DB Cargo Gives Users the Green Light to All Data with Dremio

Challenges	Solution	Results
DB Cargo's 4 large data warehouses had reached their capacity. The cost of additional storage & compute would have been immense.	Chosing an AWS cloud and Dremio, DB Cargo opted for maximum flexibility and cost control.	Dremio delivers data directly via pipelines and native connectors. This relieves the load on the system, reduces the risk of failures and saves costs.
Terabytes of data had to be moved for each job, so even a simple join operation took several minutes.	In DB Cargo's so-called "Data-as-a-Service" project, Dremio is playing the part of secure and fast data supplier.	Near-real-time results ensure more efficient transportation planning, higher shipment quality and faster decisions.
In order for the data warehouses to accommodate the latest data, older data had to be moved. Users needing this data often had to wait a week or more.	A "single source of truth" in the cloud and Dremio's semantic layer facilitate access to all data for all users.	From Excel users to Python programmers, all users can access information in no time and then focus on their tasks 100%, increasing employee productivity.

Summary

DB Cargo stores its vast amounts of data in an AWS cloud and has implemented a modern architecture that meets the requirements of multiple user groups. Dremio ensures the reliable and fast delivery of all required data in the users' tools of choice. Migrating to the cloud and employing Dremio resulted in cost and time savings, better processes and planning quality, and increased efficiency and productivity within business units.

The Business: Rail Freight Transport is the Green Solution

Deutsche Bahn Group (DB) is one of the world's leading mobility and logistics companies. The DB Cargo business unit manages DB's rail freight business. As Europe's leading rail freight company, DB Cargo has its own production network. The company's 2,600 locomotives and 80,500 freight wagons transport and shunt goods on a network that stretches from Lisbon to Russia's Nizhniy Novgorod and as far away as Shenyang in China. With the largest fleet on the European continent, 30,100 employees ensure that all kinds of cross-border transports get safely and efficiently from A to B. Green logistics are an integral part of DB Cargo, because rail transport is not only the safest but also the most eco-friendly means of freight transport. DB Cargo's transport services save 7 million tons of CO_2 per year and help to decongest the roads.

Challenges – Big Data Dilemma

No question: The more information you have, the easier it is to assess the impact of changes, identify trends, and make predictions. But what if there is so much data that it threatens to push the boundaries of the system? At DB Cargo, these huge amounts of data are generated every day. Apart from the systems used by a couple of business units, this data was stored in four large data warehouses. For each job terabytes of data had to be moved, so even a simple join operation took several minutes.

But poor performance was not the only problem. The data warehouses had reached their capacity. The IT team at DB Cargo was virtually forced to sort out old data in order to be able to ingest new data. As the cost of additional storage would have been enormous, they opted for a solution that combined maximum flexibility with cost control – the cloud.

"With its query engine, Dremio provides DB Cargo Data Lake users with an outstanding lakehouse platform: robust and scalable, particularly suited for large data volumes and a simpler access to data assets as a starting point for analysis and reporting."

Team DB Cargo Data Lake
DB Cargo

Solution – On Cloud Nine

After deciding to use Amazon Web Services, the DB Cargo IT team was now faced with the task to build a robust and reliable cloud data lake architecture and move the data to the cloud. In this process, meeting the requirements of the various users was key. SMEs in management accounting, sales, or HR should be able to work with the data just as easily and confidently as statisticians, data scientists, and application developers.

In addition to AWS tools, open source solutions from the Apache Software Foundation are also part of the new architecture. While Amazon S3 is the storage system and AWS Glue serves as data catalog, Apache Nifi takes care of ingestion and processing by managing the data traffic from the source systems, removing duplicates in the process. Spark is used to process specific key figures, for example, from punctuality measurement. In addition to ODBC/JBDC, Arrow Flight provides connectivity to various upstream systems.

