Course chapter 8:
Data Warehousing - Introduction

Some definitions

- Wikipedia:
  - **Data warehouse** is a repository of an organization's electronically stored data. Data warehouses are designed to facilitate reporting and analysis (from Inmon book they say – [1]).
  - A data warehouse houses a standardized, consistent, clean and integrated form of data sourced from various operational systems in use in the organization, structured in a way to specifically address the reporting and analytic requirements.

Some definitions

- R. Kimball (see [2, 3]):

A **data warehouse** is a copy of transactional data specifically structured for querying and analysis.

- According to this definition:
  - The form of the stored data (RDBMS, flat file) is not linked with the definition of a data warehouse.
  - Data warehousing is not linked exclusively with "decision makers" or used in the process of decision making.

Some definitions

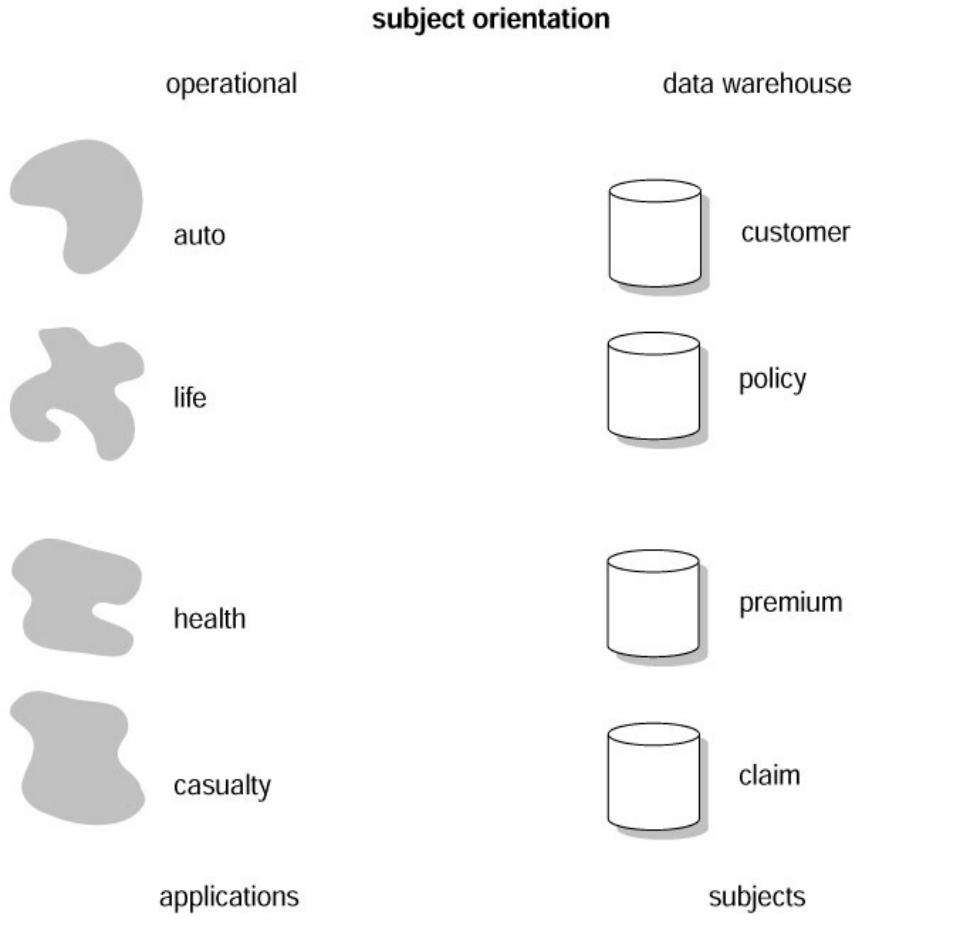
- Inmon (see [1, 3]):

A **data warehouse** is a

- subject-oriented,
- integrated,
- nonvolatile,
- time-variant

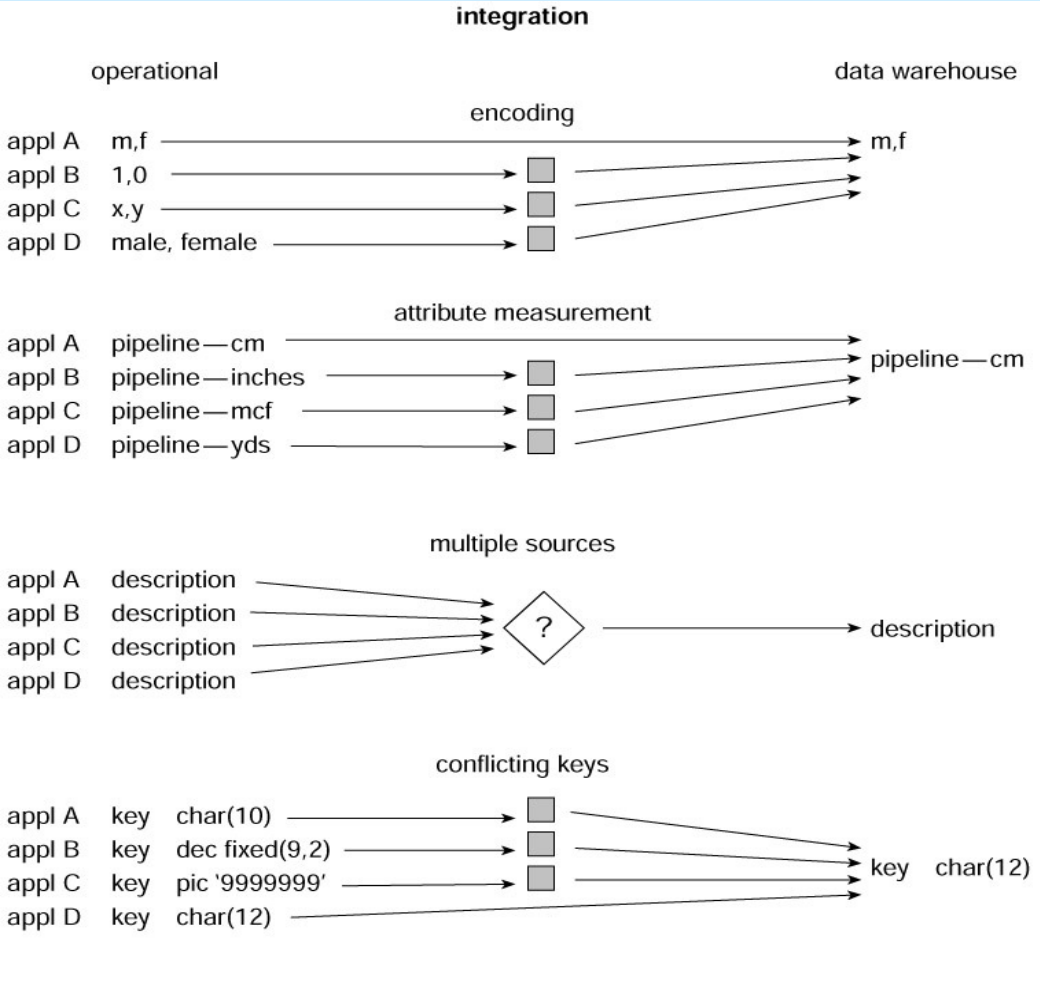
collection of data in support of management's decisions. The data warehouse contains granular corporate data.

