

教育データ活用における検討事項

東京都渋谷区教育委員会様 発表補足資料

2022年10月

日本マイクロソフト株式会社
文教営業統括本部

宝の有効活用：教育Bigデータに必要なデータ分別

MEXCBT:CBTデータ 

Microsoft365 

G WorkSpace 

OSの違い 

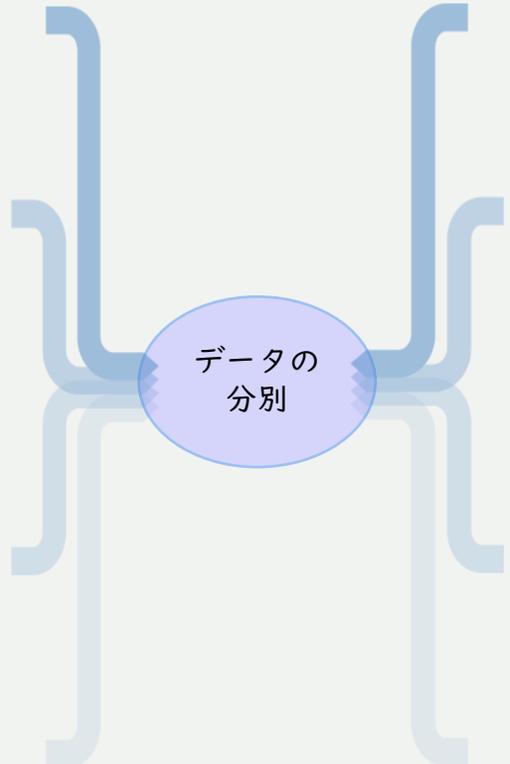
 ログとストリーム
(非構造化)

 外部データ
地域特性

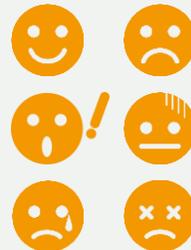
 メディア
(非構造化)

 ファイル
(非構造化)

 FORM
(アンケート)




数値主体のデータ


人の気持ちや
感情データ


人の無意識化の
行動データ

MEXCBT:CBTデータ
保健室利用など、校務支援データ
ドリルなど、教科書データ
成績など通知表データ
部活動など、児童学習外データ
教室内外の状況データ (IoTデータ)

