Data Handling for ML and Analytics

Data Handling for ML and Analytics

Carla Takahashi

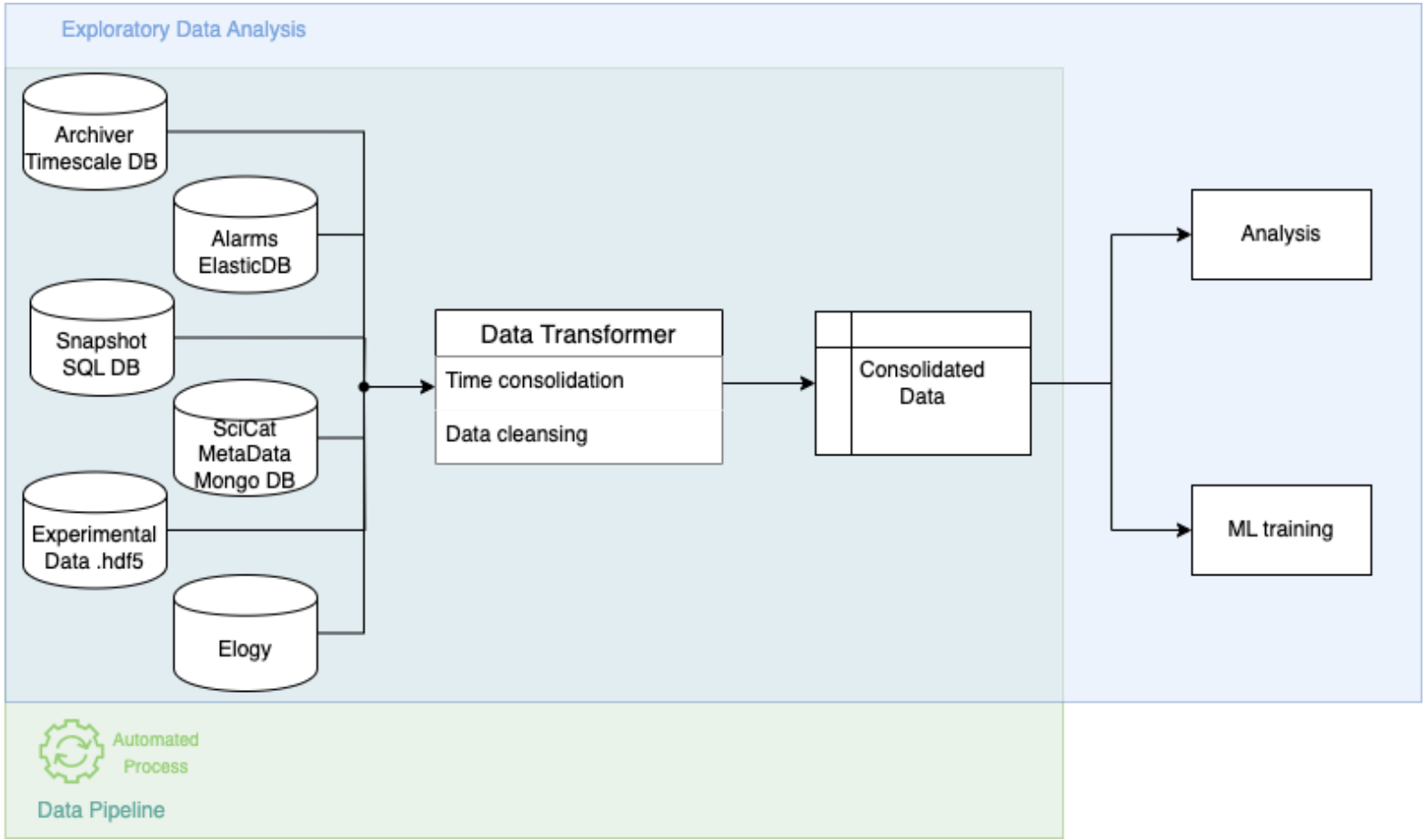
Data Sources

- Archiver DB using Timescale.
- Alarms Logs.
- Snapshot Database.
- SciCat Database.
- Experimental Data.
- Electronic Logbook (possibly?).

