AWS Builders Online Series

T4-2

データレイクプロジェクトの進め方 - 基礎から始めるデータレイクプロジェクトの考え方とそれを支える組織 -

下佐粉 昭

アマゾン ウェブ サービス ジャパン合同会社
アナリティクス事業本部 ソリューションアーキテクト部 シニアソリューションアーキテクト

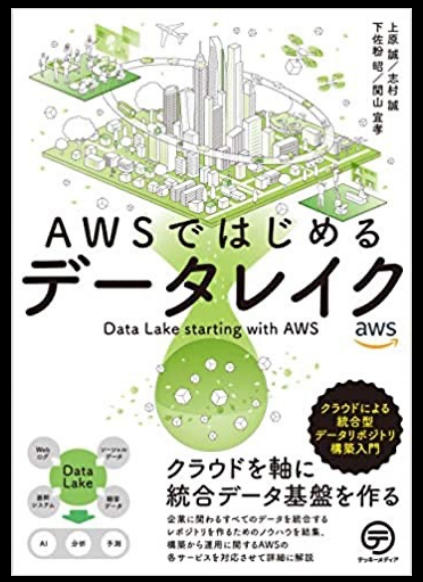
aws

自己紹介

下佐粉 昭（しもさこ あきら）
アマゾン ウェブ サービス ジャパン
シニアソリューションアーキテクト (アナリティクス)

@simosako

「AWSではじめるデータレイク」
https://techiemedia.co.jp/



週刊AWS
毎週AWSのアップデートをサマリしてお届け
（「週刊AWS」で検索してください）

内容

データ活用の課題とチャレンジ

データレイクプロジェクトの進め方① インフラ

データレイクプロジェクトの進め方② 人・組織

まとめ

データ活用の課題とチャレンジ

あらゆるユーザーが
データ分析を行う時代へ

データエンジニア/データベース開発者

データアナリスト

データサイエンティスト

BI プロフェッショナル

システム管理者

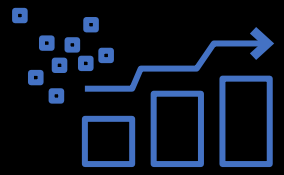
データ活用の現実

99% データによる意思決定のための積極的な投資を行っている企業の割合

24% データによる意思決定を実行できていると考える企業の数

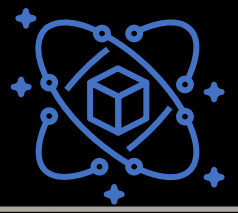
New Vantage Partners
Big Data and AI Executive Survey 2021

