

データサイエンス

すいすい会



GRI

データサイエンス企業が明かす

データ基盤の構築に際して

エンタープライズ企業が
騙されないための情報共有

2024/3/6 (水) 18:30~19:30

オンラインセミナー



古幡 征史



横井 晃広

登壇者紹介



株式会社GRI
取締役 (Ph.D. in コンピュータサイエンス)

古幡 征史

Furuhata Masabumi

お問合せご連絡先: furuhata.masabumi@gri.jp

<https://twitter.com/FuruhataMsbm>

人工知能領域の研究だけでなく、幅広い業界での実務
経験を活かした企画開発を行う

【経歴】

KPMGコンサルティング、南カリフォルニア大学、
株式会社ドワンゴを経て、2016年より現職

領域

データ解析、AI、BI、データ基盤、DX戦略、研究開発、SCM、生産、調
達、物流、マーケティング、経営、人事

業種

エンタメ、メディア、小売、建設、物流、製造、食品、通信、エネ
ルギー、インフラ、出版、広告、軍事



株式会社GRI
データサイエンス事業部

横井 晃広

Yokoi Akihiro

お問合せご連絡先: yokoi.akihiro@gri.jp

データサイエンス領域の新しい情報を常に収集し、
独自の視点と合わせて開発分析を行う

【経歴】

Matillion等のツールを駆使したデータ基盤構築に従事
2021年より現職

キーワード

データ基盤、ETL、データベース、DWH、BI、データ分析、Matillion、
Snowflake、Tableau、SQL、Python、
ももクロ、サッカー、阪神タイガース、映画、MARVEL、畑、お笑い、
ダイビング、スノボ、麻雀、ポーカー、謎解き、ポケモンカード

すいすい会の紹介

データサイエンス、データ分析、データドリブン組織 etc.

データ利活用をがんばるデータの猛者たちのための雑談会です！

悩み

を共有する

知見

を共有する

理解

を高める



実践を重視するGRIの考え方を雑談しながらお届けします

資料：[GRIホームページ](#)

Slack：[ForecastFlowチャンネル](#)

過去のすいすい会の動画 (YouTube)



GRIチャンネル

@GRIinc · チャンネル登録者数 623人 · 76本の動画

データとデザインで事業イノベーションを実現していくカンパニー。株式会社GRIの公式...

gri.jp、他2件のリンク

チャンネル登録

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

チャンネル登録 / いいね! / 高評価
よろしくお願いします!

データサイエンスすいすい会 ▶ すべて再生

データサイエンスすいすい会のアーカイブです月に一度、水曜日に開催しています



今必要なデータ人材像とキャリアエースへの道を切り開く

GRIチャンネル
48 回視聴 · 7 時間前



うつりゆく環境に適応するインテリジェンス・システム

GRIチャンネル
97 回視聴 · 2 か月前



Pythonプログラマからプロダクトオーナーへの道

GRIチャンネル
120 回視聴 · 4 か月前



事業会社におけるデータ活用拡大と分析組織強化のポイント

GRIチャンネル
179 回視聴 · 9 か月前



【GRWM】Tableau KPIダッシュボード

GRIチャンネル
862 回視聴 · 10 か月前



Attention Please! 文脈をとらえて社会を動かすAI基盤

GRIチャンネル
266 回視聴 · 1 年前

データサイエンスもくもく会 ▶ すべて再生



AI・データ活用における法律と倫理の知るべきこと

GRIチャンネル
565 回視聴 · 1 年前



DS検定対策講座の講師が指導する、業界初!?「DS検定の...

GRIチャンネル
1923 回視聴 · 1 年前



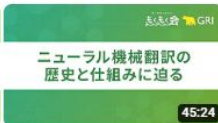
機械学習の活用には特徴量が肝心! 特徴量エンジニアリ...

GRIチャンネル
1506 回視聴 · 1 年前



ニューラル機械翻訳の歴史と仕組みに迫る / 一般物体認...

GRIチャンネル
401 回視聴 · 2 年前



ニューラル機械翻訳の歴史と仕組みに迫る

GRIチャンネル
508 回視聴 · 2 年前



畳み込みニューラルネットワーク(CNN)を解明

GRIチャンネル
2166 回視聴 · 2 年前

データサイエンス企業GRIの取り組みテーマ例



DX戦略支援/
データドリブン
経営



予測AI
(IoT予兆検知、サー
ビス解約予測)



BI
(見える化、KPI
ダッシュボード)



大規模
データ基盤構築



NLP
自然言語処理AI
(コンテキストチュア
ルターゲティング、
文脈理解、LLM)



時系列予測
(需要予測、
Marketing Mix
Model)



高度データ
人材育成



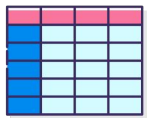
人流解析
(AIカメラ、GPS)



オープンデータ
データシェア
(クローリング、
地理空間統計、
気象)

GRIで扱うデータ

データ種別ごとに分析用に適したデータ構造が決まる



業務システム
データ

- ・ ID-POSシステム
- ・ ERPシステム
- ・ CRMシステム
- ・ SCMシステム
- ・ WMSシステム
- ・ HRシステム
- ・ ポイントシステム



