cirata | BIO EXPRESS | kobic

Kobic achieves 13x faster data transfer between heterogeneous file systems to improve performance of covid-19 research

Objective

The Korean Bioinformation Center (KOBIC) collects and manages bio-research resource information and genomic information. In 2020 they began to collect COVID-19 genome data produced worldwide. KOBIC also operates Bio-Express, a large-scale Genomic Data Analysis Cloud service, which they provide for free to bio-engineering researchers at hospitals, businesses, universities and research institutes in Korea.

Challenge

Bio-Express has high-performance data analysis requirements, but the time needed to replicate the large volume of data between their Hadoop Distributed File System (HDFS) and their Linux/Unix based Lustre file system was lengthy and impacts Bio-Express response times. KOBIC needed a way to reduce the time required to replicate data and improve the overall system performance.

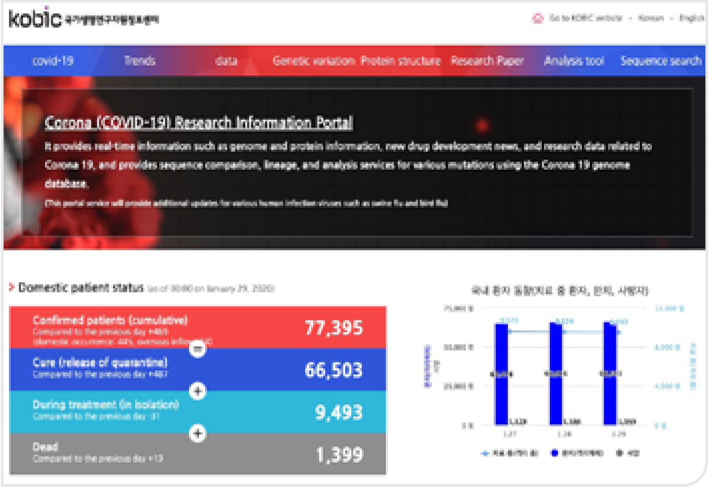
Solution

In early 2020 Cirata announced free access to their suite of cloud migration and big data tools, for teams involved in developing potential treatments and cures for the COVID-19 pandemic. Cirata provided their software along with technical resources free of charge to KOBIC to assist the organization in enhancing its architecture, developing products, and introducing Cirata's automated replication technology into KOBIC's workflow.

Results

Leveraging Cirata the Bio-Express next generation service was able to:

- Replicate files 13 times faster
- Shorten average analysis time of Bio-Express services by greater than 30%
- Provide users with faster response times and the ability to perform their research with greater efficiency



Corona (COVID-19) Research Information Portal

It provides real-time information such as genome and protein information, new drug development news, and research data related to Corona 19, and provides sequence comparison, lineage, and analysis services for various mutations using the Corona 19 genome database.

(This portal service will provide additional updates for various human infection viruses such as swine flu and bird flu)

Domestic patient status (as of 00:00 on January 29, 2020)

Confirmed patients (cumulative) Compared to the previous day +469 (domestic occurrence: 445, overseas inflow: 24)	77,395
Cure (release of quarantine) Compared to the previous day +487	66,503
During treatment (in isolation) Compared to the previous day -31	9,493
Dead Compared to the previous day +13	1,399

Company overview

The Korean Bioinformation Center (KOBIC) is the Korean national research center in bioinformatics, based in Daejeon, South Korea. KOBIC manages biological data from a number of different sources, with an emphasis on omics data. Research at KOBIC has an emphasis on next-generation sequencing methods, systems bioinformatics, biomedical informatics and structural informatics. In addition, they operate Bio-Express, a large-scale Genomic Data Analysis Cloud service, which is provided for free to bio-engineering researchers at hospitals, businesses, universities and research institutes in Korea.

In March 2020 KOBIC opened the COVID-19 research information portal (kobic.re.kr/covid19) to provide researchers with information by collecting COVID-19 related genomes and proteomic data scattered around the world for COVID-19 research.

Initial Bio-Express deployment

A common challenge faced by organizations around the world when using Hadoop is how to leverage the system in an optimal fashion across the Linux/Unix-based servers and big data analytics programs. Using the analytics programs together in a single integrated architecture requires a lot of data migration to the Hadoop distributed file system (HDFS) and storage.

