GSAW 2022

Earth Observing Data Distribution and Visualization Using the Cloud - Overview

Author:
Jay Pennington

Co-Authors:
Morgan Williams
Will Coffey
Spencer Drakontaidis
Alan Christopher
Phillip Jasper

February 23, 2022

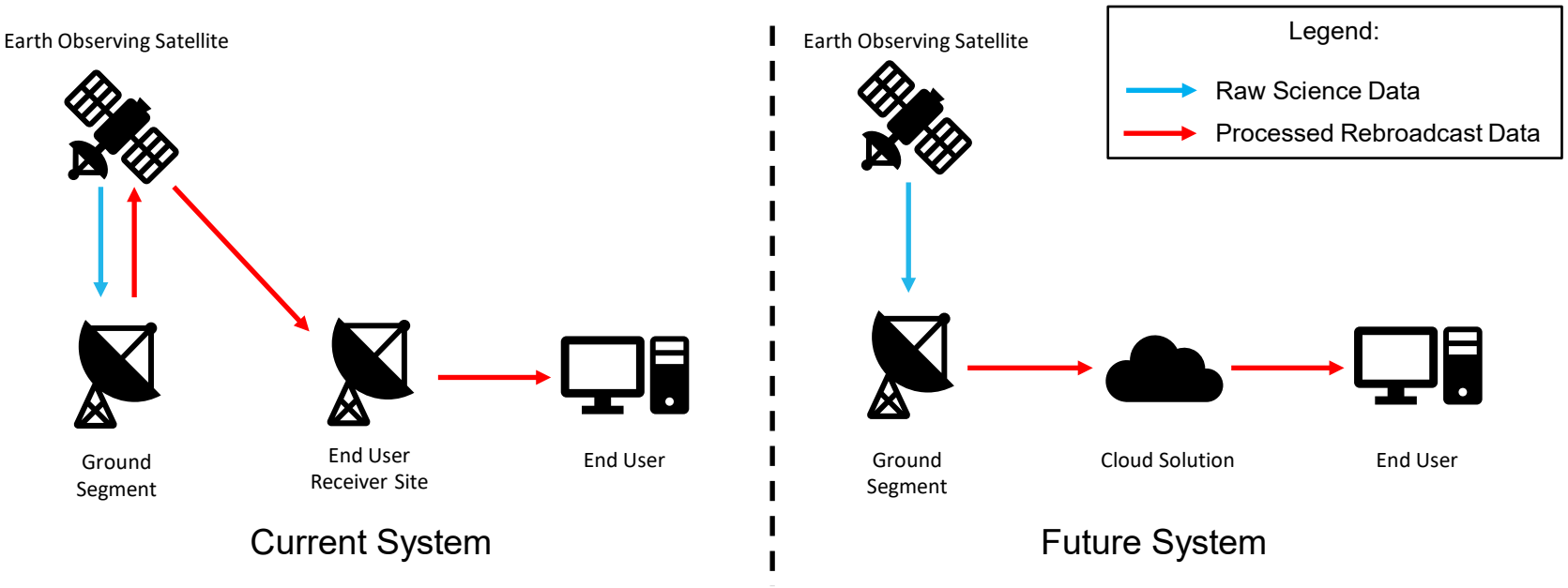
Cloud Introduction

- Cloud technologies
  - *On-demand resources that are managed for the user*
  - *"Running software on someone else's hardware"*
- Cloud is being widely adopted by many industries
  - *Tech, finance, manufacturing, automotive, healthcare, gaming, etc.*
- Many vendor options available to government sector
  - *Amazon Web Services (AWS)*
  - *Microsoft Azure*
  - *Google Cloud Platform (GCP)*
  - *IBM Cloud*
  - *Oracle Cloud*

But…Why Cloud?

