IS313 Data Warehousing

Dr.Waleed M.Ead

1

Course References

- Paulraj Ponniah, **Data Warehousing Fundamentals For It Professionals** (Second Edition) ; John Wiley & Sons Inc., NY.
- Naveen Prakash and Deepika Prakash; **Data Warehouse Requirements Engineering**; Springer Nature Singapore Pte Ltd.
- Alejandro Vaisman; **Data Warehouse Systems Design and Implementation**; Springer Nature Singapore Pte Ltd.
- Ralph Kimball and Margy Ross, **The Data Warehouse Toolkit (Second Edition),** John Wiley & Sons Inc., NY.
- W.H. Inmon, **Building the Data Warehouse (Second Edition),** John Wiley & Sons Inc., NY.
- **Practical part**
  - **SAS Viya**
  - Assigned projects

2

EXPECTATION

3

Related tools

4

Data Warehouse ,why?

- Analyzing data from databases that support line-of-business (LOB) applications is usually not an easy task.
- The normalized relational schema used for an LOB application can consist of thousands of tables.
- Naming conventions are frequently not enforced.
- Therefore, it is hard to discover where the data you need for a report is stored.

5

Data Warehouse ,why?

- Enterprises frequently have multiple LOB applications, often working against more than one database.
- For the purposes of analysis, these enterprises need to be able to merge the data from multiple databases.
- Data quality is a common problem as well.
- In addition, many LOB applications do not track data over time, though many analyses depend on historical data.
- Transactional systems deals with delivering system functionality in the hands of the user.

6

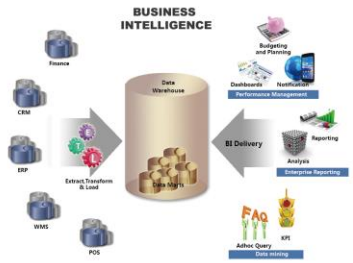
Data Warehouse ,why?

- Data warehouse systems supply information to their users who are decision-makers, so that they could take appropriate decisions.
- These decisions are made after decision-makers carry out suitable analysis of the information retrieved from the data warehouse.

A common solution to these problems is to create a *data warehouse (DW)*.

7

Data Warehousing and BI Services



8

Data Warehouse perspectives

1. The organizational

- data warehouse technology is for providing service to the organization.
- it provides Business Intelligence, BI.
- The Data Warehouse Institute considers BI in three parts, namely,
  - data warehousing,
  - tools for business analytics, and
  - knowledge management.

9

Data Warehouse perspectives

- The value of BI is realized as profitable business action.
- This means that BI is of little value if knowledge that can be used for profitable action is ignored.
- Conversely, if discovered knowledge is not realized into a value-producing action, then it is of little value.
- Thus, managers should be able to obtain the **specific information** that helps in making the **optimal decision** so that **specific actions can be taken**.
- It follows that Business Intelligence incorporates the tools, methods, and processes needed to transform data into actionable knowledge.

10

Data Warehouse perspectives

2. Technological point of view

According to Inmon definition:

A data warehouse is not just a storehouse of data but is an environment or infrastructure for decision-making.

11

Data warehouse and a database

- The data warehouse supports Online Analytical Processing, OLAP,
- Database is for Online Transaction Processing, OLTP.
- A database contains in it data of all transactions that were performed during business operations.
- Thus, for example, data of every order received is available in the database.
- If modification of the order occurred, then the modified data is available.
- In this sense, a database is an image, a snapshot of the state of the business at a given moment, T.
- Databases do not maintain historical data but reflect data at current time T only

12

Data warehouse and a database

- The purpose of the data warehouse is to provide information to facilitate making a business decision.
- Interest is in analyzing the state of the business at time t (this may include current data at t as well as historical data) so as to determine
  - what went wrong and needs correction,
  - what to promote,
  - what to optimize and, in general,
  - to decide how to make the business perform better.

13

Data warehouse and a database

- The state of the business lies in the collection of data sources of the business, the several databases, files, spreadsheets, documents, emails, etc. at time t.
- In other words, we need a different model of data than the database model.
- This OLAP model enables data to be viewed and operated upon to promote analysis of business data.
- A data warehouse provides a multidimensional view of data.
- Data is viewed in terms of **facts** and **dimensions**,
  - a **fact** being the basic data that is to be analyzed,
  - whereas **dimensions** are the various parameters along which facts are analyzed

14

Data warehouse and a database

- A data warehouse provides a multidimensional view of data.
- Data is viewed in terms of **facts** and **dimensions**,
  - a **fact** being the basic data that is to be analyzed,
  - whereas **dimensions** are the various parameters along which facts are analyzed
- Both facts and dimensions have their own attributes.

Thus, sales data expressed as number of units sold or in revenue terms (rupees, dollars) is basic sales data that can be analyzed by location, customer profile, and time.

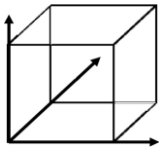
➢ location, customer profile, and time are the dimensions.

15

Data warehouse and a database

- Both facts and dimensions have their own attributes.

Thus, sales data expressed as number of units sold or in revenue terms (rupees, dollars) is basic sales data that can be analyzed by location, customer profile, and time.



- location, customer profile, and time

Each cell in the cube contains sales data, i.e., units sold or revenue.

It is possible for attributes of dimensions to be organized in a hierarchy.

