Colorado Technical University

CS683 – Data Warehouse

Colorado Technical University
Instructor: Dr. John Conklin
Unit 2 – Design Requirements

2 A Little Data Humor

3

Agenda

- Data Warehouse System Lifecycle
  - Risk Factors
  - Top-Down vs. Bottom-Up
    - Business Dimensional Lifecycle
    - Rapid Warehousing Methodology
  - Data Mart Design Phase
    - Analysis & Recommendation of Data Systems
    - Requirements Analysis
    - Conceptual Design
    - Workload Refinement
    - Logical Design
    - Physical Design
    - Data Staging Design
  - Methodological Framework
    - Data Driven Approach
    - Requirement Driven Approach
    - Mixed Approach
  - Testing the Data Mart

4

Agenda (cont.)

- Analysis & Reconciliation of Data Sources
  - Inspecting & Normalization Schemata
  - The Integration Problem
    - Different Perspectives
    - Common Concepts
  - Integration Phase
    - Preintegration
    - Schema Comparison
    - System Alignment
    - Merging & Restructuring Schemata
  - Define Mapping
- Data Warehouse and Business Intelligence Strategy
  - BI Strategy
    - Purpose
    - Usage
    - Architecture Overview

5

Agenda (cont.)

- DW Strategy
  - Usage
  - DW Architecture
    - Data Flow
    - Technical Architecture
  - Focus and Success
  - Where to Start?
    - For BI
    - For DW
  - How to Start
    - For BI
    - For DW
  - Project Phasing
- Project Resources
  - Stakeholders
  - Team Structure
  - Periodic Reviews
  - Center of Competence (CoC)

6

Agenda (cont.)

- Design Models
  - Conceptual Model
  - Logical Model
  - Physical Model

7 Data Warehouse System

Traditional Enterprise Data Warehouse Architecture

Traditionally data warehouses were built using a three-tier architecture:

- **Bottom tier**—database server used to extract data from multiple sources
- **Middle tier**—OLAP server, which transforms data to enable analysis and complex queries
- **Top tier**—tools used for high-level data analysis, querying, reporting, and data mining

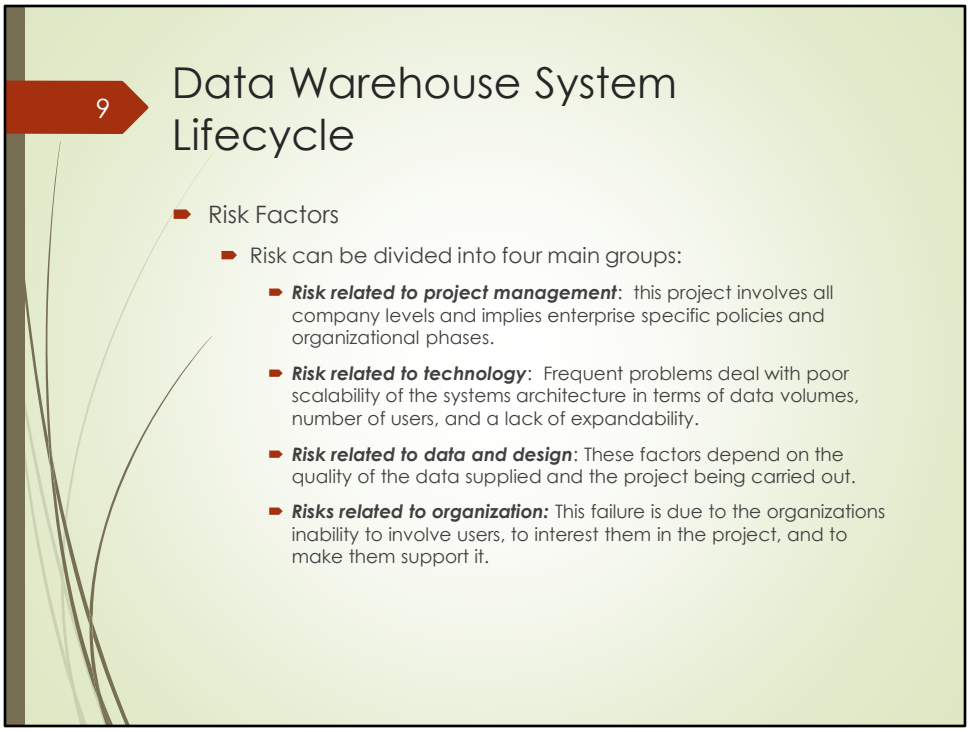
Data warehouses were structured using one of the following models:

- **Virtual data warehouse**—a set of separate databases, which can be queried together, forming one virtual data warehouse.
- **Data mart**—small data warehouses set up for business-line specific reporting and analysis. An organization's data marts together comprise the organization's data warehouse.
- **Enterprise data warehouse (EDW)**—a large data warehouse holding aggregated data that spans the entire organization.

(https://panoply.io/data-warehouse-guide/#components-of-a-data-warehouse)

Data Warehouse System Lifecycle

8



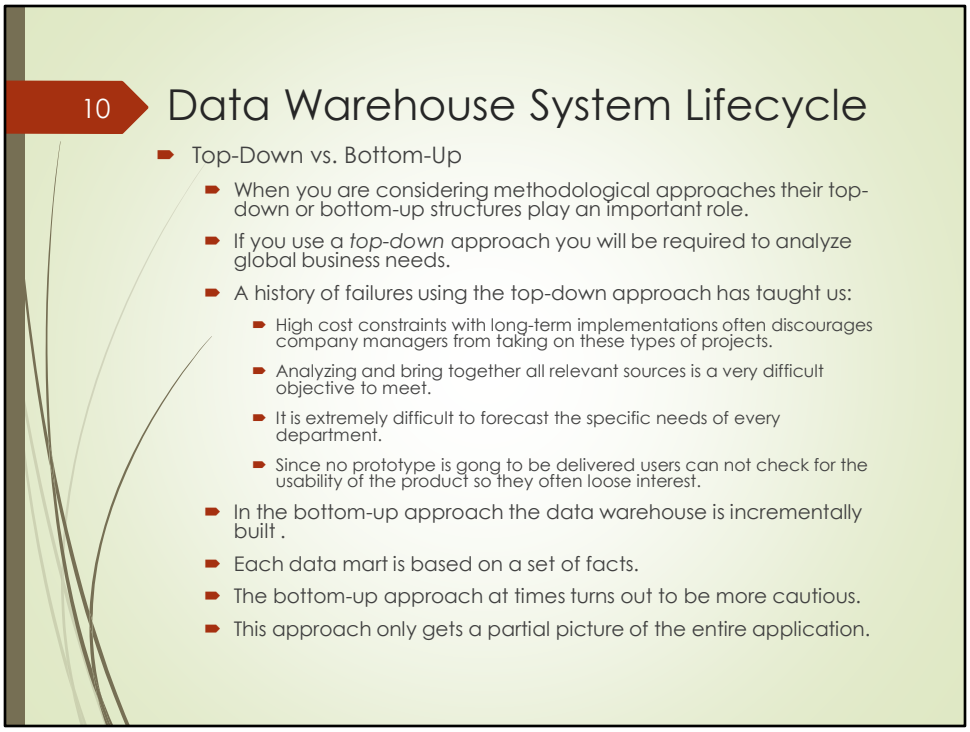
Risk can be divided into four main groups:

- ***Risk related to project management***: this project involves all company levels and implies enterprise specific policies and organizational phases. Frequent causes of failure are mainly related to scopes, sponsoring and goals of the data warehouse system.

- ***Risk related to technology***: Frequent problems deal with poor scalability of the systems architecture in terms of data volumes, number of users, lack of expandability to implement new technological solutions, insufficient technical skills, and insufficient management of meta-data exchange between components.

- ***Risk related to data and design***: These factors depend on the quality of the data supplied and the project being carried out. Another frequent cause for failure is the inability to provide users with high value added when delivering the prototypes.

- ***Risks related to organization:*** This failure is due to the organizations inability to involve users, to interest them in the project, and to make them support it. Other factors include difficulties found in radically changing business culture and the users inability to take advantage of results achieved.



Top-Down vs. Bottom-Up

When you are considering methodological approaches their top-down or bottom-up structures play an important role in their development.

If you use a top-down approach you will be required to analyze global business needs, plan how to develop a data warehouse, design and implement it as a whole.

A history of failures using the top-down approach has taught us:

- High cost constraints with long-term implementations often discourages company managers from taking on these types of projects.
- Analyzing and bring together all relevant sources is a very difficult objective to meet.
- It is extremely difficult to forecast the specific needs of every department.
- Since no prototype is gong to be delivered users can not check for the usability of the product so they often loose interest.

In the bottom-up approach the data warehouse is incrementally built and several data marts are created in stages.

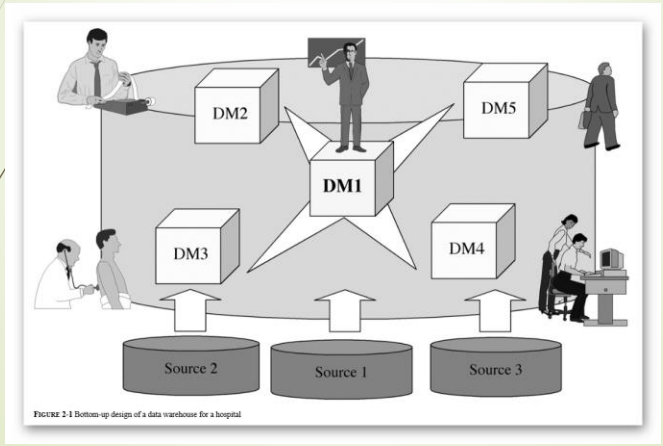
Each data mart is based on a set of facts that are linked to a specific company department.

The bottom-up approach at times turns out to be more cautious than the top-down approach and is almost always universally accepted.

This approach is not risk-free because it only gets a partial picture of the entire application.

11 Data Warehouse System Lifecycle

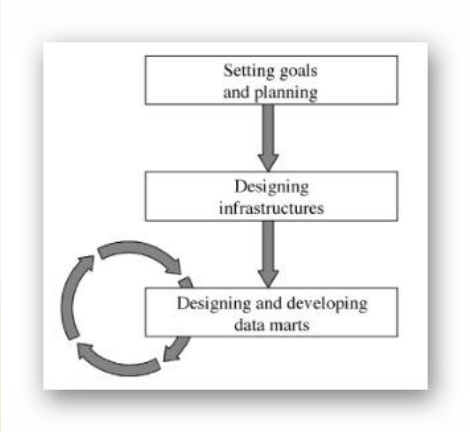
Figure 2-1 below is an example of a bottom-up approach from a hospital. DM1 stands for the first data mart implemented, then numerically in order the additional data marts are introduced to complete the entire system.



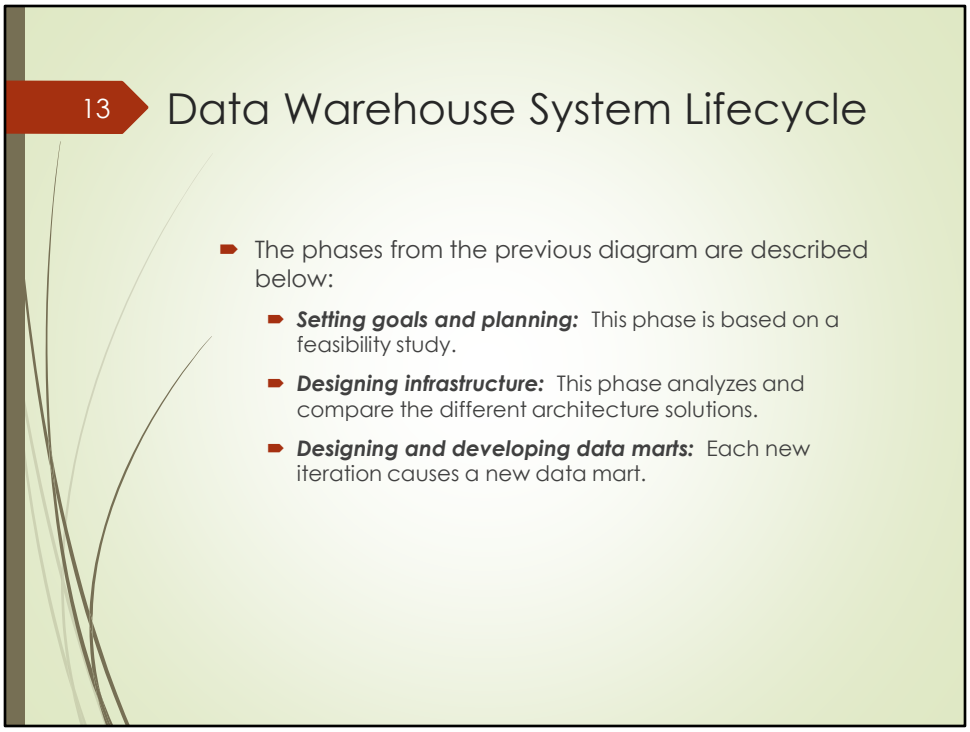
(Golfarelli & Rizzi, 2009)

12 Data Warehouse System Lifecycle

This diagram shows the basic phases of the lifecycle.



(Golfarelli & Rizzi, 2009)



The phases from the previous diagram are described below:

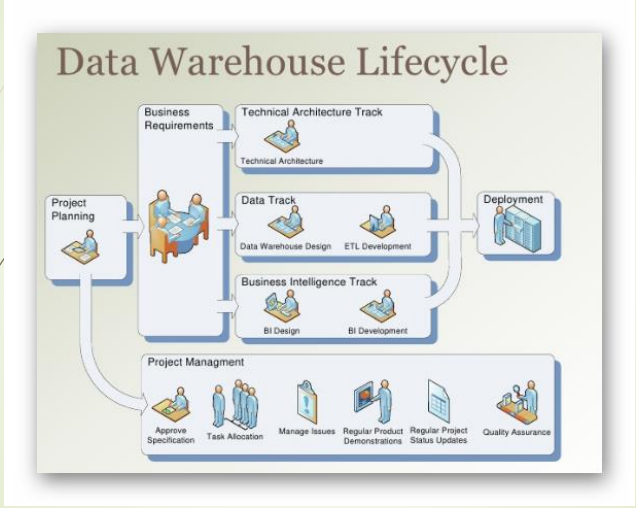
- ***Setting goals and planning:*** This phase is based on a feasibility study which is aimed at setting the system goals, properties and size estimates.

- ***Designing infrastructure:*** This phase analyzes and compare the different architecture solutions available and assess existing technology and tools to create the preliminary plan of the entire system.

- ***Designing and developing data marts:*** Each new iteration causes a new data mart and new applications to be created and added to the data warehouse system.

Data Warehouse System Lifecycle

14



(https://www.slideshare.net/bartlowe/the-data-warehouse-lifecycle)

15 Data Warehouse System Lifecycle

- Business Dimensional Lifecycle
  - This stands for the period of time needed for designing, developing, and implementing a data warehouse system.

