



データで成長する会社の 源泉データプラットフォーム 青写真としてのデータアーキテクチャ

2024-11-26 ちゅらデータ CTO 菱沼 雄太



AGENDA

- 当たり前になりつつあるデータドリブンな経営、マーケティング、プロダクトマネージメント...etc、現代の企業活動ではデータは欠かせない要素となり、データを供給するデータプラットフォームが成長の源泉として機能しています。しかしデータの利用の取り扱い量が増えるシステムの裏側では、徐々に技術的負債がたまり、アジリティが落ち、その結果、開発チームが疲弊し、最終的に経営者は式年遷宮的な大規模なシステム更改の決断を迫るようなことが起きています。このセッションではこれまでに人類が通ってきたデータプラットフォームおよびデータアーキテクチャの歴史を振り返りながら、現代においてそれらがこの問題をどう解決し、今後どうあるべきかの青写真を描こうと思います。

自己紹介

ちゅらデータ CTO
DATUM STUDIO 執行役員
菱沼雄太

SI～ゲームまでフルスタックで
やってました
今はデータエンジニアリングの人

Snowflake の DataSuperheroに
2020～2024の間で選ばれ、2021
年にはData Hero of The Yearに
も

SnowVillage の元村長

この子は

ろあちゃん



Q1、データプラットフォーム
ありますか？

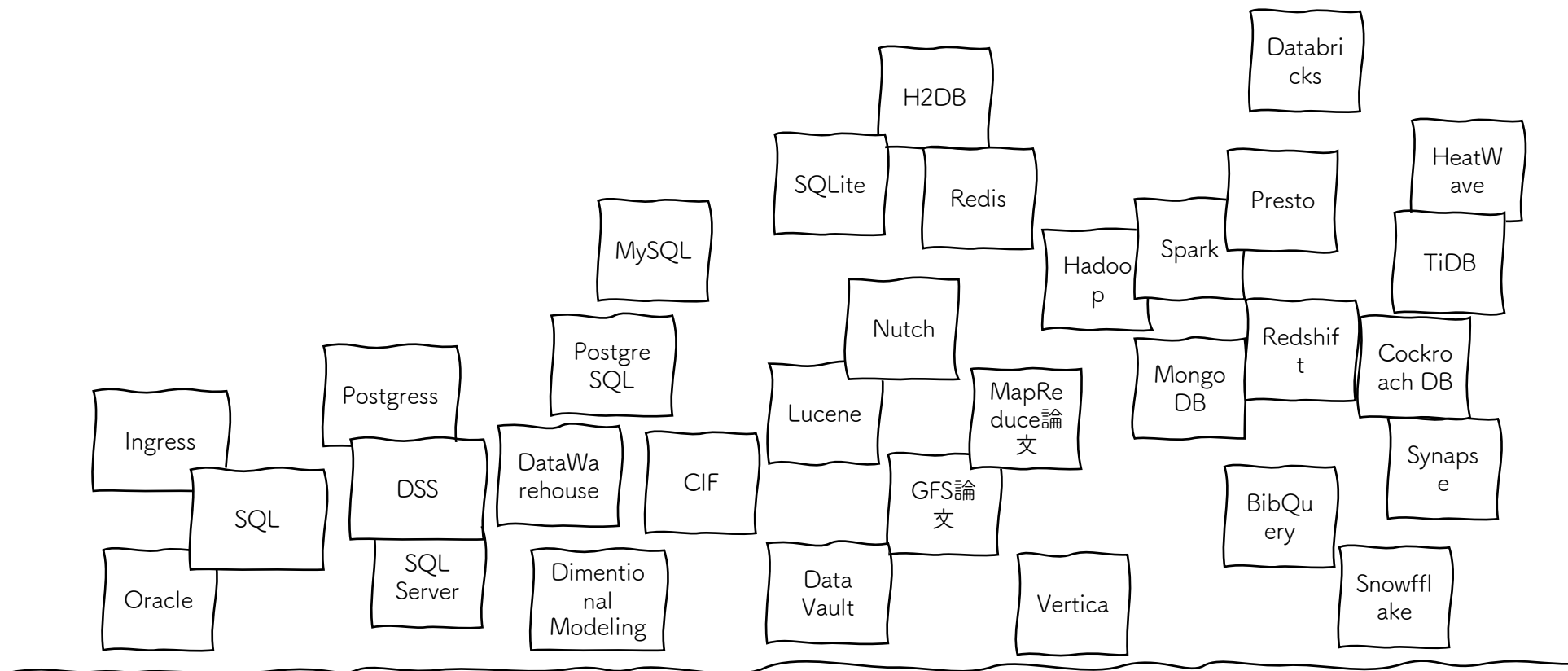
Q2、データプラットフォーム
使えていますか？

実際作ったり使ったりして
どうですか？

人類の歴史 からおさらい

人類は何をしてきたか

これまで、データとシステムについて、人類は色々やってきた



1970 RDB誕生

1990 CIF、Dimentional Modeling

2010 クラウド中心

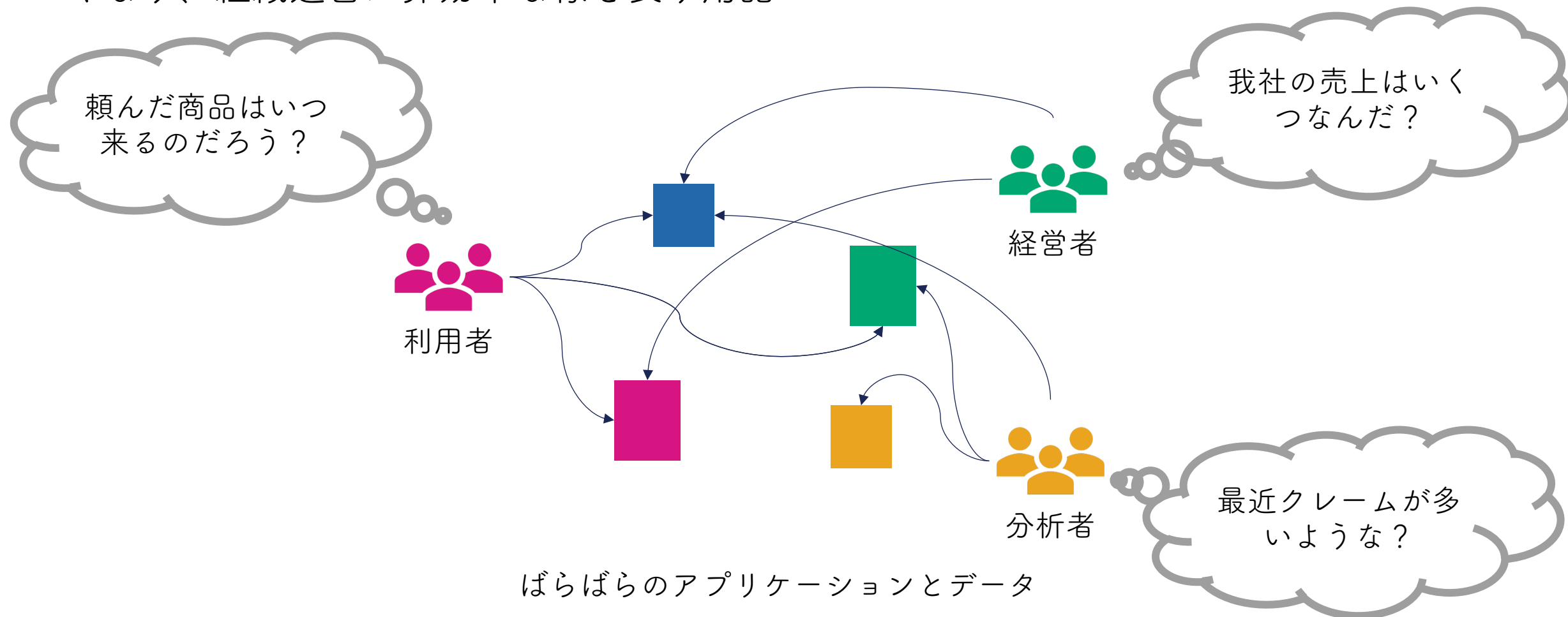
1980 データ基盤っぽい誕生

2000 ビッグデータとRDBの限界? 2020 今このあたり

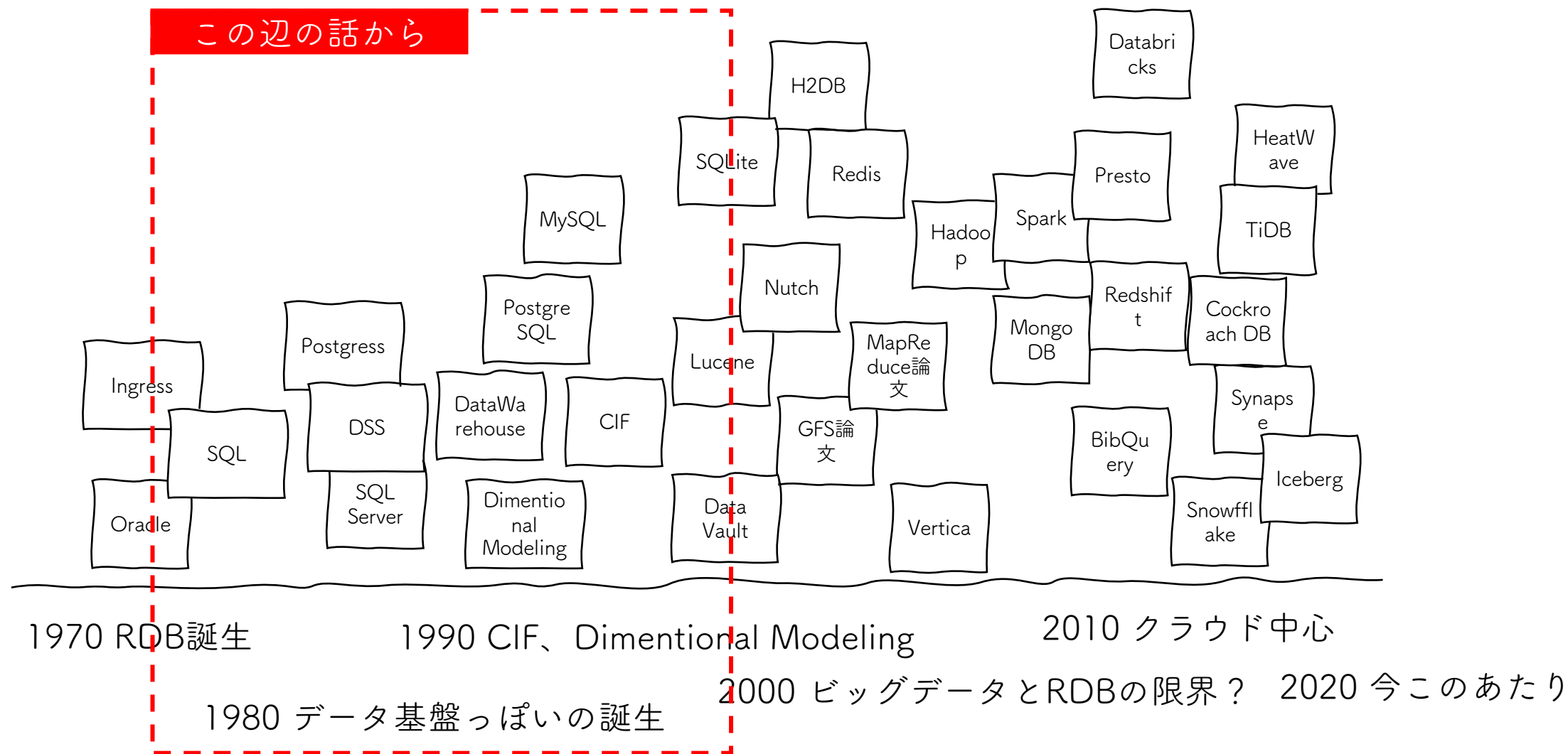
そもそもなん
でデータ基盤
が必要に？

データ基盤が必要となった大きな理由データサイロ

- たくさんのアプリケーションが乱立し、企業経営はカオスに飲み込まれた…
- これを悪名高きデータサイロと言います
 - ※サイロは穀物の腐敗を防ぐため気密性に優れた貯蔵庫、組織間の情報共有や風通しが悪くなり、組織運営が非効率な様を表す用語



