# Colorado Technical University

## CS683 – Data Warehouse

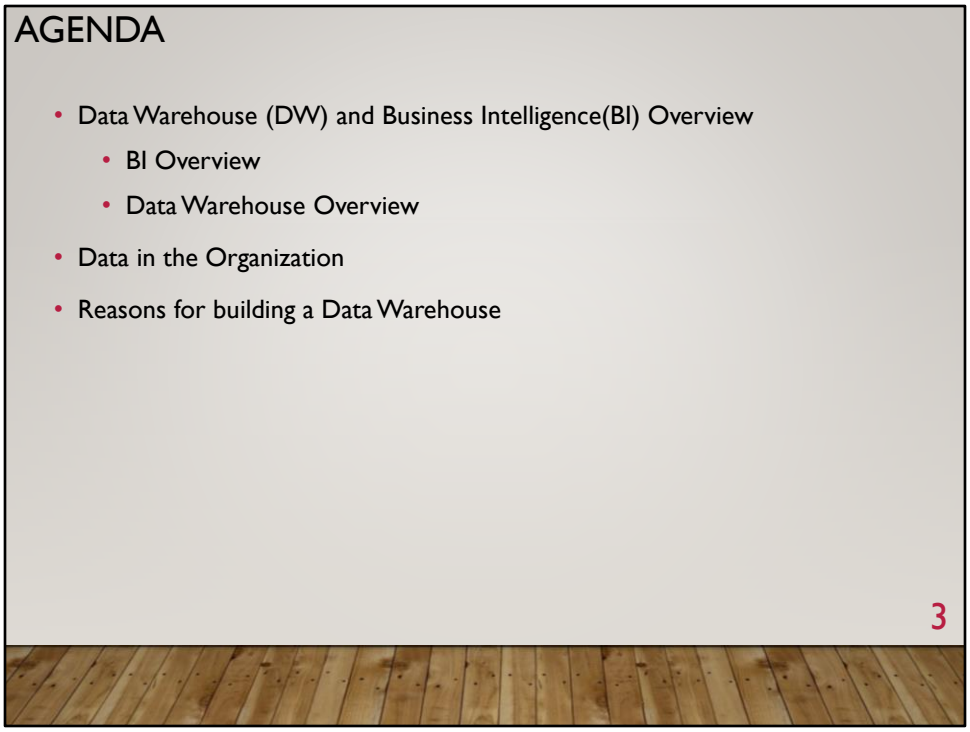
COLORADO TECHNICAL UNIVERSITY
INSTRUCTOR: DR. JOHN CONKLIN
UNIT 1 – DATA WAREHOUSE REQUIREMENTS

# DATA HUMOR


@SOURCE

https://bigdata-madesimple.com/dilberts-20-funniest-cartoons-on-big-data/

2




**Data Warehouse and BI Overview**:

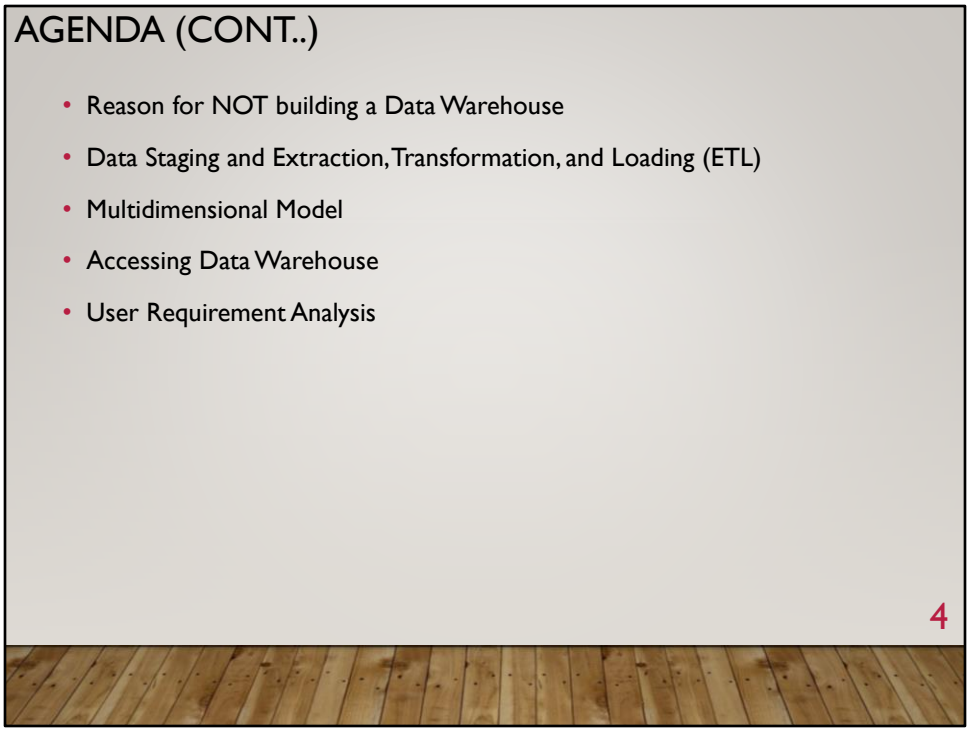
**BI Overview**: Definition of BI, Value of BI, Purpose of BI, User Presentation

**Data Warehouse Overview**: Definition, Data Warehouse System, Data Warehouse Architecture

Data Flow Terminology, Data Warehouse Purpose, Data Warehouse Value, Data Warehouse Best Practices

**Data in the Organization**: Data in Context, Data Quality, Data Vocabulary, Data Components, Data Architecture

**Reasons for building a Data Warehouse**: Platform Migration, Business Continuity, Reverse Engineer




**Reasons for NOT building a Data Warehouse**: Poor Data Quality, Lack of Business Interest, Lack of Sponsorship, Unclear Focus

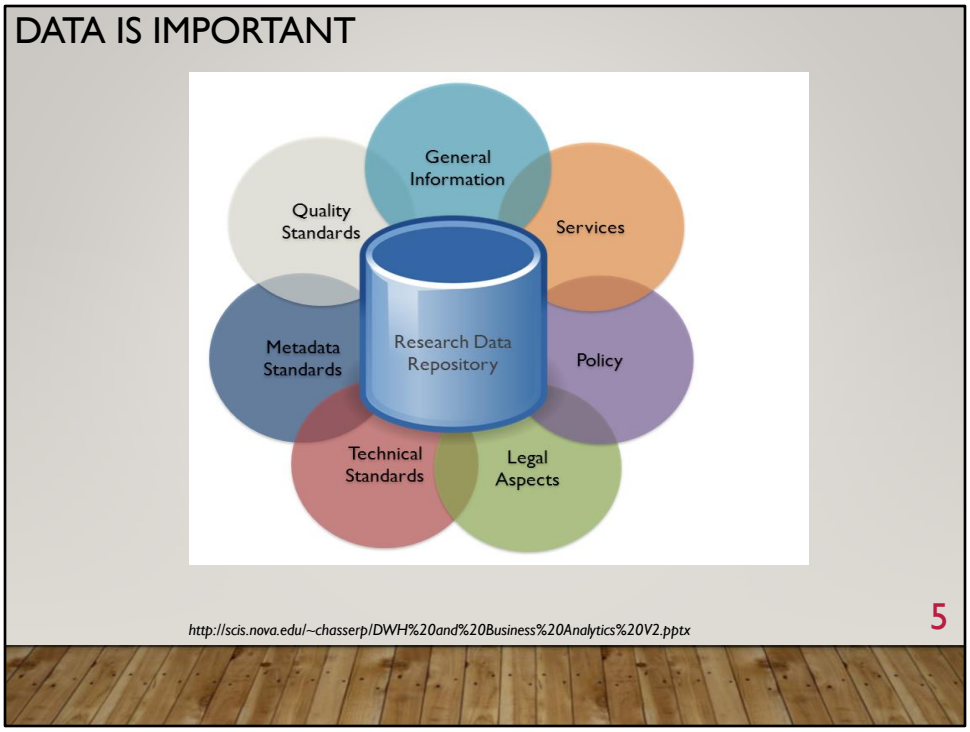
Sufficiency of Current Systems, Lack of Resources

**Data Staging and Extraction, Transformation, and Loading (ETL):** Extraction, Transformation, Loading

**Multidimensional Model:** ROLAP, MOLAP, HOLAP, DOLAP

**Accessing Data Warehouse**

**User Requirement Analysis:** Interviews




A lot of data is available to us BUT We have to make it valuable, good quality and easy to use

## WHY IS DATA QUALITY CRITICAL?

- Confidence…
- Avoids…
- Enhances…
- Better…
- Reduces…
- Add value..
- Improves…

*http://scis.nova.edu/~chasserp/DWH%20and%20Business%20Analytics%20V2.pptx*

- Confidence in making decisions
- Avoids effects of data contamination
- Enhances strategic decision making
- Better customer service
- Reduces costs in decisions
- Add value to service
- Improves productivity

## DATA QUALITY CHALLENGES

- System…
- Data…
- Heterogeneous…
- Poor…
- Incomplete…
- Input…
- Internationalization/localization
- Fraud
- Lack…

*http://scis.nova.edu/~chasserp/DWH%20and%20Business%20Analytics%20V2.pptx*

- System conversion
- Data aging
- Heterogeneous system integration
- Poor database design
- Incomplete information at data entry
- Input errors
- Internationalization/localization
- Fraud
- Lack of policies

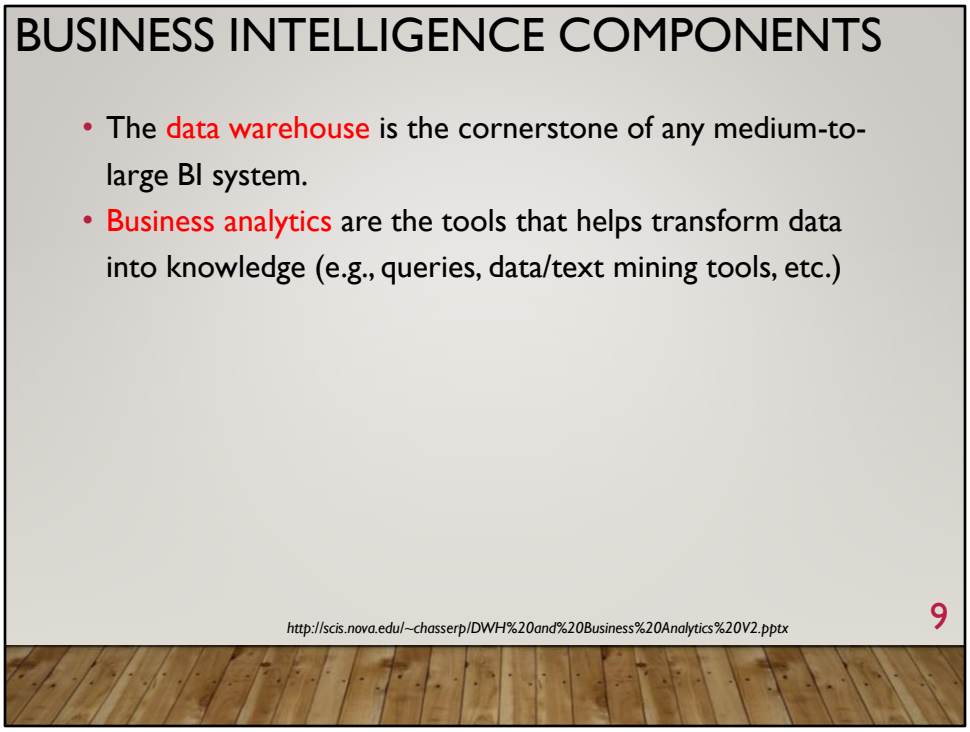
## BUSINESSES NEED STRATEGIC INFORMATION

- Formulate…
- Establish…
- Set business…
- Monitor business…
- Close the gap...
- To do this, we use…

*http://scis.nova.edu/~chasserp/DWH%20and%20Business%20Analytics%20V2.pptx*

8

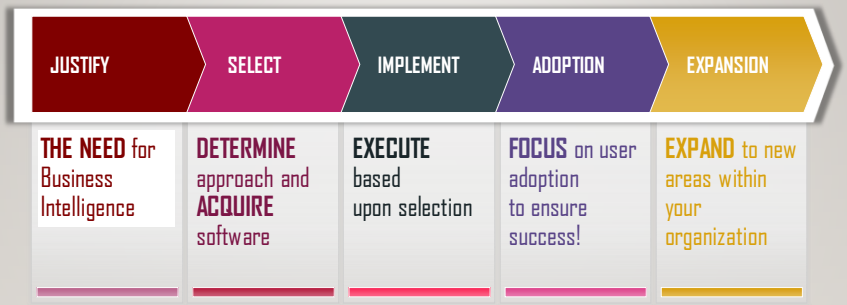
- …business strategies
- …business goals
- …objectives
- …results
- …between the current performance of an organization and its desired performance.
- …Business Intelligence (BI).




Data Warehouse:

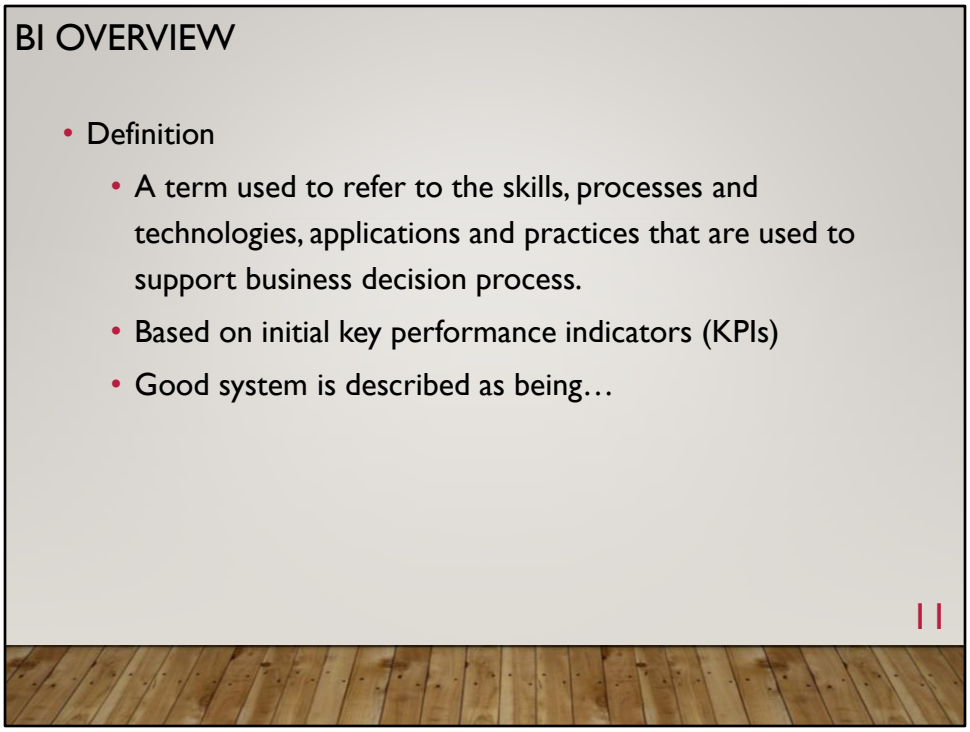
- Original data warehouses included only historical data that was organized and summarized, so end users could easily view or manipulate it.
- Today, many data warehouses include current data as well for real-time decision support

# Business Intelligence Components




10

HTTP://MAGNITUDE.COM/WP-CONTENT/UPLOADS/2014/01/WHAT-ARE-THE-STEPS-TO-BI-SUCCESS-2013-03-05.PPTX

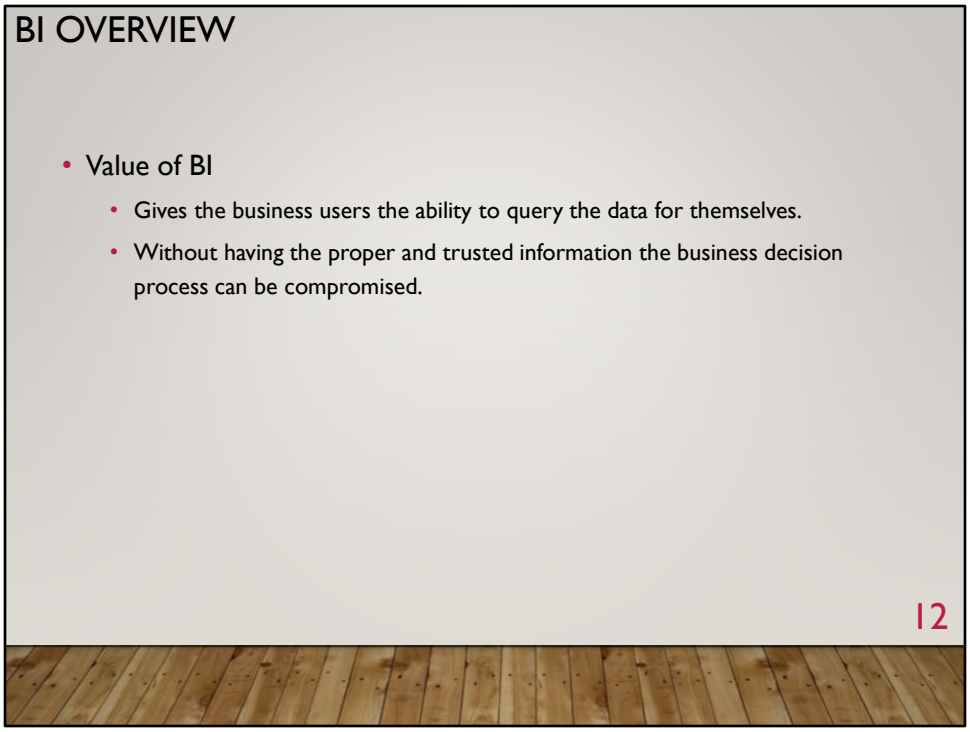



Based on initial key performance indicators (KPIs)

- KPIs can be broken down into measures, facts which normally numeric and quantifiable.

