Genomics, Drug Development, Clinical Trials and Next Generation R&D Platforms

Advanced use of big data in biopharma

The biopharmaceutical (biopharma) industry is unquestionably one of the most important, with huge potential for treating some of the most intractable medical conditions facing the human race, such as cancer and autoimmune disease. But bringing to market life-saving and life-changing drugs based on biological sources is not without its challenges.

Safety is of paramount importance, so smooth and effective clinical trials are an essential part of the process; Research and Development (R&D) is dependent on consistent, reliable and integrated data sets, while genomics — the area within genetics that focuses on the sequencing and analysis of an organism's genome — is highly dependent on the intelligent use of data and modeling.

One of Cloudera's biopharma clients — one in the sector's global top ten — embarked on a challenging project around the use of genomic data, pledging to develop no new drugs without a genetic and genomic component and to launch a next-generation R&D platform. But the genomic data it held was found in siloes all over the organization, making genome analysis a true challenge.

Key Highlights

- Deep industry understanding and powerful analytics technology combined to drive innovation and decision-making in biopharma
- One platform to integrate, simplify and unlock all R&D data, enabling safer and quicker drug production
- Prebuilt architecture to address a variety of big data challenges within biopharma

Background

Genomics first emerged in the 1980s, and it rapidly became one of the scientific disciplines to most benefit from large-scale data analysis techniques. It is used to pick up and learn from genetic patterns and is an essential part of drug development to prevent disease. As data analytics and machine learning technologies have developed, the use of genomic data has further opened-up for medical clinicians and researchers, with information gleaned much more efficiently than previously.

This project saw Cloudera and the biopharma firm agree a deal with UK Biobank to access for 10 years, the genomic data for 500,000 patients. The aim was to:

- Capture all siloed data into one data lake
- Put genomic data analytics at the heart of a new R&D platform
- Undertake no new drug development without a genomic component

Solution

The Cloudera solution comprised three broad phases — data integration, advanced analytics and machine learning. To achieve the benefits from genomic data, which is a hugely complex and large data set, data cannot be stored in siloes across the organization and must be integrated. The biopharma firm used Cloudera Enterprise to collate all the relevant data into one data lake, forming a single integrated data platform. They standardized all the thousands of clinical trials into the CDISC standard format, using the OMOP Common Data Model which allows for the analysis of disparate observational databases.

This ensured consistency, while the utilization of Trifacta offered the flexible data manipulation required by different users to visualize, analyze, query and make informed decisions. The firm also deployed Hail on Cloudera, the open-source genomic processing tool that enables the rapid extraction of insights from massive amounts of data. This was used to enable quality control and genome-wide association analysis of thousands of phenotypes across millions of variants.

This enabled them to drive new biomarkers that indicate the presence and severity of a disease and to collect all the bioassays together. This gave them much richer data that meant they could very quickly gain new targets and discoveries, adding value to R&D almost immediately.

For the machine learning element of this project, the firm used the Cloudera Data Science Workbench. This made the data science much more accessible and is a tool that can be used by researchers and clinical assistants, as well as data scientists, working in a customized environment that works just like a laptop.

Results

The next-generation R&D platform was able to work with and extract deep insight from astonishing volumes of data. It supported 2,100 structured data sources, 500K tablets, 1350 unstructured data sources, 1.3B files, 1,800 data nodes and 50PB of data overall.

The creation of this holistic data lake and data science environment, capable of working on premise and in cloud clusters, with enterprise-grade security and regulatory compliance, coupled with Trifacta for seamless data exploration and data wrangling, makes for a compelling proposition in biopharma.

It's a fast and cost-efficient approach and framework for supporting the downstream whole genome pipeline process and this project is one of the most advanced of its type in the entire industry, in terms of identifying new targets and new drug candidates. The next-generation R&D platform enables firms to bring biopharma products to market much quicker and with complete confidence in safety.

The Value in a Partnership

Harnessing the insight of big data via the use of analytics, is for many biopharma enterprises becoming a critical priority. This is used with genomics, but also with proteomics, metagenomics and epigenomics too, and is pivotal in drug development and R&D. The partnership between Atos and Cloudera, with their combined expertise across the health ecosystem and products in the big data technologies that enable the unlocking of such insight, allows biopharma firms to achieve their goals.

Atos and Cloudera can replicate the offering used with this biopharma company, combining solutions and services to offer a powerful proposition for others in the industry too. From data ingestion to data management and the data lake layer, the joint team is building an architecture and solution that can be demoed to interested biopharma firms that shows the real potential of the partnership.

Working in a hybrid multi-cloud environment, Atos and Cloudera can deliver quickly and flexibly to ensure biopharma firms can:

- Unlock and harness the insight within genomics via data analytics and machine learning
- Use that insight to address specific challenges within the industry, such as reducing the time to market and improving the drug development process
- Manage and collate data across a variety of sources and platforms to create tangible business value

Atos is certified on Cloudera Data Hub and soon to be certified on Cloudera. To find out more about Atos and Cloudera, visit our partnership page or email Daniel Nutburn dnutburn@cloudera.com.

CLOUDERA

Cloudera, Inc. | 5470 Great America Pkwy, Santa Clara, CA 95054 USA | cloudera.com

Cloudera is the only true hybrid platform for data, analytics, and AI. With 100x more data under management than other cloud-only vendors, Cloudera empowers global enterprises to transform data of all types, on any public or private cloud, into valuable, trusted insights. Our open data lakehouse delivers scalable and secure data management with portable cloud-native analytics, enabling customers to bring GenAI models to their data while maintaining privacy and ensuring responsible, reliable AI deployments. The world's largest brands in financial services, insurance, media, manufacturing, and government rely on Cloudera to be able to use their data to solve the impossible—today and in the future.

To learn more, visit Cloudera.com and follow us on LinkedIn and X.