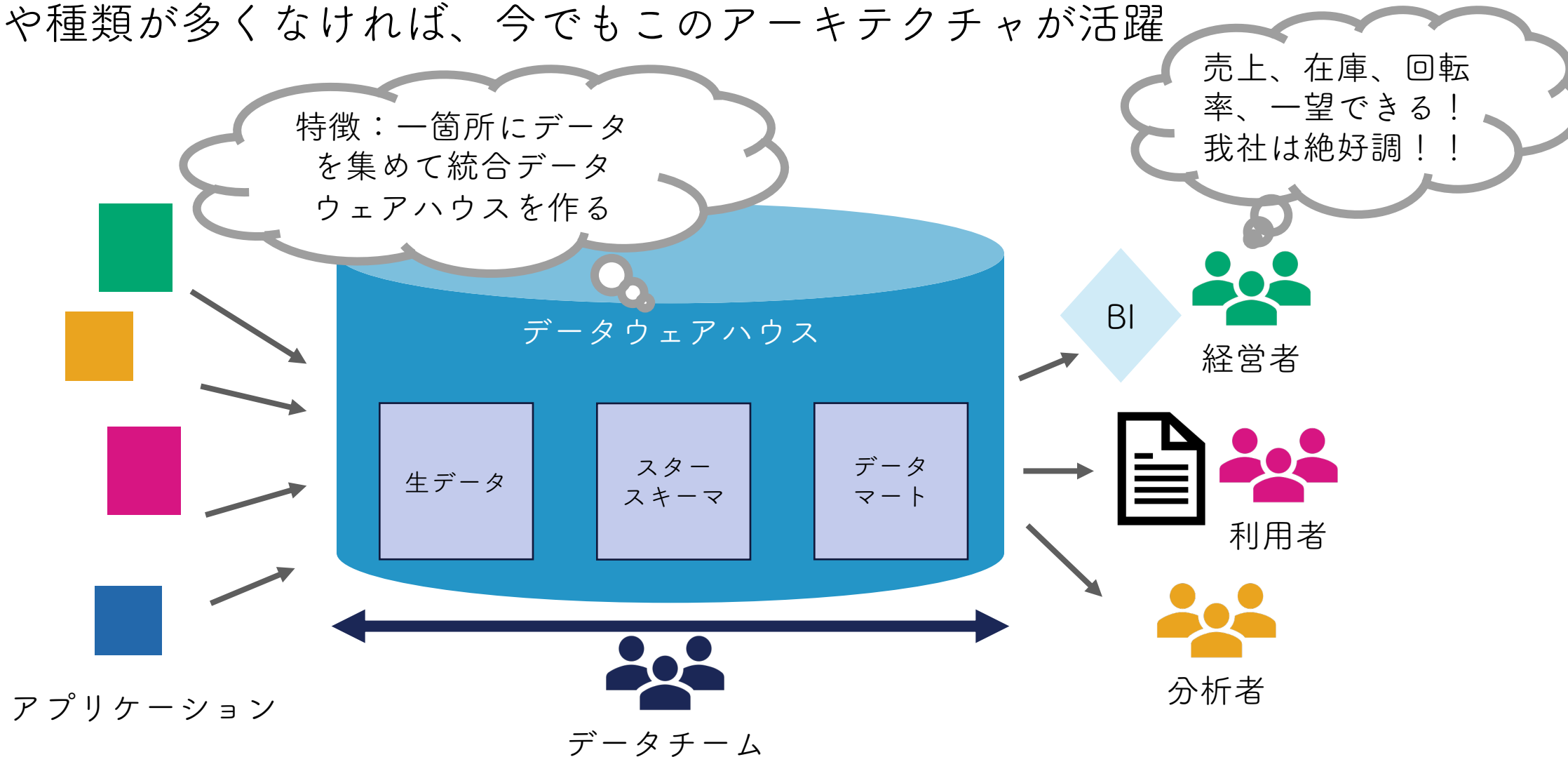
とりあえず前半部分の話をしていこう



第1世代のデー タ基盤

第1世代：エンタープライズデータウェアハウス(1970～)

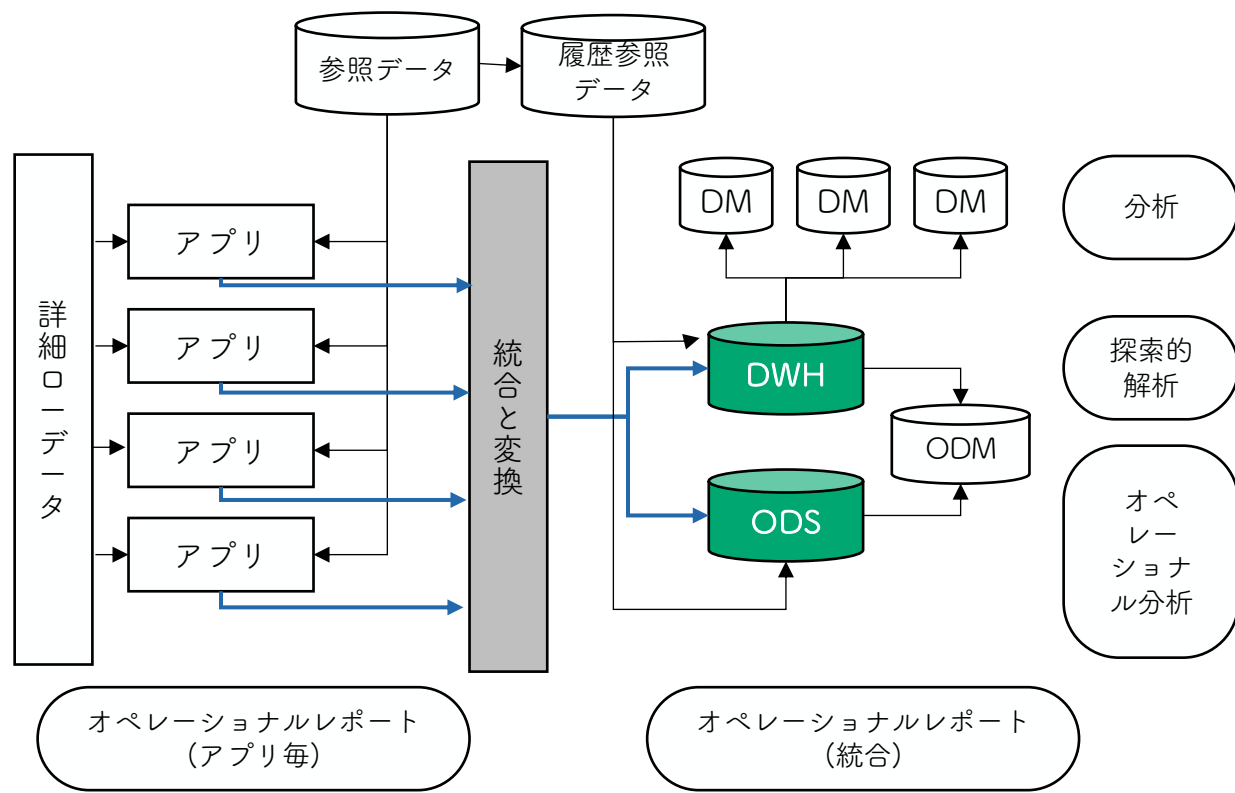
- 中央集権的なデータ基盤で、20～30テーブルくらいのデータマートから始まり、企業のデータドリブン経営、データによる意思決定を支援した
- データ量や種類が多くなければ、今でもこのアーキテクチャが活躍



その当時のアーキテクチャ

ビル・インモンが提案するデータアーキテクチャパターン

データウェアハウス自体、インモンの提唱によるものであり、正統派と言える

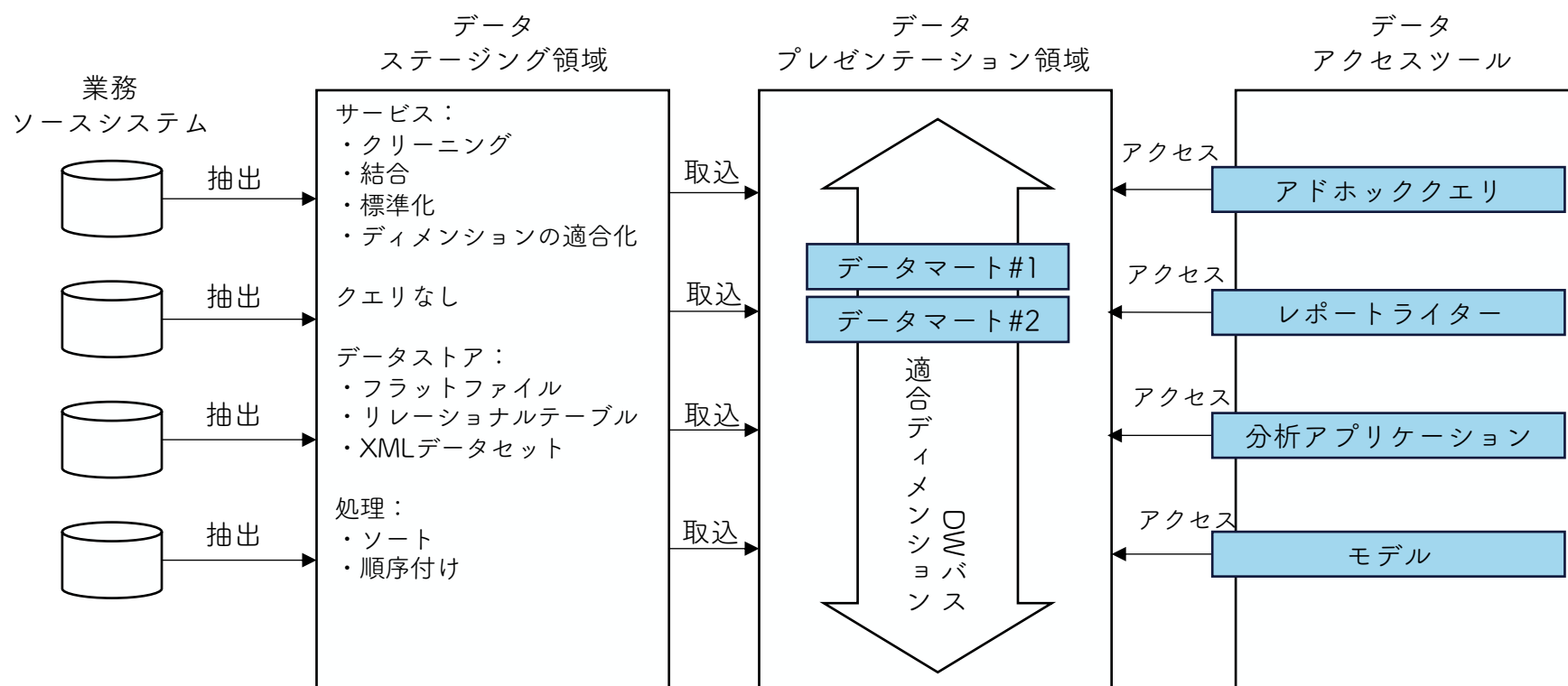


アプリ	業務アプリ、DWHとODSへデータを送るソースシステムとなる
ODS	オペレーショナルデータストア(業務DS)、業務データの統合DB、現在もしくは短期間のデータを保持
DWH	データウェアハウス、全履歴を持つ完全な正規化されたデータ統合DB、ミスデータはDWH前で修正し入れ直す
DM	データマート、分析目的に整備されたDB、DWH由来のデータをディメンショナルモデルで非正規化し保持
ODM	業務データマート、戦術意思決定用のDM、ODS由来のデータを短期間持つ
オペレーショナルレポート	業務レポート、各DBからレポートを出力する
統合と変換	アプリをソースとして、DWH、ODSをターゲットとしてデータの統合と変換を行う、ステージングとして一時的にデータを保持する場合がある

