Enterprise Strategy Group™ by TechTarget

WHITE PAPER

Create Valuable Data Lake Analytics with Starburst Galaxy on AWS

By Justin Boyer, IT Validation Analyst
Enterprise Strategy Group

August 2023

Contents

Abstract

This White Paper from TechTarget's Enterprise Strategy Group documents our evaluation of the Starburst Galaxy data lake analytics platform. We evaluated how Starburst Galaxy offers data lake management and query capabilities that provide data warehouse functionality within an AWS data lake.

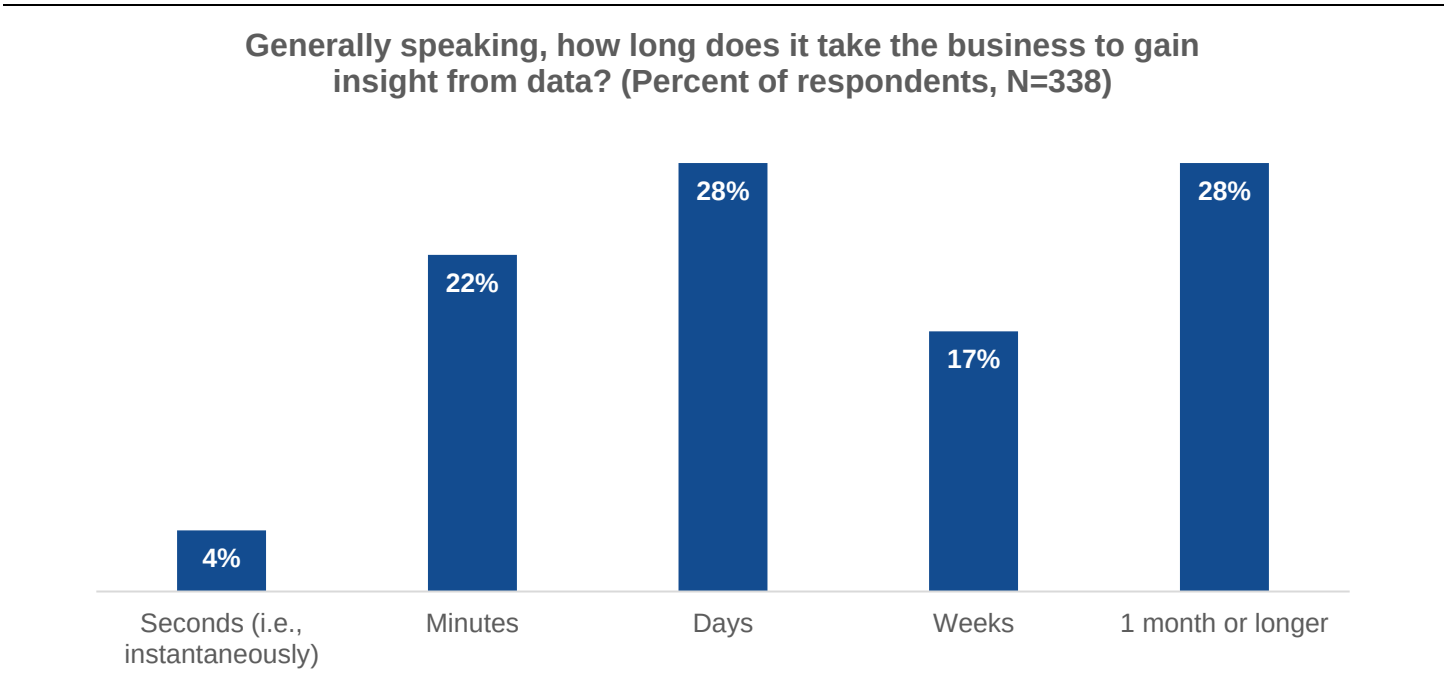
Challenges

In addition to being ubiquitous and growing exponentially, data has become a critical resource for businesses. Organizations use it to understand their customers, measure business goals, and make decisions.

