# AWS Processing Pipeline for MAVEN IUVS

J. P. Elliott, R. Meisner, B. McClellan, E.Jacobshagen, S. Jain, N. Schneider.

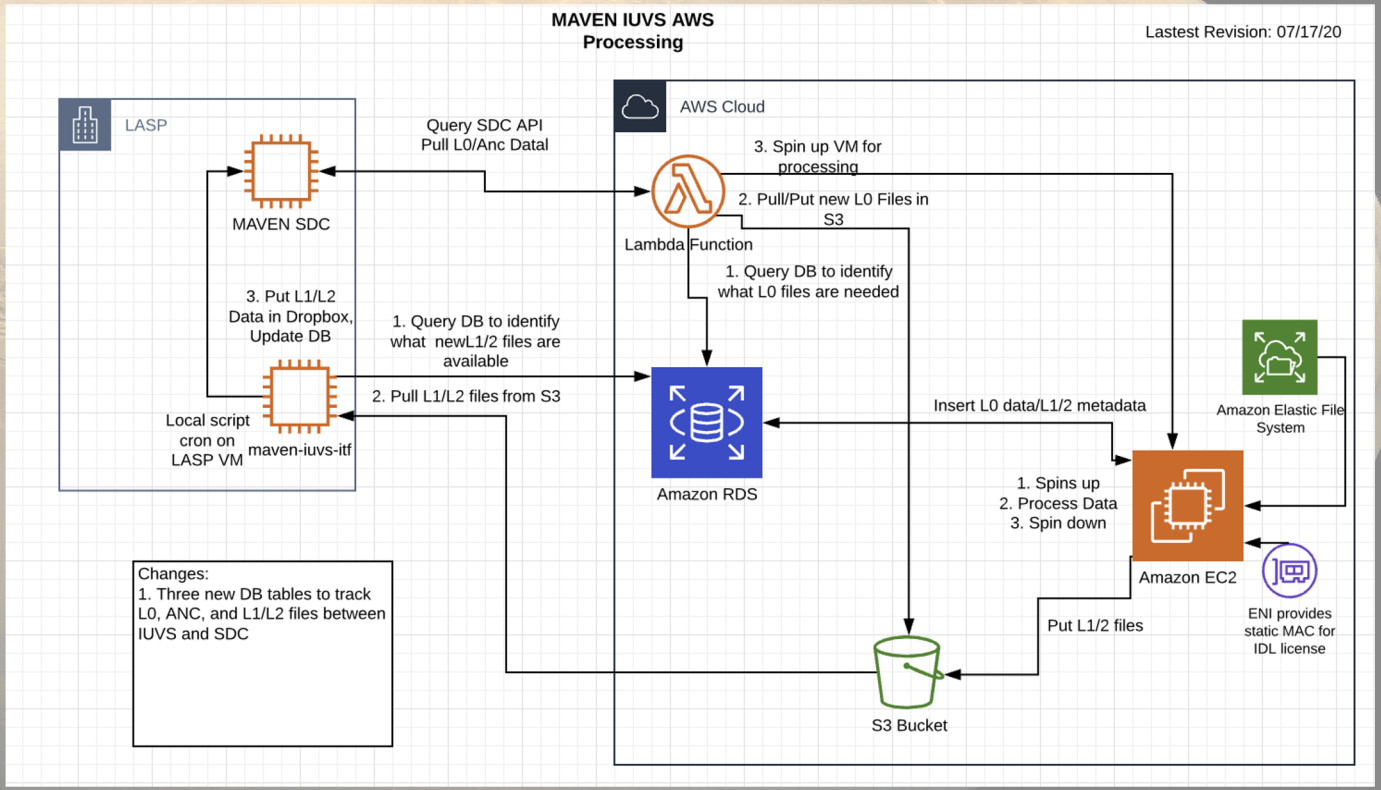
Laboratory for Atmospheric and Space Physics, University of Colorado

