Case Study:
Symphony Health

The Company

Symphony Health's business is data analytics and data science for the healthcare industry. It employs expertise and technology to ingest large volumes of anonymized health information, integrate it, and analyze it for the benefit of its customers. Founded in 2012, Symphony Health delivers high-value data, analytics, and innovative solutions with the objective of helping its customers to improve their performance, productivity, and profitability. It is owned by PRA Health Sciences, a global company with more than 13,000 employees spread over 80 countries.

Lags in Data, Analysis

Fresh data is always important, but when your business is data analytics, it is critical. Symphony Health was constrained by a legacy solution that loaded data into Oracle databases typically once a day, and into a data warehouse weekly.

In addition to data being delayed before it was available for analysis, performing new types of analysis against data in the Oracle databases took longer than it should have. If analysts needed a new schema, they had to request it from a database administrator. The request then went into a work queue, and the analysts waited for the schema to be created. This process could delay new analyses.

The delays in data availability, early data discovery, and analysis were unacceptable, and a new approach was required.

Leveraging Big Data

Symphony Health transformed its data management and analytics processes by moving to Hadoop, which has a number of benefits. For one, analysts can easily define their own data schemas in Hadoop, eliminating the need to wait for a database administrator to do it for them.

In addition, the more data Symphony Health stored in the Hadoop Distributed File System (HDFS), the less it had to store in a high-cost, proprietary RDBMS. They could use industry standard commodity hardware, rather than having to buy bigger, more expensive servers, so storage costs were drastically reduced. In fact, some industry reports state that open source data management on industry standard hardware can be as much as 90% less expensive than traditional relational databases.

The Challenge
Symphony Health needed to minimize data latency, reduce costs, and fully leverage Hadoop to increase the value of its analysis.

The Precisely Solution
Connect (ETL)

The Result
Analysis is performed more easily and quickly on data that is up-to-the-minute current in a versatile, high-performance Hadoop environment.

Industry
Data science for Healthcare

Another advantage of Hadoop is parallel processing and operations at scale. However, to take advantage of this capability, Symphony Health needed an ETL tool that would distribute processing on the Hadoop cluster. Their existing tool performed all processing on an edge node, which overloaded the single server and slowed the data processing.

The Solution
Symphony Health turned to Precisely's data integration solution, Connect, to get the best results from their new Hadoop environment, including an average of a 3-5X performance increase. In one module, runtime dropped from 20 minutes with their old solution to 20 seconds using Connect across the Hadoop cluster.

"Before, all ETL processing was done on a single server, the edge node. When Preciselycame in, the processing was done in Hadoop the way it was meant to be done," explained Robert Hathaway, Senior Manager Big Data. "The whole point of processing in Hadoop is to take instructions from the edge node and push the work to the cluster. With Precisely, we got the parallel processing that Hadoop was designed for, and no one had to write Java MapReduce or Spark code."

In addition, Connect provides the flexibility for Symphony Health to execute any task out of the workflow where and when it's needed. And if they want to do part of the work another way, like using PySpark, they can.

Benefits
Symphony Health gains a number of benefits from Connect and Hadoop. The most tangible is financial, with at least two drivers of lower costs. Not only are Connect licences less expensive than their prior tool, but also, storing data on standard hardware and Hadoop is much less expensive than storing it in the company's Oracle databases and data warehouse on expensive servers. Without the switch, the company would have had to buy more Oracle databases and accompanying high-end hardware to handle the growing data volume, and the increasing amount of ETL work.

Another benefit is speed. There are two aspects to that as well. First, data is now available for analysis much faster than it had been, making the company's analytics—the product it sells to customers —timelier. "Before, part of the data wasn't available for a day, and other parts not for a week. Now it's all available for analysis within minutes of the data arriving."

Unlike other solutions that can have poor performance because their workflows haven't been optimized, Hathaway pointed out that, "Connect is already optimized. We use its Intelligent Execution and it just performs."

The second aspect of greater speed is the ability to create queries and, therefore, perform analyses more quickly. This is due to the ability of analysts to create their own schemas on Hadoop, as well as the intuitive Connect user interface that analysts found easier to use. With just a few days of training, the team was quickly ramped up. Analysts and developers can now create new ETL workflows very quickly, and spend more time analyzing data.

> "Before, part of the data wasn't available for a day, and other parts not for a week. Now it's all available for analysis within minutes of the data arriving."
>
> — Robert Hathaway,
>   Senior Manager Big Data

"It's led to people being able to ask more questions and find things out sooner," said Hathaway. "We get the same end result, faster, cheaper, and with a bigger pool of developers to draw from who can do the work. I'm a C# and Java developer who even knows some Scala, and I still like using Connect because I can get a lot more done in the same time."

Another benefit of Connect is its flexibility. In addition to its Hadoop integration, Connect pushes data to Amazon Redshift with minimal latency. This allows data scientists to perform advanced queries on the Cloud, and makes it easy to provide analytic results to clients on Amazon, through a front-end application.

Connect's decoupled deployment option provides yet another benefit. If Symphony Health finds software that does a better job of one part of the process, the company can easily plug that new software in, without having to replace the whole solution. In short, they're not locked in. Connect is future-ready.