NIH Case Study

Curing Advanced Data Ailments Using Data Virtualization to Aid Worldwide War on Cancer

National Institutes of Health

www.nih.gov

Industry

Biomedical & Healthcare Research

Profile

NIH, the nation's medical research agency, includes 27 Institutes and Centers and is a component of the U.S. Department of Health and Human Services. NIH is the primary federal agency conducting and supporting basic, clinical, and translational medical research, and is investigating the causes, treatments, and cures for both common and rare diseases.

The Need

As two of the 27 institutes that make up the National Institutes of Health (NIH) -- the nation's medical research agency -- The National Cancer Institute (NCI) and National Human Genome Research Institute (NHGRI), recently joined forces to execute on a project known as The Cancer Genome Atlas (TCGA). The TCGA mission is to catalog the genetic mutations responsible for cancer using genome sequencing and bioinformatics. To further cancer research by making this genome data available to a larger research community, TCGA joined the International Cancer Genome Consortium (ICGC), a collaboration of the world's leading cancer and genomic researchers. As part of this partnership, TCGA would make cancer genome data available to other members of ICGC.

The NIH faced significant obstacles in reliably and efficiently moving large volumes of cancer genome data from TCGA to ICGC. This process involved transforming the TCGA data to meet ICGC format requirements and then periodically uploading the data into ICGC servers. The format changes that needed to occur to align with ICGC standards included aggregation of fields, changing field names, and various representations of the data itself. For example, TCGA may represent gender as "M or F" while ICGC may represent gender as "1 or 2". The transformation was initially accomplished using PERL scripts, but NIH faced the following challenges with this process:

- **Not Scalable** – The NIH was unable to include all TCGA genome data that comprised of hundreds of millions of rows of data across more than 25 cancer variations.
- **High Costs** - Modifying and maintaining the scripts when source and target formats changed proved to be both expensive and time consuming.
- **Inaccurate** - Limited connectivity to data sources led to redundant copies of data, slower processes and greater chance of errors.

Taking these issues into consideration, the NIH quickly realized it needed an efficient and scalable solution to transform and move TCGA data to ICGC.