児童生徒の心のデータ
関係者のアンケートデータ
授業などの満足度データ
作文、感想文データ
課外活動の成果物データ

児童生徒のコミュニケーションデータ
関係者の関わりデータ
授業などの集中度データ
課題提出に関わる時間データ
授業への関与データ

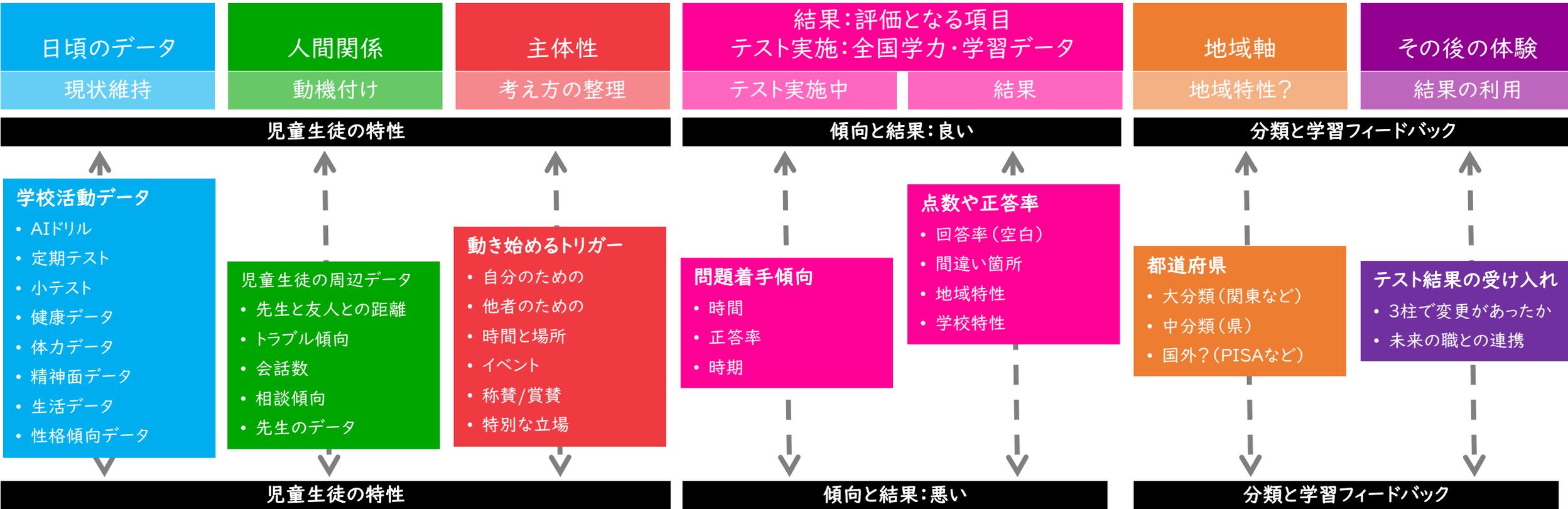



行動変異が起きた要因は?
良い方向、悪い方向


時間軸での変化


様々なデータの相関関係

データの推移について(イメージ)



上記は、例えば「全国学力・学習データ（またはテスト結果）のような、成績をあげるためにデータを活用したい」という場合、「結果：評価となる項目」だけを分析しても意味は小さくなります。当日のテストの時点で、結果はほぼ確定しているので、むしろその前段階のデータの「組み合わせ」で示唆を導き出していく必要があります。上記の結果、データの収集段階で目的の明確化と調査の計画（スライド10）、分析したい目的と項目に応じて前後の必要となるパラメータは変化します。その組み合わせが多いために、データ分析の「標準化」を定めにくくしている。

データの推移について(イメージ)



