Fully automated cloud based science data center for Emirates Mars Mission

Omran Alhammadi[1] , Bryan Harter[2], James Craft [2], Julie Barnum[2], Bryan Staley[3], Mohammad Alfalasi[1], Ransom Christofferson[2]
[1]Mohammed Bin Rashid Space Centre (MBRSC), United Arab Emirates (Omran.alhammadi, mohammad.alfalasi@mbrsc.ae),
[2]Laboratory for Atmospheric and Space Physics (LASP) at University of Colorado, United States of America (bryan.harter, James.Craft, Julie.Barnum, ransom.christofferson@lasp.colorado.edu),
[3]Net-centric Design Professionals, United States of America (bryan.staley@ndpgroup.com)

Introduction: The Science Data Center (SDC) is responsible for generating and managing 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. It is fully deployed at Amazon Web Service (AWS) utilizing different AWS managed services to build a cloud-native data processing platform for planetary missions. The Full automation in the data processing allows for generating different levels of data products for different instruments onboard the EMM automatically.

SDC Implementation: 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 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.

Data management and storage system: This system is responsible for handling all data in the SDC and ensuring its integrity. 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 that are utilized to check the integrity of received files. The component also makes sure that all data received is properly indexed in the SDC database

Data processing system: Each Instrument team uploads its processing pipeline 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. 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.

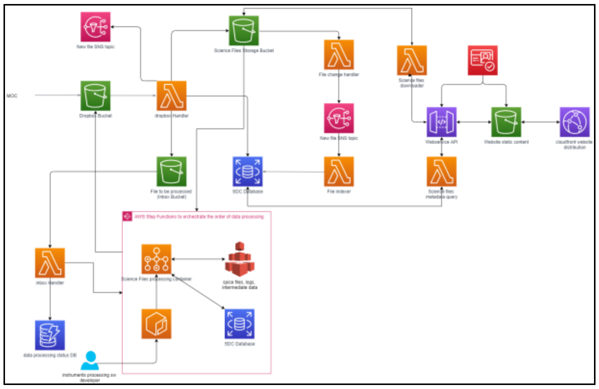
New file SNS topic
Science Files Storage Bucket
File change handler
Science files downloader
MOC
Dropbox Bucket
dropbox Handler
New file SNS topic
Webservice API
Website static content
cloudfront website distribution
File to be processed (inbox Bucket)
SDC Database
File indexer
Science files metadata query
AWS Step Functions to orchestrate the order of data processing
inbox Handler
Science Files processing container
spice files, logs, intermediate data
data processing status DB
SDC Database
instruments processing sw developer

Figure 1 SDC AWS architecture

Conclusion: 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.