Scalable Infrastructure for Galaxy Image Analysis: I. Measuring Position Angles with Radon transforms

Neo Chen, Alberto Krone-Martins
Department of Informatics, University of California, Irvine

Online Version

1. INTRODUCTION

Domain sciences, such as Astronomy, significantly rely upon image acquisition and analysis. Present and upcoming sky surveys like Vera Rubin and Euclid produce petabytes of images that must be further analyzed automatically to extract scientific insights.

In this work, we propose a prototype for a scalable software infrastructure to enable the retrieval and analysis of galaxy images from public astronomical archives. We implement and deploy our prototype to estimate the position angles of ~6 million galaxies in multiple optical to near-infrared wavelengths.

We utilize the Radon transform to determine the position angle of galaxies:

$$\check{f}(p,\emptyset) = \int_{-\infty}^{\infty} f(p\cos\emptyset - s\sin\emptyset, p\sin\emptyset + s\cos\emptyset)\, ds$$

As a fully non-parametric way of determining the rotation, this method reduces possible model-dependent bias.

We aim to use these measurements to further reveal dark matter signatures and constrain important cosmological parameters of the models we use to explain our Universe.

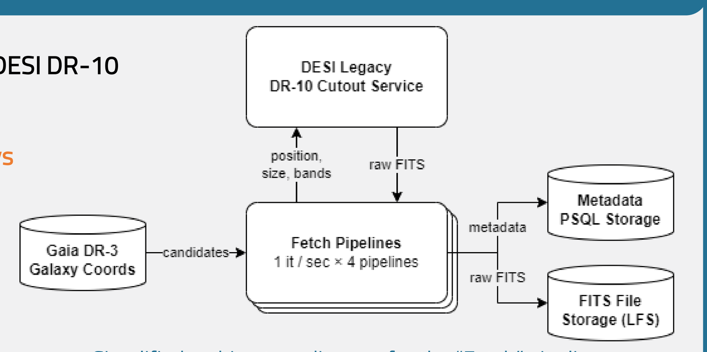
2. THE "FETCH" PIPELINE

The "Fetch" pipeline is responsible for retrieving FITS images of Gaia DR-3 galaxies from the DESI DR-10

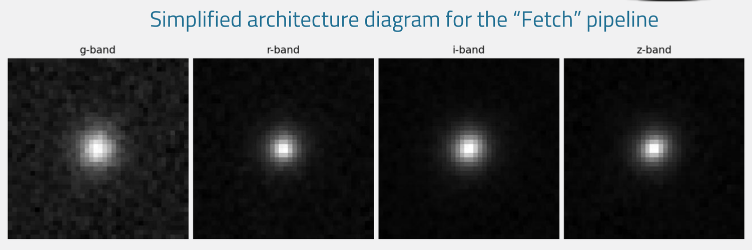
- We've retrieved FITS data from the [G, R, I, Z] optical filters
- Respecting rate limits, a total of 6,620,148 galaxy candidates was retrieved within 18 days
  - The DESI DR-10 has a recommended retrieval speed at 5 requests per second
  - But due to remote server failures, we averaged ~3.7 successful galaxies per second

However, the "Fetch" pipeline can handle much more concurrent traffic

- Designed for concurrency, this pipeline runs parallelly while ensuring data integrity
  - Utilizing Postgres's row-level locking mechanism
  - FIFO batch retrieval queue
- Failed retrievals are placed into a retry-queue with a delay for up to 3 times
- Auto-pauses when detecting an API server outage to prevent failures