データ活用に向けて – 必要となるチャレンジ

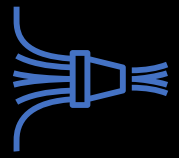


あらゆるデータに接続

だれでも簡単に分析



個人利用から大規模データまでスケール



リアルタイムデータの利用

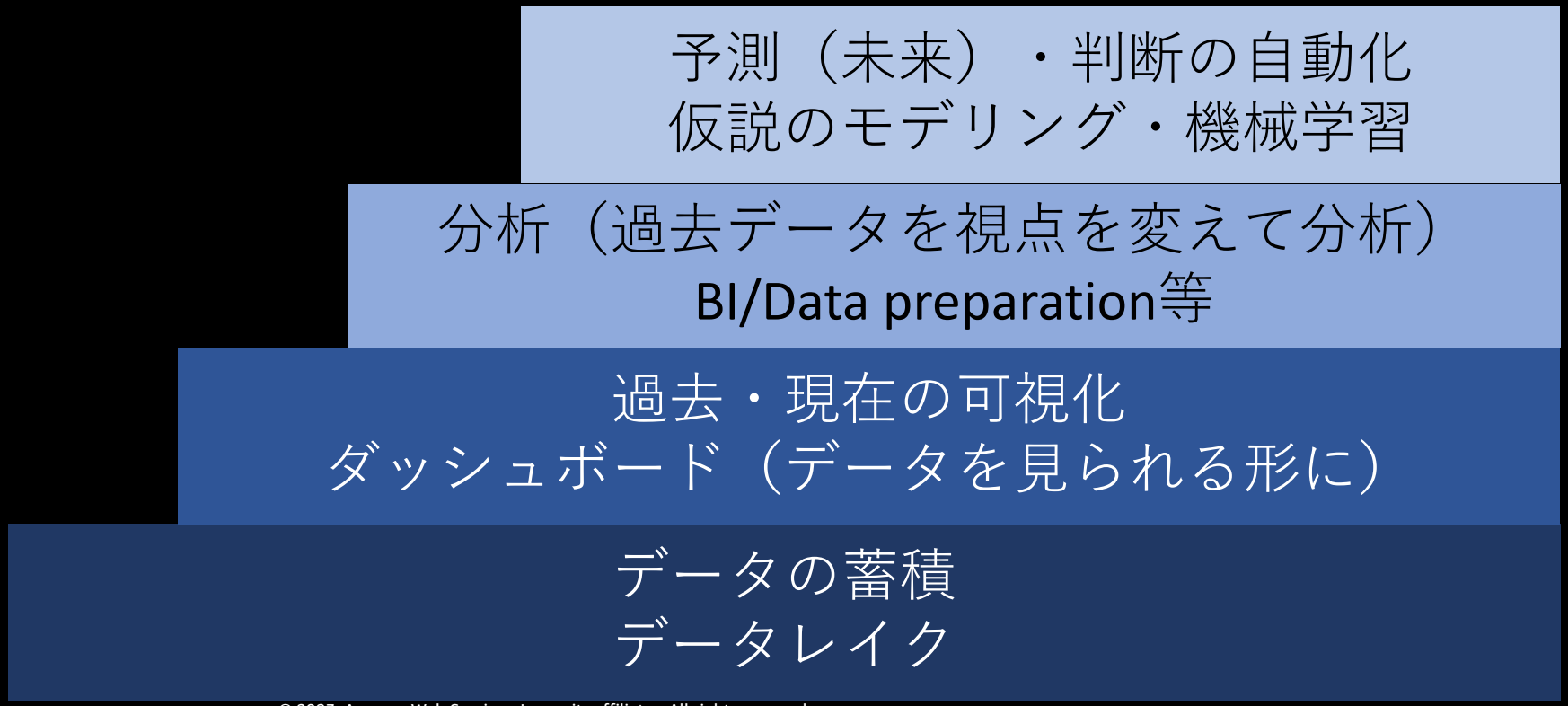
コスト最適化

セキュリティ、データ管理の一元化

画一的な分析環境によるデータ活用の限界

インフラも人も段階を踏んで高度化していく

高度な分析や予測を実化するには、その下地から作り、成長させる必要がある
ITインフラだけでなくユーザ・組織も同時に成長させていく
多様なユーザが多様なデータに触れることが当たり前になることで、
新しい発想が生まれやすい環境・文化が醸成される



データ活用が可能な環境・組織へ
データレイクのはじめかた

- 人もITも、小さくはじめて、育てていく
- 小さい目的を決めて、最小限必要なデータやITインフラから考える
- IT部門と利用部門（現場）の共同作業
- 意見を吸収して改善
- うまく動きだしたら、利用ユースケースや利用可能データを拡大していく

