Creating dimensional mode

Dimensional modeling involves the use of fact and dimension tables to maintain a record of historical data in data warehouses. Different types of data modeling techniques are optimized for different applications. Normalized Entity-Relationship models (ER models) are designed to eliminate data redundancy and quickly perform the Insert, Update, and Delete operations, and get the data inside a database.

In contrast to that, dimensional models or kimball dimensional data models – data models based on the technique developed by Ralph Kimball – are denormalized structures designed to retrieve data out of a data warehouse. They are optimized to perform the Select operation and are used in the basic design framework to build highly optimized and functional data warehouses.

Building a dimensional data model

- A business process
  A *business process* is an important operation in your organization that some legacy system supports. You collect data from this system to use in your dimensional database.
- Summary of a business process
- Determine the granularity of the fact table
  After you gather all the relevant information about the subject area, the next step in the design process is to determine the granularity of the fact table.
- Identify the dimensions and hierarchies
  After you determine the granularity of the fact table, it is easy to identify the primary dimensions for the data model because each component that defines the granularity corresponds to a dimension.
- Establish referential relationships
  For the database server to support the dimensional data model, you must define logical dependencies between the fact tables and their dimension tables.
- Fragmentation: Storage distribution strategies
  The performance of data warehousing applications can typically benefit from distributed storage allocation designs for partitioning a database table into two or more fragments. Each fragment has the same schema as the table, and stores a subset of the rows in the table (rather than a subset of its columns).

Benefits of Dimensional Modeling

Dimensional modeling is still the most commonly used data modeling technique for designing enterprise data warehouses because of the benefits it yields. These include:

Faster Retrieval of Data

Dimensional data modeling merges the tables in the model itself, which enables users to retrieve data faster from different data sources by running join queries. The denormalized schema of a dimensional model data warehouse, as opposed to normalized one in snowflake schema, is optimized to run ad hoc queries. As a result, it greatly complements the business intelligence (BI) goals of an organization.

Better Understanding of Business Processes

The principles of dimensional modeling are based on fact and dimension tables. We will cover what facts and dimensions are in the subsequent sections. This categorization of data into facts and dimensions, and the entity-relationship structure of a dimensional model, present complex business processes in an easy-to-understand manner to analysts.

Flexible to Change

Dimensional modeling framework makes the data warehousing process extensible. The design can be easily modified to incorporate any new business requirements or make any adjustments to the central repository. New entities can be added in the model or layout of the existing ones can be changed to reflect modified business processes.

ETL

What is Extract, Transform and Load? Explain

ETL stands for Extract Transform and Load. ETL is a data integration process that combines data from multiple data sources into a single, consistent data store that is loaded in a DW. It is often used to build a data warehouse.

Extract: Extract is the process of fetching (reading) the information from the database. At this stage, data is collected from multiple or different types of sources which can be structured or unstructured.

Transform: Transform is the process of converting the extracted data from its previous form into the required form. Data can be placed into another database. This can involve Filtering, Cleaning, De-duplication, Validation, Authenticating the data.

Load: Load is the process of writing the data into the target database. • ETL takes place during off-hours when traffic on the source systems and the DW is at its lowest.

Need of ETL

There are many reasons the need for ETL is arising:

• ETL helps the companies to analyze their business data for making critical business decisions.

• Data warehouse provides a shared data repository.

• ETL provides a method of moving data from various sources into a data warehouse.

• As the data sources change, the data warehouse will automatically update.

• ETL helps to migrate the data into a data warehouse.

Extract, transform, load (ETL) refers to the process of moving data from a source system into a data warehouse. Before loading into the warehouse, the data is transformed from a raw state into the format required by the enterprise data warehouse.

The 5 steps of the ETL process are:

extract, clean, transform, load, and analyze.

Extract: Retrieves raw data from an unstructured data pool and migrates it into a temporary, staging data repository

Clean: Cleans data extracted from an unstructured data pool, ensuring the quality of the data prior to transformation.

Transform: Structures and converts the data to match the correct target source Load: Loads the structured data into a data warehouse so it can be properly analyzed and used

Analyze: Big data analysis is processed within the warehouse, enabling the business to gain insight from the correctly configured data.

Each step is performed sequentially. However, the exact nature of each step – which format is required for the target database – depends on the enterprise's specific needs and requirements.

Extraction can involve copying data to tables quickly to minimize the time spent querying the source system. In the transformation step, the data is most usually stored in one set of staging tables as part of the process. Finally, a secondary transformation step might place data in tables that are copies of the warehouse tables, which eases loading.

