

Data Warehouse Definition

Data Warehouse: Perspectives from Bill Inmon and Ralph Kimball

In the realm of data warehousing, two prominent figures have shaped the field with their distinct perspectives and contributions: Bill Inmon and Ralph Kimball.

Their approaches to designing and implementing data warehouses have significantly influenced the way organizations manage and utilize their data assets.

Let's explore their definitions and understand their importance in the data warehousing world.

Definition according to Bill Inmon

Bill Inmon, often referred to as the "father of data warehousing," advocates for the top-down approach.

According to Inmon, a data warehouse is a **"subject-oriented, integrated, time-variant, and nonvolatile collection of data that serves as a single version of truth for the organization"**.

The Data Ocean fully embraces and aligns with this definition for the **Domain** definition, recognizing it as a perfect representation of its comprehensive, accurate, and unambiguous perspective. Additionally, it seeks to extend this perspective by introducing a new dimension identified as "Data-Oriented," which further enhances its understanding and utilization of data assets.

In detail:

- **Subject-oriented:**
 - The Domain data layer focuses on organizing and representing data based on specific subjects or business domains within the organization.
 - It captures the unique characteristics and requirements of each domain, enabling a more granular and targeted approach to data management.
- **Integrated:**
 - The Domain data layer integrates data from various sources, systems, and applications across the organization.
 - It provides a unified view of the data within a domain, eliminating data silos and enabling comprehensive analysis and reporting.
- **Time-variant:**
 - The Domain data layer captures and retains historical data, allowing for tracking and analysis of changes over time.
 - It enables temporal data modeling, ensuring that the evolution and historical context of data are preserved and accessible for retrospective analysis.
- **Nonvolatile:**
 - The Domain data layer stores data in a persistent and nonvolatile manner, ensuring that it remains accessible and unchanged over an extended period.
 - It supports long-term data retention, providing a reliable and consistent data source for decision-making and analysis.
- **Data-oriented:**
 - The Domain data layer places a strong emphasis on the data itself as the primary driver of its design and structure.
 - It ensures that data elements are accurately defined, categorized, and related to reflect the real-world entities and concepts within each domain and driven by the concepts (the entities) in the business source systems
 - The Domain data layer is primarily driven by the concepts, or entities, in the business source systems.
 - The Domain data layer is intricately aligned with the data representation and definition in the source systems, making it resilient to changes in volatile requirements.
 - This alignment with the source systems' data representation means that the Domain data layer only needs to be modified when there are changes in the source systems themselves.
 - It remains independent of specific and volatile requirements that may evolve over time.
 - By focusing on the data itself, the Domain data layer establishes a solid foundation that can adapt to changes in the business landscape while maintaining consistency and integrity.
 - The resilience of the Domain data layer lies in its ability to provide a stable and reliable representation of the data, regardless of external influences or evolving requirements. It serves as a robust and enduring structure that supports the organization's data management needs and facilitates effective data-driven decision-making.

Inmon emphasizes the importance of capturing and storing data from various operational systems in a structured and consistent manner. His approach focuses on building a centralized and reliable data repository that can support a wide range of reporting and analytical requirements.

Inmon's perspective places an emphasis on data integration, data quality, and data governance to ensure the reliability and accuracy of the data warehouse.

Definition according to Ralph Kimball

On the other hand, Ralph Kimball is widely acknowledged for his significant contribution to the bottom-up approach in data warehousing. He holds a prominent position as one of the most influential figures in the Data Warehousing movement, particularly through his renowned Data Modeling Methodology.

"A data warehouse is a copy of transaction data specifically structured for query and analysis, enabling decision-makers to gain insights and make informed business decisions."

The Data Ocean highly endorses and supports this concept for modeling the Data Products and believes that the Dimensional Modeling technique is the ideal option given its focus in performance and simplicity.

Key elements of this definition:

- Copy of Transaction Data:
 - A data warehouse is derived from operational systems and transactional databases.
 - It involves extracting, transforming, and loading (ETL) data from these sources into a separate environment specifically designed for analytical purposes.
 - The data warehouse is not the primary system of record but rather a replica of the transactional data.
- Structured for Query and Analysis:
 - The data in a data warehouse is organized and structured in a way that facilitates efficient querying and analysis. It involves designing dimensional models such as star schemas or snowflake schemas, which consist of fact tables representing business events or transactions and associated dimension tables providing context and attributes for analysis. This structure enables users to easily navigate and explore the data for reporting and decision-making purposes.

Kimball's definition aligns with his emphasis on designing data warehouses that are optimized for query performance and user accessibility. It highlights the purpose of a data warehouse as a repository of transactional data that is transformed and structured in a way that facilitates efficient querying, analysis, and decision-making. He believes in the importance of user-driven requirements and building data warehouses that are optimized for query performance and ease of use. His perspective values user accessibility, simplicity, and agility in delivering data to business users for decision-making purposes.

Kimball's approach emphasizes dimensional modeling and advocates for designing data marts focused on specific business areas or processes. The dimensional modeling approach focuses on creating star schemas or snowflake schemas that provide a user-friendly structure for reporting and analysis. The data warehouse serves as a reliable and consolidated source of data for users to explore, understand, and derive insights from the underlying business events or transactions.

Kimball's approach promotes a more flexible and iterative development process, where data marts are incrementally built to address specific business needs.

In summary, Ralph Kimball's definition of a data warehouse highlights its role as a structured and accessible copy of transaction data, designed specifically for querying and analysis. It underscores the importance of enabling decision-makers to gain insights and make informed business decisions based on the data stored within the warehouse.

Conclusion

Both Bill Inmon and Ralph Kimball have played instrumental roles in shaping the data warehousing landscape.

Inmon's top-down approach highlights the need for a centralized, reliable, and authoritative data source, while Kimball's bottom-up approach emphasizes user-driven requirements and flexibility.

Their perspectives have provided organizations with different options and considerations when designing and implementing data warehousing solutions, allowing them to tailor their approach based on their specific needs and priorities. The collective contributions of Inmon and Kimball have significantly advanced the field of data warehousing and continue to influence the practices and methodologies used in the industry today.