



# 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。

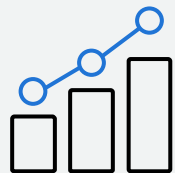
対象

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

- はじめに [ゲノム解析周辺の課題と AWS 上での全体像]
- ゲノム情報の転送
- ゲノム情報の保存
- ゲノム情報の解析
- まとめ

# ゲノムデータ利用の最近



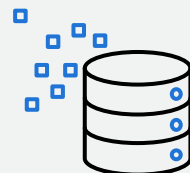
データ容量の  
爆発的増加

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

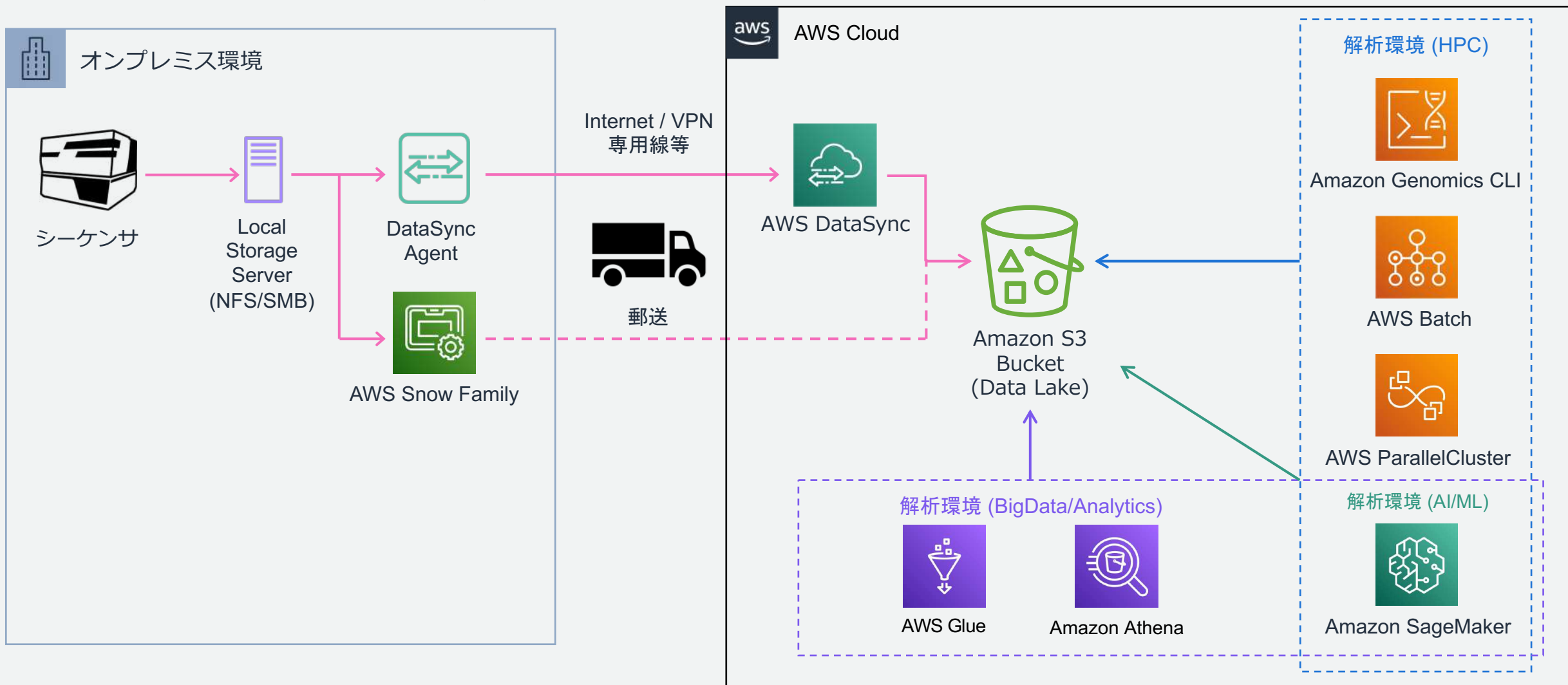
WET研究者  
DRY研究者  
データサイエンティ  
スト



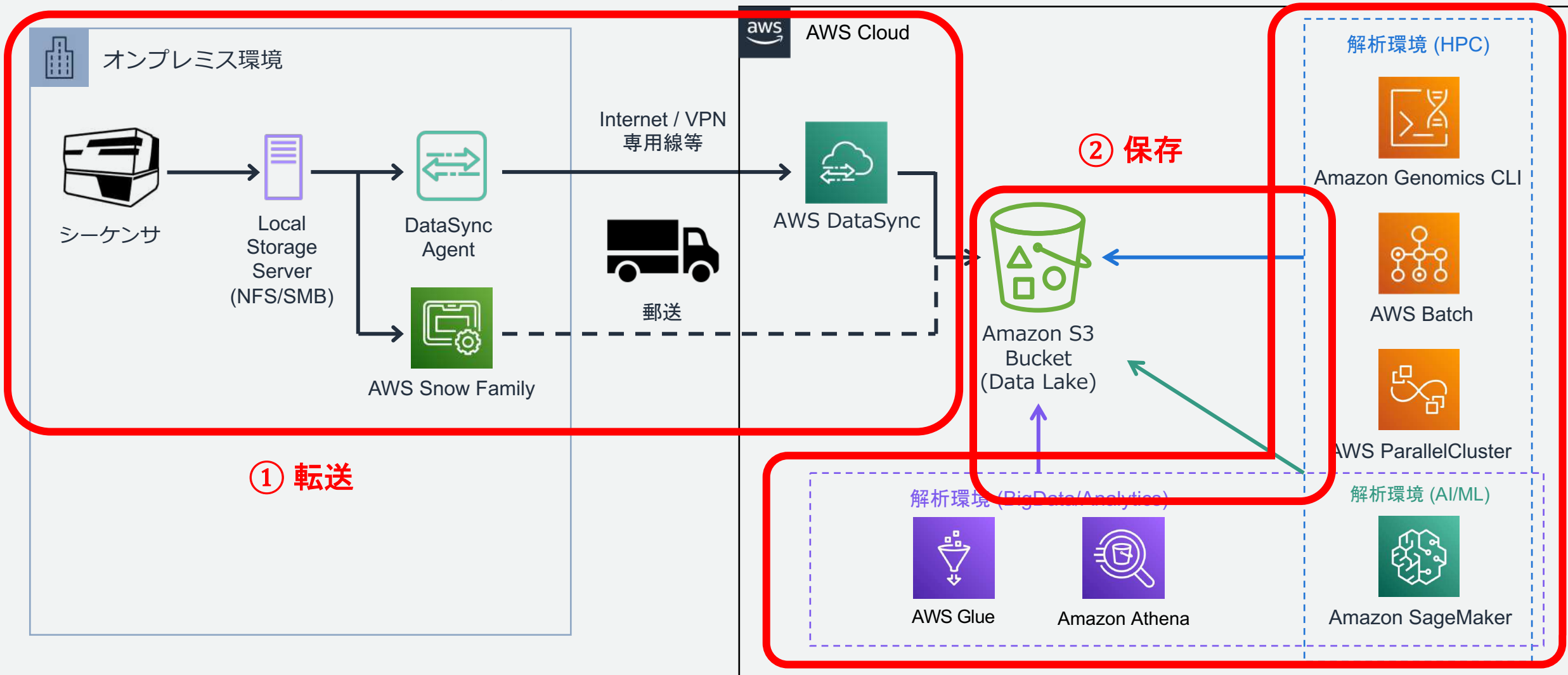
多くのアプリケーシ  
ョンからの分析

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

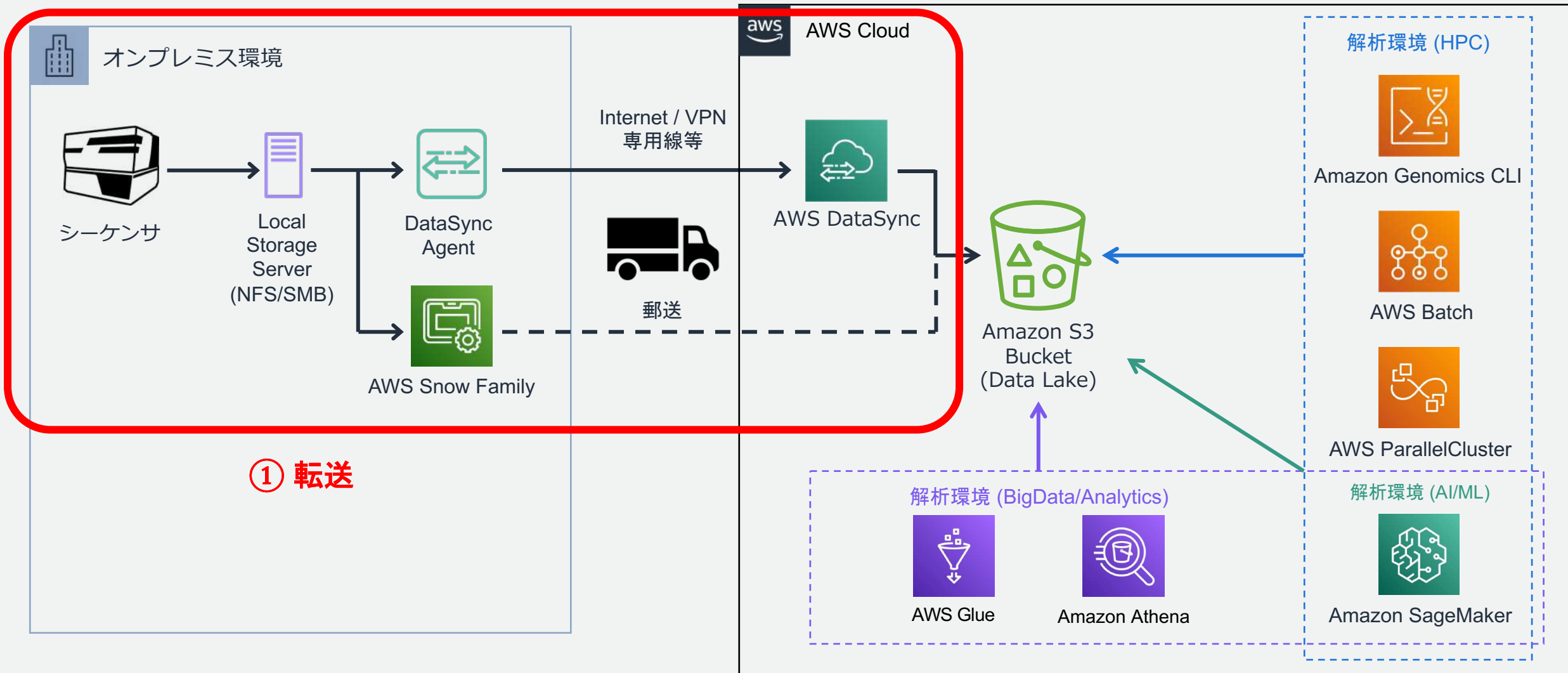


# ゲノム情報の「転送・保存・解析」アーキテクチャ



# ゲノム情報の転送

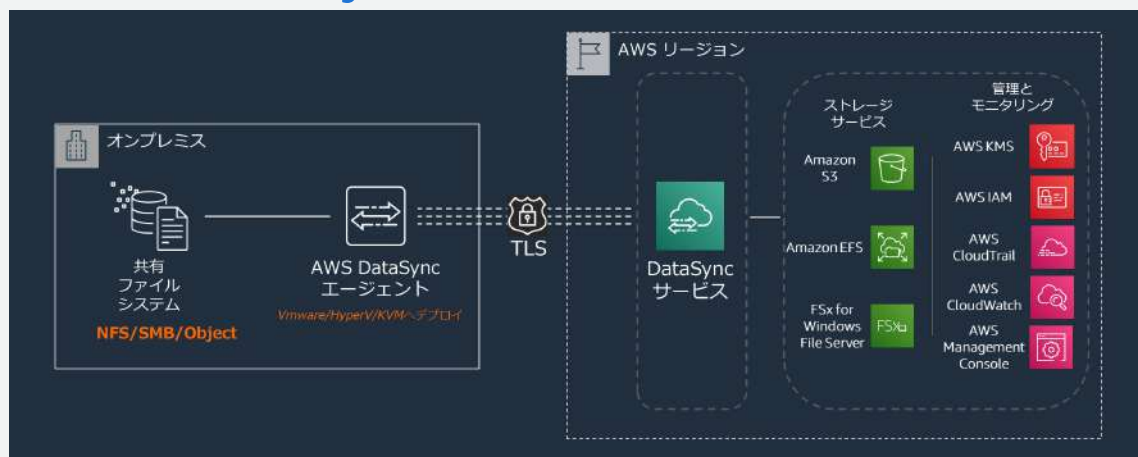
# ゲノム情報の「転送」アーキテクチャ



# 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

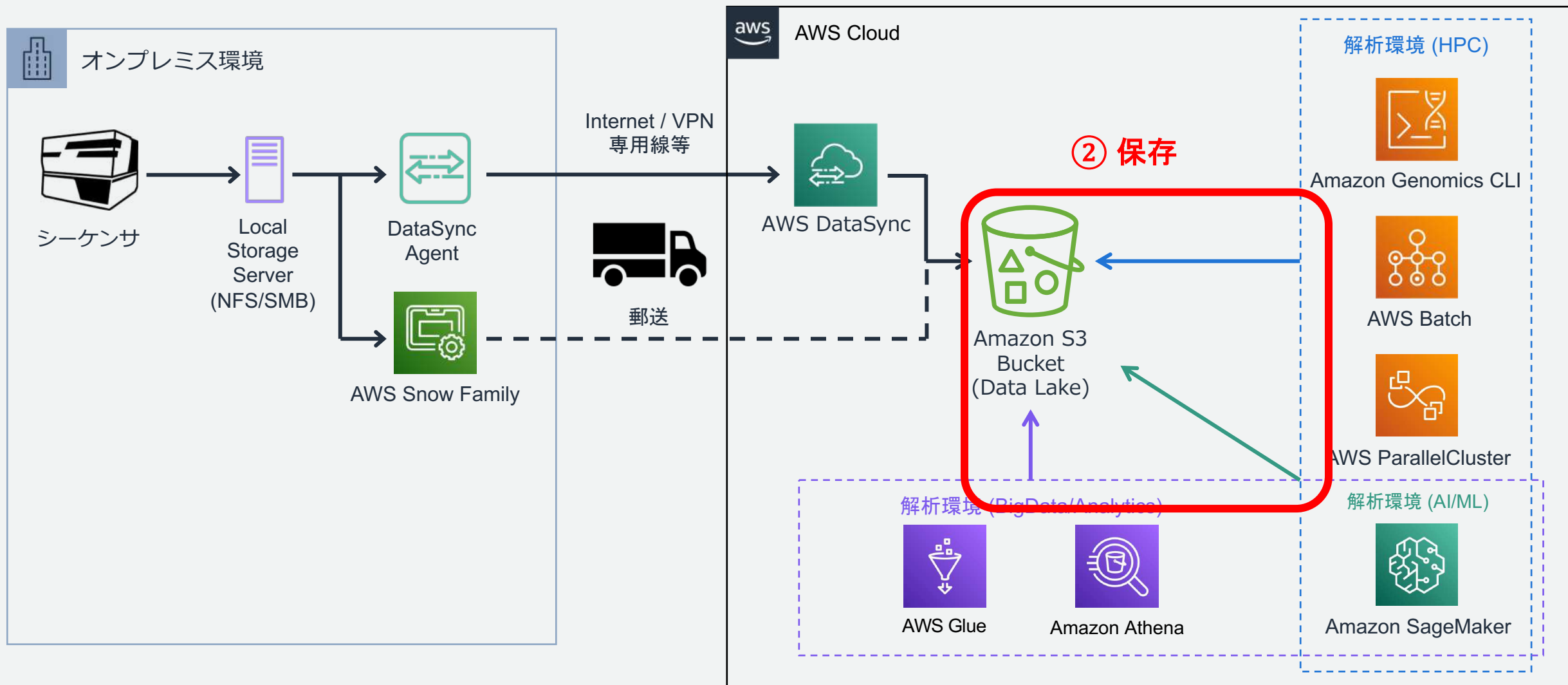


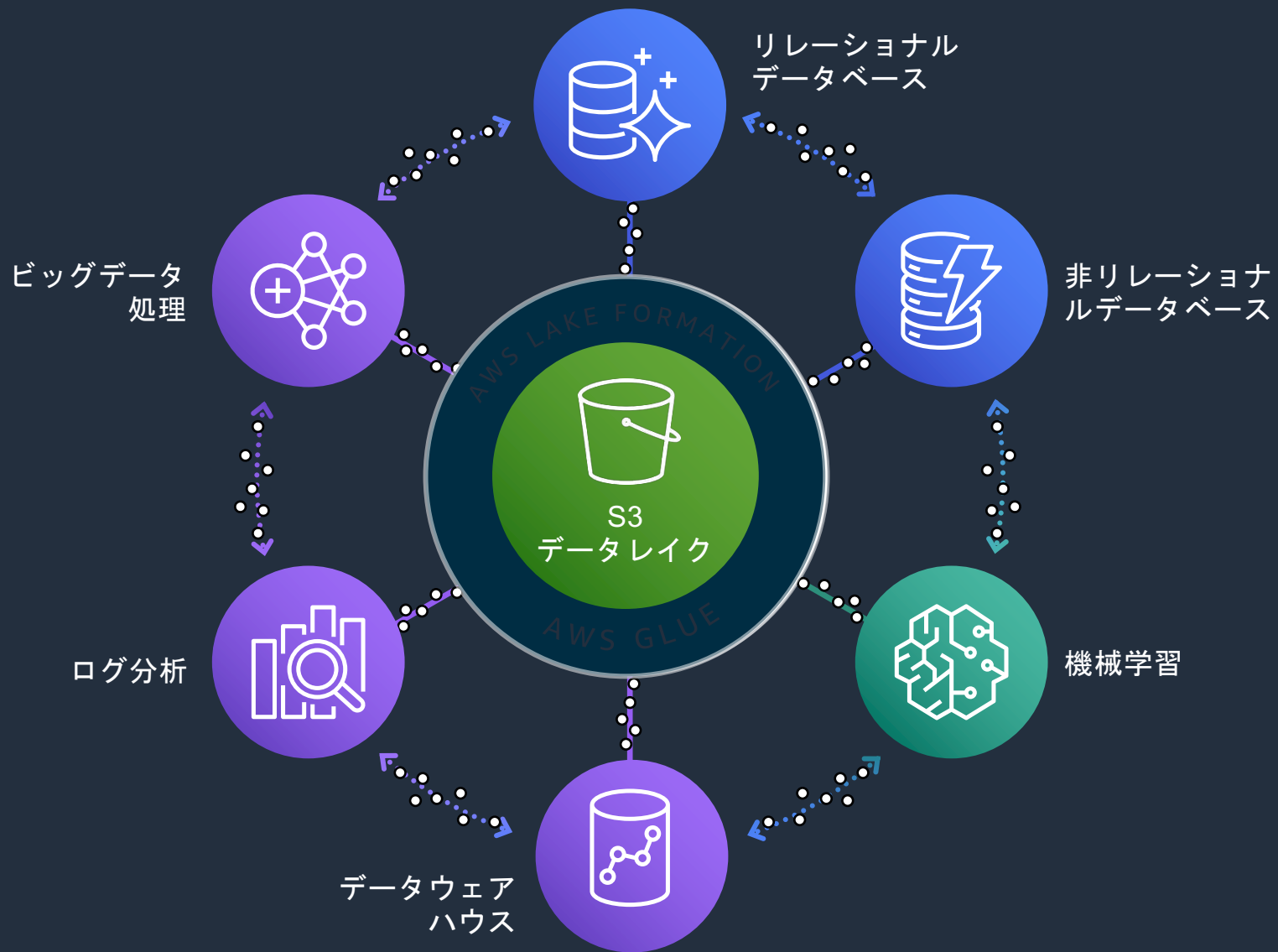
AWS Snowcone

- 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





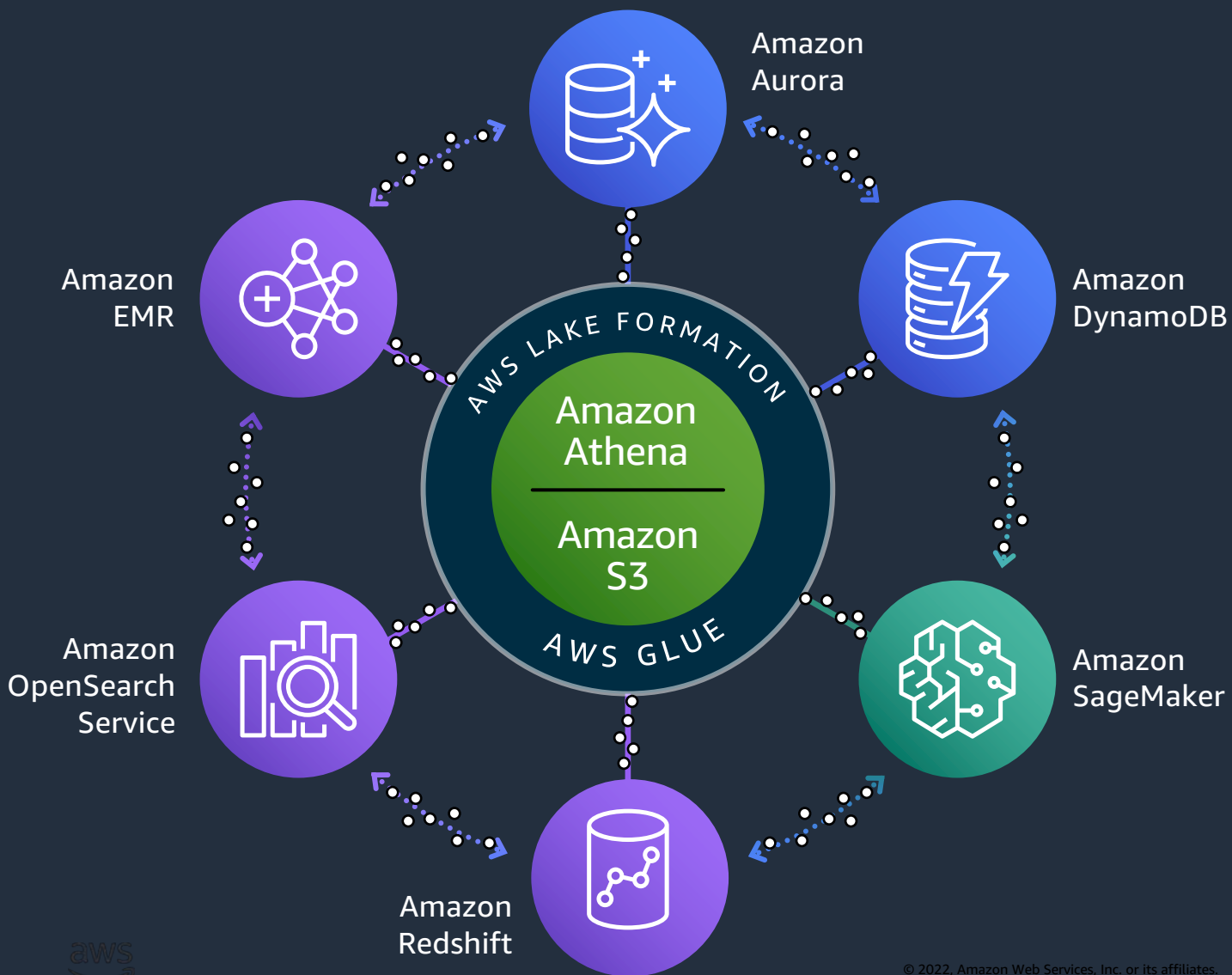
# データレイク アーキテクチャ

その中心に  
Amazon S3 が存在

# Amazon S3 データレイク



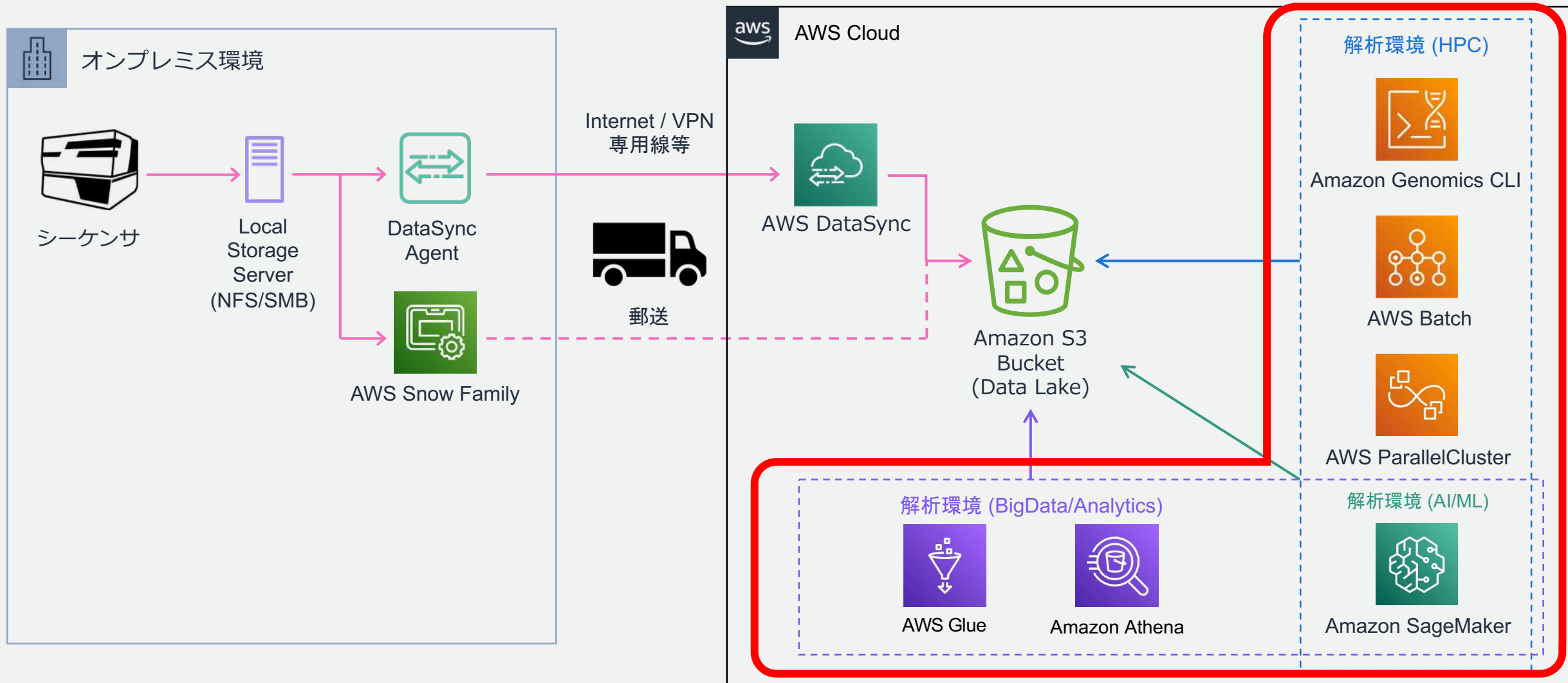
# レイクハウスアプローチと **AWS** サービス



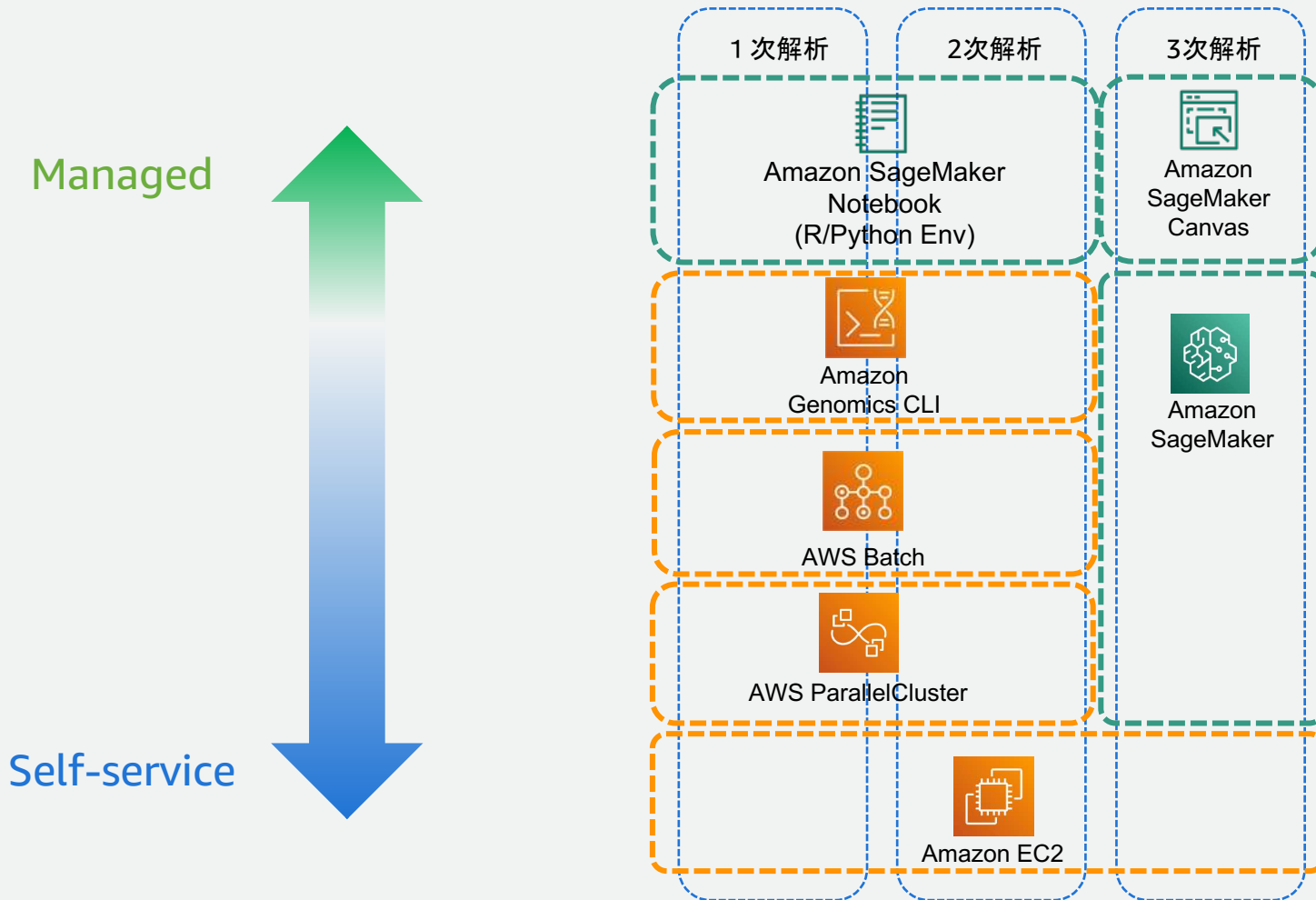
データレイク  
を囲む形で  
様々な AWS サービス  
との連携が可能です

# ゲノム情報の解析

# ゲノム情報の「解析」アーキテクチャ



# 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

アクセラレータの選択肢



Intel Xeon processor  
