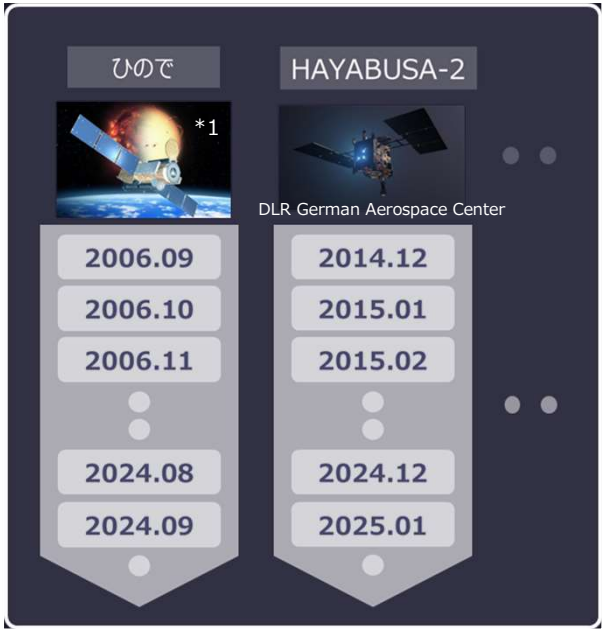
ビッグアーカイブのRAG: BA-RAG を実現する Peta Book



*1. https://www.jaxa.jp/projects/sas/solar_b/images/solar_b_main_001.jpg

*2. HK: House Keeping

*3. 数兆レコード：電話会社の通話記録、巨大IoTでの例が知られる

*4. 数万カラム：半導体製造装置、最近の宇宙機など

*5. RAG: Retrieval Augmented Generation, 検索拡張生成

1. ビッグアーカイブ（BA）とは、日々蓄積されることで巨大になる表形式データです。例えば、宇宙機のHK*2 Data（左図）が挙げられます。

こうしたBAは現代社会のあらゆる活動を記録する貴重な情報資産です。そのため、もしLLMが柔軟かつリアルタイムにBAにアクセスできたら、科学技術分野に限らず、製造、流通など幅広い分野で画期的なイノベーションを生み出すことができます。

しかし、BAは全体として数兆レコード*3あったり、数万カラム*4 あったりする巨大な表形式データです。加えて、コストや運用上の制約から、通常、多数のファイルに分割され、低コスト環境に保持されています。それらの結果、BAへのアクセスは非常に厄介になっています。

2. 3つの課題

課題1. 目的別BAの生成

多数のファイルを統合し、使用目的に沿った目的別BAを作る必要があります。目的別BAは、多数のファイルの結合、必要なカラムの抽出、変換（単位の変換など）を行って作るため、生成に長時間を要します。

課題2. 目的別BAへの高速アクセス

作った目的別BAにはそのままではインデックスがありません。巨大な目的別BAは例え高速なクラウド環境上に置いたとしてもリアルタイムのアクセスはできません。

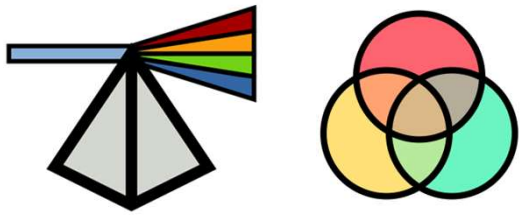
課題3. 目的別BAの保管・転送

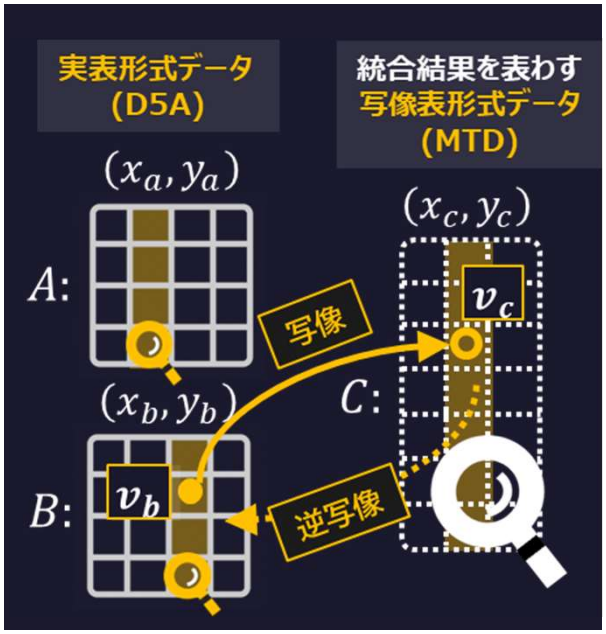
巨大な目的別BAは保管することも転送することも容易ではありません。保管ができなければ継続利用が難しく、転送ができなければ共同作業が難しくなります。