joshua.elliott@lasp.colorado.edu

Planetary Data Workshop : 28 June - 2 July 2021

# AWS Architecture

- Lambda
- EC2 (& EFS)
- RDS
- S3
- SDC API



# AWS Architecture recap

- Lambda serverless code
  - Pull L0/ancillary files from SDC API
  - Upload L0 and ancillary files to S3
  - Insert metadata into RDS
  - Spin up EC2 instance
- S3 object store
  - Storage of L0 files
  - Storage of ancillary files
  - Storage of L1 files
- Maven SDC API
  - Serve L0 files
  - Serve SPICE & ancillary files
- EC2 Virtual Machine
  - Pull L0 and ancillary files from S3
  - Process files into L1 products
  - Insert data into RDS
  - Upload L1 files to S3
- RDS Postgres database
  - Store L0 data
  - Keep track of L0 files that have been ingested
  - Keep track of L1 files that have been produced
  - Keep track of processing runs
  - Keep track of file metadata

5th Planetary Data Workshop &
2nd Planetary Science Informatics and Data Analytics Meeting

# Cloud Processing of PDS Archival Products with Amazon Web Services, Kubernetes, and Elasticsearch

Kevin Grimes – kevin.m.grimes@jpl.nasa.gov

Co-Authors: Rishi Verma, James Michael McAuley, Tariq Soliman, Anil Natha, Zachary M. Taylor

Wednesday, June 30, 2021

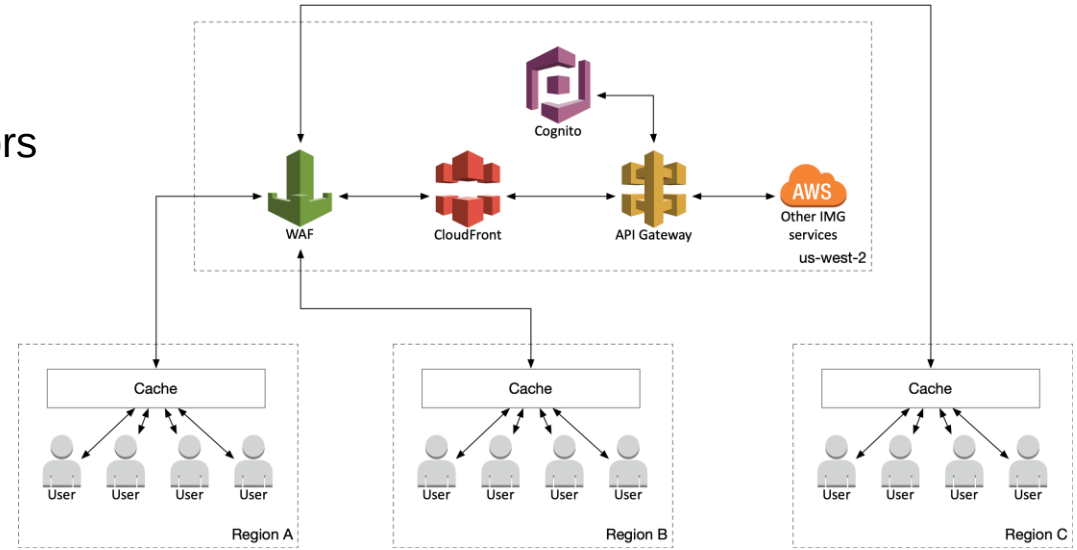
Jet Propulsion Laboratory
California Institute of Technology

# Cloud Processing of PDS Archival Products

Get the slides!