Good system is described as being:

- Accurate – data is trusted
- Timely – data is available on a regular study
- High Value – it is useful to the business users
- Actionable – information can be used in the organizations business decision process.

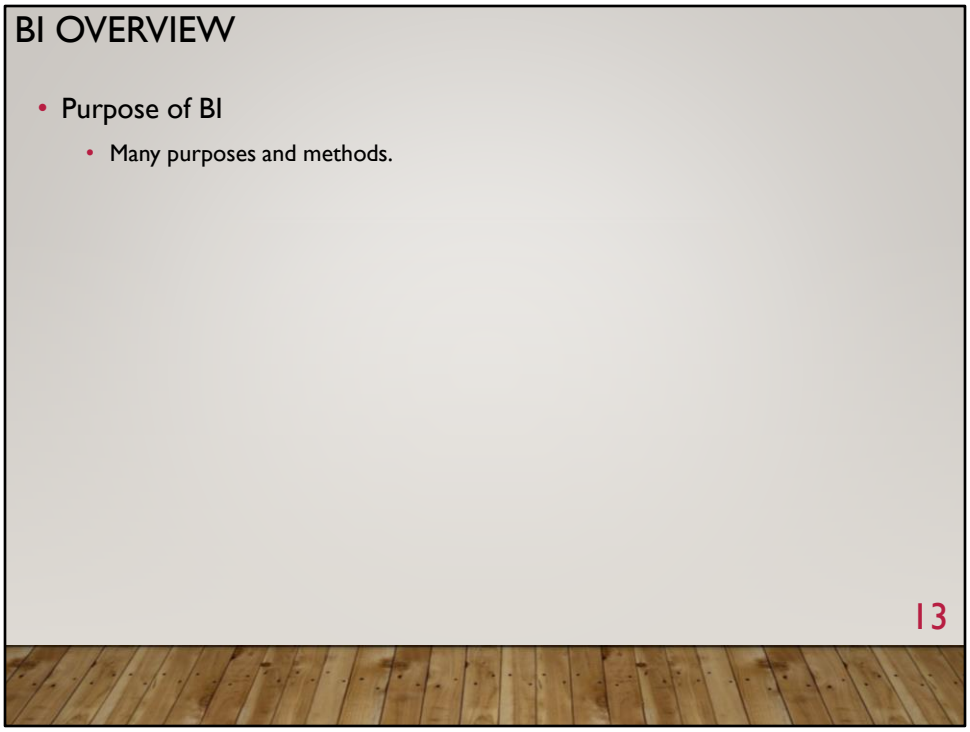



Ability:

- Provides them the opportunity to make business decisions in a shorter timeframe.
- Increases the business user's ability to process information.
- In an ideal scenario BI is:
  - Empowerment – directly usable
  - Fast – responsive
  - Timely – available
  - Accurate – trusted with quality
  - Usable – has value to the users.

Business Decision Process:

- Data is fundamental to information and it must be trusted.
- Must have a high level of reliability, aka quality, otherwise know as integrity

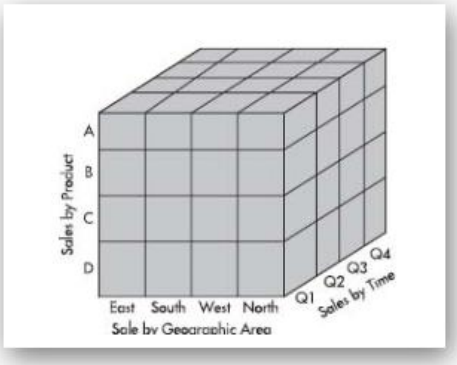



Purpose and Methods:

- Meaning that there are many different of business intelligence and business analysis.
- Some of the main purposes of BI are the following:
  - Benchmarking and/or baselining
  - View trends of predictive analysis
  - Performing market base analysis
  - Used in data mining
  - Analysis of selected subject areas.
- Each of the above items serve a specific purpose.
- The point behind these functions is to understand the purpose of business analysis and to create business intelligence based on that purpose.

## BI OVERVIEW

- User Presentation
  - Presentation of BI data can be presented in many forms.




*OLAP multidimensional cube (Laberge, 2011)*

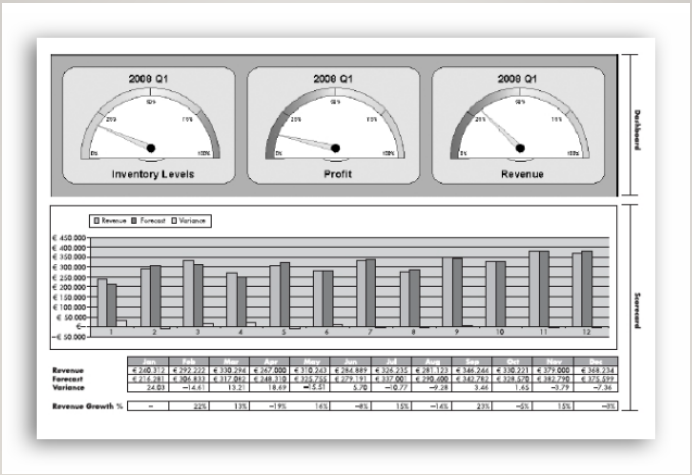
14

Presentation of BI:

- Reports – static, typically pre-run routines
- Queries – when you need to look into specific details
- OLAP – Online analytical processing – provides dynamic access to typical static reports

## BI OVERVIEW

- User Presentation

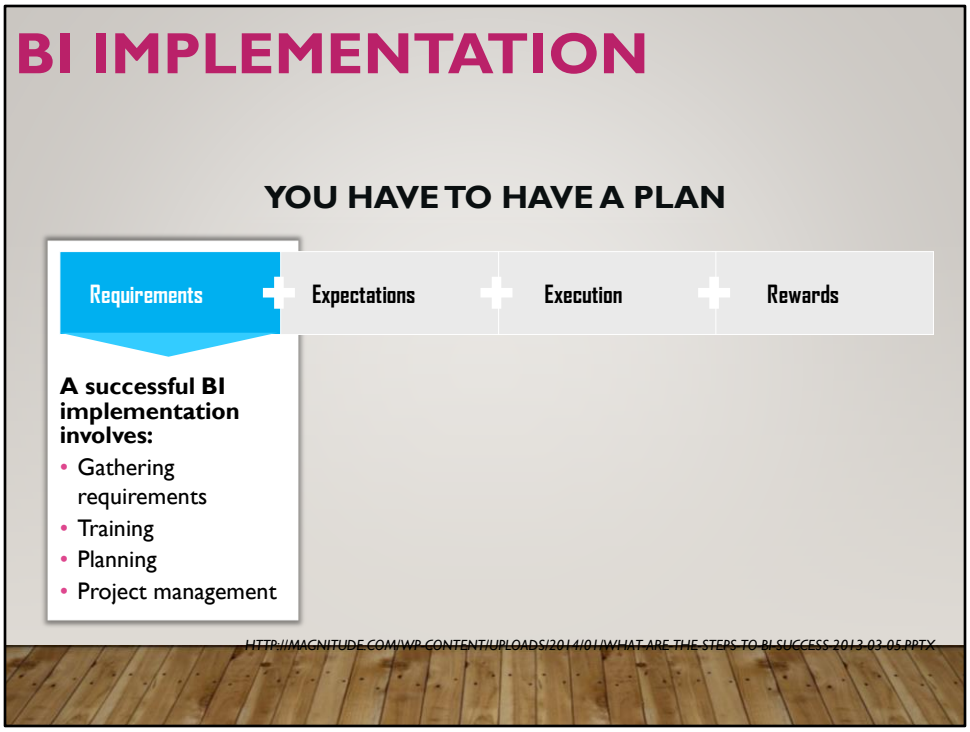



15

*Dashboard and Scorecard Example (Laberge, 2011)*

User Presentation:

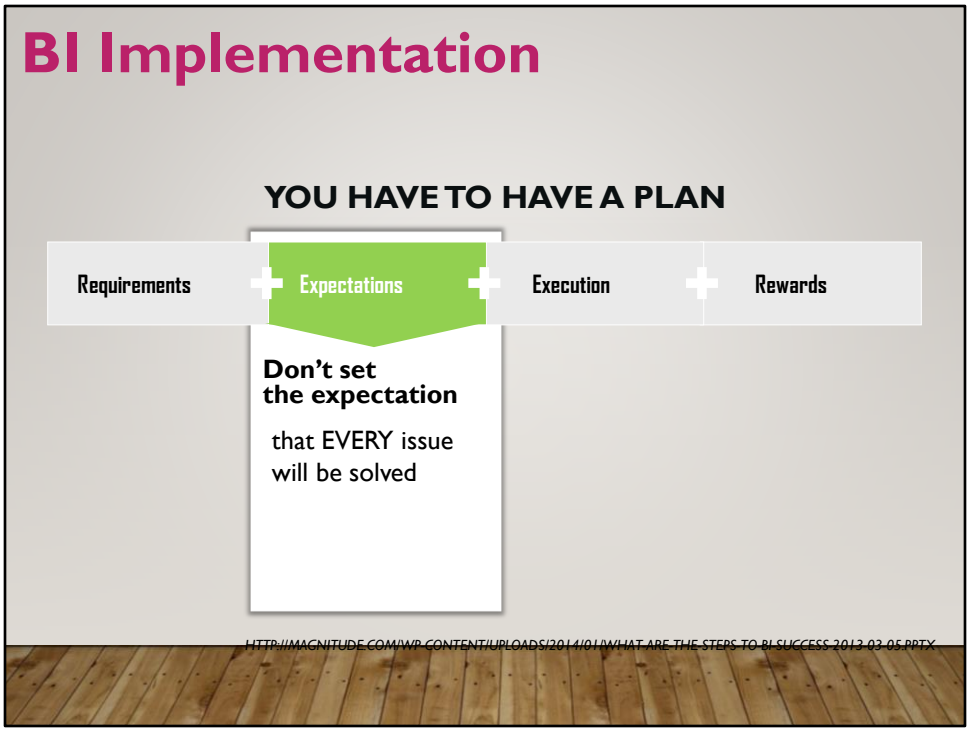
- Dashboards – Special type of report focused on visualization.
- Scorecards - Special type of report focused on visualization.
  - Both of these usually contain highly aggregated data.
  - Scorecard is similar to a school report card – based on specific key business indicator.



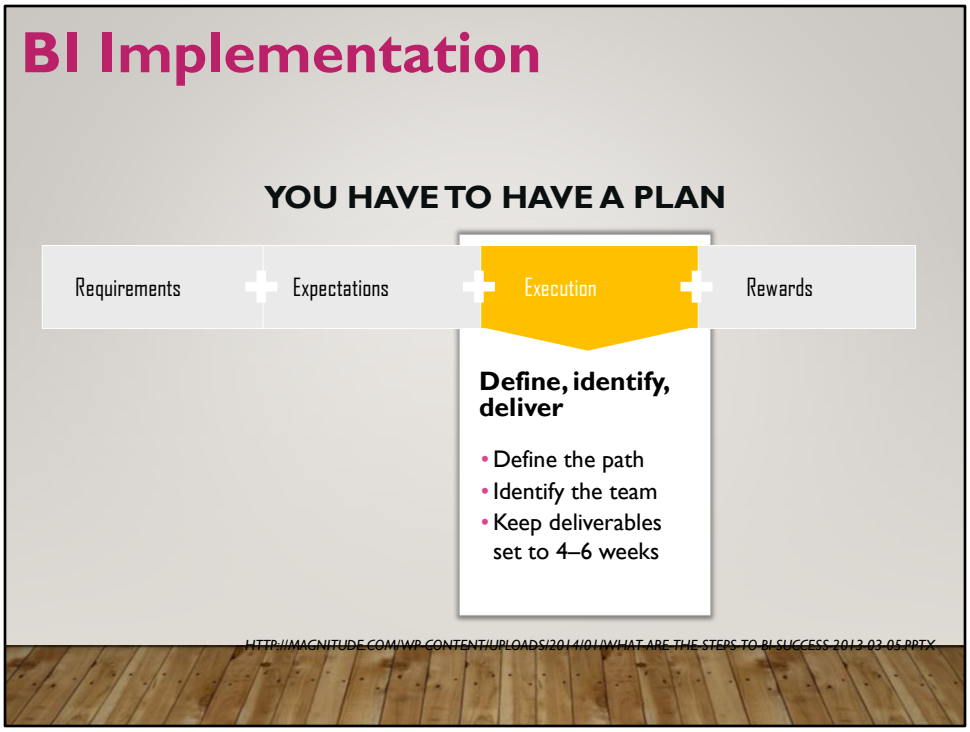

Just like any other project

You MUST have the requirements complete AS possible

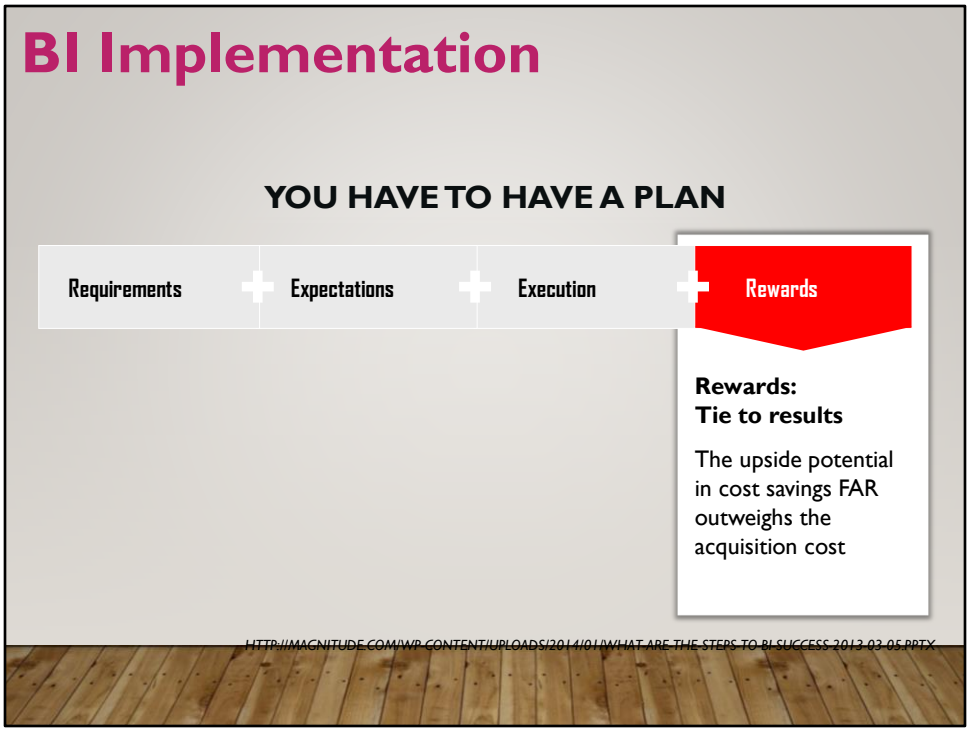
Don’t shoot for perfection


BI Implementation
YOU HAVE TO HAVE A PLAN
Requirements
Expectations
Execution
Rewards
Don't set the expectation
that EVERY issue will be solved
HTTP://MAGNITUDE.COM/WP-CONTENT/UPLOADS/2014/01/WHAT-ARE-THE-STEPS-TO-BI-SUCCESS-2013-03-05.PPTX

