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

© 2021 California Institute of Technology. Government sponsorship acknowledged.

Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not constitute or imply its endorsement by the United States Government or the Jet Propulsion Laboratory, California Institute of Technology.

Any and all permissions have been obtained and proper credit of third party material has been cited in the “References” section.

Cloud Processing of PDS Archival Products

Overview

- Introduction
- Architecture
- Deployment
- Conclusions
- Future Work
- References

© 2021 California Institute of Technology. Government sponsorship acknowledged.

Introduction

Overview

- PDS Imaging Node
- Existing backend architecture
- Motivation to evolve

© 2021 California Institute of Technology. Government sponsorship acknowledged.

Introduction

PDS Imaging Node

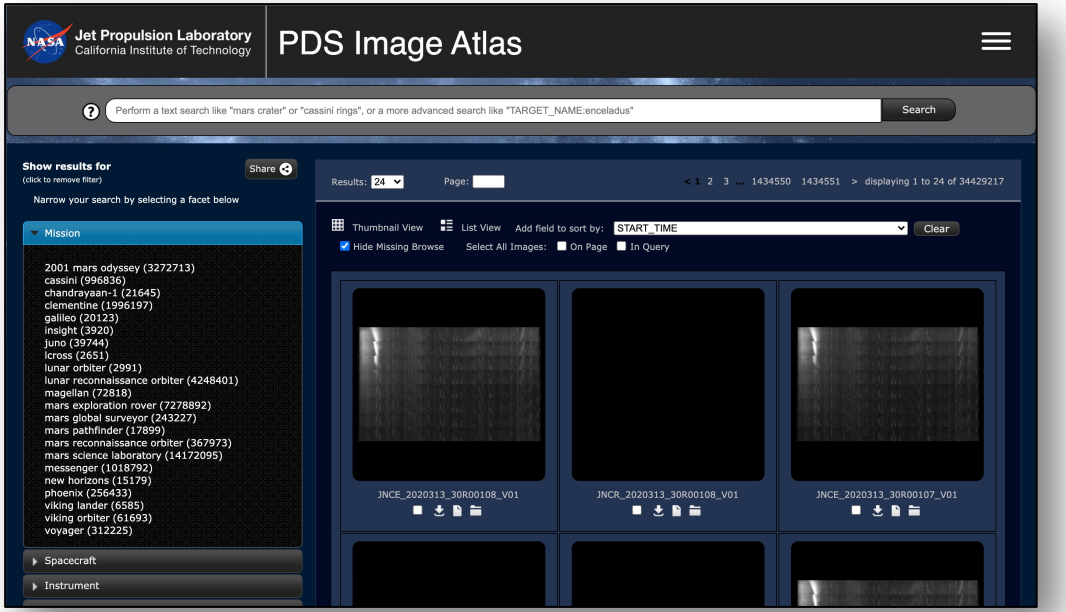
- Cartography and Imaging Sciences Node (IMG) of the NASA Planetary Data System (PDS)
- Home to nearly 2 PB of digital image archives
- Diverse collection of images
  - Both orbital and landed missions
    - Over 20 million images takes from the surface of Mars
    - Nearly 5 million images taken of Mars's surface from orbit
    - Images of Saturn, Jupiter, and Beyond
  - Original, raw experiment data and derived products
  - Differing coordinate systems

© 2021 California Institute of Technology. Government sponsorship acknowledged.

Introduction

Existing backend architecture

- **Image Atlas**
  - Primary tool for discovering data in IMG's archives
  - JavaScript webapp running on-premises
  - Interacts directly with Apache Solr backend
- *Data access*
- *Search*



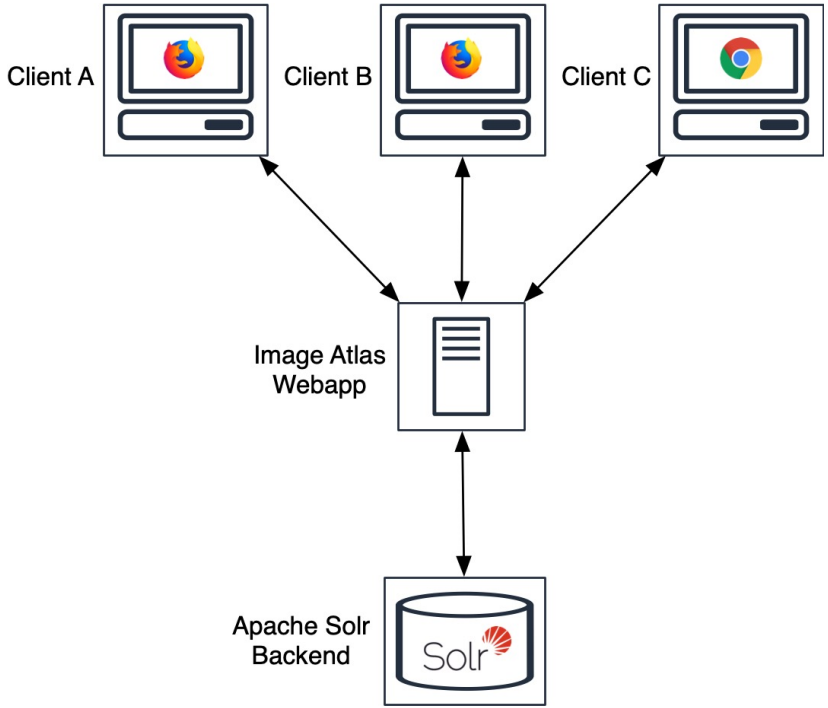
https://pds-imaging.jpl.nasa.gov/search

© 2021 California Institute of Technology. Government sponsorship acknowledged.

Introduction

Existing backend architecture

- **Image Atlas**
  - Primary tool for discovering data in IMG's archives
  - JavaScript webapp running on-premises
  - Interacts directly with Apache Solr backend
- *Data access*
- *Search*



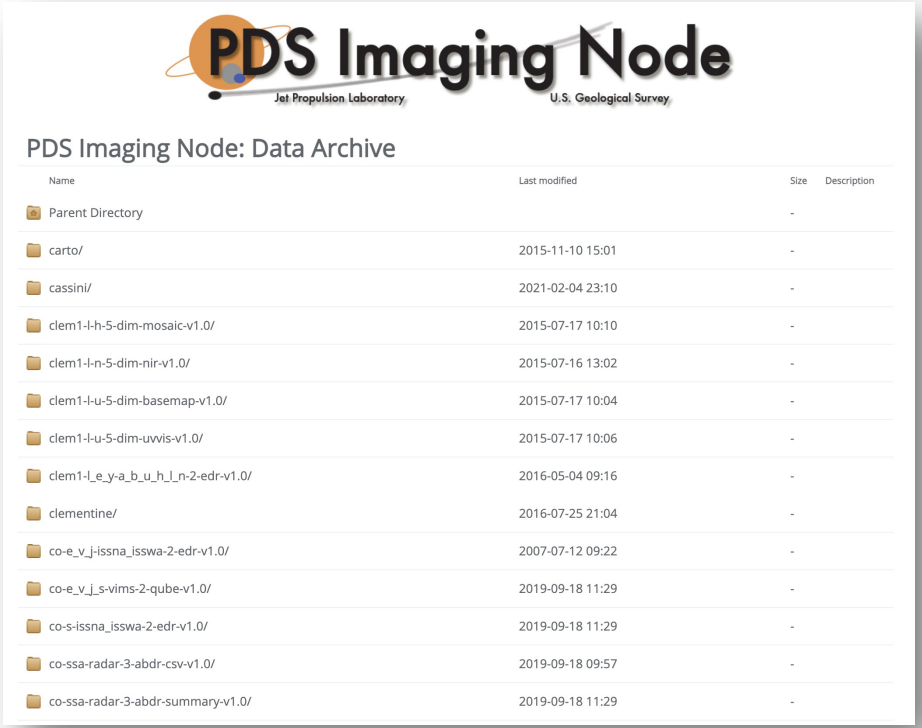
https://pds-imaging.jpl.nasa.gov/search

© 2021 California Institute of Technology. Government sponsorship acknowledged.

Introduction

Existing backend architecture

- *Image Atlas*
- **Data access**
  - Data products served over HTTPS
  - Simple HTML frontend rendering archives as they exist on disk
- *Search*



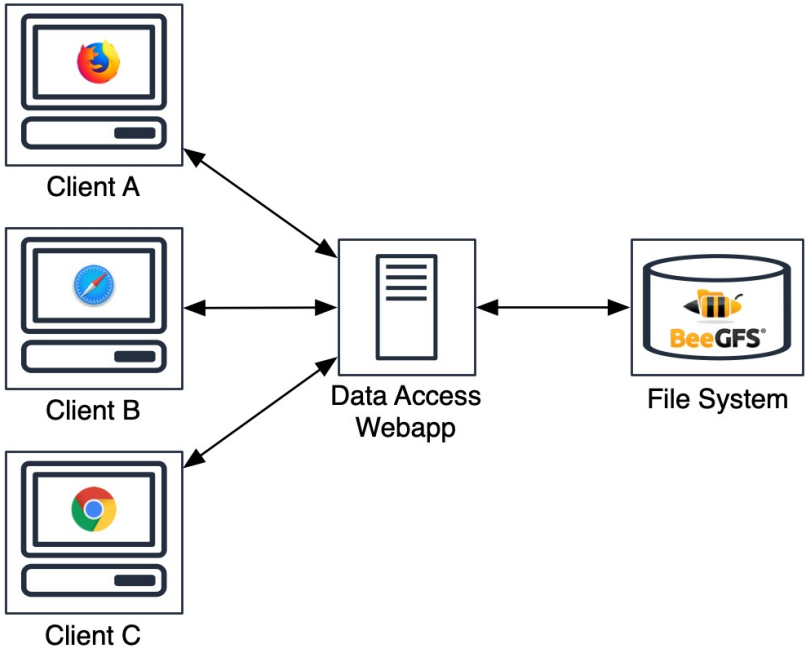
https://pds-imaging.jpl.nasa.gov/data

© 2021 California Institute of Technology. Government sponsorship acknowledged.

Introduction

Existing backend architecture

- *Image Atlas*
- **Data access**
  - Data products served over HTTPS
  - Simple HTML frontend rendering archives as they exist on disk
- *Search*



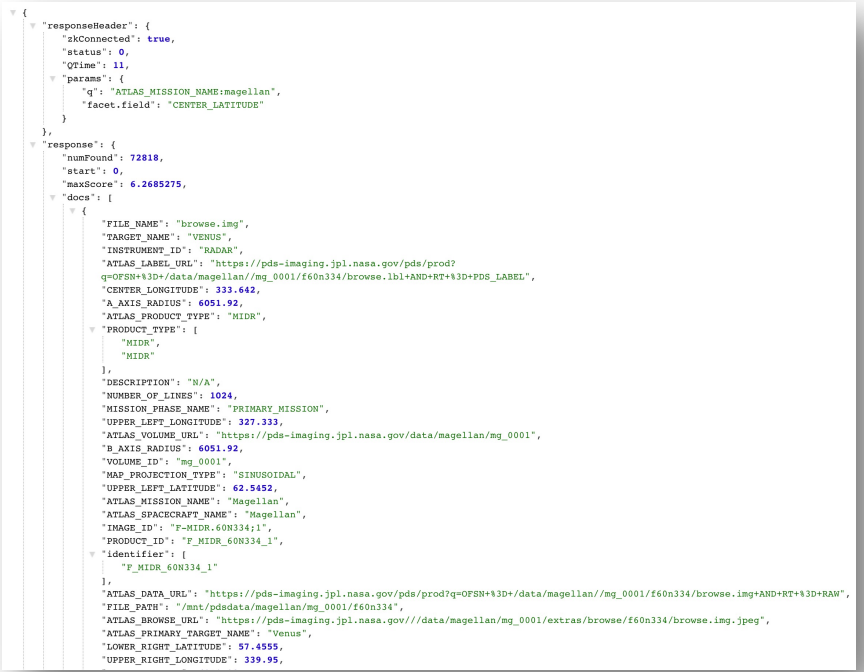
https://pds-imaging.jpl.nasa.gov/data

© 2021 California Institute of Technology. Government sponsorship acknowledged.

Introduction

Existing backend architecture

- *Image Atlas*
- *Data access*
- **Search**
  - Solr endpoint exposed to internet
  - Powerful functionality
  - Diverse content indexed
    - 40+ million PDS labels
    - “Common” metadata
    - Machine-learning enhanced metadata



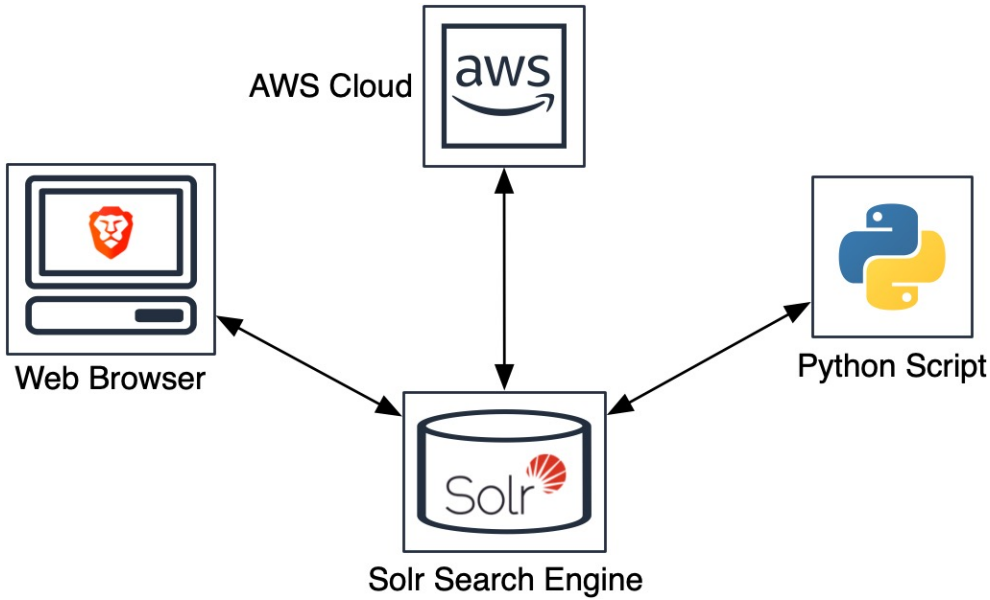
https://pds-imaging.jpl.nasa.gov/solr

© 2021 California Institute of Technology. Government sponsorship acknowledged.

Introduction

Existing backend architecture

- *Image Atlas*
- *Data access*
- **Search**
  - Solr endpoint exposed to internet
  - Powerful functionality
  - Diverse content indexed
    - 40+ million PDS labels
    - “Common” metadata
    - Machine-learning enhanced metadata



https://pds-imaging.jpl.nasa.gov/solr

© 2021 California Institute of Technology. Government sponsorship acknowledged.

Introduction

Motivation to evolve

Requirements

- Image Atlas
  - Multiple data stores
  - Availability issues
  - Cumbersome upgrades
- Data access
  - Store on multiple file systems (and object stores!)
  - Manage cloud egress costs
- Search
  - Easier than Solr API's learning curve
  - Easier to impose structure on search results

Solutions

- AWS Cloud
  - Managed services
  - Only pay for what you use
- Microservices architecture
  - Decoupled components
  - Developed in isolation
  - Communication via APIs
- Containerization
  - Easy to scale
  - Ephemeral
  - Orchestration frameworks

© 2021 California Institute of Technology. Government sponsorship acknowledged.

