I'm not robot

reCAPTCHA

Continue

I'm not robot

reCAPTCHA

Continue

What data warehousing allow organizations to achieve

Data warehouse overview The basic architecture of a data warehouse In computing, a data warehouse (DW or DWH), also known as an enterprise data warehouse (EDW), is a system used for reporting and data analysis and is considered a core component of business intelligence.[1] DWs are central repositories of integrated data from one or more disparate sources. They store current and historical data in one single place[2] that are used for creating analytical reports for workers throughout the enterprise.[3] The data stored in the warehouse is uploaded from the operational systems (such as marketing or sales). The data may pass through an operational data store and may require data cleansing[2] for additional operations to ensure data quality before it is used in the DW for reporting. Extract, transform, load (ETL) and extract, load, transform (ELT) are the two main approaches used to build a data warehouse system. ETL-based data warehousing The typical extract, transform, load (ETL)-based data warehouse[4] uses staging, data integration, and access layers to house its key functions. The staging layer or staging database stores raw data extracted from each of the disparate source data systems. The integration layer integrates the disparate data sets by transforming the data from the staging layer often storing this transformed data in an operational data store (ODS) database. The integrated data are then moved to yet another database, often called the data warehouse database, where the data is arranged into hierarchical groups, often called dimensions, and into facts and aggregate facts. The combination of facts and dimensions is sometimes called a star schema. The access layer helps users retrieve data.[5] The main source of the data is cleansed, transformed, catalogued, and made available for use by managers and other business professionals for data mining, online analytical processing, market research and decision support.[6] However, the means to retrieve and analyze data, to extract, transform, and load data, and to manage the data dictionary are also considered essential components of a data warehousing system. Many references to data warehousing use this broader context. Thus, an expanded definition for data warehousing includes business intelligence tools, tools to extract, transform, and load data into the repository, and tools to manage and retrieve metadata. IBM InfoSphere DataStage, Ab Initio Software, Informatica – PowerCenter are some of the tools which are widely used to implement ETL-based data warehouse. ELT-based data warehousing ELT-based Data Warehouse architecture ELT-based data warehousing gets rid of a separate ETL tool for data transformation. Instead, it maintains a staging area inside the data warehouse itself. In this approach, data gets extracted from heterogeneous source systems and are then directly loaded into the data warehouse, before any transformation occurs. All necessary transformations are then handled inside the data warehouse itself. Finally, the manipulated data gets loaded into target tables in the same data warehouse. Benefits A data warehouse maintains a copy of information from the source transaction systems. This architectural complexity provides the opportunity to: Integrate data from multiple sources into a single database and data model. More congregation of data to single database so a single query engine can be used to present data in an ODS. Mitigate the problem of database isolation level lock contention in transaction processing systems caused by attempts to run large, long-running analysis queries in transaction processing databases. Maintain data history, even if the source transaction systems do not. Integrate data from multiple source systems, enabling a central view across the enterprise. This benefit is always valuable, but particularly so when the organization has grown by merger. Improve data quality, by providing consistent codes and descriptions, flagging or even fixing bad data. Present the organization's information consistently. Provide a single common data model for all data of interest regardless of the data's source. Restructure the data so that it makes sense to the business users. Restructure the data so that it delivers excellent query performance, even for complex analytic queries, without impacting the operational systems. Add value to operational business applications, notably customer relationship management (CRM) systems. Make decision–support queries easier to write. Organize and disambiguate repetitive data Generic The environment for data warehouses and marts includes the following: Source systems that provide data to the warehouse or mart; Data integration technology and processes that are needed to prepare the data for use; Different architectures for storing data in an organization's data warehouse or data marts; Different tools and applications for a variety of users; Metadata, data quality, and governance processes must be in place to ensure that the warehouse or mart meets its purposes. In regards to source systems listed above, R. Kelly Rainer states, "A common source for the data in data warehouses is the company's operational databases, which can be relational databases".[7] Regarding data integration, Rainer states, "It is necessary to extract data from source systems, transform them, and load them into a data mart or warehouse".