時系列
ログデータ

- ・ Webログ
- ・ SNSデータ
- ・ センサーログ
(IoT)
- ・ TV視聴ログ
- ・ 生体信号
- ・ 金融経済データ



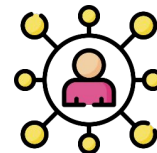
社会調査
市場調査

- ・ アンケート
- ・ 世論調査
- ・ カルテ
- ・ レビュー



非構造化データ

- ・ 自然言語
- ・ 音声
- ・ 画像
- ・ 動画



ネットワーク
データ

- ・ 交通データ
- ・ SNSデータ
- ・ 会話・文字
データ
- ・ 組織図
- ・ コミュニティ
- ・ 人間関係データ



オープンデータ

- ・ 政府統計
- ・ 天気
- ・ 人口統計
- ・ 地域
- ・ カレンダー
- ・ コーパス
- ・ スポーツ

多くのクライアントのデータ基盤を見てきた感想

- 導入支援ベンダーやIT管理企業に騙されている会社が多い
- その費用があったら、もっと分析できるのに
- 他の企業に任せるより自分たちで作った方が良い分析できる

データ基盤構築に際して、良い選定基準

データは21世紀の新しい石油

Information is the oil of the 21st century and analytics is the combustion engine
The world's most valuable resource is no longer oil, but data



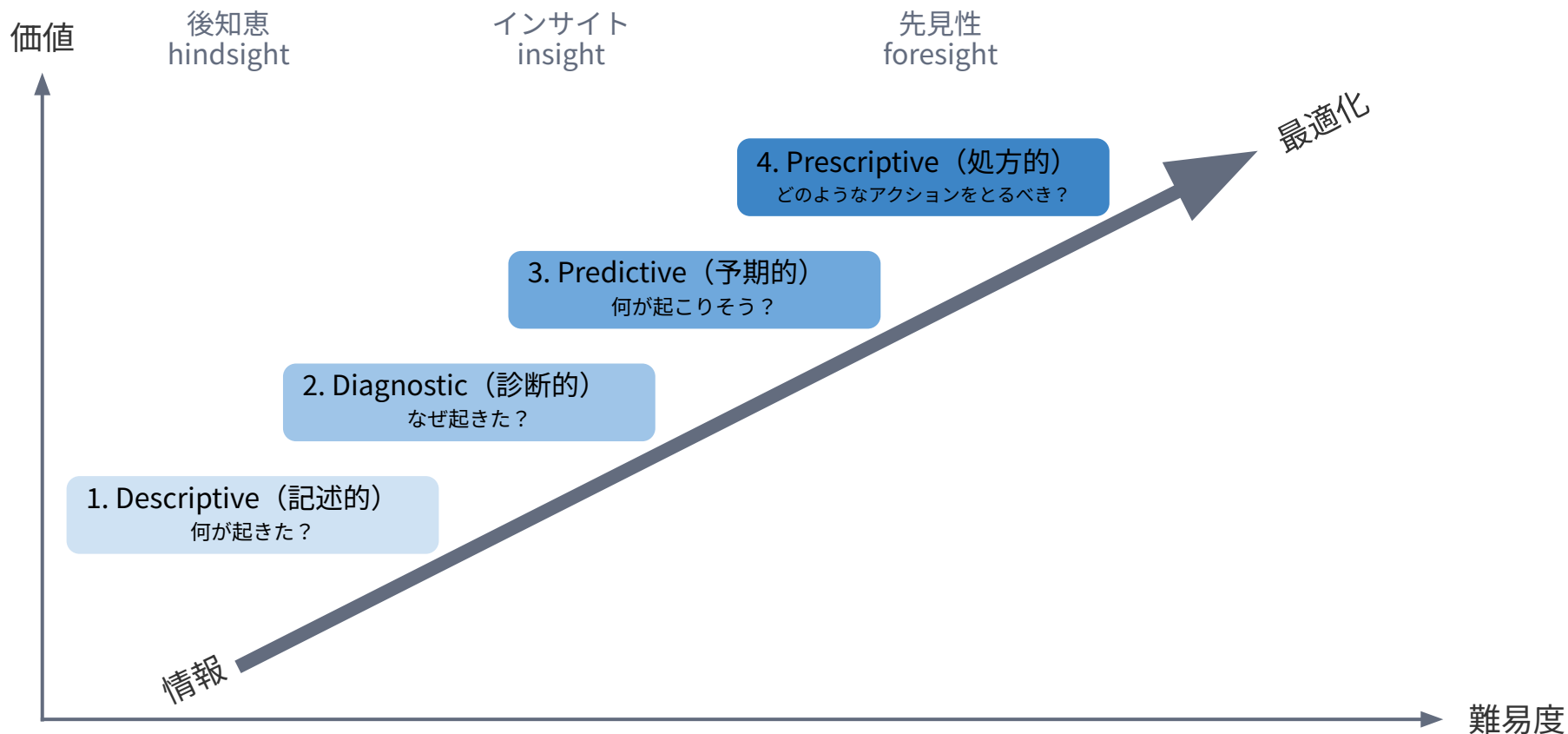
<https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data>

