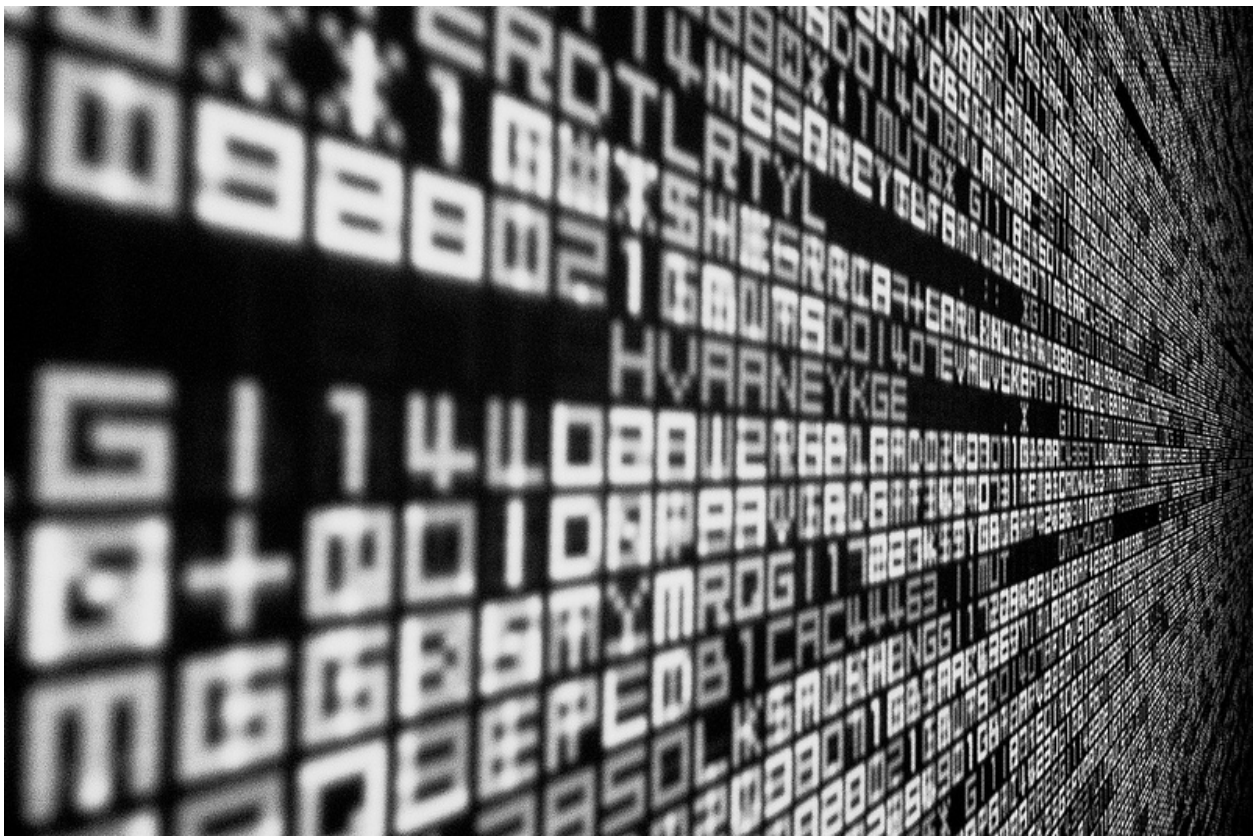
Earth Science Data Analysis in the Era of Big Data

K.-S. Kuo[1,2], T. L. Clune[2], R. Ramachandran[3]
1. Bayesics, LLC, Bowie, Maryland
2. NASA Goddard Space Flight Center, Greenbelt, Maryland
3. NASA Marshall Space Flight Center, Huntsville, Alabama



An illustration of data. Image Credit: r2hox.

