データサイエンス

すいすい会

第27回「DXを高速推進するための クラウド・データ統合ツール Matillionの紹介」

GRI
BIGDATA SCIENCE
&RESEARCH

すいすい会の紹介

■データサイエンスに関する知見を共有しあい
　みんなで実践的な理解度を高めたい

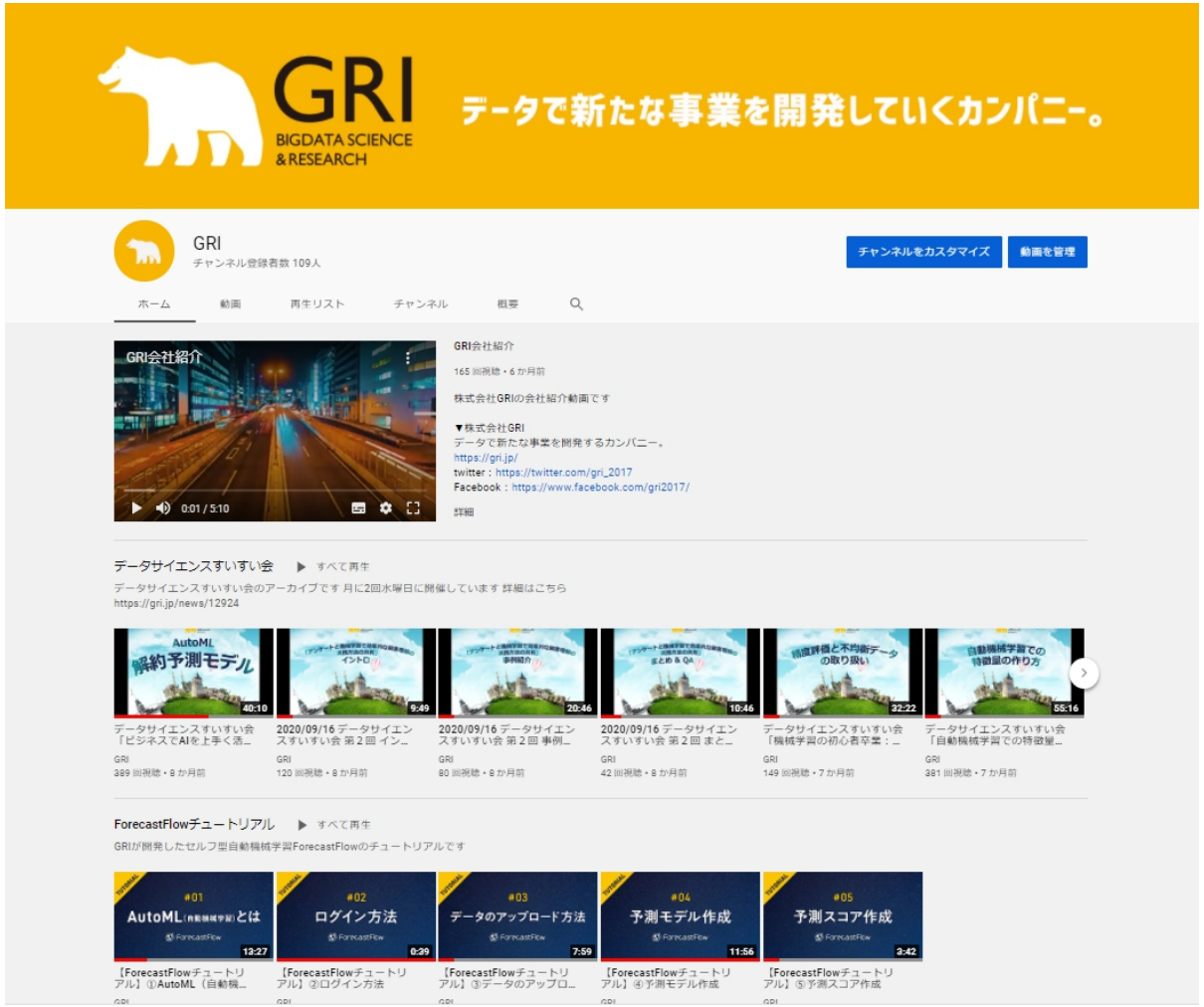
- 実践を重視しているGRIなりの考え方を共有します
- 集合知にするため、皆さんの知見をご共有いただきたいです

■資料：　GRIホームページ　https://gri.jp/news/12924

■Slack　ForecastFlowチャンネル　https://join.slack.com/t/forecastflowusers/shared_invite/enQtNTgyMjcxOTg0NzcxLTBkOWEzYWMwNDJmNTkyMDQzYmIxYWU0YWI4ZmU3ZDU0ZTMxNDUwODAxMWFmYmU1YjJiZGI0MjRhYWYyYTNlZTQ

過去のすいすい会の動画（YouTube）

https://www.youtube.com/c/GRIinc



ナビゲータの背景

古幡征史
株式会社GRI　取締役

- 株式会社GRI 取締役
- Ph.D in Computer Science
- GRIにて50以上のAI, BI, 分析基盤構築プロジェクトをリード
- KPMGコンサルティング、University of Southern California、ドワンゴを経て、2016年9月より現職

©GRI Inc.

データサイエンス

すいすい会

第27回「DXを高速推進するための クラウド・データ統合ツール Matillionの紹介」

GRI
BIGDATA SCIENCE
&RESEARCH

本日のアジェンダ

1．Matillionを使うと高速にDXを推進できる理由
　　株式会社GRI　古幡征史

2．Matillionの機能の具体例
　　株式会社GRI　高倉千紘

3．Matillionと自動機械学習ツールを連携させる応用例
　　株式会社GRI　岡部泰三

©GRI Inc.

1. Matillionを使うと高速にDXを推進できる理由

©GRI Inc.

DXを推進する際、よくある課題

- データを収集するだけで時間がかかり、分析までたどり着かない
- データが大規模すぎて、処理できない（小さいデータならば動くのに）
- ノウハウを持っている経験者が足りない
- 上司の理解が得られない
- 一度は基盤を作ってみたが、使えない

↓

現代的なデータ分析基盤のアーキテクチュアとツールを採用してください
（素質があれば未経験者が経験者にすぐになれます）

©GRI Inc.

データ分析基盤の中心は列指向データベース

- 大規模データを扱うには列指向データベースを利用
  - 超高速にSQLクエリが返ってきます
    - Tableau Onlineを例に取ると、データの行が1,000万行を超えるならばBQやSnowflakeの活用を考えはじめるべき
  - 選択肢が自由ならば、BigQueryかSnowflakeがコスパ最強
    - TreasureDataやRedshiftから乗り換えるユーザが多い
    - SQL Server、Oracle、MySQLで間違った人も乗り換え

Google BigQuery

snowflake

©GRI Inc.

列指向DBとETLツールの組み合わせ、おすすめできないです

列指向データベース

Tableau Prep
(デスクトップ)

データを一度ローカルPCに取り出し、ローカルマシーンのリソースを使って整形処理を行うので、大規模データを扱えない
※AlteryxのIn-DBは良いもの

列指向データベース

Cloud Dataprep
AWS Glue

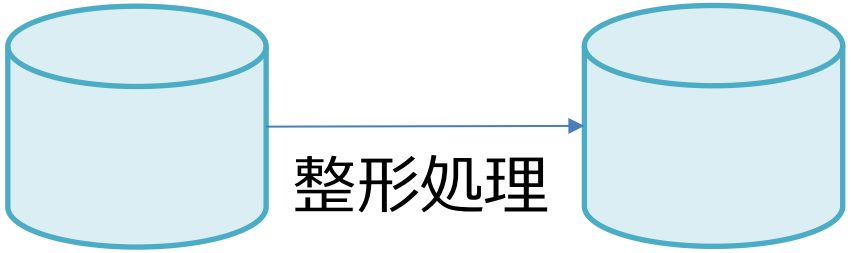
データを一度クラウドインスタンスに取り出し、バーチャルマシーンのリソースを使って整形処理を行うので、高額請求で、大規模データの処理が厳しい

©GRI Inc.

