LARGE GAMING COMPANY

AUTOMATIC OPTIMIZATION & COST SAVINGS

Tabular helped a gaming company solve an urgent problem:

securing access and adding data governance to high-volume event data across multiple query engines. The company chose Tabular's all-in-one platform that includes a modern metastore, RBAC access controls, sync to Snowflake, and automatic optimization. This combination enabled the company to solve its governance problem quickly and optimized the underlying data—now queries run twice as fast. Projections show automatic optimization can save the company nearly $2 million over the next two years in S3 costs alone.

Tabular

CASE STUDY

Executive summary

Tabular worked with a data team at a large gaming company that urgently needed to rebuild their high-volume event data pipeline. The existing solution kept more than

1,100 event types mixed together in JSON format flowing through a single Kafka topic. This strategy was costly to query and had to use the same protections for all events.

The team needed to separate events by type and track the sensitivity of columns in each type, to enforce controls for sensitive events. Each event type first needed to be described by a schema. Then events needed to be landed in separate tables for long-term storage and individual Kafka topics for stream processing.

Creating and validating the schemas for each of the 1,100 event types was a massive undertaking, in addition to transitioning production analytics to the new tables. This left the team with little time to set up infrastructure to track tables, enforce access controls, and load incoming data. The team chose to build on Tabular's warehouse platform, which met those needs on time, delivered significant cost savings, and made the tables available in Snowflake.

By using Tabular, the team was able to focus on their data problem: adding schemas and updating production analyses. The team used Tabular for critical infrastructure responsibilities that they otherwise would have needed to build internally, including:

Data loading

Loading data into 1,100 tables every 5 minutes requires more than 300,000 jobs per day that are run and monitored by Tabular

Access controls

Enforcing role-based access controls consistently by protecting the data layer instead of locking down query engines individually

Snowflake sync

Syncing all tables to Snowflake as external tables, allowing the team to quickly adapt existing pipelines and queries instead of moving to a different query engine

Choosing Tabular created additional benefits

The team wanted to modernize their data lake, and with Tabular, can now take advantage of Apache Iceberg, the open standard for analytic tables. Iceberg makes their tables compatible with many engines besides Snowflake (including Trino, Spark, and Hive), and makes queries faster and data engineers more productive. For instance, the team later needed to update event schemas, a critical feature that Iceberg natively supports.

> Automatic optimization has reduced the team's data volume by about 48% and now jobs run in half the time and resources.

48%
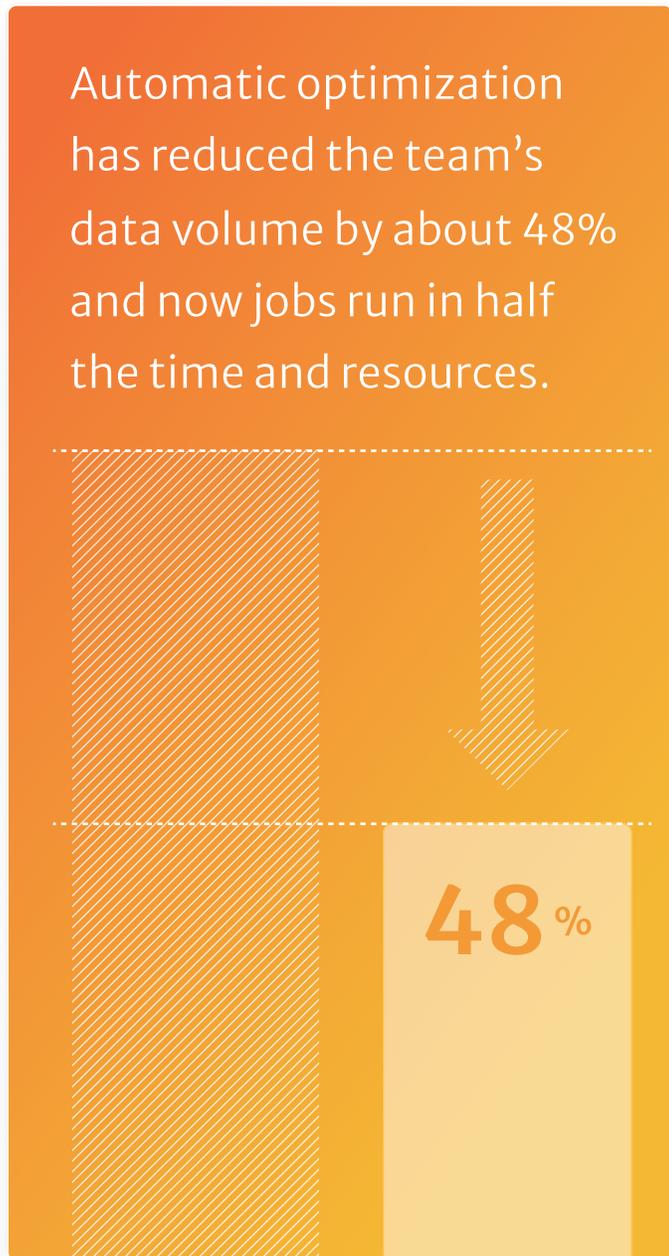
The biggest additional benefit is the impact of Tabular's automatic optimization. Tabular automatically analyzes every table to recommend tuning settings and rewrites tables using those settings to improve query performance. Automatic optimization has reduced the team's data volume by about 48% and now jobs run in half the time and resources. Optimization is projected to lower S3 storage cost alone by $1.9 million, by storing just 7.7 PB instead of 14.6 PB over two years.

Cost savings

In December 2022, Tabular loaded an average of 20.2 TB of source Parquet files each day. Optimization reduced that volume to 10.4 TB per day. Similarly, the number of files was reduced from about 4 million per day to 60,000 per day. While cost reduction was not the goal, the savings was significant in several areas:

Small files make queries run slowly and, in some cases, can cause S3 request charges to exceed storage costs. Optimization reduced the number of files—and S3 request costs—by more than 98%.

After one year, the difference in data volume reduces monthly S3 storage cost by $75,000. This rises to $150,000 after 2 years, compounding until the compacted data is deleted.

Data volume is directly related to query runtime. Reducing data by nearly half reduces query runtime and cost by half — a huge benefit to the team's productivity. Faster queries mean insights and analysis are delivered quickly and reliably.

DATA MADE SIMPLE

SALES@TABULAR.IO