

Reference ID: 11224814053_Cleveland

Reference ID: 11224814053_Cleveland

Submission Date and Time: 12/16/2019 1:15:13 AM

This contribution was submitted to the National Science Foundation in response to a Request for Information, <https://www.nsf.gov/pubs/2020/nsf20015/nsf20015.jsp>. Consideration of this contribution in NSF's planning process and any NSF-provided public accessibility of this document does not constitute approval of the content by NSF or the US Government. The opinions and views expressed herein are those of the author(s) and do not necessarily reflect those of the NSF or the US Government. The content of this submission is protected by the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (<https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode>).

Consent Statement: "I hereby agree to give the National Science Foundation (NSF) the right to use this information for the purposes stated above and to display it on a publicly available website, consistent with the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (<https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode>)."

Consent answer: I consent to NSF's use and display of the submitted information.

Author Names & Affiliations

Submitting author: William Cleveland - Purdue University

Additional authors: Wen-wen Tung; Purdue University

Contact Email Address (for NSF use only): (hidden)

Research domain(s), discipline(s)/sub-discipline(s)

statistics; machine learning; parallel, distributed environments for data analysis; earth and atmospheric sciences;

Title of Response

Divide and Recombine for Deep Analysis of Big Data

Abstract

The analysis of data depends on a number of analytic factors that are important for success.. At the most fundamental level, it requires deep analysis, which means analysis of the data in detail at their finest granularity. This includes a powerful framework for deep visualization. Deep analysis greatly

reduces the risk of missing critical information in the data. True, analysis of summary statistics is important, but by itself not sufficient. Deep analysis for small and moderate size data has been achieved, but much work is yet required for big data. The problem is computational performance. What is needed is an analytic and computational approach that provides both deep analysis and high computational performance. Divide and Recombine (D&R) is a statistical approach designed to enable parallel computation. The analyst divides the data into subsets. Each analytic method is applied to the subsets in parallel with no communication among subset processes. The subset outputs are recombined in parallel, with communication among the processes. Much research is needed to optimize statistical accuracy and computational performance.

Question 1 (maximum 400 words) – Data-Intensive Research Question(s) and Challenge(s). Describe current or emerging data-intensive/data-driven S&E research challenge(s), providing context in terms of recent research activities and standing questions in the field. NSF is particularly interested in cross-disciplinary challenges that will drive requirements for cross-disciplinary and disciplinary-agnostic data-related CI.

Sometimes the goal of D&R is one result for all of the data, such as applying logistic regression. D&R statistics research seeks statistical division methods and statistical recombination methods that optimize the statistical accuracy of the D&R result. There is already a research community carrying this out under the name "D&R", as well as "divide and conquer" and "distributed statistics". This opens up a new research domain in the field of statistics. Growth of this domain can be expected. The general division method here is "statistical division". Another general division method is "subject-matter division". The data are divided by conditioning on variables important to the analysis. One example is an analysis of 10,621,808,809 queries to the Spamhaus blacklisting service to check IP addresses. For one division, each subset was the query and response fields for all queries to one IP address. There were 206,952,971 subsets. Much work needs to be done to find effective division methods for different subjects. However, we can expect many general principles to arise that apply to many different fields.

Question 2 (maximum 600 words) – Data-Oriented CI Needed to Address the Research Question(s) and Challenge(s). Considering the end-to-end scientific data-to-discovery (workflow) challenges, describe any limitations or absence of existing data-related CI capabilities and services, and/or specific technical and capacity advancements needed in data-related and other CI (e.g., advanced computing, data services, software infrastructure, applications, networking, cybersecurity) that must be addressed to accomplish the research question(s) and challenge(s) identified in Question 1. If possible, please also consider the required end-to-end structural, functional and performance characteristics for such CI services and capabilities. For instance, how can they respond to high levels of data heterogeneity, data integration and interoperability? To what degree can/should they be cross-disciplinary and domain-agnostic? What is required to promote ease of data discovery, publishing and access and delivery?

One implementation of D&R is the R-RHIPE-Hadoop software (RRH). The user programs analyses in R, the front end. Hadoop is the back end. RHIPE (R & Hadoop Integrated Programming Environment) provides communication between front and back, and R functions to aid in programming D&R. The hardware can be a cluster or a multicore machine. Hadoop writes subsets and outputs to the Hadoop Distributed File System. For a cluster, subsets and outputs are spread across all nodes. Hadoop Map does the parallel computation of the analytic method applied to subsets without process communication. Hadoop Reduce does the parallel recombination computation of the outputs, with process communication. What are the benefits of D&R with RRH? [It enables deep analysis. This includes a powerful framework for deep visualization.] [It can provide high computational performance when the computational complexity of analytic methods is high.] [The data can have a memory size that is big. In fact, the size can be larger than the physical memory of the hardware because when subsets and outputs are analyzed, they are put in memory sequentially, not all at once.] [It provides a programming of D&R that is very efficient for users, who are protected from having to manage details of parallel computing and database management.] While D&R with RRH is succeeding today, much work needs to be done. Other parallel distributed systems need to be investigated to see if they can, like Hadoop, cover all of the tasking of deep analysis, including visualization. RRH needs a more visible home than github, perhaps as an Apache product. On the computing side, there is a desperate need for a science of Computational Performance Measurement & Analysis (CPM&A) for software systems for data analysis. CPM&A today is often lacking in rigor for parallel distributed computational environments, and not sufficiently informative. Performance tests typically use a few low-level computations such as sort, which are not informative for data analyses. Rather, testing should be based in analytic methods, which are directly what an analyst uses and wants to run as fast as possible. Benchmark testing typically fails to control for factors that are important for comparing aspects of two systems. For example, Hadoop configuration parameters can have a big impact. Pilot experiments show there are strong interactions among factors. Little attention today is given to interactions.

Question 3 (maximum 300 words) – Other considerations. Please discuss any other relevant aspects, such as organization, processes, learning and workforce development, access and sustainability, that need to be addressed; or any other issues more generally that NSF should consider.

D&R now has a base for support, both statistical and computing. However a more formal organization is needed. There needs to be a consortium of people and organizations with mechanisms for easy communication such as slack, with D&R conferences, solely, or as part of statistics and computer science conferences.

-- End Submission --