列指向DBは列指向DBの中で処理をすべき

高速にデータ整形処理を実行するには列指向DB内で
処理した後、必要に応じて取り出す

列指向データベース



ただし、以下の課題がある

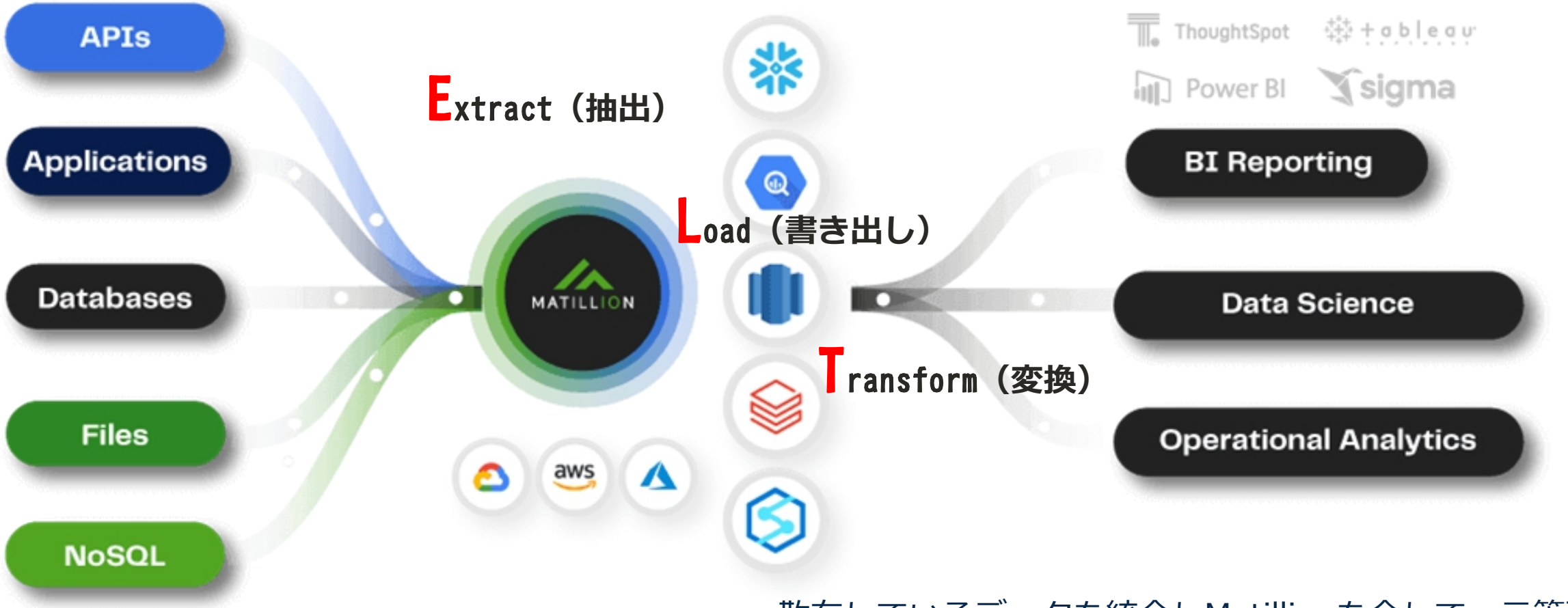
- データエンジニア／SQL職人の確保が困難
- SQLが複雑になり引き継ぎやメンテナンスが難しくなる
- エラー時のリトライなどスケジューリングの運用が難しくなる

Matillion（ELTツール）で解決

©GRI Inc.

Matillionとは

クラウドの列指向DB向けに構築されたELTツール
様々なデータソースをクラウドの列指向DBへロードするフローを簡単に作成可能
データ整形にはBigQueryの高性能リソースを利用できる



散在しているデータを統合しMatillionを介して一元管理

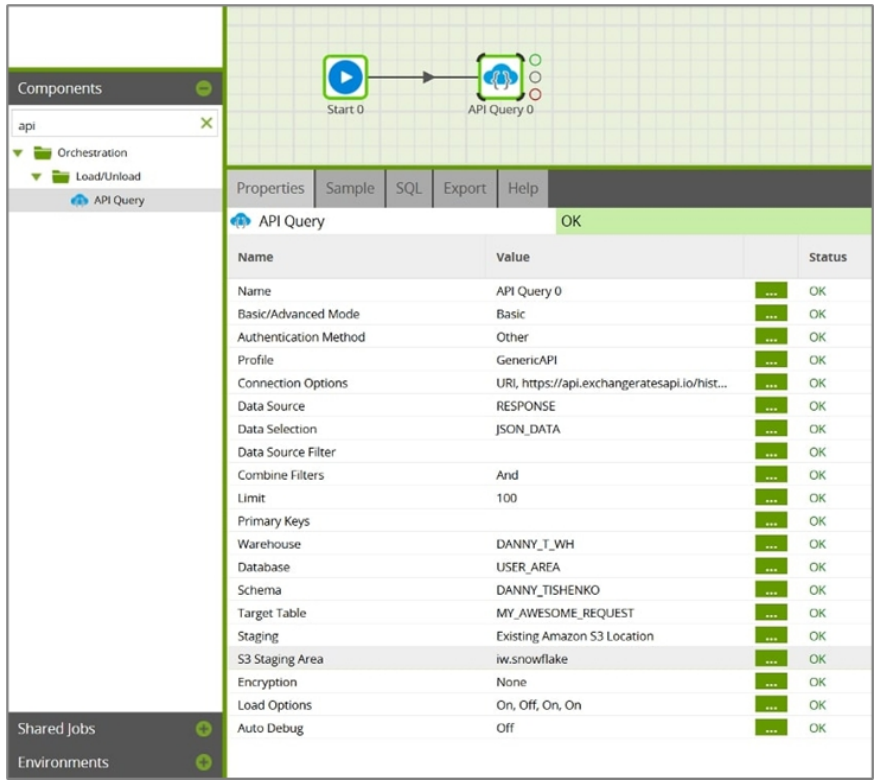
Matillion利用のメリット

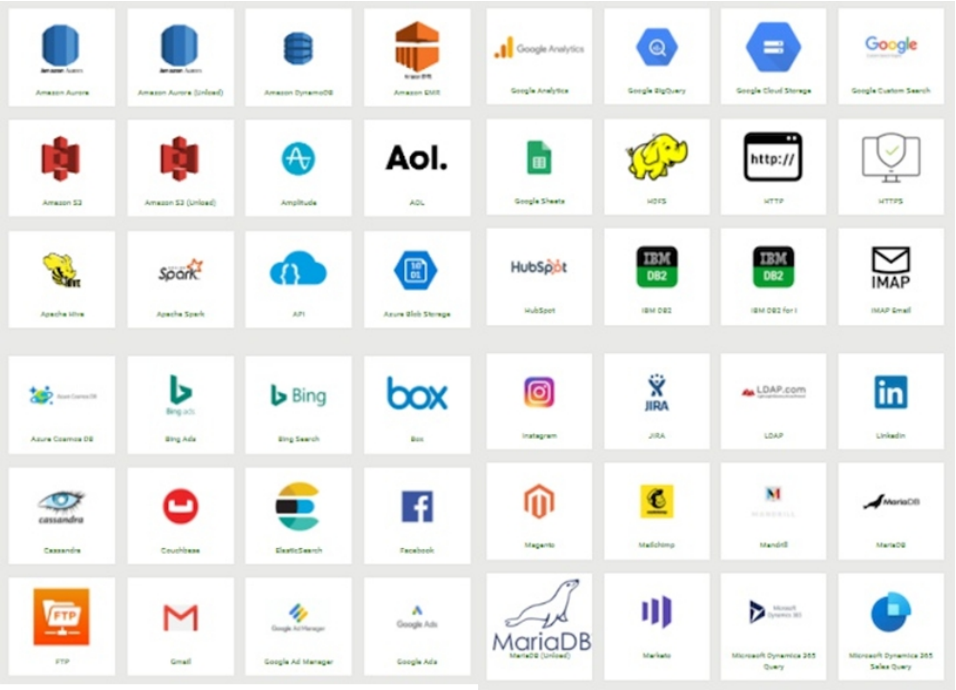
- データ分析基盤へのデータ収集機能の開発工数削減
  - 社内SE（外注管理だけになってしまっている人材）のフル活用
- データパイプラインの可視化による、引き継ぎとメンテナンス性の向上
  - SQL職人以外をデータエンジニアに活用
- ジョブフローのスケジューリング制御

©GRI Inc.