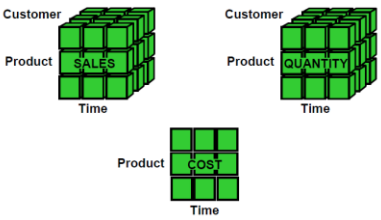
For example, the attributes month, quarter, half year, and

year of the dimension time form a hierarchy.

Monthly facts can be aggregated into quarterly, half-yearly, and yearly facts, respectively.

16

Sharing Dimensions



17

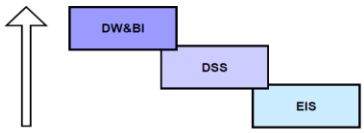
Data warehouse and a database

- A major difference between data warehouse development and database development
  - Is due to the need in the former to draw data from disparate data sources.
  - This is done in the Extraction Transformation and Loading, ETL, step where data is taken from the different sources, standardized, any inconsistencies removed, and thereafter the data is brought into multidimensional form. This "cleaned up" data is then loaded in the data warehouse.

18

Evolution of BI

- Executive information systems (EIS)
- Decision support systems (DSS)
- Data warehousing (DW) and business intelligence (BI)

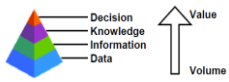


19

Business Intelligence (BI): Definition and Purpose

"Business intelligence is the process of transforming data into information and through discovery transforming that information into knowledge." – Gartner Group

The purpose of business intelligence is to convert the volume of data into business value through analytical reporting.



20

Business Intelligence (BI): Definition and Purpose

- From an information systems standpoint, BI provides users with online analytical processing or data analysis capabilities to predict trends, evaluate business questions, and so on.
- From a BI analyst viewpoint, it is the process of gathering high-quality, meaningful information about a subject, which enables the analyst to draw conclusions.
- Data warehousing creates the infrastructure for providing successful enterprise-level BI.

21

What Is Business Intelligence?

- An effective BI solution provides answers to important business questions.
- How are sales year-to-date and how do they compare to last year?
- Who is most likely to respond to my current marketing campaign and how will they impact revenue?
- What is the turnover in employees compared to the last five years?
- How is potential fraud cost being managed over time?
- What are my most profitable products by region, by year, and year-to-date?

What about the future ..?

22

Success Factors for a Dynamic Business Environment

To succeed in an ever-changing business environment, a company must:

- Know both the market that they are in and their business (internally and externally)
- Reinvent themselves to face new challenges. This may be changing product requirements, diverse and effective services, or even changes in internal organizational structures.
- Invest in research and development of new product channels
- Invest in high-value customers who contribute greater returns to the business

23

Success Factors for a Dynamic Business Environment

- Retain existing customers and attract new customers
- Invest in new technology to support business needs
- Improve access to information so that they can make rapid decisions, based on an accurate picture of the business
- Provide superior services and products to keep market share and maintain income
- Be profitable—at the same time, they must be able to invest in resources for the future, such as technology and people

24

Business Intelligence: Requirements

To address the changing requirements of today's business economy, business intelligence systems require that the following business requirements be addressed:

- Efficient design of data warehouses
- Enterprise reporting
- Ad hoc query and analysis
- Advanced analytics
- Integration with portals
- Easy administration
- Integrated environment or tools

25

Quiz 1

26

Lec2

Defining Data Warehouse Concepts and Terminology

Dr.Waleed M.Ead,Ph.D

27

Data Warehouse: Definition

"A data warehouse is a subject-oriented, integrated, nonvolatile, and time-variant collection of data in support of management's decisions."

— W.H. Inmon

"An enterprise-structured repository of subject-oriented, time variant, historical data used for information retrieval and decision support. The data warehouse stores atomic and summary data."

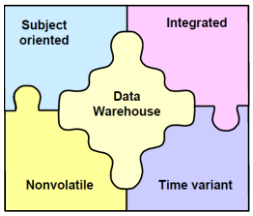
— Oracle's definition of a data warehouse

"A centralized data silo for an enterprise that contains merged, cleansed, and historical data"

-Microsoft

28

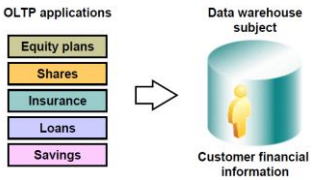
Data Warehouse Properties



29

Data Warehouse Properties Subject Oriented

- Data is categorized and stored by business subject rather than by application.



30

Data Warehouse Properties
Subject Oriented

Subject-oriented data is:

organized around major subject areas of an enterprise and is useful for an enterprise-wide understanding of those subjects. For example, a banking operational system keeps independent records of customer savings, loans, and other transactions. A warehouse pulls this independent data together to provide financial information.

You can access subject-oriented data related to any major subject area of an enterprise:

• Customer financial information

• Toll calls made in the telecommunications industry

• Airline passenger booking information

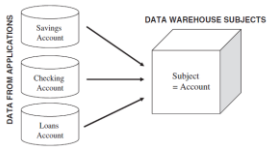
• Insurance claim data

The data is transformed so that it is consistent and meaningful for the warehouse.

31

Data Warehouse Properties
Integrated

- Data on a given subject is defined and stored once.
- Data inconsistencies are removed; data from diverse operational applications is integrated.