ひとりひとりのデータ統合の重要性について

複数のクラウドサービス (ID) の利用を前提としたデータ活用

課題1

県内の各市町村が3つOSがバラバラに採用されたため、IDが統合されていない

課題2

各社のサービスごとに独自のIDを発行しており、横串してデータを活用できない

解決策

各サービスごとに発行するIDを1個人のIDに統合して紐づけて利用

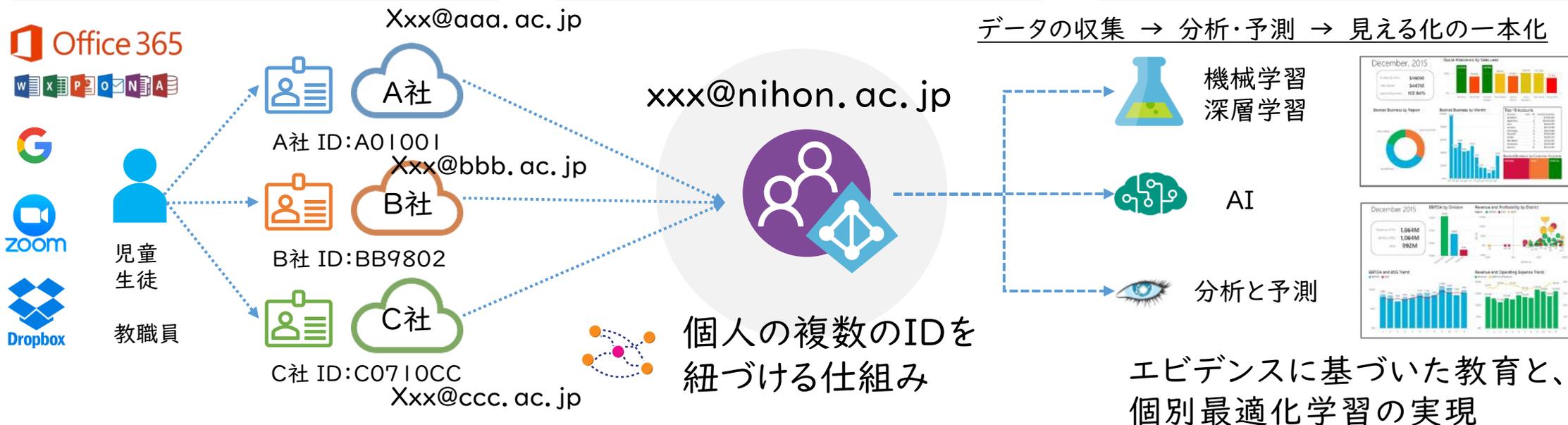
解決策

個人に紐づけされたIDに対して各サービスのすべてのデータを収集して分析

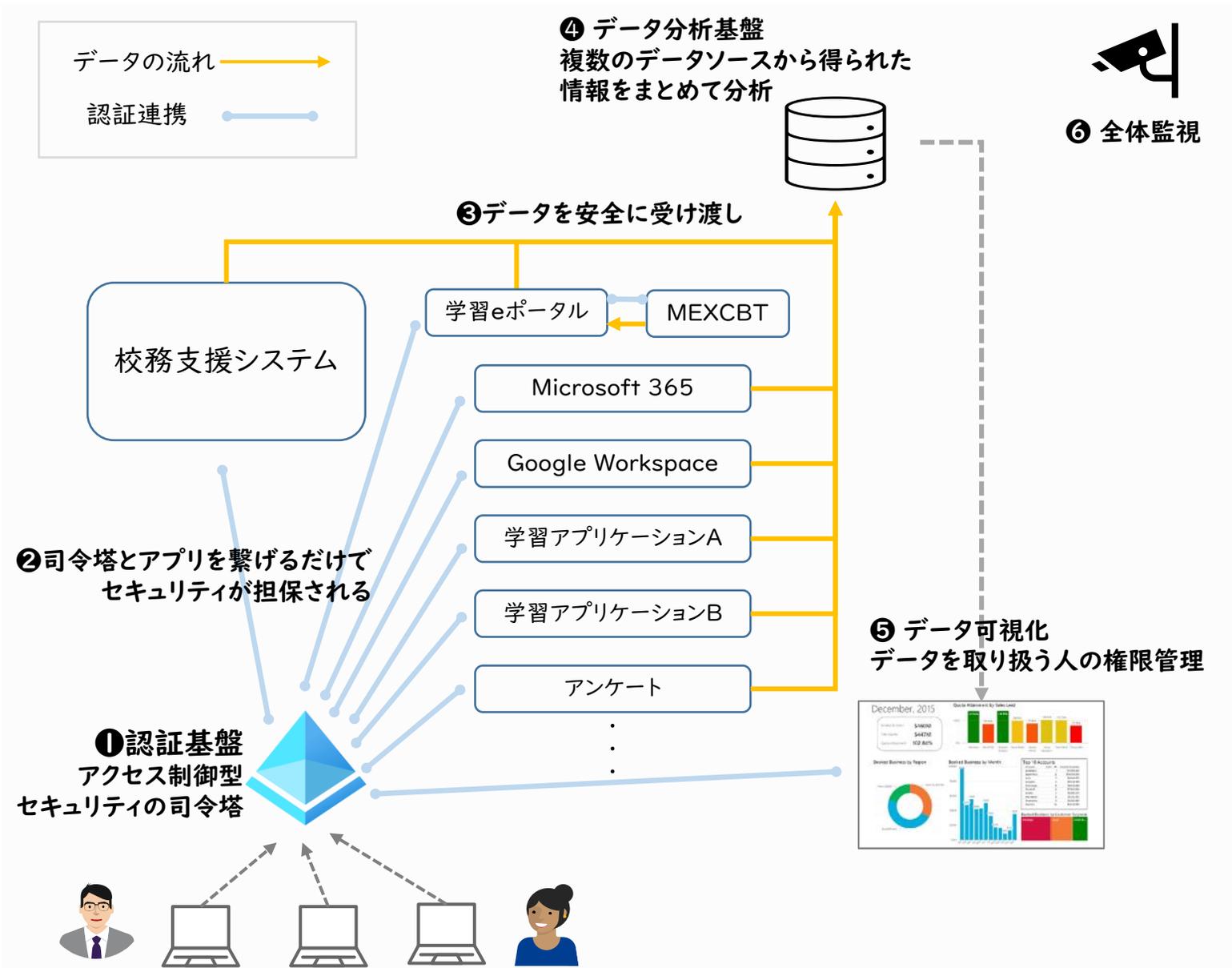
複数ある学びのデータ

個人IDの統合

複合的なデータの活用



セキュリティが担保された環境で、安全にデータ利活用を進めるために



アクセス制御型セキュリティを活用するメリット

セキュリティ対策を一か所に集約できるので

- ✓ セキュリティにかかる費用を抑えることができる。
- ✓ シンプルな構成になり、管理負荷が低い。
- ✓ データソースを増やす時にも追加投資不要。

安全にデータ利活用を進めるためのセキュリティの仕組み

- ① アクセス制御型セキュリティの司令塔の役割を果たし、全体を管理します。
- ② 認証基盤とアプリケーションを連携させることで、セキュリティを担保することができます。
- ③ アプリ・システムからデータ活用基盤への、データ受け渡し経路は暗号化されており安全です。
- ④ 安全が担保されたクラウド上にデータを貯めて、分析します。
- ⑤ 誰がどのデータを見てよいかの権限管理を行い、不適切な閲覧を防止することができます。これも司令塔（認証基盤）が担います。
- ⑥ 不正にデータが取り扱われていないか全体を監視します。

データ活用の難しさについて