収集機能の開発工数削減（1/2）

- CDPを構築したり、全社的なDXを推進するためには、多様なデータを一元管理する必要がある
- 各システムからデータを分析基盤へ収集する際、旧来の開発だと多額なコストと時間が生じていた
- Matillionのデータ収集用のコンポーネントを利用することで、最短５分ほどでデータ収集の設定が完了できる
- このコンポーネントは約100種類あり、様々なシステムと統合を容易にする

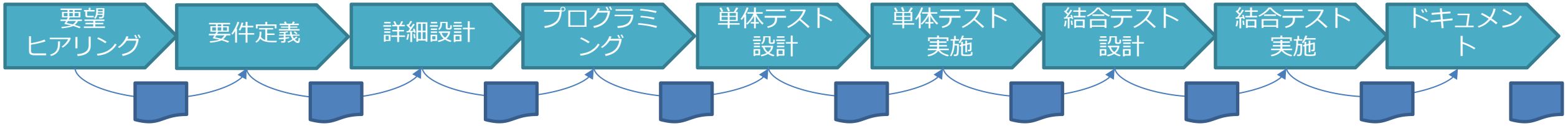




収集機能の開発工数削減（2/2）

- 従来型のデータ収集インターフェース開発では、多くの工程を必要としていた（工程ごとに職位が異なり伝言ゲームをしていた）
- Matillionのデータ収集用のコンポーネントを利用することで、これらの工程の大半をスキップできた（各工程間でタスクを受け渡すための各種書類も不要）
- また、Matillionの設定画面をドキュメントとして利用することができるので、不要なドキュメントを作る必要もなくなった

従来のデータ収集I/F開発

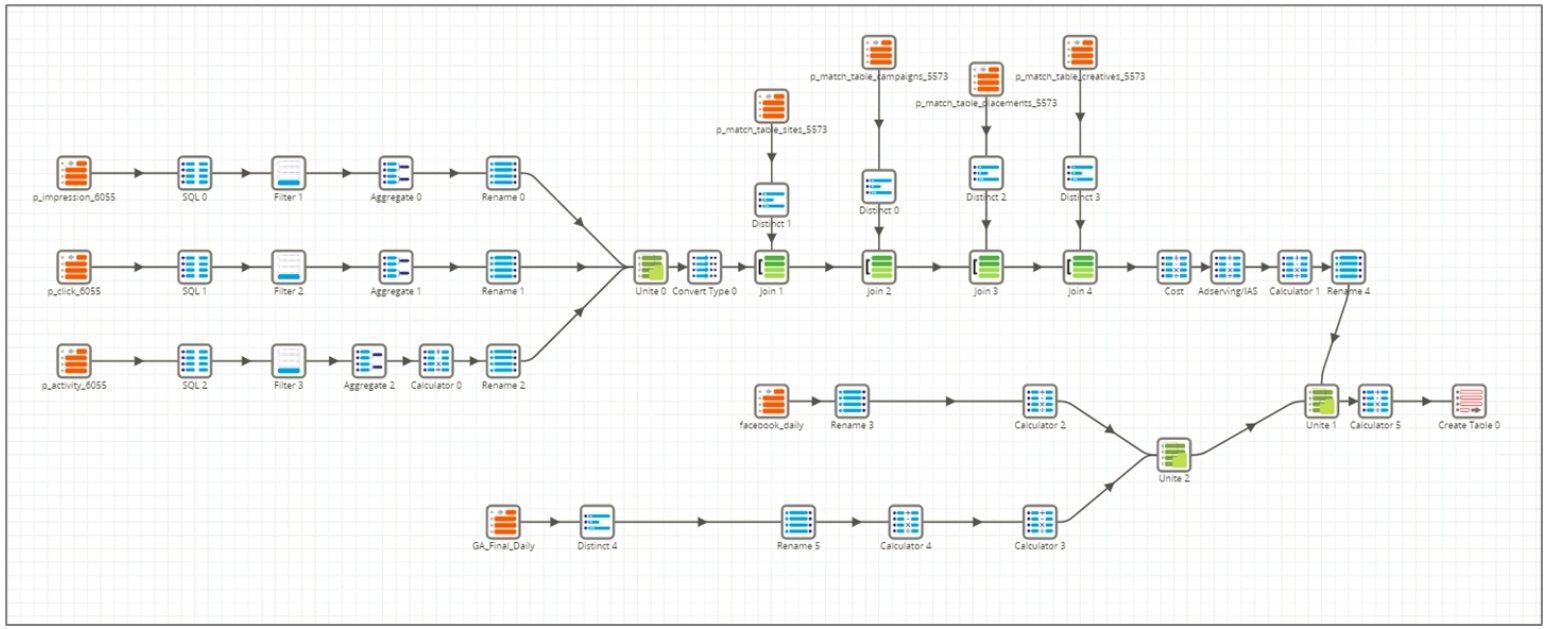


Matillionの場合

©GRI Inc.

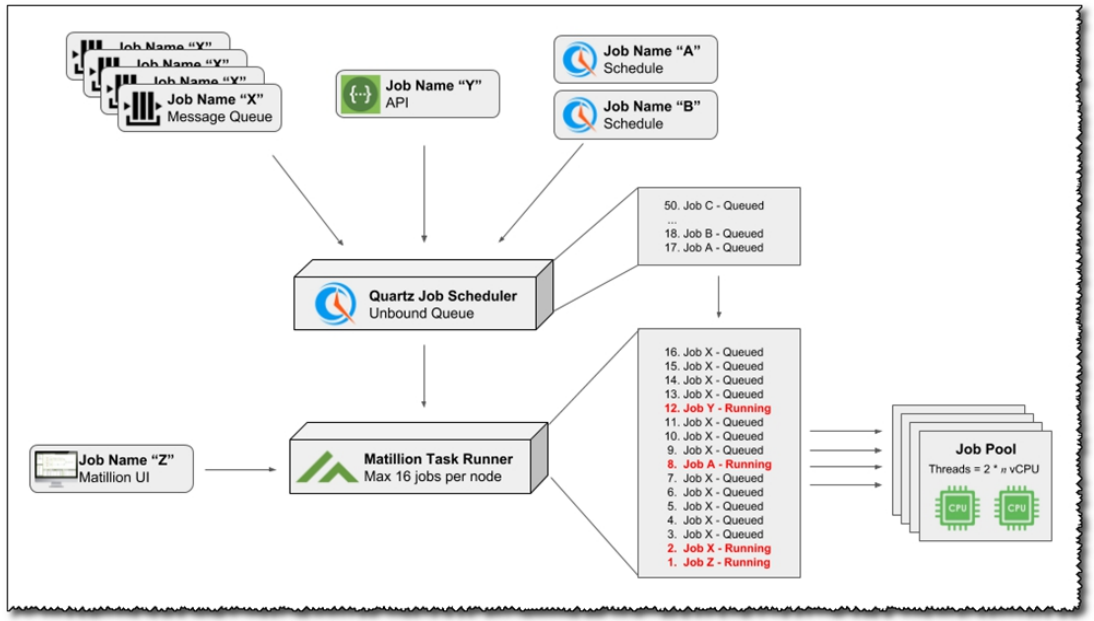
データパイプラインの可視化

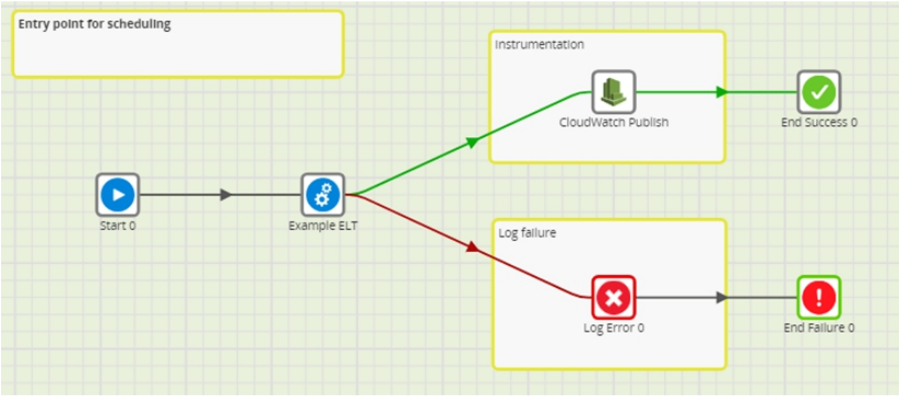
- データ分析基盤では、収集したデータを分析しやすい形に整えておく必要があり、複雑なデータ整形処理を記述してデータパイプラインの構築が必須
- Matillionのデータ整形コンポーネントを利用することで、データパイプラインを可視化することができる
- 専門家以外でも処理を理解でき、パイプライン構築と運用工数を大幅削減できる
- 引き継ぎやメンテナンス性が得られる



ジョブフローのスケジューリング制御

- データ整形の処理順序には依存関係があるため、複雑なジョブフローを制御する必要がある
- Matillionのジョブ管理機能を使い、ジョブの処理結果条件に応じたジョブフローの制御を実現できる