Modern cloud architecture makes gathering and working with data easier than ever. Instead of large, expensive, data warehouse databases and applications hosted on company-owned servers, organizations can use cloud providers to store nearly unlimited data and pay only for what they need at the time. With the power and scalability of the cloud, data lakes have become valuable resources, holding multiple types of data for analysis from disparate sources.

However, even with the flexibility and business advantages of the cloud, many organizations struggle to find insights quickly within the mountain of data they collect. As shown in Figure 1, 73% of survey respondents reported that it takes their business days or longer to gain insight from data, with 28% saying it takes a month or longer.[1]

Figure 1. Organizations Struggle to Gain Timely Insights from Data



Source: Enterprise Strategy Group, a division of TechTarget, Inc.

Organizations need to organize and operationalize their cloud data lakes so they can gain the necessary insights quickly enough to remain competitive.

[1] Source: Enterprise Strategy Group Research Report, *Cloud Analytics Trends*, October 2022.

Starburst Galaxy for Structuring Data Lake Analytics on AWS

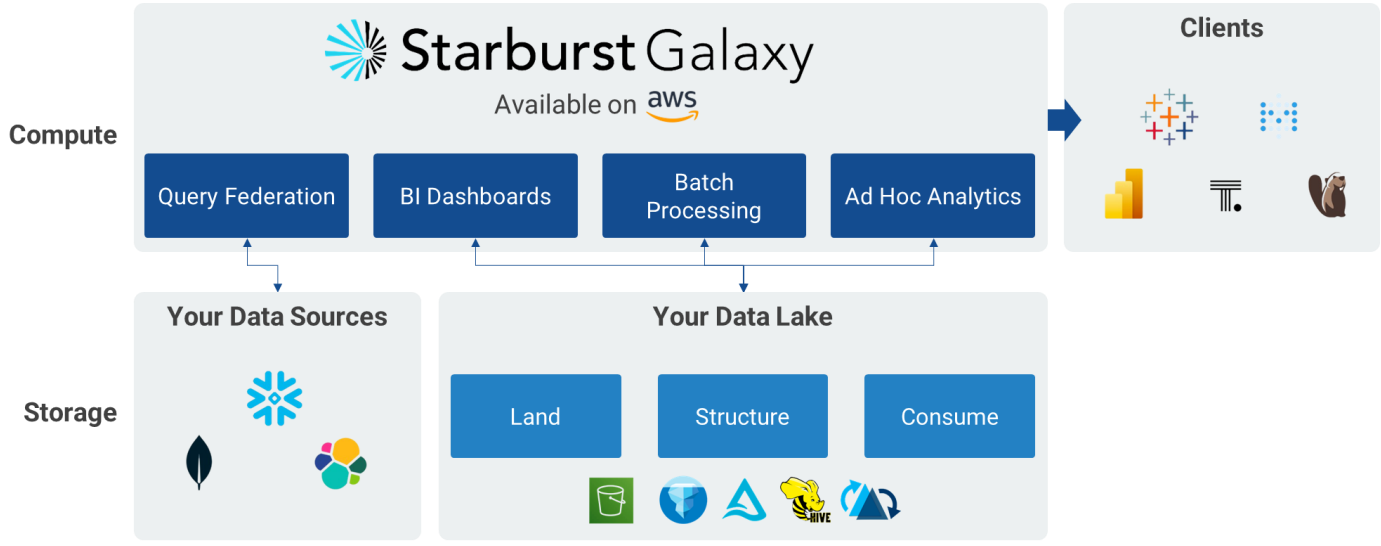
Starburst Galaxy is a software-as-a-service data lake analytics platform that helps organizations discover, govern, optimize, and analyze data. Starburst Galaxy builds upon the low-cost and scalable Amazon Simple Storage Service (Amazon S3) to store all kinds of data: structured, unstructured, and semi-structured.