エンタープライズでのデータ活用の方向性

データドリブン経営によるオペレーショナルエクセレンスの達成

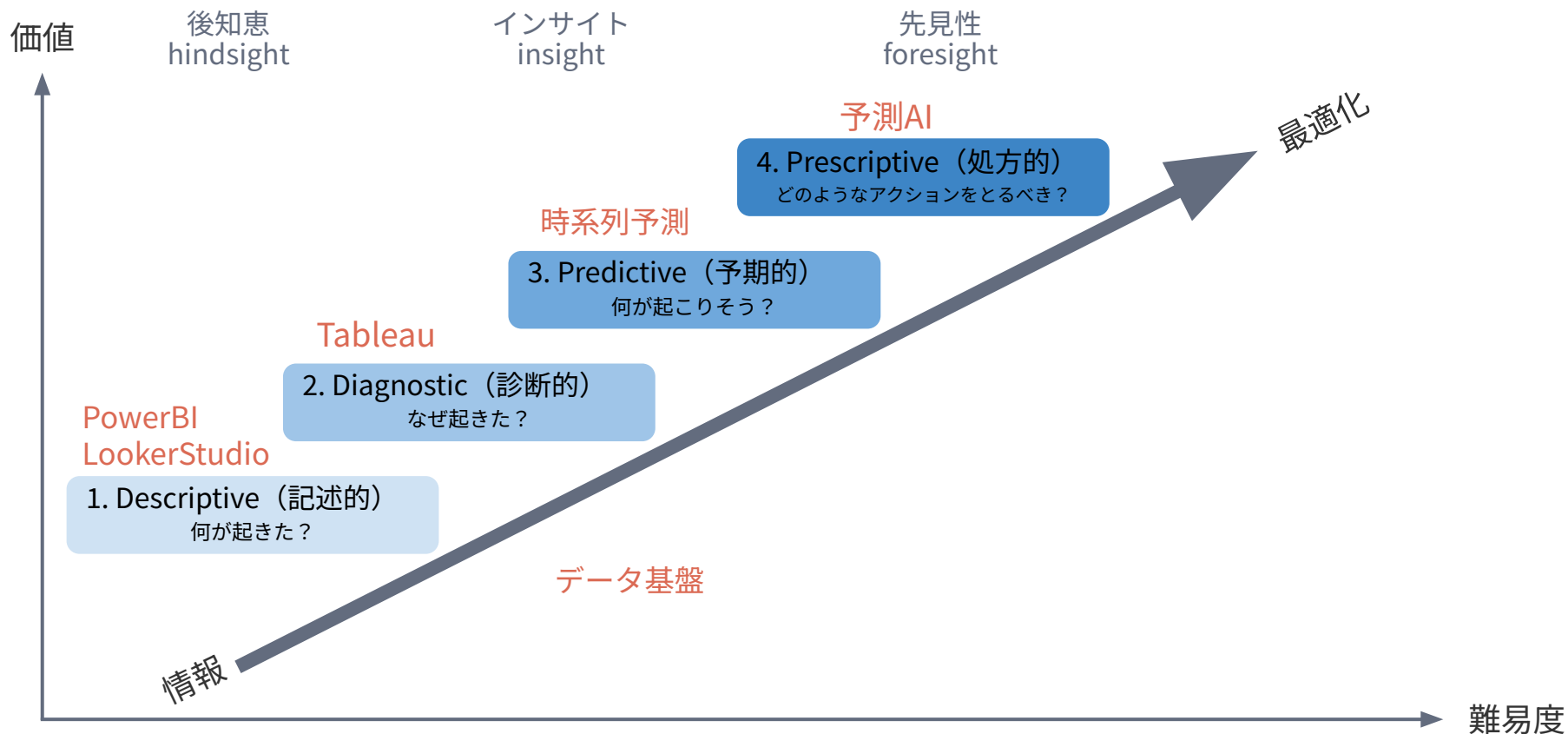
AI活用によるイノベーティブなサービスや商品開発

4段階のデータ活用レベル



<https://www.gartner.com/en>

4段階のデータ活用レベル



<https://www.gartner.com/en>

データサイエンティストの時間の使い方と典型的な方針

データサイエンティストの時間配分



- 現場で分析できる人材を増やす
- データ整備をしておき、データ分析の時間を確保できるようにする

エンタープライズ企業に立ちはだかるデータサイロという壁

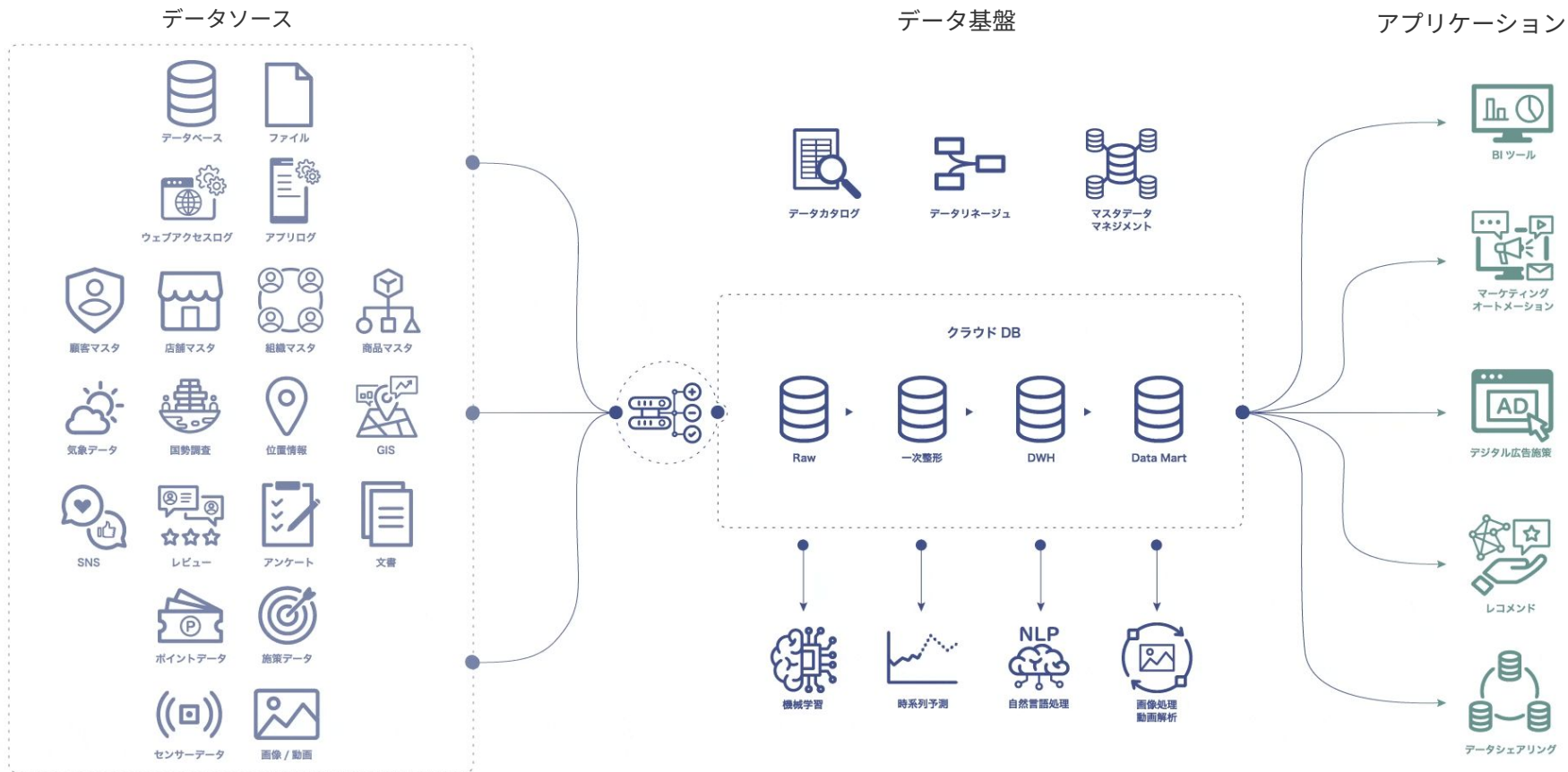
- データが分断されている状態だと、分析前にデータ収集やデータ整形が必要になり、前処理に90%の時間を使うことになる
- 分断されたデータはデータ基盤上に統合されていると、分析の成果を挙げやすい

モダンデータスタックという救世主

- モダンデータスタックとは
 - データ基盤を簡単に構築・管理するための最新のクラウドツールを組み合わせたもの
- モダンデータスタックの特徴
 - 簡単に試せる
 - 大規模データを高速処理
 - コンポーネント化で疎結合



エンタープライズのデータフローの例



モダンデータスタックのランドスケープ



Modern Data Stack

A complete landscape

ここにAIアプリ、自動機械学習、Advanced Analyticsが加わる



Storage

Object Storage

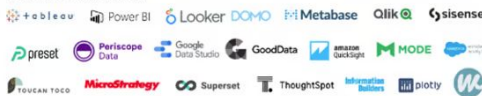


Data Warehouse



Visualize and Use

Data Visualization



SQL Editor



Notebook



Orchestrate

Workflow Scheduler



Transformation

Data Modeling



Syncing Data

ETL / ELT



Reverse ETL



Customer Data Platform



Discover and Trust

