最新事例に学ぶ製薬業界向け AWS クラウド活用セミナー2022

創薬ゲノム領域で AWS をフル活用するために

Katsuhisa Takahashi

Solutions Architect

本セッションについて

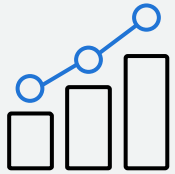
AWS上でゲノム情報を利用するためには
「転送・保存・解析」が必要です。

本セッションでは
転送・保存の基礎と
さらなる高度な解析のための AWS クラウドのフル活用を
目指す方法についてご紹介いたします。

対象 ▶ 創薬ゲノム研究部門および関連する IT 部門、パートナー様

Contents

ゲノムデータ利用の最近



データ容量の
爆発的増加

2025年までに10億を
超えるゲノムが
シーケンス

データ種類の
多様化

マルチオミクス
マルチモーダル

新たなデータソース
からの取り込み

シーケンサに加え
ヘルスケアデバイス
医療情報

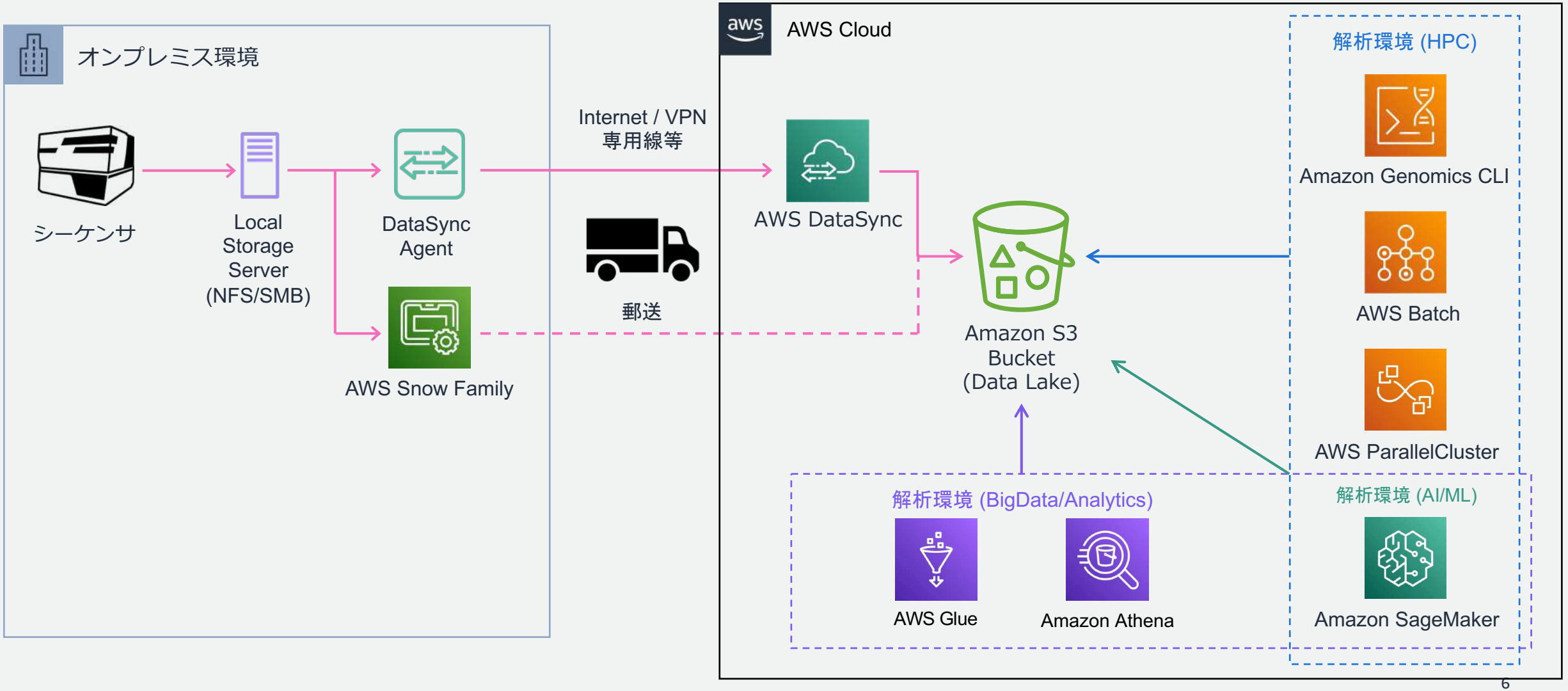
ゲノムデータを
扱う人材の多様化

WET研究者
DRY研究者
データサイエンティ
スト

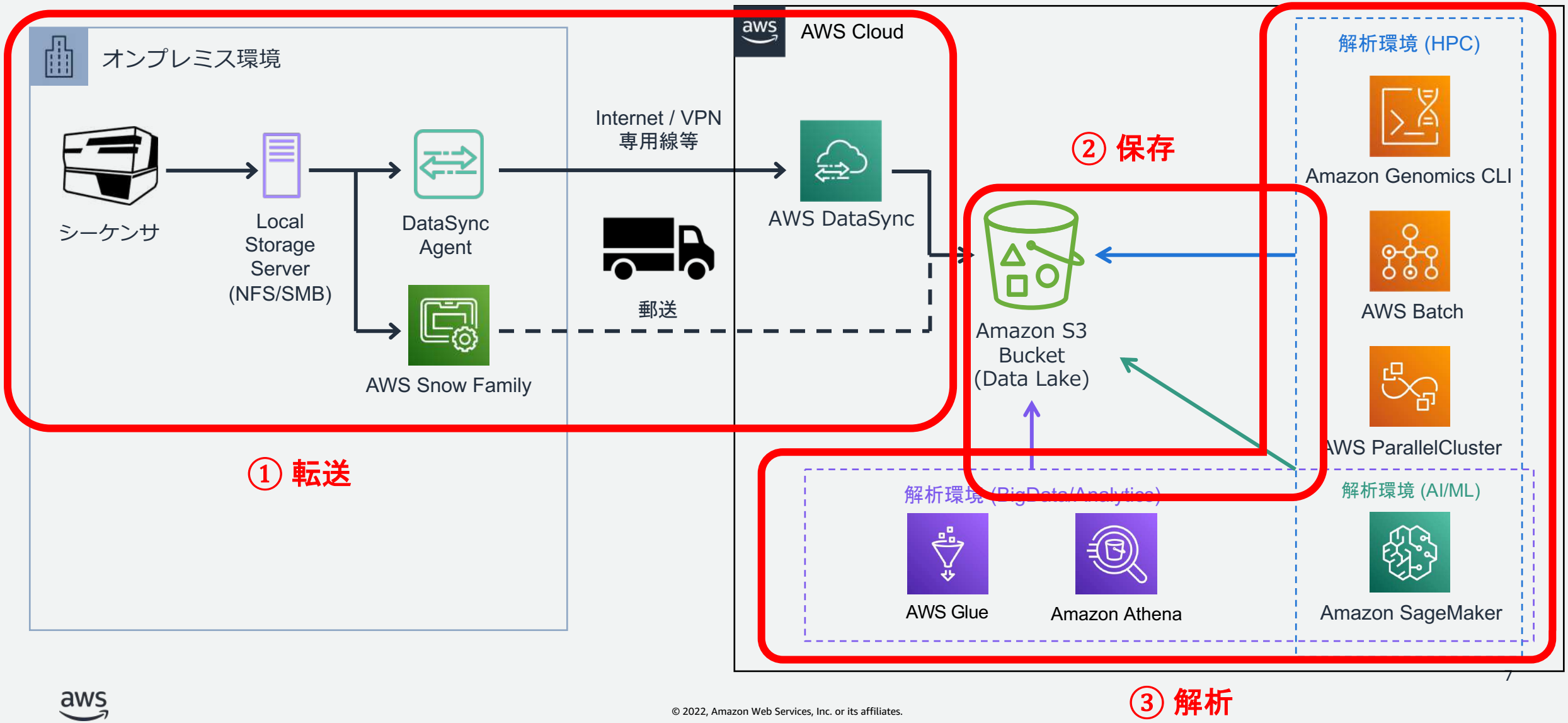
多くのアプリケーシ
ョンからの分析

ノートブック
ダッシュボード

ゲノム情報の「転送・保存・解析」アーキテクチャ

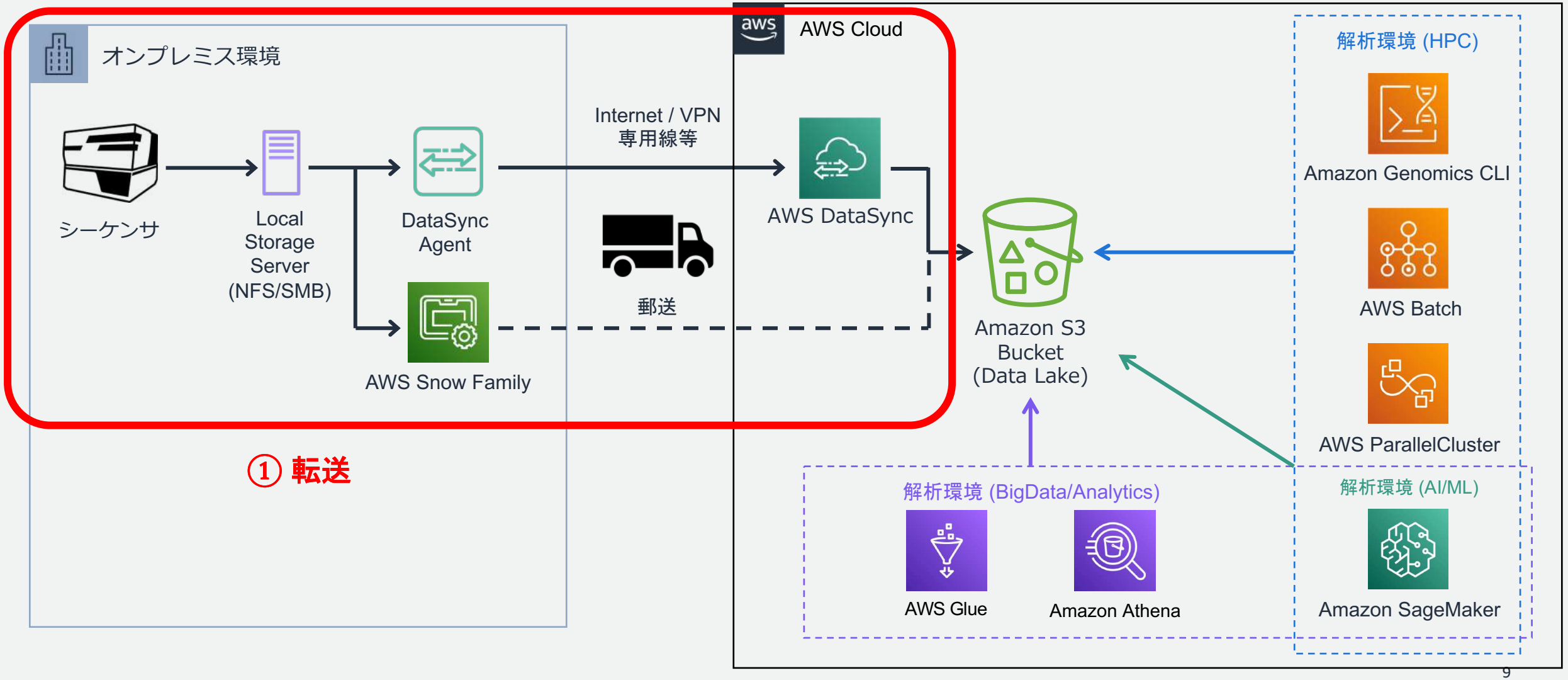


ゲノム情報の「転送・保存・解析」アーキテクチャ



ゲノム情報の転送

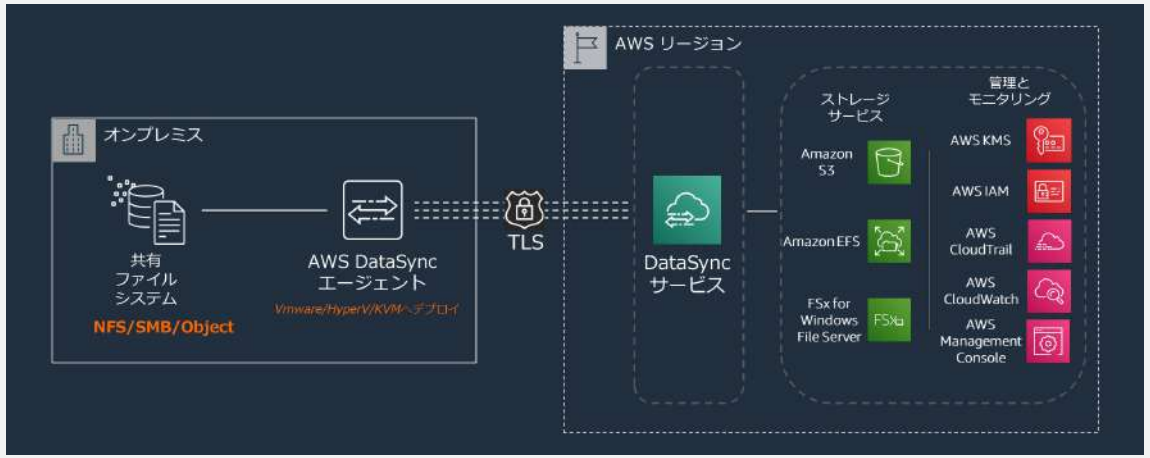
ゲノム情報の「転送」アーキテクチャ



シーケンサーから AWS へのデータ転送

オンライン転送

AWS DataSync



- オンプレミスとクラウドのデータ転送をシンプルかつ高速に自動実行するオンライン転送サービス
- 専用プロトコルによりデータ転送を高速化
- オンプレミス環境で DataSync Agent を稼働させ、共有ストレージをマウントすることで、クラウドからデータの転送を制御

