SpaceOps-2023, ID # 564

Fully automated cloud based science data processing for Emirates Mars Mission

Omran Alhammadi[a]*, Bryan Harter[b] , Mohammad Alfalasi[c] , Ransom Christofferson[d]

[a] Department of Remote Sensing, *Mohammed Bin Rashid Space Centre (MBRSC), Al Khawaneej Area UAE, DUBAI,* orman.alhammadi@mbrsc.ae
[b] *Laboratory for Atmospheric and Space Physics (LASP) at University of Colorado, 1234 Innovation Dr, Boulder, CO 80303, United States* bryan.harter@lasp.colorado.edu
[c] Department of Remote Sensing, *Mohammed Bin Rashid Space Centre (MBRSC), Al Khawaneej Area UAE, DUBAI,* mohammad.alfalasi@mbrsc.ae
[d] *Laboratory for Atmospheric and Space Physics (LASP) at University of Colorado, 1234 Innovation Dr, Boulder, CO 80303, United States* ransom.christofferson@lasp.colorado.edu
* Corresponding Author

Abstract

The science data processing system of the Emirates Mars Mission (EMM) is responsible for generating Quicklook, Level 1, and Level 2 science data products using the processing pipeline software developed by the mission Instruments Team Facility (ITF) and distributing the products to the entire EMM science team and science community. It receives Level 0 science data and ancillary data from the Mission Operation Center (MOC). The science data processing environment is well-architected and hosted at Amazon Web Service (AWS) and make use of different advanced services provided by AWS such as AWS Lambda, AWS Elastic Container Service (ECS) AWS Step Function and AWS Simple Storage Service (S3).

The development of the processing pipeline software takes place on the ITFs' institution. Once the software is of sufficient maturity, it is promoted to the ITF staging area within the science data processing environment. The cloud staging checkout ideally consists of; testing in the cloud with cloud resources (ECS/S3/etc) and packaging the processing software to be tested as a running container in the ITF staging area. After the software validation process completed at the cloud staging environment, the ITF promotes the updated version to the science data processing production platform.

Upon the reception of new Level 0 science and housekeeping files as well as the spice kernels at the specific S3 bucket, the science data processing uses an AWS Lambda function to automatically check the files integrity and trigger the AWS processing pipeline step function. AWS Step Function is utilized to organize the workflow of science processing and production for all instruments scientific data, and reflects the update of newly received Spacecraft and Planet Kernel (SPK) from the MOC. A Lambda function is automatically triggered after the generation of each level of science products to index generated products into the science data processing database. The science data processing environment provides different data access mechanisms to the available science products and ancillary files for the mission science team as well as the science community. A well-developed web application provides an easy graphical user interface to the scientists in order to retrieve the scientific products. The back-end of data access and retrieval environment uses Serverless model. The access or retrieval request is handled by AWS Abstract Programming Interface (API) Gateway in which it triggers an AWS Lambda function to retrieve the metadata of science products from the science processing database or send back the requested products as objects to the scientist.

Keywords: EMM, SDC, AWS, Data processing

1. Introduction

The Emirates Mars Mission (EMM) was launched from the Tanegashima Space Center in Japan in 19th July 2020. The mission is the first UAE's interplanetary spacecraft that is orbiting around the Mars. The probe called Hope (Al Amal, in Arabic) is on-boarded with three scientific instruments to capture a complete picture of the Martian atmosphere and its layers during different times of the day and different for a complete Marian year. The mission is led and developed by Mohammed Bin Rashid Space Center in collaboration with its knowledge transfer partners at the University of Colorado Boulder, Arizona State University, and the University of California, Berkeley. The Hope probe successfully made it Mars Orbit Insertion (MOI) on 9th of February 2021 and started its journey around the Mars. Scientific instruments mounted on one side of the spacecraft will collect information about atmospheric circulation and capture aerial images of Mars using the visible, thermal infrared and ultraviolet

wavelengths. The three instruments are: EXI (Emirates eXploration Imager) will capture high resolution images of Mars, measuring water ice and ozone in the lower atmosphere and capture colorful images of Mars, EMUS (Emirates Mars Ultraviolet Spectrometer) will measure oxygen and carbon monoxide in the thermosphere as well as hydrogen and oxygen in the upper atmosphere, which are essential for determining the loss of water from the upper atmosphere, and EMIRS (Emirates Mars Infrared Spectrometer) will measure both surface and atmospheric temperatures, as well as global distribution of water ice, water vapor and dust in lower atmosphere.

