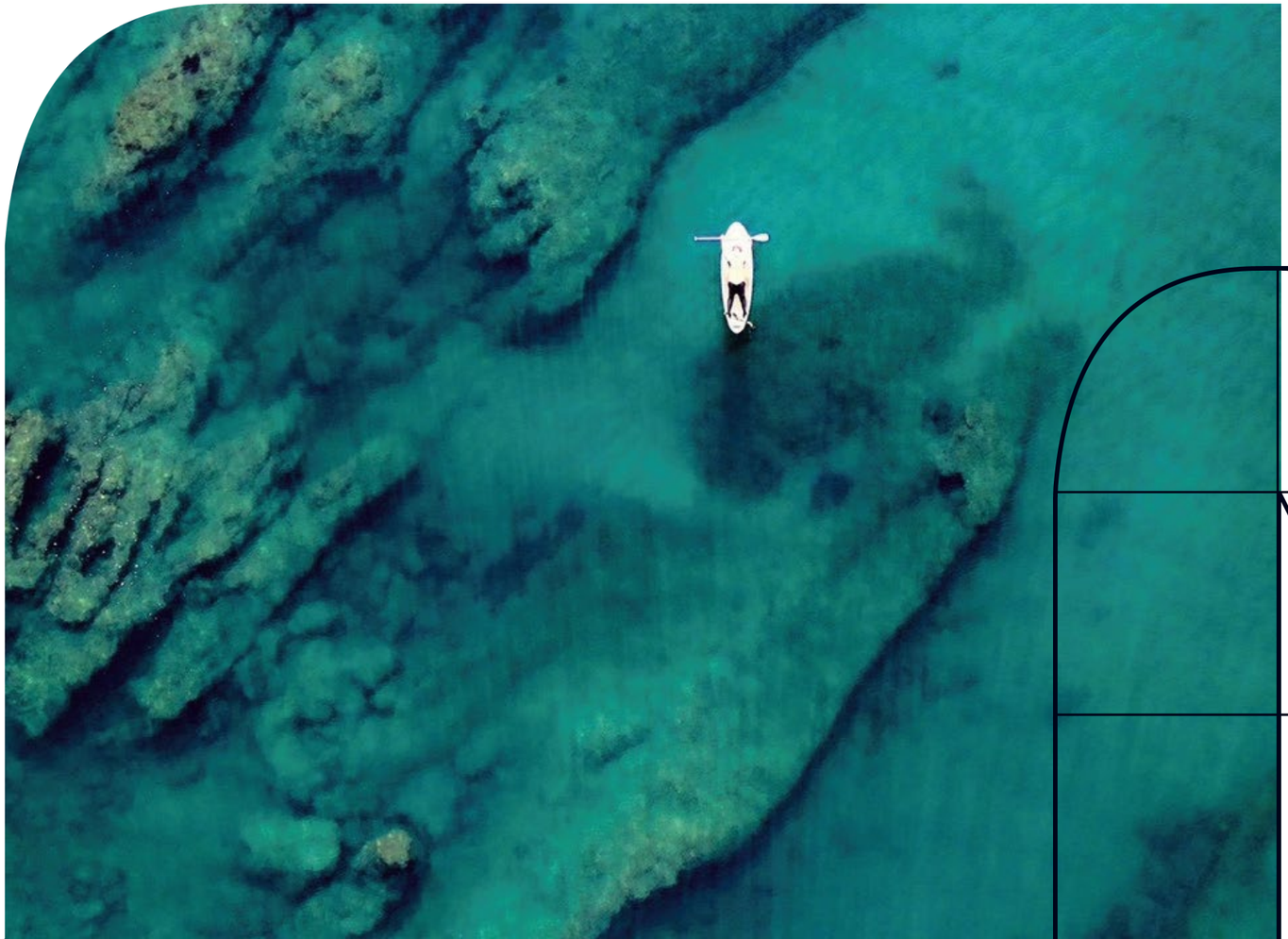
NTT DATAが考える
生成AI時代のデータ活用

～Text2SQL、セマンティックレイヤー、AIエージェントを用いたデータ活用・データマネジメント方法～



株式会社NTTデータグループ
〒135-6033 東京都江東区豊洲3-3-3豊洲センタービル
Tel: 03-5546-8051 Fax: 03-5546-2405
https://www.nttdata.com/jp/ja/

筆者

山口 永

エグゼクティブデータエンジニア

基幹系業務へのAI適用など、開発難易度が高いAIを使ったサービスを対象に、長年のMLOpsやDataOpsの実務経験を基に、AI、Data、Software、Infraの4領域を横串でのTechLeadを担っている。
昨今は生成AIを使ったAI-Agentの導入支援及び、オファリング開発に従事している。

長塚 健汰

アソシエイトデータエンジニア

データ分析基盤の開発運用から、BI製品や機械学習を用いたデータ活用まで支援。
さらにAIを利用した高度化のR&Dを通じ、最新のトレンドを現場に取り入れている。

Chapter.1 はじめに

Chapter.2 データ活用を減速させる課題

Chapter.3 データ基盤の開発迅速化に向けた取り組み

Chapter.4 “対話型”という新しいデータ活用の形

Chapter.5 データのサイロ化を防ぐ技術「セマンティックレイヤー」

Chapter.6 NTT DATAが考えるこれからの世界観

Chapter.7 おわりに

Chapter. 1

はじめに

膨大なデータの中から必要な情報を選別する「データマイニング」の活用が始まったのは、1990年代。2000年代には「ビッグデータ」という概念が定義され、同時に「機械学習」や「人工知能」の分野も発展しました。2010年代にビッグデータの技術が進化すると、それに伴って意思決定や顧客理解、新規サービスの創出など、データ分析はビジネス上ますます重要になり、現代では多くの企業が最新のデータ分析基盤を導入し、データの民主化を進めています。

一方でデータの民主化に伴って多くの課題も表出してきています。
規模の大きい企業では、データを利用し価値を出したい事業部門の人数に比べて、データを整備し提供する専門家の人数は圧倒的に少ないことがほとんどです。データの民主化により急増した利用者からの需要に、提供側のエンジニア組織が追い付けていない企業も多いのではないでしょうか。
また、多くの“市民データサイエンティスト”にとって専門的なツールの利用は難しく、目的のデータを見つけることが高いハードルになってしまっています。とはいえエンジニア組織の稼働は逼迫していて、結局手元のエクセルでお手製のテーブルやレポートを自作し、データ活用がサイロ化してしまっている状況も少なくありません。

NTT DATAでは、生成AIでデータ活用の課題を解決する取り組みを始めています。LLM※1、Text2SQL※2、セマンティックレイヤー※3など、キーとなる技術を活用してビジネス上の意思決定を支援し、社内でのデータ活用を促進します。
このホワイトペーパーでは、データ活用における課題、データ活用を促進する生成AIの技術、今後の展望などについてご紹介します。

※1 LLM（大規模言語モデル）：大量のテキストデータで学習し、自然言語の生成や理解を行う高度なAIモデル。
※2 Text2SQL：自然言語で書かれた質問や指示文をSQLクエリに自動変換する技術。
※3 セマンティックレイヤー：データとデータ分析者の間に位置し、データの意味やビジネスロジックを抽象化してデータの利活用を促進する中間層。