- Cloud services use virtual infrastructure
  - *Resources can be deployed in real time on an as-needed basis*
  - *Rapid deployments allow for dynamic scaling of capacity to meet demand*
  - *Infrastructure can be treated as software through Infrastructure as Code (IaC)*
  - *Infrastructure can be managed remotely*
- Customers should focus on competitive advantages
  - *Software development offers competitive advantage, but related operations do not*
    - Building, securing, and maintaining physical infrastructure
  - *Use of cloud services allows developers to spend more time writing software*
- New technologies are constantly being released
  - *Exploration and experimentation leads to rapid innovation*

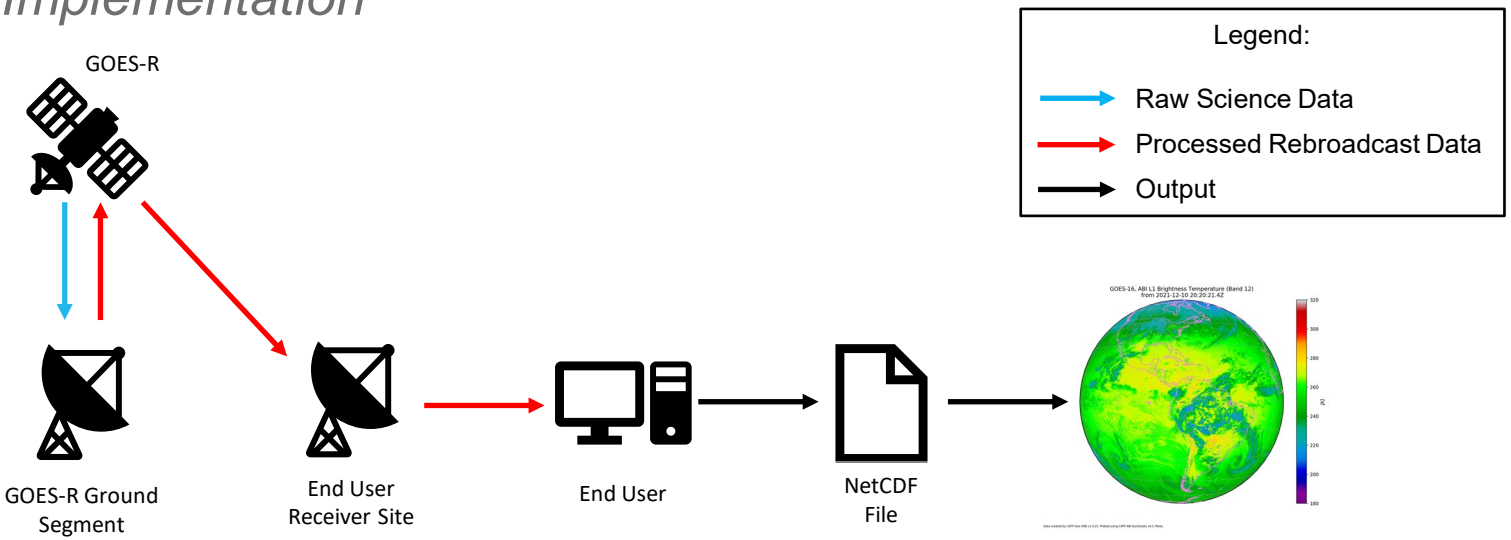
Satellite Data Rebroadcast



- Mission: earth observing science data streaming
  - *Data is collected, processed, and rebroadcast to end users*
- Problem: both satellite and end user require dedicated antennas
  - *Satellite rebroadcast and end user receiver antennas can be costly*
- Hypothesis: processed data can be rebroadcast via the cloud
  - *Remove rebroadcast antenna to reduce size, weight, and power on next-gen satellites*
  - *Remove the need for end users to procure and maintain receiver antennas*
- Aerospace is building an AWS Cloud Pilot to replicate this chain of operations
  - *Working to identify risks or concerns with cloud approach to data distribution*

