コスト最適化の柱

AWS Well-Architected フレームワーク

2020 年 7 月

This paper has been archived.

The latest version is now available at:

https://docs.aws.amazon.com/ja_jp/wellarchitected/latest/cost-optimization-pillar/welcome.html

注意

お客様は、この文書に記載されている情報を独自に評価する責任を負います。本書は、(a) 情報提供のみを目的としており、(b) AWS の現行製品と慣行ついて説明していますが、予告なしに変更されることがあり、(c) AWS およびその関連会社、サプライヤーまたはライセンサーからの契約上の義務や保証をもたらすものではありません。AWS の製品やサービスは、明示または暗示を問わず、一切の保証、表明、条件なしに「現状のまま」提供されます。お客様に対する AWS の責任は、AWS 契約により規定されます。本書は、AWS とお客様の間で締結される一切の契約の一部ではなく、その内容を修正することもありません。

目次

要約

このホワイトペーパーでは、アマゾン ウェブ サービス (AWS) Well-Architected フレームワークのコスト最適化の柱に焦点を当てます。このホワイトペーパーは、お客様が AWS 環境の設計、デリバリー、メンテナンスにベストプラクティスを適用することを支援するガイダンスとなります。

コストが最適化されたワークロードでは、すべてのリソースがフル活用され、コストの最小化を達成して、機能要件が満たされます。このホワイトペーパーは、組織内での機能の構築、ワークロードの設計、サービスの選択、サービスの構成と運用、およびコスト最適化手法の適用に関する詳細なガイダンスを提供します。

はじめに

AWS Well-Architected フレームワークは、AWS でワークロードを構築する際に行う決定を理解するのに役立ちます。このフレームワークは、信頼性が高く、安全かつ効率的で、コスト効率に優れたシステムをクラウド内で設計および運用するためのアーキテクチャ設計のベストプラクティスを提供します。アーキテクチャをベストプラクティスに照らして評価し、改善すべき領域を特定する一貫した方法を紹介します。AWS では、Well-Architected ワークロードを備えることによって、ビジネス成功の可能性が大幅に高まると確信しています。

このフレームワークは次の 5 つの柱に基づいています。

- 運用上の優秀性
- セキュリティ
- 信頼性
- パフォーマンス効率
- コスト最適化

このホワイトペーパーでは、コスト最適化の柱と、サービスやリソースを最も効果的に活用したワークロードの設計方法、最小限のコストでビジネス成果を達成する方法を重点を置いています。

皆さんは今から、コスト最適化の柱のベストプラクティスを組織に適用する方法について学習します。従来のオンプレミスソリューションでは、コストを最適化するためのハードルにぶつかる可能性があります。これは将来のキャパシティーやビジネスニーズを予測すると同時に、複雑な調達プロセスを進める必要があるためです。このホワイトペーパーのプラクティスを適用すると、皆さんの組織が以下の目標達成を支援します。

- クラウドの財務管理の実践
- 支出と使用量の認識
- 費用対効果の高いリソース
- 需要と供給を一致させる
- 継続的最適化

本ホワイトペーパーの対象者はテクノロジーとファイナンスの担当者、たとえば最高技術責任者 (CTO)、最高財務責任者 (CFO)、設計者、開発者、ファイナンシャルコントローラー、ファイナンシャルプランナー、ビジネスアナリスト、オペレーションチームメンバーなどの方々です。本ホワイトペーパーでは、運用の詳細やアーキテクチャのパターンについては説明しておりませんが、該当するリソースへの参照先が記載されています。

コスト最適化

コスト最適化とは、システムのワークロードのライフサイクル全体にわたって改良、改善する継続的プロセスです。本ホワイトペーパーのプラクティスは、組織がコストを最小限に抑えて投資利益率を最大化すると同時に、ビジネス成果の達成につながるようなコストを意識したワークロードの構築および運用を支援します。

設計原則

コスト最適化のため、以下の設計原則を検討します。

クラウドの財務管理の運用: クラウドで財務上の成功を達成し、ビジネス価値の実現を加速させるには、クラウドの財務管理に投資する必要があります。組織は、テクノロジーと使用量管理の新たなドメインで機能を構築するために必要な時間とリソースを投入する必要があります。セキュリティやシステム運用力と同様に、コスト効率の高い組織になるために、知識の構築、プログラム、リソース、プロセスを通じて機能を構築する必要があります。

消費モデルを導入する: コンピューティングリソースの使用分のみを支払い、ビジネス要件に応じて使用量を増減できます。たとえば、週の稼働日に開発環境とテスト環境を使用するのは一般的には、1 日あたり 8 時間程度にすぎません。このケースでリソースを不使用時に停止すると、コストを 75% 削減できる可能性 (168 時間から 40 時間に減少) があります。

全体的な効率を測定する: ワークロードのビジネスの成果とデリバリー関連コストを測定します。このデータを利用すると、生産性および機能性の向上とコスト削減から得られるメリットを理解することができます。

付加価値を生まない**作業への資金投入をやめる:** AWS は、サーバーの設置、積み上げ、電力供給などのデータセンターの手間のかかる運用作業を行います。また、マネージドサービスを使用することで、オペレーティングシステムやアプリケーションの管理に伴う運用上の負担も解消されます。この結果、 IT インフラストラクチャよりも顧客と組織のプロジェクトに集中できるようになります。

費用を分析し、帰結させる:クラウドでは、システムの使用状況とコストを正確に特定し、IT コストと各ワークロードの所有者との帰属関係が明瞭になります。これによって投資収益率 (ROI) を把握できるため、ワークロードオーナーはリソースを最適化してコストを削減する機会が得られます。

定義

クラウドでのコスト最適化には、5 つの重点分野があります。

- クラウドの財務管理の実践
- 支出と使用量の認識
- 費用対効果の高いリソース
- 需要と供給を一致させる
- 継続的最適化

この柱にも、Well-Architected フレームワーク内の他の柱と同様に、コスト最適化のために検討すべきトレードオフがあります。たとえば、市場投入に要する期間と、コストのどちらを優先すべきでしょうか。市場投入に要する期間の短縮、新機能導入、納期順守といったケースでは、前払いコストの投資を最適化するよりも、スピードを重視して最適化することが最善なことがあります。

長期的にコストを最適化できるワークロードのベンチマークの選定に時間を掛けるよりも、「万一の場合」の備えを過度に重視してしまう傾向が常にあるため、データではなく時間的制約によって設計上の決断が下されることがあります過大な見積もりの結果、オーバープロビジョニングになり、最適化が不十分なデプロイを行ってしまいます。ただし、オンプレミス環境からクラウド環境に「リフトアンドシフト」式にリソースを移行してから最適化する必要がある場合は、選択肢としては妥当なこともあります。

適切な労力を当初からコスト最適化戦略に投入すると、ベストプラクティスが一貫して適用され、不要なオーバープロビジョニングも回避できるため、クラウドの経済的メリットをより早く実感できます。以下のセクションでは、クラウドの財務管理の初期および継続的な実装と、ワークロードのコスト最適化のための手法とベストプラクティスを提供します。

クラウドの財務管理の実践

組織はクラウド財務管理 (CFM) を使用すると、AWS でのコストと使用状況を最適化し、スケーリングすることで、ビジネス価値と経済的成功を実現できます。

以下は、クラウドの財務管理におけるベストプラクティスです。

- 機能オーナーシップ
- 財務とテクノロジーのパートナーシップ
- クラウドの予算と予測

- コストを意識したプロセス
- コストを意識した文化
- コスト最適化によるビジネス価値の数値化

職務機能

コスト最適化担当を設定する: この担当は、コストを意識した文化を確立、維持する責任を負います。これは社内の個人でも、チームでもかまいません。組織全体から財務、テクノロジーなどの主なステークホルダーを集めてチームを新規編成することもできます。

担当者 (個人またはチーム) は、コスト管理とコスト最適化活動に必要な時間を優先順序を付けて配分します。小規模な組織の場合、大企業のフルタイムの担当と比較すると、費やす時間の割合は少ない場合があります。

担当者は、プロジェクト管理、データサイエンス、財務分析、ソフトウェアやインフラストラクチャ開発の知識などの複数分野にわたるアプローチが必要になります。担当者は、コストの最適化（集中型アプローチ）の実行、テクノロジーチームの最適化（分散型）、またはその両方の組み合わせ（ハイブリッド）の実行により、ワークロードの効率性を向上させることができます。この担当者は、コスト最適化目標 (ワークロード効率メトリクスなど) に対する実行および提供能力を評価されることになります。

この担当者には、エグゼクティブスポンサーを確保する必要があります。エグゼクティブスポンサーは、クラウド利用のコスト効率を判断する最高責任者として、担当者の考え方を上長にエスカレーションして、組織が定める優先事項としてコスト最適化活動が扱われるようサポートします。エグゼクティブスポンサーと担当者は協力し、組織のクラウド利用を効率化して、ビジネス価値の実現を継続できるようにします。

財務とテクノロジーのパートナーシップ

ファイナンス部門とテクノロジー部門のパートナーシップを確立する: 承認、調達、インフラストラクチャデプロイのサイクルが短縮されるため、テクノロジーチームのクラウドにおけるイノベーションが促進します。ファイナンス組織はこれまでプロジェクト承認時のデータセンターやオンプレミス環境の調達に大量に使用されていた時間とリソースを調整することができます。

ファイナンス部門とテクノロジー部門という重要なステークホルダー同士のパートナーシップを確立し、組織としての目標の共通理解を得て、クラウドコンピューティングのさまざまな利用モデルにおいて財務上の成功を収めるメカニズムを作り上げます。クラウドジャーニーのすべてのステージにおいて、コストと使用量に関するディスカッションに参加する必要がある組織内の関連するチームは、以下の。