Subject-Oriented Data Collections



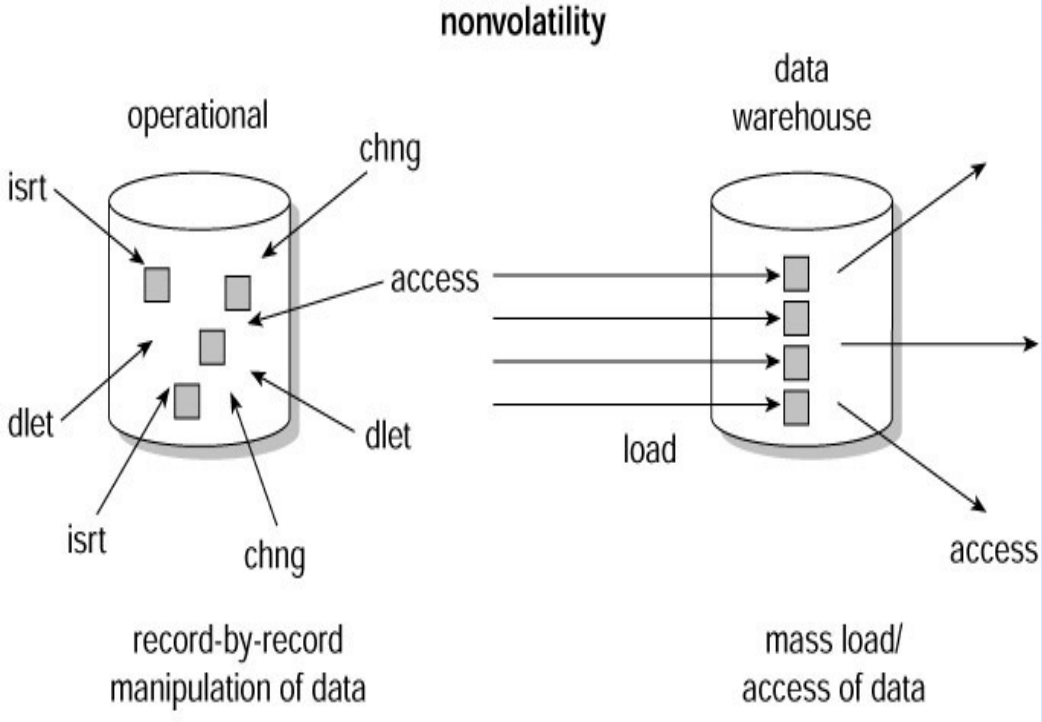
- Classical operations systems are organized around the applications of the company. For an insurance company, the applications may be auto, health, life, and casualty.
- The major subject areas of the insurance corporation might be customer, policy, premium, and claim.
- For a manufacturer, the major subject areas might be product, order, vendor, bill of material, and raw goods.
- For a retailer, the major subject areas may be product, SKU, sale, vendor, and so forth. Each type of company has its own unique set of subjects

Integrated Data Collections



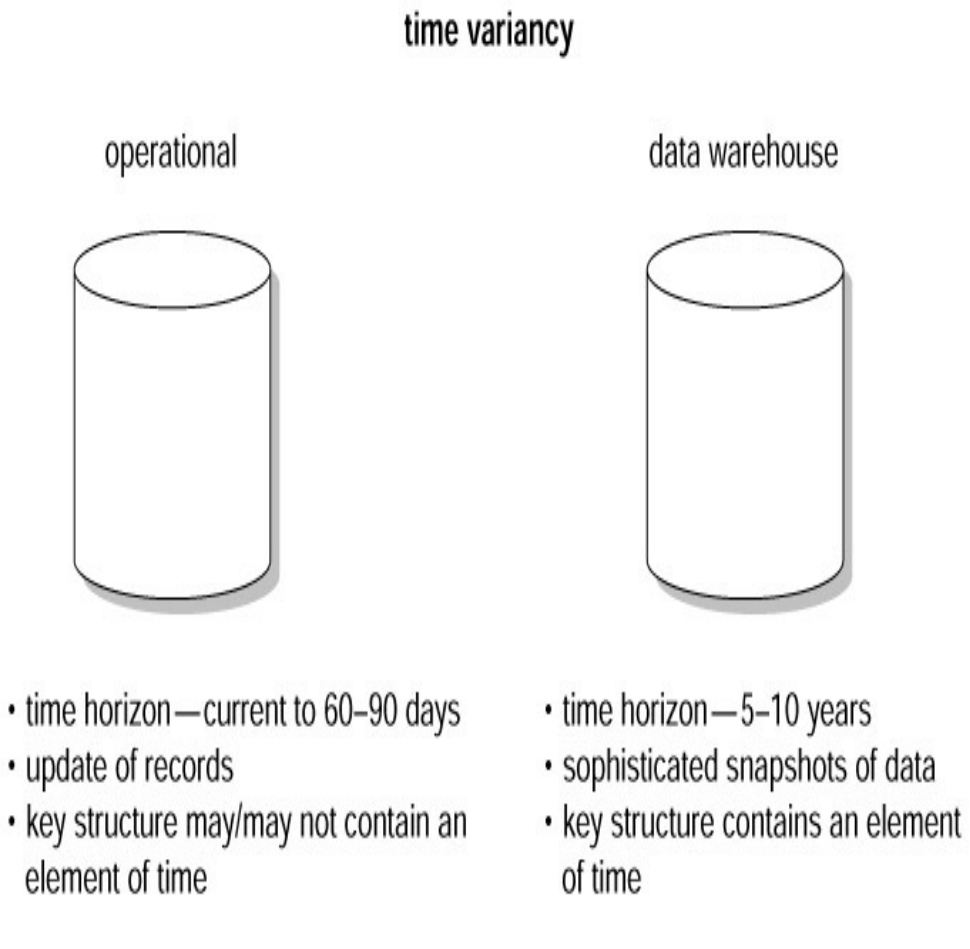
- Of all the aspects of a data warehouse, **integration** is the most important.
- Data is fed from multiple disparate sources into the data warehouse.
- As the data is fed it is converted, reformatted, resequenced, summarized, and so forth.
- The result is that data once it resides in the data warehouse has a single physical corporate image.