SET them




Use seasoned BI professionals and keep the SCOPE DOABLE

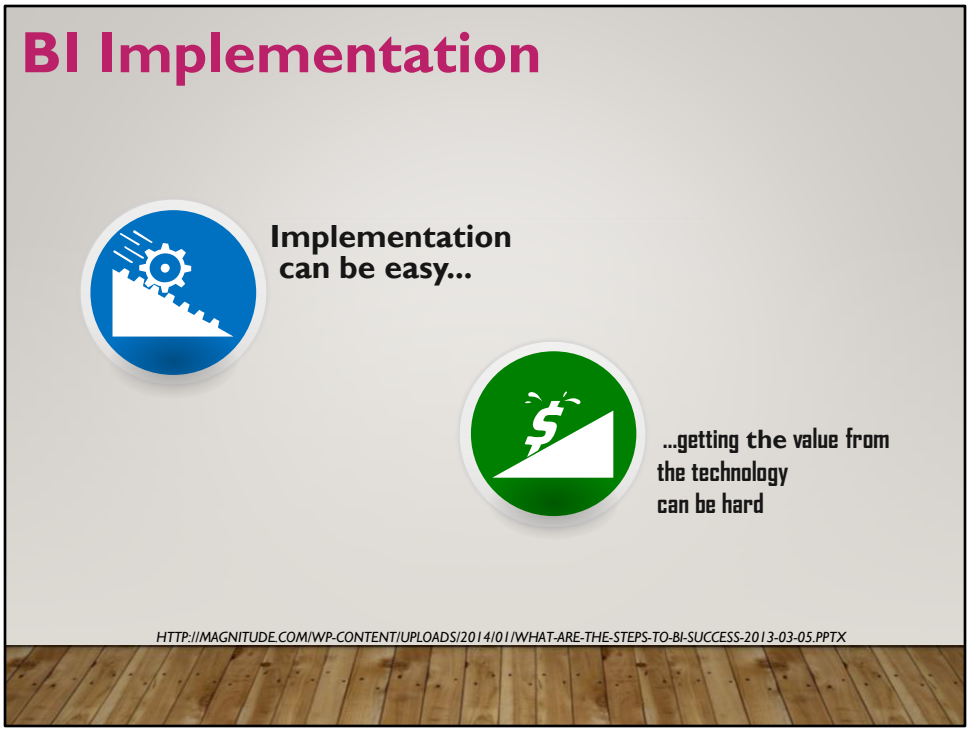



Focus on the rewards

If you simply build it … They will NOT come

You must do some screaming from the rooftops on what you did.

YOU MUST SELL IT – UP – DOWN – and ACROSS




Focus on the rewards

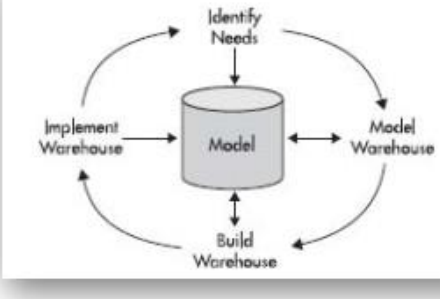
If you simply build it … They will NOT come

You must do some screaming from the rooftops on what you did.

YOU MUST SELL IT – UP – DOWN – and ACROSS

## DATA WAREHOUSE OVERVIEW

- Definition: a system for the collection, organization, holding and sharing of historical data. Data in the data warehouse comes from other systems that capture data based on their purpose.




21

*Data warehouse lifecycle (Laberge, 2011)*

- Term often used to refer to the data warehouse system and data warehouse repository. Our text uses this term to relate to the entire system.
- Used by business users for decision support.
- User query the system to extract data to aid in their business decision process.
- The classic data warehouse lifecycle is primarily to identify business needs or requirements and well as the high-level technical details.

# DATA WAREHOUSE OVERVIEW

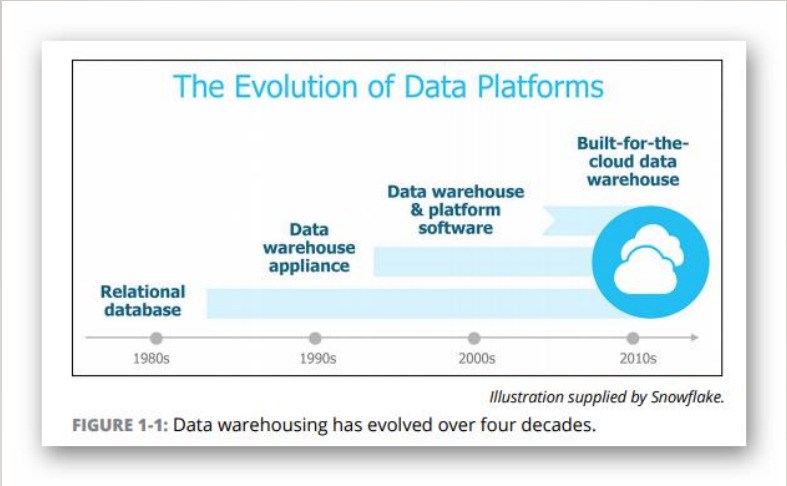



*Illustration supplied by Snowflake.*

**FIGURE 1-1:** Data warehousing has evolved over four decades.

http://info.snowflake.net/rs/252-RFO-227/images/SnowflakeDataeBook.pdf

# DATA WAREHOUSE OVERVIEW

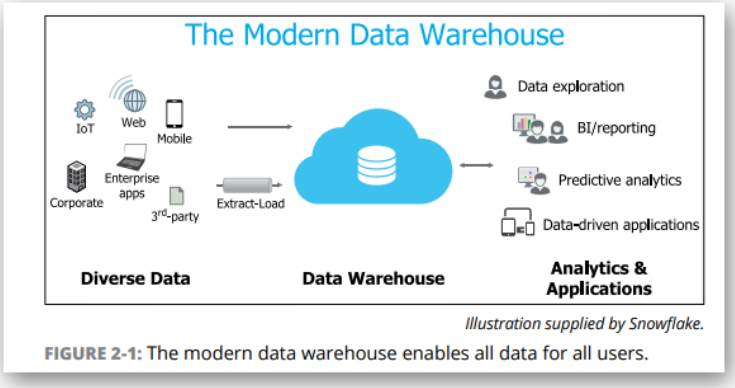



*Illustration supplied by Snowflake.*

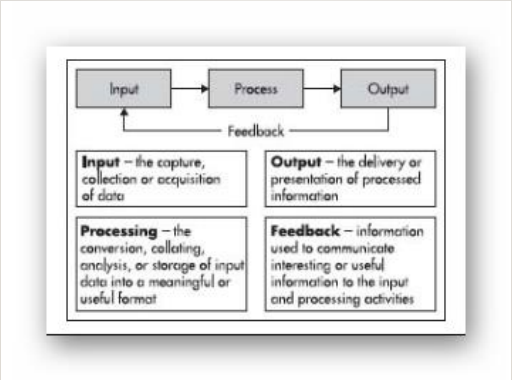
**FIGURE 2-1:** The modern data warehouse enables all data for all users.

http://info.snowflake.net/rs/252-RFO-227/images/SnowflakeDataeBook.pdf

23

## DATA WAREHOUSE OVERVIEW

- Data Warehouse System:
  - Main components of a data warehouse system are…




*Basic System Components (Laberge, 2011)*

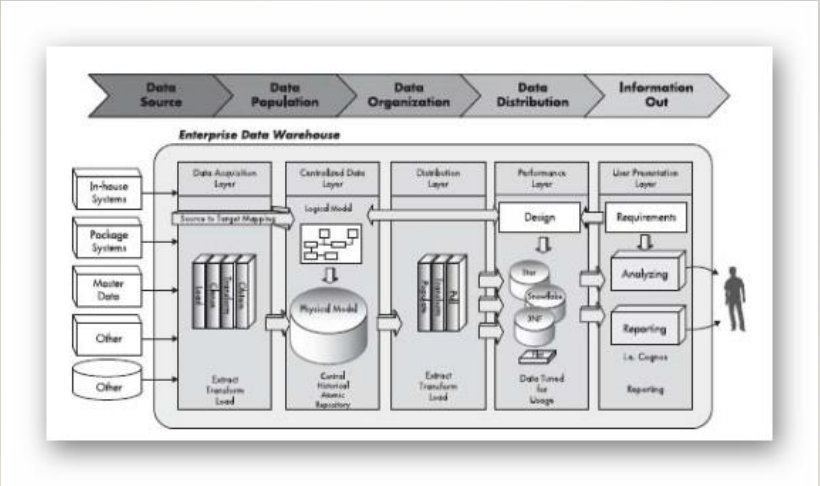
24

Main Components:

- Input – identifying and capturing the data. Data quality is critical at this stage.
- Process – Transforms and hold the data in a structured manner.
- Output – Involves transferring the data to the users that need it. (Reporting)
- Feedback – Based on the input and output functions.

## DATA WAREHOUSE OVERVIEW

- Architecture:

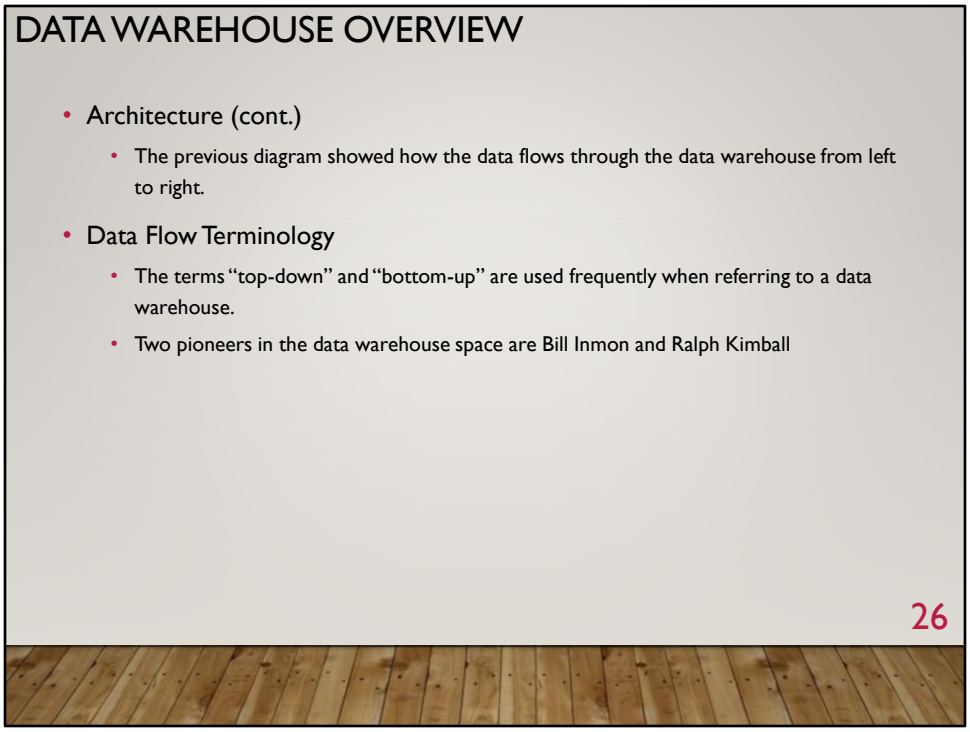



*Data Flow Diagram (Laberge, 2011)*

25

Architecture:

- This is the actual design of the data warehouse system.
- Data flow diagrams typically used to show flow of data.



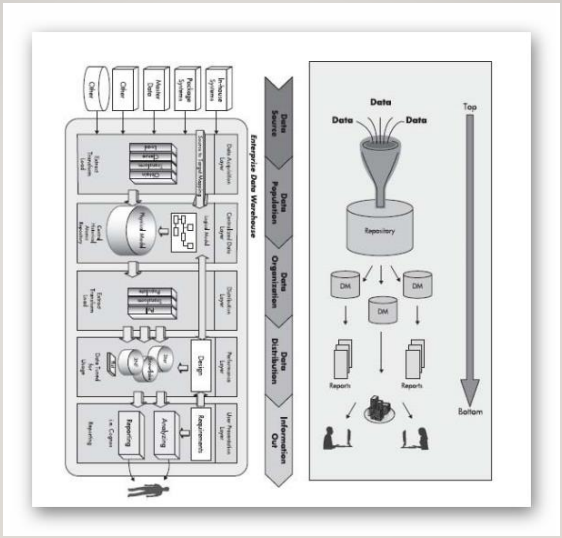

Architecture:

- These diagrams are very useful in helping users gain an understanding of the different components of the data warehouse.
- Communicates a realistic view of the overall data warehouse solution.

Data Flow Terminology Pioneers:

- Bill Inmon's methodology primarily refers to the "top-down" structure, meaning from the data point of view
- Ralph Kimball's methodology primarily refers to the "bottom-up" structure, meaning business purpose above all else with the data to support it.

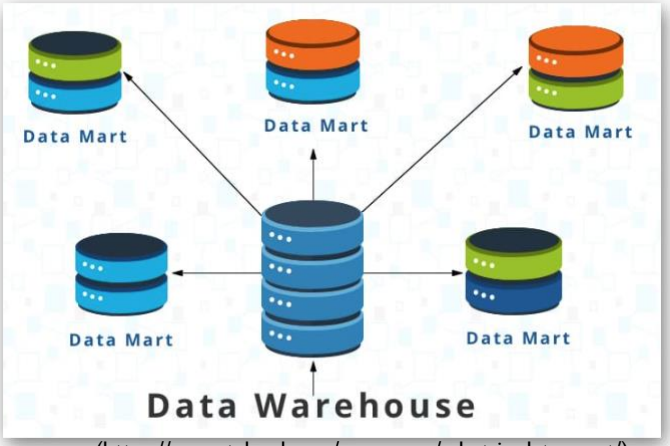
# DATA WAREHOUSE OVERVIEW




*Data Flow (Laberge, 2011)*

## DATA WAREHOUSE OVERVIEW

- Purpose?




Source: (https://www.talend.com/resources/what-is-data-mart/)

28

Purpose:

- Main purpose is to hold historical data.
- This data is integrated from many different systems.
- These operations systems are built to support specific functions such as:
  - Point of sale processing
  - Inventory control
  - Billing systems
- These separate systems are not always built to perform data analytics or data mining from, which is where the data warehouse comes in.
- The data quality could be an issue in one of the separate systems, but when merging into one system this issue is of paramount importance.
- When developing a data warehouse much consideration must be given to the structuring and organization of data.
- One advantage around the data warehouse is reporting. It allows users to aggregate data from multiple systems, something they may not have had the ability to do before.