- **ファイナンシャルリード:** CFO、ファイナンシャルコントローラー、フィナンシャルプランナー、ビジネスアナリスト、調達担当、契約業務担当、支払担当は、クラウド消費モデル、購入オプション、月次請求プロセスを理解する必要があります。クラウド運用にはオンプレミスのオペレーションと比べて根本的な違い (使用量の変動率、従量課金制やティア別課金制、料金モデル、請求明細と使用量情報など) があるため、クラウド利用が調達プロセス、インセンティブ追跡、コスト配分、財務諸表などのビジネス局面に与えるインパクトをファイナンス部門で理解することが不可欠です。
- **テクノロジーリード:** テクノロジーリード (製品およびアプリケーションオーナーを含む) は、財務要件 (予算の制約など) やビジネス要件 (サービスレベルアグリーメントなど) を認識する必要があります。これにより、組織が目指すビジネス目標を達成するワークロードの導入が可能になります。

ファイナンスとテクノロジーのパートナーシップには、以下のような利点があります。

- ファイナンスチームとテクノロジーチームは、コストと使用量をほぼリアルタイムで把握できる。
- ファイナンスチームとテクノロジーチームは、クラウドへの支出の変動をハンドリングするため、標準となる運用手順を確立する。

- ファイナンスのステークホルダーは、コミットメント割引 (リザーブドインスタンスや AWS Savings Plans など) の購入に資金がどう使用されるかや、組織拡大のためにクラウドがどう利用されるかに関して、戦略アドバイザーとして行動する。
- 既存の支払いアカウントと調達プロセスは、クラウドと併用される。
- ファイナンスチームとテクノロジーチームは共同で、将来的な AWS のコストと使用量を予測し、組織の予算を調整、編成する。
- 両者の共通言語により組織間のコミュニケーションが向上し、ファイナンスの概念の共通理解が得られる。

コストと使用量のディスカッションについて、組織内で関わるべきその他のステークホルダーは以下のとおりです。

- **ビジネスユニットオーナー:** ビジネスユニットオーナーは、ユニット内と会社全体の両方に方向性を伝えられるように、クラウドのビジネスモデルを理解する必要があります。こうしたクラウド知識は成長とワークロード使用量を予測する際に、またリザーブドインスタンスや Savings Plans などの長期購入オプションを検討する際に重要な役割を果たします。
- **サードパーティー:** 組織がサードパーティー (コンサルタントやツールなど) を利用する場合、こうしたサードパーティーが財務目標に整合し、エンゲージメントモデルと投資収益率 (ROI) の両方を通じて整合性を実証できるようにします。通常、サードパーティーは自社管理のワークロードのレポーティングと分析を担当したり、自社設計のワークロードのコストを分析したりします。

クラウドの予算と予測

クラウドの予算と予測を確立する: お客様は効率性、速度、俊敏性を求めてクラウドを利用しますが、コストと使用量は大きく変動します。コストは、ワークロードの効率性の向上や、新規ワークロードや新機能のデプロイにより削減可能です。ワークロードをスケーリングすると、サービスを提供する顧客が増えますが、その分クラウドの使用量とコストが増加します。こうした変動を折り込めるように、組織の既存の予算編成プロセスを変える必要があります。

トレンドベースのアルゴリズム (コスト履歴を入力値として使用)、ビジネスドライバーベースのアルゴリズム (新製品の発売や営業地域の拡大など)、またはこの 2 つのアルゴリズムを組み合わせて、既存の予算編成と予測プロセスをより動的なものに調整します。

AWS Cost Explorer を使用すると、コスト履歴 (トレンドベース) に適用される機械学習アルゴリズムに基づき、日次 (最大 3 か月) または月次 (最大 12 か月) のクラウドコストを予測できます。

コストを意識したプロセス

組織のプロセスにコスト意識を取り入れる: 組織の既存および新規プロセスにコスト意識を取り入れる必要があります。可能であれば、既存プロセスを変更して再利用することを推奨します。俊敏性や速度への影響が最小限に抑えられます。次の推奨事項は、ワークロードにコスト意識を実装するのに役立ちます。

- 変更管理には、変更によるファイナンスへの影響を数値化するコスト測定を含めるようにします。これは、事前対策としてコストにまつわる懸念事項に対応し、コスト削減を強調できます。
- コスト最適化が運用能力の中心の要素になるようにします。たとえば、コストと使用量に関する異常値 (コスト超過) の根本原因を調査、特定するため、既存のインシデントマネジメントプロセスを活用できます。
- オートメーションやツールにより、コスト削減とビジネス価値の実現を後押しします。運用コストについて考えるときには、ROI の要素を話題に盛り込んで、時間や費用の投資の正当な根拠とします。
- 既存のトレーニングおよび開発プログラムを拡張して、組織全体にコスト意識の高いトレーニングが及ぶようにします。これには継続的なトレーニングと認定が含まれることをお勧めしますこれにより、コストと使用量を自己管理できる組織が育成されます。

コストと使用量の最適化に関するレポートと通知を行う: 組織内のコストと使用量の最適化については、定期レポートの必要があります。コスト最適化のための専用セッションの運用や、ワークロードの通常の運用レポートサイクルにコスト最適化を盛り込むことも意味があるでしょう。AWS Cost Explorer には、ダッシュボードとレポートが用意されています。AWS Budgets Reports を使用すると、設定された予算に対するコストと使用量の進行状況を追跡できます。

また、コストと使用状況レポート (CUR) データ とともに Amazon QuickSight を使用して、より詳細なデータによる高度にカスタマイズされたレポートも生成できます。

コストと使用量に関する通知を運用して、コストと使用量の変化をすばやく反映できるようにします。AWS Budgets では、ターゲットに対する通知を送信できます。通知は、増加と減少の両方について、またワークロードのコストと使用量の両方について設定することを推奨します。

コストと使用量を事前にモニタリングする: 例外や異常がある場合に限定せず、組織内のコストと使用量を事前にモニタリングすることを推奨します。オフィスや職場環境全体を高度に可視化したダッシュボードにより、主な担当者が必要とする情報にアクセスできるようになります。また組織がコスト最適化を重視していることを示すことができます。可視化したダッシュボードを使用することで、ビジネス成果の成功に向かって推進し、その成功を組織全体に拡張できます。

コストを意識した文化

コストを意識した文化を作る: 組織全体に変更やプログラムを運用し、コストを意識した文化を作ります。まず小さく始めて、続いて機能や組織でのクラウド利用の増加に合わせて、規模を拡大していき、さまざまなプログラムを運用していくことを推奨します。

コストを意識した文化があると、組織全体で有機的かつ分散的にベストプラクティスが行われ、コストの最適化とクラウドの財務管理を規模に合わせて実行できます。これにより、厳格なトップダウンの集中型アプローチと比較すると、最小限の労力で高レベルの機能を組織全体で創出できます。

この文化のわずかな変化で、現在や将来のワークロードの効率に大きな影響を与える可能性があります。次のような例をご紹介します。

- 組織全体のコストと使用量にゲーム的要素を取り入れる。これは、公開ダッシュボードや、チーム間の標準コストと標準使用量 (たとえば、ワークロードあたりのコスト、トランザクションあたりのコストなど) を比較するレポートによって実行できます。
- コスト効率性を認識する。自発的または独断で行なったコスト最適化の成果を公開または非公開で評価して、間違いから学び、今後繰り返さないようにします。
- 事前に編成された予算で実行するワークロードについて、トップダウンの組織要件を作成する。

新しいサービスリリースに関する最新情報を入手する: 新しい AWS のサービスや機能を運用して、ワークロードのコスト効率を改善できる場合があります。新しいサービスや機能のリリースについては、AWS ニュースブログ、AWS コスト管理ブログ、AWS 最新情報を定期的にご覧ください。

コスト最適化によるビジネス価値の数値化

コスト最適化によるビジネス価値の数値化: コスト最適化による節減額を報告することに加えて、実現された付加価値を数値化することを推奨します。コスト最適化のメリットは通常、ビジネス成果に対して削減されたコストという観点で数値化されます。たとえば、Savings Plans を購入すると、オンデマンドの Amazon Elastic Compute Cloud (Amazon EC2) のコスト削減を数値化できます。これにより、コストを削減し、ワークロードの成果を維持できます。アイドル状態の Amazon EC2 インスタンスが終了した場合や、アタッチされていない Amazon Elastic Block Store (Amazon EBS) ボリュームが削除された場合は、AWS の利用料削減を数値化できます。

コスト最適化でビジネス価値を数値化することで、組織に対するメリットの全体像を把握できます。コスト最適化は必要な投資であるため、ビジネス価値を数値化することで、各ステークホルダーに投資利益率を説明できます。ビジネス価値の数値化は、将来のコスト最適化投資に関してステークホルダーからより多くの賛同を得ることができます。また、組織のコスト最適化活動の結果を測定するためのフレームワークを提供します。

コスト最適化のメリットは、コスト削減やコスト回避にとどまりません。効率性向上とビジネス価値を測定するために、その他のデータを追加で取得することを検討してください。改善例は次のとおりです。

- **コスト最適化のベストプラクティスを実行する:** たとえば、リソースライフサイクル管理によって、インフラストラクチャコストと運用コストを削減し、実験のための時間や予定外の予算を作り出すことができます。これにより、組織の俊敏性が向上し、収益創出のための新しい商機の発掘につながります。
- **オートメーション運用:** Auto Scaling は、最小限の労力で需要に対 応するようにリソースの追加や削除を行うことができ、手作業によるキャパシティープランニング作業を排除することで、スタッフの生産性を向上させます。運用の回復力の詳細については、Well-Architected 信頼性の柱に関するホワイトペーパーを参照してください。
- **将来の AWS コストの予測:** 予測により、ファイナンスのステークホルダーは、組織内外のステークホルダーと見通しを立て、組織の財務予測可能性を改善できます。AWS Cost Explorer を使用して、コストと使用量を予測できます。

リソース

予算編成とクラウド使用量の予測に関する AWS のベストプラクティスの詳細については、以下のリソースを参照してください。

- 予算レポートで予算メトリクスをレポートする
- AWS Cost Explorer による予測

- AWS トレーニング
- AWS 認定
- AWS クラウドマネージメントツールパートナー

支出と使用量の認識

組織のコストおよびコスト要因を把握することは、コストと使用量を効果的に管理し、コスト削減の機会を特定するうえできわめて重要です。組織では一般に、複数のワークロードが複数のチームによってオペレーションされています。各チームはさまざまな組織単位に属する可能性があり、そのそれぞれに独自の収益の流れがあります。リソースコストの帰属先をワークロード、各組織、製品オーナーのいずれかに割り当てると、リソースを効率的に使用し、無駄を削減できます。コストと使用量を正確にモニタリングすることで、各組織単位や製品の収益性が把握できるようになり、より確かな情報に基づいて組織内のリソース配分を決定できます。使用量が変化するとコストも変動するため、組織内のあらゆるレベルの使用量を認識することは、変化を促進する鍵となります。

