100年間利用できる ビッグデータのアーカイブ技術の取り組みを紹介

rev. 2

古庄 晋二(ESS)・山本 幸生(JAXA)・

飯沢 篤志(RICOH IT Solutions)・長尾 正(Layman's Admin)・

生座本 義勝(ESS)・小林 正英(ESS)・

早部 秀一(ESS)・佐藤 悠(ESS)

感動するビッグデータ

- Apollo計画で設置された月の地震計データは月を見る度に「あの場所の地震計で観測されたのだ」と思う由緒正しいビッグデータである。
- どれほどの努力と費用の上に得られたデータであることか！

失われるビッグデータ

- しかし月の地震計データも米国での保管期間に一部が失われた[1]。
- 月の地震計データはその後、宇宙科学研究所が多大の労力と時間をかけ引き取り保全を図ると共に、新たな活用を推進し、多くの知見を生み出し続けている[1]。
- しかし、保全措置にかかっていない危険な状況にあるビッグデータがたくさんある筈である。

[1]アポロ月地震データ公開システムの開発, 宇宙科学情報解析論文誌第１号　121-131 ページ

宇宙科学情報解析論文誌第１号
「アポロ月地震データ公開システムの開発」

概　要

1969 年から 1977 年にかけて NASA のアポロミッションで得られた月地震データは地球以外の天体で得られた最初の地震記録である．このデータは，取得以来 40 年経った現在でも解析が続けられており月の地球物理学研究においい て主要な役割を果たしている．一方で，得られた月地震データセットの全てが，現在のデータ公開機関でアーカイブ され，公開されているわけではない．また，多くの公開データのフォーマットが一般の地震学で使用されるものと異 なるため，現状，ユーザーが必要なデータと情報を取得し，解析研究を行うのに敷居の高さを伴っている．そこで， 本研究では，これまでよりも容易にユーザーが要求する月地震データとそのメタデータを取り出し，解析に供するこ とができる Apollo 月地震データ公開システムを開発した．この開発のため，まず我々はほとんど全ての月地震データ のアーカイブとデータ解析に必要となる情報の収集と整理を行った．そして，デコードしたデータから構成されるリ レーショナル型データベースとデータベースへアクセスするアプリケーションを開発し，Web 上でユーザーが要求す る月地震データを検索して取得できるようにした．本研究で開発した公開システムを通して，より多くのユーザーが 月地震データにアクセスできるようになり，解析研究を通して，月惑星科学を更に進展させていくことが期待される．

「アポロ月地震データ公開システムの開発」からの教訓

メリット

データ公開によって、そこから多くの成果がある。

課題

1. 一方、システム開発に多大の労力と時間がかかった。
2. また、大量のデータ取得のパフォーマンスが不足しがちであった。

1, 2 は一般的課題

失われる理由 ≒ 使われない

- 使われないと、
  1. ベネフィットがDOWNし、
  2. コスト　＞　ベネフィット　になり、
  3. 保全されにくくなる
- 使われないと、
  1. LOSTに気づけない
  2. BackUpがおろそかになる

使われない理由

- 公開されていないから
  - （公開により利益が得られない、不利益がある）
  - 公開システム構築・維持に手間がかかる
  - 公開システムのコストがかかる
- 利用しにくいから
  - Webサーバから公開する
  - ファイル配布で公開する

