AWS Bedrock クロスリージョン推論 調査レポート

1. はじめに

- **目的**
  - AWS Bedrockの「クロスリージョン推論」機能について、概要、利点、欠点、特にセキュリティとデータプライバシーの観点から解説します。
- **背景**
  - 最新の高性能AIモデルは、特定の海外リージョン（主に米国）でのみ利用可能になる傾向があります。
  - 日本国内からこれらの最新モデルを安全に活用するため、本機能の理解は不可欠です。

2. クロスリージョン推論とは？

- 単一のAPIエンドポイント（推論プロファイルID）を呼び出すだけで、リクエストが**複数のAWSリージョンへ自動的にルーティング**される仕組みです。
- **主な目的：**
  - 特定リージョンのトラフィック急増やリソース不足が発生しても、他のリージョンのリソースを活用し、**スループット（処理能力）と可用性（安定稼働）**を向上させます。

3. 推奨アーキテクチャ：国内データストアとの分離

データガバナンスと最新技術活用の両立

- **推論（コンピュート）のみクロスリージョン**
  - LLM/埋め込みモデルのAPI呼び出し
- **データ（ストレージ）は国内に固定**
  - Amazon S3
  - データベース (Aurora, DynamoDB)
  - アプリケーション (EC2, Lambda)

4. メリット

- **スループットと可用性の向上**
  - 複数リージョンの潤沢なリソースを利用でき、大規模リクエストに対応可能です。
- **耐障害性の強化**
  - リージョン障害時も、正常な別リージョンへ自動でフェイルオーバーします。
- **追加コスト不要**
  - 機能利用やリージョン間データ転送に追加料金はかかりません。料金は常に呼び出し元のリージョン（ソースリージョン）に基づきます。
- **実装の容易さ**
  - 複雑な負荷分散や障害対応ロジックを自前で実装する必要がありません。

5. デメリットと考慮事項

- **レイテンシー（遅延）**
  - リクエストが海外リージョンで処理された場合、通信のためにわずかな遅延（数十ミリ秒程度）が追加される可能性があります。
- **対象モデル・リージョンの制約**
  - この機能は一部のモデル・リージョンでのみ利用可能です。
  - 日本の東京リージョン（`ap-northeast-1`）は**APACグループ**に含まれ、対応する推論プロファイルを利用できます。

6. セキュリティとデータプライバシー (1/4)

AWS全体のセキュリティレベル

- AWSは、世界で最も厳しいセキュリティ要件を満たすよう設計されています。
- **ISMAP（イスマップ）**
  - 日本政府のセキュリティ評価制度に登録されており、政府機関が利用できる高いセキュリティ基準を満たしていることが証明されています。
- **国際認証**
  - `ISO 27001` (情報セキュリティ)
  - `ISO 27017` (クラウドセキュリティ)
  - `ISO 27018` (クラウドの個人情報保護)
  - その他、多数の認証を取得しています。

6. セキュリティとデータプライバシー (2/4)

Q1. データは海外に永続保存されますか？

回答：いいえ。データが永続的に保存される場所は、常にAPIを呼び出した国内のソースリージョンです。

- **解説**
  - 処理のためにプロンプトや生成結果が海外リージョンに**一時的に転送**されることはあります。
  - しかし、CloudTrail監査ログやモデル呼び出しログなどの**永続データは、常にソースリージョン（例：東京）に保存**されます。
  - 顧客はデータの保存場所を完全にコントロールできます。

6. セキュリティとデータプライバシー (3/4)

Q2. データがAIモデルの学習に利用されますか？

回答：いいえ。お客様のデータがモデルの学習に利用されることは一切ありません。

- **解説**
  - AWSはサービス規約で「お客様のコンテンツを基盤モデルの改善やトレーニングに使用しない」と明確に約束しています。
  - このポリシーはクロスリージョン推論の利用有無にかかわらず、全てのBedrock利用者に適用されます。

6. セキュリティとデータプライバシー (4/4)

Q3. 海外リージョンとの通信は安全ですか？

回答：はい。通信はAWS専用のグローバルネットワーク上で常に暗号化され、安全に保護されています。

- **解説**
  - リージョン間のデータ転送は、インターネット公衆網を経由しません。
  - 物理的に隔離・監視された**AWS専用のグローバルネットワークバックボーン**を通じて、常に暗号化されて通信が行われます。

7. 利用者側のセキュリティ対策 (1/3)

責任共有モデルに基づき、利用者側での対策も重要です。

a. IAMによる最小権限の原則

- **概要**
  - ユーザーやアプリケーションには、タスク実行に必要な**最小限の権限**のみを付与します。
- **ベストプラクティス**
  - **モデルの利用を制限：** IAMポリシーの `Resource` で、利用する推論プロファイルやモデルを明示的に許可します。
  - **管理と利用を分離：** モデルの管理を行うロールと、推論を実行するロールの権限を分離します。

7. 利用者側のセキュリティ対策 (2/3)

b. VPCエンドポイントによるプライベート接続

- **概要**
  - VPCエンドポイント（AWS PrivateLink）を利用し、BedrockへのAPI呼び出しを**AWSネットワーク内で完結**させます。
- **ベストプラクティス**
  - **セキュアな通信経路：** パブリックなインターネットを経由しないため、セキュリティが大幅に向上します。
  - **エンドポイントポリシー：** エンドポイント自体にポリシーを設定し、アクセス元をさらに厳しく制限します。

7. 利用者側のセキュリティ対策 (3/3)

c. ログの有効化と監視

- **概要**
  - 誰が、いつ、何をしたかを追跡・監査できるようにログを管理します。
- **ベストプラクティス**
  - **AWS CloudTrail：** 全てのAPI呼び出し履歴を記録し、不正操作の検知や原因調査に活用します。
  - **Bedrockモデル呼び出しログ：** プロンプトや生成結果をS3等に保存します。機密情報が含まれる可能性があるため、**保存先のアクセス権限の最小化と暗号化が必須**です。

8. まとめ

- クロスリージョン推論は、最新モデルを**高パフォーマンス・高可用性**で利用できる強力な機能です。
- **「推論はクロスリージョン、データは国内」**というアーキテクチャを採用することで、データ主権を維持しつつ、安全に利用できます。
- **セキュリティ・プライバシーの要点：**
  - データが海外に**永続保存されることはない**
  - データが**AIの学習に利用されることはない**
  - 通信は**暗号化された専用線**で保護
  - AWSは**ISMAP**に準拠
- 利用者側のベストプラクティスと組み合わせることで、**多層的なセキュリティ**を確保し、国内プロジェクトでも安心して活用可能です。

9. 引用元

- Getting started with cross-region inference in Amazon Bedrock
- Increase throughput with cross-Region inference - Amazon Bedrock
- Supported Regions and models for inference profiles - Amazon Bedrock
- Information System Security Management and Assessment Program (ISMAP) - AWS
- Japan Data Privacy - Amazon Web Services (AWS)
- Implementing least privilege access for Amazon Bedrock
- Security in Amazon Bedrock - Amazon Bedrock User Guide
- Identity and access management for Amazon Bedrock
- Enable Amazon Bedrock cross-Region inference in multi-account environments