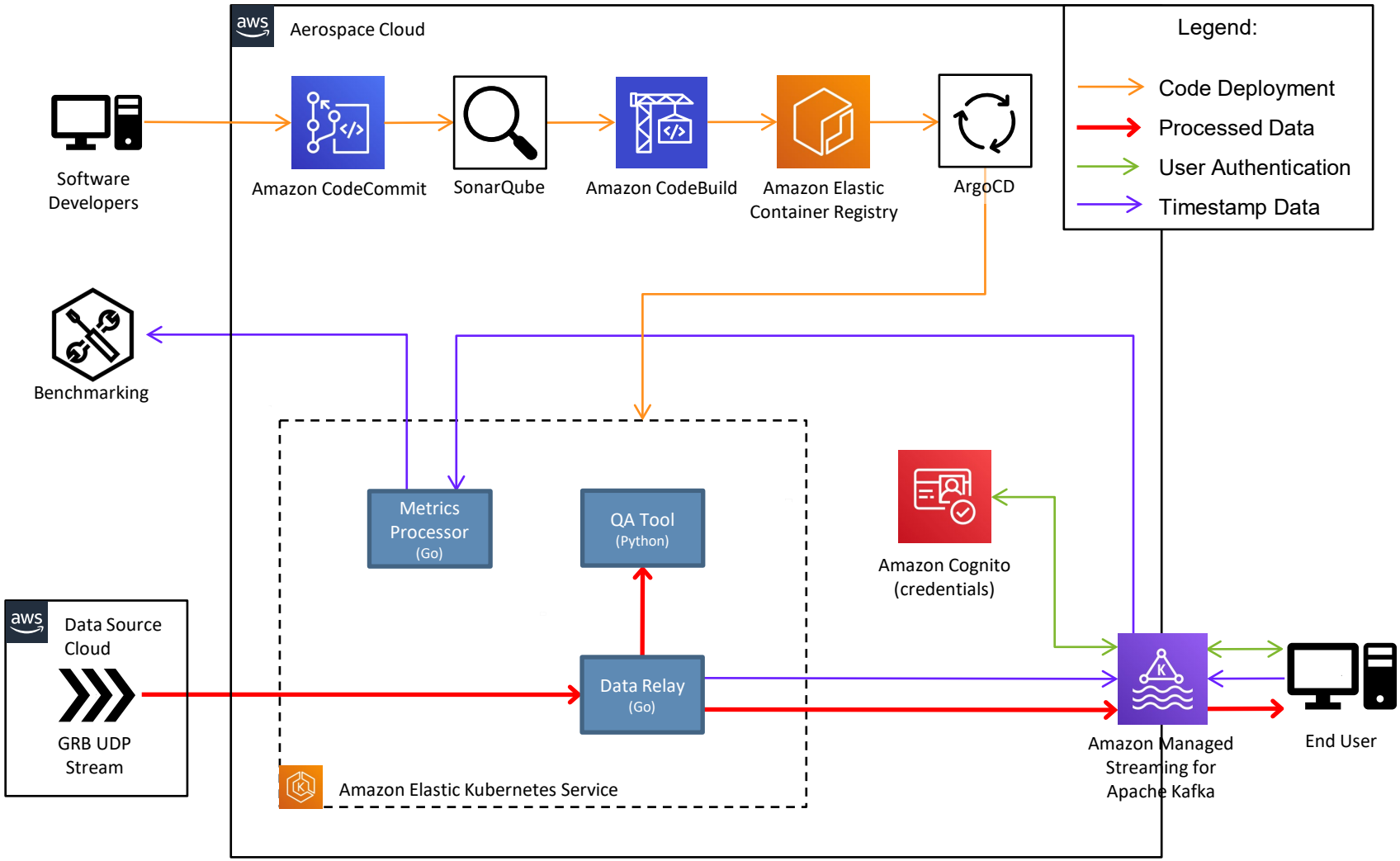
GOES-R Rebroadcast

Current Implementation



- GOES-R Rebroadcast (GRB)
  - *Data stream containing ABI, GLM, space environment, and solar data*
  - *GRB footprint covers large portions of the Western hemisphere*
  - *Users who have access to the GRB receiver antenna can capture this data*
- GRB data uses
  - *Users convert files to Network Common Data Format (NetCDF) files*
  - *NetCDF files can be used for scientific studies or converted into visualizations*
- Aerospace is using GRB to demonstrate capabilities of cloud data distribution

