をコストの安いストレージに移すこともコスト軽減に役立ちます。

日付範囲による分割を設定する方法の詳細については、FileNet P8 の資料を参照してください (**「システム管理 (System Administration)」→「コンテンツ・エンジンの管理 (Content Engine Administration)」→「コンテンツ・ベース・リトリーブ (Content-based retrieval)」を選択**)。この機能は、FileNet P8 Version 4.5.1 以降でサポートされています。また、Content Collector スタイル・セット ICC_FileSystem_PushAPI_2.1_p8cse_4.5.1 以降をこの機能と併用することが重要です。さらに、Content Collector Search Application の構成を変更して、Content Engine の索引分割で使用するように構成したのと同じデータベース・プロパティー (ICCMailDate) を使用して日付範囲を照会するように設定することも必要です。詳細については、次のアドレスを参照してください。

http://publib.boulder.ibm.com/infocenter/p8docs/v4r5m1/index.jsp?topic=/com.ibm.p8.doc/ce_help/cbr/cb_about_verity_partitions.htm

適切な日付範囲を選択する方法は、毎日のアーカイブ負荷に応じて大きく変わります。1 つの索引区画には、数百万件から数千万件の文書を保持できます (パラメーター設定や使用シナリオに応じて異なる)。E メールのアーカイブの場合、この資料の執筆時点での一般的な経験法則として、IBM Content Manager の項目タイプに収容する E メールを 5000 万件未満にし、FileNet P8 を使用して

索引範囲の構成を 500 から 800 万件の E メールごとに新しいコレクションを作成するように設定するのが最適です。この数値はハード制限ではなく、システム設計を計画する際のガイドラインです。

ウィルス・スキャナーを適切に構成する

次の 2 つの理由で、ウィルス・スキャナーが Content Collector の一時ディレクトリーと作業ディレクトリーを除外するように構成する必要があります。

- パフォーマンス。アーカイブ対象のコンテンツは、多くの場合、もっと早い段階でウィルス・スキャナーの処理を受けているため、アーカイブの段階で作成される一時ファイルをチェックすることはリソースの無駄遣いであり、アーカイブ動作が低速化します。
- ウィルスの疑いがある場合は、Content Collector の一時ファイルにロックが検出されるため、Content Collector の処理でエラーが発生します。これらのエラーについては、状況を分析して原因を究明する必要があります。

エンタープライズ・アーカイブをウィルス・スキャナーがどのように処理するか (もしも処理する場合) については、慎重に計画する必要があります。これにはテキスト索引エンジンも含まれます。テキスト索引の生成中には、深くネストしたファイル・ディレクトリー構造がファイル・ストレージ域と一時領域に作成されるため、ウィルス・スキャナーがこれらのディレクトリーを処理すると、サーバーのディスク負荷が大幅に増加します。これと同様の規則が、デスクトップのテキスト索引生成プロダクト (例えば、Microsoft Windows Search) に当てはまります。不必要なシステム負荷を生成しないために、これらのサービスをオフにするべきです。

テキスト索引エンジンが作成するテキスト索引ファイルは、ウィルス・スキャナーで処理しないようにすることを強くお勧めします。ウィルス・スキャナーがテキスト索引内でウィルスを誤検出したために、索引更新の重要な段階で索引ファイルのアクセスがブロックされてファイルが破損し、それらのファイルを再生成する必要が生じるというケースをいくつも経験しました。

高可用性およびロード・バランシングの計画

Content Collector サーバーが提供するサービスには、次の 2 種類があります。

- アーカイブ・サービス
- 対話式サービス (検索、表示、および復元操作)

アーカイブ・サービス

スケールアウト・セットアップで複数の Content Collector サーバーがインストールされている場合、Content Collector はアーカイブ・サービスに対して高可用性を提供します。

アーカイブ機能は、タスク・ルーティング・サービスによって制御されます。この機能は、内部的な概念として、1 次ノード (アーカイブ・タスクを実行するとともに、作業の分配を管理する) と拡張ノード (1 次ノードに対して 2 次ノードの役割を果たし、1 次ノードから割り当てられた項目のアーカイブだけを行う) に分類されます。

スケールアウト環境にインストールした Content Collector の場合、1 次ノード上のタスク・ルーティング・サービスにより、すべての Content Collector サーバー・インスタンス (1 次ノード自身とすべての拡張ノード) に作業が等しく分配されます。1 次ノード上のタスク経路サービスが 2 次ノードと通信不能になったことを検出すると、そのノードに対してそれ以降作業をサブミットすることをせず、現在そのノードに割り当てられている項目を失われたものとして扱います。E メールの場合で言うと、それらの E メールはアーカイブ済みとしてマークされず、送信元のメールボックスを次回に Content Collector がクロールした時点で E メールがアーカイブされます。また、現在の 1 次ノードが操作不能になった場合は、残りの 2 次ノードがその状況を検出し、2 次ノードのいずれかが 1 次ノードの役割を引き継ぐので、処理が確実に続行されます。

