Data warehousing concepts

Concise definitions, comparisons, and examples

Author: Morales, J.A.
Last update: January 13, 2023

Contents

Introduction

Concise definitions, comparisons, and examples of data warehousing concepts.

Data warehouse

A data warehouse is a data hub – an integrated storage of all business data including operational, historical, and external, from multiple systems. It serves as a comprehensive repository of aggregated data from multiple systems.

Data warehouse vs data lake vs database

Data warehouse. A data warehouse assembles data from multiple sources including databases. It is built to enable rapid analysis of large and multi-dimensional datasets. It allows users to store and access structured and semi-structured data. Data warehouse is sometimes referred to as enterprise data warehouse (EDW) in business settings.

Data lake. A data lake is a repository for raw data and tend to serve many purp**o**ses. It is built to hold back-up data and raw data. contains structured, semi-structured, and unstructured data. It allows users to store and access structured, semi-structured, and unstructured data. Data lake is sometimes referred to as enterprise data lake (EDL) in business settings.

Database. A database is an organized collection of data, generally stored in tables, and kept online. It is built to enable ingestion of large amounts of data efficiently. In the past it allowed users to store and access structured data only. Today, there are databases that allow users to store and access structured and semi-structured data. Database is sometimes abbreviated DB.

Datamart

A datamart is a subset of a data warehouse and is usually oriented to a specific team or department. We can think of a datamart as a small data warehouse with predefined scope, justified for tangible benefits from critical business-reporting needs, and has a lower development risk.

Dimensional modeling

A dimensional model is a conceptual representation of analytic data consisting of facts or measures and dimensions or descriptions.

Star schema

The star schema is a dimensional modeling approach made up of one fact table connected to multiple dimension tables. It enables denormalization of the dimensions. It makes querying simpler and aggregations faster. When sketched, it resembles a star figure with the fact table at the centre and dimension tables at the edges.

Snowflake schema

The snowflake schema is a dimensional modeling approach made up of one fact table connected to multiple dimension tables which are also connected to more dimension tables. It enables us to normalize the dimensions. It allows for more nuanced querying and efficient storage. When sketched, it

resembles a snowflake figure with the fact table at the centre and dimension tables sprawling at the edges.

…more coming soon.

References

Google (n.d.) Innovating with data and Google Cloud [Training module]. Google Cloud Digital Leader Training. (Accessed: 1 August 2022)

Han, J., Kamber, M., and Pei, J. (2012) *Data mining concepts and techniques*. Waltham, MA: Morgan Kaufmann Publishers.

Kimball, R., and Ross, M. (2013) *The data warehouse toolkit: The definitive guide to dimensional modeling*. 3rd edition. Indianapolis, IN: John Wiley & Sons, Inc.

Kopanakis, I. (2017) Data warehousing and visualization [Recorded lecture]. The University of Liverpool. (Accessed: 24 October 2019).

Simsion, G.C. (2001) *Data modeling essentials: Analysis, design, and innovation*. 2nd edition. Scottsdale, Arizona: The Coriolis Group, LLC.