2. EMM science data flow

The data flow diagram (see Fig. 1) illustrates the path that science data follows through the mission systems, from the spacecraft to the science community. Science telemetry is downlinked from the spacecraft to the ground network, and delivered to the Mission Operations Center (MOC). The MOC delivers all science data and any other information required for science data processing to the SDC, where it is indexed and stored for the duration of the mission.

The Instrument Team Facilities (ITFs) are responsible for writing the code which processes Level 0 science data into scientific data products (Level 1, 2, and science Quicklook). The ITFs upload this code and any associated calibration data and/or models needed for processing to the SDC, where it undergoes final testing before being integrated into the SDC systems. The SDC is configured to support frequent update to the processing software during the mission, specially in the early stages of science data collection.

The SDC automatically executes the ITF processing code each time new data is received from the spacecraft. Computing resources required for data processing are provided based on ITF needs as well as complexity of the algorithm. As the processing is completed, the data products are placed into the SDC's storage, along with the file metadata associated with each data product, and any logging information generated by the automated processes. The ITFs will have immediate access to all science products, metadata, and processing logs as they are placed on SDC storage. The EMM science team will also have access to this data, but via the SDC website and web services (scriptable command-line access via https) rather than direct storage access. This access will still be password-protected in order to restrict access to the EMM team only. While the SDC is responsible for producing Level 1 and 2 science data products using ITF-provided code, the Level 3 products will be generated by the ITFs using algorithms from the science team, and uploaded to the SDC by the ITFs.

After a scheduled interval allowing the ITFs to ensure that the instruments have been properly calibrated and the science processing is producing accurate results, the first dataset will be released to the public. This will involve placing links to a subset of the SDC's data storage on a public-facing website. The publicly accessible data storage and website are known as the "Emirates Science Data Center" (ESDC), and will persist past the end of the mission to provide permanent public access to the EMM science data.

As the science team and ITFs analyze the data, they will find issues with their instrument calibrations and processing algorithms; this is unavoidable when working with innovative instrumentation and new science data. This will result in revisions to their processing code, and the need to reprocess previously collected data. The SDC is developed so that it can automatically scale up to support data reprocessing without impacting the daily new data processing.

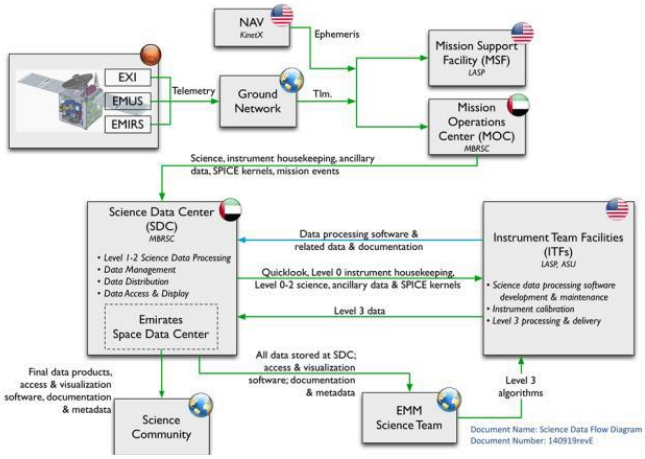

Fig. 1. EMM science data flow diagram

3. EMM Science Data Center (SDC)

The Science Data Center (SDC) is responsible for generating Quicklook, Level 1, and Level 2 science data products, distributing science data to the entire EMM team and the science community, facilitating the discovery and usage of science data, storing all science data products for the duration of the mission, and creating an archive to store data beyond the end of the mission. Standard science products include Level 1 and Level 2 data products that will be disseminated within the EMM team or to the science community. The SDC will receive Level 0 science and ancillary data from the MOC. The ITFs will develop and maintain science data processing software to be run at the SDC on that Level 0 data to generate Quicklook, Level 1, and Level 2 science data products. The science data products that are created at the SDC will be stored by the SDC for the duration of the project, and will be made available to the EMM team and the wider science community within the required timeframes. The SDC will develop and maintain interfaces for accessing science data, and basic data analysis and visualization tools, in close coordination with the EMM science team. These tools will be available to the science community via the SDC website.

