

Database Glossary

Ad-Hoc Reporting: The ability to access information from a database to meet information requirements as required. One-off requests not satisfied by pre-set reporting.

Agent: An application that searches the data and sends an alert when a particular pattern is found.

Aggregations: Information stored in a data warehouse in a summarized form. Instead of recording the date and time each time a certain product is sold, the data warehouse could store the quantity of the product sold each hour, each day, or each week. Aggregations are used for two primary reasons:

To save storage space. Data warehouses can get large. The use of aggregations greatly reduces the space needed to store data.

To improve the performance of business intelligence tools. When queries run faster they take up less processing time and the users get their information back more quickly.

Some data warehouses store both the detailed information and aggregated information. This takes even more space, but gives users the possibility of looking at all the details while still having good query performance when looking at summaries.

Some systems use aggregations for historical data. Perhaps detailed data is kept on-line for a year. After that the detailed data is kept in a less accessible, permanent storage format, and only the aggregated, summary data is kept on-line.

Aggregations are often created as the sum of the individual records. You can also have aggregations for count, distinct count, maximum value, and minimum value.

Alert: A message that is sent automatically by a computer system when a certain situation occurs.

Application: A piece of software designed to meet a specific purpose.

Attribute: A single data item related to a database object. The database schema associates one or more attributes with each database entity.

Also Known As: field, column

Example: In the following database table, the attributes are <name, ID, extension>

Name	ID	Extension
Jim	124	7075
Valeri	128	0853
Bob	192	4214

Business Intelligence Tools: Software that enables business users to see and use large amounts of complex data.

Candidate Key: A combination of attributes that can be uniquely used to identify a database record. Each table may have one or more candidate keys. One of these candidate keys is selected as the table primary key.

Examples: There are a large number of candidate keys in the sample table below. Some of these are <SSN>, <Phone Extension>, <Name, SSN>, and <Name, Age, SSN>. Note that <Age> is not a candidate key in this case because Amy and Elizabeth share the same age.

Name	Age	SSN	Phone Extension
Rob	28	123-45-6789	1242
Amy	34	987-65-4321	9281
Elizabeth	34	111-22-3333	9312
Jim	42	333-22-1111	3214
Mike	29	999-99-9999	2314

Cardinality: In set theory, cardinality refers to the number of members in the set. When specifically applied to database theory, the cardinality of a table refers to the number of rows(or tuples) contained in a table.

Chief information officer (CIO): The senior executive in a company responsible for information management and for delivering IT services.

Clickstream Data: Data regarding web browsing.

Client/server architecture: A type of network in which computer processing is distributed among many individual PCs (clients) and a more powerful, central computer (server). Clients can share files and retrieve data stored on the server.

Collaborative software: Groupware, such as Lotus Notes or Microsoft Exchange.

Data: A series of facts or statements that may have been collected, stored, processed and/or manipulated but have not been organized or placed into context. When data is organized, it becomes information. Information can be processed and used to draw generalized conclusions or knowledge.

Database: A collection of tables. It also often includes forms for entering data, rules for checking and validating data that has been entered, and the format for creating informative reports from the data in the database.

Data Cleansing: The process of manipulating stored data to make it more aligned. Implies eliminating duplication, correcting spelling and removing null fields.

Data-Based Knowledge: Knowledge derived from data through the use of Business Intelligence Tools and the process of Data Warehousing.

Data Mart: Also Known As: Local Data Warehouse or Datamart. A database that has the same characteristics as a data warehouse, but is usually smaller and is focused on the data for one division or one workgroup within an enterprise.

Data Migration: The process of physically transmitting data from one environment to another.

Data Mining: The use of automated data analysis techniques to uncover previously undetected relationships among data items. Data mining often involves the analysis of data stored in a data warehouse. Three of the major data mining techniques are regression, classification and clustering.

Data Model: The description of the contents of a database. It includes tables and the relationship between them.

Data Quality: The suitability of data for different requirements. When constructing and populating a database the data quality must be checked to ensure that it meets the needs of all the applications.

Data Scrubbing: Removing errors and inconsistencies from data being imported into a data warehouse.

