PILOT DATA ANALYSIS OF PDS GEOSCIENCES DATA IN THE CLOUD WITH THE JAVA CRISM ANALYSIS TOOL (JCAT). D. V. Polittе[1], R. E. Arvidson[1], T. C. Stein[1], L. E. Arvidson[1], [1]Department of Earth and Planetary Sciences, Washington University in Saint Louis, Saint Louis, Missouri, 63130, dvpolitte@wustl.edu.

Introduction: At the Geosciences Node of NASA's Planetary Data System (PDS), we are developing the infrastructure and tools that will allow planetary data users to perform data analysis on our holdings in a cloud-computing setting. We describe here our pilot project, which demonstrates this capability with a cloud-based analysis of data from the Compact Reconnaissance Imaging Spectrometer for Mars (CRISM) hyperspectral instrument using the Java CRISM Analysis Tool (JCAT) available from JHU/APL (crism.jhuapl.edu/JCAT) and the Geosciences Node website. The techniques and tools used in this project, as well as the rationale for using the cloud, are applicable to many other PDS datasets and the tools for analyzing them.

Rationale: There are multiple benefits in using a cloud-based data analysis workflow. Primarily, the user does not need to have as expensive a computer. Because the analysis software is running within a cloud-based data center, the user's own computer does not need to have enough storage space to hold the dataset they are using. It also does not need to have a powerful enough CPU and enough RAM to run the software, which for some software requires that the user have access to a workstation-class desktop computer.

This workflow also allows for acquiring pre-analysis data more quickly. The PDS Geosciences Node keeps a cloud copy of our archives that serve as a backup of all Geosciences Node holdings. The user can select to have their virtual machine (VM) they use for analysis in the same cloud data center. In this case, data transfer from this backup copy to the user's VM will likely be faster than downloading the data to their local computer from the main PDS Geosciences Node archive (pds-geosciences.wustl.edu). This dynamic is illustrated in Figure 3.

The co-location of archives and analysis also avoids data egress fees (i.e., a charge incurred upon download), which are often a significant portion of costs of data storage on the cloud. In general, this is financially advantageous to the provider of the data. (However, the PDS Geosciences Node benefits from an institutional "egress waiver" with Azure that already eliminates these fees, except under exceptional circumstances.)

For many data analysis projects, the required cloud computing is also inexpensive. For example, the VM used in this experiment was a Microsoft Azure Standard_D2s_v3 configuration, with 2 virtual CPUs, 8 GB of RAM, and about 100 GB of available hard drive space, which costs $0.2110/hour. (A Windows license is included in this price.) The analysis described below could be applied to 10 CRISM scenes of interest in approximately 7 hours, resulting in charges of $1.48. After our analysis was complete, we minimized recurring charges by stopping and deallocating our VM, resulting in continued billing only for the use of the virtual hard drive (about $20/month, as opposed to about $174/month without deallocation). Exact prices for other users will vary with the computing power required and institutional agreements with the cloud provider.

Methodology: For this pilot project we used a Windows Server 2019 VM, with virtual hardware as described above, in the Azure Central US zone, so that it would be co-located with the PDS Geosciences archive cloud copy (stored as Azure BlobStorage).

We transferred the installers for the non-OS software we needed to the VM using our remote desktop client's "local resources" feature, which allowed us to access local files on the remote machine. We installed Python, the Java Runtime Environment, and JCAT itself in this way.

We next used the Geosciences Node's Mars Orbital Data Explorer (ode.rsl.wustl.edu) to identify the CRISM products that we needed for our analysis. We used the ODE's search functionality to find the files. Then we copied these files' download links (which could be used to download them from the main copy of our archives). On our VM, we provided these links as input to a Python script that used them to locate & download the files from the archive cloud copy. We then performed the analysis using JCAT.

To illustrate this proof-of-concept, we chose to analyze CRISM scene HRL000040FF, which covers the western side of Jezero Crater, including the Mars 2020 rover mission (Perseverance) landing error ellipse, the delta, and the carbonate-bearing terrains up against the western Jezero Crater wall [1]. Figure 1 shows the initial JCAT screen on our virtual workstation with the IOF sensor space displayed for the full scene, together with an enlarged area. The site of the spectral retrieval shown in the plot is denoted by a cross and arrow in the enlargement. We then divided the spectrum by a volcano scan observation to remove the gas bands. The result is shown in Figure 2 with labels to demonstrate that a carbonate-bearing terrain has been identified based on diagnostic 2.3 and 2.5 micrometer absorptions. We exported these results from JCAT as PNG images

and transferred them back to our local machine via our remote desktop client, completing the analysis process.

Future Work: Having constructed a working JCAT cloud compute environment, we plan to improve the process by which data of interest are located and made available for processing. One approach is to simplify the task of locating data within the archive cloud copy by offering a "Download from cloud" capability in the PDS Geoscience's data discovery services, the Orbital Data Explorer and the Analyst's Notebook (an.rsl.wustl.edu). Alternatively, we will explore the possibility of mounting the cloud copy of our archives as a disk on a virtual machine.