# DATA WAREHOUSE OVERVIEW

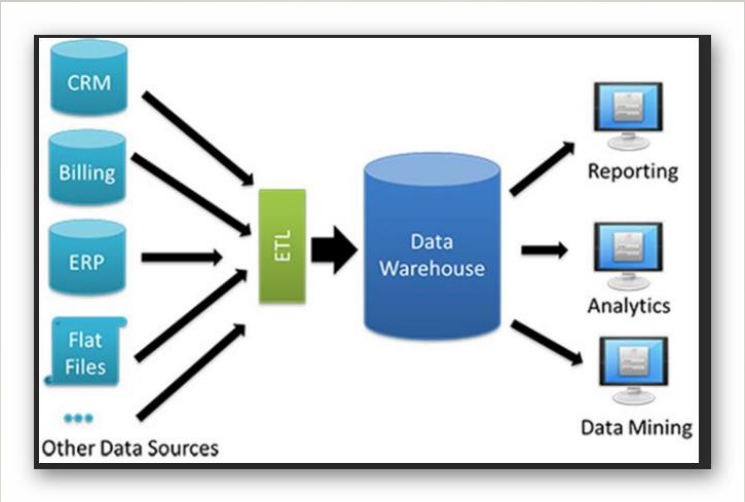



*Image Source: http://www.designandexecute.com/designs/basics-of-data-warehouse-dw/*

29

# DATA WAREHOUSE OVERVIEW

- Data Warehouse Value
- Best practices


30

Source: https://www.computerweekly.com/photostory/2240150469/Enterprise-data-warehouse-deployment-A-step-by-step-tutorial/4/Data-warehouse-implementation-best-practices)

Value:

- The main value of the data warehouse is that is creates a centralized common area for all business users to access the same underlying data.
- The context of the data can be in any manner based on the individual users' requirements
- The underlying data would be common to all users, and becomes an asset to the organization overall not just separate departments.
- Business goals are typically the same across users, and can include:
    - Deeper insight into the company's product base
    - Deeper understanding of the current business processes
    - In-dept knowledge of the company's current operations
    - Improved marketing strategies

Best Practices:

- Limit initial scope – focus on fundamental data.
- Not to start from scratch – purchase as need be.
- Ensure you have a seasoned data warehouse project manager
- Ensure you have a seasoned data warehouse architect
- With a seasoned set of staff members project should take as little as 6 months for initial stage.


Best Practices:

### Step 1 – Research

- Look into what a data warehouse and business intelligence are and how they are used.
- Become familiar with the topic and key points of each.

### Step 2 – Strategic Alignment

- Determine if a data warehouse can be useful to your organization.

### Step 3 – Focus, or Limited Scope

- Do not plan on doing everything at once.
- Focus on one area that is important to the organization

### Step 4 – Value

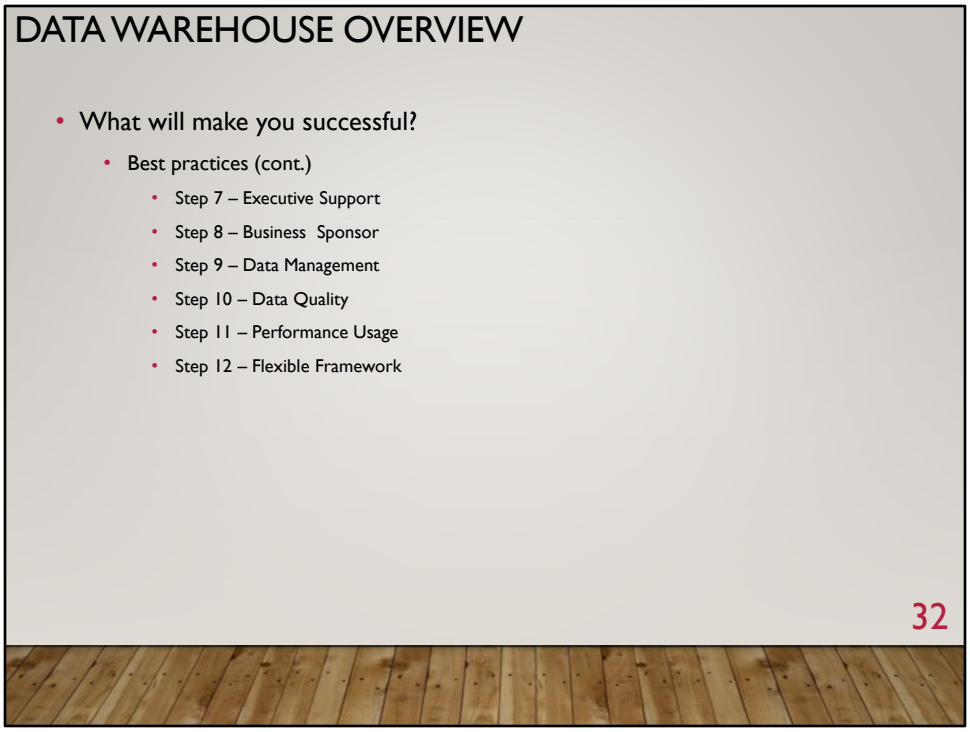
- Must have value to the organization.
- Show how data quality will be improved and how it was lacking before.

### Step 5 – Metrics

- Must be quantifiable, tangle, accountable and numeric in some form.

### Step 6 – Goals

- All in the organization must be able to see success.
- Must be coordination of goals and purpose between IT and the business.
- Keep users in the loop during the entire product development phase.




Best Practices (cont.)

Step 7 – Executive Support

- This data warehouse is an organizational strategic asset and must be supported by executive leadership.

Step 8 – Business Sponsor

- Since this data warehouse is being specifically for the business to help aid in the decision-making process it must be supported by them.

Step 9 – Data Management

- Structuring the data is paramount to a successful data warehouse systems.
- Ensure the data is organized at an enterprise level.
- Purchasing a prebuilt model can greatly help with this effort.

Step 10 – Data Quality

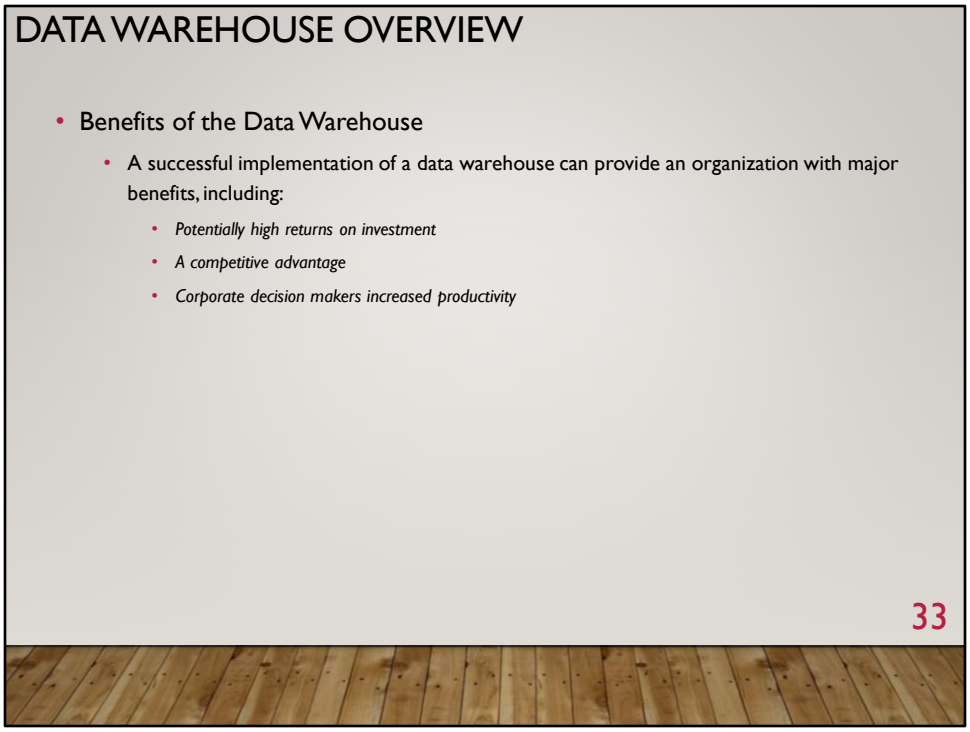
- BI is nothing if the underlying data has little or no integrity.

Step 11 – Performance Usage

- If the system lags once turned on, then you are in jeopardy of the business abandoning the data warehouse.

Step 12 – Flexible Framework

- Ensure whatever system you built is flexible to move onto the next phase.
- Must have the ability to expand at a future date.




Benefits:

*Potentially high returns on investment* – a study by the International Data Corporation (IDC) reports that data warehouse projects on average over a 3-year period a 401% ROI.

*A competitive advantage* – this is realized due to the fact that the data warehouse provides the companies decision makers access to previously unavailable, unknown and untapped information for their business.

*Corporate decision makers increased productivity* – provides decision makers access to an integrated database with access to consistent, subject-oriented, historical data.

## DATA WAREHOUSE OVERVIEW

- Problems of the Data Warehouse

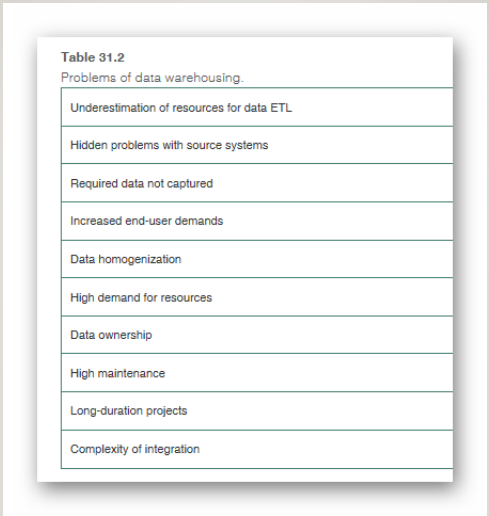


**Table 31.2**
Problems of data warehousing.

| |
|---|
| Underestimation of resources for data ETL |
| Hidden problems with source systems |
| Required data not captured |
| Increased end-user demands |
| Data homogenization |
| High demand for resources |
| Data ownership |
| High maintenance |
| Long-duration projects |
| Complexity of integration |

*(Connolly & Begg, 2015)*

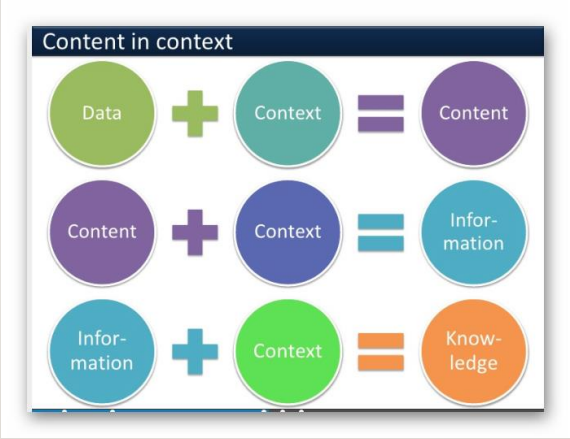
34

Although there are benefits, as with any system within an organization there are some pitfalls or problems that must be identified.

- Underestimation of resources
- Hidden problems
- Required data
- Increased end-user demands
- Data homogenization
- High demand for resources
- Data ownership
- High maintenance
- Long-duration projects
- Complexity of the integration

## DATA IN THE ORGANIZATION

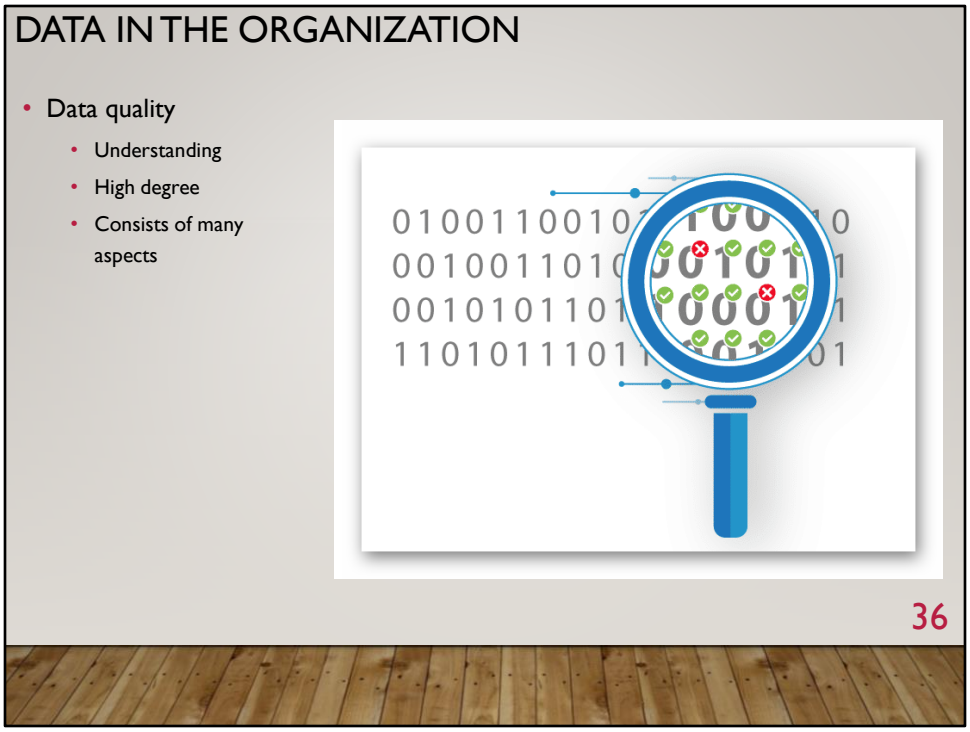
- Data in context
  - Is an asset
  - Data and context are critical
  - Organized, structured and understood.




35

Source: (https://www.slideshare.net/rahelab/big-content-content-strategy-as-a-design-framework/25-Content_in_context_Data_Context)

- Data in the organization is an asset and must be managed with due diligence, within a timely manner.
- The data asset issue is a much larger scope then just the data warehouse, but its management usually starts with a data warehouse project.
- Data and its context are critical to the data warehouse and business intelligence initiatives within the organization.
- It must be organized, structured and understood in the correct business context.




- Understanding that the data provided has a high degree of confidence.
- Should have a high degree of data quality.
- It consists of many aspects including:
  - Determining appropriate business terminology
  - Determining usage
  - Organization of the data components into manageable structures
  - Ensures the proper domain values (data type) are identified
  - Ongoing governance of the data
  - Profiling the data
  - Security for both current and historical versions.
- Managing this data is key to ensure proper data quality.
- Set high standards on data quality since your organization depends on it to be competitive in the marketplace.

## DATA IN THE ORGANIZATION

- Data vocabulary


Source: (https://www.td.org/insights/big-learning-data-vocabulary-101)

***Example:***

This example shows several fundamental terms: customer, demographics, product, store, and possibly location. These are business terms. When talking with the business users, always communicate using business terms, but clarify these using data terminology. For example, in the discovery phase, the business analyst or data modeler could discuss and break down the business term "customer" into data components as follows:

- A customer is any individual. These are two distinct concepts, as not all individuals will be customers; however, customers will form a subset of individuals.
- The individual has descriptive qualities such as: gender, age range, and so on.
- The individual may be associated to a location, which is the store location or a residential address. (Laberge, 2011)

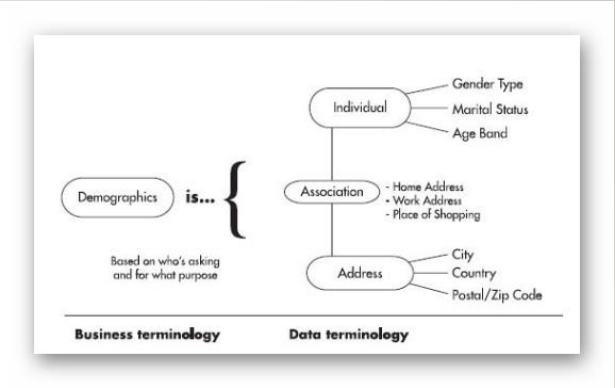
37

Data Vocabulary:

- Heart and soul of business intelligence and data warehouse.
- Must first understand what the data is.
- Cannot obtain data quality if individuals in your organization define terms differently.

