



このコンテンツは公開から3年以上経過しており内容が古い可能性があります
最新情報については[サービス別資料](#)もしくはサービスのドキュメントをご確認ください

[AWS Black Belt Online Seminar] Next Generation Redshift

サービスカットシリーズ

Solutions Architect 大藪 純平
2020/2/18

AWS 公式 Webinar
<https://amzn.to/JPWebinar>



過去資料
<https://amzn.to/JPArchive>



自己紹介

大藪 純平 (おおその じゅんぺい)

 @jostandard

アマゾン ウェブ サービス ジャパン
ソリューションアーキテクト

好きなサービス :

Amazon Redshift をはじめとした
Analytics サービス



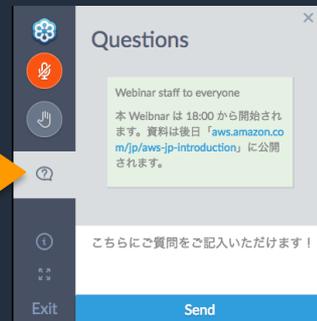
AWS Black Belt Online Seminar とは

「サービス別」「ソリューション別」「業種別」のそれぞれのテーマに分かれて、アマゾンウェブ サービス ジャパン株式会社が主催するオンラインセミナーシリーズです。

質問を投げることができます！

- 書き込んだ質問は、主催者にしか見えません
- 今後のロードマップに関するご質問は
お答えできませんのでご了承下さい

- ① 吹き出しをクリック
- ② 質問を入力
- ③ Sendをクリック



Twitter ハッシュタグは以下をご利用ください
#awsblackbelt

内容についての注意点

- 本資料では2020年2月18日時点のサービス内容および価格についてご説明しています。最新の情報はAWS公式ウェブサイト(<http://aws.amazon.com>)にてご確認ください。
- 資料作成には十分注意しておりますが、資料内の価格とAWS公式ウェブサイト記載の価格に相違があった場合、AWS公式ウェブサイトの価格を優先とさせていただきます。
- 価格は税抜表記となっております。日本居住者のお客様には別途消費税をご請求させていただきます。
- AWS does not offer binding price quotes. AWS pricing is publicly available and is subject to change in accordance with the AWS Customer Agreement available at <http://aws.amazon.com/agreement/>. Any pricing information included in this document is provided only as an estimate of usage charges for AWS services based on certain information that you have provided. Monthly charges will be based on your actual use of AWS services, and may vary from the estimates provided.

Next Generation Redshift アジェンダ

- サービス概要
- アーキテクチャの進化の歴史
- 次世代アーキテクチャ RA3/AQUA
- 新機能アップデート
- まとめ

Amazon Redshift サービス概要

AWS 公式 Webinar
<https://amzn.to/JPWebinar>



過去資料
<https://amzn.to/JPArchive>



Amazon Redshift とは？

高速、スケーラブルで費用対効果の高い
データウェアハウスおよび
データレイク分析マネージドサービス

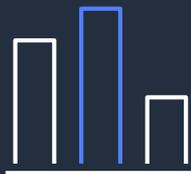


Amazon Redshift

Amazon Redshift の特長



データレイク &
AWS サービスとの親和性



ハイパフォーマンス



高いコスト効果



拡張性 & 柔軟性



セキュリティ &
コンプライアンス



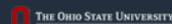
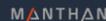
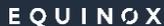
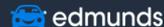
フルマネージド

数万社のお客様が Amazon Redshift を利用しています



Amazon
Redshift

多くのお客様に活用されている
クラウドデータウェアハウス



Amazon Redshift のユースケース

主に大容量データを高速に集計・分析する必要があるワークロードに活用



経営ダッシュボード



定型レポートニング



アドホック分析



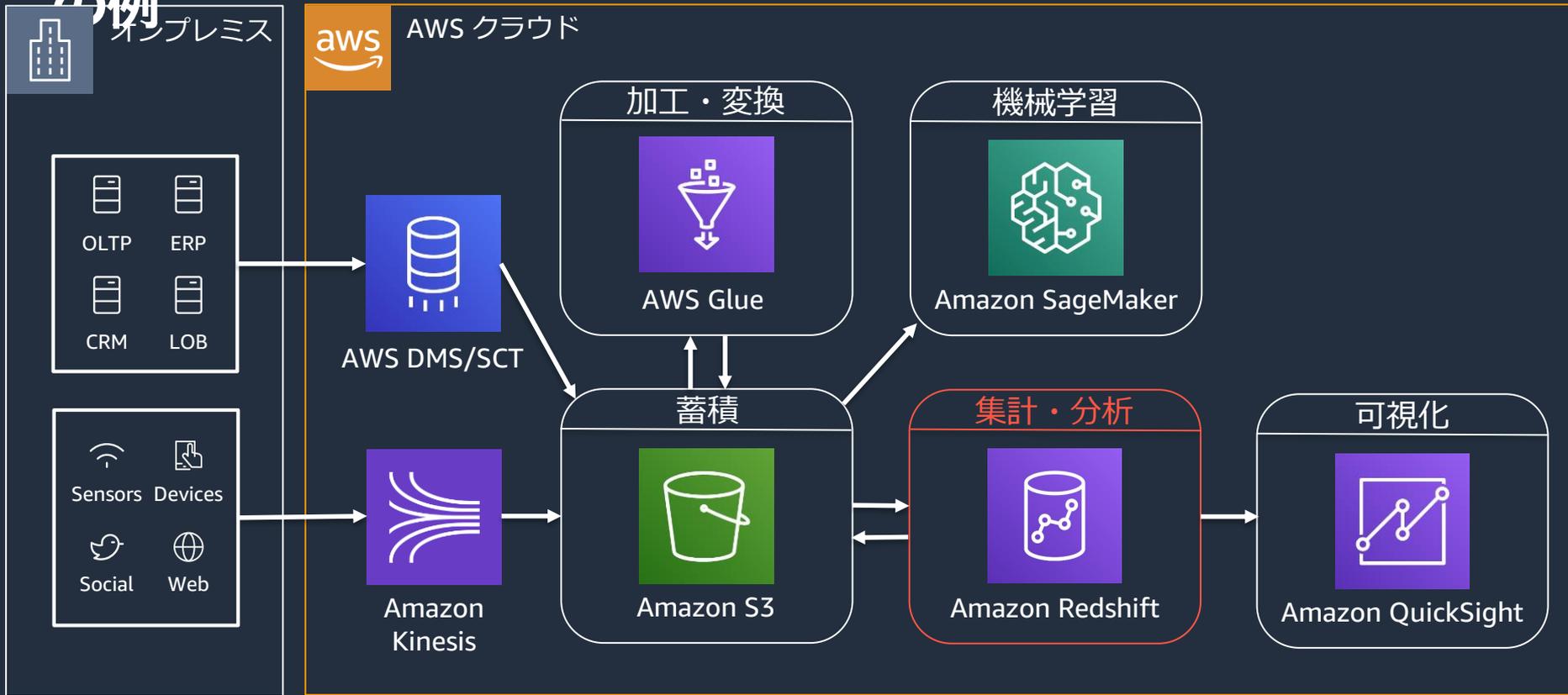
ETL/バッチ



機械学習の前処理

Amazon Redshift を中心としたデータ分析パイプライン

の例



Amazon Redshift アーキテクチャの進化の歴史

AWS 公式 Webinar

<https://amzn.to/JPWebinar>



過去資料

<https://amzn.to/JPArchive>



Amazon Redshift アーキテクチャの進化の歴史





AWS re:Invent

2012



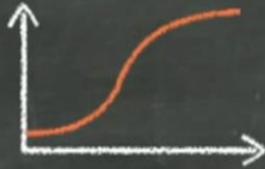
Andy Jassy
Senior Vice President, Amazon Web Services

What Would Data Warehousing Look Like If We Did It the AWS Way

Easy to provision and
scale up massively



No upfront costs,
pay as you go



Really fast performance at
a really low price



Open & flexible with
support for popular tools





AWS re:Invent 2012



Andy Jassy
Senior Vice President, Amazon Web Services

Introducing Amazon Redshift Data Warehousing the AWS Way



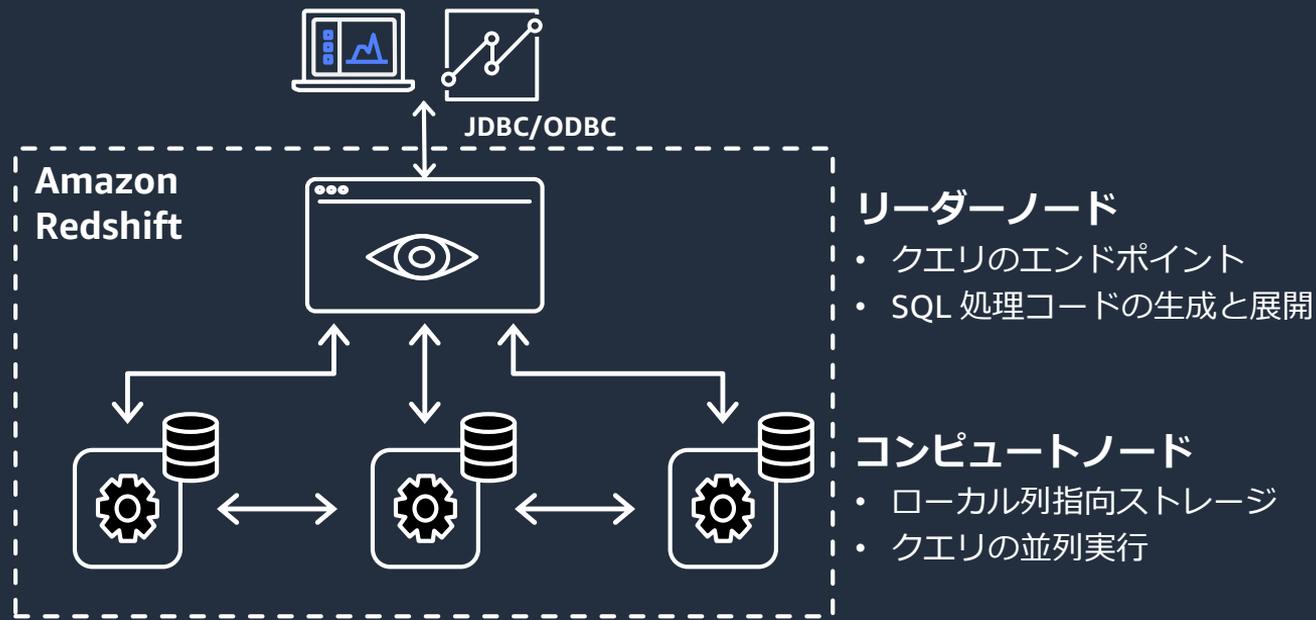
Easily and rapidly analyze petabytes of data

1/10 the cost of traditional data warehouses

Automated deployment & administration

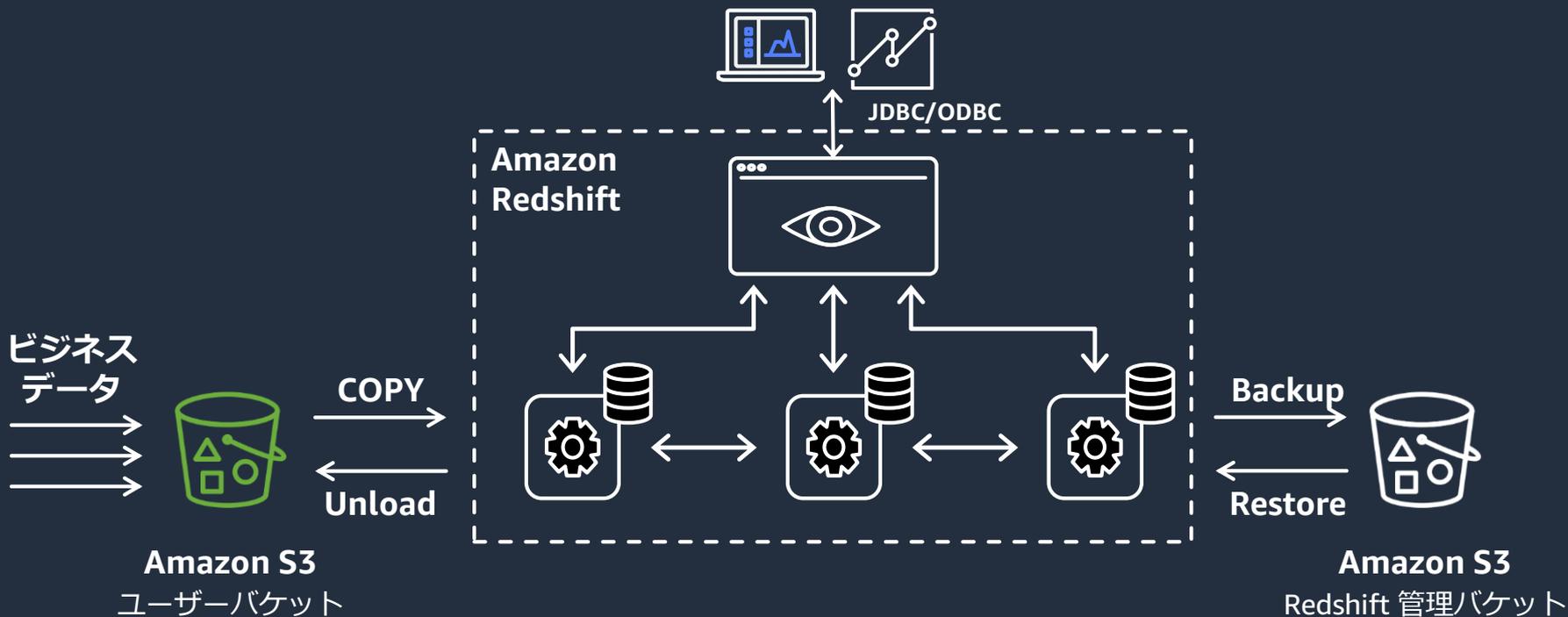
Compatible with popular BI tools

Amazon Redshift の基本アーキテクチャ



シェアードナッシング + MPP (Massively Parallel Processing) アーキテクチャ
データを複数のコンピューターノードに分散させ、各ノードで並列処理することで分析クエリに対する処理スループットを向上させるアプローチ

