Backoffice and DevOps Updates

Taylor Jacovich and The ADS Team

Introduction

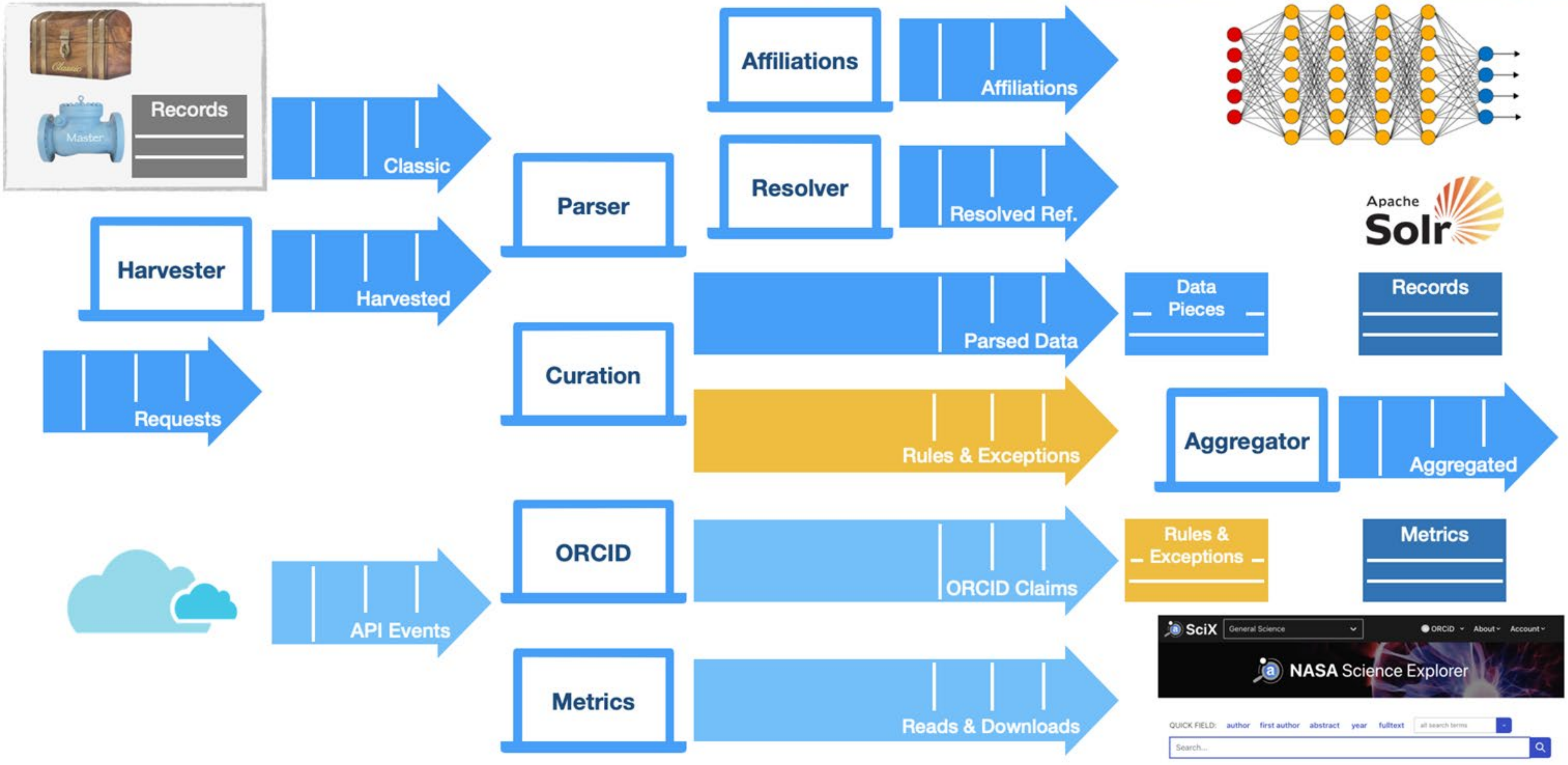
Overview

- New Architecture
  - The concept
  - Discussion of technologies
  - Roadmap
- DevOps
  - New backoffice hardware
  - Modernizing deployments
  - Extended monitoring and system resilience

