Jet Propulsion Laboratory
California Institute of Technology

Upstream Ancillary Ingest: Keep Up Best You Can

Namrata Malarout

Scientific Applications Software Engineer

Jet Propulsion Laboratory

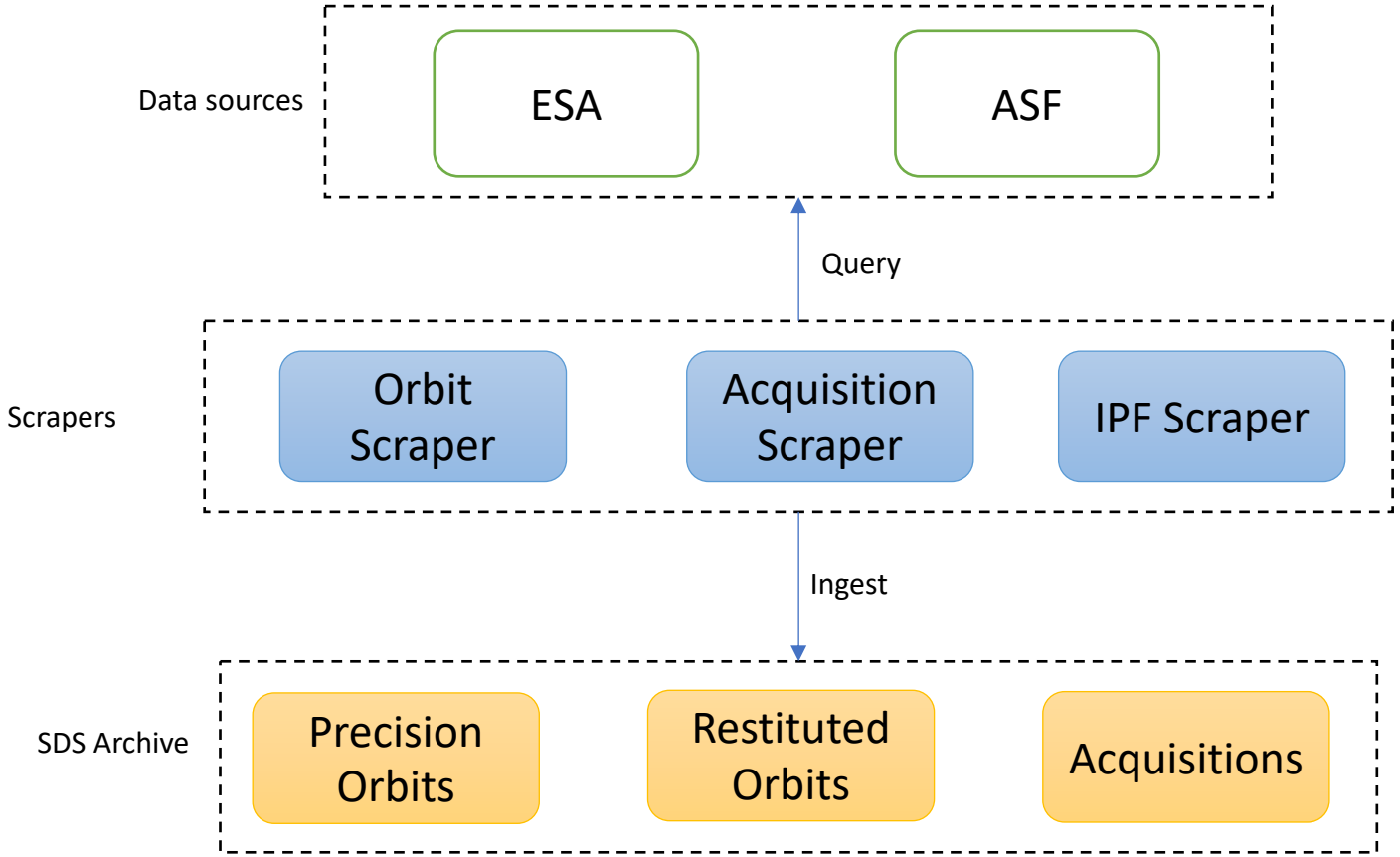
Jet Propulsion Laboratory
California Institute of Technology