使用方法と支出を認識するために多面的なアプローチを取ることを検討してください。チームは、データを収集、分析し、それに続いて報告する必要があります。考慮すべき主な要因は以下のとおりです。

- ガバナンス
- コストと使用量のモニタリング
- 廃棄

ガバナンス

クラウドのコストを管理するには、以下のガバナンス領域から使用量を管理する必要があります。

組織のポリシーを作成する: ガバナンスを実行するための最初のステップは、組織の要件を使い、クラウド使用に関するポリシーを作成することです。ポリシーでは、組織がクラウドをどのように使用するかや、リソースをどのように管理するかを定義します。ポリシーではコストや使用量に関係するリソースとワークロードのあらゆる局面、つまりリソースのライフタイム全体にわたる作成、変更、削除をカバーする必要があります。

ポリシーは理解しやすく、組織全体で効果的に利用できるように、シンプルにする必要があります。使用許可する地理的リージョンや、リソースを実行する時間帯など、幅広い高レベルのポリシーから始めます。続いてポリシーを徐々に絞り込み、さまざまな組織単位やワークロードに対応させます。一般的なポリシーの例としては、どのサービスと機能が利用できるか (たとえば、テスト環境や開発環境ではパフォーマンスが低下するストレージ)、どのタイプのリソースが各グループで使用できるか (たとえば、開発用アカウントのリソースの最大サイズはミディアム) などがあります。

目標とターゲットを設定する: 組織のコスト、目標使用量、ターゲットを設定します。目標は、期待される成果に関するガイダンスと指示を組織にもたらします。ターゲットは、具体的かつ測定可能な達成すべき結果をもたらします。目標の一例: プラットフォームの使用量を大幅に増加させ、コストは微増 (非線形) にとどまるようにする。ターゲットの一例:プラットフォームの使用率を 20% 増加させ、コスト増は 5% 未満。ワークロードを 6 か月ごとに効率化する必要があるというケースも、目標としてはよくあります。この種のターゲットとして、ワークロードの結果あたりのコストを 6 か月ごとに 5% 削減する必要があるというケースも考えられます。

ワークロードの効率を高めることは、クラウドワークロードの目標としては一般的です。これは、ワークロードのビジネス成果あたりのコストを継続的に削減するというものです。この目標と合わせて、6～12 か月ごとに効率を 5% 向上させるなどのターゲットをすべてのワークロードに設定することを推奨します。これは、クラウド内でコスト最適化の機能を構築し、新しいサービスやサービス機能のリリースを行うことで達成できます。

アカウント構造: AWS は 1 つの親アカウントと複数の子アカウントからなる構造を持ちます。このアカウント構造は、一般にマスター (親、旧称は支払者) アカウント、メンバー (子、旧称はリンク) アカウントと呼ばれます。ベストプラクティスは、組織の規模や使用量にかかわらず、1 つのメンバーアカウントを持つマスターを少なくとも 1 つ常に持つことです。すべてのワークロードリソースが存在するのは、メンバーアカウント内のみとする必要があります。

AWS アカウントの最適数は状況に応じて異なります。現在と将来の運用モデルとコストモデルを見積もり、AWS アカウントの構造が組織の目標を反映するようにします。ビジネス上の理由から複数の AWS アカウントを作成する企業もあります。次に例を示します。

- 各組織単位、コストセンター、特定のワークロード間で、管理、会計、請求の職務機能を切り離す必要がある場合。
- AWS のサービスの制限が特定のワークロードのみに設定される場合。
- ワークロードとリソース間の隔離、分離には要件があります。

AWS Organizations 内では、一括請求 (コンソリデーティッドビリング) により、1 つ以上のメンバーアカウントとマスターアカウント間で組織が構成されます。メンバーアカウントを使用すると、コストと使用量をグループ別に区別できます。一般的には、各組織単位 (財務、マーケティング、営業など)、各環境ライフサイクル (開発、テスト、本番など)、各ワークロード (ワークロード a、b、c) にメンバーアカウントをいったん分離したうえで、一括請求 (コンソリデーティッドビリング) を使用してこれらのアカウントを集約します。

一括請求 (コンソリデーティッドビリング) 機能を使用すると、複数のメンバー AWS アカウントの支払いを単一マスターアカウントに集約し、各リンクアカウントのアクティビティの可視性も維持できます。コストと使用量がマスターアカウントに集計されると、サービスのボリューム割引とコミットメント割引 (Savings Plans とリザーブドインスタンス) を最大限に活用し、割引額を最大化できます。

AWS Control Tower では、複数の AWS アカウントのセットアップと構成をすばやく行い、ガバナンスが組織要件に適合していることを確認できます。

組織のグループとロール: ポリシーを作成すると、組織内のユーザーの論理グループとロールを作成できます。これにより、アクセス許可の割り当てと使用量の制御が可能になります。高レベルの人材グループから始めます。通常これは、組織単位と役職 (IT 部門のシステム管理者、会計監査担当者など) と合致します。グループとは、類似したタスクに従事し、類似したアクセスを必要とするユーザーの集団を指します。ロールとは、グループとして義務付けられた仕事の定義を指します。たとえば、IT のシステム管理者はすべてのリソースを作成するためのアクセスが必要ですが、分析チームのメンバーは分析リソースを作成するアクセスのみで十分です。

コントロールと通知: コスト管理を導入する際の一般的な最初のステップは、ポリシー外のコストまたは使用量イベントが発生した場合に通知するように設定することです。これにより、ワークロードや新しいアクティビティを制限したり悪影響を与えたりすることなく、迅速に行動し、是正措置の必要性の有無を確認できます。ワークロードと環境の制限を理解したら、ガバナンスを適用できます。AWS では、通知を実行する AWS Budgets により、AWS のコスト、使用量、コミットメント割引 (Savings Plans とリザーブドインスタンス) の月次予算を編成できます。予算は、集計コストのレベル (たとえば、全コスト)、またはリンクアカウント、サービス、タグ、アベイラビリティーゾーンなどの特定のディメンションのみを含む詳細レベルで作成できます。現在コストまたは予測コストや使用量が設定したしきい値 (%) を超えるとトリガーされる E メール通知を予算にアタッチもできます。

コントロールと実施: 第 2 ステップとして、AWS Identity and Access Management (IAM) および AWS Organizations のサービスコントロールポリシー (SCP) により、AWS でガバナンスポリシーを実行できます。IAM により、AWS のサービスとリソースへのアクセスを安全に管理できます。IAM を使用すると、AWS リソースを作成、管理できるユーザー、作成できるリソースタイプ、リソース作成できる場所を制御できます。これにより、不要なリソースの作成が最小限に抑えられます。以前に作成したロールとグループを使用して IAM ポリシーを割り当て、正しい使用量を適用します。SCP は組織内のすべてのアカウントで利用可能なアクセス許可の上限を一元的に制御し、アカウントがアクセスコントロールのガイドライン内に収まるようにします。SCP はすべての機能が有効になっている組織でのみ使用可能で、デフォルトでメンバーアカウントのアクションの可否を SCP を設定できます。アクセス管理の導入の詳細については、Well-Architected セキュリティの柱に関するホワイトペーパーを参照してください。

コントロールとサービスクォータ: サービスクォータを管理することで、ガバナンスを導入することもできます。サービスクォータを最小オーバーヘッドに設定し、正確に維持するよう徹底することで、組織に不要なリソースの作成を最小限に抑えることができます。これを実現するには、要件がどれだけ早く変化するかを理解し、進行中のプロジェクト（リソースの作成と削除の両方）を理解し、クォータ変更をどのようにすばやく実装できるかを考慮する必要があります。サービスクォータを使用すると、必要に応じてクォータを増やせます。

AWS コスト管理サービスは、AWS Identity and Access Management (IAM) サービスと統合されます。IAM サービスをコスト管理サービスと併用して、財務データや請求コンソールに含まれる AWS ツールへのアクセスを制御します。

ワークロードのライフサイクルを追跡する: ワークロードのライフサイクル全体を確実に追跡します。これにより、ワークロードやワークロードコンポーネントが不要になった場合、削除や変更が可能になります。これは、新しいサービスや機能をリリースする際に特に便利です。既存のワークロードとコンポーネントは使用されているように見えても、顧客を新しいサービスにリダイレクトするために使用を停止する必要があります。。ワークロードの以前のステージに注目してください。ワークロードが本番稼働状態になったら、以前の環境は廃棄することと、再び必要になるまでキャパシティーを大幅に削減することも可能です。

AWS には、エンティティのライフサイクル追跡に使用できる管理およびガバナンスサービスが多数用意されています。AWS Config または AWS Systems Manager を使用すると、AWS リソースと設定の詳細なインベントリが入手可能です。プロジェクトやアセットを管理する既存のシステムを統合して、組織内のアクティブなプロジェクトや製品を追跡することを推奨します。現在のシステムを AWS が提供する豊富なイベントやメトリクスと組み合わせることにより、重要なライフサイクルイベントのビューを作成し、前もってリソースを管理し、不要なコストを削減できます。

エンティティのライフサイクル追跡運用の詳細については、Well-Architected 運用上の優秀性の柱ホワイトペーパーを参照してください。

コストと使用量のモニタリング

ワークロードに詳細な可視化を導入し、チームがコストと使用量に対しアクションを実行できるようにします。コストの最適化は、コストと使用状況の内訳、将来の支出、使用状況、機能をモデル化して予測する機能、コストと使用量を組織の目標に合わせて調整するためのメカニズムの実装をきめ細かく理解することから始まります。コストと使用量をモニタリングするために必要な領域は次のとおりです。

