Data and Machine Learning Model Management with Gypscie

Fábio Porto, Patrick Valduriez

Data and Machine Learning Model Management with Gypscie

Fabio Porto[1][0000−0002−4597−4832] and Patrick Valduriez[2,3][0000−0001−6506−7538]

[1] Laboratório Nacional de Computação Científica, Petropolis, RJ, Brazil
[2] Inria, University of Montpellier,CNRS, LIRMM, Montpellier, France
fporto@lncc.br, patrick.valduriez@inria.fr

The synergy of big data and machine learning (ML) has led to the new data science, with many benefits for data-intensive applications in terms of more accurate predictive data analysis and better decision making. For instance, in the context of the HPDaSc (High Performance Data Science) project between Inria and Brazil [3], we have shown the importance of realtime analytics using ML models to make critical high-consequence decisions, e.g., preventing an oil drill from getting stuck in the ocean subsurface salt layer based on a driller's realtime data, predicting extreme weather events and supporting realtime analytics based on the visualization of simulated data of the cardio vascular system, or the effectiveness of ML to deal with scientific data, e.g., computing Probability Density Functions (PDFs) over simulated seismic data using Spark.

As predictive analytics using ML models (or models for short) become prevalent in different stages of scientific exploration, a new set of artifacts are produced during the models' life-cycle that need to be managed [2]. In addition to the models with their evolving versions, ML life-cycle artifacts include the collected training data and pre-processing workflows, data labels and selected features, model training, tuning and monitoring statistics and provenance information. However, to realize the full potential of data science, these artifacts must be built and combined, which can be very complex as there can be many to select from. Furthermore, they should be shared and reused, in particular, in different execution environments such as HPC or Spark clusters.

In order to support the complete ML life-cycle process and produced artifacts, we have been developing the Gypscie framework, which offers collaborating researchers a common software infrastructure to develop, share, improve and publish ML artifacts. Figure 1 depicts the framework architecture.

Gypscie offers a web interface that easy the accomplishment of complex ML model tasks, even by non computer science savvy researchers. It also offers a notebook interface and an API for direct python scripts integration with the framework services. Users can build dataflows graphically to model data preprocessing tasks. Registered dataflows can be scheduled for execution and, during run time, have their activites and involved data recorded for provenance. Models can be trained using registered datasets and have stored the corresponding testing performance metrics. We integrate the Databrics MLFlow component [4] to interface Gypscie with different machine learning execution engines. Three

[3] https://team.inria.fr/zenith/hpdasc
[4] https://github.com/mlflow/mlflow

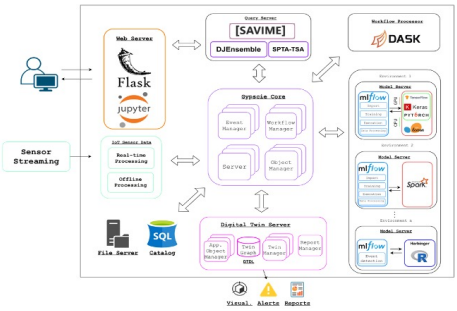

Fig. 1. Gypscie Architecture

aspects of Gypscie stand out. **Multiple-Environments:** Tasks in Gypscie can be scheduled in environments that best fit the running process. For data transformation processes, data locality is prioritized and Big Data frameworks like Apache Spark and Dask are usual choices. The Santos Dumont HPC system is a strong candidate for large scale data processing and their usage in large model training. **Model composition:** Gypscie automatically selects and allocates multiple spatio-temporal deep learning models, using the DJEnsemble approach [3]. The composition of published models is a particular motivation for sharing models within the framework. **SAVIME:** A multi-dimensional array in-memory DBMS [1] that enriches the framework with declarative, SQL-Like, query processing that is used to explore registered datasets and invoke ML models. SAVIME can process the datasets directly from their raw format with no data ingestion cost.

A first version of Gypscie has been deployed on two different applications. The first one supports oil exploration by managing models that predict the rupture of platforms stabilizers, as well as the corresponding online monitoring data. The second one refers to ML models predicting extreme rainfall events in the city of Rio de Janeiro. The latter runs on a shared-nothing cluster at LNCC and work is in progress to enable scheduling ML training using the Santos Dumont HPC system.

References

1. L. S. Lustosa, H., C. Silva, A., N. R. da Silva, D., Valduriez, P., Porto, F.: Savime: An array dbms for simulation analysis and ml models prediction. Journal of Information and Data Management **11**(3) (Feb 2021)
2. Miao, H., Li, A., Davis, L.S., Deshpande, A.: Towards unified data and lifecycle management for deep learning. In: 2017 IEEE 33rd International Conference on Data Engineering (ICDE). pp. 571–582 (2017). https://doi.org/10.1109/ICDE.2017.112
3. Pereira, R., Souto, Y., Chaves, A., Zorilla, R., Tsan, B., Rusu, F., Ogasawara, E., Ziviani, A., Porto, F.: Djensemble: A cost-based selection and allocation of a disjoint ensemble of spatio-temporal models. p. 226–231. SSDBM 2021, ACM, New York, NY, USA (2021)