Data Transformation: The modification of data as it is moved into the data warehouse.

Data Warehouse: A data warehouse is a centralized database that captures information from various parts of an organization's business processes. This information can later be analyzed to determine predictive relationships through the use of data mining techniques.

Data Warehousing Management: The on-going supervision of the data warehousing process.

Database Management System (DBMS): The software that is used to store, access, and manage data.

Decision Support System (DSS): A computer system designed to assist an organization in making decisions.

Dimension: The separation of data according to different view. A logical designation of related information that applies a hierarchy for access and reporting.

Domain: The set of all allowable values that attribute may assume.

Drill Down and Drill Up: The ability to move between levels of the hierarchy when viewing data with an OLAP browser.

i) Drill Down - Changing the view of the data to a greater level of detail.

ii) Drill Up - Changing the view of the data to a higher level of aggregation.

Enterprise resource planning: An integrated system of operation applications combining logistics, production, distribution, contract and order management, sales forecasting, and financial and HR management.

Entity: A single object about which data can be stored. It is the "subject" of a table. Entities and their interrelationships are modeled through the use of entity-relationship diagrams.

Electronic data interchange (EDI): Electronic transmission of documents through point-to-point connections using a set of standard forms, messages and data elements; this can be via leased lines, private networks or the Internet.

Fact table: In a star schema, the central table which contains the individual facts being stored in the database.

Field: The most basic structural unit of a database. It is a container for a piece of data. In most cases, only a single logical piece of data fits in each field.

Flat File: A data file that contains records with no structured relationships. Additional knowledge is required to interpret these files such as the file format properties. Modern database management systems used a more structured approach to file management (such

as one defined by the Structured Query Language) and therefore have more complex storage arrangements.

Form: A database form can be used to facilitate database data entry and/or retrieval operations. A database developer/administrator usually designs a form which can then be used by personnel without any specific database skills to perform repetitive tasks.

Front End: Tools that enable users to interact with underlying application processes or more complex programs through a familiar, easy-to-use interface.

Gateway: This is the interface between different computer network, (usually translates from one network protocol to another).

Granularity: The level of detail of the facts stored in a data warehouse.

Hardware: The magnetic, mechanical and electrical components of a computer and its peripheral devices.

Hierarchy: Organization of data into a logical tree structure.

Index: A database feature used for locating data quickly within a table. Indexes are defined by selecting a set of commonly searched attribute(s) on a table and using the appropriate platform-specific mechanism to create an index.

Example: Personnel information may be store in a Human Resource department's employee table. Clerks find that they often search the table for employees by last name but get slow query responses. Defining an index on the table consisting of the last name attribute would speed up these queries.

Information technology (IT): The hardware and software used to process information.

Key: A field that contains a unique identifier for each row in a data table. Even though each individual record represents a separate piece of data, some of those records may look identical. A key provides a completely unambiguous way to distinguish between distinct records, and more importantly, serves as a pointer to a particular record in the table. In many cases, data table keys are constructed by simply adding an additional field to function as the key.

Legacy System: A computer system that's been around for a while.

Level: The hierarchies in dimensions have levels which can be used to view data at various levels of detail.

Examples:

i) A Time dimension could have levels for Year, Quarter, Month, and Day.

ii) A Product dimension could have levels for Product Family, Product Category, Product Subcategory, and Product Name.

iii) A Customer Geography dimension could have levels for Region, Country, District, State, City, and Neighborhood.

Local Database Warehouse: A database that has the same characteristics as a data warehouse, but is usually smaller and is focused on the data for one division or one workgroup within an enterprise.

Mainframe: The central processing unit of a large computer, usually receiving input from a number of terminals.

Member: One of the data points for a level of a hierarchy of a dimension.

Example:

i) Some of the members of the Month level of the Time dimension are January, February, March, and April.

Metadata: A term that literally means "data about data." This term refers to information about data itself -- perhaps the origin, size, formatting or other characteristics of a data item. In the database field, metadata is essential to understanding and interpreting the contents of a data warehouse.

Microprocessors: Complex electronic circuits that comprise a computer's central information processing unit.