32

Data Warehouse Properties
Integrated

- **Integrated**

In many organizations, data resides in diverse independent systems, making it difficult to integrate into one set of meaningful information for analysis. A key characteristic of a warehouse is that data is completely integrated. Data is stored in a globally acceptable manner, even when the underlying source data is stored differently. The transformation and integration process can be time consuming and costly. It requires commitment from every part of the organization, particularly top-level managers who make the decisions and allocate resources and funds.

- **Data Consistency**

You must deal with data inconsistencies and anomalies before the data is loaded into the warehouse.

Consistency is applied to **naming conventions**, **measurements**, **encoding structures**, **and physical attributes of the data**.

- **Data Redundancy**

Data redundancy at the detail level in the warehouse environment is eliminated; the warehouse contains only that data, which is physically selected and moved into it; however, selective and deliberate redundancy in the form of aggregates (sums or averages) and summaries is required in the warehouse to improve the performance of queries, especially drill-down analysis.

33

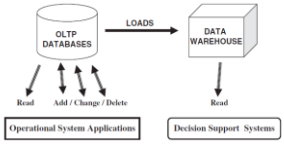
Data Warehouse Properties
Time Variant

- Data is stored as a series of snapshots, each representing a period of time.
- The time-variant nature of the data in a data warehouse
  - Allows for analysis of the past
  - Relates information to the present
  - Enables forecasts for the future

34

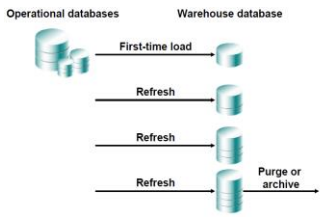
Data Warehouse Properties
Nonvolatile

- Typically, data in the data warehouse is not updated or deleted.



35

Changing Warehouse Data



36

Data Warehouse Versus OLTP

Property	OLTP	Data Warehouse
Response time	Subseconds to seconds	Seconds to hours
Operations	DML	Primarily read-only
Nature of data	30–60 days	Snapshots over time
Data organization	Application	Subject, time
Size	Small to large	Large to very large
Data sources	Operational, internal	Operational, internal, external
Activities	Processes	Analysis

37

Enterprise-Wide Data Warehouse

- Supports large-scale implementation
- Scopes the entire business
- Contains data from all subject areas
- Is developed incrementally
- Is a single source of enterprise-wide data
- Contains synchronized enterprise-wide data
- Is the single distribution point to dependent data marts.

38

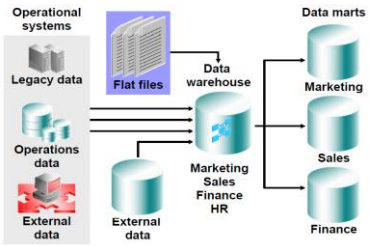
Data Warehouses Versus Data Marts

Property	Data Warehouse	Data Mart
Scope	Enterprise	Department
Subjects	Multiple	Single-subject, LOB
Data source	Many	Few
Implementation time	Months to years	Months

Note : Data mart is
a subset of data warehouse fact and summary data that provides users with information specific to their departmental requirements.

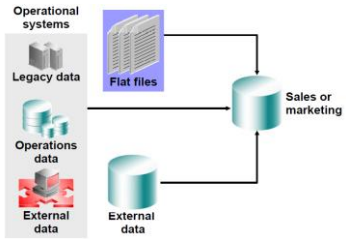
39

Dependent Data Mart



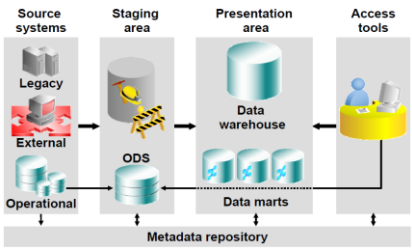
40

Independent Data Mart



41

Typical Data Warehouse Components



42

Warehouse Development Approaches

Dr.Waleed M.Ead, Ph.D

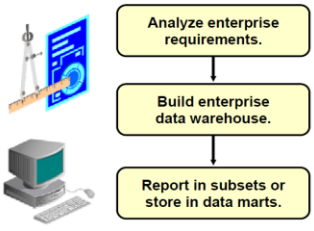
43

Warehouse Development Approaches

- "Big bang" approach
- Incremental approach:
  - – Top-down incremental approach
  - – Bottom-up incremental approach

44

Warehouse Development Approaches "Big Bang" Approach



45

Warehouse Development Approaches "Big Bang" Approach

Advantages

There are no real advantages in this approach over other approaches, and it should be avoided in most cases.

• The only real advantage is where the warehouse is being built as part of another major project or program such as reengineering and they are dependent on each other.

• You have a "big picture" of the data warehouse before starting the data warehousing project.

46

Warehouse Development Approaches "Big Bang" Approach

Disadvantages

The following are the disadvantages to this approach:

• Involves a high risk

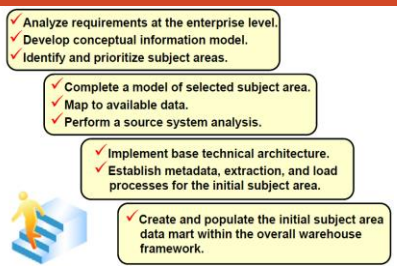
• Takes a longer time to deliver any perceived business benefit

• Runs the risk of needing to change requirements, which will change during analysis

Note: Because of dynamic evolving of business needs, requirements set at the onset of a project would no longer be viable.

47

Warehouse Development Approaches Top-Down Approach



48

Warehouse Development Approaches
Top-Down Approach

Advantages

This approach has the following advantages:

• Provides a relatively quick implementation and payback. Typically, the scoping, definition study, and initial implementation are scaled down so that **they can be completed in six to seven months**.