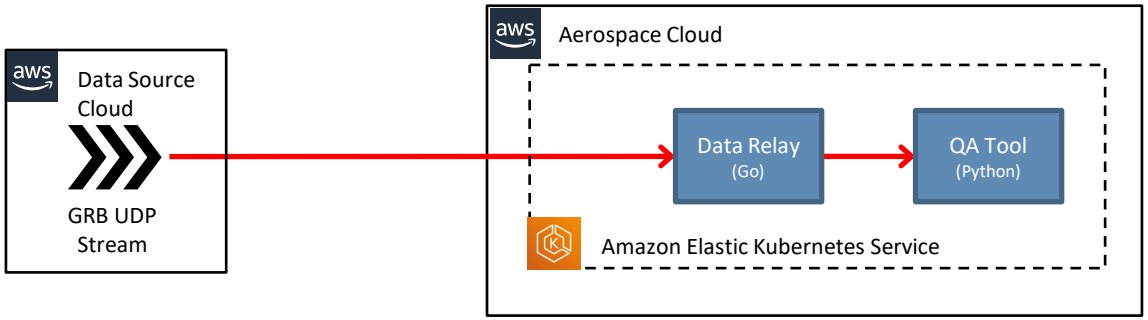
Data Streaming Prototype Architecture



Cloud Pilot Capabilities

Monitor quality of data stream

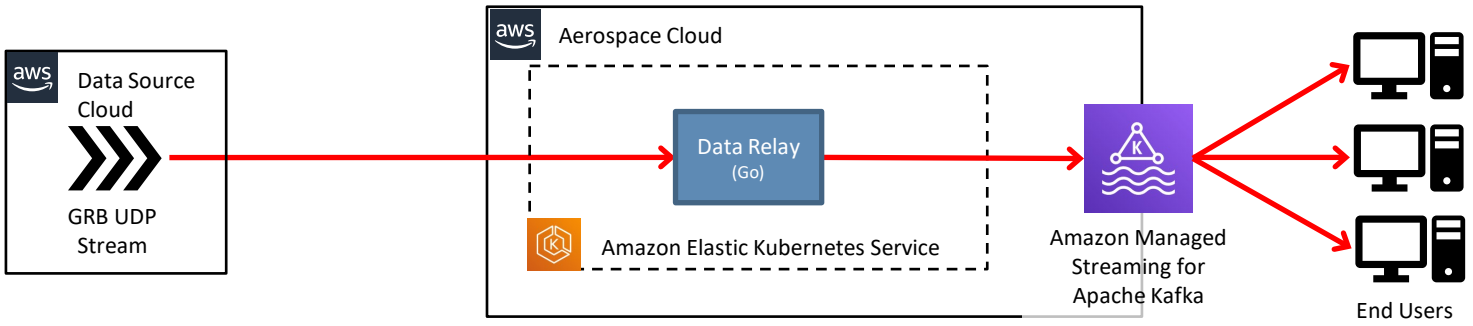
- Goal:
  - *Ensure data quality by measuring incoming GRB stream availability*
- Capability:
  - *Custom quality assurance (QA) tool tracks sequence count of packets*
    - Non-sequential packets indicates packet loss has occurred
    - Tool does not currently account for out-of-order or duplicate packets
  - *Tool can run in both cloud and on-premises environments*
  - *Tool can run at multiple points in the data stream to track quality*



Cloud Pilot Capabilities

Delivering data to multiple users

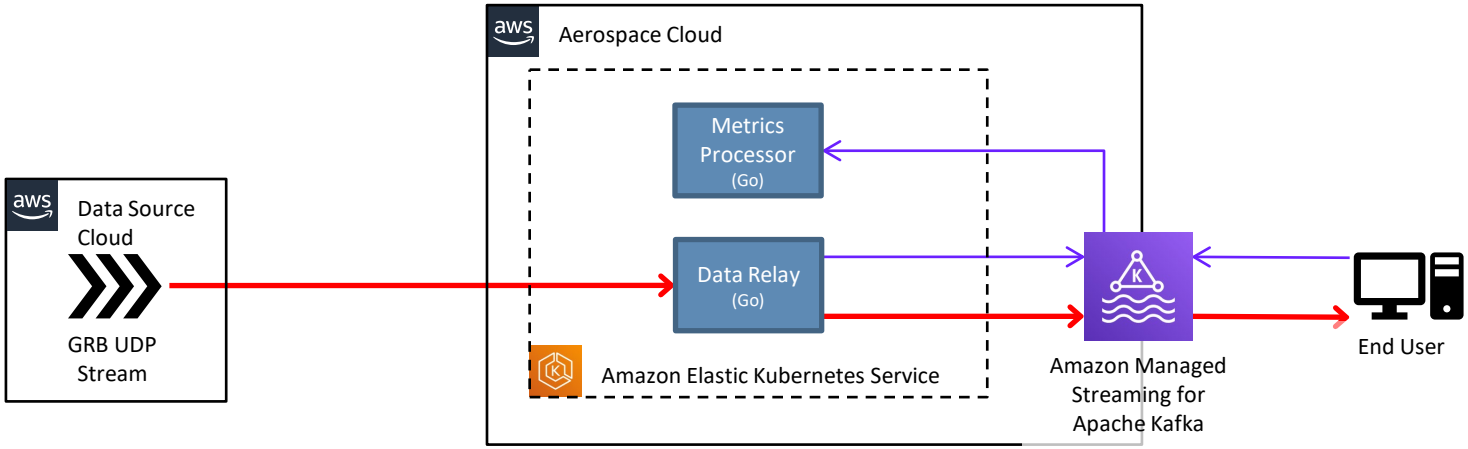
- Goal:
  - *Deliver incoming data packets to end users efficiently*
  - *Scale infrastructure for as many data streams or end users as needed*
- Capability:
  - *Amazon Managed Streaming for Apache Kafka (MSK)*
  - *Highly-scalable publish/subscribe (pub/sub) service commonly used for data transfer*
  - *Free and open-source Kafka technology provides portability between vendors*
  - *Users authenticate via Amazon Cognito to access stream*