データレイクプロジェクトの進め方① - インフラ

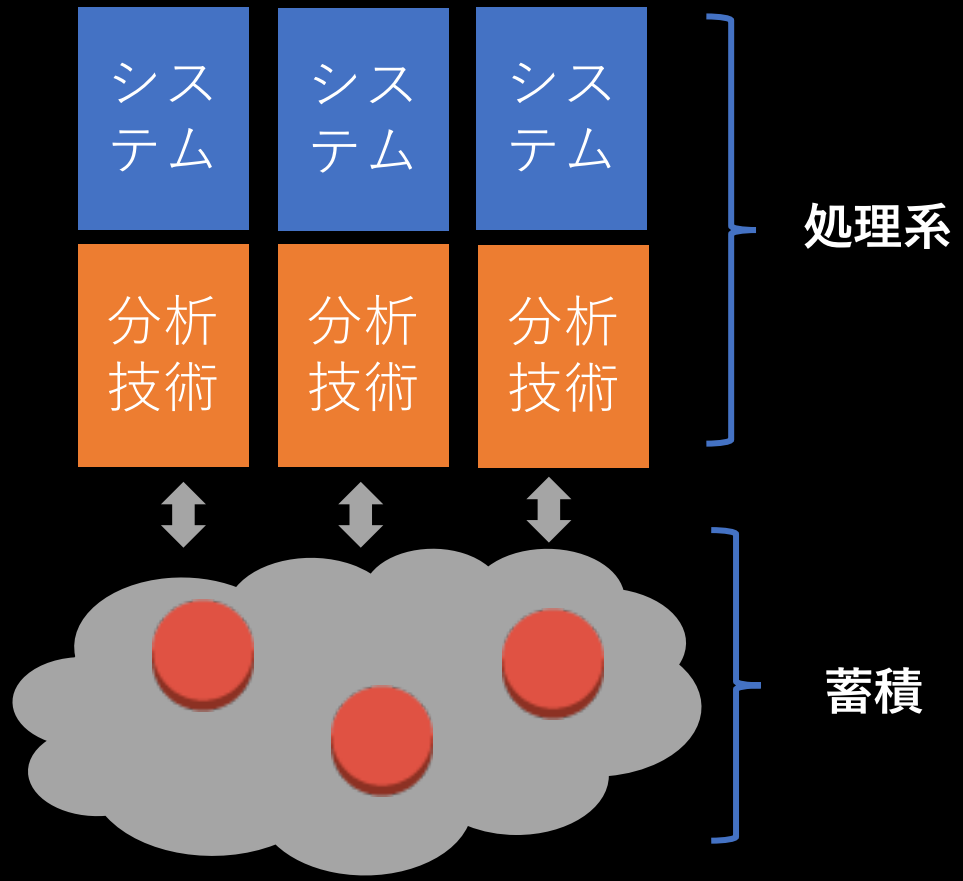
蓄積と処理系を分離することで、迅速に対応できるインフラを実現する

ライフスパンの異なる
「データ」と「処理系」を分離する

データを一か所に蓄積
- 「正」のデータが明確に
- データを捨てなくて良い

=> 新たな要件に対応できる環境

処理系が容易に切り替え可能
- 新技術に迅速に対応可能
- 社内の調整を不要に



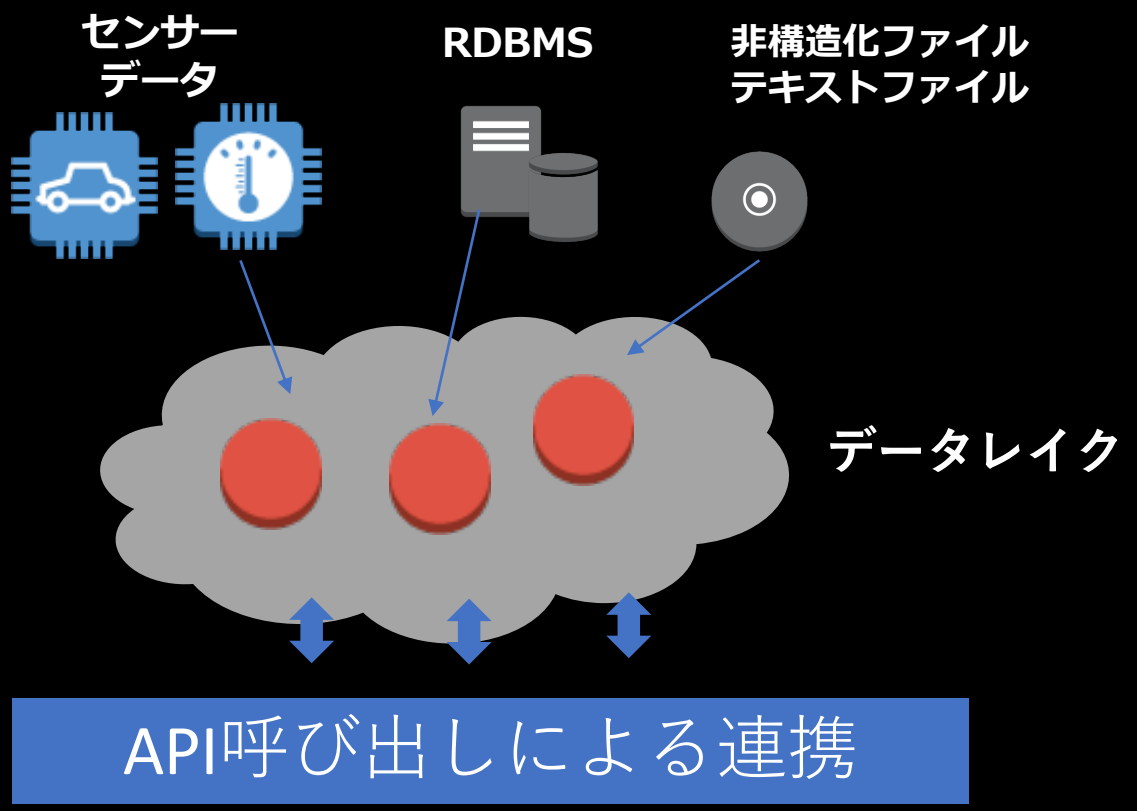
蓄積のコア機能 = データレイク

多様なデータを一元的に保存

データを失わない

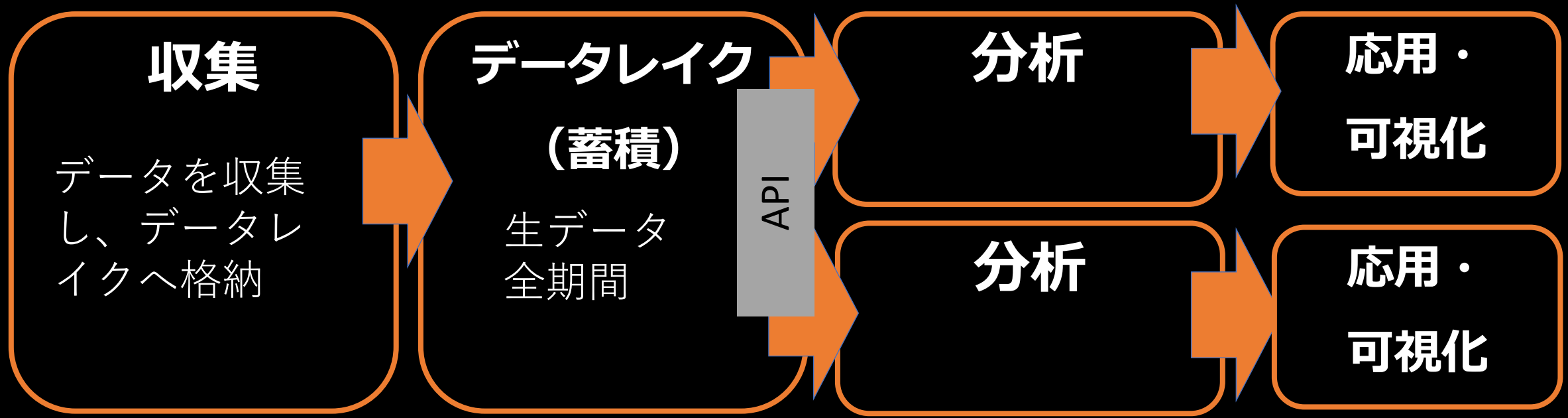
サイズ制限からの開放

決められた方法（API）で
すぐにアクセスできる



データレイクを中心とした分析環境 on AWS

- 生データをデータレイクに集め、将来的なニーズに対応
- 分析や可視化といった活用部分は取り替え可能な構成

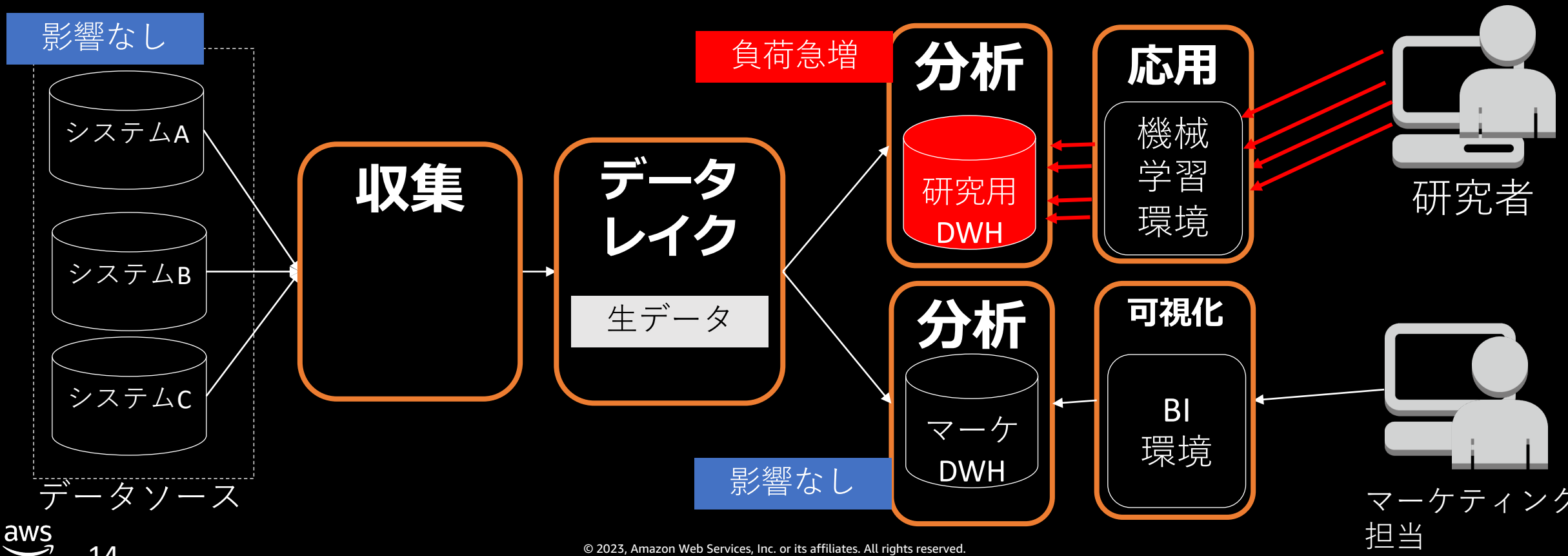


運用観点でみたデータレイク構成のメリット – 疎結合

データソースからデータレイクにデータを渡すだけなので構成が複雑化しづらい