• Offers significantly lower risk because it avoids being as analysis heavy as the "big bang" approach

• Emphasizes high-level business needs

• Achieves synergy التعاضد among subject areas. Maximum information leverage is achieved as cross-functional reporting and a single version of the truth are made possible.

49

Warehouse Development Approaches
Top-Down Approach

Disadvantages

This approach has the following disadvantages:

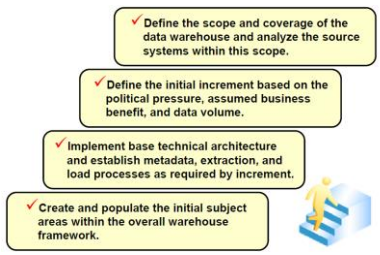
• Requires an increase in up-front costs before the business sees any return on their investment

• Is difficult to define the boundaries of the scoping exercise if the business is global

• May not be suitable unless the client needs cross-functional reporting

50

Warehouse Development Approaches
Bottom-Up Approach



51

Warehouse Development Approaches
Bottom-Up Approach

Advantages

This approach has the following advantages:

• This is a " proof of concept" type of approach; therefore, it is often appealing جذاب to IT.

• It is easier to get IT to choose this approach because it is focused on IT.

52

Warehouse Development Approaches
Bottom-Up Approach

Disadvantages

This approach has the following disadvantages:

• Because the solution model is typically developed from source systems and these source systems will have encapsulated within them the current business processes, the overall extensibility of the model will be compromised.

• IT staff is often the last to know about business changes—IT could be designing something that will be out-of-date before they complete its delivery.

53

Warehouse Development Approaches
Bottom-Up Approach

• As the framework of definition in this approach tends to be much narrower, often a significant amount of reengineering work is required for each increment.

• Because data definitions are rarely agreed upon by various lines of business for the first increment, the solution may be rejected by the next line of business to be involved.

• IT staff are used to data and not information. It is unusual for them to consider the temporal aspects of the data, thus minimizing the overall benefit to the business.

54

Warehouse Development Approaches Incremental Approach

- Multiple iterations
- Shorter implementations
- Validation of each phase

Iterative

55

Warehouse Development Approaches Incremental Approach

Benefits

- Delivers a strategic data warehouse solution through incremental development efforts
- Provides extensible, scalable architecture
- Supports the information needs of the enterprise organization
- Quickly provides business benefit and ensures a much earlier return of investment
- Allows a data warehouse to be built based on a subject or application area at a time
- Allows the construction of an integrated data mart environment

56

Data Warehousing Process Components

57

Data Warehousing Process Components

- **Methodology**
- **Architecture**
- **Extraction, transformation, and loading (ETL)**
- **Implementation**
- **Operation and support**

58

Data Warehousing Process Components Methodology

- A methodology is a set of detailed steps or procedures to accomplish a defined goal.
- To avoid failure of the warehouse implementation, you must employ a methodology and keep to it.
- Failure is generally caused in two ways.
  - The first cause of failure is that the warehouse is not delivered on time, and
  - the second is that the warehouse fails to deliver what the business users need.

A good method helps to manage expectations by identifying clear deliverables.

59

Data Warehousing Process Components Architecture

- "Provides the planning, structure, and standardization needed to ensure integration of multiple components, projects, and processes across time"
- "Establishes the framework, standards, and procedures for the data warehouse at an enterprise level" — The Data Warehousing Institute

60

Data Warehousing Process Components Architecture

- From a business and technology point of view,
  - An architecture defines a collection of components and specifies their relationships.
- The goal of the architecture activities is a single, integrated data warehouse meeting business information needs.
- Some of the components of a data warehousing architecture are:
  - Data sources -Data acquisition
  - Data management -Data distribution
  - Information directory -Data access tools

61

Data Warehousing Process Components Extraction, Transformation, and Loading (ETL)

"Effective data extract, transform, and load (ETL) processes represent the number one success factor for your data warehouse project and can absorb up to 70 percent of the time spent on a typical data warehousing project."

—DM Review

62

Data Warehousing Process Components Extraction, Transformation, and Loading (ETL)

• **Extraction:** The process of selecting specific operational attributes from various operational systems

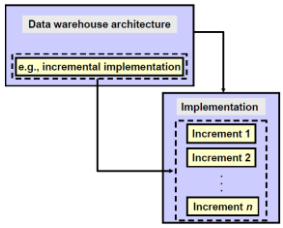
• **Transformation:** The process of integrating, verifying, validating, cleaning, and time stamping

the selected data into a consistent and uniform format for the target databases. Rejected data is returned to the data owner for correction and reprocessing.

• **Loading:** The process of moving data from an intermediate storage area into the target warehouse database

63

Data Warehousing Process Components Implementation



64

Data Warehousing Process Components Implementation

Implementation deliverables:

- **Analysis**
  - Confirm and refine requirements.
- **Design**
  - Gather specifications and prepare the blueprint for the data warehouse or data mart.
- **Construction**
  - Put in place and test the data warehouse or data mart and all required support tools.
- **Deployment**
  - Data warehouse or data mart is accepted for use in the business.