https://bit.ly/3j76cGO

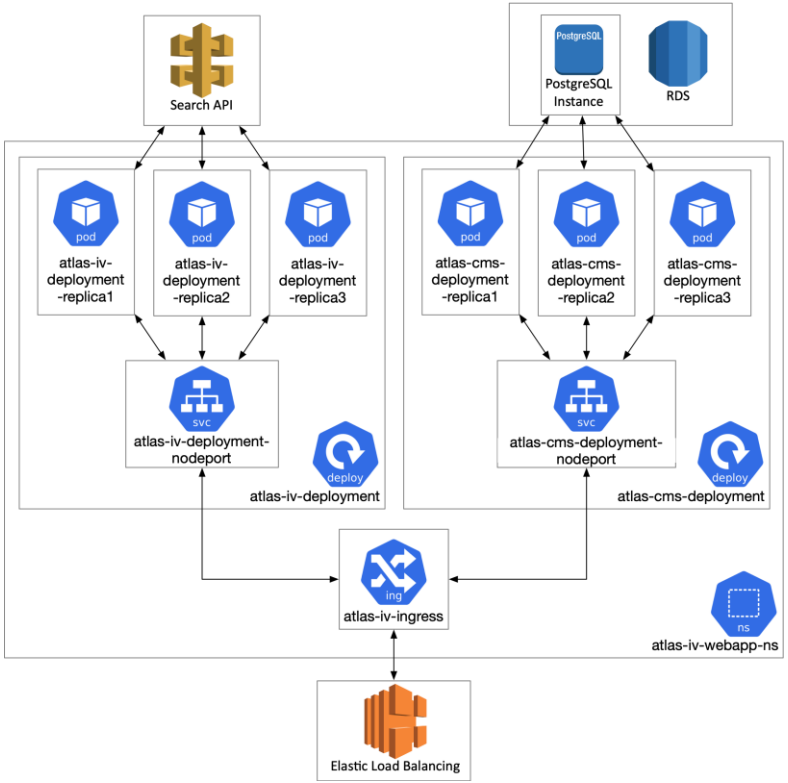
- Nearly 2 PB of data at PDS Imaging Node (and growing!)
- **Challenge**: Making large amounts of data searchable and accessible as quickly as possible
- Option A – on-prem
  - Buy more hardware
  - Hire more system administrators
- Option B – the cloud
  - AWS Cloud
  - Managed services
  - Microservices architecture
  - Containerization



# Cloud Processing of PDS Archival Products

Get the slides!

https://bit.ly/3j76cGO




**Solution: Microservices on Kubernetes**

- Build application components into Docker images
- Each “microservice” is built of multiple Docker containers
- Microservices communicate with each other via Kubernetes service interfaces
- High availability, zero-downtime upgrades
- Running on AWS’s managed Kubernetes-as-a-service, Elastic Kubernetes Service
- Replica of system on-premises for development

# References

Get the slides!

https://bit.ly/3j76cGO

"PIA24428: High-Resolution Still Image of Perseverance's Landing". https://photojournal.jpl.nasa.gov/catalog/PIA24428. NASA/JPL-Caltech.

"PDS Imaging Node". https://pds-imaging.jpl.nasa.gov. NASA/JPL-Caltech

All other icon assets provided with Omnigraffle Stencil Pack, authorized for limited use by Omni Group.