Cloud Processing of PDS Archival Products

Overview

- Introduction
- Architecture
- Deployment
- Conclusions
- Future Work
- References

© 2021 California Institute of Technology. Government sponsorship acknowledged.

Architecture

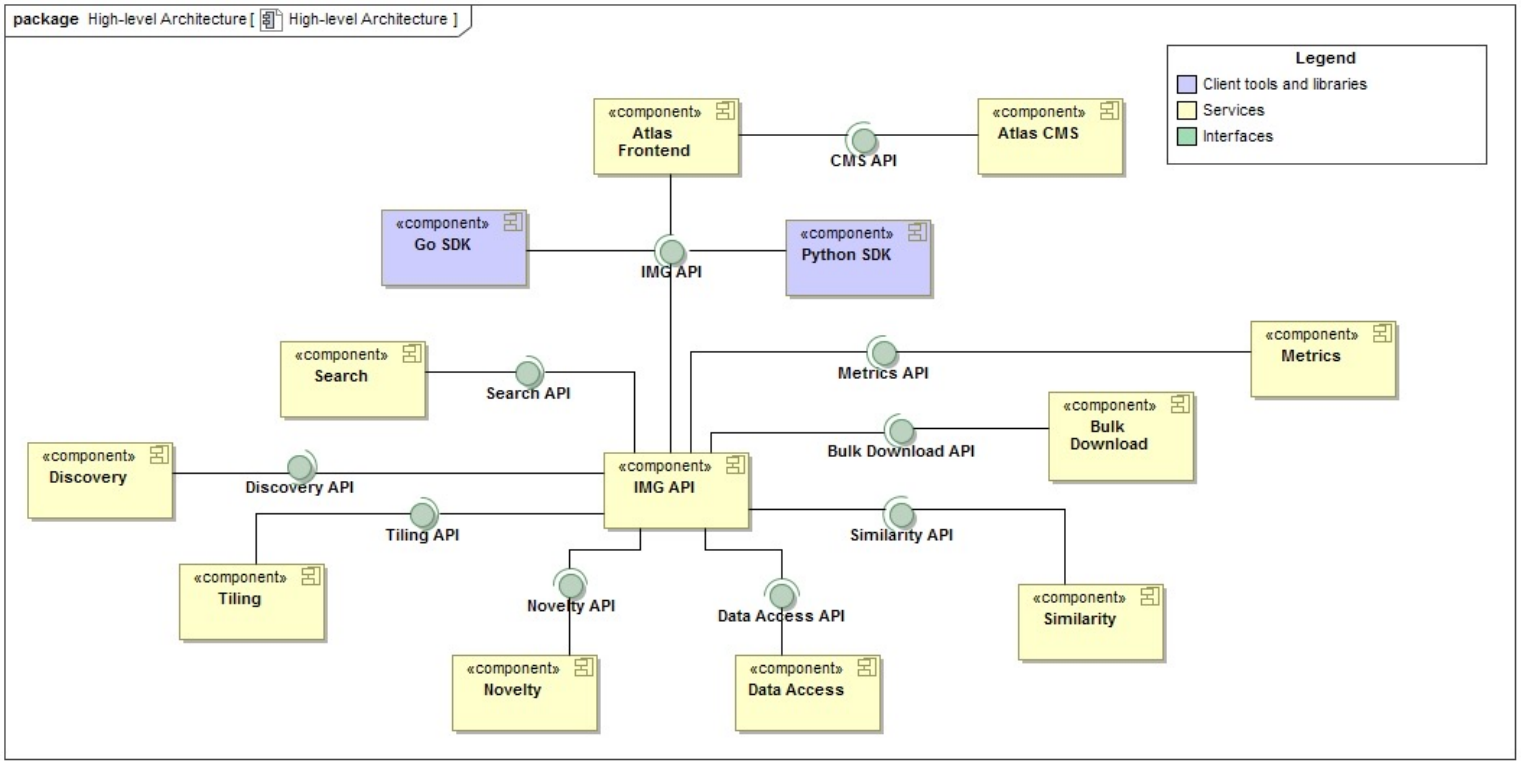
Overview

- PDS IMG API
- Data Access API
- Search API
- Image Atlas client

© 2021 California Institute of Technology. Government sponsorship acknowledged.

Architecture

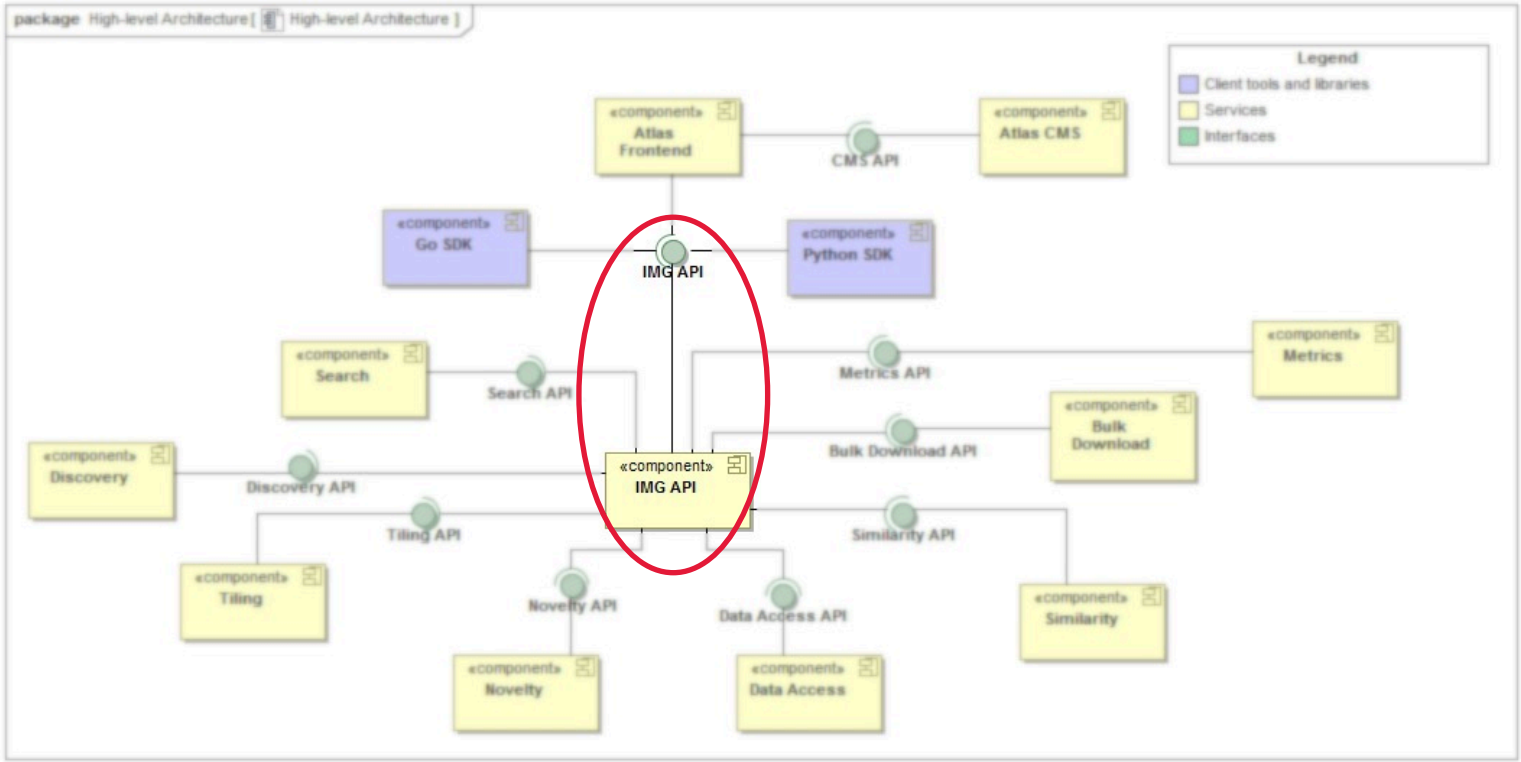
Overview



© 2021 California Institute of Technology. Government sponsorship acknowledged.

Architecture

PDS IMG API



© 2021 California Institute of Technology. Government sponsorship acknowledged.

Architecture

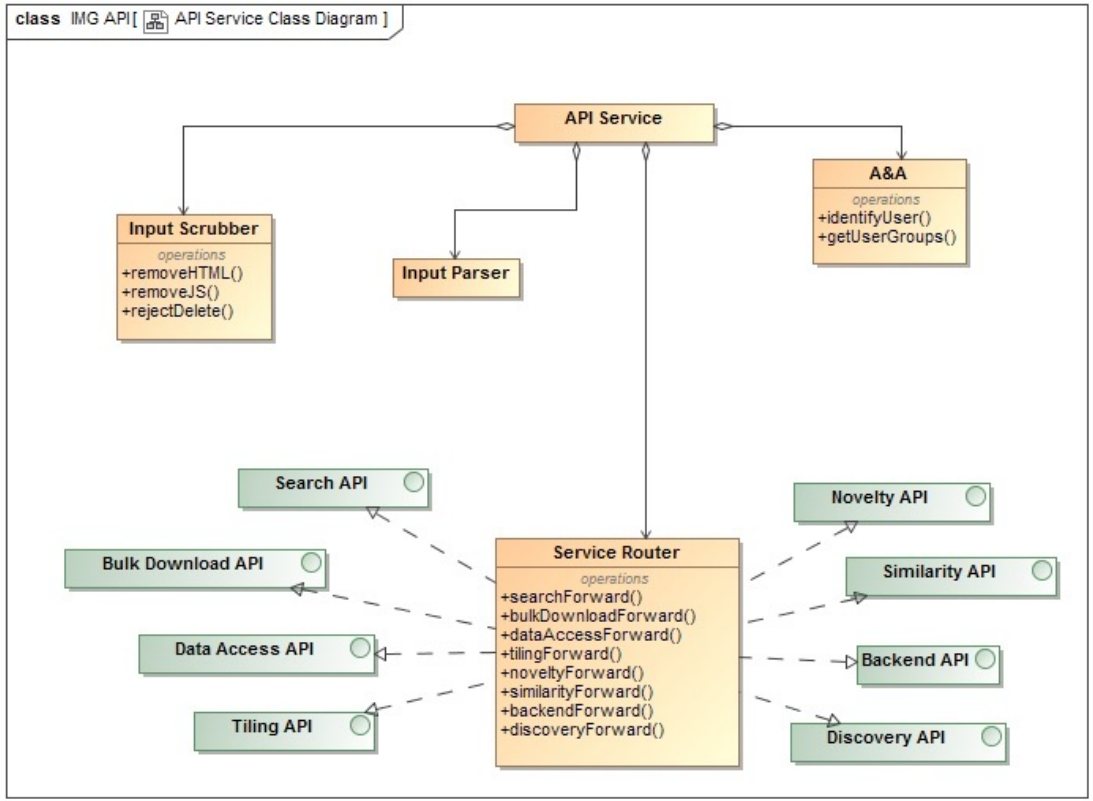
PDS IMG API

- Only service that allows direct incoming network traffic from users
- Single entry point to service network
- Imposes A&A
- Simple webapp defined by OpenAPI 3.0
- Internals of API service mesh can change without impacting user syntax

© 2021 California Institute of Technology. Government sponsorship acknowledged.

Architecture

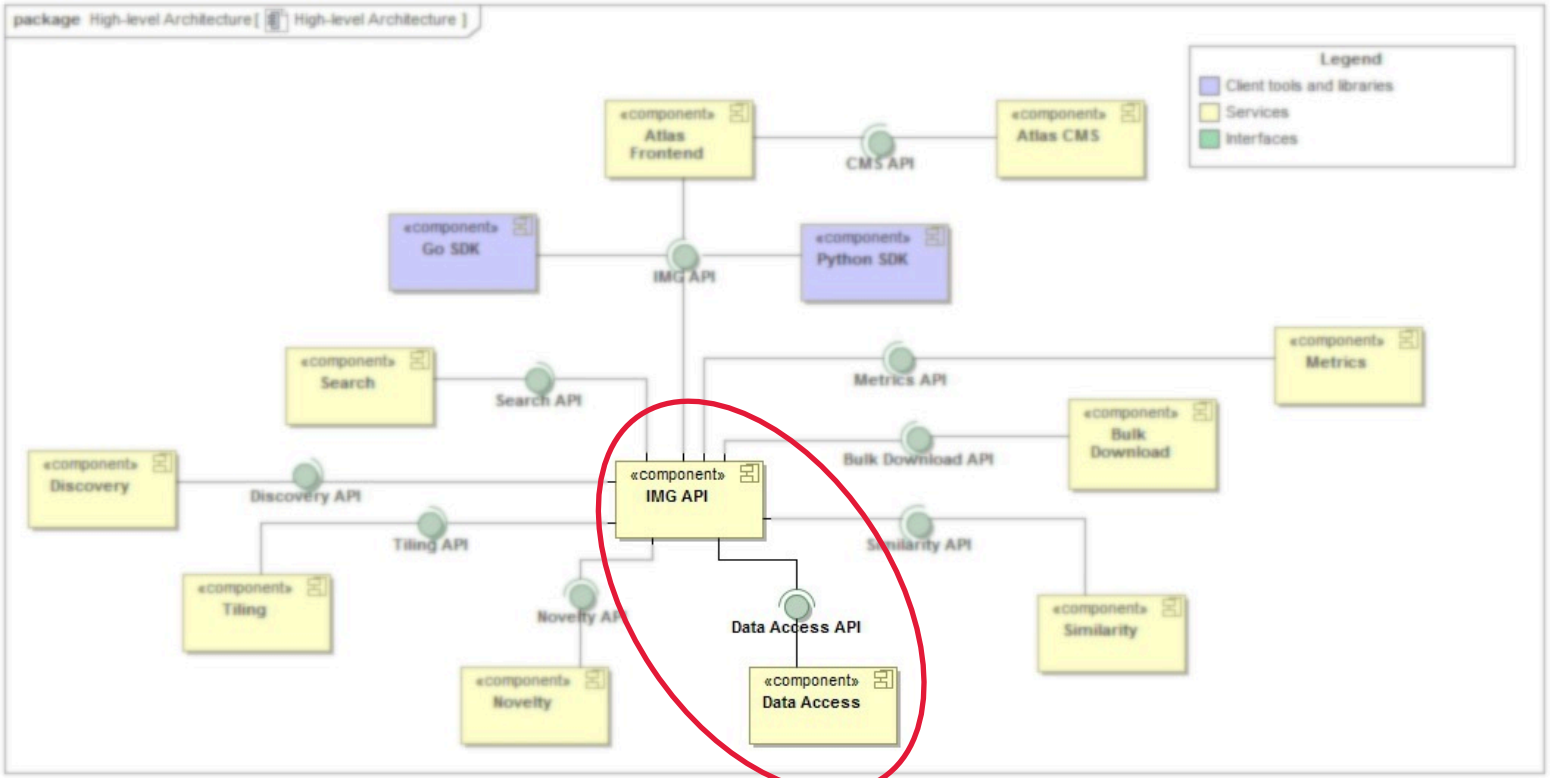
PDS IMG API



© 2021 California Institute of Technology. Government sponsorship acknowledged.