Chapter. 2

データ活用を減速させる課題

本文では、データ活用のステークホルダーである、データ提供者とデータ分析者のそれぞれに対して、データ活用にかかわる具体的な課題感をご説明します。

2.1　データ提供者の課題

ビジネス側が求めるスピード感で、データを提供できない

データ活用において最も時間や手間がかかるのは、分析それ自体ではなく、実は分析に用いるデータの準備・整形です。一般的に全体の取り組みの中でおよそ8割がこの工程に割かれるといわれており、データ活用のアジリティを高めるための肝になっています。一方で多くの企業ではビジネス側が求めるスピードでデータを提供することは難しく、一つのデータマート※4の作成に何週間も要してしまうという課題を抱えているケースが多く見られます。データ提供のスピードを改善するため、開発手法の見直しや開発補助ツールの導入を検討している組織も増えてきています。

2.2　データ分析者の課題

分析に必要なデータが見つからない

データ分析の初期フェーズでは、分析に必要となるデータに関する知見が不足していることが多々あります。どのようなデータが分析に利用できるかを知るためにはデータカタログを参照することになりますが、一般的なデータカタログ製品はn-gram（単語ベース）の検索になっており、正確なテーブル名やカラム名をあらかじめ把握しておく必要があります。一方で多くの分析者はそうした情報を必ずしも把握しておらず、これがデータ探索の障壁となっています。このような状況では、多くの時間がデータ探索に費やされ、効果的な分析に集中しづらくなってしまいます。データ分析の効率を向上させるためには、必要なデータに迅速にアクセスできる体制を整えなければなりません。

ビジネスユーザーのツールの習熟度が低く、データ分析できる人が少ない

さらに多くのビジネスユーザーが、データ分析で活用するBIツール※5や分析ツールなどの知識を十分に持っているわけではなく、分析作業に対して心理的・技術的な負担を感じています。BIツール、SQL、Jupyter Notebook※6などの分析ツールは高度な機能を提供しますが、これらを効果的に活用するためには一定の習熟度が求められます。分析のハードルを下げるためには、ビジネスユーザーが自律的に負荷の低い状態でデータ分析を行える環境を整備しなければなりません。

※4 データマート　：特定のビジネス部門やプロジェクトにおいて、用途、目的に応じて整形されたテーブルまたはビュー。
※5 BIツール　：データを分析・可視化し、ビジネスに活用するためのツール。
※6 Jupyter Notebook　：ブラウザ上でソフトウェアを開発したり、共有したりできる対話型実行環境。

2.3　データ提供者・データ分析者共通の課題

指標値管理の労力が大きい

大規模な組織では、KPIのような指標値の管理が課題になっているケースが非常に多く見受けられます。サービスの開始や終了、年度ごとの経営方針の見直しを受けて、こうした指標値は常に定義の見直しの必要があり、そのメンテナンスは非常に労力のかかる作業です。また、同じKPIでも各組織にとって必要な情報は微妙に異なっており、こうした違いをどこでどのように吸収するかは大きな課題です。

よくある指標値管理の事例

① データウェアハウス（以下、DWH）側で開発者が指標値に対応したViewを用意

これは最もよく見られる解決方法ですが、データの分析利用者が増えるほどデータ提供者の負担が大きくなってしまいます。単純な指標値一つとっても、利用者の数だけ注目する集計軸が異なるため（売上高を地域ごとに見たい、前月比を見たい、商品カテゴリごとに見たいなど）、あらかじめすべての集計軸の組み合わせに対応するViewを作ることは非現実的です。

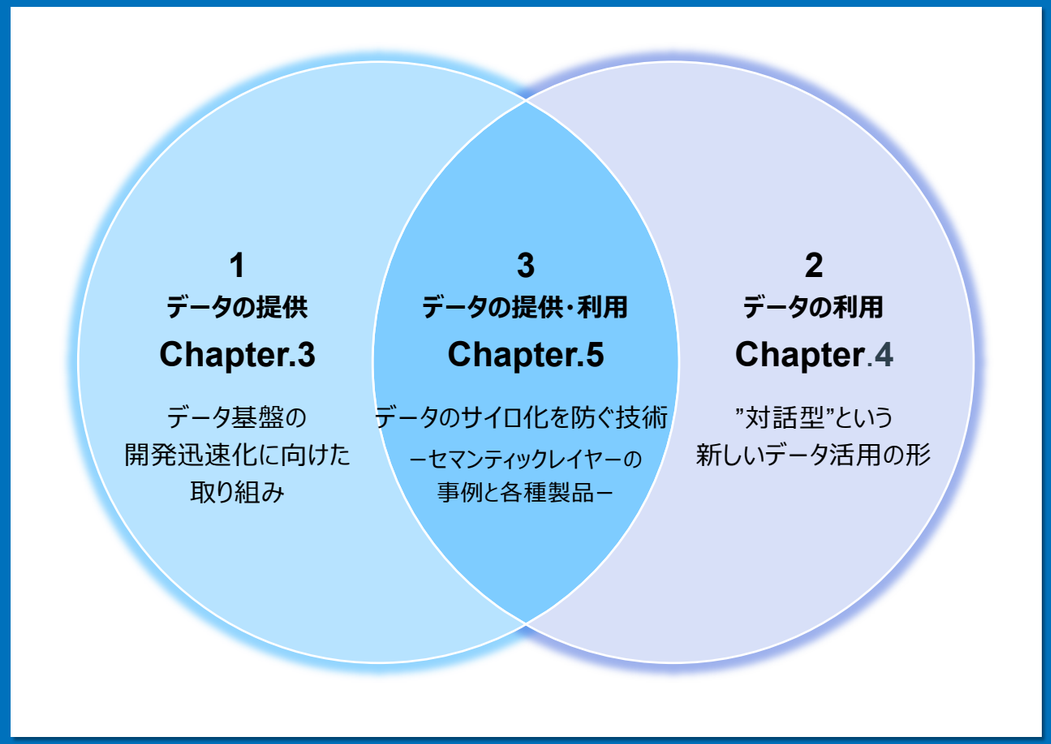
② 各種BIツールで利用者自身が指標値を定義

データエンジニアによるデータマートの提供スピードがデータ分析者の需要に追い付けない場合、データ分析者が自分たちで管理するBIツールの中で、独自のデータマートを作っているケースがあります。DWHでは、巨大な大福帳を作っておいて利用者部門で任意の切り口で集計分析をするのは、バランスが取れているように見えます。しかし、特に大規模な企業で多くの部署がそれぞれデータの集計・分析を行っている場合、KPIなどのビジネス指標をそれぞれ独自のロジックで算出してしまい、部署間で指標の数字がずれてしまうという課題が発生します。

2.4　本稿の構成

以上で挙げた課題について、本稿では3章でデータ提供者への対策、4章でデータ分析者への対策、5章でデータの提供・利用双方に共通の対策を紹介します。さらに6章では、最新のトレンドを踏まえてNTT DATAが考えるこれからのデータ活用の世界観について述べます。