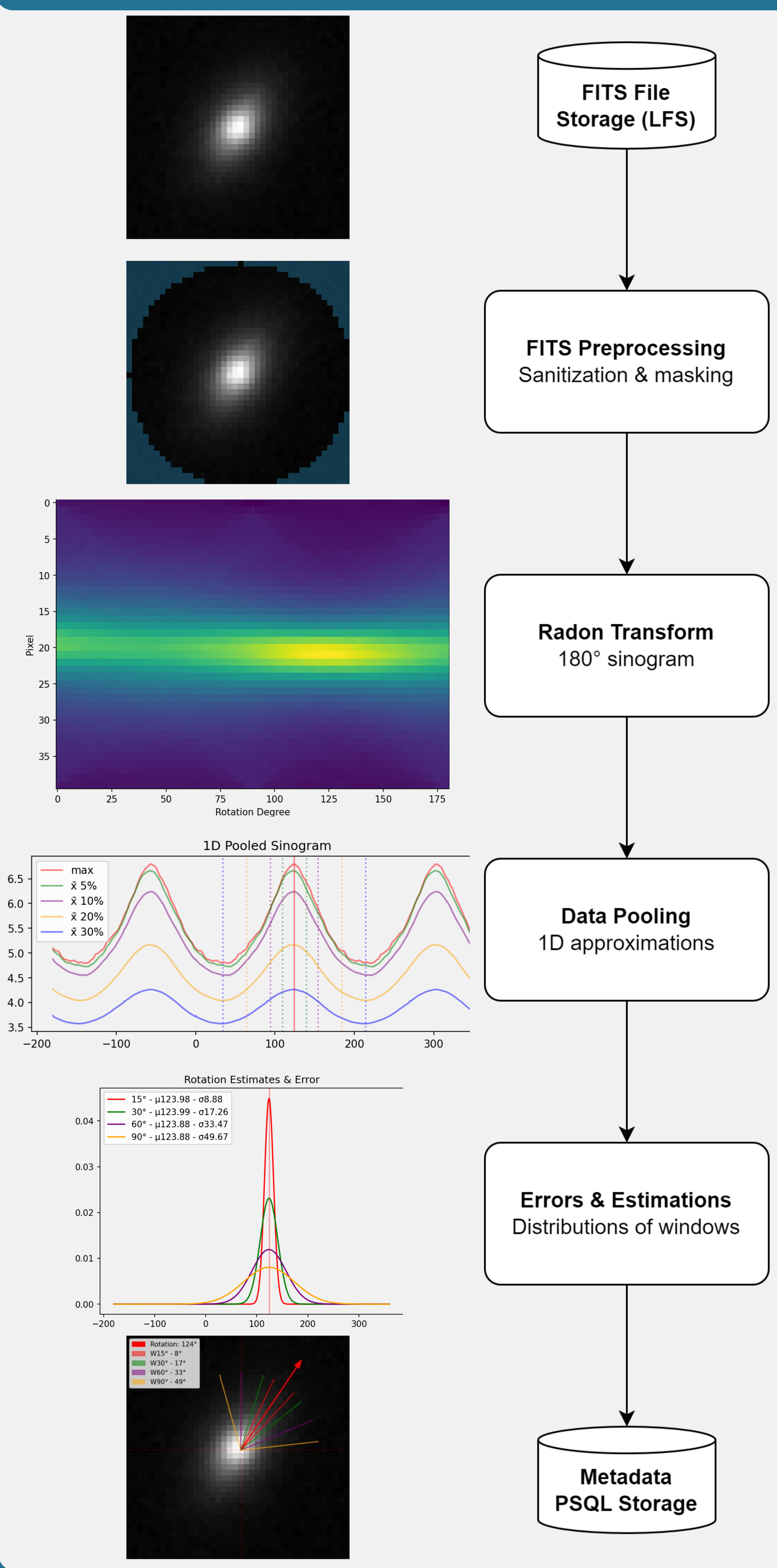


Simplified architecture diagram for the "Fetch" pipeline



Galaxy #4490125352892974336 from the DESI DR-10 survey

3. THE "RADON" PIPELINE



The "Radon" pipeline predicts the orientation of fetched galaxies by using the Radon transform and analyzing the resulting sinogram.

Step 1: The raw FITS images of a galaxy is retrieved from the FITS database

- The FITS images are converted into a 40 × 40 × 4 numpy matrix

Step 2: The raw FITS images are passed through a series of preprocessing steps

1. All invalid values (NaNs) are sanitized from the FITS data matrices
2. Each band's matrix is shifted so that the minimum value is 0
   - Surveys correct noise by subtracting a constant predicated noise
3. A circular mask (blue) is applied to each matrix, setting the exterior values to 0
   - Removes diagonal bias caused by a difference in pixel count

Step 3: Radon transform is performed on each band of the galaxy's matrix

1. All 181 affine transformation matrices is dotted with the masked image
   - Effectively rotating the image by 1° increments about its center
2. The rotated image is mapped using bicubic interpolation back to pixel coordinates
3. Pixel intensities are summed column-wise, and stacked to create the sinogram