MatillionでDXを高速に推進できる理由

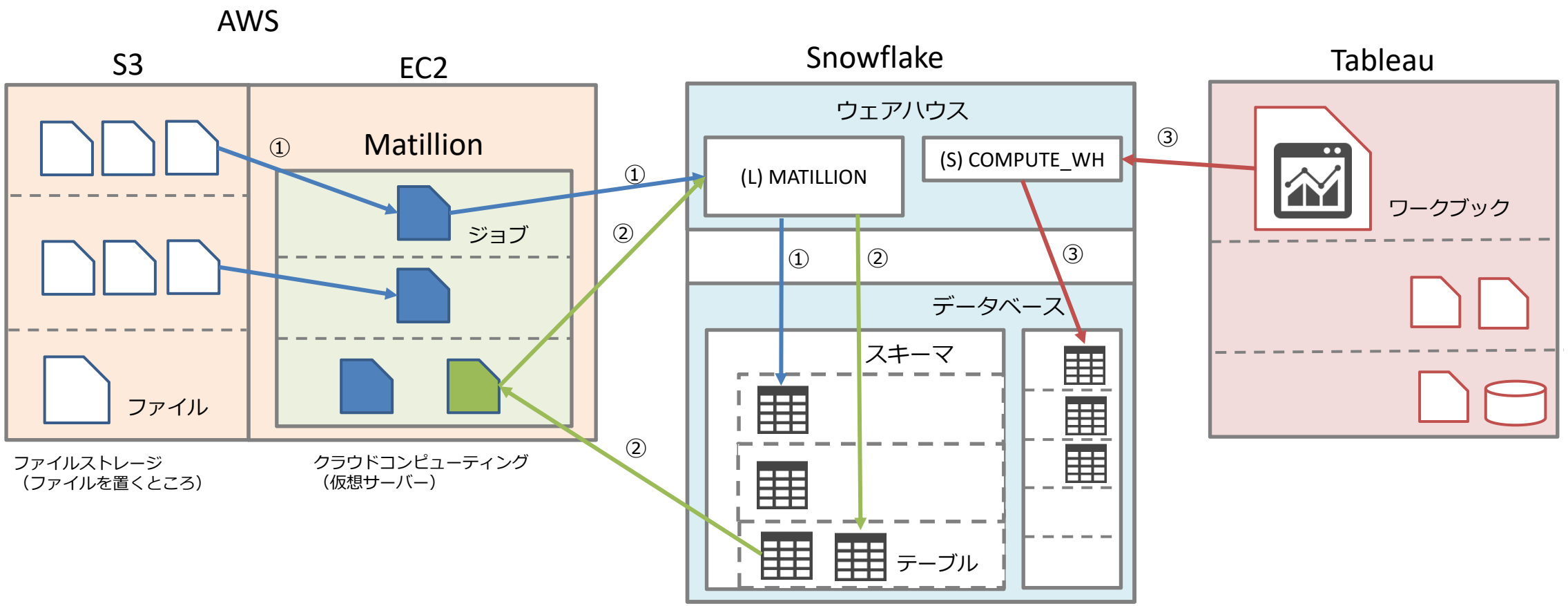
- データ収集の開発工数／時間を圧縮できる（多様なデータを一元管理できる）
  - 読売新聞に事例 https://forecastflow.jp/case/yomiuri/
- 社内SEや非エンジニアをデータエンジニアとして転向して大活躍できる
- DWHやデータマートを高速に作成できる
- 大規模データを高速に処理できる

©GRI Inc.

2. Matillionの機能の具体例

©GRI Inc.

Matillionと各サービスの関係



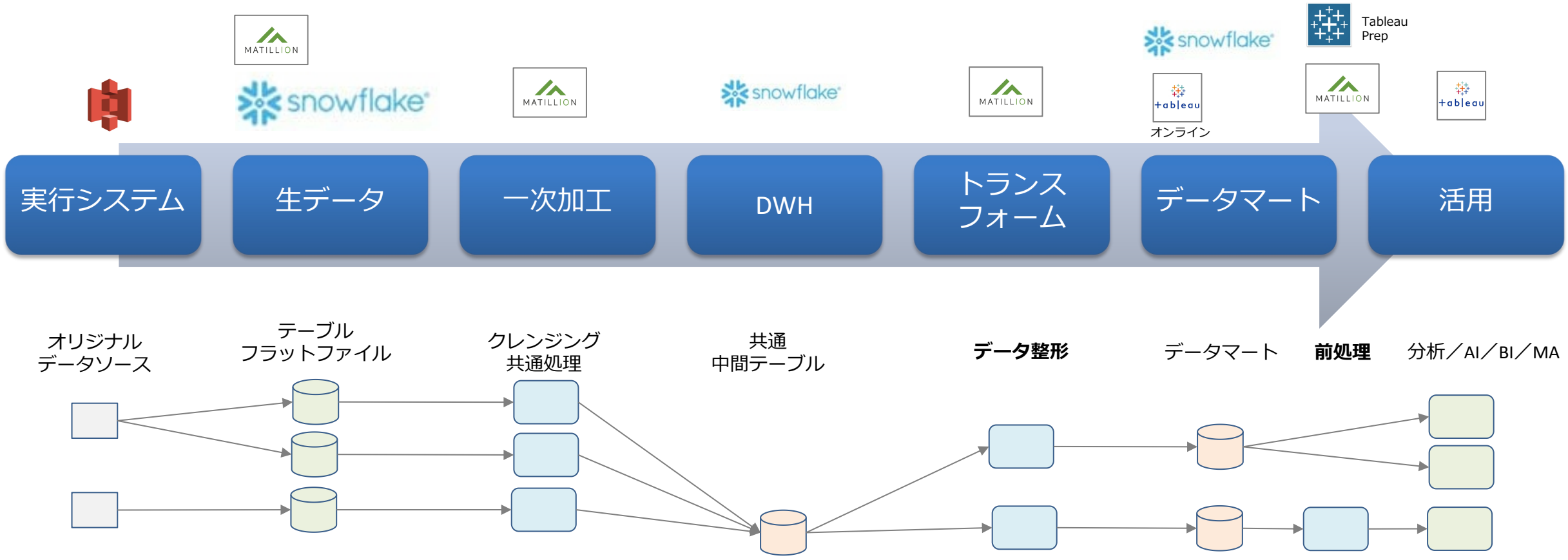
クラウドデータプラットフォーム
・データベース
テーブルを格納するところ
・ウェアハウス
Matillionから来た処理を実行するところ（①②）
Tableauから来た処理を実行するところ（③）

①S3にあるファイルを読み込んで
テーブルを作る処理をSnowfalakeに送る

②Snowfalakeのテーブルを整形し
新しいテーブルを作る処理を送る

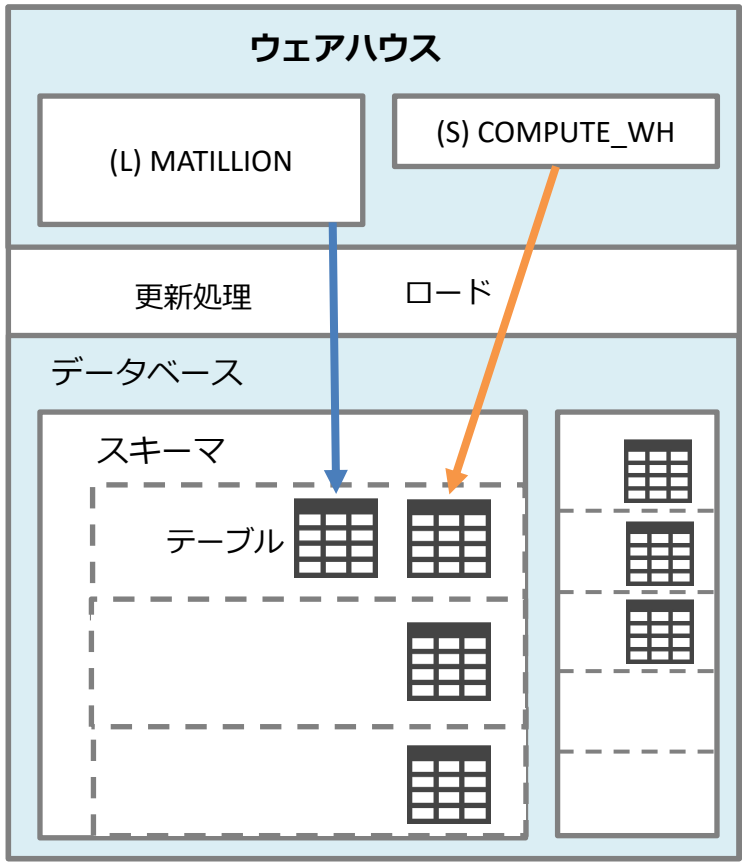
③Snowfalakeのテーブルに接続して
ダッシュボードを作る

データ活用の流れ


MATILLION
snowflake
MATILLION
snowflake
MATILLION
snowflake
tableau
オンライン
Tableau
Prep
MATILLION
tableau
実行システム
生データ
一次加工
DWH
トランス
フォーム
データマート
活用
オリジナル
データソース
テーブル
フラットファイル
クレンジング
共通処理
共通
中間テーブル
データ整形
データマート
前処理
分析／AI／BI／MA