Jet Propulsion Laboratory
California Institute of Technology

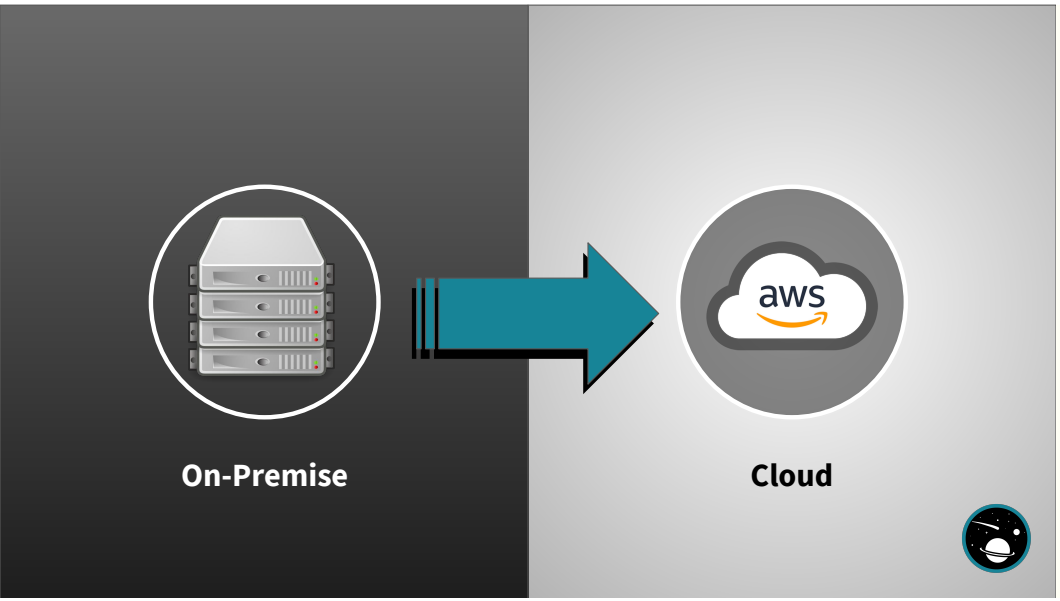
jpl.nasa.gov

# MIGRATION TO THE CLOUD: LESSONS LEARNED FROM THE PROJECT

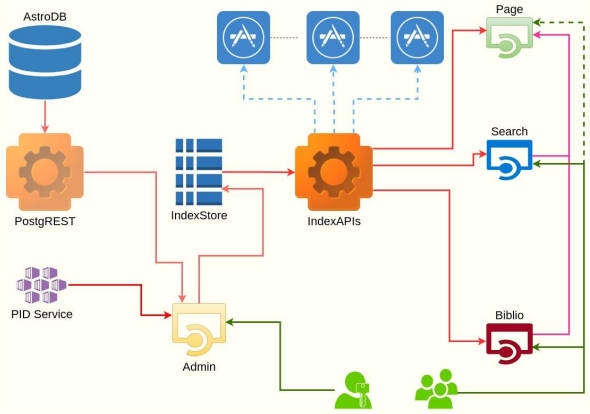
"Development and Operation of the Astromaterials Data System"

Peng Ji [1](pengji@ldeo.columbia.edu), Kerstin Lehnert [1], Dave Stern [2], Juan David Figueroa [1], Lucia Profeta [1], Jennifer Mays [1], Annika Johansson [1], Lulin Song [1],

[1] Lamont-Doherty Earth Observatory, Columbia University, New York, USA, [2] Stern Devops Group

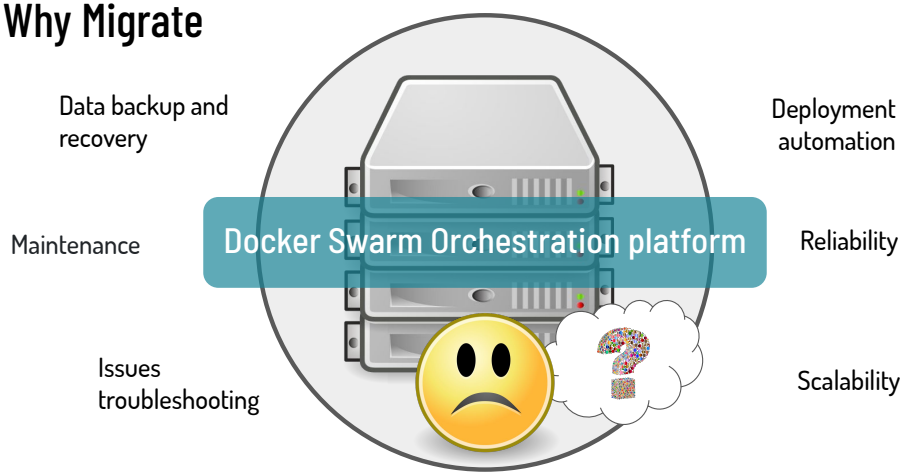



## What is Astromat

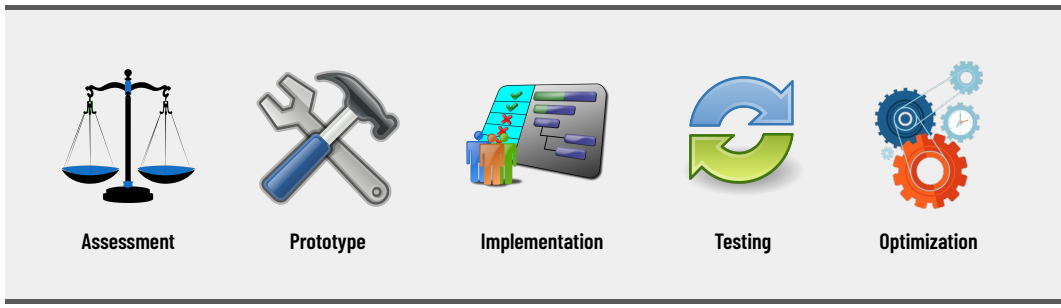



AstroMat is a comprehensive data system for laboratory analytical data generated by the study of astromaterials curated at the NASA Johnson Space Center. It is designed as an ecosystem of interconnected applications that provide human- and machine-readable interfaces to the data gathered and managed in AstroMat's databases.