Webサーバから公開するとき

1. システムの構築・維持コストが高い

   - 個人・小組織が自分のHPで公開しにくい　・・・

     従って公開されるデータに制限ができる

2. データの利用方法が固定化される

3. 異なるシステム間でのデータの組合せが困難

誰もが手軽に公開できるとは言いがたい。
利用方法も固定的。

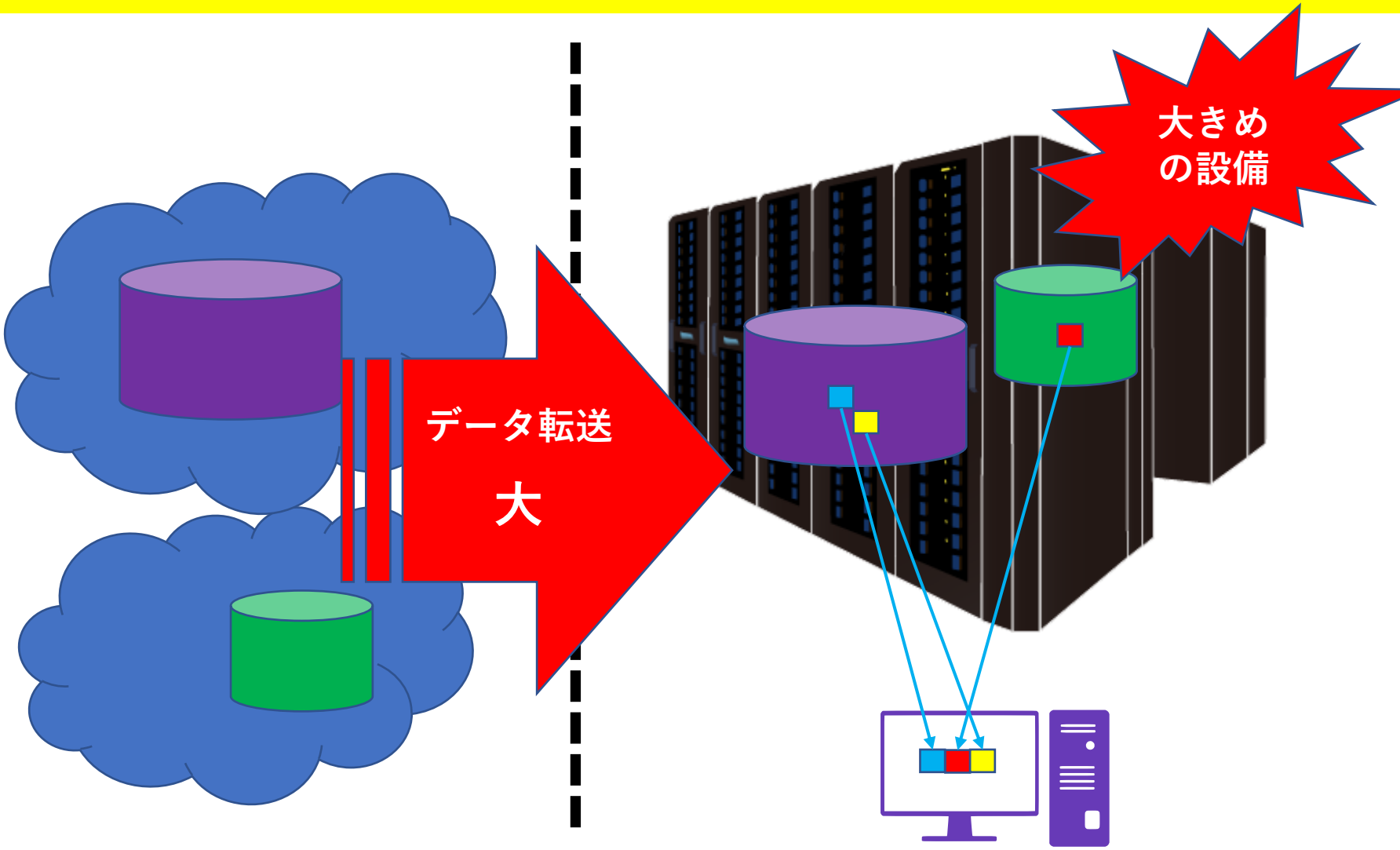
ファイル配布で公開するとき

1. 少数の大きいファイルで公開するとき
   - ダウンロードに時間が掛かる、利用時のデータの加工編集が大変
2. 多数の小さいファイルで公開するとき
   - 多数のファイルの維持管理の手間、ダウンロードが面倒
3. **ユーザ側に負担が大きい**
   - 解析処理　・・・　小さなPCで可能でも、
   - データの加工・編集処理は　・・・　大きめの設備が必要