Architecture

Data Access API



© 2021 California Institute of Technology. Government sponsorship acknowledged.

Architecture

Data Access API

- Data stored on JPL premises, other data nodes, and in Amazon S3
- Need a way to route to multiple locations from a single URL
- Manage egress costs (taxpayer money) via user access control, maintaining unfettered access to data for all legitimate use
- Implementation inspired by Earth Science's TEA: https://github.com/asfadmin/thin-egress-app

© 2021 California Institute of Technology. Government sponsorship acknowledged.

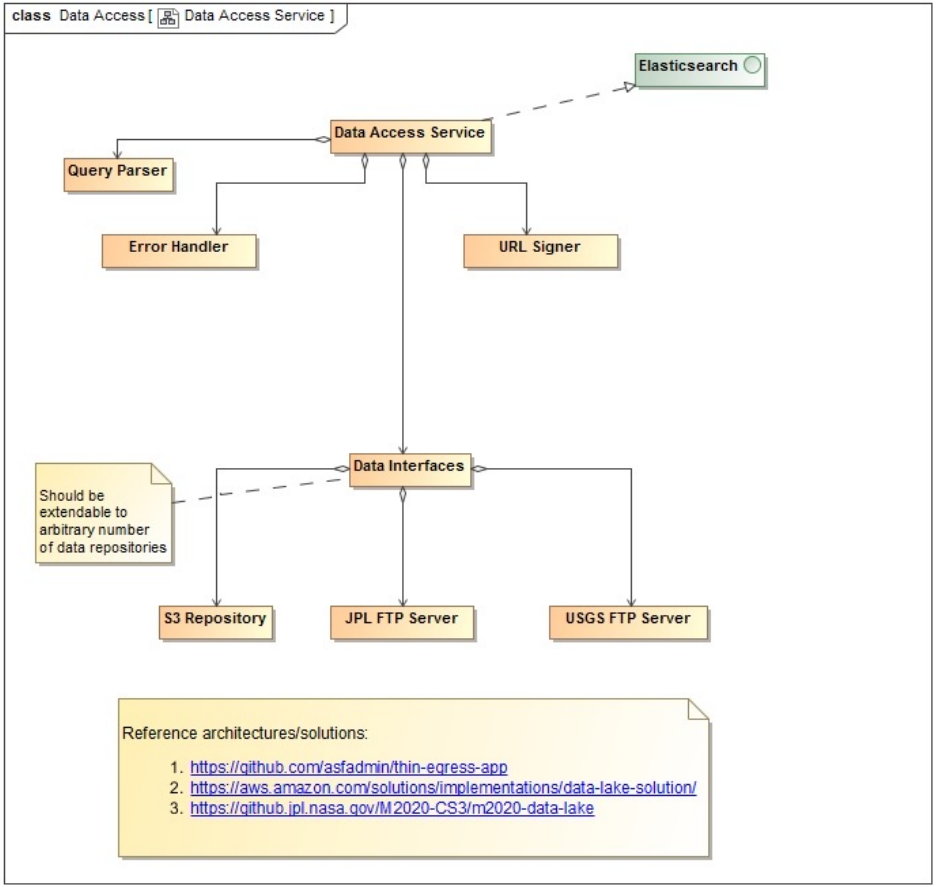
Architecture

Data Access API

1. GET /data/cassini/{}
   302
   https://pds-imaging.jpl.nasa.gov/data/cassini/{}

2. GET /data/m20/{}
   302
   https://s3.amazonaws.us-west-2.com/m20data/{}

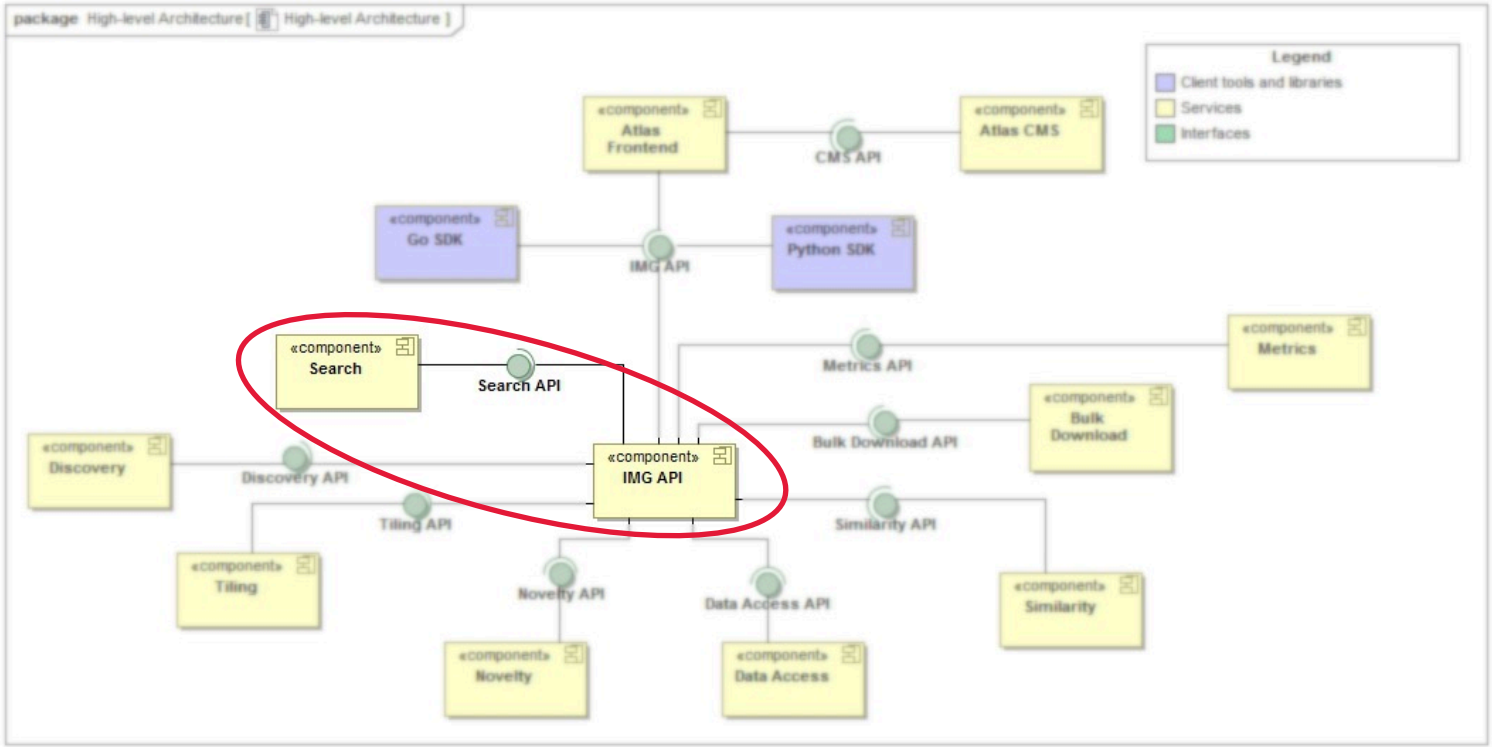
3. GET /data/lroc/{}
   302
   https://astrogeology.usgs.gov/lro/lroc/{}



© 2021 California Institute of Technology. Government sponsorship acknowledged.

Architecture

Search API

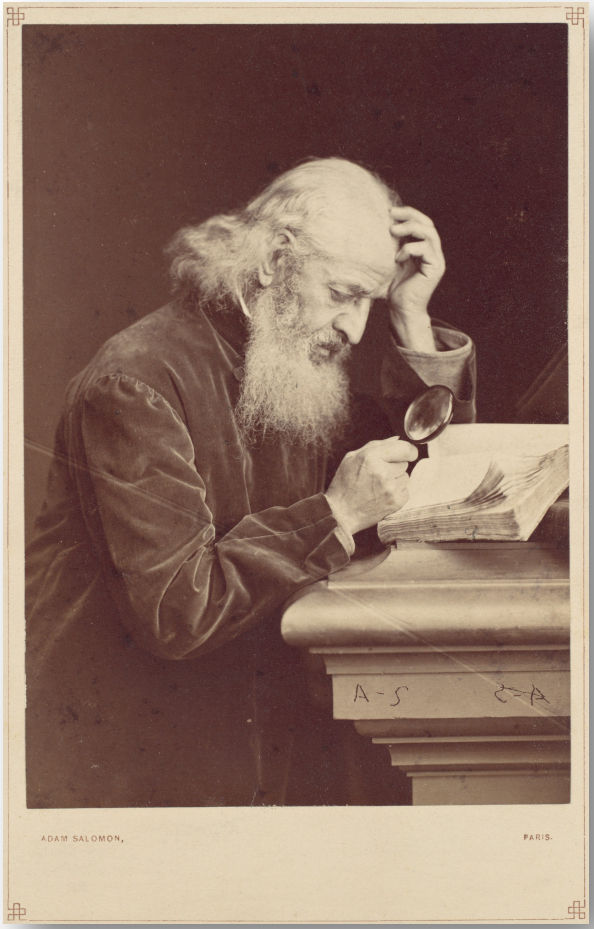


© 2021 California Institute of Technology. Government sponsorship acknowledged.

Architecture

Search API

- Multiple search indexes breaking up data
- Indexes mirroring PDS3/PDS4 label contents
- Index for Atlas search
- Indexes for machine learning metadata
- Search API abstracts away Elasticsearch indexes into a simple-to-use API



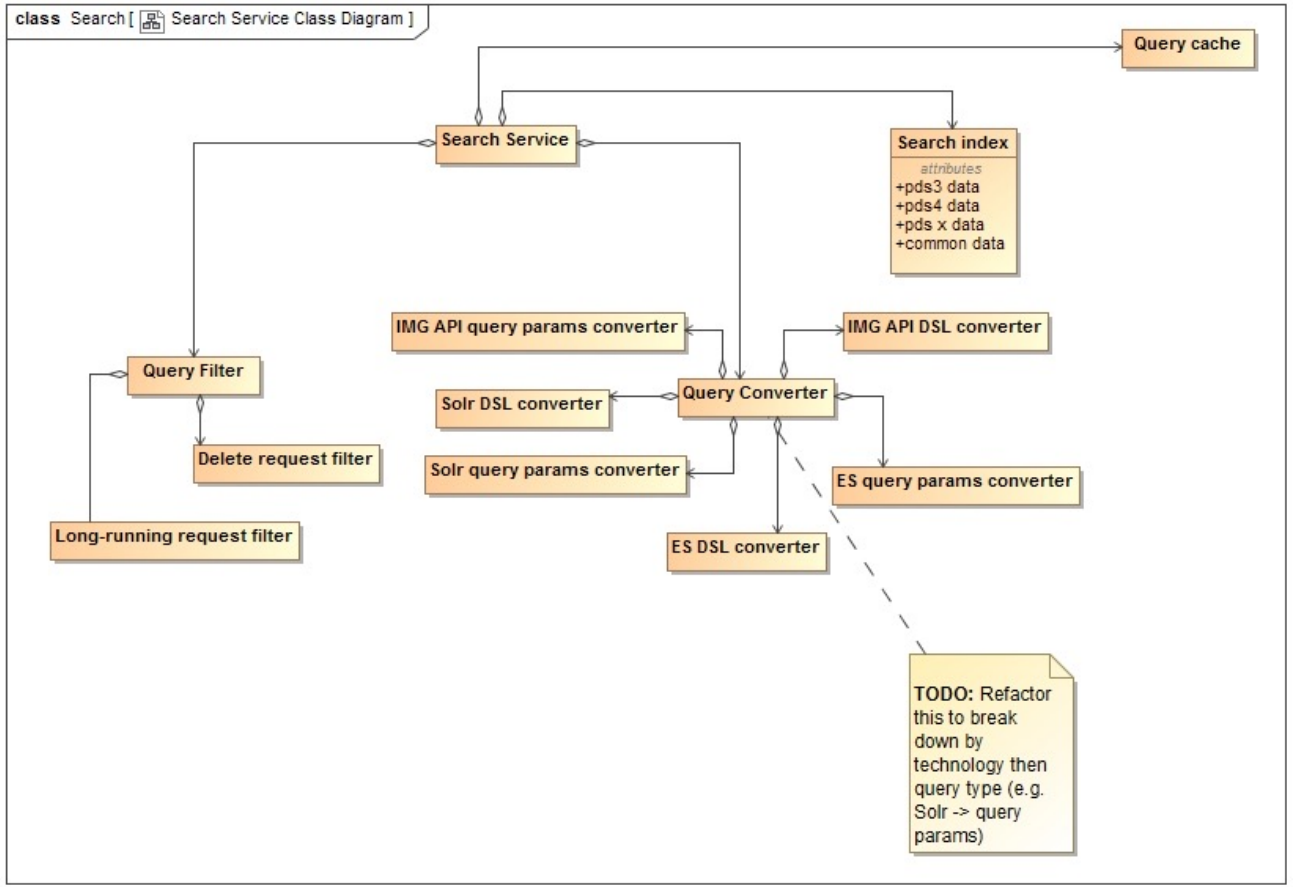
© 2021 California Institute of Technology. Government sponsorship acknowledged.

Architecture

Search API

1. `GET /search?mission:cassini → 200 https://es.amazonaws.us-west-2.com/cassini/_search`

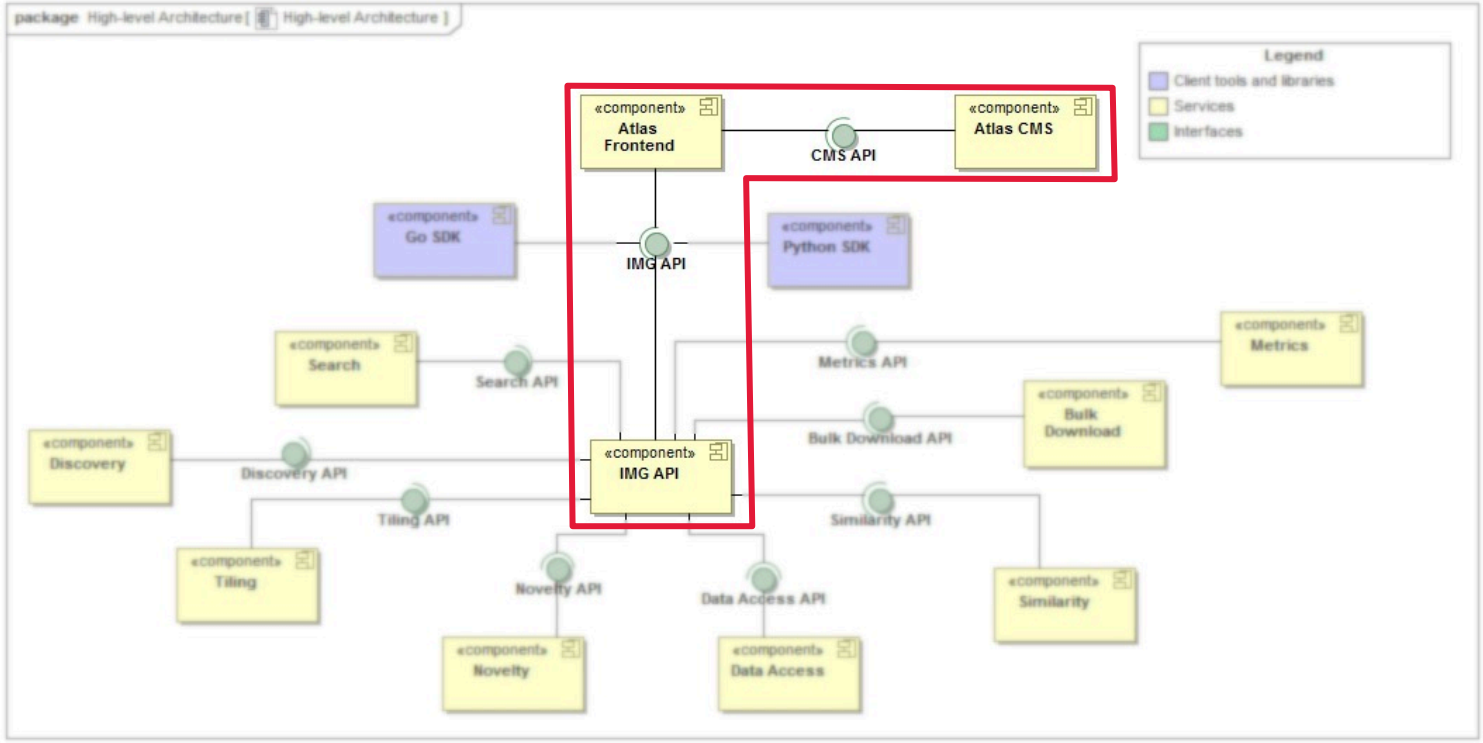
2. `GET /search?mission:m20&has:craters → 200 https://es.amazonaws.us-west-2.com/m20/_search https://es.amazonaws.us-west-2.com/ml_ldd/_search`



© 2021 California Institute of Technology. Government sponsorship acknowledged.