(Golfarelli & Rizzi, 2009)

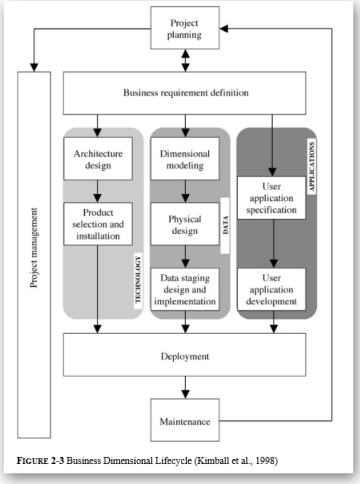
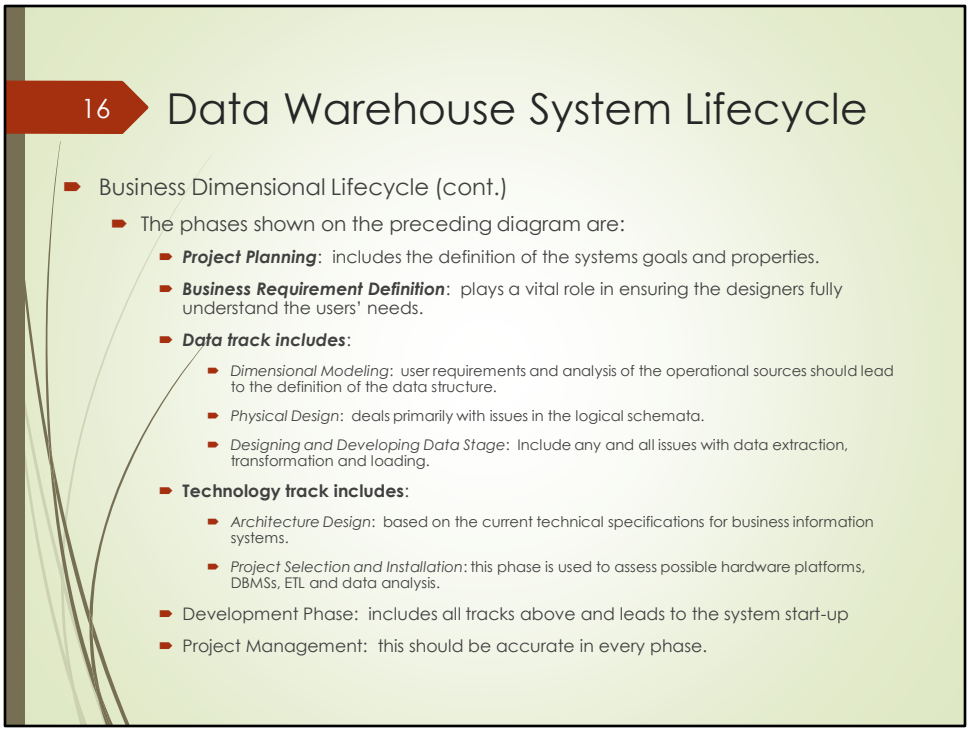


FIGURE 2-3 Business Dimensional Lifecycle (Kimball et al., 1998)



The phases shown on the preceding diagram are:

- ***Project Planning***: includes the definition of the systems goals and properties, as well as an assessment of the impact on the organizational practices and estimates of costs and benefits.
- ***Business Requirement Definition***: this phase plays a vital role in ensuring the designers fully understand the users' needs to help maximize benefits and the profitability of the system.
- ***Data track includes***:
    - *Dimensional Modeling*: user requirements and analysis of the operational sources should lead to the definition of the data structure. The final result in this phase is a set of logical schemata and a set of relationships with source schemata identified.
    - *Physical Design*: deals primarily with issues in the logical schemata to be optimized an implemented in the selected DBMS.
    - *Designing and Developing Data Stage*: Include any and all issues with data extraction, transformation and loading of data into the warehouse.
- **Technology track includes**:
    - *Architecture Design*: based on the current technical specifications for business information systems.
    - *Project Selection and Installation*: this phase is used to assess possible hardware platforms, DBMSs, ETL and data analysis.
- Development Phase: includes all tracks above and leads to the system start-up. This does not mean the system lifecycle process comes to an end, the system still need continuous maintenance.
- Project Management: this should be accurate in every phase. This allows you to keep the project on track.

Rapid Warehousing Methodology

This is an iterative approach to managing data warehouse projects. This approach was created by the SAS Institute. Each build takes advantage of the data warehouse environments previously built. It essentially adds to them to add new features and expand on the functionality of the entire system.

Includes 7 steps as outlined below:

- **Assessment**: mostly corresponds the Kimball's planning phase and it aims at ascertaining whatever an enterprise is ready to take on.
- **Requirements**: corresponds to Kimball's business requirement definition and end user application specification.
- **Design**: focus is on one project being built at a time.
- **Construction and Final Test**: one DW is implemented and populated with data from source, then the front-end applications are developed and tested.
- **Deployment**: system is delivered and start up after end user training.
- **Maintenance and Administration**: this phase is as long as the entire project.
- **Review**: each build include three review processes: implementation check, post-deployment check, and final check for costs and benefits to be assessed.

18

Data Warehouse System Lifecycle

Rapid Warehousing Methodology diagram created by the SAS Institute.

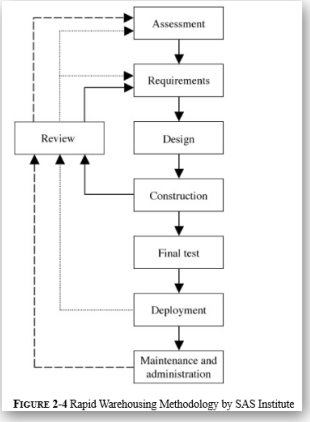
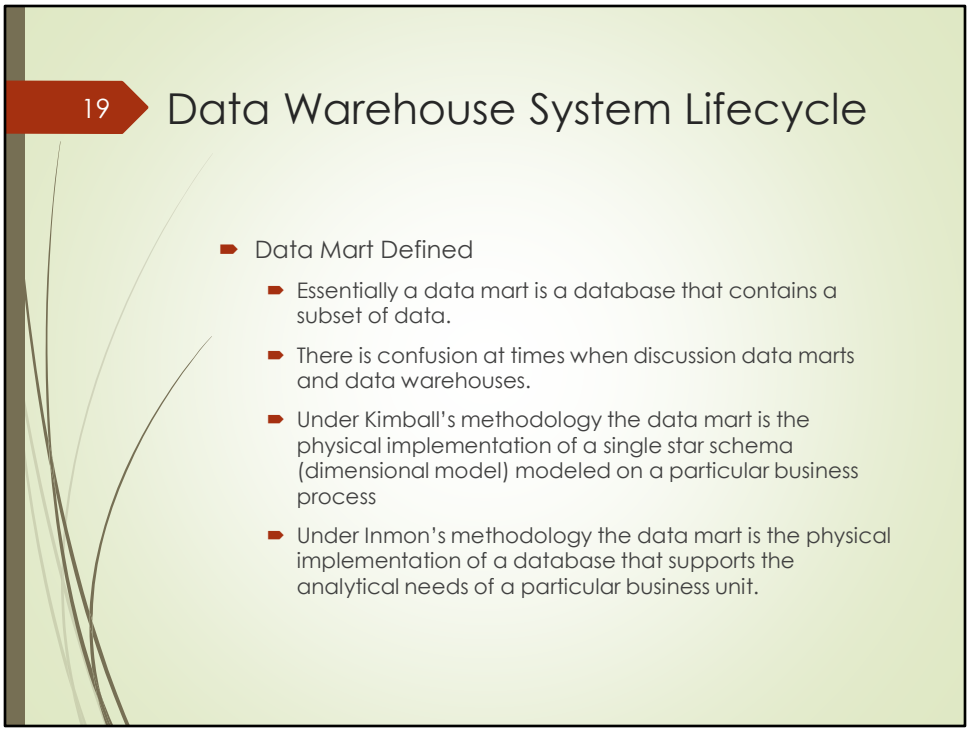


FIGURE 2-4 Rapid Warehousing Methodology by SAS Institute

(Golfarelli & Rizzi, 2009)



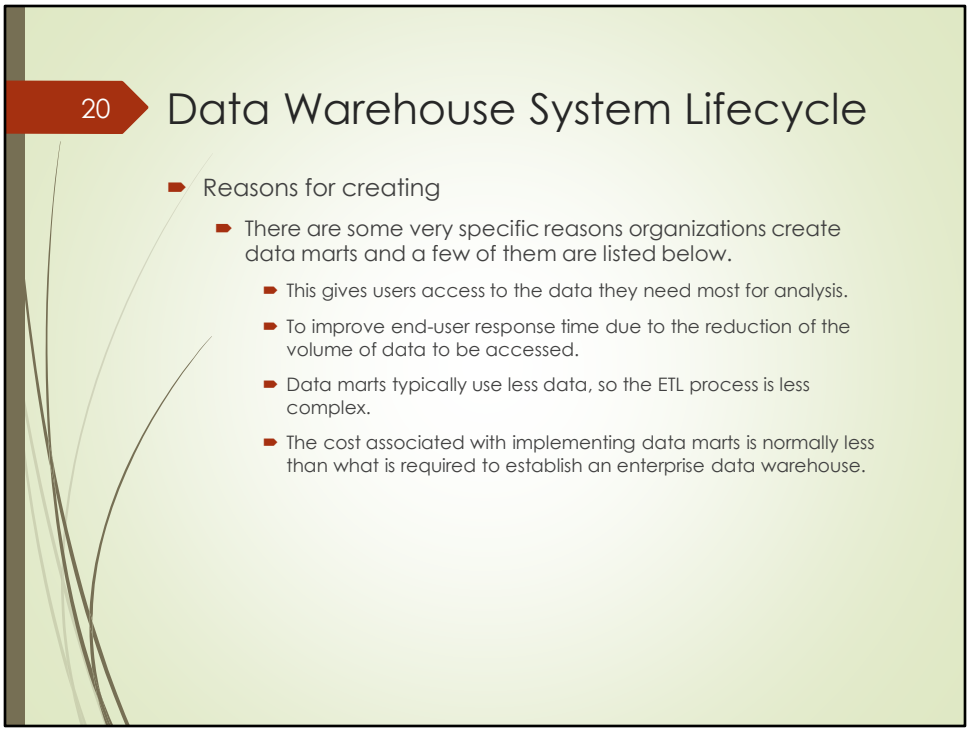
Data Mart Defined

- Essentially a data mart is a database that contains a subset of data to support the analytical needs of a specific business function.

- There is confusion at times when discussion data marts and data warehouses, particularly around what data is stored in the data mart.

- Under Kimball's methodology the data mart is the physical implementation of a single star schema (dimensional model) modeled on a particular business process. Here when all the organizations business process are represented by data marts the integration of these data marts form the enterprise data warehouse

- Under Inmon's methodology the data mart is the physical implementation of a database that supports the analytical needs of a particular business unit. Inmon's data mart receives its data from the enterprise data warehouse.



Reasons for creating

There are some very specific reasons organizations create data marts and a few of them are listed below.

- This gives users access to the data they need most for analysis.
- To improve end-user response time due to the reduction of the volume of data to be accessed.
- Data marts typically use less data, so the ETL process is less complex, and implementation and setup of the data mart is thereby less complicated and simpler then establishing an enterprise data warehouse.
- The cost associated with implementing data marts (in time, money, and resources) is normally less than what is required to establish an enterprise data warehouse.

21 Data Warehouse System Lifecycle

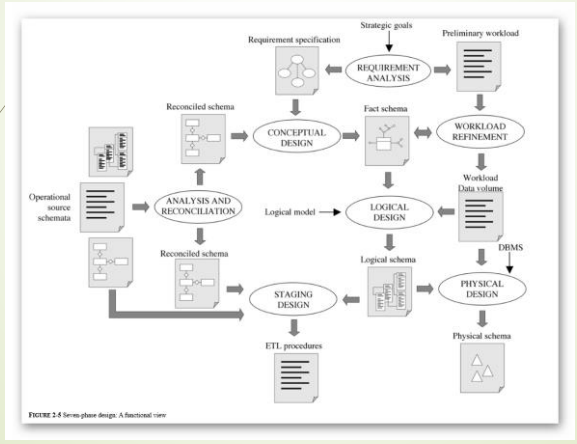
- Data Mart Design Phase
  - The chart below the seven phases of a methodological approach presented in our text.

Phase	Input	Output	People Involved
Analysis and reconciliation of data sources	Operational source schemata	Reconciled schema	Designer, data processing center staff
Requirement analysis	Strategic goals	Requirement specifications, preliminary workload	Designer, end users
Conceptual design	Reconciled schema, requirement specification	Fact schemata	Designer, end users
Workload refinement, validation of conceptual schemata	Fact schemata, preliminary workload	Workload, data volume, validated fact schemata	Designer, end users
Logical design	Fact schemata, target logical model, workload	Logical data mart schema	Designer
Data-staging design	Source schemata, reconciled schema, logical data mart schema	ETL procedures	Designer, database administrators
Physical design	Logical data mart schema, target DBMS, workload	Physical data mart schema	Designer

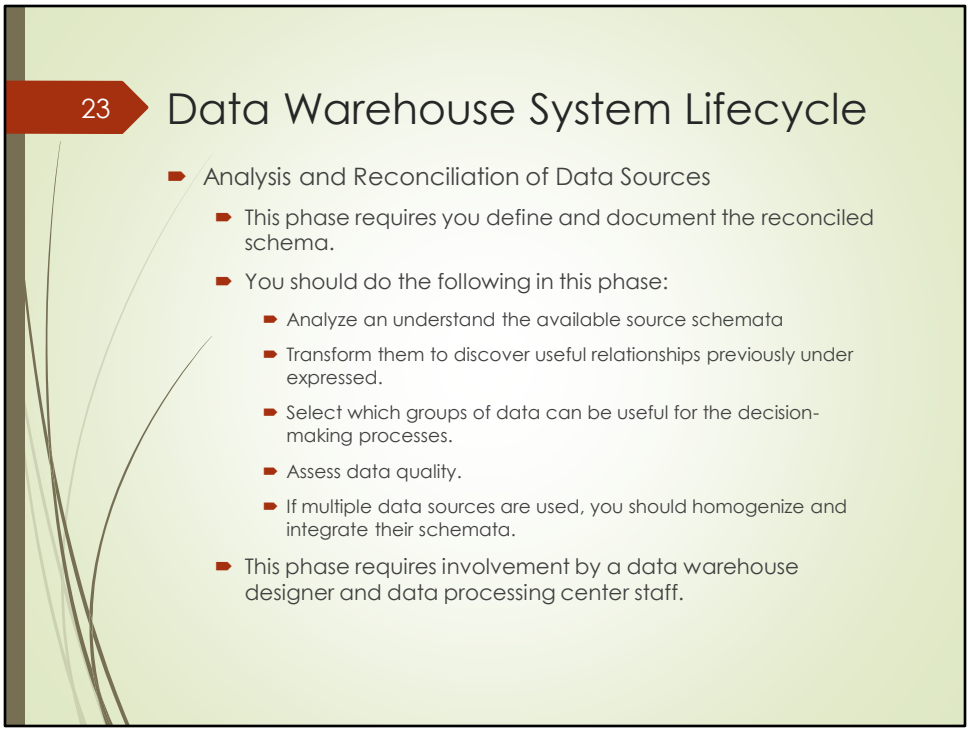
(Golfarelli & Rizzi, 2009)

22 Data Warehouse System Lifecycle

- Data Mart Design Phase
  - This diagram shows a functional view of the seven-phase design shown previously.