Snowflakeについて

- Snowflakeの構造
  - ①データベース > スキーマ > テーブルという構造になっている
  - ②テーブルやデータベースに対する処理は**ウェアハウス**が行う



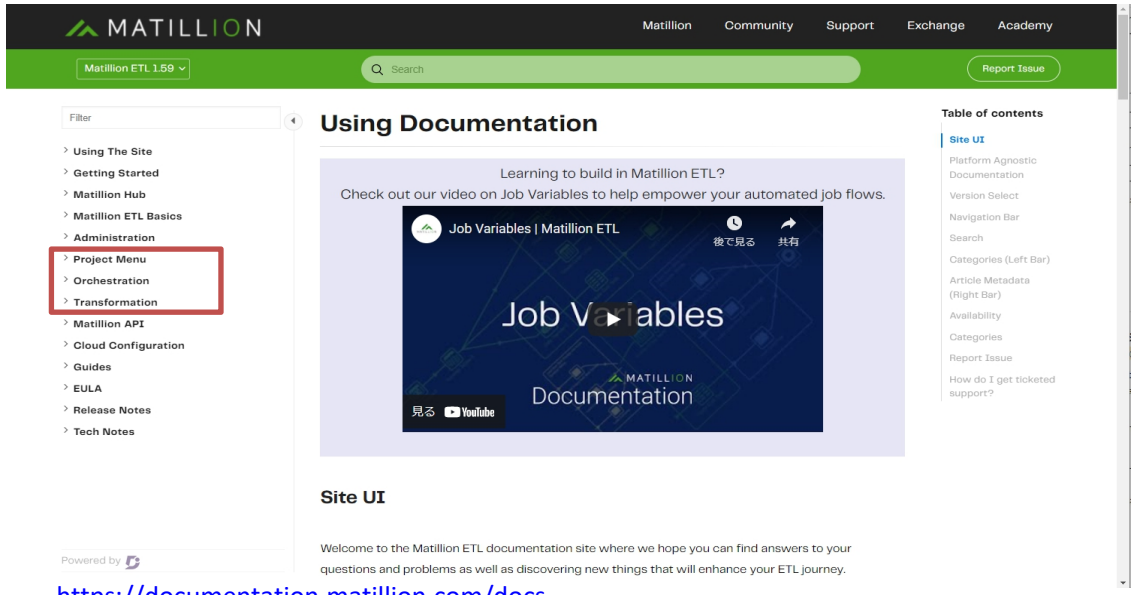
Matillionでできること

- Snowflakeに外部データをロードさせる
- Snowflakeテーブルの加工をする
  - テーブルをMatillionにロードしてきてデータを加工、新しいテーブルを作る
- ロード、加工処理のジョブ化
  - Matillionはジョブの中の処理をSnowflakeに渡して処理させる
    - データパイプラインの可視化
  - ジョブの種類
    - 種類によって使える機能が異なる
    - 青（Orchestration job）と緑（Transformation job）がある
      - » データをSnowfalakeにロードする、ジョブに対して操作を行う　→　青（Orchestration job）
      - » ロードしたデータを加工する　→　緑（Transformation job）
- ジョブの定期実行

©GRI Inc.

Transformation jobについて

- 整形処理はTransformation jobで作成する
  - 操作感はTableau Prepと同じ
    - コンポーネントと呼ばれる処理パーツを繋げていき、ジョブを作る
      - 裏ではジョブをSQLに変換してSnowflakeに送っている
  - SQL以外は書けない
  - コンポーネントの設定項目や記法は接続データベースに依存する
    - Matillion側の制限はほぼない、データベースでできることはMatillionで出来る
  - 変数を使った処理ができる
    - 関数などであらかじめ設定しておいた変数を使うことができる
    - 変数にも種類があり、後で紹介



https://documentation.matillion.com/docs

Orchestration jobについて

- 整形処理以外の操作もOrchestration jobで作成できる
  - Transformation jobと操作感は同じ
    - コンポーネントを繋げることでジョブを作成する
    - 緑と青で使えるコンポーネントが異なる
  - SQL以外のスクリプトが使用できる
    - Python,bash等が使用可能
  - 整形処理以外の処理を作る

©GRI Inc.

Orchestration jobの機能

■ コンポーネントの例

- **データベースにデータをロードする**
  - S3→Snowflakeへのロード処理ができる
  - 今回はGoogleCloud→BigQueryのロード処理を紹介
- スクリプトを書く
- ジョブ同士を繋げる
- 条件分岐、繰り返しをする

©GRI Inc.

Matillion整形機能

■ コンポーネントの例

- コンポーネントをジョブの中に配置、必要な項目を設定する
  - テーブルの読み込み（Table Input）
  - 整形系のコンポーネント
    - 計算、フィルタ、集計、Distinct、ランク、列名変更、型変更、Window関数
  - JOIN系のコンポーネント
    - Join: 1 つのコンポーネントで複数結合できる
    - Union
  - 出力（Rewrite、TableOutput）

©GRI Inc.

その他の機能について

■ コンポーネント以外の機能紹介

- 変数
  - 変数：ジョブ単位で設定する
  - 環境変数：プロジェクト単位で設定する
  - グリッド変数：テーブルからとってくる
- ジョブのスケジューリング機能
- 環境設定
  - デフォルトで使用する設定を作成できる
    - ウェアハウス、データベース、スキーマ、接続するアカウントを決める
  - Snowflakeだと、ウェアハウスの使い分けができる

©GRI Inc.

3. Matillionと自動機械学習ツールを連携させる応用例

©GRI Inc.

自動機械学習ツールとの連携

- データ基盤を構築したら次に取り組むべきなのはその活用
  - 自動機械学習ツールと連携し施策対象顧客を選定　**今回説明**
  - KPIなどの重要指標を日々モニタリング
  - DM送信顧客の選定
  - 顧客のWeb行動分析によるサービスのUI/UX改善
  - etc.

- ケーススタディを通じて、連携の流れを説明
  - ケースには第25回（2021/12/22）すいすい会で紹介した「架空の通販会社の休眠顧客の掘り起こし」を使用
  - 参考：https://gri.jp/news/12924（すいすい会のページ、動画＋資料）
- 自動機械学習ツールには弊社開発のForecastFlowを利用
- またGoogle Cloud Platformでの挙動を紹介

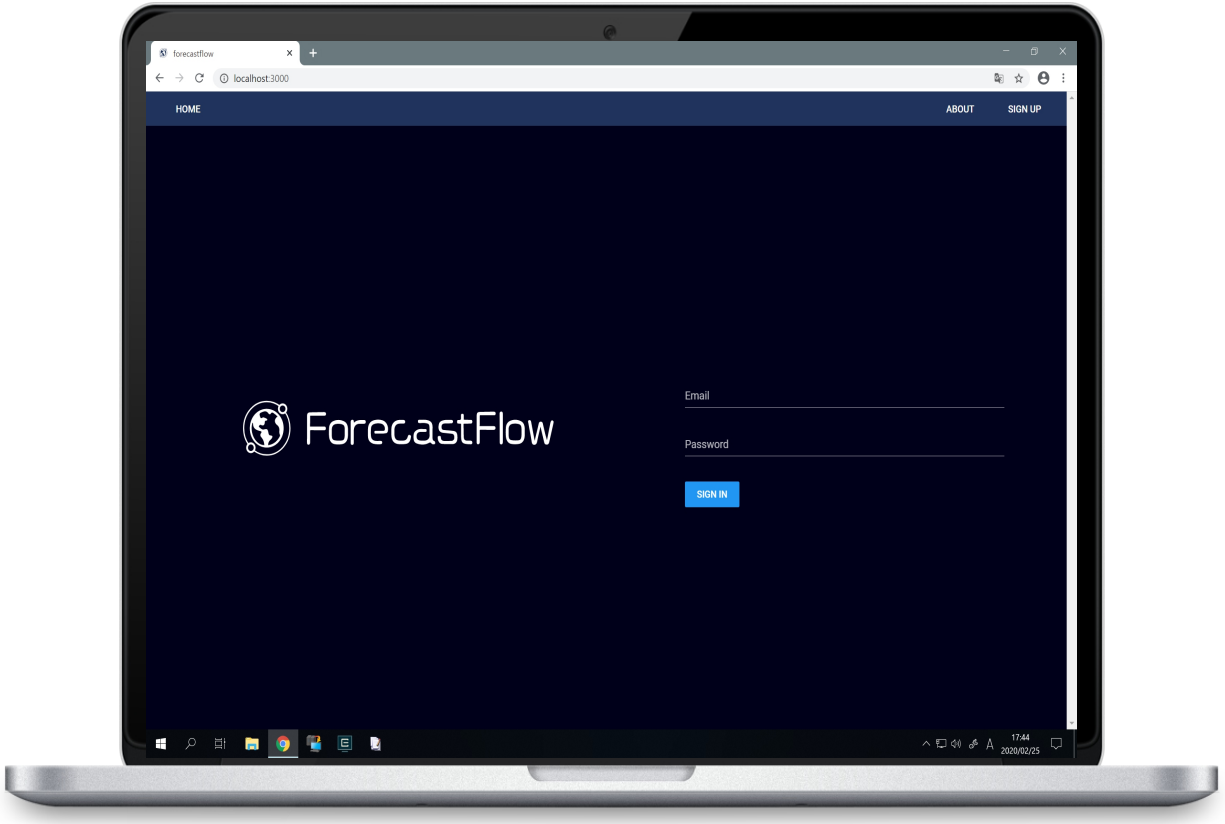
©GRI Inc.

