ESTIMATING TRANSFER TIMES OF LARGE DATASETS FOR SCIENTIFIC COMPUTING

Brovarnyk O[1]., Lassnig M[2].
[1]National Technical University «Kharkiv Polytechnic Institute», Kharkiv, Ukraine
[2]Conseil Européen pour la Recherche Nucléair, Meyrin, Switzerland

This work continues the existing research on the calculation of the transfer time of large datasets in the Rucio distributed data management environment.

The Rucio environment is very complex and there are many dynamic interactions between users and data centers, so you need to check existing results and look for new approaches in calculations.

There were used in the work the Google Collaboratory platform and the framework for modeling neural networks TensorFlow and Keras.

For the initial analysis, metadata for successful file transfers in the system was obtained, variables that affect file transfer time were transformed and highlighted.

In order to save time and reduce the number of calculations, a random sample of 10,000 records was used. In the analysis we used different samples to check whether the results are similar in all available data.

Plotting of dependence and primary analysis gave the answer that only 7 main variables should be left for further work.

A linear regression baseline is needed as the first model to estimate, something easy to understand and to compare. A linear regression model shows that the data are linearly independent.

Several models of neural networks have been created, the final one is a model based on two input levels for numerical and categorical variables, and then combined into one branch.

Prediction results are visualized using base RMSE and normalized RMSE to allow comparisons between different models and datasets, an error histogram created to make the results easier to understand when using large datasets, and a scatter plot to compare training data to predictions.

Since the Rucio system transfers files in groups, this behavior was simulated and files were grouped by 10 units, and calculations were made.

The performed calculations show good prediction results with a maximum prediction accuracy of almost 98 percent.

Continuing research on this topic will make it possible to improve the neural network model and increase the accuracy of predictions.