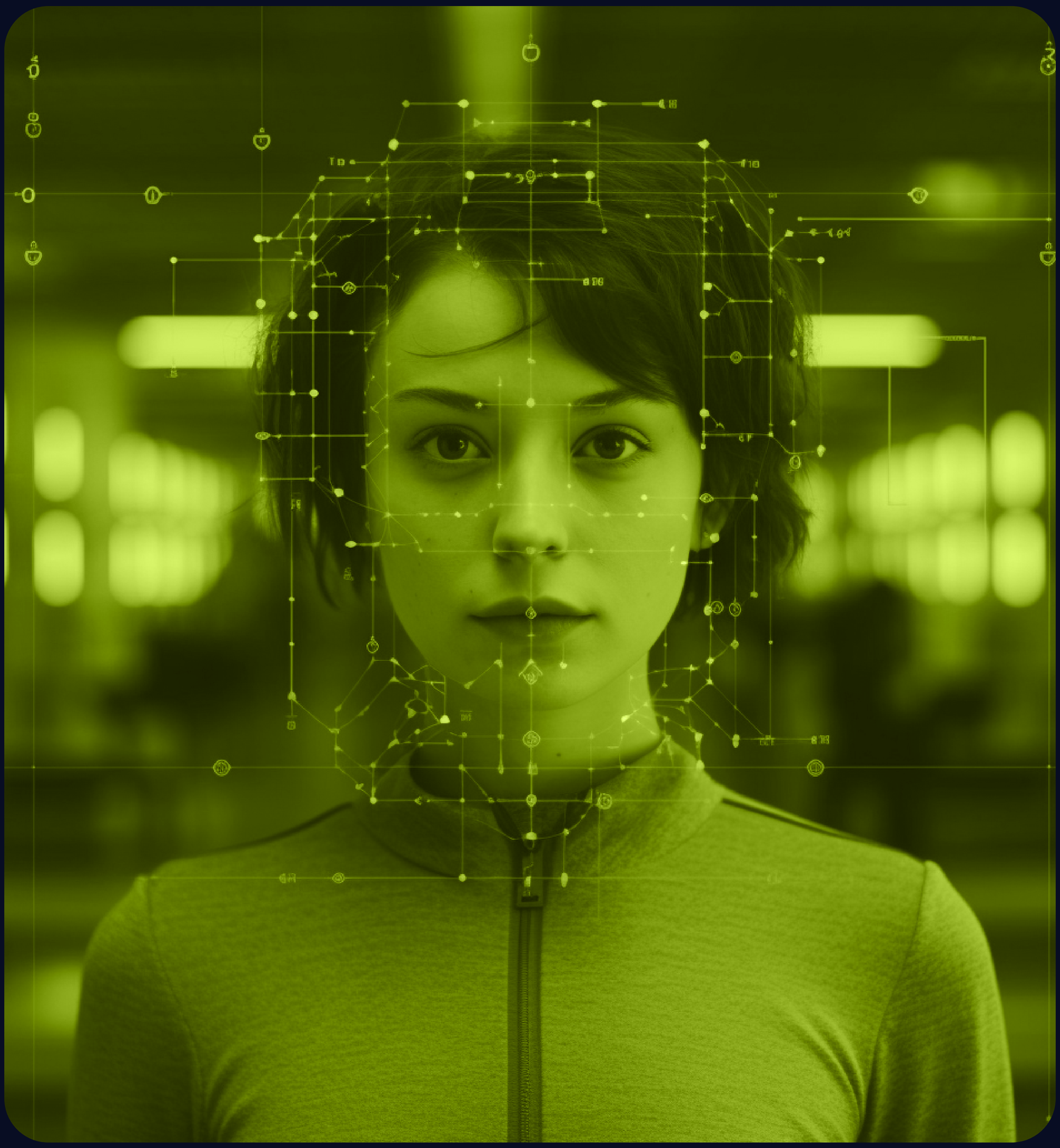

How to Build a Person Identification Pipeline Using Edgebricks

Use case: Celebrity Identification in Images

EDGE BRICKS

White Paper

Introduction

In the era of media and entertainment, identifying celebrities in images and videos is crucial for various applications, including content recommendation, copyright checking, audience targeting, branding validation and security. For instance, a celebrity may want to make sure that their face is used only along with actual brands they are sponsoring and not others. Similarly a politician may want to make sure they are not being associated with something that they do not endorse.

To address this challenge, the Edgebricks team leveraged their cutting-edge infrastructure and a comprehensive suite of AI/ML tools to build a powerful person identification pipeline specifically designed for celebrity identification. This case study explores how we leveraged various tools to build an end to end pipeline for this problem. This pipeline allows us to retrain new models as more data is available. This significantly improves the accuracy over time.

Objective

The primary objective of this case study is to showcase how Edgebricks' infrastructure, combined with best of breed AI/ML tools, facilitated the development of a robust pipeline capable of accurately identifying celebrities' faces within images or videos. Our main focus is to show a well architected pipeline to build an AI model for this problem and make it easy to improve the model over time. The accuracy of the model is bounded by the amount of data available to us and this is not the main focus of the case study.

