CASE STUDY

> For every dollar we spend on a data initiative, we are able to get $40 in return.
>
> **Andy McPhee,** Science and Enabling Units Data & Analytics Engineering Lead, AstraZeneca

Pushing the boundaries of medicine to deliver life-saving medicines

INDUSTRY

- Biopharmaceuticals

INFORMATION

- HQ: UK
- 10,001+ employees

USE CASE

- Operational efficiency: Monitoring activities across

CHALLENGE

- Bringing data together to form a single source of the truth

TALEND PRODUCTS USED

- Talend Cloud Big Data
- Talend Cloud Data Stewardship
- Talend Cloud Data Preparation
- Talend Cloud Data Quality
- Talend Cloud API Services
- Talend Data Catalog

RESULTS

- Enabled **90 percent of data to be ready for analysis within 3 minutes**
- Shortened planning cycles **from 15 days down to 3 hours**
- **$1 billion per year savings** shaving just 1 month off of every clinical trial

AstraZeneca is an English-Swedish multinational, science-led biopharmaceutical company that focuses on the discovery, development, and commercialization of prescription medicines, primarily for the treatment of diseases in four therapy areas—Oncology, Cardiovascular, Renal and Metabolism, and Respiratory. AstraZeneca operates in over 100 countries and its innovative medicines are used by millions of patients worldwide.

Like all pharmaceutical companies, AstraZeneca faces stiff competition from generics. In addition, the long, complicated, and expensive drug development process can take up to 15 years from idea to reality, with only a small portion of drugs becoming commercial products. Additionally, patents can expire after 10 years.

As Andy McPhee, Data Engineering Director at AstraZeneca, points out: "The company could be in a position where we are still testing one of our drugs and it is no longer under patent. So, we must balance this desire to speed the process with trusted data. If we do not have the quality in our data, our drugs will not be approved, and we will be affecting the lives of potential patients. Talend provides us the speed and trust we need."

Back in 2014, AstraZeneca implemented a strategic initiative of returning to growth to drive the organization forward, galvanized around data transformation and endorsed by the CEO and the CFO. Talend is at the heart of that transformation, growing 5 times the amount since the start of the AstraZeneca data journey. The company now has 50 data-related initiatives, and nearly 100 people now use Talend."

Bringing data together to form a single source of the truth

A multinational company that has grown substantially, including through mergers and acquisitions, AstraZeneca had data dispersed throughout the organization in a wide range of sources and repositories. Drawing data from CRM, HR, Finance systems, and several different versions of SAP ERP systems slowed down AstraZeneca's vital reporting and analysis projects. The company also faced a serious data challenge in the form of data silos with people creating their own versions of the truth because they did not trust the data. McPhee explains, "We were spending more time discussing the quality of the data than the business strategy. We wanted to consolidate and get a single set of global metrics so we could monitor activity across divisions and markets and do comparisons that were not previously possible."

AstraZeneca also operates in a heavily regulated environment. It must be GxP compliant, which means its products are safe, meet their intended use, and adhere to quality processes. In addition, the company must comply with the General Data Protection Regulation (GDPR), which means data privacy is at the heart of all data initiatives.

AstraZeneca decided to build a data lake to hold the data from its wide range of source systems. Cloud was a very important aspect of this architecture, providing scalability and flexibility. AstraZeneca made the strategic decision to adopt AWS as a primary cloud vendor. It also chose Talend. McPhee adds, "We selected Talend for AWS connectivity, flexibility, and its licensing model. We also valued the ability to scale rapidly without incurring extra costs from AWS or Talend."

But as McPhee explains, the data must be correct and protected. "Data without trust is useless," says McPhee. "Data Governance is critical to knowing that we can trust our data and ensuring that the data is well understood, well looked after, and only accessible to the right people."

Why Talend?

AstraZeneca is utilizing the entire Talend Cloud suite. Talend is responsible for lifting, shifting, transforming, and delivering data into the cloud, extracting from multiple sources, and then pushing that data into Amazon S3. "The Talend jobs are built and then executed in AWS Elastic Beanstalk," explains McPhee. "After some transformation work, Talend then bulk loads that into Amazon Redshift for the analytics. Talend is also being used to connect to AWS Aurora."

AstraZeneca has also deployed Talend as part of the orchestration layer in its architecture. In addition to extracting data from CRM, ERP, finance, document management, HR and other systems to load into the data lake, Talend API serves to facilitate point-to-point connections, such as those between an Amazon Redshift analytic database and a SQL database in order to add data to what AstraZeneca calls a 'conformed layer.'

Data Governance is high on the agenda. As Ian Dix, Director of Data and Analytics, explains,

Utilizing AWS cloud for scale and Talend for collecting, governing, and sharing vast amounts of data, we have built automated pipelines that would have taken years to accomplish due to the huge amounts of data involved. We want to use artificial intelligence to enable us to shorten the human trial process and improve outcomes for patients. With Talend and the data lake, these initiatives are now possible."