Architecture

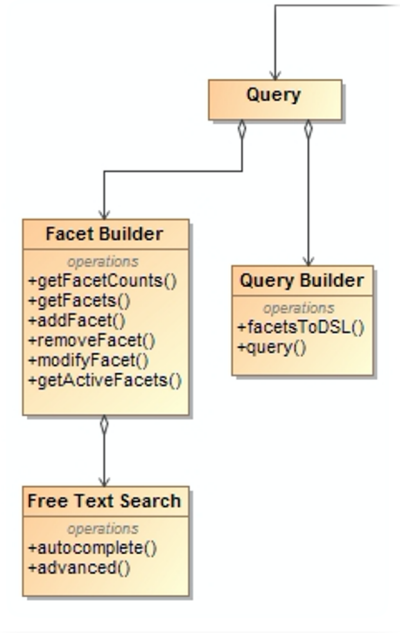
Image Atlas client

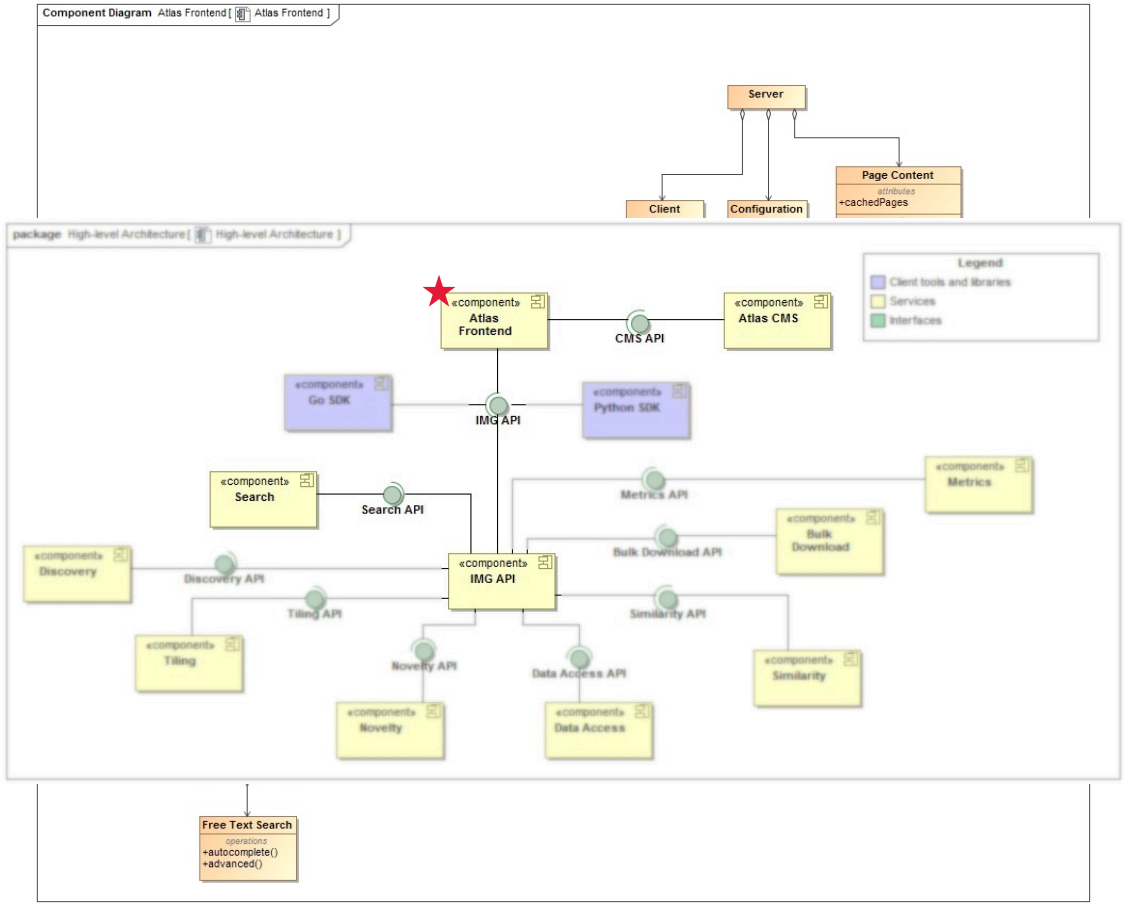


© 2021 California Institute of Technology. Government sponsorship acknowledged.

Architecture

Image Atlas Client



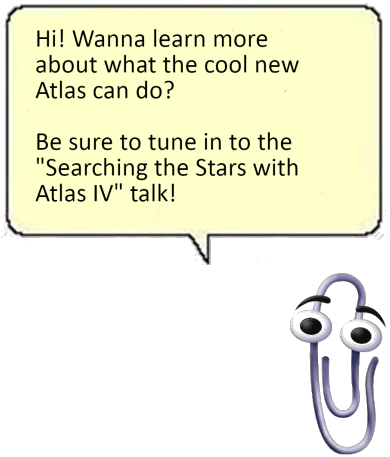


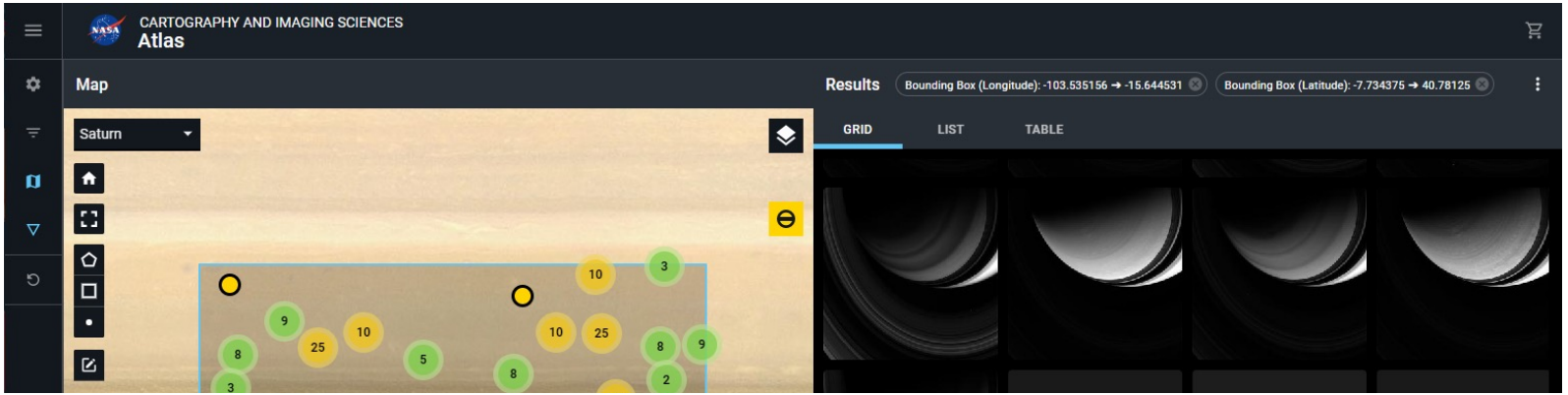
© 2021 California Institute of Technology. Government sponsorship acknowledged.

Architecture

Image Atlas client

- Redesign of existing Atlas III
- Uses PDS IMG API for search
- Stores state in database
- Just another client of PDS IMG API





© 2021 California Institute of Technology. Government sponsorship acknowledged.

Cloud Processing of PDS Archival Products

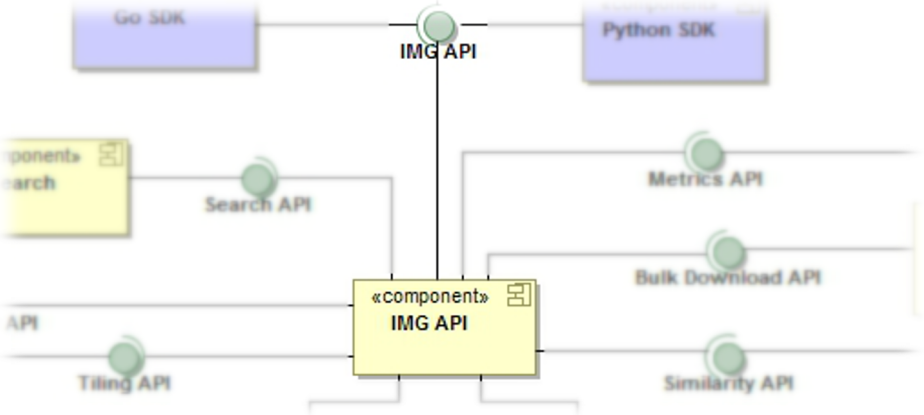
Overview

- Introduction
- Architecture
- Deployment
- Conclusions
- Future Work
- References

© 2021 California Institute of Technology. Government sponsorship acknowledged.

Deployment

PDS IMG API

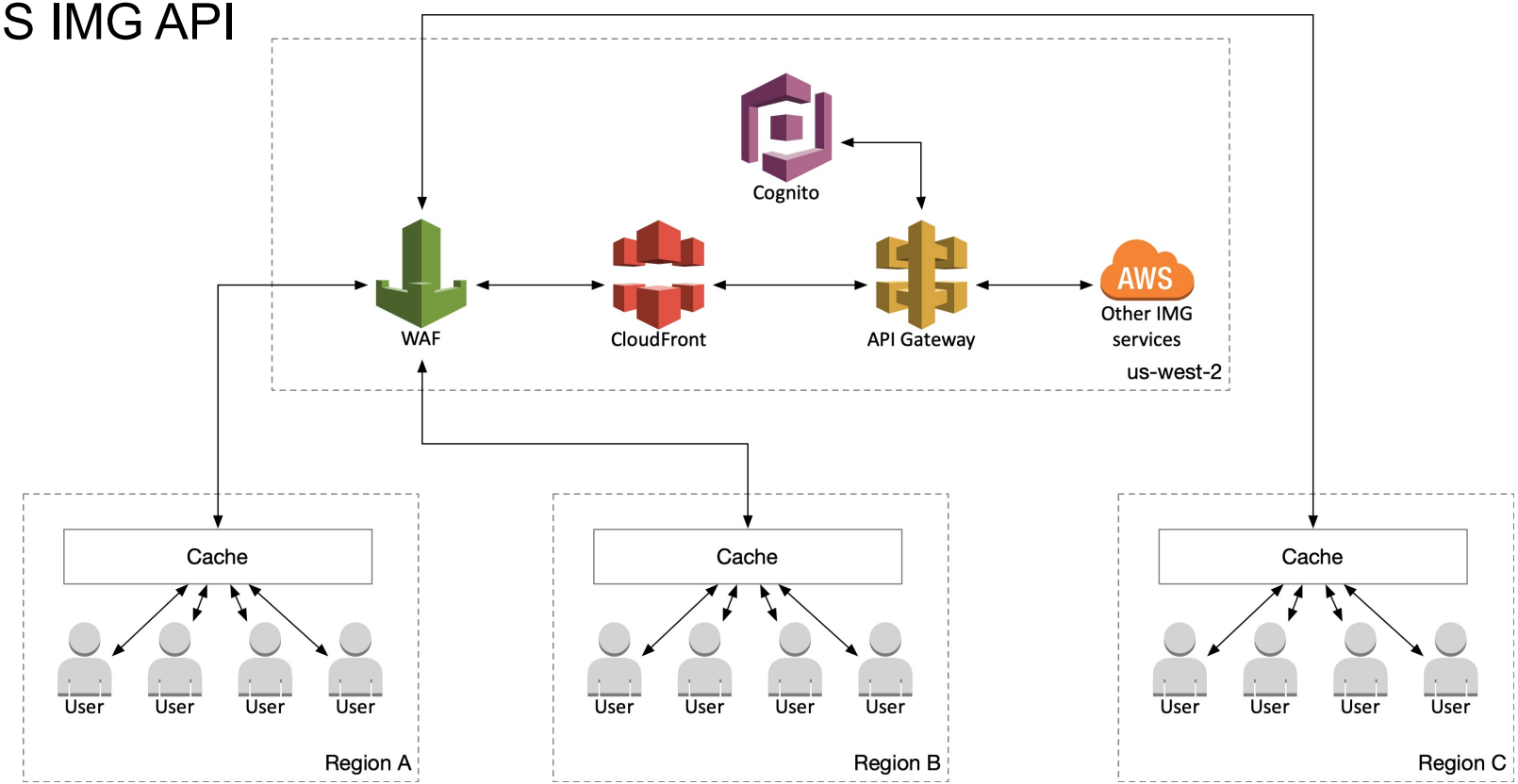


- CloudFront caching at the edge
  - Request caching
  - Per-geographical region basis
- WAF
  - Prevents application abuse
  - Configurable throttling and bypasses
- API Gateway (using OpenAPI 3.0 standard)
  - Routes requests to appropriate service
  - Supports versioning
- Amazon Cognito (tokens)
  - Integrates with different identity providers (JPL LDAP, Google, Facebook)

© 2021 California Institute of Technology. Government sponsorship acknowledged.

Deployment

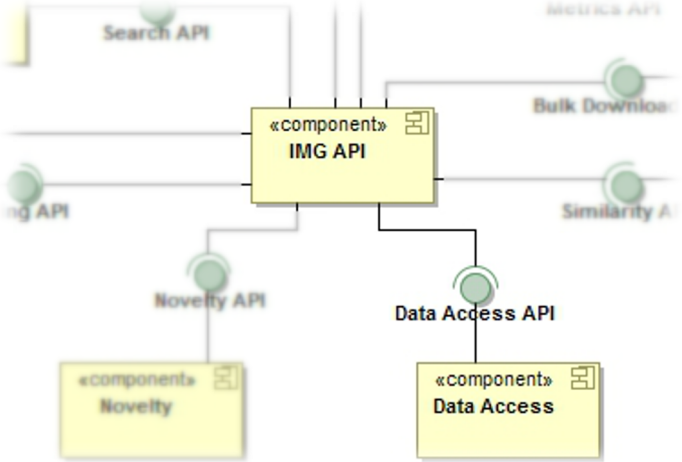
PDS IMG API



© 2021 California Institute of Technology. Government sponsorship acknowledged.

