Tucson, May 5th, 2024

Dear Committee,

Doctoral thesis review for “Hierarchical Semi-Sparse Cubes – scalable solution for combining dimensionally multi-modal big data” by Jiří Nádvorník at the Faculty of Information Technology CTU in Prague.

Timeliness of the dissertation

Large datasets are becoming the standard in astronomy and it is no longer possible to process data files in the way many people are used to. Once millions of files are involved a way to index them and access them efficiently is very important. This thesis shows one approach that could be very important in the future.

Structure of the dissertation

The thesis consists of 10 chapters and with two appendices and consists of 137 pages. Chapter 1 introduces the problems and chapter 2 describes the requirements. Chapter 3 introduces possible technologies that could be used. Chapters 4, 5, and 6 describe the architecture of the final system and chapter 7 describes the implementation. Chapters 8 and 9 describe the performance of the delivered system and compare it with a defined baseline. Finally chapter 10 provides a conclusion. Appendix A describes the process for using SDSS data in the test system and Appendix B describes the test data.

Completion of the dissertation objectives

The thesis set out with a goal of being able to efficiently visualize and analyze millions of multi-dimensional datasets with varying times, wavelengths and dimensionality. Two approaches were presented that show good performance on single nodes and HPC clusters. Enough was demonstrated to support continuing this work with an aim of supporting SKA and Rubin LSST.

Assessment of the methods used in the dissertation

The adoption of HDF5 as the core technology was justified and the implementation using HDF5 demonstrated that it was the right decision. I was impressed by the ability of the system to scale with HDF5 container sizes in the hundreds of terabytes split over multiple nodes. The use of Python is reasonable given the prevalence of Python in the astronomy community and any way to lessen the barrier to entry is good.

I worry that integrating datasets from multiple observatories will become a challenging cross-calibration and data modeling endeavor. I concede that this is out of scope of the thesis as presented.

Evaluation of the results and contributions

The performance results presented indicate that this could potentially be a very interesting approach in the era of Big Data. Whilst I do wonder if using Astropy as the FITS baseline was a good choice (it is going to be slower than using cfitsio) the performance gains were clearly significant.

Comments and questions

- Given that the system is designed with a time coordinate, what is the expectation for handling proper motions where the same object will not be in the same place during each epoch?
- How will the association process handle source confusion where a single source in one image from one instrument may consist of multiple sources in an image from another instrument?
- Since query results are materialized in the HDF5 container, is there any worry with access controls where multiple users can see the results?

Overall evaluation

I was impressed with the implementation and the performance demonstration. The author of the dissertation proved the ability to conduct research and achieve scientific results. In accordance with par. 47, letter (4) of the Law N.r 111/1998 (The Higher Education Act) I do recommend the thesis for the presentation and defense with the aim of receiving the Ph.D. degree.

Sincerely,

Dr Timothy Jenness
Data Abstraction Group Leader
Vera C. Rubin Observatory