詳細なデータソースの設定: Cost Explorer で時間単位の詳細設定を有効にして、コストと使用状況レポート (CUR)を作成します。これらのデータソースは、組織全体のコストと使用量の最も正確なビューを提供します。CUR では、課金されるすべての AWS のサービスについて、日単位または時間単位の使用量、料金、コスト、使用量といった属性が提供されます。CUR には、タグ付け、場所、リソース属性、アカウント ID など想定可能なすべてのディメンションがあります。

以下のカスタマイズ項目で CUR を設定します。

- リソース ID を含める
- CUR を自動更新する
- 時間単位の詳細
- バージョニング: 既存のレポートを上書きする
- データ統合: Athena (Parquet 形式、圧縮)

分析用データの準備には AWS Glue を、データ分析の実行には Amazon Athena を、データのクエリには SQL を使用します。また、Amazon QuickSight を使用して、カスタムの可視化や高度な可視化を行い、組織全体に配布することもできます。

コスト属性カテゴリを特定する: 財務チームや関連ステークホルダーと協力して組織内でコストを配分する方法の要件を理解します。ワークロードのコストは、開発、テスト、本稼働、廃止などライフサイクル全体にわたって配分する必要があります。学習、スタッフ育成、アイデア創出に要したコストが、どのように組織に帰属するかを理解します。この目的で使用される金額を一般的な IT コスト予算ではなく、トレーニング予算や開発の予算に正しく割り当てるのに便利です。

ワークロードメトリクスを確立する: ワークロードのアウトプットがビジネスの成功に対してどのように測定されるかを理解します。通常、各ワークロードには、パフォーマンスを示す主な成果の小さな組み合わせがあります。多数のコンポーネントを含む高度なワークロードがある場合は、リストに優先順位を付けるか、各コンポーネントのメトリクスを定義して追跡できます。チームと協力して、どのメトリクスを使用するか理解します。この単位は、ワークロードの効率または各ビジネス成果のコストを把握するために使用されます。

コストと使用量に組織としての意味を割り当てる: タグ付けを AWS に運用し、リソースに組織の情報を追加してからコストと使用量の情報に追加します。タグはキーと値のペアです。キーは定義されており、組織全体で一意である必要があります。値はリソースのグループに対して一意です。キーと値のペアの一例としては、キーが Environment で、値は Production となります。本稼働環境のすべてのリソースには、キーと値のペアがあります。タグ付けにより、関連性の高い組織情報を使用して、コストを分類、追跡できます。組織のカテゴリ (コストセンター、アプリケーション名、プロジェクト、オーナーなど) を表すタグを適用し、ワークロードやワークロードの特性 (テストや本番など) を識別して、組織全体のコストと使用量の帰属先を付与できます。

AWS リソース (EC2 インスタンスや Amazon S3 バケットなど) にタグを付け、そのタグをアクティベートすると、AWS はこの情報をコストと使用量レポートに追加します。タグ付けされたリソースとタグ付けされていないリソースに対してレポートを実行し、分析を実行することで、内部のコスト管理ポリシーへの準拠を強化し、正確な帰属を保証できます。

組織のアカウント全体に AWS タグ付け標準を作成、導入することで、AWS 環境を一元的かつ整合的に管理できます。AWS Organizations のタグポリシーを使用して、AWS Organizations アカウント内の AWS リソースに対するタグ使用ルールを定義します。タグポリシーを使用すると、AWS リソースにタグを付ける標準アプローチを簡単に導入できます。

AWS Tag Editor では、複数のリソースのタグを追加、削除、管理できます。

AWS Cost Categories を使用すると、リソースにタグを付けることなく組織としての意味をコストに割り当てることができます。コストと使用量に関する情報を、一意の内部組織構造にマッピングできます。アカウントやタグなどの請求ディメンションを使用して、コストをマッピングおよび分類するカテゴリルールを定義します。これにより、タグ付けに加えて、別のレベルの管理機能が提供されます。また、特定のアカウントとタグを複数のプロジェクトにマッピングすることもできます。

請求とコストの最適化ツールを設定する: 使用量を変更してコストを調整するには、組織内の各ユーザーがそれぞれのコストと使用量の情報にアクセスできる必要があります。クラウドを使用する場合、すべてのワークロードとチームに次のツールを設定することを推奨します。

- **レポート:** すべてのコストと使用量の情報を要約する。
- **通知:** コストまたは使用量が設定された制限を超えた場合に通知する。
- **現在の状態:** 現在のコストと使用量を示すダッシュボードを設定する。ダッシュボードはオペレーションダッシュボードと同様に、作業環境内の目に付きやすい場所で使用できるようにする必要があります。
- **傾向:** 要求した期間におけるコストと使用量の変動を、必要な詳細度で示す。
- **予測:** 将来の推定コストを示す。
- **追跡:** 設定された目標またはターゲットに対する現在のコストと使用量を表示する。
- **分析:**チームメンバーが、すべての可能なディメンションを使用して、時間単位でカスタムおよび詳細な分析を実行する機能を提供します。

この機能は、AWS Cost Explorer、AWS Budgets、Amazon Athena などの AWS ネイティブツールを QuickSight と併用することで利用できます。サードパーティー製のツールを使用することもできますが、このツールのコストが皆さんの組織にもたらす価値を事前に必ず確かめてください。

ワークロードメトリクスに基づいてコストを配分する: コスト最適化によって、最低価格でビジネス成果が提供されます。これは、ワークロードメトリクス (ワークロードの効率で測定) ごとにワークロードのコストを配分することによってのみ達成できます。定義されたワークロードメトリクスを、ログファイルまたは他のアプリケーションのモニタリングを使用してモニタリングします。このデータをワークロードのコストと組み合わせます。ワークロードのコストは、特定のタグ値またはアカウント ID でコストを確認することで取得できます。この分析は時間単位で実行することをお勧めします。リクエストレートが変化する静的なコストコンポーネント（24 時間年中無休で実行されるバックエンドデータベースなど）がある場合、通常、効率性は変化します（たとえば、使用量のピークは午前 9 時から午後 5 時で、夜間のリクエストはほとんどありません）。静的コストと変動コストの関係を理解しておくと、最適化のアクティビティに集中する一助となります。

リソースの削除

プロジェクト、従業員、テクノロジーリソースのリストを管理するようになると、使用されなくなったリソースやオーナーが不在となったプロジェクトを特定できるようになります。

ライフタイムにわたってリソースを追跡する: 不要になったワークロードリソースを削除します。一般的な例としては、テスト用途のリソースがあります。テストが完了したら、リソースは削除できます。タグでリソースを追跡する (およびそのタグのレポートを作成する) ことは、廃棄するアセットの特定に役立ちます。リソース追跡には、タグの使用が効果的な方法です。リソースにその機能か、または廃止可能になる既知の日付をラベリングできます。そうすると、これらのタグでレポートを作成できます。機能タグを付ける場合の一例として、「featureX testing」という値であれば、ワークロードのライフサイクルの観点からリソースの目的を識別できます。

削除プロセスを運用する: 組織全体の標準化プロセスで、不使用リソースを特定して廃棄します。検索の実行頻度および組織のすべての要件を確実に満たすために、このプロセスではリソースを削除するプロセスを定義する必要があります。

リソースを削除する: 使用していないリソースを検索する場合は節減額の程度によって検索頻度と投入する労力を決定する必要があるため、コスト発生額の小さいアカウントの分析は、コスト発生額が高額のアカウントよりも頻度を下げるべきです。イベントの検索および廃止は、製品が寿命を迎えた場合や交換する場合など、ワークロードの状態の変化によってトリガーされます。イベントの検索および廃止は、市況の変化や製品終了などの外部イベントによってトリガーされる場合もあります。

リソースを自動削除する: オートメーションを利用して、削除プロセスの関連コストを削減またはゼロにします。自動削除するようにワークロードを設計すると、そのライフタイム全体にわたるワークロードコストを削減できます。削除プロセスは、AWS Auto Scaling を使用して実行できます。また、API または SDK を使用してカスタムコードを運用して、ワークロードリソースの自動削除もできます。

リソース

費用の認識に関する AWS のベストプラクティスの詳細については、以下のリソースを参照してください。

- AWS Tagging Strategies
- Activating User-Defined Cost Allocation Tags
- AWS Billing and Cost Management
- Cost Management Blog
- Multiple Account Billing Strategy
- AWS SDK とツール
- Tagging best practices
- Well-Architected Labs - Cost Fundamentals
- Well-Architected Labs – Expenditure Awareness

費用対効果の高いリソース

ワークロードに適したサービス、リソース、設定の使用は、コスト節減の鍵となります。費用対効果の高いリソースを作成する場合、以下の点を考慮してください。

- サービスを選択する際にはコストを評価する
- 正しいリソースタイプ、リソースサイズ、リソース数を選択する
- 最適な料金モデルを選択する
- データ転送を計画する

AWS ソリューションアーキテクト、AWS ソリューション、AWS リファレンスアーキテクチャ、APN パートナーを利用すると、学習内容に基づいてアーキテクチャを選択できます。

サービスを選択する際にはコストを評価する

組織の要件を特定する: ワークロードのサービスを選択する場合は、組織の優先順位を理解することが重要です。パフォーマンスや信頼性などの Well-Architected の柱と、コストとのバランスが取れているようにします。十分にコスト最適化されたワークロードとは組織の要件に最も適合するソリューションであって、必ずしも最低コストのソリューションとは限りません。組織内のすべてのチームと会合し、製品、ビジネス、技術、財務などの情報を収集します。

すべてのワークロードコンポーネントを分析する: ワークロードのすべてのコンポーネントについて徹底した分析を実行します。分析コストと、そのライフサイクルで可能と考えられるワークロードの節減額のバランスを取るようにします。コンポーネントの現在の影響、および将来的に与えると考えられる影響を洗い出す必要があります。たとえば、提案されたリソースのコストが月額 10 ドルで、予測負荷が月額 15 ドルを超えない場合に、コストを 50% (月額 5 ドル) 削減するために 1 日分の労力を費やすようでは、システムの寿命全体にわたって得られると考えられる利益を超えることになるかもしれません。データに基づくより高速でより効率的な予測を使用すると、このコンポーネントの全体的な成果を最善のものにできます。

