# 3 Data Warehousing

Learning Objectives

* Understand the concept of Data Warehouse and its role in data analytics
* Appreciate the design choices in developing a DW
* Understand the top-down and bottom approaches of developing a DW and their various advantages
* Discuss the architecture of a DW
* Understand the type and range of data sources
* Understand the Extract-Transform-Load (ETL) processes
* Explain the star schema for organizing a DW
* Appreciate the key best practices for designing an effective DW

### INTRODUCTION

A Data Warehouse (DW) is an organized collection of integrated, subject-oriented databases designed to aid decision support functions. DW is organized at the right level of granularity to provide clean enterprise-wide data in a standardized format for reports, queries, and analysis. DW is physically and functionally separate from an operational and transactional database. Creating a DW for analysis and queries represents significant investment in time and effort. It must be constantly kept up to date for it to be useful. DW offers many business and technical benefits.

DW supports business reporting and data mining activities. It can facilitate distributed access to up-to-date business knowledge for departments and functions, thus improving business efficiency and customer service. DW can present a competitive advantage by facilitating decision making and helping reform business processes.

DW enables a consolidated view of corporate data, all cleaned and organized. Thus, the entire organization can see an integrated view of itself. DW thus provides better and timely information. It simplifies data access and allows end users to perform extensive analysis. It enhances overall IT performance by not burdening the operational databases used by Enterprise Resource Planning (ERP) and other systems.

#### Caselet: University Health System – BI in Healthcare

*Indiana University Health (IUH), a large academic healthcare system, decided to build an enterprise data warehouse (EDW) to foster a genuinely data-driven management culture. IUH hired a data warehousing vendor to develop an EDW which also integrates with their Electronic Health Records (EHR) system. They loaded 14 billion rows of data into the EDW—fully 10 years of clinical data from across IUH’s network. Clinical events, patient encounters, lab and radiology, and other patient data were included, as were IUH’s performance management, revenue cycle, and patient satisfaction data. They soon put in a new interactive dashboard using the EDW that provided IUH’s leadership with the daily operational insights they need to solve for the quality/cost equation. It offers visibility into key operational metrics and trends to easily track the performance measures critical to controlling costs and maintaining quality. The EDW can easily be used across IUH’s departments to analyze, track and measure clinical, financial, and patient experience outcomes. (Source: *healthcatalyst.com*)*

1. *What are the benefits of a single large comprehensive EDW?*
2. *What kinds of data would be needed for an EDW for an airline company?*

### DESIGN CONSIDERATIONS FOR DW

The objective of a DW is to provide business knowledge to support decision making. For a DW to serve its objective, it should be aligned around those decisions. It should be comprehensive, easy to access, and up to date. Following are some requirements for a good DW

*Subject oriented* To be effective, a DW should be designed around a subject domain, i.e., to help solve a certain category of problems.

*Integrated* The DW should include data from many functions that can shed light on a particular subject area. Thus, the organization can benefit from a com- prehensive view of the subject area.

*Time-variant (time series)* The data in a DW should grow at daily or other chosen intervals. This allows latest comparisons over time.

*Nonvolatile* DW should be persistent, that is, it should not be created on the fly from the operations databases. Thus, a DW is consistently available for analysis, across the organization and over time.

*Summarized* DW contains rolled-up data at the right level for queries and analysis. The process of rolling up the data helps create consistent granularity for effective comparisons. It also helps reducing the number of variables or dimensions of the data to make it more meaningful for the decision makers.

*Not normalized* DW often uses a star schema, which is a rectangular central table, surrounded by some lookup tables. The single table view significantly enhances speed of queries.

*Metadata* Many of the variables in the database are computed from other variables in the operational database. For example, total daily sales may be a computed field. The method of its calculation for each variable should be effectively documented. Every element in the DW should be sufficiently well-defined.

*Near Real-time and/or Right-time (Active)* DWs should be updated in near real-time in many high transaction volume industries, such as airlines. The cost of implementing and updating a DW in real time could be discouraging though. Another downside of real-time DW is the possibilities of inconsistencies in reports drawn just a few minutes apart.

