databricks

eBook

Big Book of Data Warehousing and BI

Contents

SECTION

01

Introduction to the Data Intelligence Platform

Data is vital to the success of every company. As organizations increasingly rely on data, maintaining efficient data management and governance becomes more crucial. Addressing this, the Databricks Data Intelligence Platform enables effective data management, utilization and access to data and AI. Built on the lakehouse architecture, it merges the best features of data lakes and data warehouses, reducing costs and speeding up data and AI initiatives. The platform provides unified governance for data and AI, along with a versatile query engine for ETL, SQL, machine learning and BI.

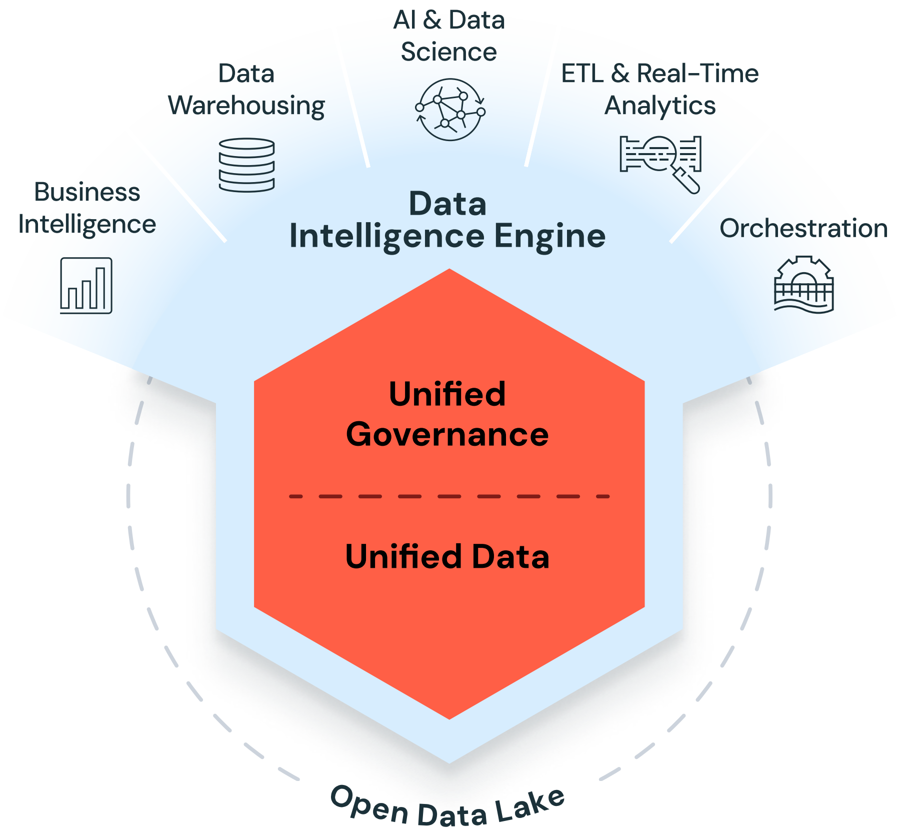
The Databricks Data Intelligence Platform enables organizations to effectively manage, utilize and access all their data and AI. The platform — built on the lakehouse architecture, with a unified governance layer across data and AI and a single unified query engine that spans ETL, SQL, AI and BI — combines the best elements of data lakes and data warehouses to help reduce costs and deliver faster data and AI initiatives.

By combining generative AI with the unification benefits of a lakehouse, the Data Intelligence Platform offers a Data Intelligence Engine called DatabricksIQ that understands the unique semantics of your enterprise data. DatabricksIQ automatically analyzes all aspects of data, including its content, metadata and usage patterns (such as queries, reports and lineage). This comprehensive analysis enables the platform to continually learn, enhance and add new capabilities, optimizing data management and AI applications. Through this deep understanding of data, the Databricks Data Intelligence Platform enables:

- **Natural Language Access:** Leveraging AI models, the Data Intelligence Platform enables working with data in natural language, tailored to each organization's jargon and acronyms. The platform observes how data is used in existing workloads to learn the organization's terms and offers a tailored natural language interface to all users — from nonexperts to data engineers.
- **Semantic Cataloging and Discovery:** Generative AI can understand each organization's data model, metrics and KPIs to offer unparalleled discovery features or automatically identify discrepancies in how data is being used.
- **Automated Management and Optimization:** AI models can optimize data layout, partitioning and indexing based on data usage, reducing the need for manual tuning and knob configuration.
- **Enhanced Governance and Privacy:** Data Intelligence Platforms can

automatically detect, classify and prevent misuse of sensitive data, while simplifying management using natural language.

First-Class Support for AI Workloads: Data Intelligence Platforms can enhance any enterprise AI application by allowing it to connect to the relevant business data and leverage the semantics learned by the platform (metrics, KPIs, etc.) to deliver accurate results. AI application developers no longer have to "hack" intelligence together through brittle prompt engineering.