オフライン転送

AWS Snow Family

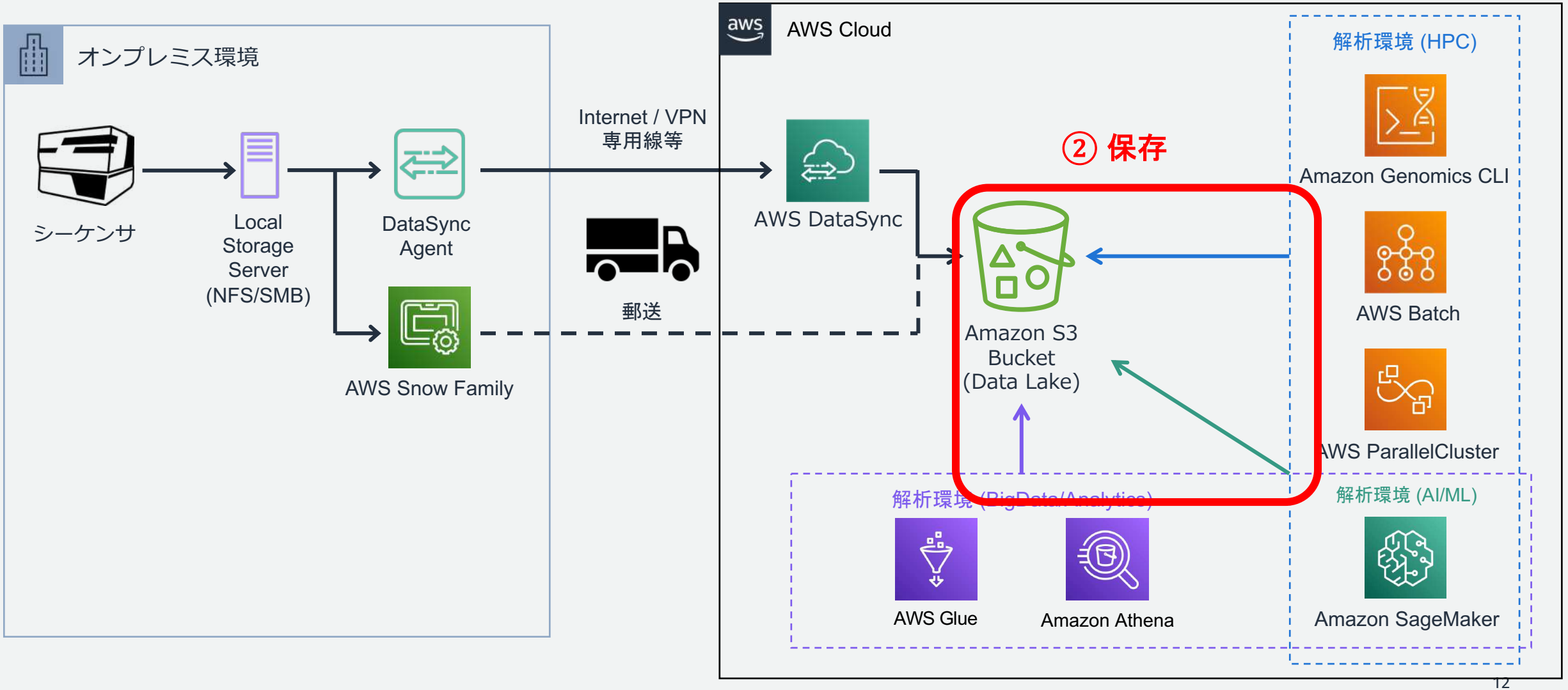
AWS Snowball Edge

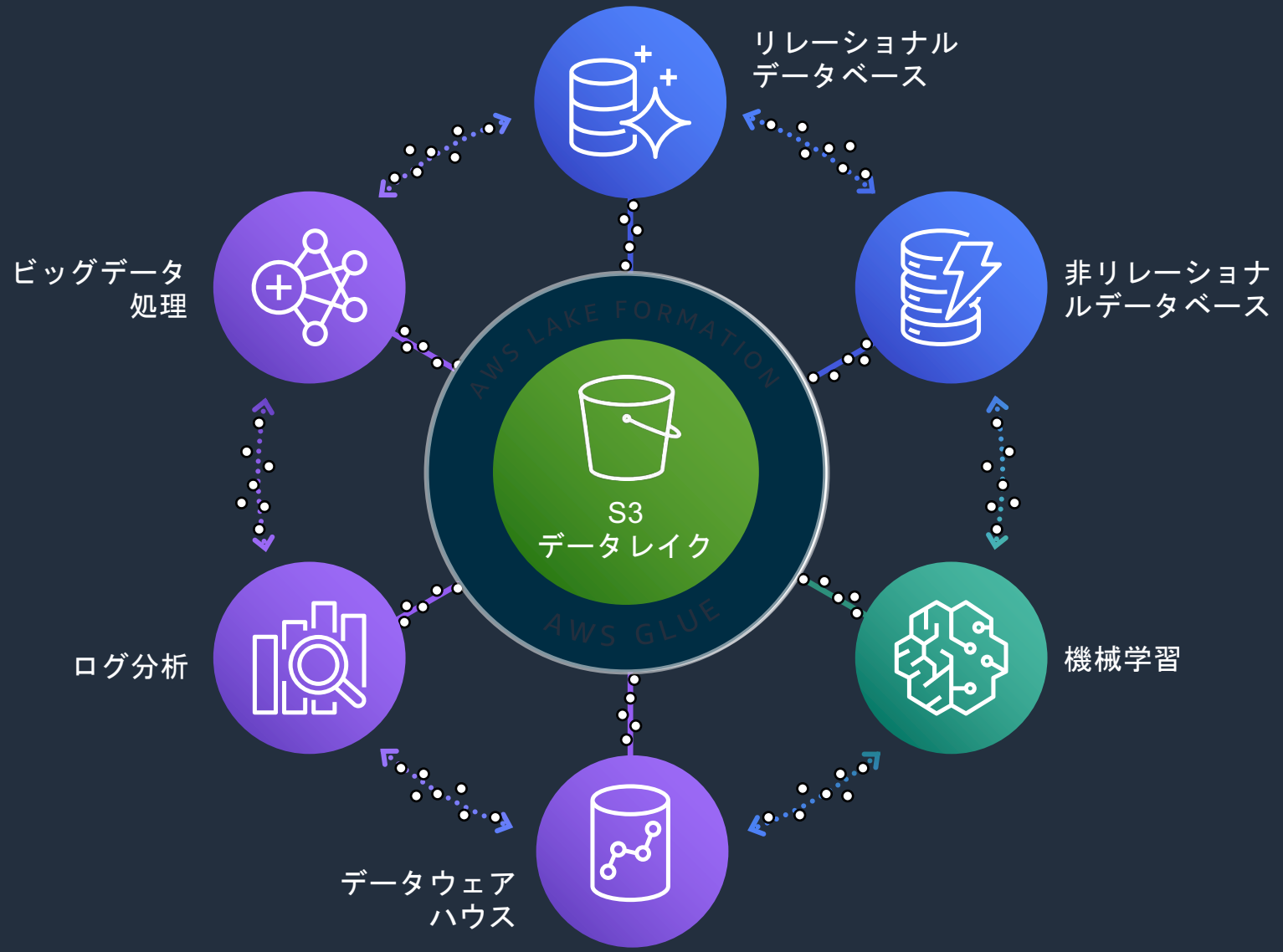
AWS Snowcone

- ハードウェアアプライアンスを郵送することでオンプレミスークラウド間のデータ移行を高速化
- データの自動暗号化や不正開封防止筐体によるセキュリティ確保
- 一度に大量のデータを送ることができるが、郵送に時間や手間がかかる点には注意が必要