対話式サービス (検索、表示、および復元操作)

これらの対話式サービスを提供している Content Collector サーバーに障害が発生した場合のフェイルオーバーのためには、外部にメカニズムを用意する必要があります。

検索、表示、復元操作のための対話式サービスは、ユーザーの検索クライアントにある Content Collector の拡張機能から、または E メール Web アクセス・サーバーから直接呼び出されます。このサービスは、各 Content Collector サーバー上に組み込まれている IBM WebSphere® Application Server で実行される Web サービス (標準インストールの場合) によって提供されます。デフォルトでは、インストール時に 1 次サーバーとして指定された Content Collector サーバーが、対話式サービスをクライアントに提供します。そのため、このマシンのホスト名が、スタブ・リンクで、およびユーザーの E メール検索アプリケーションで、ホスト名 (または別名) として使用されます。

このサーバーが利用不能になった場合、拡張ノードのいずれか 1 つが、構成されたホスト名またはホスト名の別名によって E メール・クライアントの拡張機能からアクセス可能になるというメカニズムを設定する必要があります。対話式サービスの中断は、多くの場合、例えばアーカイブ・プロセスの中断よりもはるかに影響が大きくなります (ユーザーの目にとまるので)。

Content Collector の機能は構成データベースに依存しているため、そのデータベースにもそれ自身の高可用性メカニズムが必要です。このデータベースが利用不能になると、すべての Content Collector サーバーが処理を停止します。

対話式サービスに高可用性を提供するためのソリューションはいくつかありますが、現場で使用されている 2 つの例をご紹介します。

- クラスター・サービスの利用

  Web サービスを提供するサーバーに障害が発生したとき、スタンバイ・サーバーがアクティブになります。このセットアップを選択する場合に使用可能なアーキテクチャーは、スケールアウト・セットアップによる 2 台のサーバー (アーカイブ・サービスに対して Active-Active) で構成されます。クラスター・サービスが使用されるのは、対話式サービスの高可用性の目的に対してだけです (対話式サービスに対して Active-Passive)。

- Web サービス・コンポーネントの前にロード・バランサーを配置

  ロード・バランサーの利用は、高可用性を達成するための良い方法です。それと同時に、すべての Content Collector サーバーの Web サービス・コンポーネントを並行して使用することにより、多数の対話式要求を Content Collector が処理できるようになります。

ロード・バランサーを利用するセットアップでは、Content Collector の構成で、すべての Content Collector サーバーに対して単一の別名として使用されるアーカイブ用のホスト名の別名 (例えば、archive.example.com) をシステムに割り当てます。その構成に対して、E メール・クライアント拡張機能は、すべての対話式要求を処理するために (https を使用して) このマシンにアクセスします。その後、ロード・バランサーがこれらの要求を Content Collector サーバーのクラスターに転送します。

ロード・バランサーそのものが、障害の発生する新たなポイントをシステムに持ち込むので、高可用性のためにバックアップを用意する必要があります。

さらに、https 接続を終了させるためにロード・バランサーを使用し、Content Collector サーバーにはその接続を http を使用して転送することができます。こうすることにより、負荷の高い https ハンドシェーク操作から Content Collector サーバーを解放できます。

