A Dynamic Data Middleware System for Rapidly-growing Scientific Repositories

Tanu Malik1, Xiaodan Wang2, Philip Little3, Amitabh Chaudhary3, Ani Thakar4

1 Cyber Center, Purdue University
tmalik@cs.purdue.edu
2 Dept. of Computer Science, 4 Dept. of Physics and Astronomy,
Johns Hopkins University
xwang@cs.jhu.edu, thakar@pha.jhu.edu
3 Dept. of Computer Science and Engg.,
University of Notre Dame
achaudha, plittle@cse.nd.edu

Abstract. Modern scientific repositories are growing rapidly in size. Scientists are increasingly interested in viewing the latest data as part of query results. Current scientific middleware systems, however, assume repositories are static. Thus, they cannot answer scientific queries with the latest data. The queries, instead, are routed to the repository until data at the middleware system is refreshed. In data-intensive scientific disciplines, such as astronomy, indiscriminate query routing or data refreshing often results in runaway network costs. This severely affects the performance and scalability of the repositories and makes poor use of the middleware system. We present Delta a dynamic data middleware system for rapidly-growing scientific repositories. Delta’s key component is a decision framework that adaptively decouples data objects—choosing to keep some data object at the middleware, when they are heavily queried, and keeping some data objects at the repository, when they are heavily updated. Our algorithm profiles incoming workload to search for optimal data decoupling that reduces network costs. It leverages formal concepts from the network flow problem, and is robust to evolving scientific workloads. We evaluate the efficacy of Delta, through a prototype implementation, by running query traces collected from a real astronomy survey.

Keywords: dynamic data, middleware cache, network traffic, vertex cover, robust algorithms

1 Introduction

Data collection in science repositories is undergoing a transformation. This is remarkably seen in astronomy. Earlier surveys, such as the Sloan Digital Sky Survey (SDSS) [32, 38] collected data at an average rate of 5GB/day. The collected data was added to a database repository through an off-line process; the new repository was periodically released to users. However, recent surveys such as the Panoramic Survey Telescope & Rapid Response System (Pan-STARRS) [30] and the Large Synoptic Survey Telescope (LSST) [24] will add new data at an average rate considerably more