Deployment

Data Access API

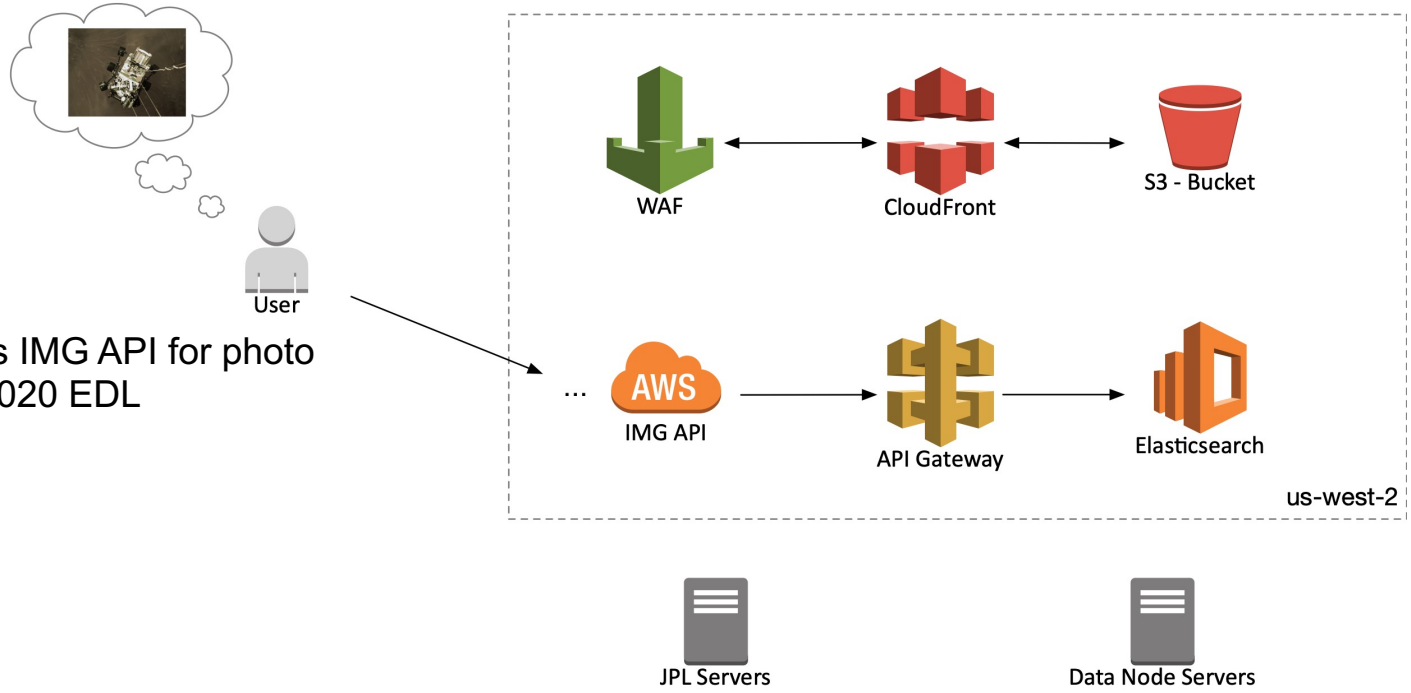


- CloudFront
  - Caching at edge
  - Data caching
  - API calls vs. multimedia transfer
  - URL signing
- WAF
  - Prevent abusive asset transfer
  - “No, you can’t just download the whole archive, sorry”
- S3 & BeeGFS for storage
- API Gateway (w/ OpenAPI 3.0) for routing

© 2021 California Institute of Technology. Government sponsorship acknowledged.

Deployment

Data Access API

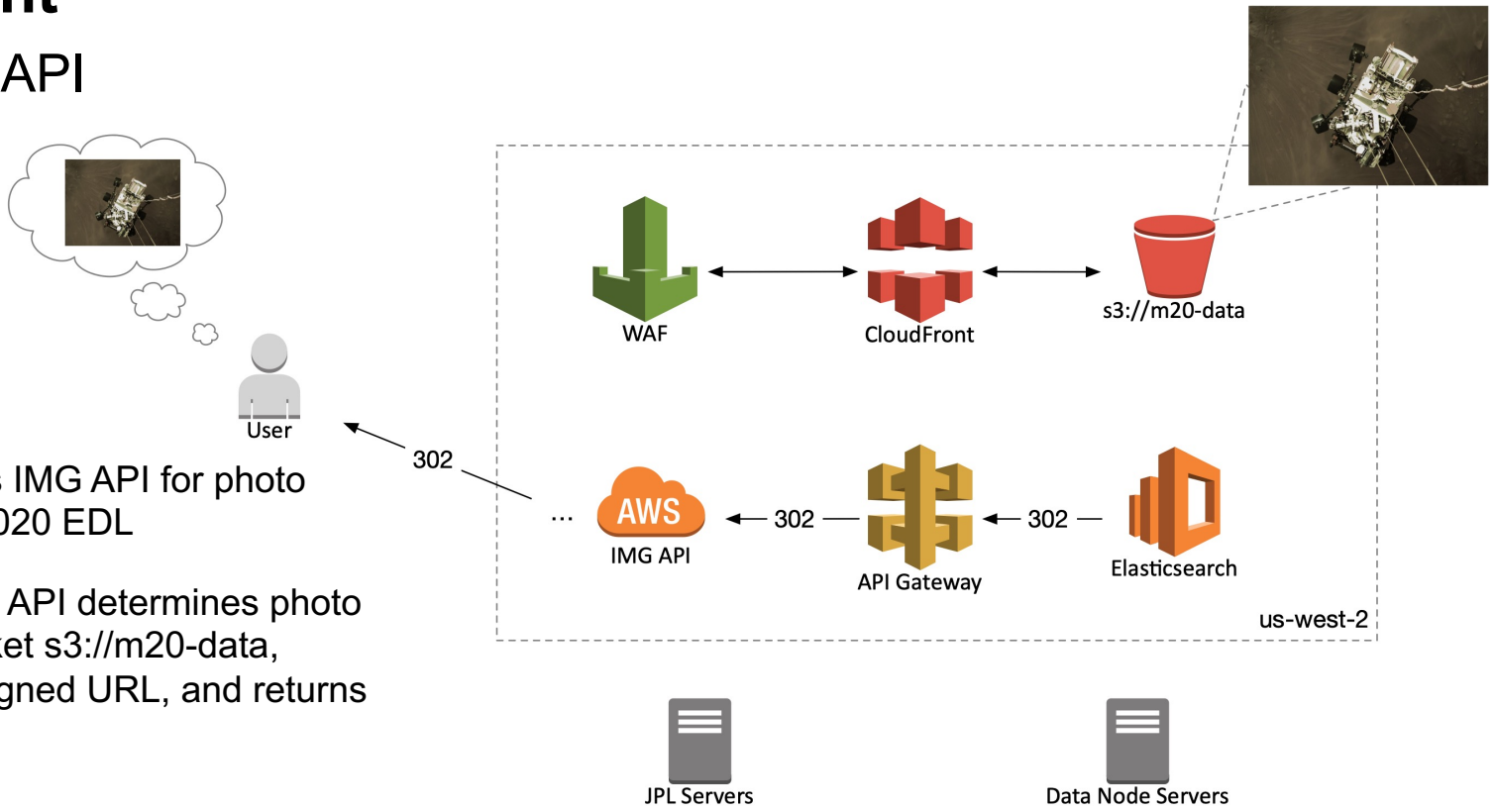


1. User queries IMG API for photo from Mars 2020 EDL

© 2021 California Institute of Technology. Government sponsorship acknowledged.

Deployment

Data Access API



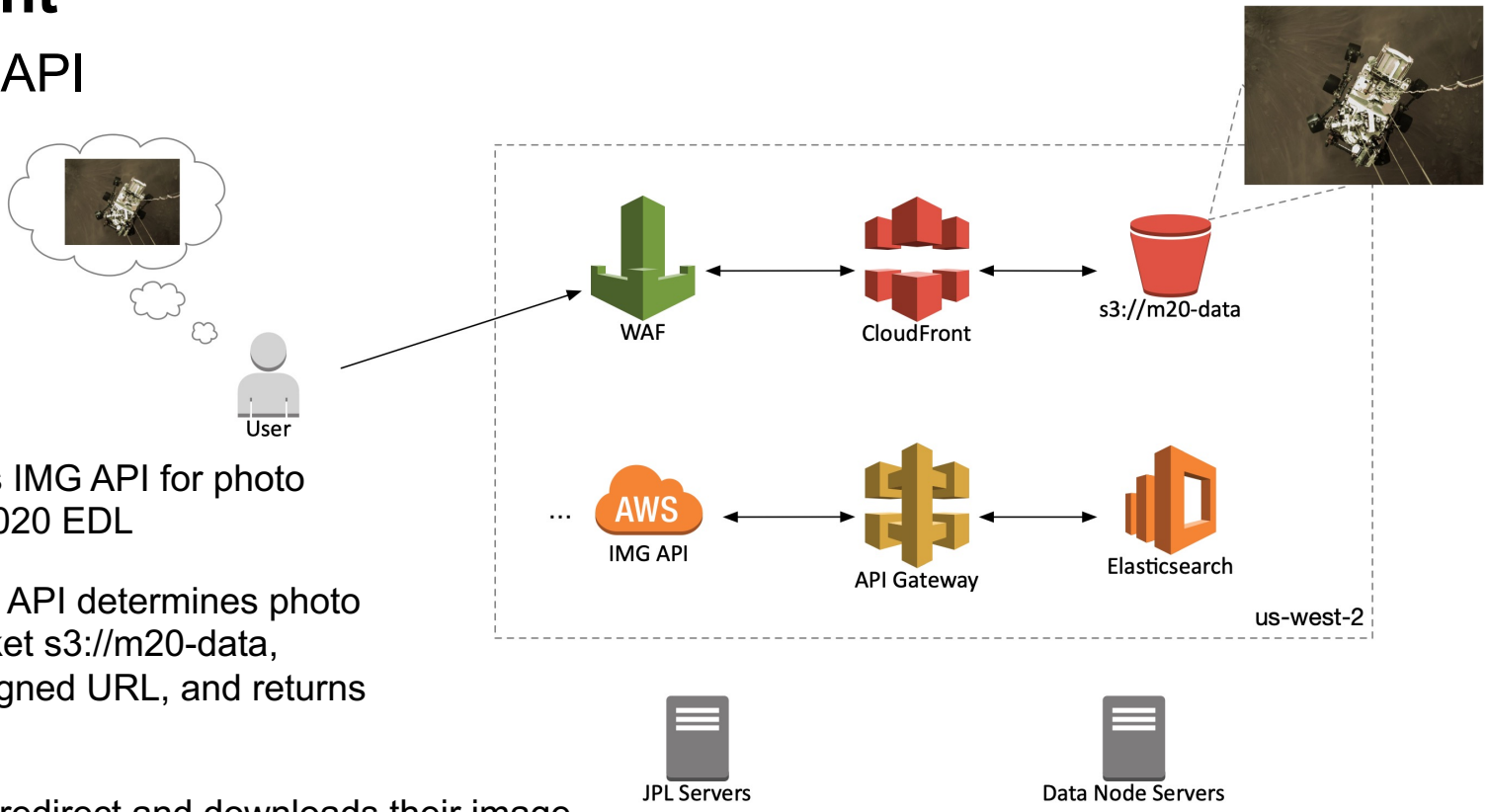
1. User queries IMG API for photo from Mars 2020 EDL

2. Data Access API determines photo is in S3 bucket s3://m20-data, generates signed URL, and returns 302 to user

© 2021 California Institute of Technology. Government sponsorship acknowledged.

Deployment

Data Access API

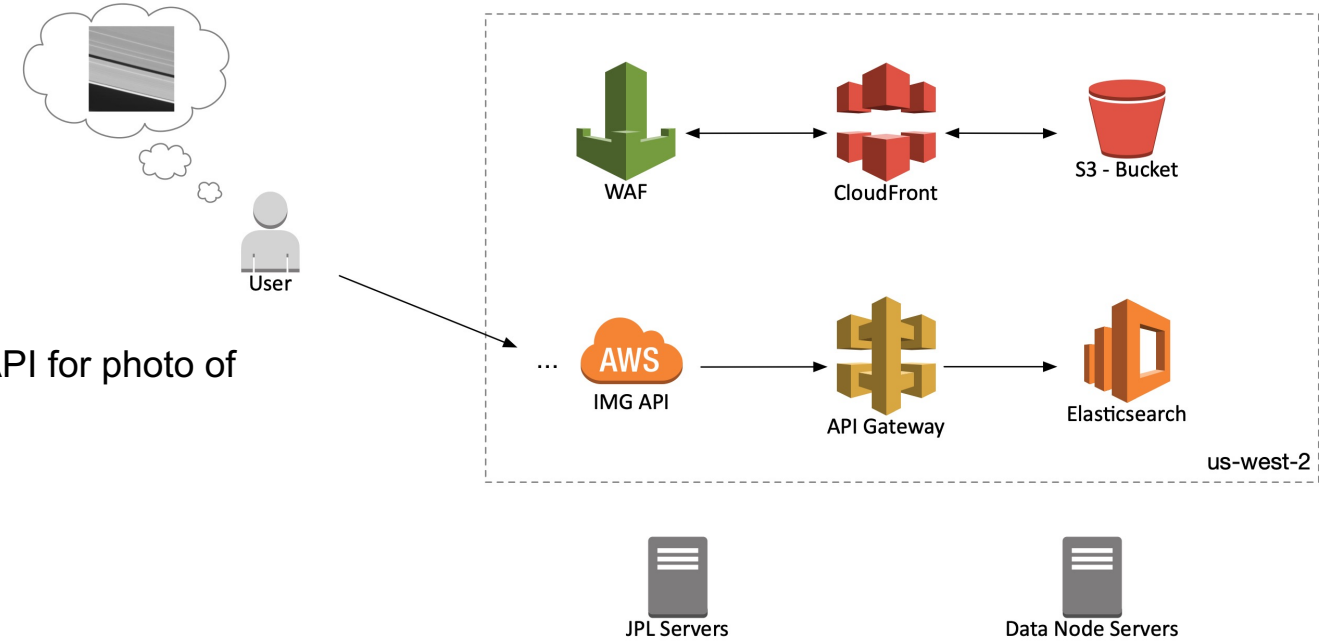


1. User queries IMG API for photo from Mars 2020 EDL
2. Data Access API determines photo is in S3 bucket s3://m20-data, generates signed URL, and returns 302 to user
3. User follows redirect and downloads their image

© 2021 California Institute of Technology. Government sponsorship acknowledged.