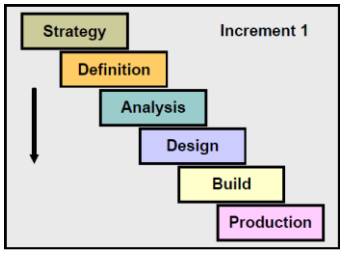
65

Data Warehousing Process Components Operation and Support

- Data access and reporting
- Refreshing warehouse data
- Monitoring
- Responding to change

66

Phases of the Incremental Approach



67

Lec 3

DATA WAREHOUSE DATABASE DESIGN PHASES

68

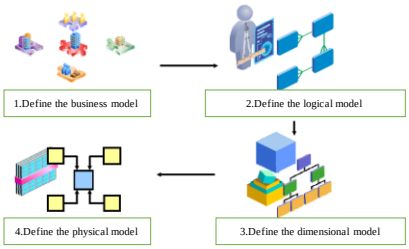
Data Warehouse Modeling Issues

Among the main issues that data warehouse data modelers face are:

- Different data types
- Many ways to use warehouse data
- Many ways to structure the data
- Multiple modeling techniques
- Planned replication
- Large volumes of data

69

Data Warehouse: Design Phases



70

71

Phase 1: Defining the Business Model

- Performing strategic analysis
- Creating the business model
- Documenting metadata

72

Defining the Business Model: Performing Strategic Analysis

- **Identify crucial business processes.**

ex. orders, invoices, shipments, inventory, sales, account administration, and the general ledger

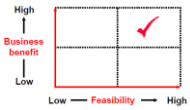
- **Understand business processes.**

-by drilling down on the dimensions that characterize each business process.

- **Prioritize and select the business processes to implement.**

-based on which one will provide the quickest and

largest return on investment (ROI).



73

Defining the Business Model: Creating the Business Model

- Defining business requirements
- Determining granularity
- Documenting metadata

74

Business Requirements Drive the Design Process

- Primary input

Interviews to collect business requirements

- Secondary input

Existing metadata

Production ERD model

Research

75

Using a Business Process Matrix

Business Dimensions	Business Processes		
	Sales	Returns	Inventory
Customers	✓	✓	
Times (Date)	✓	✓	✓
Products	✓	✓	✓
Channels	✓		
Promotions	✓	✓	

Sample of business process matrix

76

Identifying Business Measures and Dimensions

Measures

The attribute varies continuously:
- Sales
- Quantity sold
- Units sold
- Cost

Dimensions

The attribute is perceived as constant or discrete:
- Products
- Promotions
- Customers
- Countries
- Channels
- Times

77

Distinguishing Between Measures and Dimensions

During the warehouse design, you must decide whether a piece of data is a measure or a dimension.

You can use the following as a guide:

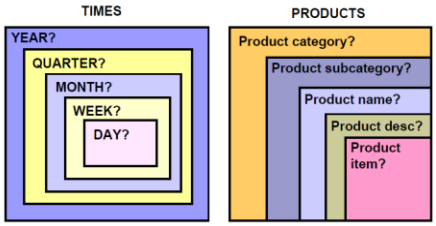
• If the data regularly changes value and is numeric, it is a measure—for example, units sold or account balances. A need or capability to summarize often identifies a measure.

• If the data is constant or it takes only a discrete number of values, it is a dimension. Typically, dimensions have descriptive, textual values—for example, the color of a product, the address of a customer, and so on.

These rules are not definitive but act as a guide where there is indecision.

78

Determining Granularity



79

Identifying Business Definitions and Rules: Example

Customer	
Credit Rating	Meaning
A+	0 bad checks or bank credit failures
A	1 bad check or bank credit failures
B	2 bad checks or bank credit failures
C	3 or more bad checks or bank credit failures

Order	
Rule 1	A customer with a credit rating of A or above will receive a 10% discount on any order totaling $500 (U.S.) or more.
Rule 2	A customer with a credit rating of A or above will receive a 5% discount on any order totaling $250 (U.S.) but less than $500.
…	…
Rule 5	A customer with a credit rating of C will not receive any discounts on purchases.

80

Documenting Metadata

Documenting metadata should include:

- Documenting the design process
- Documenting the development process
- Providing a record of changes
- Recording enhancements over time

81

Business Metadata Elements

- Name of the measure
- Business dimensions
  - Dimension attributes
- Sample data
- Business definition and rules

82

Metadata Documentation Approaches

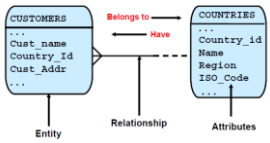
- **Automated**
  - Data-modeling tools
  - ETL tools
  - End-user tools
- **Manual**

83

Phase 2: Designing the Logical Model

The entity relationship model (ERM) uses the entity relationship diagram (ERD):

- Each CUSTOMER belongs to one COUNTRY.
- Each COUNTRY can have many CUSTOMERS.



84

?

Business dimensions:

1. Are the analytic parameters that categorize business processes for analysis purposes

2. Provide the metadata definitions for the data warehouse

3. Typically contain numeric values

4. Enable you to answer business questions

5. Are the success metrics of a business process

85

Phase 3: Defining the Dimensional Model

- **Identify fact tables:**
  - Translate business measures into fact tables.
  - Analyze source system information for additional measures.
- **Identify dimension tables.**
- **Link fact tables to the dimension tables.**
- **Model the time dimension.**

86

lec5

Data Warehouse Schemas

87

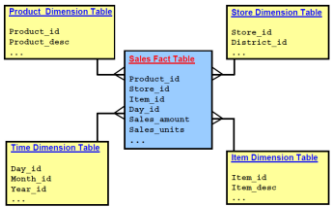
Data Warehouse Schemas

The data modeling structures that are commonly found in a data warehouse environment are:

- **Star schema**
- **Snowflake schema**
- **Third Normal Form (3NF)**

88

Star Schema Model



89

Star Schema Model

- Is easy to understand because the structure is simple and straightforward
- May provide fast response to queries with optimization and reductions in the physical number of joins required between fact and dimension tables
- Contains simple metadata
- Is supported by many front-end tools
- Is slow to build because of the level of denormalization

The star schema provides better query performance at the cost of more complex loading and transformation.

