

2022年度JUASビジネスデータ研究会 活動成果報告会

日清食品ホールディングス 西村太輔

目次

1. 研究会概要
2. 研究テーマ
3. 分科会活動報告

1. 研究会概要

【研究会名称】ビジネスデータ研究会

【実施期間】2022年6月～2023年3月

【主な活動実績】

- － 全体研究会8回、講演会3回
- － その他分科会での活動

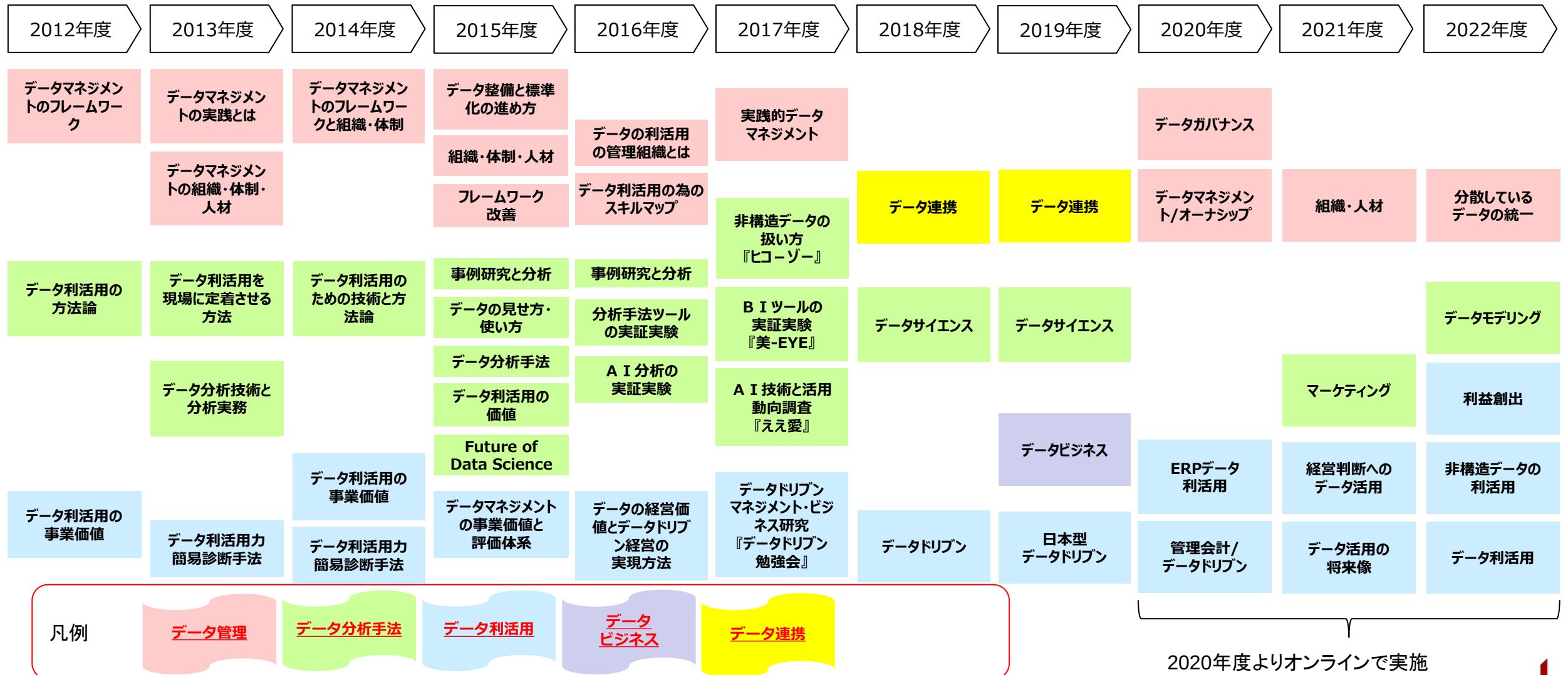
【参加者人数】39名（監事団含）

監事団（敬称略）

	氏名	会社
部会長	西村太輔	日清食品ホールディングス株式会社
副部会長	吉野太	アイエックス・ナレッジ株式会社
副部会長	三好克典	株式会社プライド
監事	石井昭紀	株式会社イージフ
監事	大下健史	ブレインズコンサルティング株式会社

2. 研究テーマ

ビジネスシーンにおけるデータの利活用について幅広く研究を実施。



2. 研究テーマ

【非構造データの利活用】

【データ利活用】

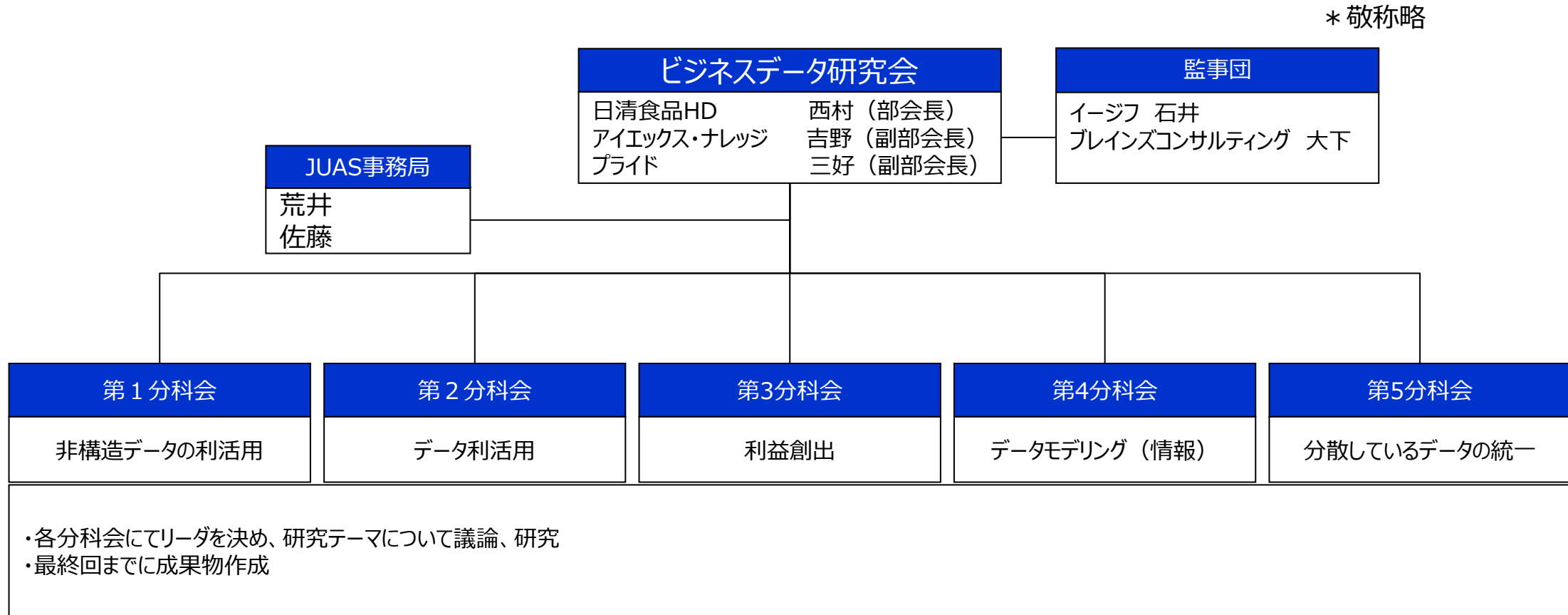
【利益創出】

【データモデリング（情報）】

【分散しているデータの統一】

2. 研究テーマ

研究体制



3. 分科会活動報告

3. 分科会活動報告

【非構造データの利活用】

非構造化データを用いた活用例について理解を深める

【非構造化データの利活用】非構造化データの定義

Q. そもそも非構造化データとは？

A. 構造化されていないデータ...

	構造化データ	非構造化データ
概要	データが列と行で集計されたもの	構造化されていないデータ
企業内の割合	2割	8割
保存方法	データベース	ファイルシステム（NAS、パソコン、スマートフォン、クラウド）
例	・顧客データ、商品データ、人事データ、会計データなど データベース化された情報 ・CSVなどのように行と列でまとまっているデータ	ワード、エクセル、PDF、メール、SNS、音声、画像、動画、センサー 情報、ログ

非構造化データのイメージを
出し合う中でのメンバーの意見

定義が広すぎて、どこまで
を非構造化データと呼べば
よいかわからない

非構造化データのイメージと
しては、画像、テキストのイメ
ージが強い

テキストも非構造化データ
に含まれるとは思わなかった

利活用に向けてどうやって
管理すれば良いのか

非構造化データであったとし
ても、データが持つメタデー
タや要素抽出によって、構
造化されたものはどうなる？

【非構造化データの利活用】事例と分析軸

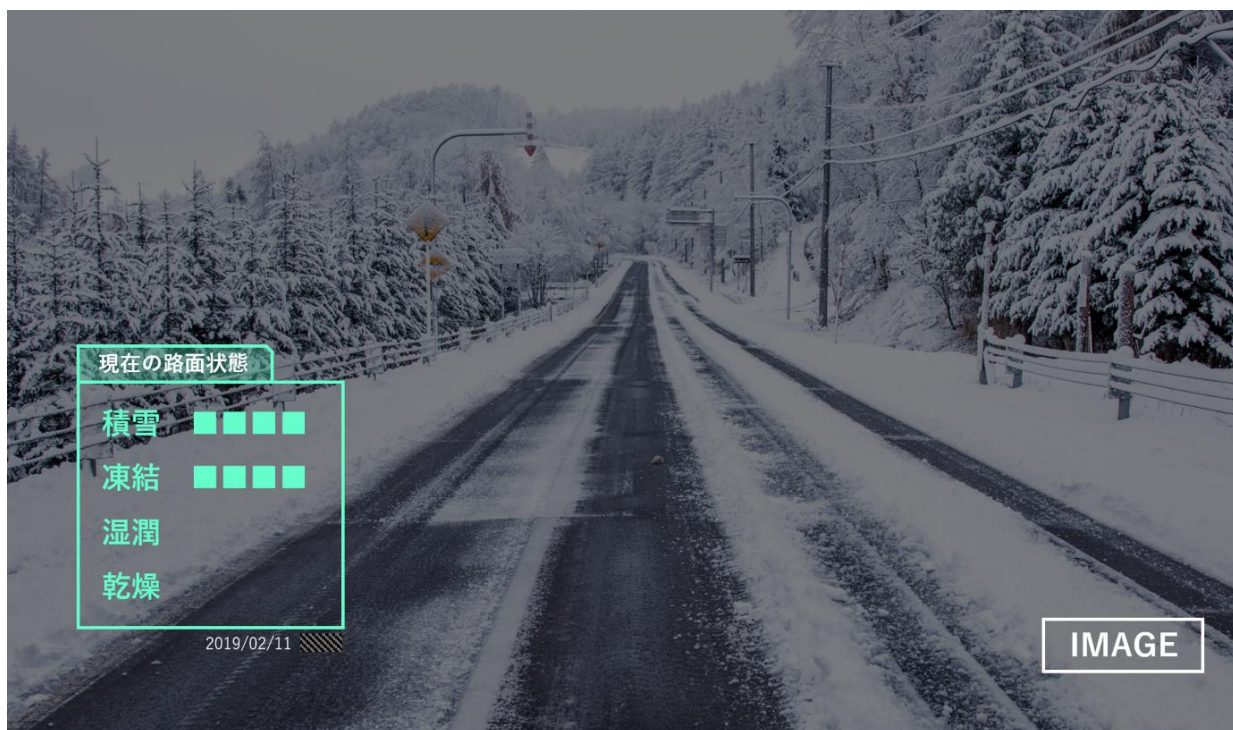
- 各メンバーで非構造化データを用いたビジネス事例を収集(データタイプ/目的/企業・団体)
 - 調査方法は、インターネット上で公開されている情報や自社事例・サービスなど
 - 最終的に全44事例を収集(事例一覧については、“(別紙)非構造化データ活用事例”を参照)
- 収集した事例を以下の“分析軸”をもとに分類し、その結果から得た“気づき”についてディスカッションを行った。

