Published by **Fierce Markets**

DATA LAKES: HOW TO GLEAN NEW INSIGHTS FROM EXISTING DATA

amazon web services

Modern biopharma companies are data businesses. Companies answer every question from drug discovery to market access by generating and analyzing rich data. This creates a wealth of data that is used once then siloed, limiting its value. There is another way, though. Managed properly, old data yields new insights.

Researchers who refer back to historic data and combine it with data from other sources are best placed to answer the questions they face

Today's data strategies recognize the value of old data and the interconnectivity of teams within companies. As activities in one functional area have implications for the broader business, organizations are eliminating silos that keep datasets generated by different teams apart. This allows companies to correlate data from each functional area across the lifecycle of the drug, from discovery through to commercialization.

The value of having ready access to historic data in its raw form is clear, too. No researcher knows all the questions to ask a dataset on the day it is generated. Science and the companies that perform it move forward, continually revealing new questions. As this happens, researchers who refer back to historic data and, better still, combine it with data from other sources are best placed to answer the questions they face, be they related to R&D, manufacturing or commercialization.

If new evidence links a gene to a phenotype, there is value in having access to historic sequencing data and accompanying medical records to build on the breakthrough. If vaccine yield at a manufacturing facility drops, there is value in having access to years of data on every aspect of the operation to look for patterns that explain the trend. When commercializing a drug, there is value in creating a real-world evidence ecosystem to demonstrate the value of the product to payers, physicians and patients. And there is value, even necessity under Quality by Design approaches, in using clinical production data to inform manufacturing scale-up.

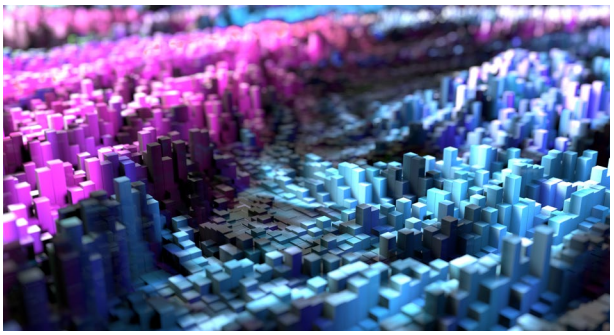
There is also value in providing intellectually-curious people with sandboxes of data. Science moves forward on a succession of "what if?" questions. Companies with centralized repositories of all their data have unparalleled capacity to answer such questions. They can also reveal "unknown unknowns", insights they never knew to look for but were able to uncover by exploring their data.

Data scientists working in fields as diverse as process development and observational research want to perform such analyses. The problem is today's data analytics pipelines are better equipped to answer predefined questions.

POOLING DATA TO DRIVE DISCOVERIES

Today, enterprise data warehouses form the backbone of analytics pipelines. Before entering the warehouse, data undergoes a process known as extract, transform and load (ETL). The goal is to pull data from source systems, transform it and load it into the warehouse. The data is then ready to use in reporting and business intelligence. This gives companies a single source of validated data everyone uses to fuel analyses.

Companies need this single source of truth but many are also realizing it cannot meet the needs of all their users. Transforming data prior to storage limits its use to business intelligence. If users want to perform



Data lakes allow you to store any type of data in any format

exploratory analyses or use the data in machine learning and predictive analytics, they need access to the raw form. Restrictions on the types of data stored in warehouses are an issue for companies interested in pulling data from devices and the internet of things, too.

The warehouse model also hives off data into disparate repositories that may run on different technologies, making it very difficult to combine them for analysis. Datasets generated in clinical trials are kept apart from those gathered by manufacturing, marketing and other functional areas. This is limiting their value.

"Historically, because systems were designed for purpose, data would effectively feed a given intended use. But the data strategy wasn't such that the findings from a specific set of signals in a clinical trial could be correlated very easily across the entire lifecycle of the drug from conception through how it's used in market, reimbursed and the messaging and positioning around it to the provider, payor and patient communities," Mark Johnston, Global Director of Healthcare and Life Sciences at Amazon Web Services (AWS), said.

Recognition of these shortcomings led to the development of a new pipeline paradigm: Data lakes. This approach stores data without first putting it through the extract, transform and load process. Instead, data generated from across the organization flows directly into a centralized repository, only stopping to be tagged to make it easy to find. Services from vendors including AWS facilitate this process.

Users can access raw or previously-transformed data. This frees people from having to start from scratch every time but gives them the flexibility to access raw data if the transformed version is ill-suited for their needs. Importantly, everything is kept in one place, making it easy to find the most appropriate data.

"Data lakes allow you to store any type of data in any format or correlative iteration. When you are looking for datasets that are relevant to your questions, if the data is properly catalogued, they are easily found in a single repository. This single repository model allows you to format, ingest and analyze these data sets into appropriately tooled compute environments," Dario Rivera, Senior Solution Architect at AWS, said.