90

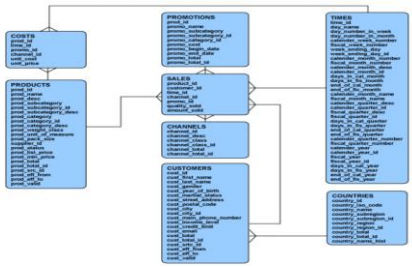
Star Schema Model

- Is easy to understand because the structure is simple and straightforward
- May provide fast response to queries with optimization and reductions in the physical number of joins required between fact and dimension tables
- Contains simple metadata
- Is supported by many front-end tools
- Is slow to build because of the level of denormalization

The star schema provides better query performance at the cost of more complex loading and transformation.

91

Star Dimensional Modeling



92

Star Dimensional Modeling

- A fact table has a multipart primary key composed of two or more foreign keys and expresses a many-to-many relationship.
- Each dimension table has a single-part primary key that corresponds exactly to one of the components of the multipart key in the fact table.

93

Advantages of Using a Star Dimensional Model

• Provides rapid analysis across different dimensions for drilling down, rotation, and analytical calculations for the multidimensional cube

• Creates a database design that improves performance

• Enables database optimizers to work with a more simple database design to yield better execution plans

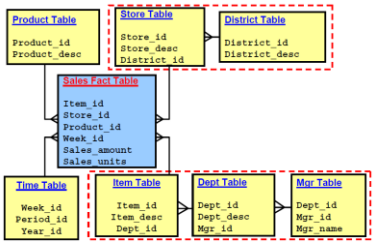
• Parallels how end users usually think of and use the data

• Provides an extensible design which supports changing business requirements

• Broadens the choices for data-access tools because some products require a star schema design

94

Snowflake Schema Model



95

Snowflake Schema Model

According to Ralph Kimball, "A dimension is said to be snowflaked when the low cardinality fields in the dimension have been moved to separate tables and linked back into the original table with artificial keys."

A snowflake model is closer to an ERD than the classic star model because the dimension data is more normalized. Developing a snowflake model means building class hierarchies out of each dimension (normalizing the data).

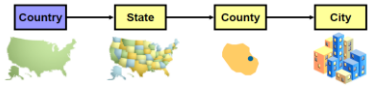
One of the major reasons why the star schema model has become more predominant than the snowflake model is its query performance advantage.

In a warehouse environment, the snowflake's quicker load performance is much less important than its slower query performance.

96

Snowflake Schema Model

- Can be used directly by some tools
- Is more flexible to change
- Provides for speedier data loading
- Can become large and unmanageable
- Degrades query performance
- Has more complex metadata



97

Third Normal Form (3NF)

- Minimizes data redundancy through normalization
- Typically has a large number of tables due to normalization and several fact tables
- Preserves a detailed record of each transaction without any data redundancy
- Allows for rich encoding of attributes and all relationships between data elements
- Requires that users typically have a solid understanding of the data in order to navigate

98

Third Normal Form (3NF)

- When compared to a star schema, a 3NF schema typically has larger number of tables (with table joins) due to this normalization process.
- 3NF schemas are typically chosen for better load performance.
- Some data warehouses use the 3NF schema design.
- Like the other schema designs, their data can also be directly accessed by using SQL code.
- They may have more efficient data storage at the price of slower query performance due to extensive table joins. Some large companies build a 3NF central data warehouse feeding-dependent star data marts for specific lines of business.

99

Third Normal Form (3NF)

When a data warehouse has larger tables or fact tables, they should be partitioned, using composite partitioning—for example, range-hash:

- A range to facilitate the data load and data elimination
- A hash on the join column to facilitate partitionwise joins
- A number of hash partitions that must be a power of 2 (#CPU X 2)

100

Fact Table: Characteristics

- Facts are the numerical measures of the business.
- The fact table is the largest table in the star schema and is composed of large volumes of data, usually making up 90% or more of the total database size.
- It can be viewed in two parts:
  - Multipart primary key
  - Business metrics
    - Numeric
    - Additive (usually)

Sales (Fact Table)
PROD_ID
CUST_ID
TIME_ID
CHANNEL_ID
PROMO_ID
QUANTITY_SOLD
AMOUNT_SOLD
...

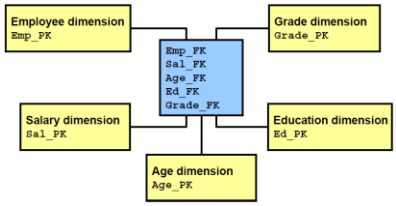
101

What Are Factless Fact Tables?