## DATA IN THE ORGANIZATION

- Data components
  - Business terminology
  - IT
  - Data analyst
  - Enterprise wide vocabulary



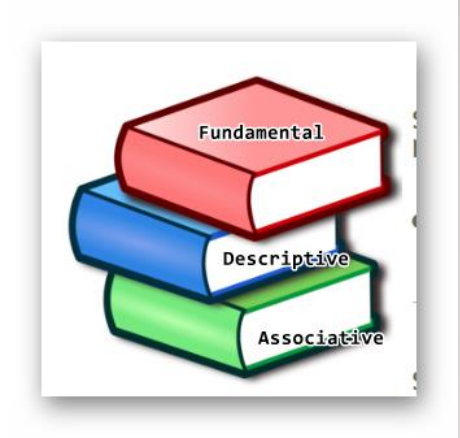

*Example of Vocabulary (Laberge, 2011)*

38

- Business users use business terminology.
- IT use data terminology.
- Useful to think in these two distinctions when building a data warehouse.
- A data analyst is then used to help translate the business terms into data terms.
- Creating an enterprise wide vocabulary with the ability to decompose the business terminology into their respective data components ensure better project success.

## DATA IN THE ORGANIZATION

- Data components
  - Three basic concepts:
    - Fundamental
    - Descriptive
    - Associative

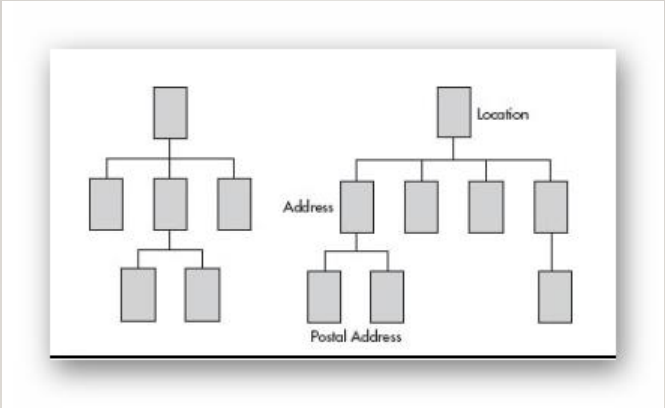



Source (https://www.sqlchick.com/entries/2017/1/9/defining-the-components-of-a-modern-data-warehouse-a-glossary) 39

- Easy method is to think of data in three basic concepts:
  - Fundamental – basically the object of the discussion or main point.
  - Descriptive – describes the fundamental data component.
  - Associative – how it is related to another fundamental data component.

## DATA IN THE ORGANIZATION

- Data components



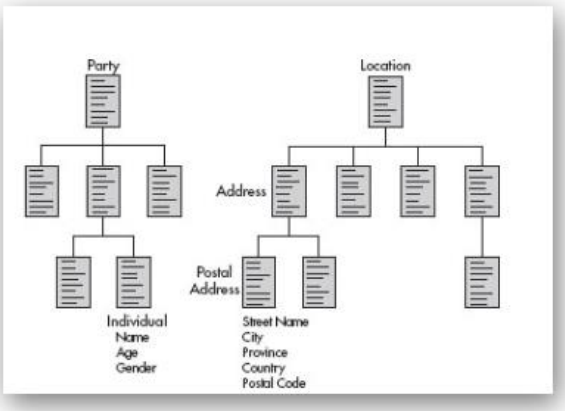

*Fundamental Data Components (Laberge, 2011)*

40

Shows the first level: Fundamental

## DATA IN THE ORGANIZATION

- Data components



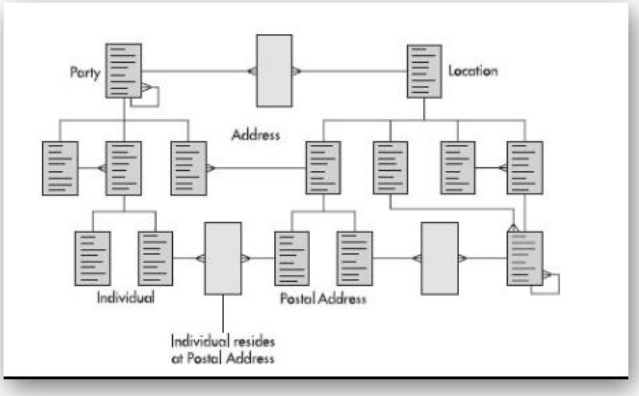

*Descriptive Data Components (Laberge, 2011)*

41

Shows the second level: Descriptive

## DATA IN THE ORGANIZATION

- Data components




*Associative Data Components (Laberge, 2011)*

42

Shows the third level: Associative

## DATA IN THE ORGANIZATION

- Organizing the data
  - Defined and structured.
  - Spans the entire organization.
  - Plan and prioritize.
- Data Models
  - Communicates data usage.
  - Different data models (Pages 53-56).

Source (https://www.iconfinder.com/icons/3251472/archiving_backing_up_data_backup_computer_data_folder_organizing_icon)

43

Organizing the data:

- To be organized it first must be defined and structured.
  - Structing data is done through design. This means you have an understanding of what the data components are and how they relate to each other.
  - Without structure the data warehouse would be very unorganized and possible useless.
- Typically the data warehouse will span the entire organization.
- Must plan and prioritize the data warehouse rollout to ensure all components are taken into consideration.

Data Models:

- Communicates the data usage in either conceptual business or data terms.
- The textbook by Laberge, 2011 shows a few different data models (Pages 53-56).

## DATA IN THE ORGANIZATION

- Data Architecture
  - Begin to understand the flow.
  - Not to be confused with the technical architecture.
  - Starts with data.
  - Some sort of business value.
  - There is a large list of areas involved when determining the under lying data.
    - Where…
    - Who owns…
    - Data…
    - Full technical…
    - Data…
    - Data…


44

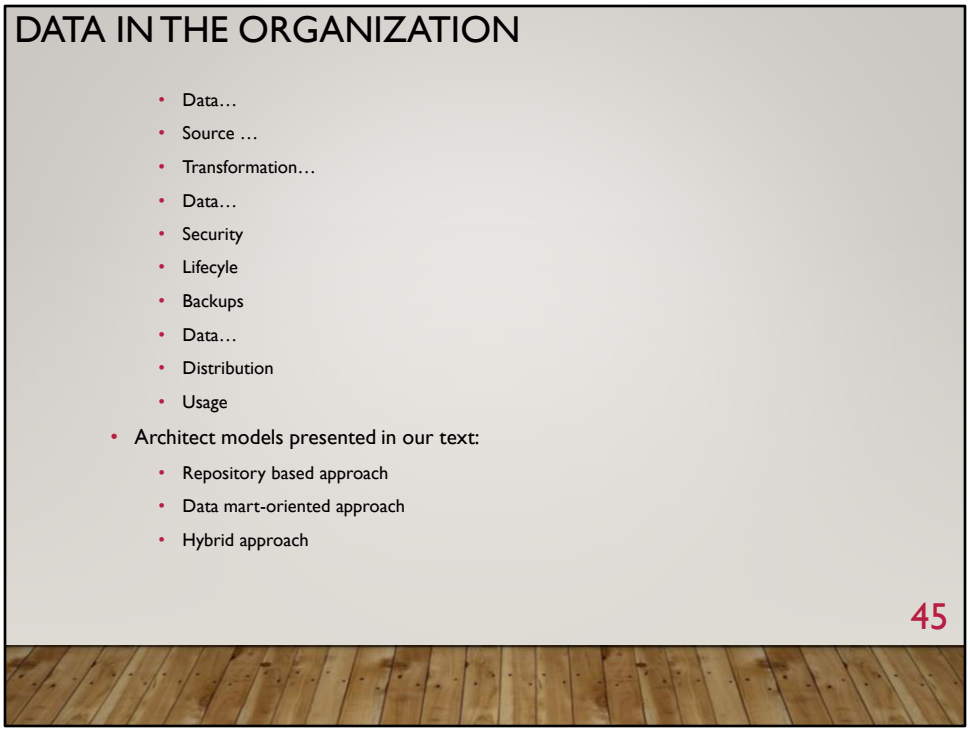
Source (https://www.dataversity.net/what-is-data-architecture/)

Data Architecture

- At this stage we begin to understand the flow of the data through the data warehouse system.
- This is not to be confused with the technical architecture of the data warehouse, which is the servers, database management system, operating systems, middleware, software, business intelligence tools and so fourth.
- Typically starts with data as is acquired by the business.
  - This helps limit the scope of the project's effort.
- Must have some sort of business value for each deliverable during the project.
  - The business expects to receive something in return for any and all expenditures laid out for the project.

Areas Involved:

- Where the data is from
- Who owns the data
- Data format
- Full technical details of the data
- Data quality
- Data availability




Areas Involved (cont.)

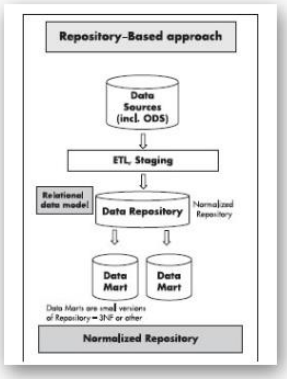
- Data availability
- Source to target mapping
- Transformation rules
- Data volumetric
- Security
- Lifecyle
- Backups
- Data models
- Distribution
- Usage

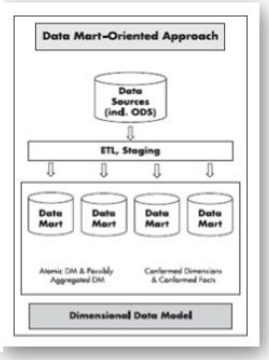
Given the exhaustive list above you can see the importance of having a data architect.

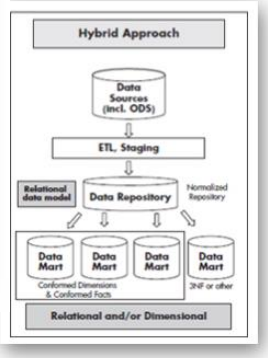
There are a few data architect models presented in our text which include:

- Repository based approach
- Data mart-oriented approach
- Hybrid approach

## DATA IN THE ORGANIZATION




*Data Architecture Models (Laberge, 2011)*

46

This diagram show all three models side-by-side for comparisons.

## DATA IN THE ORGANIZATION

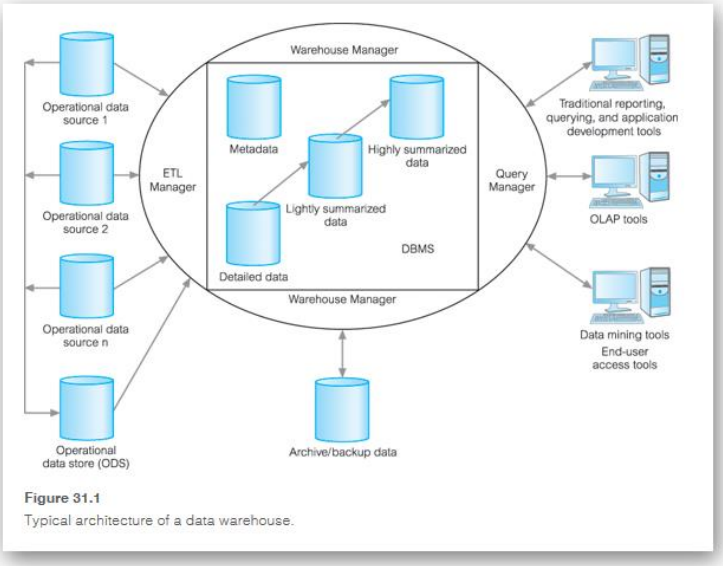



**Figure 31.1**
Typical architecture of a data warehouse.

47

*Typical architecture of a data warehouse (Connolly & Begg, 2015)*

This image represents a typical architecture design of a data warehouse.

The next slide will present information around the different components listed here.

## DATA IN THE ORGANIZATION

- Data Architecture (cont.)
  - Components noted on previous diagram
    - *ETL Manager.*
    - *Warehouse Manager.*
    - *Query Manager.*
    - *Detailed Data.*
    - *Lightly and Highly Summarized Data.*
    - *Archive/Backup Data.*
    - *Metadata.*
    - *End-user Access Tools.*


Source: (https://sungsoo.github.io/2014/06/18/big-data-architecture.html)

48

Components noted on previous diagram:

- *ETL Manager*: performs all the operations associated with the ETL of data into the warehouse.
- *Warehouse Manager*: performs all operations associated with management of data.
- *Query Manager:* performs all operations associated with management of user queries.
- *Detailed Data*: stores all the detailed data in the database schema.
- *Lightly and Highly Summarized Data*: stores all the predefined lightly and highly summarized (aggregated) data generated by the warehouse manager.
- *Archive/Backup Data*: stores the detailed and summarized data for the purpose of archive and backup.
- *Metadata*: this area stores all the metadata (data about data) definitions used by all processes in the data warehouse.
- *End-user Access Tools*: tools used by users to interact with the data warehouse for the purpose of reporting and supporting the organizations decision makers.

## DATA IN THE ORGANIZATION

- Build or Buy
  - Buy a prebuilt data warehouse.
  - Is the cost beneficial?
  - Need to consider…
    - Who?
    - What?
    - Flexibility
    - Real usage
    - Maintenance
    - Consulting services?
    - In-house resources
  - Buying a model…
  - Can also be complicated…
  - Alternatively you can…


Source: (https://www.cooladata.com/cost-of-building-a-data-warehouse/)

49

Build or Buy?

- One option is to buy a prebuilt data warehouse and business intelligence model.
- Must decide if the cost of building your own is beneficial.
- If buying you need to consider the following items:
  - Who built it?
  - What is the underlying data architecture?
  - How should it be used?
  - Is it flexible? Will it be easily adoptable to your specific business needs. What is involved in expanding this model.
  - Real Usage. Is a specific tool required and how often are releases available.
  - Is maintenance required and is it included?
  - Does the vendor offer consulting services?
  - Are in-house resources available for the project and data model.
- Buying a model can be a real time saver if used properly.
- They can also be complicated because they hold lots of information.
- Alternatively you can undertake building this model yourself, but take care to not to miss anything during the design phase.

## REASONS FOR BUILDING

- Large organizations are interested in creating one central vocabulary and essentially one version of the data environment.
- Regardless of size all organizations are interested in their ROI.
- Some popular reasons for development of the data warehouse are…

50

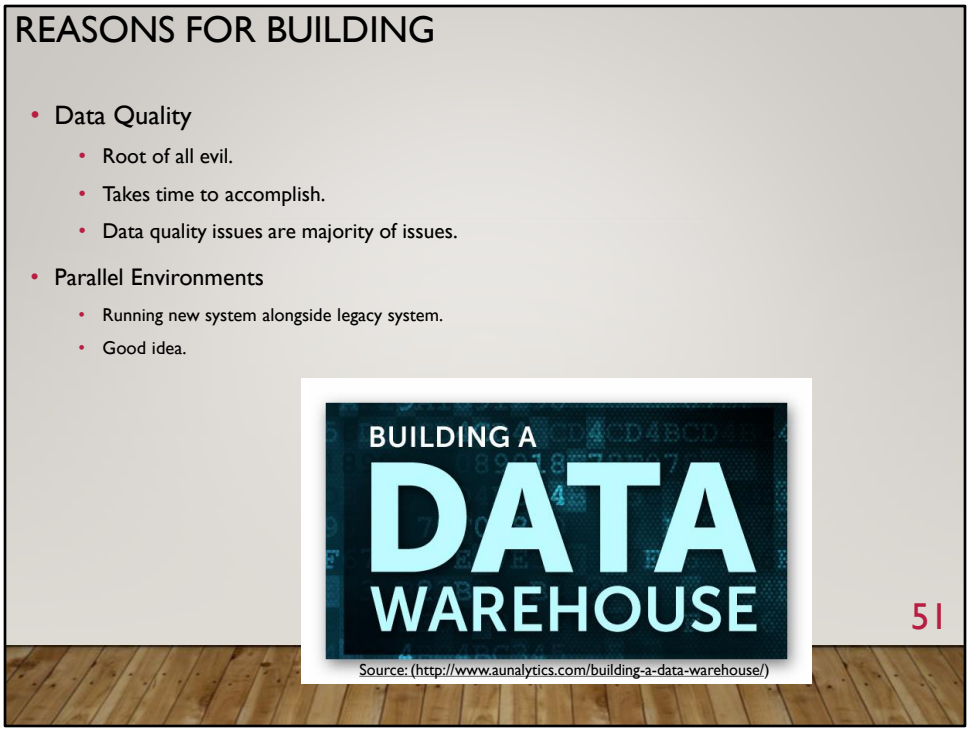
Source: (https://svitla.com/blog/data-warehouse-vs-database)

ROI:

- Depends on the strategy for building the data warehouse.
- Is it purely a business intelligence effort? Such as the need to reduce customer turn-over.
- A qualified target should be set. Examples from our text include:
  - Decrease mainframe maintenance costs by one million euros per year.
  - Decrease software costs by $250K per year for five years.
  - Increase customer base by 5 percent within one year.