ARIA: Advanced Rapid Imaging Analysis

- rapid and large-scale analysis of SAR observations
- processed to higher-level actionable information that can be used by hazards response communities
- ARIA SDS has prototyped generating data products that contain surface displacement maps (from earthquakes, volcanoes), damage maps (from building damage, landslides, fire burn scars), and flood maps
- Uses Hybrid-cloud Science Data System
  - Designed to run on public and on-premise clouds, as well as compute on legacy machines.
  - Mainly AWS and OpenStack
  - Supports cost-effective AWS *spot market*
  - Resiliency and fault-tolerant compute

Jet Propulsion Laboratory
California Institute of Technology

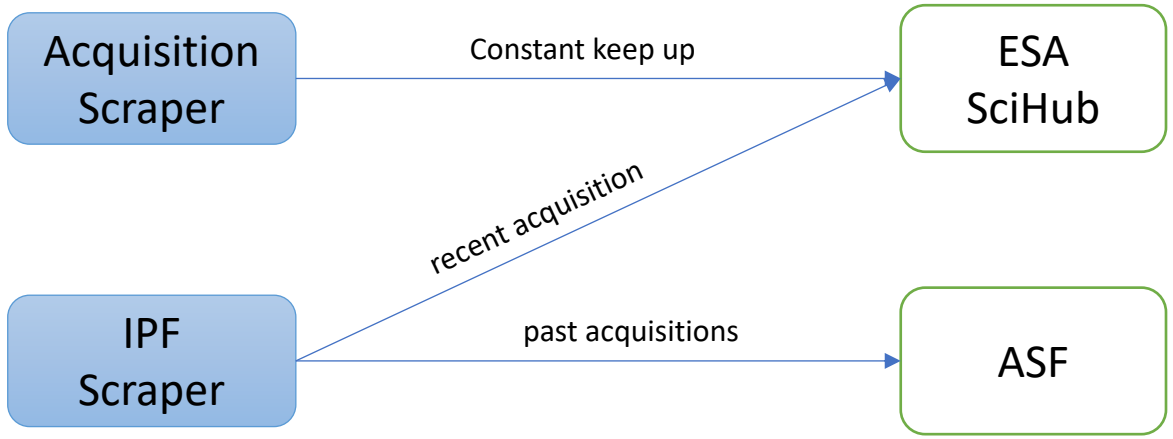
Upstream Ancillary Ingest



Jet Propulsion Laboratory
California Institute of Technology

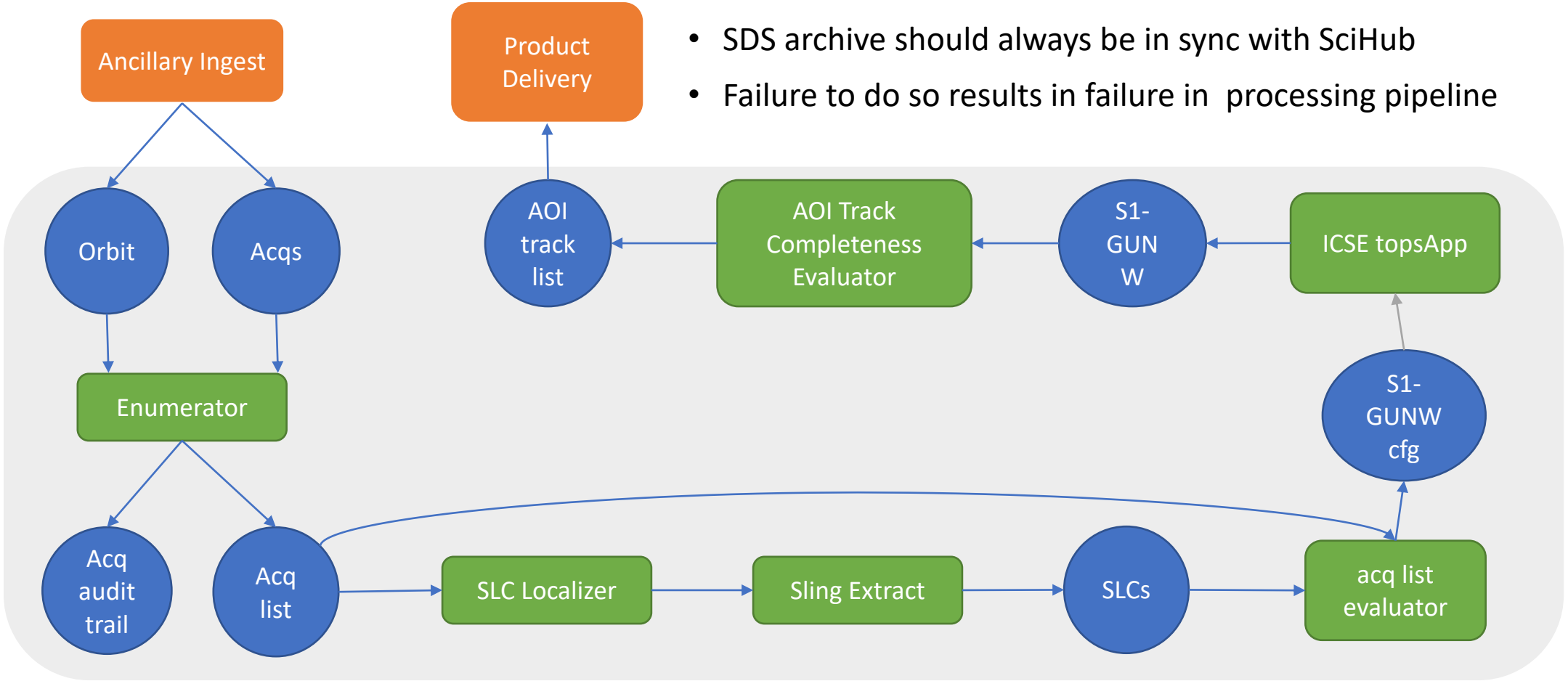
Which data provider should I go to?