ディメンショナルデータウェアハウス

90年代～

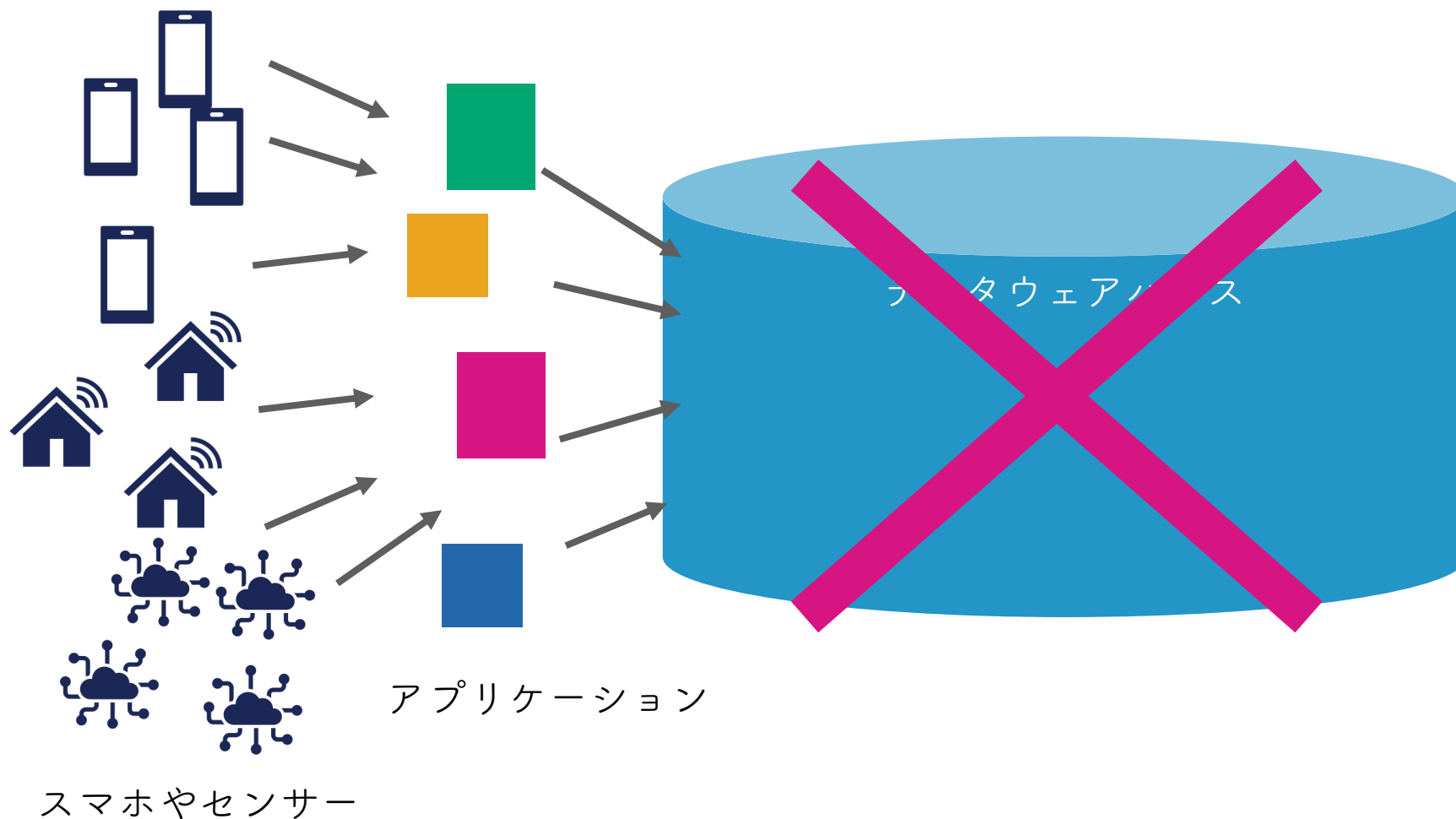
キンボールが提案するデータアーキテクチャパターン
分析のために特別に構成されるディメンショナルモデルを採用している



しかし

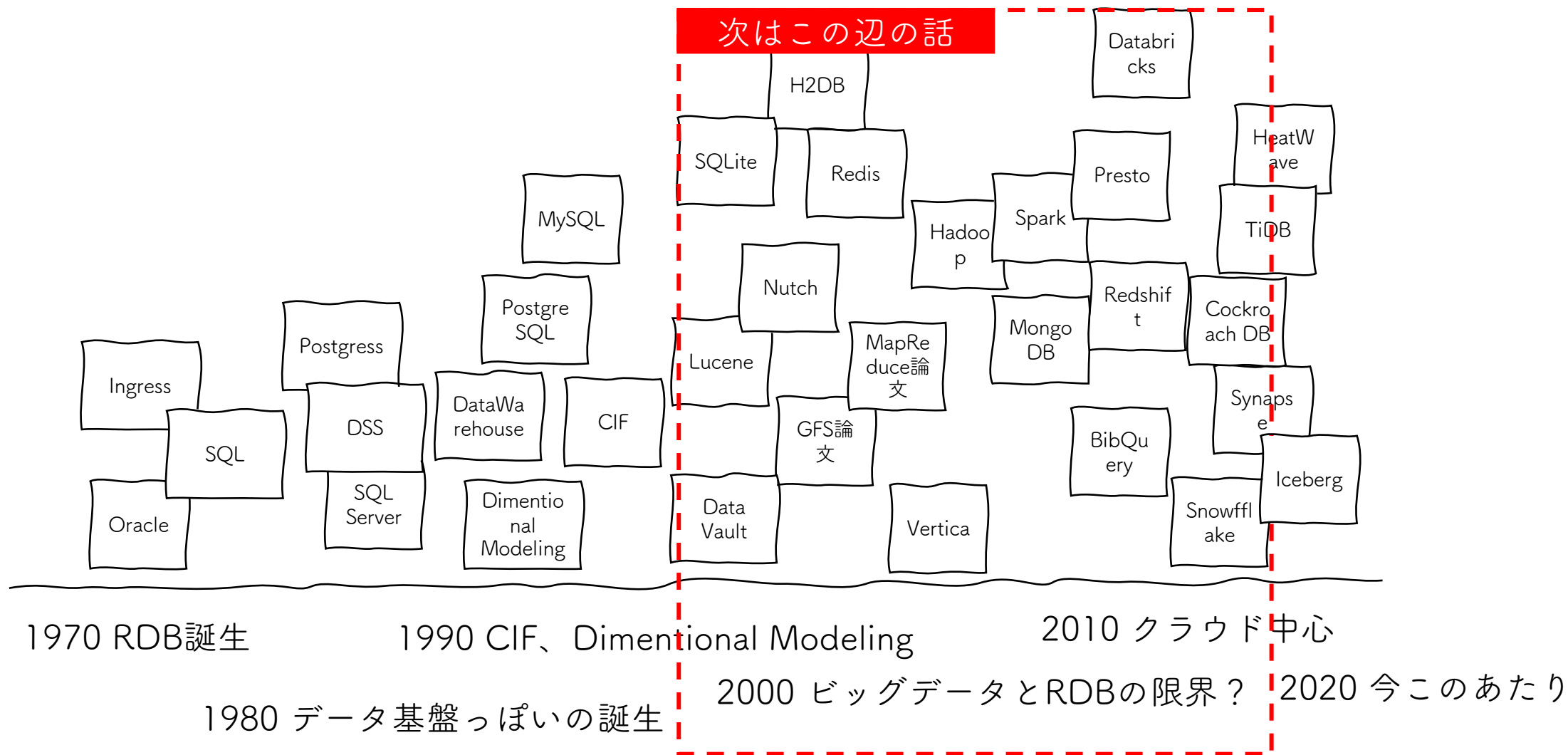
ときは2000年代…、ビッグデータが来る

- スマートフォンやWebでの活動が増え大量の行動ログが…
- 膨大なセンサーデータも扱いたくなる…
- 結果、データウェアハウスでは耐えられないデータ量に…



新たな時代に…

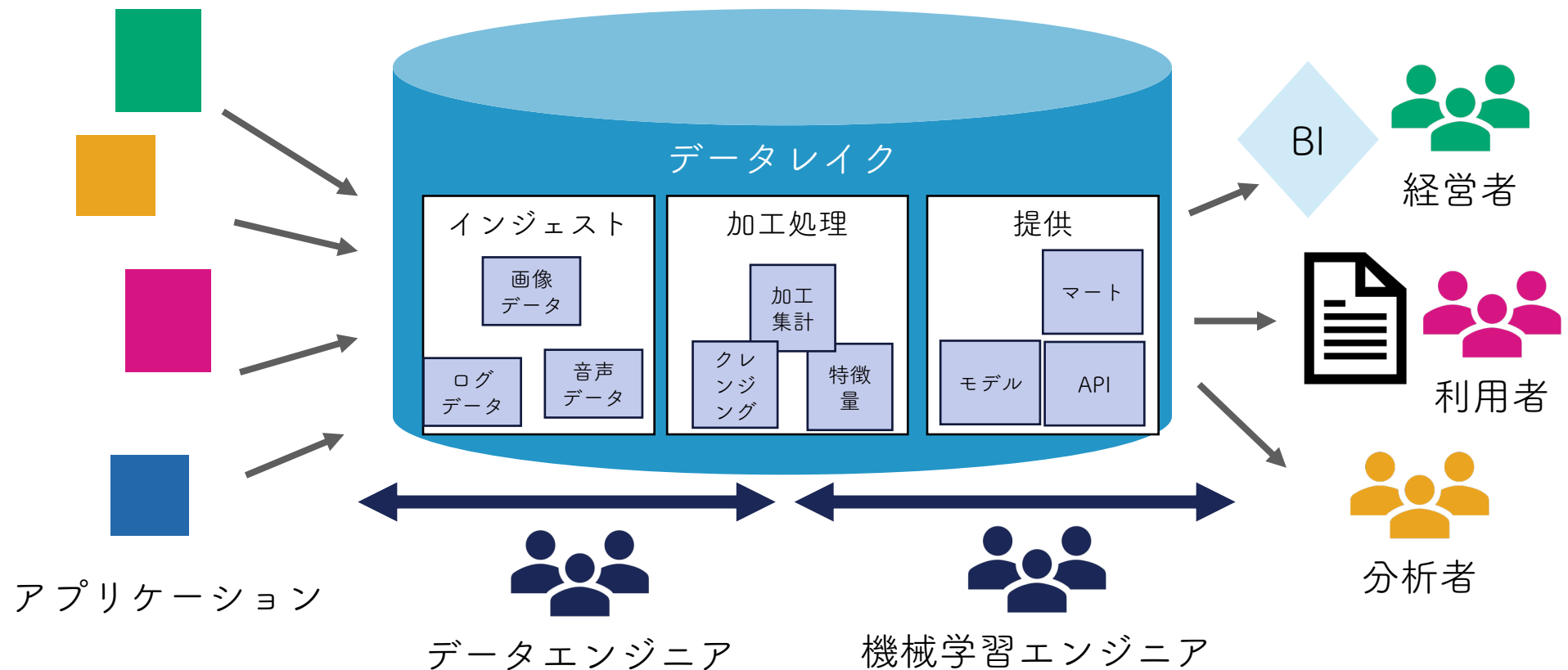
後半の話へ



第2世代のデー タ基盤

第2世代：データレイク(2010～)

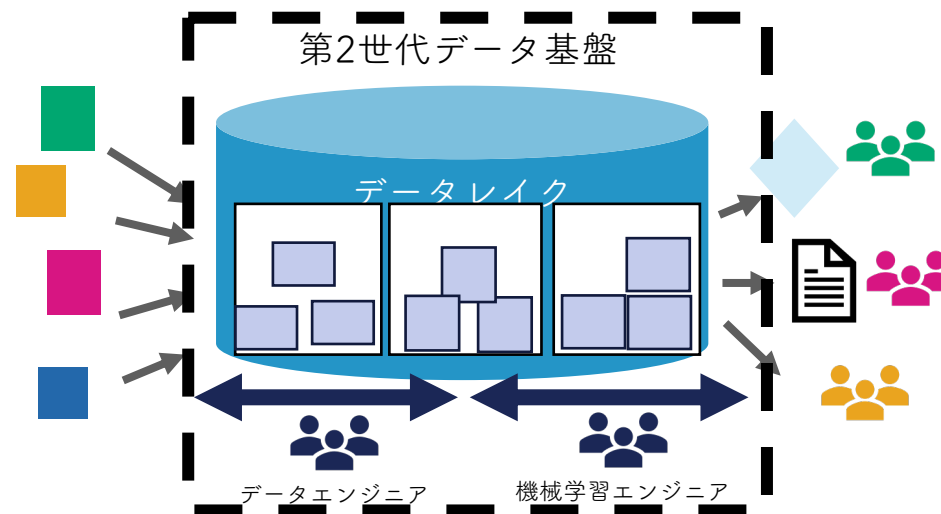
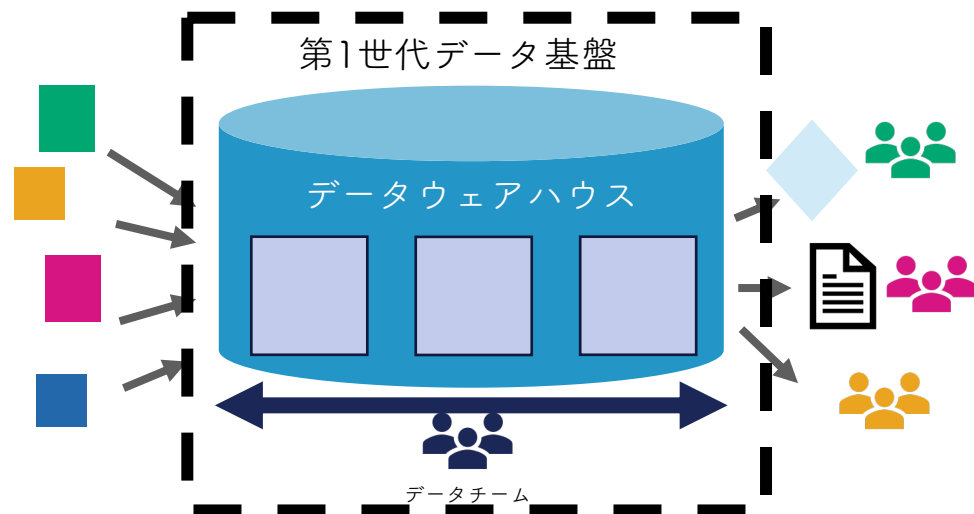
- ビッグデータの世の中ではデータレイクを中心に据えたデータ基盤が登場
- 機械学習による予測など、AIの社会実装が広まる



