

令和 5 年 6 月 10 日現在

機関番号：62616

研究種目：基盤研究(C)（一般）

研究期間：2020～2022

課題番号：20K04030

研究課題名（和文）テラバイト超データ高速可視化システム等によるALMAデータ検索機能の強化

研究課題名（英文）Enhancement of the ALMA data search system with an ultra-high-speed data visualization system that can handle TB-scale data.

研究代表者

白崎 裕治（SHIRASAKI, Yuji）

国立天文台・天文データセンター・助教

研究者番号：70322667

交付決定額（研究期間全体）：（直接経費） 3,300,000円

研究成果の概要（和文）：ALMA等により取得される多次元データは巨大化の一途をたどっており、2023年時点でファイルあたり300GBを越えるようになってきている。近い将来には1TBも超えることが想定されており、そのような巨大なデータをインターネット経由で利用するための仕組みづくりが急務となっている。本研究ではALMAのデータを複数のサーバに分割配置することにより、必要な空間領域におけるスペクトルデータを高速に抽出しユーザへ提供する機能を開発した。本機能はJVO(<http://jvo.nao.ac.jp/portal>)におけるALMAデータクイックルックシステム(FITS WebQL)として誰でも利用可能となっている。

研究成果の学術的意義や社会的意義

ALMAの高品質で巨大なデータを手元に全データをダウンロードすることなく、必要な情報のみへ高速にアクセスすることが可能となり、アーカイブデータを効率的に活用することが可能となった。これによりALMAのデータを利用した科学的成果の最大化が期待される。また、今後の巨大なデータを生み出す観測計画への適用も可能である。今回開発したシステムは研究者のみならず、だれでも利用が可能であり学校教育においても最新の天文観測データを利用した教育活動に利用することができるであろう。

研究成果の概要（英文）：Multidimensional data acquired by ALMA continues to grow in size, and exceed 300 GB as of 2023. It is expected to exceed 1TB in the near future, and there is an urgent need to create a system for using such huge data via the Internet. In this research, we developed a data visualization to extract spectral data in a specified spatial region at high speed by dividing and arranging ALMA data on multiple servers and provide it to users. This function is available to anyone as the ALMA data quick look system (FITS WebQL) on the JVO portal site (<http://jvo.nao.ac.jp/portal>).

研究分野：天文データアーカイブ

キーワード：天文データアーカイブ 分散処理 データベース 可視化 大規模データ

科研費による研究は、研究者の自覚と責任において実施するものです。そのため、研究の実施や研究成果の公表等については、国の要請等に基づくものではなく、その研究成果に関する見解や責任は、研究者個人に帰属します。

様式 C-19、F-19-1、Z-19（共通）

1 研究開始当初の背景

ALMA 等により取得される多次元データは巨大化の一途をたどっており、研究開始当初の 2020 年 4 月時点では総量 130 TB であったものが、2023 年 5 月時点では 326 TB とほぼ倍増している。また、最大ファイルサイズも 129 GB から 327 GB およそ 2.5 倍となっている。20GB 以上のファイル数で見ても 1063 から 3145 とほぼ 3 倍に増加しており、研究開始当初懸念されたように大規模なデータをいかに効率よく利用できるようにするかが、データアーカイブ開発にとっての喫緊の課題となっている。

2 研究の目的

そうした課題を克服するため、巨大なデータを手元にダウンロードせずともデータの中身を可視化し、自分の研究に必要なデータを探し出す機能の開発を行った。開発のゴールとしては 1TB のデータでも Web ブラウザ上でインタラクティブにデータの中身を大きな遅延なく表示することができることとした。また、多くのデータでは天体が検出されておらず、そうしたデータを極力排除することにより効率よく利用可能なデータを探せることも目的とした。

3 研究の方法

巨大なデータを高速で可視化するために、一つのデータを分割し複数の計算機に分散配置することにより、ユーザーからのリクエストに応じて分散並列処理により一つの画像またはスペクトルデータを再構築する機能を実装した。巨大なメモリと多数のコアを搭載可能な高性能計算機を利用するという方法では、そうした高価な計算機を購入する経費が常にあるとはかぎらず、予算が大幅削減された場合にはその機能を継続して提供することは困難となる。一方、安価な汎用計算機であればその時の予算状況に応じて少しずつ増強なりリプレースなりができるので継続したサービスの提供をしやすくなるといったメリットがある。そのため、複数の計算機で分散処理することによりデータをオンデマンドで再構築する仕組みを開発した。

