TechTarget

E-Guide

Top Data Management Terms to Know

Fifteen essential definitions you need to know

We know it's not always easy to keep up-to-date *with the latest data management terms. That's why we have put together the top fifteen terms and definitions that you and your peers need to know.*

Contents

What is OLAP (online analytical processing)

OLAP (online analytical processing) is computer processing that enables a user to easily and selectively extract and view data from different points of view. For example, a user can request that data be analyzed to display a spreadsheet showing all of a company's beach ball products sold in Florida in the month of July, compare revenue figures with those for the same products in September, and then see a comparison of other product sales in Florida in the same time period. To facilitate this kind of analysis, OLAP data is stored in a multidimensional database. Whereas a relational database can be thought of as two-dimensional, a multidimensional database considers each data attribute (such as product, geographic sales region, and time period) as a separate "dimension." OLAP software can locate the intersection of dimensions (all products sold in the Eastern region above a certain price during a certain time period) and display them. Attributes such as time periods can be broken down into subattributes.

OLAP can be used for data mining or the discovery of previously undiscerned relationships between data items. An OLAP database does not need to be as large as a data warehouse, since not all transactional data is needed for trend analysis. Using Open Database Connectivity (ODBC), data can be imported from existing relational databases to create a multidimensional database for OLAP.

Two leading OLAP products are Hyperion Solution's Essbase and Oracle's Express Server. OLAP products are typically designed for multiple-user environments, with the cost of the software based on the number of users.

Contents

What is star schema?

In data warehousing and business intelligence (BI), a star schema is the simplest form of a dimensional model, in which data is organized into facts and dimensions. A fact is an event that is counted or measured, such as a sale or login. A dimension contains reference information about the fact, such as date, product, or customer. A star schema is diagramed by surrounding each fact with its associated dimensions. The resulting diagram resembles a star.

Star schemas are optimized for querying large data sets and are used in data warehouses and data marts to support OLAP cubes, business intelligence and analytic applications, and ad hoc queries.

Within the data warehouse or data mart, a dimension table is associated with a fact table by using a foreign key relationship. The dimension table has a single primary key that uniquely identifies each member record (row). The fact table contains the primary key of each associated dimension table as a foreign key. Combined, these foreign keys form a multi-part composite primary key that uniquely identifies each member record in the fact table. The fact table also contains one or more numeric measures.

For example, a simple Sales fact with millions of individual clothing sale records might contain a Product Key, Promotion Key, Customer Key, and Date Key, along with Units Sold and Revenue measures. The Product dimension would hold reference information such as product name, description, size, and color. The Promotion dimension would hold information such as promotion name and price. The Customer dimension would hold information such as first and last name, birth date, gender, address, etc. The Date dimension would include calendar date, week of year, month, quarter, year, etc. This simple Sales fact will easily support queries such as “total revenue for all clothing products sold during the first quarter of the 2010” or “count of female customers who purchased 5 or more dresses in December 2009”.

What is fact table?

A fact table is the central table in a star schema of a data warehouse. A fact table stores quantitative information for analysis and is often denormalized.

A fact table works with dimension tables. A fact table holds the data to be analyzed, and a dimension table stores data about the ways in which the data in the fact table can be analyzed. Thus, the fact table consists of two types of columns. The foreign keys column allows joins with dimension tables, and the measures columns contain the data that is being analyzed.

Suppose that a company sells products to customers. Every sale is a fact that happens, and the fact table is used to record these facts. For example:

Time ID	Product ID	Customer ID	Unit Sold
4	17	2	1
8	21	3	2
8	4	1	1

Now we can add a dimension table about customers:

Customer ID	Name	Gender	Income	Education	Region
1	Brian Edge	M	2	3	4
2	Fred Smith	M	3	5	1
3	Salle Jones	F	1	7	3

In this example, the customer ID column in the fact table is the foreign key that joins with the dimension table. By following the links, you can see that row 2 of the fact table records the fact that customer 3, Sally Jones, bought two items on day 8. The company would also have a product table and a time table to determine what Sally bought and exactly when.

When building fact tables, there are physical and data limits. The ultimate size of the object as well as access paths should be considered. Adding

Contents

indexes can help with both. However, from a logical design perspective, there should be no restrictions. Tables should be built based on current and future requirements, ensuring that there is as much flexibility as possible built into the design to allow for future enhancements without having to rebuild the data.

What is big data analytics?