Non-volatile Data Collections



- Data is updated in the operational environment as a regular matter of course, but warehouse data exhibits a very different set of characteristics.
- Data warehouse data is loaded (usually en masse) and accessed, but it is not updated (in the general sense).
- Instead, when data in the data warehouse is loaded, it is loaded in a snapshot, static format.
- When subsequent changes occur, a new snapshot record is written.
- In doing so a history of data is kept in the data warehouse.

Time-variant Data Collections



•Time variancy implies that every unit of data in the data warehouse is accurate as of some ***one*** moment in time.

•In some cases, a record is time stamped. In other cases, a record has a date of transaction.

•But in every case, there is some form of time marking to show the moment in time during which the record is accurate.

•A 60-to-90-day time horizon is normal for operational systems; a 5-to-10-year time horizon is normal for the data warehouse.

•As a result of this difference in time horizons, the data warehouse contains *much* more history than any other environment.

Goals of a Data Warehouse

Problems that must be solved ([2]):

- "We have mountains of data in this company, but we can't access it."
- "We need to slice and dice the data every which way."
- "You've got to make it easy for business people to get at the data directly."
- "Just show me what is important."
- "It drives me crazy to have two people present the same business metrics at a meeting, but with different numbers."
- "We want people to use information to support more fact-based decision making."

Organization's information must be easily accessible

- The contents of the data warehouse must be understandable.
- The data must be intuitive and obvious to the business user, not merely the developer.
- The contents of the data warehouse need to be labeled meaningfully.
- Business users want to separate and combine the data in the warehouse in endless combinations, a process commonly referred to as *slicing and dicing*.
- The tools that access the data warehouse must be simple and easy to use. They also must return query results to the user with minimal wait times.

The data warehouse must present the organization's information consistently.

- Data must be assembled from a variety of sources around the organization, cleansed, quality assured, and released only when it is fit for user consumption.
- Information from one business process should match with information from another. If two performance measures have the same name, then they must mean the same thing. Conversely, if two measures don't mean the same thing, then they should be labeled differently.
- Consistency also implies that common definitions for the contents of the data warehouse are available for users.

The data warehouse must be adaptive and resilient to change

- User needs, business conditions, data, and technology are all subject to the shifting sands of time. The data warehouse must be designed to handle this inevitable change.
- Changes must not invalidate existing data or applications. The existing data and applications should not be changed or disrupted when the business community asks new questions or new data is added to the warehouse.

The data warehouse must protect our information assets

- An organization's informational assets are stored in the data warehouse. At a minimum, the warehouse likely contains information about what we're selling to whom at what price - potentially harmful details in the hands of the wrong people.
- The data warehouse must effectively control access to the organization's confidential information.

The data warehouse must serve as the foundation for improved decision making

- The data warehouse must have the right data in it to support decision making. There is only one true output from a data warehouse: the decisions that are made after the data warehouse has presented its evidence.
- These decisions deliver the business impact and value attributable to the warehouse.
- The original label that predates the data warehouse is still the best description of what we are designing: a **decision support system**.

The business community must accept the data warehouse

- It doesn't matter that we've built an elegant solution using best-of-breed products and platforms. If the business community has not embraced the data warehouse and has not continued to use it actively six months after training, then we have failed the acceptance test.
- Unlike an operational system rewrite, where business users have no choice but to use the new system, data warehouse usage is sometimes optional.
- Business user acceptance has more to do with **simplicity** than anything else.

Related concept: ODS

- An **operational data store** (or ODS) is a database designed to integrate data from multiple sources to make analysis and reporting easier.
- Because the data originates from multiple sources, the integration often involves cleaning, resolving redundancy and checking against business rules for integrity.
- An ODS is usually designed to contain low level or atomic (indivisible) data (such as transactions and prices) with limited history that is captured "real time" or "near real time" as opposed to the much greater volumes of data stored in the Data warehouse.

Related concept: ODS

- According to Bill Inmon, the originator of the concept, an ODS is "a subject-oriented, integrated, volatile, current-valued, detailed-only collection of data in support of an organization's need for up-to-the-second, operational, integrated, collective information."
- ODS differ from Inmon's definition of enterprise data warehouse by:
  - having a limited history,
  - more frequent update than an EDW.
- In practice ODS tend to be more reflective of source structures in order to speed implementations and provide a truer representation of production data