天体が検出されていないデータを排除できるように ALMA の全データについて天体検出ならびにエミッションラインの検出を行い、その結果をデータベース化することにより検出された天体情報によってデータを検索することができるようにした。

4 研究成果

今回開発したシステムは 5 台の計算機からなり、各計算機にはそれぞれ 8 台の 8TB HDD と 4 台の 2TB NVMe SSD が搭載されている。8 台の HDD は RAID ポリュームとして一つにまとめられ、5 台の計算機の HDD ポリュームから汎用分散ファイルシステムである GlusterFS を構成する。この領域にファイルサイズが 20GB 以上の ALMA のデータを配置した。さらに、40GB 以上のデータについては 4 台の SSD を RAID0 によりまとめた領域に配置した。ユーザからのアクセスがある度に要求された画像とスペクトルデータを分散配置されているデータから再構成し提供する。なお、20GB 未満のデータについては従来どおり一台の計算機上で動作するシステムによりデータ表示を行う。以上のように実装された可視化システムにより、SSD 上のデータについてはほぼ瞬時にユーザのブラウザ上に表示されることが確認された。また、1TB のデータでもスムーズに表示されることを確認した。HDD 上のデータについては数秒程度の遅延があるが、十分な速度で表示がされることを確認している。本機能は JVO ポータルサイト (<http://jvo.nao.ac.jp/portal>) における ALMA データクイック検索システム (FITS WebQL v5) として誰でも利用可能となっている (図 1)。

データ表示機能に加えて検出されているラインの数などによりデータを検索する機能が追加された。この機能により、ラインが検出されているデータを効率的に選択することができるようになっている (図 2)。

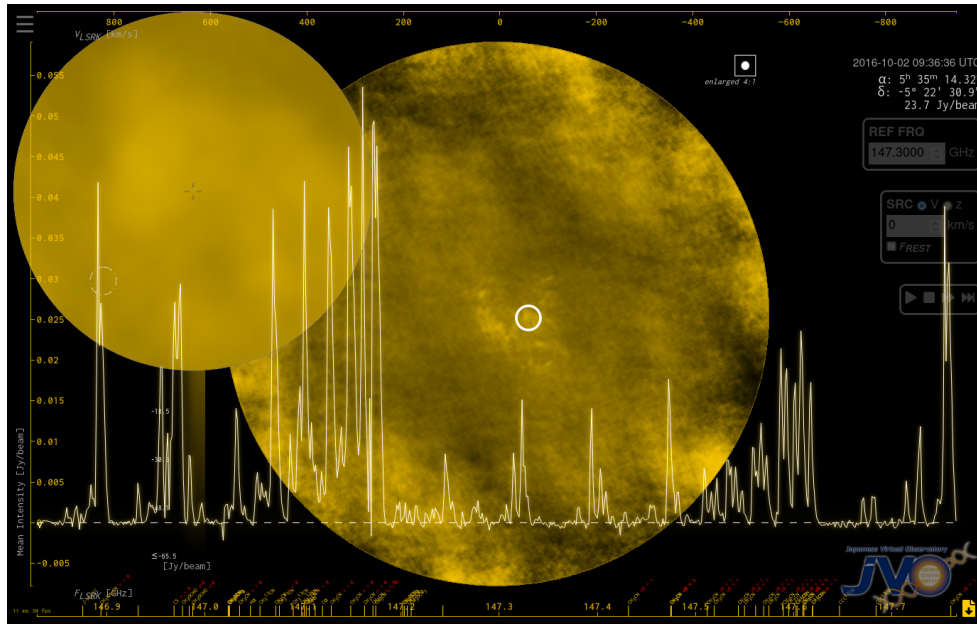


図 1: FITS WebQL v5 による ALMA データの画像とスペクトルデータの表示機能。白丸で示される領域のスペクトルデータをリアルタイムで計算し表示する。

Search

of lines per spectrum >= & <
 f_LSRK [GHz] = ±

(*) Please note that search by the rest-frame frequency properly works for data which contains enough information for determining the source velocity (or redshift). We rely on the FITS keyword "RESTFRQ" which is assumed to be set to the middle of the frequency range in the source rest-frame. For some of data source velocity/redshift retrieved from the SIMBAD database was used. The objects were selected by proximity (< 20 arcsec) to the detected sources.

f_rest [GHz] = ±

(*) Please note that "search by the species" below is provided for searching data in which emission or absorption lines are detected in the frequency range expected for given species. This doesn't guaranty that the detected lines are originated from the specified species. The result relies on the accuracy of the FITS keyword "RESTFRQ" or selection of SIMBAD objects.