(x86\_64 arch)

AMD EPYC processor\*  
(x86\_64 arch)

AWS Graviton Processor  
(64-bit Arm arch)

NVIDIA GPU

Xilinx FPGA

### 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



### 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

### C7g インスタンス

64bit Arm Neoverse V1ベース

AWS Graviton3 CPU 搭載

### 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

### 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

The screenshot shows the AWS Marketplace product page for the DRAGEN Complete Suite. At the top, there is a search bar and navigation links for 'Sign in' and 'Create a new account'. Below the navigation, there are tabs for 'Categories', 'Delivery Methods', 'Solutions', 'AWS IQ', 'Resources', and 'Your Saved List'. The product title 'DRAGEN Complete Suite' is prominently displayed, along with the provider 'Illumina Inc.' and the latest version '3.8.4'. A description states that the suite enables ultra-rapid analysis of NGS data. Pricing information shows a typical total price of \$18.40/hr. There is a 'Free Trial' button and a 'Continue to Subscribe' button. A 'Save to List' button is also visible.

## Product Overview

The DRAGEN Complete Suite\* enables ultra-rapid analysis of Next Generation Sequencing (NGS) data for large data sets, such as whole genomes, exomes, and genes/panels. This application uses the DRAGEN Platform and includes highly-optimized algorithms for mapping, aligning, sorting, duplicate marking, and haplotype variant calling. The DRAGEN CS includes a host of pipelines including our DRAGEN Germline Pipeline, DRAGEN Somatic Pipeline (T and T/N), DRAGEN Copy Number Variant (CNV) Pipeline, DRAGEN RNA Gene Fusion, DRAGEN Joint Genotyping Pipeline, and GATK Best Practices. The DRAGEN Germline and Somatic pipelines have greatly improved accuracy in calling SNPs and Indels compared to industry standard. This app also supports Illumina NovaSeq BCL conversion, download/upload of data streaming, and compressed reference hash tables for a more seamless and efficient workflow.

Note: DRAGEN license metering is on an hourly basis.

Version	3.8.4
	<a href="#">Show other versions</a>
By	<a href="#">Illumina Inc.</a>

The screenshot shows the AWS Marketplace product page for NVIDIA Clara Parabricks Pipelines. The layout is similar to the DRAGEN page, with a search bar and navigation links. The product title 'NVIDIA Clara Parabricks Pipelines' is displayed, along with the provider 'Ingram Micro' and the latest version 'Parabricks Pipelines v3.5.0'. A description states that the pipelines enable GPU-accelerated analysis of DNA and RNA based applications. Pricing information shows a typical total price of \$4.212/hr. There is a 'Continue to Subscribe' button and a 'Save to List' button.

## Product Overview

Clara Parabricks Pipelines enable GPU-accelerated analysis of DNA and RNA based applications, starting with a FASTQ file and generating a vcf or gvcf.

Version	Parabricks Pipelines v3.5.0
By	<a href="#">Ingram Micro</a>
Video	<a href="#">See Product Video</a>
Categories	<a href="#">High Performance Computing</a> <a href="#">Healthcare &amp; Life Sciences</a>
Operating System	Linux/Unix, Ubuntu 18.04
Delivery Methods	<a href="#">Amazon Machine Image</a>

The screenshot shows the AWS Marketplace product page for the Sentieon Genomics Suite. The layout is consistent with the other product pages, featuring a search bar and navigation links. The product title 'Sentieon Genomics Suite' is displayed, along with the provider 'Sentieon Inc.' and the latest version '201911.01'. A description states that the tools are computationally efficient and award-winning. Pricing information shows a typical total price of \$1.728/hr. There are 'BYOL' and 'Free Tier' options. A 'Continue to Subscribe' button and a 'Save to List' button are present.

## Product Overview

Sentieon (<http://www.sentieon.com>) supplies award-winning software tools for secondary analysis of NGS data.