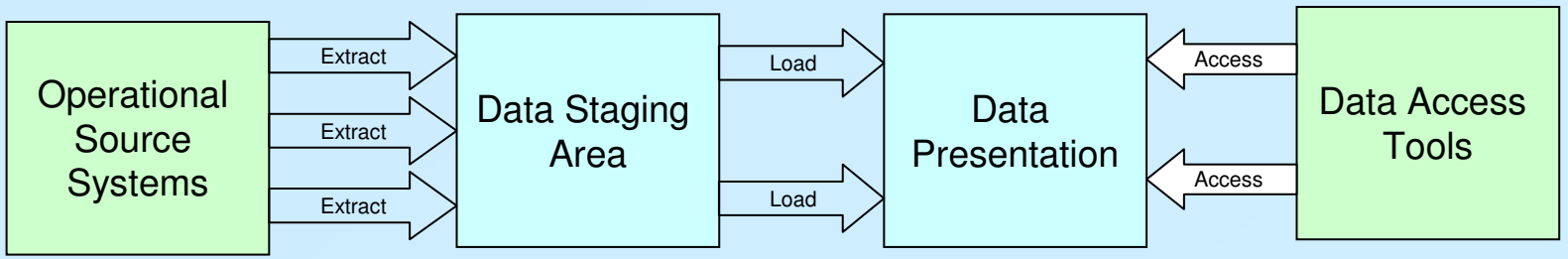
Related concept: ODS ([3])

The Operational Data Store is used for tactical decision making while the DW supports strategic decisions. It contains transaction data, at the lowest level of detail for the subject area

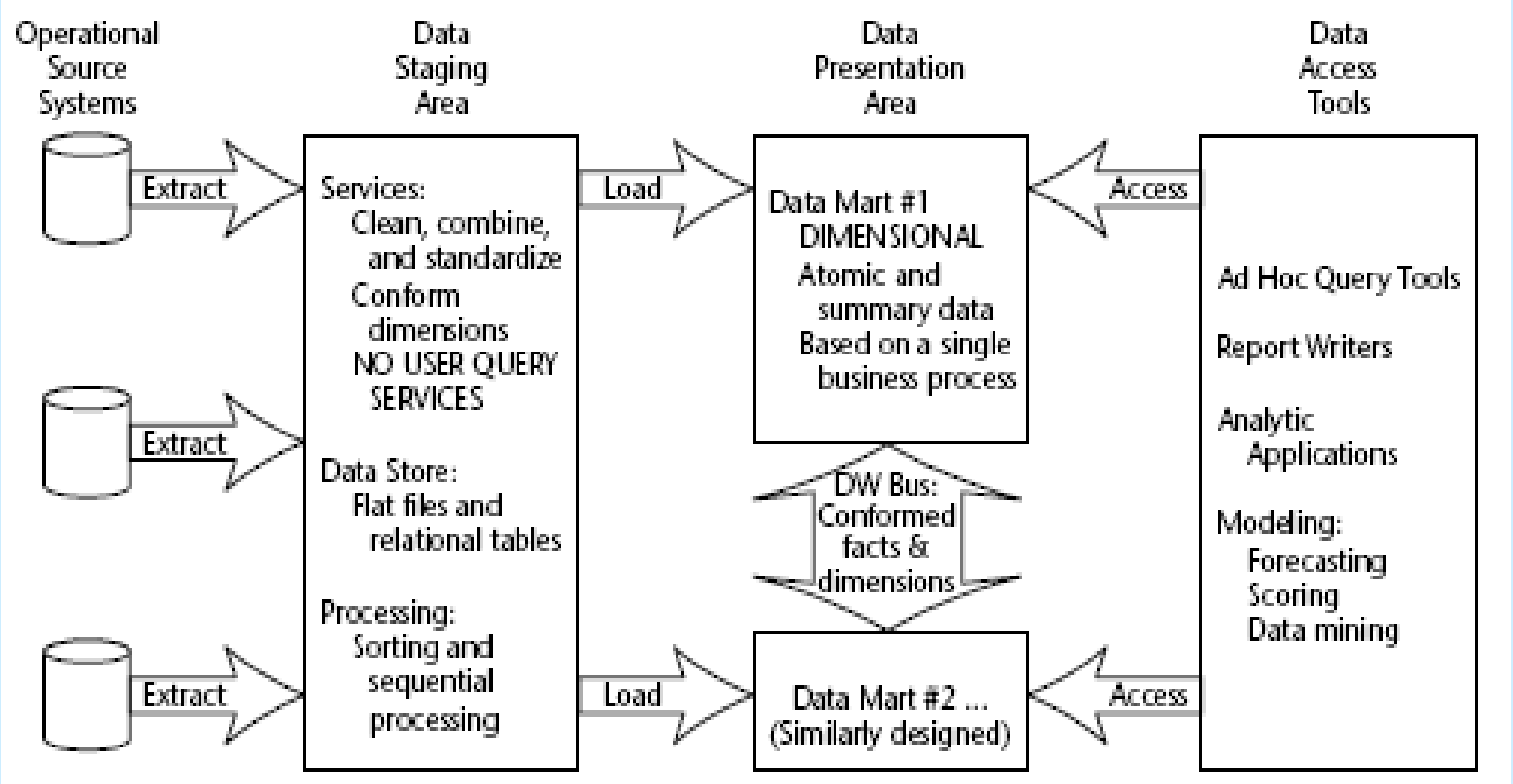
- subject-oriented, just like a DW
- integrated, just like a DW
- volatile (or updateable) , **unlike** a DW
  - an ODS is like a transaction processing system
  - information gets overwritten with updated data
  - no history is maintained (other than audit trail) or operational history
- current, i.e., not time-variant, unlike a DW
  - current data, up to a few years
  - no history is maintained (other than audit trail) or operational history