データレイクより先（応用）がどう拡張されてもデータソースには影響がない

利用者側に合わせてDWHや機械学習環境を立てることで、それぞれの環境間での影響が発生せず、個別に調整や最適化が可能



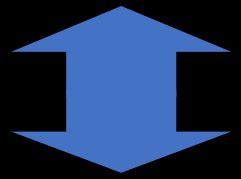
サービス選定の考え方
全てを叶える万能サービスは存在しない

例）データレイク上の標準データ作成（バッチ処理）

主眼：データ規模の増加に対応する

データサイズ：巨大（増えていく）

必要な仕組み： ITエンジニアが大規模なデータを効率的に処理可能なインフラ



例）ユーザ部門のデータ加工（データプレパレーション)

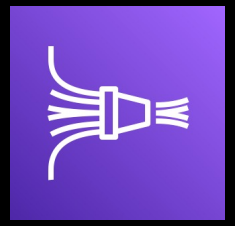
主眼：LOBの多彩なニーズに対応する

データサイズ：限定的（部門が必要な範囲）

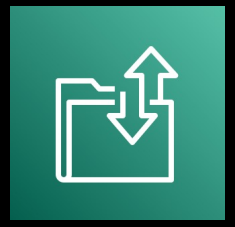
必要な仕組み：敷居が低く、誰でも必要な変形・整形処理を実行できるインフラ

AWSは多様なサービスをサーバーレスで提供
データレイクに関連するサービスもサーバーレス

データ連携



リアルタイム データ投入
Amazon Kinesis



FTP/レガシー連携