New Architecture

Sergi Blanco Cuaresma
Kelly Lockhart

NASA Science Explorer

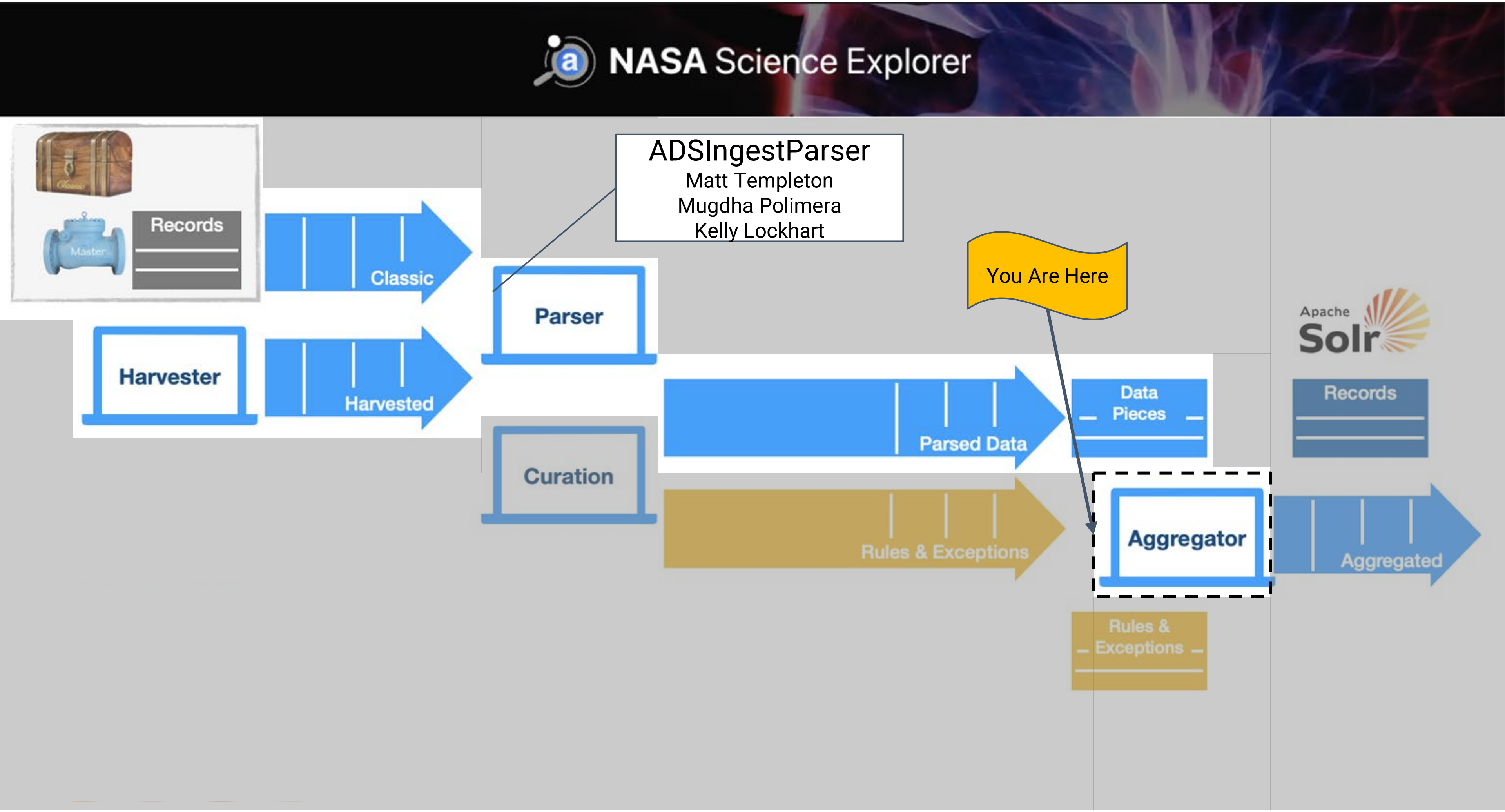


astrophysics data system

NASA

CENTER FOR ASTROPHYSICS
HARVARD & SMITHSONIAN

New Architecture



ADSIngestParser
Matt Templeton
Mugdha Polimera
Kelly Lockhart

You Are Here

New Architecture: Kafka

- Event based architecture
  - Each new piece of data produced triggers additional tasks

- Kafka Brokers are the backbone of the system
  - Pipelines will be able to operate as soon as data is available to them

  - Kafka will exist in the new API Gateway

    Allow for real time metrics calculations

    Real time ORCiD processing

New Architecture: Pipeline API

- Protocol
  - gRPC

    HTTP/2 protocol

    Built in (de)serialization and methods for defining API

    Strict type checking due to serialization

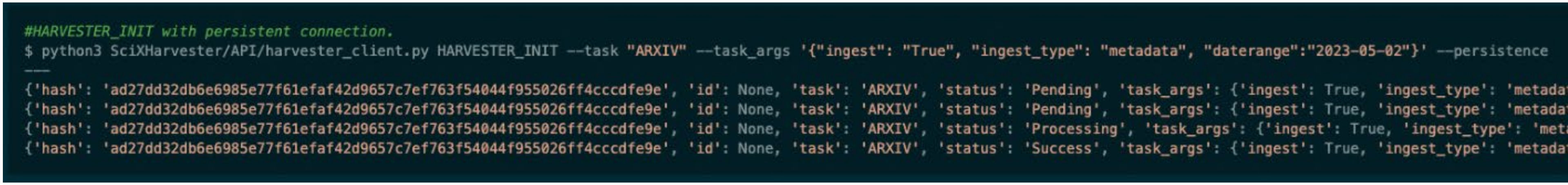
    Relatively high uptake by web developers