Sentieon DNaseq and TNseq produce results identical to the Broad Institute's BWA-GATK HaplotypeCaller/MuTect2 Best Practice Workflow by implementing the same mathematics but with more efficient computing algorithms and enterprise-strength software implementation. Furthermore, the Sentieon tools do not downsample in high-coverage regions, are able to handle arbitrary depth of coverage, and have no thread dependency, resulting in 100% consistency.

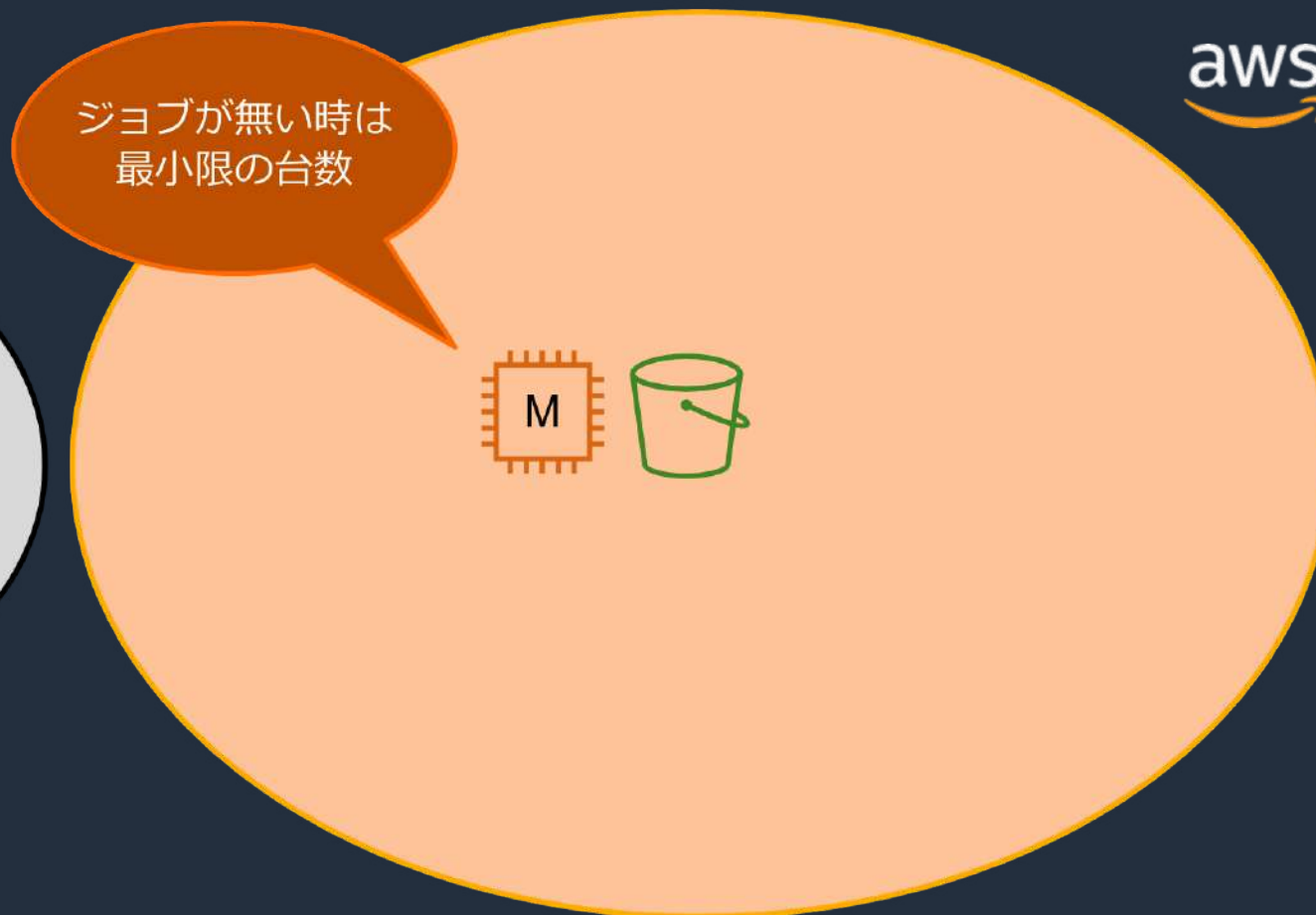
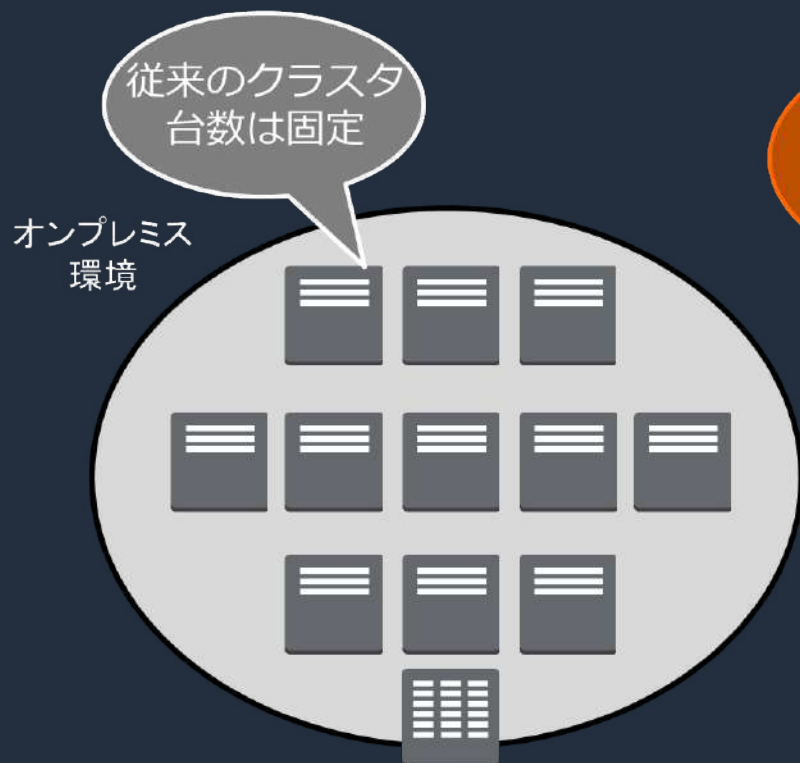
Sentieon DNAscope and TNscope build upon and improve over the mathematical models from Haplotyper/MuTect2, providing additional accuracy and supporting the use of Machine Learning models for filtering.

The Sentieon tools are enterprise-strength software tools that are inherently multi-threading and distributed-processing ready, and allow efficient processing and joint calling of extremely large cohorts. They are easily deployable, easily scalable, and easily upgradable software-only solutions running on AWS CPU instances. Typical runtime for a single sample: DNaseq will process a 100X WES sample in 15 mins, and a 30X WGS in 3 h on a c5d.8xlarge instance.