図2.4　データ活用を減速させる課題



Chapter. 3

データ基盤の開発迅速化に向けた取り組み

NTT DATAは、前章で取り上げたデータ提供者の課題に対し、生成AIを活用することでSQLを自動生成し、データマート開発の生産性の向上をサポートしています。

3.1 生成AIをデータマート開発で活用

生成AIを利用した開発効率化は様々なアプローチが取られていますが、最も効果が出やすいのがコーディング作業の効率化です。様々な製品がSQL生成機能を提供しており、検索すれば各製品のデモ動画を閲覧できますが、どちらかというとライトユーザー向けに「2024年の商品別売上」のような単純な条件でSQLを生成させているようなデモが多いように思われます。一方で商用開発において作成するデータマートは複雑な業務要件の塊であり、その要件を一つ一つプロンプトに与えることは生産性をかえって損なってしまいます。

DWH製品のSQL生成機能の紹介

Databricks Assistant

Databricksが提供しているDatabricks Assistantは、自然言語で問い合わせるだけで、コードやクエリの生成、デバッグ、最適化、説明をサポートします。具体的には、SQLやPythonコードに関する質問や指示文に対して、Unity Catalogのメタデータを基に最適なクエリの生成や説明を提供し、エラーの自動修正やインラインコードの提案を行います。また、ダッシュボードからDatabricks Assistantを使用して、データの分析・視覚化、フィルタリングを行うことも可能です。

※ NTT DATAと資本業務提携。

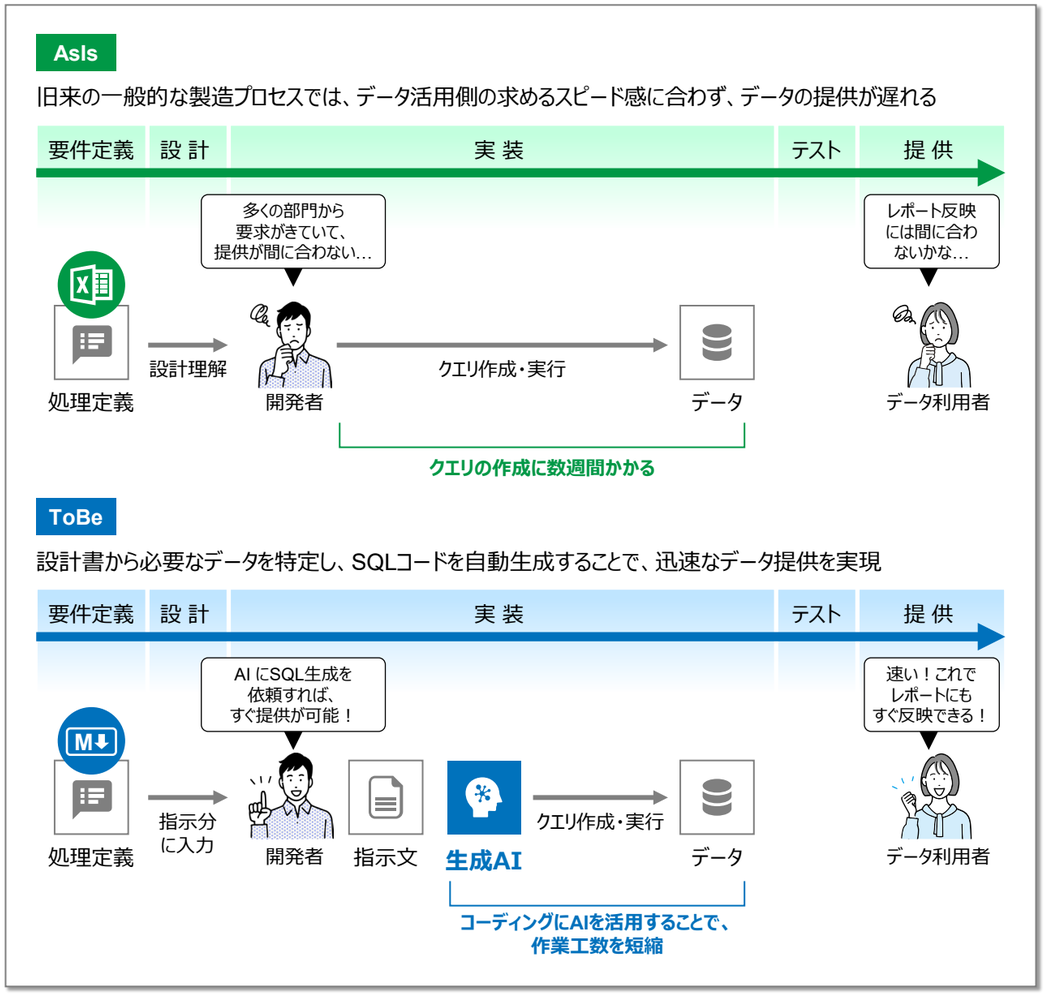
Snowflake Copilot

Snowflakeが提供しているSnowflake Copilotも自然言語で問い合わせ可能なAIアシスタントです。SQLクエリの生成および最適化をサポートします。Snowflake Copilotは、SQLクエリに関する質問や指示文に対して、テーブルスキーマ情報を基にテーブル間の関係を理解し、SQLクエリの生成や最適化、問題点の修正を行います。また、Snowflakeの機能に関する質問でも、Snowflakeのドキュメントを基にした回答が可能です。

※ NTT DATAは、Snowflakeの最上位のElite Partner。

当社では、設計書を直接こうしたSQL生成機能のプロンプトに与えるというやり方での開発効率化を進めています。製造工程の前には必ず設計工程があり、設計の成果物である設計書をそのまま利用することで、開発の流れを変えずにAIを導入することができます。この際、設計書がエクセルなどで書かれていた場合はプロンプトとして与えることができないため、Markdownなどに書き下ろす必要があります（最初からMarkdownの形式で設計書を書くこともよいでしょう）。また、シンプルに設計書だけをプロンプトに与えるよりも、Few-Shot※7形式で指示文を書けば、細かい工夫により精度向上を図ることが可能です。

図3.1　生成AIによるSQL生成のイメージ



※7 Few-Shot ： 少量の事例をプロンプト内に提示して、AIが新しいタスクに対応できるように学習させる方法。

3.2　設計書の重要性

設計書を利用することで商用開発に適用可能なSQLを生成することができつつある現在、翻って設計書の質が今後のAIを利用した開発におけるボトルネックになってくる可能性があります。当社においても、生成されたSQLの質が低い場合にプロンプトを試行錯誤してもうまくいかず、結局設計書の記載が間違っていたというケースがたびたび見受けられています。また、コーディング作業がAIの支援を受けて高速に行えるようになってきたことで、設計書の作成速度がコーディング作業に追い付かない、といったケースも今後増えてくると思われます。

こうしたトレンドを踏まえ、NTT DATAでは設計書自体もAIで作成支援するという取り組みを行っています。これは、さらに上流の要件定義フェーズの成果物をインプットにAIが設計書の作成支援を行うものです。一般的なアプリケーション開発にAIを導入しようとすると、基本的にはこのように上流工程の成果物をAIへのインプットとして利用していくことになります。一方でDWHでのデータマート・データパイプライン開発においてはまた違った観点があり、6章で将来的なトレンドとしてご紹介します。