Each ETL stage requires interaction by data engineers and developers to deal with the capacity limitations of traditional data warehouses.

ETL has been the standard for data warehousing and analytics within enterprises for some time. But as we progress further into 2021, we must view ETL not just as its own microcosm of data readiness processes within an enterprise, but also in the context of an enterprise-wide integration and enhanced business outcomes.

Define Metadata and explain metadata types.

A data of a data / Data that provides information about other data is known as Metadata.

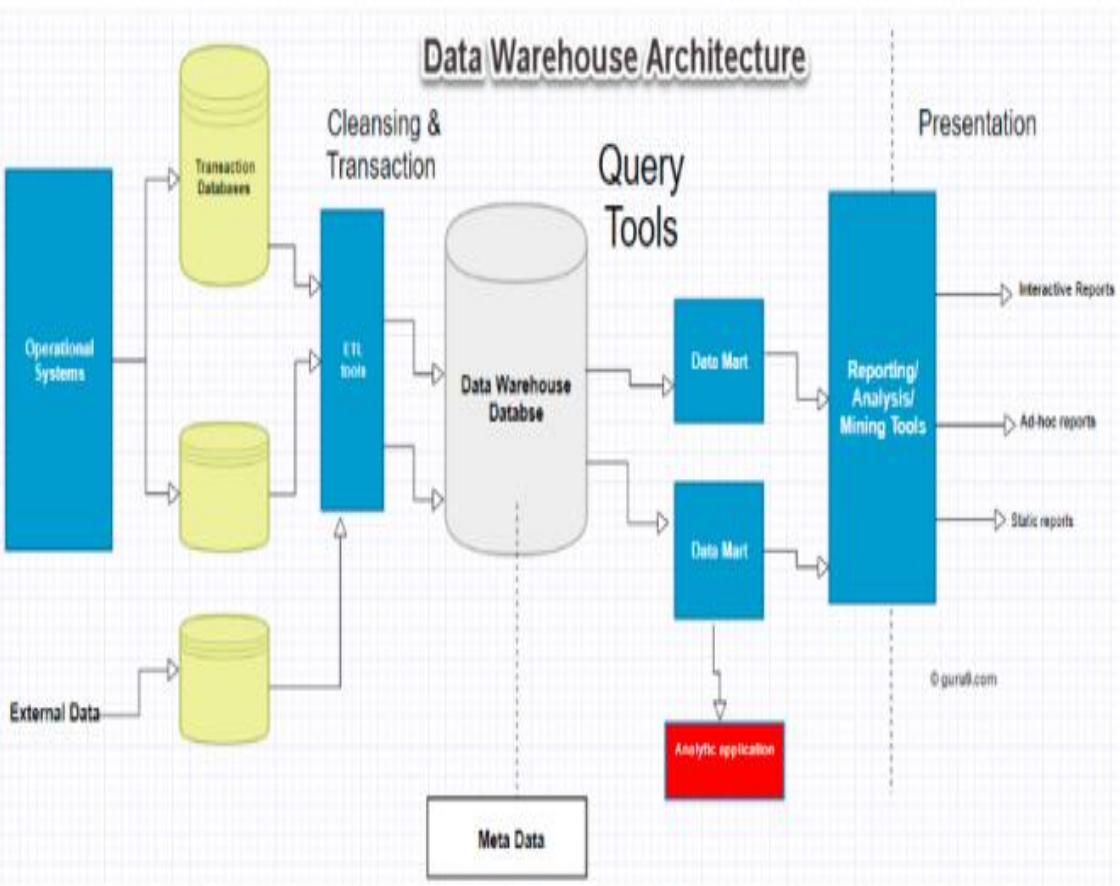
Types of Metadata

o Operational Metadata: Operational Metadata provides information about how the data is being used and the data lifecycle. It describes who has access to use it, who is using it, when it was created, and when it will be due to retire

o Extraction and Transformation Metadata Data on data removal from the source systems are included in extraction and transformation metadata, including frequencies extraction, extraction methods and data extraction regulations for business purposes. This category also provides details on all transformations in the data staging area.

o End-User Metadata The end-user metadata is the navigational map of the data warehouses. It enables the end-users to find data from the data warehouses. The end-user metadata allows the endusers to use their business terminology and look for the information in those ways in which they usually think of the business.

ARchitexcture



Data Warehouse Architecture