Amazon Redshift の基本アーキテクチャ



データはユーザー管理の S3 を経由してロード & アンロード
自動バックアップ & リストアは Redshift サービスで管理する S3 領域を活用

Amazon Redshift アーキテクチャの進化の歴史

Redshift (発表)

新しいインスタンス
DC1, DS2, DC2 (GA)

2012

2017

Amazon Redshift の料金

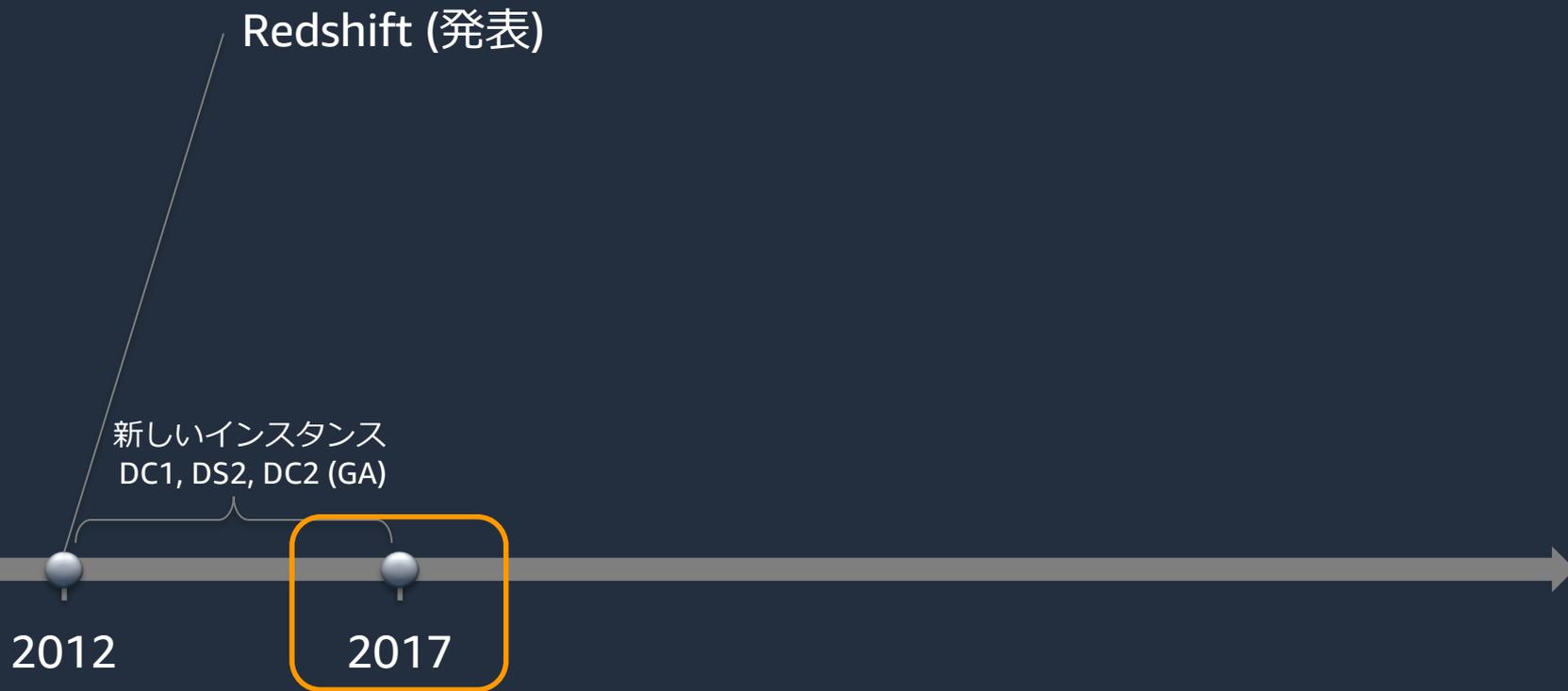
* 費用は 2020 年 2 月時点での東京リージョンのもので
コンピュータ料金は、1 時間未満の時間は 1 秒単位で請求されます

Redshift インスタンス起動時間

- コンピュートノード数 x 1 時間あたりの価格 (リーダーノードは課金対象外)
- リザーブドインスタンス (1 年、3 年) にも対応

	vCPU	メモリ	ストレージ	I/O	スライス	ノード数	インスタンス 料金 (*)
DC2 - Dense Compute							
dc2.large	2	15 GB	0.16TB SSD	0.6 GB/s	2	1~32	\$0.314 /h
dc2.8xlarge	32	244 GB	2.56TB SSD	7.5 GB/s	16	2~128	\$6.095 /h
DS2 - Dense Storage							
ds2.xlarge	4	31 GB	2TB HDD	0.4 GB/s	2	1~32	\$1.190 /h
ds2.8xlarge	36	244 GB	16TB HDD	3.3 GB/s	16	2~128	\$9.520 /h

Amazon Redshift アーキテクチャの進化の歴史



データウェアハウス環境を取り巻く状況の変化



クラウドへの
移行の加速



イベントデータの
爆発的増加



あらゆるデータを
繋げて洞察を得たい

データレイクという考え方

- 構造 / 半構造 / 非構造すべてのデータを、スケールする形で、かつ低コストで一箇所に集めて保存するためのもの
- 適切にアクセスコントロールを定義したうえで、データをセキュアに保持することができるもの
- 「民主化」した形での組織内のデータアクセスを可能とするもの
- 新しいデータ分析を、素早くかつ簡単に実行することができるもの

データレイク

AWS のデータレイク = Amazon S3

上限無し : サイジング不要

安価 :

- \$0.025/GB/月* (スタンダード)
- \$0.019/GB/月* (標準-低頻度アクセス)