①機械学習の要否 機械学習が必要かどうか。 一回限りの学習が完了するか、継続して学習させるか。	②構造化データへの変換 分析対象の非構造化データを構造化データへ変換しているかどうか。	③データ準備者 学習、分析をするためのデータは誰が用意・取得するか。	④学習の頻度 学習する頻度はどのくらいか。一時的(または不定期)に学習するか、継続的に学習するか。
⑤新規/既存 目的の達成のためにデータを新規で生成しているか、既存のデータを用いているか。	⑥人の代替/新たな価値 もともと人の代替を目的としているか、新たな価値を見出すために分析を行うか。	⑦数値化 分析結果を数値(デジタルデータ)として得られているかどうか。	

【非構造データの利活用】事例紹介

【事例紹介】路面状態の判定、予測

- 周辺気象情報、現場画像(路面)などをAIを用いて路面の状態を判定、予測する実証実験。運送事業者の安全確保など、データ駆動型社会における共通インフラ整備に期待



(出典) 日本気象協会ホームページ<画像のみ>

1	機械学習の要否	必須
2	構造化データへの変換	変換
3	データ準備者	学習データ、分析データ： 事業者、気象協会
4	学習の頻度	継続的
5	新規/既存	路面データ：新規 気象データ：既存
6	人の代替/新たな価値	新たな価値
7	数値化	数値化

【非構造データの利活用】事例研究を通じて得た気づき

【気づき①】非構造化データをどうやって管理しているか

- 多くの場合、構造化データへ変換している。
- オリジナルデータは破棄するケース，保存しておき後で別の分析に再利用するケースがある

【気づき②】分析データの統合について

- 一つのデータだけでなく複数のデータ（既存、新規含む）を組み合わせ活用する事例がある（路面状態判定：気象情報と路面の画像）
- 継続して学習させる場合、新しく教師データを追加すれば精度が上がるわけではない。（ノイズとなり精度が下がる場合も）

【非構造データの利活用】事例研究を通じて得た気づき

【気づき③】既存の非構造データを用いた活用事例が少ない

- 既存のデータの観点では、非構造化データより構造化データの活用事例が多い。
 - 認識：そもそも保有総データの全体像や中身を把握しきれていない。
 - 技術：非構造化データの取り扱いにスキルとツール、コストが求められる。
 - 手段：目的の達成手段として、新規のデータ取得に目を向けられる傾向。

【気づき④】音声データの活用事例が少ない

- 音声技術を扱う事業者/技術者が少ない傾向にある？
- 音声からテキストを生成するような技術が先行していたが、音声データの識別や感情分析などの活用領域が拡張してきている。
- NTTなどとともに音声を取り扱う事業者が先行している

【非構造データの利活用】事例研究を通じて得た気づき

【気づき⑤】テキストデータを用いた分析事例が思っていたより少ない

- テキスト分析は非構造化データの分析としては初期段階から行われており、現在は落ち着いている(他のデータタイプの分析の方が盛んになってきた)
- 日本語の特性上(分かちわけしづらい、同音異義語)分析が難しい
- ChatGPTなど新しい技術によって再度脚光を浴びる予感。(ChatGPT : OpenAI社が提供しているAIチャットサービス)

ChatGPTを試してみた結果...

Q. 日本情報システムユーザー会って楽しい？

日本語を理解して回答してくれてる。

A. JUASが楽しいかどうかは、個人の趣向や関心によって異なるでしょう。JUASは、情報システムの利用や管理、運用に関する情報交換や技術的支援、啓発活動などを主な活動としています。JUASに参加することで、情報システムに関する最新の情報や技術的な知識を得ることができたり、他の情報システムの利用者との交流や意見交換ができたりするでしょう。これらの活動に興味がある人であれば、JUASは楽しいと感じるかもしれません。

また、JUASが主催するイベントには、講演やセミナー、ワークショップなどが含まれることがあります。これらのイベントに参加することで、新しい知識やスキルを身につけたり、他の参加者との交流を楽しんだりすることができます。

ただし、JUASの活動には一定の参加費が必要になることがあるため、参加する前に費用やプログラムを確認しておくことが大切です。

どういう人が“楽しい”と感じるかの情報も提示してくれる。

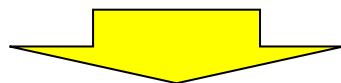
(参考)某スマートスピーカーの場合...

A. ちょっとよくわからなかったです。ごめんなさい。

【非構造化データの利活用】まとめ

非構造化データの利活用について、事例調査・研究を中心に進め、以下気づきを得た。

- 【気づき①】非構造化データをどうやって管理しているのか
- 【気づき②】分析データの統合について
- 【気づき③】既存の非構造化データを用いた事例が少ない
- 【気づき④】音声データの活用事例が少ない
- 【気づき⑤】テキストデータを用いた分析事例が思っていたより少ない



- 各メンバーの非構造化データの利用について知見を深めることができた
- 違う分析軸で事例を見ることで、また新たな気づきが出る
- 今後の研究会のテーマまたは各自の業務・個人におけるテーマとして、もっと深堀して調査を進めていきたい

3. 分科会活動報告

【データ利活用】
～データ活用におけるベストプラクティス～

【データ利活用】サブテーマ

- 「なぜ、分析と行動が結びつかないのか」を深掘り「データ活用におけるベストプラクティス」をサブテーマとした

データ利活用の課題

重点課題

「なぜ、分析と行動が結びつかないのか」

議論

ヒト

知識不足、リテラシー不足、意思決定者の関与不足、行動に結び付けようとしない、モチベーション不足、部門間の隔たり、自己満足、...

モノ

データ不足、分析結果の質が悪い、分析ツール／モデルの精度が低い、...

カネ

データ収集ツールの導入コスト、費用対効果が不明瞭、...

サブテーマの設定

データ活用におけるベストプラクティスとは？

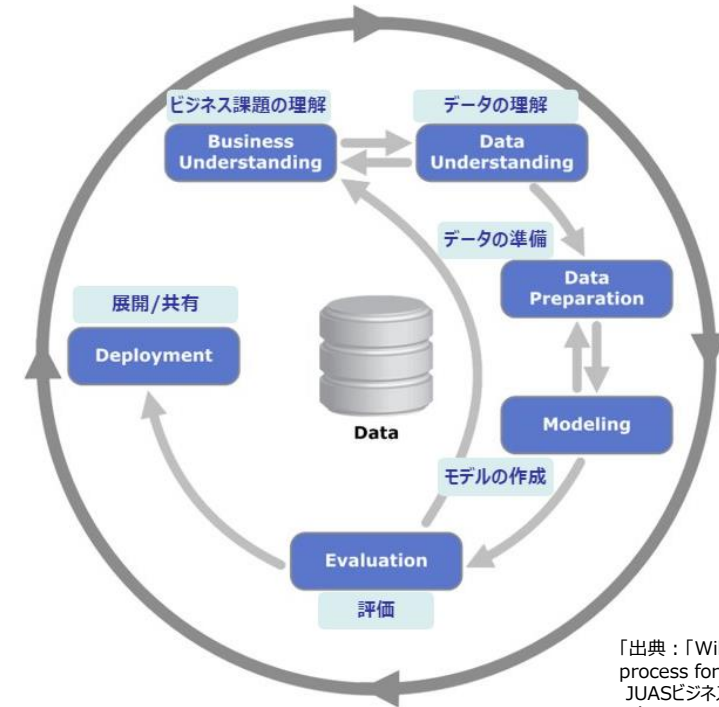
【データ利活用】要素ごとの整理

- 各自の事例を基に議論した結果、各フェーズとプロセスの組み合わせで、それぞれの注力ポイントがあると推察。

フェーズ

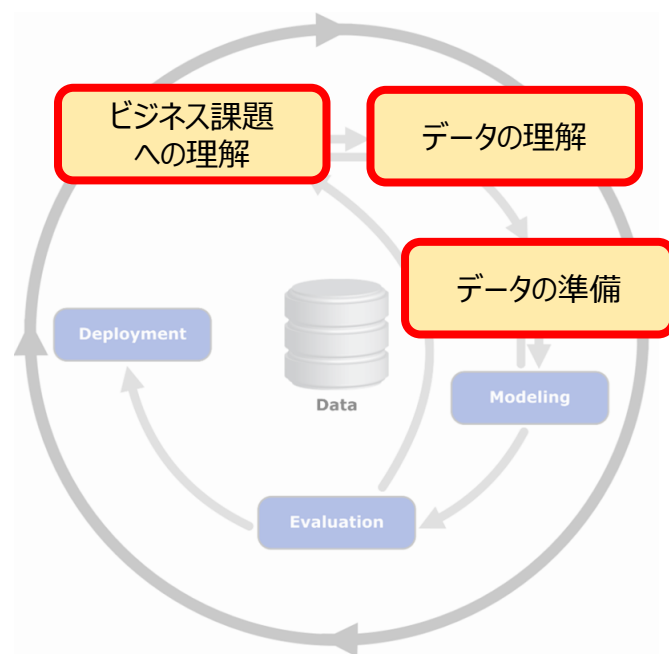


プロセス



「出典：「Wikipedia Cross-industry standard process for data mining」/
JUASビジネスデータ研究会2022で一部加工（追記）」
https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining

【データ利活用】注力ポイント： ゼロ→イチ



■フェーズ定義

- ・ できるところから小さくはじめる
- ・ スピード感をもって、成果・データ活用の実績を作る

■good practice

- ・ スピード感をもって成果を上げることが重要
- ・ ビジネスインパクトが小さくてもよいので、データに対して関心がある小さい組織・チームでスピード感をもって進める

■bad practice

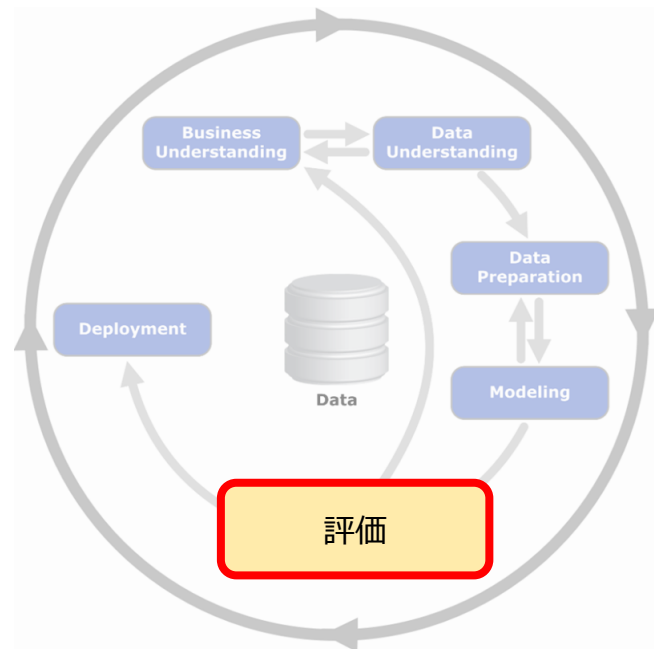
- ・ 最初からビジネスインパクトや難易度が高い対象をターゲットにしてしまう
 - ・ 結果、ステークホルダーが多くなり調整が困難になる
- ・ セキュリティが固い環境上で、進めてしまう
- ・ データマネジメントをガチガチにする
- ・ データ活用の推進役が不足している