ゲノム情報の保存

ゲノム情報の「保存」アーキテクチャ





データレイク アーキテクチャ

その中心に Amazon S3 が存在

Amazon S3 データレイク

比類のない
耐久性, 可用性, 拡張性

コストを最適化しながら
簡単に使える
インテリジェントな階層

データ取得のための
ハブの役割

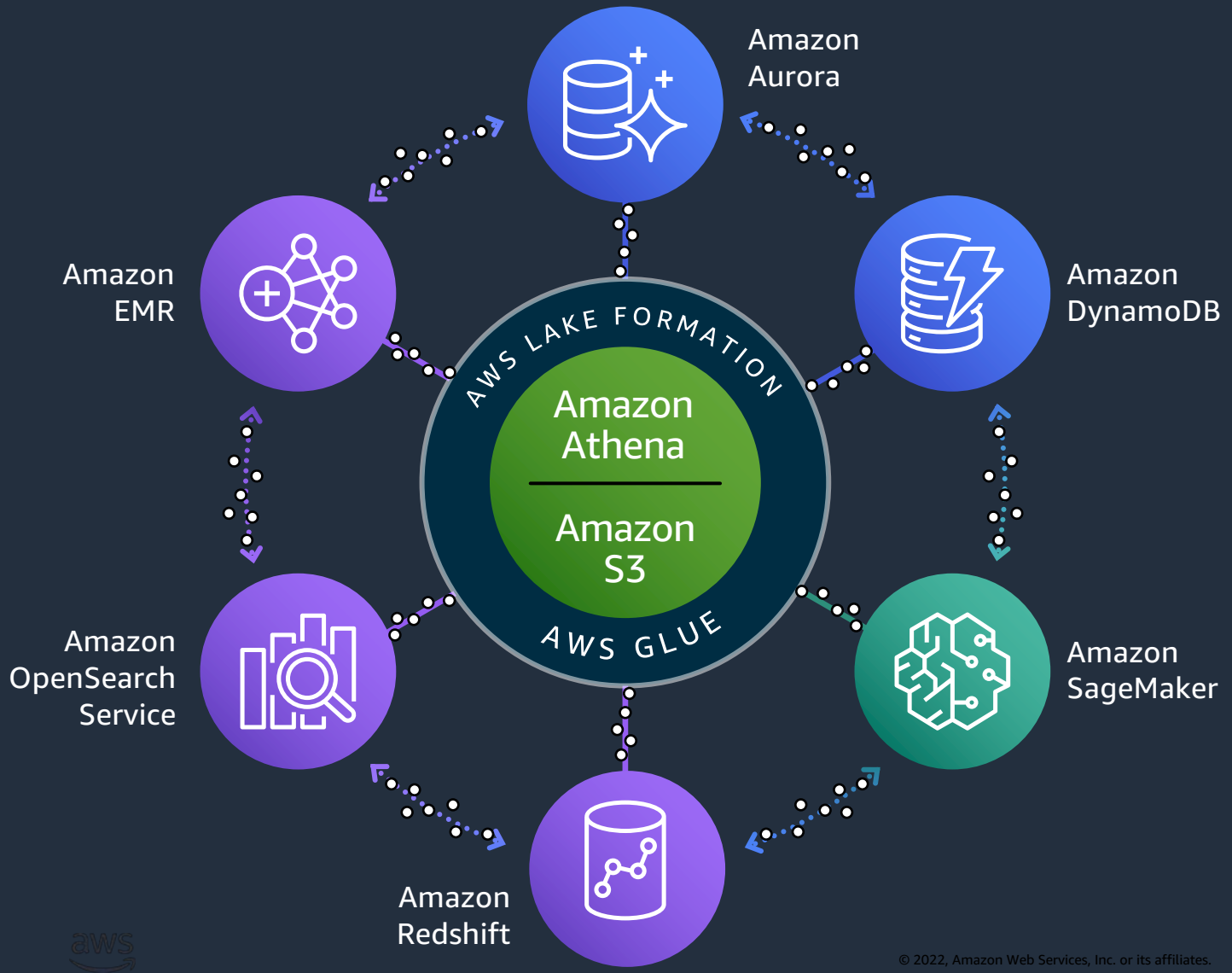
コールドストレージとアーカイブの機能

オブジェクトレベルの
アクセスコントロール

分析ツールの
幅広いポートフォリオ

ベストなセキュリティ,
コンプライアンス, 監査の機能

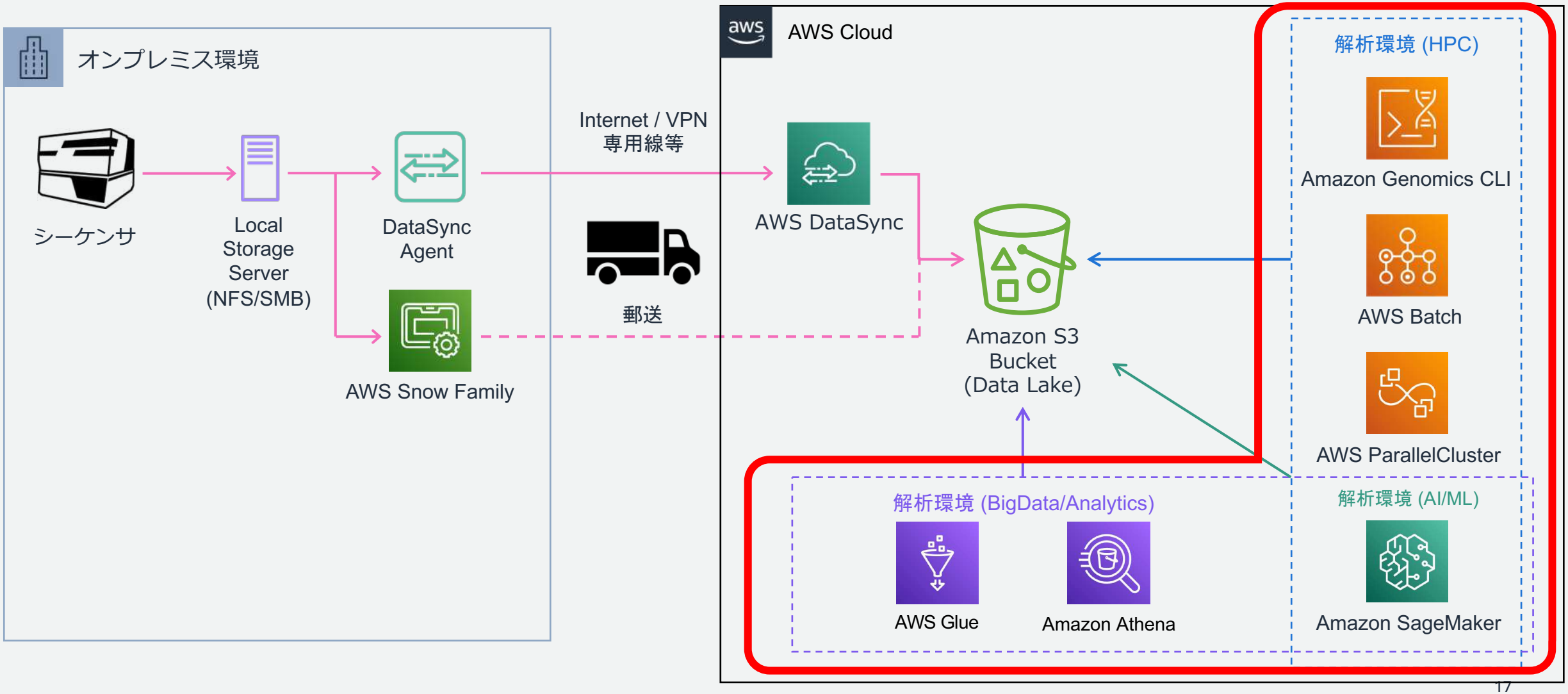
レイクハウスアプローチと **AWS**サービス



データレイク
を囲む形で
様々な AWS サービス
との連携が可能です

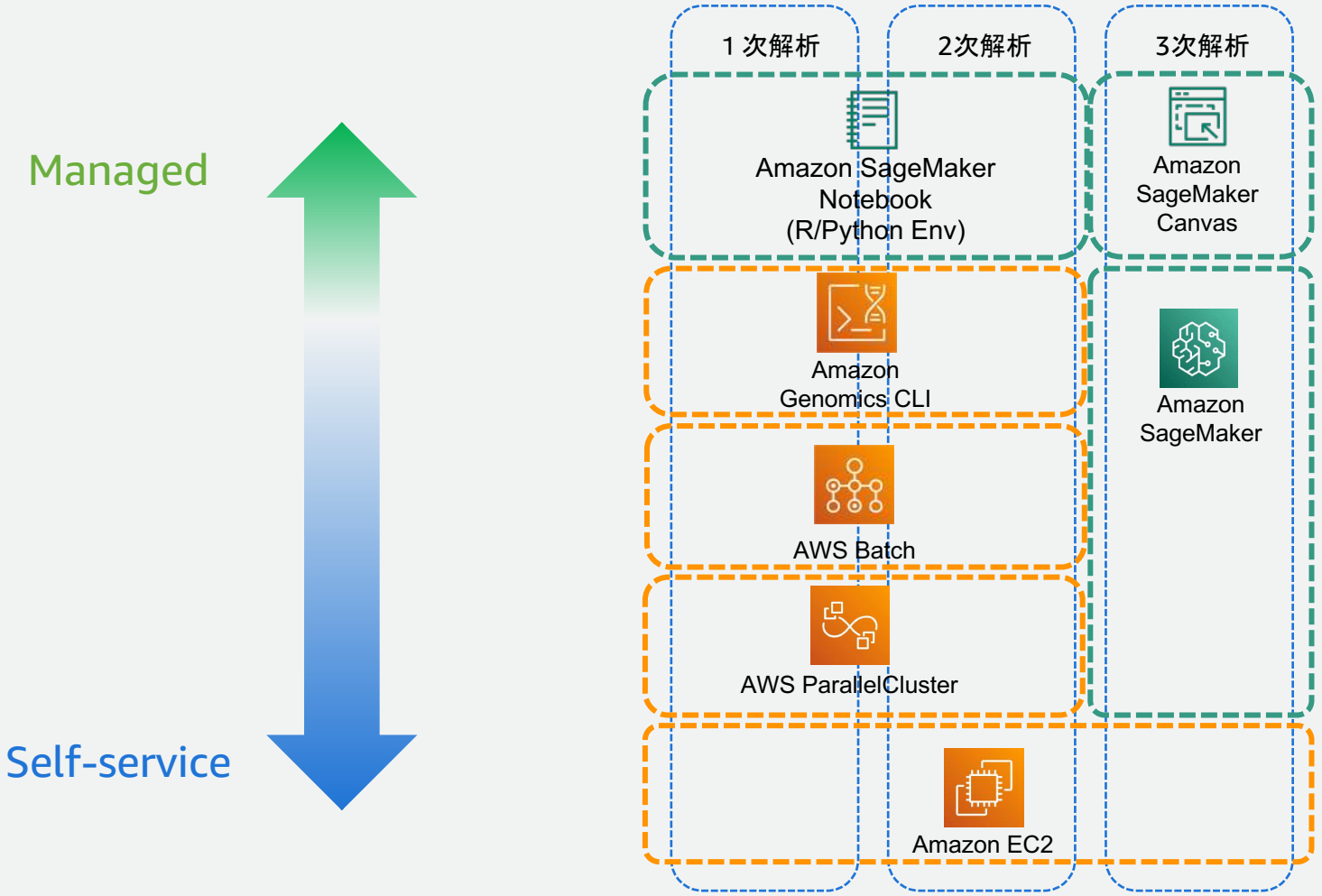
ゲノム情報の解析

ゲノム情報の「解析」アーキテクチャ



③ 解析

解析フェーズ毎にクラウドスキルの異なる利用者をサポート



Amazon EC2 (Elastic Compute Cloud)

必要なときに必要な計算リソースを確保可能な仮想サーバサービス

- 数分で起動し、秒単位の従量課金（一部タイプについては1時間単位）