Chapter. 4

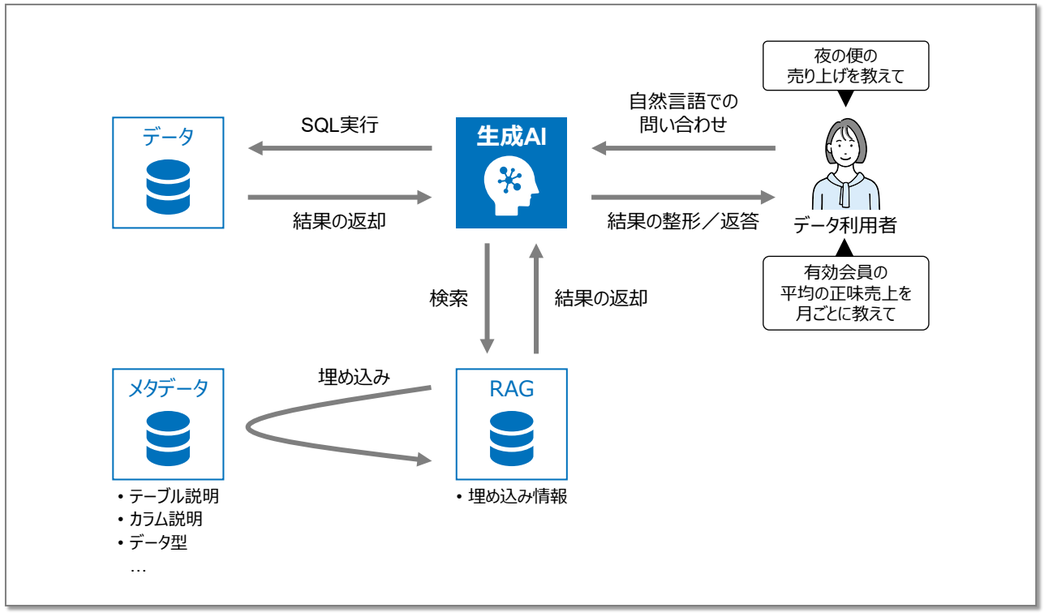
“対話型”という新しいデータ活用の形

本章では、データ分析者の課題について、生成AIを活用した解決方法を説明します。AIによるデータの探索や集計・分析のサポートと、それを実現するためのメタデータの整備について解説します。

4.1　データの民主化を支える対話型のデータ活用

データの民主化での大きな特徴が、「非専門家が主役になる」という点です。必ずしも高い技術スキルを持っているわけではない人でも使いこなせるインターフェースとして、チャット形式で対話型のデータ活用の形が注目を集めています。こうした対話型のデータ活用においても、生成AIにどのような情報を与えるかが回答精度に直結します。製造フェーズと違い、分析フェーズでは設計書のような処理の詳細な仕様があらかじめ整理されていることは稀で、代わりにデータ自体に関する情報：メタデータ※8を生成AIに与えてあげることが有効です。

図4.1　対話型のデータ活用のコンセプト



データの探索

メタデータを埋め込んでRAG※8を構築することで、LLMを用いたセマンティック検索を実現できます。セマンティック検索とは、単語の意味や文脈を理解して、より関連性の高い検索結果を提供する方法のことをいいます。
例えば、「2024年の売上データを格納しているテーブル」や「顧客の購入履歴のテーブル」という自然言語の問い合わせに対しても、テーブルやカラム名に依存せずに、適切なデータセットを迅速に見つけられます。LLMによるセマンティック検索を利用することで、探索時間を短縮でき、データ分析の効率を大幅に向上できます。

※8 メタデータ　：次ページコラム参照
※9 RAG　　：Retrieval Augmented Generationの略で、自社に蓄積された大量の業務文書・規定などの社内情報などの最新情報を活用し、それに基づいて大規模言語モデル（LLM）に回答させる方法。

データの集計・分析処理

探索して見つけたデータに対する集計・分析処理についても、同様に自然言語で指示することができます。この際もAIがメタデータにきちんとアクセスできることで、組織独自の業務事情を理解したSQLを生成させることができます。
例えば「有効会員数の集計」といった業務を行うには、その組織での各用語の定義をきちんと把握する必要があり、こういった情報を把握した上でSQLを生成する能力がこのフェーズでのAI適用には求められます。この仕組みにより、データに不慣れなビジネスユーザーも容易にデータ分析に取り組むことができ、組織全体のデータ活用を促進します。

メタデータとは

メタデータとは、主となるデータの付帯情報のことで、データの用途や制約、適用範囲などの情報を整理する考え方です。膨大なデータの中から必要な情報を取り出すには、メタデータの管理が必要となります。
データ分析に使用されるメタデータは、以下の3種類があります。
特に分析フェーズにおける生成AIの利用においては、対象データの意味を説明するビジネスメタデータが重要です。