Data Monitoring



Data Catalogs



最初に決めるべきはクラウドDB



Snowflake



- 課金に応じてスピード調整
(Storage & Computing分離)
- 普通のSQL
- データクリーンルーム



- データ投入時に工夫が必要
- 支払方法
- 各クラウドとの統合性



BigQuery



- GAなどGoogle製品と相性が良い
- スケーラブル高速処理
- 基本安い(投入含めて)



- ローカルなSQL
- クエリ課金
- IoT Core突然撤退

ツール選定は配置できる人材の特徴で決める

InterWorks社と私の見解

	Basic Starting Points	Ad Hoc and User-Driven Analytics	Data Lead	Solutions Architect	Data Engineer UI Preference	Data Engineer Code-Based	Data Engineer	Furuhata
Extraction/Load	Fivetran	Matillion	Fivetran	Matillion	Matillion	Fivetran	Matillion	Matillion
Warehousing	Snowflake	Snowflake	Snowflake	Snowflake	Snowflake	Snowflake	Google Big Query + Omni	Snowflake BigQuery
Data Transformation	dbt	Matillion	dbt Cloud	Dataiku	Matillion	Snowflake	Matillion + Google Vertex AI	Matillion dbt
Analytics	Tableau	ThoughtSpot	Looker	Tableau	Tableau	Tableau	Looker	Tableau

<https://interworks.com/blog/2022/06/13/what-makes-good-analytics-the-perfect-toolset-for-a-data-intelligent-business/>

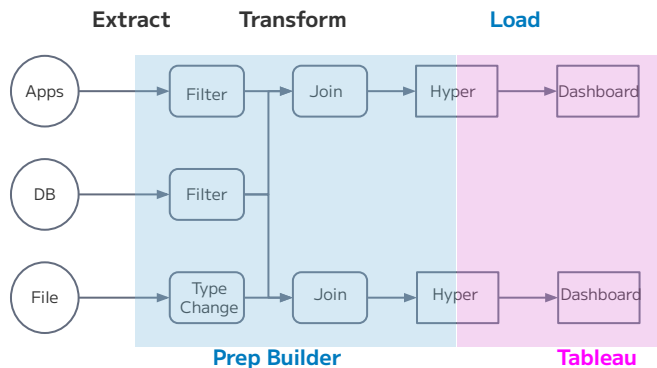
ELTの比較(Fivetran/Matillion)



