

Accelerating Bioinformatics AI use cases with Kubernetes

Alessandro Pilotti
Cloudbase Solutions
CEO



Use cases

- **Antigen-specificity**

- Given an antibody sequence, will it bind to a given antigen?
- Antigen: SARS-CoV-2 Spike protein
- **Binary classification**
 - Two labels: **yes, no**

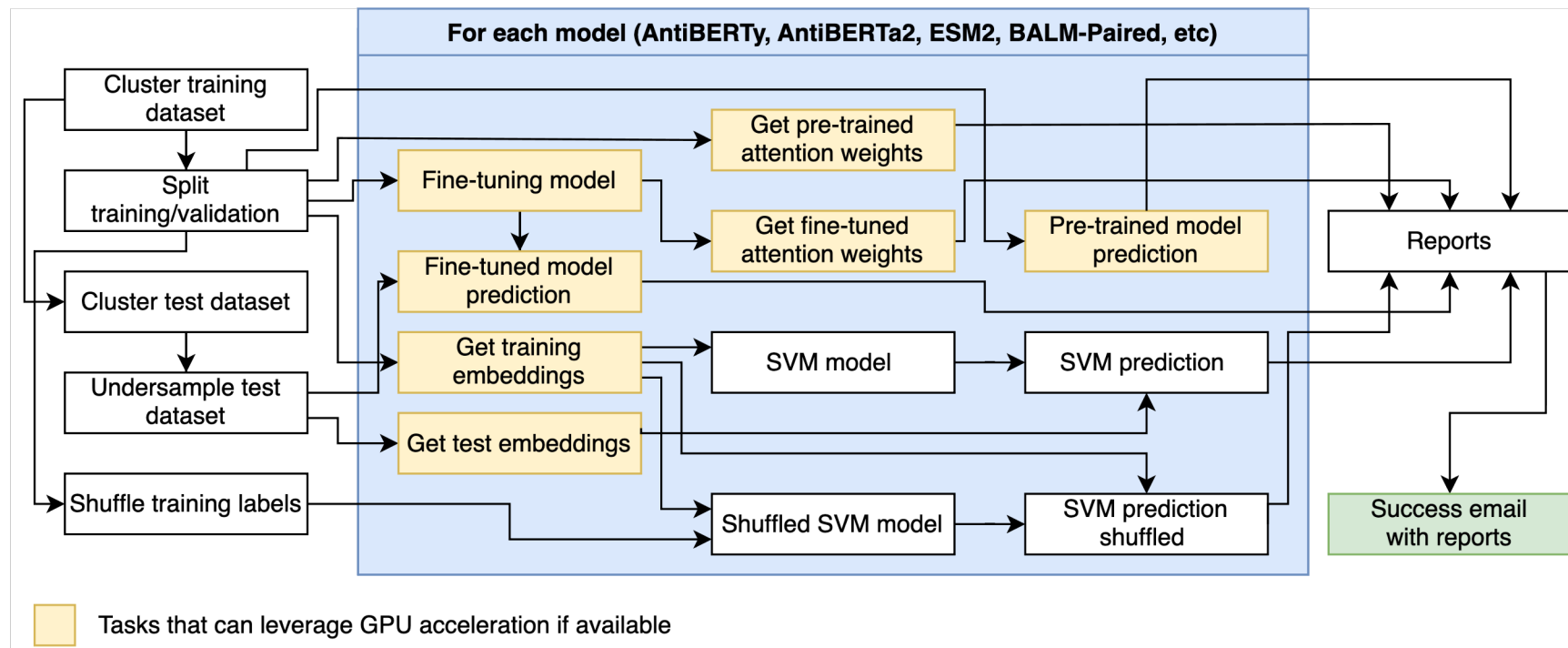
- **Paratope prediction**

- Given an antibody sequence, which positions belong to the paratope?
- **Token classification**
 - One label (yes, no) for each token