Anyone with even a cursory interest in information technology cannot help but recognize that "Big Data" is one of the most fashionable catchphrases of late. From accurate voice and facial recognition, language translation, and airfare prediction and comparison, to monitoring the real-time spread of flu, Big Data techniques have been applied to many seemingly intractable problems with spectacular successes. They appear to be a rewarding way to approach many currently unsolved problems. [1]

Few fields of research can claim a longer history with problems involving voluminous data than Earth science. The problems we are facing today with our Earth's future are more complex and carry potentially graver consequences than the examples given above. How has our climate changed? Beside natural variations, what is causing these changes? What are the processes involved and through what mechanisms are these connected? How will they impact life as we know it? In attempts to answer these questions, we have resorted to observations and numerical simulations with ever-finer resolutions, which continue to feed the "data deluge." [2]

Plausibly, many Earth scientists are wondering: How will Big Data technologies benefit Earth science research? As an example from the global water cycle, one subdomain among many in Earth science, how would these technologies accelerate the analysis of decades of global precipitation to ascertain the changes in its characteristics, to validate these changes in predictive climate models, and to infer the implications of these changes to ecosystems, economies, and public health? Earth science researchers need a viable way to harness the power of Big Data technologies to analyze large volumes and varieties of data with velocity and veracity. [3]

Beyond providing speedy data analysis capabilities, Big Data technologies can also play a crucial, albeit indirect, role in boosting scientific productivity by facilitating effective collaboration within an analysis environment. To illustrate the effects of combining a Big Data technology with an effective means of collaboration, we relate the (fictitious) experience of an early-career Earth science researcher a few years beyond the present, interlaced and contrasted with reminiscences of its recent past (i.e., the present).

In a Not-so-distant Future …

E. S. Study is a budding young atmospheric scientist. She grew up in the Northern Great Plains of the U.S. and has always been fascinated by the local winter snowstorms: the tremendous power they wield and the surreal, desolate, powdery white world they usually leave behind. Her curiosity has compelled her to learn more about these storms as she advances in her education. She is now a Ph.D. candidate at a prestigious graduate school with a top-notch reputation in atmospheric science. She has her mind set on studying blizzards as her dissertation topic. With the powerful technologies that have just become operational, she and her adviser are confident that she will be able to finish within a year both her dissertation research and the dissertation with a kind of quality and thoroughness that was previously unimaginable without these technologies.

One of these new technologies is the PARAllel Data-Intensive Analysis and Management Environment (Para-DIAME) for Earth Science, which is funded jointly by several federal agencies with stakes in Earth science research and operated by the Joint Earth Science Data Information Office, JESDIO. Para-DIAME resides on a compute cluster of hundreds of thousands of nodes with terabytes (TB) of local storage on each node, totaling millions of computing cores and multiple exabytes (EB, 2^{60} or ~10^{18} bytes) for the cluster and is capable of supporting thousands of simultaneous users with real-time analysis performance.

Unlike the traditional high-end computing (HEC) clusters that rely upon large centralized high-performance disk systems, Para-DIAME is designed to exploit a new and lower-cost architecture, often referred to as a “shared-nothing” architecture. With traditional HEC systems, computational tasks are assigned to specific processors, and the requisite data migrate between the centralized disk system and the processors as needed. This approach is well-suited for applications that do not constantly “move” large volumes of data around. However, for data-intensive applications this traditional architecture becomes far less cost-effective and may not provide optimal performance.

The shared-nothing architecture of Para-DIAME lacks a centralized disk system but instead partitions and distributes each large-volume dataset to inexpensive disks attached directly to the nodes of the cluster. When the coordinator node (or head node) of the architecture receives a computing task from the user, it sends the task instructions to the nodes where required data are located. The processors of each of these nodes execute the task instructions on its local portion of the data; parallelism is thus achieved and data

movement minimized. Since the volume of typical task instructions is exceedingly small compared to the volume of the data to be processed, moving instructions across the cluster network connecting the nodes is much faster than moving the data, allowing Para-DIAME to achieve high performance with a relatively inexpensive type of hardware.

More importantly, Para-DIAME integrates on this shared-nothing architecture 1) a parallel, distributed array-based database management system, 2) advanced analytics suites taking advantage of the underlying parallelism, and 3) a collaborative infrastructure that facilitates flexible collaboration and provides interfaces to the environment for multiple popular data-analysis languages. Each of these components brings unique advantages over the "old way" of conducting Earth science research. Combined, a force-multiplying effect is achieved and aptly exemplifies the cliché: "The whole is greater than the sum of its parts."