- SciHub maintains hot cache and long term archive
- ASF also scrapes SciHub and archives



Jet Propulsion Laboratory
California Institute of Technology

Impact of Ancillary Keep Up



- SDS archive should always be in sync with SciHub
- Failure to do so results in failure in processing pipeline

Jet Propulsion Laboratory
California Institute of Technology

Keep Ups and Hold Ups

Jet Propulsion Laboratory
California Institute of Technology

Keep Up Mode

- Global acquisition scraper
  - Best Effort Jobs
  - Frequency of runs:
    - Sliding Window Approach
    - Daily – looks back 5 days for new acquisitions
    - Hourly – looks back 5 hours for new acquisitions

- Global IPF scraper
  - Job per acquisition
  - Must succeed
  - Frequency of runs:
    - Sliding window approach
    - Daily – looks back 5 days for acquisitions missing IPFs
    - Hourly – looks back 5 hours for acquisitions missing IPFs

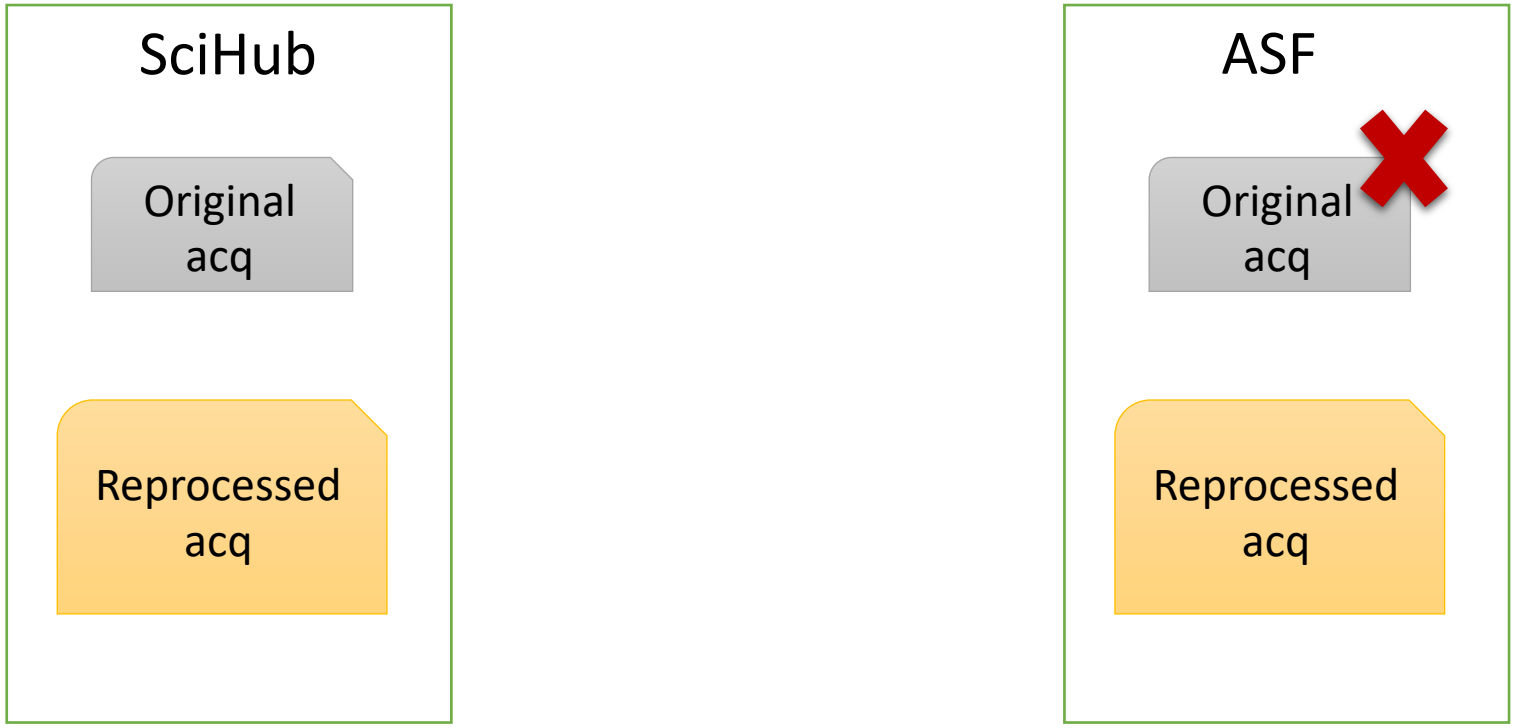
Jet Propulsion Laboratory
California Institute of Technology

Reprocessed Acquisitions

- Updated IPFs
- Updated metadata fields
- Slightly different start and end time
- Change in hash

Jet Propulsion Laboratory
California Institute of Technology

Deprecated Acquisitions



No information on which acquisition was superseded.

Jet Propulsion Laboratory
California Institute of Technology

Challenges of Dealing with Deprecated Acquisition

- Cannot retrieve it (immediately) from SciHub if acquisition is older than 2 years
- Fail to get IPF from ASF because it no longer exists
  - ASF gives response with an empty list and 200 status
  - Does an empty response mean Not Found? NO
  - Scenarios for empty response:
    - Issue in querying CMR
    - Acquisition doesn’t exist
      - Deleted
      - Not yet ingested