    Allows for persistent connections
- Serialization Schema
  - AVRO

    Consistent with Kafka serialization

    Defined in a json file

    No compilation



```
#HARVESTER_INIT with persistent connection.
$ python3 SciXHarvester/API/harvester_client.py HARVESTER_INIT --task "ARXIV" --task_args '{"ingest": "True", "ingest_type": "metadata", "daterange":"2023-05-02"}' --persistence
---
{'hash': 'ad27dd32db6e6985e77f61efaf42d9657c7ef763f54044f955026ff4cccdfe9e', 'id': None, 'task': 'ARXIV', 'status': 'Pending', 'task_args': {'ingest': True, 'ingest_type': 'metada
{'hash': 'ad27dd32db6e6985e77f61efaf42d9657c7ef763f54044f955026ff4cccdfe9e', 'id': None, 'task': 'ARXIV', 'status': 'Pending', 'task_args': {'ingest': True, 'ingest_type': 'metada
{'hash': 'ad27dd32db6e6985e77f61efaf42d9657c7ef763f54044f955026ff4cccdfe9e', 'id': None, 'task': 'ARXIV', 'status': 'Processing', 'task_args': {'ingest': True, 'ingest_type': 'meta
{'hash': 'ad27dd32db6e6985e77f61efaf42d9657c7ef763f54044f955026ff4cccdfe9e', 'id': None, 'task': 'ARXIV', 'status': 'Success', 'task_args': {'ingest': True, 'ingest_type': 'metada
```

New Architecture: State Keeping

- Postgres
  - Persistent data storage

    ie. job status, parsed metadata, file paths, etc.

- Redis
  - Real-time updates for persistent API connections

    ie. job status updates

- S3
  - Persistent object storage

    ie. text files, images, etc.

  - AWS for cloud
  - minIO for on-site

New Architecture: WEKA FS

- Parallel File System
  - Faster retrieval of data

- Redundancy
  - Data protection
  - Minimize downtime in the event of disk or node failure

- Can be deployed on-site or in the cloud
  - Gives flexibility
  - Can run on new cluster or AWS

New Architecture

Roadmap

- The Aggregator
  - Aggregation of masterDB with new Harvester
  - Major step for transitioning to new Architecture

- Additional pipeline pieces
  - We have a Template: SciXTemplatePipeline
  - Functionality related to AVRO Serialization, python configuration, S3 interaction, and UUID7 generation exists in: SciXPipelineUtils

- New Parsers
  - Many parsers are ready to go and just need a source added to the Harvester

- Additional Harvester Sources
  - ArXiv fulltext is a logical next step

DevOps Updates: New Hardware

The Brain Cluster

- New backoffice servers
- 8 Dell Servers
  - 40x 1.92TB NVME drives for WEKA
  - 16x 480GB root drive (RAID 1)
- OS installation: 9/18
- Network configuration: 9/25
- WEKA FS configuration: TBD
- Considering future modifications to support ML pipelines

DevOps Updates: Deployment

Modernizing and Automating Deployments

Fernanda de Macedo Alves
Sergi Blanco Cuaresma
Kelly Lockhart

- Replacing ADSTailor
  - Developed in-house ~2018
  - Nonfunctional due to changes to github
  - Currently considering three off-the-shelf solutions

    FluxCD

    Tekton

    ArgoCD

- Transitioning Backoffice to kubernetes
  - Unify deployment strategies between ADS cloud and backoffice
  - More fault tolerant
  - Easier to create local dev environments

DevOps Updates: Resilience

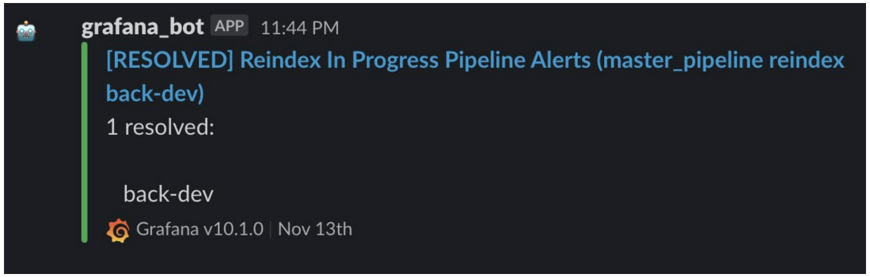
Monitoring and Redundancy

- Upgraded Grafana Dashboard to v10.1.0
  - Expanded alert methods

    individual alerts on crons

    multiple alerts per panel

    customizable alert templates



grafana_bot APP 11:44 PM

[RESOLVED] Reindex In Progress Pipeline Alerts (master_pipeline reindex back-dev)

1 resolved:

back-dev

Grafana v10.1.0 | Nov 13th

- Upgraded to Prometheus v2.4.5

  Better recovery from storage errors

- Converted RabbitMQ message broker to multi-node system
  - Redundancy in the event of broker failure

What’s Next?

- The Aggregator
- Additional pipeline pieces
- New Parsers
- Additional Harvester Sources
- Brain Cluster Configuration
  - Work with Red River to configure WEKA FS
- Deployment automation
  - Determine new deployment technology
  - Implement test system
- Backoffice Monitoring
  - Additional alerts
  - Additional logging