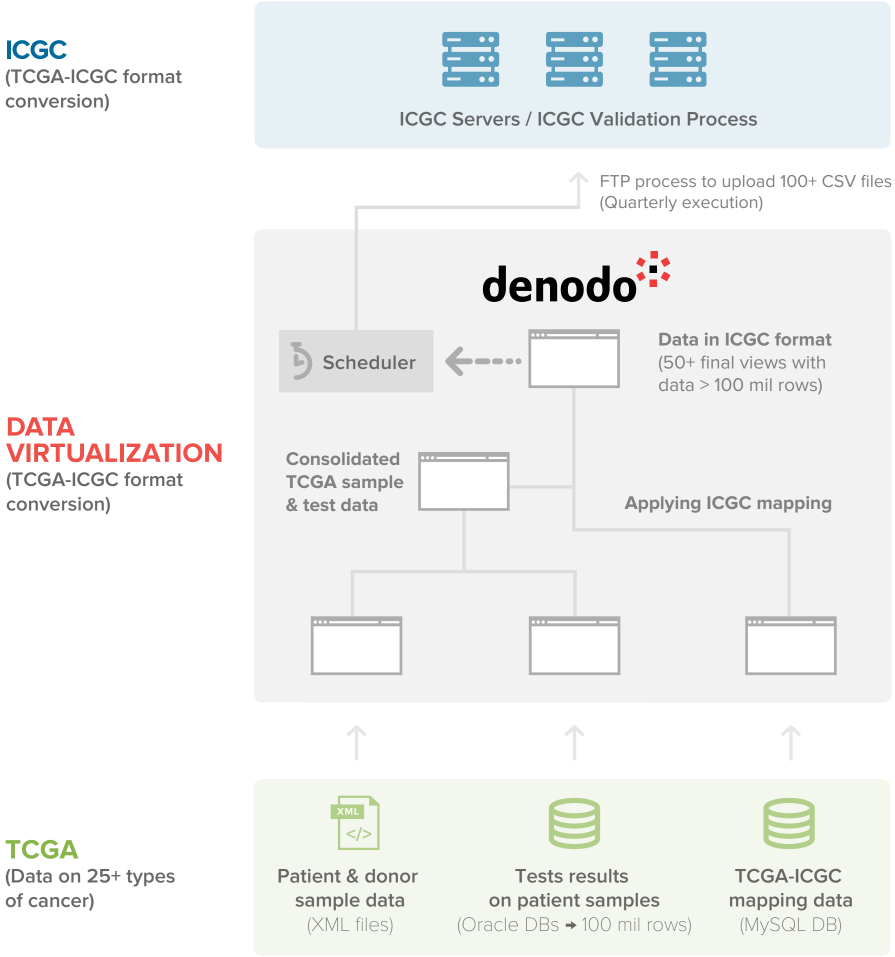
The Solution

The NIH ultimately decided that data virtualization would be the preferred method to transfer data between TCGA and ICGC systems as this technology allowed them to scale their existing process to larger data sets while maintaining efficiency and data accuracy.

The NIH used data virtualization to connect to the different sources of the genome data, apply transformations, produce the final data sets and periodically upload these data sets into the ICGC servers. The connectors within the data virtualization platform provided a normalized view of the patient and donor data stored in XML files, sample test results in Oracle and TCGA-ICGC mapping data in MySQL DB. The transformation process included three important steps: aggregating the patient and test data, converting this data into the ICGC format using the mapping information, and then creating the final output files in CSV format. Lastly, the scheduler within the data virtualization platform executed an FTP process once every quarter and then uploaded the files into the ICGC servers.

By using data virtualization, the NIH could directly connect to the different data sources and thus didn't have to replicate data in converting formats. The flexibility of the data virtualization layer enabled them to detect and propagate source changes to incorporate frequent variations in the TCGA and ICGC formats. By minimizing human intervention in these processes, they increased efficiency and data accuracy.

Data Virtualization Deployment at the NIH


ICGC
(TCGA-ICGC format conversion)
ICGC Servers / ICGC Validation Process
FTP process to upload 100+ CSV files
(Quarterly execution)
denodo
Scheduler
Data in ICGC format
(50+ final views with data > 100 mil rows)
DATA VIRTUALIZATION
(TCGA-ICGC format conversion)
Consolidated TCGA sample & test data
Applying ICGC mapping
TCGA
(Data on 25+ types of cancer)
XML
Patient & donor sample data
(XML files)
Tests results on patient samples
(Oracle DBs → 100 mil rows)
TCGA-ICGC mapping data
(MySQL DB)

Finally, the NIH was able to create a generic workflow to transform TCGA data into ICGC for one particular type of cancer and replicate this workflow to 24 other cancer types. This allowed NIH to scale their conversion process to include large cancer datasets without incurring high development and maintenance costs.

Benefits

By deploying data virtualization, the NIH was able to realize the following benefits:

- **Increased scalability:** Include larger genome data sets due to the creation of replicable generic workflows and the platform's advanced performance capabilities.
- **Increased efficiency:** Faster development and modification of TCGA – ICGC transformation processes because of the platform's diverse connectivity and publishing capabilities.
- **Increased accuracy:** Minimized replication and manual intervention led to the most current versions of data and processes being used to create the output files, leading to greater accuracy in the final data.

The success of this project led the NIH to expand the data virtualization deployment to similar projects, such as TARGET (Therapeutically Applicable Research to Generate Effective Treatments), which involved extracting pediatric cancer data from multiple XML files and then transforming them to the ICGC format. By enabling the transfer of greater volumes of cancer genome data between TCGA and ICGC, data virtualization technology is helping disseminate valuable information to cancer researchers around the globe and aiding in their efforts to cure this worldwide disease.

About Denodo

Denodo is a leader in data management. The award-winning Denodo Platform is the leading data integration, management, and delivery platform using a logical approach to enable self-service BI, data science, hybrid/multi-cloud data integration, and enterprise data services. Realizing more than 400% ROI and millions of dollars in benefits, Denodo's customers across large enterprises and mid-market companies in 30+ industries have received payback in less than 6 months.