(Golfarelli & Rizzi, 2009)



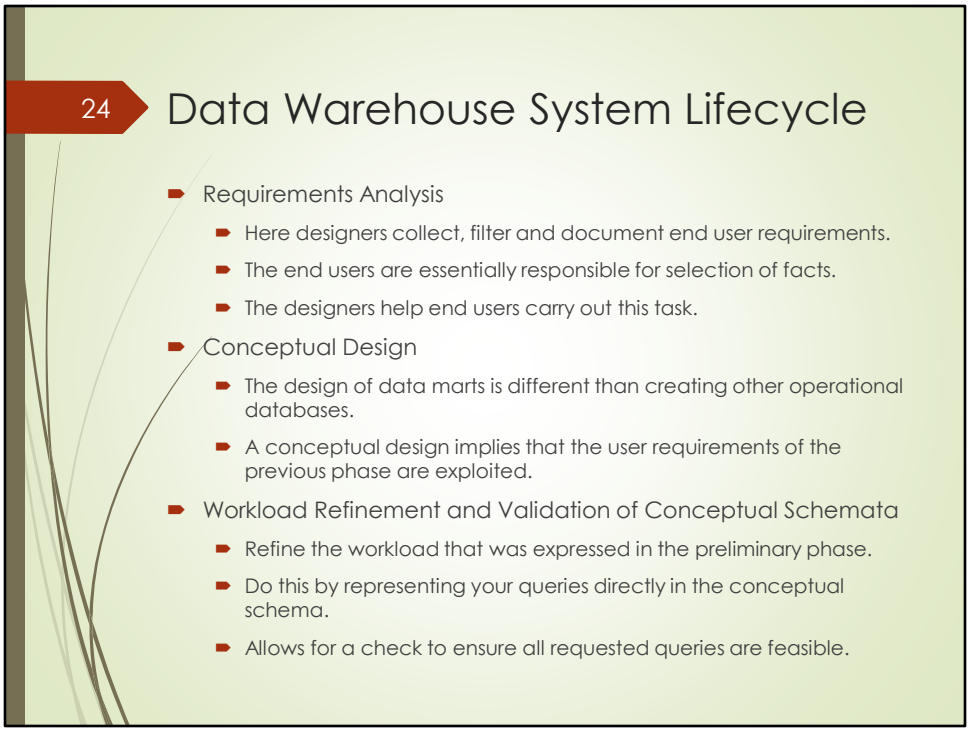
Analysis and Reconciliation of Data Sources

This phase requires you define and document the reconciled schema, the schema of the reconciled data layer used to feed your data mart.

You should do the following in this phase:

- Analyze an understand the available source schemata
- Transform them to discover useful relationships previously under expressed.
- Select which groups of data can be useful for the decision-making processes of your organization.
- Assess data quality.
- If multiple data sources are used, you should homogenize and integrate their schemata to determine common features and remove any inconsistency.

This phase requires involvement by a data warehouse designer and data processing center staff.



Requirements Analysis

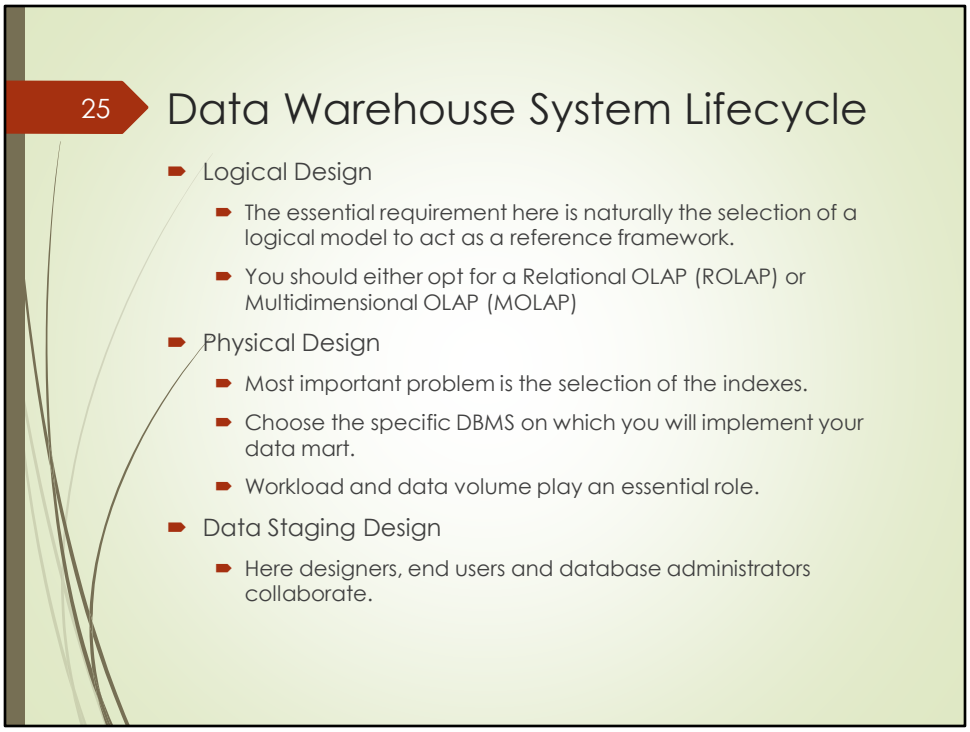
- Here designers collect, filter and document end user requirements in the requirements analysis phase to select only relevant information which will be represented in compliance with strategic goals.
- The end users are essentially responsible for selection of facts.
- The designers help end users carry out this task based on the documentation collected in previous phases of analysis and reconciliation.

Conceptual Design

- The design of data marts is different than creating other operational databases, simply because data mart workloads are different.
- A conceptual design implies that the user requirements of the previous phase are exploited to create the conceptual schema for the data mart.

Workload Refinement and Validation of Conceptual Schemata

- You should refine the workload that was expressed in the preliminary phase.
- You do this by representing your queries directly in the conceptual schema.
- Allows for a check to ensure all requested queries are feasible and leads to validation of the conceptual schema.



Logical Design

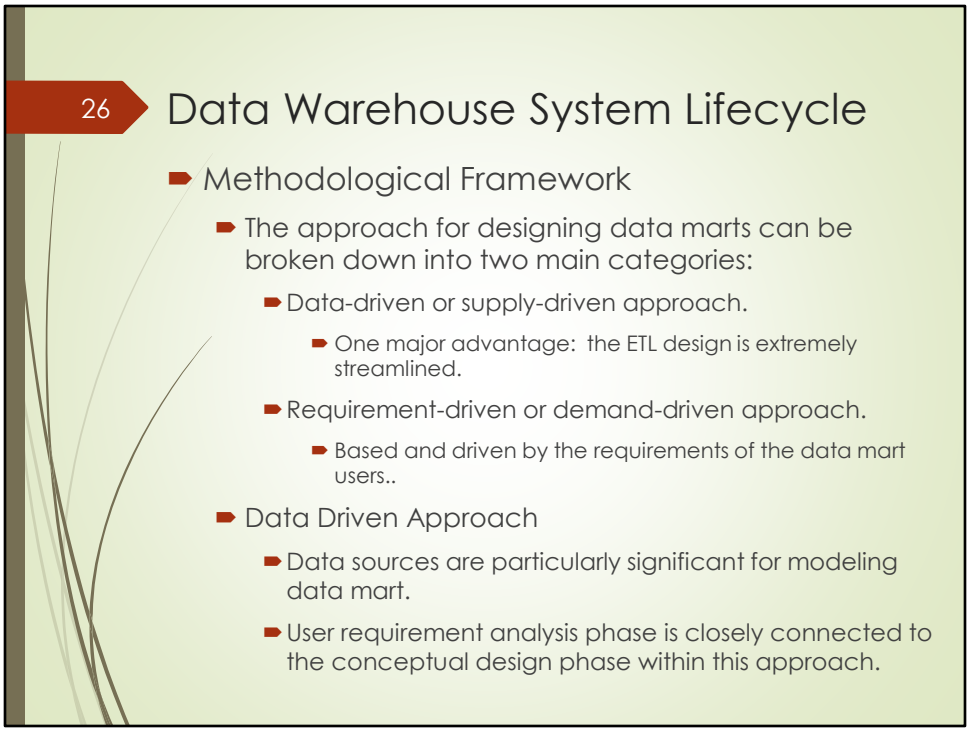
- The essential requirement here is naturally the selection of a logical model to act as a reference framework.
- You should either opt for a Relational OLAP (ROLAP) or Multidimensional OLAP (MOLAP)

Physical Design

- Most important problem with this phase is the selection of the indexes to optimize performance.
- You must also choose the specific DBMS on which you will implement your data mart.
- Workload and data volume play an essential role in the physical design of your warehouse.

Data Staging Design

- Here designers, end users and database administrators collaborate and make all the significant decisions on the population process for the reconciled layer.



Methodological Framework

The approach for designing data marts can be broken down into two main categories:

Data-driven or supply-driven approach.

- This approach has one major advantage: the ETL design is extremely streamlined because every piece stored within the data mart is directly associated with more or more attributes.

Requirement-driven or demand-driven approach.

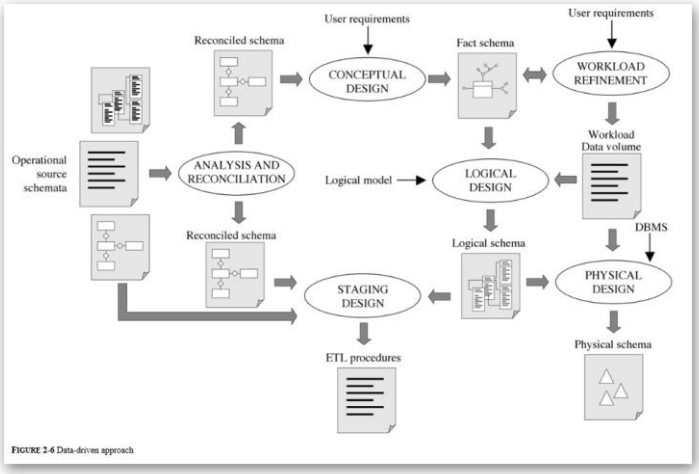
- This approach is based and driven by the requirements of the data mart users. This approach makes users' wishes play a major role in development. This approach is also typically more time expensive, simply because users don't have clear understanding of business goals and processes.

Data Driven Approach

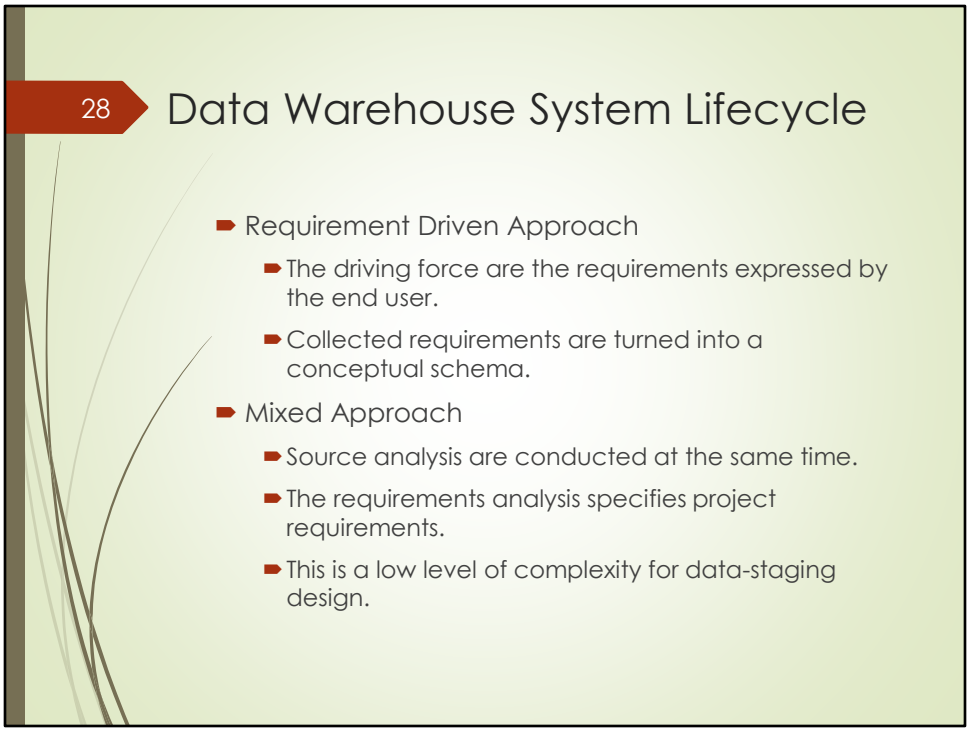
- Data sources are particularly significant for modeling data mart. For this reason data analysis and reconciliation are part of the first phase on this approach.

- User requirement analysis phase is closely connected to the conceptual design phase within this approach.

Data Warehouse System Lifecycle



(Golfarelli & Rizzi, 2009)



Requirement Driven Approach

- The driving force behind this approach are the requirements expressed by the end user.

- Collected requirements are turned into a conceptual schema, making the conceptual design phase critical to this approach.

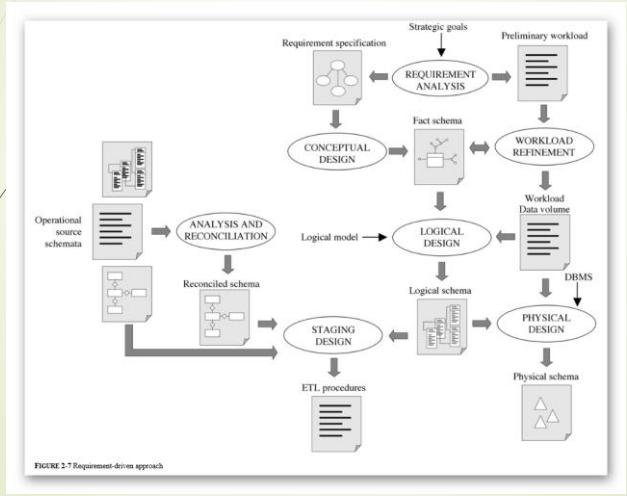
Mixed Approach

- Here requirement and source analysis are conducted at the same time.

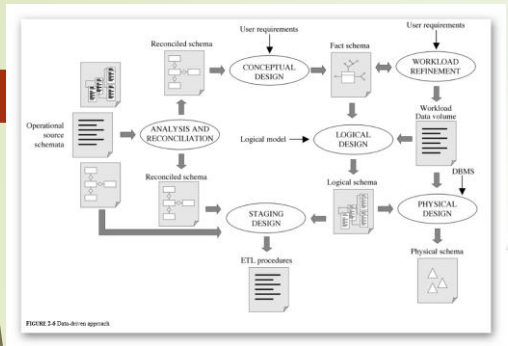
- The requirements analysis specifies project requirements.

- Here this is a low level of complexity for data-staging design described in the data driven approach.