	Characteristics	Challenges
Archiver	• From the control system side. • Our main source of historical data. • Stored on a Timescale database and allow some processing to be done on the db. • Can be accessed from the white network.	• From the machine side, there is a large amount of data available. • From the beamlines side not much data is being archived. • The data is not consolidated in time.
Alarms	• Stores persistently all alarms. • Stored in a Elastic Search.	• It is a very large dataset. • Data time resolution depends on the PLC time resolution.
Snapshot	• Control system variables are stored triggered by different processes. • For each snapshot data is consolidated in time.	• The DB lives on the same host as the Tango DB and can not be accessed on the white network. • It is uncertain the volume and quality of data stored there.
SciCat	• SciCat database can also be used to metadata as input models. • The quality of the experiment is one of the features available on the metadata.	• Scientists are still on the process of learning scicat, thus the availability of data might be an issue.
Experimental Data	• Experimental data, can be used to train models both for data analysis and for beamline alignment.	• The use of experimental data might be subject to user/beamline restrictions. • Experiments vary greatly, lead into data volume issues: too much or too little data.
Elogy	• Electronic Logbook with human inputed data. • Might contain calibration and operational information provided by scientists and operators.	• Data is not curated or structured. • Requires language processing to be used precedurally.

DATA CONSOLIDATION

- Even within the same data source, such as the Archiver, different attributes have different archiving periodes, so we can say that our data is not consolidated in time.
- Different experiments and setups have different sets of metadata and snapshots stored. Thus, even considering a long period of time, the availability of data might be inconsistent.
- For experimental data, the scope tends to be better defined, both in time and which variables need to be stored, however, depending on the experiment the volume of data might be a challenge.