AWS Transfer ファミリー

ストレージ連携
AWS DataSync

データレイク

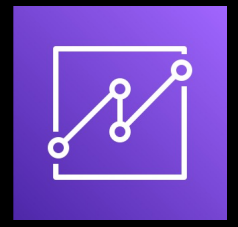
データ蓄積・ハブ
Amazon S3

ETLとデータカタログ
AWS Glue

データレイク セキュリティ
AWS Lake Formation

分析・活用

大規模DWH
Amazon Redshift Serverless



ビジネス インテリジェンス
Amazon QuickSight

Hive/Spark環境
Amazon EMR Serverless

※上記はAWSのサーバーレスサービスの一例です

補足：サーバーレスの活用による運用負荷の低減

お客様がご担当する作業

AWSが提供する機能

データレイクプロジェクトの進め方② - 人・組織

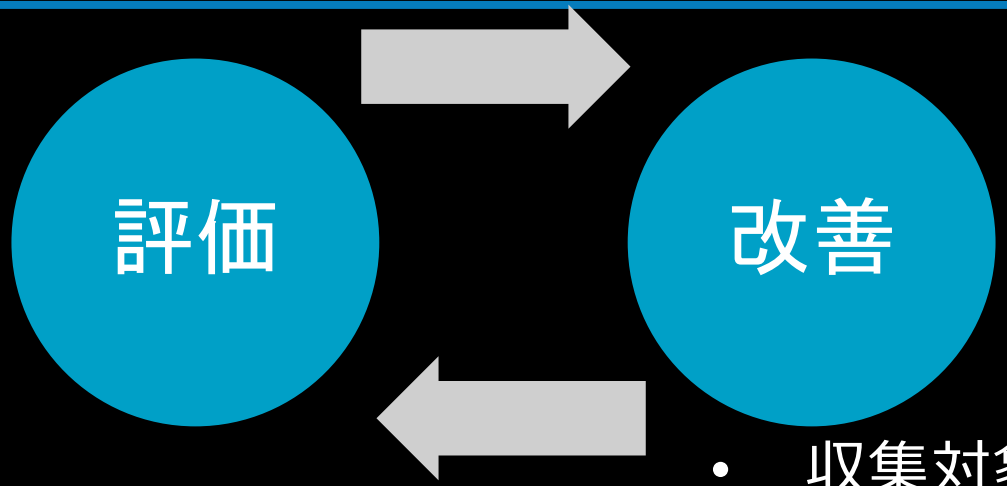
データレイクプロジェクトのポイント

データレイクは小さく、段階的につくる

- データレイクはその性質上、**最初から全てを予見するものではない**
- データレイクのコンセプト：**後から決められる、変更できる**
- 段階的に作る = **IT部門と利用部門が協力して**良いものにしていく