29 Data Warehouse System Lifecycle



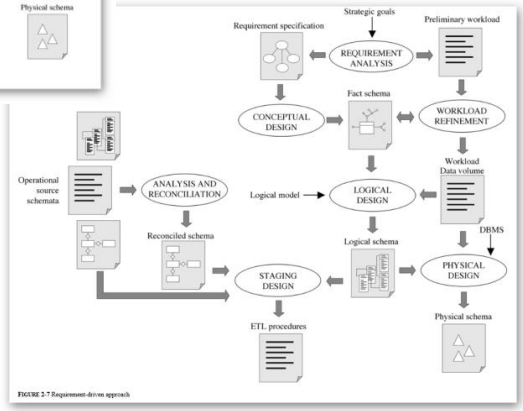
(Golfarelli & Rizzi, 2009)



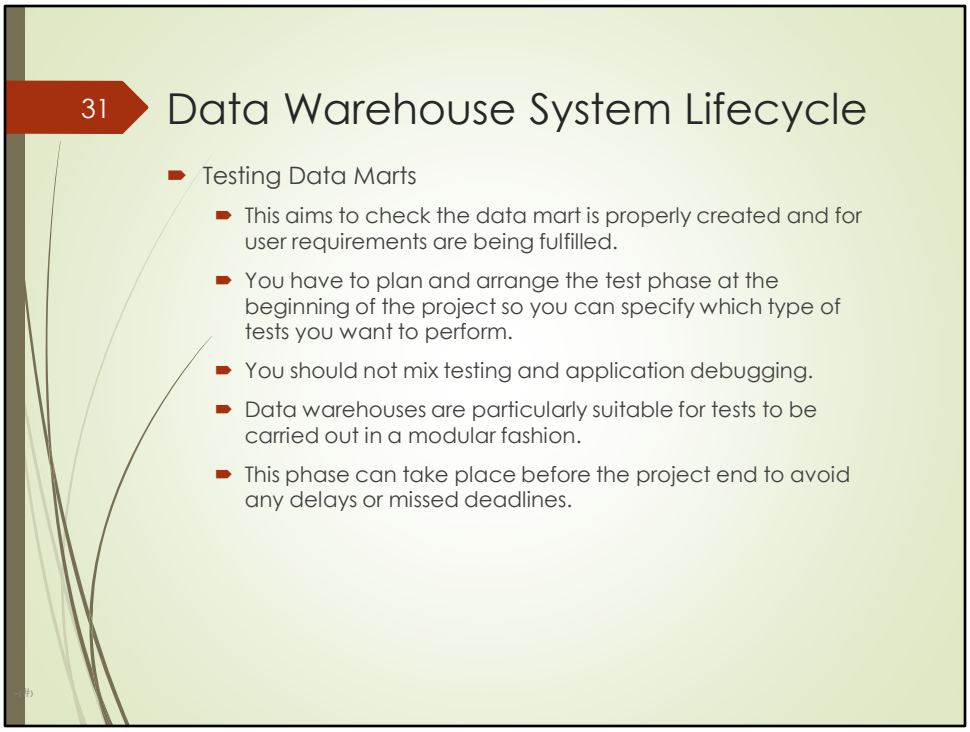
Data-driven approach

Comparison of data-driven and requirement-driven approaches

Requirement-driven approach

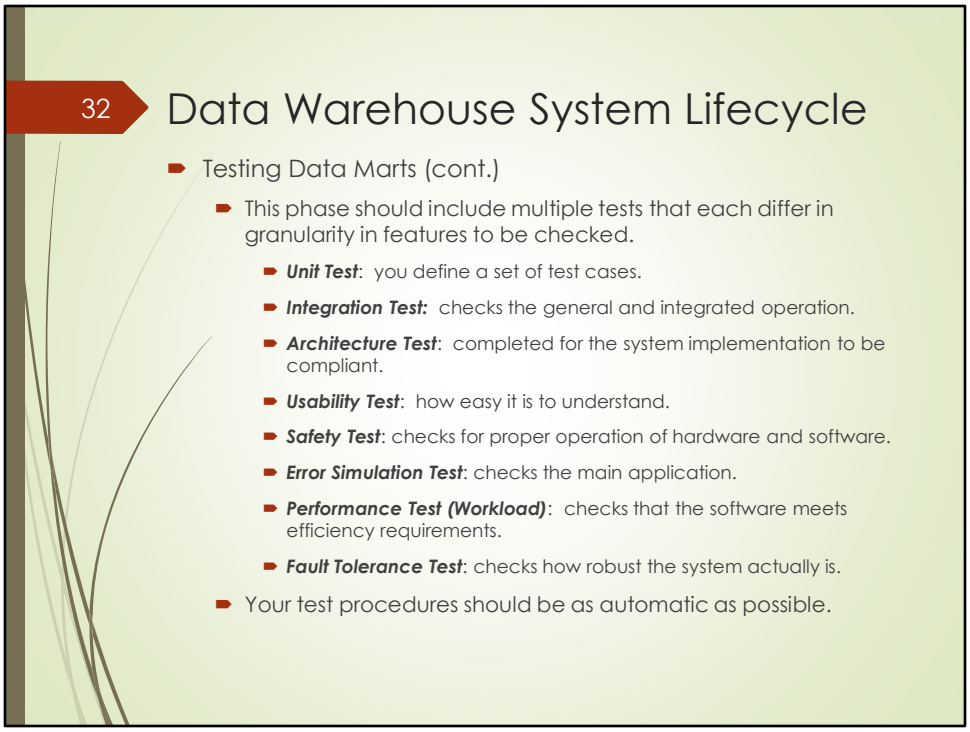


(Golfarelli & Rizzi, 2009)



Testing Data Marts

- This aims to check the data mart is properly created and for user requirements are being fulfilled.
- You have to plan and arrange the test phase at the beginning of the project so you can specify which type of tests you want to perform; which data sets need to be tested and what service level you expect.
- You should not mix testing and application debugging, which is a task usually performed by implementers who often tend to take some hypothesis for granted and do not eliminate them even when they should.
- Data warehouses are particularly suitable for tests to be carried out in a modular fashion.
- This phase can take place before the project end to avoid any delays or missed deadlines which at times forces testers to rush their tests and report successful testing for components that don't' even meet the minimum requirements.



Testing Data Marts (cont.)

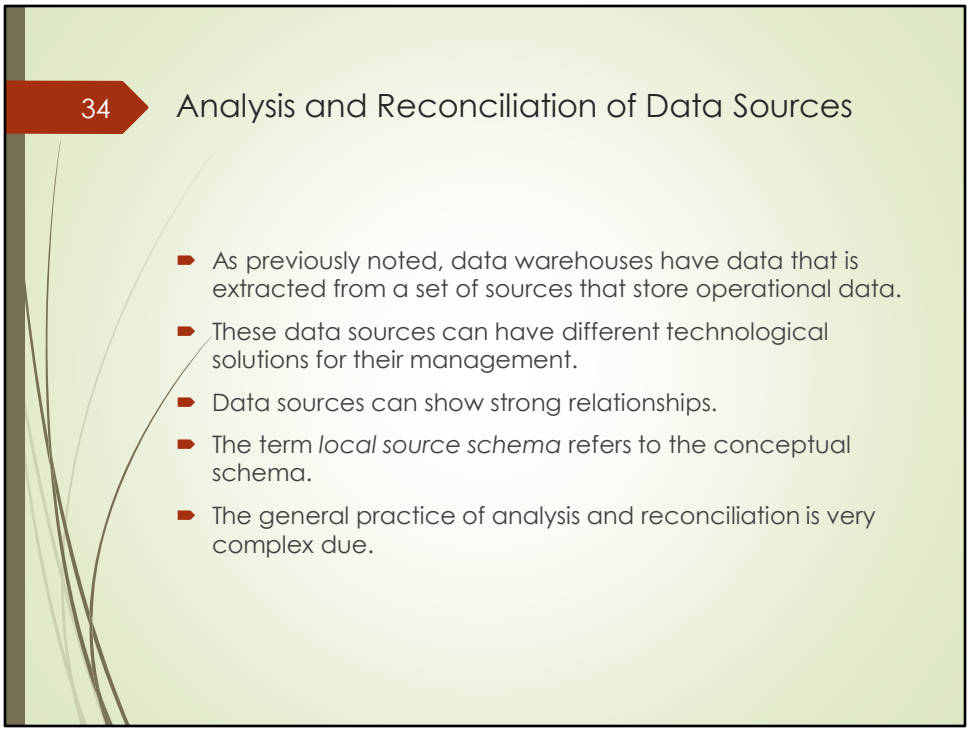
This phase should include multiple tests that each differ in granularity in features to be checked.

- ***Unit Test***: here you define a set of test cases and it includes data for every single application component.
- ***Integration Test:*** this checks the general and integrated operation of all system components.
- ***Architecture Test***: there checks are completed for the system implementation to be compliant with all specifications of the architectural schema.
- ***Usability Test***: checks how easy it is to understand, learn and use applications.
- ***Safety Test***: checks for proper operation of hardware and software mechanisms used to operate a safe system.
- ***Error Simulation Test***: checks the main application error conditions are handled properly.
- ***Performance Test (Workload)***: checks that the software meets efficiency requirements, this is typically a workload test for the ETL processes.
- ***Fault Tolerance Test***: checks how robust the system actually is. Simulates errors in one or more components and then evaluates the system response to errors.

Your test procedures should be as automatic as possible, this way the regression test just restarts all available testing procedures.

Analysis and Reconciliation of Data Sources

33



- As previously noted, data warehouses have data that is extracted from a set of sources that store operational data.

- These data sources can have different technological solutions for their management, such as database management systems (DBMSs), spreadsheets, file systems and so on.

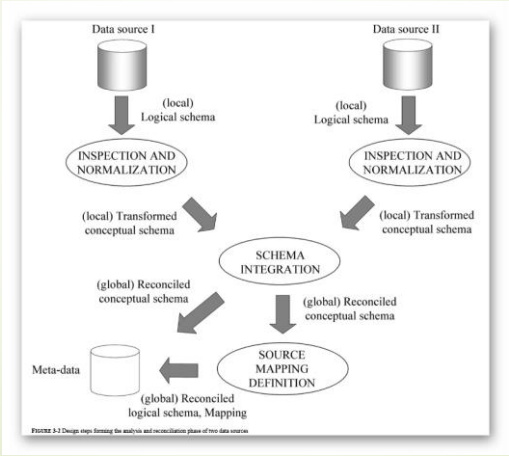
- Data sources can show strong relationships between them or be completely independent.

- The term *local source schema* refers to the conceptual schema that describes the data source without regard to the model and/or technological solutions implemented.

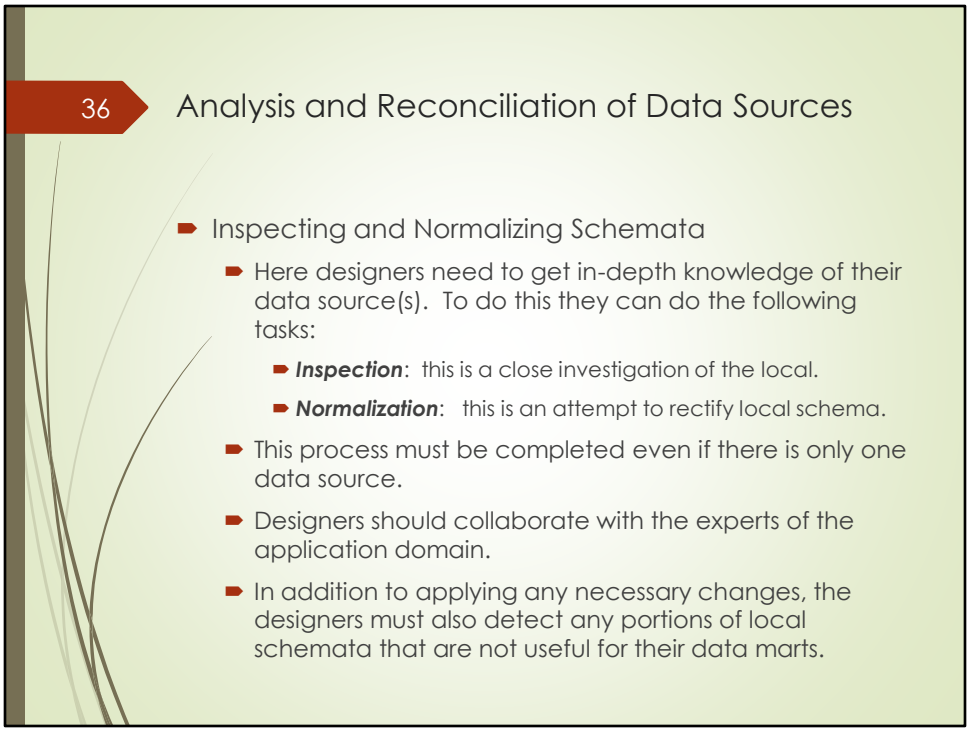
- The general practice of analysis and reconciliation is very complex due to the large number of possible combinations present in the warehouse. To help break free from this hard task you must carefully plan a list of actions to take.

Analysis and Reconciliation of Data Sources

FIGURE 3-2 Design steps forming the analysis and reconciliation phase of two data sources



(Golfarelli & Rizzi, 2009)



Inspecting and Normalizing Schemata

Here designers need to get in-depth knowledge of their data source(s). To do this they can do the following tasks:

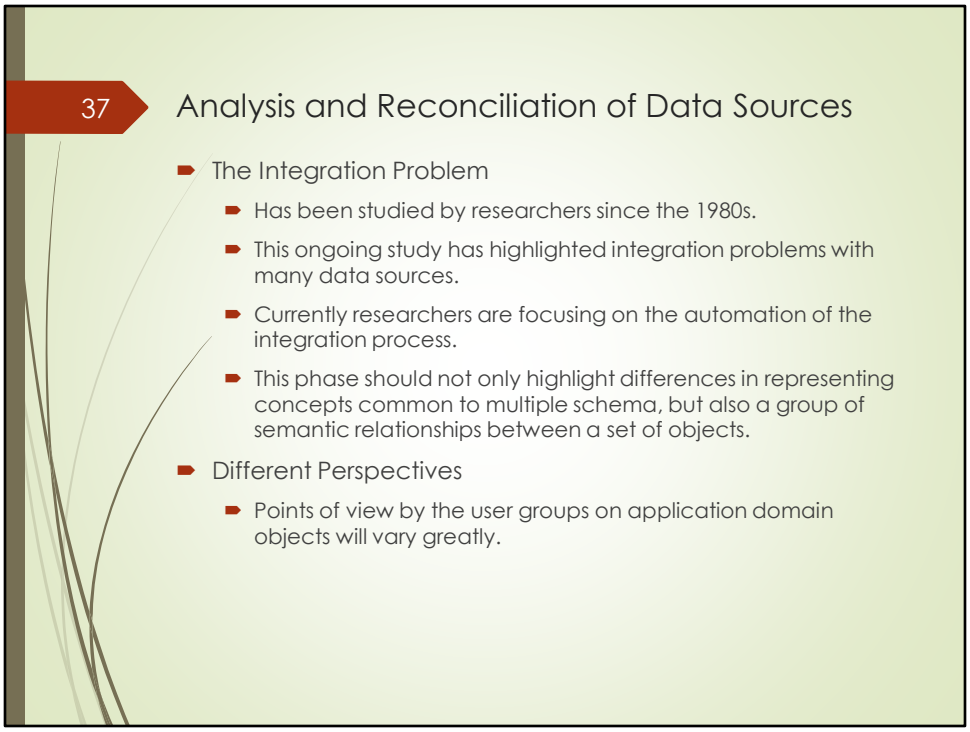
- ***Inspection***: this is a close investigation of the local schema to understand the application.

- ***Normalization***: this is an attempt to rectify local schema to model the application domain as accurately as possible.

- The inspection and normalization process must be completed even if there is only one data source.

- Designers should collaborate with the experts of the application domain, such as the data processing center staff, managers and so on in this analysis phase.

- In addition to applying any necessary changes, the designers must also detect any portions of local schemata that are not useful for their data marts.

