About this presentation

Now that we know how to draft our dimensional model, we can start to plan our low-level database design.

Agenda

Dimensional tables core concepts

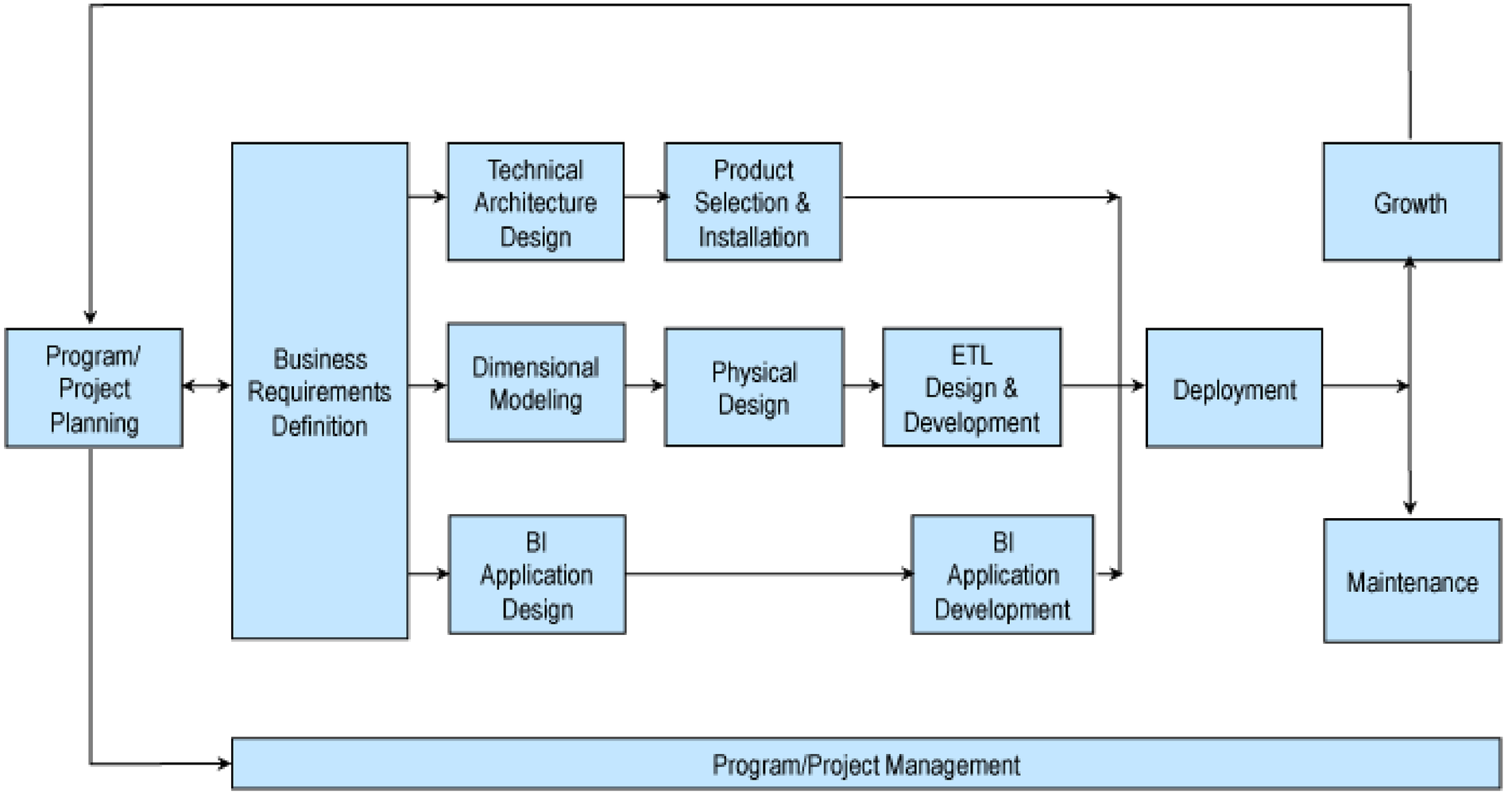
Hierarchies

Binary, Null, and Changing Values

Common Dimensions

Fall 2024

The data warehouse lifecycle



Reference: https://www.kimballgroup.com

Dimension tables

Now that we have determined our dimensions and facts at a high-level, we need to identify our dimensional attributes.