Starburst Galaxy is built on Trino, an open source, distributed query engine built to process queries against multiple data sources at scale. Galaxy improves time to insight with its ability to efficiently query multiple sources concurrently by providing a single point of access and governance regardless of where data is stored. Starburst Galaxy completely separates data storage and computation. For example, organizations store their data on Amazon S3 while Galaxy can scale the computing power necessary to perform queries with speed and accuracy.

Starburst Galaxy connects to and queries multiple modern and legacy enterprise sources, minimizing data duplication and providing a holistic view of all the data in the ecosystem, regardless of where it lives. The platform connects to popular data sources such as Snowflake, MongoDB, and Elasticsearch, enabling organizations to continue using the tools they find valuable. Starburst Galaxy also offers data products so that organizations can surface a collection of curated, high-quality related data sets to end users. Data products help to promote the visibility of critical data sets from a single interface.

Figure 2 displays a common data lake architecture used to consolidate and analyze data. Starburst enables users to structure data into reporting layers, providing a way to manage data while preparing it to be consumed by business decision-makers. With Starburst's help, users can create the tables required for analytics. Starburst Galaxy then provides query federation, business intelligence (BI) dashboards, batch processing, and ad-hoc analytics, along with the ability to serve data up to other BI tools.

Figure 2. Starburst Galaxy Data Lake Analytics Platform Architecture



Source: Starburst Data, Inc. and Enterprise Strategy Group, a division of TechTarget, Inc.

Enterprise Strategy Group Tested

Enterprise Strategy Group evaluated how Starburst Galaxy decreases time to insights and simplifies data management for engineers and analysts.

Focus on High-value Analytics Instead of Data Management

Starburst Galaxy serves as a middle layer between BI tools and data storage. The platform provides the computing power that transforms data within the reporting structure and makes it available to analysts, either via the Galaxy query interface or BI tools.

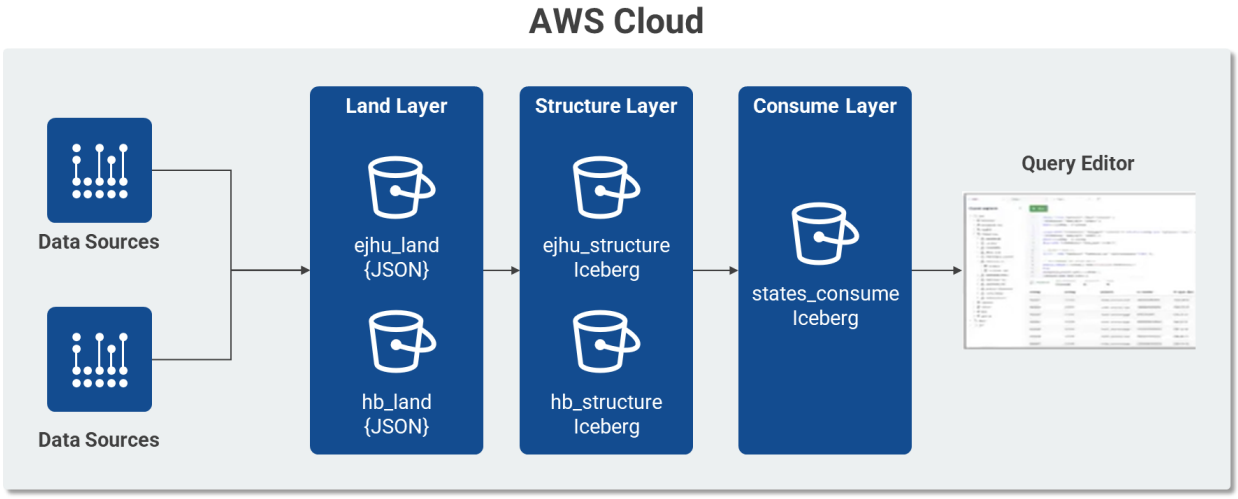
Starburst enables organizations to create a modern data lake architecture, allowing for a single point of access and governance of data. Creating three layers, which Starburst refers to as Land, Structure, and Consume, provides a vendor-agnostic and scalable architecture for data lakes. The Land layer holds raw, unmodified data; the Structure layer joins, enriches, and cleans the data, focusing on what analysts need; and the Consume layer holds the aggregated data ready to be consumed using BI tools.