Jet Propulsion Laboratory
California Institute of Technology

Handling Deprecation

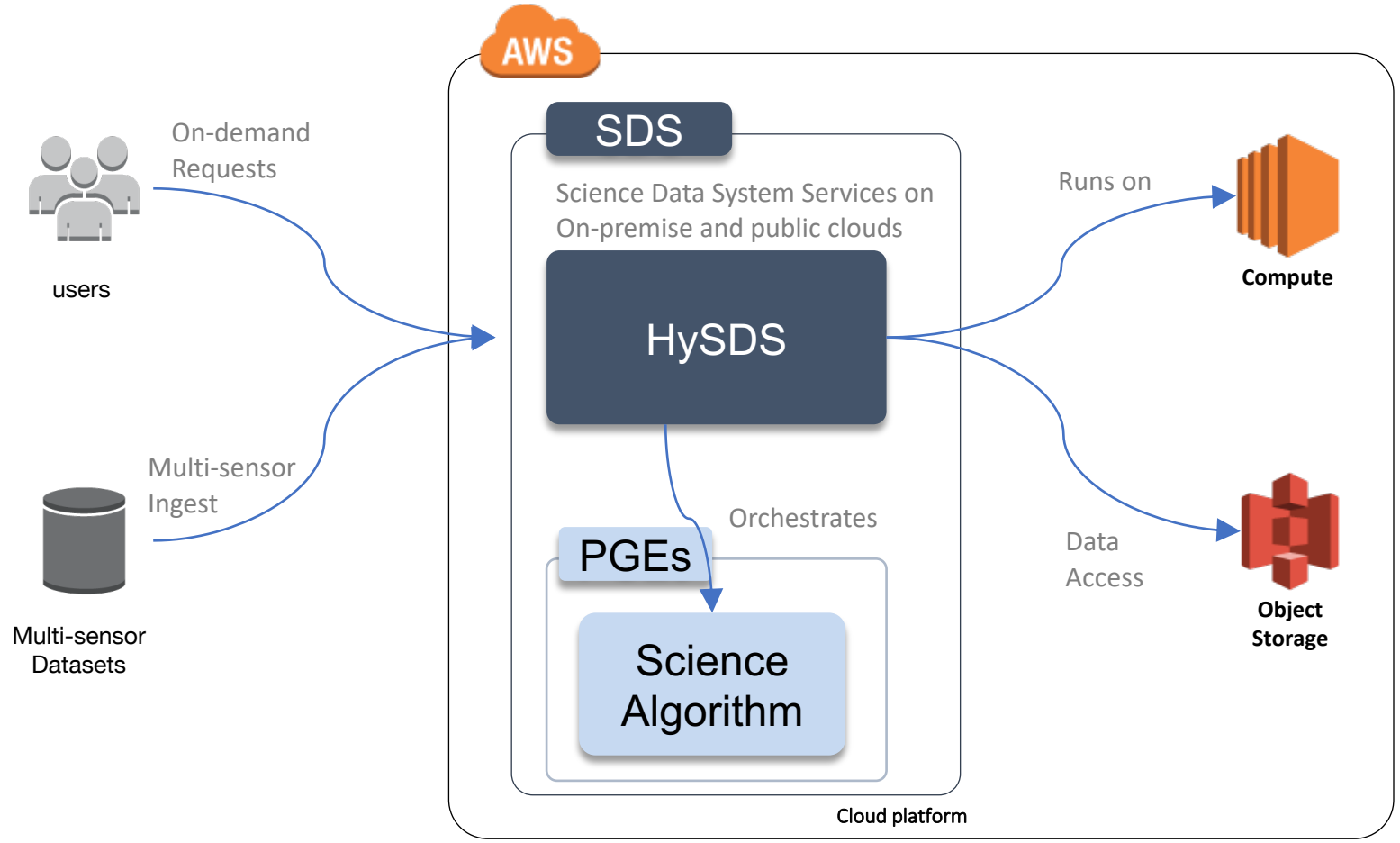
- Work in Progress
- Can handle cases where only hash of acquisition changes
- Yet to handle reprocessed acquisition with update timestamps