ビジネスメタデータ

- 業務ルール、テーブルやカラムの定義と説明など、業務に関する情報
- カテゴリ値の意味、0など特殊な値の解釈

テクニカルメタデータ

- カラムのプロパティやデータベースオブジェクトのプロパティなど、ITに関する情報

オペレーショナルメタデータ

- バッチプログラムのジョブ実行ログ、データの抽出とその結果などの履歴など、システム運用の過程で生成される情報

4.2　メタデータの重要性

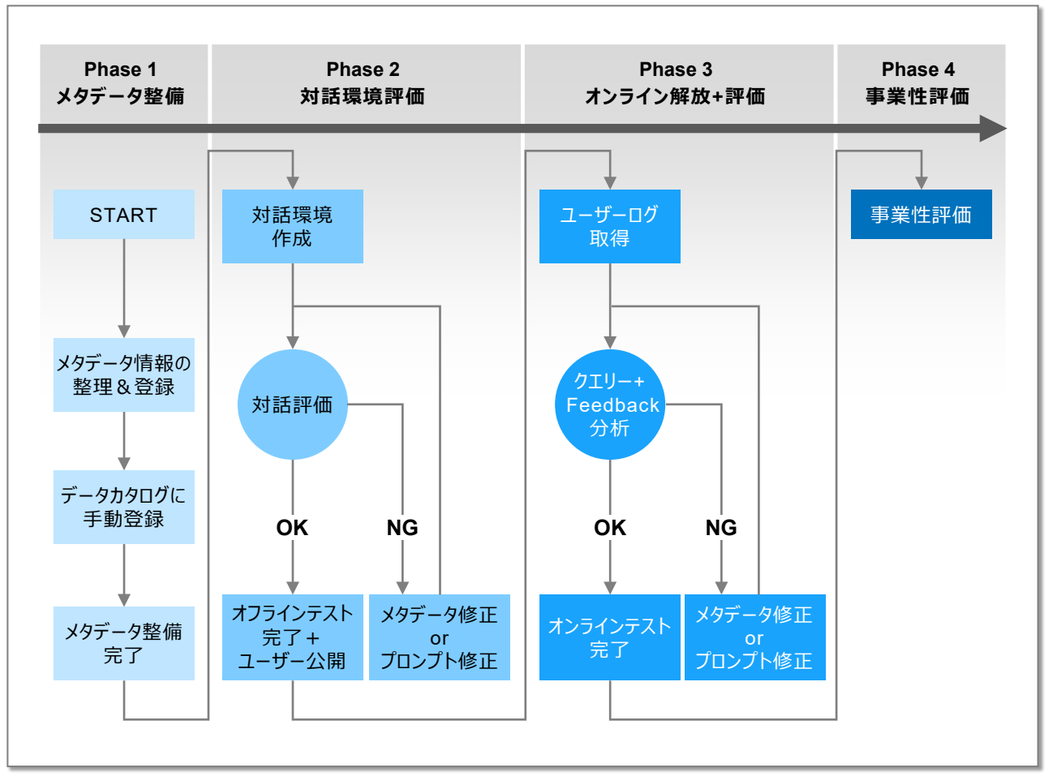
生成AIは決して銀の弾丸（開発の生産性をすぐに向上させる技術や実践）ではなく、メタデータが肝となります。とはいえ、メタデータも銀の弾丸ではありません。AIに活用可能な形で十分にメタデータを整備できている企業はいまだ少ないのが現状です。カラムの論理名やテーブルの更新頻度など、最低限の情報はデータマート開発の一環でカタログに登録されることが多いのですが、カラムの詳細な説明の記載は後回しにされてしまうことが少なくありません。これは単に開発時の優先度が低いというだけでなく、用意するデータマートの使われ方を事前に100%想定することができないため、どのような情報を記載すればよいのかをあらかじめ知ることが難しいという事情があります。
特に分析フェーズでの生成AI利用において重要なビジネスメタデータは、データ提供者とデータ分析者双方の業務プロセスの中で、能動的に育てていく必要があります。このような考え方は、従来のパッシブで静的な方法と対照的に「Active Metadata Management※10」と呼ばれており、近年のメタデータ管理の新しいトレンドになっています。

4.3　開発・導入・事業性評価の流れ

対話分析環境の開発・導入が通常のSI開発と大きく異なるのは、メタデータの質が機能要件の達成に大きく影響するという点です。したがって回答精度のテスト結果をもとに、メタデータの追加・修正のサイクルを回していくことが非常に重要です。開発者自身によるオフラインテストはもちろんのこと、対話分析は実際にどのように使われるかが事前に把握しづらいため、ユーザーによるオンラインテストも実施することが望ましいでしょう。 また、メタデータとしては一般的なデータカタログに登録するもの以外にも、FAQの形で質問と回答のセットを直接登録することも可能です。特にユーザーからのフィードバックを受けて改善していくフェーズでは、質問に対して適切な回答をしたものに対して評価をしてもらい、評価の高いものをFAQに登録していくことで継続的な精度改善が見込めます。
これらの取り組みを行った後、データ利用者を対象に分析時間短縮や利便性向上など、アンケートやインタビューを通して事業性を評価します。その上で、データ利用者視点で利便性向上をインセンティブに、利用者が能動的にメタデータを整備していけるような仕組み整備と啓蒙を行っていきます。

※10 ：従来のトップダウンの指示に基づく受動的な方法では無く、自分から主体的にメタデータを整備・収集を可能とする方法論
Gartner Research「クイック・アンサー：アクティブ・メタデータとは何か」
https://www.gartner.com/en/documents/4556899

図4.3 【AI×データ活用】対話分析環境整備



Text２SQLの精度向上テクニック

回答精度を向上させるためには、以下のようなメタデータを追加する事が有効です。

		（例）
①データモデリング	**A） エンティティ間の関係**	AテーブルとBテーブルの複合主Keyと外部Keyの関係
	B） 適切なデータ型	日付形式は日付型とする ∵ 日付に関する計算を、AIに判断しやすくするため
②メタデータ	**C） ドメイン定義**	コード値に関する、個別コード値それぞれの業務的な意味
	D） ドメイン固有情報	ビジネスメタデータや企業固有の用語定義
	E） フィルター条件	最新の在庫棚卸日付等のドメイン知識に基づくフィルター条件
③対話例	**F） クエリ例**	分析環境で頻出する問い合わせクエリ
④複雑なテーブル生成	**G） マテリアライズドビュー**	複雑なロジックから生成される一時View ∵ ハルシネーション防止のため、AI側で生成させない
	H） ユーザー定義テーブル関数（UDTF）	複雑な集計指標定義 ∵ ハルシネーション防止のため、AI側で生成させない

自然言語でデータ分析が行えるAI／BI Genie

Databricks社が提供しているAI／BI Genieは、ビジネスユーザーが自然言語を使用してセルフサービスのデータ分析や可視化を行うことが可能です。
具体的には、ビジネスドメインの質問や指示文に対して、Unity Catalogのメタデータを基に分析クエリに変換し、表や図を用いて視覚的に回答します。また、データの変更や新たな質問に応じて、メタデータを継続的に更新します。間違った回答に対するチューニングも可能なため、より正確な分析情報をユーザーに提供できます。
さらに、ベンチマークを使用すると、ユーザーが頻繁に行う質問内容に対して回答精度を評価するためのテスト問題を作成できます。

Chapter. 5

データのサイロ化を防ぐ技術「セマンティックレイヤー」

データ提供者、データ分析者共通の課題である「指標のずれ」を解決する方法として、近年注目を集めているのが「セマンティックレイヤー」という技術です。本章ではセマンティックレイヤーの概要を歴史とともに紹介し、さらにこれが前章の対話型分析にも活かされることを確認していきます。一方で新しい技術分野であることから、現段階での課題についても整理します。