■成功事例

- ・ データを集めつつ、週報の作成作業を分析ダッシュボード化など
- ・ 業務改善などによる、データ活用の実績作り



【データ利活用】注力ポイント： 類似事象で再現性担保



■フェーズ定義

- ・ ゼロ→イチで実施した内容を他の類似業務領域に適用し、同様の結果が出ることを確認

■good practice

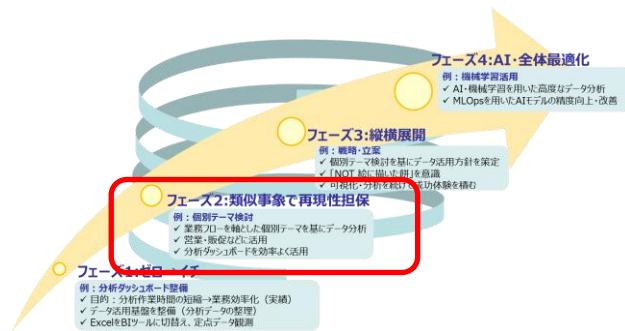
- ・ データ分析前のビジネス課題設定の強化

■bad practice

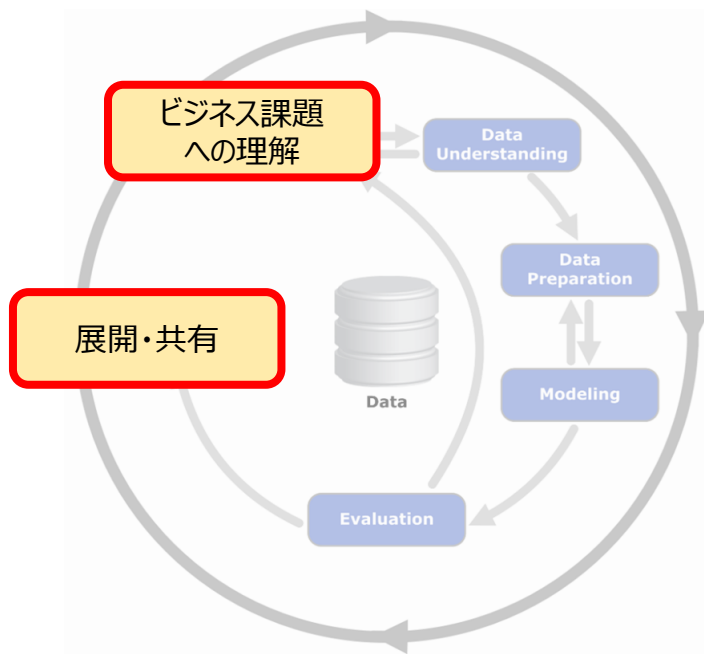
- ・ たまたまいい結果を取り上げてしまう
- ・ 分析結果から施策に結びつけられない

■成功事例

- ・ 業務フロー軸などの個別テーマを基にデータ分析
 - ・ 営業・販促などの領域に活用
 - ・ 分析ダッシュボードを効率よく活用



【データ利活用】注力ポイント： 縦横展開



■フェーズ定義

- ・ (横)他部門へ情報展開・共有を実施し、業務ルールも同時に整備する
- ・ (縦)ビジネスインパクト・難易度がより高い対象にして、広めていく
- ・ データ分析を考えていない組織、業務にも適用していく

■good practice

- ・ 推進フェーズに合ったデータガバナンスの体制構築
- ・ データベース/データ規則（結合キーなど）の整理

■bad practice

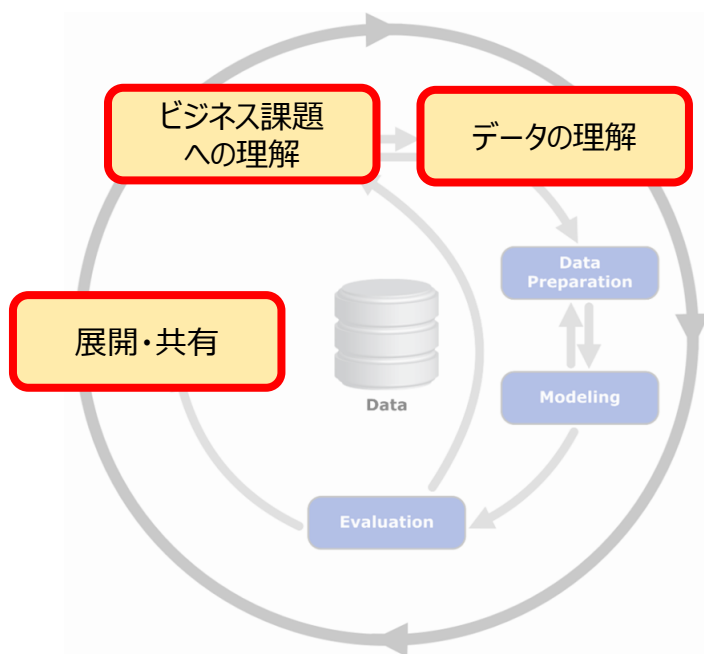
- ・ データマネジメントを考慮せずに、データ基盤を拡大させてしまう
- ・ 部門内の個別最適の弊害により、データマネジメントを複雑化し、データ運用コストを増加させる

■成功事例

- ・ 知見がない人にも定着できるよう、フォロー体制を整備する
- ・ ゼロ→イチの際には発生しなかった問題などにも対応
 - ・ ビジネスインパクト・難易度がより高くなり複数組織での連携が必要となるため



【データ利活用】注力ポイント： AI・全体最適化



■フェーズ定義

- ・ AI・機械学習を用いた高度なデータ分析
- ・ データ活用が当たり前の文化を醸成し、継続的な改善活動を行う
- ・ データガバナンスの強化

■good practice

- ・ 分析結果を業務まで落とし込む（業務適用数を重視する、等）
- ・ 勉強会や自己研鑽の支援の実施をする

■bad practice

- ・ AIのモデル精度が悪いまま適用しようとする
- ・ 縦割による、協力体制の不確立、データ活用が進まない

■成功事例

- ・ 事例はなさそう



- 本研究で検討したベストプラクティスを国内の企業に対してデータ活用を推進するための参考に頂きたい
- ある程度予想できるアンチパターン（bad practice）と同じ失敗をすることを避けるため、本資料を参考に対策を考えることが重要である
- 本研究の次に取り組むべきテーマとして、本研究のアウトプットを基にデータ活用・検証を行い、ケーススタディを蓄積していく

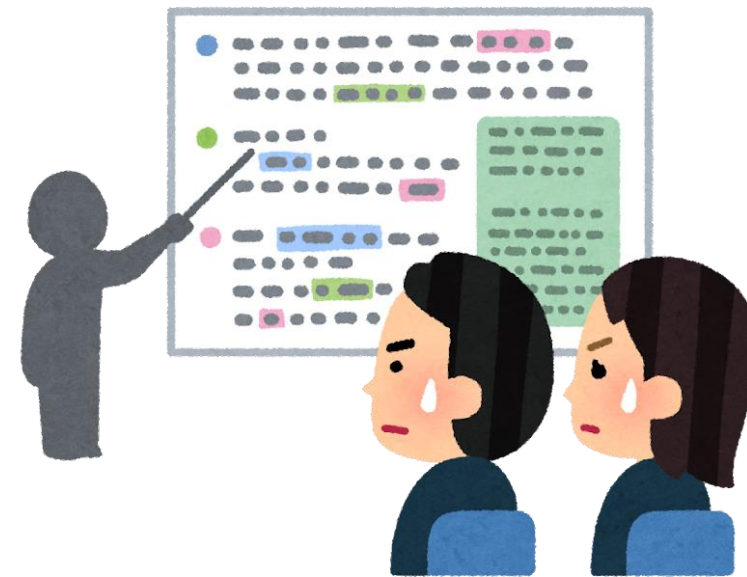
3. 分科会活動報告

【利益創出】
～カネで時を買う～
データ利活用による利益可視化検討

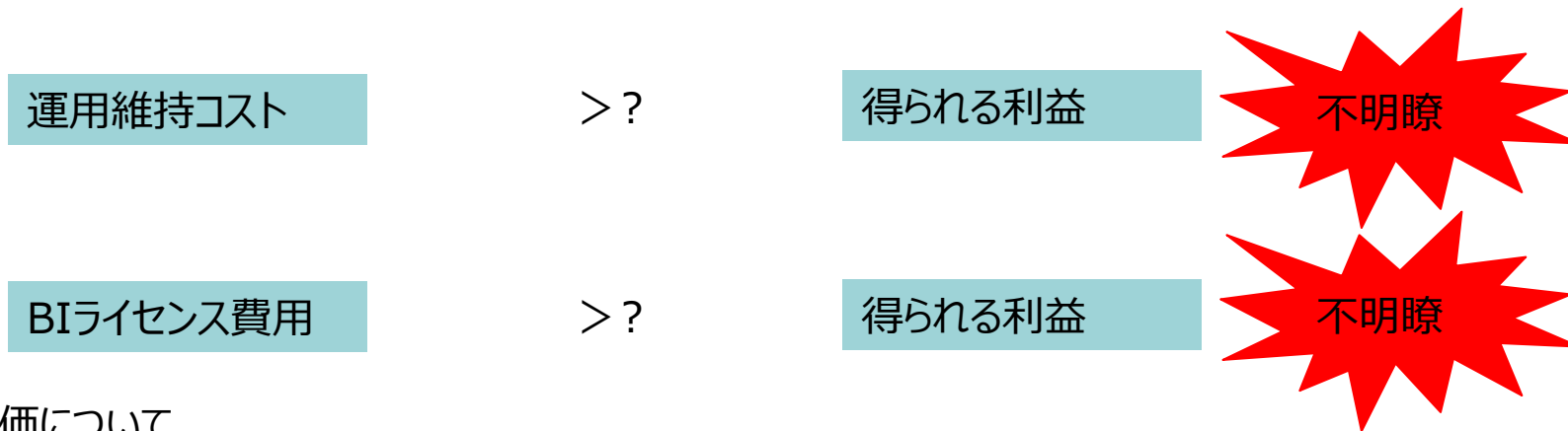
データ分析基盤を構築、運用にあたって、必ず**カネ**の問題が必ず出てくる。

費用対効果/妥当性を説明する際に直面しやすい3つについて検討した。

- ①運用維持のコストともたらす利益が比べづらい
- ②導入における効果を数字で示しづらい
- ③システムの評価計測をどのように回すべきか示せない



【利益創出】①運用維持コストと、もたらす利益が比べづらい



■ BIの定量的な評価について

そもそもBIツールとは..人間が主体となって、意思決定する際の支援をするもの。
あくまで人間の判断の根拠を補助するツールということで、そこから得られる利益については、
あくまでもその判断によって実際に行動した内容に帰属すると考える。

エクセルやワードのようなツールの利益を評価をしないように、BIツールについても得られる利益も評価しない。
BIツールの利用での資料のクオリティの向上といった定性的な評価や、資料作成の時間の削減といった定量的な
評価をすることは出来るがあくまでもBIツールに関わるコストは必要経費としてとらえ、
そのインプットのデータが最新であることも同じく必要経費になると考える。