The Databricks Platform also simplifies the development of enterprise AI applications. The platform makes it easy for enterprises to build AI applications that understand their data. The Databricks Platform offers multiple capabilities to directly integrate enterprise data into AI systems, including:

- End-to-end RAG (retrieval augmented generation) to build high-quality conversational agents on your custom data
- Training custom models either from scratch on an organization's data, or by continued pretraining of existing models such as MPT and Llama 2, to further enhance AI applications with deep understanding of a target domain
- Efficient and secure serverless inference on your enterprise data, with unified governance and quality monitoring functionality
- End-to-end MLOps based on the popular MLflow open source project, with all produced data automatically actionable, tracked and monitorable in the lakehouse

This how-to reference guide showcases data warehousing best practices on the Databricks Data Intelligence Platform through end-to-end use cases from real-world examples. Discover how the platform helps businesses of all sizes translate raw data into actionable data using SQL — from data ingestion to data processing, AI and LLMs, analytics and BI. We'll arm you with reference architectures and code samples so you can explore all aspects of the data lifecycle on the Data Intelligence Platform.

To learn more about Databricks SQL, read our eBook Why the Data Lakehouse Is Your Next Data Warehouse.

SECTION

02

Data Warehousing Built With Intelligence and Automation

Success in lakehouse-based data and AI initiatives hinges on a simplified data architecture, data quality and governance, and scalability and performance. These pillars collectively provide the foundation upon which organizations build their data strategies, guiding them through the intricate maze of modern data management and analytics.

Simplified Data Architecture

The data lakehouse architecture solves the problems of data silos while incorporating the capabilities of a data warehouse. The approach starts by centralizing data within a cloud-based data lake. This foundation, supported by **Delta Lake**, allows analytics and AI use cases to operate on a single data source — reducing storage expenses and streamlining data engineering. The lakehouse architecture integrates a unified governance and security structure with **Unity Catalog**, ensuring granular data control and timely access for respective teams.

The lakehouse architecture's holistic approach encompasses the entire data lifecycle, transformation and impact across various analytics and AI workloads. Because all workloads share the same data while adhering to uniform security and governance policies, you can feel confident knowing you can rely on the accuracy of the data. Functional silos diminish, paving the way for seamless collaboration and, consequently, enhanced productivity in delivering data products.

These are some of the additional benefits of a simplified data architecture:

- Built on open source and standards, a lakehouse removes data silos, simplifying the data landscape and facilitating data and AI operations
- One platform serves integration, storage, processing, governance, sharing, analytics and AI. It offers a unified approach for handling both structured and unstructured data, a complete view of data lineage and provenance, and a consolidated toolset for Python and SQL, notebooks, IDEs, batch, streaming and all primary cloud providers.
- Automated optimization for performance and storage delivers optimal TCO, setting performance benchmarks for data warehousing and AI — including advanced processes like large language models (LLMs)

Data Governance and Quality

Data, being fundamental to organizations, requires stringent quality and governance, especially with increasing volumes and variety. Organizations must prioritize accuracy, reliability and compliance to maximize their lakehouse's potential. Subpar data quality can distort insights, while weak governance can lead to regulatory and security lapses. The Databricks Data Intelligence Platform, with its Unity Catalog, addresses these issues, providing an integrated framework for managing and improving the quality of data across its lifecycle. With governance on the lakehouse architecture, you can:

- Discover, classify and consolidate data and AI assets from various platforms on any cloud, enhancing data exploration and insight extraction using natural language, all from a single access point
- Simplify access management through a unified interface, ensuring consistent and secure access across clouds and platforms, with enhanced fine-grained control and scalable low-code policies
- Harness AI to automate data and ML model monitoring, receive proactive alerts for issues, streamline debugging, and achieve holistic observability of your lakehouse operations using built-in system tables
- Efficiently share data and AI assets across clouds, regions and platforms using open source Delta Sharing in Unity Catalog, enabling secure collaboration and value creation without the need for complex processes or costly replication

Scalability and Performance

With growing data volumes, a lakehouse architecture distributes computing features, independent of storage, aiming to maintain consistent performance at optimal costs. The Databricks Data Intelligence Platform is designed for elasticity, allowing organizations to scale their data operations as needed. Scalability extends across various dimensions:

SERVERLESS

The Databricks Platform utilizes serverless cloud-based computing resources, enabling workloads to adjust and scale elastically based on the required computing capacity. Such dynamic resource allocation guarantees rapid data processing and analysis, even during peak demand.

CONCURRENCY

Leveraging serverless compute and AI-driven optimizations, the Databricks Platform facilitates concurrent data processing and query execution. This ensures that multiple users and teams can undertake analytics tasks concurrently without performance constraints.

STORAGE

The platform seamlessly integrates with data lakes, facilitating the cost-effective storage of extensive data volumes while ensuring data availability and reliability. It also optimizes data storage for enhanced performance, reducing storage expenses.