- 独自の仮想化基盤 Nitro System により、仮想化オーバーヘッドを極小化
- ワークロードに応じて様々なインスタンスタイプを選択可能

高性能計算向けインスタンスタイプの例

高性能 CPU の選択肢

Intel Xeon processor
(x86_64 arch)

M6i インスタンス
Ice Lake
最大時全コア 3.5 GHz 駆動
M5zn インスタンス
Cascade Lake
最大全コア 4.5 GHz 駆動

AMD EPYC processor※
(x86_64 arch)

M6a インスタンス
EPYC Milan
最大 3.3 GHz 駆動
Hpc6a インスタンス
EPYC Milan
HPC特化

AWS Graviton Processor
(64-bit Arm arch)

C7g インスタンス
64bit Arm Neoverse V1ベース
AWS Graviton3 CPU 搭載

アクセラレータの選択肢

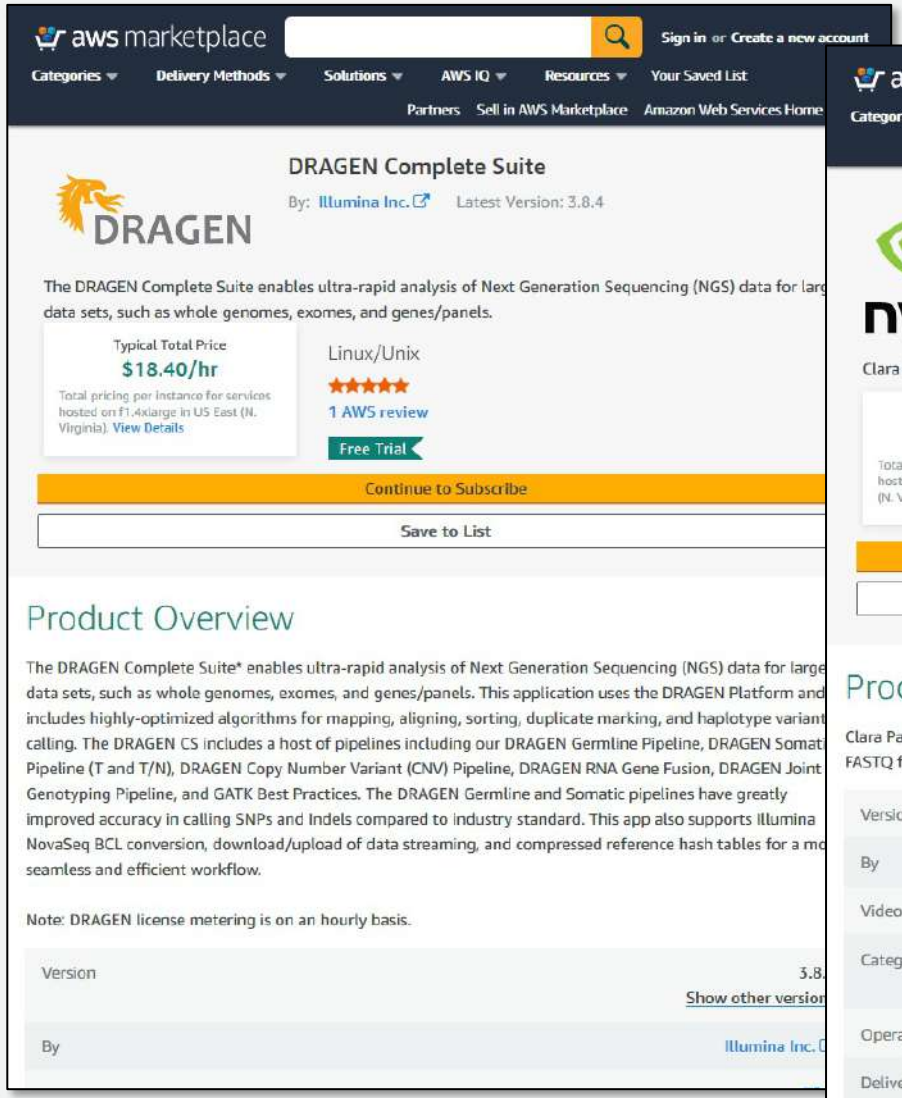
NVIDIA GPU

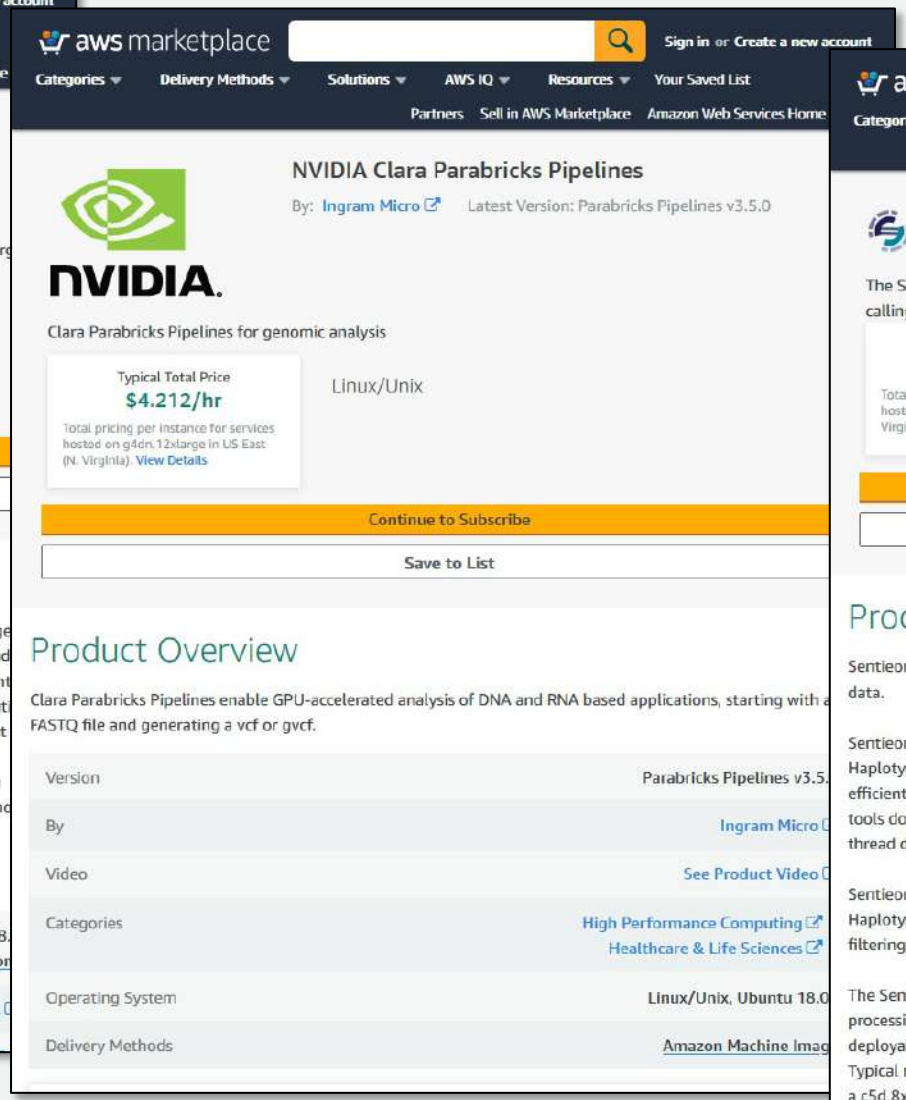
P3 インスタンス
V100 GPU 搭載
P4d インスタンス
A100 GPU 搭載
G5 インスタンス
A10G GPU 搭載