1990年代に入り計算機科学の分野から、大規模データを分析する技術として、データマイニングが提案されました。データマイニングという言葉自体は1970年代から使われてきましたが、大量のデータからパターン認識、人工知能、統計学等の理論を駆使し、有用な情報（知識）を導き出すプロセス、という意味合いで広く一般に用いられるようになったのは1990年代に入ってからです。

統計学や計算機科学の専門知識がなくても使える操作性のよいツールが開発され、実務の現場にも導入されました。結果としてデータマイニングにより「データ=知識」という概念が一般のものになりました。

しかしながら、多くのビジネス分野において、新たな知識発見という当初に期待されたような意味で成功を収めているとは言いがたい状況が続いております。

GIGAスクール構想の下での校務の情報化に係る論点整理（中間まとめ）を拝見させて頂き、BIに関係しそうで、この辺りが整理されると、一気に議論が進むかと想定しますポイントを下記にまとめました。



データのサイロ化

ネットワーク分離環境



種々雑多な
データ型の不一致

データ活用を前提に
なっていない



ソリューションの複雑さ

データ活用前提で
無い環境構成



マルチクラウド
環境

これからどんな
サービスが増えるかも
分からない



コストの増大

ハブ環境を考えるだけで
めまいが起きる

データ活用の難しさはテクノロジーで解決できます

膨大な非構造化データを蓄積し、リアルタイムに操作することは先端技術等の技術により可能となり、これらのITを活用すれば何か新しい知識が発見できるかもしれない、という期待が高まっています。

しかし、教育において、知識はデータの中ではなく人間の頭の中にも存在します。

その意味で教育データ活用に関しては、複雑な現象から機械的に知識を得るというのではなく、まずは単純な現象（仮説）を複雑なデータから検証するというアプローチから始めることが望まれると経験から理解いたしました。