Scalability, though essential, is complemented by performance. In this regard, the Databricks Data Intelligence Platform stands out, offering a variety of AI-driven optimizations:

OPTIMIZED QUERY PROCESSING

The platform utilizes machine learning optimization techniques to accelerate query execution. It leverages automatic indexing, caching and predicate pushdown to ensure queries are processed efficiently, resulting in rapid insights.

AUTOSCALING

The Databricks Platform intelligently scales serverless resources to match your workloads, ensuring that you pay only for the compute you use, all while maintaining optimal query performance.

PHOTON

The new native massively parallel processing (MPP) engine on the Databricks Platform provides extremely fast query performance at low cost — from data ingestion, ETL, streaming, data science and interactive queries — directly on your data lake. Photon is compatible with Apache Spark™ APIs — no code changes and no lock-in.

DELTA LAKE

Delta Lake with Unity Catalog and Photon offers the best price/performance out of the box without manual tuning. The Databricks Platform uses AI models to solve common challenges with data storage, so you get faster performance without having to manually manage tables, even as they change over time.

- Predictive I/O for updates optimizes your query plans and data layout for peak performance, intelligently balancing read vs. write performance. So you can get more from your data without needing to decide between strategies like copy-on-write vs. merge-on-read.
- Liquid clustering delivers the performance of a well-tuned, well-partitioned table without the traditional headaches that come with partitioning, such as worrying about whether you can partition high-cardinality columns or expensive rewrites when changing partition columns. The result is lightning-fast, well-clustered tables with minimal configuration.
- Predictive optimization automatically optimizes your data for the best performance and price. It learns from your data usage patterns, builds a plan for the right optimizations to perform and then runs those optimizations on hyper-optimized serverless infrastructure.

Having established the foundation of the lakehouse architecture, it's pertinent to explore the delivery of data warehouse and analytics capabilities on Databricks with appropriate data structures and management functionalities facilitated by Databricks SQL (DB SQL).

Databricks SQL Serverless: The best data warehouse for a lakehouse architecture

Databricks SQL was introduced to enhance data warehousing capabilities and offer premier SQL support on the Databricks Data Intelligence Platform. It streamlines SQL-based data transformation, exploration and analysis, catering to users of diverse technical backgrounds. From BI analysts and data architects to analytics engineers, its intuitive SQL interface facilitates queries and complex data operations without necessitating specialized programming. This broadened data access fosters an organization-centric culture, empowering more teams to base decisions on data insights.

Databricks SQL distinguishes itself with its ability to handle massive data sets with speed and efficiency. Utilizing Databricks' next-gen engine, Photon with AI-driven optimizations, ensures rapid data processing and analysis, notably decreasing query execution durations. High performance is crucial for organizations facing data challenges, guaranteeing insights from an extensive variety of data sets. Moreover, Databricks SQL champions collaboration, providing a workspace where data professionals can instantaneously share queries, outcomes and understandings. This shared setting promotes knowledge exchange and hastens resolution, allowing organizations to capitalize on their teams' collective intelligence.

Additionally, Databricks SQL incorporates advanced provisions for data governance, security and compliance, empowering organizations to uphold data quality, implement access restrictions, oversee data activities, safeguard sensitive data and adhere to regulatory standards. To sum up, Databricks SQL provides:

- **Faster Time to Insights**
  Use plain English to access data, and it will automatically create the SQL queries for you. This makes it faster to refine your queries and is available to everyone in the enterprise.
- **Best Price/Performance**
  Serverless compute combined with AI-optimized processing achieves top-tier performance and scale at lower costs, without the need for cloud infrastructure management
- **Unified Governance**
  Establish one unified governance layer across all data and AI assets no matter where they live
- **Reduce Complexity**
  Unify all your data, analytics and AI on one platform that supports SQL and Python, notebooks and IDEs, batch and streaming, and all major cloud providers
- **Rich Ecosystem**
  Utilize SQL and favorite tools such as Power BI, Tableau, dbt and Fivetran with Databricks for BI, data ingestion and transformation

Conclusion

The Databricks Data Intelligence Platform represents a significant advancement in the realm of data warehousing and analytics. It addresses the critical need for efficient data warehousing workloads in today's data-driven business landscape. The platform's simplified data architecture centralizes data with complete end-to-end data warehouse capabilities, leading to cost savings and speeding up time to turn raw data into actionable insights at scale — and unify batch and streaming. Moreover, it ensures data quality and governance through the Unity Catalog, enabling organizations to easily discover, secure and manage all their data with fine-grained governance with data lineage across clouds.

Scalability and performance are key foundational differentiators of the Databricks Platform, with serverless computing, concurrency support and optimized storage strategies. AI-driven optimizations enhance query processing, improve performance and optimize data for best performance and price, making data analysis faster and more cost-effective.

Databricks SQL further enhances the platform's capabilities by providing premier SQL support, facilitating data transformation and analysis for many users. It promotes collaboration, data governance and a rich ecosystem of tools, breaking down data silos and enabling your organization to harness the full potential of your data. Now, let's take a look at a few use cases for running your data warehousing and BI workloads on the Databricks Platform.