THE FIRST WAVE OF DATA LAKE APPLICATIONS

Multiple biopharma companies have shown the value of bringing disparate datasets into a central repository.

Merck delivered an early validation of the data lake approach in 2013 when it set out to understand why it was discarding a higher proportion of certain vaccines than usual. The affected site had data on all aspects of the manufacturing workflow, such as minute-by-minute temperature readings from across the facility and process control records for each batch. What it lacked was a way to quickly analyze these resources in search of an answer to its discard problem.

That changed when the vaccine team loaded its data into a platform running on AWS. Within three months of starting to centralize its data on

the platform, the team had a conclusive answer to why discard rates for one vaccine were higher than expected.

Merck made the breakthrough by plotting data from every vaccine batch ever produced at the plant on a heat map. This revealed patterns that led Merck to identify fermentation performance traits that correlated closely to yield. Merck ran 15 billion calculations and 5.5 billion batch-to-batch comparisons in its search for the answer.

Merck ran 15 billion calculations and 5.5 billion batch-to-batch comparisons in its search for the answer

Other companies followed Merck in setting up data lakes to glean new insights from old data. Amgen embraced the idea after its process development and observational research groups asked for capabilities that were beyond the scale and functions of its existing data warehouses.

Members of the process development group wanted to use the growing output of data from Amgen's labs, production lines and bioreactors to optimize processes. Similarly, statisticians and epidemiologists on the observational research team had access to ever more real-world evidence but lacked tools to effectively mine the data for insights into the safety, efficacy and economic value of Amgen's products.

Amgen responded to both situations by centralizing its existing repositories to create a data lake. The biotech then added the means for users to automatically spin up environments with the tools they need to quickly uncover insights in data housed in the central repository.

BUILDING AND HOSTING DATA LAKES IN THE CLOUD

The fast expansion in the data available to manufacturing teams and growing need for biopharma companies to empirically demonstrate the value of their products mean process development and observational research groups are two of the big beneficiaries of data lakes. They are far from the only functional areas to adopt the model, though. The emergence of population-scale sequencing means data lakes are equally valuable to genomic researchers.

"We have customers that have put together data lakes of close to 20 petabytes of mostly sequencing data," Patrick Combes, Global Technical Leader, Healthcare and Life Sciences at AWS, said.

The size of these data lakes shows why the cost of storage is critical to the concept. The whole point of a data lake is that it houses all of an organization's data. If concerns about cost force an organization to start being selective about what goes into the data lake, the value of the system diminishes as it is less well equipped to answer unforeseen questions.

When coupled to the fact data lake technology needs to be scalable, extensible and flexible, these cost considerations mean the cloud is well suited to the concept. The cloud offers low cost storage that scales automatically in line with user needs. On premise storage, in contrast, forces companies to estimate their future needs and pay for more capacity than they currently use.

The cost effectiveness of a cloud data lake is further improved by using a technologies that handle compute and storage separately, such as Amazon EMR for the former and Amazon S3 for the latter. This allows storage to be scaled independently of compute capacity. As data lakes use more storage

than compute capacity, the use of Amazon EMR and S3 frees companies from paying for more compute capacity than they need.

AWS customers have realized these benefits when setting up petabyte-scale data lakes that combine in-house sequencing data with publicly-available genomic repositories, annotation information, phenotype data and other resources. This pooling of data creates lakes with the scale to identify rare variants and the context to start understanding the significance of sequencing results.

Companies are also using data lakes at the opposite end of the biopharma value chain, for example to analyze how to improve adherence to drug regimens. Having access to a pooled repository of data allows companies to ask questions about what actions improve adherence.

The answer may differ depending on the drug and nature of the adherence problem. But the process of analyzing pooled data to identify problems and potential responses remains constant, as does the impact of improving adherence on health outcomes and company financials.

HOW TO IMPLEMENT DATA LAKES

The barriers to setting up data lakes have come down since early pioneers such as Merck used the approach. Initially, companies had to create the structure, metadata system and governance of their data lakes from scratch. Mistakes at this stage have serious implications. Failure to marry metadata to an effective search function makes it difficult to locate data. This is a common enough problem for the resulting hard-to-search repositories to have their own name: Data swamps.

Data lake solutions from vendors such as AWS have eliminated this danger by ensuring websites point to where datasets are and describe what they contain.

These solutions also simplify security and compliance by using the strong safeguards and administrative controls built into cloud services.

The upshot is biopharma companies can now create secure and searchable data lakes in minutes.

Companies that are seizing this opportunity recognize it is insufficient to simply generate data across their businesses. They know it is also essential to be smart about how that data is captured, characterized, shaped and given meaning.

Data lakes are an enabler of this way of thinking. It is a way of thinking that goes beyond generating data to answer known questions today. Instead, it gathers, organizes and explores data to break new ground, answering questions nobody knew to ask and making discoveries for which nobody knew to look.