KOBIC's hybrid data pipeline architecture faced similar challenges. Given the large volume of genomic data and rapidly increasing number of Bio-Express users, KOBIC encountered data processing performance issues. Response times for processing the genomic data, which utilized semi-structured and unstructured data, continued to increase. The data needed to be replicated bidirectionally between their HDFS cluster used for their big data analytics programs, and the Linux-based Genomic analysis cluster called Lustre. The larger the dataset, the longer the data transmission time. For KOBIC the time taken to replicate data between clusters began to significantly impact the overall computation time. It was found that Bio-Express consumed more than 40% of the processing time due to an average of about 20TB of data replication per day.

> "KOBIC uses Cirata to automate file transfer 13 times faster in both directions between Hadoop-based Big Data Analysis Program Execution Cluster (HDFS) and Linux-based Genomic Analysis Program Execution Cluster (Lustre). We were able to reduce the overall average time to analyze user genomic data of Bio-Express service by more than 30%."
>
> *Kun-Hwan Ko, Researcher at Kobic's Computational Development Team*

Summary of Bio-Express service challenges

- Bidirectional replication of 70 genomic datasets with an average of 20TB between heterogeneous file systems per day
- Data Analysis Flow: When a service user performs a large-scale analysis using genomic data stored in Hadoop, it is replicated to Lustre for data quality management analysis with a Linux/Unix-based analysis program and the results are computed. The results are then replicated back to the Hadoop distributed file system and further analyzed through Spark.
- As a result of requirements to support heterogeneous file system analysis programs, continuous data movement occurs frequently in both directions.
- DistCp, which is provided for free by Hadoop itself, is difficult for this organization to apply to replication between clusters because they analyze sensitive genomic data, which is erased (EraseCoding), and KOBIC's own scripts also use the get/put command provided by Hadoop, resulting in operational overhead on expensive IT resources, data inconsistencies, and inefficiencies of IT manpower.

Cirata contribution

To help with research needed to overcome COVID-19, in the first quarter of 2020 Cirata announced free access to their suite of cloud migration and big data tools, for teams involved in developing potential treatments and cures for the COVID-19 pandemic.

