IIT School of Applied Technology
ILLINOIS INSTITUTE OF TECHNOLOGY
information technology & management

526 Data Warehousing

February 12, 2018
Week 5 Presentation

Visualization Tips Presentation

- Sign up your slot in Google Sheet:
  https://docs.google.com/spreadsheets/d/17hWmYrKqLj_lzhceXuAHCAe2GksNQr0374iXejbCmFc/edit?usp=sharing
- Upload your TDWX or PBIX:
  https://drive.google.com/drive/folders/1MFd3NcxnePkLfTgZ05ofTkUylFSK-WrG?usp=sharing
- Submit the file in BB as well
- Give a quick demo during class (Not more than 3 minutes per person, up to 3 persons)

Dimensional Modeling Core Concepts – Dimension Tables

- We will discuss
  - Business-driven goals of DW/BI
  - Deliverables of DW/BI
  - Dimensional modeling core concepts
  - Dimension Tables
  - Slowly Changing Dimensions

Business-Driven Goals of DW/BI

- Business needs we hear frequently:
  - "We collect tons of data, but we can't access it."
  - "We need to slice and dice the data every which way."
  - "Business people need to get at the data easily."
  - "Just show me what is important."
  - "We spend entire meetings arguing about who has the right numbers rather than making decisions."
  - "We want people to use information to support more fact-based decision making."

Goals of DW/BI

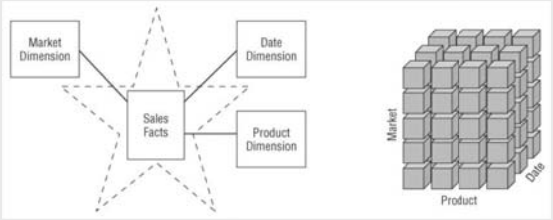
- Make information accessible and understandable to the business users
- Present information consistently in a timely manner
- Ensure DW/BI deliverables accepted by business community to support their decision making

Deliverables that Achieve the DW/BI Goals

- Dimensional models
  - Relational star schemas
    - Fact tables for measurements
    - Dimension tables for descriptors
  - Multidimensional online analytical processing (OLAP) cubes
- Business Intelligence applications
  - Reporting, analytic, visualization tools, etc.

Star Schema VS. OLAP Cubes

- At a logical level, there is no difference
- It is a matter of physical database implementation.
- **Star schema** is implemented in a relational database and is queried through SQL
- **OLAP Cubes** (multidimensional databases) are implemented for extreme performance and are queried through MDX.

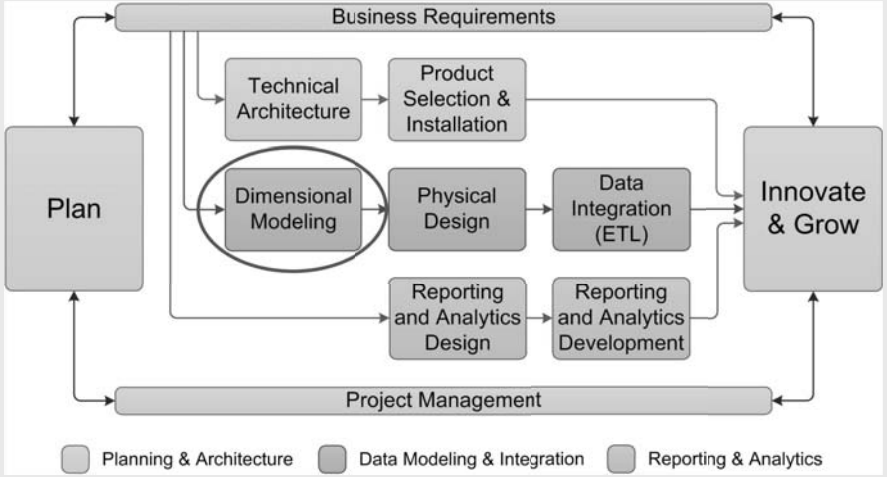


Star Schema VS. OLAP Cubes

- The star schema can store large amounts of detailed data.
- OLAP Cubes provide higher performance with pre-calculated summary data.
- In general, OLAP cubes are populated from the star schema