## Why Migrate

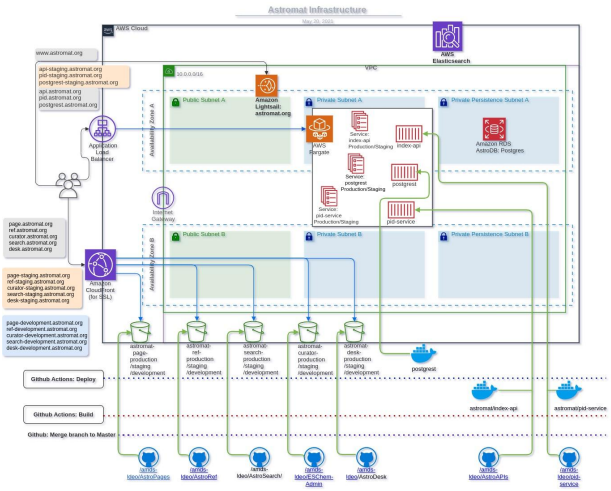



## Five Steps to AWS



## Live in AWS

- Infrastructure-as-Code (IaC)
- Document-as-Code (DaC)
- Continuous Integration(CI)
- Continuous Delivery (CD)




## Resource Mapping Checklist

| | Definition | Tech Options | AWS Resource |
|---|---|---|---|
| Front-end tier | Applications with UI | Angular/React/Vue | S3 + CloudFront |
| Middle tier | The glue between data and UI | NodeJS/... | Fargate service + ELB<br>Lambda |
| Back-end tier | Data layer | Postgresql/MySQL/<br>ElasticSearch | RDS<br>ElasticSearch service |

## What We Learned

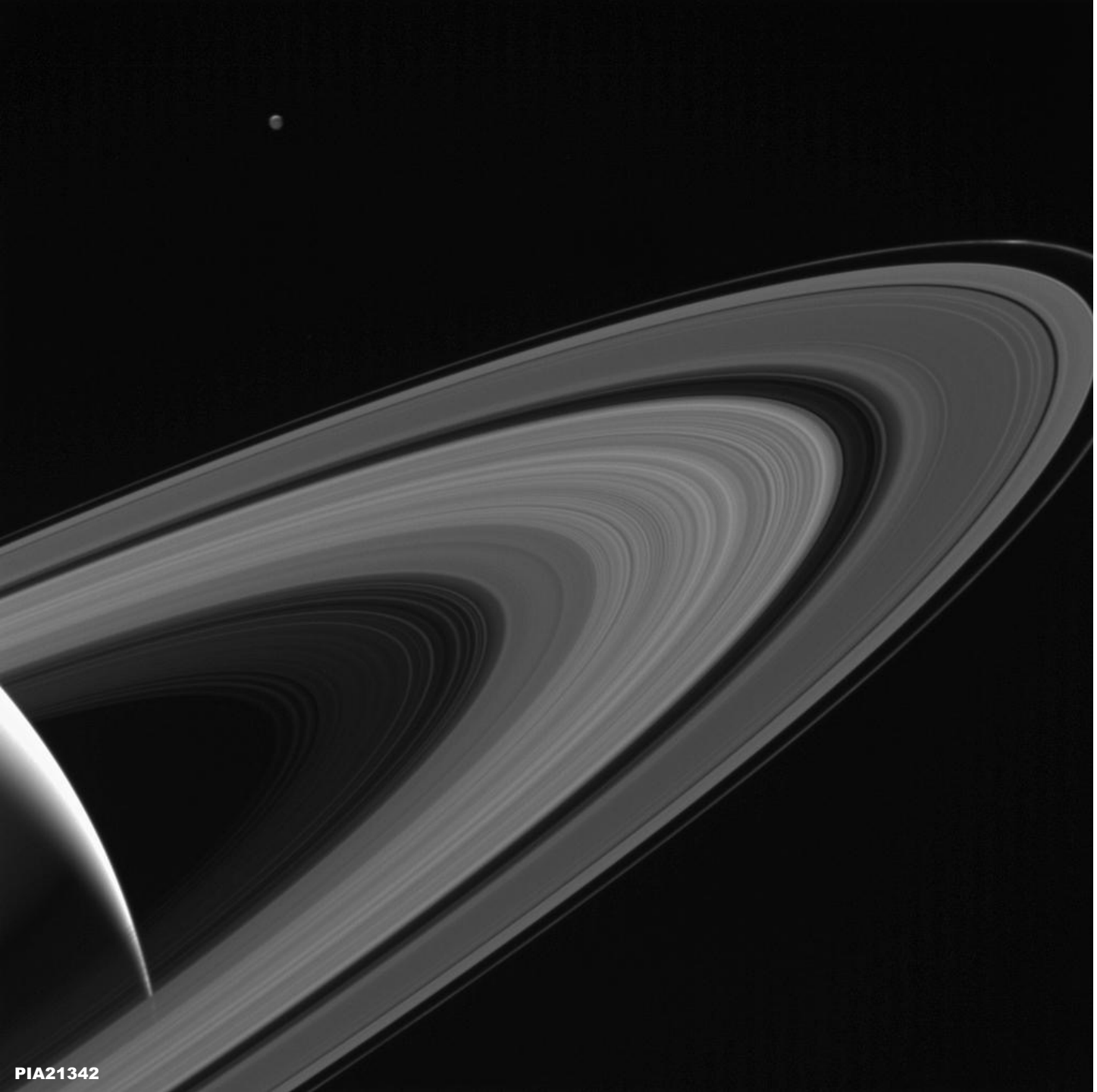
### Challenges:

- Tech changes
- Responsibility and Role changes
- Development approach changes

### Solutions:

- Cloud knowledge and skills
- Guidelines and workflows for development, deployment and management
- Collaboration and code reviews

CONTINUOUS FLOW!




5th Planetary Data Workshop

# NASA PDSIMG KDP: A Distributed Pipeline Processing Tool for Kubernetes

Zachary M Taylor
zachary.m.taylor@jpl.nasa.gov

Co-Authors
Kevin M Grimes, Brad Lunsford

Wednesday, June 30, 2021

Jet Propulsion Laboratory
California Institute of Technology

# KDP: A Distributed Pipeline Processing Tool for Kubernetes

## Overview

- Introduction
  - What is PDS IMG
  - Data Processing Use-Cases
  - Operational Challenges
  - Problem Statement
- KDP: “Kubernetes Does Pipelines”
  - Overview
  - Kubernetes Primer
  - KDP Architecture
  - Case Study – Mars 2020
- Conclusion

# KDP: "Kubernetes Does Pipelines"

## Overview

- **Distributed processing made easy**
  - Processing pipelines described in lightweight YAML[1] syntax
  - Reusable processing steps allows for creation of complex pipelines
- **Built on Kubernetes[2]**
  - State-of-the-art container orchestration framework
  - Open-source software (OSS)
  - Deployable virtually anywhere
- **Infinite horizontal scalability**
  - Scale processing jobs to thousands of CPUs with a single command
  - Negligible overhead achieves near-linear scaling
- **Data-location agnostic**
  - Data can be anywhere with an S3[3] interface
  - Most cloud providers, on-prem using services like MinIO[4]
- *Still a work in progress.*

[1]: https://yaml.org/
[2]: https://kubernetes.io/
[3]: https://aws.amazon.com/s3/
[4]: https://min.io/

Jet Propulsion Laboratory
California Institute of Technology

# EXPERIMENTS IN TRANSFERRING, VALIDATING, AND RELEASING MARS 2020 MISSION ARCHIVAL MULTI-MEDIA & IMAGERY DATA DELIVERIES IN THE CLOUD