Popular reasons for development:

- Migration from one platform to another.
- The centralization of diverse data warehouses.
- The consolidation of diverse data marts.
- New Initiatives.
- IT just-build-it scenario




Data Quality:

- At times data quality can be the root of all evil in the data warehouse.
- This effort takes time to accomplish, since profiling must be done on the data, then an owner found and then management must decide on a corrective action plan.
- In many cases migrations from legacy systems to another platform data quality issues are a majority of the issues and requires a large effort to remediate.

Parallel Environments:

- Entails running the new system along side the legacy system from 3 months up to 6 months, depending on the reporting regularity (monthly, quarterly, yearly).
- This is a good idea and should be followed by all organization. Allows time to ensure the new system is working properly and reporting information correctly as the old system did.

## REASONS FOR BUILDING

- Platform Migration
  - Has grown from a legacy system.
  - Help reduce costs.
  - Time to purchase.
- Business Continuity
  - Devastating to a business.
  - May lead to changes.
  - New system should match.
- Reverse Engineering
  - Done to understand...
  - Done when the system lacks documentation.
  - End results is to have the old legacy system migrate.
  - Can be very costly.


https://hexaware.com/blogs/a-proactive-approach-to-building-an-effective-data-warehouse/

52

Platform Migration:

- Usually occurs when the data warehouse has grown from a legacy system, most likely some sort of mainframe operating system.
- Usually done to help reduce costs due to the decommission of the mainframe system.
- Good time to purchase prebuilt data warehouse model, disks, and database.

## Business Continuity

- Since any deviation from the current level of data integrity and system reliability can be devastating to a business there must remain continuity with the business during the development.
- Change many times results in many technical adjustments, which may lead to changes in data reporting and those must be account for and addressed.
- End result is that any numbers reported in the new system should match those reported in the old system.

## Reverse Engineering

- Done to understand how the current data warehouse is structured, and how it was built.
- At times this must be done when the system lacks documentation so you can understand what it is doing.
- End results is to have the old legacy system migrate to the new system with all processes in place and documented fully.
- Can be very costly and at certain times simply impossible.


## Poor Data Quality

- If the historical system has bad data then building a new data warehouse will not fix this issue.

## Lack of Business Interest

- If the business is not interested in this project, then acceptance by them will be difficult to obtain, ultimately costing the company dearly.

## Lack of Sponsorship

- If the company's management does not support this project, as with the lack of business interest, then taking on this project is not recommended.

## Unclear Focus

- If the company can't come up with a tangible ROI, or they are not clear on their goals then the project is doomed from the start.

## Sufficiency of Current Systems

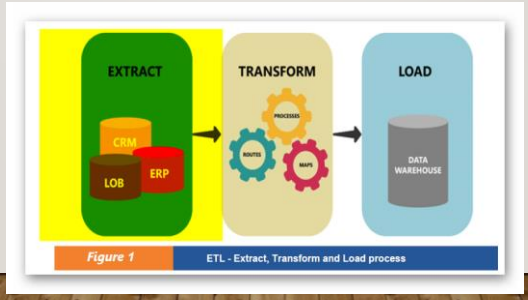
- If the current system gives the business what it needs, and there really is not need to upgrade then why undertake this costly process.

## Lack of Resources

- If you don't have the proper skilled resources in your organization then taking on this project might not be in your best interest.

## DATA STAGING AND ETL

- **Extraction**
  - Relevant data is extracted.
  - Normally internal.
  - Initially you use static extraction.
  - Switch to incremental extraction.
  - Normally copied to temporary storage.




54

Image Source: (https://datawarehouseinfo.com/etl-vs-elt-transform-first-or-transform-later/)

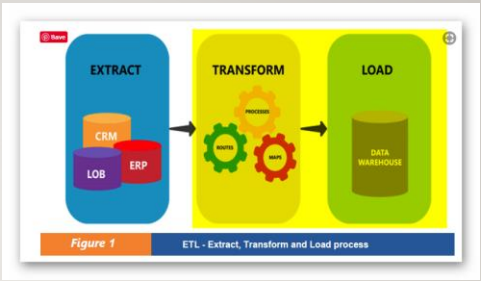
The data staging level houses the ETL processes.

ETL process takes place once the data warehouse has been populated for the first time.

**Extraction:**

- In this stage the relevant data is extracted from the available sources.
- These sources are normally internal but can also be external sources such as suppliers and customers.
- Initially you can use static extraction, which essentially looks like a snap-shot of the data.
- Once populated you switch to incremental extraction, which is the process for updating the data in the warehouse. This process is based on the database log maintained by the source database.
- Normally the data is copied to temporary storage often referred to as the operational data store (ODS) or staging area (SA).

## DATA STAGING AND ETL




- **Transformation**
  - Core of reconciliation phase.
  - Applies a series of rules.
  - Converts the data.
  - Data is taken from a normalize state.
- **Loading**
  - Last step.
  - Can only occur after transformation.
  - At loading additional constraints applied.

55

Image Source: (https://datawarehouseinfo.com/etl-vs-elt-transform-first-or-transform-later/)

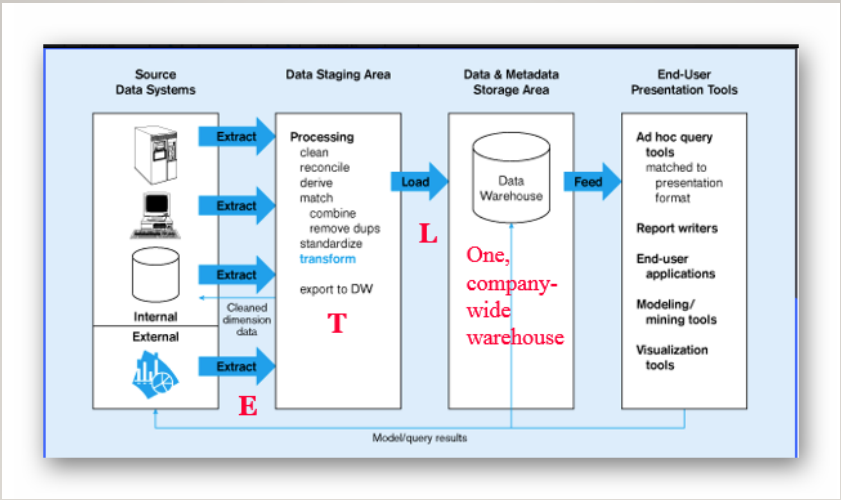
## Transformation

- This is the core of the data reconciliation phase.
- This process applies a series of rules or functions to the extracted data, which in turn determines how the data will be used for analysis. This can often include transformations such as data calculations and the creation of surrogate keys.
- Converts the data from its operational source format into the specific data warehouse format.
- During this phase the data is taken from a normalize state to a denormalized state since most data in the data warehouse is typically denormalized. It is commonly recommended that the data be held at the lowest level of granularity as possible.

## Loading

- Last step and can be accomplished by either refreshing data or updating data.
- This step can only occur after all the transformation processes have completed.
- At loading additional constraints which are defined in the database schema can then be applied to the data, as well as any documented triggers be enacted.

## DATA STAGING AND ETL



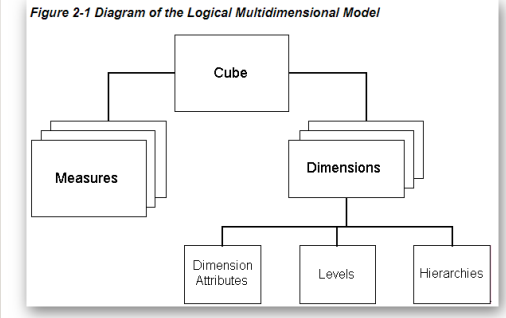

56

Image Source: (https://etltestingbyanupray.blogspot.com/)

Show ETL process.

## MULTIDIMENSIONAL MODEL

- Fundamental.
- Paradigm of data warehouse.
- Productivity tools.
- Begins with observation.
- Typically facts.
- Typically mappings.




57

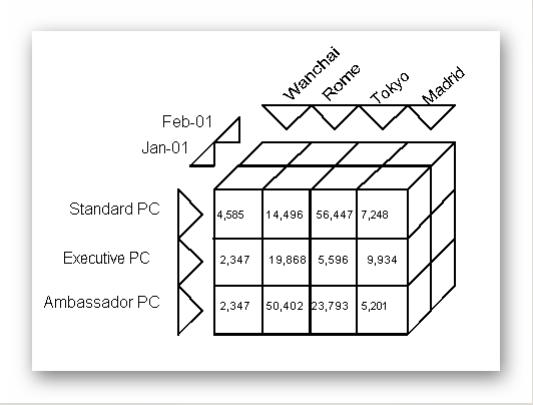
Image Source: (https://docs.oracle.com/cd/B13789_01/olap.101/b10333/multimodel.htm)

- Fundamental to many decision-making support systems.
- Used as a paradigm of data warehouse data representation.
- Linked to the widespread use of productivity tools, like spreadsheets which adopt this multidimensional model effectively as a visualization paradigm.
- Begins with observation of the facts affecting the company's decision-making process.
  - Each fact is described by values of a relevant measure that provide quantitative descriptions of events.
- Data in this model is typically facts (numeric measurements) such as property sales revenue data and the association of this data with dimensions such as location (of the property) and time (of the property sale)
- These dimensions are typically mappings from a set of lower-level concepts to high-level concepts.

Diagram of the logical multidimensional model. Cubes consist of measures and dimensions. Dimensions consist of levels, hierarchies, and dimension attributes.

## MULTIDIMENSIONAL MODEL

- Concept of dimensions.
- Cube hinges on a ***fact***.



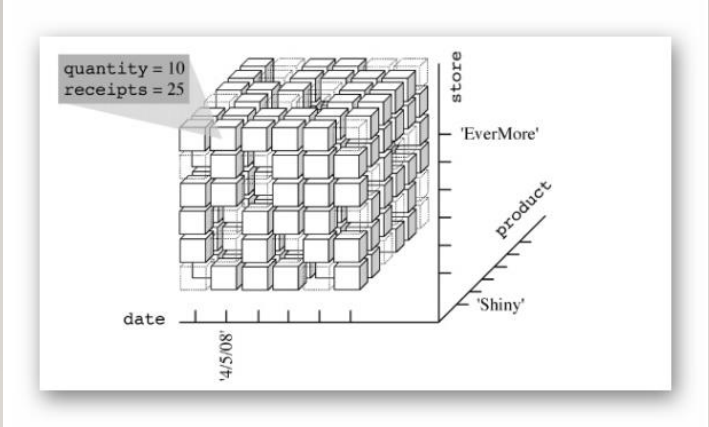

58

Image Source: (https://docs.oracle.com/cd/B13789_01/olap.101/b10333/multimodel.htm)

- This concept of dimensions gave birth to widely used metaphor of ***cubes*** to represent this multidimensional data.
  - Each cube cell is given a value to represent each measure.
- The multidimensional cube hinges on a ***fact*** relevant to the decision-making process.
  - Shows a set of ***events*** with numeric ***measures*** that provide a quantitative description.
  - The terms fact and cube are often used interchangeably.
  - All agree that the term dimensions to specify the coordinates.

Image: A sales cube with products down one edge, time periods across another edge, and geographic areas along a third edge.

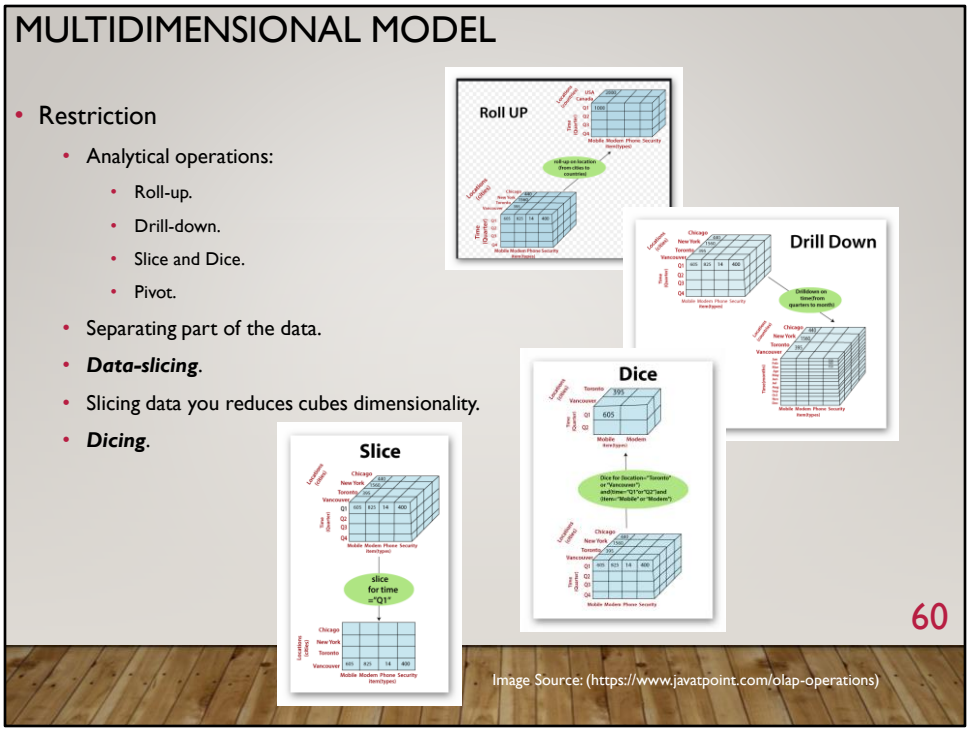
# MULTIDIMENSIONAL MODEL




*The three-dimensional cube modeling sales in a store chain: 10 packs of Shiny were sold on 4/5/2008 in the EverMore store, totaling $25.*

*(Golfarelli & Rizzi, 2009)*

59




Analytical Operations:

- Roll-up – performs aggregations on the data by moving up the dimension.
- Drill-down – reverse of the roll-up and moves down the dimension.
- Slice and Dice – ability to look at the data form different view points.
- Pivot – ability to rotate the data to provide an alternative view.

- Means separating part of the data from the cube to help mark out a field for analysis.
- Simplest type is called ***data-slicing***.
- When slicing data you are reducing the cubes dimensionality.
- Generalization of slicing is called ***dicing***. It poses some constraints on the cubes dimensional attributes to help scale down/reduce the size of the cube.

# MULTIDIMENSIONAL MODEL

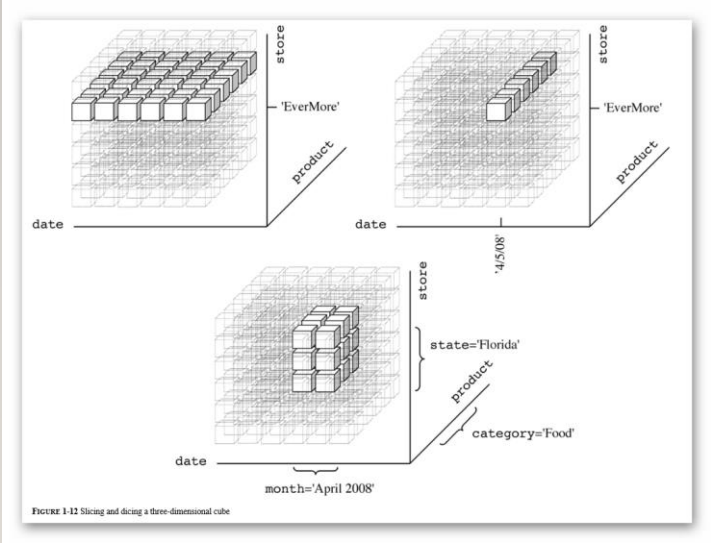



FIGURE 1-12 Slicing and dicing a three-dimensional cube

*(Golfarelli & Rizzi, 2009)*

61

## MULTIDIMENSIONAL MODEL


- Metadata
  - Data to define data.
  - Purpose to show the pathway.
  - Plays an essential role.
  - Most interested in *internal meta-data*.
  - Concerned with *external meta-data*.
  - Accessed by all architecture components.