Significance of the Array-model Database Technology

In such a Para-DIAME, with a number of long-term high-resolution global climate datasets cataloged in the array-based database, E. S. Study is able to quickly identify the potential blizzard events globally for the entire length of the records. She further finds in-situ and remote-sensing data, also cataloged in the database, offering more comprehensive and detailed observations for the events. Obviously, she needs to first formulate the appropriate criteria in order to query the database for these events, but that is part of what her science training is about and what scientists are supposed to do, unlike the chores of downloading, managing, and backing up data.

Prior to the advent of the array-based databases, the great majority of database management systems (DBMSs) were based on the table data model of columns and rows. Unfortunately, the table model is quite inept at accommodating multidimensional arrays, the data model in which most scientific data are expressed. This is the primary reason that, while a significant portion of contemporary business data has become "structured," scientific data largely remains only "semi-structured."

Data become structured when they are stored in, and managed by, database systems that support "queries" to conveniently extract only the part of the data relevant to users' interest at the time. For example, when we shop online, more often than not, we are interacting with structured data in databases through a Web-browser interface. If we want to buy a solid-state drive of 1-3 TB capacity, we would start with a search engine or point our browser directly to an online electronic or computer store and type "solid state drives" in a search field. Solid state drives of various capacities, form factors, manufacturers, etc., would show up in the search results. Often we would be offered the option to narrow down the selection by clicking on some checkboxes: certain capacity range, certain form factor, certain manufacturers, certain price range, etc. As we apply these filtering conditions, the search results refresh instantly and accordingly. We have the backend parallel databases to thank for this convenience and real-time performance.

Before databases based on the array data model became available, scientists conducting data analysis were deprived of the kind of convenience common to online shoppers. Scientific data were mostly packaged in files. If a dataset was particularly large in volume, it would be broken up and packaged into files of manageable sizes. Only the metadata of these files were likely to be stored and managed by a database system. Thus they became semi-structured, or only partially structured. This paradigm created much duplication and inefficiency, because querying only the metadata was much less targeted than

querying the data themselves. Moreover, one needed to download the whole file even when only a subset was needed for analysis.

Significance of the Integrated Analytics Suites

After finding the potential blizzard events, E. S. Study proceeds to obtain various statistics about these events. The advanced analytics suites integrated with the parallel database system in this Para-DIAME offer many data-analysis capabilities that leverage the environment's parallelism and are able to churn out many of the statistics needed by E. S. Study almost instantaneously. Nevertheless, she encounters a problem: In her literature survey, she has come across a technique, nonlinear dimensionality reduction (NDR) that she feels can help her analysis and understanding of blizzards. But this technique has not yet been implemented in this para-DIAME and cannot be enacted by constructing a workflow using existing capabilities of the analytics suites.

She faces a choice: either she implements NDR with her limited programming skill on her local computer, downloads the data, and performs the analysis, or she solicits help from JESDIO to have a professional software engineer implement NDR to take the unique advantages provided by Para-DIAME's software stack and architecture. Since she has no confidence in her skills for implementing such a complicated algorithm, and the data volume involved is horrendous it becomes an easy decision for her to make.

She contacts the user-services department of JESDIO, explains what she needs, and requests help. User-services promptly assigns a software engineer who then contacts her and talks to her in length to elicit detailed requirements. Within a couple of weeks, and a few more iterations with the software engineer, she is informed by user-services that a custom NDR operator has been completed and added to one of the analytics suites for her and, of course, for all other users of Para-DIAME. Indeed, NDR proves to be useful and she is able to obtain a unique insight from the analysis of her blizzard events that has never before been reported in the literature.

Before such data-analysis environments became mainstream, scientists often-implemented data-analysis techniques themselves. This caused much duplication of effort: The same techniques were likely to have been implemented multiple times by different individuals or research groups. Moreover, since the great majority of these scientists were not trained software engineers, their implementations often were not taking full advantage of the hardware available, were not constructed based on the best software engineering practices, and were of suspect qualities. In contrast, trained software engineers working together with scientists in the new paradigm offer several advantages in the implementation of data-analysis algorithms and techniques: 1) full exploitation of the environments' capabilities, 2) better software quality, and 3) immediate reusability for other users. In addition, strict adherence to revision tracking/control as well as meticulous documentation practice for both data and algorithms, enforced by the management of such environments and carried out by professional information technologists, not only relieves scientists of burdens unrelated to scientific research but also ensures reproducibility, which is a hallmark of rigorous scientific research.