LEARN MORE

→ Why the Data Lakehouse Is Your Next Data Warehouse: 2nd Edition

→ What's new in Databricks SQL?

→ Introducing Lakehouse Federation Capabilities in Unity Catalog

→ Introducing AI Functions: Integrating Large Language Models with Databricks SQL

SECTION

03 The Next Wave of Business Intelligence

The Next Wave of Business Intelligence

Introduction to Databricks AI/BI

For Databricks customers, AI/BI represents a significant advancement in the area of BI by integrating AI directly into the analytics process. It allows users to ask questions in natural language and receive trusted, AI-generated insights. What sets this system apart is its native integration within the Databricks Data Intelligence Platform, which provides a unified environment for analytics, data management and governance. This tight integration offers distinct advantages over traditional, siloed BI systems that often operate separately from the underlying data platforms. By learning from the entire data estate, Databricks AI/BI delivers accurate insights that reflect both the complexity and the specific context of the organization's data.

Moving Beyond Traditional BI Systems

Traditional BI tools typically operate on static datasets or predefined reports, often housed in disconnected systems. This fragmented approach leads to inefficiencies, with challenges such as duplicated data and governance efforts and limited real-time analytics, resulting in slower time to insight. Databricks AI/BI circumvents these issues by embedding AI capabilities directly within the platform's native architecture.

Key benefits of this native integration include:

- A single, trusted source of data for all analytics, eliminating the need for maintaining separate BI systems
- Unified governance and security through Unity Catalog, ensuring consistent data access policies and robust compliance mechanisms
- High-performance querying and real-time insights are made possible through the seamless use of Databricks SQL

This approach streamlines operations, reduces data inconsistencies and allows organizations to scale their analytics capabilities without the typical technical bottlenecks seen in siloed BI systems.

AI/BI Dashboards: A New Paradigm for Data Analysts

Databricks AI/BI Dashboards provide a powerful environment where data analysts can leverage Databricks Assistant to streamline the development of analytical datasets, optimize SQL queries and leverage natural language to build effective visualizations. By integrating AI directly into the dashboard creation process, AI/BI enables analysts to develop complex dashboards faster and with greater accuracy. Whether it's improving SQL efficiency or automatically generating data visualizations from natural language queries, AI/BI Dashboards significantly reduce the time and effort required to create insightful analytics.

Notable features of AI/BI Dashboards include:

- **AI-Assisted BI Development:** By helping analysts optimize SQL, fix code errors and build visualizations using natural language, AI accelerates dashboard development, enabling faster decision-making and reducing manual work
- **Customizable Visualizations:** A wide range of visualization types allows users to present data in ways that are most relevant to their business context
- **Interactive Filters and Query-Based Parameters:** These features enable more granular analysis, giving users the ability to adjust their reports in real time to reflect specific business questions
- **Real-Time Data Exploration:** Dashboards operate on live data, ensuring that the insights reflect current business conditions

AI/BI Genie: The Future of Self-Service Analytics

The introduction of AI/BI Genie — a conversational AI analyst — is one of the most transformative elements of Databricks AI/BI. Designed to bridge the gap between nontechnical business users and complex data systems, Genie allows users to ask questions about their enterprise data in natural language and receive immediate, contextually relevant answers. This generative AI capability goes beyond basic query translation by learning from user interactions, continuously refining its understanding of business data and context.

Genie's ability to understand the full data estate allows it to return highly accurate results, even when dealing with complex queries. This enhanced understanding is referred to as *data intelligence*. Unlike many AI models that might generate incorrect or ambiguous answers, Genie also proactively seeks clarification from the user in instances of uncertainty, learning from feedback to ensure the integrity of the insights it delivers.

The Role of Generative AI in Enabling Self-Service Analytics

The power of generative AI within Databricks AI/BI lies in its capacity to extend self-service analytics to nontechnical users. Traditional BI tools often require users to possess technical skills, particularly in querying data through SQL or navigating predefined reports. Genie's natural language interface changes this dynamic by enabling business users to directly engage with data.

However, Genie's effectiveness isn't just due to its conversational interface — its awareness of the entire data platform is what truly sets it apart. Because Genie learns from the organization's data ecosystem, including usage patterns and business semantics, it can deliver insights with a high degree of accuracy, ensuring that even complex data queries are handled with precision.

Governance and Security: A Unified Approach

Databricks AI/BI inherits the governance and security capabilities of the Databricks Platform, particularly through Unity Catalog. This native integration ensures that all analytics are governed by consistent access controls and security policies. By operating within a unified platform, organizations benefit from:

- Consistent governance, as all data, from raw datasets to final visualizations, is subject to the same access policies
- Simplified compliance, as the platform supports auditing, lineage tracking and role-based access controls
- A reduction in risks associated with data silos, as all analytics activities occur within the same governed environment

By combining AI-driven analytics with robust governance, Databricks AI/BI provides users with flexibility and administrators with confidence in maintaining data security and integrity.