### DW DEVELOPMENT APPROACHES

There are two fundamentally different approaches to develop a DW – top-down and bottom-up. The top-down approach is to make a comprehensive DW that covers all the reporting needs of the enterprise. The bottom-up approach is to produce small data marts, for the reporting needs of different departments or functions, as needed. The smaller data marts will eventually align to deliver comprehensive EDW capabilities. The top-down approach provides consistency but takes more time and resources. The bottom-up approach leads to healthy local ownership and maintainability of data.

Table Comparing Data Mart and Data Warehouse

|  |  |  |
| --- | --- | --- |
|  | Functional Data Mart | Enterprise Data Warehouse |
| Scope | One subject or functional area | Complete enterprise data needs |
| Value | Functional area reporting and insights | Deeper insights connecting multiple functional areas |
| Target organization | Decentralized management | Centralized management |
| Time | Low to medium | High |
| Cost | Low | High |
| Size | Small to medium | Medium to large |
| Approach | Bottom-up | Top-down |
| Complexity | Low (fewer data transformations) | High (data standardization) |
| Technology | Smaller scale servers and databases | Industrial strength |

### DW ARCHITECTURE

DW has four key elements (Figure 3.1). The first element is the data source that provides the raw data. The second element is the process of transforming that data to meet the decision needs. The third element is the method of regularly and accurately loading of that data into EDW or data marts. The fourth element is the data access and analysis part, where devices and applications use the data from DW to deliver insights and other benefits to users.

Data Source Operations



–ERP systems

–Legacy systems

–Point of Sale

–RFID systems

–Web usage

External Source

–Suppliers

–Customers

–Government

Data Transformation

–Select Data

–Extract Data

–Cleanse Data

–Compute Data fields

–Integrate Data

–Load data

Data Mart or Warehouse

Rectangularized data ready for analysis

–One for each department, Or

–One for the whole enterprise

Accessing Users and Applications

–OLAP tools

–Reporting Tools

–Dashboards

–Queries

–Mobile Devices

–Data Mining

–Custom apps

FIGURE 3.1 Data Warehousing Architecture

### Data Sources

Data warehouses are created from structured data sources. Unstructured data such as text data would need to be structured before inserted into the DW.

*Operations Data* This includes data from all business applications, including from ERPs systems that form the backbone of an organization’s IT systems. The data to be extracted will depend upon the subject matter of the data warehouse. For example, for a sales/marketing datamart, only the data about customers, orders, customer service, and so on would be extracted.

*Specialized Applications* This includes applications such as Point of Sale (POS) terminals, and e-commerce applications that also provide customer-facing data. Supplier data could come from Supply Chain Management systems. Planning and budget data should also be added as needed for making comparisons against targets.

*External Syndicated Data* This includes publicly available data such as weather or economic activity data. It can also be added to the DW, as needed, to provide good contextual information to decision makers.

### Data Loading Processes

The heart of a useful DW is the process that populates the DW with good quality data. This is called the Extract-Transform-Load (ETL) cycle.

* Data should be extracted from the operational (transactional) database sources, as well as from other applications, on a regular basis.
* The extracted data should be aligned together by key fields and integrated into a single dataset. It should be cleansed of any irregularities or missing values. It should be rolled-up together to the same level of granularity. De- sired fields, such as daily sales totals, should be computed. The entire data should then be brought to the same format as the central table of the DW.
* This transformed data should then be uploaded into the DW.

This ETL process should be run at a regular frequency. Daily transaction data can be extracted from ERPs, transformed, and uploaded to the database the same night. Thus, the DW is up to date every morning. If a DW is needed for near- real-time information access, then the ETL processes would need to be executed more frequently. ETL work is usually done using automated programming scripts that are written, tested, and then deployed for periodically updating the DW.

### Data Warehouse Design

Star schema is the preferred data architecture for most DWs. There is a central fact table that provides most of the information of interest. There are lookup tables that provide detailed values for codes used in the central table. For ex- ample, the central table may use digits to represent a salesperson. The lookup table will help provide the name for that salesperson code. Here is an example of a star schema for a data mart for monitoring sales performance (Figure 3.2).