3. FACT を提供する BA-RAG

上記の3課題は自然数インデックス（NNI）を用いることで解決できます。その結果、実現可能になるのがLLMがリアルタイムアクセスできるBAである **BA-RAG** です。

BA-RAGは、既存のRAGと異なり、埋め込みベクトルによる「知識」ではなく、そのBAに蓄積された正確かつ詳細な「FACT（例：製品の販売記録）」を提供します。





4. 自然数インデックス(NNI)

すべての表形式データは、値の情報(文字列や数値など)と位置の情報(自然数)で構成されています。前者は1つ、後者は異なる特性を持った複数の成分に一意に分解することができます。

位置の成分はさまざまに組み合わせることが可能で、多様な自然数ならではの性質を利用した、高速化アルゴリズム群を生成できます。この成分の構成は全てのカラムで共通ですから、これらの高速化アルゴリズム群は全てのカラムで利用できます。同様に全てのカラムの組合せ、全ての部分集合に対しても高速化アルゴリズムが存在します。

これらの高速化アルゴリズム群はインデックスと見なせます。このように自然数の性質を使って多様かつ幅広い高速化を達成するアプローチが自然数インデックス(NNI)です。

5. D5A[*1]、実表形式データ

NNIを使って多数に分割されたBAの個々のファイルを保持するフォーマットがD5Aです。また、そのファイルフォーマットで保持された実表形式データもD5Aと呼びます。

D5Aは **値とインデックスのオリジン** です。

6. 写像表形式データ、MTD[*2]

写像表形式データ(MTD)は、ソース表形式データ(D5Aまたは他のMTD)からの写像で定義される表形式データです。

(遅延転送)

値が必要になった際に、逆写像を辿ってD5Aから値を取得します。これを遅延転送と呼びます。遅延転送により、リアルタイムに目的別BAを生成でき、課題1を解決できます。加えて何兆レコードのデータでも滑らかに表示できるようになります。

(インデックス継承)

値と同様に、インデックスもD5Aから継承できます。これにより、MTDは生成と同時に全ての場所にインデックスを備え、課題2を解決できます。

(ポータビリティ)

値もインデックスも持たないので極めてコンパクトなファイルになります。保管や転送が容易になり、課題3を解決できます。

このようなMTDは あらゆる場所にインデックスがある、**目的別BA** であると言えます。

1. D5A: 第5世代アーカイブ
2. MTD: Mapped Table formatted Data

7. 分散テーブルファイル機構

値とインデックスのオリジンであるD5Aファイル群と、あらゆる場所にインデックスがある目的別BAのMTD群を組み合わせると、フルインデックスな分散テーブルファイル機構を構築することができます。

この分散テーブルファイル機構は前記3つの課題を解決できる、一種のデータベースシステムと見なすことができます。そこで、既存のシステム(Big Query)との長短を考えてみましょう。

- **動作環境**：ファイルシステムなのでPCや小規模システムでも構いません。
- **レスポンス**：あらゆる場所でインデックスを使用できることから、多くの場合、リアルタイムになります。このレスポンスはクラウド環境でも実現困難だったものです。
- **コスト**：低くなります。
- **可能な処理**：インデックスで高速化できることと、その応用処理に限られます。

このような分散テーブルファイル機構は、LLMからの多様なクエリーにリアルタイムに応答しBAへのフリーアクセスを可能にします。

8. Peta Book

Peta Book(PB)とは、分散テーブルファイル機構に以下の機能群を加えたものです。

1. メタ情報管理
2. データの表示・可視化
3. アカウント・セキュリティ管理
4. 業務に適合させるためのアプリ

このようなPBは BA-RAGを具現化したソフトウェアと言うことができます。

9. NNIアライアンス

NNIアライアンス(JAXAの共同研究会、NNIテクノロジーズ、株式会社セック、株式会社エー・スター・クォンタム)は、PBの推進を通じて、LLMの可能性を異次元の領域に引き上げることに挑戦しています。

	BigQuery	実＋写像 表形式データ
仕組み	分散並列コンピューティング	フルインデックスな分散テーブルファイル機構
動作環境	クラウド上（限定的な実行環境）	ファイルシステム上（幅広い実行環境）
可能な処理	幅広い	インデックスで加速可能な処理とその応用に限定
コスト	高い（多数のCPUを使う）	低い（SSD上に格納するだけ）
レスポンス	多くの処理が何秒、それ以上	多くの処理が1秒以下、しばしばミリ秒単位
データ更新	可能	追記のみに限定
データ統合（結合・抽出・変換）	長時間	リアルタイム
データの保管・転送	長時間	リアルタイム