すなわち、データマイニングと同様に、教育データ活用のポイントも問題の特定化と仮説検証型アプローチにあると考えています。

その意味で実務家、すなわち教職員と教育ドメインに関しての知識と問題意識を持ち、仮説を構築できる人材の参加がますます重要となってくると想定しております。教育データの活用にはIT、分析の専門家だけでなく、より多くの実務家の参加が必要です。

データのサイロ化



すべてのデータを
1つに吸い上げます

種々雑多な
データ型



多様なデータ型の
サポート

ソリューションの複雑さ



オンプレとクラウド両
方への対応、複数
システムへの対応

マルチクラウド
環境



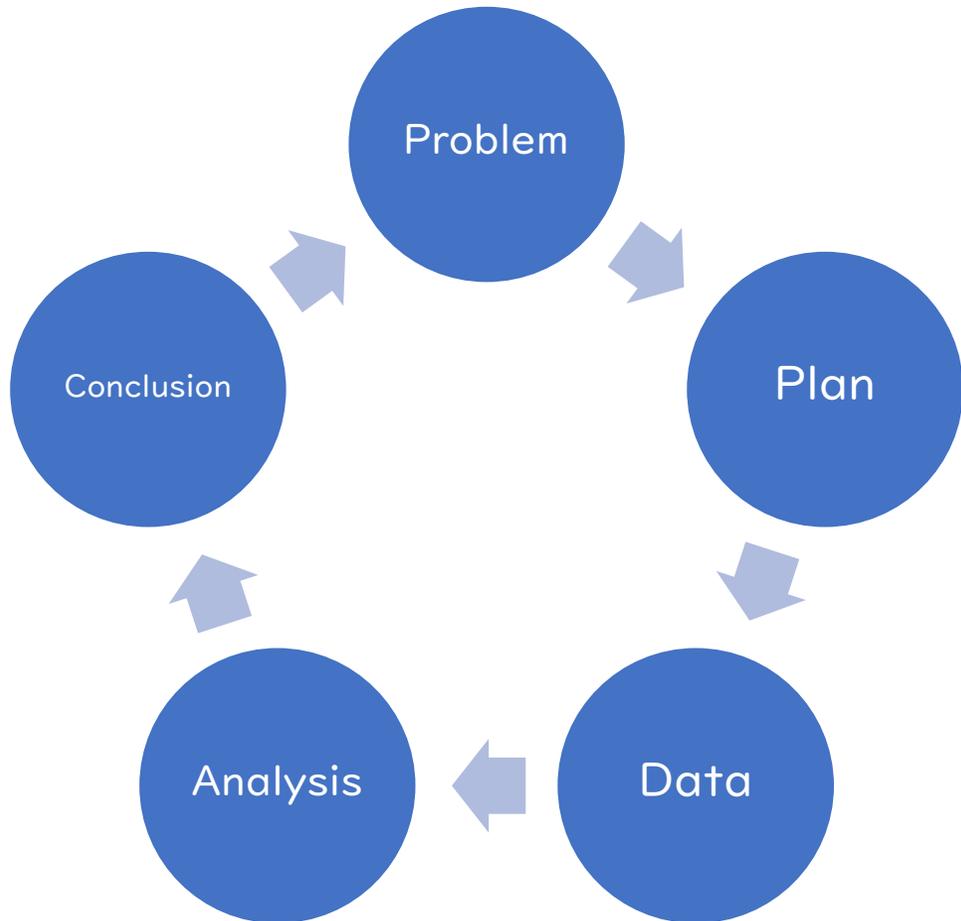
どのクラウドを利用して
いてもOK

コストの増大



自治体が独自分析し
ない限り、GIGA環境
を利用想定

なぜデータ標準化が難しいか：育てる前提が必要



Problem: 問題の明確化