ライセンス料のコストが高いのに使われていないのであれば、使われるように社内での研修や啓蒙活動をすることで、BIツールを当たり前を使う
人達を増やすことを優先するべき=人の問題と考え、そのような認識を経営層、ユーザーに持たせる必要がある。

【利益創出】①対策

◆対策

BIツールは「あって当たり前」、それであればデータ分析基盤もあって当たり前ではないのか？

データ分析基盤

=

必須のインフラ

インフラの運用維持についてはキャッシュフローの算定は難しいといわれている。

それはあって当然のもので、運用維持が必須だからである。

そのため、従来インフラの運用維持に関わるコストの効果は非財務的な指標を用いて評価されてきた。

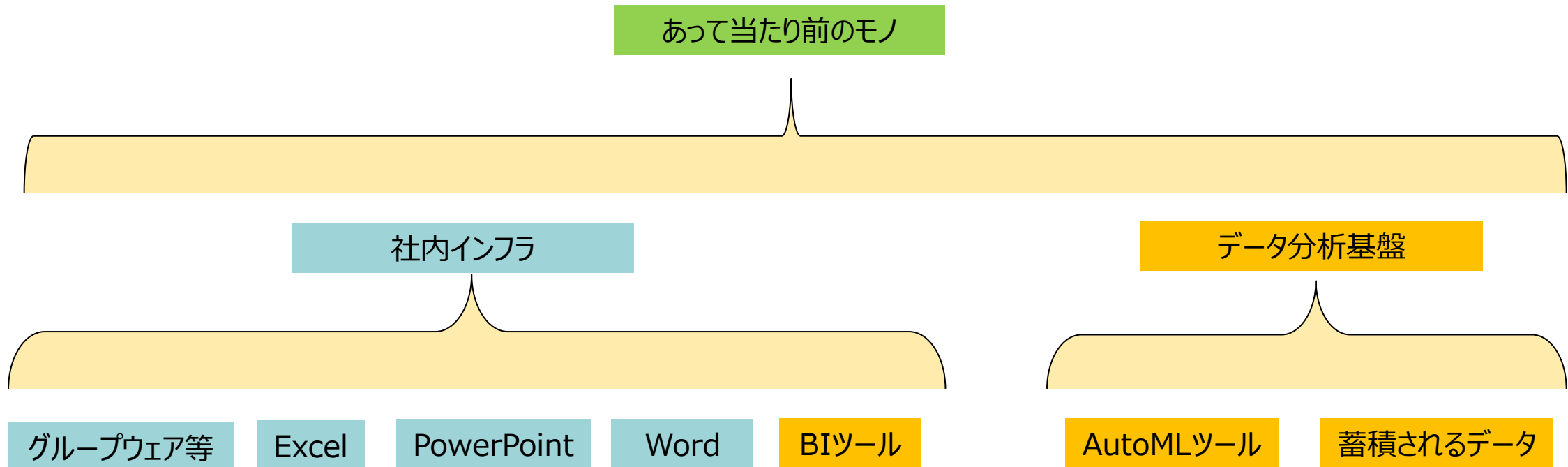
データ分析基盤においても、以下のような非財務的な指標で評価することもできるのではないか。

- ・システムの利用率
- ・テーブルへのアクセス回数
- ・データ分析基盤ユーザーへの満足度調査

【利益創出】①まとめ

大前提として、データ分析基盤を必須のインフラとして捉えさせることを試してみる

- 非財務的な指標を用いて評価させる。
- データ分析基盤で使用するツールは必要経費として考える



【利益創出】②導入における成果を数字で示しづらい

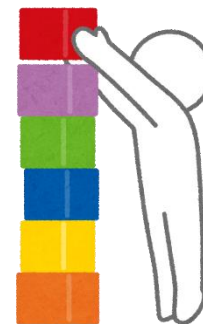
データ分析基盤構築が必要だと経営者やステークホルダーに説明するにしても、成果・効果を数字で示し、ROIを示さないと納得・理解してもらえない。構築の必要性があることは理解してもらえても、投下した費用に見合う効果があるのか、判断材料を示していない提案では予算を取得するのは難しい。

その為の解決策、アプローチとして、まずはスモールスタートからの**実績の確立**が有効だと考える。



【利益創出】②対策

具体的な方法として、
取りあえず、データ収集・活用を**手作業**で始める。
当然、かなりの時間がかかるので**ムダ・ムラ・ムリ**が表面化する。



そこで、省力化・効率化作業としてその過程を自動化することで、どの程度マンパワーや時間を削減できるのか、その部分を数値化することで、費用対効果を**数字**で示すことが可能になる。

例えば、

手作業では2人×20時間かかっていたことが、作業を自動化することで1人×5時間になったとする。

(2×20) 時間 - (1×5) 時間 = 35時間の削減になる

担当者の時給を2,600円とした場合、1度の実施で**2,600円×35時間=約9万円**の効果がある。

※500万(年収)÷12(月)÷160(時間)≒2,600円で計算

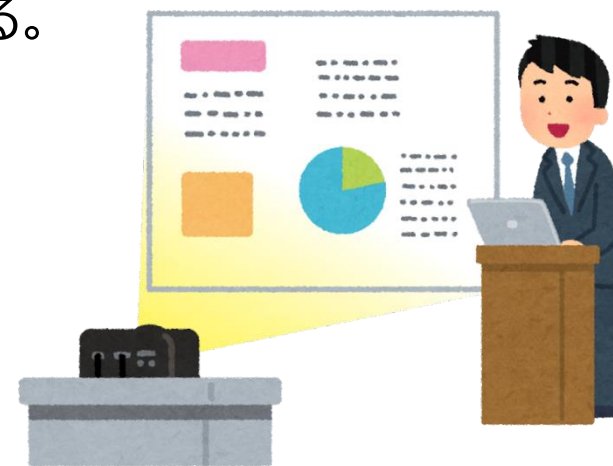
【利益創出】②まとめ

データ収集を毎週 1 回実施するとした場合、

月額だと9万×4回で36万円、年間だと36万円×12か月で**約430万円**の削減効果となる。

自動化するとその実行環境の運用費用がかかるので、実際は丸々費用が浮くということはないが、費用回収にかかる期間を見積もることが出来、経営層へ説得の材料になる。

そこで、この費用内に収まるプロジェクトを提案し、結果を出してから次のプロジェクトへと**段階的**に進めるのはどうだろうか。一度に全てのデータを対象にせず、一部を少しずつ進めていけば、最終的には全てのデータ基盤が構築できる。



既に手作業で実施している部署やチームがあれば数値化の協力を仰ぐのもあり。自動化できれば彼らからも感謝され、更に評価が上がる。

【利益創出】③システムの評価計測をどのように回すべきか示せない

“3無い地獄”



成果を評価計測しづらい背景

① 目標や仮説がまったく無い = 目標やシナリオ設定のノウハウが無い

→目標や仮説が曖昧なまま“とりあえず”データを分析、活用している。次の“行動”に繋がりにくい。

👉 データ分析とビジネス課題が結びついていない

② 目標や仮説があっても、データ分析に対する前例が無い = 評価軸が無い

→データ分析による成功事例やPDCA経験が無いため、比較検討がしにくい。

👉 目標や仮説と異なる思わぬ方向で結果が出たときに、成果かどうか判断出来ない

③ 最終目標から、事業に合った施策レベルの数値目標まで落とし込めていない = フレームワークが無い

→KGIがあったとしても、KGIを正確に要素分解出来ておらず、数値目標（KPI）の設定がずれている。

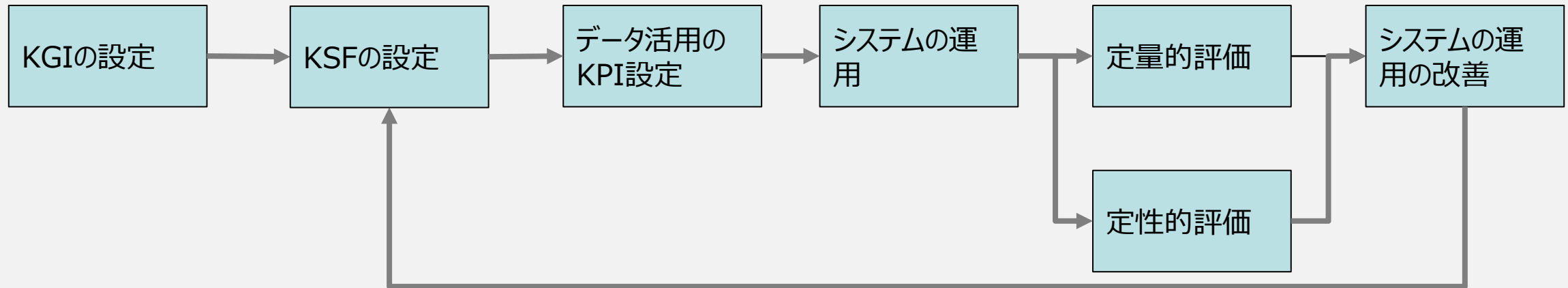
👉 KGIはあっても、KPIにうまく落とし込めていないケース



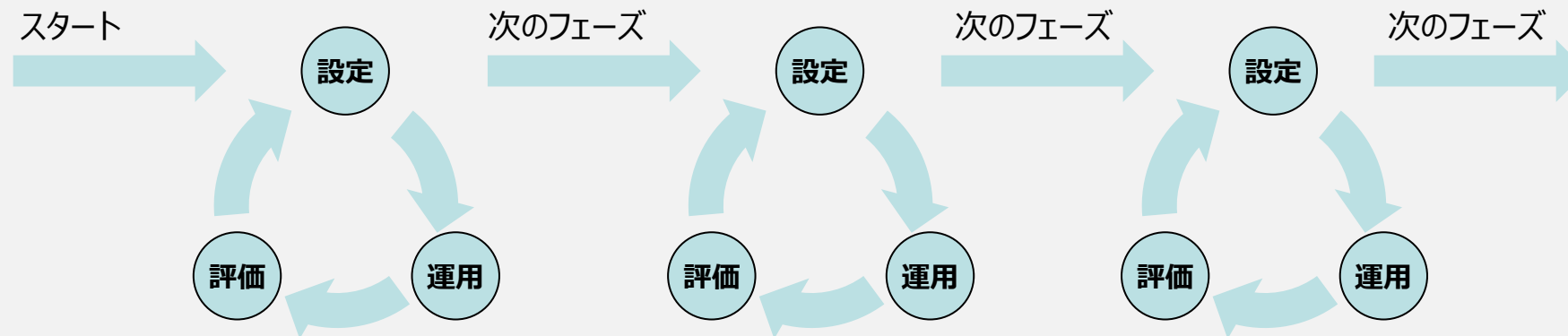
どうしよう...