- There are two types of factless fact tables: event-tracking and coverage.
- Event-tracking tables record and track events that have occurred, such as college students' class attendance,
- coverage factless tables support the dimensional model when the primary fact table is sparse متناثر (for example, a sales promotion factless table). In the latter case, the events did not occur.
- The factless fact table represents the many-to-many relationships between the dimensions so that the characteristics of the event can be analyzed.

102

Examples



103

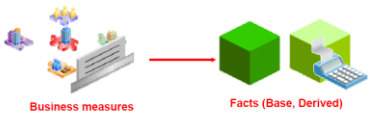
- **Human resources**: Studies of the labor force composition are often conducted for reporting and planning purposes. Analysis of employees with different characteristics can be conducted using the illustrated star. Most of the resulting information from this kind of table is a series of counts. In the example illustrated, selecting COUNT(EMP_FK) gives the number of employees, whereas selecting COUNT(SAL_FK) gives the number of employees on a specified salary grade.

- **Retail store**: Promotions are typical within the retail environment. An upscale retail chain wants to compare its customers who do not respond to direct mail promotion to those who make a purchase. A factless fact table supports the relationship between the customer, product, promotion, and time dimensions.

- **Student attendance**: Factless fact tables can be used to record student class attendance in a college or school system. There is no fact associated with this; it is a matter of whether the students attended.

104

Identifying Base and Derived Measures



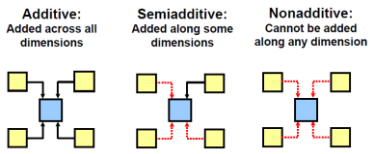
Business Measures
Quantity sold
Amount sold
Profit

Sales Fact Table	Measure Type
Quantity sold	Base
Amount sold	Base
Profit	Derived

105

Fact Table Measures

Fact table measures can be:



106

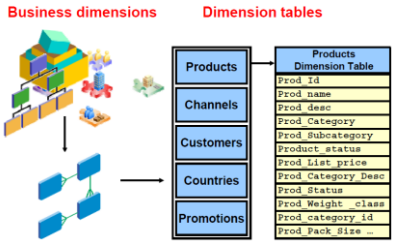
Dimension Table: Characteristics

Dimension tables:

- Contain textual information that represents the attributes of the business
- Contain relatively static data
- Are joined to a fact table through a foreign key reference

107

Translating Business Dimensions into Dimension Tables



108

Phase 4: Defining the Physical Model

Tasks

- Translate the dimensional design to a physical model for implementation.
- Update the metadata document with the physical model information.
- Determine the hardware architecture.
- Define the storage strategy for tables and indexes.
- Perform database sizing.
- Define the partitioning strategy.
- Define the initial indexing strategy.
- Define the security strategy.

109

Lec 7

Database Sizing, Storage, Performance, and Security Considerations

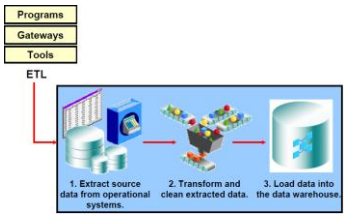
110

Lec9

The ETL Process:
Extracting Data

111

Extraction, Transformation, and Loading (ETL) Process



112

ETL: Tasks, Importance, and Cost

ETL Tasks: ETL involves a series of tasks that:

- Extract data from source systems

- Transform and clean up the data

- Index the data

- Summarize the data

- Load data into the warehouse

- Track the changes made to the source data required for the warehouse

- Restructure keys

- Maintain the metadata

- Refresh the warehouse with updated data

113

ETL Importance:

- Relevant and useful to the business users

- Of high quality

- Accurate

- Easy to access so that the warehouse is used efficiently and effectively by the business users

114

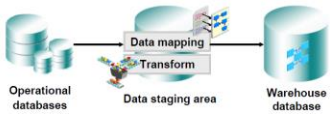
ETL Cost: Building the ETL process is potentially one of the biggest tasks of building a warehouse; it is complex and time consuming.

In some implementations, it can take more than half of the total warehouse implementation effort.

115

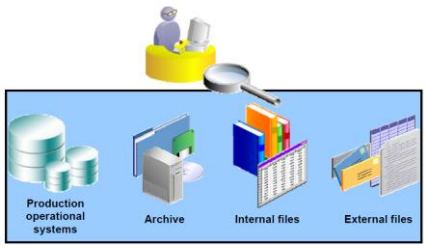
Extracting Data

- Source systems:
  - Data from various data sources in various formats
- Extraction routines:
  - Are developed to select data fields from sources
  - Consist of business rules, audit trails, and error correction facilities



116

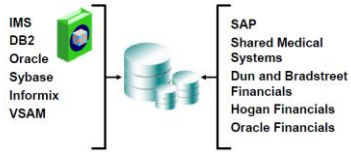
Examining Data Sources



117

Production Data

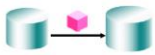
- Operating system platforms
- File systems
- Database systems and vertical applications



118

Extraction Methods

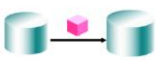
- Logical extraction methods:
  - Full extraction
  - Incremental extraction
- Physical extraction methods:
  - Online extraction
  - Offline extraction
- Your logical choice influences the way the data is physically extracted.