利用者は
やりたくない！

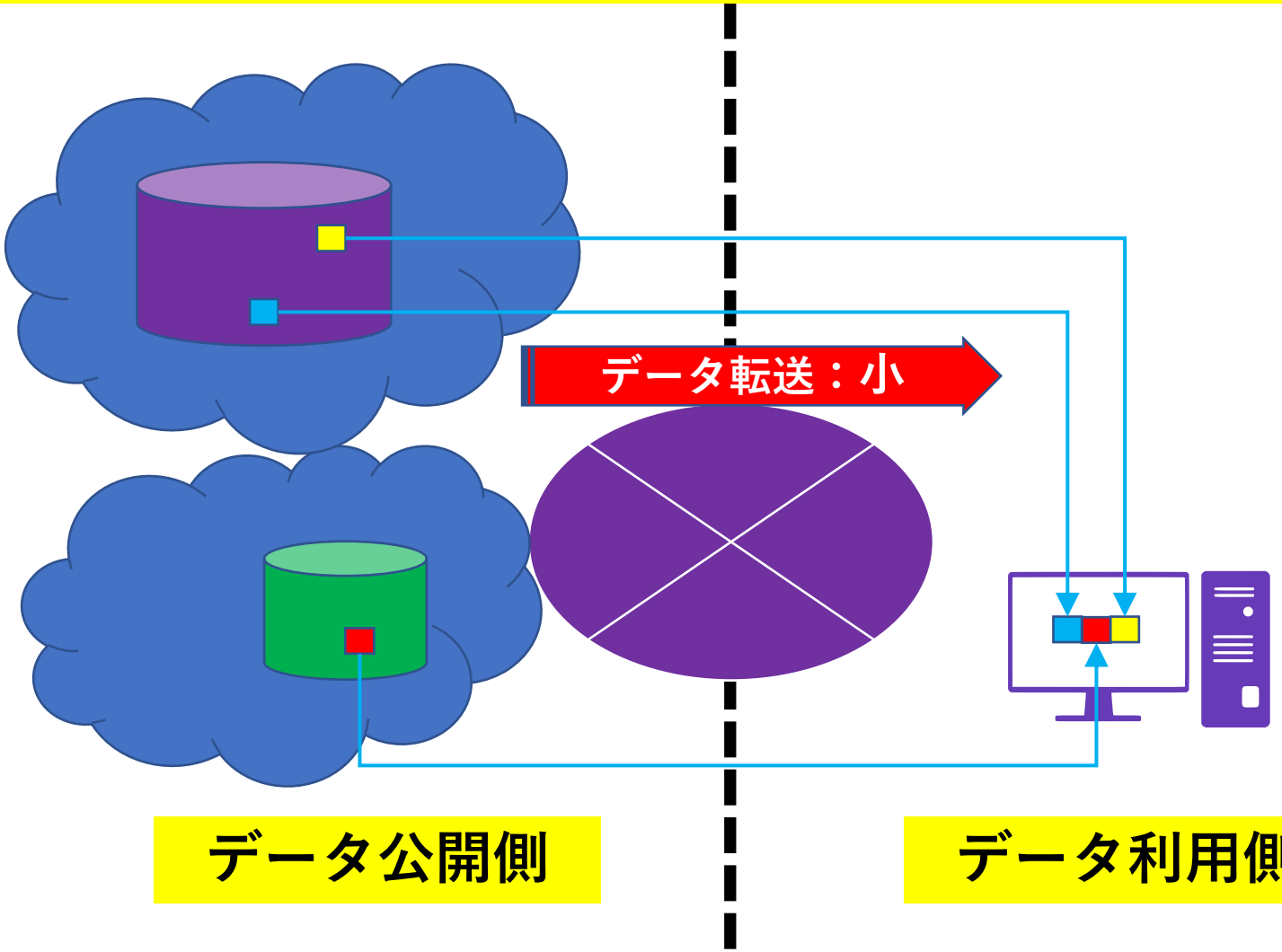
利用に手間と時間と大きめの設備が必要

ファイル配布で公開時の利用形態例



データの加工・編集処理を

ネットワーク上で行えれば解決できる



ネットワーク上での
データの加工・編集処理とは？？？

1. 大きい表形式データを

   最大１兆レコード、１０万カラム

2. どのカラムでも等しく検索・集計・ソート可能

   ⇒　予め使い方、データを指定する必要が無い

3. 多数を組み合わせても（UNION、JOIN）使える

   どのカラムでも等しく検索・集計・ソート可能

1．どのカラムでも、
2．組み合わせても、
検索・集計・ソート出来る理由

1. どのカラムも同じ構造である

   データを自然数にマッピングする、自然数インデックス