【利益創出】③対策：評価の全体の流れ

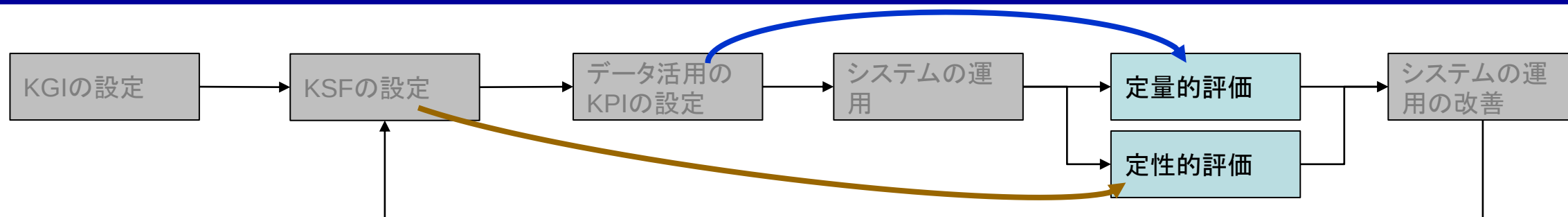
➤ 事業フェーズの成長度合い（グロース）に合わせて設定する



➤ 事業フェーズの成長度合い（グロース）に合わせて設定する



【利益創出】③対策：定量的評価に求められる要件



定量的評価

取得方法

- システム上の測定可能なデータにより、定期的にデータを取得して評価する

ポイント

- 基本的に設定したKPIが評価の対象となる
- KPIを説明できる根拠の数値を用いる

定性的評価

取得方法

- ユーザに対するアンケートやヒアリングにより取得する

ポイント

- 変化を確認する質問を入れておくことよい
例：システムにより業務の変化はあったか、
ビジネスアクションに変化はあったか
- 当初のKSF(システムの目的)が達成されているか確認

問題点

成果を評価計測しづらい背景にあるもの…“**3無い地獄**”

- ①目標や仮説がまったく無い = **目標やシナリオ設定のノウハウが無い**
- ②目標や仮説があっても、データ分析に対する前例が無い = **評価軸が無い**
- ③最終目標から、事業に合った施策レベルの数値目標まで落とし込めていない = **フレームワークが無い**

まとめ

- ・KGIを起点とし、事業フェーズ毎に**KSFとKPI**を定める
- ・KSF、KPIに対して**定量的評価と定性的評価の両方**を実施する
- ・実施に当たっては、プロジェクトキャンバスなどの**フレームワーク**に当てはめる

データ利活用による利益創出のために…



データ基盤導入前

データ収集・活用を**スモールスタート**で行い、発生する（している）『ムリ・ムダ・ムラ』を、**データ基盤導入によりどれだけ省力化・効率化できるかを算出**する。特に、運用維持コストと基盤がもたらす利益の比較は、プロト的にスモールモデルを作成して説明材料とするなど、**データ基盤は『必須インフラ』であることを明文化**する。

データ基盤導入



データ基盤導入後

導入、実装後の評価計測は、**目的・全体像と定性評価、KPIと定量評価の結びつき**をしっかりおさえる。特に、成果の評価を難しくする背景には『**3無い地獄**』がある。「KGIを起点とし、事業フェーズ毎にKSFとKPIを定める」、「KSF、KPIに対して定量的評価と定性的評価の両方を実施する」といったPDCAを実践していく。

3. 分科会活動報告

【データモデリング】

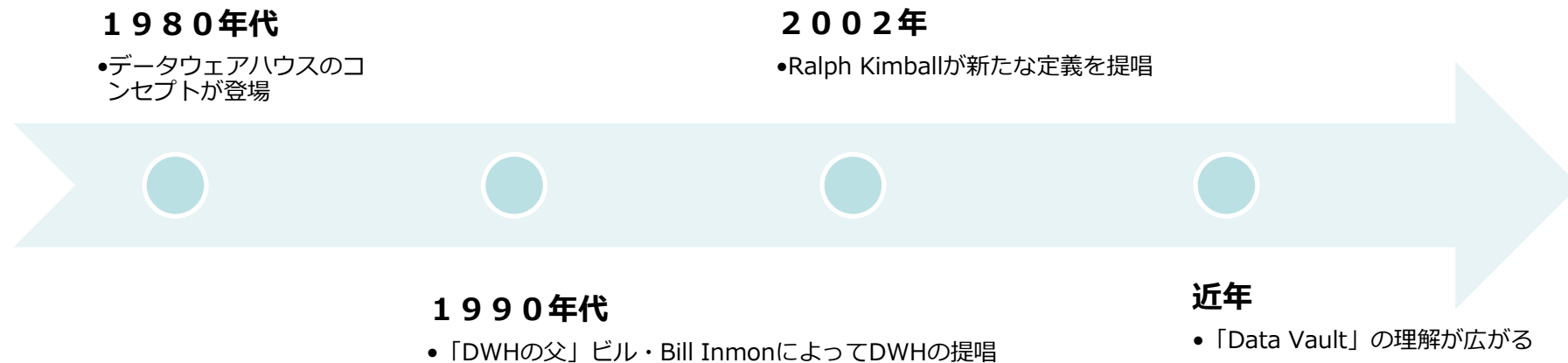
データモデルの必要性を理解し、
技法調査結果をユースケース毎に整理する。

一ステップ

- ① データモデリングの必要性をデータ活用のどこで重要なのかを理解するために全体像を捉える。
- ② データ分析におけるデータモデリング技法を調べる。
 - ✓ Bill Inmonの正規化モデリング
 - ✓ Ralph Kimballのディメンショナルモデリング
 - ✓ Data Vault2.0
 - ✓ 大福帳モデリング
- ③ ユースケースなどに合わせてどのモデリングが最適かを調べる。

【データモデリング】DWHの歴史

- **1980年代**
データウェアハウスのコンセプトが登場。様々なソースからのデータを共通のデータモデルに統合できる技術が生まれた。
- **1990年代初頭**
Bill InmonがDWHの概念を提唱、DWHの定義の生みの親「DWHの父」と呼ばれる。
また、データウェアハウスが本格的に構築され始める。
- **2002年**
Ralph Kimballによって、Bill Inmonとは異なる新たな定義が提唱される。
- **近年**
Bill Inmon・Ralph Kimballとは異なるアプローチ、「Data Vault」の理解が広がる。



【データモデリング】 Bill Inmonの正規化モデリング

Bill Inmonはデータウェアハウスを「**経営陣の意思決定プロセスを支え、サブジェクト（主題）別に統合化された、時系列な不変データの集合**」と定義している。
データの保存と管理には、**正規化されたリレーショナルモデル**が使われる。

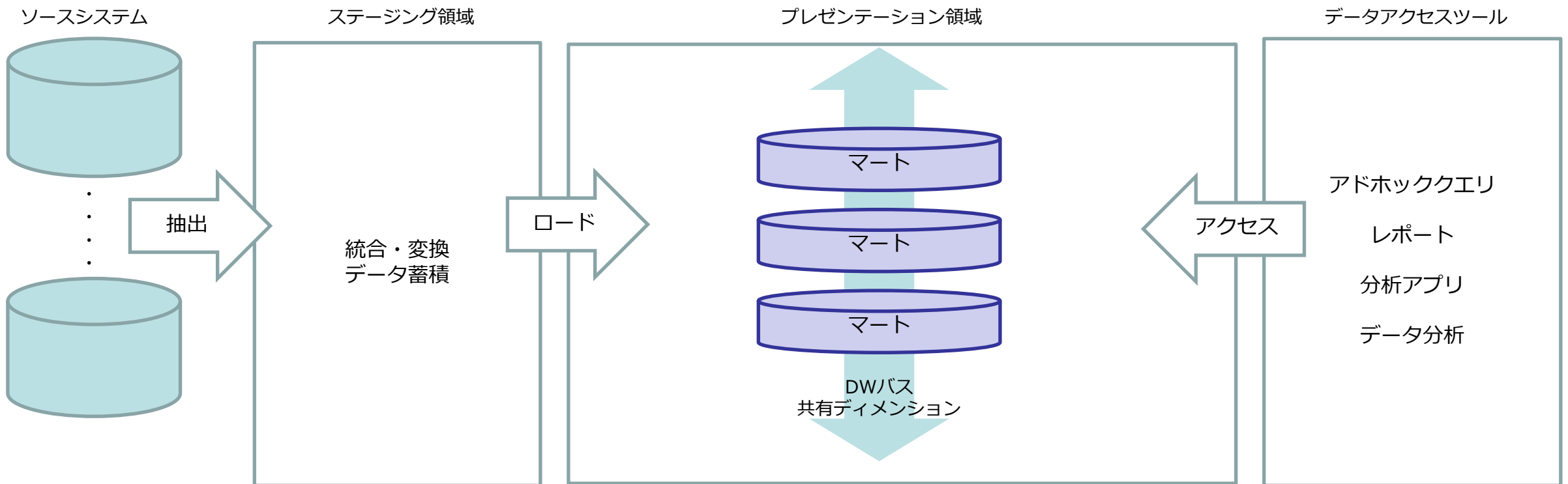
Bill Inmonが提唱したアーキテクチャの構造：コーポレートインフォメーション・ファクトリ（CIF）
データウェアハウジングの概念（全体像）を表現しており、以下の特徴がある。

- ・サブジェクト指向：主要な**業務実態に対して編成**される ※システムごとではなく、業務ごとに構成
- ・統合化された : **データは統一され、一貫性がある**
 (同じキー構造、データ構造の記号化と解読、データ定義、命名規則が全体で一貫して適用される)
- ・時系列 : **ある時点に存在したデータ**が保存される（それぞれの時点におけるデータの状態を反映する）
- ・不変的 : レコードは更新されず、**新しいデータは既存のデータに追加**される。
- ・集計・詳細データ：トランザクションを構成する最小レベルの明細とその要約データが含まれる
- ・履歴 : 現在のデータ+履歴データが含まれる

【データモデリング】Ralph Kimballのディメンショナルモデリング

Ralph Kimballはウェアハウスを「クエリと分析用に特別に構成されたトランザクションデータのコピー」として定義している。このアプローチではディメンショナルモデルが必要である。

Ralph Kimballが提唱したアーキテクチャの構造：ステージング領域とプレゼンテーション領域が定義されていて、ステージング領域の役割は比較的小さく、プレゼンテーション領域のデータマートが主体となっている。DWバスとも呼ばれる共有ディメンション（複数のスタースキーマが共有するディメンション）によってそれぞれのDMを全社レベルで統合することによって、エンタープライズデータモデルが形作られる。データプラットフォームの全体アーキテクチャとデータモデルが一体となっている。

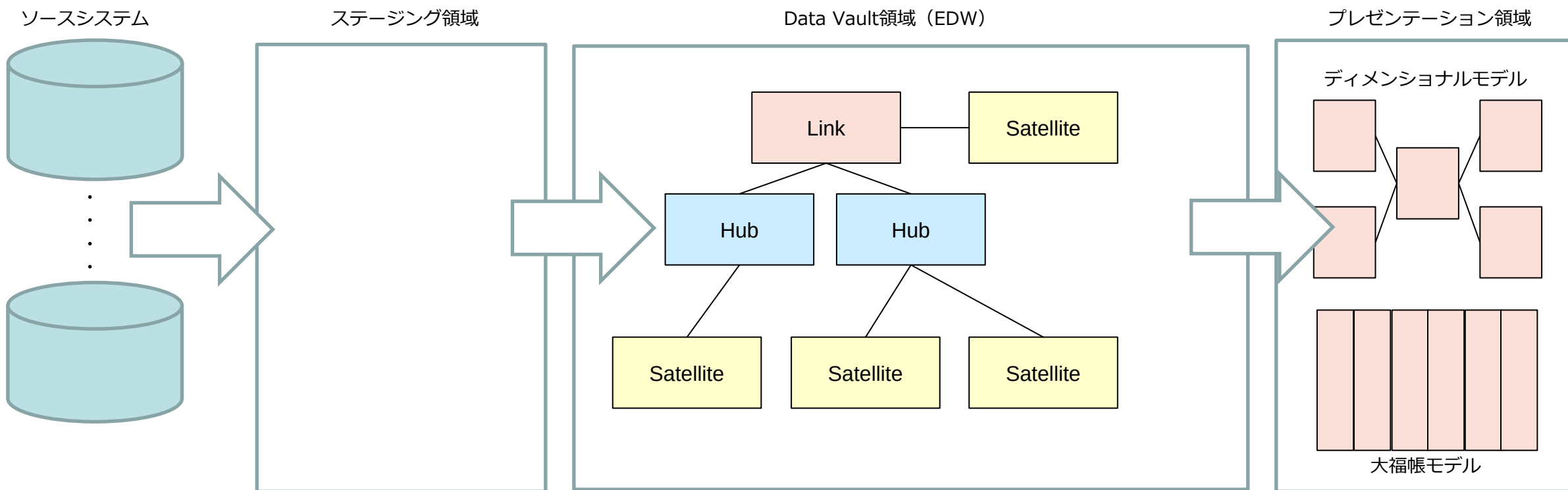


【データモデリング】Data Vault2.0

ステージング領域とプレゼンテーション領域の間にHub、Satellite、Linkと呼ばれる3種類のテーブルで作られたエンタープライズデータモデルの領域を作る手法