62

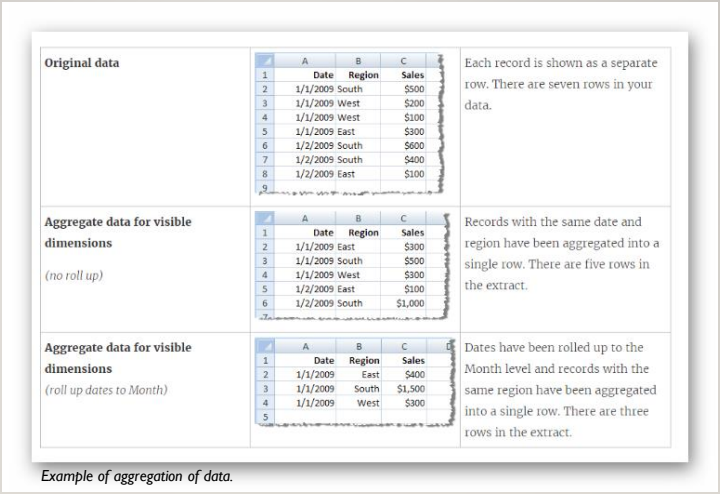
Image Source: (https://www.nodegraph.se/what-is-metadata-and-why-does-it-matter/)

Metadata

- This is a term applied to the data used to define other data.
- Major purpose of metadata is to show the pathway back to where the data began, so the history of any item in the data warehouse is known to its administrators.
- Plays an essential role because it specifies the source, values, usage and features of the data warehouse data and also defines how data can be changed and processed.
- System administrators are most interested in *internal meta-data* because it defines the data sources, any transformation processes, population policies and physical schema.
- End users are more concerned with *external meta-data* since it defines definitions, quality standards , units of measure and any relevant aggregations of data.
- This data is stored in a meta-data repository which can be accessed by all the other architecture components.

## MULTIDIMENSIONAL MODEL

- Aggregation



| | | |
|---|---|---|
| **Original data** | | Each record is shown as a separate row. There are seven rows in your data. |
| **Aggregate data for visible dimensions** *(no roll up)* | | Records with the same date and region have been aggregated into a single row. There are five rows in the extract. |
| **Aggregate data for visible dimensions** *(roll up dates to Month)* | | Dates have been rolled up to the Month level and records with the same region have been aggregated into a single row. There are three rows in the extract. |

| | A | B | C |
|---|---|---|---|
| 1 | Date | Region | Sales |
| 2 | 1/1/2009 | South | $500 |
| 3 | 1/1/2009 | West | $200 |
| 4 | 1/1/2009 | West | $100 |
| 5 | 1/1/2009 | East | $300 |
| 6 | 1/2/2009 | South | $600 |
| 7 | 1/2/2009 | South | $400 |
| 8 | 1/2/2009 | East | $100 |
| 9 | | | |

| | A | B | C |
|---|---|---|---|
| 1 | Date | Region | Sales |
| 2 | 1/1/2009 | East | $300 |
| 3 | 1/1/2009 | South | $500 |
| 4 | 1/1/2009 | West | $300 |
| 5 | 1/2/2009 | East | $100 |
| 6 | 1/2/2009 | South | $1,000 |
| 7 | | | |

| | A | B | C |
|---|---|---|---|
| 1 | Date | Region | Sales |
| 2 | 1/1/2009 | East | $400 |
| 3 | 1/1/2009 | South | $1,500 |
| 4 | 1/1/2009 | West | $300 |
| 5 | | | |

*Example of aggregation of data.*

63

Image Source: (https://td.unh.edu/TDClient/KB/ArticleDet?ID=1629)

Aggregation:

- This plays a very fundamental role in the multidimensional database.
- Every aggerate event will essentially sum up the data available in that particular events aggregates

# MULTIDIMENSIONAL MODEL

You can aggregate along various dimensions at the same time. For example, Figure 1-14 shows that you can group sales by month, product type, and store city, and by month and product type. Moreover, selections and aggregations can be combined to carry out an analysis process targeted exactly to users' needs.

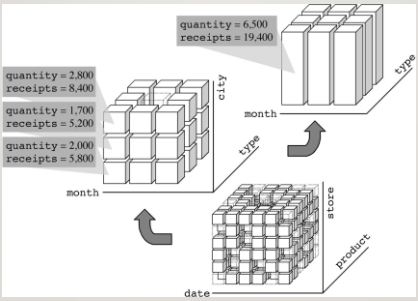
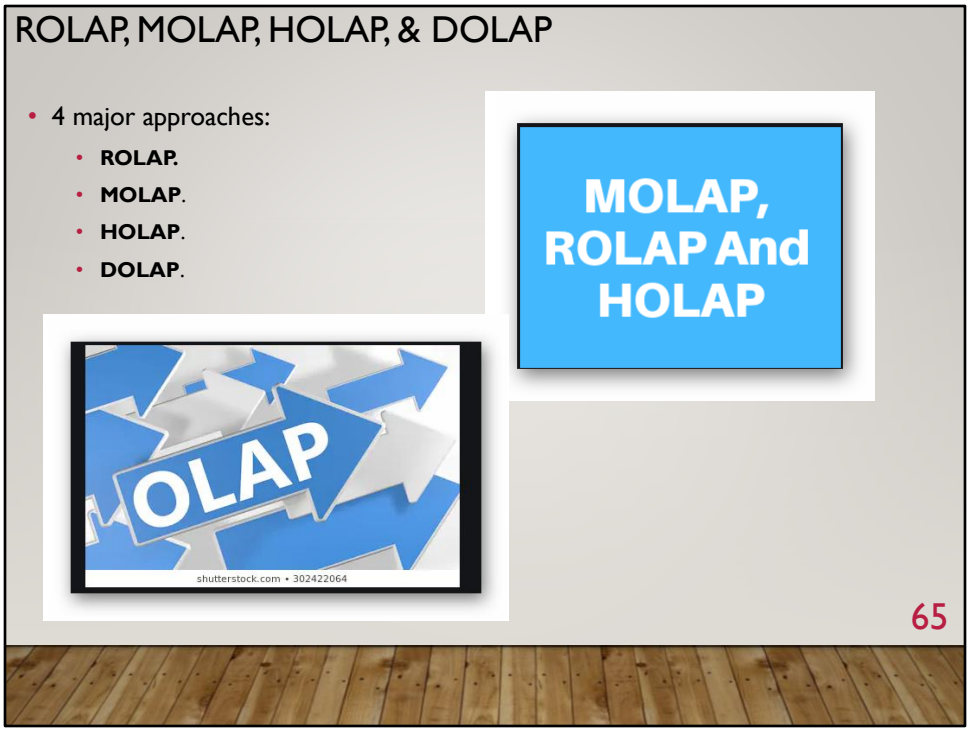



***FIGURE 1-14*** *Two cube aggregation levels. Every macro-event measure value is a sum of its component event values.*

*(Golfarelli & Rizzi, 2009)*

64




These acronyms represent 4 major approaches to implanting a data warehouse, and are related to the logical model used to represent the data.

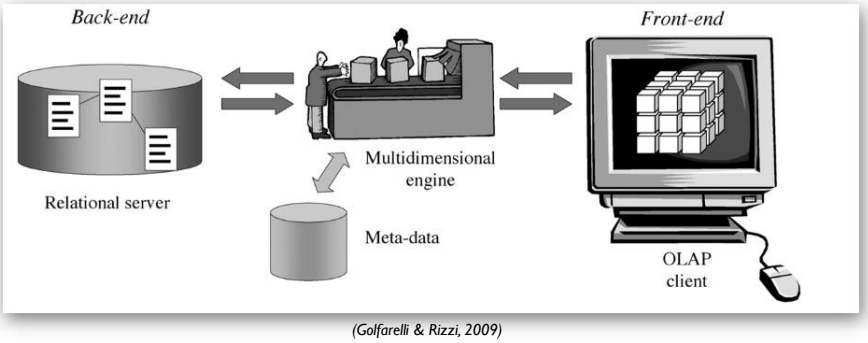
**ROLAP**: Relational OLAP and is based on relational DBMSs.

**MOLAP**: Multidimensional OLAP, based on multidimensional DBMSs.

**HOLAP**: Hybrid OLAP, uses both relational and multidimensional techniques.

**DOLAP**: Desk OLAP, stores the OLAP data in client-based files.

## ROLAP

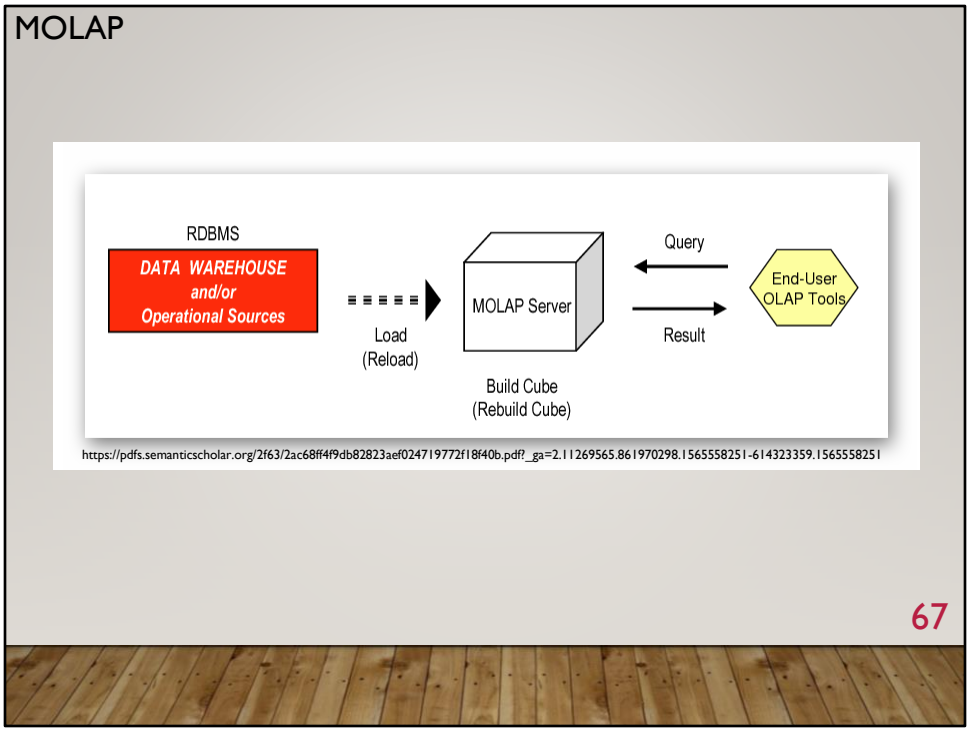



*(Golfarelli & Rizzi, 2009)*

66

ROLAP:

- Main problem with this approach is the performance hit due to costly joins.
- To reduce this problem ROLAP often utilizes the process of denormalization, which is a clear breach of the 3NF orientation of database structure.
- This often requires specialized middleware, called a multidimensional engine between the back-end relational servers and front end.




MOLAP:

- Based on an ad-hoc logical model which can be used to represent a multidimensional data and operations directly.
- Greatest advantage compared to ROLAP is you can perform multidimensional operations in an easy, natural way without the need for complex join operations.

## HOLAP

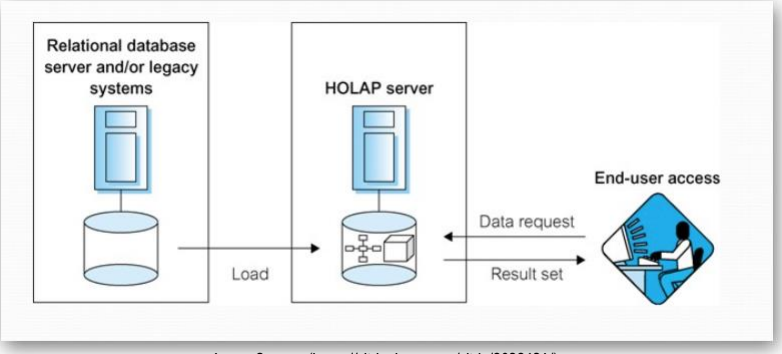



Image Source: (https://slideplayer.com/slide/9039631/)

68

HOLAP:

- Aim is to mix the advantages of both basic solutions (ROLAP & MOLAP).
- Takes advantage of standardization levels and the ability to manage large amounts of data from ROLAP implementations with the query speed of the typical MOLAP implementation.
- Implies that the largest amount of data should be stored in the RDBMS to avoid problems caused by sparsity.

DOLAP

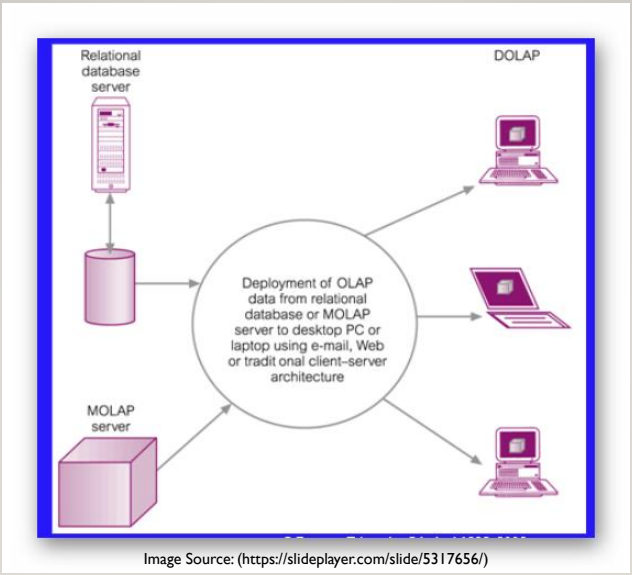



Image Source: (https://slideplayer.com/slide/5317656/)

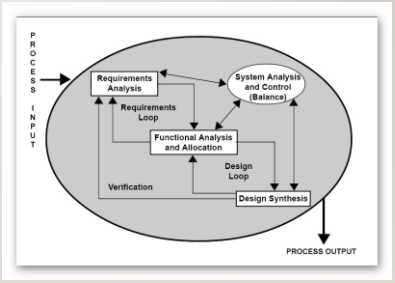
69

DOLAP:

- Supports multidimensional processing by using a client multidimensional engine.
- Typically administered by a central server or processing routine that prepare the data cubes and sets of data for each user.

## USER REQUIREMENT ANALYSIS

- Attempts to collect end users needs.
- Strategic significance.
- Delivers ambiguous, incomplete and short-lived requirements
- Two different approaches are presented by Golfarelli & Rizzi (2009):
  - Informal Approach.
  - Formal approach based on ***Tropos formalism.***




70

Tropos formalism: https://www.researchgate.net/publication/225198353_Tropos_An_Agent-Oriented_Software_Development_Methodology/link/02bfe50f9559e5c8da000000/download

This phase attempts to collect the end users needs.

Holds a strategic significance for designing data marts/warehouses.

This phase often delivers ambiguous, incomplete and short-lived requirements because of the following reasons:

- These are long lived projects
- The information requirements for the data warehouse applications are very difficult to explain because of the flexibility in the decision-making process.
- The decision-making process requirements often make references to information that is not available in a format suitable for the needs to be derived from.