[7] Rainer discusses storing data in an organization's data warehouse or data marts.[7] Metadata is data about data. "IT personnel need information about data sources; database, table, and column names; refresh schedules; and data usage measures".[7] Today, the most successful companies are those that can respond quickly and flexibly to market changes and opportunities. A key to this response is the effective and efficient use of data and information by analysts and managers.[7] A "data warehouse" is a repository of historical data that is organized by the subject to support decision-makers in the organization.[7] Once data is stored in a data mart or warehouse, it can be accessed. Related systems (data mart, OLAPS, OLTP, predictive analytics) A data mart is a simple form of a data warehouse that is focused on a single subject (or functional area), hence they draw data from a limited number of sources such as sales, finance or marketing. Data marts are often built and controlled by a single department within an organization. The sources could be internal operational systems, a central data warehouse, or external data.[8] Denormalization is the norm for data modeling techniques in this system. Given that data marts generally cover only a subset of the data contained in a data warehouse, they are often easier and faster to implement. Difference between data warehouse and data mart Attribute Data warehouse Data mart Scope of the data enterprise-wide department-wide Number of subject areas multiple single How difficult to build difficult easy How much time takes to build more less Amount of memory larger limited Types of data warehousing: 1960s – General Mills and Dartmouth College, in a joint research project, develop the terms dimensions and facts.[11] 1970s – ACNielsen and IRI provide dimensional data marts for retail sales.[11] 1970s – Bill Inmon begins to define and discuss the term Data Warehouse.[clarification needed] 1975 – Sperry Univac introduces MAPPER (MAintain, Prepare, and Produce Executive Reports), a database management and reporting system that includes the world's first 4GL. It is the first platform designed for building Information Centers (a forerunner of contemporary data warehouse technology). 1983 – Teradata introduces the DBC/1012 database computer specifically designed for decision support.[12] 1984 – Metaphor Computer Systems, founded by David Liddle and Don Massaro, releases a hardware/software package and GUI for business users to create a database management and analytics system. 1985 – Sperry Corporation publishes an article (Martyn Jones and Philip Newman) on information centers, where they introduce the term MAPPER data warehouse in the context of information centers. 1988 – Barry Devlin and Paul Murphy publish the article "An architecture for a business and information system" where they introduce the term "business data warehouse".[13] 1990 – Red Brick Systems, founded by Ralph Kimball, introduces Red Brick Warehouse, a database management system specifically for data warehousing. 1991 – James M. Kerr authors The IRM Imperative, which suggests data resources could be reported as an asset on a balance sheet, furthering commercial interest in the establishment of data warehouses. 1991 – Prism Solutions, founded by Bill Inmon, introduces Prism Warehouse Manager, software for developing a data warehouse. 1992 – Bill Inmon publishes the book Building the Data Warehouse.[14] 1995 – The Data Warehousing Institute, a for-profit organization that promotes data warehousing, is founded. 1996 – Ralph Kimball publishes the book The Data Warehouse Toolkit.[15] 2000 – Dan Linstedt releases in the public domain the Data vault modeling, conceived in 1990 as an alternative to Inmon and Kimball to provide long-term historical storage of data coming in from multiple operational systems, with emphasis on tracing, auditing and resilience to change of the source data model. 2008 – Bill Inmon, along with Derek Strauss and Genia Neushloss, publishes "DW 2.0: The Architecture for the Next Generation of Data Warehousing", explaining his top-down approach to data warehousing and coining the term, data-warehousing 2.0. 2012 – Bill Inmon develops and makes public technology known as "textual disambiguation". Textual disambiguation applies context to raw text and reformats the raw text and context into a standard data base format. Once raw text is passed through textual disambiguation, it can easily and efficiently be accessed and analyzed by standard business intelligence technology. Textual disambiguation is accomplished through the execution of textual ETL. Textual disambiguation is useful wherever raw text is found, such as in documents, Hadoop, email, and so forth. Information storage Facts A fact is a value, or measurement, which represents a fact about the managed entity or system. Facts, as reported by the reporting entity, are said to be at raw level; e.g., in a mobile telephone system, if a BTS (base transceiver station) receives 1,000 requests for traffic channel allocation, allocates for 820, and rejects the remaining, it would report three facts or measurements to a management system: tch_req_total = 1000 tch_req_success = 820 tch_req_fail = 180 Facts at the raw level are further aggregated to higher levels in various dimensions to extract more service or business-relevant information from it. These are called aggregates or summaries or aggregated facts. For instance, if there are three BTS in a city, then the facts above can be