**Presenter:** Rishi Verma

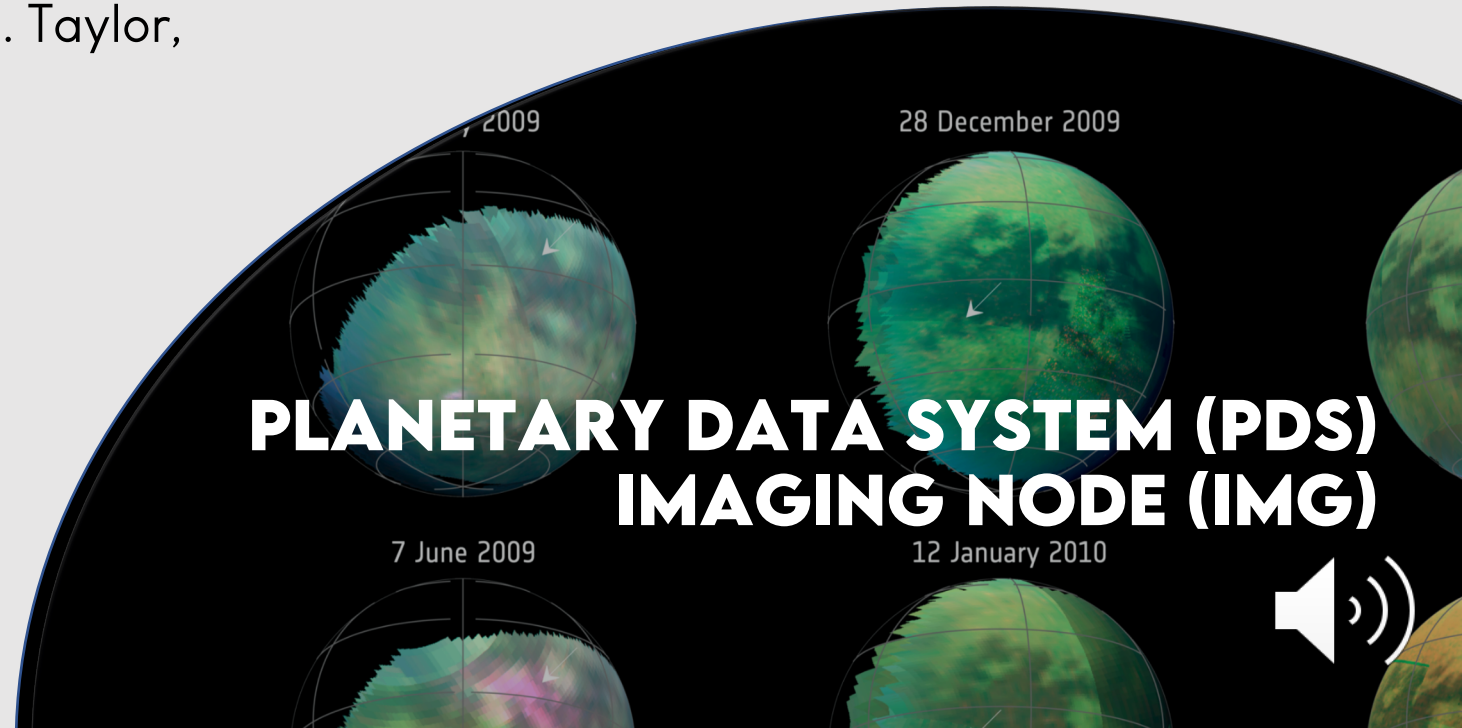
**Co-Authors:** Kevin M. Grimes, Zachary M. Taylor, Michael McAuley, Minh Le, Rafael Alanis

**Additional Support**: Kate Crombie, Brad Lunsford, Tariq Soliman, Anil Natha



5th Planetary Data Workshop & 2nd Planetary Science Informatics and Data Analytics Meeting

Wednesday, June 30th 2021


PLANETARY DATA SYSTEM (PDS)
IMAGING NODE (IMG)