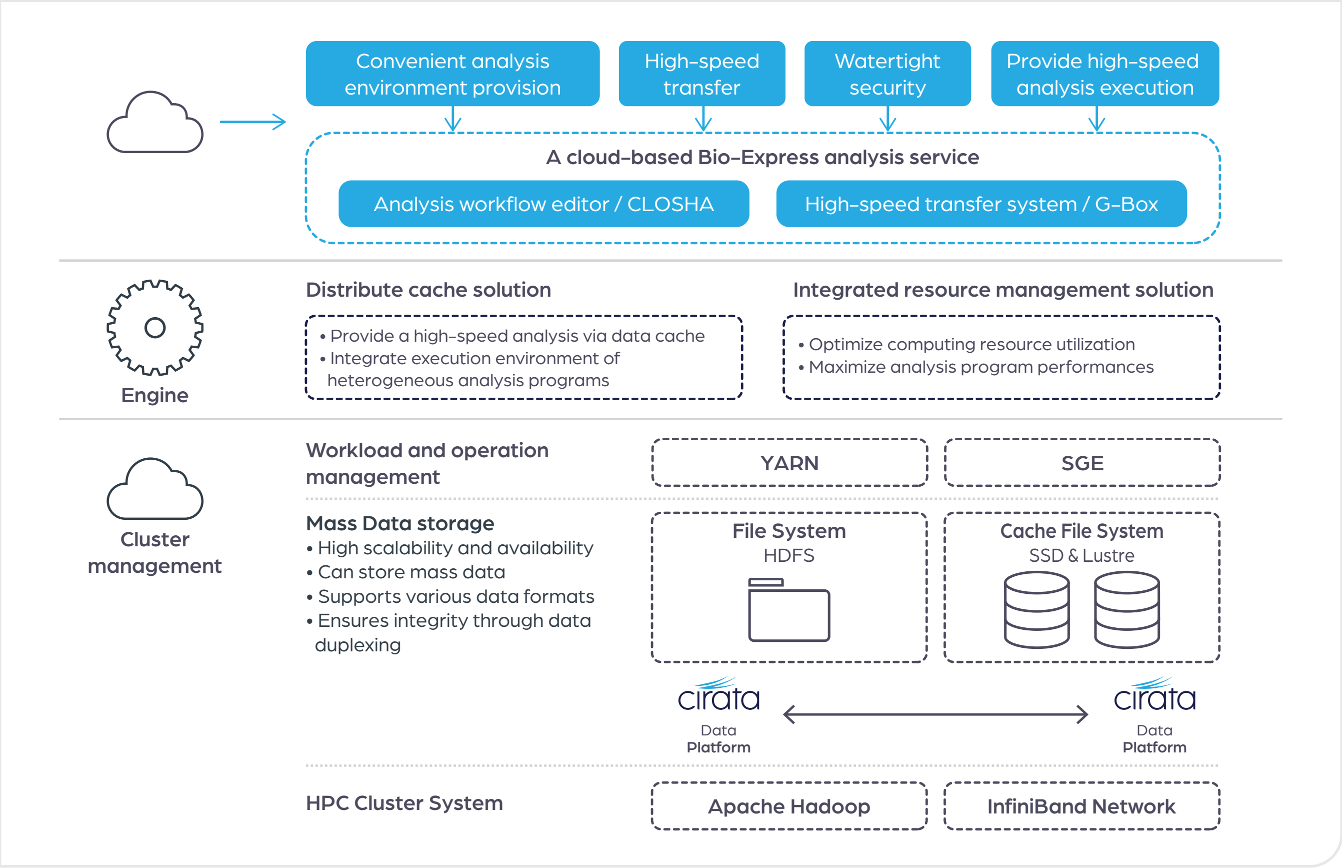
In June 2020 KOBIC and Cirata reached a cooperative agreement to support researchers operating the Bio-Express platform as well as the platform's external users. Since that time, Cirata has provided KOBIC with a permanent license to Cirata's solution to support active-active replication of live data. Cirata also provided technical support personnel to assist KOBIC in enhancing its architecture, developing products, and applying automated replication into the workflow. Due to the Coronavirus the entire process was carried out through remote conferences and remote support.

Next generation Bio-Express service

As both organizations began to understand each other's business and products they worked as a team to establish an optimal architecture improvement plan. After deploying the changes, they validated it with the previous version, and in December 2020 verified the improvements in speed of replication and automation of the process. With Cirata the new solution was able to replicate files 13 times faster and shorten the overall average analysis time of Bio-Express services by more than 30%. As a result, KOBIC is ready to apply the next generation Bio-Express service and plan to have it widely available in February 2021.

About KOBIC

The Korea Bioinformation Center (KOBIC) is a national bio-resource information center for general management of domestic bioresource information and research in the field of bioinformation. KOBIC helps domestic research institutes, hospitals, companies, and universities to research genomic data and COVID-19 for free. One of KOBIC's main missions is to develop and operate a system that can analyze large-scale genomic data using the state-of-theart information technology.


Convenient analysis environment provision
High-speed transfer
Watertight security
Provide high-speed analysis execution
A cloud-based Bio-Express analysis service
Analysis workflow editor / CLOSHA
High-speed transfer system / G-Box
Engine
Distribute cache solution
• Provide a high-speed analysis via data cache
• Integrate execution environment of heterogeneous analysis programs
Integrated resource management solution
• Optimize computing resource utilization
• Maximize analysis program performances
Cluster management
Workload and operation management
YARN
SGE
Mass Data storage
• High scalability and availability
• Can store mass data
• Supports various data formats
• Ensures integrity through data duplexing
File System
HDFS
Cache File System
SSD & Lustre
cirata Data Platform
cirata Data Platform
HPC Cluster System
Apache Hadoop
InfiniBand Network

5000 Executive Parkway, Suite 270, San Ramon, CA 94583
US +1 877 (926-3472) EMEA +44 (0) 114 3039985
APAC +61 2 8211 0620 All other +1 925 380 1728
www.cirata.com