Norbert Podhorszki

PO Box 2008
Oak Ridge, TN 37831-6057
(865) 574-7159 | pnorbert@ornl.gov

Dr. Norbert Podhorszki
Distinguished Research Scientist
Computer Science and Mathematics Division
Oak Ridge National Laboratory
Oak Ridge, TN, 37831, USA

May 22, 2024

Doctoral Thesis Review

Hierarchical Semi-Sparse Cubes - scalable solution for combining dimensionally multi-modal big data
Submitted by Ing. Jiří Nádvorník

The thesis has been submitted to the Faculty of Information Technology, Czech Technical University in Prague, in the doctoral study programme Informatics, in partial fulfilment of the requirements for the degree of Doctor in 2024.

Up-to-datedness of the dissertation

The timeliness of this dissertation is very high. The field of astronomy has produced large amount of data in the past, and scientist have already been struggling to process all data so that no phenomena is left undiscovered in the data. The technology for astronomy, however, far outpaced the progress in the development of data management for astronomy. The new and future instruments, like LSST, SKA, James Webb, the Roman Telescope, and others, will produce unprecedented amount of data, for which no one is ready yet. The thesis sets out to transform the speed and easiness of processing, both for visual exploration by a human and for machine learning algorithms, a specific subset of optical observational data, defined as “dimensionally multi-modal data”, which requires efficient and scalable storage and processing of semi-sparse data sets. On the other hand, the author has been using the latest features of the HDF5 I/O middleware still in development, and provided feedback to the developers, to achieve the scalability required to achieve the goals of the dissertation.

Formal structure and organization of the dissertation

The thesis is organized into 10 chapters spanning 97 pages, and 40 more pages for the appendix. The thesis starts with an introduction to astronomy surveys and explains the main goals of the dissertation to provide fast visualization and fast data processing to data combined from multiple data sources. The introduction also provides the definitions (dimensionally multi-modal data) that are used later in the thesis and introduces the HiSS-Cube framework.
Chapter 2 describes the requirements for HiSS-Cube, including the image processing, uncertainty computation, combining images with spectral measurements, creating multiple resolutions and so on.

Chapter 3 is dedicated to review state-of-the-art data management solutions to similar problems and to evaluate them for the requirements for this work, and to explain why the Hierarchical Data Format (HDF5) was chosen as the basis for data storage in HiSS-Cube. The HiSS-Cube overall architecture is described in Chapter 4. A sequential design is presented in Chapter 5. The design of the parallel solution is detailed in Chapter 6. Chapter 7 provides the implementation details for both the serial and parallel version of HiSS-Cube. Chapter 8 focuses on the evaluation of the serial version, showing the performance of the visualization tasks, one of the aims of the dissertation. Chapter 9 evaluates the performance of the parallel version, focusing on the second aim of the dissertation. The conclusions are formulated in Chapter 10. Appendix A includes the documentation of the demonstrated spectra and images preprocessing needed for their combination with the full Python codes.

Completion of the dissertation objectives

The overall objective of the dissertation is to create a framework for fast processing of combined dimensionally multi-modal big data and to apply this framework to combine astronomical spectroscopic and photometric data. There are two major goals specified that the author set out to achieve:

1. Fast visualization of combined multi-dimensional big data.
2. Fast machine access to combined multi-dimensional big data.

The thesis introduces the concept of “semi-sparse data”, which is the result of when one combines dimensionally multi-modal data from multiple data sources. Then it presents the design and implementation of the HiSS-Cube framework, the software developed by the author. HiSS-Cube is designed for future astronomical data sources and for multi-petabytes of data.

The framework is tested on data obtained from the Sloan Digital Sky Survey (SDSS), which contains about 60 TB of images and almost 1TB of spectral data. Detailed examples show the ease of use of HiSS-Cube by a human for visual exploration and fast querying of the data. The performance scalability tests also prove that the entire SDSS database can be processed within an hour by machine learning algorithms if a sufficiently large compute cluster with a sizable and fast parallel file system is available. The performance of HiSS-Cube is excellent. The work in the thesis has also been published in peer-reviewed journals, aimed at both the astronomical community as well as the data management computer science community. Therefore, I consider the aims of the dissertation to be fulfilled.

Assessment of the methods used in the dissertation

It is very commendable that the author has considered many relevant data organization and management paradigms and evaluated them for their applicability for the designated goal for managing semi-sparse, dimensionally multi-modal data with interlinked properties, uncertainties, and multiple resolutions. Both relational (row-oriented) databases and newer, column-oriented databases (SciDB, MonetDB), array databases (Rasdaman, TileDB) as well as Hadoop and the file-based hierarchical data model (HDF5) were considered. The author evaluated them based on the following main issues he foresaw for building HiSS-Cube: scalability, ability to combine dimensionally multi-modal data, efficient storage of semi-sparse data, flexibility, and modularity.

The performance evaluation of HiSS-Cube was conducted on the Karolina cluster of the university and with the full SDSS astronomical survey dataset. In all aspects, this is a very good, realistic use case and a data size is very large considering the available resources, and therefore the performance evaluation is a good indicator of HiSS-Cube general capabilities and scalability.