Significance of the Integrated Collaborative Infrastructure

Because data are uniformly "structured" in a para-DIAME and an extensive collection of sampling, averaging, convolution, and inter- and extrapolation techniques is available from the advanced analytics suites, "data fusion" becomes much easier. E. S. Study takes advantage of this and is able to analyze data from a number of sources. For example, she is able to bring together model data and observations of blizzards from a spectrum of models and various satellite instruments, such as visible and infrared radiometers, microwave radiometers, and precipitation radars, as well as ground-based observations from weather radar networks, and occasionally, in-situ or remotely sensed observations acquired during field campaigns. This enables her to analyze and understand the phenomenon much more holistically.

Probably the most amazing thing about this is that she does not have to perform many of the difficult tasks herself. She is often able to leverage the tools of other fellow scientists in the para-DIAME. In the past, collaboration had been difficult to carry out. Personal preferences in computing platforms, in programming languages, and in data-management practices all conspired against effective collaborations. In the uncommon occasions when software was shared for collaboration, the frequent negligence for ensuring software quality, the inconsistency in revision tracking and control, and the general lack of adequate documentation further exacerbated the ineffectiveness. These deficiencies in conducting effective collaboration seriously impeded scientific productivity because the study of Earth science, due to innumerable interconnected processes in the Earth system, called for the cooperation and collaboration of many disciplines.

As anticipated, E. S. Study finishes her dissertation and obtains her Ph.D. according to the original one-year estimate. Her dissertation is of such high quality and potential impact that she is promptly hired by one of the most prominent climate research institutes. Dr. Study has since become aware of several other researchers investigating various aspects of blizzards. In the same para-DIAME, they have seamlessly exchanged and deliberated their definitions of blizzards and winter storms in general. They have arrived at a classification scheme for this phenomenon that they deem to be the most reasonable, useful, and applicable, which is gaining recognition and becomes the de facto standard.

Dr. Study is now working with researchers of other Earth science phenomena and data-mining specialists. She believes that, in the interconnected systems of the Earth system, there must be some relations or correlations between the characteristics of blizzard and those of other phenomena. She is devising a strategy, in collaboration with these other researchers and data-mining specialists, to leverage the capability of para-DIAME and to uncover these potentially game-changing relations or correlations.

Epilogue

The idyllic research scenario described above is, unfortunately, not yet the reality experienced by any investigator within any institution. However, a number of promising projects are pioneering changes in the Earth science data-analysis landscape and bringing us closer to the day when scientists can routinely harness and direct the power of "Big-Data technologies" to distill compelling results from the incredible volumes of data that have been generated and collected by NASA and other agencies. One such project is the Automated Event Service (AES), which cradles the fledglings of the three components in a para-DIAME, i.e., 1) a next-generation array-based database management system designed to store, manage,

and manipulate multidimensional scientific data in parallel, 2) a core analytics suite leveraging the underlying parallelism of the database system, and 3) a collaborative portal to improve the effectiveness of scientific collaboration. A major objective of AES is to allow researchers to scan decades of data to identify all occurrences of various types of phenomena, ranging from the familiar (e.g., hurricanes, blizzards, heat waves) to the lesser known (e.g., Somali jets, tropopause folds), that are important for understanding our home planet, Earth.

Acknowledgement

We wish to thank the Advanced Information Systems Technology (AIST) program of the NASA Earth Science Technology Office (ESTO) for funding the development of AES. We also wish to express our gratitude to the NASA High End Computing (HEC) program for providing a development platform. We are in debt to Lara Clemence for her assistance in preparing this article.

References

[1] V. Mayer-Schönberger and K. Cukier, *Big Data: A revolution that will transform how we live, work, and think*. An Eamon Dolan Book, Houghton Mifflin Harcourt, Boston, 2013.

[2] P. McFedries, The coming data deluge, *IEEE Spectrum*, 2011. (http://spectrum.ieee.org/at-work/innovation/the-coming-data-deluge)

[3] http://www.ibmbigdatahub.com/infographic/four-vs-big-data