ワークロードは時間の経過とともに変化する可能性があり、ワークロードのアーキテクチャや使用方法が変化すると、適切なサービスの組み合わせが最適にならない場合があります。サービスの選択に関する分析には、現在および将来のワークロードの状態と使用量レベルが組み込まれる必要があります。将来のワークロードの状態や使用量に合わせてサービスを運用すると、今後の変更に必要な労力を軽減または削除できることになり、全体的なコストを削減できます。

AWS Cost Explorer と CUR では、PoC (概念実証) コスト、または実行環境コストを分析できます。また、AWS 簡易見積りツールや AWS 料金計算ツールを使用して、ワークロードコストの見積もりができます。

マネージドサービス: マネージドサービスにより、サービス維持に伴う運用上および管理上の負担が軽減されるため、イノベーションに集中できます。さらに、マネージドサービスはクラウドのスケールで運用されるため、トランザクション単位またはサービス単位でコストを削減できます。

チームが技術的な負債の返済、イノベーション、付加価値機能に集中できるように、時間短縮を検討します。たとえば、オンプレミス環境からクラウド環境に「リフトアンドシフト」式に可能な限り速やかに移行して、最適化は後回しにならざるを得ない場合があります。ライセンスコストを排除または削減するマネージドサービスを利用して、コスト削減が可能なところを模索することには時間をかけるだけの価値があります。

マネージドサービスには通常、十分なキャパシティーを確保できるように設定できる属性があります。この属性を設定およびモニタリングして、余剰キャパシティーを最小限に抑え、パフォーマンスを最大化する必要があります。AWS マネジメントコンソールまたは AWS API および SDK を使用して AWS マネージドサービスの属性を変更し、需要の変化に合わせてリソースのニーズを調整できます。たとえば、Amazon EMR クラスター (または Amazon Redshift クラスター) のノード数を増減して、スケールアウトまたはスケールインできます。

AWS リソースの複数のインスタンスを圧縮して、高密度使用もできます。たとえば、1 個の Amazon Relational Database Service (Amazon RDS) DB インスタンスに複数の小規模データベースをプロビジョニングします。使用量が増えたら、スナップショットや復元プロセスを使用して、そのデータベースの 1 つを専用 RDS DB インスタンスに移行できます。

マネージドサービスでワークロードをプロビジョニングする際は、サービスキャパシティーの調整要件を理解する必要があります。主な要件としては、時間、労力、通常のワークロードオペレーションへの影響などが一般には考えられます。プロビジョニングされたリソースでは変更が発生するまでの時間が許容され、このために必要なオーバーヘッドをプロビジョニングする必要があります。システムに統合する API と SDK や Amazon CloudWatch などのモニタリングツールを使用することで、サービス変更に求められる継続的な努力は実質ゼロにできます。

Amazon Relational Database Service (RDS)、Amazon Redshift、Amazon ElastiCache が提供するのは、マネージドデータベースサービスです。Amazon Athena、Amazon Elastic Map Reduce(EMR)、Amazon Elasticsearch が提供するのは、マネージド型分析サービスです。

AWS マネージドサービス (AMS) は、エンタープライズカスタマーやパートナーに代わって AWS インフラストラクチャを運用するサービスです。コンプライアンスに準拠したセキュアな環境で、ワークロードをデプロイできます。AMS では、エンタープライズクラウド運用モデルとオートメーションを使用して、組織の要件を満たし、クラウド移行を高速化し、オンゴーイングの管理コストを削減できます。

サーバーレスまたはアプリケーションレベルのサービス: AWS Lambda、Amazon Simple Queue Service (Amazon SQS)、Amazon Simple Notification Service (Amazon SNS)、Amazon Simple Email Service(Amazon SES) などのサーバーレスサービスやアプリケーションレベルのサービスを使用できます。これらのサービスではリソースを管理する必要がなく、コード実行、キューサービス、メッセージ配信の機能を利用できます。もう 1 つの利点は、使用量に応じてパフォーマンスとコストをスケールするため、コスト配分とコストの帰属が効率的になることです。

サーバーレスの詳細については、Well-Architected サーバーレスアプリケーションレンズホワイトペーパーを参照してください。

使用量が時間の経過とともに変化をするワークロードの分析: AWS で新しいサービスや機能がリリースされるたびに、ワークロードに最適なサービスが変わっていく可能性があります。求められる労力は、潜在的な利点が反映されたものである必要があります。ワークロードレビューの頻度は、組織の要件によって異なります。ワークロードにかなりのコストがかかっている場合、新しいサービスの運用が早いほどコスト削減が最大になるため、レビュー頻度が高い方が有利です。レビューのトリガーとしては、使用パターンの変化も挙げられます。使用量が大幅に変化した場合は、別のサービスを使った方がよい場合もあります。たとえば、データ転送速度が高い場合、Direct Connect サービスのほうが VPN よりも安価で、必要な接続性能を提供できます。サービス変更時に起こりうる影響を予測すると、使用量レベルのトリガーをモニタリングできるため、費用対効果が最も高いサービスを速やかに運用できます。

ライセンスコスト: ソフトウェアライセンスのコストは、オープンソースソフトウェアを使用することで削減できます。オープンソースソフトウェアへの変更は、ワークロードサイズが拡大するにつれ、ワークロードコストに大きな影響を与える可能性があります。ライセンスを取得したソフトウェアの利点を総コストと比較して測定し、ワークロードを最適化します。ライセンス変更とその変更がワークロードコストに与える影響をモデリングします。あるベンダーがデータベースライセンスのコストを変更したなら、それがワークロードの全体的な効率にどのような影響を与えるかを調査します。ベンダーの過去の価格アナウンスを検討して、ベンダー製品全体のライセンス変更の傾向を検討してください。ライセンスコストは、ハードウェアごとにスケールするライセンス (CPU バウンドライセンス) など、スループットや使用量とは関係なくスケールされる場合があります。こうしたライセンスは、それに伴う成果が見られないままコストが急増する可能性があるため、避けてください。

AWS License Manager を使用すると、ワークロードのソフトウェアライセンスを管理できます。ライセンス規則を設定し、必要な条件を適用することで、ライセンス違反を防ぎ、ライセンス超過で発生するコストを削減できます。

正しいリソースタイプ、リソースサイズ、リソース数を選択する

最適なリソースタイプ、リソースサイズ、リソース数を選択することで、最低限のリソースコストで技術要件を満たすことができます。適切なサイズ変更アクティビティは、ワークロードのすべてのリソース、各リソースのすべての属性、および適切なサイジング操作に関連する作業が考慮されます。適切なサイズ変更は、使用パターンや、AWS の値下げや 新しい AWS リソースタイプなどの外部要因の変化により、反復プロセスになることがあります。適切なサイズ変更作業のコストがワークロードの寿命全体にわたって削減されると考えられる額を上回る場合は、正しいサイズ設定は 1 回に限って設定もできます。

AWS には以下のようなさまざまなアプローチがあります。

- コストモデリングを実行する
- メトリクスまたはデータに基づいたサイズを選択する
- サイズを自動選択する (メトリクスに基づく)

コストモデリング: ワークロードと各コンポーネントのコストモデリングを実行して、特定のパフォーマンスレベルに応じて、リソース間のバランスを把握し、ワークロードにある各リソースの適切なサイズを見つけます。さまざまな予測負荷におけるワークロードのベンチマークアクティビティを実行し、コストを比較します。モデリングの際には、費やされた時間がコンポーネントのコストまたは予想される削減額に比例しているといった潜在的な利点を織り込む必要があります。このプロセスのベストプラクティスについては、AWS Well-Architected フレームワークのパフォーマンス効率の柱に関するホワイトペーパーのレビューセクションを参照してください。

AWS Compute Optimizer は、ワークロードの実行におけるコストモデリングを支援します。使用履歴に基づき、コンピューティングリソースの正しいサイズ設定に関するレコメンデーションを提供します。これがコンピューティングリソースにとって理想的なデータソースである理由は、リスクレベルに応じて複数のレコメンデーションを作成できる機械学習が使われている無料サービスだからです。Amazon CloudWatch と CloudWatch Logs をデータソースとしてカスタムログと併用して、他のサービスやワークロードコンポーネントの適切なサイズ設定のオペレーションもできます。

以下は、コストモデリングのデータおよびメトリクスのレコメンデーションです。

- モニタリングにはエンドユーザーエクスペリエンスを正確に反映させること。対象期間を正確に選択して、平均ではなく最大パーセンタイルまたは 99 パーセンタイルのいずれかを選択します。
- いかなるワークロードサイクルであってもカバーするために必要な分析期間を正確に選択すること。たとえば、2 週間の分析を実行した場合、使用率の高い月次サイクルを見落として、プロビジョニング不足につながる可能性があります。

メトリクスまたはデータに基づく選択: ワークロードとリソースの特性 (たとえば、コンピューティング、メモリ、スループット、書き込み頻度) に基づいて、リソースサイズやリソースタイプを選択します。この選択は通常、コストモデリング、以前のバージョンのワークロード (オンプレミスバージョンなど)、ドキュメント、ワークロードに関するその他の情報ソース (ホワイトペーパー、公開ソリューション) を用いて行います。

メトリクスに基づく自動選択: ワークロード内にフィードバックループを作成します。このループでは、実行中のワークロードのアクティブなメトリクスを使用して、そのワークロードを変更します。AWS Auto Scaling などのマネージドサービスを使用して、正しいサイズ設定を実行できるように設定できます。AWS には、API、SDK のほかにも、最低限の労力でリソースを変更できる機能も用意されています。EC2 インスタンスの停止と起動のワークロードをプログラムして、インスタンスサイズやインスタンスタイプを変更できます。これにより、正しいサイズ設定による利点が得られるだけでなく、変更に必要なほぼすべての運用コストを削減することもできます。

一部の AWS サービスには、S3 Intelligent-Tiering など、タイプやサイズの自動選択が組み込まれています。S3 Intelligent-Tiering では、使用パターンに基づいて、高頻度アクセスと低頻度アクセスの 2 つのアクセスティア間でデータが自動的に移動します。