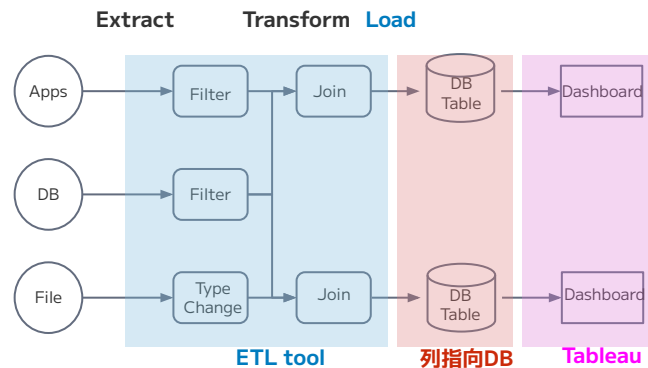
クラウドDBへのデータ投入	コンポーネントで ポチポチ(150+)	コンポーネントで ポチポチ(100+)
データ整形	SQLやdbtを 管理しやすい	コンポーネントで データ整形 (SQLやdbtも使える)
SQLエンジニア	複数以上いる	用意できない
課金体系	冪等性を担保するデータ入替が 不利な課金体系	稼働時間課金で予算組みやすい

ETLとELTの違い

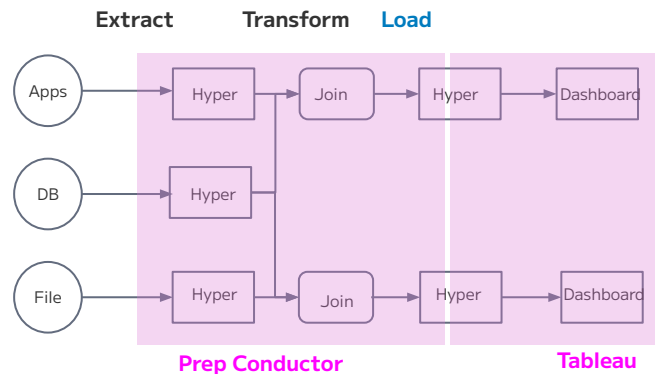
Tableau中心のデスクトップETL



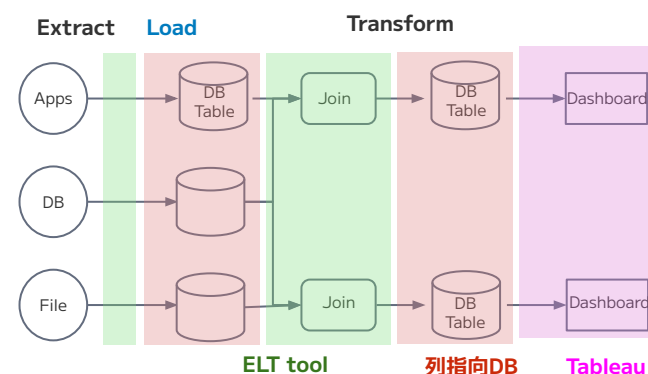
列指向DB中心のETL