例) 10TB の保存で約 2.1万円/月**

高い耐久性 : 99.999999999%

API アクセス

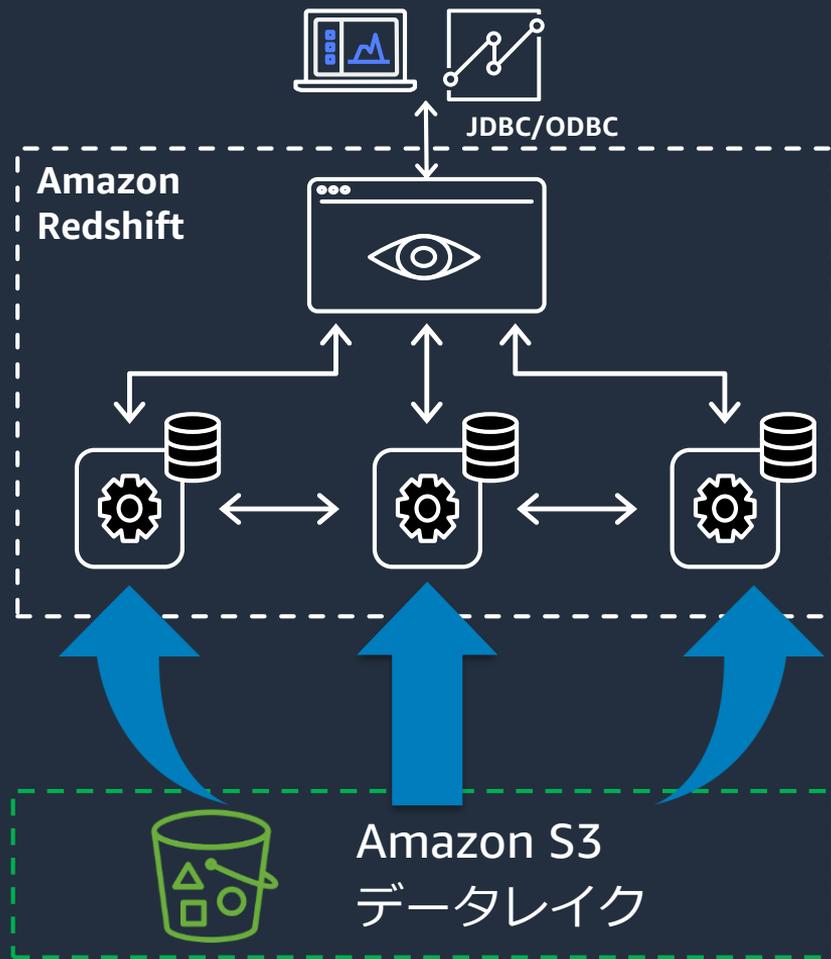
- 多様な言語のライブラリを提供
- AWS 各種サービスと連携

* 費用は 2020 年 2 月時点での東京リージョンでの価格です

** 1USドル = 110円で、標準-低頻度アクセスでの試算



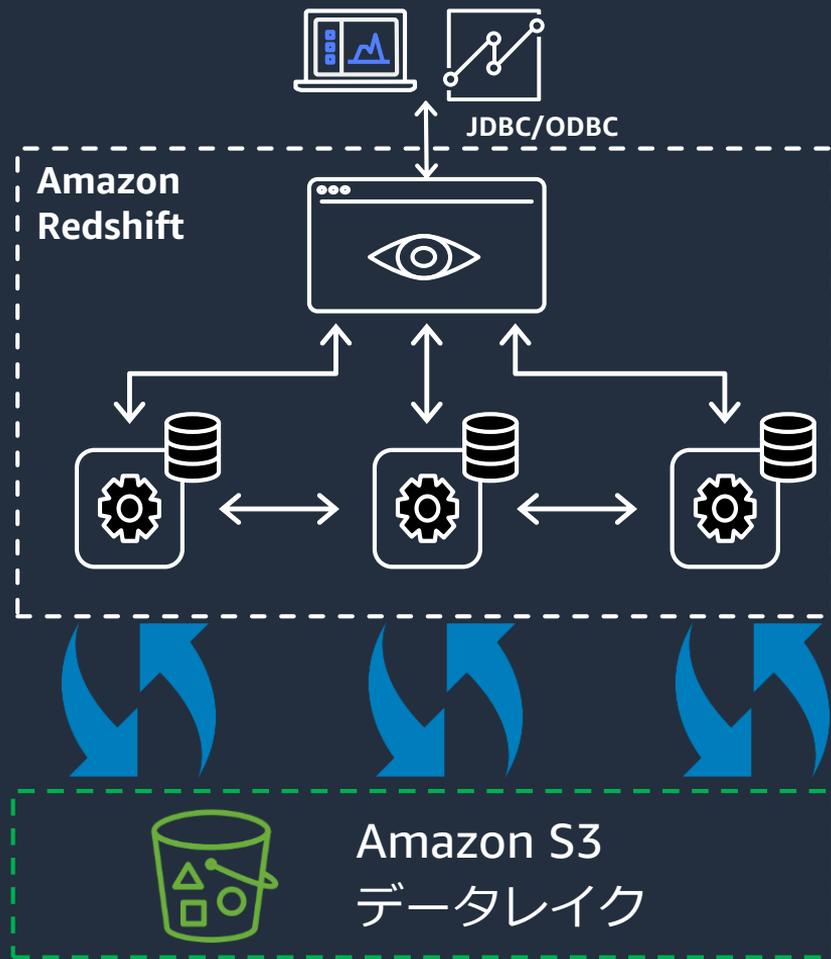
データウェアハウスとデータレイクの関係



データレイクにはあらゆるデータが格納される

必要なデータをデータウェアハウスにロードして分析

データウェアハウスとデータレイクの関係



データウェアハウスには

- 格納が難しい
- 格納したくない

データもある

データをデータレイクに置いたまま、分析ができないだろうか？

Redshift Spectrum でアーキテクチャをデータレイクに拡張



JDBC/ODBC

Amazon
Redshift



Amazon Redshift Spectrum

S3 ファイルへのクエリ実行エンジン



Amazon S3
ユーザーバケット

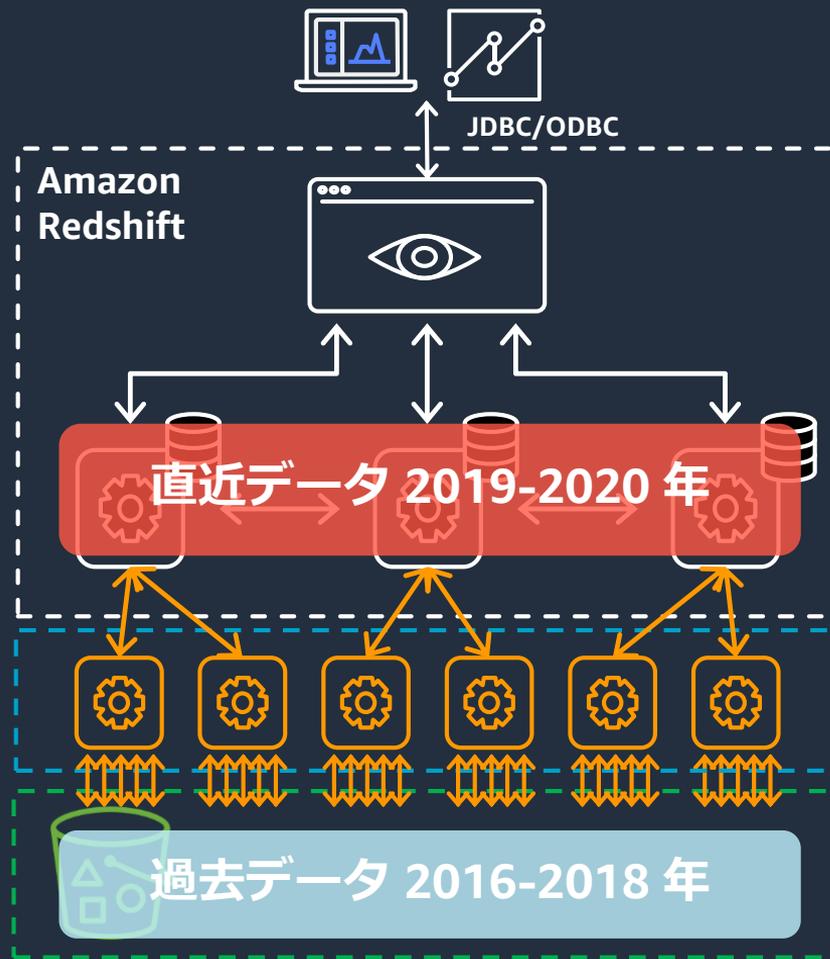


オープンフォーマットファイル
(Parquet, ORC, JSON, CSV etc)

アプリケーションは、
データウェアハウスと
データレイクの双方の
データに透過的に
アクセスできる

Redshift Spectrum の代表的ユースケース

- 頻繁にアクセスされる直近データを Redshift データウェアハウスに置き、アクセス頻度の落ちる過去データは S3 データレイクに置く
- それぞれのデータを UNION, JOIN などで繋いでビュー化することで、データの格納先を意識せずに透過的に分析することが可能に
- 直近データの保存に必要なだけのノード数を維持しながら、より多くのデータを分析対象にできる



Redshift Spectrum の料金

* 費用および課金体系は 2020 年 2 月時点での東京リージョンのものです

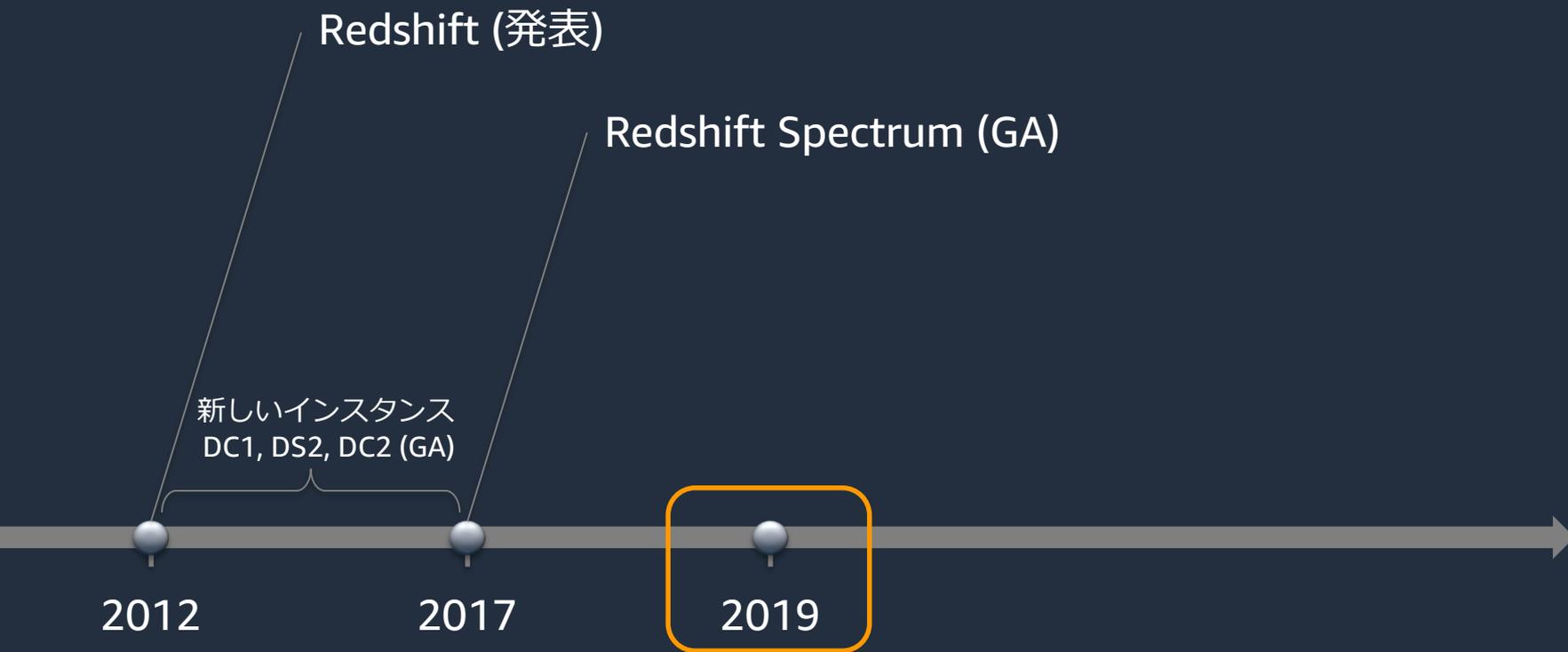
Amazon S3 データレイクへのクエリ容量

- S3 上の「圧縮済み」データ 1 TB スキャンあたり \$5 (*)

<Spectrum の料金を抑えるための Tips>

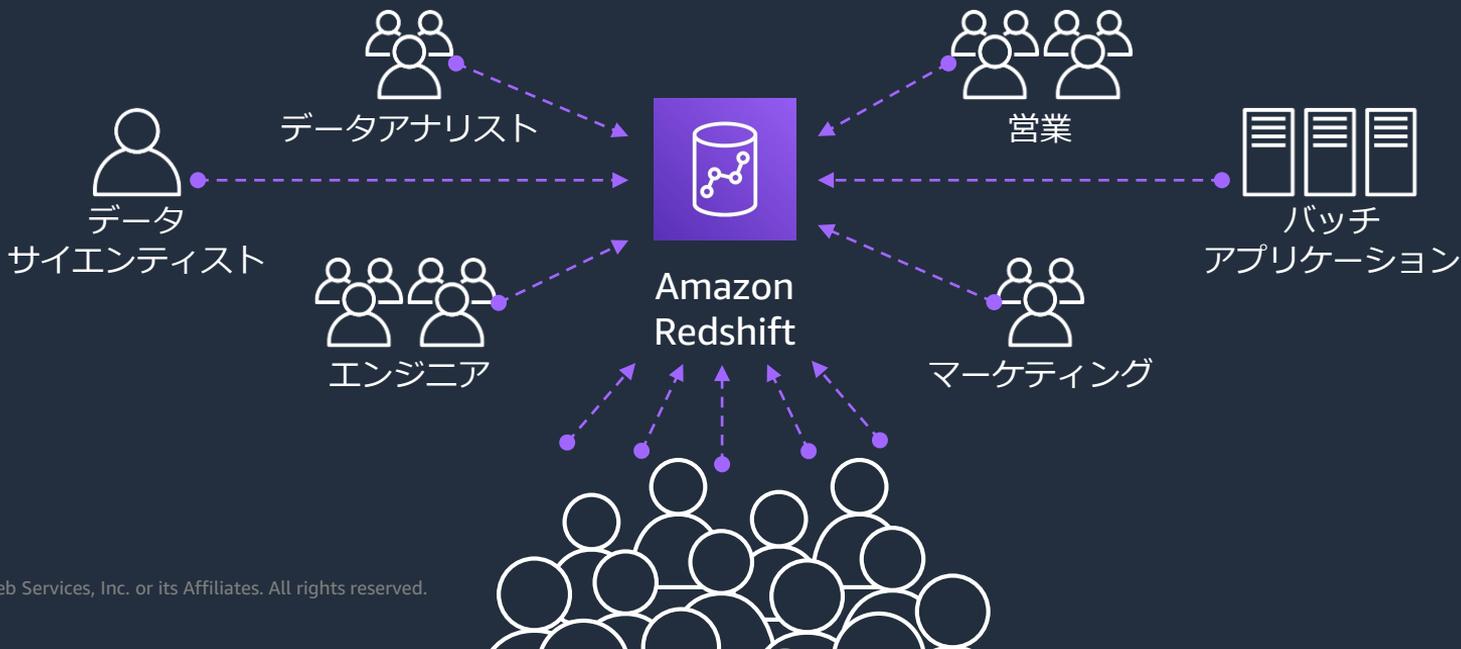
- 配置するファイル
 - Parquet や ORC などのカラムナフォーマットでファイルを保存し、クエリによるアクセス範囲を小さくする
 - ファイルは圧縮し、100MB~1GB 程度に分割
 - パーティショニングする
- 効果的なクエリの記述
 - 必用な列だけ取得する
 - 集計、フィルターで戻すデータを最小にする

Amazon Redshift アーキテクチャの進化の歴史



頭を悩ませる並列ワークロードの課題

- 評判のよいデータウェアハウスは成長していく
- 多様なユーザーが様々なタイミングで様々なツールを使って接続してくる
- このような環境下で、アクセスの集中するピークの時間帯では
クラスター全体のパフォーマンス低下を引き起こす可能性があった



ピーク時にコンピュートを自動拡張する **Concurrency**

Scaling

メインクラスター



Amazon Redshift



ディスパッチ

追加クラスター(1~10)



