IBM Cloud
Case Study

Life Sciences

GigaScience

Fast transfers improve research data accessibility for life sciences community

Overview

The need

Uploading and downloading large data sets from accompanying manuscripts to and from the *GigaScience* database, GigaDB.

The solution

Content submitters, reviewers, and research users use the IBM® Aspera® Connect plug-in to upload and download large data sets at maximum speed.

The benefit

Large data sets are uploaded in hours instead of days.

With high-speed transfers, *GigaScience* can review, accept, and publish manuscripts more quickly and return their decisions to the submitter within their target of two weeks.

GigaScience is an online open-access, open-data life sciences journal, co-published by BGI and BioMed Central, that publishes "big-data" articles covering the full spectrum of biological and biomedical sciences, including fields based on difficult-to-access data such as imaging studies, neuroscience, and systems biology. All of the manuscripts that are accepted and published in the journal focus on the use, analysis, or tool-development for large-scale data sets.

Perceiving a problem with the reproducibility of data-heavy scientific studies, *GigaScience* set out to provide a solution. With a goal of making research reproducible and reusable, research articles transparent, and large-scale data easily accessible and citable, *GigaScience* hosts the complete data sets associated with each published article in a comprehensive public database, GigaDB. It further provides each dataset with a "digital object identifier," which makes it easier for people to locate the files they are looking for and also provides the means for people to directly cite the data when reusing or reproducing research.

To handle the transfer of such enormous datasets, *GigaScience* has adopted a suite of IBM Aspera software products to provide authors, reviewers, and other users with the tools to upload and download all the large data sets that accompany manuscripts at maximum speed.

"We wanted a solution that would work for a broad range of people, not just those with technological expertise. Using Aspera, anyone can upload or download data easily," says Laurie Goodman, PhD Editor-in-Chief, GigaScience.

Solution components

Software

- IBM® Aspera® Connect Server
- IBM® Aspera® Connect browser plug-in
- IBM® Aspera® Console
- IBM® Aspera® Cargo

Moving TB size research data faster and further

The data sets submitted in support of articles published in the *GigaScience* journal can reach multiple terabytes in size. *GigaScience* found that FTP was not suitable for moving large files because transfers were often exceedingly slow, and if a user encountered a network problem, the transfer would have to be restarted from the beginning. Plus, transfers over long distances were particularly time-consuming and unreliable due to the high latency on the network.

On one occasion, *GigaScience* was presented with the challenge of uploading a 15 TB liver-cancer data set[1]. Ruling out FTP, *GigaScience* had to load the data onto 8 hard drives and physically transport it from the submitter to the journal, a costly and time-intensive process.

GigaScience, and researchers, cannot wait weeks to finish uploading large data sets attached to manuscripts – the journal typically likes to return reviews to authors within two weeks and publish immediately upon acceptance; and researchers need rapid download times to be most effective. *GigaScience* sought a transfer solution to provide a fast and efficient mechanism for authors, reviewers, and content consumers to upload and download large data sets to and from GigaDB while also providing ease of use for users with varying levels of technological expertise.

Fast sharing of data between authors, reviewers, and readers

After reviewing their options, *GigaScience* selected IBM® Aspera® Connect Server to rapidly transfer all the data sets that accompanied submitted manuscripts to the *GigaScience* database and IBM® Aspera® Console to manage and monitor the entire end-to-end transfer process. Authors use Aspera's free downloadable Aspera Connect plug-in to submit manuscript-associated data sets to a private data storage site at *GigaScience*. Staff reviewers then access the files, using the browser plug-in to download and upload files at high-speed. If a paper is accepted for publishing, the data is then transferred to the journal's public database, GigaDB, via Aspera, where it is readily available for journal readers to view and download, again using the Aspera Connect plug-in.

Better access to and reproduction of research and articles

In the past, it could take an entire week for a user to download just a small portion of data. With Aspera, uploads and downloads to GigaDB are fast, reliable, and simple, regardless of the data set size.

"People want to use this data; they don't want to sit and wait for a week while the data is downloading," said Laurie Goodman, Editor-in-Chief of *GigaScience*. "Aspera is the only solution currently out there that can meet the journal's needs to provide a reasonable way for people to access data in a timely manner."