ForecastFlowとは

第11回すいすい会

入力データに応じて最適な機械学習モデルを自動的に構築する
AutoML（自動機械学習） サービス

- 予測することが得意
- 誰でも予測が「かんたん」にできる



過去のすいすい会でたびたび紹介しているので、
詳細はそちらをご覧ください

機械学習を利用した架空の通販会社の休眠顧客の掘り起こし

第25回すいすい会

■単品リピート通販のビジネスを想定します。
■消費者は、この会社の商品をリピートで購入する想定ですが、途中で解約する方もいます。
■この会社では、休眠顧客を復帰させることに様々な施策を打っていますが、下記の課題があります
– 休眠顧客全員ではなく、対象を絞って費用対効果を改善したい
– どのような施策をいつ打つべきかを知りたい
■そのため、機械学習の効果的な利用方法を検討しています
■復帰する顧客と復帰しない顧客の違いを下記のような特徴量と呼ばれる説明変数だけで判別する法則を機械学習に探させます

架空の通販会社の休眠顧客の掘り起こし【予測フェーズ】

第25回すいすい会

■最新データを予測モデルへ投入し、顧客ごとの復帰予測スコアを算出する
■施策担当者は予測スコアに応じて施策対象の選定を行う

本デモのゴール

最新データ

予測スコアに応じた
施策対象の選定を行う

自動機械学習ツールの役割

訓練フェーズ

すでに結果がわかっている**過去の顧客の属性・行動データ**から、復帰しそうな顧客のパターンを見つけ出す

やっていることは膨大なif-then

予測フェーズ

現在休眠中の顧客に対して**最新の顧客データ**を投入し、復帰の可能性をスコアとして計算

それぞれのフェーズごとに連携イメージをデモ

©GRI Inc.

自動機械学習ツールの役割

訓練フェーズ

すでに結果がわかっている**過去の顧客の属性・行動データ**から、復帰しそうな顧客のパターンを見つけ出す

やっていることは膨大なif-then

予測フェーズ

現在休眠中の顧客に対して**最新の顧客データ**を投入し、復帰の可能性をスコアとして計算

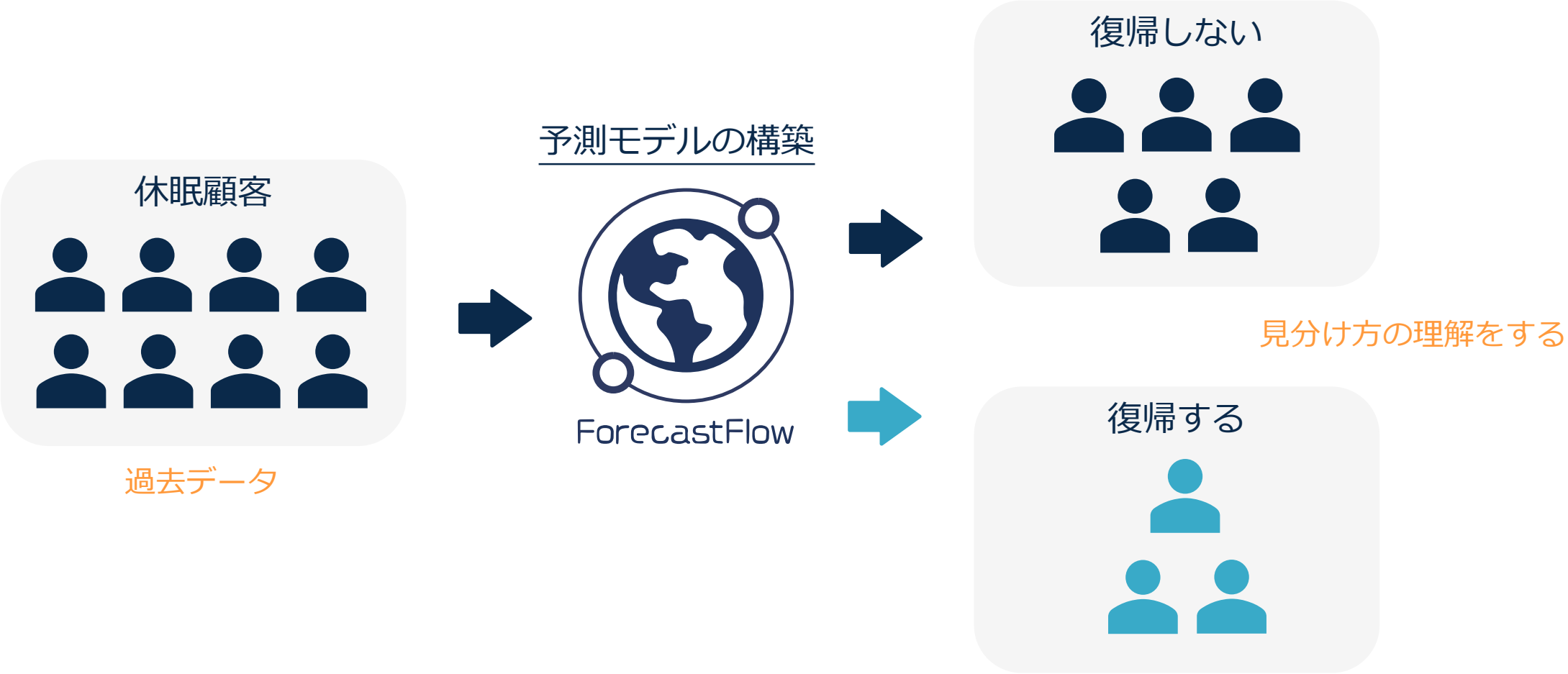
©GRI Inc.

架空の通販会社の休眠顧客の掘り起こし【訓練フェーズ】

第25回すいすい会

■過去実績データの復帰したかどうかの状態を隠して、機械学習の訓練により復帰する顧客と復帰しない顧客を見分ける法則を学習（訓練による機械学習モデルの構築）

■施策担当者は見分ける法則を理解する



訓練用データの準備

例） https://drive.google.com/drive/folders
gX3ZA25UQqBspZin
復帰予測_モデル構築用データ（1000行）.