図 4 に、ロード・バランサーを使用した Content Collector システムのセットアップを示します。

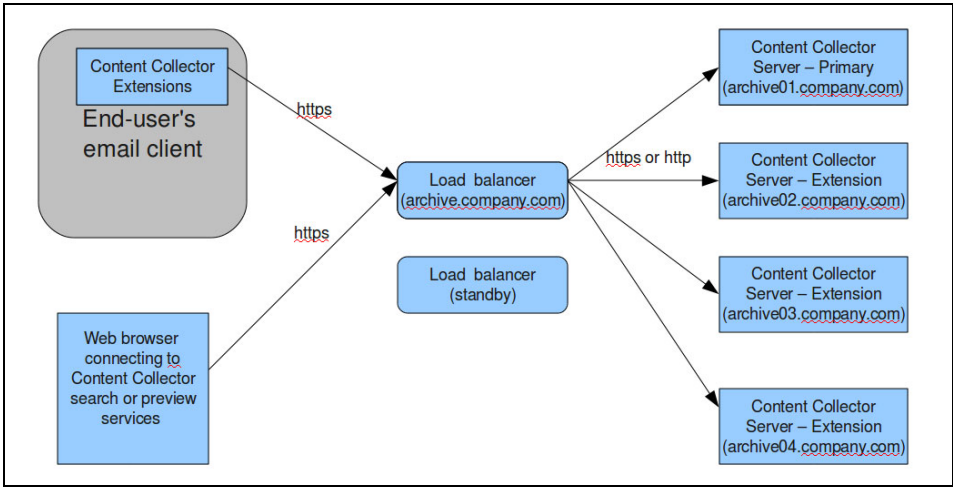

図 4　ロード・バランサーを使用する Content Collector システム

この構成 (図 4) では、Content Collector Web アプリケーションと構成 Web サービスに対するすべての要求を、Content Collector サーバーの前に位置するロード・バランサーが受け取ります。その要求はホスト名 archive.company.com に対するものです。すべての要求はポート 11443 上の SSL 経由で受信されます。SSL トンネルは、ロード・バランサーのところで終了します。その後、要求は HTTP 経由で Content Collector サーバーのポート 11080 に転送されます。

セッションは*ネゴシエーションで設定されたセッション ID（cookie）をベースにしてスティッキー* であることが必要です。セッションをオープンして Content Collector サーバーに認証されたユーザーが、引き続き同じサーバーに対してセッションを維持するためです。

各 Content Collector サーバー上で Content Collector Web アプリケーションが活動状態であり、操作可能であることを確認するには、ロード・バランサーから定期的に Content Collector Web アプリケーションの次の URL を呼び出すことができます。

`http://<Content CollectorServer>:11080/AFUConfig/status`

構成 Web サービスは、すべての Content Collector サーバー上で実行されている必要があります。その Web サービスは、構成データベースにアクセスするために JDBC を使用します。すべてのノード上で JDBC アクセスを構成するには、次の手順を実行します。

1. 1 次ノードのセットアップを完了した後、….¥ContentCollector¥ctms¥scripts から `jdbcInstall.cmd` ファイルをコピーします。
2. アーカイブ・ユーザーに合わせて、パスワード、パス、およびセル名を編集します。

3. jdbcInstall.cmd ファイルを、すべての 2 次ノードの同じディレクトリーにコピーします。
4. すべてのノード上で **jdbcInstall.cmd** を実行します。

これで、すべてのノードから構成アクセスが可能になります。このことを検証するには、すべてのノードで仮想状況 URL にアクセスします。

このセットアップを構成する最も簡単な方法は次のとおりです。すべてのサーバーを 1 つずつ独立して操作可能な状態にセットアップします。次に、すべてのトラフィックを一度に 1 台のサーバーにリダイレクトしてテストします。最後に、ロード・バランサーの構成を変更して、すべてのサーバー・インスタンスに対する負荷の平衡化をアクティブにします。

フェイルオーバー・シナリオでの Configuration Manager の使用

Configuration Manager コンポーネントは、デフォルトで、Content Collector デプロイメントの 1 次ノードに対する書き込みアクセスが可能な状態でインストールされます。構成データベース内の表により、構成に対する書き込みアクセスが制御されます。

1 次ノードに物理的な障害またはオペレーティング・システムの障害が発生した場合には、いずれかの拡張ノードの Content Collector セットアップを使用して構成を変更する必要が生じます。その場合は、次の手順を実行してください。

1. システム・レジストリーの HKEY_LOCAL_MACHINE¥SOFTWARE¥IBM¥Content Collector¥Server で、インストール・タイプの値を NODE_B (拡張ノードを表す) から NODE_A (1 次ノードを表す) に変更します。
2. 前のステップでキー値を変更した後、「IBM Content Collector GUI Components」サービスを開始します。

この手順を実行した後、Configuration Manager を始動すれば、構成データベースへの書き込みアクセスが可能になります。同時に複数のサーバーの書き込みアクセスを可能にするのは危険です。複数の Configuration Manager インスタンスが構成データベースを書き込みアクセスのために同時にオープンすると、構成が無効になることがあります。

バックアップおよび災害復旧の計画

Content Collector コンポーネントのバックアップ、エンタープライズ・コンテンツ管理アーカイブのバックアップ、そしてテキスト検索索引のバックアップは、それぞれ独立して実行できます。しかし、理想的には、これらのアーカイブ環境全体をまとめてバックアップおよびリストアする戦略を構築しておく必要があります。このシナリオが重要な理由は、アプリケーション間の状態を保存するため、すなわちデータを失う危険性を回避するためです。