Basic elements of a DW

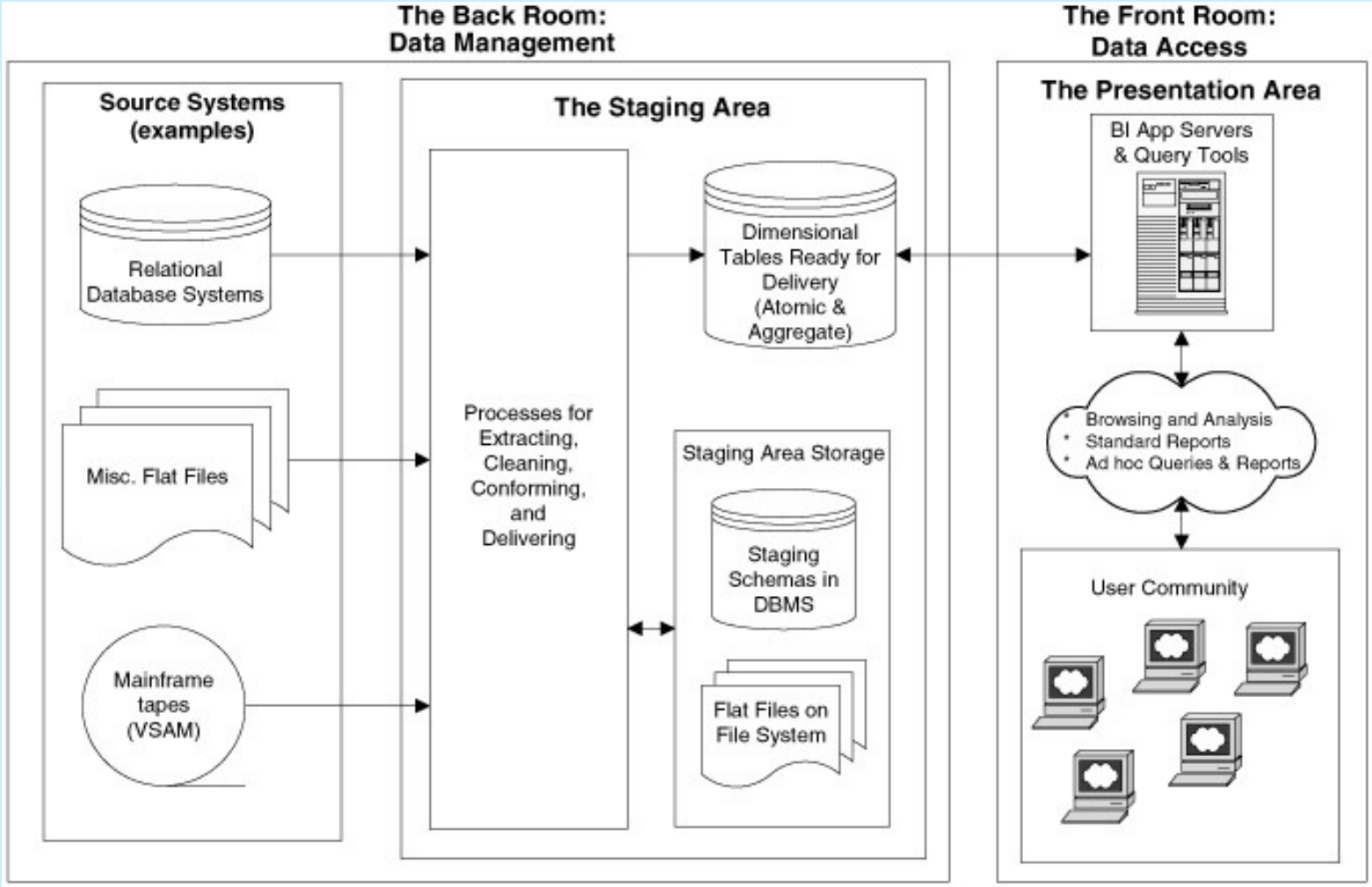
- Operational Source Systems
- Data Staging Area
- Data Presentation
- Data Access Tools



Basic elements of a DW ([2])



Basic elements of a DW ([3])



Operational Source Systems

- Also called Operational Database Layer
- These are the operational systems of record that capture the transactions of the business.
- The source systems should be thought of as outside the data warehouse because presumably we have little to no control over the content and format of the data in these operational legacy systems.
- The main priorities of the source systems are processing performance and availability.
- Queries against source systems are narrow, one-record-at-a-time queries that are part of the normal transaction flow and severely restricted in their demands on the operational system.

Operational Source Systems

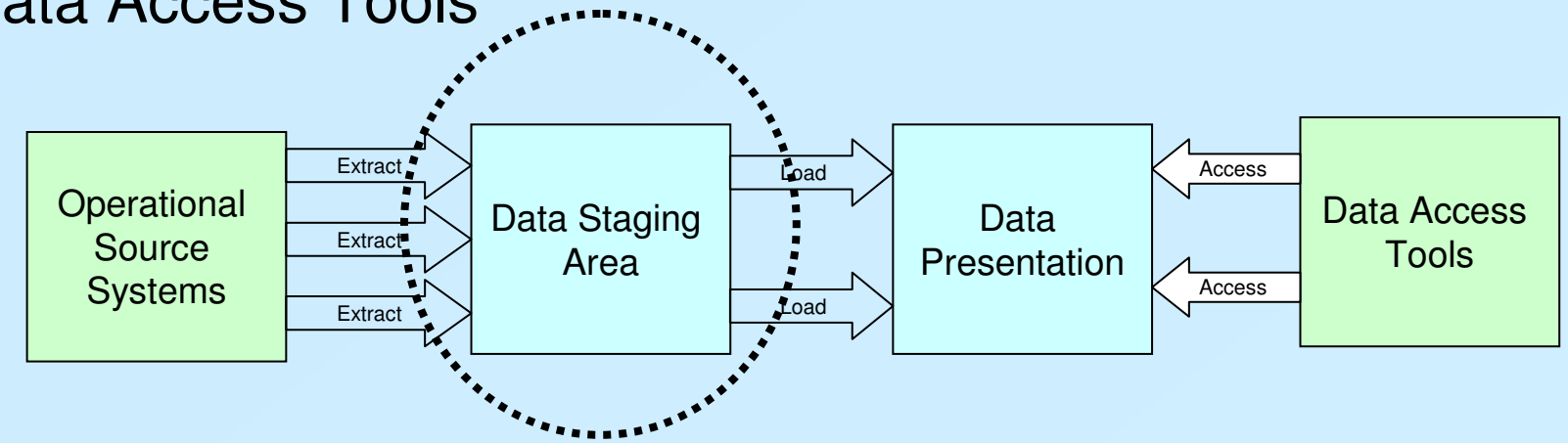
- The source systems maintain little historical data, and with a good data warehouse, the source systems can be relieved of much of the responsibility for representing the past.
- Each source system is often a natural stovepipe application[1], where little investment has been made to sharing common data such as product, customer, geography, or calendar with other operational systems in the organization.