This results in a 40 × 181 × 4 sinogram numpy matrix

Step 4: The 2D sinogram is transformed into 1D arrays through pooling filters

- Max-pooling and top x% average-pooling filters reduces the dimensionality
  - Max pooling: the brightest pixel of each column is selected
  - Top x% average pooling: the average of the top x% pixels is selected
- Reducing dimensionality facilitates further processing
- When extended, the 1D sinograms resemble sinusoidal patterns
  - Enables approximations by fitting sine-like functions

Step 5: Normal distributions are fitted to approximate errors

- Errors are estimated in proportion to the difference of the maximum value and its nearby values at various windows
  - The larger the difference, the smaller the error, and vice versa
- A normal distribution can be fitted to visualize the error estimations

Step 6: Rotation-related data is saved as metadata into the Postgres database

- Rotation, error estimates, and sinusoidal-approximations are saved per band
- The state of the galaxy is updated as "Transformed"
- Further analysis can be performed utilizing these metadata entries

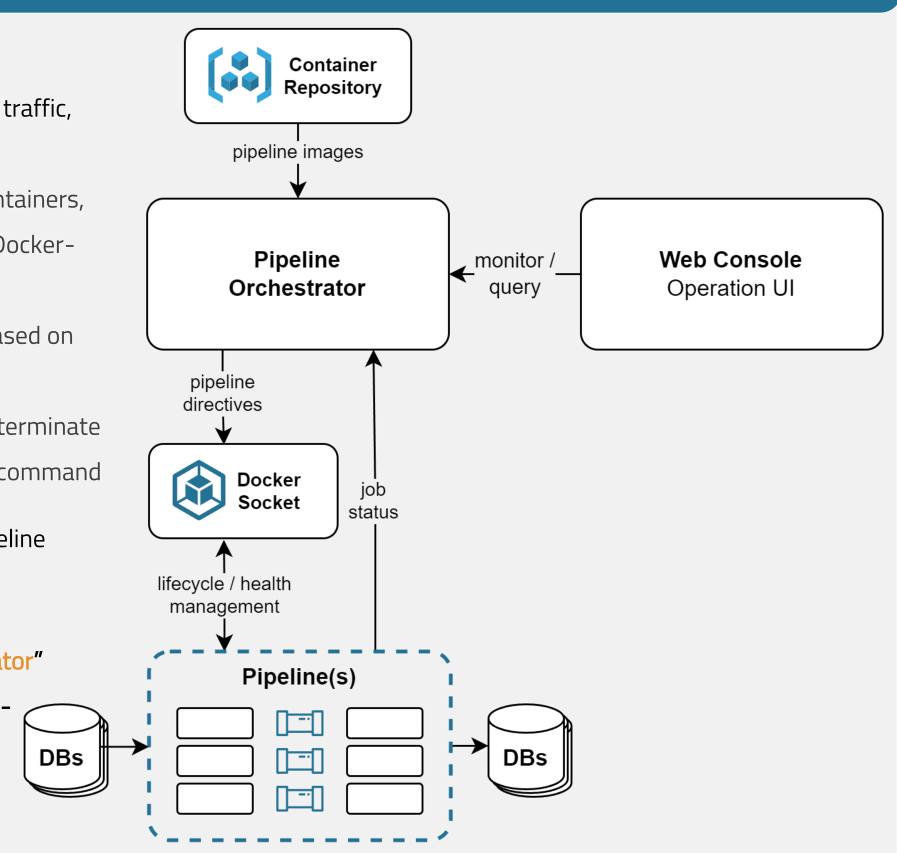
4. PIPELINE ORCHESTRATION

We architected a Docker-based system to streamline our data processing workflows. At our current scale, a single host is sufficient to support all our traffic, and Docker brings several advantages:

- **Environment-agnostic:** Our infrastructure is encapsulated in Docker containers, making it platform-independent and ensures compatibility across any Docker-supported operating systems
- **Adaptive scaling:** Pipeline containers can be easily scaled up or down based on workload requirements, providing flexibility in resource allocation
- **Deployment Efficiency:** With docker-compose, we can easily initiate or terminate all required resources, such as containers and networks, using a single command

To facilitate future data processing needs, we also designed a common pipeline interface that allows seamless integration of new pipelines.