3.1 SDC implementation

The SDC fully deployed at Amazon Web Service (AWS) as a native cloud solution utilizing different cloud based managed services. Current SDC design has processes running on "Serverless" architecture, i.e. no permanent virtual machines are needed in order for the SDC to function. Instead, the architecture relies primarily on S3 buckets, lambda functions, and AWS Batch for data processing. The system orchestrated so that it supports full end-to-end automation from receiving the raw science products to the delivery of higher level of scientific products to the science community.

The mission will release scientific products to the science community every three months. This requires a quicker and automated data processing environment that can scale up based on processing demands. Accordingly, the mission science team has more time to analyze the generated products and produce higher level of scientific products that they are responsible for it.

3.2 SDC Components

The SDC can be divided into three components: data management and storage, data processing system, and data dissemination systems. These components and their major sub-components are described below.

3.2.1 Data management and storage system

This system is responsible for handling all data in the SDC and ensuring its integrity. The data management architecture diagram (see Fig. 2) shows orchestration of data management and storage infrastructure. AWS S3 storage bucket is used to store all scientific products and the SDC Rational Database Service (RDS) which holds all science products metadata. The data management activities are scripted into different inter-connected Lambda functions. Various functions are utilized to check the integrity of received files, make sure they are as per the defined standards. The component also makes sure that all data received is properly indexed in the SDC database and has different tools scheduled to routinely run to ensure all files in the storage repository are well indexed and aligned with the information available in the database. Once the management tasks are performed successfully, the files will be prepared for the data processing component and notify different teams accordingly.

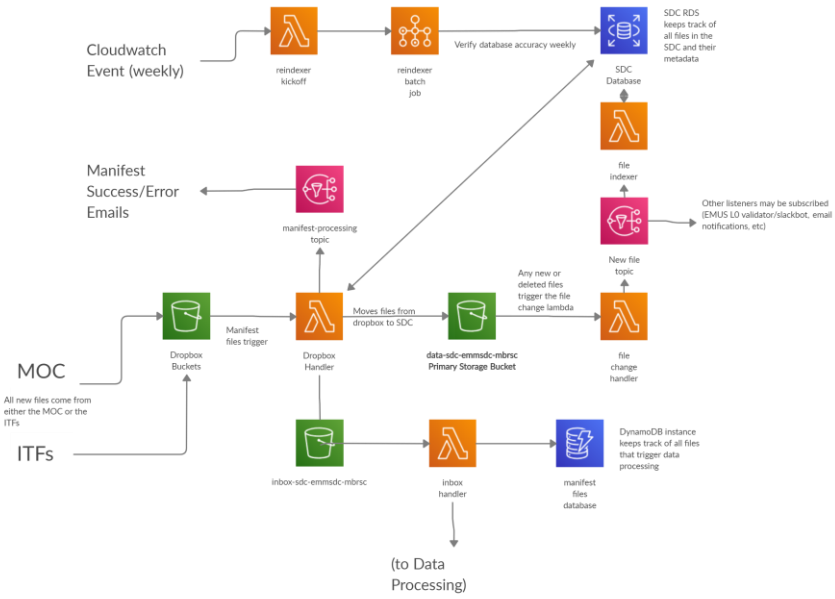

Fig. 2. Data management and storage architecture

3.2.2 Data processing system

Each ITF uploads its processing code in a Docker container to the Elastic Container Registry which is the managed container repo provided by AWS. The environment designed to support the upload of updated version of the data processing software and tested on the system staging platform. Once the updated container tested successfully on the staging platform, it can be deployed to the production environment.

The routine automatic processing job is triggered by the data management component by sending a manifest file to a specific S3 bucket location. This manifest triggers a Lambda function that starts all of the Step Function processing pipeline. The use of AWS Step Function helps orchestrating the order that processing jobs are running on the SDC. AWS Batch service is utilized to run the data processing pipeline by fetching the appropriate docker image from ECR repository. The batch compute environments are pre-defined and describe the environment variables to pass it, the CPU/RAM specifications needed to run the image and the IAM role should be assigned to successfully communicate with different integrated services. Each ITF processing environment is equipped with RDS service to store any kind of processing information and metadata. it is also integrated with Elastic File System service that contains SPICE data, as well as all processing logs and any other arbitrary data that the ITFs require for their processing pipeline (e.g. calibration files).

The reprocessing of scientific products are expected to happen regularly and the SDC designed to scale up automatically to handle the production data processing as well as reprocessing jobs simultaneously. Fig. 3 shows the architecture of data processing system.