第1,2世代、 共通の悩み

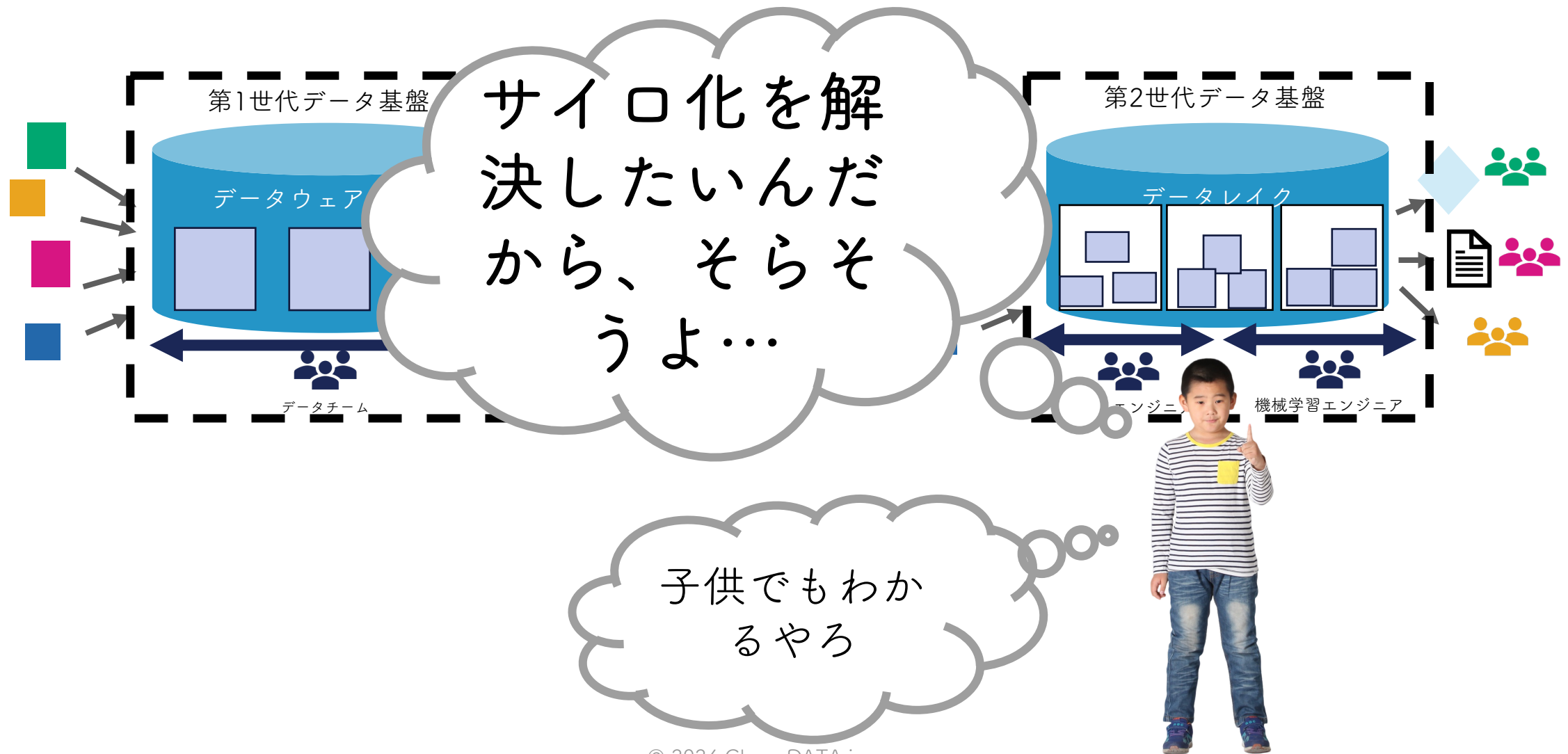
共通の構造

- 両世代はどちらも真ん中のデータ基盤がモノリシックになっている



共通の構造

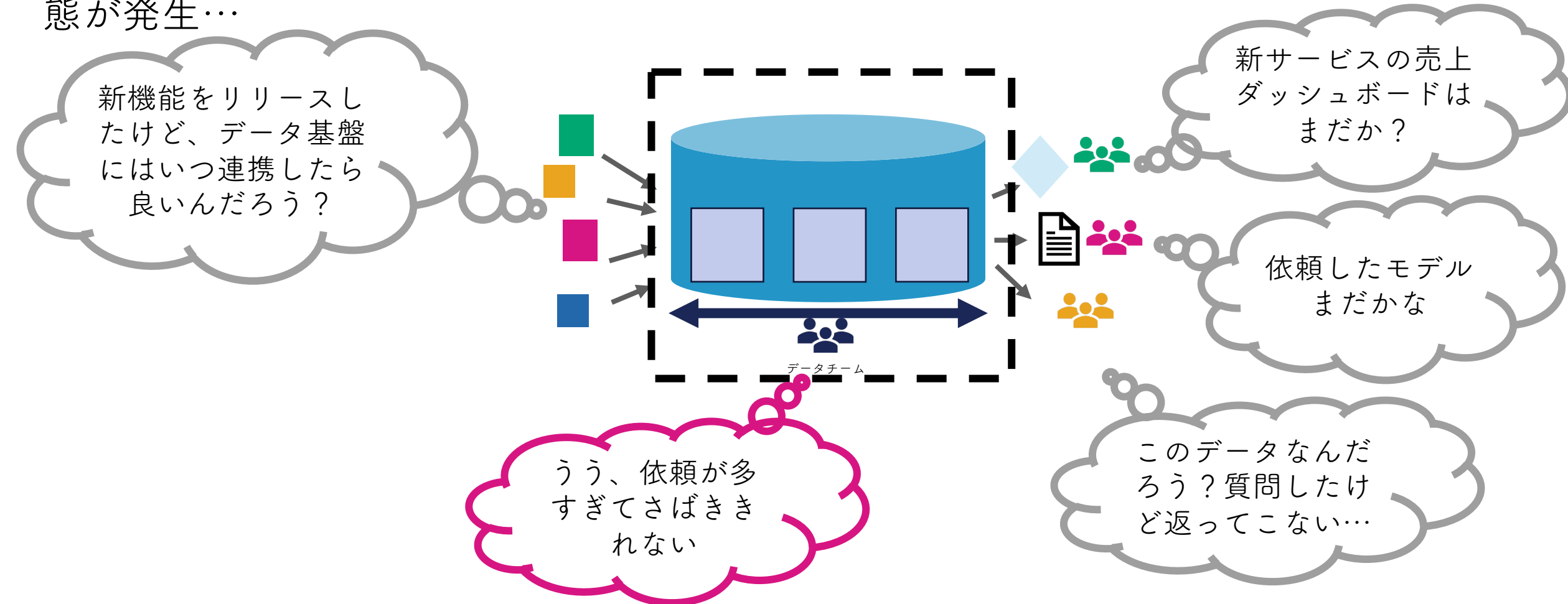
- 両世代はどちらも真ん中のデータ基盤がモノリシックになっている



モノリシッ
クの悩み

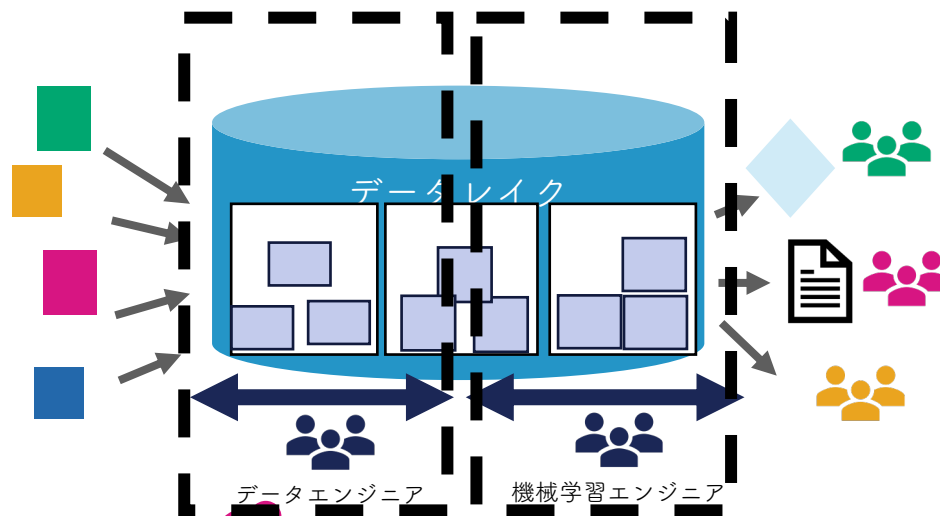
モノリシックの問題

- モノリシックであるがゆえ、データチームに負荷が集中し、全体の足を引っ張ることに…
- データ利用者の不満が増大したり、アプリケーションの開発についていけなくなる事態が発生…



データレイクの特有の問題

- さらにデータレイクでは、データチームがデータエンジニアと機械学習エンジニアに分断され、自分の担当領域で局所最適化が行われる結果、さらにボトルネックが顕著に…
- これがデータレイクモンスターと呼ばれる現象



急なデータ変更を検知
するのにデータ品質ア
ラートを作らなきゃ

本番データはまだかな？
こっちはPoC終わって
るんだけど…

さらに問題を悪化させるビジネス環境（VUCAな時代）

象徴的なコロナの時期のように、ビジネスはどんどん先が読みにくい時代へ突入していきっており、当然データプラットフォームもそれに答えることが求められる



Volatility(変動性)

Uncertainty(不確実性)

Complexity(複雑性)

Ambiguity(曖昧性)