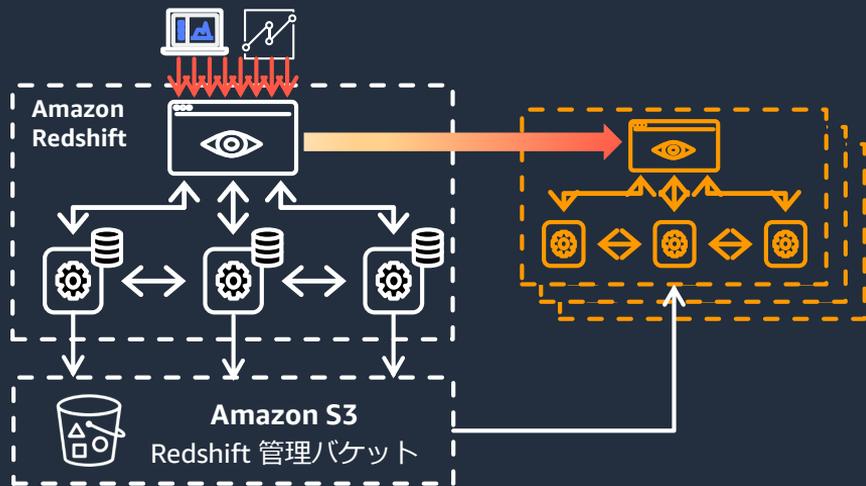
Backup



Amazon S3
Redshift 管理バケット

データ読み取り

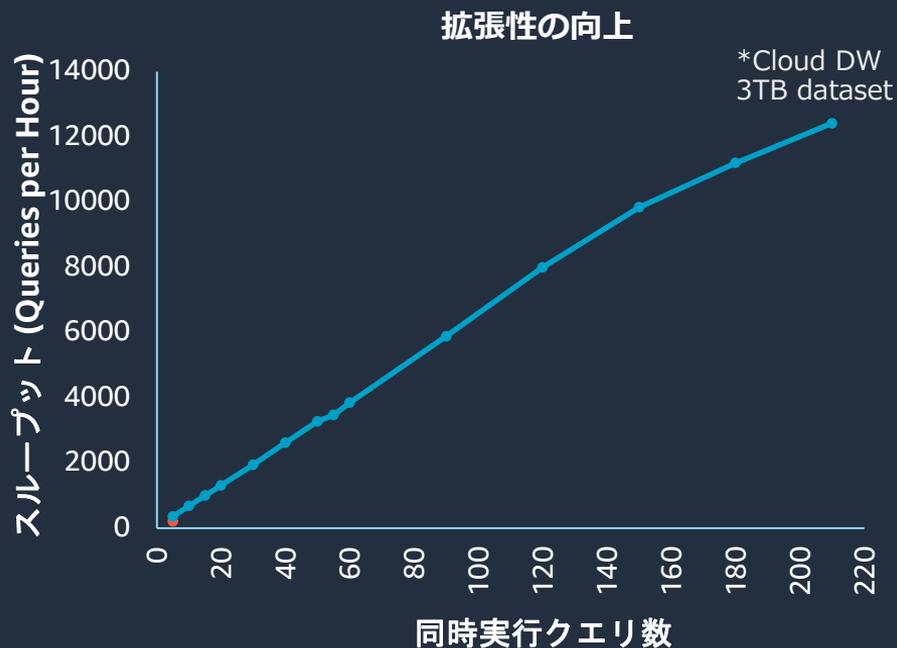
Concurrency Scaling の進化



メインクラスターでクエリのキュー待ちが発生すると、バックグラウンドで別の独立したクラスターを自動的に追加し並列処理

継続的にパフォーマンスを向上

GA 後 1 年で性能は 35 倍



Concurrency Scaling の料金

* 課金体系は 2020 年 2 月時点での東京リージョンのものです

Redshift Concurrency Scaling クラスターでのクエリ実行時間

- 各追加クラスターでクエリが実行された期間（秒）
- 1 日あたり 1 時間分の**無料クレジット**が付与（最大 30 時間）



Amazon Redshift 次世代アーキテクチャ RA3/AQUA

AWS 公式 Webinar

<https://amzn.to/JPWebinar>

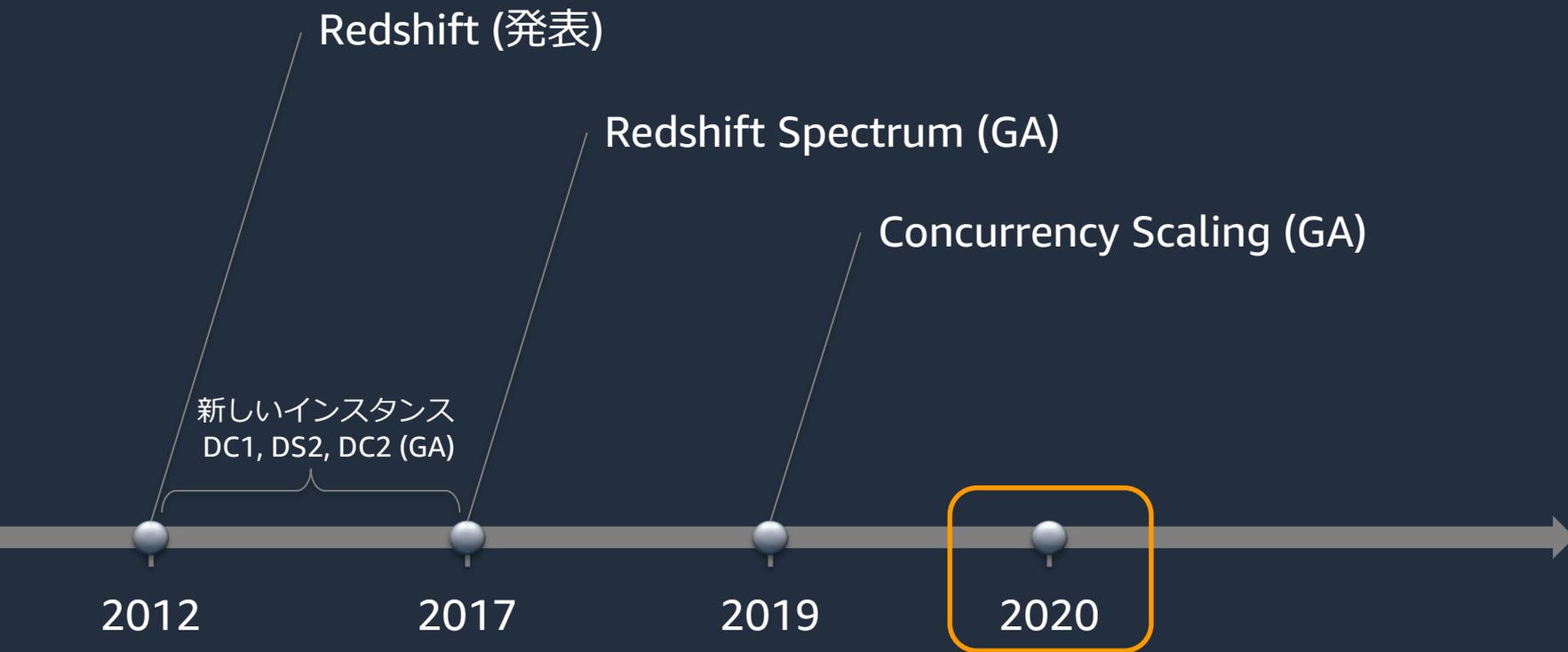


過去資料

<https://amzn.to/JPArchive>



Amazon Redshift アーキテクチャの進化の歴史



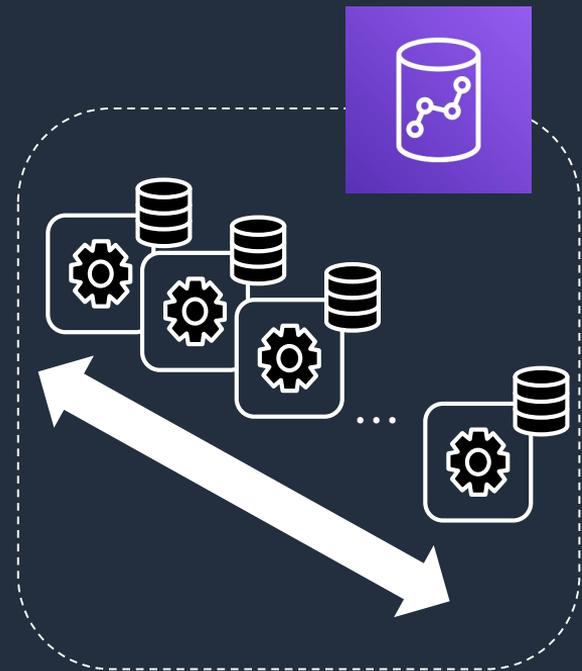
これまでの Redshift アーキテクチャの課題

Amazon Redshift が採用している
シェアードナッシング + MPP アーキテクチャは
コンピューとストレージのセットを並列に構えて
大容量データを高速分析するための優れたアーキテクチャ

しかし、コンピューとストレージが密結合ゆえに、

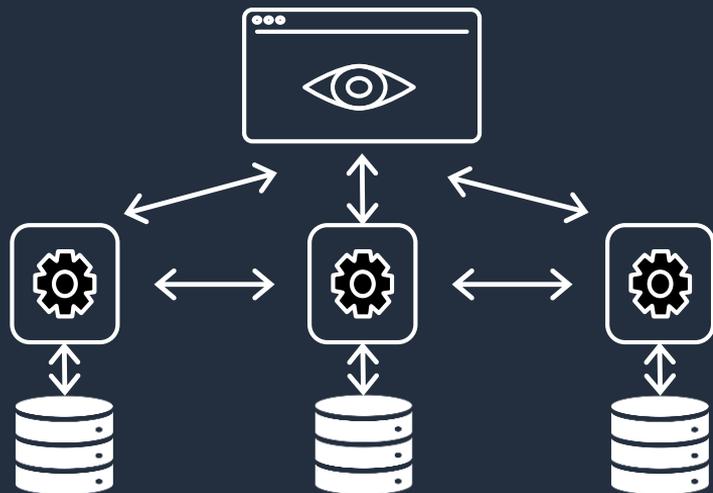
- コンピューとストレージを別々にスケールできない
- データ増に伴うノード構成変更時にデータの再分散が発生し、時間がかかる

という課題があった



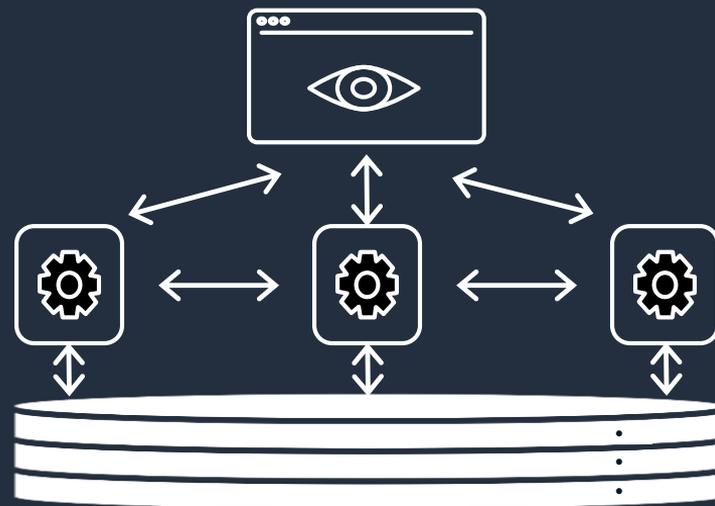
2つのデータウェアハウスアーキテクチャ (一般論)

シェアードナッシング型



利点 : 高いパフォーマンス
課題 : 柔軟かつ高速に基盤を
スケールすることが難しい

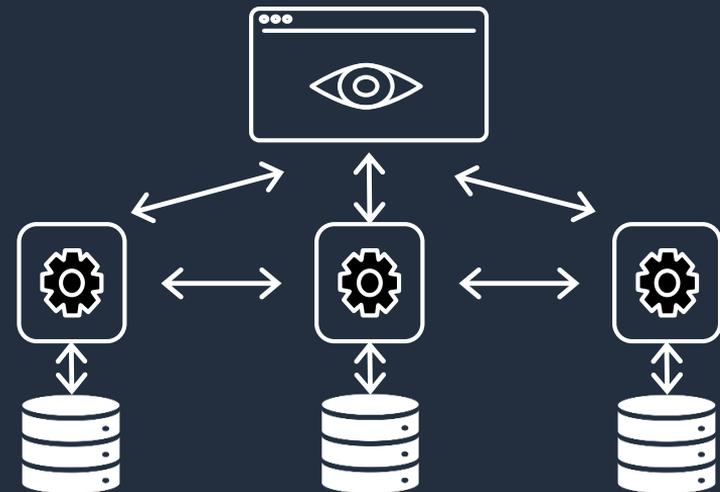
シェアードストレージ型



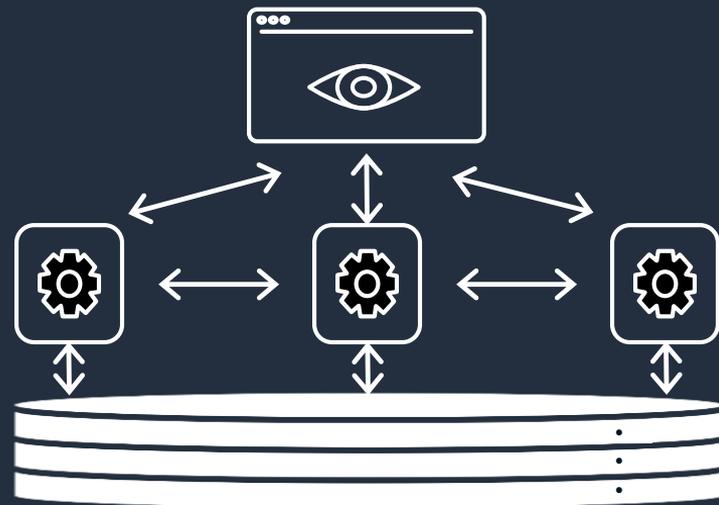
利点 : 高い柔軟性と耐障害性
課題 : ディスクやネットワークの
ボトルネックが発生しやすい