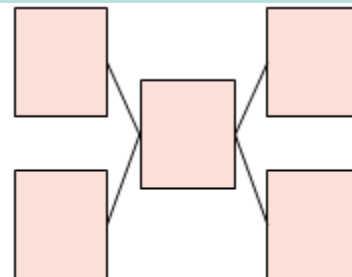
★特長

- ✓ **アジャイル性**：変更容易性、拡張性を重視
- ✓ **柔軟性**：きれいに正規化せず、元データソースの保持

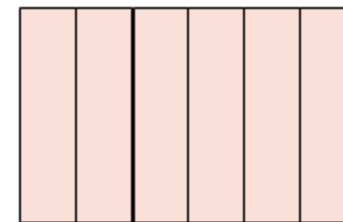


【データモデリング】BIツール向けの大福帳モデリング

テーブルが分割され、参照時に必ず結合が発生するディメンショナルモデリングは、列指向でストレージが安価なクラウドDWHにとっては、コスト対パフォーマンスが良くない、という意見があります。その代わりに、すべてのカラムが一つのテーブルにまとまっている横長のテーブル、いわゆる「大福帳」型のテーブルを提供すべき、という意見です。



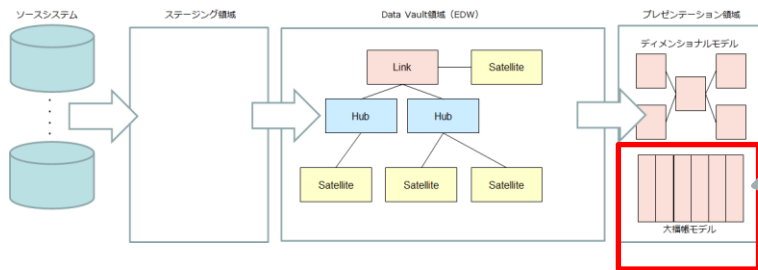
テーブルが分割されている
ディメンショナルモデル



関連するデータを一つの
テーブルにまとめる大福帳

これは性能を高めるためだけでなく、セルフBIの普及等によりデータリテラシーの比較的低いユーザが増加している現代において、可視性を高めるという意味でも有効なテクニックとして注目すべき技法です。

ただし、InmonやKimballのアプローチは、最終的にユーザに提供するデータモデルの形を定義しているだけではないことを忘れてはなりません。そのデータをどのように作るのか、どうやって拡張・メンテナンスしていくのか、DWH全体のモデルとしてどのように整合性をとっていくのか。こういったことを考えたうえで、**大福帳は最終的にエンドユーザに提供するデータモデルの一つとしてとらえるべき。**

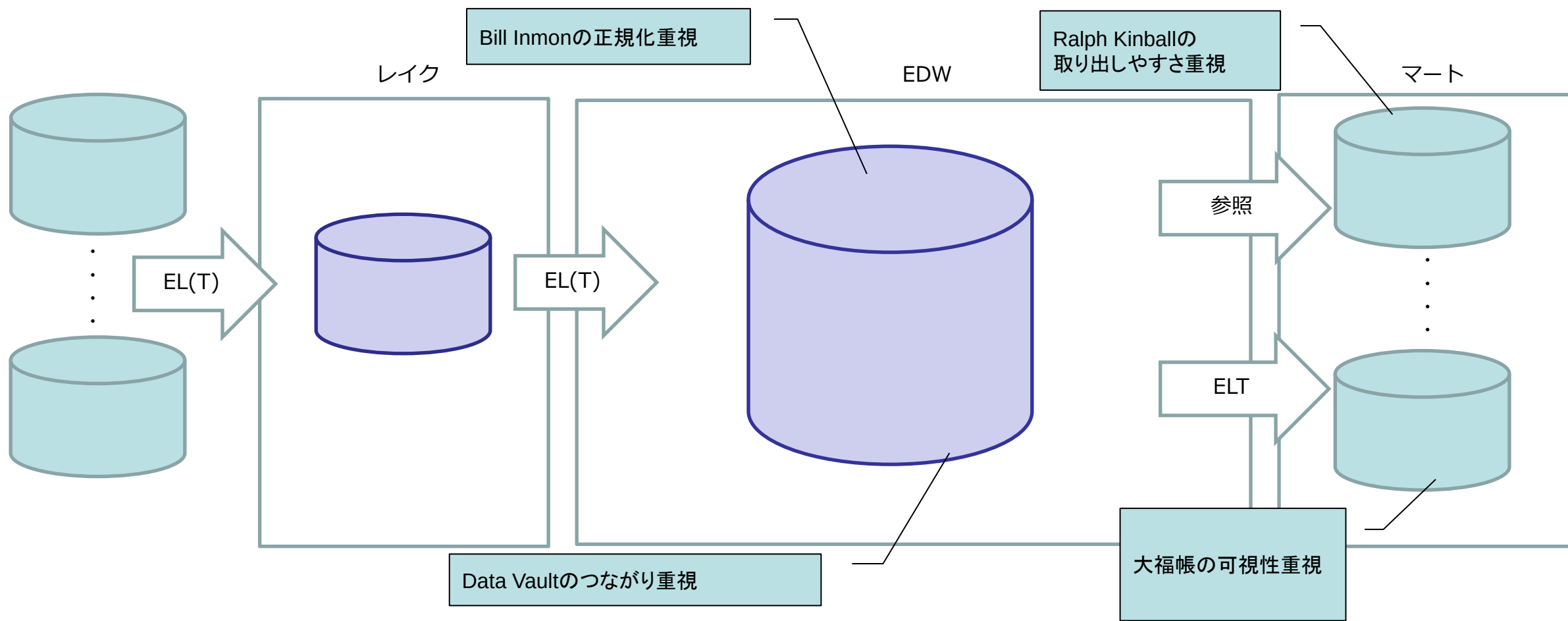


DataVaultでモデリングされたDWHからBIツール向けにデータ変換するプレゼンテーション領域で大福帳を作成するのが良いのではないかな？

実質的に無制限のデータ容量をもつクラウドDWHの登場により、データレイクとDWHの境界があいまいになった現代において、どのようなデータも最低限の加工をしたうえで一旦DWHに蓄積しておき、必要に応じて柔軟にデータマートを提供するというData Vaultの考え方は、時代に合っていてスマートです。また、大量の時系列データを様々なソースシステムから集めて利用する上で、必ずぶち当たるような問題に対し丁寧に対応が考えられており、実践的なモデリング技法でもあります。

【データモデリング】ユースケース合わせて選ぶもの

- 各技法の重視するポイントより、利用される場所は以下の通り。



【データモデリング】考察

- データ活用におけるデータモデリングは目指す姿に対してアプローチを変えたり、組み合わせる必要がある。

Bill Inmonの提唱する「統合化された状態」がベースであり理想的であると考ええる。

しかしコーポレートレベルの「統合化」には、多かれ少なかれ困難が伴う。

「統合化」出来なければビジネスインテリジェンスが成立しなくなると、多くの企業でデータ分析をあきらめることになる。

Ralph Kimballもステージング領域で統合化は行うものの、プレゼンテーション領域に重きを置いてと思われる。

しかし分析結果の説明責任のことを考えると、どのソースデータから、どう統合・変換されているかを把握する必要がある。

統合変換→EDW→マート
(Bill Inmon)

統合変換→マート (スタースキーマ)
(Ralph Kimball)

このBill Inmonのアーキテクチャにおける問題を解消しようとしたアプローチがDataVaultではないだろうか。

あらためて正規化モデルを作るのではなく、一旦データを投入し中間で見える化を図りながら改善を続ける。そこがアジャイル的と言われる所以であり、完全な正規化ができなくとも、データ分析を止めずに説明責任を果たそうとするアプローチなのだろうと考える。

Bill Inmon
(EDW)

Data Vault
(EDW)

また、大福帳はRalph Kimballのスタースキーマとは別のアプローチ（フラットスキーマ）として出現してきたものである。Bill Inmonの正規形をベースとしたマートを含め、どれが最適なのか、一概に言い切れるものではなく、分析要件やデータ統合状態によって最適解を検討する必要がある。

正規形→マート
(Bill Inmon)

スタースキーマ
(Ralph Kimball)

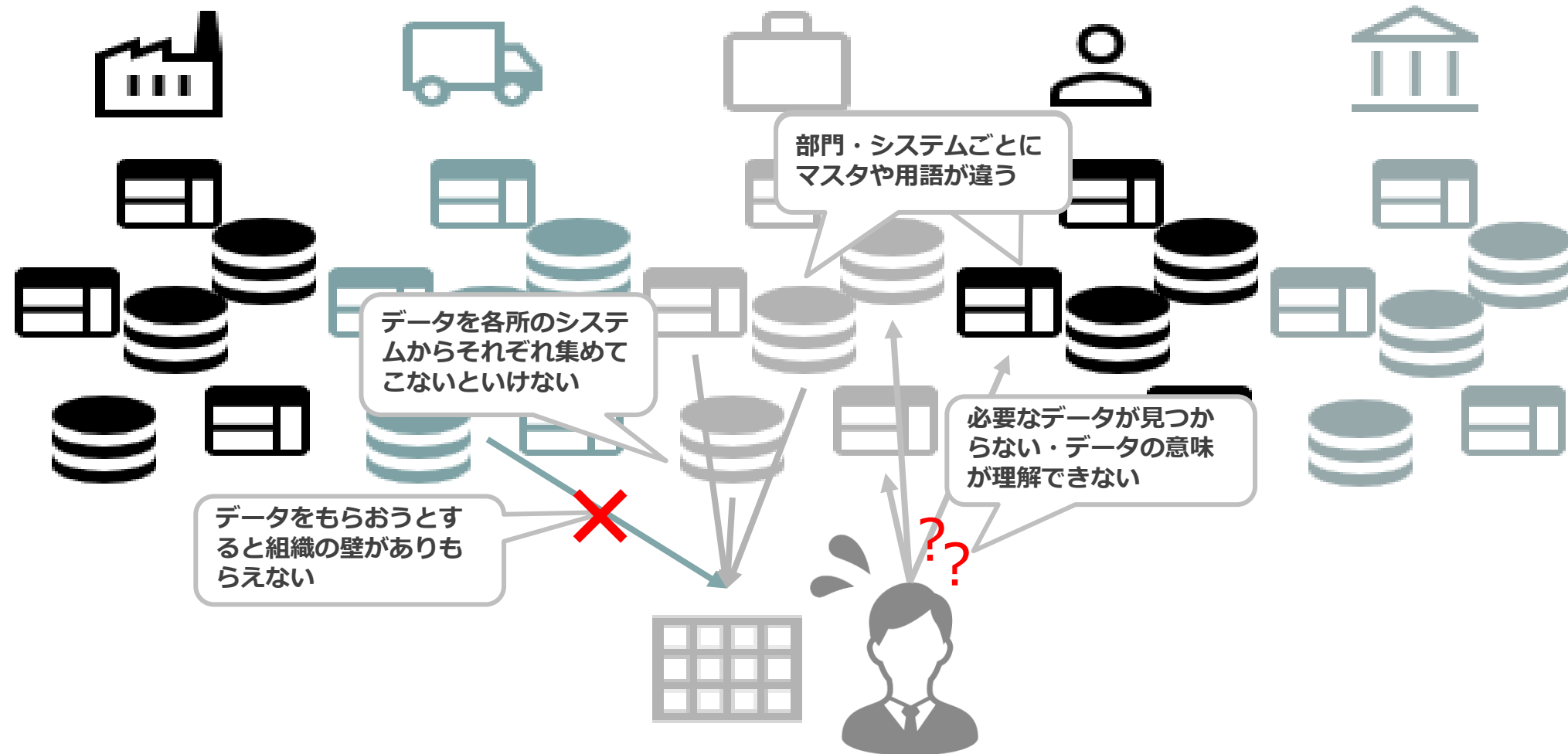
大福帳 (フラットスキーマ)