Dimension tables

Surrogate key: This is a primary key for a record that is unique, but has no business meaning. For example, an auto-increment value (1,2,3,4,5).

Natural key/Business key: A unique value that can be used as a primary key for a record, but has a business meaning. For example, a Social Security Number.

Dimension tables should almost always use a surrogate key as a primary key.

Why? Natural keys that have business meaning can change. For example, if an employee quits and is re-hired, their employee ID number may change.

Can’t we just use the primary key from the transactional system? No. When you combine data from multiple transactional systems, you may end up with multiple records with the same primary key (which can’t happen).

Reference: Kimball, Ralph, and Margy Ross. *The Data Warehouse Toolkit : The Definitive Guide to Dimensional Modeling*, John Wiley & Sons, Incorporated, 2013.

Keys to success

Employee system (North America):

Employee# (pk)	Name	Hire Date	Department
1	Dominic Sellitto	1/1/21	Frozen
2	Jane Smith	2/3/21	Customer Service
3	John Doe	2/4/21	Management

Employee system (Europe):

Employee# (pk)	Name	Hire Date	Department
1	Brad Pitt	1/2/21	Dairy
2	James Bond	4/3/21	Customer Service
3	John Doe	2/4/21	Management

What happens if we use Employee# as the PK for our dimension table?

Reference: Kimball, Ralph, and Margy Ross. *The Data Warehouse Toolkit : The Definitive Guide to Dimensional Modeling*, John Wiley & Sons, Incorporated, 2013.

Keys to success

Employee dimension:

Employee# (pk)	Name	Hire Date	Department
1	Dominic Sellitto	1/1/21	Frozen
1	Brad Pitt	1/2/21	Dairy
2	Jane Smith	2/3/21	Customer Service
3	John Doe	2/4/21	Management
2	James Bond	4/3/21	Customer Service
3	John Doe	2/4/21	Management

Primary keys must be unique!

Reference: Kimball, Ralph, and Margy Ross. *The Data Warehouse Toolkit : The Definitive Guide to Dimensional Modeling*, John Wiley & Sons, Incorporated, 2013.

Keys to success

Instead, try this:

EmpSK#(pk)	Employee# (nk)	Name	Hire Date	Department
1	1|NA	Dominic Sellitto	1/1/21	Frozen
2	1|EU	Brad Pitt	1/2/21	Dairy
3	2|NA	Jane Smith	2/3/21	Customer Service
4	3|NA	John Doe	2/4/21	Management
5	2|EU	James Bond	4/3/21	Customer Service
6	3|EU	John Doe	2/4/21	Management

Affixing the employee number with some source system info can help.

Reference: Kimball, Ralph, and Margy Ross. *The Data Warehouse Toolkit : The Definitive Guide to Dimensional Modeling*, John Wiley & Sons, Incorporated, 2013.

Dimension tables

Degenerate dimensions: A dimension with no data. Degenerate dimensions are typically placed in a fact table.

Invoice #12345

Customer: John Smith

Date: 1/1/2021

Store: 123 ABC St.

Products:

Cheese x 1 : $12.95
Milk x 1 : $3.95
Eggs x 1 : $4.35

Reference: Kimball, Ralph, and Margy Ross. *The Data Warehouse Toolkit : The Definitive Guide to Dimensional Modeling*, John Wiley & Sons, Incorporated, 2013.

Dimension tables

Degenerate dimensions: A dimension with no data. Degenerate dimensions are typically placed in a fact table.

Surrogate Key	Product (dim)	Customer (dim)	Date (dim)	Store (dim)	Invoice (ddim)
1	1 (Cheese)	1 (John Smith)	1234 (1/1/21)	46 (123 ABC St.)	12345
2	2 (Milk)	1 (John Smith)	1234 (1/1/21)	46 (123 ABC St.)	12345
3	3 (Eggs)	1 (John Smith)	1234 (1/1/21)	46 (123 ABC St.)	12345

Reference: Kimball, Ralph, and Margy Ross. *The Data Warehouse Toolkit : The Definitive Guide to Dimensional Modeling*, John Wiley & Sons, Incorporated, 2013.