2つのデータウェアハウスアーキテクチャ (一般論)

シェアードナッシング型



シェアードストレージ型



それぞれのアーキテクチャの利点を享受できないだろうか？

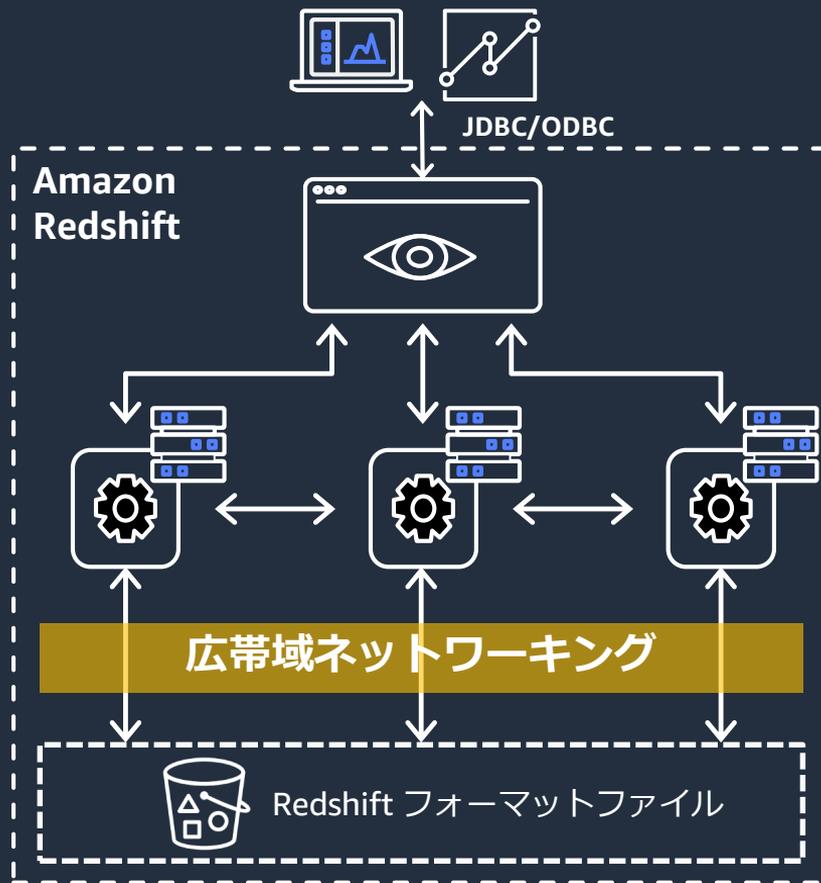
次世代 Amazon Redshift のアーキテクチャ (RA3インスタンス)

RA3 インスタンス

シェアードナッシングの利点を活かしつつ

シェアードストレージの利点を享受する

ハイブリッドアーキテクチャを採用



リーダーノード

- クエリのエンドポイント
- SQL 処理コードの生成と展開

コンピューターノード

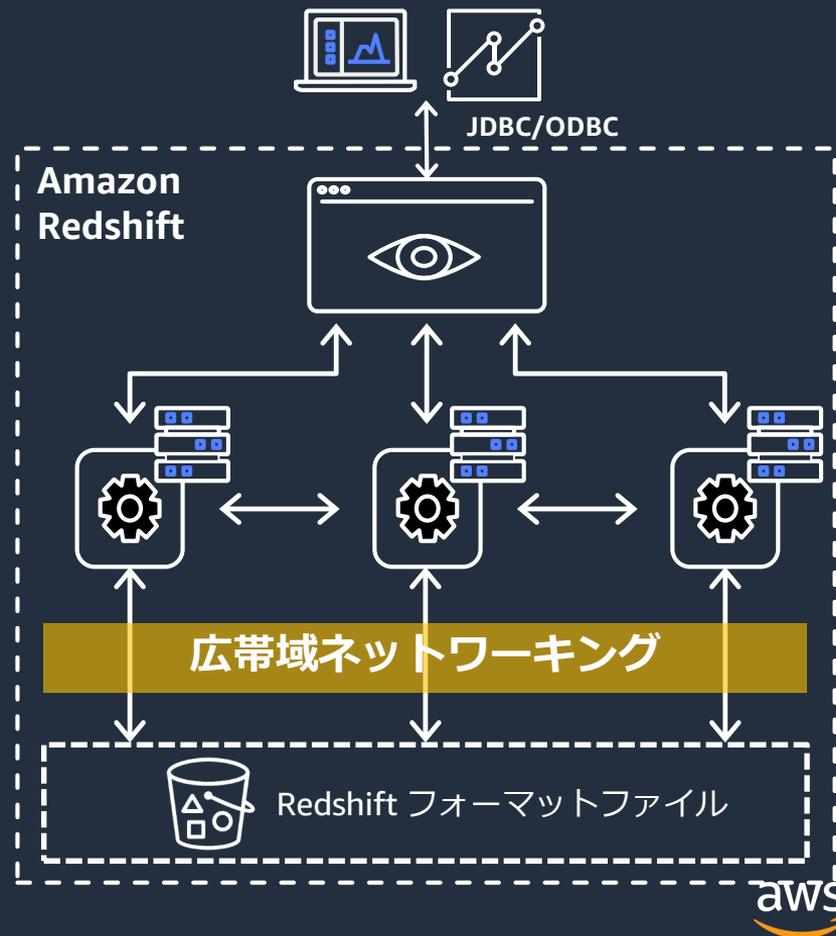
- 高速ローカル SSD キャッシュ + 大容量 RAM 搭載
- 広帯域ネットワーク
- クエリの並列実行

マネージドストレージ

- Redshift 管理 S3 バケット

次世代 Amazon Redshift のアーキテクチャ (RA3インスタンス)

- コンピュートとストレージを分離しスケーリングと支払いを独立
- データは、永続ストレージとしての S3 とキャッシュとしてのローカル SSD に格納される
- アクセス頻度の高いブロックはキャッシュにとどまり、あまりアクセスされないブロックは自動的にキャッシュアウト

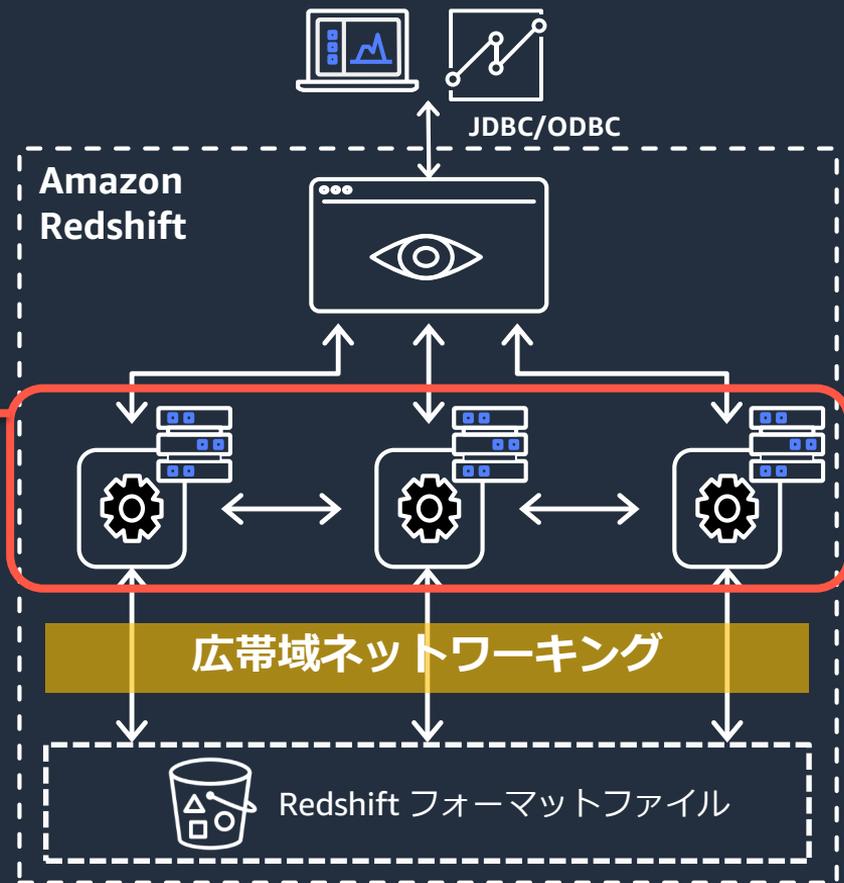


RA3 インスタンスの料金

* 費用は 2020 年 2 月時点での東京リージョンのもので
コンピュータ料金は、1 時間未満の時間は 1 秒単位で請求されます

RA3.16xlarge

- コンピュート料金
 - \$15.347/ノード/時間
 - 3年リザーブドインスタンスで60%以上ディスカウント

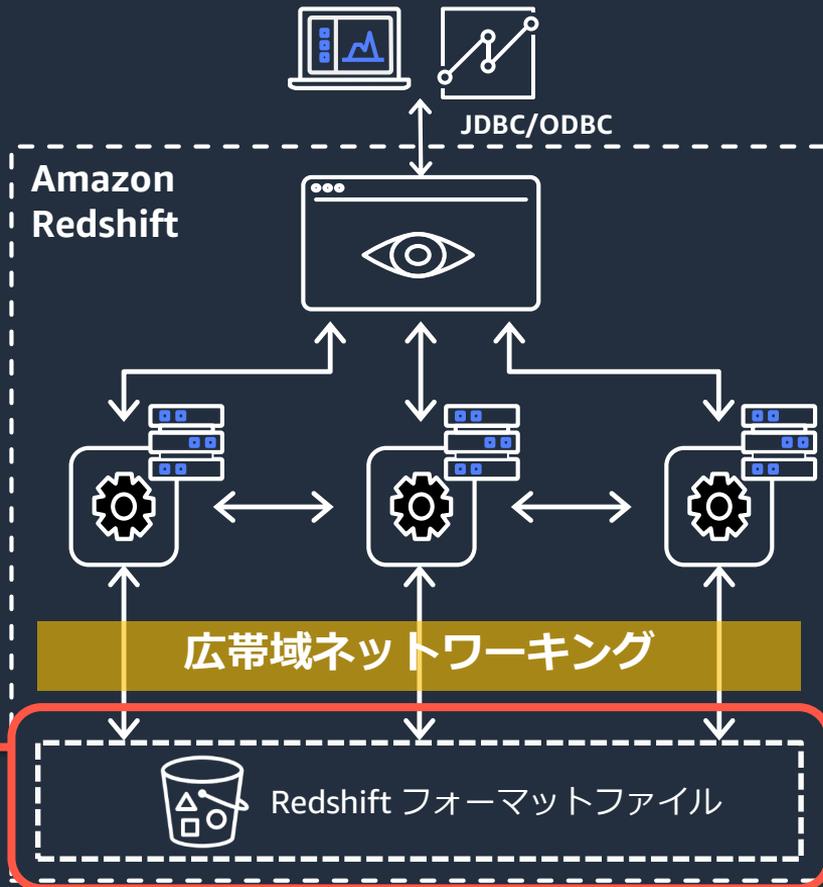


RA3 インスタンスの料金

* 費用は 2020 年 2 月時点での東京リージョンのもので
コンピュータ料金は、1 時間未満の時間は 1 秒単位で請求されます

RA3.16xlarge

- コンピュート料金
 - \$15.347/ノード/時間
 - 3年リザーブドインスタンスで60%以上ディスカウント
- ストレージ料金
 - 1 ノードあたり最大 64TB
 - 格納したデータ容量分のみ
 - \$0.026/GB-月