We want to leverage as much of open source tools as possible for the pipeline creation so that our customers can build it without incurring any extra cost. We want to allow customers to build such pipelines and finally a production model using their own data along with existing AI/ML base models. Also, we plan to allow creation of a pipeline starting from data labeling, data cleaning, model training, testing and deployment. We also want to support production critical features like model versioning.

Challenges

When organizations want to build an AI model for doing face or person identification in a large pool of images, they have to think through various stages of such a pipeline and deal with many challenges. Some of these include:

- **Variability in Image Quality:** Images sourced from various platforms and devices often exhibit diverse qualities such as resolution, size, lighting, and background variations. These challenges necessitated a robust data preparation phase to extract meaningful features from the celebrity's face despite these inconsistencies.
- **Labeling and Large-scale Data Management:** Labeling a lot of images and managing a vast amount of labeled data efficiently while ensuring quick access for training and validation purposes posed a significant logistical challenge for the Edgebricks team. We also want to minimize data copying from one stage to another.

- **Model Training and Testing:** When it comes to model training, we want to enable a couple of critical features. Some of them include:
    a. Using one of the many base models designed for such a task
    b. Start from scratch or a previously trained model
    c. Do training using the new data only and not with all the data
    d. Divide data set automatically into training and testing
    e. Do validation and testing of model on test data

- **Model Optimization and Version Control:** Streamlining the model optimization process using hyperparameters tuning, changing various input parameters, and maintaining a versioned record of trained models were essential aspects of building a reliable and scalable solution for celebrity identification.

Objective

The Edgebricks team meticulously devised a solution by leveraging a suite of tools tailored to each step of the pipeline, addressing the aforementioned challenges and ensuring exceptional accuracy and efficiency throughout the process.

- **Data Labeling with Label Studio:** To create a high-quality training dataset, Edgebricks employed Label Studio, a versatile data labeling tool. This facilitated the annotation of celebrity faces within the images, enabling the extraction of accurate ground truth for subsequent model training. This also allows for multiple labelers to work together to do the labeling task.

- **Efficient Data Storage and sharing with NFS Storage:** In an AI/ML pipeline, it is critical to have a way to share data across various stages of the pipeline. There are two ways to do it in most cases. First is to use an object store and second is to use a NFS share. In the case of object store, the model labeling and training code has to download the data to local storage and then work with it. This data movement is very inefficient and slow. Edgebricks utilized NFS Storage, a reliable and scalable storage solution, to manage the vast amount of labeled data. This ensured seamless access to the labeled dataset for model training, validation, and future iterations of the pipeline.

- **Code Development with Jupyter Notebook:** The Edgebricks team harnessed the power of Jupyter Notebook, a popular development environment, to build and refine their code for various pipeline stages, including model training. This is the same tool used by AI developers on various cloud environments like AWS, Azure and GCP. This versatile tool provided an interactive interface, allowing for rapid prototyping, experimentation, and fine-tuning of the machine learning algorithms.

- **Model Training and Parameter Tuning with ClearML:** In order to do model training we need a fleet of machines typically with GPUs and we need to orchestrate various pipeline stages on that fleet. In some cases, we also need to schedule such jobs submitted from different developers or pipelines on the same set of underlying infrastructure. We used ClearML as our advanced experiment orchestration tool. ClearML played a vital role in automating the model training jobs, running various pipeline stages and tuning hyperparameters.

By systematically tracking experiments, managing resources, and visualizing performance metrics, ClearML accelerated the training and refinement of the celebrity identification model.

- **Seamless Model Versioning with MLflow:** To maintain a well-documented and versioned record of their trained models, the Edgebricks team integrated MLflow, a comprehensive model lifecycle management tool. MLflow facilitated model tracking, experimentation reproducibility, and deployment readiness, enabling the team to deploy the best-performing models confidently.
- **Model Deployment and Inference:** In practice once a model is built, it may get deployed on various cloud or edge platforms. Currently we did a deployment on a local web server using streamlit, so that our team and others can try out the model by uploading any image that they want to test.