→データを集めることによって確かめることができるような問題を定義

Plan: 実験・調査の計画

→Problemで明確になった問題に対して、どのように実験や調査を実施するかを決定（統計手法、分析アルゴリズム特定）

→主語が重要で、誰に対してどのような測定をおこなうか？

Data: データの収集→ここからスタートするとほぼ失敗

→Planで策定した計画に基づいて実施

→データ収集の際に生じる、欠測値の問題や回答の誤りなどに対しても適切に対応が必要

→測定値の有効桁数の設定や、測定に際して生じる誤りの修正

→ある程度「整形化」するのでデータは少し汚くても良い

Analysis: データの分析: BI

→データを集計した結果を表としてまとめたり、グラフを使って表現

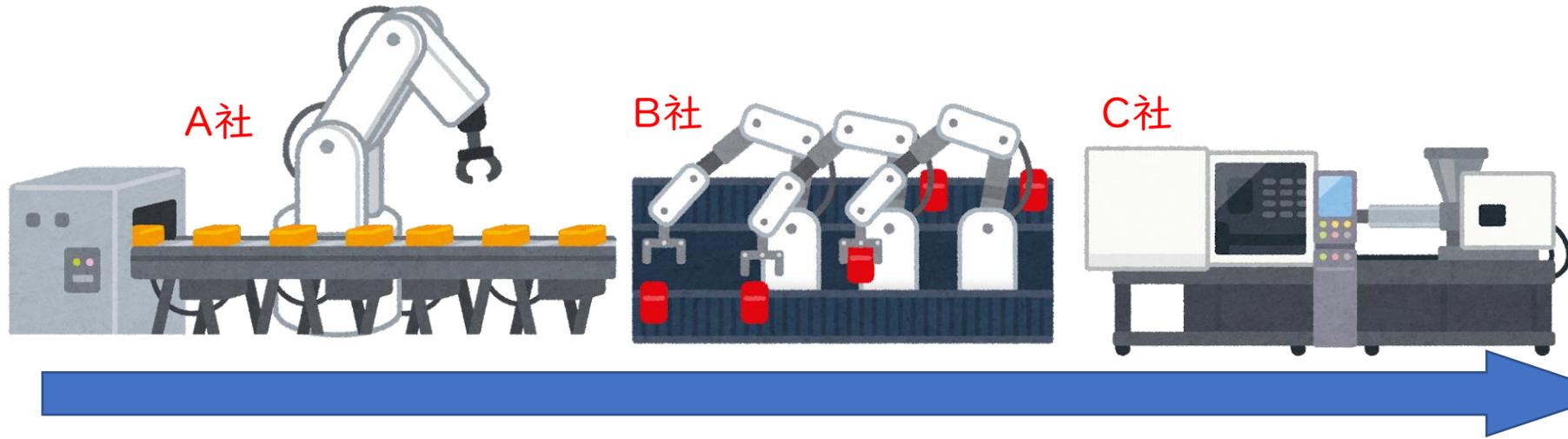
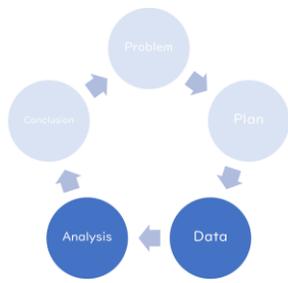
Conclusion: 問題の解決