Deployment

Data Access API

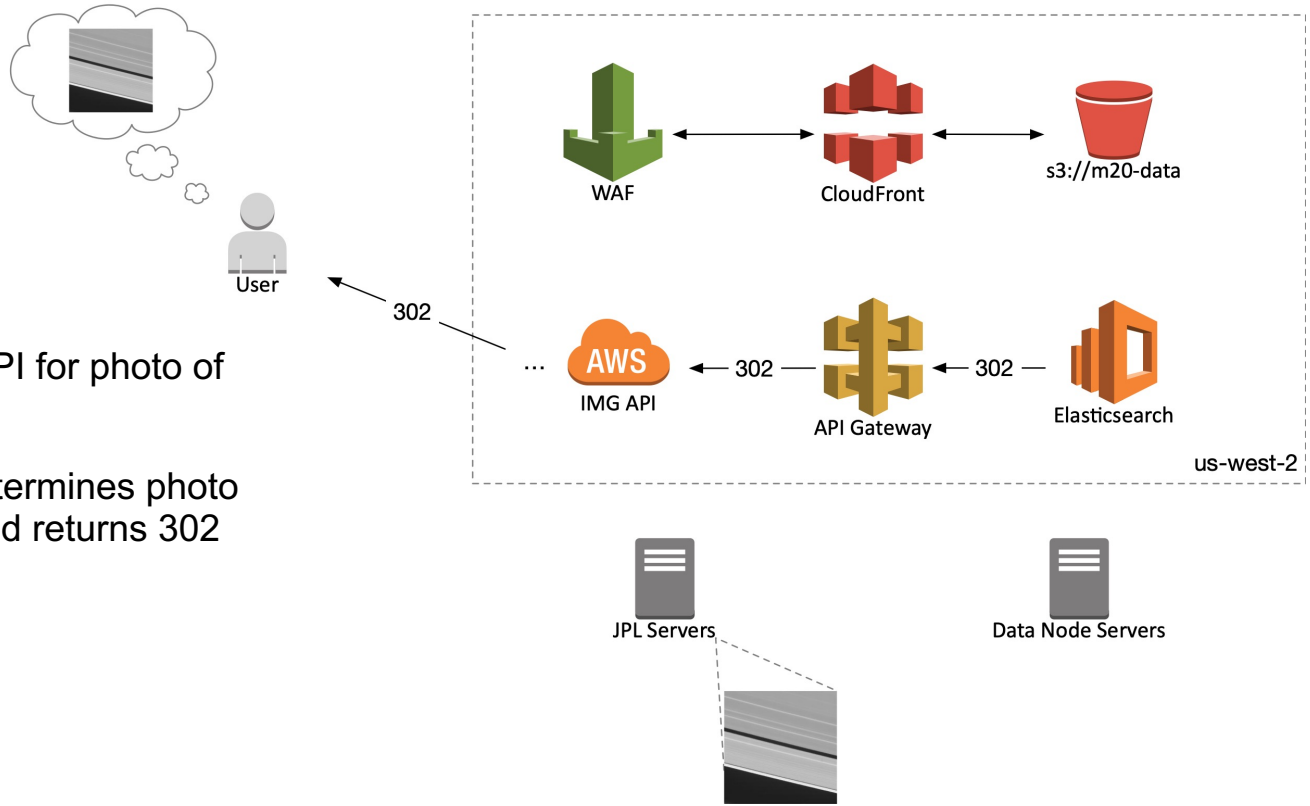


1. User queries IMG API for photo of Saturn's a-ring

© 2021 California Institute of Technology. Government sponsorship acknowledged.

Deployment

Data Access API

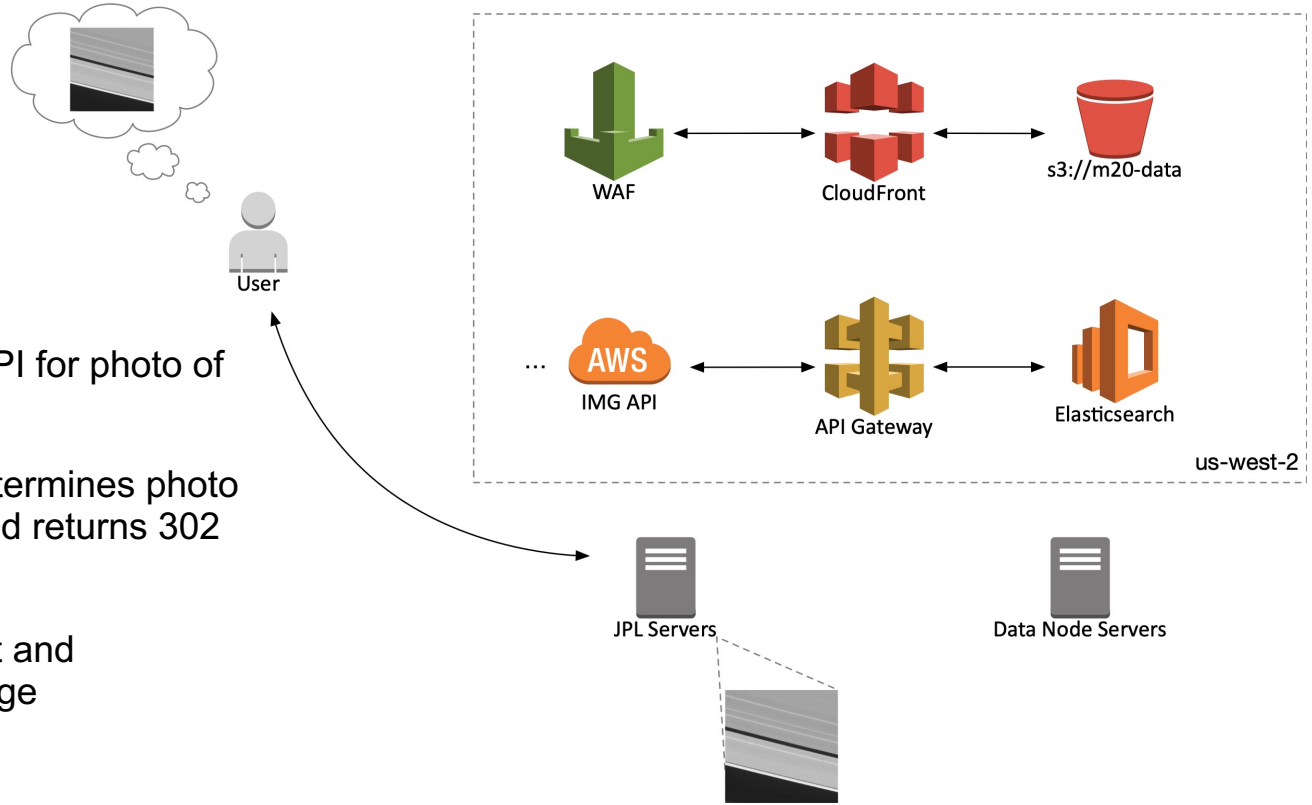


1. User queries IMG API for photo of Saturn's a-ring
2. Data Access API determines photo is on JPL servers and returns 302 to user

© 2021 California Institute of Technology. Government sponsorship acknowledged.

Deployment

Data Access API



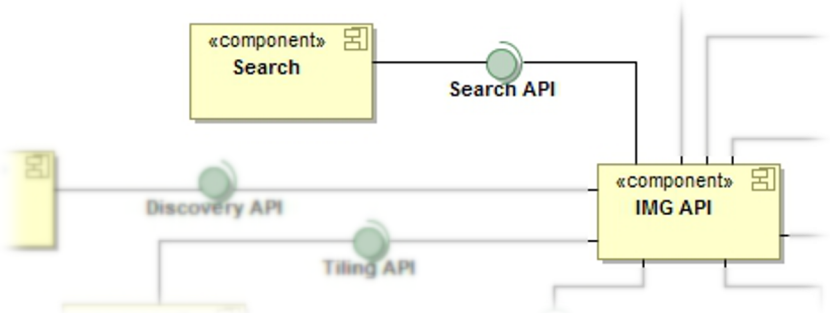
1. User queries IMG API for photo of Saturn's a-ring
2. Data Access API determines photo is on JPL servers and returns 302 to user
3. User follows redirect and downloads their image

© 2021 California Institute of Technology. Government sponsorship acknowledged.

Deployment

Search API

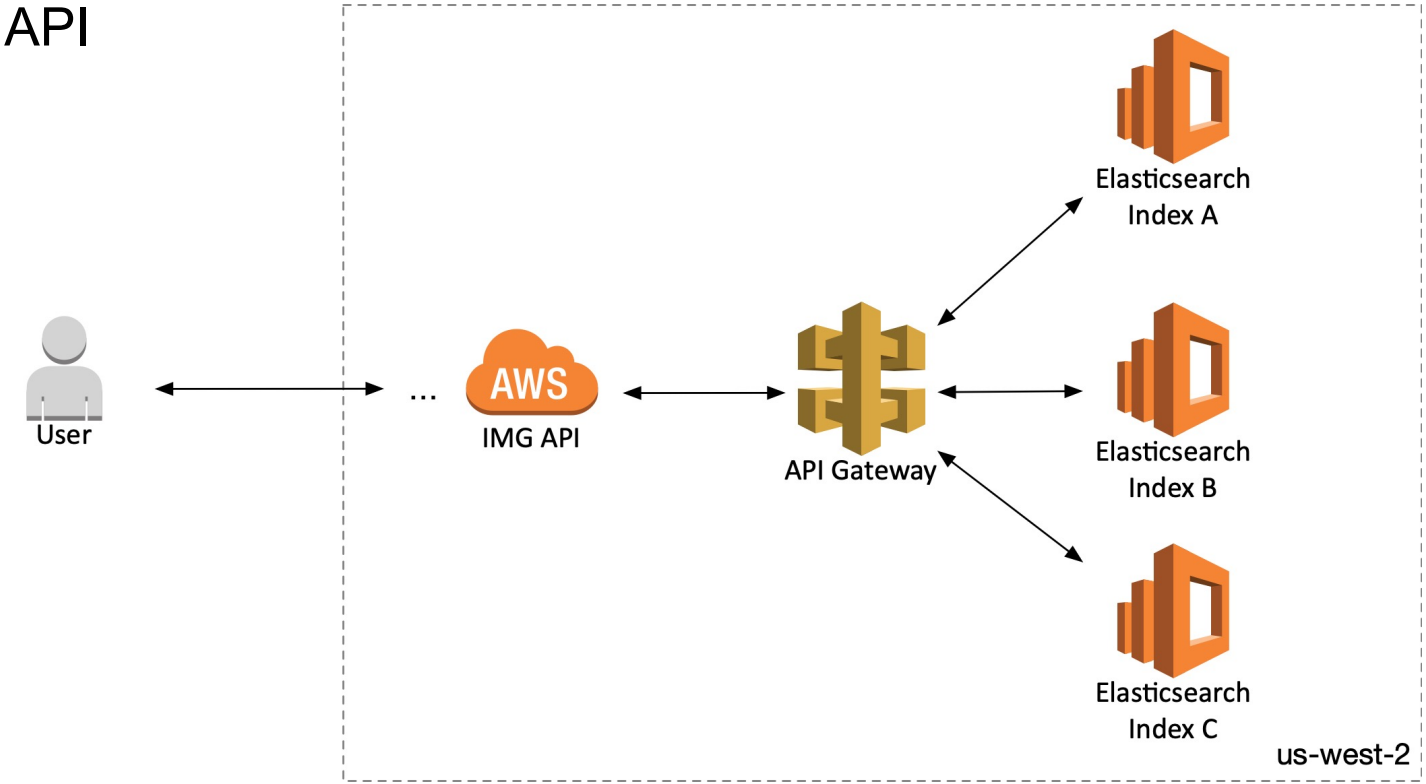
- CloudFront
- API Gateway
  - Routes to appropriate ES index
- Managed Elasticsearch
  - Single-button upgrades
  - Easy to setup and maintain



© 2021 California Institute of Technology. Government sponsorship acknowledged.

Deployment

Search API

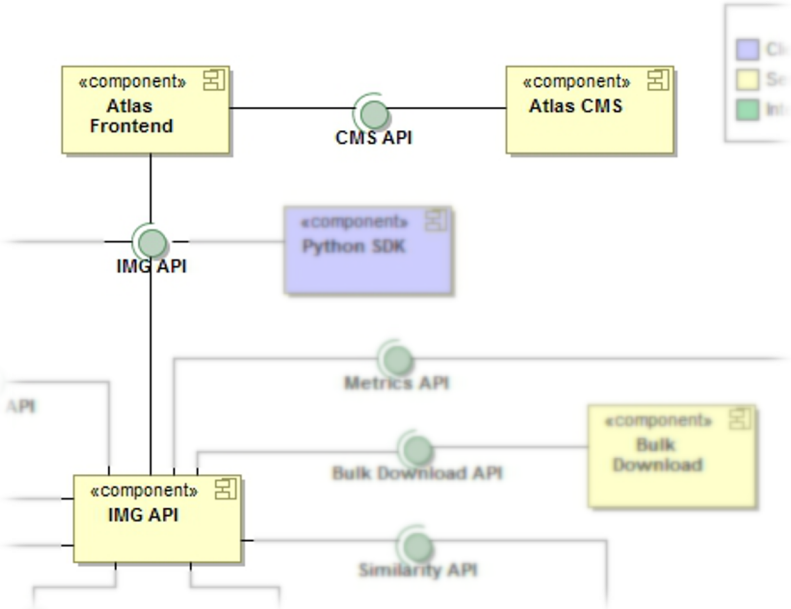


© 2021 California Institute of Technology. Government sponsorship acknowledged.

Deployment

Image Atlas Client

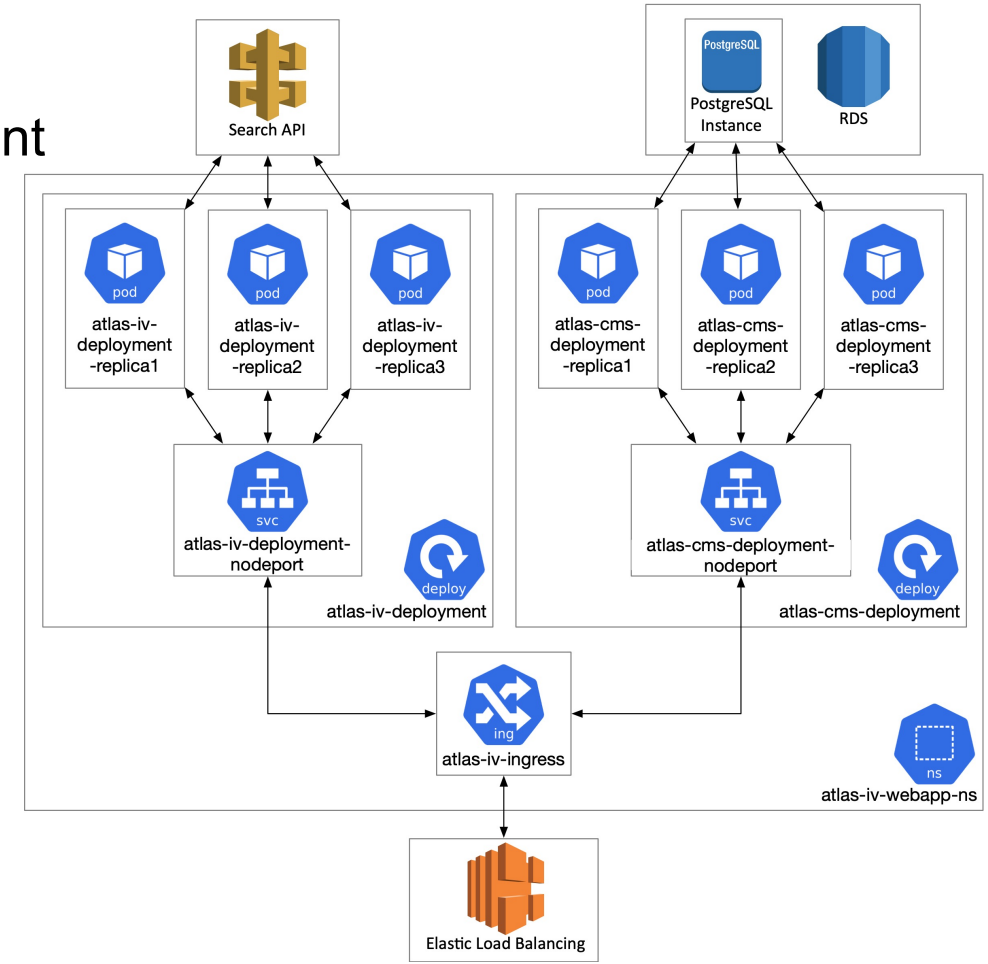
- ReactJS application built into a Docker image
- Wrapped in a Kubernetes Deployment
- Stores some state in Strapi CMS
- Queries Search API for image products to render
- Kubernetes manifests packaged as a Helm chart
- Deployed using AWS's Elastic Kubernetes Service (EKS)



© 2021 California Institute of Technology. Government sponsorship acknowledged.