> *"Aspera is the only solution currently out there that can provide a reasonable way for people to access data in a timely manner."*
>
> — Laurie Goodman, PhD Editor-in-Chief, *GigaScience*

With Aspera, *GigaScience* provides journal readers with a much faster, more reliable, and more pleasant download experience. *GigaScience* has received positive feedback from users, including one author who was very impressed when a 1.2 TB set of data was fully uploaded in a matter of hours rather than days, a sharp improvement over the previous results seen with FTP.

Plus, Aspera provides ease of use for users with all different levels of computational capability, accommodating computer-savvy individuals and non-technical users alike. "We wanted a solution that would work for a broad range of people, not just those with technological expertise," added Goodman. "Using Aspera, anyone can upload or download data easily."

Aspera's speed, reliability, and ease of use make it possible for *GigaScience* to realize its goal of improving the accessibility and reproducibility of research and articles for the life science community.

Other notable benefits include the following:

- **Fast transfers:** Using Aspera's Connect Server, uploads and downloads to GigaDB are accomplished at maximum speed, regardless of file size, transfer distance, or network conditions.
- **Ease of use:** With an intuitive web-based interface and the self-installing Aspera Connect plug-in, Aspera provides ease of use for every user of GigaDB, no matter the level of computational expertise.
- **Reliability:** With automatic resume and retry for partial or failed transfers, *GigaScience* and its users are confident their transfers will complete dependably.

About GigaScience

GigaScience (http://www.gigasciencejournal.com) aims to revolutionize data dissemination, organization, understanding, and use. An online open-access open-data journal, we publish “big-data” studies from the entire spectrum of life and biomedical sciences. To achieve our goals, the journal has a novel publication format: one that links standard manuscript publication with an extensive database, GigaDB, that hosts and provides a citable format for all associated data and downloadable data analysis tools, as well as Galaxy platform resources. *GigaScience* is based out of BGI, the world’s largest genomics institute, which carries out research relevant to human diseases, prenatal care, agriculture, and the environment. BGI co-publishes *GigaScience* with BioMed Central, the world’s largest open-access publisher.

GigaDB is available at http://gigadb.org. GigaGalaxy (http://galaxy.cbiit.cuhk.edu.hk) is a joint project with Chinese University of Hong Kong-and, BGI, at the CUHK-BGI Innovation Institute of Transomics. The Galaxy platform and GigaDB are supported by BGI and the China National Genebank.

About Aspera, an IBM Company

Aspera, an IBM company, is the creator of next-generation transport technologies that move the world’s data at maximum speed regardless of file size, transfer distance and network conditions. Based on its patented, Emmy® award-winning FASP® protocol, Aspera software fully utilizes existing infrastructures to deliver the fastest, most predictable file-transfer experience. Aspera’s core technology delivers unprecedented control over bandwidth, complete security and uncompromising reliability. Organizations across a variety of industries on six continents rely on Aspera software for the business-critical transport of their digital assets.

For more information

For more information on IBM Aspera solutions, please visit ibm.com/cloud-computing/products/high-speed-data-transfer/ and follow us on Twitter @asperasoft.

aspera
an IBM® company

IBM Corporation
Route 100
Somers, NY 10589

Produced in the United States of America
January 2017

This document is current as of the initial date of publication and may be changed by IBM at any time. Not all offerings are available in every country in which IBM operates.

The performance data and client examples cited are presented for illustrative purposes only. Actual performance results may vary depending on the specific configurations and operating conditions. It is the user's responsibility to evaluate and verify the operation of any other products or programs with IBM product and programs. THE INFORMATION IN THIS DOCUMENT IS PROVIDED “AS IS” WITHOUT ANY WARRANTY, EXPRESS OR IMPLIED, INCLUDING WITHOUT ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND ANY WARRANTY OR CONDITION OF NON-INFRINGEMENT. IBM products are warranted according to the terms and conditions of the agreements under which they are provided.

[1] Kan Z, et al. (2012) Hepatocellular carcinoma genomic data from the Asia Cancer Research Group *GigaScience*, http://dx.doi.org/10.5524/100034

♲ Please Recycle