→問題に対して明確な判断ができない場合には次の問題を考える必要がある。

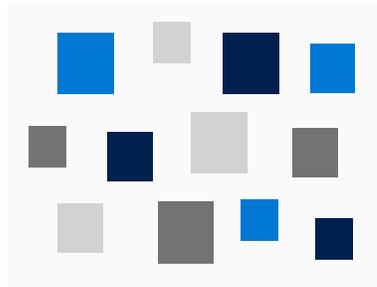
→問題に対してある程度の示唆がでた場合には**データ項目を微調整**する。

→Problemに戻る

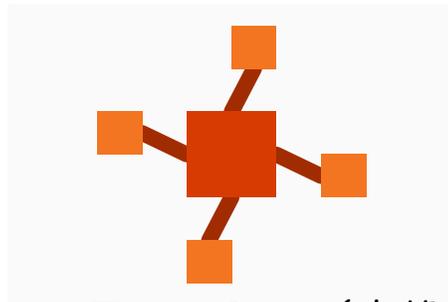
標準化は難しいので…ETLを活用しデータの変換を実施



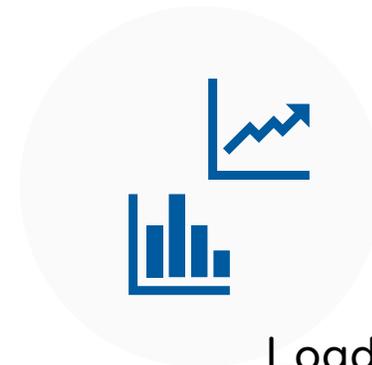
データ項目の標準化について出来ている業界は無い…。



Extract (抽出)



Transform (変換:加工)



Load (書き出し:格納)

基幹システムやデータベースからデータの抽出を行います。どのような利用目的でデータを抽出するのかを明確にして作業に臨みます。利用目的の明確化により、不要な情報は抽出されず、その後の作業を効率よく進めることが可能です。なお、自組織にデータを保管し、ETL処理を業者に委託する場合は、データの抽出作業を自社で行います。セキュリティ面を考慮し、委託先がデータ抽出すること推奨されません。

データ格納先に書き出しやすいよう、一定の規則に従いデータの変換・加工を行います。「抽出」のプロセスで書き出しやすいデータ形式になっている場合は、このプロセスを省くことが可能です。しかし、適したデータ形式を把握しておらず、別の形式で抽出してしまうと変換・加工に大幅な時間がかかります。そのため、各プロセスで担当者が異なる場合や抽出・変換を委託する場合、情報の共有が大切になります。

変換・加工のプロセスで作成したデータファイルをデータ格納先へ書き出します。

教育データ利活用に関して必要と考えられる機能群



GIGA
Office 365/G-workのデジタル
アクティビティ

校務支援システム
欠席情報
保健室利用情報

アンケート等
非認知情報

テスト等
学力

クラス運営で先生が見る

クラス運営内のデジタルアクティビティを可視化

- ✓ 課題
- ✓ コミュニケーション
- ✓ ファイルの使用
- ✓ 会議
- ✓ OneNote
- ✓ リフレクト(心の天気)

学校全体で教職員が見る

テナント(学校単位、学年単位等)のデジタルアクティビティの可視化

- ✓ 課題
- ✓ コミュニケーション
- ✓ ファイルの使用
- ✓ 会議
- ✓ OneNote
- ✓ リフレクト(心の天気)