Best Practices for Maximizing the Potential of AI/BI

Organizations looking to maximize the potential of Databricks AI/BI should focus on establishing a culture of data literacy and self-service analytics. This can be achieved through:

- Training business users to effectively interact with Genie, ensuring they can independently query data and understand the results
- Leveraging AI/BI Dashboards to monitor critical business metrics, with regular reviews to adjust dashboards to reflect evolving business needs
- Utilizing Unity Catalog to enforce data governance across all departments, ensuring compliance and data security while providing seamless access to insights

For data practitioners wanting to create effective Genie spaces in Databricks AI/BI, our advice is to start small and iteratively expand based on feedback, using focused tables and well-annotated columns. Involve domain experts, define the audience and purpose, and thoroughly test with sample questions to ensure Genie interprets domain-specific language. Conduct user testing for continuous refinement, and address issues such as ambiguous jargon, table selection and performance bottlenecks. For details on troubleshooting and additional strategies, please see the full guidelines on Genie best practices.

Real-World Applications of AI/BI

Databricks AI/BI is already driving innovation across several industries, including:

- **Retail and Consumer Goods:** AI/BI enables retailers to optimize inventory levels by predicting consumer behavior and demand fluctuations in real time
- **Healthcare and Life Sciences:** Healthcare providers are leveraging AI/BI's capabilities to analyze patient data, enhancing diagnostics and improving treatment outcomes
- **Financial Services:** In the financial sector, AI/BI is being used to assess risk more effectively and detect fraud, thanks to its ability to analyze vast datasets with high accuracy

Examples like these highlight how AI/BI's deep integration with Databricks, combined with its AI-first approach, is transforming industries by enabling faster, more accurate decision-making.

The Future of AI-Driven Business Intelligence

As data continues to grow in volume and complexity, AI-first BI tools like Databricks AI/BI are poised to lead the future of analytics. By combining AI-powered insights with seamless integration into the data platform, Databricks AI/BI ensures that organizations are equipped to handle the demands of modern data environments. The evolution of self-service analytics, driven by generative AI models, will continue to push the boundaries of what business users can achieve, empowering them to uncover actionable insights without the need for technical expertise.

LEARN MORE

- Introducing AI/BI: Intelligent Analytics for Real-World Data
- Onboarding Your New AI/BI Genie
- Building Confidence in Your Genie Space With Benchmarks and Ask for Review
- How to Share AI/BI Dashboards With Everyone in Your Organization
- How to Embed AI/BI Dashboards Into Your Websites and Applications

SECTION

04

Data Warehousing Best Practices on the Lakehouse

SECTION 4.1

Data Warehousing Modeling Techniques and Their Implementation on the Lakehouse

Using Data Vaults and Star Schemas on the Lakehouse

by Soham Bhatt and Deepak Sekar

The lakehouse is a new data platform paradigm that combines the best features of data lakes and data warehouses. It is designed as a large-scale enterprise-level data platform that can house many use cases and data products. It can serve as a single unified enterprise data repository for all of your:

- data domains,
- real-time streaming use cases,
- data marts,
- disparate data warehouses,
- data science feature stores and data science sandboxes, and
- departmental self-service analytics sandboxes.

Given the variety of the use cases, different data organizing principles and modeling techniques may apply to different projects on a lakehouse. Technically, the lakehouse architecture can support many different data modeling styles. In this article, we aim to explain the implementation of the Bronze/Silver/Gold data organizing principles of the lakehouse and how different data modeling techniques fit in each layer.