species include all the checked (and input) species
 (popular) 13CN 13CO 13CS C17O C18O C34S CCH CN CO CS DCN DCO H13CO
 H13CO H2CO HC3N HCN HCO HDO HN13C HNC N2H SiO SO
 (atom) H He C N O
 (diatomic) AlCl AlF AlO CF CH CN CO CP CS FeO HD HCl HF OH KCl NH
 O2 PN PO SH SO SiC SiN SiO SiS TiO
 (Triatomic) AlNC CCH C2S C2P H3 CH2 H2Cl H2O H2S HCN HCO HCP HCS KCN
 MgCN NH2 N2O NaCN OCS SO2 SiC2 SiCN TiO2
 (Four atoms) I-C3H C3N H3O H2CN H2CO H2CS HCCN HOCO HCNO HOCN HNCO HNCS
 NH3 HNCS c-SiC3 HMgNC
 (Five atoms) C3H2 H2C3 CH2CN H2C2O H2COH C4H HC3N HCOOH NH2CN CNCHO
 SiC4
 (Six atoms) H2C3O HNCHCN CH3CN CH3OH CH3SH HC3NH NH2CHO C5H C5N HC2CHO
 HC4N H2CCNH C5S
 (Seven atoms) H2COCH2 CH3CCH CH3NH2 CH2CHCN H2CHCOH C6H HC5O CH3CHO CH3NCO
 (Eight atoms) HCCC2CN CH2OHCHO CH3OCHO CH2CHCHO H2CCCHCN C7H H2NCH2CN
 (Nine atoms) CH3C4H CH3OCH3 CH3CH2CN CH3CH2OH C8H HC7N
 c-C6H5CN
 (free text) =
 match score >=
 match fraction >=

図 2: JVO ALMA FITS Archive のライン情報にもとづく検索画面。検出されているラインの情報によりデータを検索することができるようになっている。

5. 主な発表論文等

〔雑誌論文〕 計0件

〔学会発表〕 計5件（うち招待講演 0件 / うち国際学会 1件）

1. 発表者名 Christopher Zapart, Yuji Shirasaki, Masatoshi Ohishi, Yoshihiko Mizumoto, Wataru Kawasaki, George Kosugi, Eisuke Morita, Akira Yoshino, Youhei Hayashi
2. 発表標題 Julia meets BIG DATA: JVO experience with distributed computing
3. 学会等名 Astronomical Data Analysis Software and Systems (ADASS) XXXI (国際学会)
4. 発表年 2021年

1. 発表者名 ザパート クリストファー、白崎裕治、大石雅寿、水本好彦、川崎渉、小杉城治、森田英輔、吉野彰、林洋平
2. 発表標題 FITSWEBQLSE とその周辺(C/C++, FORTRAN, Rust, Julia, Python 等)
3. 学会等名 日本天文学会2022年春季年会
4. 発表年 2022年

1. 発表者名 白崎裕治、ザパート クリストファー、大石雅寿、水本好彦
2. 発表標題 JVO updates
3. 学会等名 ALMA/45m/ASTE Users Meeting 2021
4. 発表年 2021年

1. 発表者名 Christopher Zapart、白崎裕治、大石雅寿、水本好彦、川崎渉、小杉城治、森田英輔、吉野彰、林洋平
2. 発表標題 FITSWebQL SE (Supercomputer Edition)の開発状況
3. 学会等名 日本天文学会2020年秋季年会
4. 発表年 2020年

1. 発表者名 Christopher Zapart、白崎裕治、大石雅寿、水本好彦、川崎涉、小杉城治、森田英輔、吉野彰、林洋平
2. 発表標題 分散クラスター対応の FITS WebQL SE v5 の試験運用開始
3. 学会等名 日本天文学会2023年春季年会
4. 発表年 2023年

〔図書〕 計0件

〔産業財産権〕

〔その他〕

-

6. 研究組織

	氏名 (ローマ字氏名) (研究者番号)	所属研究機関・部局・職 (機関番号)	備考
研究協力者	ザパート クリstoffァー (Zapart Christopher)		
研究協力者	大石 雅寿 (Ohishi Masatoshi)		
研究協力者	小杉 城治 (Kosugi George)		

7. 科研費を使用して開催した国際研究集会

〔国際研究集会〕 計0件

8. 本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関
---------	---------