INCREASING THE SPEED AND QUALITY OF THE DRUG DISCOVERY PIPELINE

Pharmaceutical organizations place high importance on the speed and quality of the drug discovery pipeline. It can take from six to 12 years and nearly $3 billion dollars to conduct all the steps necessary—from research and testing to clinical trials and regulatory approvals—to bring a new drug or vaccine to market. Once a new product goes to market, pharmaceutical companies have a small window of opportunity to recoup development costs before their patent expires. Adding to the challenge, the cost to produce drugs has remained static in recent years, leading to a considerable reduction in profitability.

This global pharmaceutical company is on the forefront of research and wants to accelerate safe medicine delivery to market. To deliver on these ambitious goals and maintain leadership in the industry, the company needed to rethink its data architecture and strategy.

IMPACT

- $22.8 million net present value over 3 years; 10 months to break-even with CDP Private Cloud investment
- Reduced time for genome-wide association studies analysis from 80 years to a few weeks
- 97% of data is visible to users within R&D
- Improved success rate for drug development process through predictive analytics

Challenges

The key goal of their envisioned data platform was to enable strategic business value by unifying their distributed and siloed data sets, such as clinical, lab, and production data, across different legacy systems. The platform would also address the lack of self-service access to data for R&D departments and scientists, reduce costs involved with existing processes, and meet quality and compliance requirements. They determined that adopting a hybrid cloud architecture would be most effective.

This undertaking came with some new and intensified data and technical challenges including:

Noisy neighbor problem - So many analysts working on shared data sets means an unpredictable nature of workloads - spiky, with times of huge contention. Erroneous data leads to huge pipeline reruns. Compute was shared across multiple tenants that even spanned different clusters in some instances. This caused a major delay of 4 weeks in analysis of scientific hypotheses for drug discovery and added to the operational costs

Triple the ETL workloads - For large organizations it can sometimes take two to three months to turn around the procurement and provisioning of hardware, which makes it difficult to expand compute resources in time for seasonal workloads.

Run anywhere - the company wanted the flexibility of running the workloads where the data was and in a way that optimizes resource utilization - i.e. in a cloud native way, be it on premises or in the public cloud.

Maximize investment - It was impossible to maximize their capital expenditure on hardware assets as resources were severely underutilized. In one example, Impala ran on 150 nodes with 30% static reserved memory irrespective of actual usage. This rendered these resources idle for large amounts of time. To add insult to injury, key Spark jobs were left waiting for resources to complete other tasks. Horizontal scaling wasn't possible with their legacy architecture.

Longer SLAs - It took three to four months to perform an upgrade, including applying patches and running end-to-end testing, in order to minimize impact on downstream applications. As a result the platform team needed to institute longer SLAs with business users.

23M

Nearly a $23M net present value over 3 years

Double the data - They wanted to double the number of daily workload, the amount of researchers accessing data ad hoc from 1,500 to 3,000, and also anticipated doubling their data quantity from 25 PB to 50 PB within one year.

Solutions

After careful evaluation, this global pharmaceutical company turned to a hybrid approach led by Cloudera's CDP Private Cloud, partner technologies, and homegrown tools to deliver a holistic view of all data within R&D and give researchers an immense analytic advantage. The platform serves up information in the form of assets, where data is ingested from a variety of sources and then curated, housed, stored, cleansed, governed, and optimized based on popularity.

The platform combines all data from across the organization. As a result, researchers can combine and analyze data, ad-hoc, regardless of when, how, and where it was generated. With its new platform, researchers are using these pharma analytics to gain insights that help streamline every aspect of the R&D process. For example, it previously could take several months to identify the optimal mix of participants, assemble and analyze data from across multiple clinical trials. Now, with the clinical trial data standardized and analytics-ready, the same analysis can be done in minutes.

With privacy and security of vital importance in the healthcare industry, the company needed to confirm that the platform addressed rigorous industry and internal standards, including the Health Insurance Portability and Accountability Act (HIPAA). By leveraging Cloudera's Shared Data Experience (SDX) capabilities, the company can manage all the metadata and policy information in a centralized fashion.

Cloudera Data Science Workbench (CDSW) is another important part of the stack. CDSW is being utilized for data discovery, exploration, ad-hoc analytics, with secure self-service access to governed data for exploration and analysis.

Results

Through adopting CDP Private Cloud, this global pharmaceutical company has been able to address horizontal scaling with OCP and data in HDFS, solve for the "noisy neighbor" problem, improve memory utilization, and implement a common security interface. Additionally they have seen the following benefits:

Increased speed and quality of the drug discovery pipeline: The timeline and price tag for the drug discovery pipeline are both much too high, notwithstanding, the success rate of a new drug is only 12%. This is why the speed and quality of the drug discovery pipeline is so important. With each phase of the development process, there's a drop-off in the odds of success. This global pharmaceutical company has now been able to speed up the drug discovery pipeline. In one example, traditional genomic data analysis tools would have taken 80 years to complete. By leveraging Cloudera's platform, the company was not only able to run genome-wide association studies analysis, but the timeline for completing this analysis was reduced to a few weeks. In addition, all research data was made more easily available to more researchers, through the fully integrated search engine in CDP - based on rich and powerful Solr. The quality and confidence in discovery paths have been accomplished and allowed for more experiments to run in parallel while being able to reduce duplicated efforts.

About Cloudera

At Cloudera, we believe that data can make what is impossible today, possible tomorrow. We empower people to transform complex data into clear and actionable insights. Cloudera delivers an enterprise data cloud for any data, anywhere, from the Edge to AI. Powered by the relentless innovation of the open source community, Cloudera advances digital transformation for the world's largest enterprises.

Learn more at cloudera.com

Enabled self-service access to data: Another one of the major wins the R&D group has seen was delivering on the goal of lowering the barrier of accessing data. Now 97% of data is visible to users within R&D. This encompasses thousands of clinical trials that have occurred, where the data is brought in and provided back to the business. Parts of the discovery and data curation process have been automated through the rich APIs that come with Solr and other open source components of CDP. Lowering the barrier to accessing data gives scientists the capability at their fingertips to deep dive on pharma analytics.

Leveraging Artificial Intelligence and Machine Learning: Artificial intelligence (AI) and machine learning (ML) are both being used to mine genetics and genomics data. For data to be available and analytics-ready, it needs to be ingested, curated, discoverable, and access needs to be governed. By using predictive analytics to identify targets with a higher propensity for seeing the program through, the drug development process has an improved success rate.

The R&D team contributes to that mission by leveraging AI and ML on genetics and genomics data to identify and validate targets better than they ever have. The Cloudera platform is a key component for them to succeed. As this global pharmaceutical company achieves greater efficiency and new insights across its many R&D processes, executives expect to ultimately move the needle in terms of time-to-market, bringing new drugs and vaccines to market more quickly and less expensively to help patients.

CLOUDERA