- Kimball focuses on the star schema rather than the OLAP cubes
- Star schema usually has 15 dimensions
- OLAP usually has 8-10 dimensions

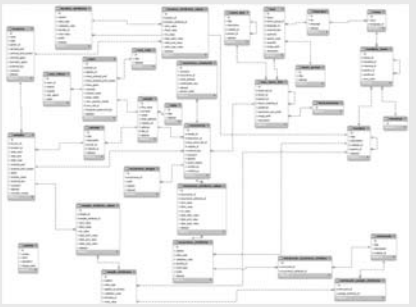
Kimball Lifecycle Approach



Dimensional Modeling Fundamental Concepts: Make it Simple

- 3NF: Immensely useful in operational processing (OLTP)
- However, it's too complicated for users to use.
- Unpredictable queries cause performance problems

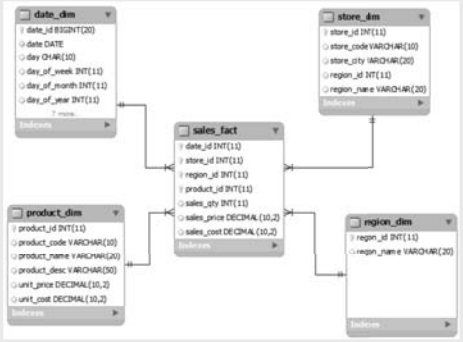
3NF



Dimensional Modeling Fundamental Concepts: Make it Simple

- Dimensional modeling came to a rescue in the presentation layer
- Dimensional modeling provides
  - Understandability
  - Query performance
  - Resilience to change
- Making it simple via **Denormalization**

Dimensional Model



Dimensional Modeling Fundamental Concepts: Make it Simple

- Widely accepted model for presenting analytic data
- Techniques for making database simple through denormalization
- "*We sell products in various **stores** in different **regions** and measure our performance over **time***": emphasis on
  - Products, Stores, Regions, and Time
  - Performance: sales volume, profit
  - → Make a simple modeling
- Simple modeling is important
  - Resist temptation to over-engineer

ITMD - 526

Dimensions – Physical Table Elements

- Contain descriptive attributes that are typically textual fields
- Shallow and wide
- Corresponds to entities that business interacts with
  - Customer, Employee, Products, Accounts
- Surrogate key as a single column PK

Product Dimension
Product Key (PK)
SKU Number (Natural Key)
Product Description
Brand Name
Category Name
Department Name
Package Type
Package Size
Abrasive Indicator
Weight
Weight Unit of Measure
Storage Type
Shelf Life Type
Shelf Width
Shelf Height
Shelf Depth
...

ITMD - 526

Descriptive Dimension Attributes

- Describe "Who, What, Where, When and Why"
- Consists of words rather than cryptic abbreviation

- DW is only as good as the dimension attributes
- Embedded meaning within codes as separate attributes

ITMD - 526

Descriptive Dimension Attributes

- Denormalized many-to-one hierarchies

Product Key	Product Description	Brand Name	Category Name
1	PowerAll 20 oz	PowerClean	All Purpose Cleaner
2	PowerAll 32 oz	PowerClean	All Purpose Cleaner
3	PowerAll 48 oz	PowerClean	All Purpose Cleaner
4	PowerAll 64 oz	PowerClean	All Purpose Cleaner
5	ZipAll 20 oz	Zippy	All Purpose Cleaner
6	ZipAll 32 oz	Zippy	All Purpose Cleaner
7	ZipAll 48 oz	Zippy	All Purpose Cleaner
8	Shiny 20 oz	Clean Fast	Glass Cleaner
9	Shiny 32 oz	Clean Fast	Glass Cleaner
10	ZipGlass 20 oz	Zippy	Glass Cleaner
11	ZipGlass 32 oz	Zippy	Glass Cleaner

Product, Brand and Category to Product Dimension

- Operational natural key as an attribute, not as a primary key

ITMD - 526

Calendar Date Dimension

- All dimensional models need a time component
- Generally the calendar date dimension with the granularity of a single day
- Surprisingly has many attributes
  - Holidays, work days, fiscal periods, week numbers, last day of month flags, etc.