Cloud Pilot Capabilities

Benchmark latency and availability metrics

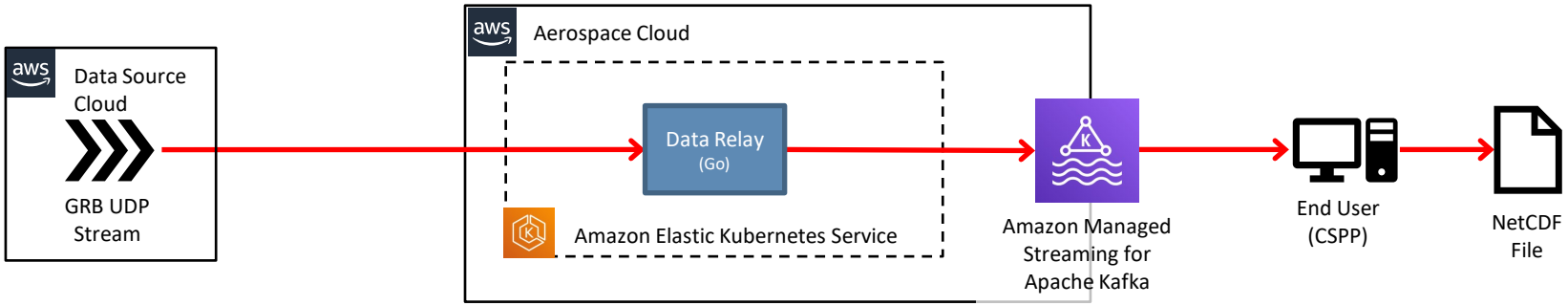
- Goal:
  - *Develop software to measure availability and latency of data delivery via cloud*
  - *Measure cloud performance to understand adoption risks and concerns*
- Capability:
  - *Metrics collected, aggregated, and processed in real time for each GRB packet*
  - *Metrics streaming performed via Amazon MSK to simplify architecture*



Cloud Pilot Capabilities

Delivering data in useful formats

- Goal:
  - *Deliver pre-processed data to end user in a useful format*
  - *Leverage publicly available software commonly used by end users*
- Capability:
  - *University of Wisconsin's Community Satellite Processing Package (CSPP)*
  - *Common free and open-source Java application deployed by nearly all end users*
  - *Converts RF GRB streaming data into usable NetCDF files in real-time*
  - *Modified to convert non-RF GRB stream into useable NetCDF files in real-time*