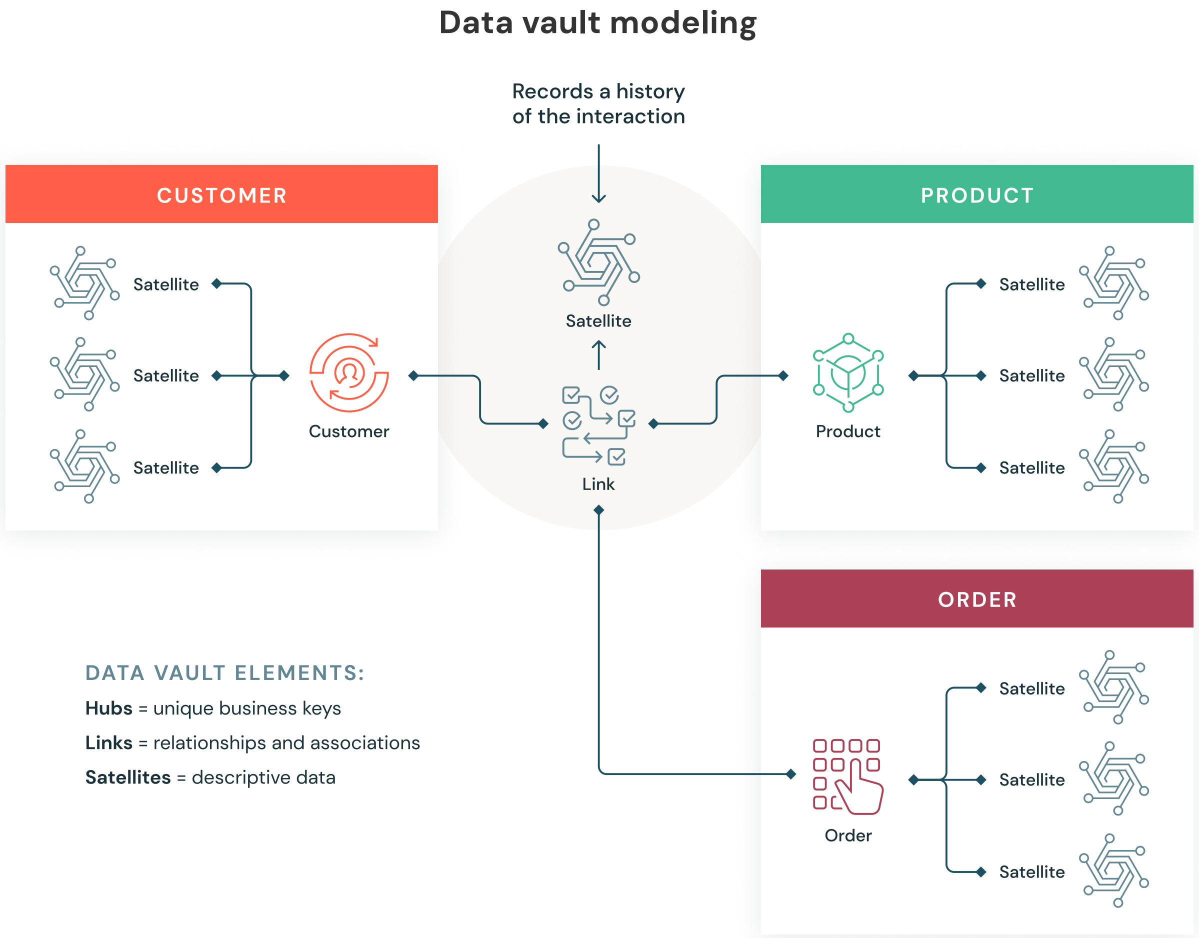
What is a Data Vault?

A Data Vault is a more recent data modeling design pattern used to build data warehouses for enterprise-scale analytics compared to Kimball and Inmon methods.

Data Vaults organize data into three different types: hubs, links, and satellites. Hubs represent core business entities, links represent relationships between hubs, and satellites store attributes about hubs or links.

Data Vault focuses on agile data warehouse development where scalability, data integration/ETL and development speed are important. Most customers have a landing zone, Vault zone and a data mart zone which correspond to the Databricks organizational paradigms of Bronze, Silver and Gold layers. The Data Vault modeling style of hub, link and satellite tables typically fits well in the Silver layer of the lakehouse architecture.

Learn more about Data Vault modeling at Data Vault Alliance.



DATA VAULT ELEMENTS:

Hubs = unique business keys

Links = relationships and associations

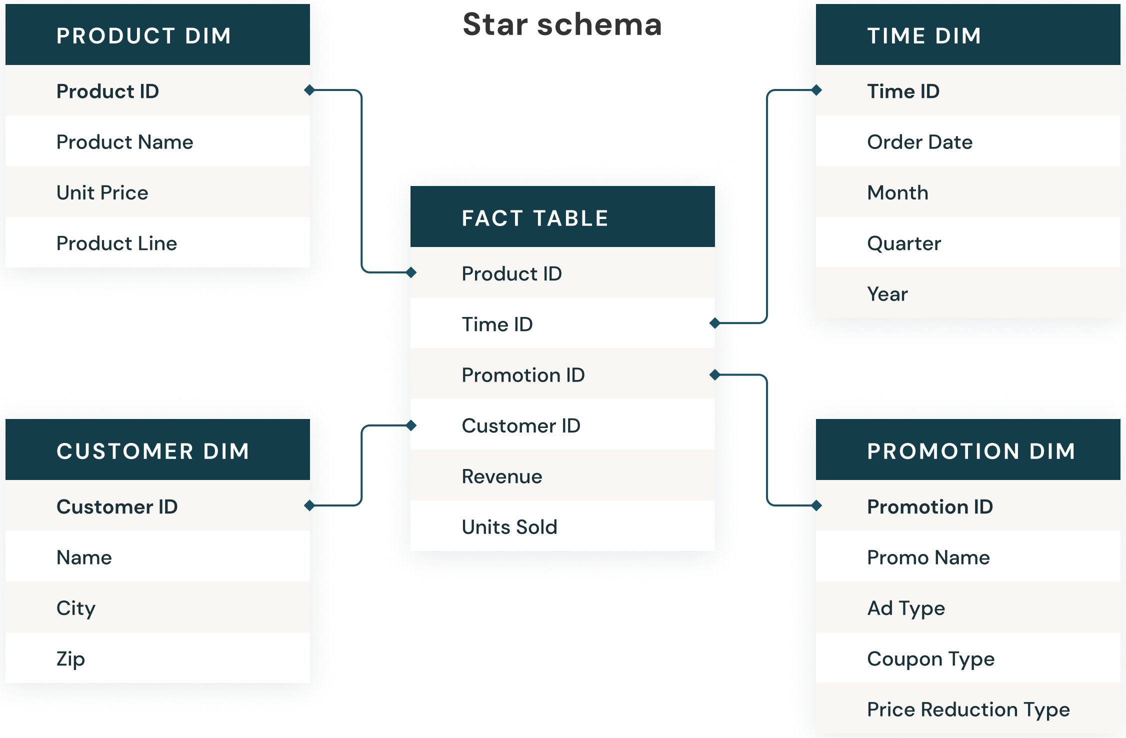
Satellites = descriptive data

A diagram showing how Data Vault modeling works, with hubs, links, and satellites connecting to one another.

What is Dimensional Modeling?

Dimensional modeling is a bottom-up approach to designing data warehouses in order to optimize them for analytics. Dimensional models are used to denormalize business data into dimensions (like time and product) and facts (like transactions in amounts and quantities), and different subject areas are connected via conformed dimensions to navigate to different fact tables.

The most common form of dimensional modeling is the star schema. A star schema is a multi-dimensional data model used to organize data so that it is easy to understand and analyze, and very easy and intuitive to run reports on. Kimball-style star schemas or dimensional models are pretty much the gold standard for the presentation layer in data warehouses and data marts, and even semantic and reporting layers. The star schema design is optimized for querying large data sets.



PRODUCT DIM
Product ID
Product Name
Unit Price
Product Line

TIME DIM
Time ID
Order Date
Month
Quarter
Year

FACT TABLE
Product ID
Time ID
Promotion ID
Customer ID
Revenue
Units Sold

CUSTOMER DIM
Customer ID
Name
City
Zip

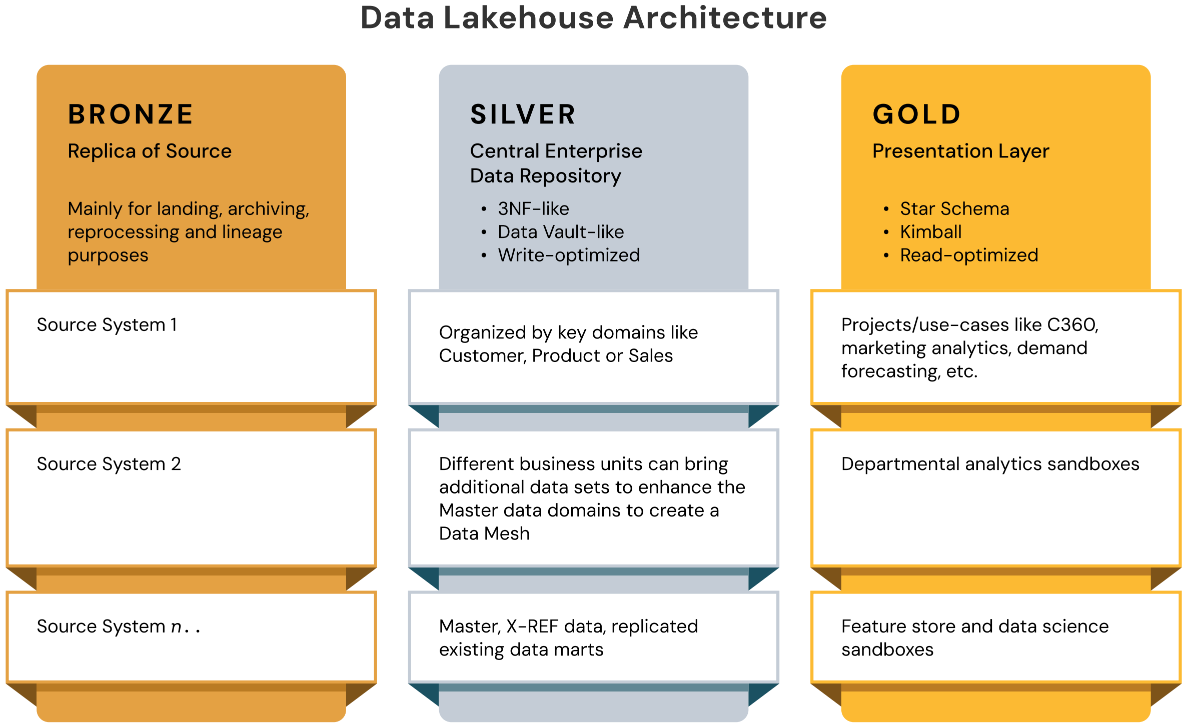
PROMOTION DIM
Promotion ID
Promo Name
Ad Type
Coupon Type
Price Reduction Type

A star schema example

Both normalized Data Vault (write-optimized) and denormalized dimensional models (read-optimized) data modeling styles have a place in the Databricks Data Intelligence Platform. The Data Vault's hubs and satellites in the Silver layer are used to load the dimensions in the star schema, and the Data Vault's link tables become the key driving tables to load the fact tables in the dimension model. Learn more about dimensional modeling from the Kimball Group.

