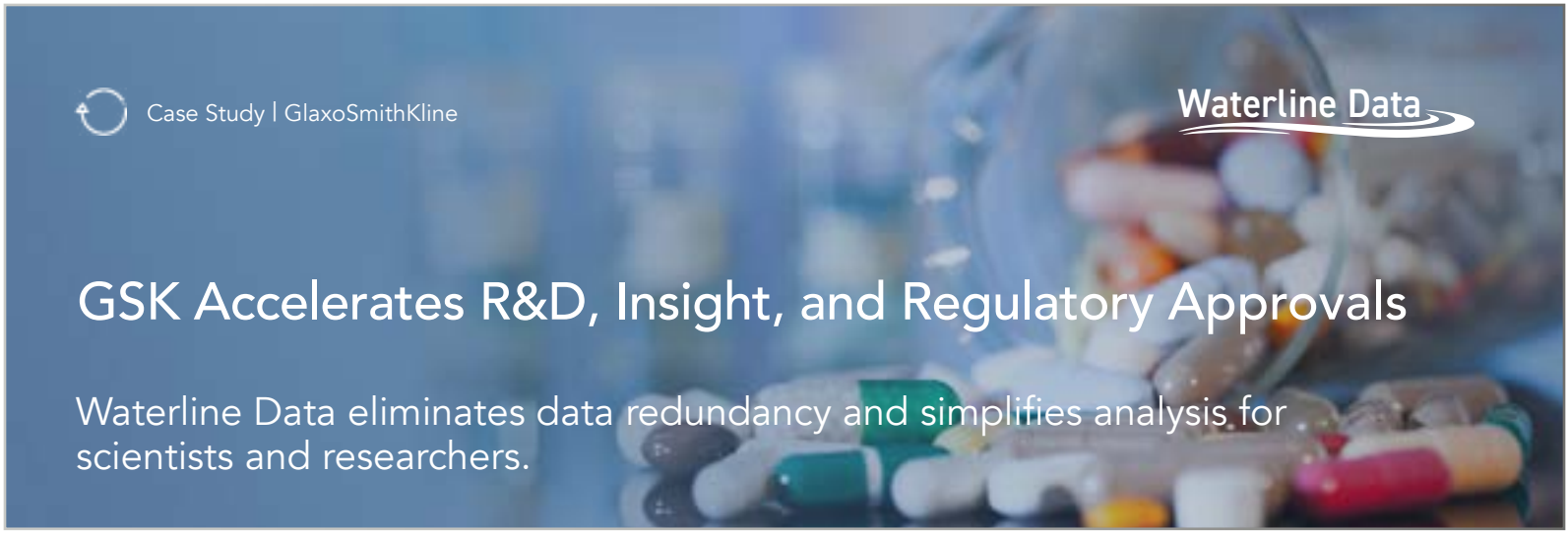

Case Study | GlaxoSmithKline

Waterline Data

GSK Accelerates R&D, Insight, and Regulatory Approvals

Waterline Data eliminates data redundancy and simplifies analysis for scientists and researchers.

GSK Backstory

GSK (GlaxoSmithKline) is a leader in global healthcare. In 2017, GSK sold 1.9 billion packs of medicine and 798 million vaccines. Their advanced R&D approach fuels the organization's innovation and success.

The Challenge

GSK is a science-led global healthcare company with practices in pharmaceuticals, vaccines, and consumer healthcare. Of the company's 98K+ employees, 8,700 are scientists that create and work with data to inform decisions on product research and development (R&D), clinical trials, manufacturing, regulatory approvals and more.

These scientists generate an immense amount of data. Teams rely on a multi-petabyte data lake deployed on Cloudera to share and analyze clinical trial data in preparation for regulatory approvals on products. The data lake collects various types of data from several different sources, which results in excessive data duplication.

"The quantifiable number that comes to mind is 2,100 data sources across a multitude of different tools and technologies. It's not a hundred terabyte problem. It's not a 200 terabyte problem. It's millions of files across petabytes worth of data," said Ranjith Raghuanath, Director of Big Data Solutions at GSK.

This redundant data clutter hindered the productivity and decision-making of researchers and scientists while also increasing the cost and complexity of storage and compute resources needed to work with the data. Delayed clinical trials resulted in slower regulatory approvals, which eventually trickled down to reduced productivity and slower time to revenue.

Challenges

- Too much time spent manually searching cluttered, redundant data
- Drawn out clinical trials and regulatory approvals

Results

- Highlighted existing duplicate and overlapping data sets
- Automated duplicate identification to avoid future redundancy

Benefits

- ✓ Reduced time to analysis and accelerated regulatory approval
- ✓ Reduced storage and server usage associated with redundant data
- ✓ Lowered costs of Oracle and Teradata licenses associated with duplicate data

Integrations Connected

Oracle – Teradata – Tamr – Cloudera

Waterline Data

The Solution

GSK now uses Waterline Data Catalog to identify duplicate or overlapping data sets so that they can be merged or deleted using Tamr. After a massive culling of existing data to eliminate redundancy, the process is now automated and ongoing. As scientists continue to conduct clinical trials, all duplicate data data sets are automatically eliminated.

According to Mark Ramsey, R&D Chief Data Officer at GSK, "The data catalog is extremely important because our model is to focus on self-service. It's really a matter of connecting the researchers and the scientists directly to the data." GSK now has a catalog of information that is easy to search and has been grounded in business terminology through auto tagging so that scientists can interact with the data and understand the results.

To gain further insight into new and existing data, GSK leverages Waterline Data Fingerprinting™ to record metadata and identify its lineage. According to Raghuanath, "We're almost tracing back the fingerprints of that data. How has it evolved? What does that look like?"

"By getting access to the data, each business can make decisions that they otherwise couldn't make," said Ramsey. GSK views being able to visualize data relationships at this level as providing long-term business benefits such as:

- **Revenue generation**: By improving access to information, the business can now make decisions based on science-based insights.

GlaxoSmithKline

> "By getting access to the data, each business can make decisions that they otherwise couldn't make.
>
> Mark Ramsey
> R&D Chief Data Officer, GlaxoSmithKline

- **Cost reduction**: Automated de-duplication addresses IT complexity issues and, for GSK, reduces the cost of Oracle and Teradata licenses.
- **Increased productivity**: Making data accessible using terms familiar to researchers and scientists accelerates insight and encourages innovation.

Waterline has helped GSK reduce time to analysis in clinical trials and accelerate its regulatory approval process, saving millions of dollars each year.

But it's also delivering an unquantifiable value by enhancing the understanding of linkages between clinical data and test data, or between clinical data and target selection data: Scientific insight that has never before been available until now.

Waterline Data automates data discovery, compliance and the ability to take action on data by using a combination of artificial intelligence, machine learning, ratings and reviews, and tribal knowledge to deliver a Smart Data Catalog. Our customers spend less time searching for data and more time using it to derive value while complying with data governance mandates such as GDPR. Our data catalog is best of breed for large enterprises with big data implementations from multiple sources or migrations to the cloud.