Big data analytics is the process of examining large amounts of data of a variety of types (big data) to uncover hidden patterns, unknown correlations and other useful information. Such information can provide competitive advantages over rival organizations and result in business benefits, such as more effective marketing and increased revenue.

The primary goal of big data analytics is to help companies make better business decisions by enabling data scientists and other users to analyze huge volumes of transaction data as well as other data sources that may be left untapped by conventional business intelligence (BI) programs. These other data sources may include Web server logs and Internet clickstream data, social media activity reports, mobile-phone call detail records and information captured by sensors. Some people exclusively associate big data and big data analytics with unstructured data of that sort, but consulting firms like Gartner Inc. and Forrester Research Inc. also consider transactions and other structured data to be valid forms of big data.
Big data analytics can be done with the software tools commonly used as part of advanced analytics disciplines such as predictive analytics and data mining. But the unstructured data sources used for big data analytics may not fit in traditional data warehouses. Furthermore, traditional data warehouses may not be able to handle the processing demands posed by big data. As a result, a new class of big data technology has emerged and is being used in many big data analytics environments. The technologies associated with big data analytics include NoSQL databases, Hadoop and MapReduce. These technologies form the core of an open source software framework that supports the processing of large data sets across clustered systems.

Contents

Potential pitfalls that can trip up organizations on big data analytics initiatives include a lack of internal analytics skills and the high cost of hiring experienced analytics professionals, plus challenges in integrating Hadoop systems and data warehouses, although vendors are starting to offer software connectors between those technologies.

What is data modeling?

Data modeling is the formalization and documentation of existing processes and events that occur during application software design and development. Data modeling techniques and tools capture and translate complex system designs into easily understood representations of the data flows and processes, creating a blueprint for construction and/or re-engineering.

A data model can be thought of as a diagram or flowchart that illustrates the relationships between data. Although capturing all the possible relationships in a data model can be very time-intensive, it's an important step and shouldn't be rushed. Well-documented models allow stake-holders to identify errors and make changes before any programming code has been written.

Data modelers often use multiple models to view the same data and ensure that all processes, entities, relationships and data flows have been identified. There are several different approaches to data modeling, including:

Conceptual Data Modeling - identifies the highest-level relationships between different entities.

Enterprise Data Modeling - similar to conceptual data modeling, but addresses the unique requirements of a specific business.

Logical Data Modeling - illustrates the specific entities, attributes and relationships involved in a business function. Serves as the basis for the creation of the physical data model.

Contents

Physical Data Modeling - represents an application and database-specific implementation of a logical data model.

What is ad hoc analysis?

Ad hoc analysis is a business intelligence process designed to answer a single, specific business question. The product of ad hoc analysis is typically a statistical model, analytic report, or other type of data summary. According to Merriam-Webster Dictionary, ad hoc means "for the particular case at hand without consideration of wider application." The purpose of an ad hoc analysis is to fill in gaps left by the business' static, regular reporting.

Ad hoc analysis may be used to create a report that does not already exist, or drill deeper into a static report to get details about accounts, transactions, or records. The process may be also used to get more current data for the existing areas covered by a static report.

OLAP dashboards are specifically designed to facilitate ad hoc analysis by providing quick, easy access to data from the original report. Allowing the user (typically a manager or executive) access to data through a point-and-click interface eliminates the need to request data and analysis from another group within the company. This capacity allows for quicker response times when a business question comes up, which in turn should help the user respond to issues and make business decisions faster.

Although most ad hoc reports and analyses are meant to be run only once, in practice they often end up being reused and run on a regular basis. This relatively common practice can lead to unnecessary reporting processes that impact high-volume reporting periods. Reports should be reviewed periodically for efficiencies to determine whether they continue to serve a useful business purpose.

Contents

What is data visualization?

Data visualization is a general term that describes any effort to help people understand the significance of data by placing it in a visual context. Patterns, trends and correlations that might go undetected in text-based data can be exposed and recognized easier with data visualization software.

Today's data visualization tools go beyond the standard charts and graphs used in Excel spreadsheets, displaying data in more sophisticated ways such as infographics, dials and gauges, geographic maps, sparklines, heat maps, and detailed bar, pie and fever charts. The images may include interactive capabilities, enabling users to manipulate them or drill into the data for querying and analysis. Indicators designed to alert users when data has been updated or predefined conditions occur can also be included.