---

[1] In engineering and computing, a **stovepipe system** is a legacy system that is an assemblage of inter-related elements that are so tightly bound together that the individual elements cannot be differentiated, upgraded or refactored. The stovepipe system must be maintained until it can be entirely replaced by a new system. The term is also used to describe a system that does not interoperate with other systems,

Basic elements of a DW

- Operational Source Systems
- Data Staging Area
- Data Presentation
- Data Access Tools



Data Staging Area

- The data staging area of the data warehouse is both a storage area and a set of processes commonly referred to as *extract-transformation-load* (ETL).
- The data staging area is everything between the operational source systems and the data presentation area.
- The key architectural requirement for the data staging area is that it is off-limits to business users and does *not* provide query and presentation services.
- Compared in [2] with the kitchen of a restaurant.

Data Staging Area - ETL

- ***Extraction*** is the first step in the process of getting data into the data warehouse environment.
- Extracting means reading and understanding the source data and copying the data needed for the data warehouse into the staging area for further manipulation.
- Once the data is extracted to the staging area, there are numerous potential ***transformations***, such as cleansing the data (correcting misspellings, resolving domain conflicts, dealing with missing elements, or parsing into standard formats), combining data from multiple sources, deduplicating data, and assigning warehouse keys.
- These transformations are all precursors to ***loading*** the data into the data warehouse presentation area.

Data Staging Area - Integration

- Data are not simply extracted
- Data across operational systems are severely un-integrated:
  - same data, different name
  - same name, different data
  - different keys, same data, etc
  - non-standard data encodings (male/female, 0/1, etc) and field transformations
  - different measurements, measurements at different levels of detail
  - different source system technologies (DB2, SQL Server, OS/390 DB2, VSAM, IMS, etc)

Data Staging Area - Storage

- There are two leading approaches to storing data in a data warehouse:
    1. The normalized approach (Inmon – [1])
    2. The dimensional approach (Kimball – [2])
- These approaches are not mutually exclusive, and there are other approaches. Dimensional approaches can involve normalizing data to a degree.
- This presentation is based on the dimensional approach and Kimball & Ross book– see [2].

Data Staging Area - Storage

- 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 main advantage of this approach is that it is straightforward to add information into the database.

Data Staging Area - Storage

- A disadvantage of this approach is that, because of the number of tables involved, it can be difficult for users both to
  1. join data from different sources into meaningful information and then
  2. access the information without a precise understanding of the sources of data and of the data structure of the data warehouse.

Data Staging Area - Storage

- In a ***dimensional approach***, transaction data are partitioned into ***facts*** (numeric transaction data), and ***dimensions*** (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 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.

Data Staging Area - Storage

- 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.
- The main disadvantages of the dimensional approach are:
  1. In order to maintain the integrity of facts and dimensions, loading the data warehouse with data from different operational systems is complicated, and
  2. It is difficult to modify the data warehouse structure if the organization adopting the dimensional approach changes the way in which it does business.

Data Staging Area

- The data staging area is dominated by the simple activities of sorting and sequential processing.
- In many cases, the data staging area is not based on relational technology but instead may consist of a system of flat files.
- After data validation for conformance with the defined one-to-one and many-to one business rules, it may be pointless to take the final step of building a fullblown third-normal-form physical database.

Data Staging Area

- However, there are cases where the data arrives at the doorstep of the data staging area in a third-normal-form relational format (remember 3rdNF?).
- In these situations, the managers of the data staging area simply may be more comfortable performing the cleansing and transformation tasks using a set of normalized structures.

Data Staging Area

- It is acceptable to create a normalized database to support the staging processes; however, this is not the end goal.
- The normalized structures must be off-limits to user queries because they defeat understandability and performance.
- As soon as a database supports query and presentation services, it must be considered part of the data warehouse presentation area.
- By default, in the dimensional approach normalized databases are excluded from the presentation area, which should be strictly dimensionally structured.

Data Staging Area - Load

- Regardless of the storage (a series of flat files or a normalized data structure) in the staging area, the final step of the ETL process is the ***loading*** of data.
- Loading in the data warehouse environment usually takes the form of presenting the quality-assured dimensional tables to the bulk loading facilities of each ***data mart***.
- The target data mart must then index the newly arrived data for query performance.

Data Staging Area - Load

- When each data mart has been freshly loaded, indexed, supplied with appropriate aggregates, and further quality assured, the user community is notified that the new data has been published.
- Publishing includes communicating the nature of any changes that have occurred in the underlying dimensions and new assumptions that have been introduced into the measured or calculated facts.
- The Data Marts are part of the Presentation Area (as showned in the figure from slide 20).

Data Marts ([5])

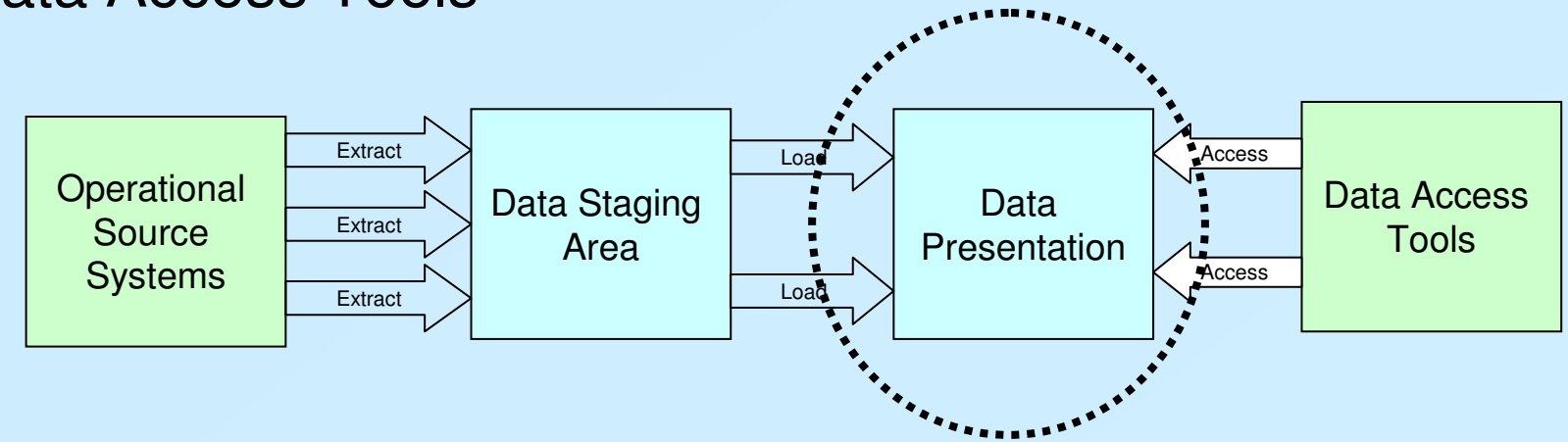
- A ***data warehouse*** is a central repository for all or significant parts of the data that an enterprise's various business systems collect. It enables ***strategic decision*** making.
- A ***data mart*** is a repository of data gathered from operational data and other sources that is designed to serve a particular community of knowledge workers.
- In scope, the data may derive from an enterprise-wide database or data warehouse or be more specialized. The emphasis of a data mart is on meeting the specific demands of a particular group of knowledge users in terms of analysis, content, presentation, and ease-of-use.