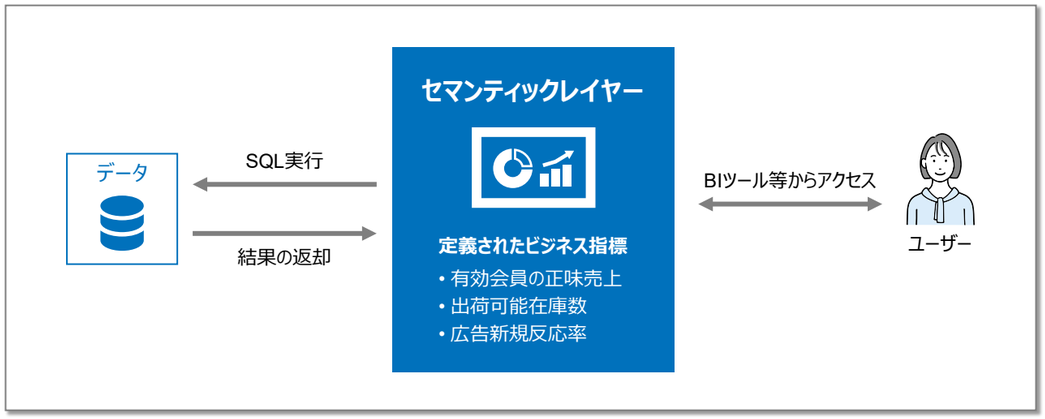
5.1　セマンティックレイヤーの歴史

まったく新しい技術領域という印象を持つ方もいるかもしれない「セマンティックレイヤー」という言葉ですが、これはもともとBIツールにルーツを持つものであり、普段からBIを触っている人には馴染みやすい概念です。類似する機能としてPower BIでは「セマンティックモデル」、Lookerではそのまま「セマンティックレイヤー」と呼ばれています。
BIではダッシュボードを作る前にまずテーブル間の関係性を定義し、指標の計算のために新しいカラム（BIでは「メジャー」と呼ばれることが多い）を作成します。こうしたデータの意味（= セマンティクス）をあらかじめ定義しておくことで、異なるダッシュボードでも同じ指標の数字をずれなく表示することができるようになります。
近年注目されているセマンティックレイヤーは、厳密には「（ユニバーサル）セマンティックレイヤー」と呼ばれており、もともと各BIツール内に閉じた機能として提供されてきた（ローカルな）セマンティックレイヤーを、様々なツールが共通してアクセスできるように独立して存在させたものです。
近年、データの民主化とそれに伴うサイロ化を受けてその注目度は上がっており、2024年には第4回Semantic layer Summitが開催され、DWHの父といわれるBill Inmon氏も出席しています。

5.2　セマンティックレイヤーとは

（ユニバーサル）セマンティックレイヤーは、データとデータ分析者の間に位置し、複雑なデータの意味やビジネスロジックを理解可能なビジネスの概念に変換・翻訳する中間層のことをいいます。

図5.2　セマンティックレイヤーの概念図



BIツールの場合と同様にテーブル間の関係性を定義し、KPIなどのビジネス指標のロジックも再利用できるよう定義しておきます。例えば「有効会員」と「正味売上」、それらを合体した「有効会員の正味売上」といったようなロジックを定義しておくことで、セールス部門では営業店舗ごとに月次で集計したり、マーケティング部門では商品カテゴリごとに四半期で集計したりできます。こうした異なる切り口の分析に対して、従来は都度データマートを丸ごと新しく作るケースが多かったのですが、それでは開発者の負担が増えてしまいます。注目する指標のコアとなるロジックはセマンティックレイヤーで一元的に管理する一方で、分析の切り口は利用者側が適宜柔軟に選べるようにしておくことで、データガバナンスを効かせつつデータ活用のアジリティを高めていくことができます。このように開発者にも分析利用者にも嬉しい一石二鳥的な技術領域がセマンティックレイヤーなのです。

5.3　セマンティックレイヤー×生成AI

セマンティックレイヤーはデータの意味（セマンティクス）を定義し、データの分析・利用において一貫したロジックを提供します。これは非常に貴重なメタデータであり、前章と同じようにこのメタデータを利用した対話型分析がセマンティックレイヤー製品の一機能として提供されるようになってきています。

製品事例

Cortex Analyst

Snowflakeが提供するCortex Analyst※11は、ビジネスユーザーが自然言語で質問し、高精度かつ信頼度の高い回答を得られるデータ分析用途のアプリケーション開発支援機能です。具体的には、セマンティックモデルで事前定義したメジャーやディメンション情報を基に、ビジネスドメインの質問や指示文を高精度なSQLクエリに変換し、信頼性の高い回答を提供します。

また、APIを通じて既存のビジネスツールとシームレスに統合できるだけでなく、データはSnowflakeのセキュリティとガバナンス枠内で安全に処理されます。

Ask dbt

dbtが提供するAsk dbt※12は、セマンティックレイヤー機能とLLMを組み合わせ、自然言語でデータの問い合わせが可能なアプリケーション開発支援機能です。dbtセマンティックレイヤーでは、テーブルからセマンティックモデルおよびビジネス指標であるメトリクスを定義できます。Ask dbtは、自然言語での質問や指示文に基づいて、dbtセマンティックレイヤーにて定義されたメトリクスやディメンションを選択および実行し、結果をアウトプットします。これにより、複雑なメトリクスの取得に対しても、一元管理されたメトリクスを使用して高い精度の返答を実現します。

※11 Cortex Analyst ：https://docs.snowflake.com/en/user-guide/snowflake-cortex/cortex-analyst
※12 Ask dbt ：https://docs.getdbt.com/docs/cloud-integrations/snowflake-native-app

5.4　普及に向けての課題

データのサイロ化を防いでくれる上に、AIと連携してデータの分析までサポートしてくれる夢のような技術ですが、普及に向けた大きな課題として、情報の参照範囲の狭さがあります。Ask dbtもCortex Analystも、現状はセマンティックレイヤー用に定義したメタデータの情報しか参照することができません。したがって、例えばテーブルやカラムといったオブジェクトにDDL文でコメントをつけても、その情報を利用することができません。そのようなメタデータを利用するには、セマンティックレイヤー用のメタデータの定義ファイルにその情報を書き写す必要があり、二重管理で手間もかかってしまいます。定義ファイルの作成のサポート機能が提供されている場合もありますが、それでも根本的にメタデータが分散してしまうため、現場での運用上の負担は大きくなってしまいます。
この点については、改めて次章で最新のトレンドとNTT DATAの構想をご紹介します。