3. 分科会活動報告

【分散しているデータの統一】

【分散しているデータの統一】困りごと

データの分散により、分析が困難になっている



【分散しているデータの統一】データ統一のための5つの仮説

データの分散解決のために5つの仮説を検討

データ分散による問題

組織の壁によってデータが入手できない

データを各所のシステムから
それぞれ集めてこないといけない

必要なデータが見つからない
データの意味が理解できない

部門・システムごとにマスタや用語が異なる

データ統一のための課題

データ活用型の組織を構成する

全てのデータを一つの場所から
利用できる環境を整備する

データの所在・意味を一元的に管理する

全社共通のマスタデータを管理する

組織横断的な
データ活用推進
組織

データレイク
データウェア
ハウス

データ仮想化

データカタログ

マスターデータ
マネジメント

【分散しているデータの統一】組織横断的なデータ活用推進組織

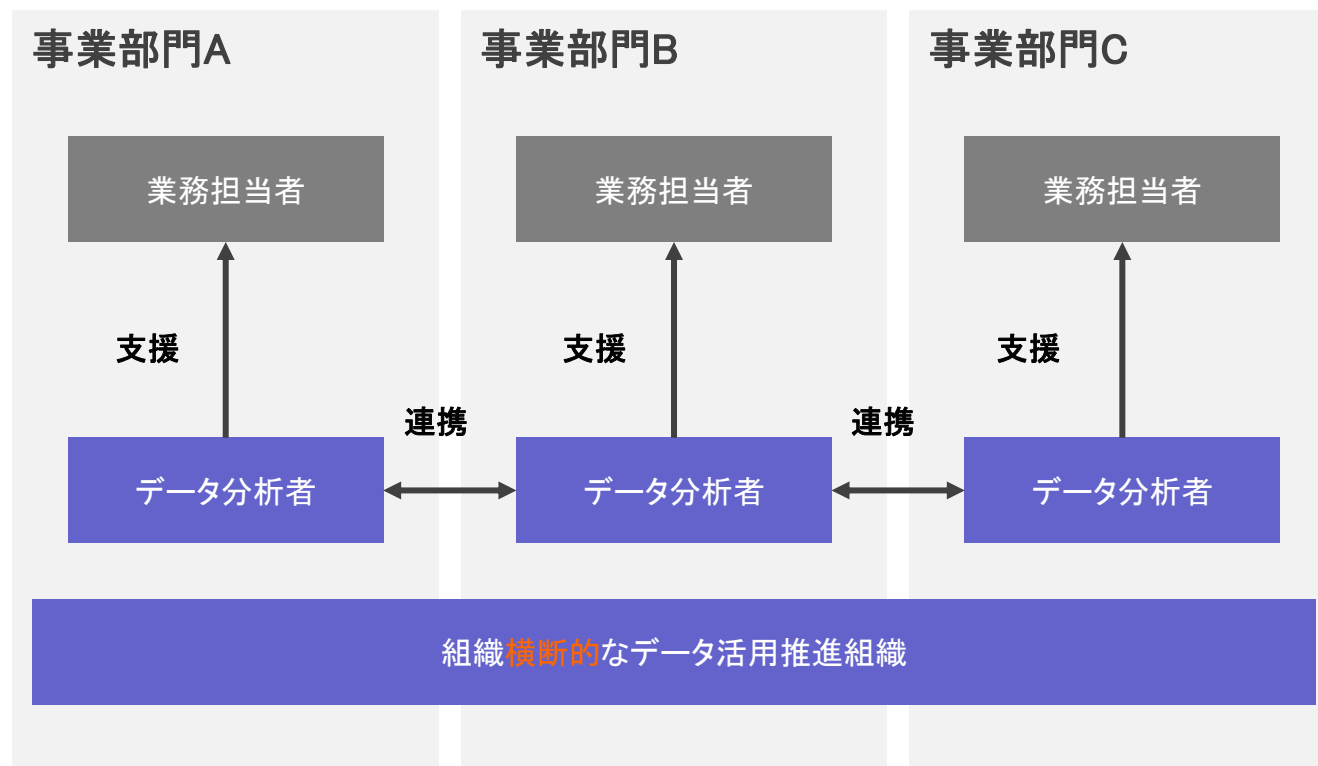
データ活用推進に向けた組織の在り方

共通して求められるのは部門を横断した組織であること
ただし、企業によって、最適なデータ活用組織の在り方は異なる

データ活用推進組織の役割

- 業務担当者へのデータ分析支援
- 組織間の利害の調整
- 専任的に全社のデータの維持・管理を行う
- 全社のデータの所在を横断的に把握する

データ活用組織のイメージ図

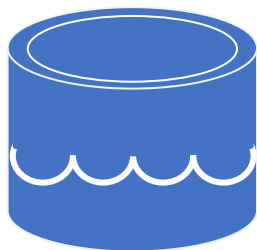


【分散しているデータの統一】データレイク/データウェアハウス

データレイク/データウェアハウスによる保存場所の統一

データレイク・データウェアハウスに全社データを統一して保存することでデータの連携負荷を軽減する

データレイク



データ保存場所の統一

データウェアハウス



データマート



概要

データの構造や大きさ、性質、種類などにかかわらず、あらゆるデータを
そのままの形で集めて保存する

様々なフォーマットのデータを分析しやすいよう集約・整理し、
検索が容易な形式で保存する

利用部門や用途、目的などに応じて必要なものだけを抽出・加工し、
利用しやすい形式で格納する

メリット

加工のコストをかけず、大量データを保存し、全社を横断したデータ共有が可能

データの時系列を把握できるため、総合的な分析が可能。

迅速かつ容易にデータを取り出すことができる

デメリット

未加工データを保存しているため、分析に工数がかかる

あらかじめ目的をもって設計されることが多いため、定型的な分析に陥りやすい

抽出するデータが限られるため、重層的・多角的な分析に不向き

【分散しているデータの統一】データの仮想化

データ仮想化とは

データを仮想的に結合し、アクセスポイントを一つにすることで
組織・システムを横断したデータ活用が少ないリソースで実施できる

データ仮想化とは？

様々な場所・形式で分散するデータを物理的に複製することなく、仮想的なデータベースとして、統合的に参照できるようにする技術