"Using Talend's data lineage functionality, we have started to look at data movement across systems to bring transparency to how data is being used across our organization."

Standardizing data across multiple systems was key. "We make use of Talend Data Catalog to harness the metadata that underpin the information in our data lake. We looked at the data models and the metadata to get a common language for attributes like product, indication, and patient," says McPhee.

It was also important to ensure clear control over data access. "Defining data owners and who can access what data for what kind of processes is critical to compliance with GDPR and making sure we are managing our PII data properly," comments Dix.

Business owners also must manage the quality of their data through data stewards. With Talend Data Quality and Talend Data Stewardship, users can decide how to structure and format the data so they can apply data quality rules and bring transparency to where the data quality issues are.

Andre de Jong, Business Intelligence Engineer, explains, "We do not want to develop the same pipeline across different functions repeatedly. Talend has enabled us to standardize our data movement and transformation. Now we create generic Talend jobs that can extract data from any type of data source, allowing us to deploy those jobs in Docker containers and deploy them on a serverless infrastructure such as AWS Fargate, making them ultimately scalable across business units."

Shortening the drug development timeframe and increasing efficiency in back office functions

The focus of Talend has been initially around the Enabling Units, which are the back-office functions such as finance, HR, legal, compliance, and insurance.

AstraZeneca has undertaken a multi-year program to transform its **finance processes.** "The financial hub brings together more than 25 data sources that arrive at different intervals and have complicated compliance and reporting requirements," explains Peter Wolstencroft, Finance Data Hub Platform Lead. "While the hub is only 1.5 terabytes, with Talend we are able to process data through the platform within five minutes of it landing in the data lake. We established a target to have 90 percent of data ready for analysis within three minutes. Talend enabled us to meet that goal. In addition, planning cycles that previously took 10 to 15 days now take only 3 hours."

In the **HR workforce space,** analyzing HR data from Workday cloud-based application provides the ability to report quarterly on the number of hires in different locations and the number and type of terminations, to identify patterns in employee recruitment and retention, and to adapt the HR strategy accordingly. Also, extracting data from two CRM systems and a document management system helps monitor how **commercial and marketing** functions interact with healthcare providers. This gives them the capability to segment customers and tie marketing to sales in order to market more effectively to key decision makers.

In procurement, integrating the company's iBuy system makes it possible to track purchase-to-pay data. If necessary, procurement can re-segment suppliers as the company's strategic priorities change and continue to manage all relationships in a consistent manner. It is also looking for patterns in travel and expense for policy enforcement when integrating Concur data in the data lake.

AstraZeneca is now expanding and exploiting the usage of Talend across its early science and late science units. Shortening the drug development timeframe is important to AstraZeneca because it gets lifesaving and life-improving drugs into the hands of patients faster and enables the company to complete the process before the patent expires.

In early science, the company is rethinking how to create a drug from scratch, utilizing groundbreaking technologies such as genomic sequencing to understand the effect of that drug will have on a human being. Talend and the data lake are also enabling the work of AstraZeneca's data scientists by providing them with a single source of truth for their work and enabling artificial intelligence and machine learning. The search capability provided by Talend also enables data scientists to exploit the massive amounts of data provided by the drug discovery process, imaging, and genome projects. For the imaging work, AstraZeneca has access to years of images of particular diseases and their associated metadata. With Talend, the company can use this metadata, digitize it, and push it through specific workloads and pipelines to allow data scientists to write algorithms to learn what is in the images. In the future, when they get a new image, they can predict with high degree of certainty if an image includes the disease.

In late science, completing the clinical trials is one of the most expensive parts of drug development, costing millions of dollars. Shaving just one month off of each clinical trial will save AstraZeneca $1 billion per year.

Stuart Charles, Clinical Control Tower Technical lead explains, "We have approximately 20 drugs in the pipeline at any one time. Clinical trials are complicated and regulations vary per country so that each drug will go through separate trials across 100 different countries. For any clinical trial we are conducting, we try to project how many subjects we need to recruit on to a study and then track the actuals versus plans. Are we recruiting fast enough for this site, on this country, for this study? How are we performing against that? Every time you are dealing with data, there are real people behind the data. If there is a safety issue, speed is essential as we need to report it quickly to stop further patients from receiving that treatment."

McPhee explains the company's next steps with Talend. "We want to put things out into the hands of patients and gather data that way. This would involve taking a digital and data approach to improving patients' experiences, such as placing more emphasis on IoT technologies. We are starting to use deep learning on top of full body CT scans to determine whether someone has lung cancer. In the future we will use more of these innovative technologies to accelerate our growth."