aggregated from the BTS to the city level in the network dimension. For example: tch_req_success_city = tch_req_success_bts1 + tch_req_success_bts2 + tch_req_success_bts3 avg (tch_req_success_city = ((tch_req_success_bts1 + tch_req_success_bts2 + tch_req_success_bts3) / 3 Dimensional versus normalized approach for storage of data There are three or more leading approaches to storing data in a data warehouse – the most important approaches are the dimensional approach and the normalized approach. The dimensional approach refers to Ralph Kimball's approach in which it is stated that the data warehouse should be modeled using a Dimensional Model/star schema. The normalized approach, also called the 3NF model (Third Normal Form), refers to Bill Inmon's approach in which it is stated that the data warehouse should be modeled using an E-R model/normalized model.[16] Dimensional approach In a dimensional approach, transaction data is partitioned into "facts", which are generally numeric transaction data, and "dimensions", which are the reference information that gives context to the facts. For example, a sales transaction can be broken up into facts such as the number of products ordered and the total price paid for the products, and into dimensions such as order date, customer name, product number, order ship-to and bill-to locations, and salesperson responsible for receiving the order. A key advantage of a dimensional approach is that the data warehouse is easier for the user to understand and to use. Also, the retrieval of data from the data warehouse tends to operate very quickly.[15] Dimensional structures are easy to understand for business users, because the structure is divided into measurements/facts and context/dimensions. Facts are related to the organization's business processes and operational system whereas the dimensions surrounding them contain context about the measurement (Kimball, Ralph 2008). Another advantage offered by dimensional model is that it does not involve a relational database every time. Thus, this type of modeling technique is very useful for end-user queries in data warehouse. The model of facts and dimensions can also be understood as a data cube.[17] Where the dimensions are the categorical coordinates in a multi-dimensional cube, the fact is a value corresponding to the coordinates. The main disadvantages of the dimensional approach are the following: To maintain the integrity of facts and dimensions, loading the data warehouse with data from different operational systems is complicated. It is difficult to modify the data warehouse structure if the organization adopting the dimensional approach changes the way in which it does business. Normalized approach In the normalized approach, the data in the data warehouse are stored following, to a degree, database normalization rules. Tables are grouped together by subject areas that reflect general data categories (e.g., data on customers, products, finance, etc.). The normalized structure divides data into entities, which creates several tables in a relational database. When applied in large enterprises the result is dozens of tables that are linked together by a web of joins. Furthermore, each of the created entities is converted into separate physical tables when the database is implemented (Kimball, Ralph 2008). The main advantage of this approach is that it is straightforward to add information into the database. Some disadvantages of this approach are that, because of the number of tables involved, it can be difficult for users to join data from different sources into meaningful information and to access the information without a precise understanding of the sources of data and of the data structure of the data warehouse. Both normalized and dimensional models can be represented in entity-relationship diagrams as both contain joined relational tables. The difference between the two models is the degree of normalization (also known as Normal Forms). These approaches are not mutually exclusive, and there are other approaches. Dimensional approaches can involve normalizing data to a degree (Kimball, Ralph 2008). In Information-Driven Business,[18] Robert Hillard proposes an approach to comparing the two approaches based on the information needs of the business problem. The technique shows that normalized models hold far more information than their dimensional equivalents (even when the same fields are used in both models) but this extra information comes at the cost of usability. The technique measures information quantity in terms of information entropy and usability in terms of the Small Worlds data transformation measure.[19] Design methods This section needs additional citations for verification. Please help improve this article by adding citations to reliable sources. Unsourced material may be challenged and removed. (July 2015) (Learn how and when to remove this template message) Bottom-up design In the bottom-up approach, data marts are first created to provide reporting and analytical capabilities for specific business processes. These data marts can then be integrated to create a comprehensive data warehouse. The data warehouse bus architecture is primarily an implementation of "the bus", a collection of conformed dimensions and conformed facts, which are dimensions that are shared (in a specific way) between facts in two or more data marts.