利用されるデータレイクの実現：評価と改善のサイクルを回し続ける

データレイクは塩漬けにせず
継続的に進化させることが大切



- 利用状況の分析
- フィードバックの収集

- 収集対象データの拡充
- 利用エンジンの拡充
- ドキュメント整備
- 利用ポリシーの改善

Conway の法則

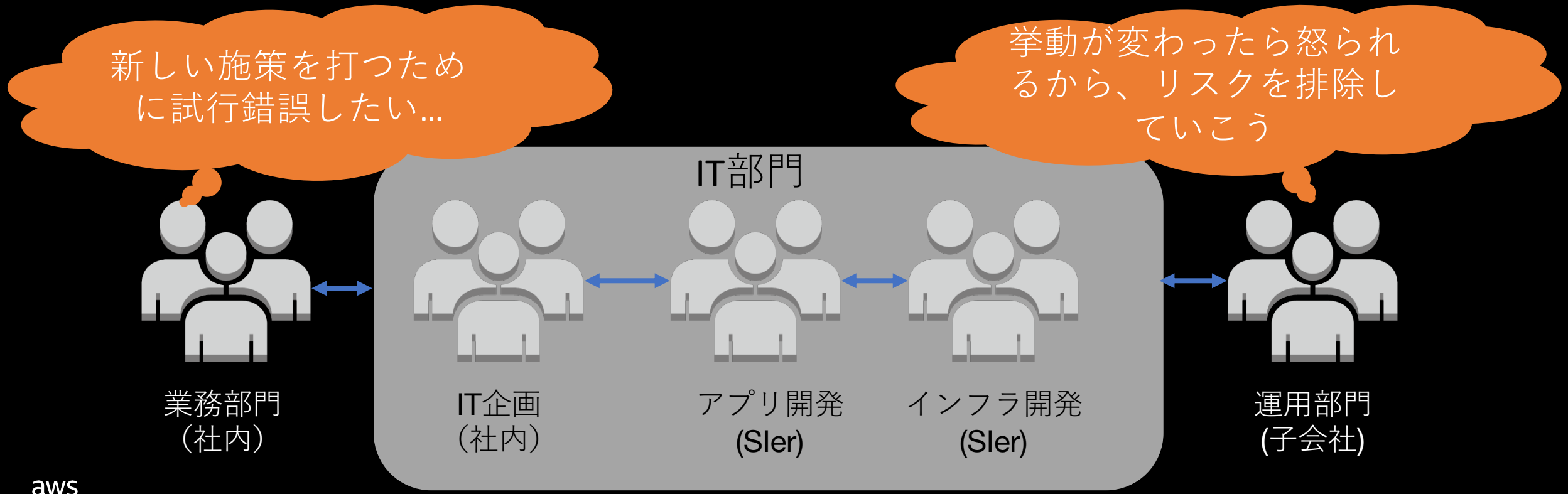
いかなる組織のシステムにおいても
生成される設計は、
組織のコミュニケーション構造のコピーとなる。
Melvin Conway, 1967

新しい試みを実施するには、
組織構造にも目を向け、改善していく必要があります

改善サイクルを支えるための組織構造の検討

利用者のニーズを吸い上げ、改善のループをまわすことができる
コミュニケーション構造になっているか？

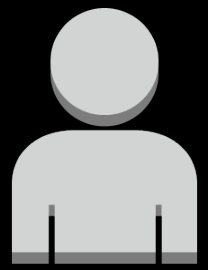
例）よくある企業の組織構造



データ利用を促進させる組織設計の例：データスチュワードが利用者とITをつなぐ

組織内のデータの統制を保つ特殊な任務を担当

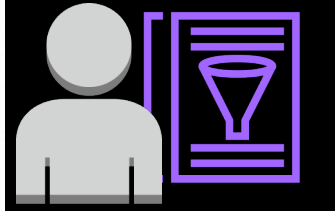
- データの内容（コンテンツ）とクオリティを管理
- ニーズに応じてデータの追加を支援する
- 利用ユーザのニーズとIT（データレイク）をつなぐ
- データレイク、周辺データのアクセス権限管理を統括