Evaluation of the results and contributions of the dissertation

The contributions of the thesis are original work and has high practical value for two areas. First, the astronomy community should accept that parallel processing is unavoidable for their data challenges, and the pressure from the majority of scientist demanding a serial solution is no longer valid. The product in this thesis demonstrates that one can create scalable solution that requires large computing resources to manage the data but still be accessible and easy to use by a human explorer. Second, the work in this thesis is pushing the limits of the latest HDF5 features, which have been developed in the U.S. Exascale Computing Project concurrently with this work, and it provides invaluable feedback to the HDF5 developers. The two reviewed publications of the author, relevant to this thesis, are also a good example of this. One was published in the Astronomy and Computing journal of Elsevier, describing the HiSS-Cube product for astronomers, while the other was published in IEEE Access, describing the parallel framework.

Remarks, objections, notes and questions for the defense

First of all, I need to state that I have been the lead developer of the ADIOS I/O Framework for the last fifteen years, which is, by now, the only viable competitor for extreme scale I/O to the HDF5 I/O middleware. These two products have been supported in the U.S. Exascale Computing Project for supporting extreme scale data producers and consumers. Therefore, I have had strong bias in the evaluation of the thesis. On one hand, it saddens me that the author has not considered ADIOS at all for this work. I would have been very interested to see an independent comparison of HDF5 and ADIOS features. The fact that HiSS-Cube only become scalable at the end when using the latest (unproven, unstable) feature of subfiles in HDF5, which is an implementation of one of the basic features of ADIOS' file-format designed foremost for scalability, validates my feeling that the author should have started with ADIOS in the first place. On the other hand, I acknowledge, that this work provides a much-needed real-world success story for the competition.

The only notable error in the dissertation is that Figure 9.2 on page 85, “The HiSS-Cube running times without Global database query phase” still includes the global DB query phase. The figure in the published IEEE Access paper (Figure 11) should have been placed here instead.

My main criticism of the thesis is the following. The thesis specifically states that *“The performance of HiSS-Cube is bounded by the I/O bandwidth and I/O operations per second of the underlying parallel file system, and it scales linearly with the number of I/O nodes.”*
This is technically true, but in my opinion, it is because it cannot achieve optimal performance at lower scales. And therefore, it is not the important metric for HiSS-Cube. Rather, I would say that HiSS-Cube is not running as fast as it should, all but due to the underlying problems of the interplay of HDF5 and Lustre. This is not the fault of the author at all, but rather the consequence of the choice (HDF5), and that performance with HDF5 is always a work in progress at large scale.

Let me explain my reasoning here. Figure 9.7 shows Write time and Other time based on measurements using the Darshan tool. However, the total of those numbers is far below the total runtime length of the image construction phase, and therefore I cannot trust them as valid values. I must estimate times from Figure 9.4, on 18432 CPU cores, which is the fastest time for constructing the database. The image construction time (purple color, label 5 image data) is about 1/5.33 of 10,000 seconds in the chart, that is 1875 seconds. Considering that the Write and Other times in Figure 9.7 is about 50-50% of the total time, I will assume 1875/2 = 937 seconds used for Writing in the best of all runs (18432 CPU cores with I/O aggregators). The output image data size is 120 TB from the SDSS survey as stated in section 9.2. This gives me a 130 GB/sec production speed of the image data. This speed contrasts with the theoretical write bandwidth of 730 GB/s of the test cluster’s Lustre file system. That is, HiSS-Cube utilizes 18% of overall bandwidth. Not bad at all, considering how the previous single shared-file approach of HDF5 stayed below a few percent (e.g. 2% of total bandwidth of Frontier supercomputer’s Lustre file system, which is about 10 TB/s). For everyone but ADIOS users, this is an excellent result.

This result is achieved with 1 subfile per compute node, and 1 I/O aggregator per node used by the subfile driver in HDF5. Incidentally, this is ADIOS’ default setup for file I/O, and is known to underachieve in terms of write bandwidth at moderate scales. For example, exascale simulations (WarpX, XGC, S3D, PIConGPU) on Frontier need to run on about the half of the system (4096 nodes and above) to let ADIOS default setup utilize 50% or more of the Lustre file system’s bandwidth. If someone wants to write faster on less nodes, one need to change the setup to have more I/O aggregators per node.

Moreover, I believe that setup used in the thesis was not optimal for increasing the write bandwidth and therefore to minimize the total runtime of the database construction. The author used all CPU cores and tested the scalability by using more and more nodes. However, the author did not measure the effect of using less-then-available cores. Since the workload is not computationally bound but I/O bound, it would probably be better to use fewer cores per node, and more nodes to increase the writing bandwidth and hence decrease the image data phase writing time, the dominating part of the runtime.

Question for the defense: the subfile driver of HDF5 was supposed to utilize the performance of the file-per-process pattern on Lustre, i.e. up to 350 GB/s on the Karolina cluster at 72 nodes and above. However, the test results indicate below 20% of that. What is the reason for this mismatch?

Overall evaluation of the dissertation

The dissertation is well written and is easy to follow. The specific goals for optical astronomy to combine images and spectra helped throughout the thesis to understand the design, the algorithms and the challenges. Despite my criticism of some details, I believe this thesis is valuable. It contains new and original contributions to an important aspect of processing astronomy observational data for current and future telescopes, as well as valuable evaluation and feedback to the HDF5 developers to further improve their product for extreme scale data management.

Recommendation

> **The author of the dissertation has proved his ability to conduct research and to achieve scientific results. In accordance with par. 47, letter (4) of the Law Nr. 111/1998 (The Higher Education Act), I do recommend the thesis for the presentation and defense with the aim of receiving a Ph.D. degree.**

Sincerely,
Norbert Podhorszki, PhD

Podhorszki Norbert

Distinguished Research Scientist
Workflow Systems Group
Computer Science and Mathematics Division
Oak Ridge National Laboratory