2. 多数を組み合わせても（UNION、JOIN）

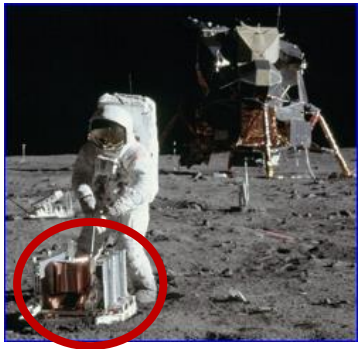
   同義の構造を再生成できるか、

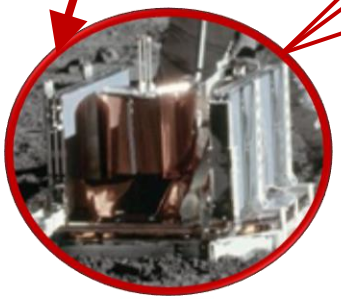
   もしくは容易に複数の処理結果をまとめることが出来る

デモ

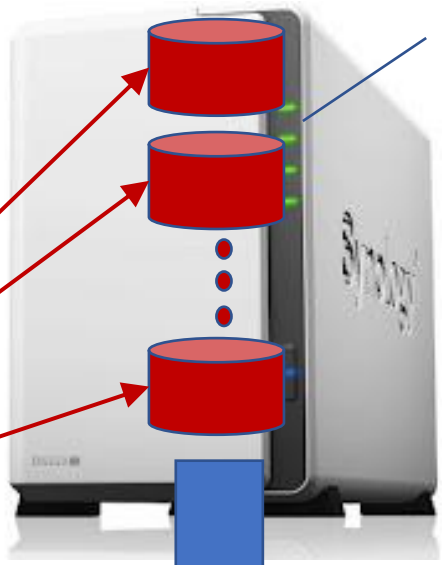
デモムービーの実行環境

Apollo 11～17号は
月面に地震計を設
置してきました。





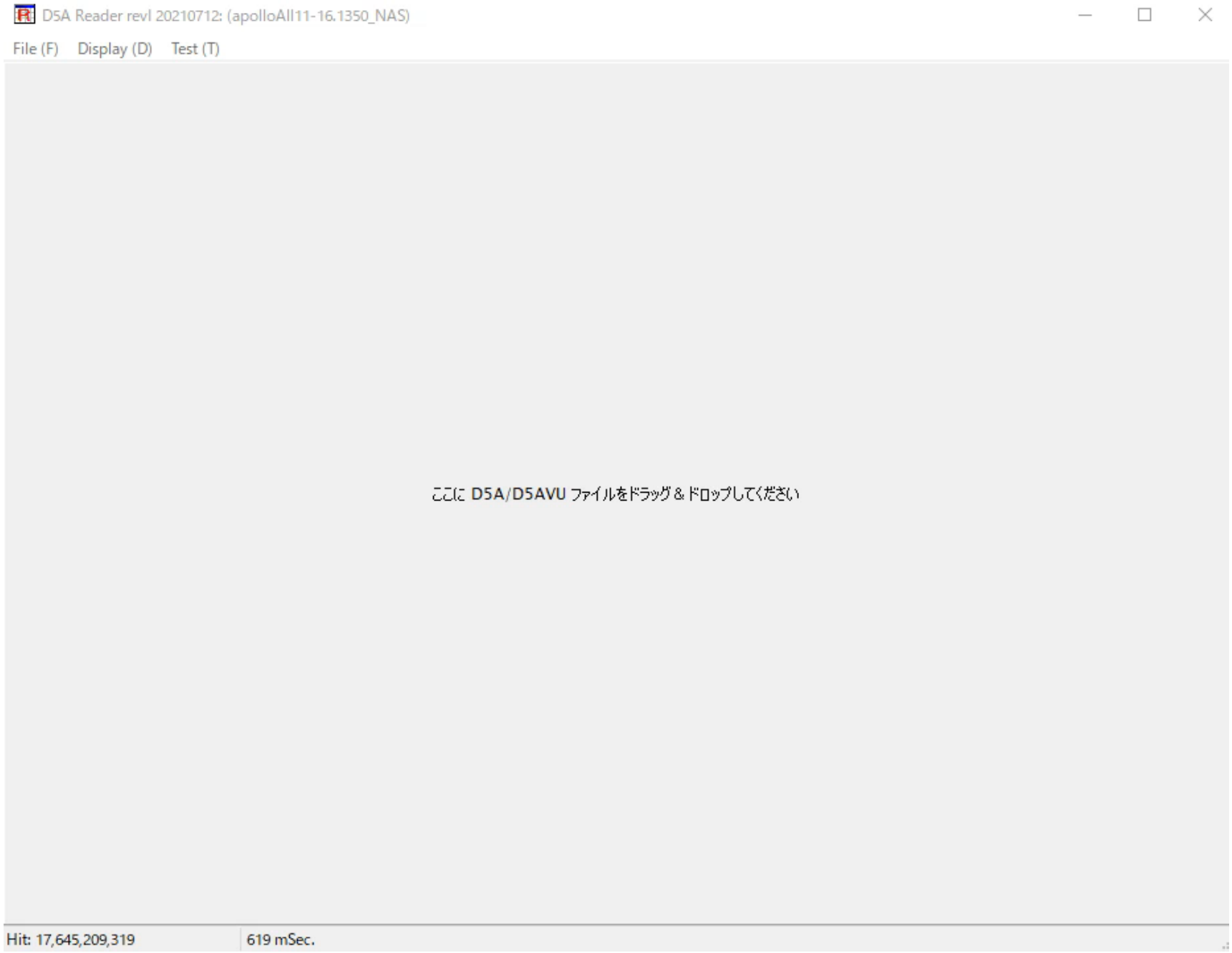
月面設置の
地震計データ



SSDを搭載したNAS上に、
カラム構造の異なる**40個**の
データファイル(D5A)

LAN (1Gbps)

全1350億レコードの
検索とソート


D5A Reader revl 20210712: (apolloAll11-16.1350_NAS)
File (F)
Display (D)
Test (T)
ここに D5A/D5AVU ファイルをドラッグ＆ドロップしてください
Hit: 17,645,209,319
619 mSec.

開発の歴史

これまでの経緯

	開発機能	論文・特許
2011 … 2018	検索・集計・ソート 仮想UNION	DEIM2018 特許出願4件
2019	ファイル圧縮	- 特許出願1件
2020	座標検索	宇宙科学情報解析論文誌10号 特許出願1件
2021	JOIN (COMBINE)、他	宇宙科学情報解析論文誌11号 特許出願3件
2022	**階層化**	DEIM2022 (TOD,宇宙科学情報解析論文誌12号,...) 特許出願1件