Other schemas include the snowflake architecture. The difference between a star and snowflake is that in the latter, the lookup tables can have their own further lookup tables.

There are many technology choices for developing DW. This includes selecting the right database management system and the right set of data management tools. There are a few big and reliable providers of DW systems. The provider of the operational DBMS may be chosen for DW also. Alternatively, a best-of-breed DW vendor can be used. There are also a variety of tools out there for data migration, data upload, data retrieval, and data analysis.

Item id Item name

Region id Name

...

...

...

Orders Data

item id Sales

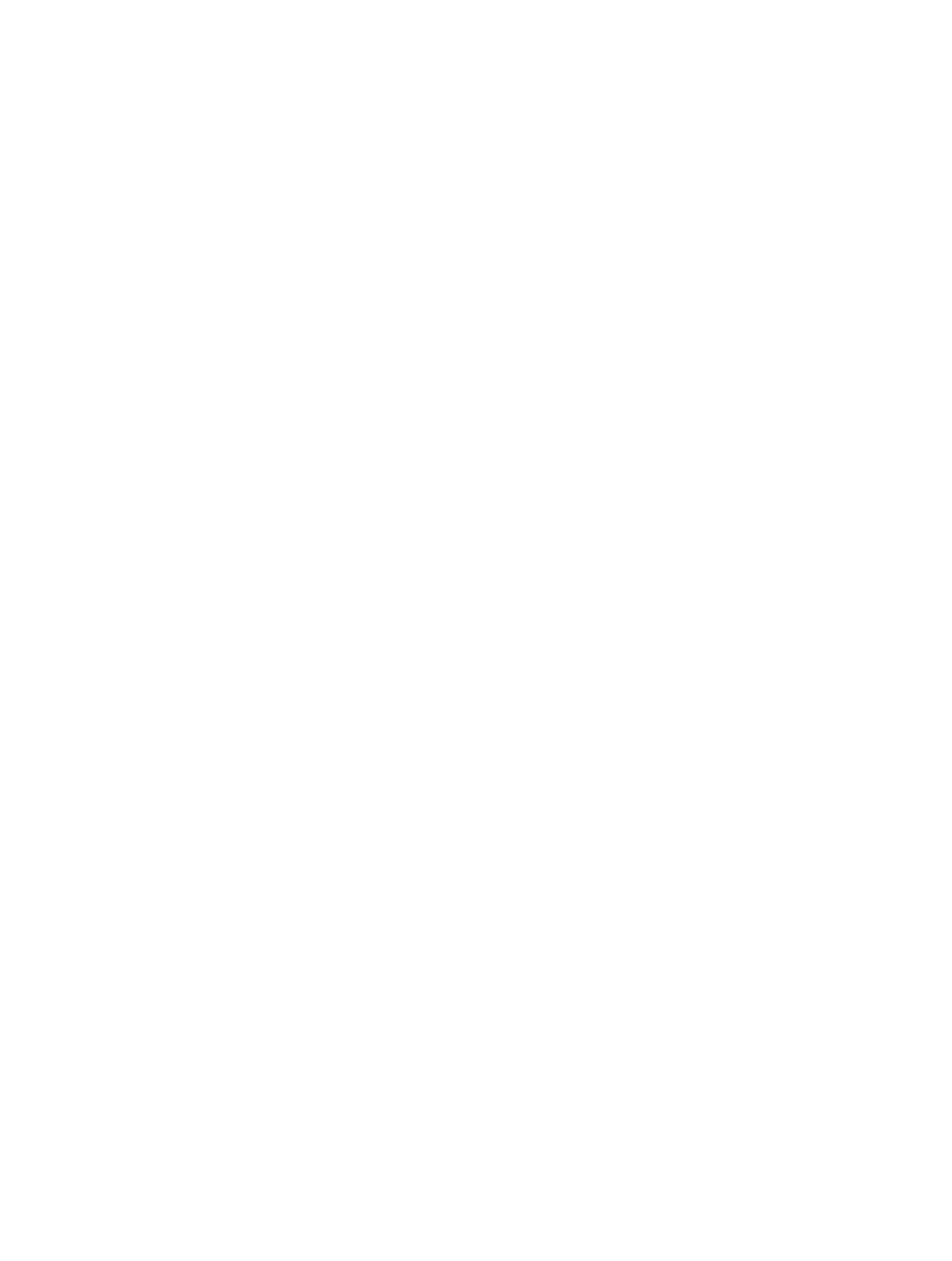
Person ID

Region ID

Period id

Total Sales

Sales Quota



Period id

...