Dimension tables

Degenerate dimensions: A dimension with no data. Degenerate dimensions are typically placed in a fact table.

In some cases, the degenerate dimension *Invoice* could also be a natural key.

Surrogate Key	Product (dim)	Customer (dim)	Date (dim)	Store (dim)	Invoice (ddim)
1	1 (Cheese)	1 (John Smith)	1234 (1/1/21)	46 (123 ABC St.)	12345
2	2 (Milk)	1 (John Smith)	1234 (1/1/21)	46 (123 ABC St.)	12345
3	3 (Eggs)	1 (John Smith)	1234 (1/1/21)	46 (123 ABC St.)	12345

Reference: Kimball, Ralph, and Margy Ross. *The Data Warehouse Toolkit : The Definitive Guide to Dimensional Modeling*, John Wiley & Sons, Incorporated, 2013.

Hierarchies

Hierarchies: Defines aggregation paths for data. In star schema data warehouses, hierarchies are ***reporting*** structures, not normalization mechanisms.

Hierarchies support drill-down and drill-up functionalities in reporting.

- Address -> City -> State -> Region -> Country -> Continent -> Earth
- 123 Fake st. -> Buffalo -> NY -> Northeast -> USA -> North America -> Earth

Surrogate Key	Address	City	State	Region	Country
1	123 Fake St.	Buffalo	NY	Northeast	USA
2	143 ABC Rd.	Buffalo	NY	Northeast	USA
3	234 Pavement	Philadelphia	PA	Northeast	USA

Reference: Kimball, Ralph, and Margy Ross. *The Data Warehouse Toolkit : The Definitive Guide to Dimensional Modeling*, John Wiley & Sons, Incorporated, 2013.

Fixed Hierarchies

Fixed Hierarchies: Each attribute has a One-to-Many relationship with the next attribute and always have the same number of attribute levels.

- Products -> Subcategories -> Categories -> Departments (4 levels)
  - If this hierarchy is fixed, then each product will have a subcategory, category, and department
  - Each department will be associated with many categories, subcategories, and products

Surrogate Key	Product Name	Subcategory	Category	Department
1	Hershey Bar	Chocolate	Candy	Snacks
2	Crunch Bar	Chocolate	Candy	Snacks
3	Lays Chips	Potato	Chips	Snacks
4	Sun Chips	Multigrain	Chips	Snacks

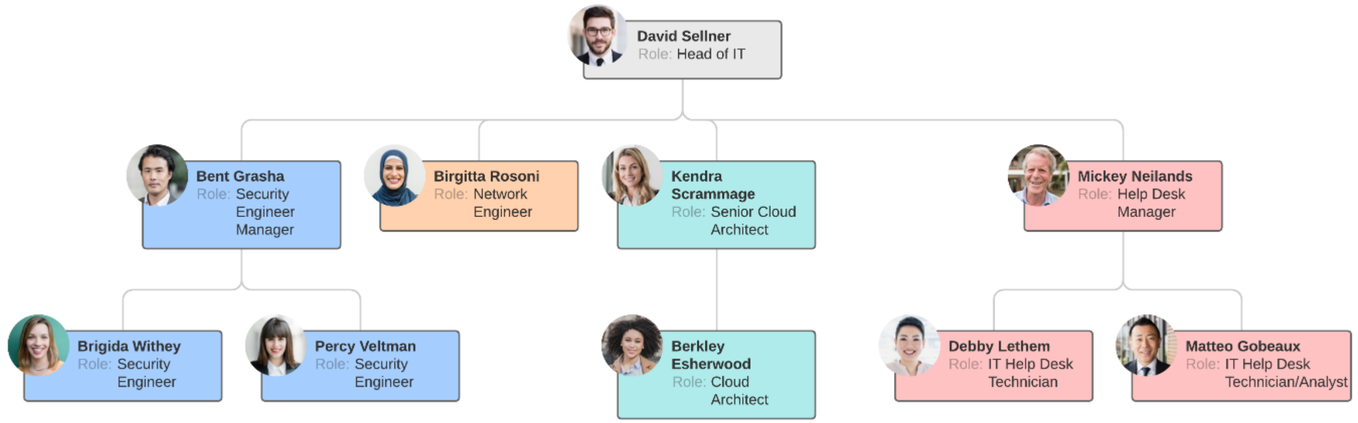
Reference: Kimball, Ralph, and Margy Ross. *The Data Warehouse Toolkit : The Definitive Guide to Dimensional Modeling*, John Wiley & Sons, Incorporated, 2013.