Content Collector サーバーは一時データだけを保持します。一時データはバックアップの必要がありません。サーバーのインストール時やアップグレード時に 1 回バックアップを行えば十分です。Content Collector で唯一バックアップが必要なコンポーネントは、構成データベースです。このデータベースは、エンタープライズ・コンテンツ管理のアーカイブ・データベース・システム上に置くことをお勧めします。アーカイブ・データベースのバックアップ・メカニズムを活用できるからです。

災害復旧のためには、ハードウェア・レベルのレプリケーションにより、書き込み順序が保持される方法で重要なデータ・セットをバックアップすることをお勧めします (この後の説明を参照)。しかし、災害復旧サイトへの複製は、従来のバックアップの代用にはなりません。1 次サイトのデータが破損するエラーが発生した場合に、その破損も災害復旧サイトに複製されてしまうからです。そのレプリケーションが唯一のバックアップであるとすれば、復旧が不可能になります。

それぞれのビルディング・ブロックによって維持されているデータは、他のビルディング・ブロックとの参照整合性の上に成り立っています。アーカイブ環境でアプリケーション相互間の関係が失われると、不整合が発生する恐れがあります。そのような不整合には、解決が困難なものと、解決不可能なものがあります。バックアップ戦略では、この関係を保持する必要があるため、データを別にした環境のバックアップを実行しなければなりません。この操作を毎日のバックアップ時間枠の中で行うのは、磁気テープのパフォーマンス上の特性から、実行が困難です。

SAN および NAS ストレージのスナップショット機能を使用するか、論理ボリューム・マネージャーを使用すると、バックアップに要するダウン時間を最小限に抑えられます。エンタープライズ・コンテンツ管理アプリケーションの変更可能なデータ (メタデータ用のデータベース、全文索引のコレクション、およびファイル・ストレージ域) をすべて SAN 上に格納すると仮定すると、スナップショットを利用して、アプリケーションのポイント・イン・タイム・コピーをファイル・システム・バックアップに適した状態で取り込むことができます。アプリケーション・コンポーネントがシャットダウンしている間に、変更可能なデータのスナップショットをとります。大部分のスナップショット技術はほとんど瞬時に実行されるので、磁気テープへの転送を待つ必要がなく、シャットダウンしたアプリケーション・コンポーネントを短時間のうちに元に

戻せます。その後、スナップショットを代替ロケーションにマウントし、バックアップ・ソフトウェアを使用してファイル・システムをバックアップできます。

固定コンテンツ・デバイスを使用するには、スナップショットをとることに加えて、連続的な差分バックアップが必要になります。孤立したデータ (対応するメタデータのないコンテンツ) が発生することは、それとは逆の状態 (対応するコンテンツの欠落したメタデータの発生) より見逃されることが多いようです。ユーザーの目に見えるエラーが起きないからです。この理由により、固定ストレージ・バックアップは、変更可能なデータのスナップショットをとってから行う必要があります。この方法の場合も、すべてのサービスをシャットダウンし、すべてのデバイスのバックアップを行い、その後にサービスを復元するという方法に比べて、バックアップ時間枠を最小限に抑えることができます。

このバックアップ戦略は、次のいずれかの状況が当てはまる場合に適切です。

- アプリケーションによってサポートされている部門や技術部門が長時間のバックアップに耐えられないか、営業時間帯の制約でバックアップ時間枠の延長が不可能な場合。
- アプリケーションによってサポートされている部門が長期間のリカバリー・ポイントを設定した場合の財務上の責務に対応できず、データの損失を 1 日以内にすることを要求している場合。
- アプリケーションによってサポートされている部門が、一般用途のためにアプリケーションが利用不能になる期間が日単位で延長されることに耐えられない場合。

スナップショットを利用すると、バックアップ時間枠が短くなります。また、すべてのアプリケーション・コンポーネントの間でバックアップの参照整合性が確保されるため、復元に成功する確率が大幅に上がります。

変更可能なデータすべてのスナップショットを整合性のある状態で作成するには、全データを 1 つのストレージ・アレイに格納するか、IBM SAN ボリューム・コントローラーなどのストレージ・バーチャリゼーション製品を使用する必要があります。さらに、スナップショットは、ファイル・システム・レベルでのバックアップが必要ありません。最高のパフォーマンスを達成するために、ボリュームをブロック・レベルでバックアップする (元のコンテンツをビット単位で複製する) ことができます。

このレッドペーパーの著者チーム

この資料の著者は、世界各国で業務にあたっている International Technical Support Organization, Rochester Center の専門家たちです。