Data organization principles in each layer of the lakehouse

A modern lakehouse is an all-encompassing enterprise-level data platform. It is highly scalable and performant for all kinds of different use cases such as ETL, BI, data science and streaming that may require different data modeling approaches. Let's see how a typical lakehouse is organized:


Data Lakehouse Architecture
BRONZE
Replica of Source
Mainly for landing, archiving, reprocessing and lineage purposes
Source System 1
Source System 2
Source System n . .
SILVER
Central Enterprise Data Repository
• 3NF-like
• Data Vault-like
• Write-optimized
Organized by key domains like Customer, Product or Sales
Different business units can bring additional data sets to enhance the Master data domains to create a Data Mesh
Master, X-REF data, replicated existing data marts
GOLD
Presentation Layer
• Star Schema
• Kimball
• Read-optimized
Projects/use-cases like C360, marketing analytics, demand forecasting, etc.
Departmental analytics sandboxes
Feature store and data science sandboxes

A diagram showing characteristics of the Bronze, Silver and Gold layers of the data lakehouse architecture.

Bronze Layer — the Landing Zone

The Bronze layer is where we land all the data from source systems. The table structures in this layer correspond to the source system table structures "as-is," aside from optional metadata columns that can be added to capture the load date/time, process ID, etc. The focus in this layer is on change data capture (CDC), and the ability to provide an historical archive of source data (cold storage), data lineage, auditability, and reprocessing if needed — without rereading the data from the source system.

In most cases, it's a good idea to keep the data in the Bronze layer in Delta format, so that subsequent reads from the Bronze layer for ETL are performant — and so that you can do updates in Bronze to write CDC changes. Sometimes, when data arrives in JSON or XML formats, we do see customers landing it in the original source data format and then stage it by changing it to Delta format. So sometimes, we see customers manifest the logical Bronze layer into a physical landing and staging zone.

Storing raw data in the original source data format in a landing zone also helps with consistency wherein you ingest data via ingestion tools that don't support Delta as a native sink or where source systems dump data onto object stores directly. This pattern also aligns well with the autoloader ingestion framework wherein sources land the data in landing zone for raw files and then Databricks AutoLoader converts the data to Staging layer in Delta format.

Silver Layer — the Enterprise Central Repository

In the Silver layer of the lakehouse architecture, the data from the Bronze layer is matched, merged, conformed and cleaned (“just-enough”) so that the Silver layer can provide an “enterprise view” of all its key business entities, concepts and transactions. This is akin to an Enterprise Operational Data Store (ODS) or a Central Repository or Data domains of a Data Mesh (e.g. master customers, products, non-duplicated transactions and cross-reference tables). This enterprise view brings the data from different sources together, and enables self-service analytics for ad-hoc reporting, advanced analytics and ML. It also serves as a source for departmental analysts, data engineers and data scientists to further create data projects and analysis to answer business problems via enterprise and departmental data projects in the Gold layer.

In the lakehouse data engineering paradigm, typically the (Extract-Load-Transform) ELT methodology is followed vs. traditional Extract-Transform-Load(ETL). ELT approach means only minimal or “just-enough” transformations and data cleansing rules are applied while loading the Silver layer. All the “enterprise level” rules are applied in the Silver layer vs. project-specific transformational rules, which are applied in the Gold layer. Speed and agility to ingest and deliver the data in the lakehouse is prioritized here.

From a data modeling perspective, the Silver layer has more 3rd-Normal Form like data models. Data Vault-like write-performant data architectures and data models can be used in this layer. If using a Data Vault methodology, both the raw Data Vault and Business Vault will fit in the logical Silver layer of the lake — and the Point-In-Time (PIT) presentation views or materialized views will be presented in the Gold layer.

Gold Layer — the Presentation Layer

In the Gold layer, multiple data marts or warehouses can be built as per dimensional modeling/Kimball methodology. As discussed earlier, the Gold layer is for reporting and uses more denormalized and read-optimized data models with fewer joins compared to the Silver layer. Sometimes tables in the Gold layer can be completely denormalized, typically if the data scientists want it that way to feed their algorithms for feature engineering.

ETL and data quality rules that are “project-specific” are applied when transforming data from the Silver layer to Gold layer. Final presentation layers such as data warehouses, data marts or data products like customer analytics, product/quality analytics, inventory analytics, customer segmentation, product recommendations, marketing/sales analytics, etc., are delivered in this layer. Kimball style star-schema based data models or Inmon style Data marts fit in this Gold layer of the lakehouse. Data Science Laboratories and Departmental Sandboxes for self-service analytics also belong in the Gold layer.