[20] Top-down design The top-down approach is designed using a normalized enterprise data model. "Atomic" data, that is, data at the greatest level of detail, are stored in the data warehouse. Dimensional data marts containing data needed for specific business processes or specific departments are created from the data warehouse.[21] Hybrid design Data warehouses (DW) often resemble the hub and spokes architecture. Legacy systems feeding the warehouse often include customer relationship management and enterprise resource planning, generating large amounts of data. To consolidate these various data models, and facilitate the extract transform load process, data warehouses often make use of an operational data store, the information from which is parsed into the actual DW. To reduce data redundancy, larger systems often store the data in a normalized way. Data marts for specific reports can then be built on top of the data warehouse. A hybrid DW database is kept on third normal form to eliminate data redundancy. A normal relational database, however, is not efficient for business intelligence reports where dimensional modelling is prevalent. Small data marts can shop for data from the consolidated warehouse and use the filtered, specific data for the fact tables and dimensions required. The DW provides a single source of information from which the data marts can read, providing a wide range of business information. The hybrid architecture allows a DW to be replaced with a master data management repository where operational (not static) information could reside. The data vault modeling components follow hub and spokes architecture. This modeling style is a hybrid design, consisting of the best practices from both third normal form and star schema. The data vault model is not a true third normal form, and breaks some of its rules, but it is a top-down architecture with a bottom up design. The data vault model is geared to be strictly a data warehouse. It is not geared to be end-user accessible, which, when built, still requires the use of a data mart or star schema-based release area for business purposes. Data warehouse characteristics There are basic features that define the data in the data warehouse include subject orientation, data integration, time-variant, nonvolatile data, and data granularity. Subject-oriented Unlike the operational systems, the data in the data warehouse revolves around subjects of the enterprise. Subject orientation is not database normalization. Subject orientation can be really useful for decision making. Gathering the required objects is called subject-oriented. Integrated The data found within the data warehouse is integrated. Since it comes from several operational systems, all inconsistencies must be removed. Consistencies include naming conventions, measurement of variables, encoding structures, physical attributes of data, and so forth. Time-variant While operational database characters current values as they support day-to-day operations, data warehouse data represents a long time horizon (up to 10 years) which means it stores mostly historical data. It is mainly meant for data mining and forecasting. (E.g. if a user is searching for a buying pattern of a specific customer, the user needs to look at data on the current and past purchases.)[22] Nonvolatile The data in the data warehouse is read-only, which means it cannot be updated, created, or deleted (unless there is a regulatory or statutory obligation to do so).[23] Data warehouse options Aggregation In the data warehouse process, data can be aggregated in data marts at different levels of abstraction. The user may start looking at the total sale units of a product in an entire region. Then the user looks at the states in that region. Finally, they may examine the individual stores in a certain state. Therefore, typically, the analysis starts at a higher level and drills down to lower levels of details.[22] Data warehouse architecture The different methods used to construct/organize a data warehouse specified by an organization are numerous. The hardware utilized, software created and data resources specifically required for the correct functionality of a data warehouse are the main components of the data warehouse architecture. All data warehouses have multiple phases in which the requirements of the organization are modified and fine-tuned.[24] Versus operational system Operational systems are optimized for the preservation of data integrity and speed of recording of business transactions through use of database normalization and an entity-relationship model. Operational system designers generally follow Codd's 12 rules of database normalization to ensure data integrity. Fully normalized database designs (that is, those satisfying all Codd rules) often result in information from a business transaction being stored in dozens to hundreds of tables. Relational databases are efficient at managing the relationships between these tables. The databases have very fast insert/update performance because only a small amount of data in those tables is affected each time a transaction is processed. To improve performance, older data are usually periodically purged from operational databases. Data warehouses are optimized for analytic access patterns. Analytic access patterns generally involve selecting specific fields and rarely if ever select *, which selects all fields/columns, as is more common in operational databases. Because of these differences in access patterns, operational databases (loosely, OLTP) benefit from the use of a row-oriented DBMS whereas analytics databases (loosely, OLAP) benefit from the use of a column-oriented DBMS. Unlike operational systems which maintain a snapshot of the business, data warehouses generally maintain an infinite history which is implemented through ETL processes that periodically migrate data from the operational systems over to the data warehouse. Evolution in organization use These terms refer to the level of sophistication of a data warehouse: Offline operational data warehouse Data warehouses in this stage of evolution are updated on a regular time cycle (usually daily, weekly or monthly) from the operational systems and the data is stored in an integrated reporting-oriented database. Offline data warehouse Data warehouses at this stage are updated from data in the operational systems on a regular basis and the data warehouse data are stored in a data structure designed to facilitate reporting. On-time data warehouse Online Integrated Data Warehousing represent the real-time Data warehouses stage data in the warehouse is updated for every transactional performed on the source data Integrated data warehouse These data warehouses assemble data from different areas of business, so users can look up the information they need across other systems.[25] References ^ Dedić, Nedim; Stanier, Clare (2016). Hammoudi, Slimane; Maciaszek, Leszek; Missikoff, Michele M. Missikoff; Camp, Olivier; Cordeiro, José (eds.). An Evaluation of the Challenges of Multilingualism in Data Warehouse Development". Mike2.0, the open source methodology for Information Development". Mike2.openmethodology.org. Retrieved 2013-06-14. ^ "The Bottom-Up Misnomer - DecisionWorks Consulting". DecisionWorks Consulting. Retrieved 2016-03-06. ^ Gartner, Of Data Proceedings of the 18th International Conference on Enterprise Information Systems, 25–28 April 2016, Rome, Italy (PDF). Proceedings of the 18th International Conference on Enterprise Information Systems (ICEIS 2016). 1. SciTePress. pp. 196–206. doi:10.5220/0005858401960206. ISBN 978-989-758-187-8. ^ a b "9 Reasons Data Warehouse Projects Fail". blog.rjmetrics.com. Retrieved 2017-04-30. ^ "Exploring Data Warehouses and Data Quality". spotlessdata.com. Archived from the original on 2018-07-26. Retrieved 2017-04-30. ^ "What is Big Data?". spotlessdata.com. Archived from the original on 2017-02-17. Retrieved 2017-04-30. ^ Patil, Preeti S.; Srikantha Rao; Suryakant B. Patil (2011). "Optimization of Data Warehousing System: Simplification in Reporting and Analysis". IJCA Proceedings on International Conference and Workshop on Emerging Trends in Technology (ICWET). Foundation of Computer Science. 9 (6): 33–37. ^ Marakas & O'Brien 2009 ^ a b c d e f Rainer, R. Kelly; Cegielski, Casey G. (2012-05-01). Introduction to Information Systems: Enabling and Transforming Business, 4th Edition (Kindle ed.). Wiley. pp. 127, 128, 130, 131, 133. ISBN 978-1118129401. ^ "Data Mart Concepts". Oracle. 2007. ^ "OLTP vs. OLAP". Datawarehouse4u.Info. 2009. We can divide IT systems into transactional (OLTP) and analytical (OLAP). In general, we can assume that OLTP systems provide source data to data warehouses, whereas OLAP systems help to analyze it. ^ "The Story So Far". 2002-04-15. Archived from the original on 2008-07-08. Retrieved 2008-09-21. ^ a b Kimball 2013, pg. 15 ^ Paul Gillin (February 20, 1984). "Will Teradata revive a market?". Computer World. pp. 43, 48. Retrieved 2017-03-13. ^ Devlin, B. A.; Murphy, P. T. (1988). "An architecture for a business and information system". IBM Systems Journal. 27: 60–80. doi:10.1147/sj.271.0060. ^ Inmon, Bill (1992). Building the Data Warehouse. Wiley. ISBN 0-471-56960-7. ^ a b Kimball, Ralph (2011). The Data Warehouse Toolkit. Wiley. ISBN 978-0-470-14977-5. ^ Golfarelli, Matteo; Maio, Dario; Rizzi, Stefano (1998-06-01). "The dimensional fact model: a conceptual model for data warehouses". International Journal of Cooperative Information Systems. 07 (02n03): 215–247. doi:10.1142/S0218843098000118. ISSN 0218-8430. ^ dbd/cs831/notes/dcubes/dcubes.html ^ Hillard, Robert (2010). Information-Driven Business. Wiley. ISBN 978-0-470-62577-4. ^ "Information Theory & Business Intelligence Strategy - Small Worlds Data Transformation Measure - MIKE2.0, the open source methodology for Information Development". Mike2.openmethodology.org. Retrieved 2013-06-14. ^ Warehouses, Operational Data Stores, Data Marts and Data Outhouses, Dec 2005 ^ a b Paulraj., Ponniah (2010). Data warehousing fundamentals for IT professionals. Ponniah, Paulraj. (2nd ed.). Hoboken, N.J.: John Wiley & Sons. ISBN 9780470462072. OCLC 662453070. ^ H., Inmon, William (2005). Building the data warehouse (4th ed.). Indianapolis, IN: Wiley Pub. ISBN 9780764599446. OCLC 61762085. ^ Gupta, Satinder Bal; Mittal, Aditya (2009). Introduction to Database Management System. Laxmi Publications. ISBN 9788131807248. ^ "Data Warehouse". Further reading Davenport, Thomas H. and Harris, Jeanne G. Competing on Analytics: The New Science of Winning (2007) Harvard Business School Press. ISBN 978-1-4221-0332-6 Ganczarski, Joe. Data Warehouse Implementations: Critical Implementation Factors Study (2009) VDM Verlag ISBN 3-639-18589-7 ISBN 978-3-639-18589-8 Kimball, Ralph and Ross, Margy. The Data Warehouse Toolkit Third Edition (2013) Wiley, ISBN 978-1-118-53080-1 Linstedt, Graziano, Hultgren. The Business of Data Vault Modeling Second Edition (2010) Dan linstedt, ISBN 978-1-4357-1914-9 William Inmon. Building the Data Warehouse (2005) John Wiley and Sons, ISBN 978-81-265-0645-3 Retrieved from "