Most business intelligence software vendors embed data visualization tools into their products, either developing the visualization technology themselves or sourcing it from companies that specialize in visualization.

What is extract, transform, load (ETL)?

In managing databases, extract, transform, load (ETL) refers to three separate functions combined into a single programming tool. First, the extract function reads data from a specified source database and extracts a desired subset of data. Next, the transform function works with the acquired data - using rules or lookup tables, or creating combinations with other data - to convert it to the desired state. Finally, the load function is used to write the resulting data (either all of the subset or just the changes) to a target database, which may or may not previously exist.

ETL can be used to acquire a temporary subset of data for reports or other purposes, or a more permanent data set may be acquired for other purposes such as: the population of a data mart or data warehouse; conversion from one database type to another; and the migration of data from one database or platform to another.

Contents

Contents

What is association rules (in data mining)?

Association rules are if/then statements that help uncover relationships between seemingly unrelated data in a relational database or other information repository. An example of an association rule would be "If a customer buys a dozen eggs, he is 80% likely to also purchase milk."

An association rule has two parts, an antecedent (if) and a consequent (then). An antecedent is an item found in the data. A consequent is an item that is found in combination with the antecedent.

Association rules are created by analyzing data for frequent if/then patterns and using the criteria support and confidence to identify the most important relationships. Support is an indication of how frequently the items appear in the database. Confidence indicates the number of times the if/then statements have been found to be true.

In data mining, association rules are useful for analyzing and predicting customer behavior. They play an important part in shopping basket data analysis, product clustering, catalog design and store layout.

Programmers use association rules to build programs capable of machine learning. Machine learning is a type of artificial intelligence (AI) that seeks to build programs with the ability to become more efficient without being explicitly programmed.

What is relational database?

A relational database is a collection of data items organized as a set of formally-described tables from which data can be accessed or reassembled in many different ways without having to reorganize the database tables. The relational database was invented by E. F. Codd at IBM in 1970.

Contents

The standard user and application program interface to a relational database is the structured query language (SQL). SQL statements are used both for interactive queries for information from a relational database and for gathering data for reports.

In addition to being relatively easy to create and access, a relational database has the important advantage of being easy to extend. After the original database creation, a new data category can be added without requiring that all existing applications be modified.

A relational database is a set of tables containing data fitted into predefined categories. Each table (which is sometimes called a relation) contains one or more data categories in columns. Each row contains a unique instance of data for the categories defined by the columns. For example, a typical business order entry database would include a table that described a customer with columns for name, address, phone number, and so forth. Another table would describe an order: product, customer, date, sales price, and so forth. A user of the database could obtain a view of the database that fitted the user's needs. For example, a branch office manager might like a view or report on all customers that had bought products after a certain date. A financial services manager in the same company could, from the same tables, obtain a report on accounts that needed to be paid.

When creating a relational database, you can define the domain of possible values in a data column and further constraints that may apply to that data value. For example, a domain of possible customers could allow up to ten possible customer names but be constrained in one table to allowing only three of these customer names to be specifiable.

The definition of a relational database results in a table of metadata or formal descriptions of the tables, columns, domains, and constraints.

Contents

What is denormalization?

In a relational database, denormalization is an approach to speeding up read performance (data retrieval) in which the administrator selectively adds back specific instances of redundant data after the data structure has been normalized. A denormalized database should not be confused with a database that has never been normalized.

During normalization, the database designer stores different but related types of data in separate logical tables called relations. When a query combines data from multiple tables into a single result table, it is called a join. Multiple joins in the same query can have a negative impact on performance. Introducing denormalization and adding back a small number of redundancies can be a useful for cutting down on the number of joins.

After data has been duplicated, the database designer must take into account how multiple instances of the data will be maintained. One way to denormalize a database is to allow the database management system (DBMS) to store redundant information on disk. This has the added benefit of ensuring the consistency of redundant copies. Another approach is to denormalize the actual logical data design, but this can quickly lead to inconsistent data. Rules called constraints can be used to specify how redundant copies of information are synchronized, but they increase the complexity of the database design and also run the risk of impacting write performance.

What is master data management (MDM)?

Master data management (MDM) is a comprehensive method of enabling an enterprise to link all of its critical data to one file, called a master file, that provides a common point of reference. When properly done, MDM streamlines data sharing among personnel and departments. In addition, MDM can facilitate computing in multiple system architectures, platforms and applications.