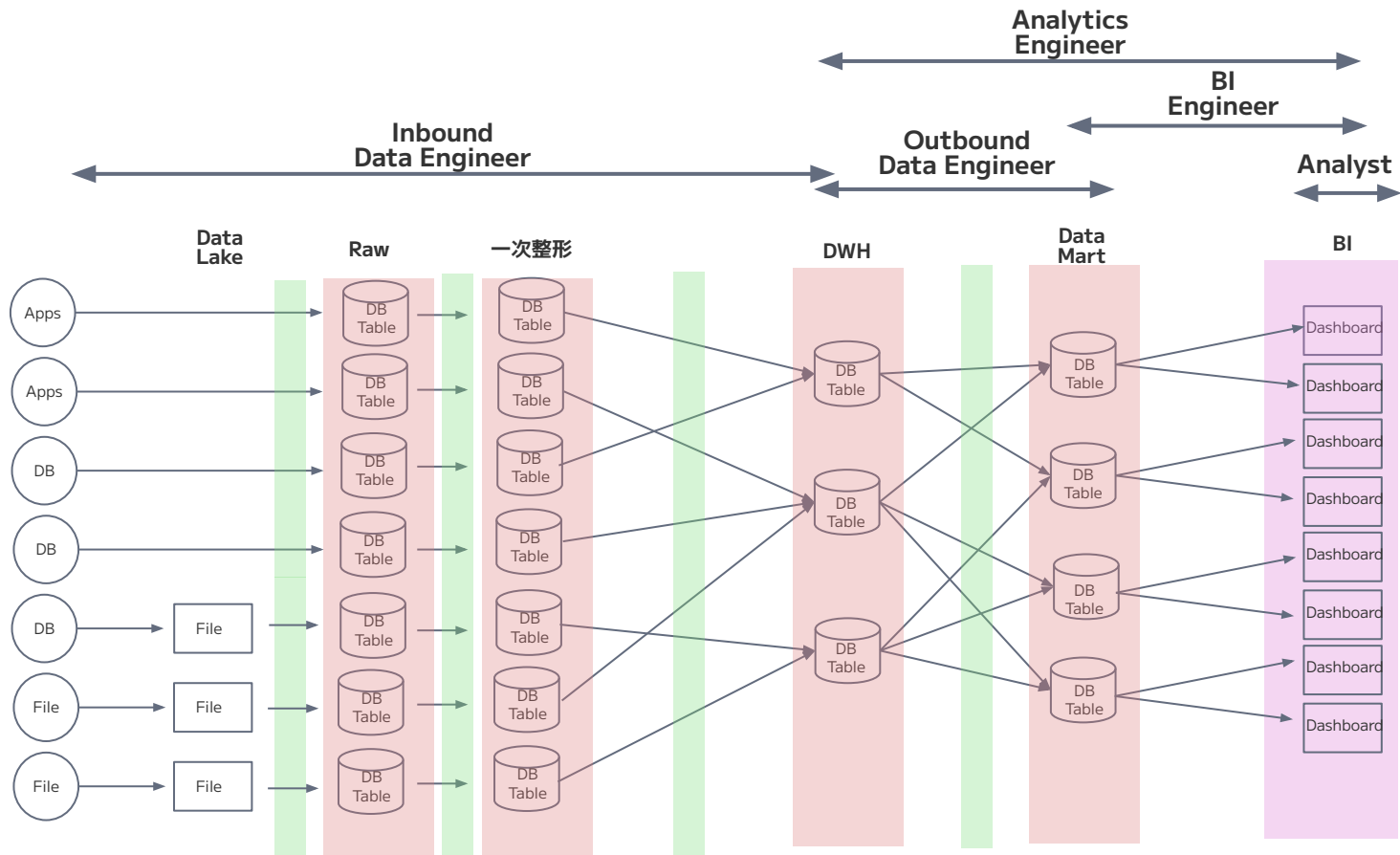
Tableau中心のサーバETL



列指向DB中心のELT



列指向DB中心のELTでのデータフロー、人材、ツール

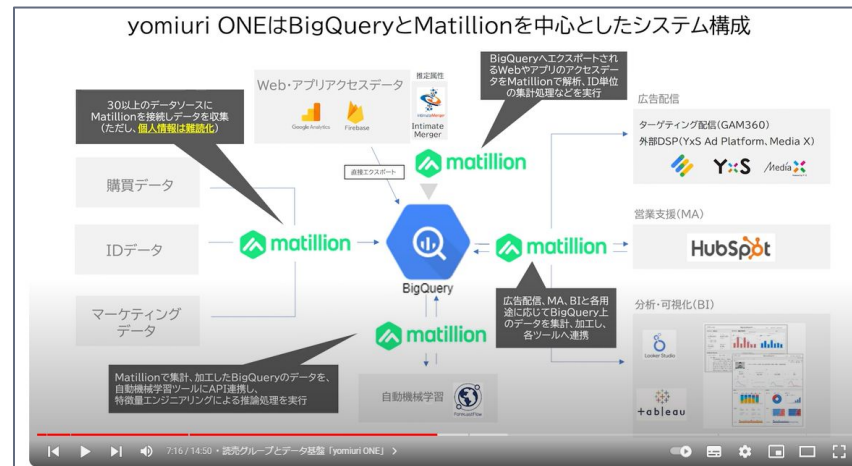
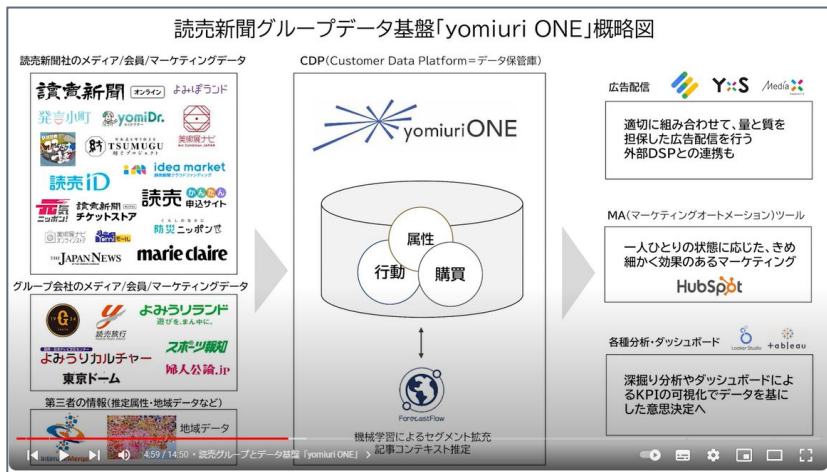


ダーティデータへの対応

- エンタープライズ企業にダーティデータがあるのが普通
- 初期分析でダーティデータの特徴を把握
 - データ定義書や数行のサンプルデータを信じない
→ データ基盤が完成してからデータを見て修正しない
 - 大規模データで確認
- 対応方針を決める
 - データ・クレンジング → データ整形で対応
 - データ・エンリッチメント → 外部データやAI活用