Taking advantage of this common interface, our central "Pipeline Orchestrator" service taps into the Docker socket to manage the lifecycle of any interface-compatible pipelines.



Simplified components diagram of our software architecture

We designed the orchestrator service to have the following capabilities:

- Dynamically spin up and shut down pipelines based on task load
- Consistently monitor the health of all containers in the pipelines, and replace unhealthy ones gracefully
- Regularly pull pipeline image updates from DockerHub, and roll out deployments as needed
- Expose an API to the web console, allowing manual deployments overrides from authenticated operators

5. GALAXY AUGMENTATIONS

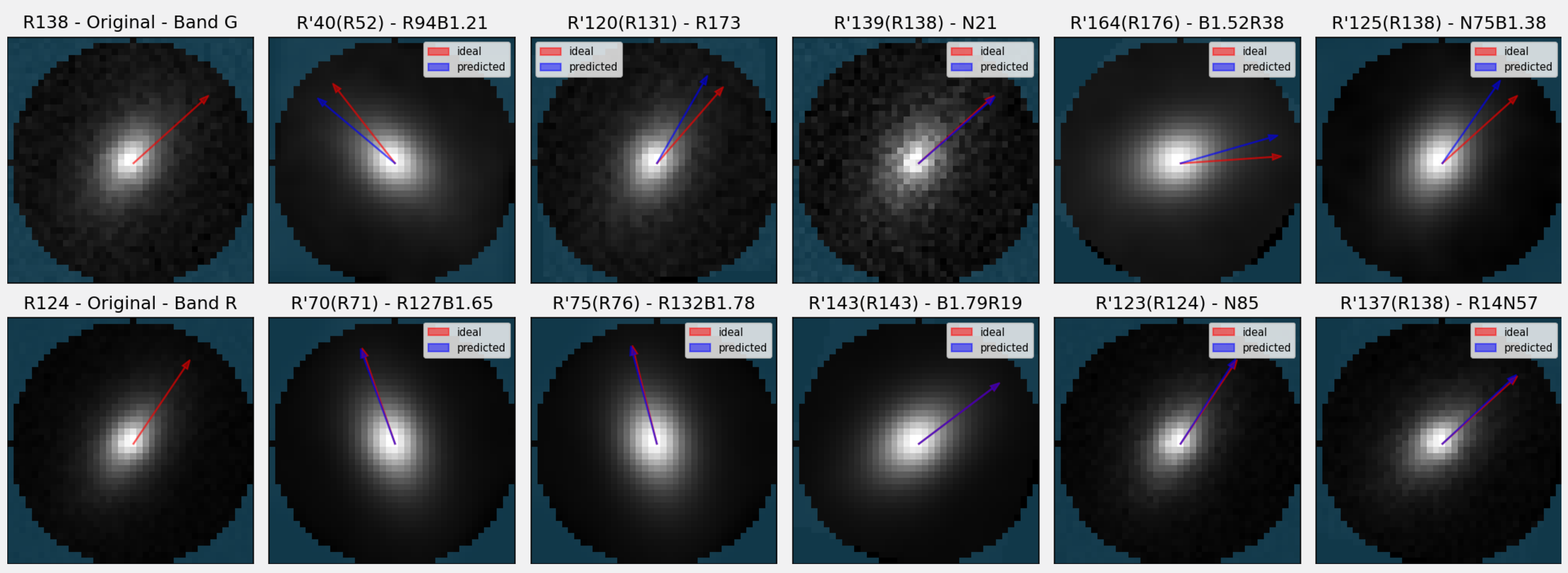
Due to the lack of a definitive oracle for galaxy rotations, we sought alternative methods to measure our model's accuracy. Given the complexities in galaxy observations – ranging from varied orientations to image clarity challenges – we introduced a set of augmentations, aiming to simulate different conditions that the model could face.

The methods included:

- Rotation (R): rotates the image by a random degree via dotting a transformation matrix and applying bicubic interpolation
  - This augmentation will change the predicted rotation by the degree of rotation; R' = (R + rotation) % 180
- Gaussian Blur (B): blurs the image by a random strength using the gaussian blur method
- Resampling (N): reconstructs the image at a random clarity by sampling repeatedly in the Poisson distribution

Using the predicted rotation of the original image as our benchmark, we applied these augmentations randomly, recalculated the Radon transform, and measured the deviation in the model's rotation prediction.