Two different approaches are presented by Golfarelli & Rizzi (2009):

- Informal Approach which requires glossaries to support designers during the conceptual design phase.
- Formal approach which is based on ***Tropos formalism*** for user requirements.
- Visit this site to learn more of the Tropos formalism: https://www.researchgate.net/publication/225198353_Tropos_An_Agent-Oriented_Software_Development_Methodology/link/02bfe50f9559e5c8da000000/download

## USER REQUIREMENT ANALYSIS

Informal Approach

- Data-driven design
- First create a *derivation table* and *usage table.*
- Listing all existing functional dependencies.
- *Structure table.*
- Conducted simultaneously.


71

Image Source: (https://performanceculture.com/providing-informal-feedback/)

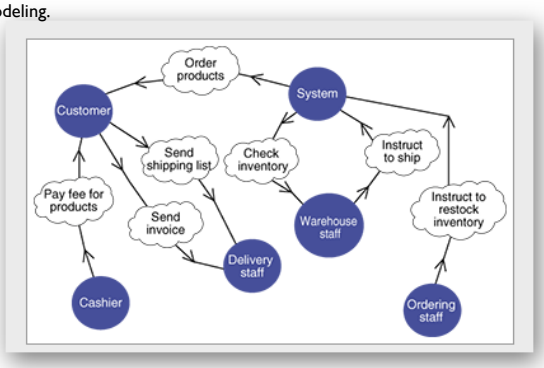
# Informal Approach (Glossary Based Requirement Analysis)

- Normally part of a data-driven design framework.
- Recommendation is to first create a *derivation table* and *usage table.*
  - *Derivation table:* establishes every attribute relationship with operational sources by specifying a schema attribute or procedures for values to be extracted.
  - Usage table: links textual descriptions to attributes and specifies roles, such as the analysis dimension and/or measure.
- Next a table listing all the existing functional dependencies between the attributes should be created.
- Finally it is suggested that a *structure table* be created which specifics whether the attributes should be modeled as dimensions, or attributes linked to dimensions or measures.
- This phase is conducted simultaneously with the conceptual design.

## USER REQUIREMENT ANALYSIS

Formal Approach

- Calls for the requirements analysis on goals.
  - Adopts two approaches:
    - Decision-making modeling.
    - Organizational modeling.




72

Image Source: (https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr200807sf3.html)

# Formal Approach (Goal Oriented Requirement Analysis)

- This approach calls for the requirements analysis to be based on the goals of the decision-makers in the organization.
- This approach further adopts two different approaches to conduct requirement analysis:
  - Decision-making modeling – focus is on the requirements of the organization's decision makers.
  - Organizational modeling – this approach targets the stakeholders, those that take part in managing the enterprise.

## USER REQUIREMENT ANALYSIS

- Interviews/Facilitated Sessions
  - Main source gain system requirements.
  - Gained from the business users.
  - Two procedures:
    - Interviews.
    - Facilitated Sessions.
  - Activities involved:
    - Pre-interview research
    - Interviewee selection
    - Interview question development
    - Interview scheduling
    - Interviewee preparation
  - Generally three types of questions:
    - Open-ended
    - Closed questions
    - Evidential questions

73

Image Source: (https://www.reed.co.uk/career-advice/group-interview-tasks-and-activities/)

## Interviews/Facilitated Sessions

- This is the main source from which to gain system requirements.
- These requirements are gained from the business users, or end users as they are known.
- Two basic procedures can be used:
  - Interviews: conducted with single users or small groups. Advantage is that everyone can contribute and participate in the discussion.
  - Facilitated Sessions: often involve large groups led by a facilitator, who is in charge of setting up a common language for all the interviewees.
- Some main activities involved in interviews are:
  - Pre-interview research
  - Interviewee selection
  - Interview question development
  - Interview scheduling
  - Interviewee preparation
- There are generally three types of questions you should ask:
  - Open-ended
  - Closed questions
  - Evidential questions

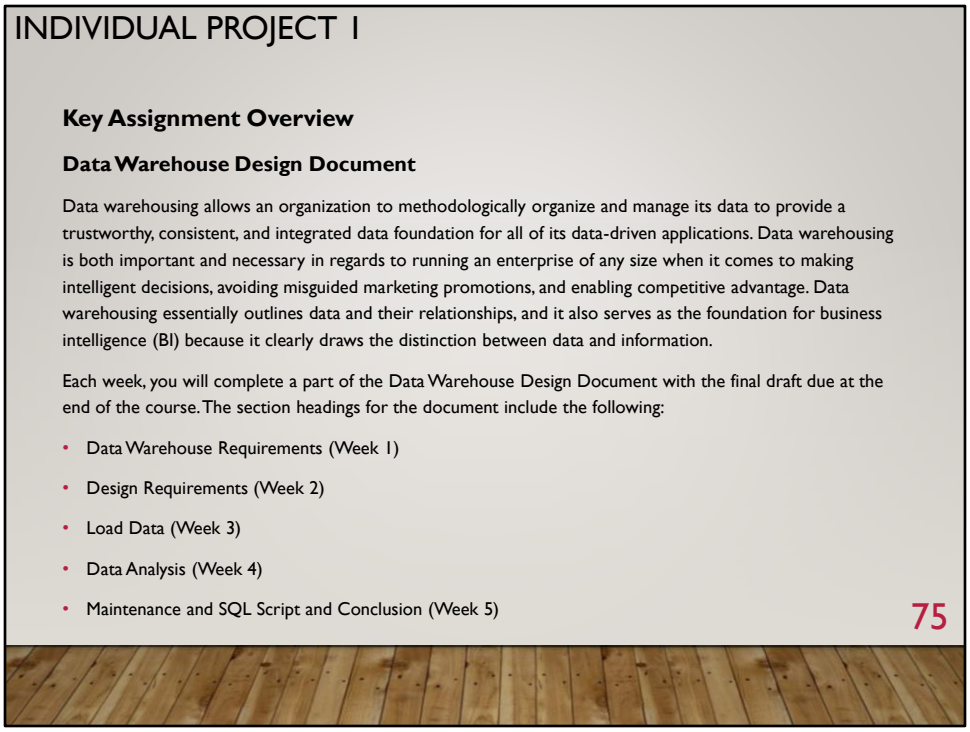
# USER REQUIREMENT ANALYSIS

TABLE 4-1 Advantages and Disadvantages of Three Types of Questions (Kendall & Kendall, 2002).

| Questions Asking For | Advantages | Disadvantages |
|---|---|---|
| Open-ended answers | They give the interviewer the opportunity to learn participants' vocabulary that proves their educational level, attitudes, and opinions.<br>Answers are rich in details.<br>They allow the interviewer to explore new possibilities that were not found in the interview preparation phase.<br>They get interviewees more involved. | Their answers can be long and rich in useless details.<br>They can result in digressions diverting from the interview goals.<br>They can be very time-demanding.<br>They can make interviewees think that the interviewer is not well-trained. |
| Closed answers | They shorten the interview time.<br>They make various interviewees' answers comparable.<br>They allow interviewees to focus on relevant elements. | They can sound boring to interviewees.<br>They do not allow interviewer to begin a real dialog with interviewees.<br>They assume that the interviewer has already guessed key factors. |
| Evidential answers | They allow interviewer to understand interviewees' knowledge level.<br>They show the interviewer's interest in understanding the interviewees' opinions. | They can make interviewees nervous because the questions are probing. |

- **Open-ended questions** Such as *What do you think of data source quality?* and *What are the key objectives your unit has to face?*
- **Closed questions** Such as *Are you interested in sorting out purchases by hour?* and *Do you want to receive a sales report every week?*
- **Evidential questions** Such as *Could you please give me an example of how you calculate your business unit budget?* and *Could you please describe the issues with poor data quality that your business unit is experiencing?*

*(Golfarelli & Rizzi, 2009)*




Throughout this course, you will work on several aspects of data warehousing that will result in a Data Warehouse Design Document. This course is comprised of a series of Individual Project assignments that will contribute to a Key Assignment submission at the end of the course. Each week, you will complete a part of a Data Warehouse Design Document. You will use an organization of your choice and apply your research to the development of the Data Warehouse Design Document. Appropriate research should be conducted to support the development of your document, and assumptions may be made when necessary. The goal of this course is to design a Data Warehouse Design Document that would reflect an actual data warehouse implementation in an enterprise.

# INDIVIDUAL PROJECT 1

## Organization and Project Selection

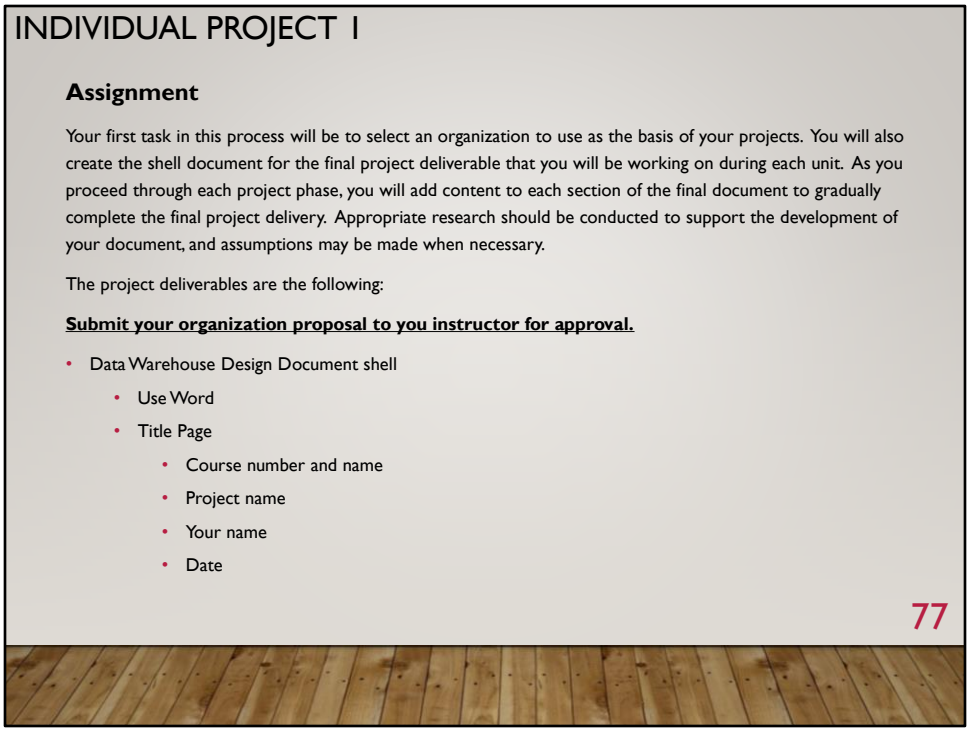
The first step will be to select a **real** or **hypothetical organization** as the target for your Data Warehouse Design Document. This organization will be used as the basis for each of the assignments throughout the course, and it should conform to the following guidelines:

- **Sensitivity:** The selected organization should be large and should contain sensitive data requiring the implementation of security measures.
- **Familiarity:** You should be familiar enough with the organization and typical security needs without significant time required for security research and education.
- **Accessibility:** You should have good access to security officers and management or incident response personnel in the organization, because these resources will provide direction as they progress throughout the development of the report.

The selected organization must have a need for some kind of data warehouse because of poor data quality modeling in its operations. Therefore, feel free to identify a hypothetical organization that meets the requirements. Any necessary assumptions may be made to fulfill the requirements of organization selection.

**Select an organization that fits these requirements, and submit your proposal to your instructor before proceeding further with the assignments in the course**. Approval should be sought within the first several days of the course. Your instructor will tell you how to submit this proposal and what notification will be given for project approval.

76




For the assignments in this course, you will not be implementing a Data Warehouse Design Document, but you will be proposing a requirements elicitation process for a data warehouse (DW) that identifies its information contents. These contents support the set of decisions that can be made. Thus, if the information that is needed to take every decision is elicited, then the total information determines the DW's contents.

# INDIVIDUAL PROJECT 1

**Assignment**

- Table of Contents (TOC)
  - Use an autogenerated TOC.
  - This must be on a separate page.
  - This must be a maximum of 3 levels deep.
  - Be sure to update the fields of the TOC before submitting your project.
- Section Headings (***Create each heading on a new page with "TBD" as content, except for the section for Week 1.)***
  - Data Warehouse Requirements (Week 1)
  - Design Requirements (Week 2)
  - Load Data (Week 3)
  - Data Analysis (Week 4)
  - Maintenance and SQL Script and Conclusion (Week 5)

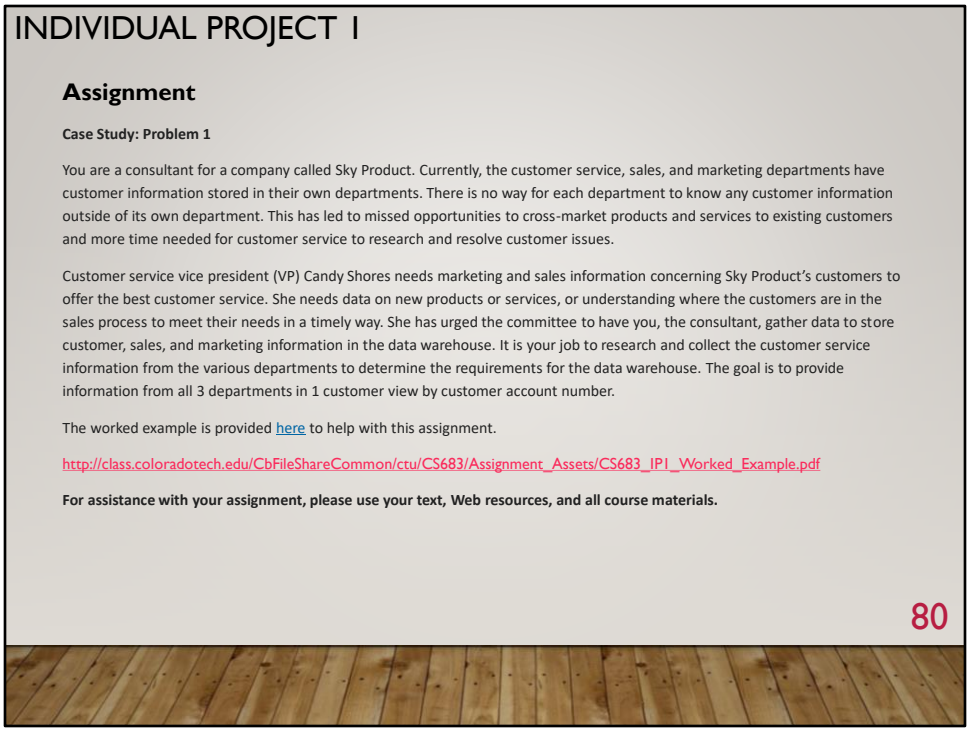
# INDIVIDUAL PROJECT 1

## Assignment

**Week 1:** Data Warehouse Requirements

- Give a brief description of the company (can be hypothetical) where the Data Warehouse Design Document will be implemented.
  - Include the company's size, location(s), and other pertinent information.
- Describe the process that you will use to gather data warehouse requirements during data acquisition (source data and data staging), data storage (data warehouse database management system, data marts and metadata), and information delivery (master data database, data mining, online analytical processing [OLAP], and report query).
- The requirements should include the following:
  - Data Acquisition
    - Source data
    - Data staging
  - Data Storage
    - Data warehouse (DW) database management system (DBMS)
    - Data marts
    - Metadata
  - Information Delivery
    - Multidimensional database (MDDB)
    - Data mining
    - OLAP
    - Report and query
  - Management and Control
- Name the document "yourname_CS683_IP1.doc."
- **Be sure that this project is approved by the instructor**.

79




**Worked Example**

Please refer to the Worked Example for an example of this assignment based on the Problem-Based Learning Scenario. The Worked Example outlines the Key Assignment using a fictional company called Sky Product. The worked example is not intended to be a complete example of the assignment, but it will illustrate the basic concepts 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.

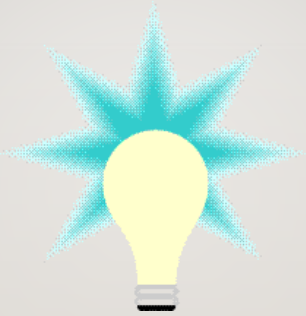
# CONTACT INFORMATION

- 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



82

(Image Source: 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

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.