Chapter. 6

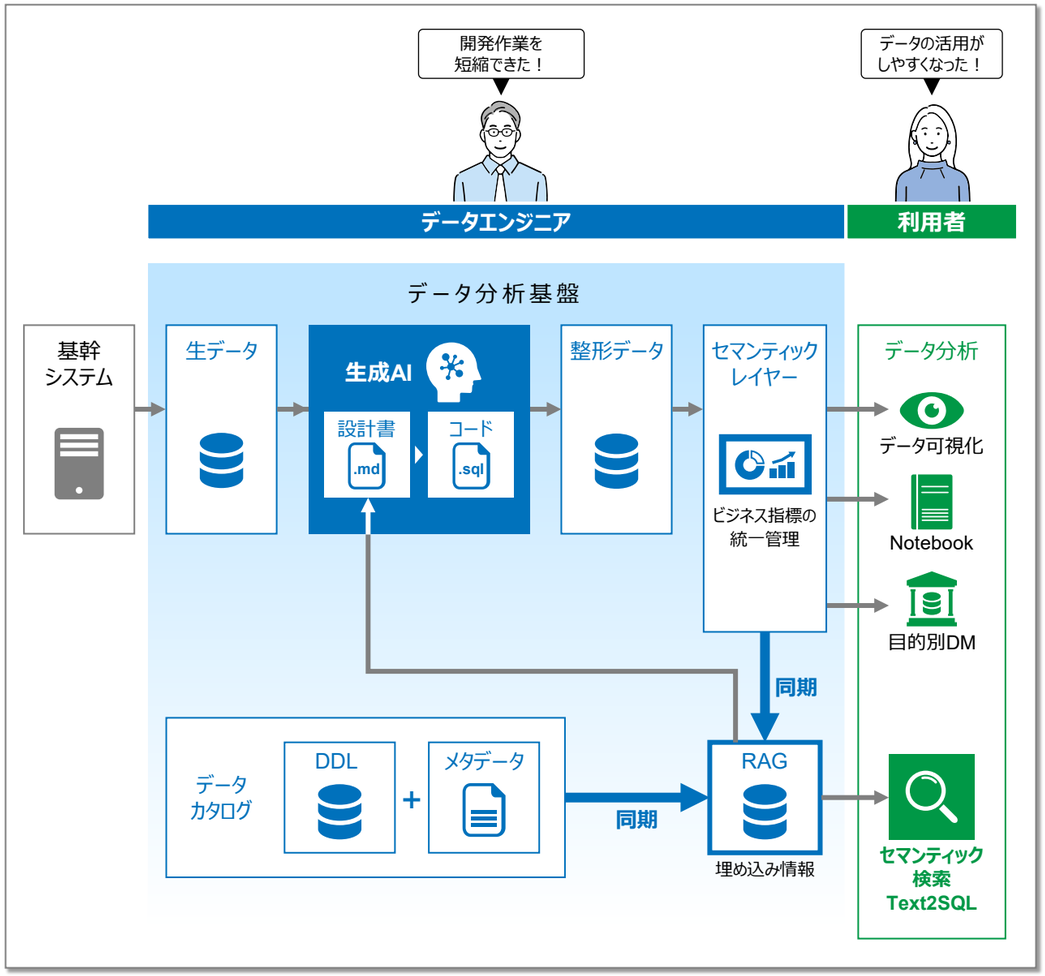
NTT DATAが考えるこれからの世界観

データ活用領域で生成AIを活用するためには、生成AI活用を前提としたデータマネジメントが必要となります。
従来からあるデータマネジメントで扱っているメタデータの整備だけではなく、データ提供業務におけるETL開発時の設計書やデータ分析業務で利用するセマンティックレイヤーなど、新たなメタデータを導入します。
その上で、当該メタデータを使って、データ活用領域の業務効率化が行えるような、データ活用に特化した AI-Agent を導入することで業務の自律実行を実現し、データ活用全体のプロセスを加速化&高度化を図ります。

6.1　メタデータの統合

これまでの章で、生成AIの活用にあたっては「いかにAIに十分な背景情報を与えるか」が重要であることをご紹介しました。 さらに一歩踏み込めば、とりわけDWHでのAI活用においてはその中心であるデータそのものに関わる情報の提供、すなわち「いかにAIに十分なメタデータを与えるか」が重要だというのが、近年のDWHにおけるAI活用の大きなテーマです。
現状の生成AIのコンテキスト（人間で言うところの記憶領域）長は有限であることから、膨大なメタデータから都度必要な分だけAIに渡す技術であるRAGが急速に注目を浴びていますが、ここで重要なことは、RAG基盤が参照できる情報の範囲です。いかにRAGが優れた技術で、例えば経理の基幹システムから送られてきた生データの情報を持っていても、営業部門が作成したデータマートについては知らないという状態であれば、データの探索も集計もやりづらいことは想像に難くないでしょう。
5章ではセマンティックレイヤーが有用なメタデータ整備ツールであると同時に、それを活かした対話分析機能も提供している一方で、データカタログなど他製品のメタデータは参照できないという課題を挙げました。実際にこのような課題は、様々なツールで見られます。BIでもデータカタログでもドキュメント管理ツールでも、ここ数年の生成AIブームに乗り遅れまいと様々な機能が次々とリリースされていますが、こうした生成AI機能が利用するメタデータは基本的に各製品内に閉じており、ほかの製品との連携まで考えられたものはまだ普及していないのが現状です。しかし将来的には、AIに真に「適切な」メタデータを渡すことのできる、製品の枠を越えて組織内のすべてのメタデータにアクセスできるRAGを備えた生成AIサービスが登場し、普及していくと当社は考えています。

図6.1　組織内のメタデータが一つのRAGに集約されている理想図



実際こうした方向を示す製品も出てきています。例えばDatabricks社はData+AI Summit 2024において「Unity Catalog Metrics」というサービスを発表しました。これはDatabricksのデータカタログ機能であるUnity Catalogに新たにセマンティックレイヤー機能を追加するものであり、さらにdbtやcube※13といった代表的なセマンティックレイヤー製品で定義されたメタデータとの連携も見据えたもので、まさに製品や技術領域をまたいでメタデータを統合しようという動きが見えてきています。

この10年、データ活用の主役はデータを管理するDWHでした。生成AI時代のデータ活用の主役は、メタデータを管理する製品になっていくでしょう。当社ではこの領域の最新動向を調査・検証し、お客様とともにビジネスの現場で価値を創出しています。

※13 Cube ： DWHとそれぞれのBIツールの間に独立して存在し、各BIツール内に閉じた機能として提供されてきたセマンティックレイヤーを、ツールをまたいでロジックの集約を図るサービス。

6.2　設計書・SQL生成はメタデータでより高精度に

3章でAIによるSQL生成では設計書の質が重要であること、またこうしたトレンドを踏まえ、当社でも設計書自体のAIによる作成支援に取り組んでいることを紹介しました。一方でDWHでのデータマート・データパイプライン開発においては、単に上流の要件定義工程の成果物をインプットするだけでなく、RAGの形で統合されたメタデータ基盤からの情報も同様にインプットとして利用していくことが、将来的なトレンドになる可能性があります。従来のAIを用いない開発でも既存テーブルの情報（カラム名、データ型、null値の解釈、更新頻度とタイミング、下流で使われているダッシュボードなどへの影響など）を調査しながら設計を行ってきたことを考えると、こうした情報は設計工程における必須要素であり、AIによる設計書の作成支援においても、精度向上に重要な役割を果たすものと思われます。AIの技術発展につれて、こうした情報をAIが利用可能な形で整備できていることが、組織の開発生産性を大きく左右するようになっていくでしょう。

6.3　統合されたメタデータ基盤がデータ利用の攻めと守りをサポート

データの民主化の文脈で、4章では技術がなくてもデータを活用できる対話型分析を、5章ではデータやロジックのサイロ化を防ぐセマンティックレイヤーを紹介しました。これまでは専門部署のデータサイエンティストが高度な分析を行い、したがってメタデータもその専門部署のメンバーが把握していれば大きな問題にはなりませんでした。しかし、事業部門・営業部門の一人ひとりが日々の業務でデータから価値を見出すためには、メタデータも閉じずに広く開放され、そしてサイロ化せず統合されたものである必要があります。部門間の横断的なデータ分析をより速く・正確に回していけるという攻めの効果が期待できるのはもちろん、組織内に存在するデータマートやダッシュボードを把握し、車輪の再発明を防ぎながら組織横断の最適化を図るという守りの意味でも、メタデータの統合基盤への需要は今後高まっていくと思われます。