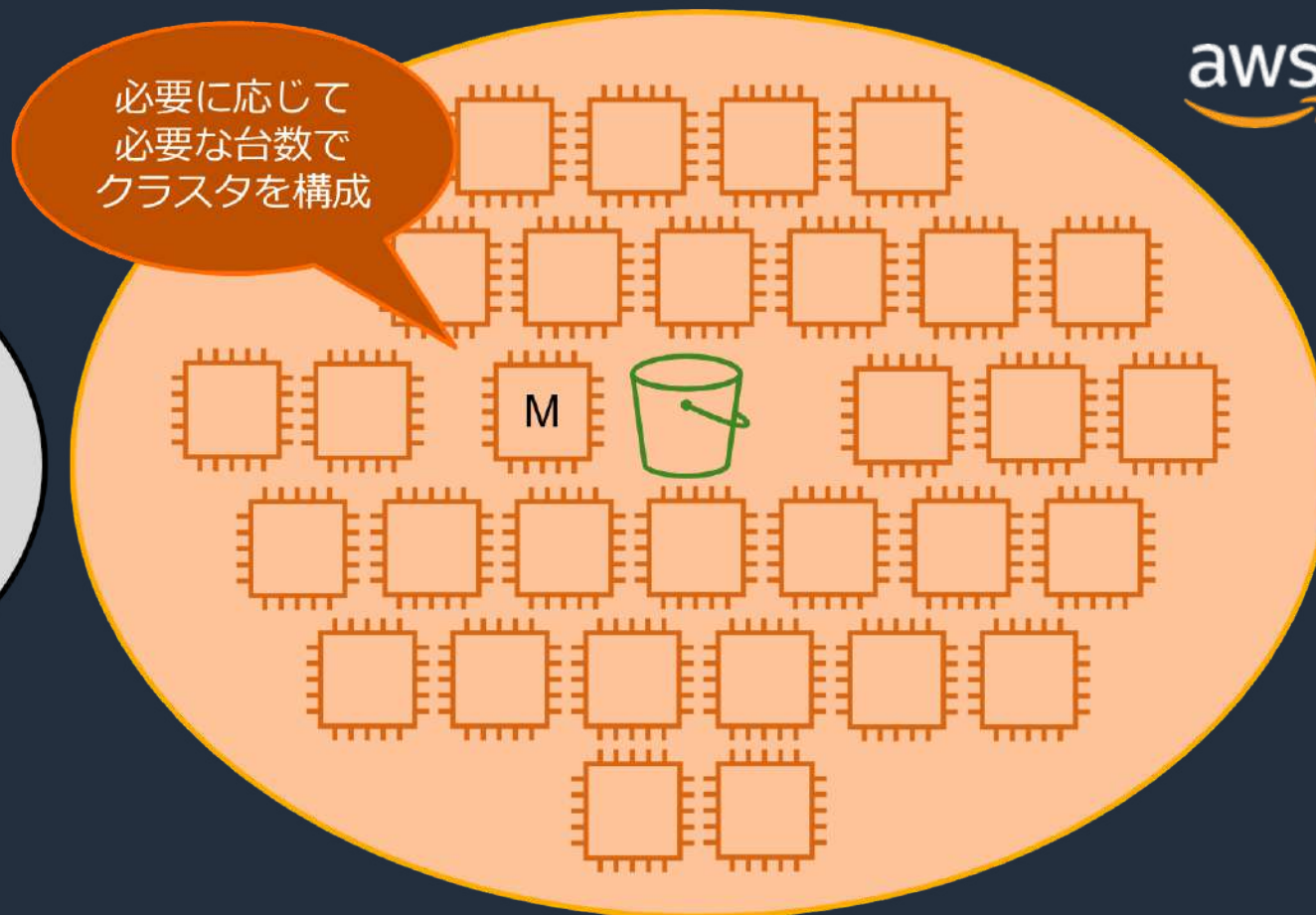
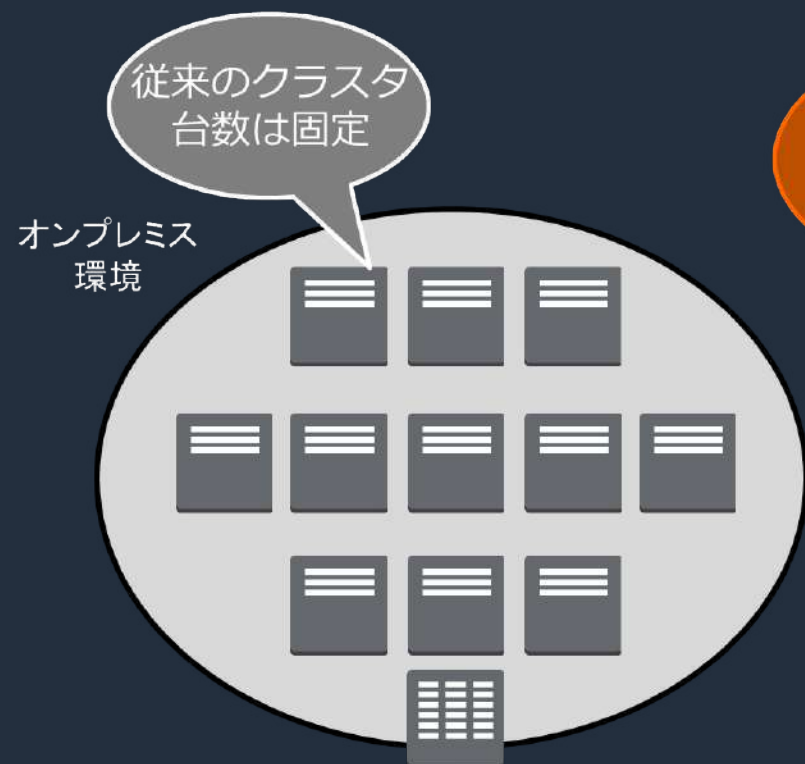
# 必要な時に必要な計算リソースを活用