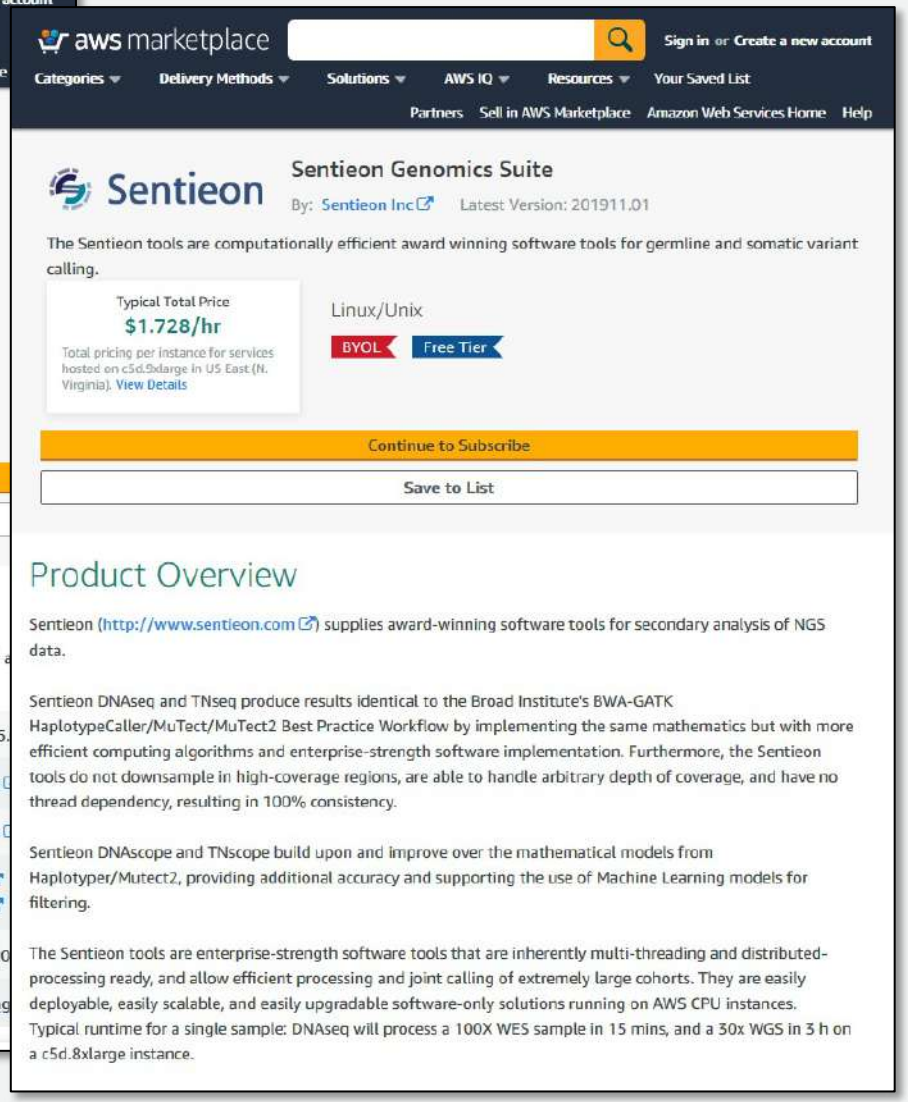
Xilinx FPGA

F1 インスタンス
Virtex UltraScale+
VU9P 搭載

AWS Market Place：様々な製品を選択可能

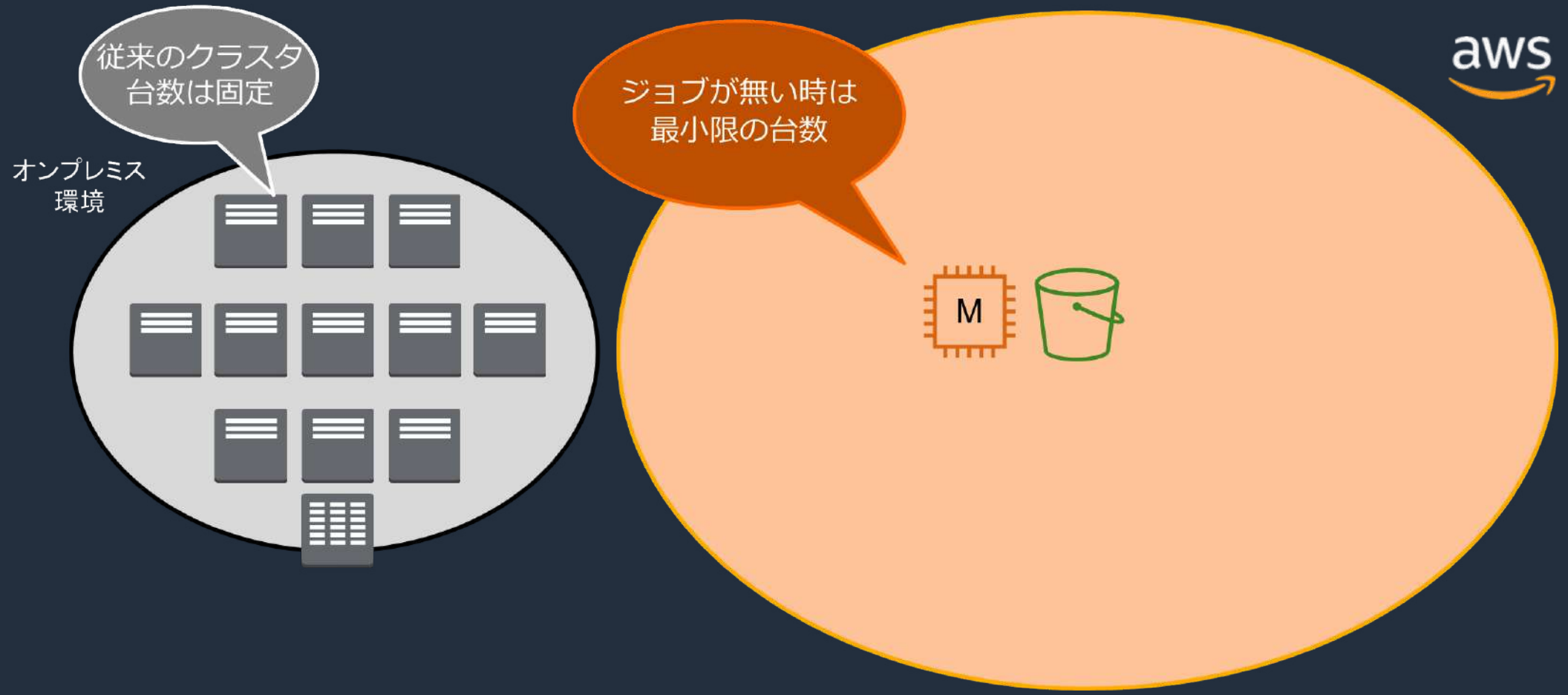






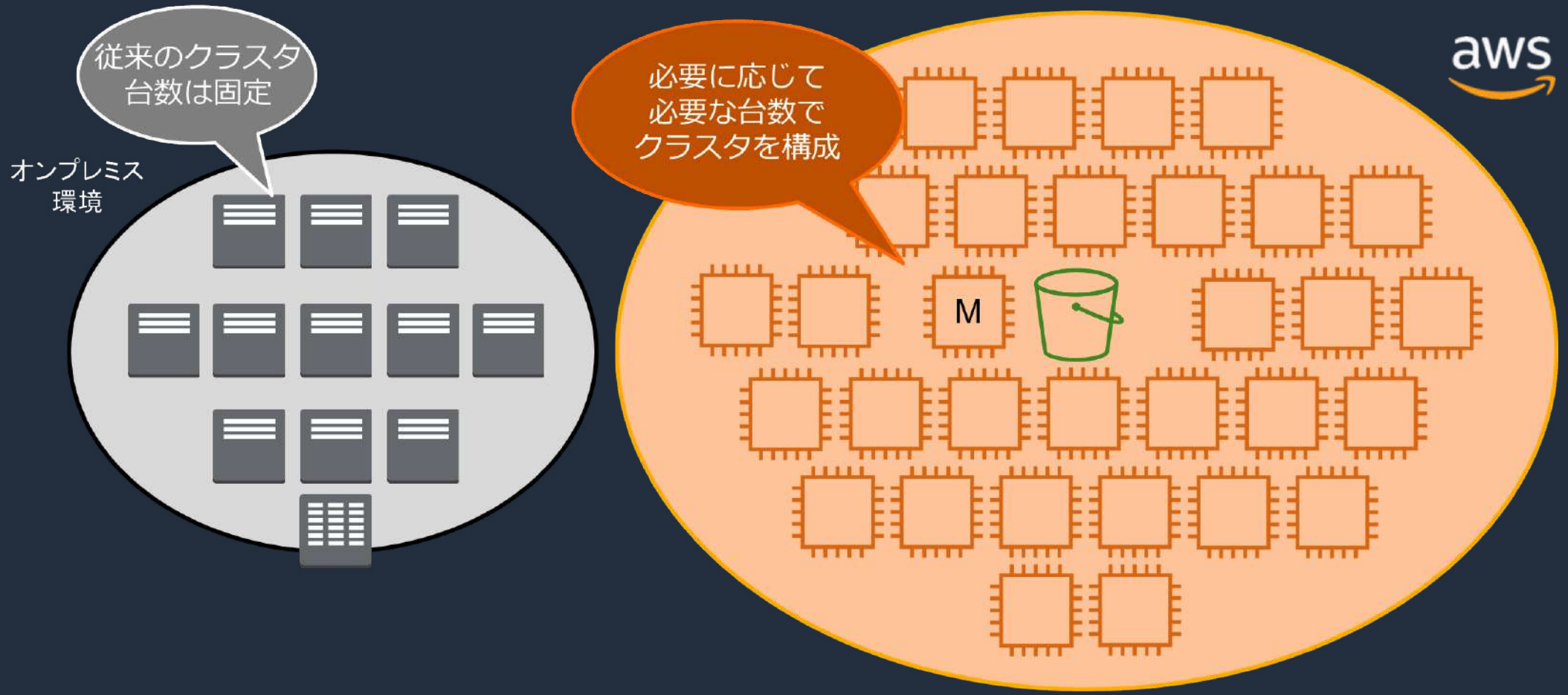
必要な時に必要な計算リソースを活用

スケーラブルなリソースによりコスト効率よく大規模な処理を実行



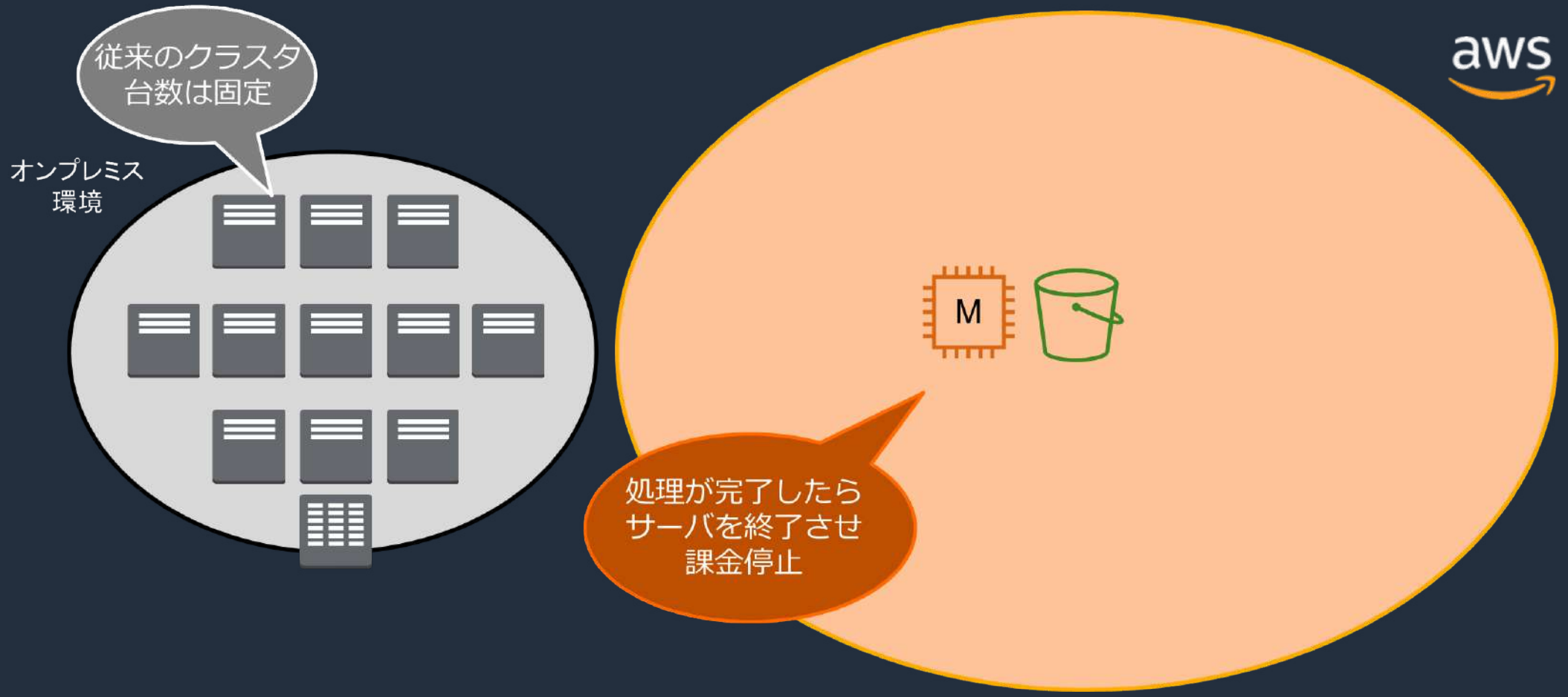