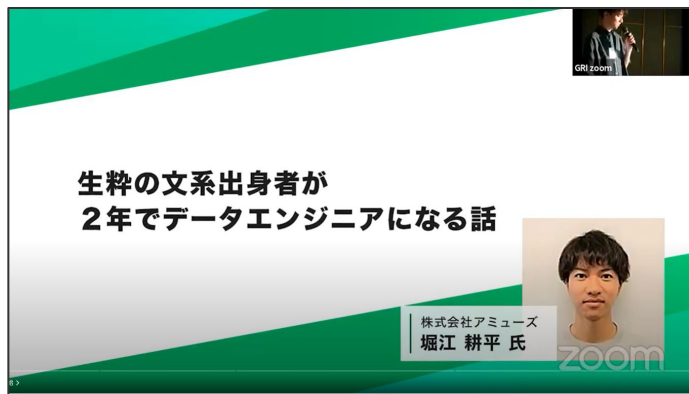
Dirty DATA

データエンリッチメントの例（読売新聞グループ様）

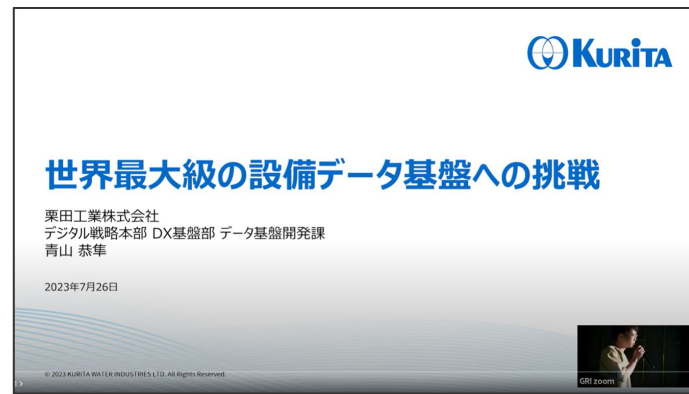


<https://www.youtube.com/watch?v=uOZeI51842I>

先ほどのMatillionがエンタープライズに向いている理由



https://www.youtube.com/watch?v=G_5IH0dv6Uo



<https://www.youtube.com/watch?v=vveRFROvQps>

Matillionについて



Matillionの主な特徴

- **DWHの強力なマシンパワーが使えるELTツール**
 - IoTセンサーなどの大規模データにも強い
- **非エンジニアでも扱いやすいノーコードかつGUIによる開発**
 - ボトルネックになりがちな社内浸透の壁を超えやすい
- **コネクタが豊富であらゆるシステムと連携可能**
 - 社内に乱立するシステムを統合して一括管理
- **費用は従量課金制**
 - 必要最低限のコストかつ予算が組みやすい

Matillion×Snowflakeの大規模データ基盤構築の事例

- **依頼内容**

- 工場内に設置したセンサーから計測値を集計し、監視する
- 1工場につき計測項目は200～300、工場数は10,000箇所
- 毎分データを取得
- 取得したデータは時間単位に集計
- 3年分のデータを保持
- 定期的に3年分のデータを再計算

- **検証内容**

- Matillion×Snowflakeでパイプラインを構築
- Snowflakeに1兆行（約2TB）と10兆行（約29TB）のテーブルを用意
- 6XLのウェアハウス（通常の512倍）で再計算処理を実行
- 依頼から検証まで約2ヶ月