RA3 インスタンスは最新の Nitro 高性能ハードウェアで構成

RA3.16xlarge

ノードあたりスペック

- 48 vCPUs
- 384 GiB RAM
- 8GB/s の I/O スループット

組める構成

- 最小2ノード (128TB) から最大128ノード (8PB) までスケールが可能

※ RA3.4xlarge Coming Soon

クラスター設定

ニーズに最適な構成を計算する

ノードの種類
CPU、RAM、ストレージ容量、およびドライブタイプの要件を満たすノードの種類を選択します。

推奨

RA3

スケーラブルなマネージド型ストレージによる高パフォーマンス

<input checked="" type="radio"/> ra3.16xlarge	\$15.347/ノード/時
マネージド型ストレージ: 最大 64 TB/ノード	\$0.026/GB/月

DC2

固定ローカル SSD ストレージによる高パフォーマンス

<input type="radio"/> dc2.large	\$0.314/ノード/時
ストレージ: 160 GB/ノード	
<input type="radio"/> dc2.8xlarge	\$6.095/ノード/時
ストレージ: 2.6 TB/ノード	



dc2.8xlarge
32 vCPU (2 世代)

▶ レガシー高密度ストレージノードタイプを表示

ノード
必要なノードの数を入力します。

範囲 (2~128)

Amazon Redshift の料金 Update

*費用は 2020 年 2 月時点での東京リージョンのもので
コンピュータ料金は、1 時間未満の時間は 1 秒単位で請求されます

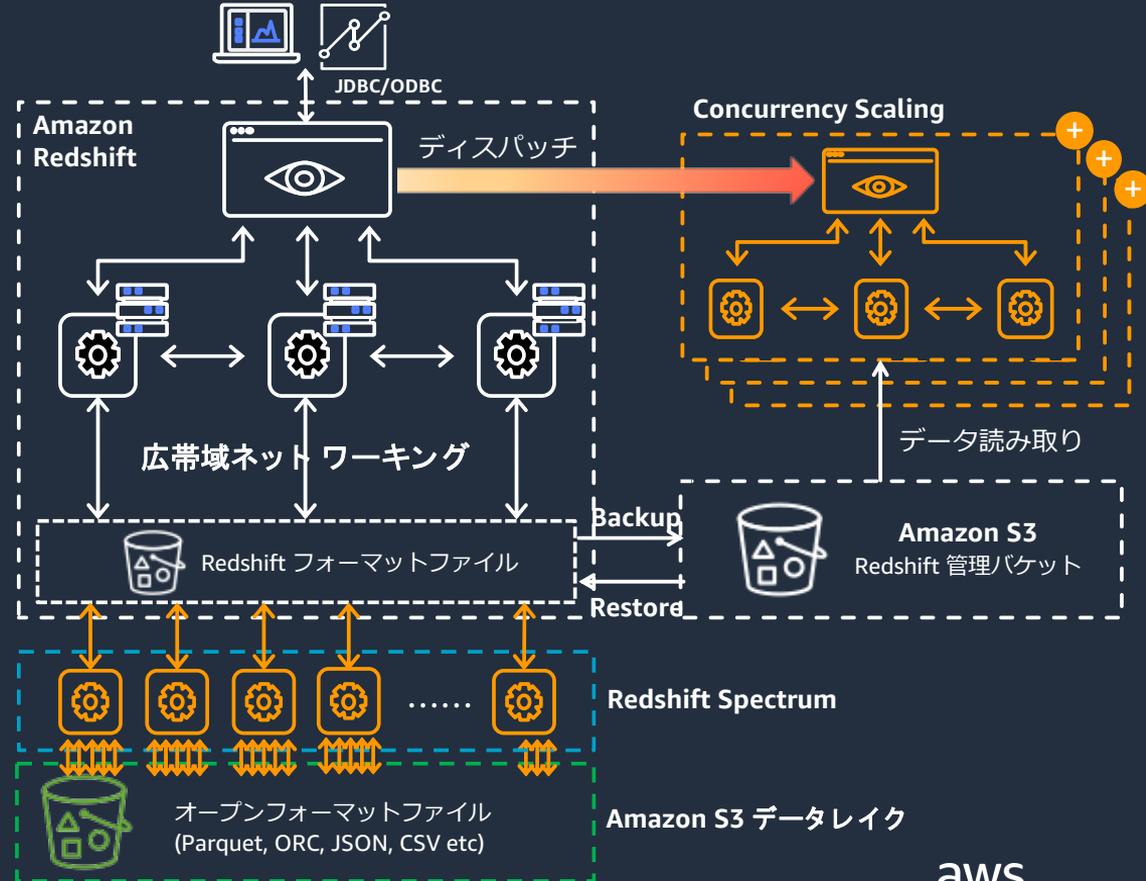
Redshift インスタンス起動時間 + ストレージ使用量(RA3のみ)

- コンピュートノード数 x 1 時間あたりの価格 (リーダーノードは課金対象外)
- リザーブドインスタンス (1 年、3 年) にも対応

	vCPU	メモリ	ストレージ	I/O	スライス	ノード数	コンピュート料金 (*)	ストレージ料金 (*)
DC2 - Dense Compute								
dc2.large	2	15 GB	0.16TB SSD	0.6 GB/s	2	1~32	\$0.314 /h	
dc2.8xlarge	32	244 GB	2.56TB SSD	7.5 GB/s	16	2~128	\$6.095 /h	
DS2 - Dense Storage								
ds2.xlarge	4	31 GB	2TB HDD	0.4 GB/s	2	1~32	\$1.190 /h	
ds2.8xlarge	36	244 GB	16TB HDD	3.3 GB/s	16	2~128	\$9.520 /h	
RA3 with Redshift Managed Storage(RMS) New!								
ra3.4xlarge	Stay tuned !							
ra3.16xlarge	48	384 GB	64TB RMS	8.0 GB/s	16	2~128	\$15.347 /h	\$0.026 /GB-Month

RA3 インスタンスと従来インスタンス (DC2/DS2) の互換性

- RA3 と従来の DC2/DS2 はアプリケーションレベルで互換性がある
- Redshift Spectrum や Concurrency Scaling などの機能についても引き続き使用できる



RA3 と Redshift Spectrum との違い・使い分け

	RA3	Redshift Spectrum
ストレージ	Redshift 管理の S3 領域を使用	ユーザー管理の S3 領域を使用
データフォーマット	Redshift 独自フォーマット	オープンフォーマット
データへのアクセス	Redshift からのアクセスのみ	他の AWS サービスからもアクセス可
データの更新	DML UPDATE に対応	DML UPDATE には非対応
コスト	Redshift インスタンス料金 + マネージドストレージ料金	Redshift インスタンス料金 + S3ストレージ格納料金 + Spectrum スキャン容量料金
特徴	RA3 のマネージドストレージは Redshift ストレージの拡張であり、その背後で S3 の技術が使われているが、利用時に S3 を意識する必要はない	データレイクとしての S3 にアクセスする機能であり、データレイクを介した他サービスとの連携のために今後も重要な機能であり続ける
使い分け	これまで Redshift 内のデータ増加を抑制する目的で過去データを削除したり、S3 にデータをオフロードして Spectrum を活用していた場合は、RA3 ストレージを活用	すべてのデータがデータレイクにあり、Redshift だけではなく他のサービスからも同じデータにアクセスする要件がある場合は Spectrum を活用

DS2 インスタンスから RA3 インスタンスへの移行

- HDD ベースの DS2 インスタンスが、最も RA3 インスタンスへの移行が推奨されるインスタンスタイプ
- RA3.16xlarge は DS2.8xlarge 2台につき 1台が目安 (2:1)
 - 同等価格で性能 2倍 & データ容量 2倍
 - RA3.16xlarge の最小ノード構成は 2台からのため、DS2.8xlarge 3台以下の場合は RA3.4xlarge(*) も検討
 - DS2.xlarge は 8ノードで DS2.8xlarge x 1ノード と考えて試算
- **上記はあくまで目安であり、実際には PoC による性能確認を強く推奨**

DC2 インスタンスから RA3 インスタンスへの移行

- SSD ベースの DC2 インスタンスから、RA3 インスタンスへの移行については以下の点に留意する
- RA3.16xlarge は DC2.8xlarge 3台につき 1台が目安 (3:1)
 - ストレージの観点では 8倍 (7.5TB vs 64TB) のキャパシティーになるが vCPU 数の観点では 0.5 倍 (96vCPU vs 48vCPU) になる (注意: CPU世代も異なるためコンピュート能力の単純比較はできない)
 - DC2.8xlarge 環境での CPU 使用率が 50% より高い場合はより多くの RA3 インスタンスが必要になる可能性がある
- データ容量が 10TB 以下の場合、RA3 インスタンスへの移行は非推奨

既存 Redshift クラスターの RA3 インスタンスへの移行方法

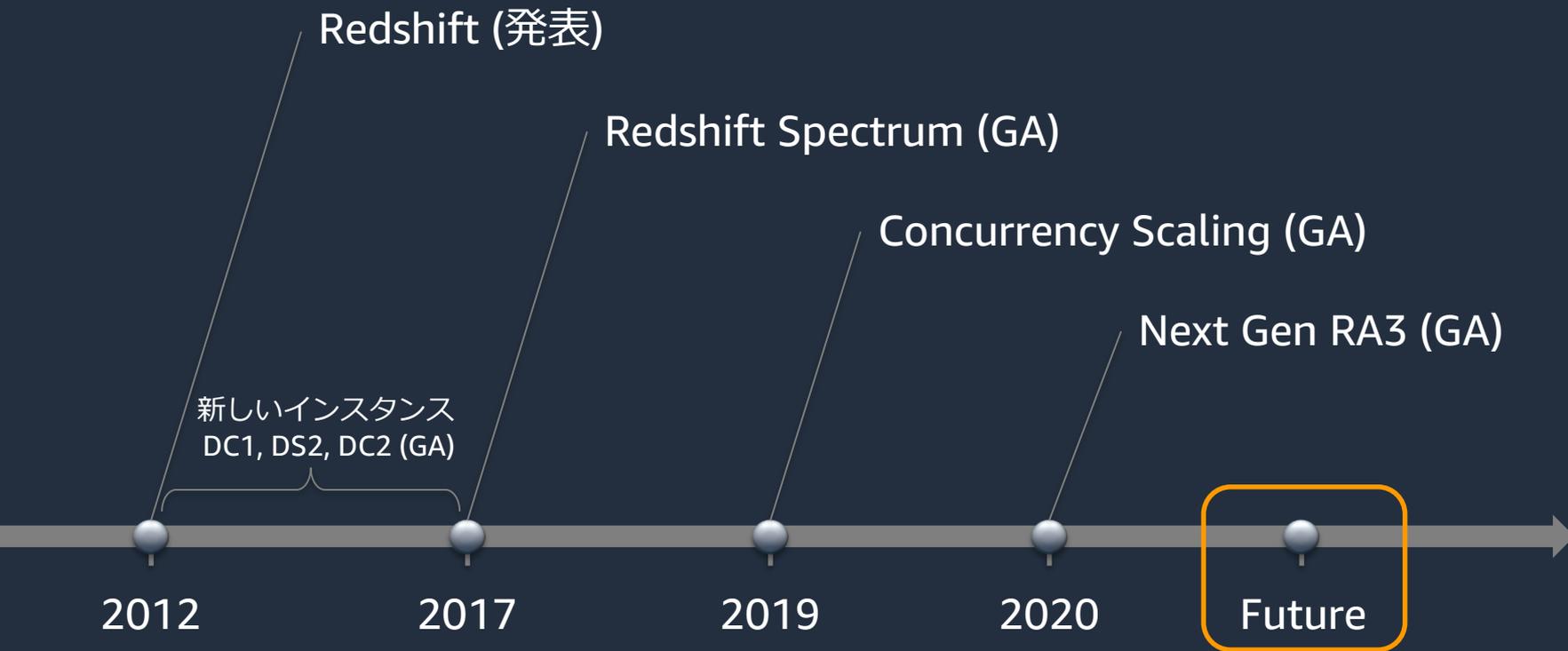
1. スナップショットからの復元

- RA3 クラスターを数分で起動
- 新規クラスターの動作確認を行い、旧クラスターを削除
- 「変更クラスター」より新規クラスター名を旧クラスター名に変更

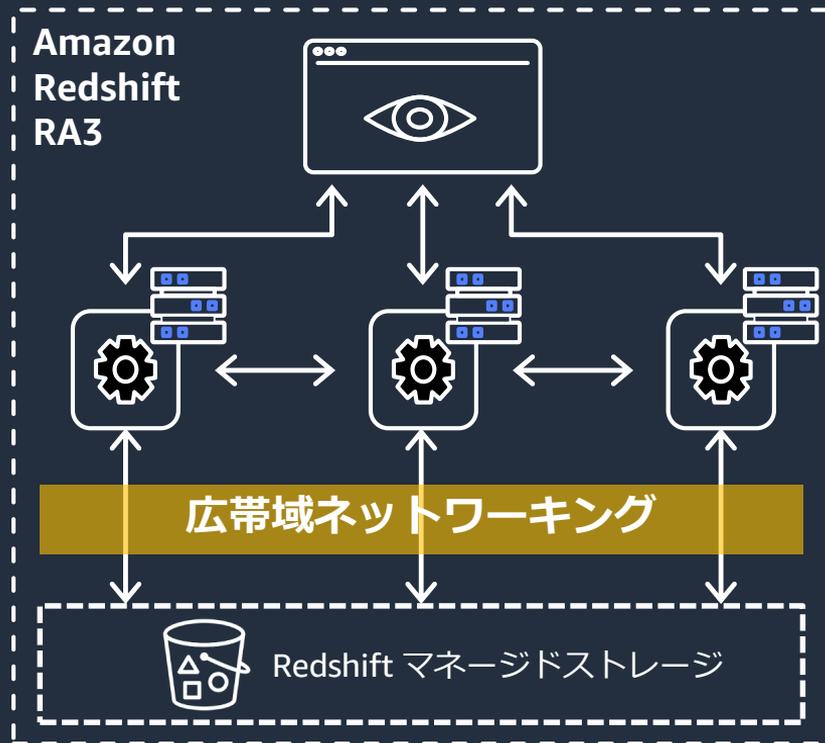
2. Classic Resize

- 「サイズ変更クラスター」より実行
(スナップショットの復元で行うオペレーションを一気通貫で実施。
ただしスナップショットからの復元に比べると時間を要する)