Variable Depth Hierarchies

Variable Depth Hierarchies: Do not have a fixed number of levels. These fall under two categories:

- Slightly ragged
- Ragged

Think about an organizational chart…



Reference: Kimball, Ralph, and Margy Ross. *The Data Warehouse Toolkit : The Definitive Guide to Dimensional Modeling*, John Wiley & Sons, Incorporated, 2013.
https://www.lucidchart.com/blog/types-of-organizational-structures

Force-Fitting Slightly Ragged Hierarchies

A slightly ragged hierarchy is called ***slightly ragged*** because the number of levels only has minor variations.

In these cases, you can “force” a slightly ragged hierarchy to be a fixed hierarchy, by populating unknown values with another standard value (e.g., the parent value).

Simple Loc
Loc Key (PK)
Address+
City
City
City
State
Country
...

Medium Loc
Loc Key (PK)
Address+
City
City
Zone
State
Country
...

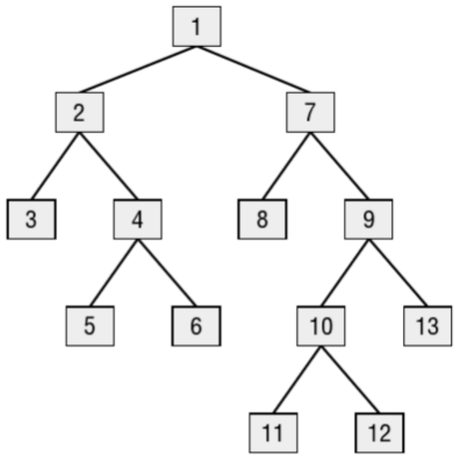
Complex Loc
Loc Key (PK)
Address+
City
District
Zone
State
Country
...

Reference: Kimball, Ralph, and Margy Ross. *The Data Warehouse Toolkit : The Definitive Guide to Dimensional Modeling*, John Wiley & Sons, Incorporated, 2013.

Ragged Hierarchies

Ragged hierarchies have higher variation in levels and, sometimes, no clear pathway upward in the hierarchy. If they can't easily be flatted into fixed hierarchies, there are a few techniques to address…

Recursion: A method of representing a parent-child relationship by having the parent surrogate key as an attribute of the child record.

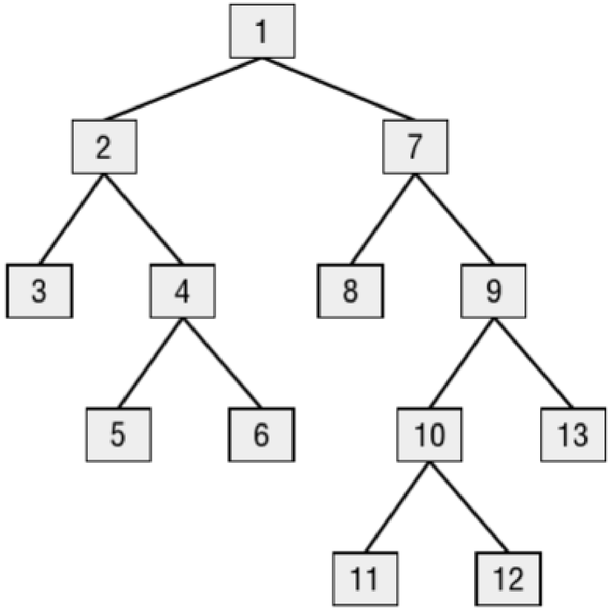


Employee_Key	Name	Manager_Key
1	Sarah	NULL
2	Jeff	1
3	Millie	2
4	John	2
5	Jimmy	4
6	Sam	4
7	Sally	1
8	Peter	7
9	Bob	7
10	Meghan	9
11	Timothy	10
12	Harold	10
13	Arnold	9