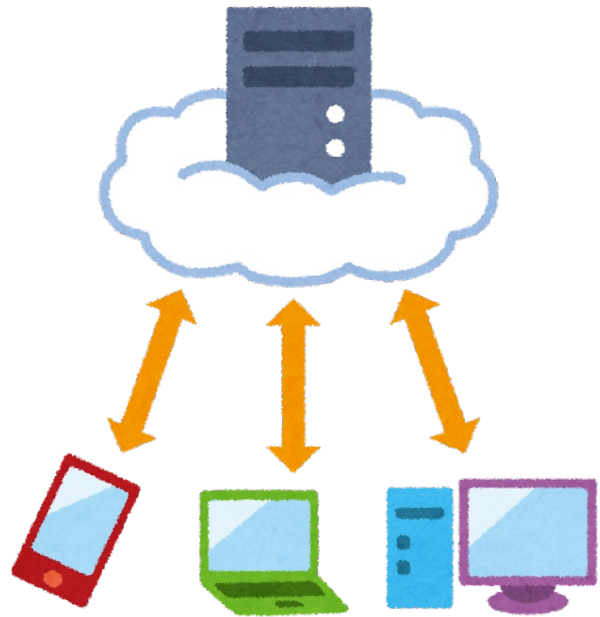
VUCAな時代

大まかな歴史の話でした

ビジネスの方にも改めて

今どきの組織の勝ちパターン

組織（≡会社）はデータを用いて意思決定することで、競合よりも良い意思決定を繰り返し競争に勝ち残れます



データを集める



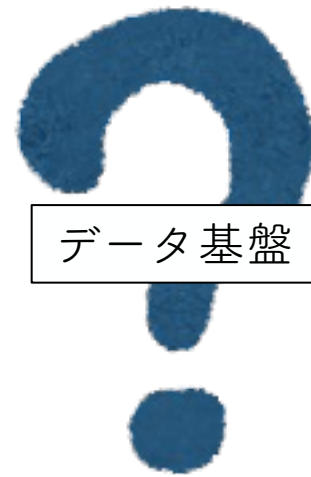
データで意思決定



大成功

組織の意思決定にはデータドリブンできる基盤が……

データドリブンに意思決定するには、一般的に「データ基盤」と呼ばれる組織のデータを集約し戦略・戦術レベルに活用可能なシステムが必要



組織のビジネス拡大を阻害する要因

組織は成長する中で、様々なイベントを経験しますが、それを乗り越えていく必要があります

顧客増加

顧客
チャネル増加

顧客種類増加

サービス増加

システム増加

ガバナンス対応増加

社員増加

部署増加

役職増加

オタクしか使わないサービスを一般人が使うとトラブル起こるよなあ

成長自体が成長を阻む要因となる

データ基盤はスケーラブルであることが求められる

組織の意思決定を支えるデータ基盤には、組織の成長に伴い生まれる成長阻害要因に対抗するスケーラブルな特性が要求されます

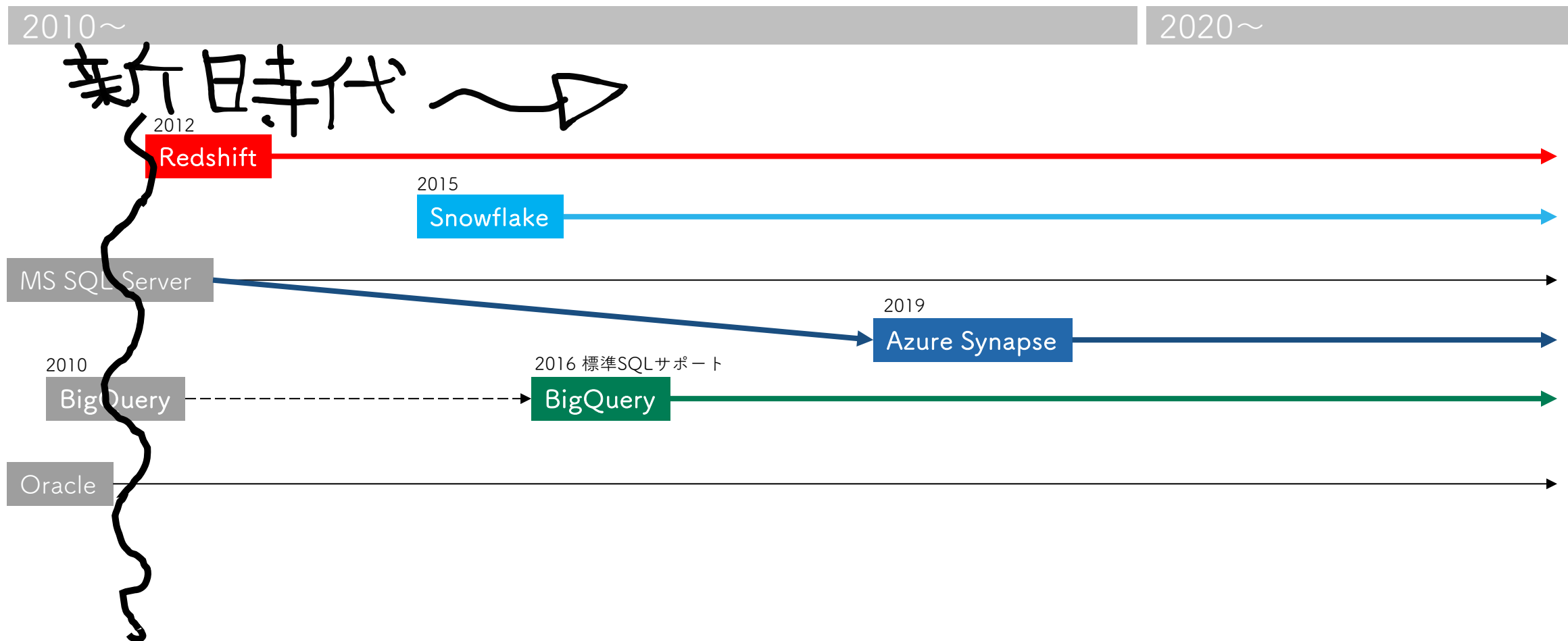


インフィニットスケーラブル
データ基盤

※スケーラブル＝拡張可能

インフィニットスケールなクラウドデータウェアハウス

2012年に登場したRedshift(AWSのOLAP型のDB、データウェアハウス)はこの世界に事実上無限にスケールするデータ基盤の可能性を示した



でも、システムの的にはできるかもしれないが

世界はエンジニアにとって相変わらず厳しい

※非エンジニアもそうせざるを得ないだけなので理解はする

新しいビジネスが増えたんだ

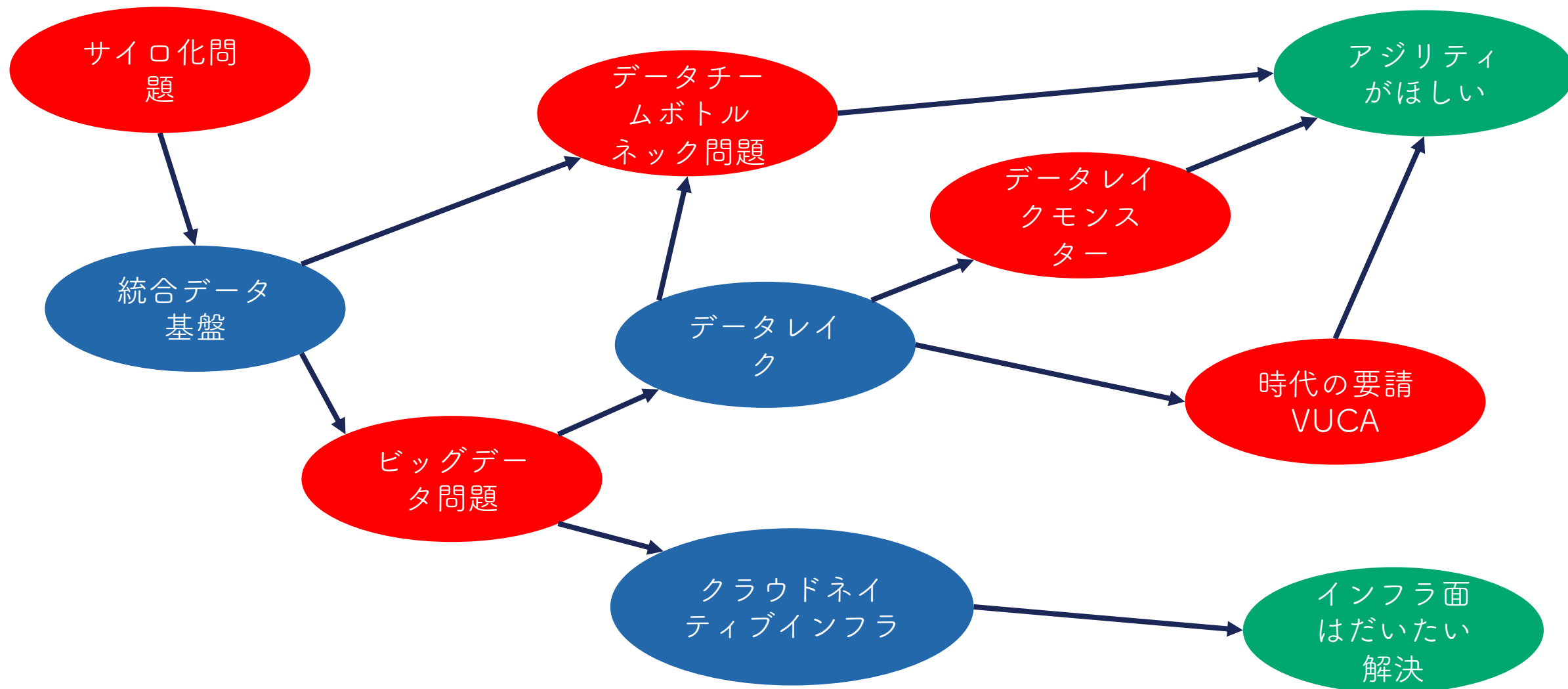


新しいデータ持ってきたよ



データ基盤システムの歩み

こうして人類はさまざまな問題を試行錯誤しながら解決してきた



どうやってアジリティの高い
データ基盤を作っていくか

どうやってアジ
リティの高い
データ基盤を
作っていか

の 高い

データ基盤

そんな今
青写真があると良いと思いませんか？

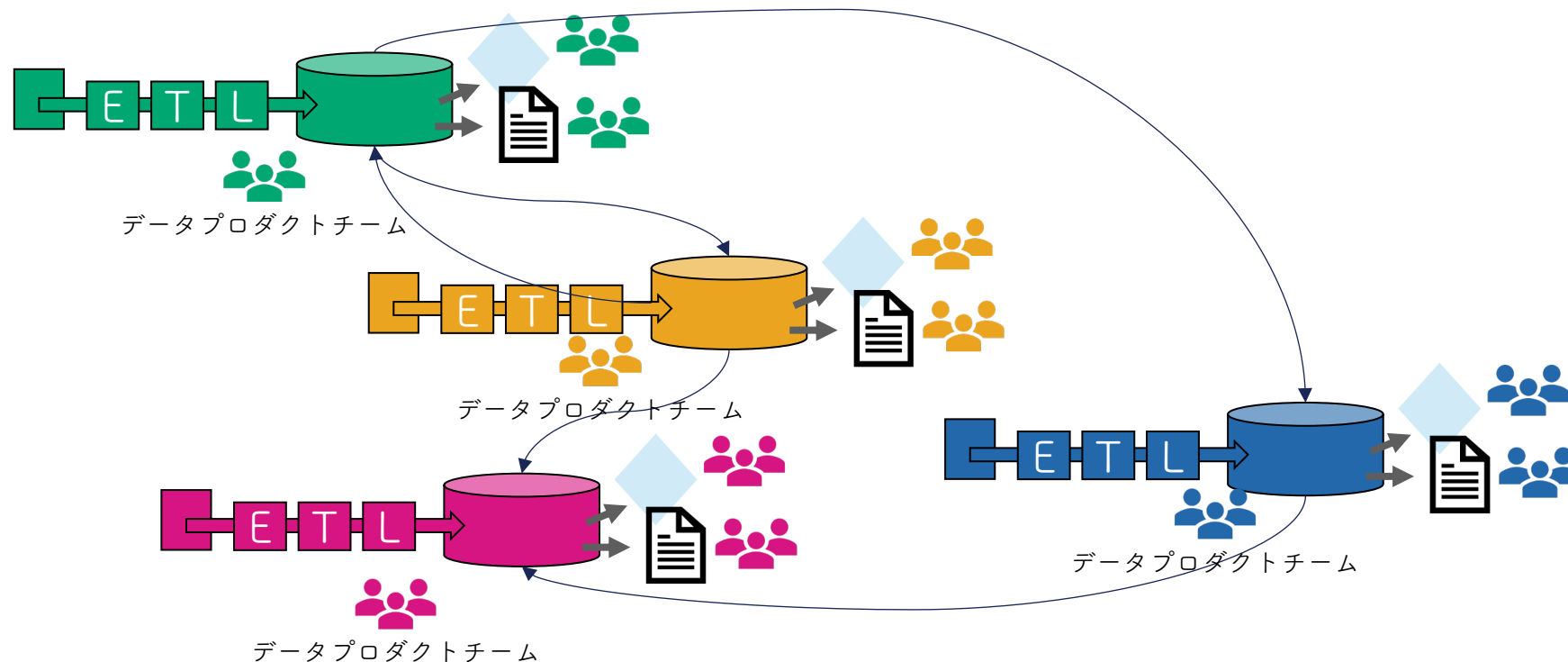
青写真とは
図面などの複写に使う写真の一種。図形や文字が青地に白く出る。比喩的に、**見取り図的な将来計画。**