119

Extraction Techniques

- Programs: C, C++, PL/SQL, or Java
- Gateways: Transparent database access
- ETL tools: Oracle Warehouse Builder
- Data Pump import and export
- External tables



120

Designing Extraction Processes

- Analysis:
  - Sources, technologies
  - Data types, quality, owners
- Design options:
  - Manual, custom, gateway, tools
  - Replication, full, or delta refresh
- Design issues:
  - Volume and consistency of data
  - Automation, skills needed, resources

121

Maintaining Extraction Metadata

- Source location, type, structure
- Access method
- Privilege information
- Temporary storage
- Failure procedures
- Validity checks
- Handlers for missing data

122

Possible ETL Failures

- A missing source file
- A system failure
- Inadequate metadata
- Poor mapping information
- Inadequate storage planning
- A source structural change
- No contingency plan
- Inadequate data validation

123

Maintaining ETL Quality

- ETL must be:
  - Tested
  - Documented
  - Monitored and reviewed
- Disparate metadata must be coordinated.

124

Quiz

The data source systems may comprise data existing in:

1. Production operational systems
2. Archives
3. Internal files
4. External data from sources outside the company

125

The ETL Process: Transforming Data

126

Transformation

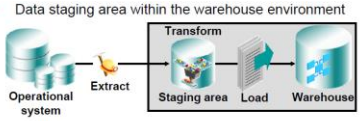
Transformation eliminates anomalies from operational data:

- Cleans and standardizes data
- Presents subject-oriented data

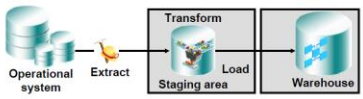
127

Remote Staging Model

- Data staging area within the warehouse environment



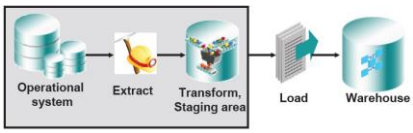
- Data staging area in its own environment



128

On-Site Staging Model

Data staging area within the operational environment, possibly affecting the operational system



129

Data Anomalies

- No unique key
- Data naming and coding anomalies
- Data meaning anomalies between groups
- Spelling and text inconsistencies

CUSNUM	NAME	ADDRESS
90233479	Oracle Limited	100 N.E. 1st St.
90233489	Oracle Computing	15 Main Road, Ft. Lauderdale
90234889	Oracle Corp. UK	15 Main Road, Ft. Lauderdale, FLA
90345672	Oracle Corp UK Ltd	181 North Street, Key West, FLA

130

Transformation Routines

- Cleaning data
- Eliminating inconsistencies
- Adding elements
- Merging data
- Integrating data
- Transforming data before load

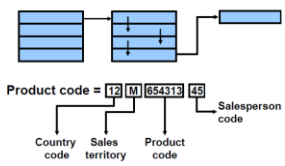
131

Transforming Data:
Problems and Solutions

132

Multipart Keys Problem

Multipart keys



133

Solution

The program or tools you use must be capable of identifying on a character-by-character (or position-by-position) basis the individual values, length of value, and the meaning of the resulting information. In the example cited, it is important that the code can extract the M and know that this is a territory code that identifies "Midwest," "Manchester," or "Moscow."

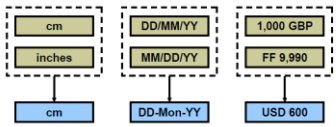
You may need to build a series of transformations to evaluate the results fully. For example, these steps may be appropriate:

1. Extract the third character position.
2. Evaluate the character against a master lookup table.
3. Evaluate the meaning of M.
4. Store the meaning (Moscow) in a field for insertion into the data warehouse.

134

Multiple Local Standards Problem

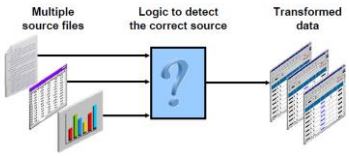
- Problem: There are multiple local standards.
- Solution: Use the tools or filters to preprocess the data.



135

Multiple Files Problem

- Problem: There is an added complexity of multiple source files.
- Solution: Start simple as shown in the following diagram.

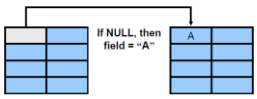


136

Missing Values Problem

Solution:

- Ignore
- Wait
- Mark rows
- Extract when time-stamped



137

Duplicate Values Problem

Solution:

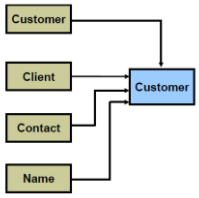
- SQL self-join techniques
- RDBMS constraints

138

Element Names Problem

Solution:

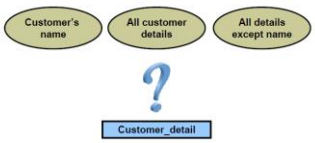
Common naming conventions



139

Element Meanings Problem

- Problem: This is a complex solution.
- Solution:
  - Avoid misinterpretation.
  - Document the meaning in metadata.



140

Referential Integrity Constraints Problem

Solution:

- SQL antijoin
- Server constraints
- Dedicated tools

Department
10
20
30
40

Emp	Name	Department
1099	Smith	10
1289	Jones	20
1234	Doe	50
6786	Harris	60

141

Name and Address Problem

- Single-field format

Mr. J. Smith,100 Main St., Bigtown, County Luth, 23565

- Multiple-field format

Name	Mr. J. Smith
Street	100 Main St.
Town	Bigtown
Country	County Luth
Code	23565

Database 1	
NAME	LOCATION
DIANNE ZIEFELD	N100
HARRY H. ENFIELD	M300

Database 2	
NAME	LOCATION
ZIEFELD, DIANNE	100
ENFIELD, HARRY H	300

142

The ETL Process: Loading Data

143