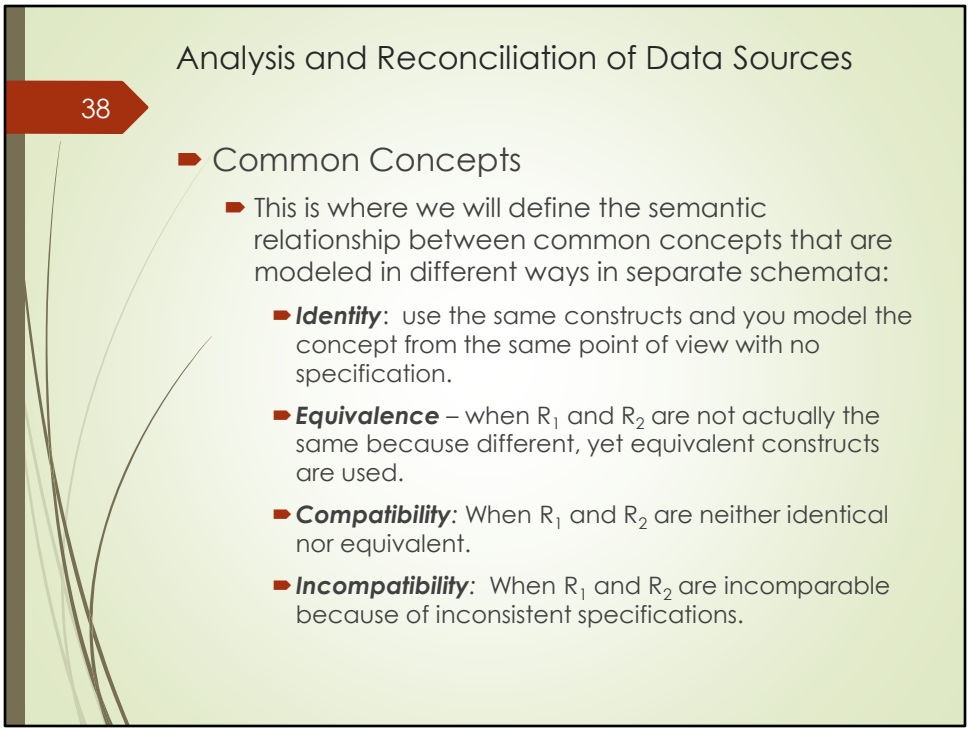


The Integration Problem

- This problem has been studied by researchers since the 1980s.
- This ongoing study has highlighted integration problems with many data sources, such as relational databases, object-oriented databases, and semi-structured sources.
- Currently researchers are focusing on the automation of the integration process.
- This phase should not only highlight differences in representing concepts common to multiple schema, but also a group of semantic relationships between a set of objects.

Different Perspectives

- Points of view by the user groups on application domain objects will vary greatly depending on the points relevant for the tasks the users have to perform.



Common Concepts

This is where we will define the semantic relationship between common concepts that are modeled in different ways in separate schemata. Four types of relationships exists between separate representations R_1 and R_2 for the same concept:

- ***Identity***: use the same constructs and you model the concept from the same point of view with no specification mistake – when R_1 and R_2 exactly coincide.

- ***Equivalence*** – when R_1 and R_2 are not actually the same because different, yet equivalent constructs are used and neither mistakes nor different perspectives exists.

- ***Compatibility:*** When R_1 and R_2 are neither identical nor equivalent but the constructs used and designers' view points are comparable.

- ***Incompatibility:*** When R_1 and R_2 are incomparable because of inconsistent specifications. This meant the set of objects in the real world modeled by R_1 denies the set of objects in the real world modeled by R_2.

Analysis and Reconciliation of Data Sources

For example, both modeling representations shown in Figure 3-6 could be the result of an error because it is very unlikely that a professor is not allowed to teach more than one course at the same time (on the left) or that he is required to teach more than one course (on the right). (*Golfarelli & Rizzi, 2009, p. 68)*

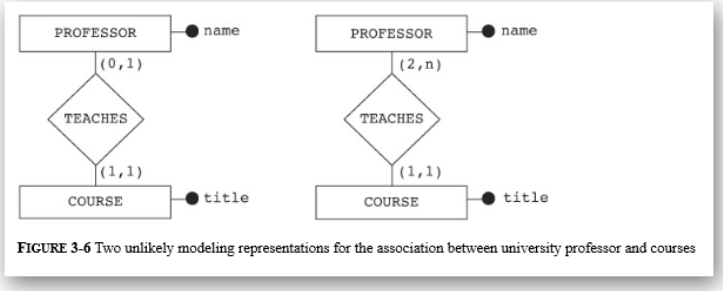


FIGURE 3-6 Two unlikely modeling representations for the association between university professor and courses

Analysis and Reconciliation of Data Sources

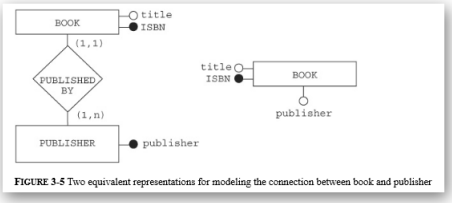


FIGURE 3-5 Two equivalent representations for modeling the connection between book and publisher

Figure 3-5 shows two schemata modeling the relationship between a book and its publishing house. The right-hand side schema uses just one attribute, but the left-hand side one aims at emphasizing the same concept creating the PUBLISHER entity. *(Golfarelli & Rizzi, 2009, p. 68)*

Figure 3-7 shows two possible instances of the schemata that are compliant with the previous definition.*(Golfarelli & Rizzi, 2009, p. 70)*

BOOK

ISBN	title	publisher
123445	DFM	McGraw-Hill
4354543	Apparently Logical	Read All
4566454	The Right Measure	Not Only Books
.....	.....	.....

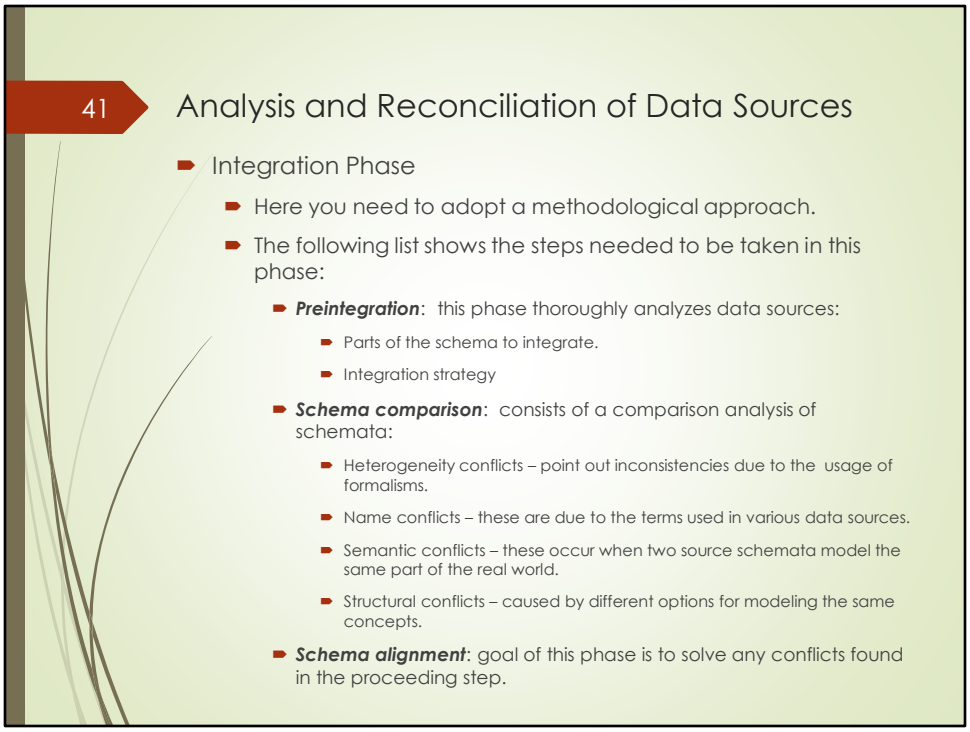
PUBLISHER

publisher
McGraw-Hill
Read All
Not Only Books
.....

BOOK

ISBN	title	publisher
123445	DFM	McGraw-Hill
4354543	Apparently Logical	Read All
4566454	The Right Measure	Not Only Books
.....	.....	.....

FIGURE 3-7 Two instances for the equivalent schemata shown in Figure 3-5

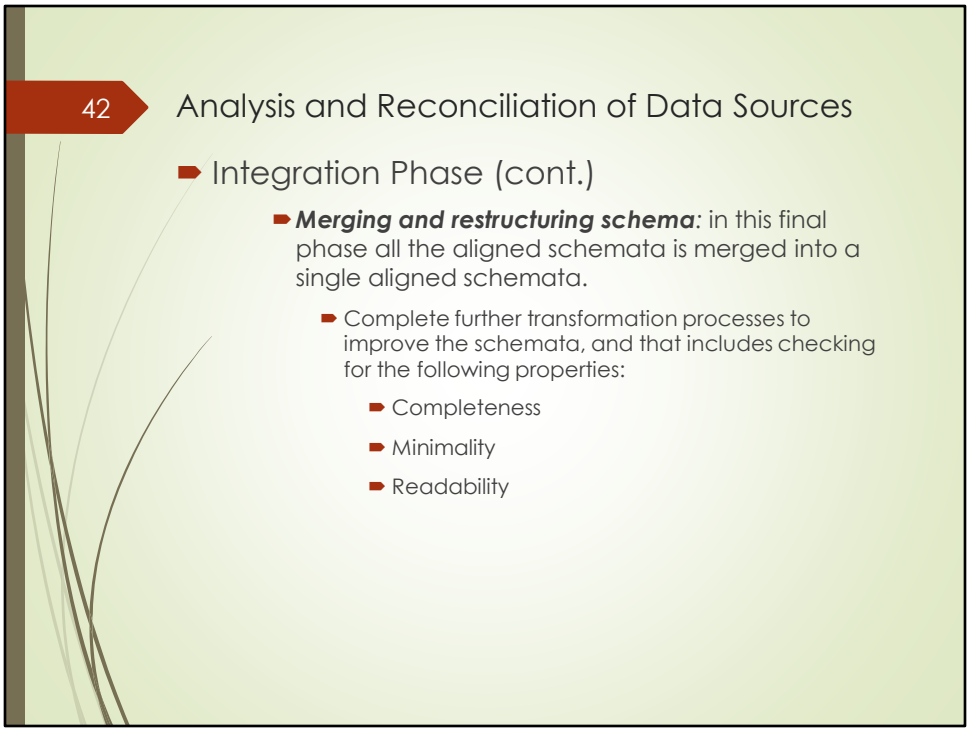


Integration Phase

Here you need to adopt a methodological approach.

The following list shows the steps needed to be taken in this phase:

- ***Preintegration***: this phase thoroughly analyzes data sources and leads to the definition of the general integration standards. The main decision here concern the following:
    - Parts of the schema to integrate.
    - Integration strategy
- ***Schema comparison***: consists of a comparison analysis of schemata. The goal in this phase is to discover relationships and conflicts in the schemata. There are types of conflicts that can be detected in this phase and they are:
    - Heterogeneity conflicts – point out inconsistencies due to the usage of formalisms with different expressive powers in source schema.
    - Name conflicts – these are due to the terms used in various data sources because of different perspectives adopted by designers.
    - Semantic conflicts – these occur when two source schemata model the same part of the real world at different abstraction and detail levels.
    - Structural conflicts – caused by different options for modeling the same concepts.
- ***Schema alignment***: goal of this phase is to solve any conflicts found in the proceeding step. To correct those conflicts you need to apply transformation primitives to the source schema. They typically deal with changes in names, attribute types, functional dependencies and existing constraints applied to the schemata.



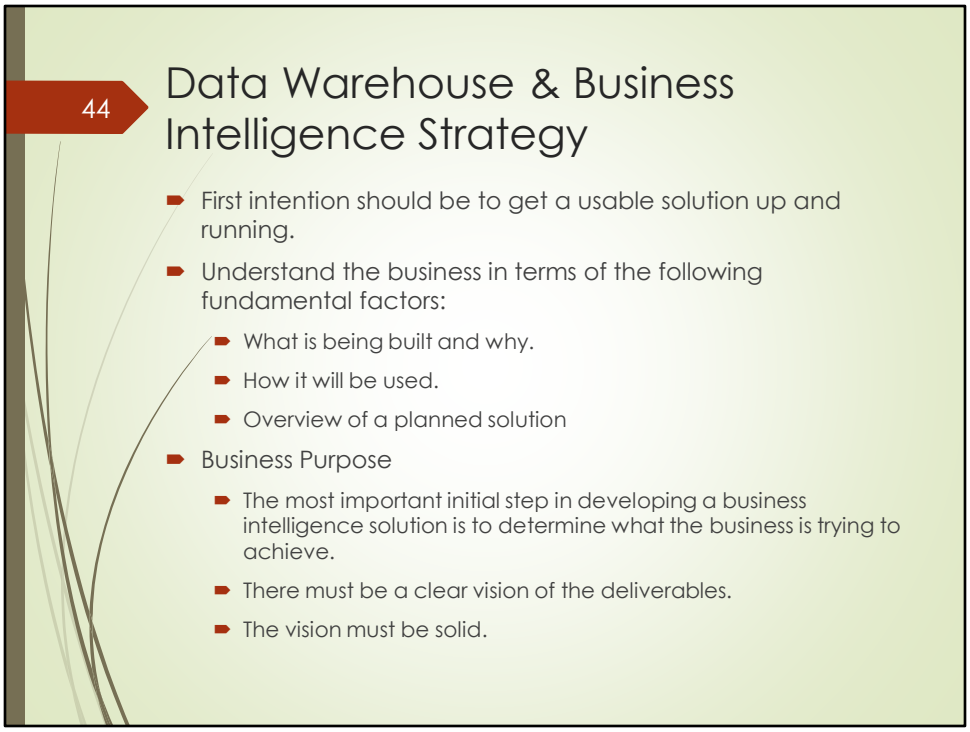
Merging and restructuring schema: in this final phase all the aligned schemata is merged into a single aligned schemata.

After this operation you should complete further transformation processes to improve the schemata, and that includes checking for the following properties:

- Completeness
- Minimality
- Readability

Data Warehouse & Business Intelligence Strategy

43



First intention should be to get a usable solution up and running as fast as possible to help support the business and ensure that it follows a data architecture strategy.

It is good to understand the business in terms of the following fundamental factors:

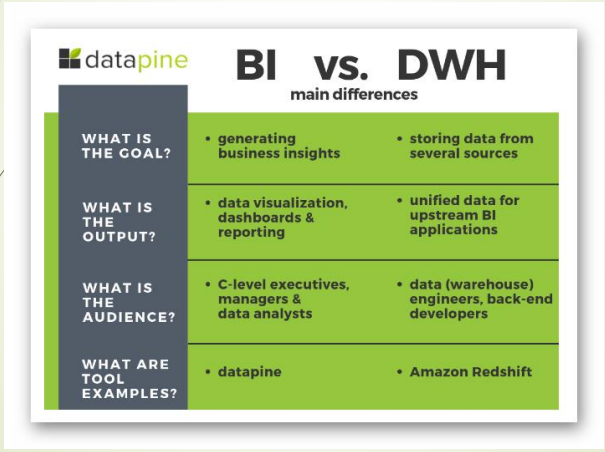
- What is being built and why.
- How it will be used.
- Overview of a planned solution

Business Purpose

- The most important initial step in developing a business intelligence solution is to determine what the business is trying to achieve and conceptually how this solution will be used in support of the business decision-making strategy.
- There must be a clear vision of the deliverables at the onset of this project.
- The vision must be solid.