Amazon Redshift アーキテクチャの進化の歴史



RA3 インスタンスを更に強化

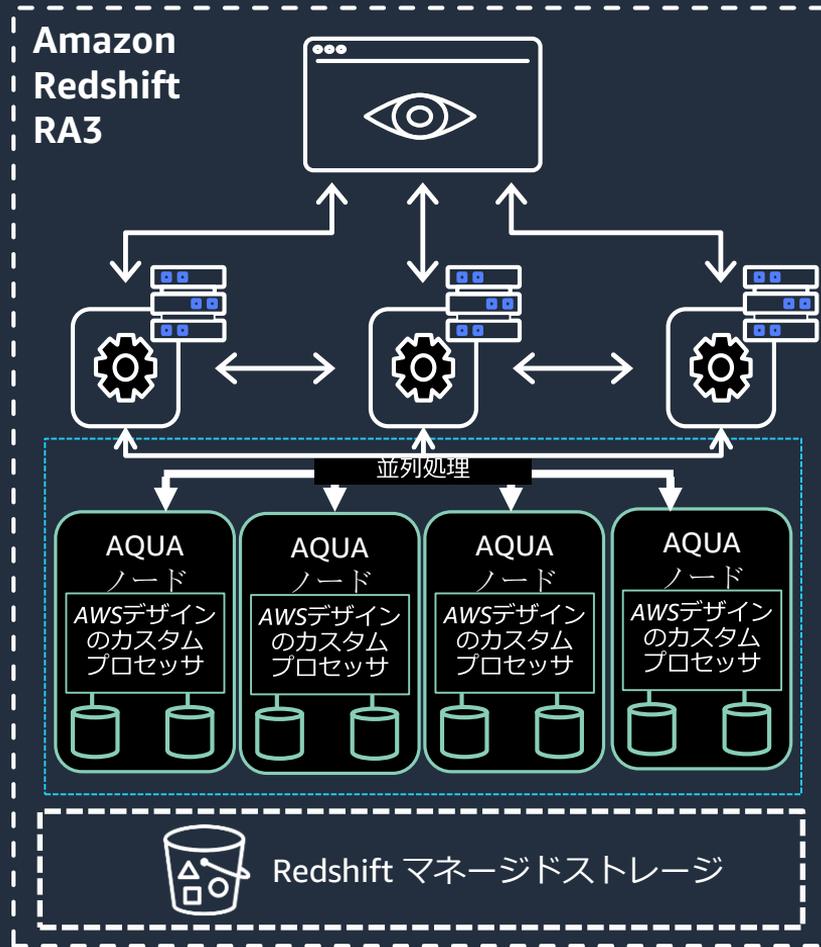


コンピュータードと
マネージドストレージ間の
ネットワークの
パフォーマンスペナルティ
を防ぐには？

AQUA (Advanced Query Accelerator) for Amazon

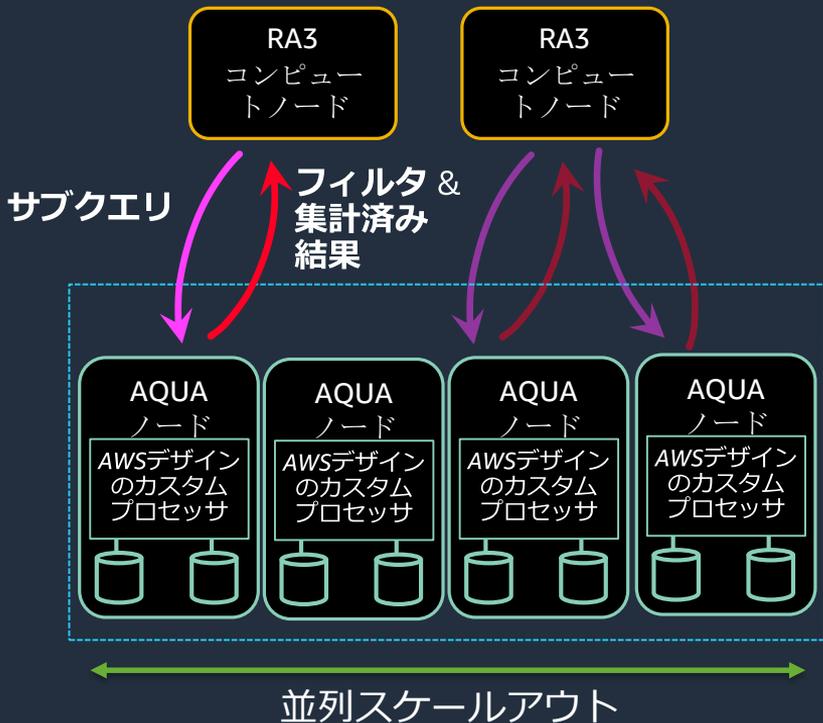
Redshift

コンピュータードと
マネージドストレージ
間に AQUA という
新しい分散型ハード
ウェアアクセラレー
ション処理レイヤ
が組み込まれる



AQUA を活用することで、
Amazon Redshift は他の
クラウドデータウェアハウ
スに比べて 10 倍高速に

AQUA - プッシュダウン & スケールアウト計算レイヤー



AQUA 上の Nitro チップがデータの
圧縮、暗号化を高速化

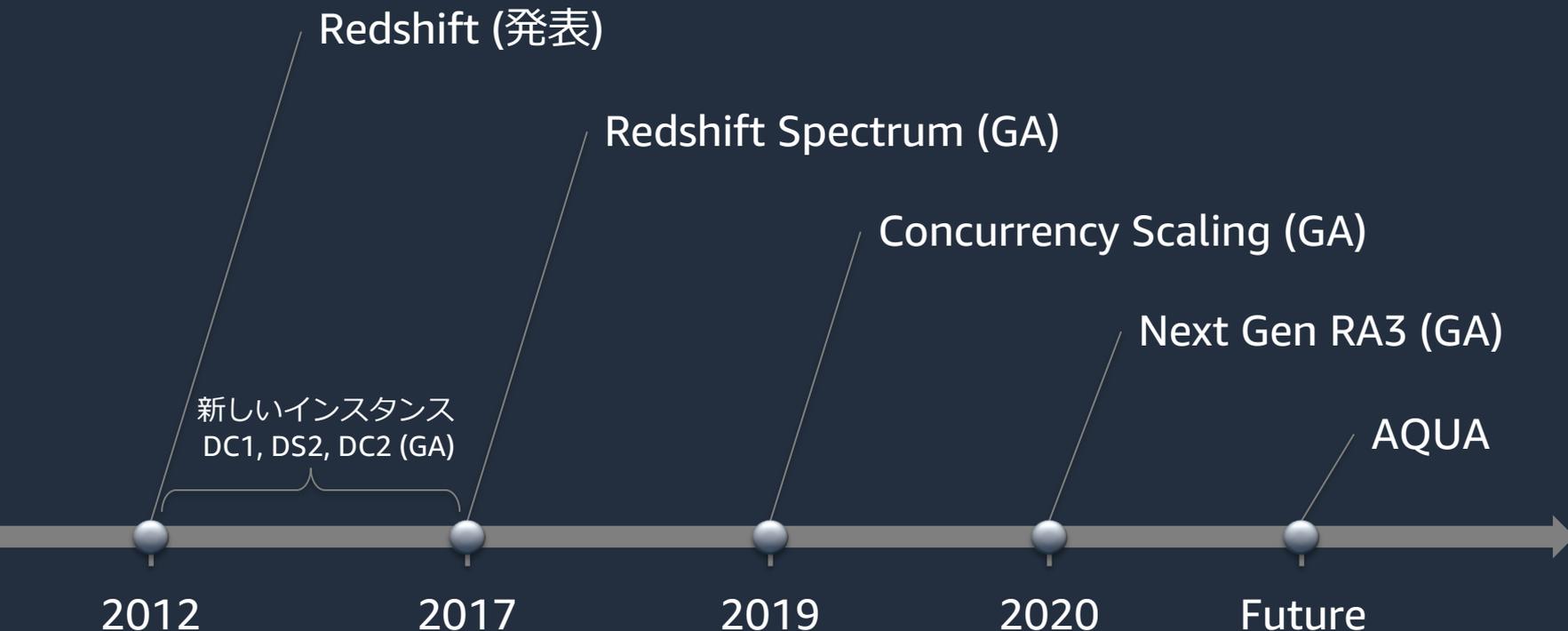
AWS がカスタムデザインした
AQUA ノード上の分析処理向け
プロセッサが、データの
フィルタリングや集計を高速に実行

コンピュータノードから
AQUA ノードに処理を
プッシュダウンすることで
ネットワーク転送データを最小化

AQUA その他留意点

- AQUA は RA3 インスタンスを前提とした機能拡張であり、他のインスタンスタイプでは利用することはできない
- RA3 インスタンスに AQUA が搭載されても、
現行の Redshift の SQL やその他オペレーションは変更する必要なく
そのまま利用が可能
- 現在パブリックプレビュー受付中

Amazon Redshift アーキテクチャの進化の歴史



Amazon Redshift 新機能アップデート

AWS 公式 Webinar

<https://amzn.to/JPWebinar>



過去資料

<https://amzn.to/JPArchive>



近年の Amazon Redshift のめざましい進化

Robust result set caching

Large # of tables support
~20000

Copy command support
for ORC, Parquet

IAM role chaining

Elastic resize

Groups

Redshift Spectrum: date formats, scalar json and ION file formats support, region expansion, predicate filtering

Auto analyze

Health and performance monitoring w/Amazon Cloud watch

Automatic table distribution style

Cloud watch support for WLM queues

Performance enhancements— hash join, vacuum, window functions, resize ops, aggregations, console, union all, efficient compile code cache

Unload to CSV
Auto WLM

~25 Query Monitoring Rules (QMR) support

AQUA

Concurrency Scaling

200+

DC1 migration to DC2

Resiliency of ROLLBACK processing

Manage multi-part query in AWS console

Auto analyze for incremental changes on table

Spectrum Request Accelerator

Apply new distribution key

Redshift Spectrum: Row group filtering in Parquet and ORC, Nested data support, Enhanced VPC Routing, Multiple partitions

Faster Classic resize with optimized data transfer protocol

過去 18 ヶ月にリリースされた新機能の数

Performance: Bloom filters in joins, complex queries that create internal table, communication layer

Redshift Spectrum: Concurrency scaling

Amazon Lake Formation integration

Auto-Vacuum sort, Auto-Analyze and Auto Table Sort

Auto WLM with query priorities

Snapshot scheduler

Performance: join pushdowns to subquery, mixed workloads temporary tables, rank functions, null handling in join, single row insert