Matillion×Snowflakeの大規模データ基盤構築の事例

- **検証結果**

- **なんと無事最後まで処理が完了！Snowflakeすごい！**

	処理時間	1回の処理にかかった費用
1兆行(約2TB)	約7分	約25,000円
10兆行(約29TB)	約50分	約170,000円

Snowflakeがすごいのはわかった

Matillionで開発期間が短縮できるのもわかった

ELTツールならMatillionじゃなくてもいいのでは、！？

Matillionのデモ



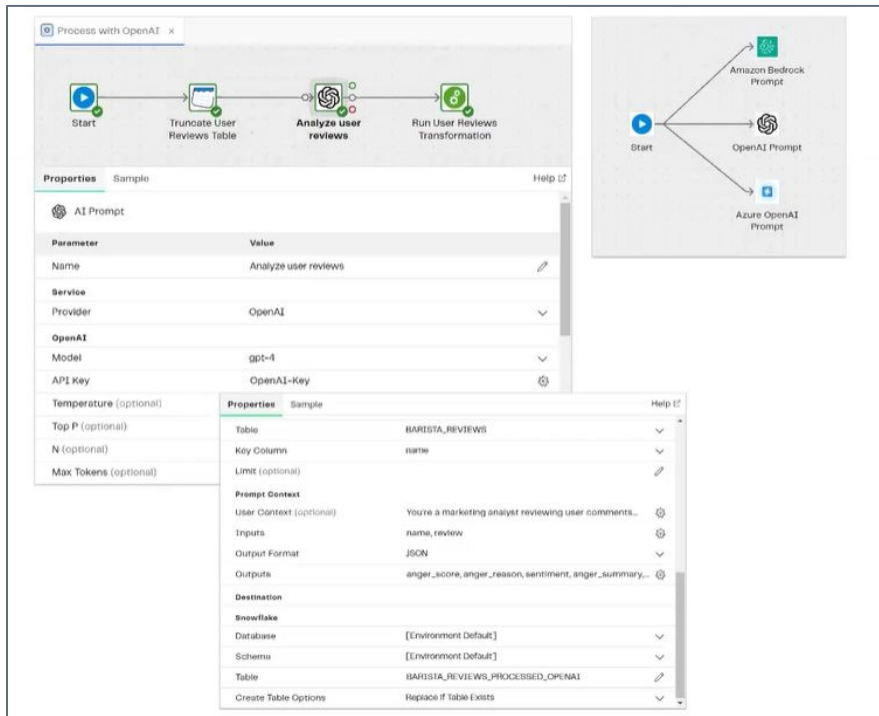
Matillionであなたのデータ基盤はこう変わる！

- **わかりやすいGUIで属人化を防ぐ！**
 - コンポーネントの配置が自由で処理の分岐も作りやすい
 - 日本語対応かつ自由に配置できるコメント機能で伝わりやすい
 - 加工途中のデータサンプルを確認できるので処理の内容がわかりやすい
- **ジョブのカスタム性能が優れており開発効率爆上げ！**
 - 単純なスケジュール実行だけでなくジョブチェーンが作れる
 - よくある処理はSharedJobで共通化
 - SQLやPythonのスクリプトも書けちゃう
- **地味に困っていたところにも手が届く機能で業務改善を図る！**
 - 複数のTransformationの処理も一括でSQL変換可能
 - 高額なツールを使わずともリバーズETLを実現

Matillionの将来像



今後のMatillionはLLM連携にシフト



The screenshot shows a workflow in Matillion titled "Process with OpenAI". The workflow consists of four steps: Start, Truncate User Reviews Table, Analyze user reviews, and Run User Reviews Transformation. The "Analyze user reviews" step is highlighted, and its configuration is shown in the "Properties" pane.

AI Prompt Configuration:

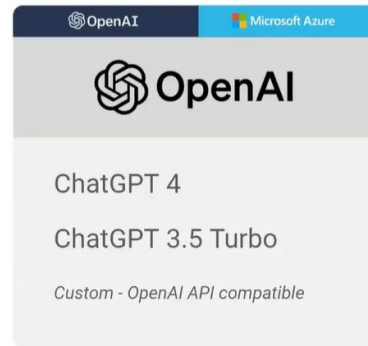
Parameter	Value
Name	Analyze user reviews
Service	
Provider	OpenAI
OpenAI Model	gpt-4
API Key	OpenAI-Key
Temperature (optional)	
Top P (optional)	
N (optional)	
Max Tokens (optional)	

Properties Pane (Analyze user reviews):

Property	Value
Table	BAHISTA_REVIEWS
Key Column	name
Limit (optional)	
Prompt Context	
User Context (optional)	You're a marketing analyst reviewing user comments...
Inputs	name, review
Output Format	JSON
Outputs	anger_score, anger_reason, sentiment, anger_summary...
Destination	
Snowflake	
Database	[Environment Default]
Schema	[Environment Default]
Table	BAHISTA_REVIEWS_PROCESSED_OPENAI
Create Table Options	Replace if table exists

AI Prompt Component

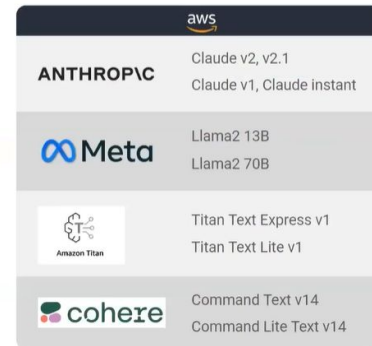
Supported Providers/Models



OpenAI Microsoft Azure

OpenAI

- ChatGPT 4
- ChatGPT 3.5 Turbo
- Custom - OpenAI API compatible



aws

- ANTHROPIC**
 - Claude v2, v2.1
 - Claude v1, Claude instant
- Meta**
 - Llama2 13B
 - Llama2 70B
- Amazon Titan**
 - Titan Text Express v1
 - Titan Text Lite v1
- cohere**
 - Command Text v14
 - Command Lite Text v14

例えば、「顧客の声」や「商品カテゴリ」の適切な分類をデータパイプラインの中でLLMを活用して実施

LLMを正しく使う & 検算する能力 / 体制

The screenshot shows the Microsoft Copilot interface within a data tool. On the left, the Copilot chat window is highlighted with a red box, containing a text input field and a 'Submit' button. The main area displays a pipeline diagram with a 'Netflix series' component connected to a 'Calculate Movie Age' component. To the right, a configuration panel for 'Calculate Movie Age' is shown, with a red box around it, containing 'Configuration', 'SQL', 'Include Input Columns' (set to 'Yes'), and 'Calculations' (set to '(2023 - "YEAR")'). Below the pipeline, a 'Sample data' table is displayed, also highlighted with a red box, showing columns for ID, TITLE, YEAR, IMDBRATING, SYNOPSIS, UPDATED_DATE, and MOVIE_AGE. The table contains two rows of data.

やりたいことを適切に思い付き
適切なプロンプトを入れる能力

検算する能力

ID	TITLE	YEAR	IMDBRATING	SYNOPSIS	UPDATED_DATE	MOVIE_AGE
70113305	Dev.D	2009	8.8	Because they come from different castes, the son of a tax collector and his true love are not allowed to marry, sending them down divergent paths.	2023-12-20 19:57:02.951	14.00000
81218363	Sillu Karuppatti	2019	8.8	From first	2023-12-20	4.00000

今後の予定



今後の予定

Matillionユーザ会

<https://mug-japan.studio.site/>

お問合せ先： GRIのホームページより

(採用強化中、案件のご相談)

<https://gri.jp/contact/?id=9773>

次回のすいすい会

Xにて告知 https://twitter.com/gri_2017

- 自動機械学習やオープンデータ活用を想定
- 「ここだけのデータ基盤の話をぶっちゃける会」
オフラインでエンタープライズ同士の情報共有会を企画中