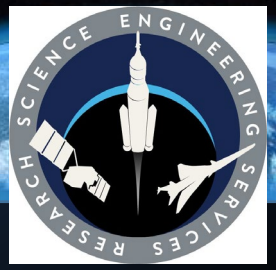
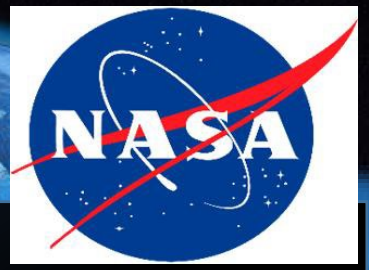




ASDC's Python-Based Metadata Extraction Pipeline for Suborbital Campaigns

Abraham Porter[1,3], Nathan Jester[2,3], Megan Buzanowicz[1,3], Sean Leavor[1,3], Gabriel Mojica[1,3], John Kusterer[3], Gao Chen[3]

[1]ADNET Systems inc.; [2]Booz Allen Hamilton; [3]NASA Langley Research Center (LaRC)

Atmospheric Science Data Center

- One of twelve Distributed Active Archive Centers (DAACs)

- Established in 1991 at NASA Langley Research Center

- Over 1000 archived collections

- We work to make data products that translate findings into meaningful knowledge to inspire scientists and educators, decision makers, and the public.

Atmospheric Science Data Center

Let's talk about FAIR data

Findable

Metadata and data should be easy to find for both humans and computers.

Accessible

Users need to know how data can be accessed once it is found.

Interoperable

Data can be integrated with other data; interoperate with applications or workflows for analysis, storage, and processing.

Reusable

Optimize reuse of the data; metadata and data should be well-described so that they can be replicated and/or combined in different settings.

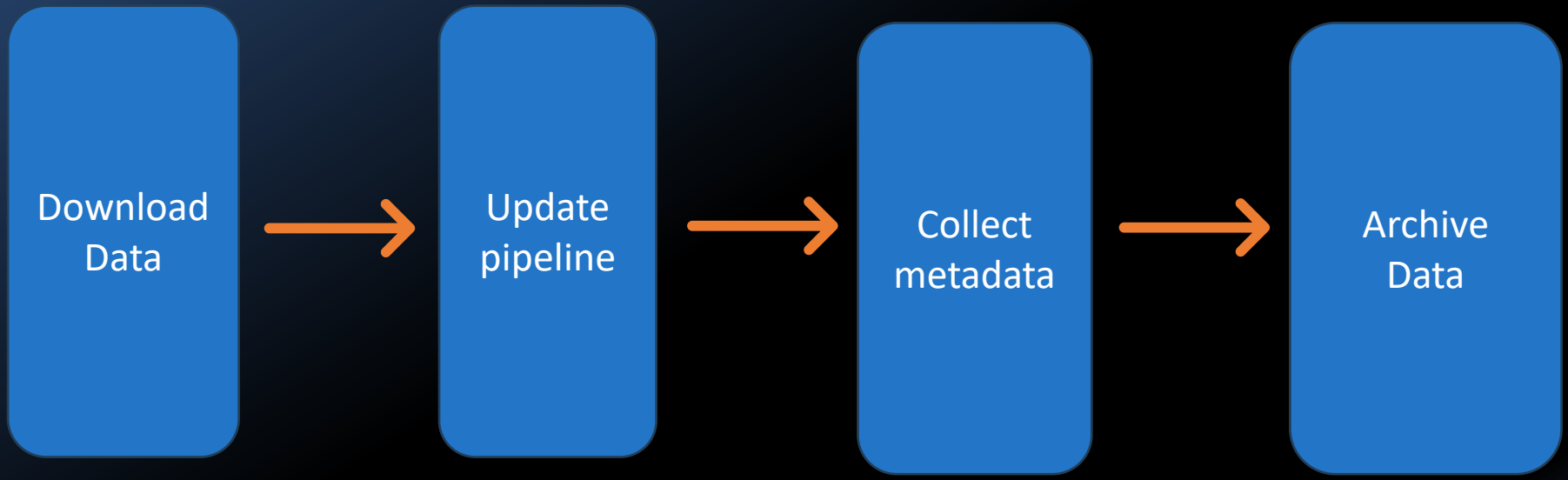
Why build a pipeline?

- The ASDC has a high quantity of suborbital data - need to streamline the ingest process to get data available the public in a timely manner

- Maintain a consistent flow of data and metadata

- Ensure that the ASDC meets the needs of the suborbital science community.

Pipeline Flow



The Metadata Extraction Pipeline

- Download campaigns from sftp and ftps servers; can also process data sent directly by the science team

- Update the pipeline in order to handle the unique aspects of the campaign

- Ingest the data and collect metadata

- Pass the original data and the metadata on to ASDC ingest and archive system – data can then be made publicly available

Plug in Based Model

- Allows adaptability of file formats and data types

- New parsers can be created as needed without needing to rewrite the entire code base

- Enables non-developers to make changes in order to run the pipeline

What does the data look like?

ICARTT: (International Consortium for Atmospheric Research on Transport and Transformation)

HDF: (Hierarchical Data Format)

netCDF: (network Common Data Form)

Ames File Format: (Multiple variations)

Files without easily extractable metadata

PDF (Portable Document Format)

GIF (Graphics Interchange Format)

What does the metadata look like?

- Start and end date times

- Flight paths

- Location

-Revision number

Atmospheric Science Data Center

Flight path for Macpex (2011)

Challenges we face

- Legacy campaign data is far less standardized and can lack metadata

- Formats and templates have drastically changed throughout the years

- Metadata is diverse for each campaign

Questions?

Resources

- Earthdata

- NASA Airborne Science Program

- ASDC About Page