Jet Propulsion Laboratory
California Institute of Technology

Challenges of Running in Cloud Environment

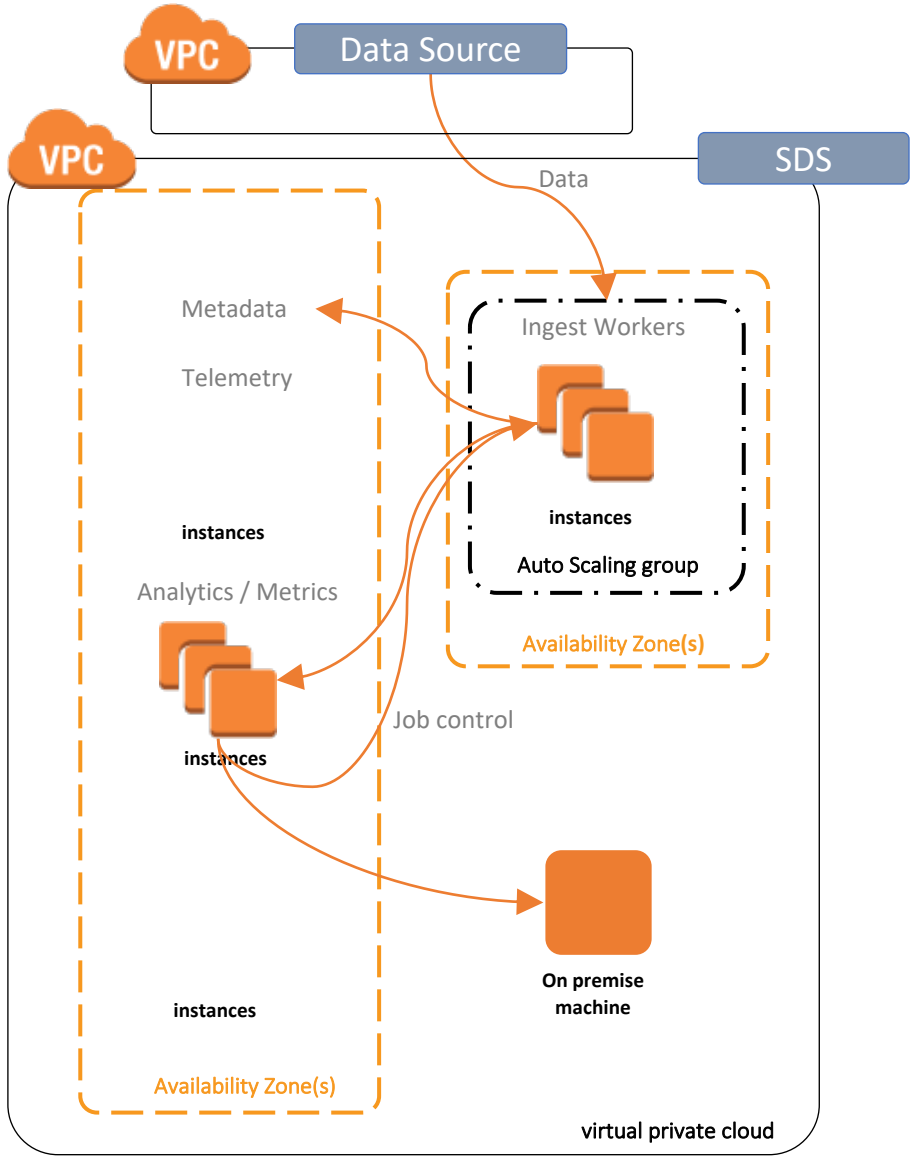
Jet Propulsion Laboratory
California Institute of Technology

Top-Level Software Components



Limits on Connections

- No limit in number of connections to ASF
- Limited to 2 accounts per worker per account



Jet Propulsion Laboratory
California Institute of Technology

IPF Common Errors

Scenario	Error	
Exception: Got code 503 (scihub, exceeding concurrent connections to scihub)	Exceeding max concurrent SciHub connections.	→ Retry later
Scihub 202 Response	Acquisition moved to Long Time Archive	→ Query ASF instead
Retrieved IPF value is null	Got null IPF	→ Retry job
Empty Response from ASF	Acquisition not found at ASF	→ Hope next hourly will catch it
Null XML Manifest	IPF not found in XML	→ Transient Error, retry

Jet Propulsion Laboratory
California Institute of Technology

Duplicate Jobs in System

Problem

- The retry rule keeps re-queuing failed IPF scrape jobs.
- HySDS has a feature to detect duplicate jobs
- It does not “dedup” against failed job
- The global IPF scrape submitter would submit jobs for the same acquisition every hour
- Multiple IPF scrape jobs for same acquisition
- Introduces race condition since it runs on auto scaling instances

Duplicate Jobs in System

Solution: Short Circuit Jobs

Every time the job runs, there are several checks:

- Is the IPF still null for the acquisition? then proceed else exit.
- Query Job Manager for duplicate jobs. Check if current job has the highest retry count. If YES, then short circuit. Other duplicates with lesser retry attempts exist in the system.
- Make sure to check if IPF is filled and not a null retrieved value before calling document update

Take away: Robustness and Resilience

- Make system robust enough to custom handle specific failures (i.e trigger rules)
- Make jobs resilient to external changes and unreliable responses.
- Use sliding window approach to account for service downtimes
- Use domain knowledge and experience to mature your jobs (handling to edge cases)
- To handle duplicate jobs, use short circuit logic such that the last man standing successfully finishes the job

Jet Propulsion Laboratory
California Institute of Technology

Questions?