Deployment

Image Atlas Client



© 2021 California Institute of Technology. Government sponsorship acknowledged.

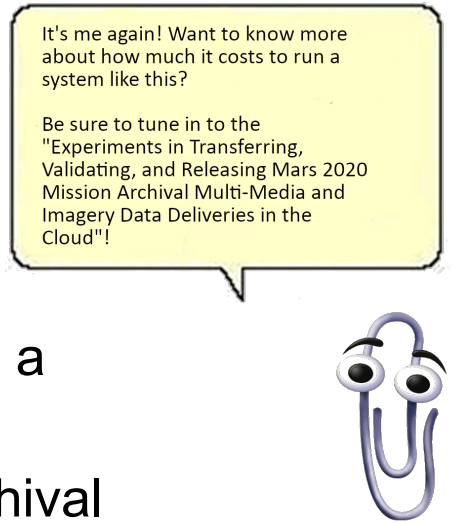
Cloud Processing of PDS Archival Products

Overview

- Introduction
- Architecture
- Deployment
- Conclusions
- Future Work
- References

© 2021 California Institute of Technology. Government sponsorship acknowledged.

Conclusions



- Microservice architectures enable diverse applications to interact with each other and form a complete system
- Cloud computing is entirely feasible for data archival workloads
- Serving data from the cloud can be cost-efficient if the proper safeguards are put in place

© 2021 California Institute of Technology. Government sponsorship acknowledged.

Cloud Processing of PDS Archival Products

Overview

- Introduction
- Architecture
- Deployment
- Conclusions
- Future Work
- References

© 2021 California Institute of Technology. Government sponsorship acknowledged.

Future Work

- Integrate with PDS API being developed by ENG
- Develop SDKs to interact with various API features (Python libraries, Go modules, etc)
- Additional API work
  - Bulk download
  - Metrics

© 2021 California Institute of Technology. Government sponsorship acknowledged.

Cloud Processing of PDS Archival Products

Overview

- Introduction
- Architecture
- Deployment
- Conclusions
- Future Work
- References

© 2021 California Institute of Technology. Government sponsorship acknowledged.

References

"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

"Firefox logo". https://commons.wikimedia.org/wiki/File:Firefox_Logo,_2017.png. The Mozilla Foundation, MPL 2 <https://www.mozilla.org/en-US/MPL/2.0/>, via Wikimedia Commons.

"Chrome logo". https://commons.wikimedia.org/wiki/File:Google_Chrome_icon_(September_2014).svg. Google, Public domain, via Wikimedia Commons.

"Apache Solr logo". https://commons.wikimedia.org/wiki/File:Apache_Solr_logo.svg. Apache Software Foundation, Apache License 2.0.

"Safari logo". https://commons.wikimedia.org/wiki/File:Safari_browser_logo.svg. Apple Inc.Vectorization: CMetalCore, Public domain.

"BeeGFS logo". https://commons.wikimedia.org/wiki/File:BeeGFS-Logo.png. ThinkParQ, CC BY-SA 4.0.

"Brave logo". https://commons.wikimedia.org/wiki/File:Brave_icon_lionface.png. derivative work: Kreuzschnabel, Public domain.

"Python logo". https://commons.wikimedia.org/wiki/File:Python-logo-notext.svg. www.python.org, GPL.

"Drawbridge and portcullis". https://commons.wikimedia.org/wiki/File:The_drawbridge_and_portcullis_of_Castle_Rushen_-_geograph.org.uk_-_2109775.jpg. Shazz, CC BY-SA 2.0.

"Signpost". https://commons.wikimedia.org/wiki/File:Signpost_(3199024949).jpg. Tim Green from Bradford, CC BY 2.0.

"Thinking with magnifying glass". https://commons.wikimedia.org/wiki/File:-Bearded_Man_with_Magnifying_Glass_Examining_a_Manuscript-_MET_DP111349.jpg. Metropolitan Museum of Art, CC0.

"Clippy". Microsoft. Public domain.

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

© 2021 California Institute of Technology. Government sponsorship acknowledged.

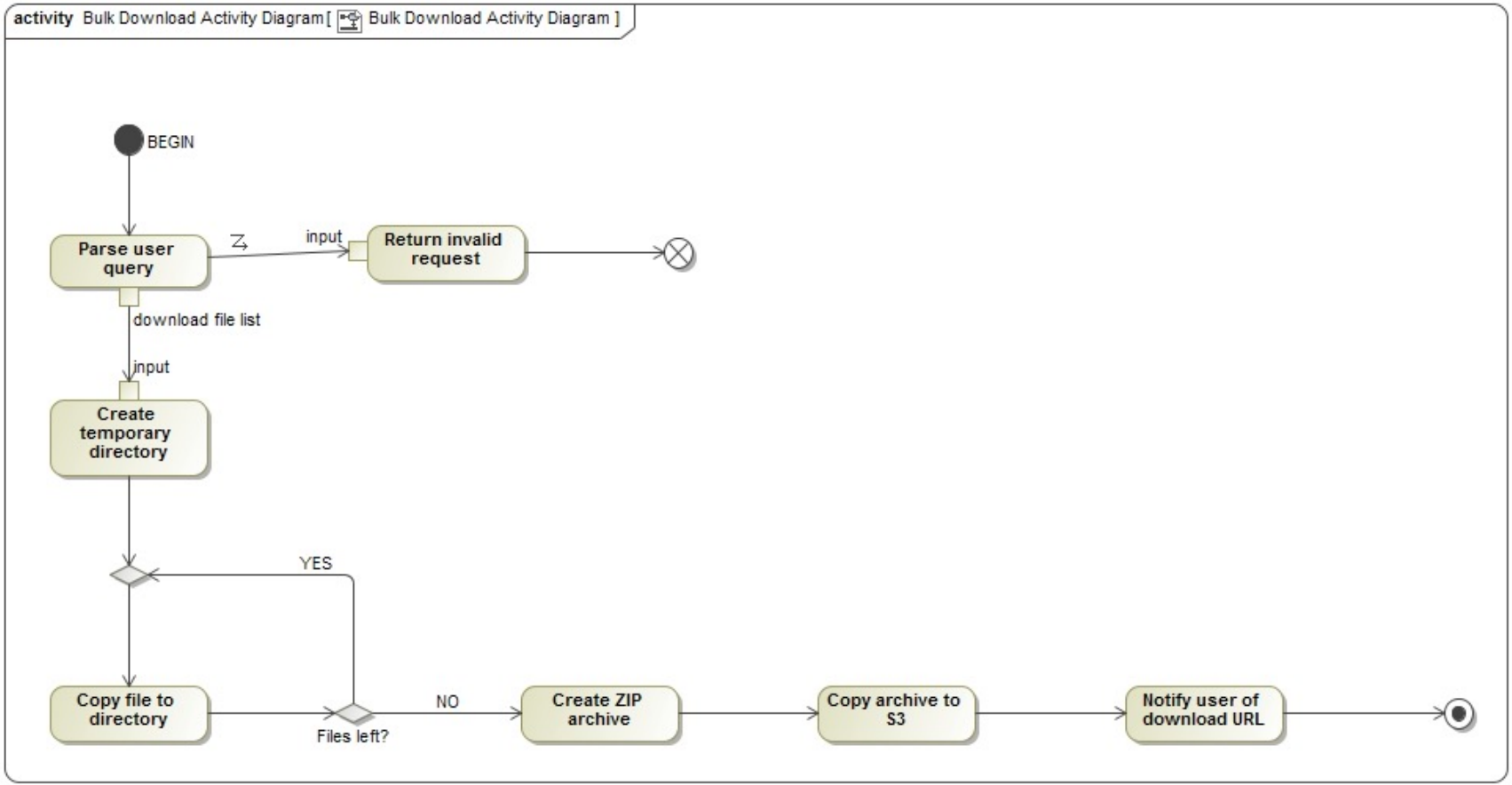
Jet Propulsion Laboratory
California Institute of Technology

jpl.nasa.gov

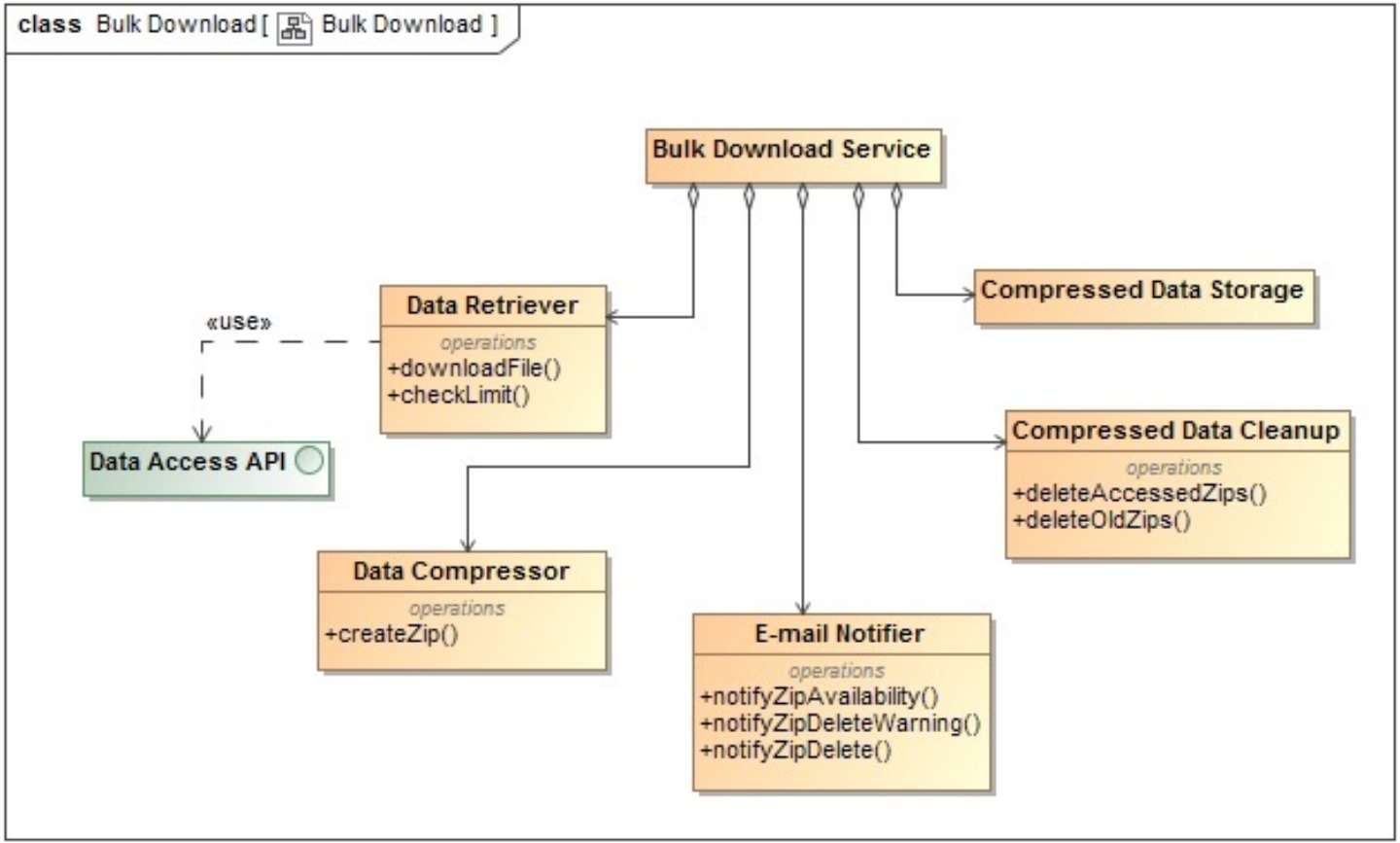
© 2021 California Institute of Technology. Government sponsorship acknowledged.

Backup

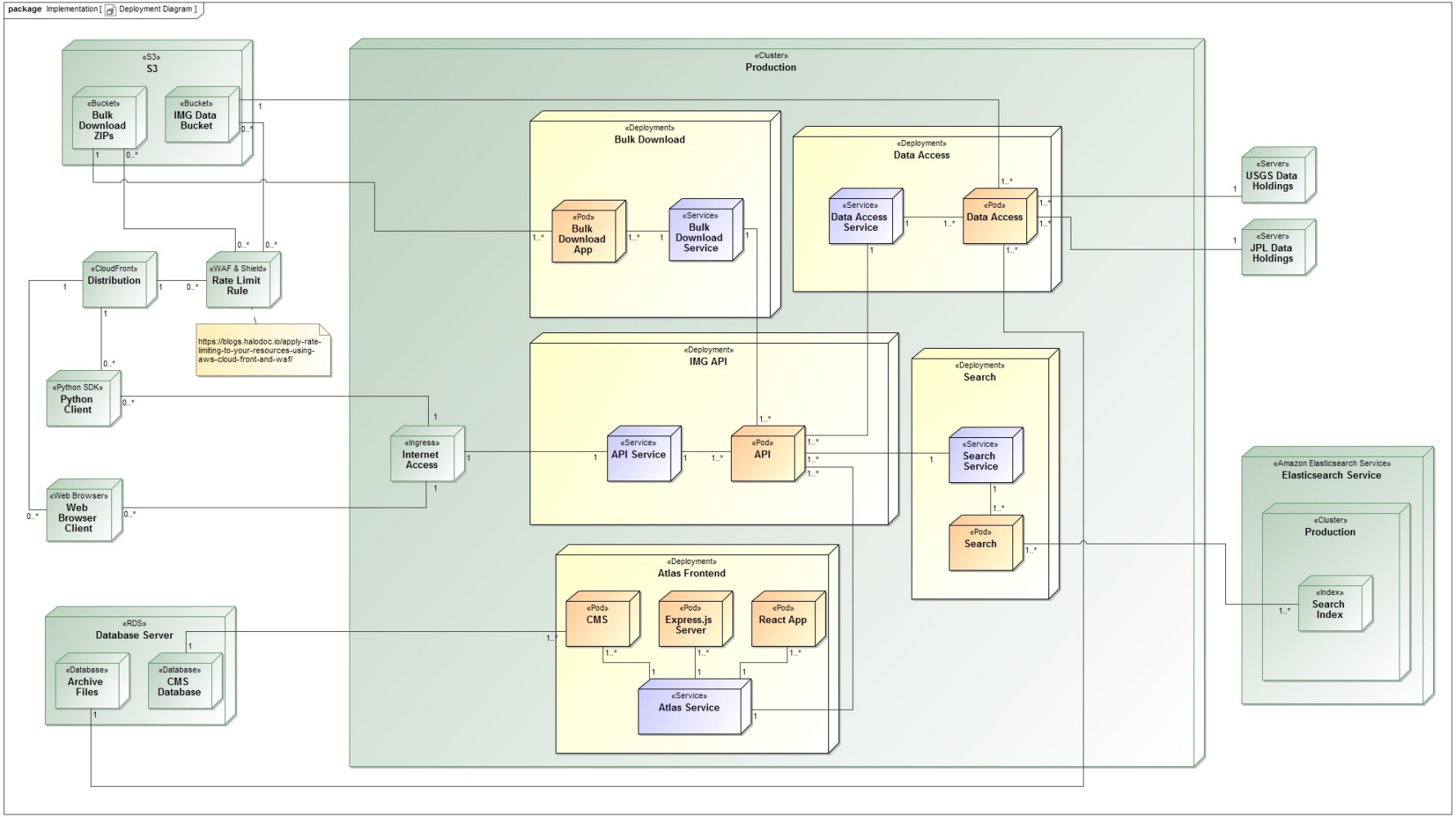
© 2021 California Institute of Technology. Government sponsorship acknowledged.