We ran the test across 500,000 augmented galaxy images, and the results was promising. Our model was able to achieve an average error of ~6.17 degrees, highlighting its robustness under diverse conditions.



Multiple augmentations of Galaxy #2311426402275174016, showcasing both the ideal (R) and model-predicted (R') rotations

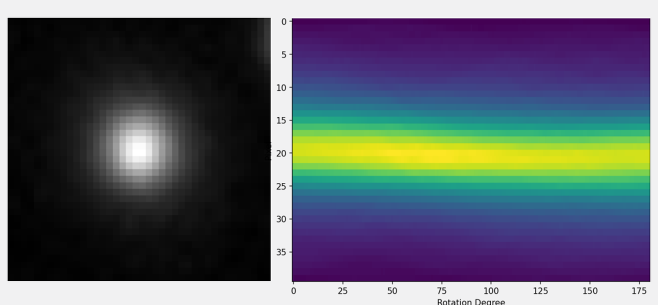
6. ELLIPTICITY

The study of the ellipticity of many galaxies can offer insights into unseen phenomena, such as the presence of dark matter. The apparent shape of a galaxy could be influenced by dark matter in the line of sight between the galaxy and the observer.

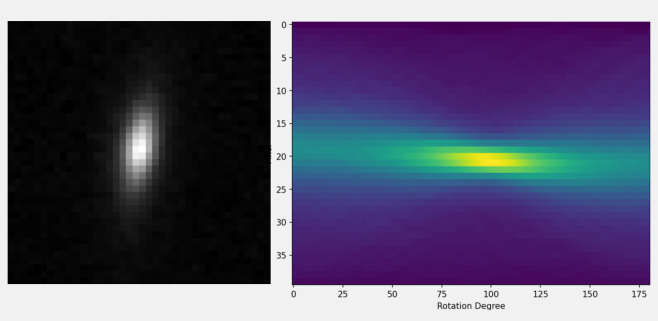
Ellipticity can be observed from the sinogram obtained from Radon transform, assuming that the galaxy is the sole object and relatively centered in the FITS image:

- A circular galaxy will produce a sinogram that is uniformly bright, with the brightest region near the horizontal center. This is similar to how perfect circle's sinogram would appear
- An elliptical galaxy will produce a sinogram with varying brightness across the horizontal center. As an elliptical shape condenses more bright pixels into a narrower area, the Radon transform's projection at certain angles will be more intense, leading to brighter regions in the sinogram

We also found a relationship between ellipticity and the predicted orientations from various optical filters for the same galaxy: a more significant variance in these predicted orientations usually suggests a more circular galaxy.



Circular galaxy #2266288529379361280 and its sinogram



Elliptical galaxy #2266306946198752000 and its sinogram

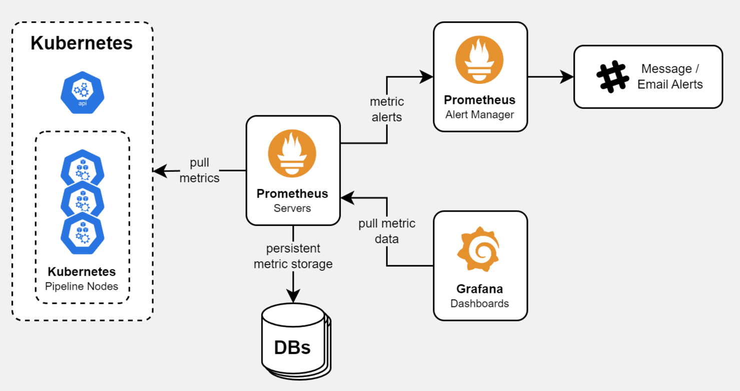
7. SCALING UP & OUT

While our current scope of ~6 million galaxies is relatively modest in comparison to industry projects, our infrastructure is designed for scalability. This design allows for smooth expansion across machines or various cloud platforms. To support this, we recommend using scalable services or cloud platforms.