クラス運営で先生が独自視点を入れて見る

端末で先生が可視化できるBIによるアクティビティ

- ✓ Office 365 の各サービスの利用状況
- ✓ 監査ログの可視化

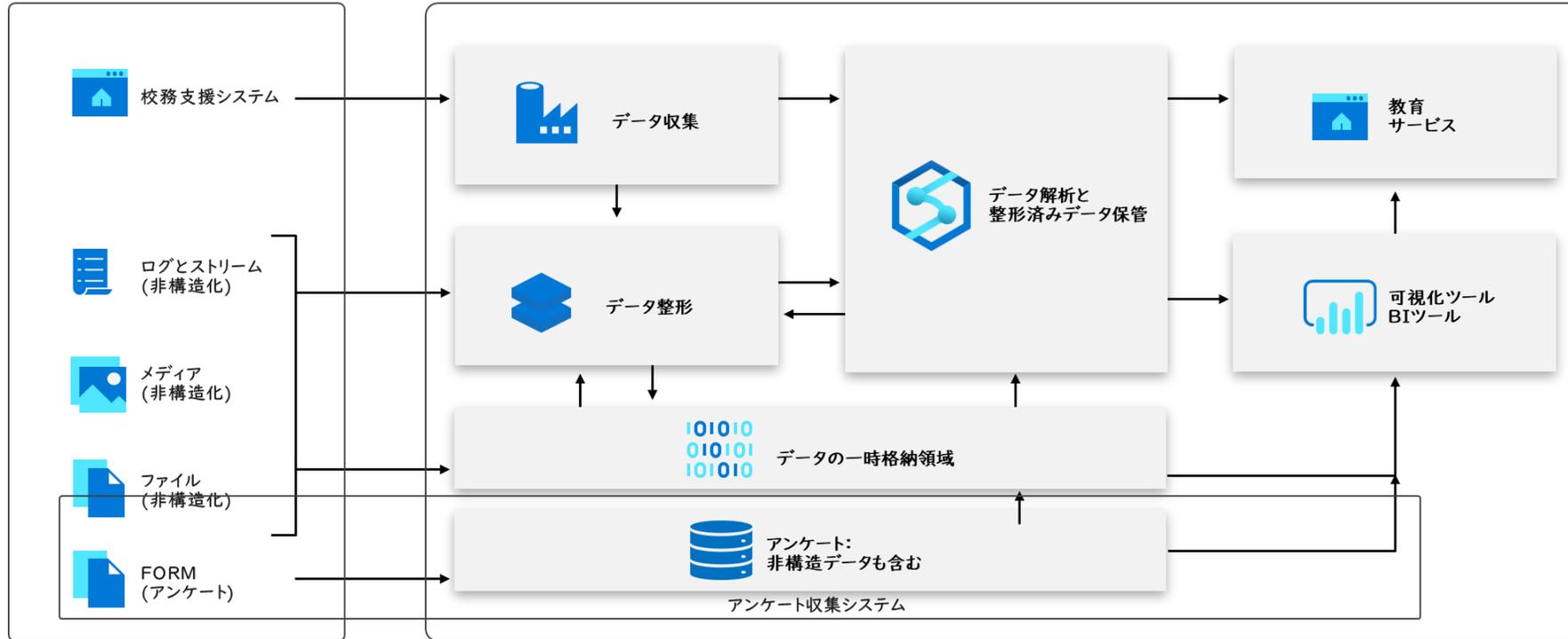
**自治体、行政、自由な単位
自由な分析で見える**

校務支援やアンケートを含むすべてのデータの可視化
学習ログ分析基盤

どんな状況にも対応する、最新のデータ活用基盤

データ取得元（教育現場）

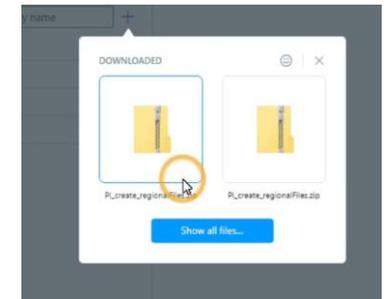
データ分析基盤（例：自治体）



ETL設定操作は、マウス & GUIで



作った設定はテンプレートに



全てがPaaSで構築されており、データ活用基盤上のシステムについて個別のバックアップやバージョンアップは不要。また、リソースも「自動スケール」で動作します。使うか分からない、または、加工が必要なデータは“一時格納領域”に、活用が推奨されてるデータは、活用領域に統計処理などを得て保管されます。

「将来これくらい大きくなる」という想定があったとしても、利用開始時点ではその想定データ量に応じた金額を支払う必要はありません。

まとめ：教育データ活用のBIについて

- 生徒個人の複数のIDを紐づける統合IDの活用（纏められない場合は名寄せ）
Security的にもアクセス制御型ベースでの管理ができる
- 分析/活用したいデータの前後との組み合わせが重要
教職員様とBI担当者が一緒に活動する（BI作り手が寄り添う）
- データの受け渡しを制御するETLが重要
自治体独自分析がある場合はETL同士の連携、無い場合はGIGA環境利用
- データ分析環境自体はそんなに高くない（傾向）
教育データの分析は、様々なデータを掛け合わせる次の工程が大切
教育データから示唆を導き出す方を組織の中で育てることも大切