6.4　AIエージェント×メタデータによる開発・分析の自動化

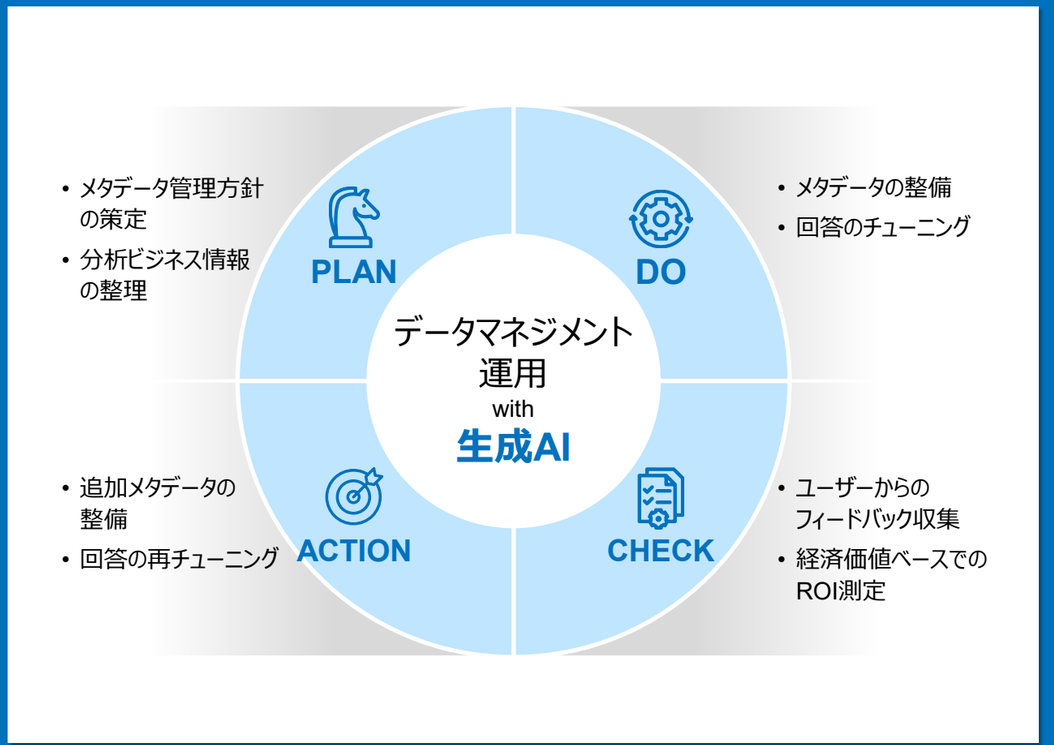
生成AIを利用したデータ活用業務の生産性と品質向上のアプローチをご紹介させて頂きましたが、まだまだ人が介在しなければならない作業は多い状況です。
そのような状況を解決するべく、今後は開発と分析を自律的に自動実行可能なデータ活用領域に特化したAIエージェントの導入が加速していくことが期待されます。
例えばデータ活用業務においては、データ提供者が開発するETLを対象に、要求仕様を元に、設計書作成、レビュー、テスト、BIツールへの反映までを自動で実行する事が期待されます。
データ分析業務においては、仮説検討や分析サイクルの実行に関しても、AIエージェントが自動的にインサイトのある分析レポートを作成してくれる事が期待されます。
ただしAIエージェントに渡すメタデータが不足している場合、十分な業務知識をAIエージェントに与えられない事から、AIエージェントはタスクを完遂する事は出来ません。
そのような事を踏まえると、メタデータを整備していくプライオリティはますます上がっていくことが予想され、AIエージェントが業務を自立的に実行出来るレベルまで、業務知識をメタデータに反映していくアプローチが必要となります。

6.5　能動的に自らメタデータを整備するようなデータマネジメントの未来

本稿では一貫してメタデータの重要性を強調してきました。メタデータの整備・管理は、一般的にデータマネジメントという大きい枠組みの一環として語られることの多いテーマです。データマネジメントは企業のデータ活用の成否を左右する非常に重要な取り組みである一方で、進め方がわからないというご相談を当社にいただくことも多い状況です。
データマネジメントのいろはを体系的にまとめているDMBOK※14を参照してみると、メタデータ管理、データ品質、ドキュメントとコンテンツ管理などの章が並んでいますが、概念としてのデータマネジメントと具体としての組織への導入方法との間に乖離を感じている企業が多いのが実情ではないでしょうか。
またデータマネジメント導入後の効果測定を行う場合、定量的に利用実態を把握し、経済価値としてROIがどのように改善するかを算出するのが難しい状況です。こうした難しさがデータマネジメントの現場での進め方、また組織内での予算獲得や成果のアピールがしづらい一因でしょう。
しかし、近年の技術発展でデータ提供者もデータ分析者も生成AIによってメタデータの価値を最大限に引き出し、自身の業務に直接・素早く役立てることができるようになってきました。この強いインセンティブは、組織内でのデータマネジメントの取り組みを進める上での大きな旗印になると同時に、AIによる業務効率化という経済指標を通して企業のデータマネジメントを定量的に評価するヘルスメーターとして機能することができます。このヘルスメーターを羅針盤にしてPDCAのサイクルを回していくことで、従来の受動的なデータマネジメントから、利用者参加型の能動的なデータマネジメント（Active Metadata Management※10）へ進めていくことが、生成AI時代のデータマネジメントの姿であると当社は考えています。

※14 DMBOK ：データマネジメントに関する知識を体系立ててまとめた書籍・フレームワーク。DMBOKでは、メタデータの種類としてビジネスメタデータ、テクニカルメタデータ、オペレーションメタデータの3種類が定義されている。

図6.5　生成AI前提でのメタデータ運用



当社では、お客様のデータマネジメント支援を提供しています。生成AIを含む最新トレンドを踏まえて、現状のアセスメントからコンサル施策立案のコンサルティング、また計画した施策の伴走支援までお客様のデータ活用の成功をサポートしています。

Chapter. 7

おわりに

本ホワイトペーパーでは、データ活用におけるデータ提供者、データ活用者それぞれの課題を説明した上で、解決方法として、Text2SQLによるSQL自動生成、精度向上のためのメタデータ整備とセマンティックレイヤー導入に関して紹介させて頂きました。

また発展的な話題として、データ活用領域に特化したAIエージェントや、これらAIによる業務効率化や利便性向上をインセンティブとした、利用者参加型の能動的なデータマネジメント方法（Active Metadata Management※10)）に関してもご紹介させて頂きました。

NTT DATAでは生成AIやセマンティックレイヤー、データ活用に特化したAIエージェント、生成AIを前提としたデータマネジメントの導入事例や支援実績が多数あります。

これらデータ活用に関してお困りごとやご相談などがありましたら、是非ご相談・ご連絡ください。

問い合わせ先

NTTデータグループ　技術革新統括本部　Apps&Data技術部
Mail: dioffering-contact@kits.nttdata.co.jp

連絡先

濱方 大伸　ゼネラルマネージャー

データ源泉システムからのデータ収集・蓄積、大規模データ分散処理、データプリパレーション、データ提供、データマネジメント、データガバナンスと、広範囲にわたりコンサルティングから実装まで広範囲な知見を活用し、お客様のビジネスを推進。

八木　香充　マネージャー

大手メガバンク子会社の大規模更改開発に従事した後、R&D部門にて開発チームリーダーとして従事。
現在は、データサイエンス・アジャイル・AIガバナンスの専門性を活かして多くのデータ活用案件をリードしている。