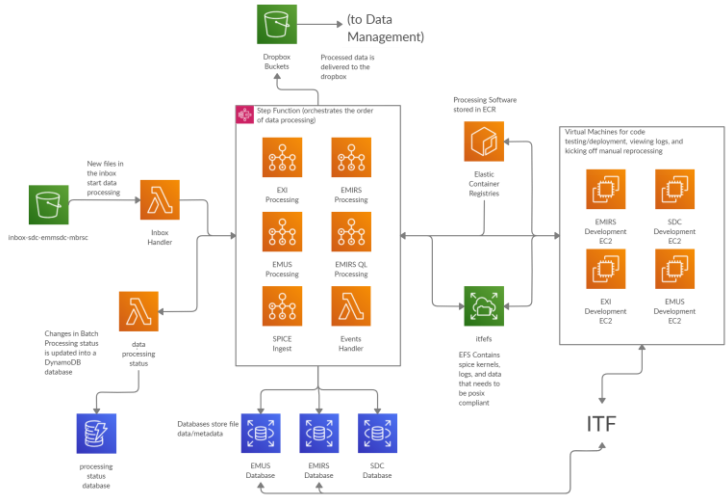

Fig. 3. Data processing architecture

3.2.3 Data dissemination system

The SDC takes advantage of AWS's capabilities to host static website content in an S3 bucket, which ensures that the website is highly available and highly scalable. The website is served by AWS Cloudfront to ensure low latency access to users around the globe.

APIs are built in API Gateway to query the SDC database for specific files or download relevant files to the end user. These APIs are used by the website to power their tools, but can also be accessed directly for users that want programmatic access. API requests are handled by either Lambda functions or batch jobs to either query the SDC website or prepare for file downloads and the responses are sent back to the requester.

The APIs are secured by AWS Cognito which is used as the user authentication/authorization tool. Users must register with a username and password on the website. From there, they supply a username/password to Cognito, which will then grant them a token that enables access to the website or the APIs.

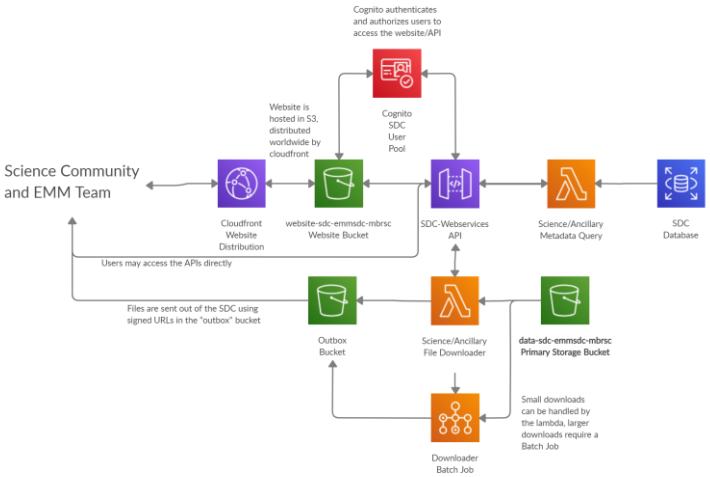

Fig. 4. Data dissemination architecture

4. Conclusions

The Science Data Centre for Emirates Mars Mission is responsible for managing the mission scientific products, performing science data processing to produce all higher scientific products based on the mission raw data received from the Mission Operation Center, and disseminate the generated science products to the mission science team as well as the science community.

The SDC fully deployed on Amazon Web Service as a cloud-native Serverless solution. The solution utilizes different AWS managed service such as AWS S3, Batch, Cloudfront, API Gateway, Lambda and Step Function. Different managed service innovatively orchestrated to perform the three main roles of the science data centre: data management and storage, data processing, and dissemination.

The deployment of SDC on the cloud allows for better system scalability, agility, reliability and backup and restore. Furthermore, it dramatically reduces the CapEx cost and allows for better utilization of the infrastructure budget. Data processing jobs completed faster than the required allocated timeframe which helps the instrument teams and mission scientists to quickly look at the instrument observations and processed products. Data reprocessing jobs are expected especially during the early stage of the mission due to enhancement of the calibration mechanisms and algorithms, accordingly, the developed SDC allows for new data processing and exsisting data reprocessind to be performed simultaneously.