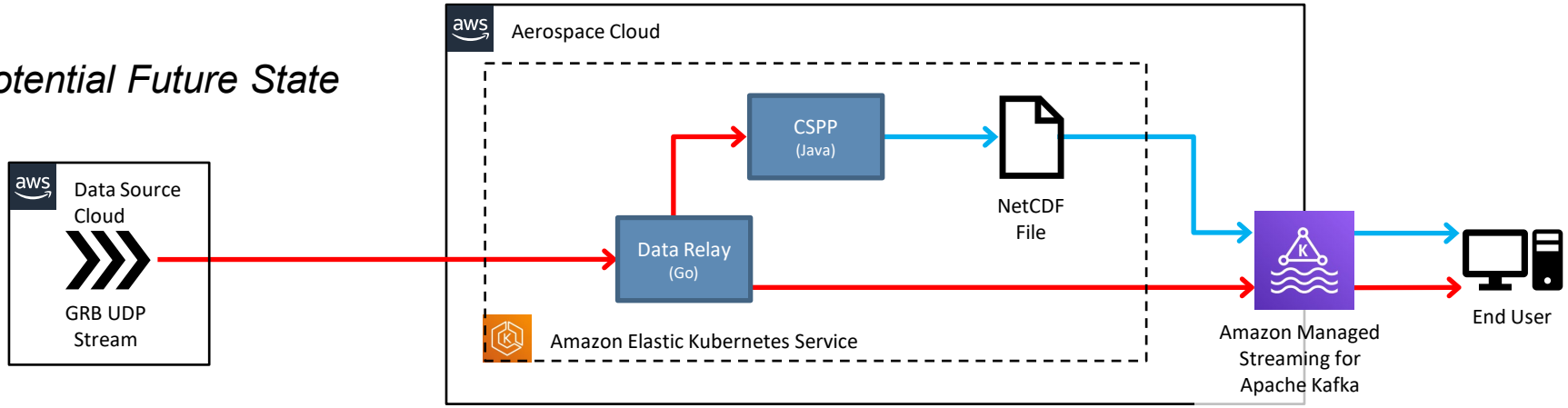


Cloud Pilot Capabilities

Delivering data in useful formats

- Goal:
  - *Deliver pre-processed and processed data to end user in a useful format*
  - *Leverage cloud technologies to simplify data processing for end user*
- Capability (future):
  - *Deploy CSPP Java software into cloud infrastructure*
  - *Offer delivery of both GRB packets and NetCDF files to end users*

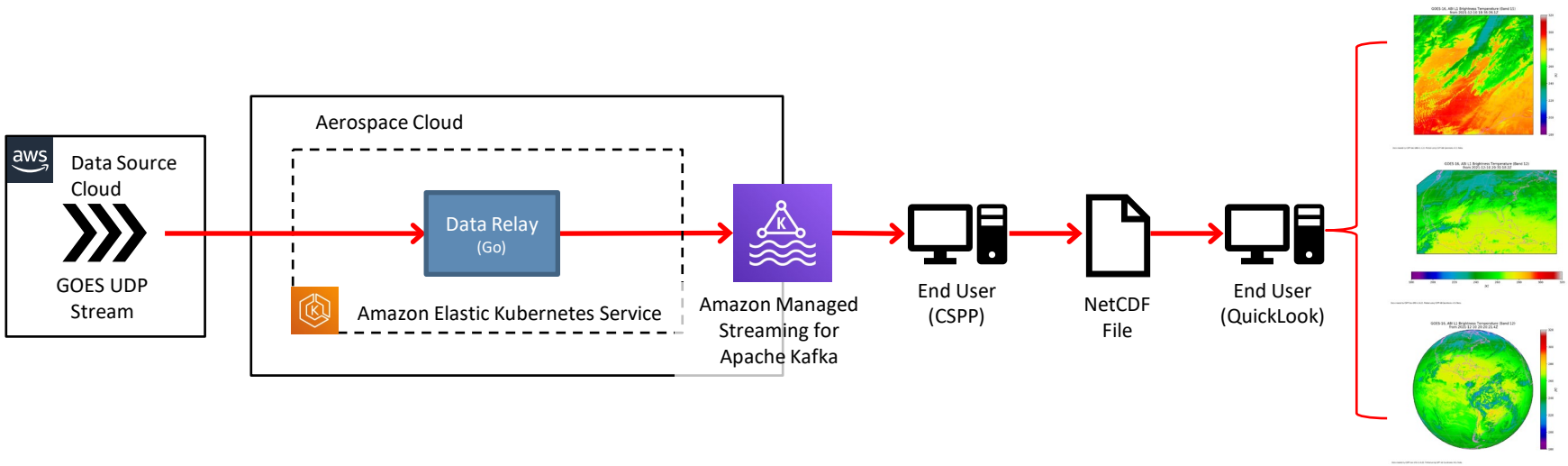
Potential Future State



Cloud Pilot Capabilities

Science Image Generation

- Goal:
  - *Convert NetCDF files to visualizations*
  - *Use existing tools to create a sample of visualizations to verify data stream quality*
- Capability (current):
  - *University of Wisconsin's Sounder QuickLook tool*
  - *Free and open-source Python application uses NetCDFs to create PNG images*
  - *End users must manually convert NetCDF files to visualizations*