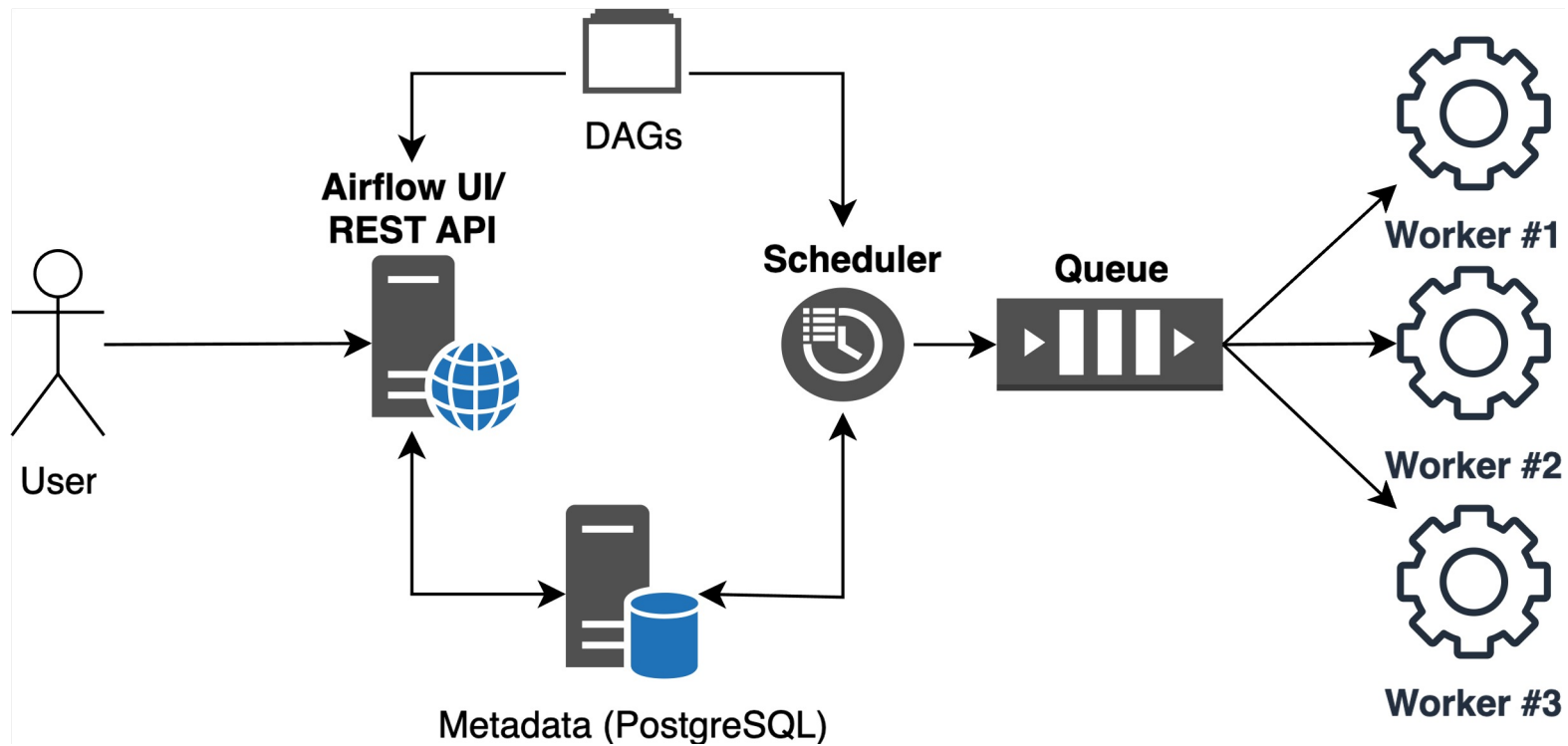
The need for an automate pipeline

- ~600 tasks between both sequence and token classification
- Most tasks require GPUs
- Tasks are connected in a directed acyclic graph (**DAG**) for the correct order of execution
 - **Apache Airflow**
 - **Nextflow**
- Resources can be spread across multiple clusters, e.g.,
 - **Kubernetes**
 - **SLURM**
 - **SGE**

The antigen affinity pipeline



Apache Airflow



Storage considerations

- All pods need to access the same shared storage
- A **ReadWriteMany** CSI is needed
 - **CephFS** (great for production, multiple nodes needed)
 - **NFS** (great for PoCs, easy to deploy)
 - If running on hyperscalers, evaluate native options e.g., AzureFile
- A single PV + PVC is created and attached to all pods during pipeline runs

Otimisations

- 1 CUDA: MIG vs Time Slicing vs MPS
 - **MIG** (+multitenancy, good for VMs)
 - **Time Slicing** (easiest, no isolation)
 - **Multi-process service (MPS)** (in between, still experimental in device plugin)
- 2 Scaling a large model across multiple GPUs (e.g. NVLink) and hosts (TCP/IP)
 - Huggingface accelerate
 - **DeepSpeed (ZeRo)**
 - 3 levels: increased optimisation outcome and complexity from 1 to 3



DAG	Yes	Yes
Year	2014/10	2013/03
Language	Python	Groovy
Mandates DSL use	No (very flexible)	Yes (very opinionated)
GitHub stars	44.1k	3.3k
Learning curve	Easy	Fairly steep
UI	Yes	No
HPC executors	K8s	SLURM, K8s, SGE, etc
Cloud executors	Yes	Yes
Easy to extend	Yes (providers)	To a point (plugins)
REST API	Yes	No
Multitenant	Not fully	No (it's just a process)
Can it be integrated in 3rd party solutions	To a point (strict architecture)	Definitely (it's just a process)
Governance	Apache Foundation	Largely driven by a single company
License	Apache 2	Apache 2
Strong Bioinformatics community	No	Yes (nf-core)

Repositories

- <https://github.com/alexpilotti/bbk-mres>
 - All the atomic building blocks of the pipeline
 - Latest version always checked out during pipeline executions
- <https://github.com/alexpilotti/bbk-mres-airflow>
 - Apache Airflow DAG definitions
- <https://github.com/alexpilotti/docker-bbk-mres>
 - Dockerfiles
 - One for Python workloads with CUDA support
 - One for R workloads (reports)