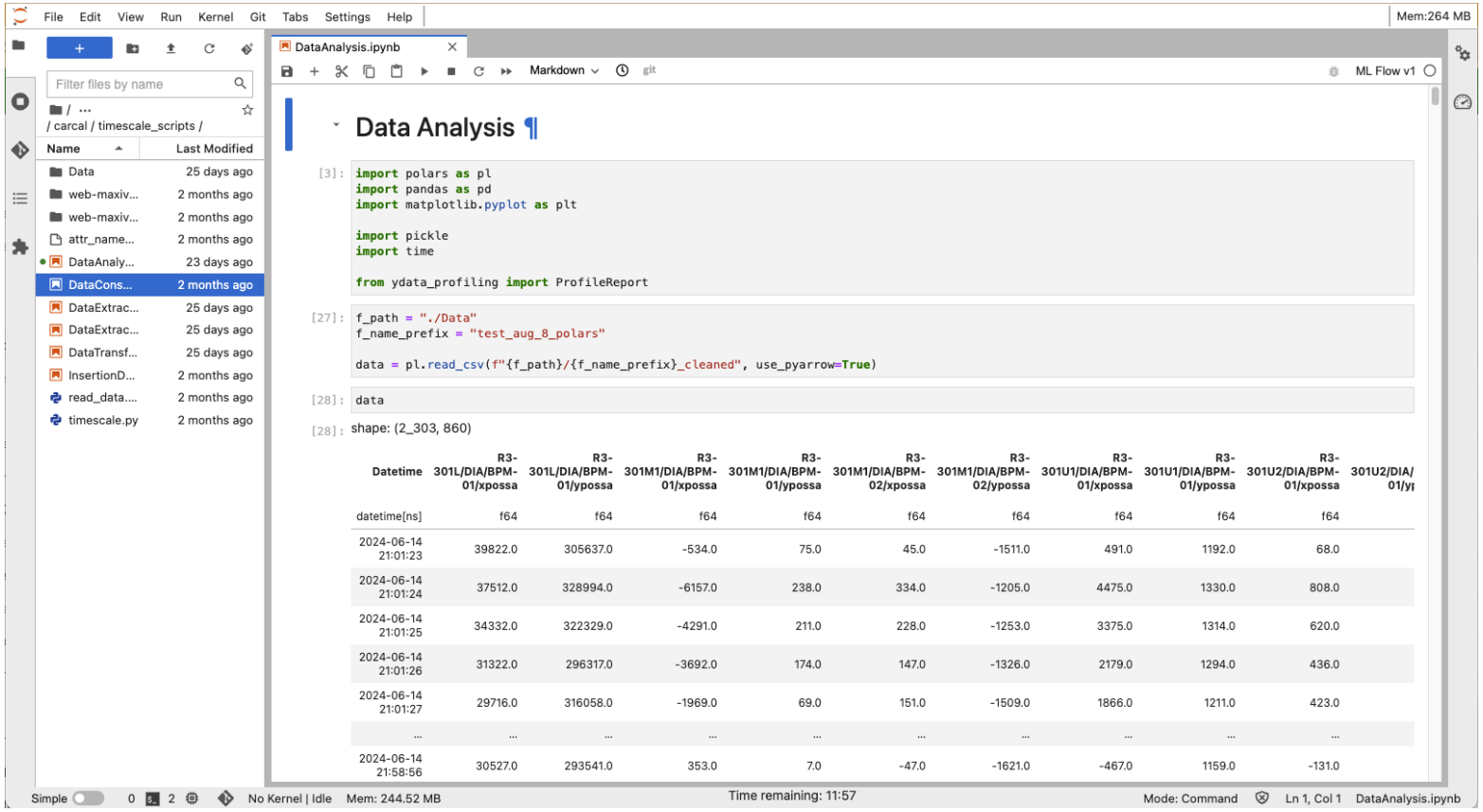


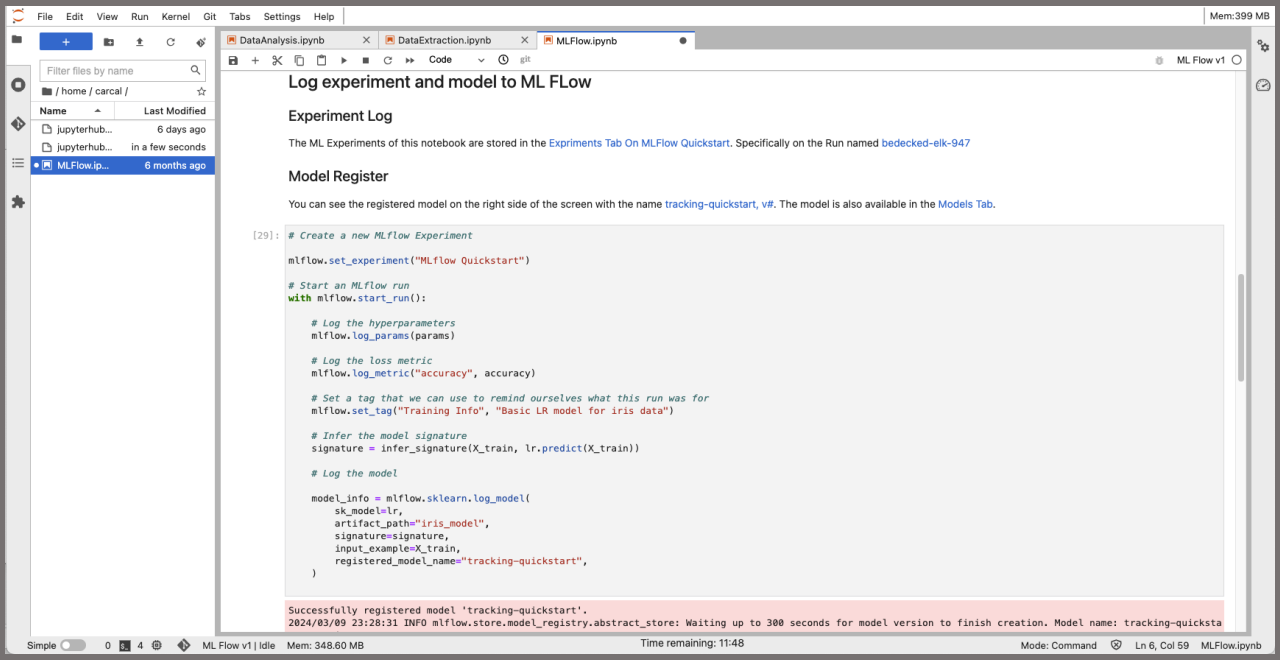
Data Pipeline Challenges

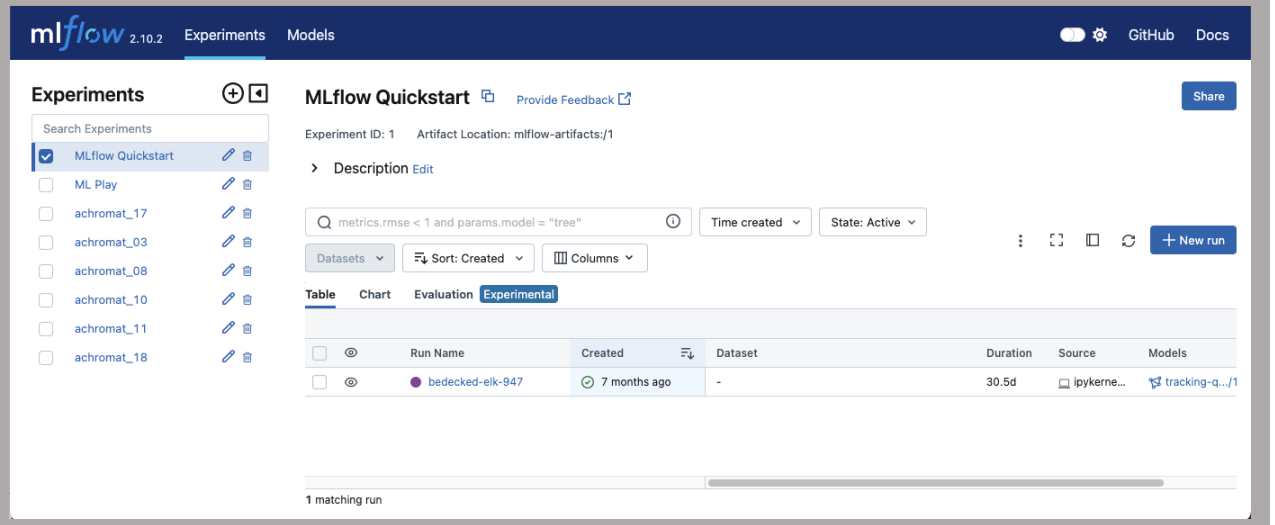
- Currently, there is no standardized location for automated data processing. As a result, the ETL process is performed directly on the high-performance cluster and manually triggered through Jupyter notebooks.
- Similarly, there is no defined storage location for final training data blobs, which is important for auditing the final models. For now, these data are stored either locally or in Active Directory storage.
- Alternatively, the process used to generate the data could be stored, allowing for data regeneration if needed.
- Implementing an Airflow-based ETL solution is advisable.

HIGH PERFORMANCE CLUSTER

- Jupyter Hub runs on the HPC.
- Has GPU nodes available.
- Can connect to the Archiver Timescale DB.
- Can connect to ML Flow to track experiments.
- Can create specific Kernels for different experiments.





