Contribution ID : **32** Type : **Workshop**

Scalable Scientific Analysis in Python using Pandas and Dask

Thursday, 30 August 2018 13:00 (300)

Pandas is a Python package that provides data structures to work with heterogenous, relational/tabular data. It provides fundamental building blocks for a powerful and flexible data analysis. Pandas provides functionality to load a wide set of data formats, manipulate the resulting data and also visualize it using various plotting frameworks. We will show in the workshop how to clean and reshape data in Pandas and use the concept of split-apply-combine to do exploratory analysis on it. Pandas provides powerful tooling to do data analysis on a single machine and is mostly mostly constrained to a single CPU. To parallelize and distribute these tasks, one can use Dask.

Dask is a flexible tool for parallelizing Python code on a single machine or across a cluster. We can think of dask at a high and a low level: Dask provides high-level Array, Bag, and DataFrame collections that mimic NumPy, lists, and Pandas but can operate in parallel on datasets that don't fit into main memory. Dask's high-level collections are alternatives to NumPy and Pandas for large datasets. In the low level, Dask provides dynamic task schedulers that execute task graphs in parallel. These execution engines power the high-level collections mentioned above but can also power custom, user-defined workloads. In the tutorial, we will cover the high-level use of dask.array and dask.dataframe.

Summary

Presenter(s) : NEUBAUER, Sebastian (Blue Yonder); KORN, Uwe

Session Classification : Tutorials