Month/yr

...

...

...

Sales Name Person id

...

...

FIGURE: Star Schema Architecture for DW

### DW Access

Data from a DW can be accessed for many purposes, by many users, through different devices and systems. The primary use of a DW is to produce routine management and monitoring reports.

* OLAP (Online Analytical Processing) systems are query and reporting tools that help to slice the multidimensional data in a DW and produce desired reports. For example, a sales performance report would show sales by many dimensions compared with the plan; or one can request a list of all the products sold in a region in a certain price range during a certain time period and have them displayed in a report format.
* A dashboarding system uses data from the warehouse and present analysis to the users. The data from a DW can be used to populate customized performance dashboards for executives. The dashboard could include drill-down capabilities to analyze the performance data for root cause analysis.
* The data from a DW can be used for ad-hoc queries and any other applications that make use of the internal data.
* Data from a DW is used to provide data for mining purposes. Parts of the data would be extracted and then combined with other relevant data for data mining.

### DW Best Practices

A data warehousing project reflects a significant investment into information technology (IT). All the best practices in implementing any IT project should be followed.

* A data warehouse project should *align with the corporate strategy*. Top management should be consulted for setting objectives. Financial viability (ROI) should be established. The project must be managed by both IT and business professionals. The DW design should be carefully tested before beginning the development work. It is often much more expensive to redesign after development work has begun.
* It is important to *manage user expectations*. The data warehouse should be built incrementally. Users should be trained in using the system so they can absorb many features of the system.
* *Quality and adaptability* should be built-in from the start. Only relevant, cleansed, and high-quality data should be loaded. The system should be able to adapt to new tools for access. As business needs change, new data marts may need to be created for new needs.

## Conclusion

Data Warehouses are special data management facilities intended for creating reports and analysis to support managerial decision making. They are designed to make reporting and querying simple and efficient. The sources of data are operational systems and external data sources. A data warehouse needs to be updated with new data regularly to keep it useful. Data from a DW provides a useful input for data mining activities.

## Questions

1. What is the purpose of a data warehouse?

**Data Warehouse’s purpose is for creating reports and analysis to support managerial decision making. They are designed to make reporting and querying simple and efficient.**

1. What are the key elements of a data warehouse? Describe each one of them.

Data Warehouse has 4 key elements

* + - Data sources that provide the raw data.
    - The process of transforming that data to meet the decision needs.
    - The methods of regularly and accurately loading of that data into EDW or data marts.
    - Data access and analysis part, where devices and applications use the data from DW to deliver insights and other benefits to users

1. What are the sources and types of data for a data warehouse?

**Operational systems and External data sources**

1. How will data warehousing evolve in the age of social media?

Due to increased number of people connected to internet mobile phones and computers, more and more information and being available. The signals, data is so vast that the need for data warehousing will also increase in requirements. Software's will be more powerful and intuitive to filter out the noise and prepare the data for better structured format information.

## True/False

1. A data warehouse can help the entire organization to see an integrated view of itself. **True**
2. There are three approaches to build a data warehouse – top-down, bottom- up, and global. **False**
3. ETL stands for Extract, Transfer and Label the data into the DW. **False**
4. A star schema consists of a central fact table and smaller lookup tables.

**True**

1. A data warehouse is an expensive project and should align with corporate strategy. **True**
2. Once created, the data warehouse should be protected from adding data into it. **True**
3. Social media data such as customer comments should be directly transferred into the data warehouse to allow for quick insights. **False**
4. A data warehouse contains metadata about how data is organized and how to use it effectively. **False**
5. POS is not a data source to a data warehouse. **False**
6. Data warehouse is a valuable source of data for data mining. **True**