ID	Date	Day	DaySuffix	DayOfWeek	DayOfYear	WeekOfYear	WeekOfMonth	Month	MonthName	Quarter	QuarterName	Year
20090327	2009-03-27 00:00:00.000	27	27th	Friday	86	13	4	3	March	1	First	2009
20090328	2009-03-28 00:00:00.000	28	28th	Saturday	87	13	4	3	March	1	First	2009
20090329	2009-03-29 00:00:00.000	29	29th	Sunday	88	14	5	3	March	1	First	2009
20090330	2009-03-30 00:00:00.000	30	30th	Monday	89	14	5	3	March	1	First	2009
20090331	2009-03-31 00:00:00.000	31	31st	Tuesday	90	14	5	3	March	1	First	2009
20090401	2009-04-01 00:00:00.000	1	1st	Wednesday	91	14	1	4	April	2	Second	2009
20090402	2009-04-02 00:00:00.000	2	2nd	Thursday	92	14	1	4	April	2	Second	2009
20090403	2009-04-03 00:00:00.000	3	3rd	Friday	93	14	1	4	April	2	Second	2009
20090404	2009-04-04 00:00:00.000	4	4th	Saturday	94	14	1	4	April	2	Second	2009
20090405	2009-04-05 00:00:00.000	5	5th	Sunday	95	15	2	4	April	2	Second	2009
20090406	2009-04-06 00:00:00.000	6	6th	Monday	96	15	2	4	April	2	Second	2009
20090407	2009-04-07 00:00:00.000	7	7th	Tuesday	97	15	2	4	April	2	Second	2009

Slowly Changing Dimensions (SCD)

- Dimensions contains relatively static data such as
  - Geo locations, customers, or products
- Data in the dimensions change slowly and in unpredictable time
- This can cause referential integrity issue between a fact and dimension tables (e.g. an employee left the job)
- SCD is a mechanism to deal with these changes in dimensions

ILLINOIS INSTITUTE OF TECHNOLOGY School of Applied Technology

ITMD - 526

SCD Type 1

- Overwrite the old with new data
  - Pre-existing facts now refer to the updated Dimension
  - May cause inconsistent reports

Supplier_Key	Supplier_Code	Supplier_Name	Supplier_State
123	ABC	Acme Supply Co	CA

↓

Supplier_Key	Supplier_Code	Supplier_Name	Supplier_State
123	ABC	Acme Supply Co	IL

ILLINOIS INSTITUTE OF TECHNOLOGY School of Applied Technology

ITMD - 526

SCD Type 2

- Insert a new dimension row with the new data and new effective start date
- Update the effective end date on the prior row
- Maintains the historical context of the data
- Fact tables reference to the correct snapshot information of a SCD type 2 dimension
- Results in multiple dimension rows for a given natural key

Supplier_Key	Supplier_Code	Supplier_Name	Supplier_State	Start_Date	End_Date
123	ABC	Acme Supply Co	CA	2000-01-01	2004-12-21
124	ABC	Acme Supply Co	IL	2004-12-21	NULL

ILLINOIS INSTITUTE OF TECHNOLOGY School of Applied Technology

ITMD - 526

Dimensional Modeling Fundamental Concepts

Questions?

Class Exercises

- Kettle: Jobs and Environmental Variables – Your ETL code must use environmental variables for DB user/password and folder location
- Normalization
- Denormalization
- Tableau/PowerBI Installation and Tutorials