45

Data Warehouse & Business Intelligence Strategy



(https://www.datapine.com/blog/data-warehousing-and-business-intelligence-architecture/)

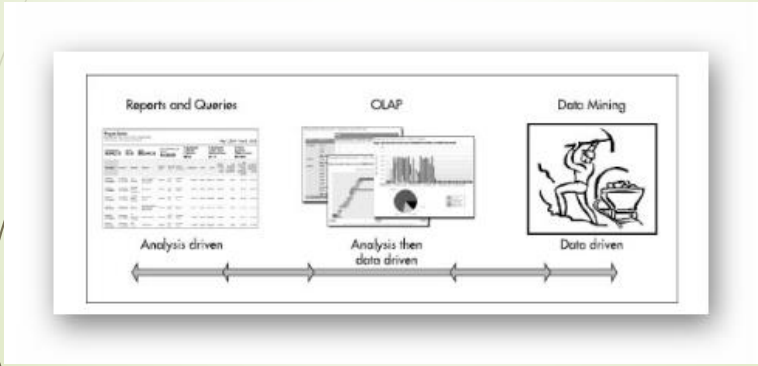
Business Usage

- The business usage is the second influencing factor toward the overall solution.
- The driving factor here is to have an understanding of the business purpose from their goal to the practical use of the system.
- The system must provide the users with a way to view the data in a communicative fashion which directly applies to the OLAP perspective, graphs, pie charts, etc..
- Advanced users should also have the ability to mine the data independently. This is called "data mining" and is purely a data-driven effort in hopes of discovering associations between data.
- All solutions are designed and planned to help empower the business users with data in such a manner that will help support their decision-making process.
- "One of the BI architecture components is data warehousing. Organizing, storing, cleaning, and extraction of the data must be carried by a central repository system, namely data warehouse, that is considered as the fundamental component of business intelligence." (Durcevic, 2019)

47

Data Warehouse & Business Intelligence Strategy

This diagram from out text shows a comparison analysis driven versus data driven business intelligence usage

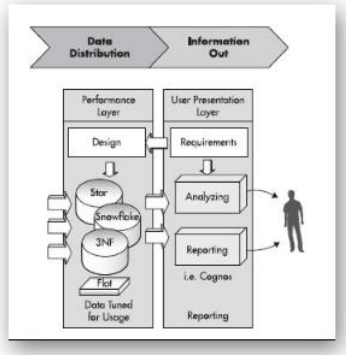


(Laberge, 2011)

48

Data Warehouse & Business Intelligence Strategy

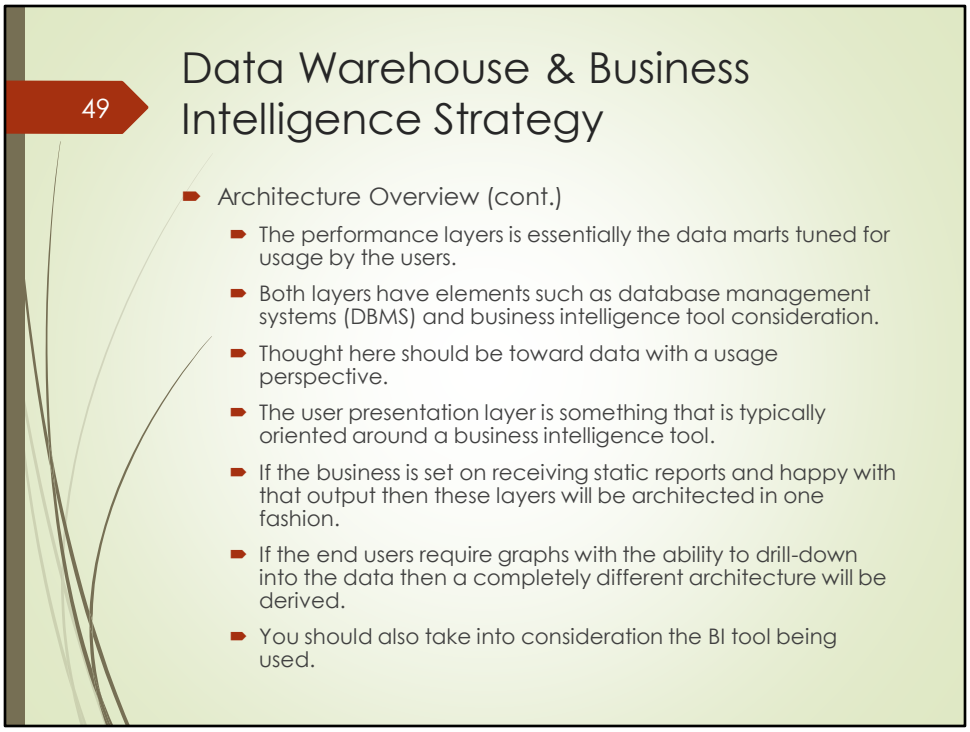
- Architecture Overview
  - How the business want to access the information can give great insights.



(Laberge, 2011)

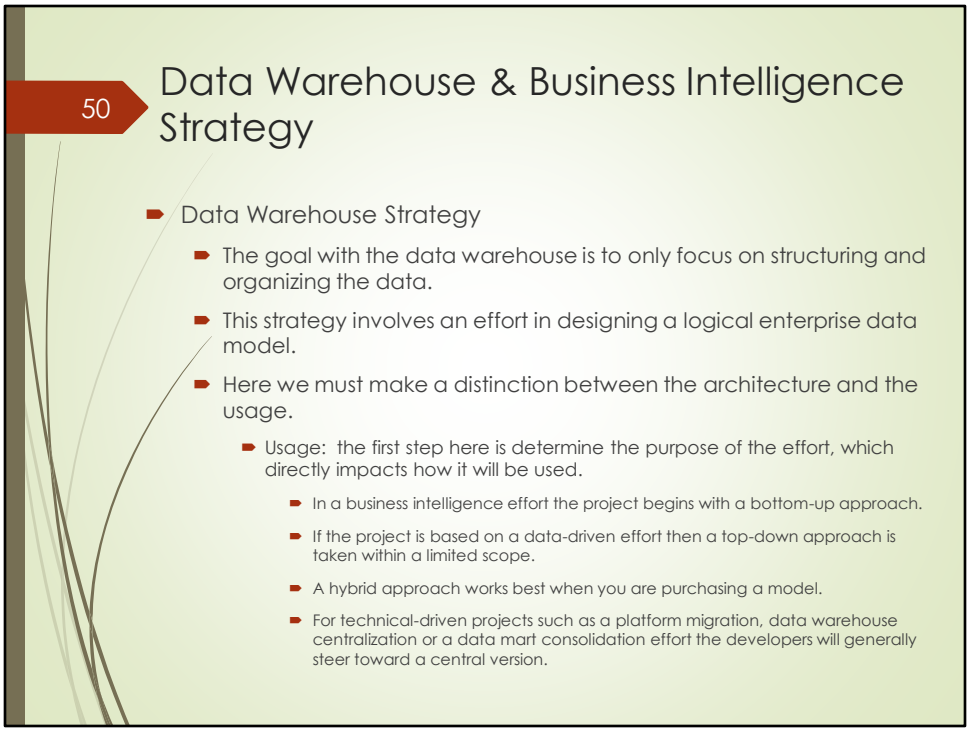
Architecture Overview

- How the business want to access the information can give great insights into the business intelligence architecture in the performance and user presentation layers as depicted below.



Architecture Overview (cont.)

- The performance layers is essentially the data marts tuned for usage by the users.
- Both layers have elements such as database management systems (DBMS) and business intelligence tool consideration.
- Thought here should be toward data with a usage perspective.
- The user presentation layer is something that is typically oriented around a business intelligence tool such as Cognos, Business Objects, MicroStrategy, or Crystal Reports.
- If the business is set on receiving static reports and happy with that output then these layers will be architected in one fashion.
- If the end users require graphs with the ability to drill-down into the data then a completely different architecture will be derived.
- You should also take into consideration the BI tool being used, since most of these tools work well in aggregating a lot of data. A star schema with dimensions and a lot of fact rows is great for most tools.



Data Warehouse Strategy

- The goal with the data warehouse is to only focus on structuring and organizing the data, which is in contrast with the business intelligence strategy which has a main focus of producing business user deliverables.
- This strategy involves an effort in designing a logical enterprise data model which should aid in gaining insights into the business data components.
- Here we must make a distinction between the architecture and the usage.
- Usage: the first step here is determine the purpose of the effort, which directly impacts how it will be used.
- In a business intelligence effort the project begins with a bottom-up approach. This means the business usage and purpose requirements are understood first.
- If the project is based on a data-driven effort then a top-down approach is taken within a limited scope. This scope is limited to help set a boundary on the breadth of the effort, and usually gives a realistic perspective to the project.
- A hybrid approach works best when you are purchasing a model because the data is already structured and defined.
- For technical-driven projects such as a platform migration, data warehouse centralization or a data mart consolidation effort the developers will generally steer toward a central version of data on a specific data base.

Data Warehouse Architecture

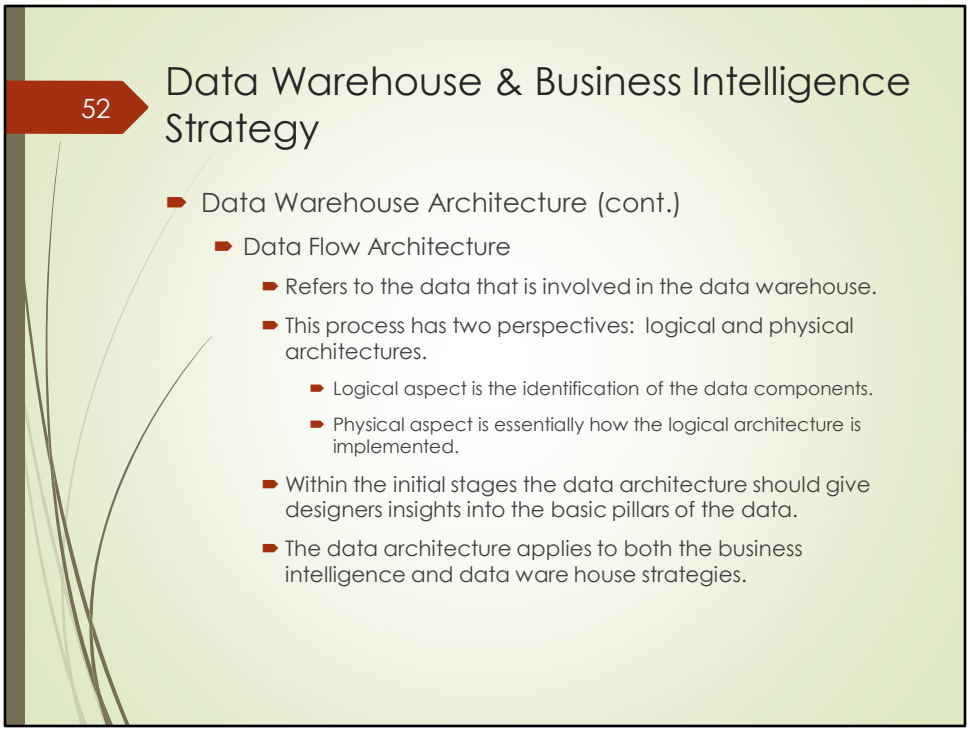
- This as to do with choosing a strategy to design, develop and possible build the repository and not around actually building the data warehouse.

- The initial step is to set a base foundation and understand the goal. Typically the purpose of any data warehouse is to centralize all corporate information.

- Here the reliability of the data is then referred to as the "trusted and central single version of the truth" for the entire organization.

- Data warehousing initially began as database environments but has since evolved into a system that includes not only the architecture but the business intelligence, aka the business usage, aspect of data.

- As with building anything with structure you must first start with a blueprint. The blueprint in this aspect consists of two levels: data flow architecture and technical architecture.



Data Warehouse Architecture (cont.)

Data Flow Architecture

- Refers to the data that is involved in the data warehouse and its anticipated route, flow or motion.

- This process has two perspectives: logical and physical architectures.

- Logical aspect is the identification of the data components and how the inter-related within one data model and between data-models.
- Physical aspect is essentially how the logical architecture is implemented. This physical aspect and implementation are highly dependent on the usage and logical design.

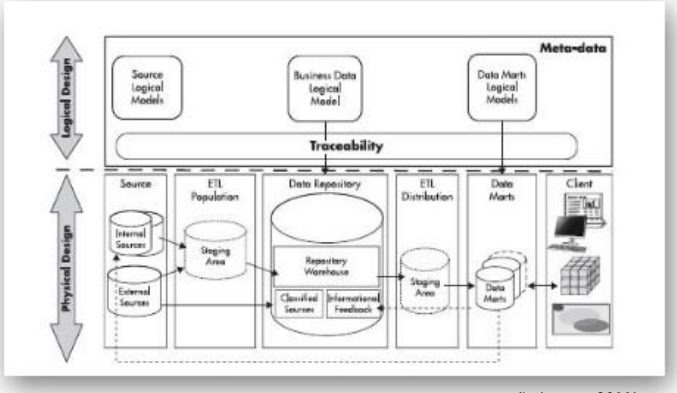
- Within the initial stages the data architecture should give designers insights into the basic pillars of the data for this first perceived phase of the project.

- The data architecture applies to both the business intelligence and data ware house strategies.

53

Data Warehouse & Business Intelligence Strategy

This diagram depicts a typical simple data warehouse architecture flow.

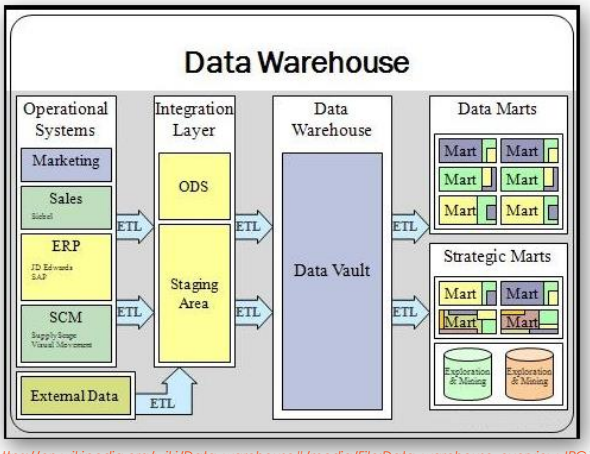


(Laberge, 2011)

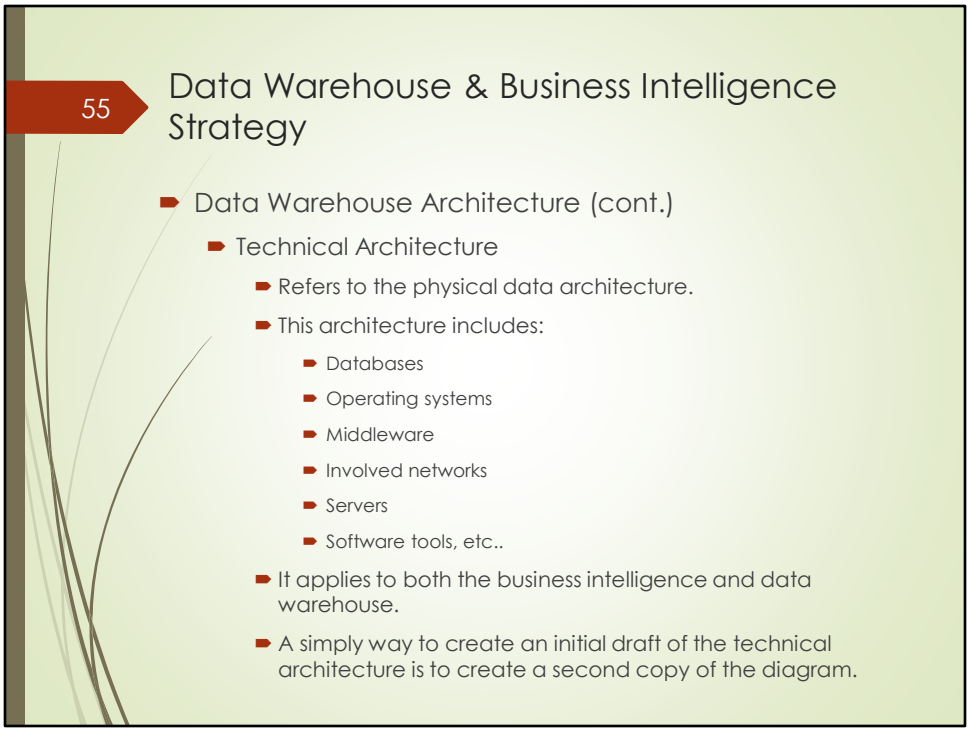
Data Warehouse & Business Intelligence Strategy

54

Another pictorial representation of a data warehouse architecture flow.