第25回すいすい会

■今回使用するデータセットはこちらです

ID　ターゲット　特徴量

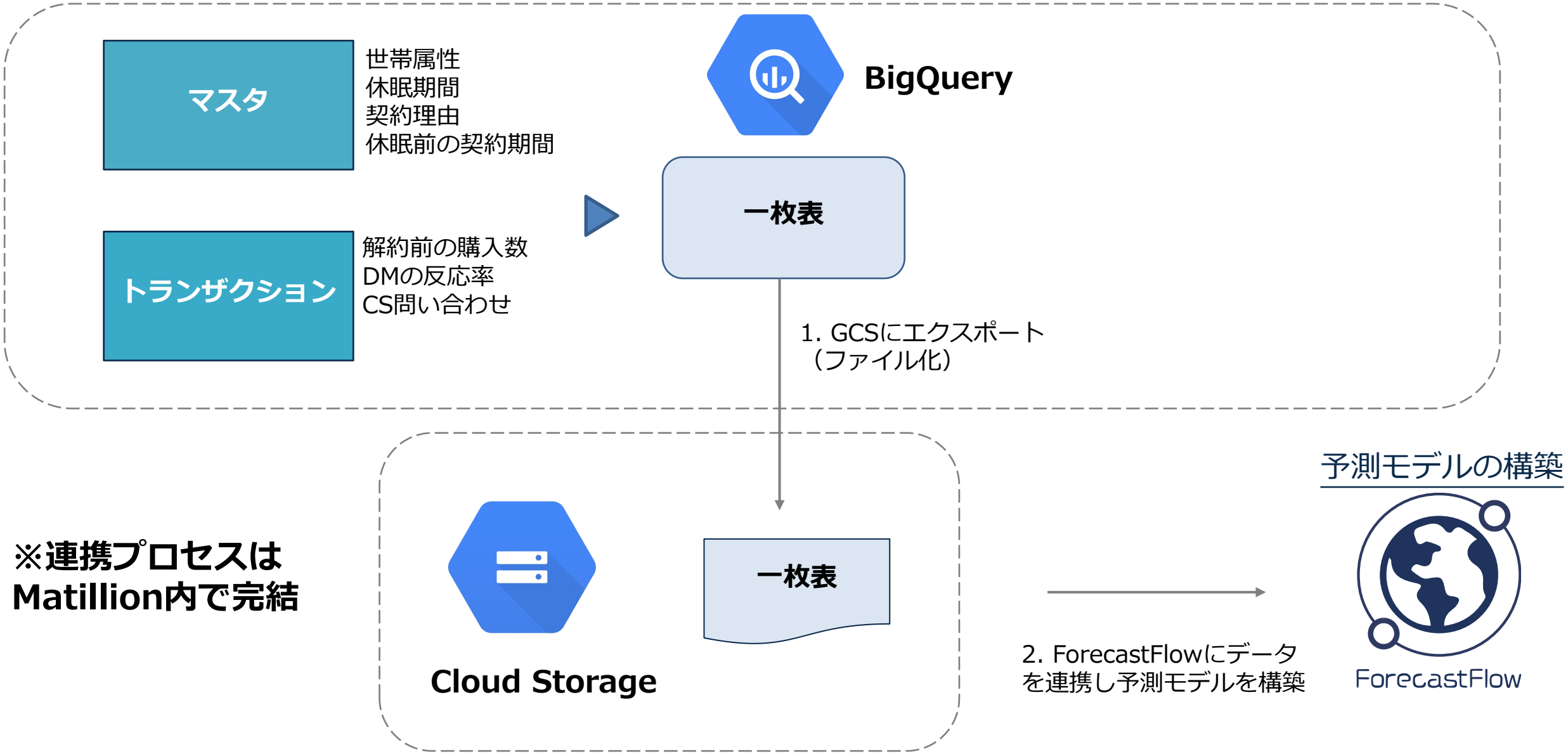
ID	復帰した？	都道府県	年齢	休眠期間(月)	休眠前契約期間(月)	休眠前累計購入数量	休眠前累計売上(円)	自社の別商品購入経験	販促メール種別
CUS4153948861	False.	茨城県	29	3	13	17	31880	yes	B
CUS4153949220	False.	愛知県	34	10	13	20	32880	yes	B
CUS4153949825	False.	佐賀県	26	3	14	15	36880	no	B
CUS4153949934	False.	岐阜県	31	5	13	18	33860	yes	B
CUS4153951380	False.	宮崎県	37	0	15	22	39380	yes	B
CUS4153951718	False.	福井県	34	2	18	20	45000	yes	B
CUS4153953852	False.	富山県	43	4	14	25	35720	no	B
CUS4153955015	False.	愛知県	31	4	15	18	37520	no	B
CUS4153955783	False.	滋賀県	25	4	9	15	21980	yes	B
CUS4153956030	False.	三重県	30	4	11	18	28540	yes	B
CUS4153956149	False.	神奈川県	24	5	9	14	24100	yes	B

特徴量

販促メール開封率	休眠開始日からのLINEプレゼント送付日数	販促DM種別	販促DMの画像サイズ	販促DM開封率(%)	販促DM累計送付回数	アンケート累計実施回数	コールセンター累計通話回数	コールセンター累計通話時間(秒)	サイト累計訪問回数	サイト累計滞在時間(秒)
0.0215	0	A	0	1.16	15	4	0	100	5	119.5
0.035	0	A	0	1.89	8	6	2	117	11	254.4
0.063	23	A	2	3.4	16	5	0	91	7	161.3
0.09	0	A	1	4.86	7	6	0	113	5	120.2
0	0	A	3	0	10	6	3	128	8	185.3
0.046	0	A	2	2.48	13	5	1	101	5	111.4
0.0595	33	A	3	3.21	9	6	0	98	4	95
0.0645	31	A	2	3.48	10	6	1	121	11	236.3
0.026	0	A	1	1.4	8	7	0	109	8	186.4
0.043	0	A	3	2.32	8	5	0	65	8	178.9

©GRI Inc.

ForecastFlow連携のアーキテクチャ



※連携プロセスは Matillion内で完結

デモ

©GRI Inc.

自動機械学習ツールの役割

訓練フェーズ

すでに結果がわかっている**過去の顧客の属性・行動データ**から、復帰しそうな顧客のパターンを見つけ出す

やっていることは膨大なif-then

予測フェーズ

現在休眠中の顧客に対して**最新の顧客データ**を投入し、復帰の可能性をスコアとして計算

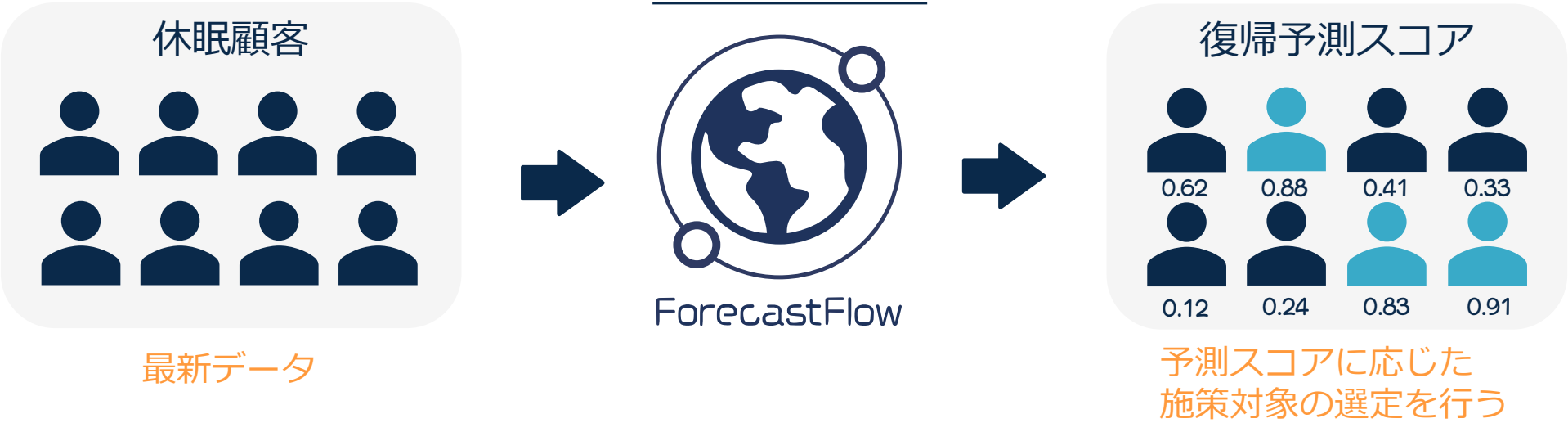
©GRI Inc.

架空の通販会社の休眠顧客の掘り起こし【予測フェーズ】

第25回すいすい会

■最新データを予測モデルへ投入し、顧客ごとの復帰予測スコアを算出する
■施策担当者は予測スコアに応じて施策対象の選定を行う

本デモのゴール



予測用データの準備

例） https://drive.google.com/drive/folders
gX3ZA25UQqBspZin
復帰予測_モデル構築用データ（1000行）.