© 2021 California Institute of Technology. Government sponsorship acknowledged.



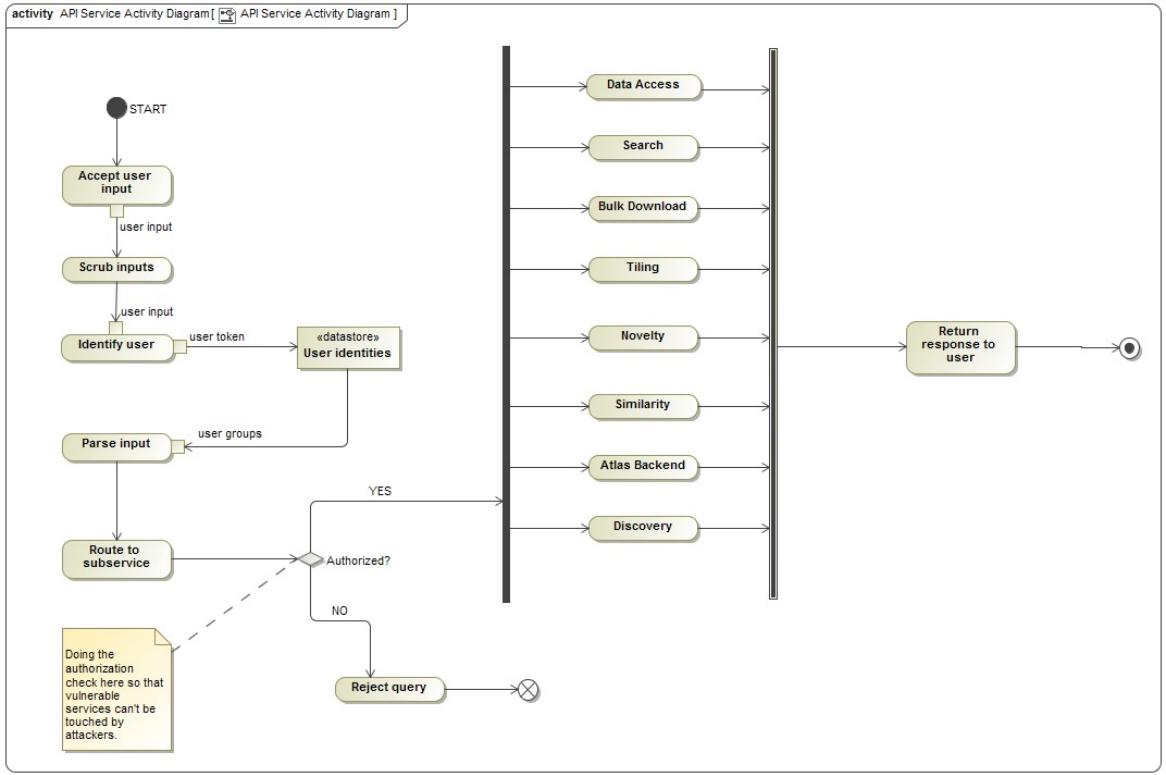
© 2021 California Institute of Technology. Government sponsorship acknowledged.



© 2021 California Institute of Technology. Government sponsorship acknowledged.

Architecture

PDS IMG API



© 2021 California Institute of Technology. Government sponsorship acknowledged.

Architecture

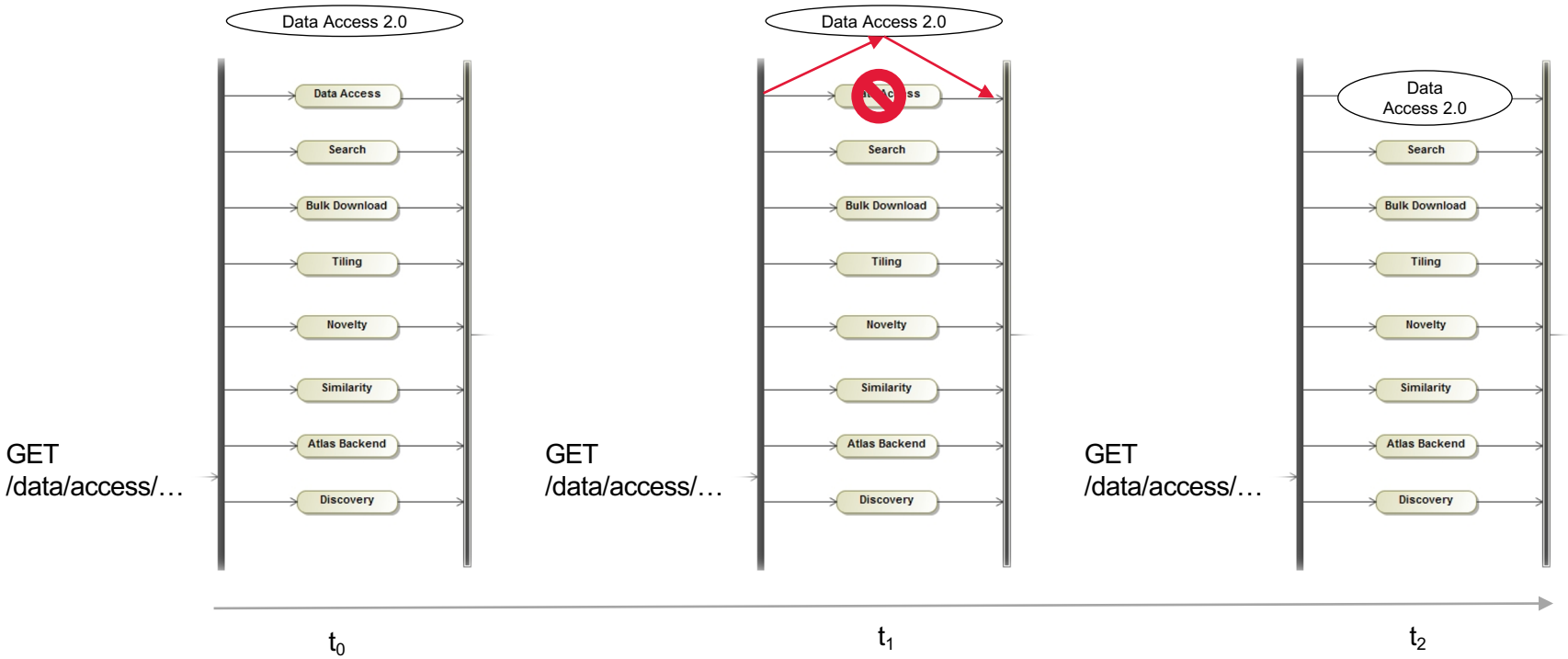
PDS IMG API

Scenario: upgrade “Data Access Service” from v1.0 to v2.0

© 2021 California Institute of Technology. Government sponsorship acknowledged.

Architecture

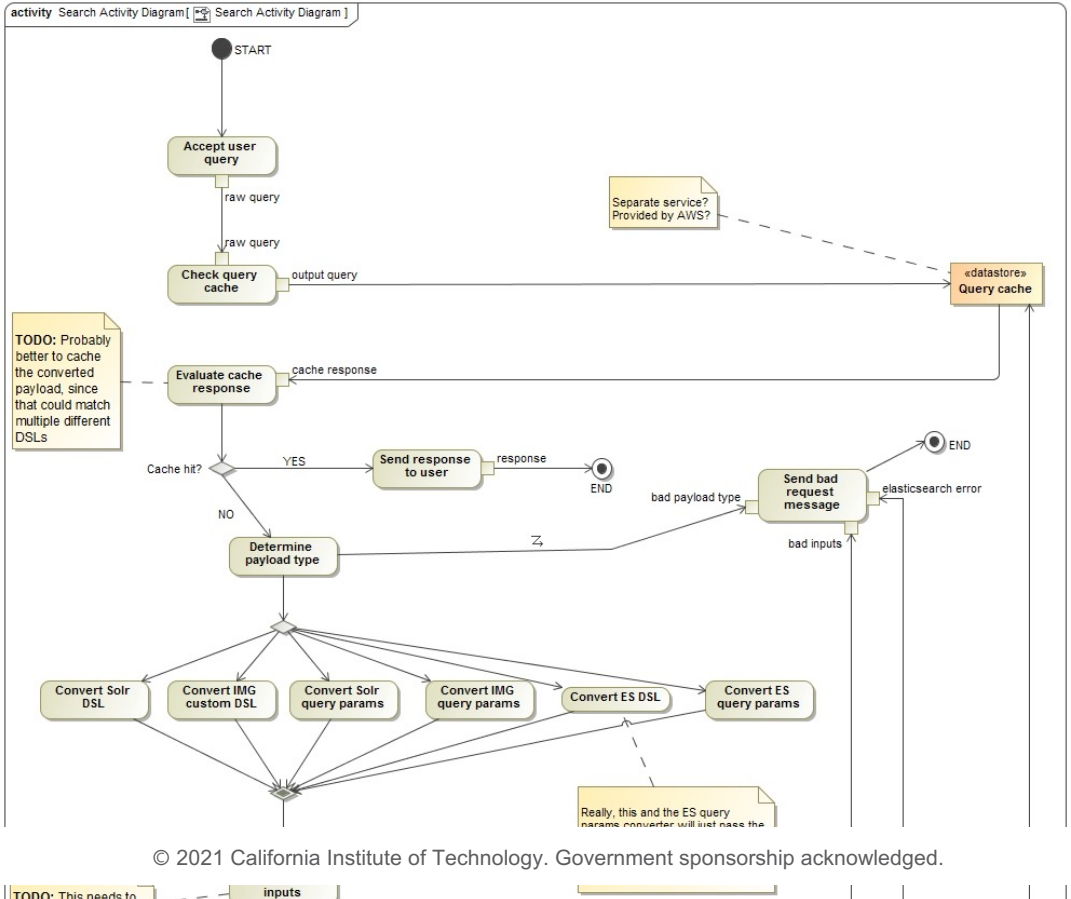
PDS IMG API



© 2021 California Institute of Technology. Government sponsorship acknowledged.

Architecture

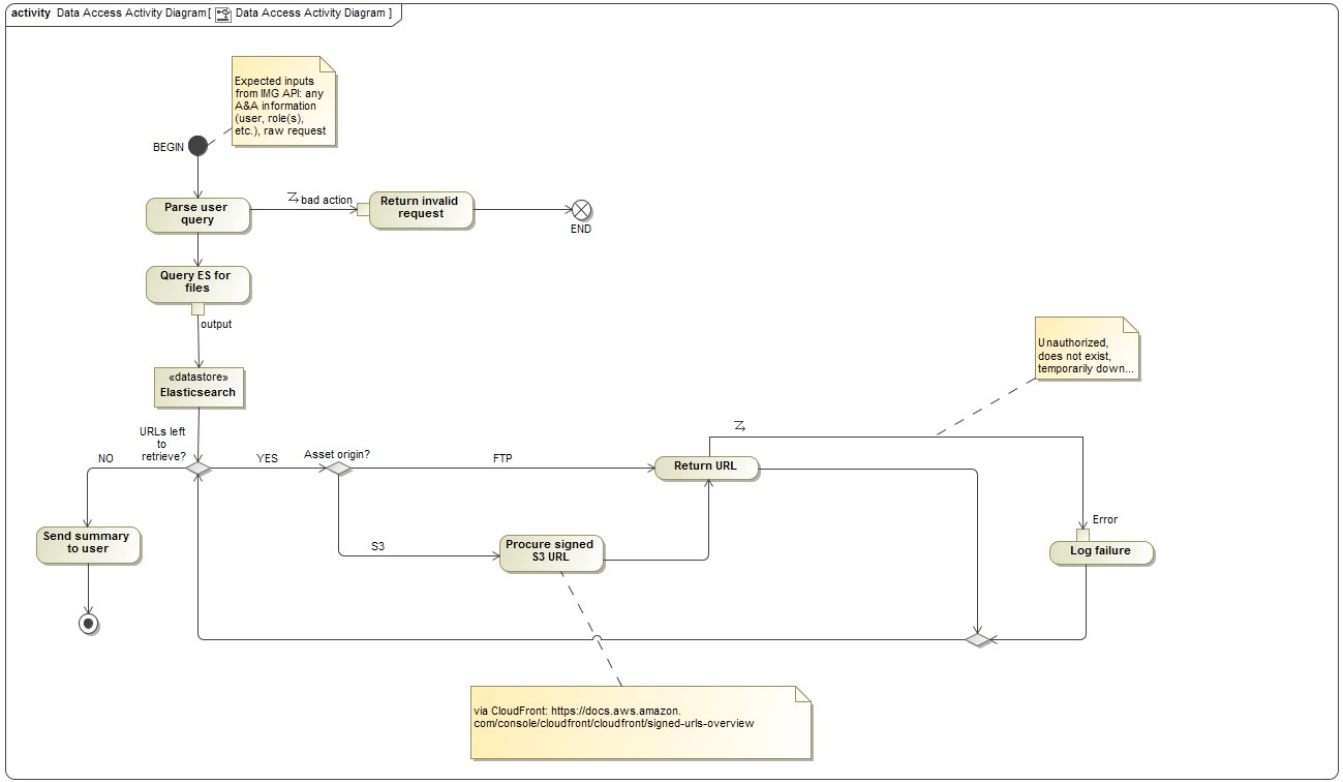
Search API



© 2021 California Institute of Technology. Government sponsorship acknowledged.

Architecture

Data Access API



© 2021 California Institute of Technology. Government sponsorship acknowledged.

Jet Propulsion Laboratory
California Institute of Technology

jpl.nasa.gov

© 2021 California Institute of Technology. Government sponsorship acknowledged.