必要な時に必要な計算リソースを活用

スケーラブルなリソースによりコスト効率よく大規模な処理を実行



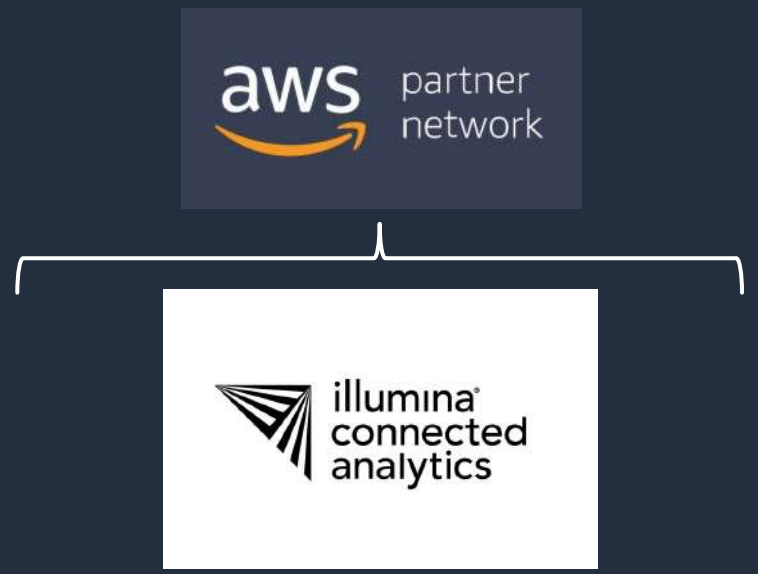
必要な時に必要な計算リソースを活用

この仕組みは1から作ることも可能ですが、AWS サービスを利用することで構築の手間なくすぐにご利用を開始できます



AWS におけるスケーラブルな計算環境

AWS のインフラを抽象化して便利にお使いいただけるサービスから、
Genomics に特化し、さらに便利にお使いいただける、広範なパートナーソリューションまで、
様々な選択肢を活用して AWS 上でのゲノミクスワークロードが実行可能