- リアルタイム分析

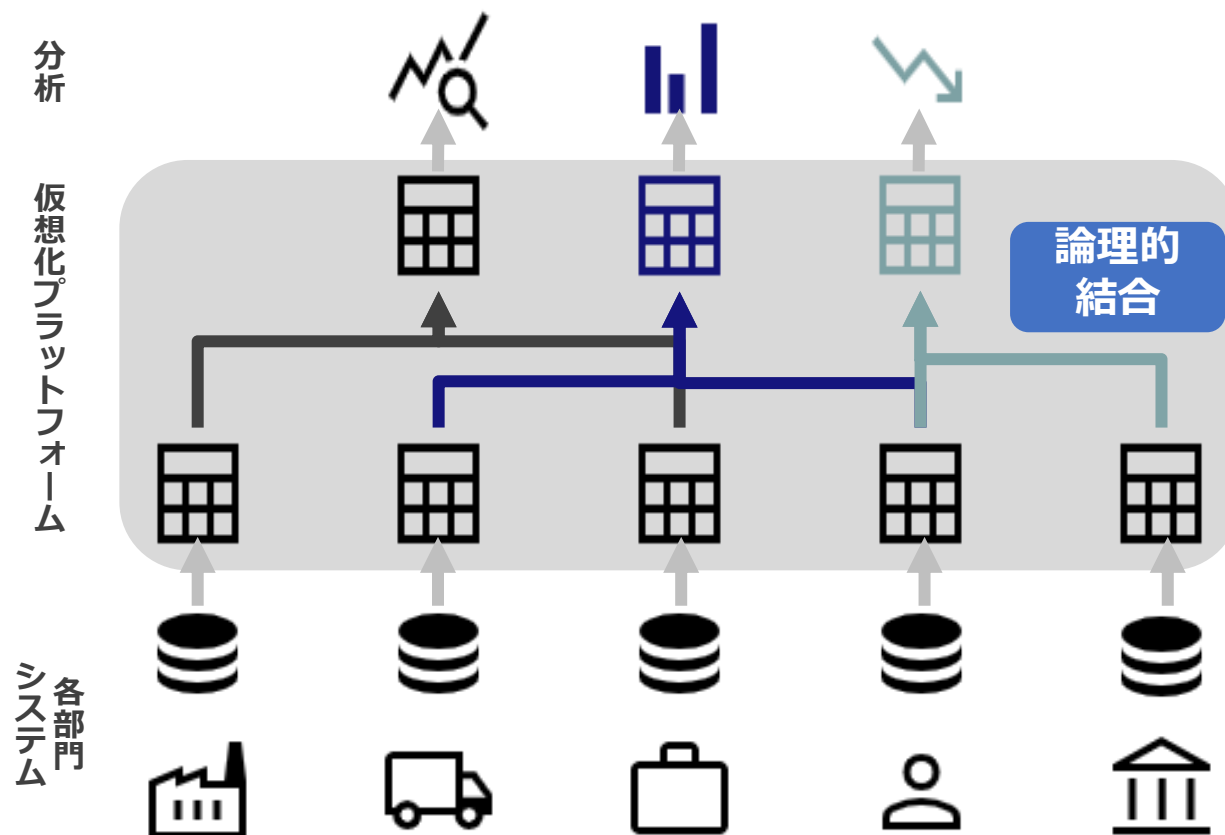
データを複製することなく、
各々のシステムを直接参照

- 開発リソース削減

データを複製しないため、
テーブル設計などの開発コストを削減

- セキュリティの一括管理

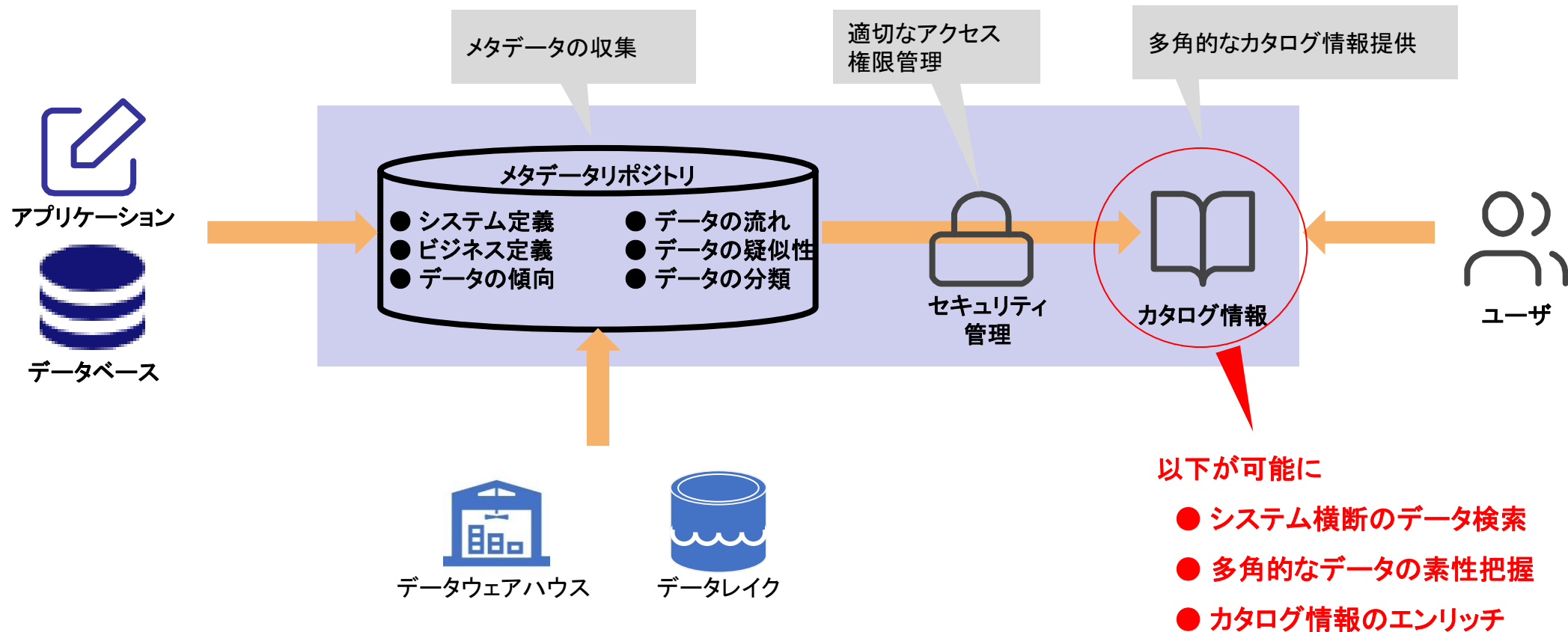
単一のアクセスポイントによる一元管理



【分散しているデータの統一】データカタログ

データカタログとは

各種データ資産をカタログ化し、あらゆるユーザーによる検索と把握を支援



【分散しているデータの統一】マスタデータマネジメント

マスタデータマネジメント

マスタデータ管理 (Master Data Management = MDM) とは、業務で扱う基本データ、つまりマスタデータを全社の観点で統合させ、品質を維持する活動である。マスタデータマネジメントを行うことで、部門やシステムの違いを意識せず横断的に分析することが出来る。

MDM=MDMシステムでは無い。

重要なのは「人」そして「プロセス」であり、MDMシステムはその後に位置するものである。

目的・要件の定義

「業務横断での売り上げ拡大」「情報漏洩の危険性を防ぐセキュリティ強化」等、優先的に実現したいビジネス目的を検討する。

⇒漠然と「統合」を進めようとしても、企画が進まない。

現状分析

現在、どのシステム(もしくはシステム外)で、誰が、いつ、マスタデータを生成・管理している課の現状を確認する。

⇒データの更新頻度、粒度、過去データ有無、例外データ、フォーマット等を可能な限り確認していきたい。

管理対象データの定義

どのシステムの、どのデータを管理対象とすべきかを検討する。

⇒古いデータ、使えないデータを対象としても活用は困難。影響範囲・活用想定を踏まえて優先順位をつけて決定していく。

実装方式・アーキテクチャ検討

ITツールの活用も含めて、どのようにMDMを実装するかを検討する。

⇒「マスタオブマスタによる中央集権」「各システムからの集信」「各システム間のキーのマッピングによる論理集約」等、いくつかの実装方式から、適した方式及びツールを検討する。

運用プロセス定義

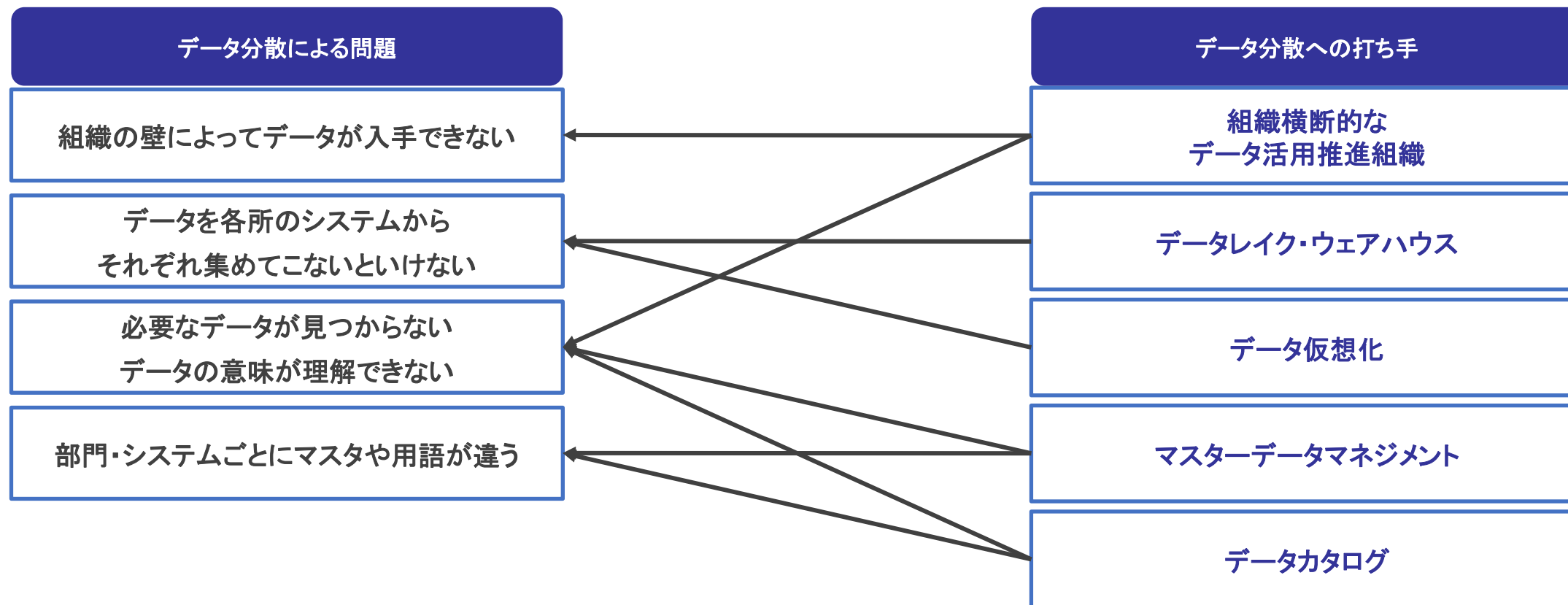
データ登録・収集/更新を誰が、いつ、どのように業務プロセスとして実現していくを検討する。

⇒データの初期登録だけではなく、継続的な品質維持を実現するための体制・プロセスを合わせて検討する。

全てのマスタデータを同じタイミングで統合していくのは非常に困難である。
システム刷新・再構築などに合わせて、優先順位を決めて段階的に進めていくのが現実解

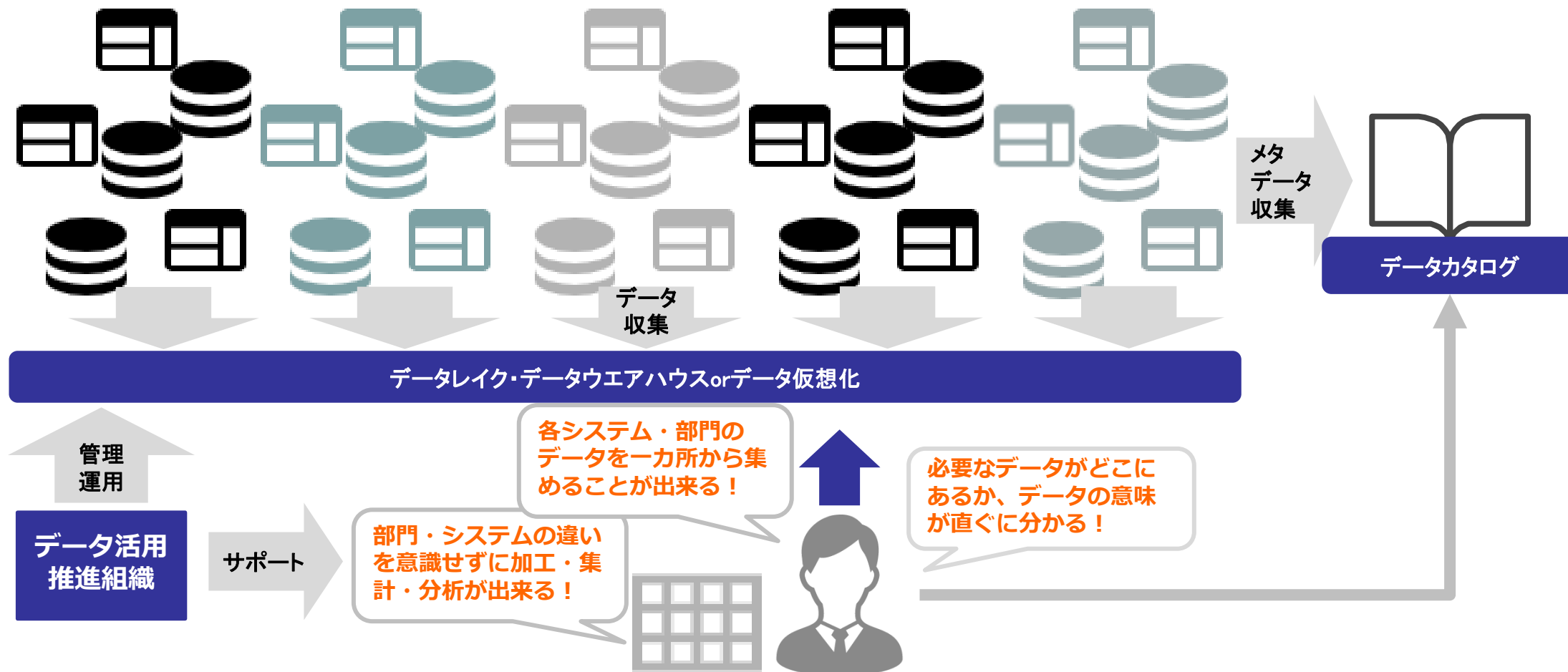
【分散しているデータの統一】問題とデータ統一への打ち手

データの統一観点からの打ち手の効果



【分散しているデータの統一】データ統一後の姿

データの統一によりデータの分析にかかるハードルが取り除かれる



以上