第25回すいすい会

ID　**ターゲット列がない（まだわからない）**　特徴量

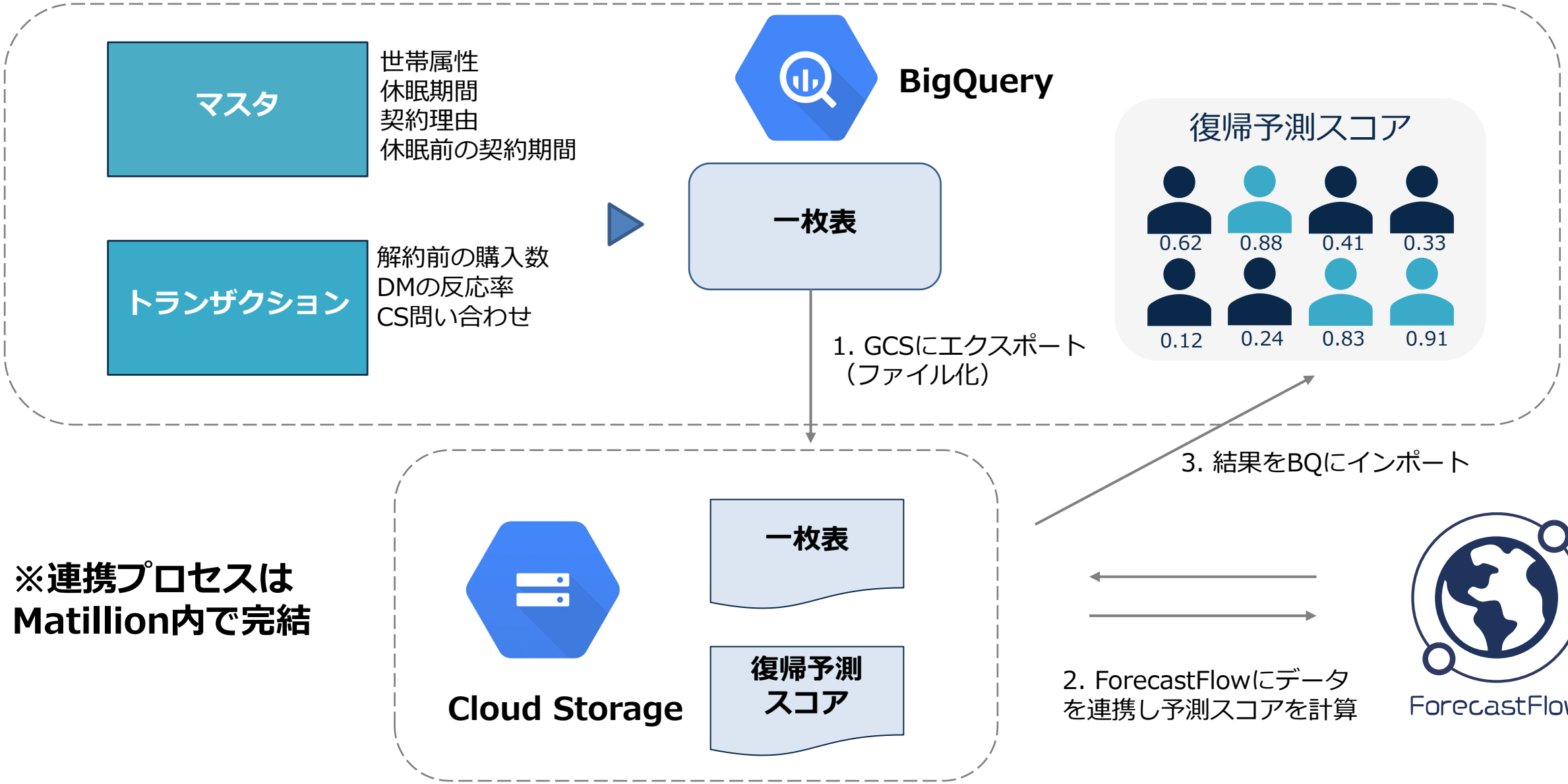
ID	[redacted]	都道府県	年齢	休眠期間(月)	休眠前契約期間(月)	休眠前累計購入数量	休眠前累計売上(円)	自社の別商品購入経験	販促メール種別
CUS4153948861		茨城県	29	3	13	17	31880	yes	B
CUS4153949220		愛知県	34	10	13	20	32880	yes	B
CUS4153949825		佐賀県	26	3	14	15	36880	no	B
CUS4153949934		岐阜県	31	5	13	18	33860	yes	B
CUS4153951380		宮崎県	37	0	15	22	39380	yes	B
CUS4153951718		福井県	34	2	18	20	45000	yes	B
CUS4153953852		富山県	43	4	14	25	35720	no	B
CUS4153955015		愛知県	31	4	15	18	37520	no	B
CUS4153955783		滋賀県	25	4	9	15	21980	yes	B
CUS4153956030		三重県	30	4	11	18	28540	yes	B
CUS4153956149		神奈川県	24	5	9	14	24100	yes	B

特徴量

販促メール開封率	休眠開始日からのLINEプレゼント送付日数	販促DM種別	販促DMの画像サイズ	販促DM開封率(%)	販促DM累計送付回数	アンケート累計実施回数	コールセンター累計通話回数	コールセンター累計通話時間(秒)	サイト累計訪問回数	サイト累計滞在時間(秒)
0.0215	0	A	0	1.16	15	4	0	100	5	119.5
0.035	0	A	0	1.89	8	6	2	117	11	254.4
0.063	23	A	2	3.4	16	5	0	91	7	161.3
0.09	0	A	1	4.86	7	6	0	113	5	120.2
0	0	A	3	0	10	6	3	128	8	185.3
0.046	0	A	2	2.48	13	5	1	101	5	111.4
0.0595	33	A	3	3.21	9	6	0	98	4	95
0.0645	31	A	2	3.48	10	6	1	121	11	236.3
0.026	0	A	1	1.4	8	7	0	109	8	186.4
0.043	0	A	3	2.32	8	5	0	65	8	178.9

©GRI Inc.

ForecastFlow連携のアーキテクチャ



**※連携プロセスは
Matillion内で完結**

デモ

©GRI Inc.

予測結果

顧客ID　　1-復帰予測スコア　　復帰予測スコア

id	FALSE	TRUE
CUS5103872799	0.88134452097833100	0.1186554790216690
CUS5104076391	0.95611305229587700	0.04388694770412350 0
CUS5103905470	0.94701266357120100	0.05298733642879910 0
CUS5103907863	0.9177454421822810	0.08225455781771940
CUS5104006454	0.97469856908159700	0.02530143091840340
CUS5104186455	0.20075862206145000	0.79924137793855
CUS5104029558	0.69858409673040100	0.30141590326959900
CUS5103985788	0.95451557920239400	0.04548442079760580
CUS5103872604	0.59895544489271000	0.40104455510729000
CUS5103938736	0.79625458987234900	0.2037454101276510 0
CUS5104111261	0.93153895282847700	0.06846104717152340 0
CUS5104205990	0.76955368438498900	0.2304463156150110 0
CUS5104152935	0.9129144472924810	0.08708555270751960 0

・本デモはここまで

・この後は復帰スコアごとにコミュニケーションの仕方を変えていくなどが考えられる

©GRI Inc.

Matillion x ForecastFlowが特に有用となるケース例

大規模データ

頻繁なデータ更新

明確な
予測ターゲット

- Webメディアやスマホアプリなどを運用しており、万単位あるいはそれ以上のDAUを抱えている
  →施策対象となるIDが多いケース

- IoTセンサーデータで各種デバイスを管理している、ECサイトを運用している
  →IDは多くないとしても、トランザクションデータが膨大に貯まる

- サブスク型ビジネスを展開しており、購買契約/解約申し込みや問い合わせなど日々顧客の行動データが入ってくる
  →顧客の態度変容をデータとして補足できる

©GRI Inc.

Matillion（BigQuery） x ForecastFlowのパワー目安（実績）

- 数億行のデータ（数百GB）のデータをハンドリングして数千万行x100列（数十GB）の一枚表を作成、ForecastFlowに連携し推論スコアを計算、結果をBigQuery格納、の一連のプロセスを1-2時間程度で完了
- さらに、その推論を20個以上並列で回すことが可能（契約体系によりますが）

©GRI Inc.

備考

・pythonスニペットは現時点では以下のブログを参照
https://forecastflow.jp/datascicene/report3/

©GRI Inc.

まとめ

©GRI Inc.

まとめ

- Matillionを使ってDXを高速に推進できる理由を説明
- デモを通してMatillionの利用イメージを共有
- 自動機械学習（ForecastFlow）との連携例を共有

©GRI Inc.

ForecastFlowの無料トライアル

- 90日間、無料でForecastFlowを試すことができます

- 下記より、お申込み
- https://forecastflow.jp/

©GRI Inc.

DX推進やMatillionの導入支援

- GRIでサポートすること可能ですので、お問い合わせください
  - https://gri.jp/contact/?id=9774

©GRI Inc.

次回のすいすい会

■2022年3月16日（水）17:00-19:00　（少し時間が変わり拡大版）

■【アミューズ様　ご登壇】エンタメDX先進事例のご紹介

- アミューズ社、Tableau社、Snowflake社、GRI社共催

- Matillion, Tableau, Snowflakeを活用したエンタメビジネスでのDXの推進事例をお伝えします
  - データ分析基盤へのテーブル投入数が数百
  - データ分析者やデータエンジニアをどのように育成したか

©GRI Inc.

残っている質問（今後、説明していきたい内容）

- ビジネスの実践例を知りたい
- どのくらい正解データの数があるべきか
- 未来の施策と機械学習の実行タイミングの関係性
- ダウンサンプリングとProbability Calibration
- ForecastFlowのアルゴリズムを知りたい
- 様々な自動機械学習ツールの比較
- 様々なETL/ELTツールの比較
- Feature Store、分析基盤、CDPの説明

©GRI Inc.