Brought to you by:

Starburst

Modern Data Lakes

for dummies®
A Wiley Brand

Store and process big data

Extract insights with ease

Avoid data swamps

Tom Nats

Starburst Special Edition

About Starburst

For data-driven companies, Starburst offers a full-featured data lake analytics platform, built on open source Trino. Our platform includes the capabilities needed to discover, organize, and consume data without the need for time-consuming and costly migrations. We believe the lake should be the center of gravity, but support accessing data outside the lake when needed. With Starburst, teams can access more complete data, lower the cost of infrastructure, use the tools best suited to their specific needs, and avoid vendor lock-in. Trusted by companies like Comcast, Grubhub, and Priceline, Starburst helps companies make better decisions faster on all their data.

Get started with data lakes: https://www.starburst.io/learn/data-fundamentals/data-lake/

Sign up for a free trial: starburst.io/platform/starburst-galaxy/start/

starburst.io

Modern Data Lakes

Starburst Special Edition

by Tom Nats

for dummies®
A Wiley Brand

Modern Data Lakes For Dummies®, Starburst Special Edition

Published by
John Wiley & Sons, Inc.
111 River St.
Hoboken, NJ 07030-5774
www.wiley.com

For general information on our other products and services, or how to create a custom *For Dummies* book for your business or organization, please contact our Business Development Department in the U.S. at 877-409-4177, contact info@dummies.biz, or visit www.wiley.com/go/custompub. For information about licensing the *For Dummies* brand for products or services, contact BrandedRights&Licenses@Wiley.com.

ISBN 978-1-394-22650-4 (pbk); ISBN 978-1-394-22651-1 (ebk)

Printed in the United States and Great Britain.

Publisher's Acknowledgments

Development Editor: Rachael Chilvers

Project Editor: Pradesh Kumar

Acquisitions Editor: Traci Martin

Editorial Manager: Rev Mengle

Business Development Representative: Matt Cox

IN THIS CHAPTER

- Knowing why to use a modern data lake
- Enjoying the benefits of using a modern data lake

Chapter 1
Exploring a Modern Data Lake

The term "data lake" — a repository that can store vast amounts of data — may elicit different reactions and definitions from different folk in the data analytics space. Those who lived through the Hadoop era are often skeptical of the value data lakes provide an organization. Luckily, technology has advanced, and many issues that data lakes encountered in the past have been resolved. Modifying data, high performance, and, most importantly, data quality and security are now features encompassed in data lakes.

Leveraging a single data store for all your analytical needs without being locked into one vendor is a wish come true for many organizations. After all, data is data — so the concept applies to companies of all sizes and all industries. Having a seemingly unlimited, fully managed storage area is great; however, you must define, transform, catalog, quality check, and structure the data before it can easily be consumed by a variety of technical and non-technical end users.

Why Use a Modern Data Lake?

A modern data lake provides data warehouse functionality without the constraints of legacy Hadoop-based data lakes. Additionally, modern data lakes are open, which alleviates single vendor and technology lock-in when architecting, building, and accessing data in your data lake. Companies leveraging modern data lakes own their data, period.

Table 1-1 summarizes the requirements that a modern data lake fulfills.

TABLE 1-1 How a Modern Data Lake Can Meet Your Requirements

Requirement	Modern Data Lake
Single storage platform and multiple engines	Highly performant object stores with multiple engines reading, writing, and managing the data
Can serve a majority of analytical use cases	Business intelligence (BI) reporting, ad-hoc queries, machine learning/AI model building and serving, and so on
Can accommodate different types of data	From JSON to CSV to Parquet and table formats such as Apache Iceberg and Delta Lake
ANSI SQL support	Fully ANSI SQL compliant supporting a variety of programming languages and most BI tools
Ability to modify individual records	Full DML (Data Manipulation Language — update, delete, merge)
Data quality checks	Constraints such as "not null" and valid values for table columns
High performance	Seconds to millisecond querying is possible using different engines' indexing and caching mechanisms
Efficient joins	With Hadoop, joins between tables were discouraged. Joins are now a common pattern and encouraged in modern data lakes

Requirement	Modern Data Lake
Schema evolution	Changing data structures was challenging in legacy data lakes. Table formats, such as Iceberg, Delta Lake, and Apache Hudi, enable these changes just like a traditional database
Affordable and maintainable	Cloud-based modern data lakes benefit from an economic, fully managed, scalable storage repository

As you can see from Table 1-1, the modern data lake has come a long way. It's an exciting time for companies as they can finally simplify their analytics architecture, avoid vendor lock-in, and provide a single storage layer for all of their diverse data.

After the introduction of Hadoop, companies used the technology to land massive amounts of disparate data from various sources. From there, the data was analyzed in its raw form by data scientists. BI professionals using standard SQL and reporting tools struggled to get value from the unorganized data in the lake. Performance paled in comparison to traditional data warehouses. Once the data had been processed, it was copied to a data warehouse. This increased complexity and cost for companies who now had to maintain two data systems and employ staff with multiple skill sets, as shown in Figure 1-1.

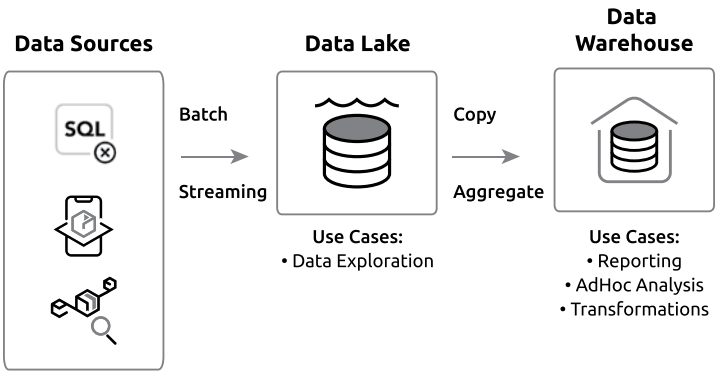

FIGURE 1-1: The bad ol' days of data lakes.

With a modern data lake, also known as a data lakehouse, organizations can serve all of their analytical use cases from a single storage platform; see Figure 1-2. Additionally, you can use different