Oxford Languagesより

アーキテクチャとかデザインパターンって先人たちの知恵
だから、そういうのはふんだんに取り入れていきたい

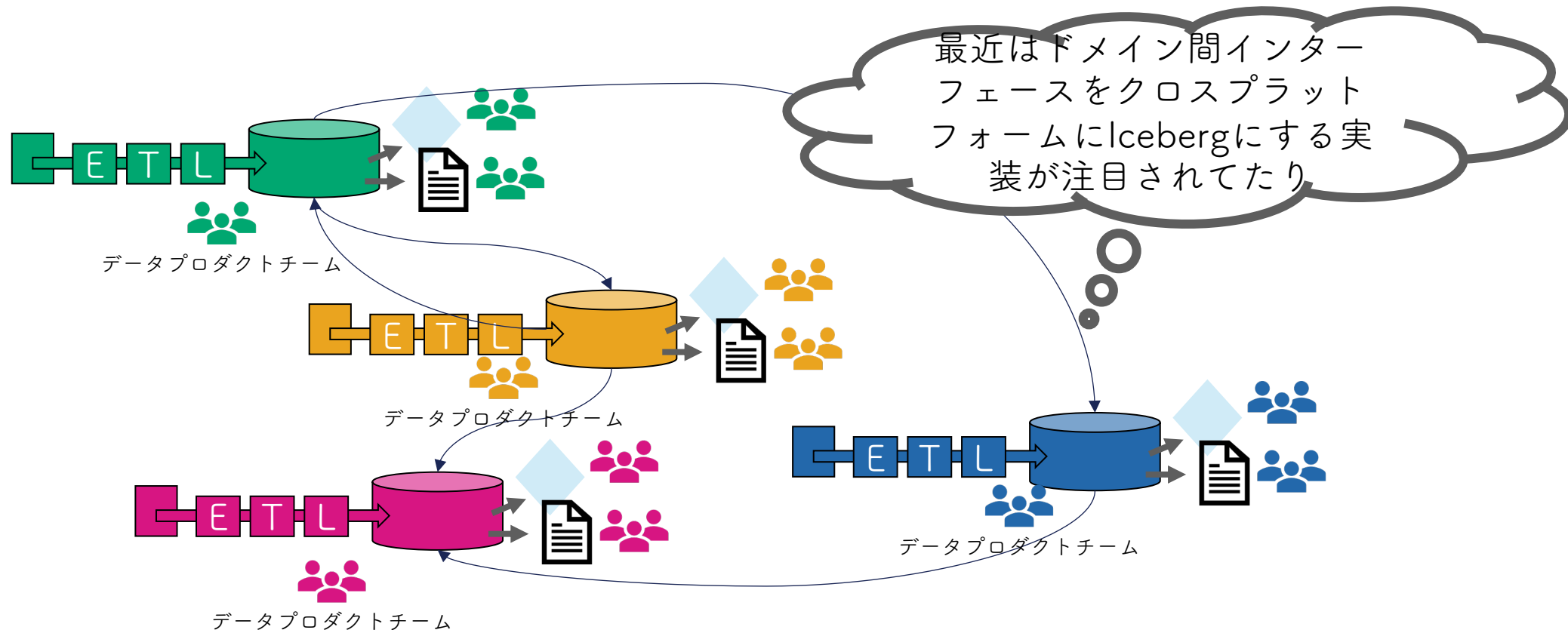
第3世代：データメッシュ(2019～)

- データメッシュはビジネスドメインごとに自身で完結したデータ基盤を持つアーキテクチャ
- ビジネスドメイン間はインターフェース(API、GraphQL、SQLなど)でつながる
- 各ビジネスドメインにはデータプロダクトチームを設置し製品と同様の開発を行う



第3世代：データメッシュ(2019～)

- データメッシュはビジネスドメインごとに自身で完結したデータ基盤を持つアーキテクチャ
- ビジネスドメイン間はインターフェース(API、GraphQL、SQLなど)でつながる
- 各ビジネスドメインにはデータプロダクトチームを設置し製品と同様の開発を行う

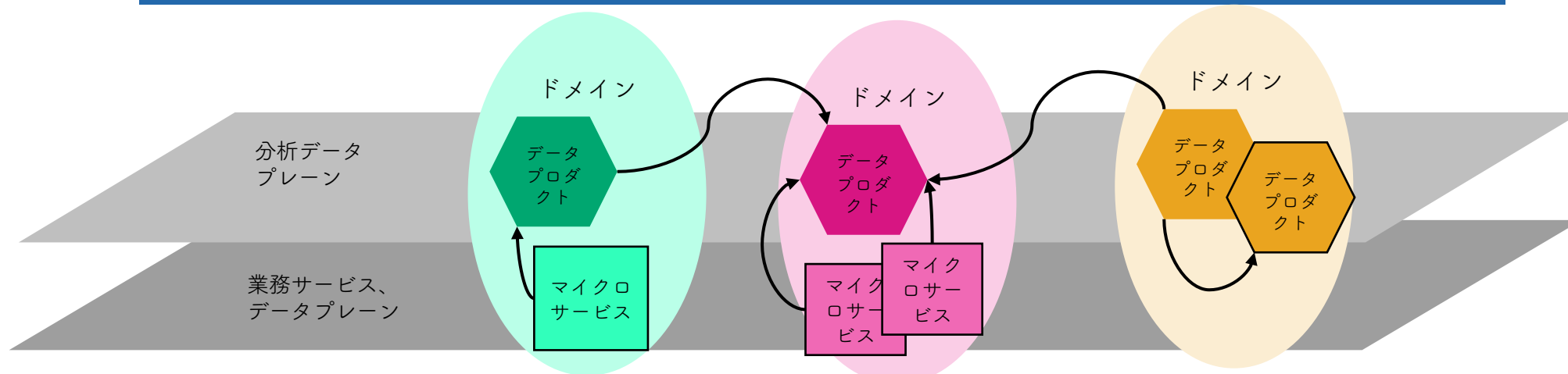


統合から分散へ

データメッシュの原則

- Domain Ownership … データを所有するのはビジネスドメインだ
- Data as a Product … データは製品とおなじだ
- Self-Serve Data Platform … クソ便利なデータプラットフォーム
- Federated Computational Governance … 連合型で自動化されたガバナンス

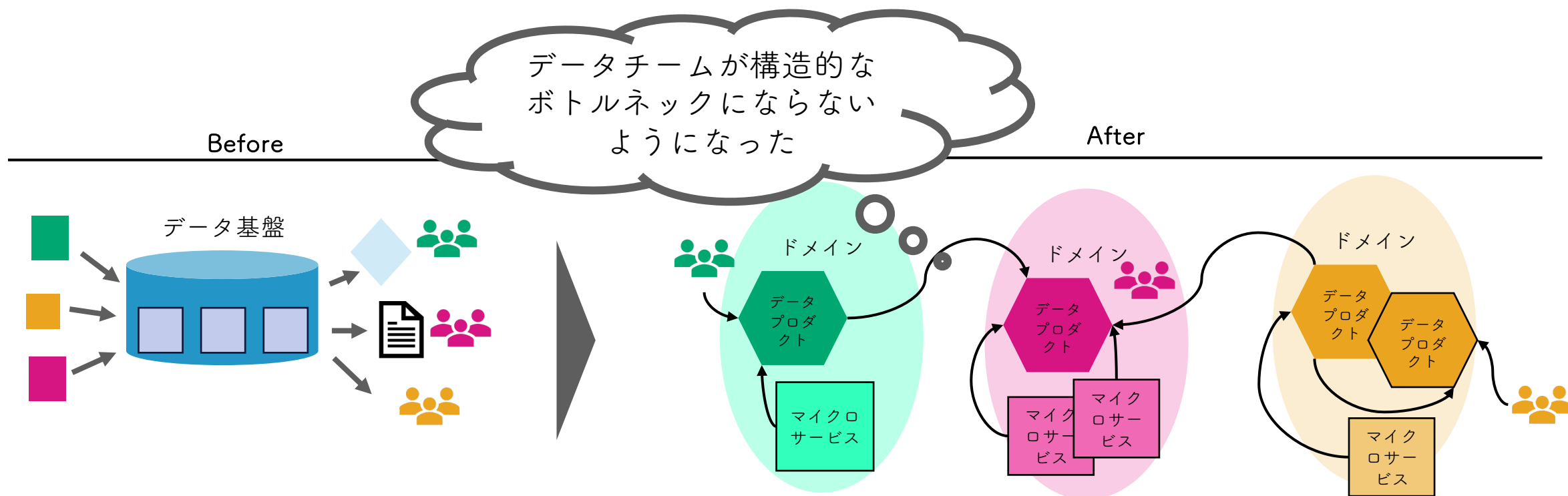
連合型ガバナンス
(アクセス制御、データカタログ…)



(ストレージ、オーケストレーション、モニタリング…)
共有データプラットフォーム

Domain Ownership

データメッシュにおいてはデータ基盤はビジネスドメインごとに分割され、それぞれのドメインがデータ基盤を所有する



- 各ドメインからデータ基盤へデータパイプラインを構築
- データ基盤をデータチームがメンテナンス

- ドメインチームが、それぞれデータ基盤(データプロダクト)をメンテナンス
- ドメイン内でデータパイプラインを構築
- マイクロサービスの開発に合わせてドメインチームがデータ基盤を継続的に開発

チームトポロジ

チームトポロジと各チーム

説明

チームトポロジはチーム間の依存を減らすことで、認知負荷を下げ、チームをセルフサービス化することでフローを最大化する開発チームの編成方法論

ストリームアラインドチーム

ビジネス要求に沿って配置するチームで、フロー最大化を目指す、職能横断型チームで自分たちですべてデリバリできる

イネイブリングチーム

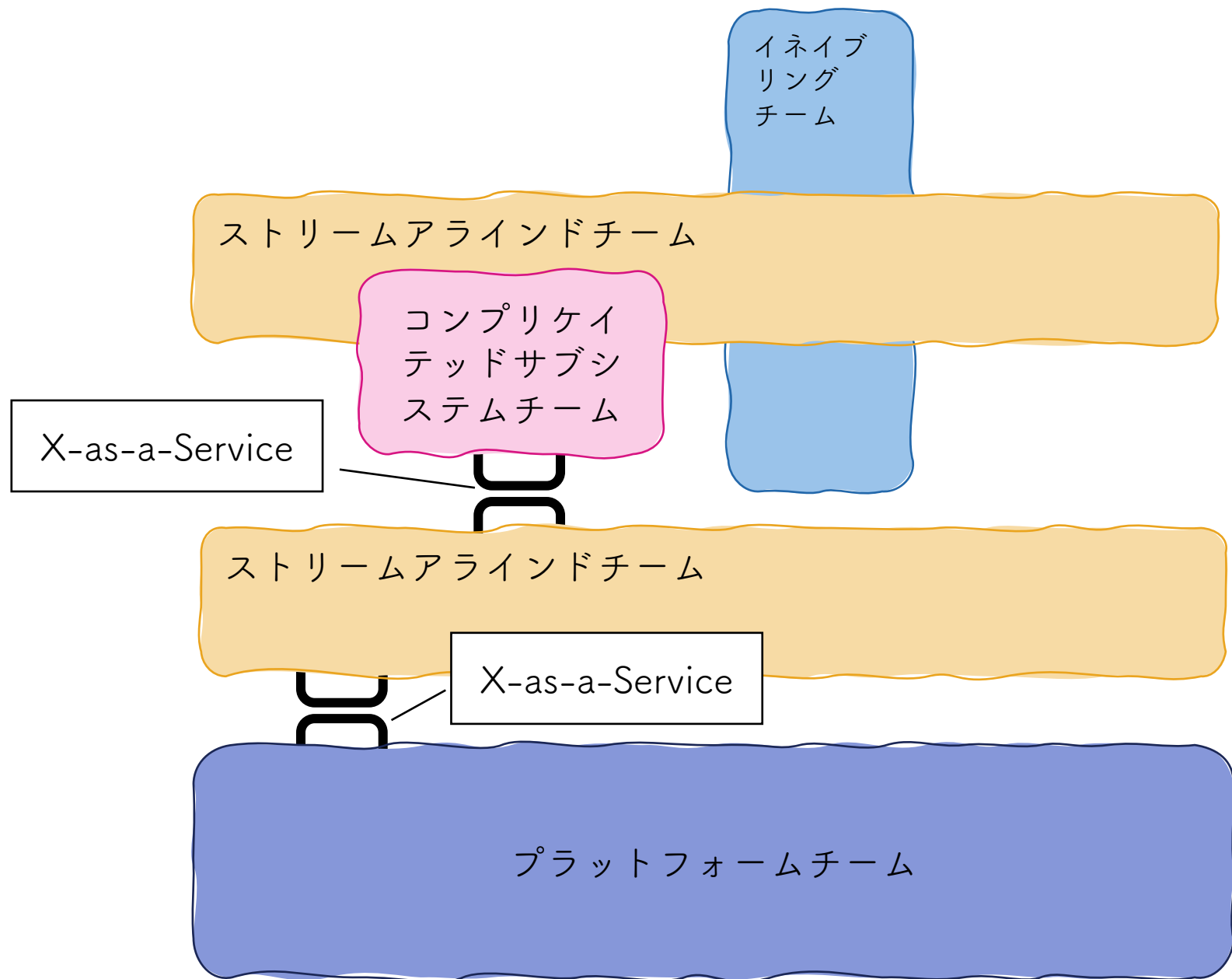
他のチームの支援やテクニカルコンサルをするチームで、特定領域の技術に強かったりする