The benefits of the MDM paradigm increase as the number and diversity of organizational departments, worker roles and computing applications expand. For this reason, MDM is more likely to be of value to large or complex enterprises than to small, medium-sized or simple ones. When companies merge, the implementation of MDM can minimize confusion and optimize the efficiency of the new, larger organization.

For MDM to function at its best, all personnel and departments must be taught how data is to be formatted, stored and accessed. Frequent, coordinated updates to the master data file are also essential.

What is predictive modeling?

Predictive modeling is a process used in predictive analytics to create a statistical model of future behavior. Predictive analytics is the area of data mining concerned with forecasting probabilities and trends.

A predictive model is made up of a number of predictors, which are variable factors that are likely to influence future behavior or results. In marketing, for example, a customer's gender, age, and purchase history might predict the likelihood of a future sale.

In predictive modeling, data is collected for the relevant predictors, a statistical model is formulated, predictions are made and the model is validated (or revised) as additional data becomes available. The model may employ a simple linear equation or a complex neural network, mapped out by sophisticated software.

Predictive modeling is used widely in information technology (IT). In spam filtering systems, for example, predictive modeling is sometimes used to identify the probability that a given message is spam. Other applications of predictive modeling include customer relationship management (CRM), capacity planning, change management, disaster recovery, security management, engineering, meteorology and city planning.

What is text mining analytics?

Text mining is the analysis of data contained in natural language text. The application of text mining techniques to solve business problems is called text analytics.

Text mining can help an organization derive potentially valuable business insights from text-based content such as word documents, email and postings on social media streams like Facebook, Twitter and LinkedIn. Mining unstructured data with natural language processing (NLP), statistical modeling and machine learning techniques can be challenging, however, because natural language text is often inconsistent. It contains ambiguities caused by inconsistent syntax and semantics, including slang, language specific to vertical industries and age groups, double entendres and sarcasm.

Text analytics software can help by transposing words and phrases in unstructured data into numerical values which can then be linked with structured data in a database and analyzed with traditional data mining techniques. With an iterative approach, an organization can successfully use text analytics to gain insight into content-specific values such as sentiment, emotion, intensity and relevance. Because text analytics technology is still considered to be an emerging technology, however, results and depth of analysis can vary wildly from vendor to vendor.

What is Hadoop cluster?

A Hadoop cluster is a special type of computational cluster designed specifically for storing and analyzing huge amounts of unstructured data in a distributed computing environment.

Contents

Contents

Such clusters run Hadoop's open source distributed processing software on low-cost commodity computers. Typically one machine in the cluster is designated as the NameNode and another machine the as JobTracker; these are the masters. The rest of the machines in the cluster act as both DataNode and TaskTracker; these are the slaves. Hadoop clusters are often referred to as "shared nothing" systems because the only thing that is shared between nodes is the network that connects them.

Hadoop clusters are known for boosting the speed of data analysis applications. They also are highly scalable: If a cluster's processing power is overwhelmed by growing volumes of data, additional cluster nodes can be added to increase throughput. Hadoop clusters also are highly resistant to failure because each piece of data is copied onto other cluster nodes, which ensures that the data is not lost if one node fails.

As of early 2013, Facebook was recognized as having the largest Hadoop cluster in the world. Other prominent users include Google, Yahoo and IBM.

Want more? View our Data Management Glossary:

http://whatis.techtarget.com/glossary/Data-and-Data-Management

TechTarget

Contents

Free resources for technology professionals

TechTarget publishes targeted technology media that address your need for information and resources for researching products, developing strategy and making cost-effective purchase decisions. Our network of technology-specific Web sites gives you access to industry experts, independent content and analysis and the Web's largest library of vendor-provided white papers, webcasts, podcasts, videos, virtual trade shows, research reports and more —drawing on the rich R&D resources of technology providers to address market trends, challenges and solutions. Our live events and virtual seminars give you access to vendor neutral, expert commentary and advice on the issues and challenges you face daily. Our social community IT Knowledge Exchange allows you to share real world information in real time with peers and experts.

What makes TechTarget unique?

TechTarget is squarely focused on the enterprise IT space. Our team of editors and network of industry experts provide the richest, most relevant content to IT professionals and management. We leverage the immediacy of the Web, the networking and face-to-face opportunities of events and virtual events, and the ability to interact with peers—all to create compelling and actionable information for enterprise IT professionals across all industries and markets.

Related TechTarget Websites

SearchDataManagement

SearchBusinessAnalytics