IRIS-HEP Summer Fellow Proposal 2022

"Estimating transfer times of large datasets for scientific computing"

About me

I am Oleksii Brovarnyk, a fifth-year student at Kharkiv Polytechnic University. I'm studying computer science.

I'm interested in Big Data and Data Science. I completed courses on the Coursera platform: "Inferential Statistical Analysis with Python", "Fitting Statistical Models to Data with Python". I also study Python and improve my skills in this area.

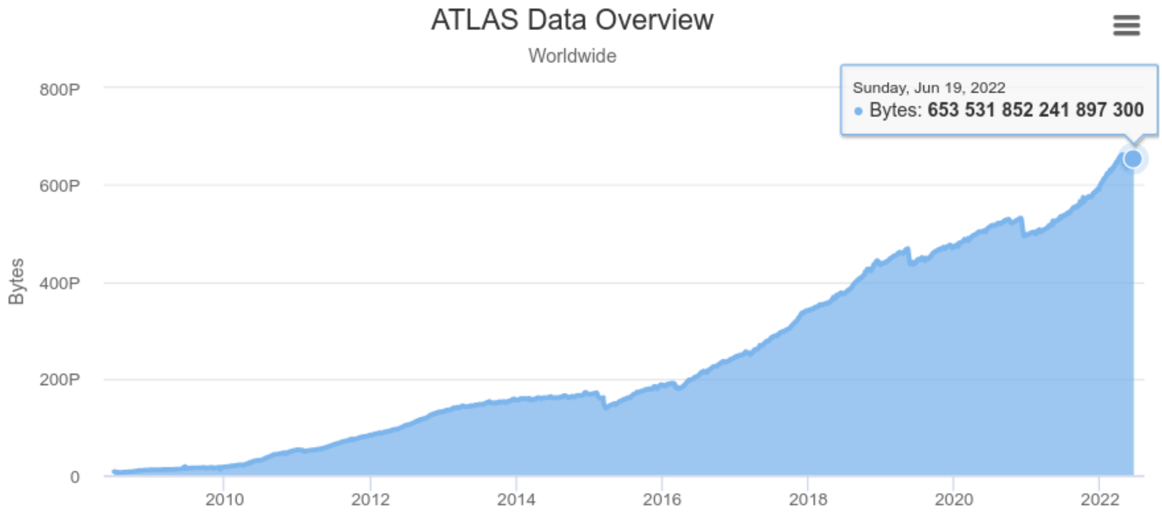
I have also worked with Arduino and compatible components and C-like languages.

I took part in a project that was provided by NTU "KhPI" and Temple University. Its objective was to unite students from different faculties with a common goal – to create a startup idea and bring it to implementation. Our team was working on a "quadcopter artist" that paints an image on surfaces and minimizes the danger of working in the field of industrial mountaineering. I learned how to work in a team, allocate the roles, and gained experience with robotics.

I believe the IRIS-HEP Fellowship will help me take the next step by giving me a good start and an unforgettable research and development experience.

About Rucio

Rucio is an open-source software framework that provides functionality to scientific collaborations to organize, manage, monitor, and access their distributed data and dataflows across heterogeneous infrastructures. Rucio was originally developed to meet the requirements of the high-energy physics experiment ATLAS, has been adopted by CMS, and is continuously enhanced to support diverse scientific communities, such as LIGO, SKA, Xenon, and many more. Since 2016 Rucio has orchestrated multiple Exabytes of data access and data transfers globally.



Project description

This project will continue the existing research of the Rucio team on the estimation of the duration of file transfers for large scale sciences. The distributed data management environment for scientific experiments forms a complex ecosystem with dynamic interactions between users and data centers. The accuracy of the predictions of the models will be limited by the amount of data about the system available at a given moment, and by the stochastic processes involved in certain parts of this system. Rucio's central role as the data management system, and the rich amount of data gathered about the transfers and data rules life cycles will help in creating machine learning for transfer time estimation.

To start, the research needs data on successful file transfers. This is based on the Rucio events (event_type: transfer-done) which are available in the Rucio Elasticsearch instance. Access to this data will be provided by CERN. The idea is then to generate time series from this data, including event metadata like "started_at" (when did the transfer start), "transferred_at" (when did the transfer finish), "bytes" (how big was the file), "rule_id" (to which user request belongs this file), and many more. There are more than 30 different variables that can be used for this model. The Rucio events database also contains other events, such as deletion events, or transfer failure events. These can be used for more complex analysis.

To start the estimation, the inputs for a first model will be the number and size of files, and the outputs are the duration in seconds. From this, a first linear regression model will be learned, which should answer the question: For a given dataset (number of files, and gigabytes), how long will it take for this dataset to finish transferring. After this first model, we will enter into a cycle, which we use to add (or remove) variables from the available input data to see which ones improve the prediction the best, and validate against the history for correctness. We will then repeat this cycle as often as possible to get the best possible prediction.

The study uses the Python programming language, the PyTorch machine learning framework, and Google Colaboratory.

Project plan and milestones

Week 1-2

Get access to Rucio events on Elasticsearch. Write Python code to extract data from Elasticsearch and convert from JSON to PyTorch native tensor. The data for the research for the last 7 days are taken, with the possibility to expand to the last 30 days. Use the time to understand how Rucio events work and the specific meaning of the variables in the data.

Week 3

Build a basic machine learning model with PyTorch (most likely a straightforward linear regression neural network).

Week 4

Validate the model. Take the history of the events, and compare with model output. Automate the display of the histogram of errors.

Weekly cycle (5, 6, 7, 8)

Further development and improvement of the model in short cycles. Validation of the model with additional parameters, and continuous improvement.

Week 9

Finish documentation, prepare the results and prepare a presentation.

References

[1] Barisits, M., Beermann, T., Berghaus, F. et al., Rucio: Scientific Data Management, Springer CSBS (2019), https://doi.org/10.1007/s41781-019-0026-3

[2] J. Bogado, M. Lassnig, F. Monticelli, J. Diaz, Modelling large-scale scientific data transfers, accepted for publication, Springer CSBS 2022

[3] J. Bogado, M. Lassnig, F. Monticelli, J. Diaz, T. Beermann, Zenodo (2020), 10.5281/zenodo.4320937

[4] M. Lassnig, W. Toler, R. Vamosi, J. Bogado, Journal of Physics: Conference Series 898, 062009 (2017), 10.1088/1742-6596/898/6/062009

[5] V. Begy, M. Barisits, M. Lassnig, E. Schikuta Journal of Computational Science 44, 101158 (2020), https://doi.org/10.1016/j.jocs.2020.101158

[6] J. Bogado, F. Monticelli, J. Diaz, M. Lassnig, I. Vukotic, in 2018 IEEE 14th International Conference on e-Science (e-Science) (2018), pp. 334–335, 10.1109/eScience.2018.00081