CASE STUDY

Amazon Accelerates Supply Chain Decision Making by Implementing an Innovative Analytics Architecture using Dremio

Monish Balasundaram, Head of Analytics, Amazon
Jeremiah Morrow, Product Marketing Director, Dremio

At a Glance

The Customer

amazon

Challenge

Amazon's Supply Chain Finance Analytics team had a complex architecture, and needed to deliver high performance Business Intelligence reporting on billions of rows of data to enable supply chain for decision making.

Solution

The team developed a new analytics architecture with Dremio to simplify ETL processes, accelerate queries, and provide analytics on complete and unified view of the data.

Results

- 10x increase in query performance, from 60 seconds to 4-6 seconds.
- 90% reduction in setup time.
- Auto query planning eliminated as much as 60 hours of work per project
- Supported thousands of concurrent users
- No limitations to data available to data consumers

The Customer:

Amazon's Supply Chain Finance Analytics team is responsible for providing controllership for Amazon's consumer supply chain systems and serves as a strategic partner for the Supply Chain Optimization Technologies (SCOT) team. The team's work is data-intensive, and they have to stitch together hundreds of different tables produced by various business units to create a unified financial view of the results from supply chain systems that are responsible for making science and machine learning driven decisions on tens of billions of dollars in supply chain spend. The team faced significant challenges in streamlining Extract, Transform, & Load (ETL) processes and in reducing their data engineering workloads while also providing reliable, consistent, and accurate insights to internal users with high-quality analytics products.

The Challenge:

The SCOT Finance Analytics team at Amazon had to create financial reporting that provided a unified financial view of supply chain systems for decision-making. The team had to deal with enormous volumes of data, with some datasets consisting of billions of rows and hundreds of columns.

The insights had to be generated on various dimensions and filters, including date ranges, geographical regions (country, state, city, zip), business categories, and product categories. However, the data source was only performant for 1-3 dimensions of filters, and the BI tools' extracts usually failed after approximately 100 million rows, degrading analytics products performance severely, especially with high numbers of columns. Queries took very long (four minutes or more for complex queries) to complete, even with sample data.

To try to solve this challenge, the team initially created 20+ materialized views of the same table in the data warehouse that were partitioned, sorted, and distributed by different dimensions, and they used parametrization functions in the BI tool to select the right data source based on the filter and value selected by the user. However, this approach was not performant or scalable, and it required significant

amounts of ongoing maintenance and management. It also created numerous and complex data pipelines, introducing more chances for human error, and varying system responsiveness based on the Data Warehouse cluster's load. They required a better and more scalable solution to meet their standards.

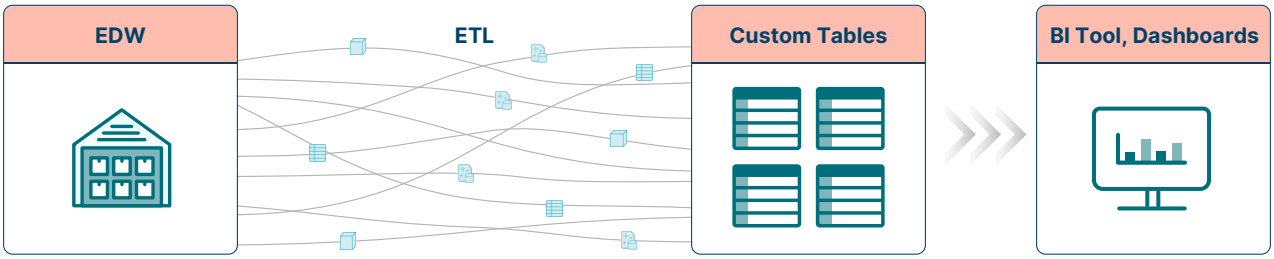
The Solution:

The team researched several commercial and open-source solutions, including Apache Kylin, a commercial big data Online Analytical Processing (OLAP) provider, and Dremio for in-depth evaluation. They had several requirements: it needed to produce views with less than a 10-second refresh time for each user click or filter selection, deliver consistent completion of daily backend data appends, query the entire dataset (with more than three years of historical data), without reducing scope, offer low setup and maintenance labor, and scale compute elastically without bottlenecking resources.

