Case Study

Stackable
your data, your platform

Data about all the different kinds of life

Global Biodiversity Information Facility builds upon Open Source with Stackable

GBIF Global Biodiversity Information Facility

Initial Situation and Challenges

GBIF, the Global Biodiversity Information Facility, is an international network that aggregates and provides **open access to biodiversity data from various global sources**, supporting scientific research, and conservation efforts. Scientists, conservationists, policymakers, educators and citizen scientists use it for studies, conservation strategies, policy-making and education. The private sector also uses GBIF data for research, development and environmental impact assessments.

Over time, GBIF faced significant issues with its ageing big data infrastructure, which was running an outdated version of Cloudera. This setup limited scalability and flexibility. Some parts of the existing codebase relied on deprecated components and unsupported Java versions, which made them difficult to maintain.

A key challenge was to migrate this environment to a modern, robust platform without disruption.

> *GBIF's transition to the Stackable Data Platform has been transformative; a much-needed modernization of our infrastructure. It has enabled us to seamlessly migrate live services, standardize diverse data, while continuing our commitment to following open data and open-source principles. Stackable's modular architecture allows us to efficiently integrate biodiversity data from across the globe and provides us the flexibility to adapt as technologies evolve in the future.*
>
> Tim Robertson, Head of Informatics at GBIF

GBIF Global Biodiversity Information Facility

> *Tim and I have shared the same passion for open-source software for many years. I am very pleased that he and his team have decided in favour of the open software stack from Stackable for GBIF.*
>
> Lars Francke, CTO and Co-Founder at Stackable

Stackable

Solution

In response to these challenges, GBIF initiated a **comprehensive migration to the Stackable Data Platform** as part of a broader data center reallocation. This involved provisioning new hardware and deploying the Stackable environment to establish a robust foundation for the new system. Before dismantling and reformatting the old infrastructure, approximately **two petabytes** of replicated data were synchronised into this environment. This data was stored in various formats, including Apache Avro, Parquet, HFiles, Hive and Iceberg. The older nodes were then reintegrated into the Stackable-based setup to ensure continuity and the **cost-effective reuse of existing hardware**.

Key tools and technologies adopted during this migration included upgrading the codebase from Apache Spark™ 2 to Spark 3 and transitioning from Apache Oozie to Apache Airflow for workflow orchestration. For ad hoc analytics tasks, GBIF replaced Hue with a more advanced Trino and Apache Superset setup. The platform managed all data through the Hadoop Distributed File System (HDFS).

Result and Successes

The migration process began with a proof of concept, which involved designing and validating configurations. This was followed by **extensive performance tests** to evaluate workload compatibility. A shadow environment was tested in parallel with **mirrored production traffic** to ensure reliability. Once verified, it seamlessly replaced the previous user test environment before finally moving to production.

Migrating to the Stackable Data Platform has significantly enhanced GBIF's data infrastructure, which is now based on the **latest software versions** that are also systematically checked for vulnerabilities. The new setup offers greater availability, improved scalability and flexibility, enabling GBIF to manage biodiversity data more securely and efficiently. By optimising infrastructure and reducing operational complexity, GBIF has adopted a more **adaptable and future-proof architecture**.

The result is a **stable, maintainable, and modular platform** that lays the foundation for future growth and innovation.

Highlights at a glance

Flexible, scalable platform using open-source technologies like Apache Spark and Airflow

Transition to Trino and Apache Superset for advanced analytics

Migrated two petabytes of data to Stackable Data Platform on HDFS

Reallocated data center, reusing old hardware for cost-effective operations

New setup with current software ensures greater platform resilience

Modern platform for future growth and innovation in biodiversity data management

> *As a DevOps Engineer at GBIF, transitioning to the open-source Stackable platform has been liberating. Compared to Cloudera, Stackable offers unparalleled flexibility and consistency, significantly facilitating our daily operations.*
>
> Aleksander Nilsson, DevOps Engineer at GBIF

GBIF Global Biodiversity Information Facility

About GBIF

Global Biodiversity Information Facility

GBIF

The Global Biodiversity Information Facility is an international network and data infrastructure funded by governments worldwide aimed at providing anyone, anywhere, open access to data about all types of life on Earth.
Coordinated through its Secretariat in Copenhagen, the GBIF network of participating countries and organizations, working through the participant nodes, provides data-holding institutions around the world with common standards, best practices and open-source tools enabling them to share information about where and when species have been recorded. This knowledge derives from many different kinds of sources, including everything from museum specimens collected in the 18th and 19th centuries to DNA barcodes and smartphone photos recorded in recent days and weeks.

More at www.gbif.org

About Stackable

Stackable

With its open source data platform, Stackable stands for professional data sovereignty in the corporate context. The company attaches particular importance to data security, openness and transparency and offers first-class support at fair conditions.

The company, based in Wedel, Germany, was founded in 2020 out of the open source community. Stackable relies on open source code to support companies in dealing with big data. As an innovative and internationally active provider, Stackable puts the community at the center, promotes cooperation instead of competition and supports companies in the development of their data architecture.

More at www.stackable.tech