Wei-Dong Zhu (ジャッキー)。International Technical Support Organization のエンタープライズ・コンテンツ管理のプロジェクト・マネージャー。会計、イメージ・ワークフロー処理、デジタル・メディア配布といった分野のソフトウェア開発業務に 10 年を超える経験。サザン・カリフォルニア大学のコンピューター・サイエンス学科で修士号を取得。IBM への入社は 1996 年。IBM Content Manager の公認ソリューション・デザイナー。エンタープライズ・コンテンツ管理関連の多数の IBM Redbooks® の制作でマネージャーとして参加。

Markus Lorch。ドイツ IBM の顧問ソフトウェア・エンジニア。エンタープライズ・コンテンツ管理ソフトウェアおよびコンテンツ・ディスカバリー・ソフトウェアの開発分野で 5 年を超える経験。IBM Content Collector のパフォーマンスおよびスケーラビリティー・エンジニアリング活動のリーダー。バージニア工科大学でコンピューター・サイエンスの博士号を取得。専門分野は、ソフトウェアのパフォーマンス、コンテンツ・アーカイブ、テキスト分析、エンタープライズ・サーチ、分散システムのセキュリティー。権限とセキュリティー、グリッドおよびクラスター・コンピューティング、セマンティクス・サーチ・メカニズムの分野で、論文審査のある 20 を上回る文書を執筆。

このプロジェクトに貢献した下記の方々に謝意を表します。

Dieter Schieber
Silke Wastl
Thorsten Hammerling
IBM ソフトウェア・グループ、ドイツ IBM

あなたも著者になれます！

あなたのスキルに光を当て、キャリアを育成し、著者にもなれる。このすべての目標を同時に達成できます。ITSO の常駐プロジェクトに参加して、あなたの専門分野の資料の執筆を補佐しながら、最先端のテクノロジーでキャリアを磨いてください。製品の評判とお客様の満足度を高める手伝いをしながら、技術面での人的ネットワークを広げることができます。常駐プロジェクトの期間は 2 週間から 6 週間です。個人として、または自宅で遠隔スタッフとしての参加が可能です。

常駐プログラムについて詳しくは、下記の常駐インデックスを参照のうえ、オンラインでご応募ください。

ibm.com/redbooks/residencies.html

IBM Redbooks の最新情報

- Facebook:

  http://www.facebook.com/IBMRedbooks

- Twitter:

  http://twitter.com/ibmredbooks

- LinkedIn:

  http://www.linkedin.com/groups?home=&gid=2130806

- IBM Redbooks 週刊ニュースレター (Redbooks 資料の最新情報。常駐プロジェクトとワークショップに関する情報):

  https://www.redbooks.ibm.com/Redbooks.nsf/subscribe?OpenForm

- RSS フィードで Redbooks 資料の最新情報を入手 :

  http://www.redbooks.ibm.com/rss.html

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、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。

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。

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。

著作権使用許諾 :
本書には、様々なオペレーティング・プラットフォームでのプログラミング手法を例示するサンプル・アプリケーション・プログラムがソース言語で掲載されています。お客様は、サンプル・プログラムが書かれているオペレーティング・プラットフォームのアプリケーション・プログラミング・インターフェースに準拠したアプリケーション・プログラムの開発、使用、販売、配布を目的として、いかなる形式においても、IBM に対価を支払うことなくこれを複製し、改変し、配布することができます。このサンプル・プログラムは、あらゆる条件下における完全なテストを経ていません。従って IBM は、これらのサンプル・

プログラムについて信頼性、利便性もしくは機能性があることをほのめかしたり、保証することはできません。お客様は、IBM のアプリケーション・プログラミング・インターフェースに準拠したアプリケーション・プログラムの開発、使用、販売、配布を目的として、いかなる形式においても、IBM に対価を支払うことなくこれを複製し、改変し、配布することができます。

本書、SG88-4064-00 は 2011 年 11 月 22 日に作成 / 更新されました。

ご意見は、以下のいずれかの方法でお送りください。
- オンラインの「**Contact us**」Redbooks レビュー・フォームを使用する。
  ibm.com/redbooks
- ご意見を E メールで以下の宛先に送信する。
  redbook@us.ibm.com
- ご意見を以下の宛先に郵送で送付する。
  IBM Corporation, International Technical Support Organization
  Dept. HYTD Mail Station P099, 2455 South Road
  Poughkeepsie, NY 12601-5400 U.S.A.

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 Corp. 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。

以下は、International Business Machines Corporation の米国およびその他の国における商標です。

AIX®	Lotus Notes®	Redbooks (ロゴ) ®
DB2®	Lotus®	Solid®
Domino®	Notes®	WebSphere®
FileNet®	Redbooks®	
IBM®	Redpaper™	

Microsoft、Windows 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。

SG88-4064-00