We will also consider providing a pre-made VM image for cloud data analysis which has basic enabling software, such as Python and the Java Runtime Environment, already installed. This would make it easier and faster to set up an analysis VM.

We will provide instructions on our website (pds-geosciences.wustl.edu) once this workflow is more refined.

Acknowledgments: Funding for this work, as part of the PDS Geosciences Node's mission, comes from NASA. The CRISM data used can be found in the PDS datasets MRO-M-CRISM-3-RDR-TARGETED-V1.0 and MRO-M-CRISM-6-DDR-V1.0.

References: [1] Goudge T. A. et al. (2015) *JGR Planets, 120,* 775-808.

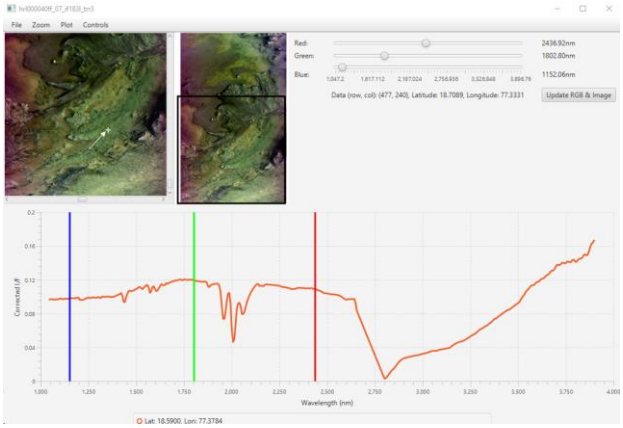

Figure 1: JCAT's initial view of CRISM scene HRL000040FF's data upon loading it. The displayed IOF spectrum was chosen from the candidate carbonate-bearing terrain. Vertical lines show the spectral bands used to display the RGB false color images in the upper left. Carbon dioxide band triplet at ~2 micrometers is evident, along with the deep 3 micrometer band and thermal rise at longer wavelengths. North is toward the bottom of the image views.

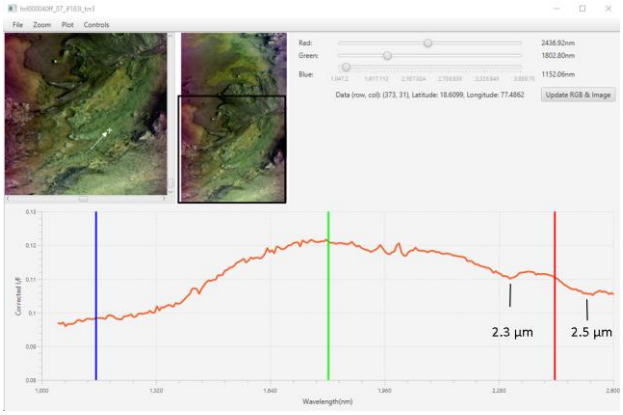

Figure 2: IOF spectrum shown after dividing by a volcano scan to remove gas bands, with the X and Y ranges of the spectral plot configured to better show absorption features at 2.3 and 2.5 micrometers associated with one or more carbonate bearing minerals.

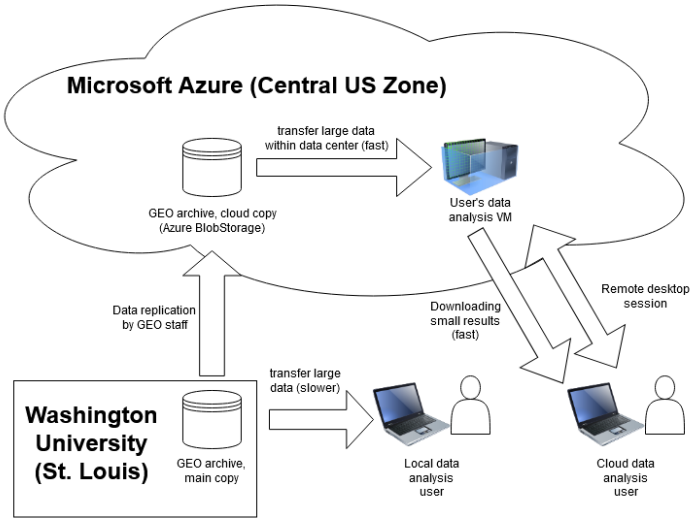

Figure 3: Diagram showing differences between local and cloud data analysis of PDS Geosciences Node (GEO) data. In local analysis, large files must be transferred to the user's local machine across the public internet, potentially leading to large delays. In cloud analysis, however, large files are transferred within a cloud data center's local network and so can be ready for analysis sooner. Only the smaller result files need to be transferred back to the user's local machine.