コンプリケイテッドサブシステムチーム

複数のサービスから利用されるような複雑度の高いコンポーネントをあつかうスペシャリストチーム

プラットフォームチーム

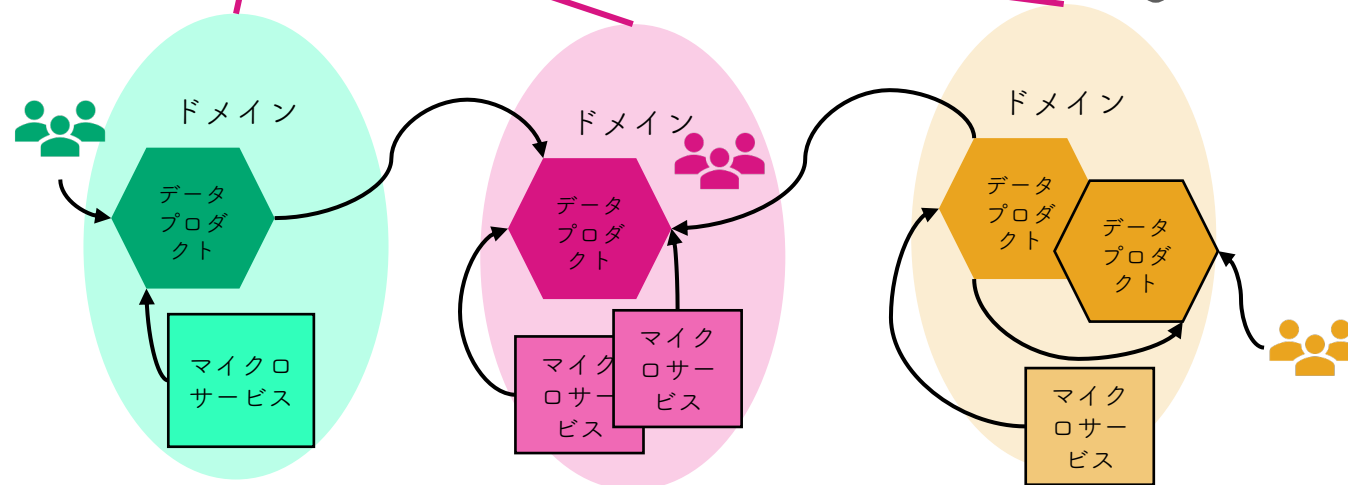
インフラや共通サービスを内部向けに提供するチームで、プラットフォームエンジニアリングを行なっている



データメッシュとチームトポロジ

ストリームアラインドチーム
ビジネス要求に沿って配置するチームで、フロー最大化を目指す、職能横断型チームで自分たちですべてデリバリできる

データメッシュとチームトポロジを組み合わせる



(ストレージ、オーケストレーション、モニタリング...)
共有データプラットフォーム

プラットフォームチーム
インフラや共通サービスを内部向けに提供するチームで、プラットフォームエンジニアリングを行なっている

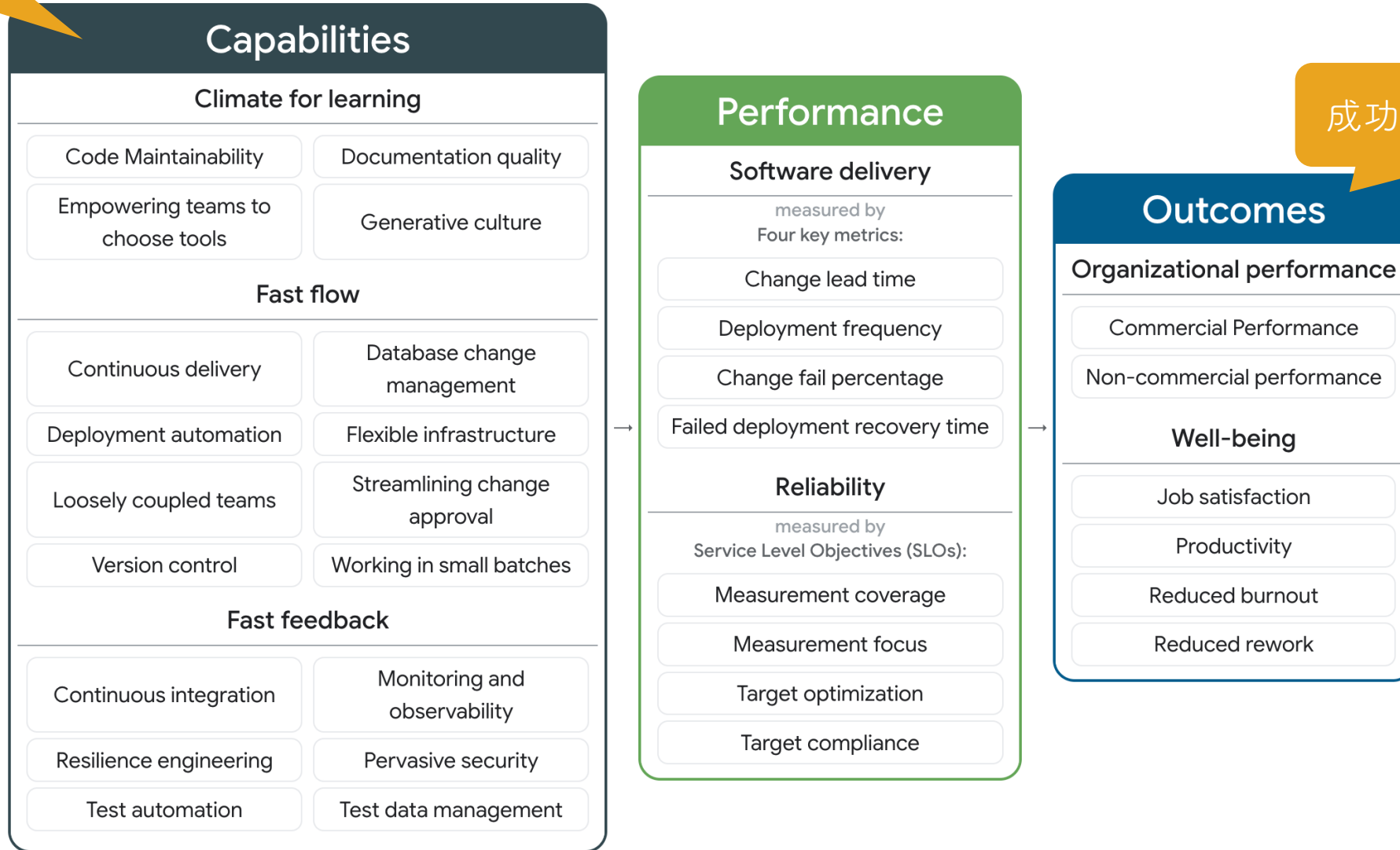
そしてDevOps

DORA(DevOps Research and Assessment) Core Model

こういう取り組みをしている組織が...

DORA Core Model

View mode: ☐ summary ☒ detail



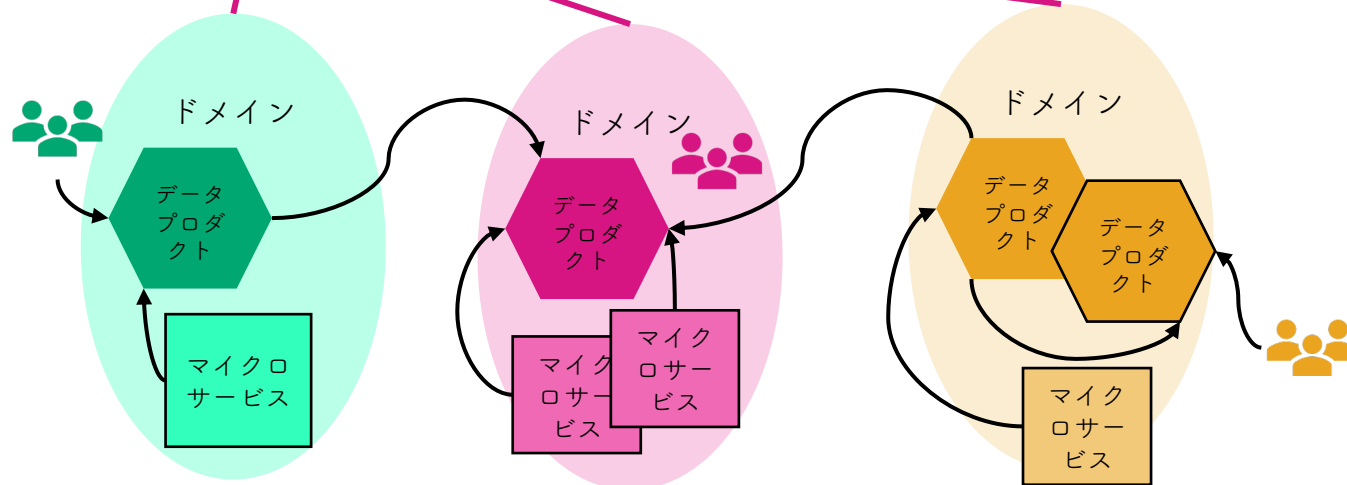
成功を収めてるよ

DORA Core model v2.0.0 download: [PNG](#) [PDF](#)