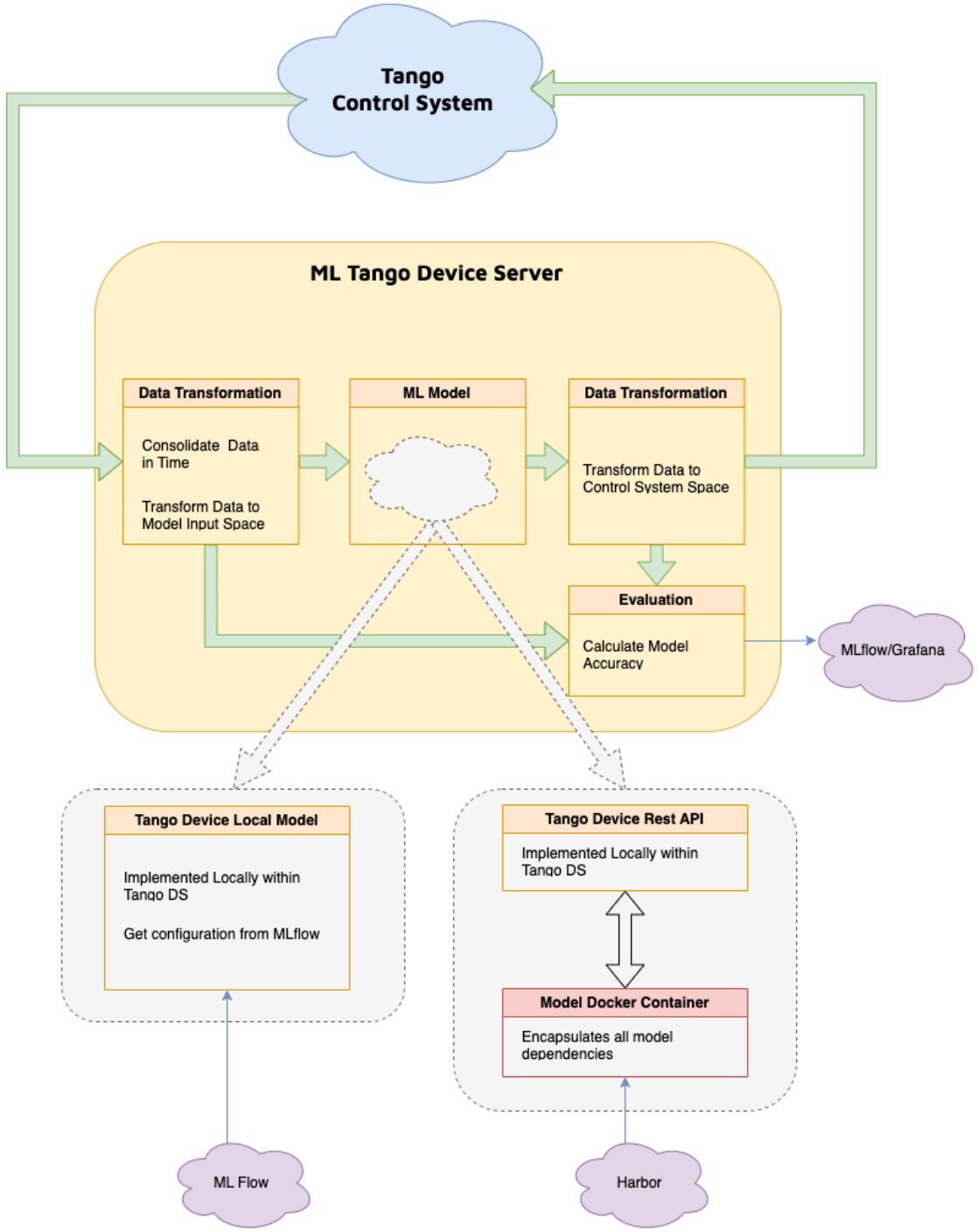


ML Flow

- Can be accessed from the HPC Jupyter Hub.
- Using the MLflow api within the jupyter notebook, it is possible to compare different runs for the same experiment.
- MLflow can publish models, in order to create a traceable repository of stable models.
- The Web Interface can also be used to compare models and publish them.

REAL TIME DATA PROCESSING

- For use cases on control system applications (e.g. trajectory or alignment adjustment).
- Tango device deployed on the control system hosts.
- Should have access to the trained model.
- Should consolidate data collected on real time.
- Should apply model solutions on real time.
- Should have quality evaluation over time.
- Model deployment should be done via Ansible.



Current State

- There is currently no standardized solution for data handling.
- Most challenges are being addressed as they arise.
- A machine learning project focused on Insertion Devices Compensation is currently underway, and we are using this opportunity to learn how MAX IV should approach ML-based development.