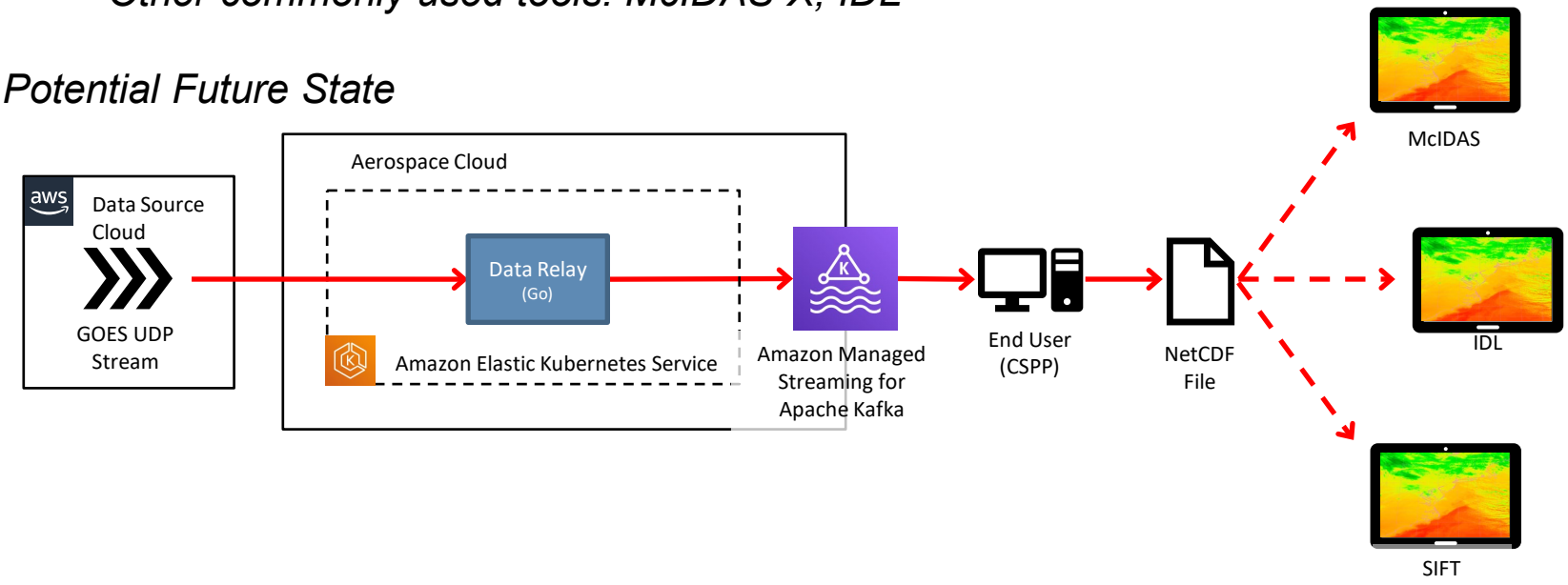


Cloud Pilot Capabilities

Science Image Generation

- Goal:
  - *Convert NetCDF files to visualizations in real-time*
  - *Use existing tools to create all real-time visualizations to verify data stream quality*
- Capability (future):
  - *Investigate 3rd party tools that automatically convert NetCDF files to visualizations*
  - *Tools automatically convert NetCDF files into static and animated visualizations*
  - *Commonly used free and open-source tools: McIDAS-V, SIFT*
  - *Other commonly used tools: McIDAS-X, IDL*

Potential Future State



Lessons Learned

- Cloud offers a vast number of technologies
  - *Explore and experiment to find the best solutions*
- DevSecOps processes improve security and quality
  - *Holistic integration of developers, security experts, and operations experts*
  - *CI/CD pipeline allows for rapid, repeatable, automated deployments*
- Important to have accurate tools to measure data quality
  - *Undetected upstream quality issues caused baffling issues with visualization*
- Important to understand ease of use vs portability
  - *Managed services make it easier to create and use infrastructure*
  - *Each managed service has a different level of vendor lock in*
  - *Find balance between decreased management vs increased vendor lock in*

Pilot Next Steps

- Conduct capability demonstrations
  - *Prove power/utility using visualizations, quality measurements, and benchmarking*
  - *Real-time cloud distribution end-to-end capability*
  - *Measurements of long-term data delivery performance*
  - *Performance during increased scale of end users*
- Evaluate follow-on capabilities and priorities
  - *Additional imagery processing via 3rd party toolsets (McIDAS-V, IDL, etc.)*
  - *Distribution of additional incoming data streams*


Thank You!