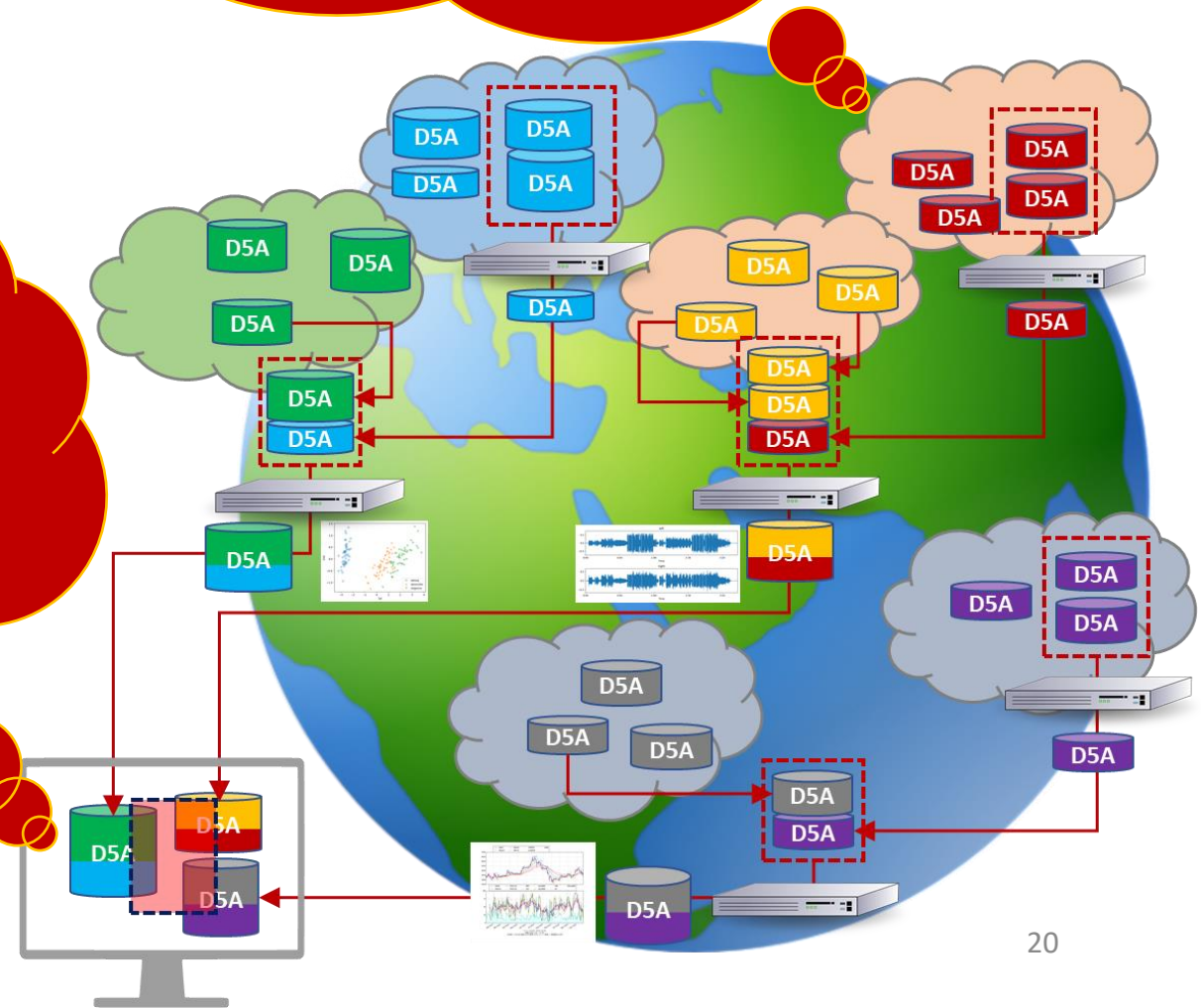
JAXAとの共同研究
(巨大時系列データの高速アクセスに関する共同研究)

これから

D5A で

表形式ビッグデータを
D5Aファイルにして
ファイルサーバに格納するだけ

世界のD5Aファイルを
自由に組み合わせ、
検索・集計・ソートし
て必要部分を特定、
直ちにダウンロード



開発スケジュール

- ２０２２年の末を目処に、宇宙科学関係の希望者がD5Aデータを試用できる環境を整えたい。

最後に

- Apollo計画で設置された月の地震計データは月を見る度に「あの場所の地震計で観測されたのだ」と思う由緒正しいビッグデータである。
- どれほどの英知、努力と費用の上に得られたデータであることか！
- そんなビッグデータは工学分野にも強い開発の動機付けを与えてくれた。
- 我々の取り組みが、１００年後の誰かがそのようなビッグデータを活用してくれることにつながるかもしれないのが楽しみ。