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



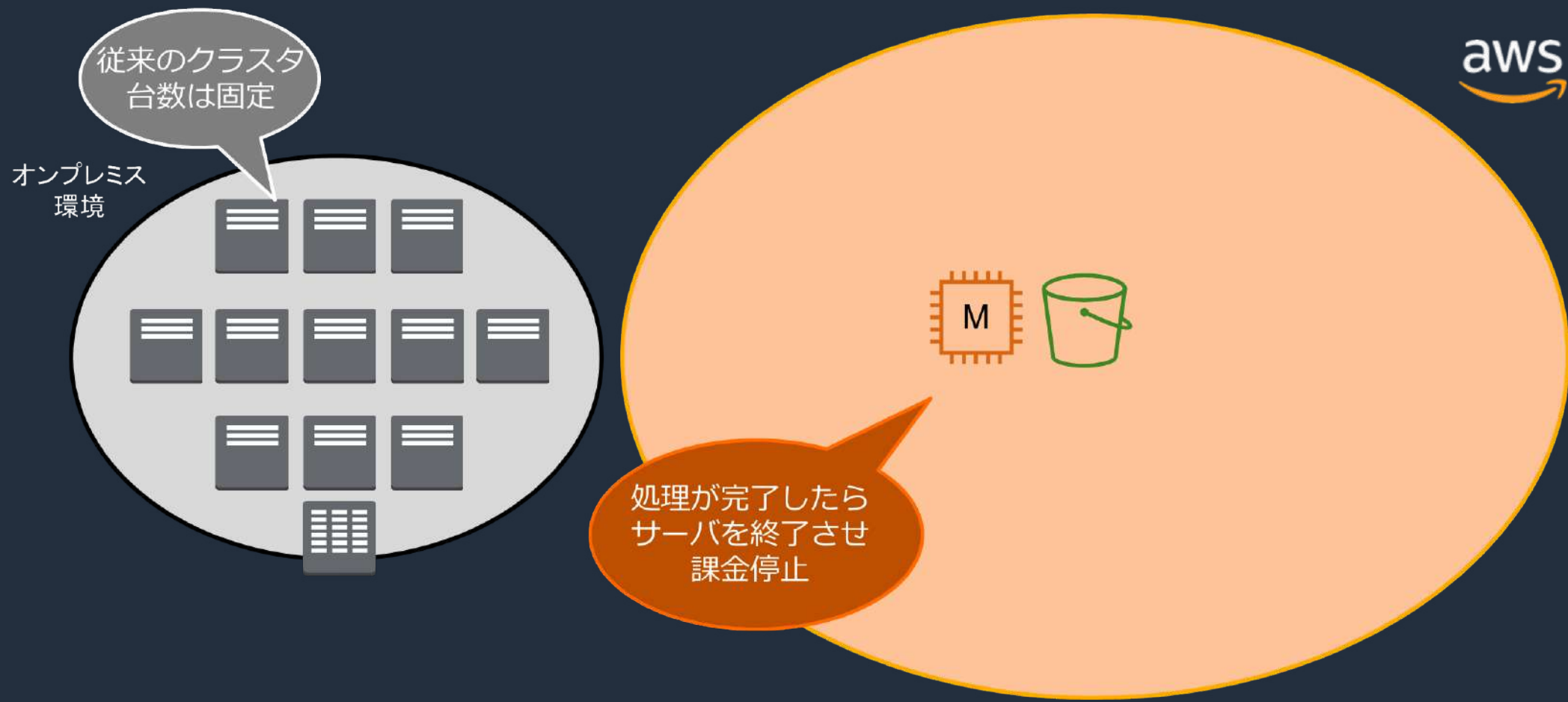
# 必要な時に必要な計算リソースを活用

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



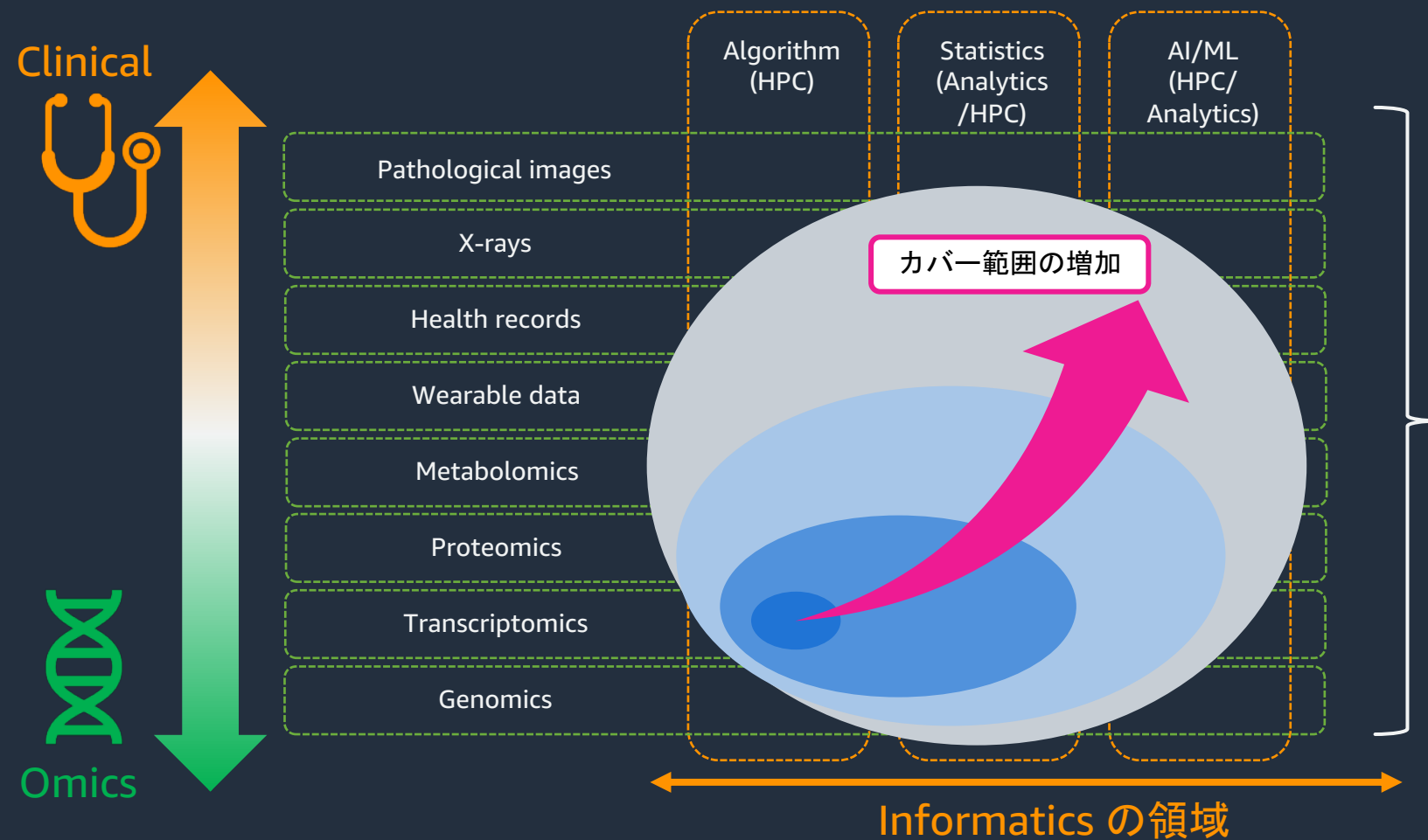
# AWS におけるスケーラブルな計算環境

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 上でのゲノムワークロードが実行可能



# 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

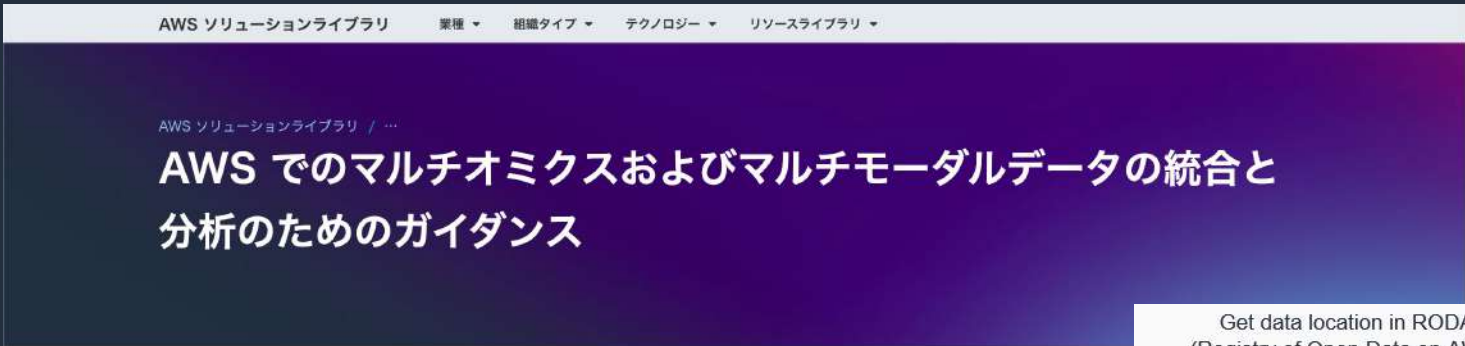


各 Biology の領域で様々な Informatics 技術の利用がされてきた

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

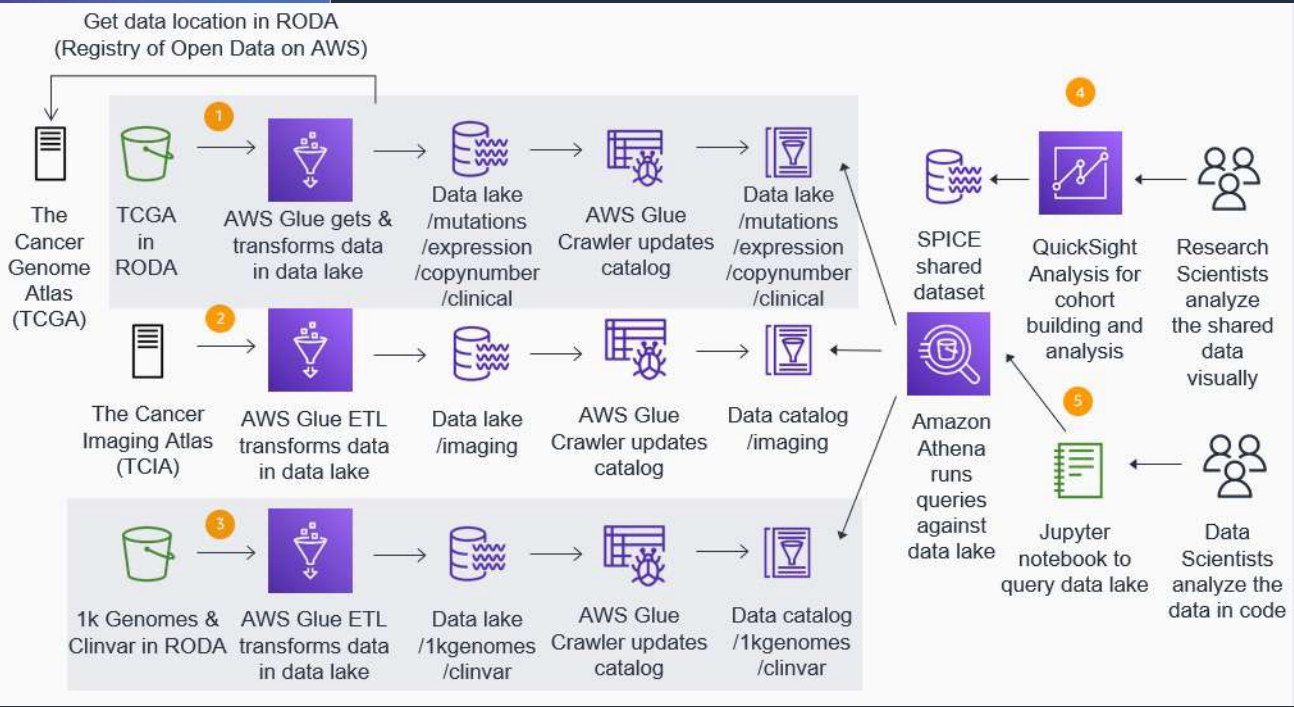
# AWS でのマルチオミクスおよびマルチモーダルデータの統合と分析のためのガイダンス



このページをナビゲートする

- アーキテクチャ図
- Well-Architected Pillars
- その他の考慮事項

このガイダンスは、ユーザーが大規模な分析のためにゲノム、臨床、変異、発現、および、データレイクに対してインタラクティブなクエリを実行するために役立ちます。Infrastructure as Code の自動化、データを変換するための取り込みパイプライン、分析のためのノートブックとダッシュボードなどを説明します。このガイダンスは Bic 構築されました。