データスチュワード

- データの健全性を保つ
- データ連携の定義管理
- 利用者サポート

データレイク



データカタログ管理（メタデータ管理）

アクセス権限管理

データレイクの運用の考え方①
利用が進むよう、拡張を続ける

データの拡張

データソースの追加

データ量の増大への対応

データの整形/修正

ユースケースの拡張

アプリケーションの追加

分析エンジンの追加

PoC の実施

ルールの拡張

アクセス制御の追加/更新

データ分類定義の更新

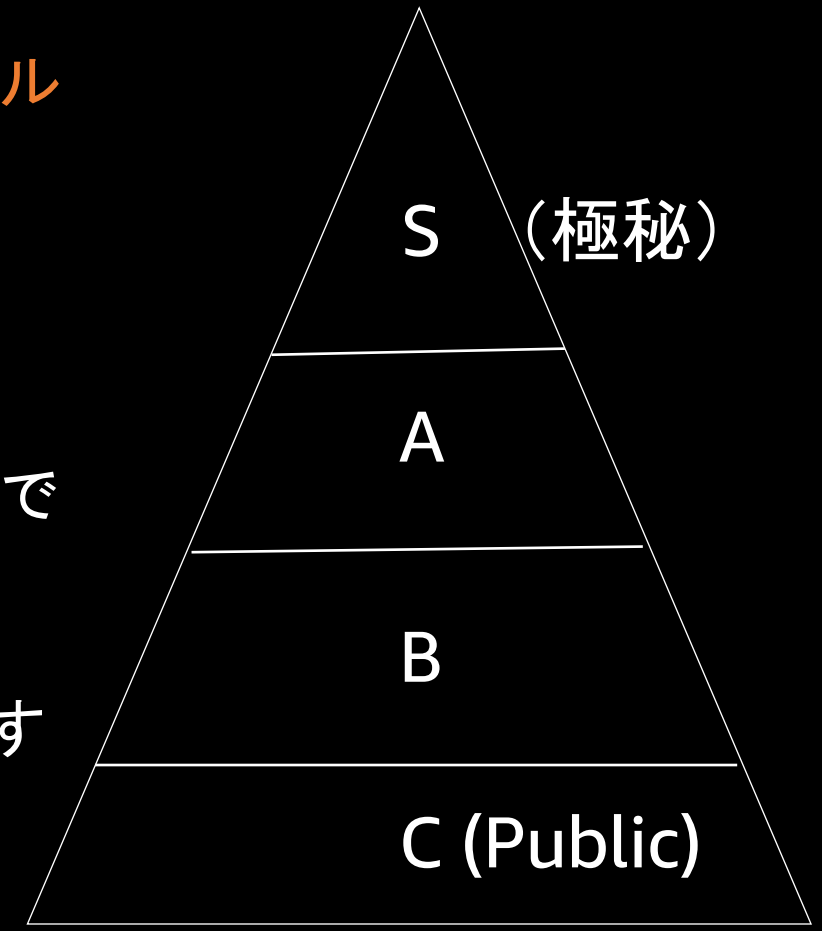
データ利用を促進させるルールの拡張の例：データのクラス分けと、その継続的な改善を支援

ここまでならこのデータを利用して良い、というルールが明確な環境ではデータ利用が加速します

＝＞データのクラス分け（ルール）を検討

クラス分けしたあとは、どうすれば下位クラスに移動できるかのルール整備と改善を継続することが重要です

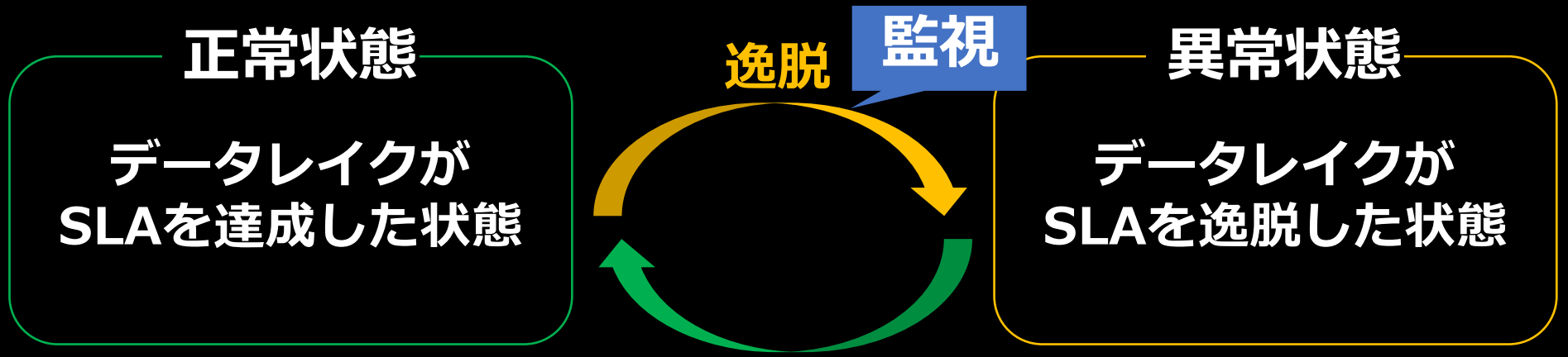
例：Sが外部に出せないとして、SをAにするにはどうすれば良いか？等を検討・改善する
（例：住所と名前をマスクする）