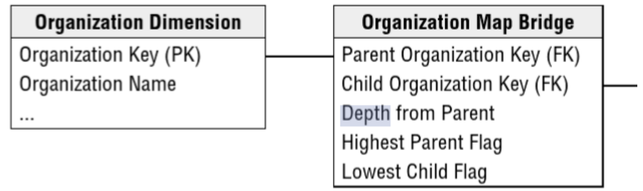
Reference: Kimball, Ralph, and Margy Ross. *The Data Warehouse Toolkit : The Definitive Guide to Dimensional Modeling*, John Wiley & Sons, Incorporated, 2013, https://www.nuwavesolutions.com/ragged_hierarchical_dimensions/

Ragged Hierarchies

Hierarchical Bridge Table: An “intermediary” table that connects a ragged dimension with a fact table and provides information about all of the rollup possibilities for the ragged hierarchy.



Parent Organization Key	Child Organization Key	Depth from Parent	Highest Parent Flag	Lowest Child Flag
1	1	0	TRUE	FALSE
1	2	1	TRUE	FALSE
1	3	2	TRUE	TRUE
1	4	2	TRUE	FALSE
1	5	3	TRUE	TRUE
1	6	3	TRUE	TRUE
1	7	1	TRUE	FALSE
1	8	2	TRUE	TRUE
1	9	2	TRUE	FALSE
1	10	3	TRUE	FALSE
1	11	4	TRUE	TRUE
1	12	4	TRUE	TRUE
1	13	3	TRUE	TRUE
2	2	0	FALSE	FALSE
2	3	1	FALSE	TRUE
2	4	1	FALSE	FALSE
2	5	2	FALSE	TRUE
2	6	2	FALSE	TRUE
3	3	0	FALSE	TRUE
4	4	0	FALSE	FALSE
4	5	1	FALSE	TRUE
4	6	1	FALSE	TRUE
5	5	0	FALSE	TRUE



Reference: Kimball, Ralph, and Margy Ross. *The Data Warehouse Toolkit : The Definitive Guide to Dimensional Modeling*, John Wiley & Sons, Incorporated, 2013

Binary and Abbreviated Attributes

Transactional systems often have attributes that are designed to simplify operational activities. For example, health insurance codes or SAP transaction codes.

Code	Description
92507	Speech/hearing therapy
92508	Speech/hearing therapy
92520	Laryngeal function studies
92521	Evaluation of speech fluency
92522	Evaluate speech production

Which one is easier to understand?

Reference: Kimball, Ralph, and Margy Ross. *The Data Warehouse Toolkit : The Definitive Guide to Dimensional Modeling*, John Wiley & Sons, Incorporated, 2013,
2021 DHS Code List (https://www.cms.gov/license/ama?file=/files/zip/list-codes-effective-january-1-2021-issued-december-1-2020.zip

Binary and Abbreviated Attributes

These attributes should be transformed / translated into ones that are simple for end-users to understand.

Name	Status
John Smith	1
James Bond	1
Sally Sampson	0
Marge Simpson	0
Indiana Jones	1

VS

Name	Status
John Smith	Active
James Bond	Active
Sally Sampson	Inactive
Marge Simpson	Inactive
Indiana Jones	Active

Reference: Kimball, Ralph, and Margy Ross. *The Data Warehouse Toolkit : The Definitive Guide to Dimensional Modeling*, John Wiley & Sons, Incorporated, 2013

NULL values

Sometimes, the source data is incomplete. In these cases, we should be adding “Unknown” or “Not Applicable” into a field. Avoid using the terminology “NULL” with empty data, in favor of an easier-to-read method.

Additionally, some database management systems and business intelligence software have special handling considerations for “NULL” values, which may cause issues for you down the road.

Reference: Kimball, Ralph, and Margy Ross. *The Data Warehouse Toolkit : The Definitive Guide to Dimensional Modeling*, John Wiley & Sons, Incorporated, 2013

Conformed Dimensions

A conformed dimension is a dimension that can be tied to more than one fact table. Conformed dimensions create flexibility and scalability, as the data in the dimension is all in the “same language” and is fit for multiple contexts.