最適な料金モデルを選択する

ワークロードコストモデリングの実行: ワークロードコンポーネントの要件を考慮し、考えられる料金モデルを理解します。コンポーネントの可用性要件を定義します。ワークロードで関数を実行する複数の独立したリソースの有無、ワークロードの継続的に必要となる要件を確認します。デフォルトのオンデマンド料金モデルと他の適用可能なモデルを使用して、リソースのコストを比較します。リソースまたはワークロードコンポーネントで変更可能なものはすべて考慮します。

アカウントレベルの分析を定期的に実行: コストモデリングを定期的に実行すると、複数のワークロードにまたがって最適化を運用できます。たとえば、複数のワークロードでオンデマンドを使用している場合、集計レベルでは変更リスクが低くなり、コミットメントベースの割引を運用すると全体的なコストが低くなります。2 週間から 1 か月の定期的なサイクルで分析を実行することを推奨します。これにより、調整のための小口購入が可能になり、ワークロードやコンポーネントの変更に合わせて料金モデルの調整を続けることができます。

AWS Cost Explorer のレコメンデーションツールを使用して、コミットメント割引を適用する機会を見つけます。

スポットのワークロードを実行する機会を見つけるには、使用量全体の 1 時間ごとのビューを使用して、使用量や伸縮性の変化の定期的な期間を探します。

料金モデル: AWS には複数の料金モデルがあり、組織のニーズに合った最も費用対効果の高い方法でリソース料金を支払うことができます。以下のセクションでは、各購入モデルについて説明します。

- オンデマンド
- スポット
- コミットメント割引 - Savings Plans

- コミットメント割引 - リザーブドインスタンス/キャパシティー
- 地理的選択
- サードパーティーの契約と料金

オンデマンド: これはデフォルトの従量制料金モデルです。EC2 インスタンスなどのリソースや DynamoDB オンデマンドなどのサービスを利用する際は定額料金を支払うだけで、長期のコミットメントはありません。リソースやサービスのキャパシティーは、お使いのアプリケーションの需要に合わせて増減できます。オンデマンドには時間単位の料金がありますが、サービスによっては 1 秒単位での利用も可能です (たとえば、AWS Lambda、Linux EC2 インスタンス)。オンデマンドが適するケースとして、定期的な急増や予想外な急増が避けられない短期的なワークロード (たとえば、4 か月間のプロジェクト) を伴うアプリケーションが挙げられます。ほかにもオンデマンドが適しているのは、中断のないランタイムを必要としていて、本番稼働前の環境のように実行時間がコミットメント割引 (Savings Plans またはリザーブドインスタンス) を受けられるほど長くないワークロードです。

スポット: スポットインスタンスは、長期のコミットメントを必要とせず、オンデマンド料金の最大 90% 割引で利用できる予備的な EC2 コンピューティングキャパシティーです。スポットインスタンスを使用すると、アプリケーションの実行コストの大幅な削減や、同じ予算でアプリケーションのコンピューティングキャパシティーをスケーリングできます。スポットインスタンスはオンデマンドとは異なり、EC2 のキャパシティーを元に戻す必要がある場合や、料金が設定価格を上回っている場合に中断されることがあります。その際は中断の 2 分前に警告通知が送信されます。スポットインスタンスの平均的な中断頻度は時間全体の 5% 未満です。

キューまたはバッファがある場合や、独立した動作でリクエスト処理をする複数のリソース (Hadoop データ処理など) がある場合には、スポットインスタンスは最適です。バッチ処理、ビッグデータと分析、コンテナ化された環境、ハイパフォーマンスコンピューティング (HPC) などのワークロードは一般に、耐障害性があり、ステートレスで柔軟性があります。テスト環境や開発環境などの重要性の低いワークロードも、スポットインスタンスの利用に適しています。

スポットインスタンスは、EC2 Auto Scaling グループ (ASG)、Elastic MapReduce (EMR)、Elastic Container Service (ECS)、AWS Batch などの複数の AWS のサービスにも統合されています。

スポットインスタンスを回収する必要があるときには、スポットインスタンスの中断通知を経て、CloudWatch Events とインスタンスメタデータに 2 分前警告が EC2 から送信されます。この 2 分間を使って、アプリケーションの状態を保存したり、実行中のコンテナを空にしたり、最終ログファイルをアップロードしたり、ロードバランサーからインスタンスを削除したりできます。2 分が経過すると、スポットインスタンスの休止、停止、終了の選択肢があります。

ワークロードにスポットインスタンスを採用するときは、以下のベストプラクティスに従ってください。

- **上限価格をオンデマンドレートとして設定:** これにより、支払いは現在のスポットレート (最安料金) となり、金額はオンデマンド料金を上回ることはありません。現在のレートと過去のレートは、コンソールと API から入手できます。
- **できるだけ多くのインスタンスタイプに柔軟に対応:** インスタンスタイプのファミリーとサイズに柔軟に対応し、目標とするキャパシティー要件を満たす可能性を高め、コストを可能な限り低減し、中断の影響を最小限に抑えます。
- **ワークロードの実行場所を柔軟に調整:** アベイラビリティーゾーンによって利用可能なキャパシティーが異なる場合があります。複数の予備キャパシティープールを使用してターゲットキャパシティーを実現する可能性が向上し、コストを可能な限り低く抑えます。
- **継続性を考慮した設計:** ステートレスかつ耐障害性のあるワークロードを設計し、EC2 キャパシティーの一部が中断されても、ワークロードの可用性やパフォーマンスに影響がないようにします。
- スポットインスタンスをオンデマンドおよび Savings Plans/リザーブドインスタンスと組み合わせて使用し、そのパフォーマンスによってワークロードのコストを最適化することを推奨します。

コミットメント割引 – Savings Plans: AWS にはコストを削減するさまざまな方法があります。一定量のリソースの使用を予約またはコミットすると、リソースが割引されます。Savings Plan では、1 年または 3 年の期間で時間あたりの利用料金をコミットすることにより、リソース全体で割引価格が適用されます。Savings Plans では、EC2、Fargate、Lambda などの AWS コンピューティングサービスに割引が適用されます。コミットすると、そのコミットメント額に対して毎時間支払いが発生します。また、オンデマンド使用料から所定の割引率が差し引かれます。たとえば、1 時間あたり 50 USD をコミットしていて、オンデマンド使用料が 1 時間あたり 150 USD であるとします。Savings Plans 料金により、この使用料には 50% の割引が適用されます。つまり、50 USD のコミットメントにより 100 USD のオンデマンド使用料がカバーされます。支払うのは、50 USD (コミットメント) と 残りのオンデマンド使用料の 50 USD となります。

Compute Savings Plans は最も柔軟なプランで、利用料金を最大 66% 割引できます。アベイラビリティーゾーン、インスタンスサイズ、インスタンスファミリー、オペレーティングシステム、テナンシー、リージョン、コンピューティングサービスの全体にわたって、割引が自動適用されます。

Instance Savings Plans の柔軟性は Compute Savings Plans よりも低くなりますが、割引率は高くなります (最大 72%)。アベイラビリティーゾーン、インスタンスサイズ、インスタンスファミリー、オペレーティングシステム、テナンシーの全体にわたって、割引が自動適用されます。

お支払い方法は 3 つあります。

- **前払いなし:** 前払いがなく、時間料金の割引が適用された当月の合計利用時間分を毎月支払います。
- **一部前払い:** 前払いなしよりも割引率が高くなります。利用料金の一部を前払いし、時間料金の割引 (率は低くなる) が適用された合計利用時間分を毎月支払います。
- **完全前払い:**期間全体の使用量を前払いします。コミットメントの対象となる残存期間の残りの期間に他のコストは発生しません。

この 3 つの購入オプションを任意に組み合わせてワークロードに適用できます。

Savings Plans ではまず、割引率の高いものから低い順に、購入に使用されたアカウントの使用量に適用されます。次に、割引率の高いものから低い順に、すべての連結アカウントの使用量に適用されます。

Savings Plans はすべて、マスターアカウントのように使用量もリソースもないアカウントで購入することを推奨します。Savings Plan により、すべての使用量において最も高い割引率が適用され、割引額が最大になります。

ワークロードと使用量は通常、経時変化します。長期間にわたって、Savings Plans のコミットメントを少額ずつ継続的に購入することを推奨します。これにより、割引を最大化するための高いカバレッジレベルを維持できるうえに、コスト削減計画をワークロードや組織で求められる要件と常に一致させることができます。

割引は変動するため、アカウントにはターゲットカバレッジを設定しないでください。カバレッジのレベルが低いからといって、必ずしも削減額が大きくはなりません。アカウントのカバレッジが低い場合でも、ライセンスを取得したオペレーティングシステムで、スモールインスタンスで構成して使用する場合は、割引率がわずか数パーセントにしかならないこともあります。その場合は、Savings Plan レコメンデーションツールで、削減可能なものを追跡してモニタリングします。Cost Explorer で Savings Plans のレコメンデーションを頻繁に確認して (定期的な分析を実行)、削減率の見積りが組織で必要な割引率を下回るまで、コミットメントを引き続き購入します。たとえば、削減率の想定見積りが 20% を下回っている状態を追跡、モニタリングしているとします。20% を超えたらコミットメントを購入する必要があります。

使用率とカバレッジをモニタリングする際は、変更の検出のみ行います。特定の使用率 (カバレッジ率) を目標にしないでください。節減の結果必ずしもスケールできるとは限らないからです。Savings Plans を購入してカバレッジが増加していることを確認して、もしカバレッジや使用率が減少した場合は数値化して公開するようにします。たとえば、ワークロードリソースを新しいインスタンスタイプに移行すると、既存プランの使用率が減少しますが、パフォーマンスで得られる利点の方が使用率減少の利点よりも大きくなります。

コミットメント割引 – リザーブドインスタンス/コミットメント: Savings Plans と同様に、リザーブドインスタンスでは、最小限のリソースを実行するコミットメントに対して最大 72% の割引があります。リザーブドインスタンスは、RDS、Elasticsearch、ElastiCache、Amazon Redshift、DynamoDB で利用できます。Amazon CloudFront と AWS Elemental MediaConvert では、最低利用料金のコミットメントに対して割引が適用されます。リザーブドインスタンスは現在 EC2 で利用できますが、Savings Plans では柔軟性が高く、管理諸経費なしで利用できる同レベルの割引があります。