Advisor recommendations for distribution keys

AZ64 compression encoding

Console redesign

Stored procedures

RA3

Performance of Inter-Region Snapshot Transfer

Federated Query

Materialized Views

Spatial Processing

Column level access control with AWS lake formation

AZ64 - 高い圧縮率と性能を実現する新しい列エンコーディング

高いデータ圧縮率と性能を両立

データを小さなグループで効率的に圧縮

内部で SIMD 命令による並列処理を活用

数値および日付型データに対して有効で、
デフォルトで設定されるようになっている

AZ64 と既存の圧縮エンコーディング
との比較(*)

	AZ64 圧縮率	AZ64 性能
RAW	60-70% ストレージ容量削減	25-30% 高速
LZO	35% ストレージ容量削減	40% 高速
ZSTD	5-10% ストレージ容量削減	70% 高速

* [TPC-DS \(v2.10\)](#), 30TBを使用、クエリの修正なし

機械学習ベースの自動最適化でクエリパフォーマンス向上

テーブルメンテナンスの自動化

機械学習を活用し、クラスターの状態やクエリワークロード状況を見ながら処理を実行

チューニングアドバイザにより
リコメンデーションの提供



Automatic Analyze



Automatic Table Distribution Style



Distribution/Sort key advisors



Automatic Vacuum Delete



Automatic Table Sort

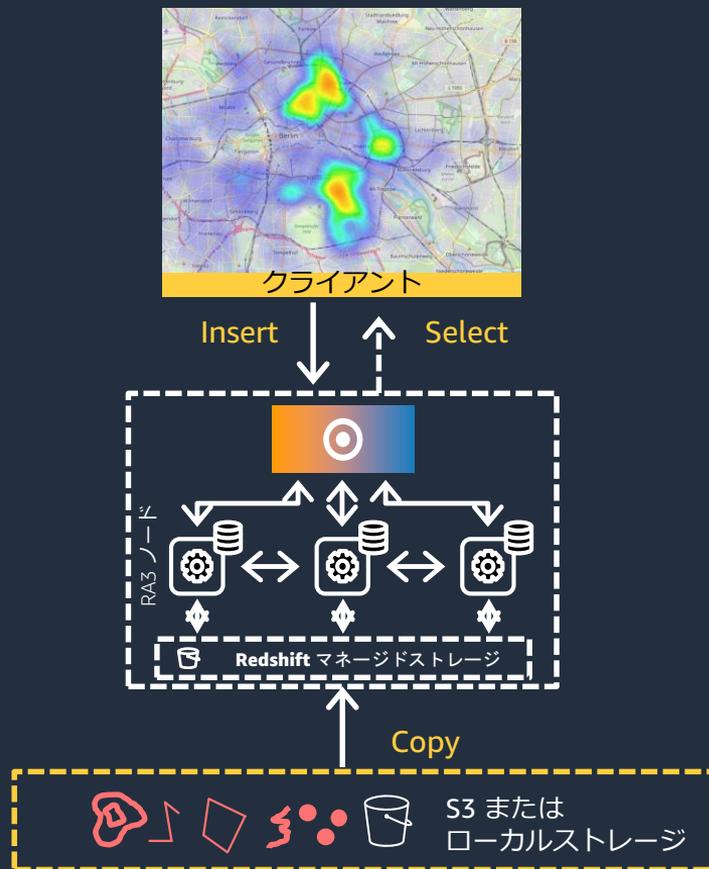
空間サポート - 新しいデータ型 Geometry

空間データとビジネスデータを
シームレスに統合

Geometry データ型がサポートする空間タイプ:
MultiPoint, MultiLinestring, MultiPolygon,
GeometryCollection など

40以上の Spatial 関数を提供

空間データのインポート、エクスポート、
アクセスおよび処理を実行可能



マテリアライズドビュー (プレビュー中)

頻繁に実行するクエリパターンを高速化

- 結合、フィルタ、集計、射影

ETL/BI パイプラインの簡素化

- 差分リフレッシュ
- ユーザーによるメンテナンス

Redshift へのより簡単で迅速な移行

マテリアライズドビュー

loc_sales	
loc	total_sales
SF	12.00
NY	10.00

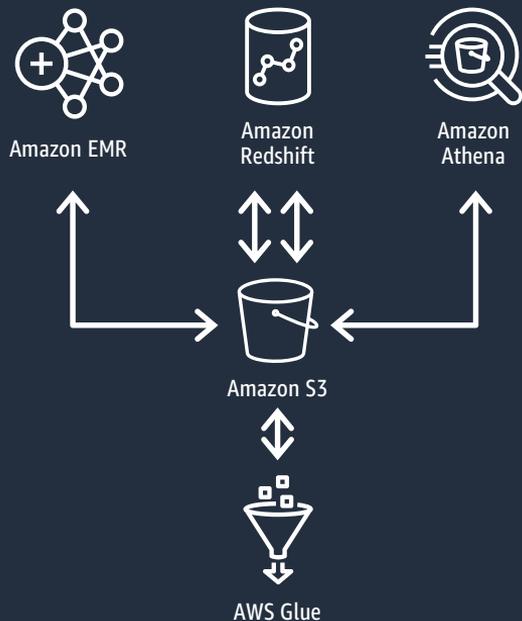


“地域ごとの
総売上は?”

store_info		
store	owner	loc
s1	Joe	SF
s2	Ann	NY
s3	Lisa	SF

sales			
item	store	cust	price
i1	s1	c1	12.0
i2	s2	c1	3.0
i3	s2	c2	7.0

データレイクエクスポート - Parquet でデータを共有



Redshift テーブルデータを Amazon S3
上へ Parquet でエクスポートすること
が可能に

Parquet は分析ワークロード向けの
オープンな列指向ファイルフォーマット

Redshift Spectrum だけでなく、
Amazon Athena や Amazon EMR など
AWS の他の分析サービスでも
すぐに分析に活用することが可能に

データレイクエクスポートのユースケース①



Amazon
Redshift

Redshift テーブル上の過去データを
Parquet ファイルに変換

Sales
(Table)

A white icon representing a table with 3 columns and 3 rows, with a black outline.

UNLOAD

```
('select * from sales where  
sales_date = YYYYMMDD')  
TO 's3://mybucket/unload/sales/'  
FORMAT as PARQUET  
PARTITION BY (sales_date);
```



Amazon S3

Parquet

データレイクエクスポートのユースケース②



Amazon
Redshift

S3 上の CSV ファイルを外部表経由で
Parquet ファイルに変換

Event
(External table)

UNLOAD

```
('select * from Event')
```

```
TO 's3://mybucket/unload/Event/'
```

```
FORMAT as PARQUET
```

```
PARTITION BY (Event_date);
```



Amazon S3

CSV

Parquet

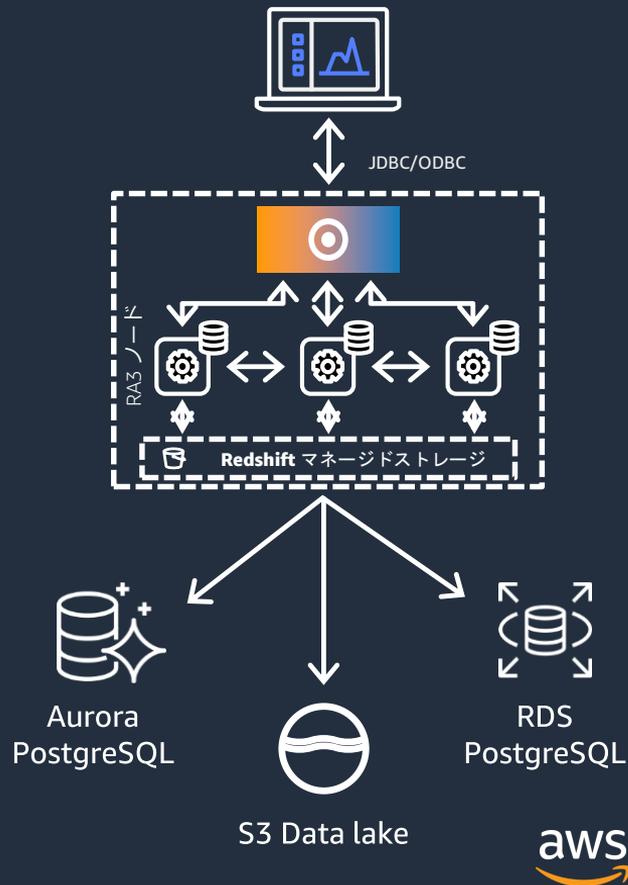
フェデレーテッドクエリ (プレビュー中)

RDS/Aurora PostgreSQL
に対して直接クエリ可能に

データ移動なしにライブデータを分析

データウェアハウス、データレイク、
オペレーショナルデータベースの
データを統合して分析

高い性能でセキュアにデータアクセス



フェデレーテッドクエリのユースケース

データレイク、データウェアハウス、OLTP データベースのデータを統合して分析

```
-- Aurora Postgres has Hot Data (2019)
-- Redshift has Recent Data (2016-2018)
-- S3 has Archival Data (1992-1998)
-- Declare a view across all backends
```

```
CREATE VIEW lineitem_all AS
SELECT * FROM s3.lineitem_1t_part -- データレイク
UNION ALL
SELECT * FROM public.lineitem -- データウェアハウス
UNION ALL
SELECT * FROM apg.lineitem -- OLTP データベース
WITH NO SCHEMA BINDING
```

```
-- Find #sales with 1 item in Jan of each year
-- Predicates are being pushed down
-- Partition pruning on the S3 data
-- Aggregates are being pushed down
-- Very intuitive syntax
```

```
SELECT EXTRACT(year FROM l_shipdate) AS year,
       EXTRACT(month FROM l_shipdate) AS month,
       COUNT(*) AS orders
FROM lineitem_all
WHERE extract(month FROM l_shipdate) = 1
AND l_quantity < 2
GROUP BY 1,2
ORDER BY 1,2;
```

Amazon Redshift まとめ

AWS 公式 Webinar
<https://amzn.to/JPWebinar>



過去資料
<https://amzn.to/JPArchive>



まとめ

- Redshift は、高速、スケーラブルで費用対効果の高いデータウェアハウスおよびデータレイク分析マネージドサービス
- お客様のニーズに応じて柔軟にアーキテクチャを進化させてきており、現在も進化の途中
 - Redshift Spectrum
 - Concurrency Scaling
 - RA3/AQUA
 - ...
- 今後もお客様のリクエストをもとに、多くの新機能が追加される予定です

Q&A

お答えできなかったご質問については

AWS Japan Blog 「<https://aws.amazon.com/jp/blogs/news/>」にて

後日掲載します。

AWS の日本語資料の場所「AWS 資料」で検索



日本担当チームへお問い合わせ サポート 日本語 ▼ アカウント ▼

コンソールにサインイン

製品 ソリューション 料金 ドキュメント 学習 パートナー AWS Marketplace その他 🔍

AWS クラウドサービス活用資料集トップ

アマゾン ウェブ サービス (AWS) は安全なクラウドサービスプラットフォームで、ビジネスのスケールと成長をサポートする処理能力、データベースストレージ、およびその他多種多様な機能を提供します。お客様は必要なサービスを選択し、必要な分だけご利用いただけます。それらを活用するために役立つ日本語資料、動画コンテンツを多数ご提供しております。(本サイトは主に、AWS Webinar で使用した資料およびオンデマンドセミナー情報を掲載しています。)

[AWS Webinar お申込 »](#)

[AWS 初心者向け »](#)

[業種・ソリューション別資料 »](#)

[サービス別資料 »](#)

<https://amzn.to/JPArchive>



AWS Well-Architected 個別技術相談会

毎週“W-A個別技術相談会”を実施中

- AWSのソリューションアーキテクト(SA)に
対策などを相談することも可能

- **申込みはイベント告知サイトから**

(<https://aws.amazon.com/jp/about-aws/events/>)

AWS イベント

で[検索]



ご視聴ありがとうございました

AWS 公式 Webinar

<https://amzn.to/JPWebinar>



過去資料

<https://amzn.to/JPArchive>



参考情報

AWS 公式 Webinar
<https://amzn.to/JPWebinar>



過去資料
<https://amzn.to/JPArchive>



RA3 インスタンスを評価されたケースの例

- DS2.8xlarge 14 ノードからの移行
 - ほぼ同等のコスト : RA3.16xlarge 7ノードでクエリが ~2.1 倍高速に
- DS2.8xlarge 16 ノードからの移行 (ETL インテンシブなワークロード)
 - ほぼ同等のコスト : RA3.16xlarge 8ノードで ETL ロードが ~1.3 倍高速に
- DC2.8xlarge 15 ノードからの移行
 - ほぼ同等のコスト : RA3.16xlarge 5ノードでクエリが ~1.25 倍高速に