Data Marts ([4]) – Another definition

- A **data mart** is a subset of an organizational data store, usually oriented to a specific purpose or major data subject, that may be distributed to support business needs.
- Data marts are analytical data stores designed to focus on specific business functions for a specific community within an organization.
- Data marts are often derived from subsets of data in a data warehouse, though in the *bottom-up* data warehouse design methodology the data warehouse is created from the union of organizational data marts.

Basic elements of a DW

- Operational Source Systems
- Data Staging Area
- Data Presentation
- Data Access Tools



Data presentation

- The data presentation area is where data is organized, stored, and made available for direct querying by users, report writers, and other analytical applications.
- Since the backroom staging area is off-limits, the presentation area *is* the data warehouse as far as the business community is concerned. It is all the business community sees and touches via data access tools.
- We typically refer to the presentation area as a series of integrated data marts.
- A data mart is a wedge of the overall presentation area pie. In its most simplistic form, a data mart presents the data from a single business process.

Data presentation – Dimensional schemas

- In the presentation area, data must be presented, stored, and accessed in dimensional schemas.
- Most people find it intuitive to think of this business as a cube of data, with the edges labeled product, market, and time.
- We can imagine slicing and dicing along each of these dimensions. Points inside the cube are where the measurements for that combination of product, market, and time are stored.
- The ability to visualize something as abstract as a set of data in a concrete and tangible way is the secret of understandability.

Data presentation – Dimensional schemas

- Dimensional modeling addresses the problem complex schemas.
- A dimensional model contains the same information as a normalized model but in other format.
- Normalized modeling is immensely helpful to operational processing performance.
- Normalized models, however, are too complicated for data warehouse queries.
- Users can't understand, navigate, or remember complex normalized models (are you?).

Data presentation – Atomic data

- The presentation area data marts must contain detailed, atomic data.
- Atomic data is required for evaluating unpredictable ad hoc user queries.
- The data marts also may contain performance-enhancing summary data, or aggregates,
- It is not sufficient to deliver these summaries without the underlying granular data in a dimensional form.
- In other words, it is completely unacceptable to store only summary data in dimensional models while the atomic data is locked up in normalized models.

Data presentation – Conformed

- All the data marts must be built using common dimensions and facts, which we refer to as *conformed*.
- Without shared, conformed dimensions and facts, a data mart is a standalone stovepipe application.
- When data marts have been designed with conformed dimensions and facts, they can be combined and used together.

Data presentation – Conformed

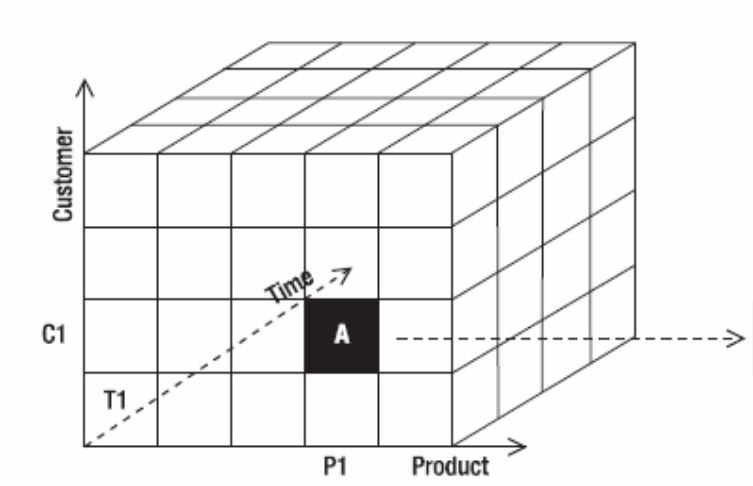
- The data warehouse presentation area in a large enterprise data warehouse ultimately will consist of 20 or more very similar-looking data marts.
- The dimensional models in these data marts also will look quite similar.
- Each data mart may contain several fact tables, each with 5 to 15 dimension tables.
- If the design has been done correctly, many of these dimension tables will be shared from fact table to fact table.

Data presentation – Star Schemas

- If the presentation area is based on a relational database, then these dimensionally modeled tables are referred to as star schemas (or snowflake schemas).
- If the presentation area is based on multidimensional database or online analytic processing (OLAP) technology, then the data is stored in cubes (or OLAP cubes).
- Dimensional modeling is applicable to both relational and multidimensional databases. Both have a common logical design with recognizable dimensions; however, the physical implementation differs.

Multidimensional database? ([6])

- A multidimensional database (abbreviated as MDB, MDD, or MDDB) is a form of database where the data is stored in cells and the position of each cell is defined by a number of hierarchical called *dimensions*.
- Each cell represents a business event, and the value of the dimensions shows when and where this event happened.