Multidimensional Database Management System (MDBMS): A database management system that organizes data multidimensionally.

Non-Volatile: Data that does not change.

Normalization: The process of structuring relational database schema such that most ambiguity is removed. The stages of normalization are referred to as normal forms and progress from the least restrictive (First Normal Form) through the most restrictive (Fifth Normal Form). Generally, most database designers do not attempt to implement anything higher than

OLAP (On-Line Analytical Processing): The use of computers to analyze an organization's data.

OLAP Browser: A tool used for multidimensional (OLAP) browsing.

OLAP System: Term that is used as a synonym for datawarehousing system.

OLTP (OnLine Transaction Processing): The use of computers to run the on-going operation of a business.

Program: A set of digitally coded definitions and instructions that enables a computer to perform a particular task.

Protocol: The language that one computer uses to communicate with another.

Query: The primary mechanism for retrieving information from a database and consist of questions presented to the database in a predefined format. Many database management systems use the (Structured Query Language)standard query format.

Record or a row: The fields in a record provide a complete description of each item in a collection. A record is a unique instance of data about an object or event.

Relationships: The connections between records in different data tables are provided by *relationships*.

i) **one-to-many relationship:** The most common relationship between two tables. In this situation, precisely one record in data table A is related to a number of records in data table

B. The primary key of table A is inserted as a field into table B where it serves as a foreign key. Relationships between tables are always made through keys.

ii) **one-to-one relationship:** Each record in one table is linked to one and only one record in another table. In many cases, one of the tables is set of data about a subset of the entities in the main table.

iii) **many-to-many relationship:** Table design rounds out the possible ways of designing data relationships. In some situations, multiple entries in a data table are related to multiple entries in another data table. The classic example is a database of classes and students. Each class consists of many students and each student can take many classes, so in designing a database to track student and class information, a many-to-many design is necessary.

The distinguishing characteristic of many-to-many relationships between two tables is that they require a third table to make the relationship. Simply putting the key from one table into the other table would result in a lot of duplicated information, so a linking table is used to connect the tables. The linking table simply stores the primary key from one table with the primary key of its related entry from the second table, along with any other information unique to the relationship.

Relational Database Management System (RDBMS): A Database Management System is based on relational theory. Most modern Database Management Systems (Oracle, Sybase, Microsoft SQL Server) are relational databases. These databases support a standard language - SQL (Structured Query Language).

Replication: The physical copying of data from one database to another.

Scale, Scalable, and Scalability: Having to do with the ability of a computer system or a database to operate efficiently with larger quantities of data. Scalability is often discussed in situations when multiple processors are joined together. The system scales well (or is scalable) if doubling the number of processors also doubles the speed at which the system performs its tasks. The extra work involved in coordinating larger systems usually prevents them from being fully scalable - so that going from one to two processors would increase the total speed by less than a factor of two.

Schema: The logical organization of data in a database.

Shared Dimension: In Microsoft Analysis Services, a dimension used by more than one cube.

Slowly Changing Dimensions (SCD): A dimension that has levels or attributes that are changing on an occasional basis.

Software: The programs that are run on a computer system.

Star Schema (Business Definition): A method of organizing information in a data warehouse that allows the business information to be viewed from many perspectives. The star is a picture of the way the data is being stored. The basic factual information is in the middle of the star. The points of the star represent various perspectives from which the factual information can be viewed.

Structured Query Language (SQL): An industry-standard language used for manipulation of data in a relational database. The major SQL commands of interest to database users are Select, Insert, Join and Update.

Summary Tables: Tables used to store summarized or aggregated data.

Supercomputer: An extremely powerful computer; designed to deal with large amounts of data at very high speed, often used for military or scientific tasks.

Table: The formal name given to the group of records that contain the elements of the collection. A table normally represents a distinct object (business clients or library books), or an event (product orders or stock prices).

Time-variant data: Data that is identified with a particular time period. Time-variant is one of the original defining characteristics of a data warehouse.

XML: XML is the eXtensible Markup Language -- a system created to define other markup languages. For this reason, it can also be referred to as a metalanguage. XML is commonly used on the Internet to create simple methods for the exchange of data among diverse clients.