Bioinformatics, 33(4), 2017, 627–628
doi: 10.1093/bioinformatics/btw679
Advance Access Publication Date: 28 November 2016
Applications Note

Databases and ontologies

Big Data Smart Socket (BDSS): a system that abstracts data transfer habits from end users

Nicholas A. Watts[1] and Frank A. Feltus[2,*]

[1]Clemson Computing & Information Technology and [2]Clemson University Department of Genetics & Biochemistry, Clemson, SC 29634, USA

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on June 15, 2016; revised on August 30, 2016; editorial decision on October 24 2016; accepted on October 25, 2016

Abstract

Motivation: The ability to centralize and store data for long periods on an end user's computational resources is increasingly difficult for many scientific disciplines. For example, genomics data is increasingly large and distributed, and the data needs to be moved into workflow execution sites ranging from lab workstations to the cloud. However, the typical user is not always informed on emerging network technology or the most efficient methods to move and share data. Thus, the user defaults to using inefficient methods for transfer across the commercial internet.

Results: To accelerate large data transfer, we created a tool called the Big Data Smart Socket (BDSS) that abstracts data transfer methodology from the user. The user provides BDSS with a manifest of datasets stored in a remote storage repository. BDSS then queries a metadata repository for curated data transfer mechanisms and optimal path to move each of the files in the manifest to the site of workflow execution. BDSS functions as a standalone tool or can be directly integrated into a computational workflow such as provided by the Galaxy Project. To demonstrate applicability, we use BDSS within a biological context, although it is applicable to any scientific domain.

Availability and Implementation: BDSS is available under version 2 of the GNU General Public License at https://github.com/feltus/BDSS.

Contact: ffeltus@clemson.edu

1 Introduction

The increasing size of datasets used in scientific computing has made it difficult or impossible for a researcher to store all their data at the primary compute site. For biologists, the advent of Big Data has necessitated that the data transfer step is a key consideration in design of the high performance computing (HPC) workflow. Accordingly, scientific data repositories like the National Center for Biotechnology Information (NCBI) have begun to offer services such as dedicated data transfer nodes (DTNs) and advanced transfer clients, such as Aspera (asperasoft.com), while others use Globus (Allcock *et al.*, 2005). Despite these major advances in technology (reviewed in (Feltus *et al.*, 2015)), many researchers continue to use familiar but suboptimal data transfer practices using slow transfer clients such as the HTTP protocol via a web browser or secure shell copy (*scp*). Often, data transfers occur over improperly tuned networks.

The Big Data Smart Socket (BDSS) system, aims to alleviate this problem by shifting the burden of discovering alternative file mirrors, transfer clients, tuning parameters, etc., from the end user to a group of 'network data curators'. BDSS has the ability to take a data file manifest and look-up an alternate host from a curated database and select an optimal data transfer method for the source and destination computers. If no alternatives paths or methods are found, BDSS will default to the user request. BDSS allows a researcher, who may be unaware of available technologies, to perform faster transfer by asking for data in a familiar ways. In an effort to disseminate BDSS to biologists, we have integrated BDSS as a tool in Galaxy, a platform for execution of data intensive workflows

(Hillman-Jackson *et al.*, 2012). Additionally, BDSS will be tied to Tripal (Sanderson *et al.*, 2013) a toolkit for construction of online genomic and genetic databases as part of an NSF funded initiative. BDSS, Galaxy and Tripal will allow online genome databases to provide custom, large data workflows to their user-base with potentially improved data transfer rates between user, site and the computational facility.

2 Methods

2.1 Metadata repository

The metadata repository (MR) contains a curated map of alternate locations and data transfer mechanisms. It consists of a web based application for managing that information and a service for the client to access it. In our use case, MR data was curated after empirical data transfers and hardware tuning of endpoints relevant to the agricultural genomics community. Multiple MR instances for multiple communities are possible and MR databases could be merged. To support this, an MR's configuration can be exported to a file and imported into another MR.

The MR contains a set of data sources, which store information about the location of data files and the method for retrieving them from that location. For example, files from NCBI's SRA archive can be retrieved using either FTP or Aspera from the same URL. Each data source is configured with one or more URL matchers, which are patterns used to determine which source a requested data file comes from. For example, a URL matcher could be a regular expression or a specific combination of URL scheme and hostname.

Data sources are related to each other via URL transforms. These are generalized mappings between data file URLs at different data sources. For example, a data source may be related to a site mirroring its content by a transform that replaces the hostname of the original URL to get the URL of the same file at the mirror site.

2.2 BDSS Client

The BDSS client retrieves alternate URLs from the MR for an optimal data transfer. When a client requests alternate sources for a file, it sends the file URL to the MR. The MR determines if any known data sources match the URL using the data sources' URL matchers. If a source does match, the file URL is rewritten using the URL transform configured for the source. The transform output provides an alternate URL for the file the modified transfer mechanism.

- *Transfer mechanisms.* The BDSS client invokes other programs to transfer files. BDSS currently supports Aspera, scp, curl and GridFTP-Lite.
- *MR Configuration.* A BDSS client maintains a configuration for a default MR which may be overridden. Thus, end users can switch between multiple MRs.
- *Usage.* The BDSS client is invoked as follows: 'bdss transfer file-manifest'. Alternatively, the client can be used to list transfer mechanisms available on the machine and search data sources in the MR.

3 Results

As a test case, we compared the time to transfer 871 Mb of NCBI SRA data into a Galaxy server at Washington State University using the default Galaxy *Upload File* tool and BDSS. Over nine trials with a temporal spread, the Galaxy built-in tool transferred the data in 98.6s ± 6.9s, while the BDSS tool used Aspera and ran concurrent transfers to transfer the data in 47.0s ± 2.1s, for an average speedup of 2.1x (two tailed t test, $P = 1.6\text{E-}6$). Thus, integration of BDSS into a Galaxy workflow can result in improved transfer rates for users without any knowledge of how the data was transferred.

4 Discussion

BDSS is unique in that it is tied to a curated inventory of file data transfer nodes, networks and data transfer methods, and allows scientists to take advantage of advances in data transfer technologies even when they are unaware of them. Integration with Tripal and Galaxy will increase use of BDSS, thus increasing its impact. Our approach is to create a community of institutions whose researchers know where their data is stored, and to map the best practices for data transfer into the metadata repository. We are working with networking engineers from multiple academic institutions to tune network paths leading to Galaxy instances. Our pilot community is primarily agricultural genomics database users (Tripal) and we are coding the data transfer information between Tripal databases (Sanderson *et al.*, 2013) and Galaxy instances. The integration of this community will serve as a model for others.

Funding

This work was supported by NSF Award #1443040.

Conflict of Interest: none declared.

References

Allcock,W. *et al.* (2005) The Globus Striped GridFTP Framework and Server. In: *Proceedings of the 2005 ACM/IEEE SC05 Conference)*. Seattle, WA.

Feltus,F.A. *et al.* (2015) The widening gulf between genomics data generation and consumption: a practical guide to big data transfer technology. *Bioinf. Biol. Insights*, **9**, 9–19.

Hillman-Jackson,J. *et al.* (2012) Using Galaxy to perform large-scale interactive data analyses. *Curr. Protoc. Bioinf.*, Chapter 10:Unit10 15.

Sanderson,L.A. *et al.* (2013) Tripal v1.1: a standards-based toolkit for construction of online genetic and genomic databases. *Database (Oxford)*, **2013**, bat075.