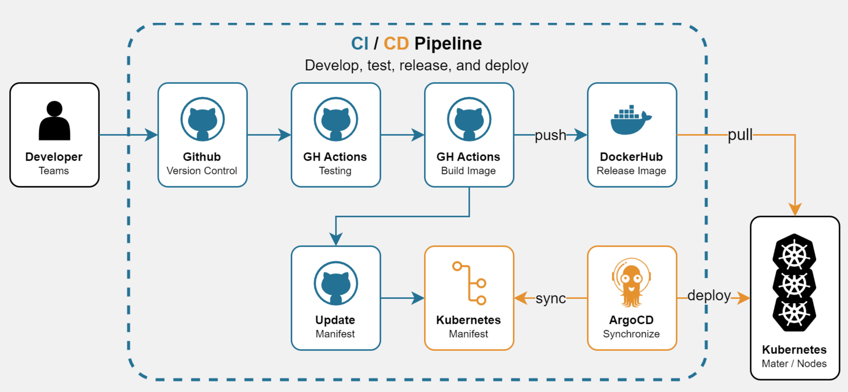
To facilitate scalability across different hosts or regions, we can utilize large-scale container orchestration platforms such as Kubernetes. Kubernetes allows us to deploy our pipelines to different fleets of machines, bolstering the system's robustness. When paired with monitoring solutions like Grafana or Prometheus, we can gain real-time insight into our pipelines, enabling proactive management and maintenance. Moreover, many infrastructure-as-code frameworks such as Terraform allows programmatic infrastructure management, effectively minimizing human errors.

In a distributed environment, the choice of database is also critical. Depending on data reliability and scalability requirements, a separate database hosting service can be considered. These services often support features like replication, sharding, and automated backups, ensuring data integrity and robustness.

Lastly, a CI/CD pipeline could be implemented to streamline the release processes. This allows changes to propagate through various testing stages before being deployed automatically, enabling release efficiency and warranting system reliability.



Example monitoring infrastructure using Prometheus and Grafana



Example CI/CD pipeline using GitHub, Argo CD, and Kubernetes

8. REFERENCES

[1] Z. Ivezi ˇ c and et al., "LSST: From Science Drivers to Reference Design ´ and Anticipated Data Products," The Astrophysical Journal, vol. 873, no. 2, p. 111, Mar. 2019.

[2] R. Laureijs and et al, "Euclid Definition Study Report," arXiv e-prints, p. arXiv:1110.3193, Oct. 2011.

[3] P. Schneider and C. Seitz, "Steps towards nonlinear cluster inversion through gravitational distortions. I. Basic considerations and circular clusters." Astronomy and Astrophysics, vol. 294, pp. 411–431, Feb. 1995.

[4] L. Whittaker, M. L. Brown, and R. A. Battye, "Weak lensing using only galaxy position angles," Monthly Notices of the Royal Astronomical Society, vol. 445, no. 2, pp. 1836–1857, Dec. 2014.

[5] L. Whittaker, "Constraining cosmology using galaxy position angle-only cosmic shear," Monthly Notices of the Royal Astronomical Society, vol. 502, no. 1, pp. 728–749, Mar. 2021.

[6] J. Varga, I. Csabai, and L. Dobos, "Refined position angle measurements for galaxies of the SDSS Stripe 82 co-added dataset," Astronomische Nachrichten, vol. 334, no. 9, p. 1016, Nov. 2013.

[7] J. Radon, "Bestimmung von Funktionen durch ihre Integralwerte langs ¨ gewisser Mannigfaltigkeiten," Berichte der Sachsischen Akadamie der ¨ Wissenschaft, vol. 69, p. 262–267, 1917.

[8] A. Krone-Martins, C. Ducourant, R. Teixeira, L. Galluccio, P. Gavras, S. dos Anjos, R. E. de Souza, R. E. G. Machado, and J. F. Le Campion, "Pushing the limits of the Gaia space mission by analyzing galaxy morphology," Astronomy and Astrophysics, vol. 556, p. A102, Aug. 2013.

9. ACKNOWLEDGEMENTS

We acknowledge support from the Portuguese Fundação para a Ciência e a Tecnologia grants UIDB/FIS/00099/2020 and EXPL/FIS-AST/1368/2021. This work made use of DESI Legacy Imaging Surveys; the complete acknowledgments can be found at https://www.legacysurvey.org/acknowledgment/.

This work has made use of data from the European Space Agency (ESA) mission Gaia (https://www.cosmos.esa.int/gaia), processed by the Gaia Data Processing and Analysis Consortium (DPAC, https://www.cosmos.esa.int/web/gaia/dpac/consortium).