FACILITATING FREE AND OPEN ACCESS TO BIODIVERSITY DATA IN REAL-TIME

10K

Records updated per second

Impact

- Provides consistent view of critical data that was previously dispersed and fragmented
- Data platform supporting near real time ingestion with filtered search and analysis across 1.4B records
- Saved the effort of about one full-time employee, or 20% of team's capacity

GBIF is an international network and research infrastructure headquartered in Copenhagen that provides free and open access to biodiversity data for use in scientific research and policy. Nearly 1,600 institutions from more than 130 countries worldwide share data through GBIF.org, making it one of the world's largest sources of information about where and when all organisms—plants, animals, fungi, and microbes—have been observed or collected over the past four centuries.

GBIF provides institutions holding data about the natural world with common standards and open-source tools that enable them to share information about where and when species occur on Earth. This knowledge derives from myriad sources, including everything from centuries-old museum specimens to DNA sequence-based samples to geotagged smartphone photos snapped by amateur naturalists just last week.

To fulfill its mission of giving anyone anywhere in the world free and open access to biodiversity data via the Internet and support large-scale knowledge generation and data analysis, the GBIF Secretariat in Copenhagen relies on the flexibility of the Cloudera tools and systems to deploy its data lake.

Sharing is caring

Like other sciences, biology has become a computational area of research. While the traditional image of scientists conducting field expeditions to collect the information underpinning their insights still holds some truth, biological researchers around the world today fundamentally rely on data—their own and others'—to carry out their research. But without a collective framework for storing and sharing it, much of the data about life on earth would be used once and then sit discarded and forgotten on disconnected computers without contributing to wider knowledge.

"The value of biodiversity data is not exhausted by use—it can inform later research that performs deeper analyses or re-investigates aspects of the same topic," stated Oliver Meyn, a Hadoop and big data consultant who is a former senior developer and current "Biodiversity Open Data Ambassador" for GBIF. "Our aim is to eliminate the silos that prevent data from being shared and champion global access to data, which in turn leads to better research results."

GBIF's global infrastructure required a platform that would make it easier for scientists, researchers, and institutions both to share and to access data. Early versions of the infrastructure were built around MySQL which quickly proved inefficient for dealing with large amounts of data. The facility needed a platform to handle demand leading to real-time data collection and classification. Introduced initially to keep heavy processing load from MySQL, the Hadoop ecosystem now provides the core platform for GBIF to integrate, process, index and analyse all incoming data.

"Before deploying a Hadoop-based solution, users were limited to accessing a few hundred thousand records in data exports. Today users can filter across 1.4 billion records in a few minutes with unrestricted data exports. This database could simply not support such extensive use without the help of Cloudera."

Tim Robertson, Informatics team lead, GBIF

Creating a global index of the living world

GBIF's 100 formal members and 1,600 data publishers form a distributed network of both experts and infrastructure. When a data publisher creates or updates a dataset in a GBIF-connected repository, the crawling infrastructure brings these changes into the data lake. Newly arrived data goes through a series of formatting, quality control and enrichment and made available to the data analysts through Hive and to the public through search indexes. With MySQL, the team was having to stop the crawlers from performing their task as the sheer volume of data was crushing the database.

With its small informatics team and limited resources, the GBIF Secretariat turned to Cloudera to set the foundation of a modern data architecture. "Switching from MySQL to Hadoop provided us the scalability and flexibility we needed to process, analyse and distribute the volumes of data we see," said Tim Robertson, who leads the GBIF informatics team. "Perhaps most importantly, though, Cloudera enables us to maximize the resources of a small team. The Cloudera distribution provides us compatible versions of the products (HBase, Solr, Hive etc), easy deployment and updating along with monitoring, alerting and diagnostic tools. In fact, we were able to save the effort of about one full-time employee, or 20% of our capacity. This means we are able to focus our effort on building software specific to biodiversity data and spend far less time on the internal plumbing of the platform."

After implementing the enterprise data platform, GBIF was able to support the weight of its volumes, regularly updating indexes at 10,000 records per second.

Results

Through Cloudera, GBIF was able to maximize the resources of its small team. In doing so, it enabled data sharing and access and significantly improved operational efficiencies, freeing up time to focus on other challenges. The open-sourced platform is critically important in balancing out the team's resources for developing, managing and supporting GBIF's open data repository. Without the community and the ability to look at the source code, the team would have been unable to achieve results at this scale and speed.

"Before deploying a Hadoop-based solution, users were limited to accessing a few hundred thousand records in data exports. Today users can filter across 1.4 billion records in a few minutes with unrestricted data exports," said Robertson. "Our system today enables scientists to use that data to build models about changes to life on earth. Every day, researchers publish two peer-reviewed papers that make use of data mediated through the GBIF network. This database could simply not support such extensive use without the help of Cloudera."

GBIF achieves real-time data analysis, recovery and index updates, all managed within the enterprise data platform. Whether a scientist would like to know how climate change or invasive species will affect patterns of life on earth or alter the benefits we depend on from natural systems, GBIF provides a framework for making the latest data readily available and easily accessible.

Cloudera, Inc. 395 Page Mill Road Palo Alto, CA 94306 USA cloudera.com