Multidimensional database? ([6])

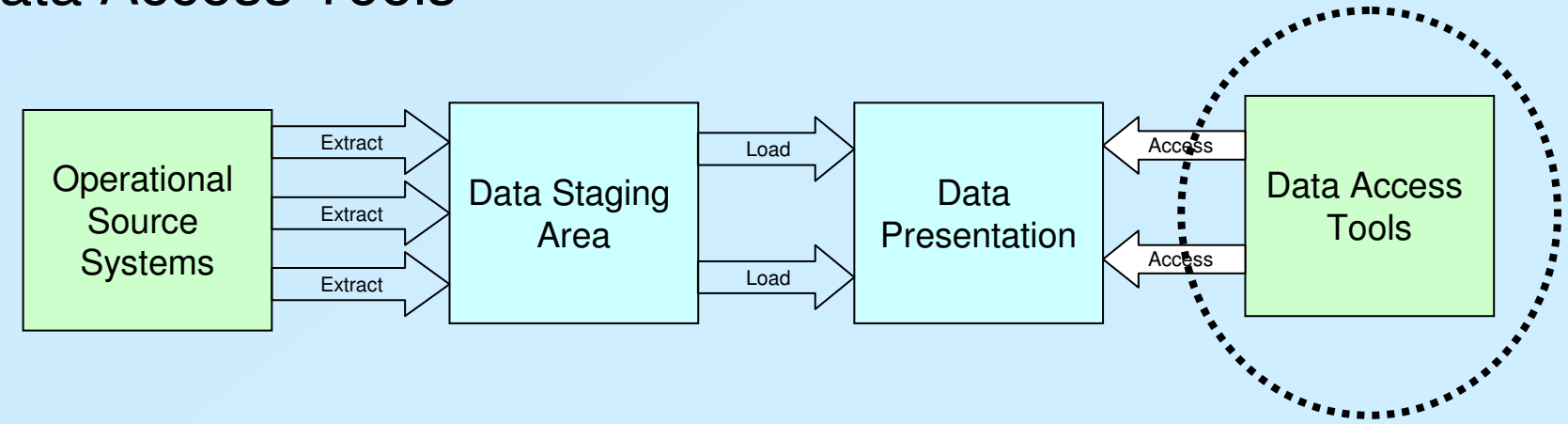
- The structure stores the aggregate values as well as the base values, typically in compressed multidimensional array format, rather than in RDBMS tables.
- Aggregate values are precomputed summaries of the base values.
- Other terms used for MDB: hypercube (more than 4 dimensions), cube, OLAP cube, and multidimensional data store (MDDS).

Multidimensional database? ([6])

- Physically, a MDB is a file.
- A multidimensional database occupies less disk space compared to a relational *dimensional* database because:
  - it is compressed and
  - it does not use indexing like a DDS.
- Instead, it uses multidimensional offsetting to locate the data.

Basic elements of a DW

- Operational Source Systems
- Data Staging Area
- Data Presentation
- Data Access Tools



Data Access Tools

- By definition, all data access tools query the data in the data warehouse's presentation area for analytic decision making.
- Querying, obviously, is the whole point of using the data warehouse.
- A data access tool can be as simple as an ad hoc query tool or as complex as a sophisticated data mining or modeling application.

Data Access Tools

- Ad hoc query tools can be understood and used effectively only by a small percentage of the potential data warehouse business user population.
- The majority of the business users will access the data via prebuilt parameter-driven analytic applications.

Data Access Tools

- Approximately 80 to 90 percent of the potential users will be served by these canned applications that are essentially finished templates that do not require users to construct relational queries directly.
- Some of the more sophisticated data access tools, like modeling or forecasting tools, actually may upload their results back into operational source systems or the staging/presentation areas of the data warehouse.

Dimensional Modeling Myths

- Myth #1: Dimensional models and data marts are for summary data only
  - can't predict all the queries users will ask, therefore information must be stored at the detail level in a DW and summarized levels for a DM
- Myth #2: Dimensional models and data marts are departmental, not enterprise, solutions
- Myth #3: Dimensional models and data marts are not scalable
- Myth #4: Dimensional models and data marts are only appropriate when there is a predictable usage pattern
- Myth #5: Dimensional models and data marts can't be integrated and therefore lead to stovepipe solutions.

Dimensional Modeling Pitfalls

- **Pitfall 10.** Become overly enamored with technology and data rather than focusing on the business's requirements and goals.
- **Pitfall 9.** Fail to embrace or recruit an influential, accessible, and reasonable management visionary as the business sponsor of the data warehouse.
- **Pitfall 8.** Tackle a galactic multiyear project rather than pursuing more manageable, while still compelling, iterative development efforts.
- **Pitfall 7.** Allocate energy to construct a normalized data structure, yet run out of budget before building a viable presentation area based on dimensional models.
- **Pitfall 6.** Pay more attention to backroom operational performance and ease of development than to front-room query performance and ease of use.

Dimensional Modeling Pitfalls

- **Pitfall 5.** Make the supposedly query-able data in the presentation area overly complex. Database designers who prefer a more complex presentation should spend a year supporting business users; they'd develop a much better appreciation for the need to seek simpler solutions.
- **Pitfall 4.** Populate dimensional models on a standalone basis without regard to a data architecture that ties them together using shared, conformed dimensions.
- **Pitfall 3.** Load only summarized data into the presentation area's dimensional structures.
- **Pitfall 2.** Presume that the business, its requirements and analytics, and the underlying data and the supporting technology are static.
- **Pitfall 1.** Neglect to acknowledge that data warehouse success is tied directly to user acceptance. If the users haven't accepted the data warehouse as a foundation for improved decision making, then your efforts have been exercises in futility.

Bibliography

1. W.H. Inmon - Building The Data Warehouse. Third Edition, Wiley & Sons, 2002
2. Ralph Kimball, Margy Ross - The Data Warehouse Toolkit, Second Edition, Wiley & Sons, 2002
3. Dimitra Vista, CS 680 Course notes
4. Wikipedia – Pages on Data Warehouse, etc.
5. http://searchsqlserver.techtarget.com/
6. Vincent Rainardi, Building a Data Warehouse with Examples in SQL Server, Springer, 2008