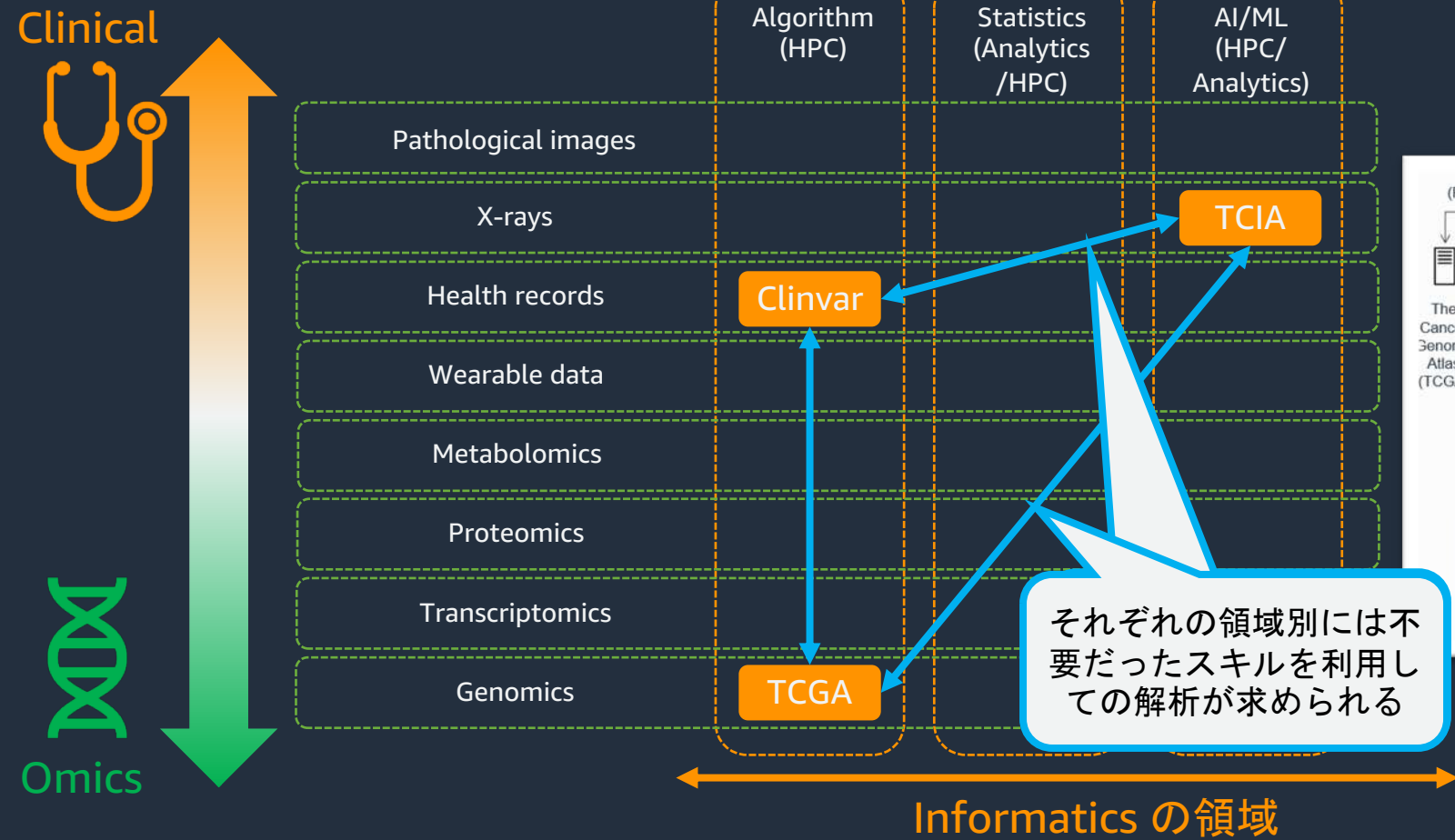


<https://aws.amazon.com/jp/solutions/guidance/multi-omics-and-multi-modal-data-integration-and-analysis/#>

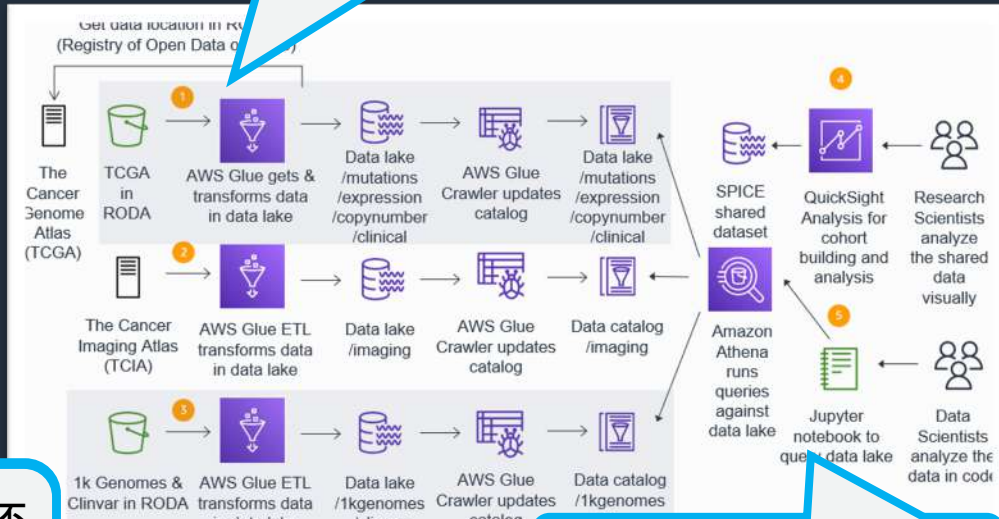


# マルチオミクス/マルチモーダル解析へ

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  
 を変換



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

分析サービスを連携

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

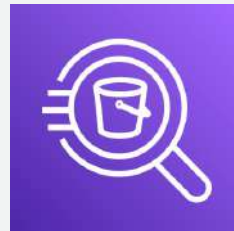
BigData処理のSpark、AI/ML の高度な処理を組み合わせ、複雑なマルチオミクス/マルチモーダル解析を実現

## BigData/Analytics



AWS Glue

データカタログ/サーバーレスSpark環境  
マルチオミクス/マルチモーダル解析を効率化



Amazon Athena

社内のS3上のデータレイクや公開S3上の  
データに対してSQLクエリを実行

## AI/ML



Amazon SageMaker

マネージドJupyter環境に加え様々な拡張/API/コラボ  
レーション機能を有した機械学習プラットフォーム

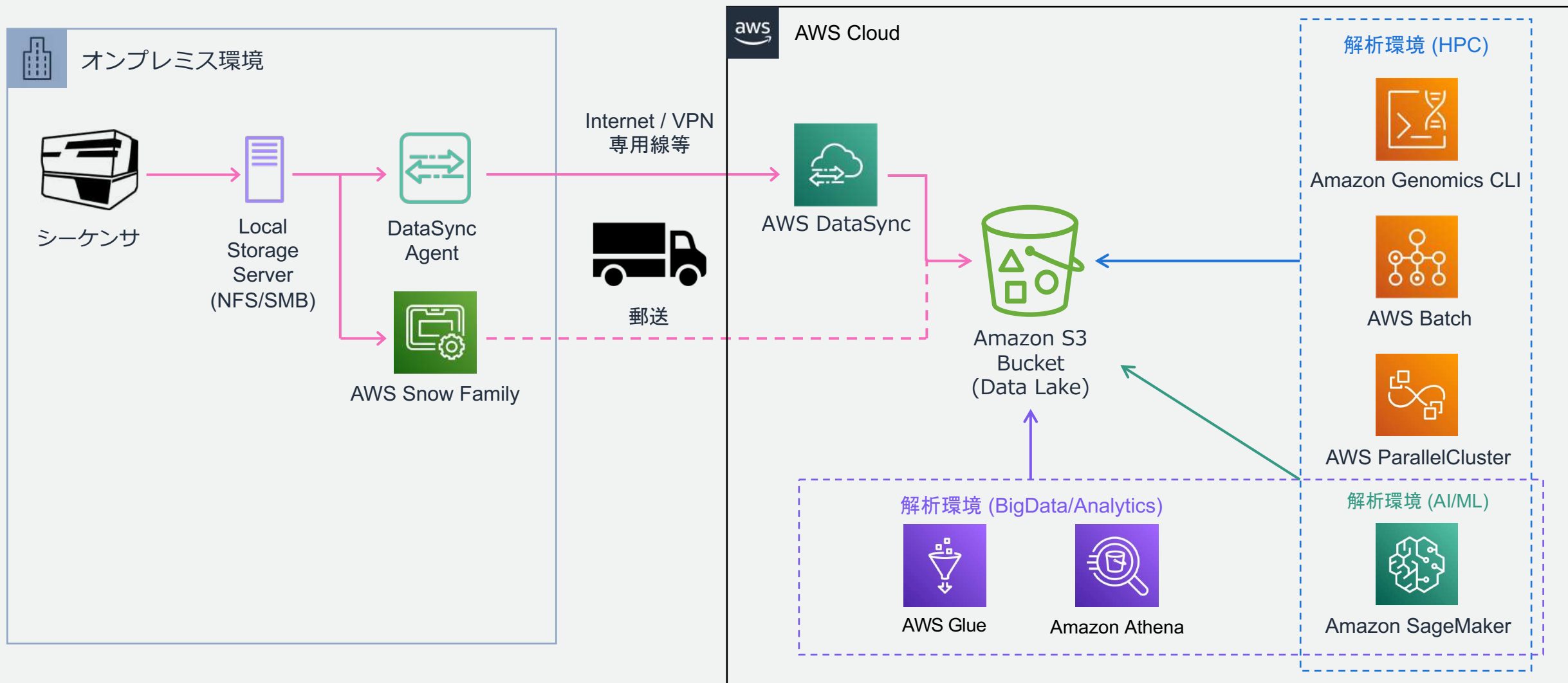


Data lake

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 Genomics 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。

お問い合わせはこちらから ↓

<https://aws.amazon.com/jp/contact-us/sales-support/>





**Thank you!**