Enterprise Strategy Group observed Starburst Galaxy being used to build a reporting structure using Amazon S3. Hospital bed usage and confirmed case data gathered during the COVID-19 pandemic was used to provide a reasonable data set for the test.

Figure 3 outlines the data lake architecture created during the test. Working from left to right, we started by using Starburst's connection capabilities to load the data files. Starburst Galaxy's Great Lakes connector can query multiple table formats, including Hive, Delta Lake, and Iceberg.

For this test, we loaded the test data into Amazon S3 buckets. Using Amazon S3 buckets for the Land, Structure, and Consume layers provided superior scalability and ease of use. We created a cluster within Galaxy to provide computing power and then created a catalog to connect to the data. Galaxy can handle both interactive and long-running queries. Using the catalog, we created Land tables and then Structure tables to hold data regarding hospital bed usage and the location of those hospitals. Finally, we created Consume tables to provide aggregated data. Once these tables were created, they were ready to use for reporting or business insights by Starburst Galaxy and other BI tools.

Figure 3. Starburst Galaxy Data Lake Implemented in Amazon S3



Source: Amazon Web Services, Inc. and Enterprise Strategy Group, a division of TechTarget, Inc.

Starburst Galaxy features role-based and attribute-based access controls to ensure proper access to the tables by users, groups, or applications. Organizations can give users direct access to Starburst Galaxy or use an external identity provider to provide single sign-on capabilities. At the time of this writing, Starburst Galaxy supports Okta, Azure Active Directory, Google Workspace, and any provider that supports the SAML protocol.

Conclusion

The growth of data and cloud adoption within businesses has created a new set of challenges. Companies need data to better their customer experiences and make critical business decisions. The cloud provides scalability and a pay-as-you-go structure that saves organizations time and money. However, despite the many benefits of using cloud services for data storage and processing, many companies struggle to find valuable insights from within the data and sources they accumulate over time.

Starburst Galaxy seeks to answer these challenges by providing organizations with a managed analytics platform that offers a holistic view of all the data in the ecosystem, regardless of where it lives. Its Great Lakes connector provides connectivity to many different sources of data and supports multiple table formats, such as Hive, Delta Lake, and Iceberg. Using Starburst Galaxy, organizations don't need to change what they already use but are still able to gather, organize, and analyze data through a modern data lake architecture, allowing for a single point of access and governance. Role and attribute-based access controls allow the right people to view the right data, and setting up users is made easier by integrating with single sign-on providers.

Enterprise Strategy Group validated Starburst Galaxy's process of creating a modern data lake architecture. We saw that Starburst enables companies to create a layered analytics architecture and host it on Amazon S3 to provide scalability and ease of use. Users can query data sources via the Galaxy interface to create data transformation routines and regularly refresh tables used for analysis. Galaxy serves as the computing engine for data analytics as well as a source for BI tools. Enterprise Strategy Group observed Galaxy's ability to quickly set up a robust reporting structure, enabling engineers and analysts to focus on delivering business insights instead of spending large amounts of time and money building complicated transformation functions.

If your organization needs a scalable data lake architecture to provide key business insights, no matter where the data lives, we suggest you take a serious look at Starburst Galaxy, available on AWS marketplace.

About Enterprise Strategy Group
TechTarget's Enterprise Strategy Group provides focused and actionable market intelligence, demand-side research, analyst advisory services, GTM strategy guidance, solution validations, and custom content supporting enterprise technology buying and selling.

contact@esg-global.com
www.esg-global.com