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 Ph.D. study programme Informatics in December 2023.

Formal structure and organization of the doctoral thesis

The manuscript is 137 pages long and is organized into 10 chapters: a bibliography with 75 references, 2 reviewed publications of the author relevant to the thesis, 5 remaining author's publications relevant to the thesis (including posters and personal communications), and 2 appendices (39 pages in total).

The structure of the thesis is well-organized and easy to follow. The introductory chapter sets the stage for the research, followed by a focus on the requirements and definition of research problems in Chapter 2. Chapter 3 provides an overview of the state-of-the-art, while Chapters 4 to 9 are dedicated to the core proposal. Finally, Chapter 10 concludes the thesis by summarizing the contributions and suggesting future research directions. The six chapters that describe the author's proposal are thoughtfully divided into two parts: the architecture design and main components are presented in a progressive manner, followed by the practical implementation and evaluation of the two proposed variants of the solution.

Up-to-dateness of the doctoral thesis

The objective of this thesis is to develop efficient storage and access methods that can handle multi-modal big data gathered from various scientific instruments. The challenge of dealing with observational big data of various types is significant due to the discrepancies and diverse scales. Thus, this thesis addresses an important issue that is crucial in several scientific applications, such as astronomical surveying and remote sensing in the realm of earth observation.

The topic pertains to array databases, which play a key role in the realm of scientific data management and continue to remain an active area of research. Array databases are a crucial tool for managing and analyzing large datasets, especially in scientific applications, and their effective utilization is critical for advancing scientific knowledge. Given their importance, researchers have been exploring various aspects of array databases, including their performance, scalability, and usability. However, no efficient and scalable system combining various types of array data exists.

I believe that the topic provided holds significant value and is sufficient for verification of the originality of the solution within the field of scientific data management.

Completion of the dissertation objectives

The dissertation created a framework that aims to process dimensionally multi-modal scientific data in a highly parallel manner. The data organization in HiSS-Cube wisely leverages the HDF5 file format while adapting it to semi-sparse and uncertain data. Through experiments on large spectroscopic and photometric data, the dissertation has demonstrated the scalability and efficiency of HiSS-Cube, and its potential in handling Petabyte-scale big astronomical. Additionally, the proposed framework can be easily extended to other scientific data types by modifying the preprocessing code.

As such, the initial objectives of the dissertation have been reached.

Assessment of the methods used in the thesis

The dissertation employs scientific methods, techniques, and tools consistent with the state-of-the-art.

The purpose of the thesis is to achieve efficient query processing and machine learning over dimensional multi-modal big data by providing efficient, flexible, and scalable access to the data. It is motivated by the importance of combining multiple modalities originating from different instruments and points out the induced data semi-sparsity and uncertainty problems. In the candidate's analysis, there is no efficient software available to combine multi-dimensional multimodal big data with different incompatible dimensions or densities. Hence, the thesis thoroughly examines the issues that need to be addressed, including scalability, multimodality integration, hierarchical access, and uncertainty management. The dissertation introduces the proposal of a framework called Hierarchical Semi-Sparse Cube (HiSS-Cube).

A critical aspect of the proposed framework is its modularity. This means that the framework should be based on well-supported and industrial-standard technology. The state-of-the-art section focuses on mature systems and evaluates them based on the requirements and implementation difficulty of HiSS-Cube. Three systems, namely RasDaMan, TileDB, and SciDB, which specialize in managing multidimensional array data, are reviewed, along with some relational-based solutions and Apache Parquet and HDF5 formats and libraries. After a thorough examination of different array data management, relational, and columnar systems, it was found that none of them meet all the necessary requirements. However, HDF5 has proven to be a versatile and widely used storage format for managing large scientific data. It is also flexible enough to be adapted for the implementation of HiSS-Cube. Therefore, the elaborated solution builds on this assumption and covers the entire data process workflow.

Overall, the methodology and tools used in this research field are appropriate and reliable. They are sufficient for achieving the research objectives.

Evaluation of the results and contributions of the thesis

The dissertation demonstrates that HDF5, along with its I/O library, is a viable option for the primary storage format. The author proposes several contributions based on this, including the overall pipeline within the HiSS-Cube framework, optimized data models for both sequential and parallel settings, and an in-depth study of implementation options. The experiments were conducted on both a simple laptop and a powerful, highly parallel infrastructure. They are highly detailed and the results are convincing. The proposal's flexibility is discussed with respect to different scenarios, but a thorough evaluation with diverse data types would help ensure its effectiveness and applicability in various scientific domains. However, this is a minor remark compared to the important achievements made in this work.

Note that the developed software is open-source and available on GitHub, following the policy of open science.

Comments and questions for the defense

- The model addresses dense and semi-dense (i.e. semi-sparse) data. But what about the support of sparse data?
- The uncertainty formula (Eq. 4.1 in Chapter 4) is not a common standard deviation definition and should be clarified. Also, the visualization is based on another value derived from the pixel uncertainty, but without specifying its exact definition.
- In general, formal definitions of some features and measures are missing, which hinders the deep understanding of the approach for researchers who are not specialists in astrometry (e.g., flux density).
- As the thesis is in computer science, it would be useful to provide a technical description of HDF5, HEALPiX, HiPS, STMOC, and STEMOC, especially the features such as "region reference", "Virtual File Drivers", BITPIX, NAXIS, NAXIS in FITS and HDF5. This would make the dissertation self-contained.
- How are the different indexes combined and how is this evaluation done by the query benchmark?
- Most data processing steps depend heavily on the characteristics of the two datasets used to exemplify the method. There is no evidence to suggest that this can be extrapolated to other input data. If so, how difficult would it be?
- In Chapter 6, Compression is not utilized to avoid the write operations that create a bottleneck for HiSS-Cube files. However, the last sentence suggests that this might be reconsidered, albeit without explaining how to work around the performance issue.

Overall evaluation and recommendation

Jiří Nádvorník's thesis delves into an interesting topic concerning the management of scientific data. The author identifies the challenges and the lack of mature solutions and proposes novel methods and software components that have proven to be effective in addressing the raised issues. Specifically, the thesis presents a framework that enables an end-to-end combination of big scientific array data, multi-resolution storage, and processing based on HDF5. This framework has been demonstrated on combined photometric - spectrometric data. The results have been published in two peer-reviewed journal articles and shared at various scientific events. In addition, the software has been made available to the community as an open source.

For these reasons, **I do recommend** the thesis for the defense with the aim of obtaining the Degree of Ph.D.

Versailles, Mai 13, 2024

Karine Zeitouni, Professor