ジョブスケジューラ (Slurm) を中心とした HPC計算環境

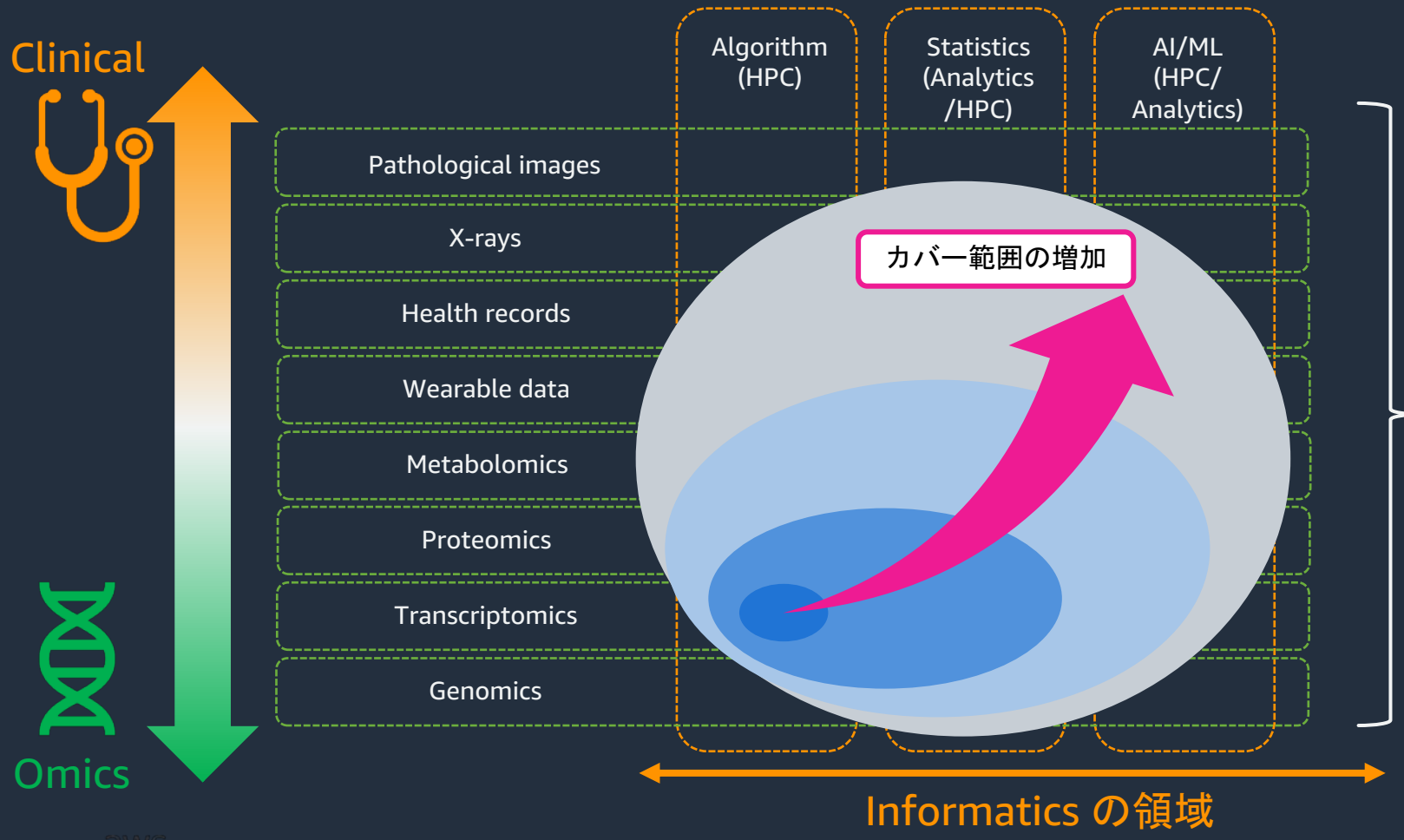
ジョブをコンテナとして定義する AWS独自のフルマネージドジョブスケジューラ

ゲノム解析ワークフロー実行環境を AWS 上で数分で構築する CLI ツール

この後、
11:25 ~「製薬業界におけるゲノム解析・解釈のためのイルミナインフォマティクス」
セッションにて
ご紹介いただきます！

マルチオミクス/マルチモーダル解析へ

クラウドの利用により大量・多種のデータ収集が可能になった



各 Biology の領域で様々なInformatics技術の利用がされてきた

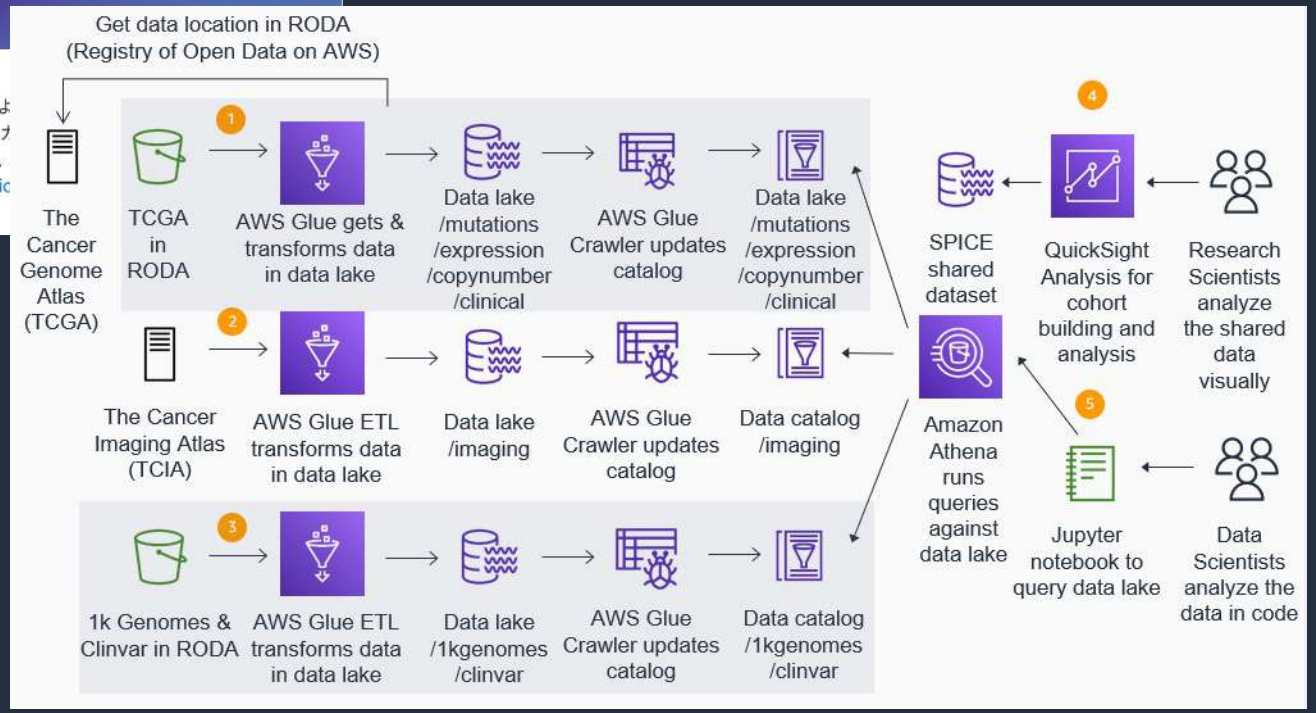
今後は、
マルチオミクス/マルチモーダル解析により、
新たな価値を創出していく

→ 各領域でさえ大量だったデータ量がさらに増加する。
データ前処理が今まで以上に困難に。

→ Jupyter Notebook 上でのPandas等を利用した探索的解析はスケールしない

→ BigData/Analytics 領域で利用されてきた並列分散処理環境の利用で解決

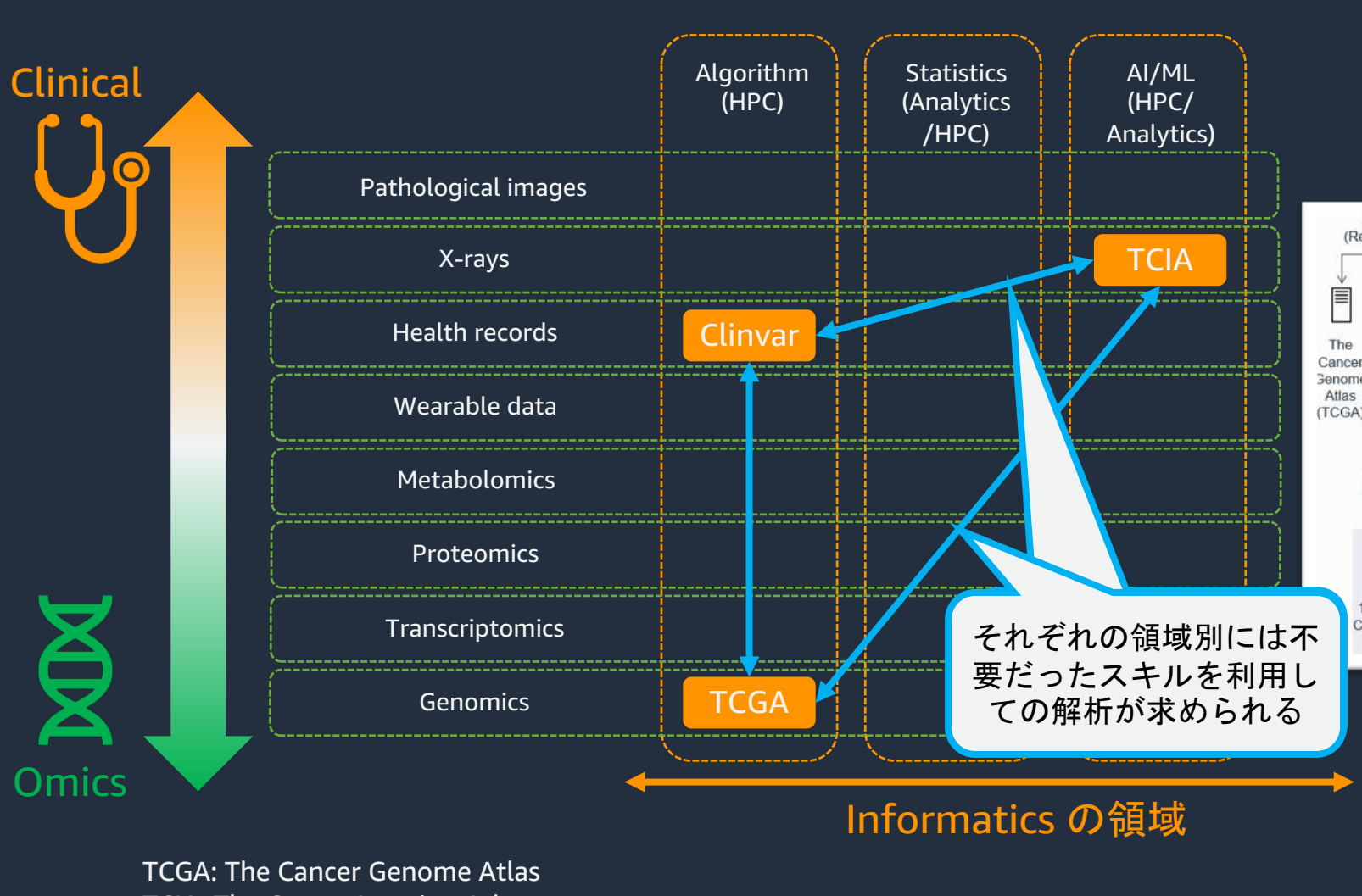
AWS でのマルチオミクスおよびマルチモーダルデータの統合と分析のためのガイダンス



https://aws.amazon.com/jp/solutions/guidance/multi-omics-and-multi-modal-data-integration-and-analysis/#