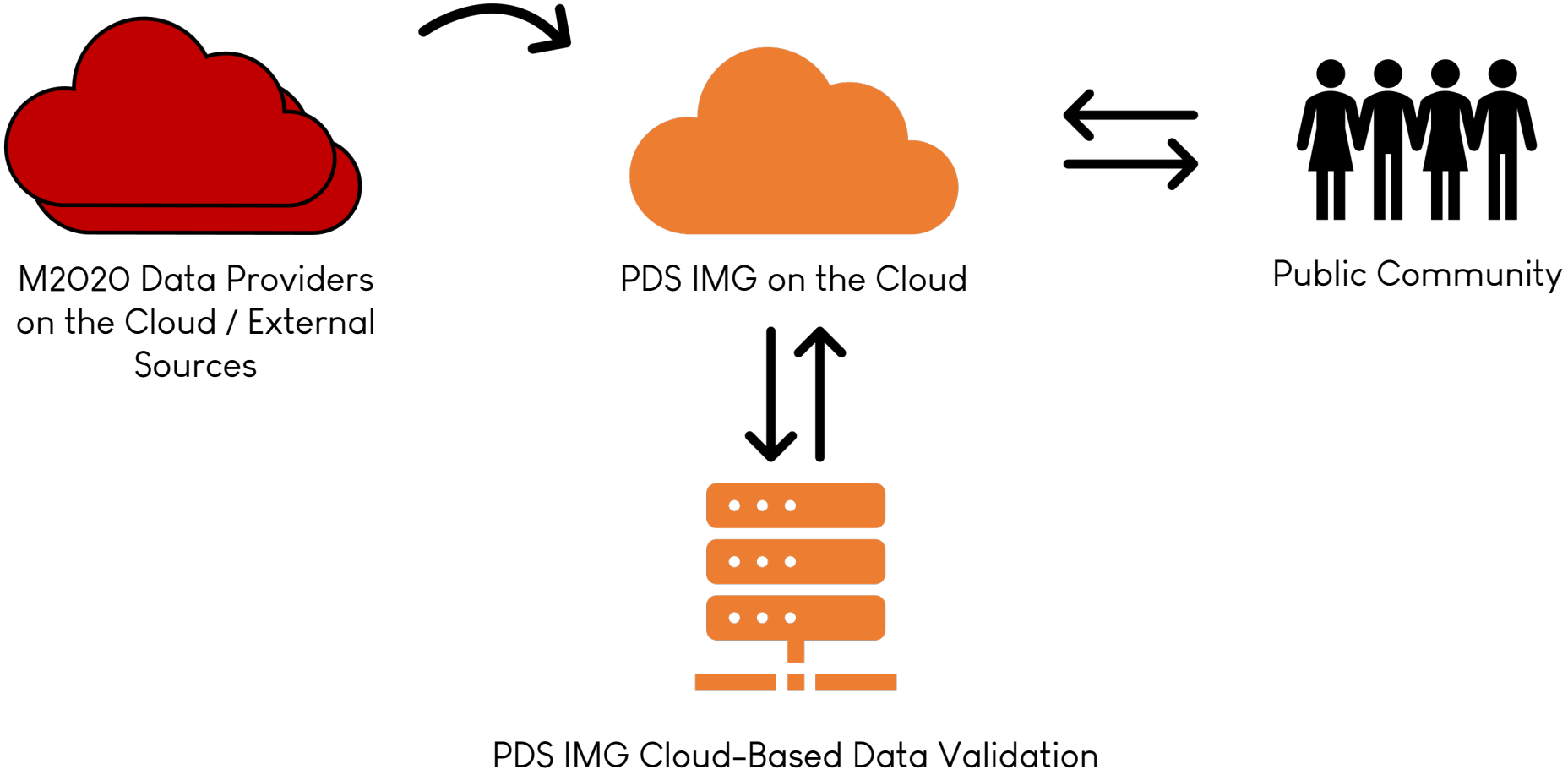
# MOTIVATION

Problem Statement

"How do we *reduce* the time it takes to accept, validate, and release mostly cloud-provided M2020 bundle(s) to the public *while reducing* system-wide data transfer, processing, and backup costs?"

# MOTIVATION

Strategy: Bring IMG Data Releases into the Cloud



# DATA FLOW

## Nominal State: Delivering and Validating New M2020 Data Bundles while Permitting Access Only to Previously Delivered Bundles