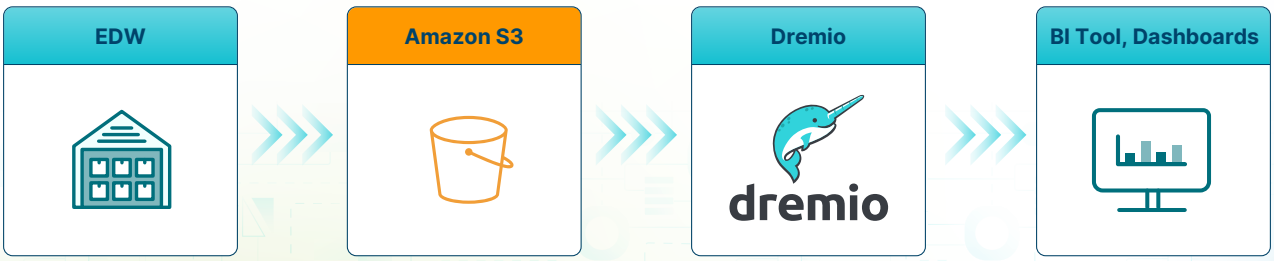
Ultimately, the team chose Dremio. They quickly set up a Dremio instance using an Amazon Web Services CloudFormation Template (AWS CFT), and they were able to scale compute up and down as needed. Dremio allowed them to do everything using a modern User Interface for both SQL and no code analytics. Dremio had seamless integrations with existing BI tools and a built-in SQL Runner (SQL IDE) for ad hoc query analysis and exploration. The team was able to set up a reflection build trigger based on new data ingestion into the Amazon Simple Storage Service (Amazon S3) bucket, simplifying their ETL process. Dremio could build multiple combinations of reflections using a GUI interface. After the evaluation process, they chose Dremio over Kylin because it had a faster setup time, it was more user-friendly, and it offered seamless integration with most BI tools.

Results:

The team selected Dremio because it met or exceeded all of their requirements. They were able to achieve a 10x increase in query performance, reducing response time from 60 seconds to only 4-6 seconds. In addition to improved analytics product performance, they were also



Before Dremio



After Dremio

able to run multiple concurrent queries, consolidate them, and return all results in less than 4-6 seconds. They were also able to connect their reporting to all data without reducing the scope or using subsets of the data. They also achieved consistent, quick, and easy updates of reflections based on new data in Amazon S3.

The SCOT Finance Analytics team were able to easily build multiple reflections in the GUI interface, and the Dremio query planner chose the best reflection(s) and/or combination of datasets and reflections to deliver the lowest-cost query plan. The auto-query plan eliminated more than 60 hours per project, a tremendous amount of savings for the team. They also eliminated resource bottlenecks they were experiencing with their data warehouse solution, as Dremio existed outside of it as a standalone installation, and Dremio did not slow down other production pipelines. Their initial setup took 90% less time versus other solutions.

Conclusion:

Amazon's Supply Chain Optimization Technology (SCOT) Finance Analytics team faced challenges in managing their data pipeline while providing reliable, consistent, and accurate insights to internal users with high-quality analytics products. They were able to achieve their goals by using Dremio to accelerate queries, streamline ETL, and reduce their data engineering workloads, getting high-quality insights into the hands of their end users fast. Dremio was able to meet all of their requirements, and they were able to deliver the end result exactly as imagined with no compromise on their vision.

ABOUT DREMIO

Dremio is the easy and open data lakehouse, providing self-service analytics with data warehouse functionality and data lake flexibility across all of your data. Use Dremio's lightning-fast SQL query service and any other processing engine on the same data. Dremio increases agility with a revolutionary data-as-code approach that enables Git-like data experimentation, version control, and governance. In addition, Dremio eliminates data silos by enabling queries across data lakes, databases, and data warehouses, and by simplifying ingestion into the lakehouse. Dremio's fully managed service helps organizations get started with analytics in minutes, and automatically optimizes data for every workload. As the original creator of Apache Arrow and committed to Arrow and Iceberg's community-driven standards, Dremio is on a mission to reinvent SQL for data lakes and meet customers where they are on their lakehouse journey.

Hundreds of global enterprises like JPMorgan Chase, Microsoft, Regeneron, and Allianz Global Investors use Dremio to deliver self-service analytics on the data lakehouse. Founded in 2015, Dremio is headquartered in Santa Clara. CNBC recognized Dremio as a Top Startup for the Enterprise and Deloitte named Dremio to its 2022 Technology Fast 500. To learn more, follow the company on GitHub, LinkedIn, Twitter, and Facebook, or visit www.dremio.com.