データメッシュとチームトポロジとDevOps

ストリームアラインドチーム

ビジネス要求に沿って配置するチームで、フロー最大化を目指す、職能横断型チームで自分たちですべてデリバリできる



(ストレージ、オーケストレーション、モニタリング…)
共有データプラットフォーム

プラットフォームチーム

インフラや共通サービスを内部向けに提供するチームで、プラットフォームエンジニアリングを行なっている

DevOpsのケイパビリティ



イネイブリングチーム

他のチームの支援やテクニカルコンサルをするチームで、特定領域の技術に強かったりする

要するに

データメッシュアーキテクチャにして、チームトポロジで
組織作りして、DevOpsすれば良いんだよ

体=アーキテクチャ
データメッシュ

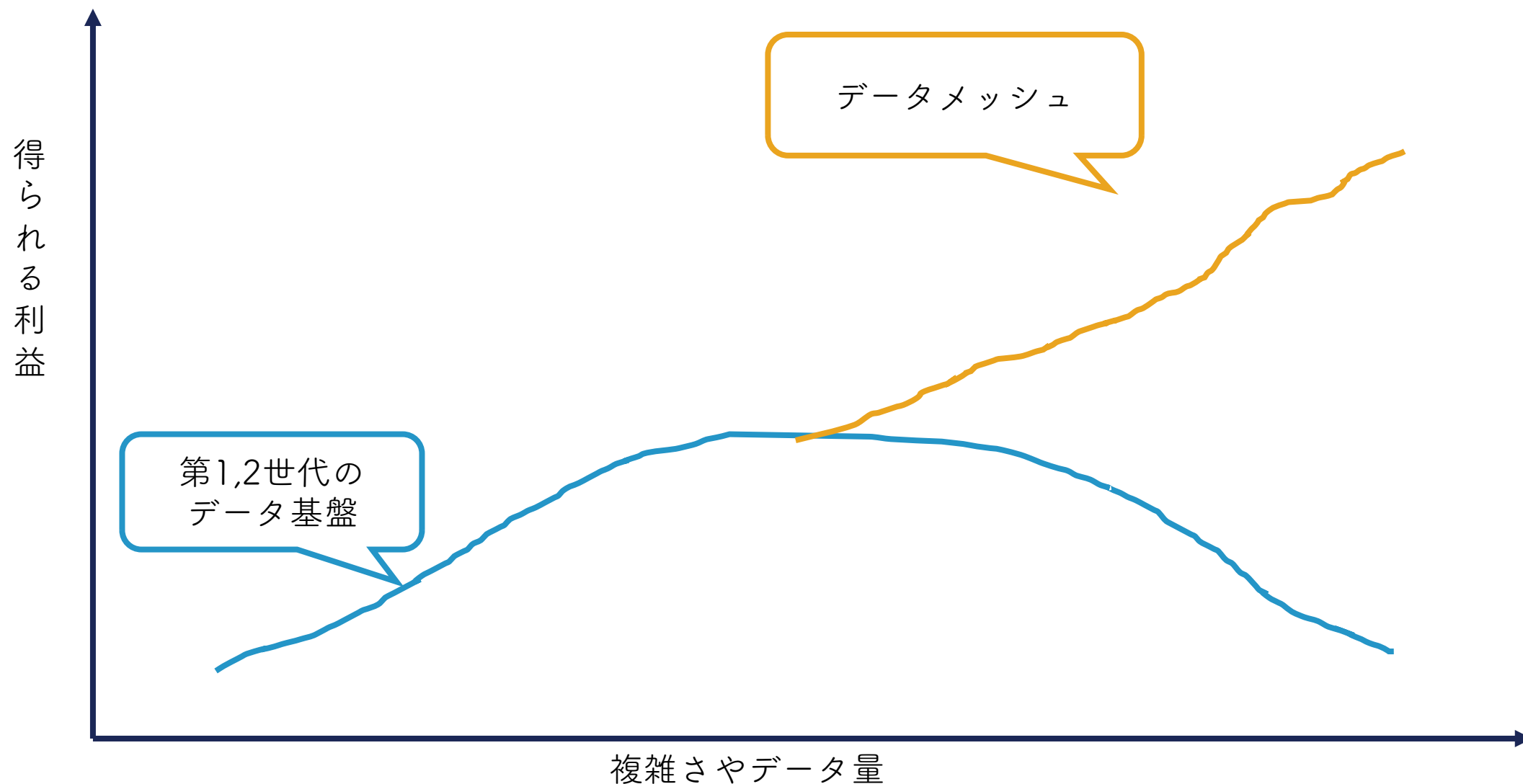
心技体

心=組織
チームトポロジ

技=やること
DevOps

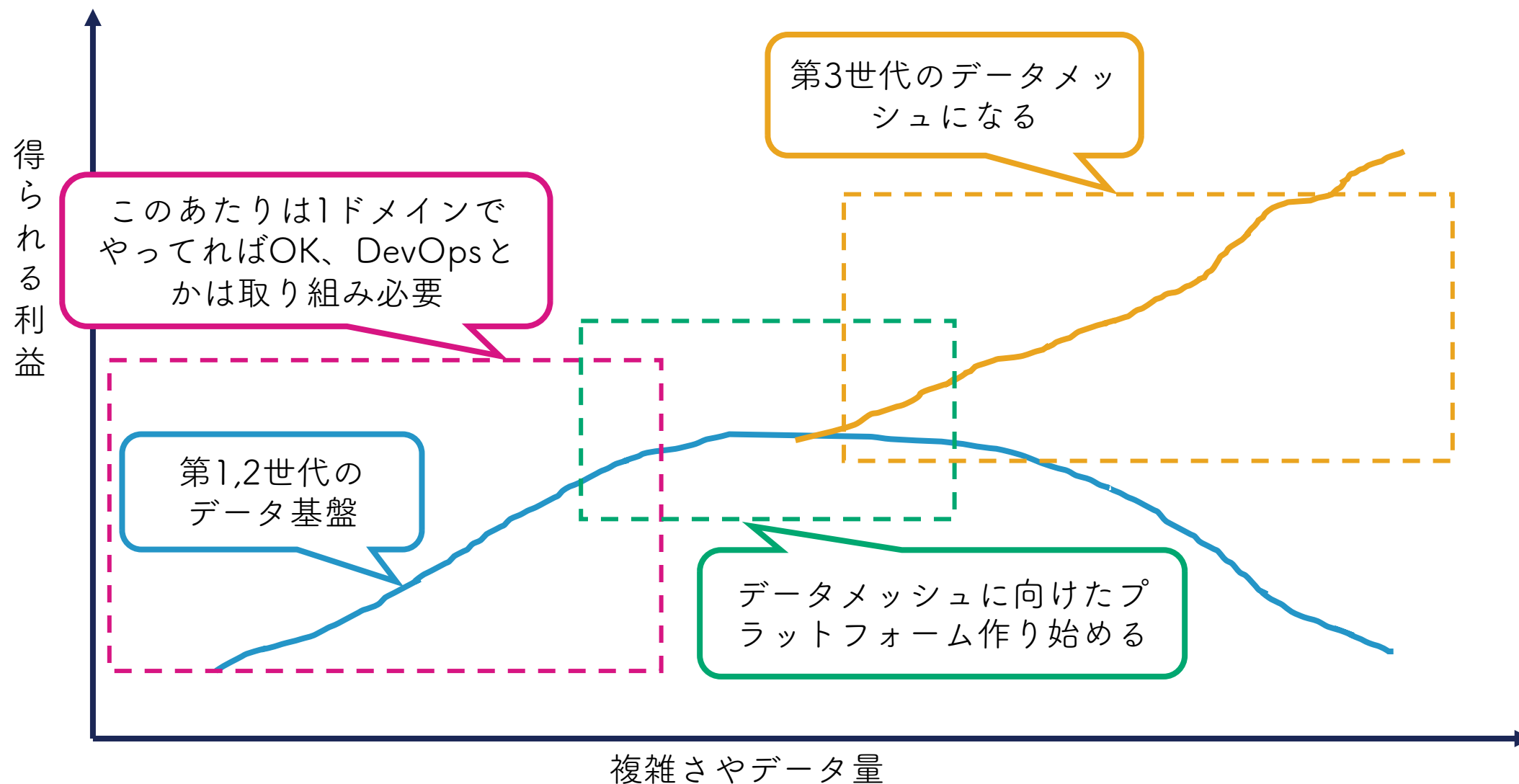
データメッシュのつかいどころ

データメッシュは既存のデータ基盤が複雑さやデータ量が増してきて、パフォーマンスが上がらなくなってきたから、更に上にいくためのアーキテクチャ



データメッシュのつかいどころ

データメッシュは既存のデータ基盤が複雑さやデータ量が増してきて、パフォーマンスが上がらなくなってきたから、更に上にいくためのアーキテクチャ



データエコシステムの成熟度

	ステージ1 データで始める	ステージ2 データで拡大する	ステージ3 データ基盤のゴール
目指すデータ基盤の状態	堅牢なデータを供給するデータ基盤を作って運用してる状態	スケーラブルなデータ基盤が作られている状態	自動化されたパイプライン、セルフサービスなMLと分析が可能、許可された人がアクセスできるデータが実現されたデータ基盤
ステージの説明	利用者に信用できる堅牢なデータを提供するために、データを集め、標準化するなどし品質を高めることで、スモールチームで挑戦を続けるステージ MLへの挑戦よりもまずは経営指標の可視化などへ取り組み、定時・速報レポートなど、ビジネス上の意思決定に利用できることを目指す	堅牢なデータが供給されるようになった結果、データを用いた意思決定が組織の一部で始まり、カバーする領域が経営指標の元となる大きなデータから、アプリケーションログのような細かいデータにまで徐々に増えていく、その結果MLを用いた機能開発や、利用者によるセルフ分析が始まり、データチームはそれを支えるデータ基盤を運用するステージ	データドリブンな意思決定が当たり前の組織になり、経営層にもデータサイエンティスト出身者などデータリテラシの高いメンバーが増える、新規・既存事業間のデータシェアリングだけでなく、M&Aやアライアンスによるデータエクステンションをいかに高いアジリティを保って進めるかがデータチームに課せられた課題、ビジネスドメインに特化した部分は各ドメインに移譲も求められる
データチームの状態	10人以下、データエンジニアはフルスタックの何でも屋	10～20人、データエンジニアは特定技術の専門家に、拡大期においてはチームがボトルネックになることも	20人～、データチームは分割され、チームトポロジーの採用
採用されやすい代表的なアーキテクチャ	エンタープライズデータウェアハウス、データレイク、モダンデータスタックなどシンプルなアーキテクチャから始まる	データチームは生産性、アジリティが重要になりモダンデータスタック系になっていく、スケールのためにDataVaultを採用する場合も	Scaled Architecture、DataMesh、DataFabricなど、大規模にデータ管理可能なアーキテクチャへ移行していく

成熟度によって徐々に取り組んでいくという
ことを忘れないようにしつつ…

青写真

青写真としての心技体

体=アーキテクチャ
データメッシュ

心=組織
チームトポロジ

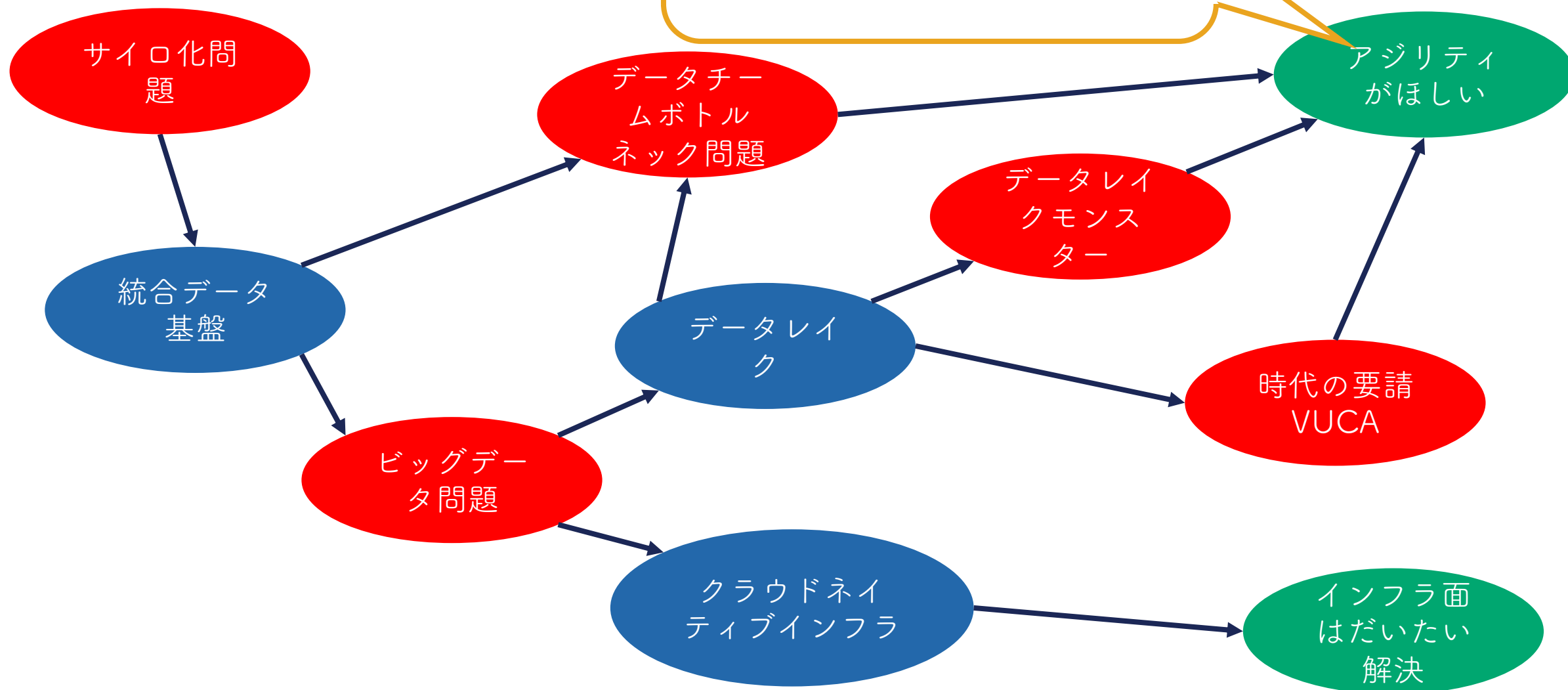
技=やること
DevOps

で、問題はどうか？

データ基盤システムの歩み

こうしてさまざまな問題を試行錯誤し

分散データ基盤アーキテクチャであるデータメッシュ(+チームトポロジ+DevOps)は十分なアジリティをもたらすか



まとめ

- 人類はいろいろがんばってきた
- 人類が獲得した叡智を3つを青写真にして進もう
 - 心(組織)=チームトポロジ
 - 技(とりくみ)=DevOps
 - 体(アーキテクチャ)=データメッシュ(に代表される分散データ基盤アーキテクチャ)
- ~~十分なアジリティは競合との差として現れるので最後は組織の結果論...~~

続きはAsk the Speakerで！

※本当は喋るべきトピックがもっとある