リザーブドインスタンスでも、前払いなし、一部前払い、全前払い、1 年または 3 年という同じ料金オプションがあります。

リザーブドインスタンスは、リージョンまたは特定のアベイラビリティーゾーンで購入できます。アベイラビリティーゾーンで購入すると、キャパシティーを予約できます。

EC2 はコンバーティブル RI を備えていますが、柔軟性の向上と運用コストの削減のために、すべての EC2 インスタンスで Savings Plans を使用する必要があります。

リザーブドインスタンスの追跡と購入には、同じプロセスとメトリクスを使用する必要があります。アカウント全体の RI のカバレッジは追跡しないことを推奨します。また、使用率 (%) をモニタリングや追跡せずに、Cost Explorer で使用量レポートからテーブルの [Net savings] 列を使用することを推奨します。純削減額が大幅にマイナスである場合、未使用の RI を修正するための措置を講じる必要があります。

EC2 フリート: EC2 フリートは、ターゲットとするコンピューティングキャパシティーを定義し、インスタンスタイプやフリートのオンデマンドとスポットのバランスを指定できる機能です。EC2 フリートによって、リソースの最低料金の組み合わせが自動的に起動され、定義されたキャパシティーが達成されます。

地理的選択: ソリューションを設計するときのベストプラクティスは、コンピューティングリソースをユーザーに近い場所に配置し、レイテンシー低下とデータ主権の強化を図ることです。グローバルな利用者のニーズに応えるためには、複数のロケーションを使用する必要があります。コストを最低限に抑えられる地理的ロケーションを選択する必要もあります。

AWS クラウドのインフラストラクチャは、リージョンとアベイラビリティーゾーンを中心として構築されます。リージョンとは世界の物理的なロケーションを意味し、各リージョンには複数のアベイラビリティーゾーンが配置されています。アベイラビリティーゾーンは少なくとも 1 か所の独立したデータセンターで構成されています。各データセンターには、冗長性のある電源、ネットワーキング、接続性能が備えられ、それぞれが別々の設備に収められています。

各 AWS リージョンは、ローカルの市場状況において運用されるため、リソースの料金はリージョンごとに異なります。 ソリューションのコンポーネントまたはソリューション全体を運用するための特定のリージョンを選択して、最低限の料金でグローバルに実行できるようにします。AWS 簡易見積りツールを使用すると、さまざまなリージョンのワークロードのコストを見積もることができます。

サードパーティーの契約と料金: クラウドでサードパーティーのソリューションまたはサービスを利用する場合、料金構造とコスト最適化の結果を連動させることが重要です。料金は、コスト最適化の結果とサービスの価値に合わせてスケールする必要があります。この一例として、削減率を使用したソフトウェアがあります。削減率 (結果) が高くなるほど請求額も上がるというものです。請求額に合わせてスケールする契約は、特定の請求書のあらゆる部分で結果が得られない限り、ほとんどの場合コストの最適化と連動していません。たとえば、EC2 のレコメンデーションが提供され、請求全体のある割合が課金されるソリューションでは、利点がない他の複数のサービスを使用をしている場合に、請求額が上がります。もう 1 つの例は、管理するリソースのコストの割合に応じて課金されるマネージドサービスです。インスタンスサイズを大きくすると常に管理作業が増えるわけではありませんが、請求額が高くなります。これらのサービス料金設定に、コスト最適化プログラムまたはサービス機能が含まれていることを承知のうえで効率性を向上させます。

データ転送を計画する

クラウドの利点は、マネージド型のネットワークサービスであることです。スイッチ、ルーター、その他の関連するネットワーク機器などのフリートの管理や運用は不要になります。クラウド内のネットワーキングリソースは CPU とストレージと同じように消費され、同じように実際に使用した分だけを支払うことになります。クラウドでコストを最適化するには、ネットワーキングリソースを効率的に使用する必要があります。

データ転送モデリングを実行: ワークロードでデータ転送が発生する場所、転送コスト、データ転送の利点を理解します。これにより、十分な情報に基づいてアーキテクチャ設計上の変更や承諾の決定ができます。たとえば、アベイラビリティーゾーン間でデータをレプリケートするマルチアベイラビリティーゾーンを設定したとします。構造のコストをモデリングし、これが許容可能なコスト (両方のアベイラビリティーゾーンにおけるコンピューティングコストとストレージコストと同様のもの) であると判断されると、必要な信頼性と耐障害性が達成されます。

さまざまな使用レベルでコストをモデリングします。ワークロード使用量は経時変化します。また、サービスの種類ごとに異なるレベルで費用対効果が向上する場合があります。

AWS Cost Explorer またはコストと使用状況レポート (CUR) を使用して、データ転送コストを把握し、モデリングします。PoC (概念実証) を設定するか、またはワークロードをテストして、現実的な条件でシミュレートされた負荷を用いてテストを実行します。ワークロードのさまざまな需要に応じてコストをモデルリングできます。

データ転送を最適化: データ転送向けのアーキテクチャでは、データ転送コストを最低限に抑えられます。このアーキテクチャでは、コンテンツ配信ネットワークを使用してユーザーに近いデータを特定したり、お客様のプレミスと AWS をつなぐ専用ネットワーク接続が使用される場合があります。WAN の最適化やアプリケーションの最適化によって、コンポーネント間で転送されるデータ量を減らすこともできます。

データ転送コストを削減するサービスを選択: Amazon CloudFront は、データを低レイテンシーで高速転送するグローバルなコンテンツ配信ネットワークです。世界のエッジロケーションでデータをキャッシュするため、リソースの負荷が軽減されます。CloudFront を使用することで、世界の多数のユーザーにコンテンツを配信するための管理労力が軽減され、レイテンシーも最低限に抑えることができます。

AWS Direct Connect を使用すると、お客様のオンプレミスから AWS への専用ネットワーク接続を確立できます。ネットワークコストが削減され帯域幅が増加するほか、インターネット経由の接続よりも安定したネットワーク接続が可能になります。

AWS VPN を使用すると、プライベートネットワークと AWS グローバルネットワークとの間にセキュアでプライベートな接続を確立できます。迅速かつ簡単な接続とフルマネージド型の伸縮自在なサービスは、小規模なオフィスやビジネスパートナーに最適です。

VPC エンドポイントにより、プライベートネットワークを利用して AWS のサービス間の接続が可能になり、パブリックデータ転送のコストと NAT ゲートウェイのコストを削減できます。ゲートウェイ VPC エンドポイントには時間単位の課金はなく、Amazon S3 と Amazon DynamoDB がサポートされています。インターフェイス VPC エンドポイントは AWS PrivateLink により提供され、時間単位の料金と GB あたりの使用料がかかります。

リソース

費用対効果の高いリソースに関する AWS のベストプラクティスの詳細については、以下のリソースを参照してください。

- AWS マネージドサービス: エンタープライズトランスフォーメーションジャーニーの動画
- Cost Explorer によるコストの分析
- リザーブドインスタンスのレコメンデーションへのアクセス
- 正しいサイズ設定のレコメンデーションの開始方法
- スポットインスタンスのベストプラクティス
- スポットフリート
- リザーブドインスタンスの仕組み
- AWS グローバルインフラストラクチャ
- スポットインスタンスアドバイザー
- Well-Architected ラボ - 費用対効果の高いリソース

需要と供給を一致させる

クラウドに移行すると、必要な分だけを支払うことになります。必要な時にワークロードの需要に合わせたリソースを供給できるため、コストがかかる無駄なオーバープロビジョニングを排除できます。また、スロットル、バッファ、キューを使用して需要を変更すると、少ないリソースで需要を円滑に処理できます。

ジャストインタイム供給による経済的利点を得るには、リソース障害、高可用性、プロビジョニング時間を考慮したプロビジョニングの必要性とのバランスを保つことが必要です。需要の変動性の有無に応じて、スケーリング中であっても、環境の管理を最小限に抑えるためにメトリクスおよびオートメーションを計画します。需要を変更するときは、ワークロードが許容できる最大遅延を把握しておく必要があります。

AWS には、需要管理とリソース供給に使用できるさまざまなアプローチがあります。以下のセクションでは、アプローチの使用方法を説明します。

- ワークロードの分析
- 需要管理
- 需要ベースの供給
- 時間ベースの供給

ワークロードの分析: ワークロードの要件を把握します。組織の要件に、リクエストに対するワークロードの応答時間を含める必要があります。応答時間は、需要が管理されているかどうか、または需要を満たすためにリソースの供給が変化するかどうかを判断するために使用できます。

分析には、需要の予測可能性と再現性、需要の変化率、需要の変化量を含める必要があります。分析は、月末処理や休日のピークなどの時季的な変動が組み込まれるように、十分な期間にわたって実行されるようにします。

分析作業では、スケーリングの運用により潜在的な利点が反映されていることを確認します。コンポーネントの予想される合計コスト、ワークロードのライフタイムにおける使用量の増減およびコストの増減に注目します。

AWS Cost Explorer または Amazon QuickSight を CUR またはアプリケーションログと併用すると、ワークロードの需要を可視化して分析できます。

需要管理

需要管理 – スロットリング: 需要元のソースに再試行機能がある場合は、スロットリングを設定できます。現在リクエストを処理できない場合は、後で再試行する必要があることがスロットリングによってソースに通知されます。ソースでは一定時間待機してから、リクエストが再試行されます。スロットリングの運用には、リソースの最大量およびワークロードのコストを制限できるという利点があります。AWS では、スロットリングは Amazon API Gateway によって運用できます。スロットリングの運用の詳細については、Well-Architected 信頼性の柱に関するホワイトペーパーを参照してください。

需要管理 – バッファベース: バッファはスロットリングと同様に、リクエスト処理を延期し、アプリケーションが異なる動作速度で実行されていても効果的に通信できるようにします。バッファベースのアプローチでは、キューを使用してプロデューサーからメッセージ (作業単位) を受信します。メッセージはコンシューマーによって読み取られ、処理されるため、コンシューマーのビジネス要件を満たせる動作速度でメッセージを実行できます。プロデューサーがデータの耐久性やバックプレッシャーなどのスロットルの問題に対処する必要があることを心配する必要はありません（コンシューマーの動作が遅いためにプロデューサーが遅くなります）。