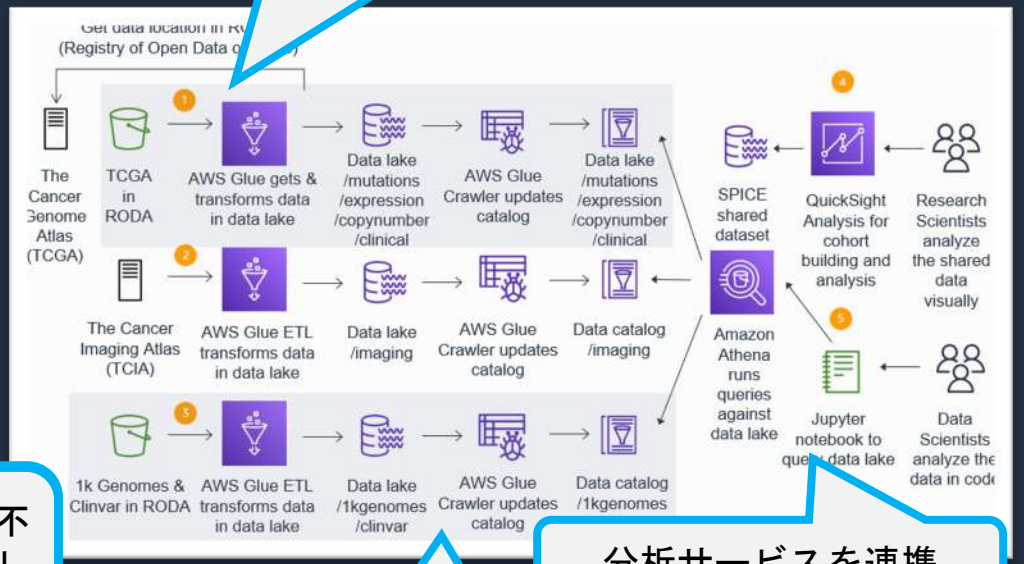
マルチオミクス/マルチモーダル解析へ

大量・多種のデータを収集しデータレイクを構築、分析・機械学習サービスと連携



並列分散処理のためのETL処理

hail VCF, tsv, gtf, bedなどを変換



分析サービスを連携

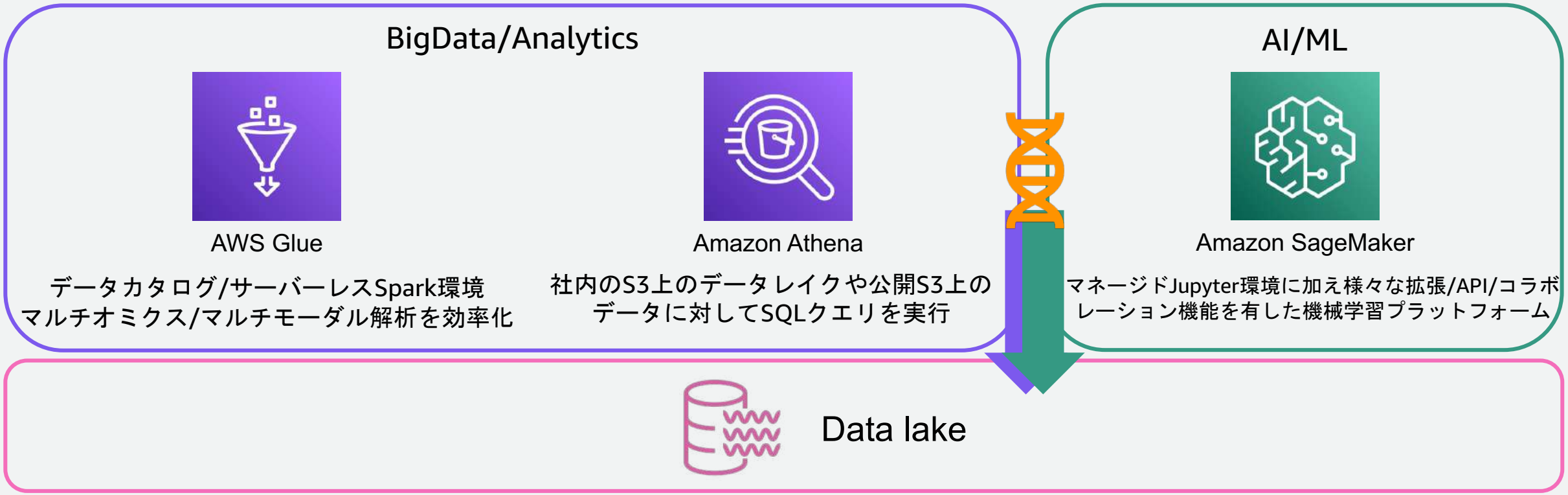
それぞれの領域別には不要だったスキルを利用しての解析が求められる

データレイクを作成
分析サービスを連携させるアーキテクチャ

TCGA: The Cancer Genome Atlas
TCIA: The Cancer Imaging Atlas
Clinvar: ヒトゲノムの変異と疾患の関連性データ

進む Analytics と AI/ML の融合

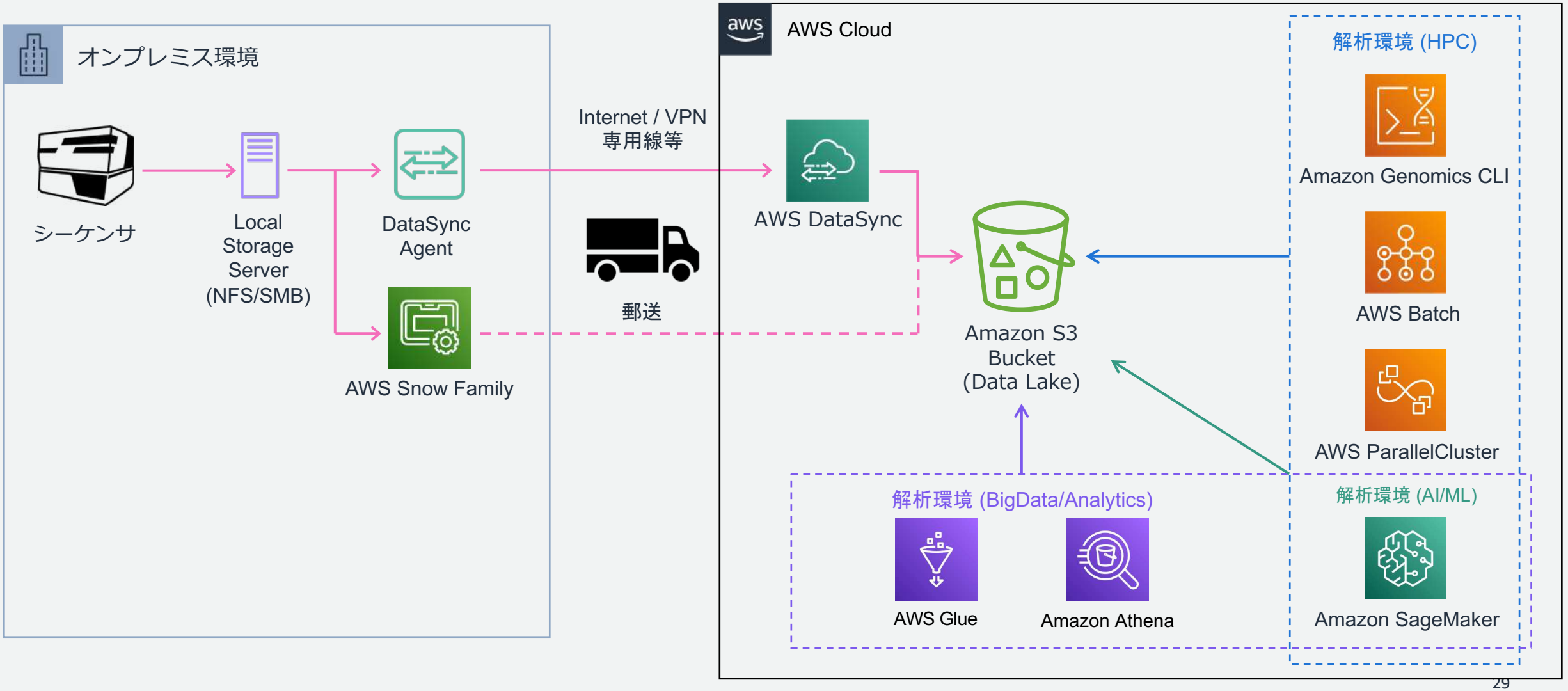
BigData処理のSpark、AI/ML の高度な処理を組み合わせ、複雑なマルチオミクス/マルチモーダル解析を実現



データレイクに収集した大量/多種データに対して、スキルセットによらず高度な解析を手軽に手に入れる

- Jupyter 環境(SageMaker)から、アドホッククエリや Spark によるAI/MLのための訓練データ作成
- SQL から AI/ML モデルの学習(AutoML 機能が利用可能)・推論

再掲: ゲノム情報の「転送・保存・解析」アーキテクチャ



まとめ: 創薬ゲノム領域で AWS を活用しきるために

転送・保存・解析の中でも特に”解析”に重点をおいてご紹介をしました

- シーケンサーから出力される大容量データをどのように転送するか？
  - 効率的なネットワーク転送を実現する AWS DataSync に加え、AWS Snow Family による物理筐体の郵送での転送も可能

- 増え続けるゲノムデータをどのように楽に保存し、価値に繋げるか？
  - 高い耐久性かつ安価なオブジェクトストレージである Amazon S3 を活用し、データレイクを構築
  - Amazon S3 データレイクにあらゆる種類のデータを集積、連携サービスによる解析に繋げる

- 様々な解析を行うための大量かつ多様な計算リソースをどのように確保するか？
  - AWS ParallelCluster/AWS Batch/Amazon Genomisc CLI/Amazon SageMaker 等、
    ワークフローツールやジョブスケジューラなど環境に応じてスケーラブルに計算リソースを確保する方法が提供されている
  - HPC、AI/MLといった領域に加え、今後はBigData/Analytics領域での並列分散処理を適材適所で用いる。各領域の高度なスキルセット習得なしに利用できる機能をフル活用してマルチオミクス/マルチモーダルを実現する

最後に

- AWS では今回紹介したサービス以外にも創薬ゲノム領域に関連する様々なワークロードでお使いいただけるサービスを提供しています。
- お客様の AWS ご利用のために、アーキテクチャレビュー・ハンズオン等の技術支援を行なっています。
- 今回のご紹介内容や創薬領域にとどまらず、様々な業界で得た知見を元にしたご支援も可能です。

お問合せはこちらから ↓
https://aws.amazon.com/jp/contact-us/sales-support/

Thank you!