In DB Cargo's so-called "Data-as-a-Service" project, Dremio is playing the role of secure and fast data supplier. Dremio's open lakehouse platform can provide Excel users with analytics-ready data just as quickly and reliably as BI users or R and Python programmers. Dremio's semantic layer facilitates data access for all users.
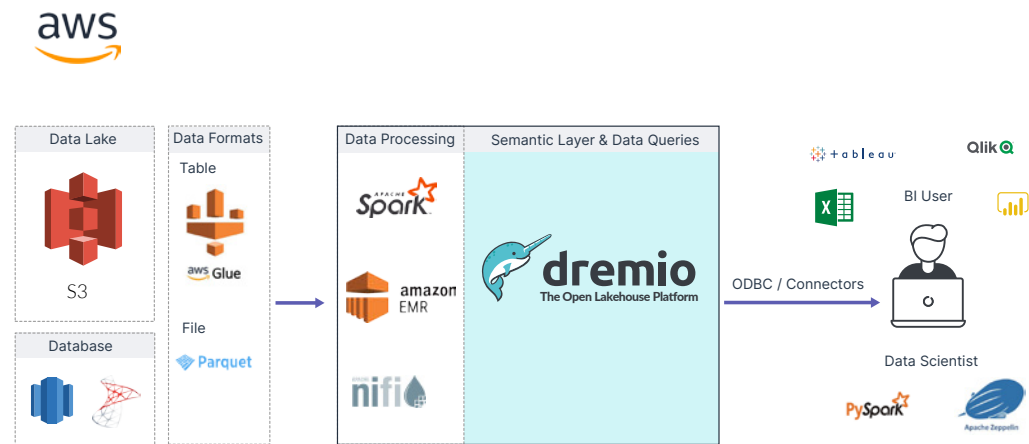
Copyright: Deutsche Bahn AG / Oliver Lang

Results

The new system has been live since 2020 and is gaining more and more users. Currently, 600 are registered and 300 users are actively exploring the entire spectrum of DB Cargo data with the tools of their choice. Business units and IT benefit from the advantages, including:

Flexibility

The new system is not only much faster than the old data warehouses, but also more agile. This is an important plus for the IT team, because now data is processed only when needed, while Dremio ensures that it ends up directly in the user applications.

Efficiency

The cloud data lake promises unlimited storage at foreseeable costs and enabled the IT team to separate storage and compute. With Dremio, data no longer needs to be actively collected. Instead, it is practically delivered to the door via direct pipelines and native connectors, thus eliminating the formerly frequent access of the production system. This relieves the load on the system, reduces the risk of failures and saves costs.

Time to Value

Where is my data? The users had to spend 60% of their time on answering this question. When the data was finally found, more often than not a table was missing from the dataset. With Dremio this tedious and frustrating search has ended. Today, subject matter experts have all crucial information at their fingertips and can focus 100% on their analyses.

Speed

To be able to include the latest data in their data warehouses, the IT team had to move older data first. Anyone needing this historic data had to wait a week or more. This delay was particularly annoying for statistics, as all the data that was not available on the reference date became obsolete. Today, all data can be accessed without limitations. This is possible using Reflections, a unique Dremio feature. Reflections store datasets in already optimized formats. They eliminate the need to access source systems and provide immediate results.

Use Cases Operational Excellence in Transportation Planning

In the capacity control application, a shipment is planned from the initial enquiry to the delivery at the destination and the shipment quality is monitored. Every change to the plan, for example whether the contract has been signed or the shipment is already on its way, is logged in JSON.

dremio

The total data volume is 10 terabytes per year. And with Dremio's virtual datasets up to 200 gigabytes of JSON data per day can be conveniently analyzed any time in near-real time in tools like Qlik. So the analyst can identify a need for action, for example when a route is frequently blocked during storms due to fallen branches.

Successfully Combating Noise with IoT Sensor Data

As the use of particularly noisy freight wagons was banned in Germany on December 13, 2020, measuring and evaluating sensor data and making the appropriate constructional changes allowed DB Cargo to increase the number of quiet freight wagons to 90 percent. These wagons are 10 decibels quieter, which is a perceived halving of the noise. Dremio played the key role in delivering these 50 terabytes of IoT data.

Next Steps

Currently, there is 1 petabyte of data in the DBC Data Lake, but that is by no means all the data available. Therefor, the data lake team is planning to include all product, sales, and financial data in the near future. The next step will be to extend self-service. The SMEs should soon be able to handle data preparation, BI queries, and analytics on their own.

ABOUT DREMIO

Dremio is the lakehouse company. Hundreds of enterprises around the world, including brands like Allianz Global Investors, FactSet, Knauf Insulation, Nutanix and OTP Bank, use Dremio's SQL engine to deliver mission-critical BI on the lake. As the original creator of Apache Arrow, Dremio is on a mission to reinvent SQL for data lakes and meet customers where they are in their cloud journey. Dremio was founded in 2015 and is headquartered in Santa Clara. To learn more, follow the company on GitHub, LinkedIn, Twitter, and Facebook, or visit www.dremio.com.

Deploy Dremio
Learn more at dremio.com

CONTACT SALES
contact@dremio.com

dremio