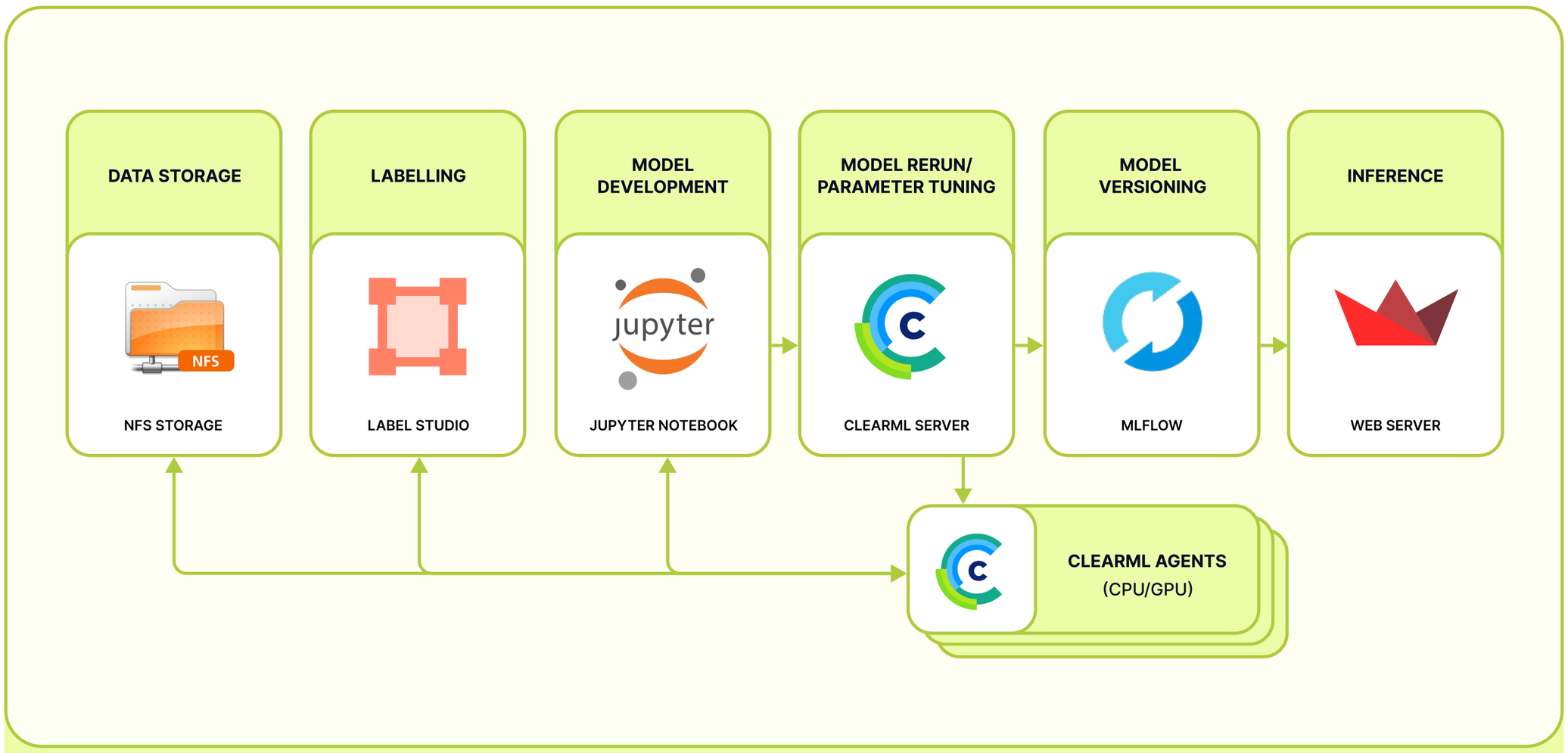

Figure 1 (Pipeline Architecture)

Figure 1 shows the overall architecture and how these tools are stitched together. Note that each box in the diagram is represented by a VM or container on Edgebricks cloud. Many of these stages are easy to scale since they use the same underlying NFS storage.

Final Outcome

By harnessing the capabilities of Edgebricks' cloud infrastructure and utilizing the aforementioned tools, the Edgebricks team achieved a very robust and well architected celebrity identification pipeline. Here are some key features and advantages of our approach:

- **Get a working pipeline within a few hours:** The first main advantage is that customers can get a fully developed end to end pipeline within a day by leveraging all these components from our AI/ML app store. This app store is built into the Edgebricks cloud platform itself.
- **No need for proprietary tools:** In building this pipeline, we have leveraged all the open source tools and the customers don't need to buy any extra licenses to use these tools.
- **Ability for quick experimentation:** The developed pipeline is highly configurable and our customers can choose between various base AI models, train a new model from scratch or use an existing model, do hyperparameter tuning and use shared data storage for quick data access. Using just a small amount of data, we were able to get more than 80% accuracy but given the small sample set, we don't claim that the model would work well on a very diverse input.
- **Data Privacy and Efficiency:** First all data stays within the Edgebricks cloud which can be deployed in a colo or a private datacenter by a customer. Second, labeling tools are built into the platform, so no data needs to be sent outside. Labelers can simply get access and accounts on the labeling tool and work remotely. Finally, with the aid of NFS Storage and streamlined data management practices, Edgebricks is able to efficiently handle large-scale labeled datasets, enabling rapid model training iterations and future expansion.
- **Accelerated Model Development:** The integration of Jupyter Notebook, ClearML, and MLflow significantly expedited the model development cycle, enabling the Edgebricks team to iterate quickly, optimize hyperparameters, and deploy high-performing models with confidence.

Conclusion

Edgebricks' person identification pipeline for celebrity detection is a really powertool tool that can be easily deployed by customers and adapted to using their own data. By seamlessly integrating tools like Label Studio, NFS Storage, Jupyter Notebook, ClearML, MLflow and Streamlit server, Edgebricks was able to quickly build a robust and versatile solution capable of accurately identifying celebrities' faces across various media platforms. Through this case study, it is evident that Edgebricks' cloud infrastructure, App store and toolset offer a powerful foundation for tackling complex computer vision challenges, driving innovation, and transforming industries wanting to do accurate and cost efficient person detection.