データレイクの運用の考え方②
データレイク全体のSLAベースで運用する

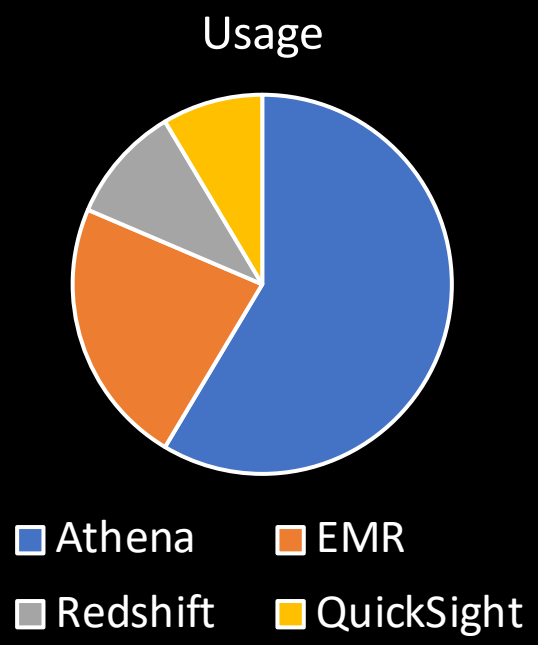
正常状態を定義し、その状態を維持する

- 正常性指標 (SLA: Service Level Agreement)ベースで運用する
- サーバーの稼働率がxx.x%といったSLAではなく、データレイクがどのような状態であれば正常かをSLAとして定義する
- 運用：監視し、逸脱した場合は正常に戻す



SLAの計測：
機械的に計測可能な部分とそうでない部分の双方で検討

継続的改善には計測(KPI化)が重要



- アクティビティ種別ごとの利用量
- 組織、職種、ユーザーごとの利用量
- 日付、時間帯ごとの利用量

AWSでは機械的に計測可能

- ユーザーからのアンケート（数値化）
- ユーザーからのフィードバック

ユーザとのコミュニケーション

正常状態のSLA定義（例）

管理対象	正常状態	SLA の例
データの鮮度	データ発生から利用可能になるまでの時間が規定内	アプリ上で出力されたログデータが**XX時間以内に利用可能**になっている
	所定の日時までに対象データが利用可能となる	レポーティング用のデータの**前処理が午前xx時までに完了**している
データの信頼性	データ起因のエラーや結果不正の発生率が規定内	分析クエリの**エラーレートがXX%以下**におさまっている
データレイクのコスト	データレイク全体のコストが規定内	ストレージ、ETL、クエリ等の**総コストがXX円/月以下**におさまっている
ユーザの活用度	ユーザ評価（5段階アンケート）等で一定の評価を得る 目標ユーザ数に到達する	四半期アンケートで**平均評価4.0以上**を維持する

まとめ：変化を織り込んだ基盤を構築し、人とITがともに成長していく

- データレイクの目的は、データと処理系を分離する構成により、変更に強いインフラを維持すること
- 未来は見通せないという前提で進めるのが、データレイクプロジェクトの特徴
- 小さく作り、フィードバックを得ながら進める
- インフラだけでなく、人・組織もいっしょに成長していく

内容についての注意点

- 本資料では2022年12月9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います。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.

AWS TRAINING & CERTIFICATION

AWS Skill Builder の 500+ の無料デジタルコースで学ぼう

30以上のAWSソリューションの中から、自分に最も関係のあるクラウドスキルとサービスにフォーカスし、自習用のデジタル学習プランとRamp-Upガイドで学ぶことができます。

- 自分のペースでAWSクラウド上を活用した未来を切り開く
- 学習プランでスキルや知識を向上
- AWS認定資格でクラウドの専門知識を証明する

自分に合ったスキルアップ方法をで学びましょう

EXPLORE.SKILLBUILDER.AWS »

AWS Builders Online Series に
ご参加いただきありがとうございます

楽しんでいただけましたか? ぜひアンケートにご協力ください。
本日のイベントに関するご意見/ご感想や今後のイベントについての
ご希望や改善のご提案などがございましたら、ぜひお聞かせください。


aws-apj-marketing@amazon.com


twitter.com/awscloud_jp


facebook.com/600986860012140


https://www.youtube.com/user/AmazonWebServicesJP


https://www.linkedin.com/showcase/aws-careers/


twitch.tv/aws

Thank you!