AWS でバッファベースのアプローチを運用する際は、複数のサービスから選択できます。Amazon SQS は、1 人のコンシューマーが個別のメッセージを読むことができるキューを提供するマネージドサービスです。Amazon Kinesis は、多数のコンシューマーが同じメッセージを読むことができるストリームを提供します。

バッファベースのアプローチで設計する場合は、必要な時間内にリクエストを処理するワークロードが設計されていて、さらに作業の重複リクエストをハンドリングできるようにします。

動的供給

需要ベースの供給: クラウドの伸縮自在性を活用して、需要の変化に対応するリソースを提供します。API やサービス機能を活用すると、アーキテクチャ内のクラウドリソースの量をプログラムで動的に変更できます。これにより、アーキテクチャ内のコンポーネントをスケールしたり、需要急増の際にはリソース数を自動的に増加させてパフォーマンスを維持したり、需要低下の際には需要キャパシティーを減らしてコストを削減したりできます。

Auto Scaling により、キャパシティーを調整して、最低限のコストで安定かつ予測可能なパフォーマンスを維持できます。これは、Amazon EC2 インスタンス、スポットフリート、Amazon ECS、Amazon DynamoDB、Amazon Aurora と統合されたフルマネージド型の無料サービスです。

Auto Scaling では、リソースの自動検出によってワークロード内に設定可能なリソースを検出できます。また、パフォーマンス、コスト、この両者のバランスを最適化するためのスケーリング戦略が組み込まれており、予測スケーリングによって定期的に発生する急増に対応できます。

Auto Scaling では、手動スケーリング、スケジュールに基づくスケーリング、需要ベースのスケーリングのいずれかを運用できます。また、Amazon CloudWatch のメトリクスとアラームを使用して、ワークロードのスケーリングイベントをトリガーできます。一般的なメトリクスは、CPU 使用率、ネットワークスループット、ELB で確認されたリクエスト/レスポンスのレイテンシーなど、Amazon EC2 の標準メトリクスです。可能な場合は、カスタマーエクスペリエンスの指標となるメトリクスを使用する必要があります。このメトリクスは一般には、ワークロード内のアプリケーションコードから生成されるカスタムメトリクスです。

需要ベースのアプローチで設計する場合は、主に 2 つの点を考慮する必要があります。第 1 に、新しいリソースをどれだけ早くプロビジョニングする必要があるかを理解することです。第 2 に、需要と供給の差異が変動することを理解することです。需要の変動ペースに対処できるようにしておくだけでなく、リソースの不具合にも備えておく必要があります。

Elastic Load Balancing (ELB) は、複数のリソースに需要を分散して、スケーリングを支援します。運用するリソースが増加すると、ロードバランサーにリソースを追加し需要を分散させます。AWS ELB では、EC2 インスタンス、コンテナ、IP アドレス、Lambda 関数がサポートされています。

時間ベースの供給: 時間ベースのアプローチでは、リソースのキャパシティーを予測可能な需要、または時間ごとに明確に定義された需要に合わせます。このアプローチは通常、リソースの使用率レベルに依存しません。リソースが必要な特定の時間にそのリソースを確保します。また、起動手順、およびシステムや一貫性のチェックにより、遅延なくリソースを提供できます。時間ベースのアプローチでは、繁忙期に追加のリソースを供給したり、キャパシティーを拡大したりできます。

時間ベースのアプローチは、スケジュールされた Auto Scaling によって運用できます。営業開始時など、特定の時間にスケールアウトまたはスケールインするようにスケジュールできるため、ユーザーがアクセスしたときや需要が発生したときにリソースを利用可能にしておくことができます。

また、AWS API、SDK、AWS CloudFormation を活用して、必要に応じて自動的にプロビジョニングしたり、環境全体を削除したりできます。このアプローチは、所定の営業時間や一定期間にのみ実行される開発環境またはテスト環境に適しています。

API を使用した環境内のリソースサイズのスケーリング (垂直スケーリング) にも対応しています。たとえば、インスタンスのサイズやクラスを変更して、本番稼働ワークロードをスケールアップできます。これには、インスタンスを停止、起動し、別のインスタンスサイズやインスタンスクラスを選択します。このテクニックは、インスタンス使用中にサイズの拡大、パフォーマンスの調整 (IOPS)、ボリュームタイプの変更が可能な EBS Elastic Volumes などのリソースにも適用できます。

時間ベースのアプローチを設計する際は、主に 2 つの点を考慮する必要があります。第 1 に、使用パターンの一貫性です。 第 2 に、パターンを変更した場合の影響です。 予測精度は、ワークロードをモニタリングし、ビジネスインテリジェンスを使用することで高めることができます。使用パターンに大幅な変更がある場合は、時間を調整して予測対象範囲に収まるようにします。

動的な供給: AWS Auto Scaling を使用するか、AWS API または SDK を使用してコードにスケーリングを組み込むことができます。これにより、環境を手動変更していた運用コストがなくなり、その結果、全体的なワークロードコストが削減され、実行速度が向上します。また、ワークロードリソースと需要を常に一致させることができます。

リソース

需要管理とリソース提供に関する AWS のベストプラクティスの詳細については、以下のリソースを参照してください。

- API Gateway のスロットリング
- Amazon SQS の開始方法
- Amazon EC2 Auto Scaling の開始方法

継続的最適化

AWS では、新しいサービスをレビューし、それをワークロードに運用することによって、時間をかけて最適化していきます。

新しいサービスをレビューして運用する

AWS で新しいサービスと機能がリリースされたときは、ベストプラクティスとして、既存のアーキテクチャの決定事項をレビューし、費用対効果が維持されるようにすることが大切です。要件の変化に応じて、不要になったリソース、コンポーネント、ワークロードを積極的に排除します。継続的に最適化を行うには、次の点を考慮します。

- ワークロードのレビュープロセスの開発
- サービスのレビューと運用

ワークロードレビューのプロセスの開発: ワークロードの費用対効果が常に最大になるようにするには、ワークロードを定期的にレビューし、新しいサービス、機能、コンポーネントを運用する機会の有無を考慮する必要があります。全体的なコスト削減を達成するには、潜在的なコスト削減量に比例したプロセスを行う必要があります。たとえば、支出全体の 50% を占めるワークロードは、支出全体の 5% を占めるワークロードよりも定期的かつ徹底的にレビューする必要があります。外部要因または変動性を考慮します。ワークロードにより特定の地域、特定の市場セグメントにサービスが提供されていて、その領域での変化が予測される場合、レビュー頻度を高くすることでコスト削減につながる可能性があります。レビューで考慮すべきもう 1 つの要因は、変更を運用する労力です。変更のテストおよび検証に多大なコストがかかる場合は、レビューの頻度を下げる必要があります。

時代遅れのレガシーコンポーネントやリソースには維持するために継続的にコストがかかることや、新しい機能を運用できないことを考慮します。テストと検証にかかる現在のコストが、提案されている利益を上回っている場合があります。しかし、ワークロードと現在のテクノロジーとのギャップが時間の経過とともに大きくなるにつれて、変更にかかるコストが増加し、結果として巨額のコストになることがあります。たとえば、新しいプログラミング言語に移行するときの費用対効果は現時点で低いとします。しかし、5 年後には、その言語に精通した人材のコストが増加する可能性があります。ワークロードが増加すると、さらに大規模なシステムを新しい言語に移行することになり、結果的にこれまでよりもさらに多大な労力を要します。

ワークロードをコンポーネントに分割し、コンポーネントのコストを割り当て (コスト見積もりで可)、各コンポーネントにその要因 (労力や外部市場など) を一覧表示します。この指標を使用して、各ワークロードのレビュー頻度を決定します。たとえば、ウェブサーバーが高コストで、変更の労力が低く、外部要因が高い場合は、レビュー頻度が高くなります。中央データベースが中程度のコストで、変更の労力が高く、外部要因が低い場合は、レビューの頻度は中程度になります。

ワークロードの確認とサービスの運用: 新しい AWS のサービスと機能の利点を得るには、ワークロードでレビュープロセスを実行し、必要に応じて新しいサービスや機能を運用する必要があります。たとえば、ワークロードをレビューし、メッセージングコンポーネントを Amazon Simple Email Service (SES) に置き換えできます。これにより、すべての機能を低コストで提供しながら、インスタンスのフリートの運用と維持にかかるコストを削減できます。

まとめ

コスト最適化とクラウドの財務管理は、継続的な取り組みです。財務チームやテクノロジーチームと定期的に協力し、アーキテクチャのアプローチをレビューし、コンポーネントの選択を更新していく必要があります。

AWS が目指すのは、弾力性、応答性、順応性に優れたデプロイを最低限のコストで実現することです。デプロイコストを適切に最適化するには、本ドキュメントでご紹介したツール、テクニック、ベストプラクティスを活用してください。

寄稿者

本ドキュメントの寄稿者は以下のとおりです。

- Philip Fitzsimons、Well-Architected シニアマネージャー、アマゾン ウェブ サービス
- Nathan Besh、Well-Architected コストリード、アマゾン ウェブ サービス
- Levon Stepanian、アマゾンウェブサービス
- Keith Jarrett、ビジネス開発リード – コスト最適化
- PT Ng、コマーシャルアーキテクト、アマゾン ウェブ サービス
- Arthur Basbaum、ビジネス開発者マネージャー、アマゾン ウェブ サービス
- Jarman Hauser、コマーシャルアーキテクト、アマゾン ウェブ サービス

その他の資料

詳細については、以下のソースを参照してください。

- AWS Well-Architected フレームワーク

ドキュメント改訂履歴

日付	説明
2020 年 4 月	CFM、新しいサービス、Well-Architected との統合を追加。
2018 年 7 月	AWS の変更内容を反映し、カスタマーレビューから学習した
2017 年 11	AWS の変更内容を反映し、カスタマーレビューから学習した
2016 年 11	初版発行