(https://en.wikipedia.org/wiki/Data_warehouse#/media/File:Data_warehouse_overview.JPG)



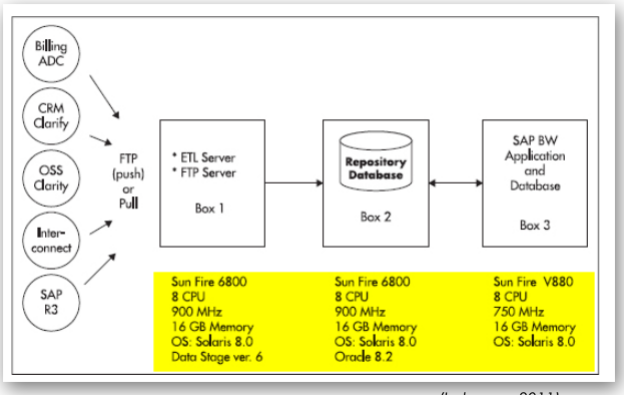
Data Warehouse Architecture (cont.)

Technical Architecture

- Refers to the physical data architecture and any tangible aspects of the physical system.
- This architecture includes:
    - Databases
    - Operating systems
    - Middleware
    - Involved networks
    - Servers
    - Software tools, etc..
- It applies to both the business intelligence and data warehouse strategies equally.
- A simply way to create an initial draft of the technical architecture is to create a second copy of the diagram which is identical to the data architecture diagram and just add the technical aspects as the chart from our text on the next slides shows.

Data Warehouse & Business Intelligence Strategy

This diagram depicts a simple technical architecture diagram.



(Laberge, 2011)

Focus and Success

Enterprise or Line of Business?

- Is this data warehouse initiative IT-driven or business driven?
- Both groups are involved in building this solution, and if one side tries to build this initiative by them selves the success will be one-sided.
- Both groups need to work tighter to ensure the business has a usable system, which will greatly enhance their business decision-making process. This will help ensure that the data assets is managed at the corporate perspective.
- A good strategy with the best probability of success for the organization is when both sides work together.

Goal Focused

- During the first phase of the project it is important not to get too carried away.
- The idea should be to figure out the story line for developing the data warehouse taking into consideration the resources, timeframe, budget, management support, data quality issues and political or cultural environment.

Focus and Success

Success: When are we done?

- Most projects begin without having a properly defined target goal.
- The best idea is to set a definition for success up front.
- The point here is break down the project into milestones for IT and business. Think of the project as being completed in chunks, not just with one deliverable.
- The project is completed only once the agreed-to-goals are attained by all the parties involved, and only at this point can the project as a whole be considered a success.

Where to start?

- The first step in trying to determine where to start is to determine what is to be built.
- Does the business want analytics, or is the purpose of the system to organize and structure the data throughout the organization, or both?
- The whole idea is to cultivate the corporate data asset.
- For BI: the first phase is to determine the type of BI environment based on usage.
- Batch reports – static reports
- Dynamic reports – having ability to drill down into the data
- OLAP reporting – same as dynamic reports but with more communicative presentations
- Data mining – advance uses can query the data
- Specialized: dashboards and/or scorecards – pre-planned specialized graphic reporting

Where to start? (cont.)

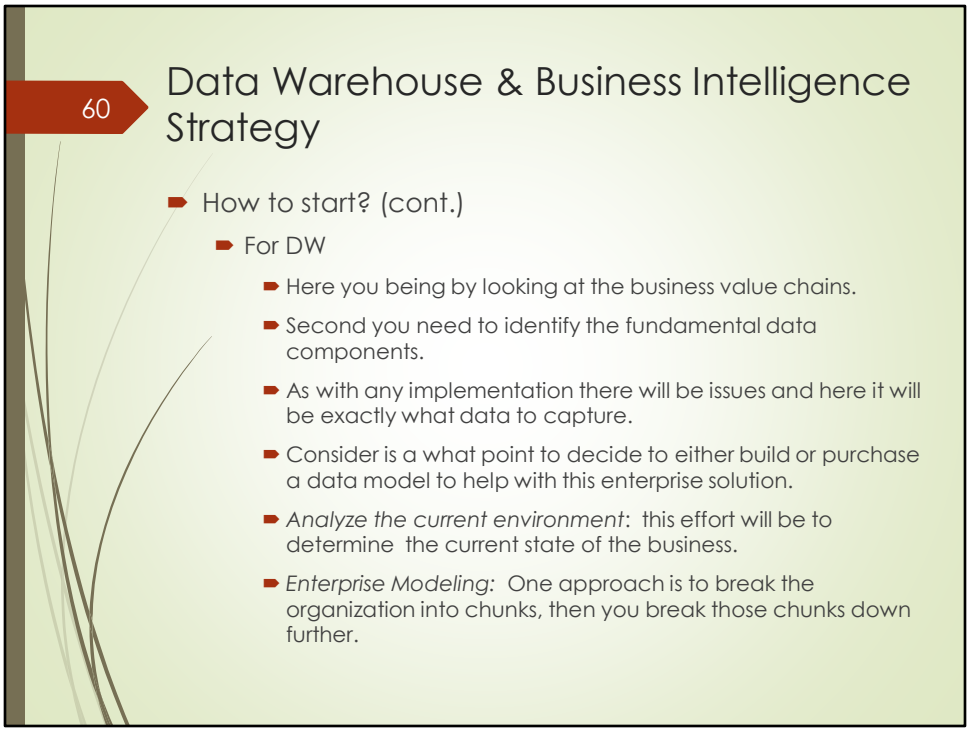
For DW: the first phase for projects that are not business intelligence-focused is to decide on purpose. What is the effort?

- To create a central enterprise data vocabulary?
- To model the business data to determine how all interconnects?
- To logically model a data warehouse repository to capture and hold all historical data?
- A combination of several of these?
- Since this scenario is data-driven the goal is to determine the fundamental data elements and how they interassociate along with their descriptive aspects.

How to start?

For BI

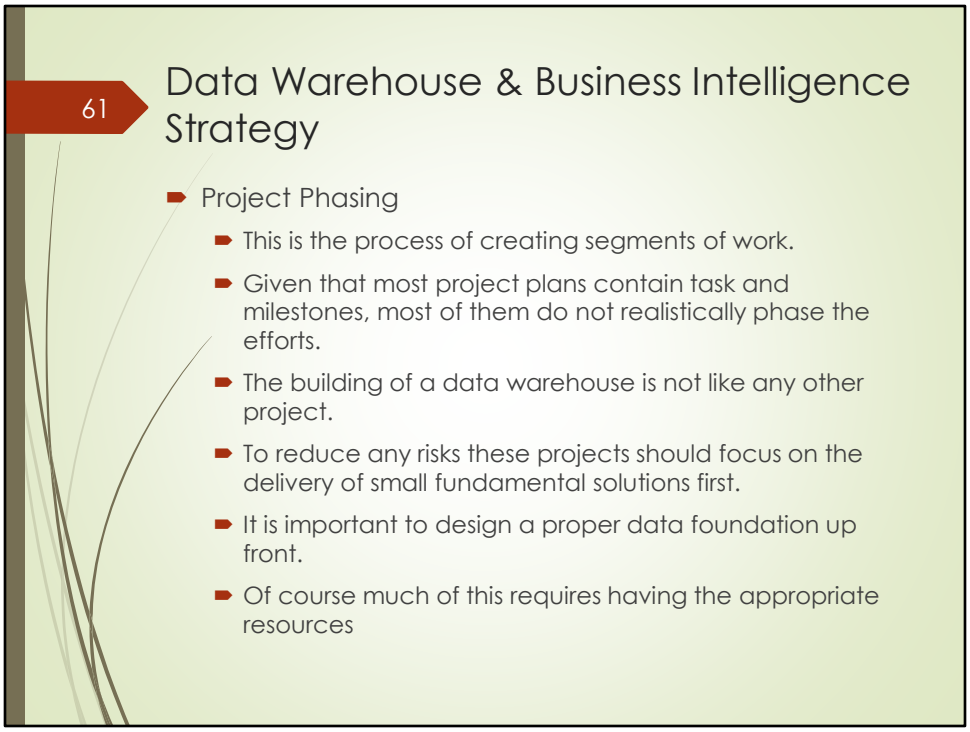
- If the business has existing report then you should obtain those and decompose them to extract them into their data components.
- Next, you should map the discovered data dimensions and measures to the source systems, basically to find out where the data came from.
- There are often gaps, which is why conducting a gap-analysis is important. This is used to help determine whether all the data is where it should be.



How to start? (cont.)

For DW

- Here you being by looking at the business value chains, which are the main business processed within the organization.
- Second you need to identify the fundamental data components such as customers, product and the main events for the business.
- As with any implementation there will be issues and here it will be exactly what data to capture. Without specific business usage requirements these types of questions will continually be an annoyance to the project.
- Something else you should consider is a what point to decide to either build or purchase a data model to help with this enterprise solution.
- *Analyze the current environment*: this effort will be to determine the current state of the business regarding data sources, the technical environment and the available resource base to fulfil the project tasks and goals.
- *Enterprise Modeling:* One approach is to break the organization into chunks, then you break those chunks down further. These chunks are known as subject areas, and the further broken-down sections are concepts. Having very high-level anchors is the key.



Project Phasing

- This is the process of creating segments of work, each which have a set start, effort and end with a defined deliverable that moved the entire project forward.
- Given that most project plans contain task and milestones, most of them do not realistically phase the efforts.
- The building of a data warehouse is not like any other project given that this type of project spans a lot more of the organization.
- To reduce any risks these projects should focus on the delivery of small fundamental solutions first, followed then by tangible end-user solutions.
- It is important to design a proper data foundation up front for the enterprise expected business requirements.
- Of course much of this requires having the appropriate resources, and their ability to understand the business, the data components, and how these two overlap.

Project Resources: Roles and Insights

62

Project Teams

- These are the lowest level of the hierarchy, given they usually take on the fundamental building task for any organization.

- Since projects have a start and an end date individuals are taken, or borrowed from their respective functional group to work with others on a project, typically referred to as a *matrix reporting scenario*, simply because these individuals report to their functional managers and to the project manager.

- Individuals within the organization which have the most years of experience within their respective areas of expertise tend to take on a more authoritative role.

Senior Expertise

- Only senior-level IT expertise should join a data warehouse project.

- These projects require a strategic level design since it covers the entire enterprise.

- This term "senior-level expertise" refers to an individual within the organization with senior level ability to innovate, and has the experience of having built a data warehouse before.

Leadership

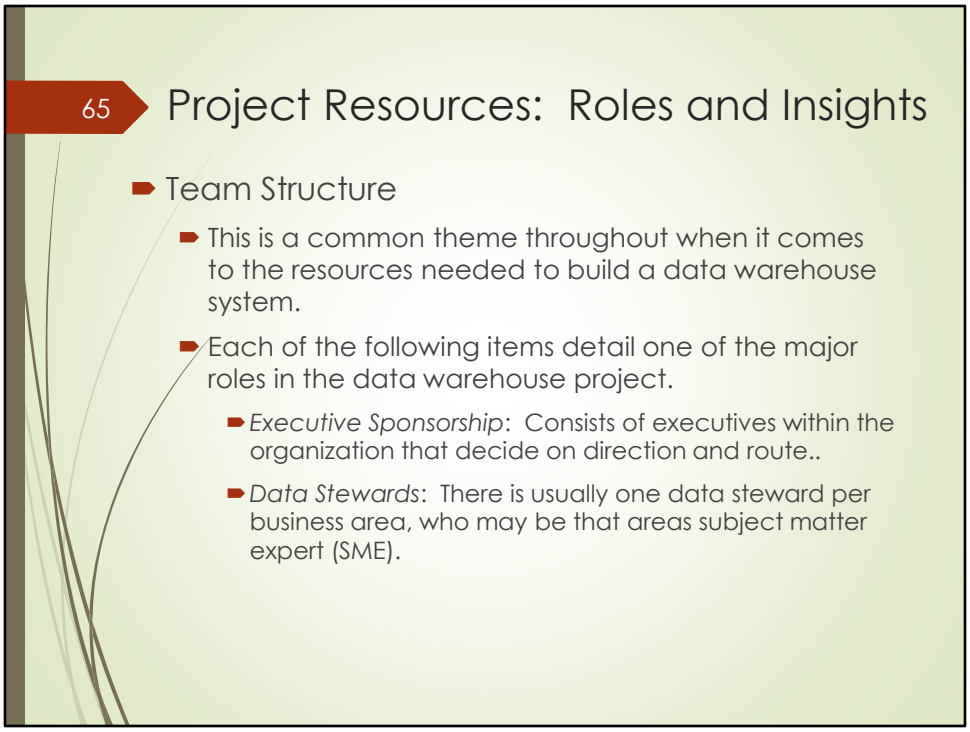
- This is where we ensure the team members play well together.
- Senior level resources tend to lean inwards and can build based on their beliefs and experiences which at times can go against the grain of the group.
- So with this in mind the project manager must take on a leadership role without seeming authoritative and only manage based on common sense and the best approach for the scenario at hand, while empowering the team members.
- Managing a data warehouse can be quite difficult, especially with cross-social cultures and globalization.
- As is true with any management role it is the managers task to ensure they instill a culture suitable for the project goals give the team members.

Project Sponsor

- This role is a must to ensure the project is moving along appropriately.

Data Warehouse Executive

- Many organizations, especially financial ones, are creating vice president level roles specifically charged with owning the data warehouse system, and creating optimized business intelligence solutions.



Team Structure

- This is a common theme throughout when it comes to the resources needed to build a data warehouse system.

- Each of the following items detail one of the major roles in the data warehouse project.

- *Executive Sponsorship*: Consists of executives within the organization that decide on direction and route. Without a sponsor there is no real push for business usage, which essentially defeats the entire purpose of building the data warehouse. Given the system is built by IT and used by the business there needs to be sponsorship from both areas.

- *Data Stewards*: There is usually one data steward per business area, who may be that areas subject matter expert (SME). The respective data marts are overseen by the data stewards. They are the main point of contact for their specific business areas. They normally take responsibility in being aware of changes to the system and the impacts those may have on the data warehouse system. They can also assist with resolving data quality issues, while relying heavily on the executive sponsors for support.

Project Resources: Roles and Insights

- Team Structure (cont.)
  - *Basic Resources*: For this type of initiative there are fundamental resources that need to be on the project team:
    - ***Project manager*** – this position is a must have and is dedicated to the development and coordination of efforts for the data warehouse.
    - ***Data warehouse architect*** – this role works hand-in-hand with the project manager and is responsible for the detailed development of the final product.
    - ***Business analyst*** – this role helps determine the business requirements via business user interviews and then organizes the information for the team. They are ultimately responsible for development the functional specification for the final project.
    - ***Business lead (SME)*** – this role will work side-by-side with the business analyst to relay business requirements, business processes and business rules to the team.

Project Resources: Roles and Insights

- Team Structure (cont.)
  - ***Data modeler*** – this role is responsible for modeling the data warehouse repository and the data marts logically (design), therefor they must understand the implications of the physical design.
  - ***Database administrator (DBMS and application level)*** – required to transform the data model design into a physical or optimized data model. The DBMS DBA is tasked with ensuring that the database system is operational, as well as ensuring and administrating the infrastructure (servers). The application level DBA is more involved in ensure optimized data loads and retrieval for specific business usage.
  - ***ETL programmer*** – this roles main focus is capturing the data from the source system, transforming it to load into the database.
  - ***Business intelligence (report and analytics) programmer*** – they use the underlying data marts to support the business reports determined by the business analyst.
  - ***Application tester and user trainer*** – since someone must test the system to ensure it operates as designed there is the application tester. They report back the project manager any defects that need to be reworked.

68

Project Resources: Roles and Insights

- Periodic Reviews: Progress Audit
  - It is highly recommended to review the progress regularly.
- Center of Competence (CoC)
  - This is typically setup to manage and enhance the business intelligence system as well as handling ongoing business ad-hoc query request.
  - Must also contain the business intelligence report and query writers.
  - It is the call center for business users in this forum.
  - It can help with questions regarding the data components, ad-hoc queries, or any other IT help regarding the developed system.
  - This entity works towards:
    - Data management
    - Information delivery
    - BI coordination efforts

Project Resources: Roles and Insights

When is a COC not the right approach?

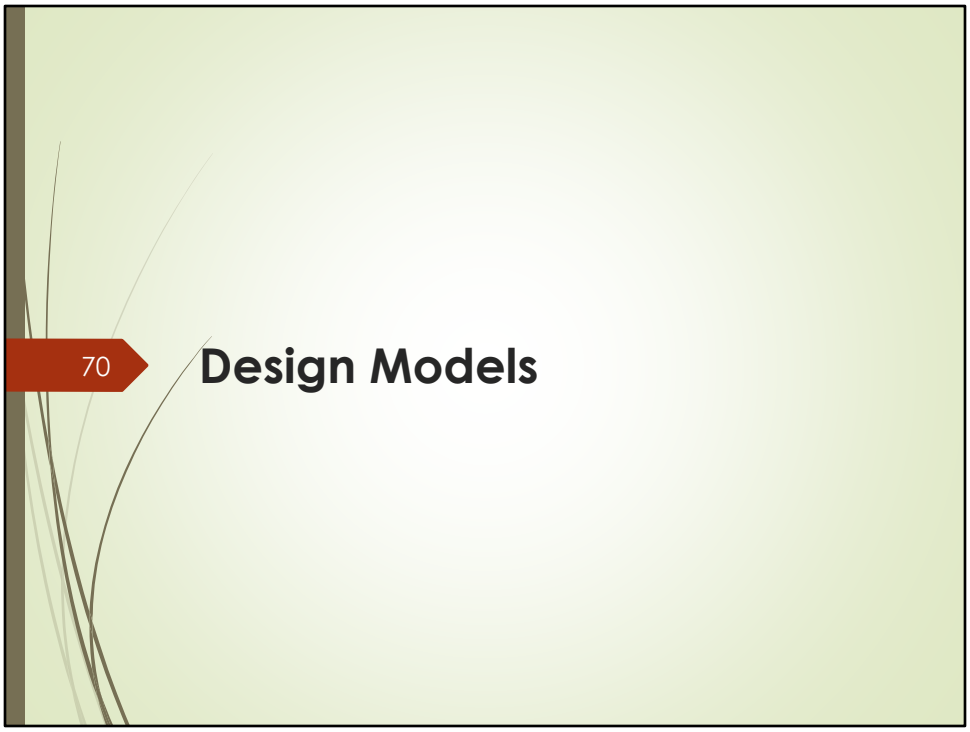
COCs aren't perfect. In fact, there are several situations in which they become absolutely counterproductive. The following situations are examples of when a COC is not the right approach.

- The COC becomes part of the problem rather than the solution: COC members need a special type of attitude and personality. They need to be able to both reassure developers that new ways of working are possible, and be sensitive to the disruption that new technologies cause, while at the same time push teams on to new heights. COC members can't become advocates for why things can't change. If a COC team member starts identifying too much with a particular project's problems and isn't providing new ideas and solutions, that member must be disengaged in favor of a new COC team member.

- The COC becomes a political battleground or "retirement home:" A COC must have specific goals and a specific funding model that is supported preferably at the CTO or CIO level. If a COC is not meeting those goals and can't consistently demonstrate value to the organization, the COC must be disbanded before it becomes another asset to be acquired in a turf war among executives.

  Just as bad as a COC that spends all of its time in funding or political battles is one that is seen as a sinecure for architects or project managers that are no longer involved in day-to-day development. A COC must be on top of the absolute latest technologies and development principles. The best way to achieve that is to have a mix of experience and youth that can cross-pollinate and generate ideas.

https://www.ibm.com/cloud/garage/practices/culture/practice_center_of_competency/


70
Design Models

71

Different data models have different benefits.

- A conceptual model is primarily for understanding and communicating its components.
- A logical data mart model typically communicates an analysis focus and
- A physical data model represents tables and columns for a specific DBMS. (Laberge, 2011)

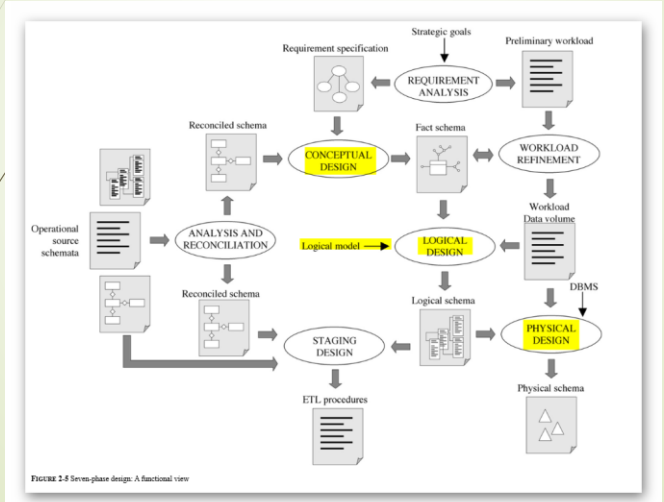


FIGURE 2-5 Seven-phase design: A functional view

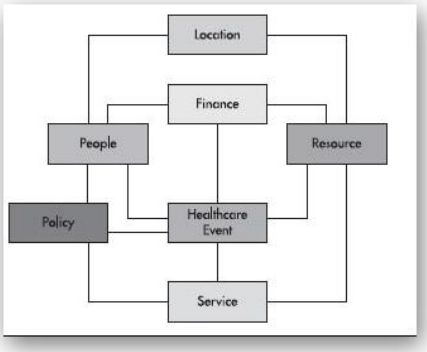
Different data models have different benefits.

- A conceptual model is primarily for understanding and communicating its components.
- A logical data mart model typically communicates an analysis focus and
- A physical data model represents tables and columns for a specific DBMS. (Laberge, 2011)

72

Conceptual Design

- According to our approach, a conceptual design implies that the user requirements of the previous phase are exploited
- This model is used to communicate the business fundamentals.



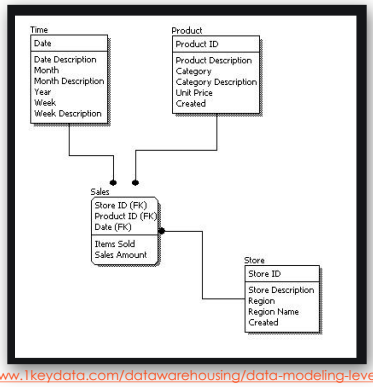
- According to our approach, a conceptual design implies that the user requirements of the previous phase are exploited to create a conceptual schema for a data mart on the basis of a reconciled schema (Golfarelli & Rizzi, 2009)

- This model is used to communicate the business fundamentals and how the business operates from a business and data perspective within the organization. (Laberge, 2011)

73

Logical Design

- The essential requirement for logical design is naturally the selection of a logical model
- Data models can be logical, which is similar to blueprints for a house



https://www.1keydata.com/datawarehousing/data-modeling-levels.html

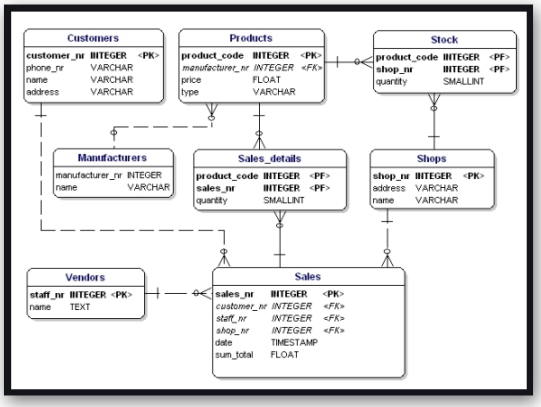
- The essential requirement for logical design is naturally the selection of a logical model to act as a reference framework. (Golfarelli & Rizzi, 2009)

- Data models can be logical, which is similar to blueprints for a house, or physical, which represents the optimized logical data model tuned for creating database tables for specific databases. (Laberge, 2011)

74

Physical Design

- The most important problem of the physical design phase is the selection of the indexes .
- A physical database model shows all table structures.

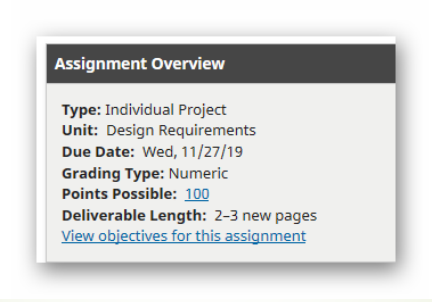


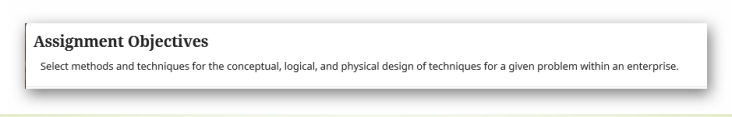
https://www.datanamic.com/support/lt-dez005-introduction-db-modeling.html

- The most important problem of the physical design phase is the selection of the indexes to optimize performance. (Golfarelli & Rizzi, 2009)
- A
- physical database model shows all table structures, including column name, column data type, column constraints, primary key, foreign key, and relationships between tables.

75

Unit 2 – Individual Project





Unit 2 – Individual Project

76

Assignment Description

For this assignment, you will need to design the data warehouse requirements.

- Update the Data Warehouse Design Document title page with a new date and project name.
- Update the previously completed sections based on your instructor's feedback.
- **Week 2: Design Requirements**
  - Design the conceptual, logical, and physical aspects of the data warehouse based on requirements.
- Update your table of contents before submission.
- Name the document "yourname_CS683_IP2.doc."
- Submit the document for grading.

Worked Example

Please refer to the Worked Example below for an example of this assignment based on the Problem-Based Learning Scenario. The worked example is not intended to be a complete example of the assignment, but **it will illustrate the basic concepts that are required for completion of the assignment and can be used as a general guideline for your own project.** Your assignment submission should be more detailed and specific and should reflect your own approach to the assignment rather than just following the same outline provided in the worked example.

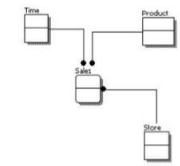
77

Worked Example

Design Requirements (Week 2)

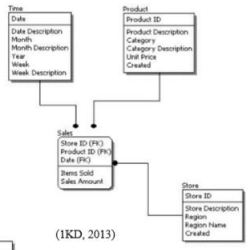
Design the conceptual, logical, and physical design based on the requirements. An example of a sales conceptual, logical, and physical design for Sky Products may look like the following.

Conceptual Model Design



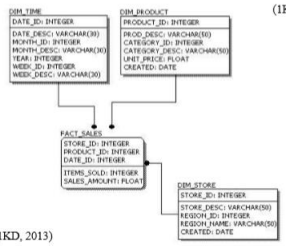
(1KD, 2013)

Logical Model Design



(1KD, 2013)

Physical Model Design



(1KD, 2013)

78

Websites

Conceptual Model Design

https://en.wikipedia.org/wiki/Conceptual_model

https://sites.google.com/site/thedesignofeverydaythings/home/conceptual-model

https://www.sciencedirect.com/topics/computer-science/conceptual-modeling

Logical Model Design

https://www.1keydata.com/datawarehousing/logical-data-model.html

https://www.guru99.com/data-modelling-conceptual-logical.html

https://www.sciencedirect.com/topics/computer-science/logical-data-model

Physical Model Design

https://www.1keydata.com/datawarehousing/physical-data-model.html

https://www.visual-paradigm.com/support/documents/vpuserguide/3563/3564/85378_conceptual.l.html

https://www.sciencedirect.com/topics/computer-science/physical-data-model

79

Unit 2 – Individual Project

Case Study: Problem 2

The data warehouse committee has received the first part of the data warehouse document, the requirements definition. The requirements will now be used to design the conceptual, logical, and physical aspects of the data warehouse, with the goal of providing information from all 3 departments in one customer view by customer account number.

Note that the worked example includes material from previous worked examples. The new material will be found under the Week 2 sections of the Table of Contents.

The worked example is provided here to help with this assignment.

Please submit your assignment.
For assistance with your assignment, please use your text, Web resources, and all course materials.

Unit 2 – Individual Project

Model Answer

The students must complete the following within the design requirements section:

Design the conceptual, logical, and physical aspects of the data warehouse based on requirements.

The report should be formatted according to APA guidelines. Grammar, spelling, punctuation, and format should be correct and professional.

81

Assignment Due Dates

Assignment Name	Due Date	Type
Unit 1 - Discussion Board	Fri, 4/10/20	Discussion Board
Unit 1 - Discussion Board 2	Tue, 4/14/20	Discussion Board
Unit 1 - Individual Project	Wed, 4/15/20	Individual Project
Unit 2 - Discussion Board	Tue, 4/21/20	Discussion Board
Unit 2 - Individual Project	Wed, 4/22/20	Individual Project
Unit 3 - Discussion Board	Tue, 4/28/20	Discussion Board
Unit 3 - Individual Project	Wed, 4/29/20	Individual Project
Unit 4 - Discussion Board	Tue, 5/5/20	Discussion Board
Unit 4 - Individual Project	Wed, 5/6/20	Individual Project
Unit 5 - Discussion Board	Tue, 5/12/20	Discussion Board
Unit 5 - Individual Project	Wed, 5/13/20	Individual Project

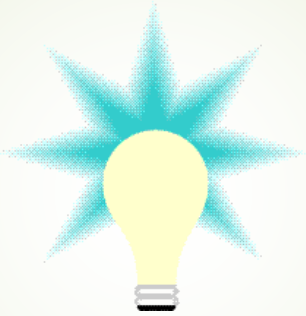
82

Contact Information

- My e-mail address- JConklin@coloradotech.edu
- Office Hours - Wednesday 6:00 P.M. – 7:00 P.M. CST
  Saturday 11:00 A.M. – 12:00 P.M. CST
- Live Chats - Thursday 7:00 P.M. – 8:00 P.M. CST

* Please note that only one live chat session per week is required for this course. However, optional live chat sessions may be held sporadically throughout the course.

Questions / Comments



(Ideas/Think Web Graphics, 2019).

References

Colorado Technical University. (2019). Instructor's guide for CS 683-1903B-01. Retrieved from Colorado Technical University Online, Virtual Campus, Course Overview: https://campus.ctuonline.edu

Connolly, T. and C. Begg (2015). Database systems; a practical approach to design, implementation, and management, 6th ed. Portland, Pearson Education

Durcevic, S. (2019). The Role Of Data Warehousing In Your Business Intelligence Architecture. Retrieved from https://www.datapine.com/blog/data-warehousing-and-business-intelligence-architecture/

Golfarelli, M., Rizzi, Stefano (2009). Data Warehouse Design: Modern Principles and Methodologies McGraw Hill.

Ideas/Think Web Graphics. (2019). In *Desktop Publishing*. Retrieved from: http://desktoppub.about.com/od/freeclipart/l/blidea1.htm

Kraynak, J. (2017). Cloud Data Warehouse for Dummies. Hoboken, NJ, John Wiley & Sons, Inc.

Laberge, R. (2011). The Data Warehouse Mentor: Practical Data Warehouse and Business Intelligence Insights McGraw Hill.