For example, a **date** dimension should contain all of the attributes needed by the organization in a consistent format (irrespective of source transactional systems):

- Year
- Quarter (fiscal quarter)
- Season
- Month
- Day of the month
- Day of the week
- Holiday
- Etc.

Reference: Kimball, Ralph, and Margy Ross. *The Data Warehouse Toolkit : The Definitive Guide to Dimensional Modeling*, John Wiley & Sons, Incorporated, 2013

Shrunken Dimensions

A Shrunken Dimension represents a subset of a conformed dimension used when the source data doesn’t contain all of the information.

For example, a **shrunken date dimension** may relate to a fact table that only captures sales by quarter and year, but not at a lower granularity:

- Year
- Quarter (fiscal quarter)
- ~~Season~~
- ~~Month~~
- ~~Day of the month~~
- ~~Day of the week~~
- ~~Holiday~~
- ~~Etc.~~

Reference: Kimball, Ralph, and Margy Ross. *The Data Warehouse Toolkit : The Definitive Guide to Dimensional Modeling*, John Wiley & Sons, Incorporated, 2013

Changing dimensional attributes

Dimension attributes (fields) may need to be changed over time. Values may be updated, added, or deleted. We need to consider the **type and context** of a change before deciding on how to update our dimension tables.

There are three main types of slowly changing dimensions:

- Type 1: Overwrite
- Type 2: Add new row
- Type 3: Add new attribute

Note: Type 0 is for unchanging dimensional attributes.

Reference: Kimball, Ralph, and Margy Ross. *The Data Warehouse Toolkit : The Definitive Guide to Dimensional Modeling*, John Wiley & Sons, Incorporated, 2013,
https://www.kimballgroup.com/2008/08/slowly-changing-dimensions/

Changing dimensional attributes

Type 1 changing dimensional attributes involve you overwriting an attribute in an existing record in the dimension table.

While this is the simplest method, it is also the most dangerous in terms of data quality.

Let’s take an example: Customer #1, Dominic Sellitto, has a current “City” attribute set to Buffalo in the Customer dimension table.

Surrogate Key	Customer Name	Address	City	State
1	Dominic Sellitto	123 Fake St.	Buffalo	NY

This City was added in error, and the customer updates the City to the correct city, Lancaster:

Surrogate Key	Customer Name	Address	City	State
1	Dominic Sellitto	123 Fake St.	Lancaster	NY

Reference: Kimball, Ralph, and Margy Ross. *The Data Warehouse Toolkit : The Definitive Guide to Dimensional Modeling*, John Wiley & Sons, Incorporated, 2013, https://www.kimballgroup.com/2008/08/slowly-changing-dimensions/

Changing dimensional attributes

In this scenario, it might make sense to **overwrite** the city. After all, the customer address was incorrect, and you want all **past** activity related to the proper corrected address.

Let’s switch up the context. Instead of updating the address, let’s say Dominic moves.

Here’s the old data:

Surrogate Key	Customer Name	Address	City	State
1	Dominic Sellitto	123 Fake St.	Buffalo	NY

And the new data:

Surrogate Key	Customer Name	Address	City	State
1	Dominic Sellitto	123 Cool St.	Lancaster	NY

Now we have a problem.

Reference: Kimball, Ralph, and Margy Ross. *The Data Warehouse Toolkit : The Definitive Guide to Dimensional Modeling*, John Wiley & Sons, Incorporated, 2013, https://www.kimballgroup.com/2008/08/slowly-changing-dimensions/

Changing dimensional attributes

Overwrites are **destructive.** That means, all history of the previous value is lost. If Dominic had orders tied to his previous address, they now will all be **incorrectly** associated with his new address.

Surrogate Key	Customer Name	Address	City	State
1	Dominic Sellitto	123 Cool St.	Lancaster	NY

Overwrites can also have serious regulatory and legal consequences, depending on the organization and its reporting requirements.

Reference: Kimball, Ralph, and Margy Ross. *The Data Warehouse Toolkit : The Definitive Guide to Dimensional Modeling*, John Wiley & Sons, Incorporated, 2013, https://www.kimballgroup.com/2008/08/slowly-changing-dimensions/

Changing dimensional attributes

Type 2 changing dimensional attributes involve you adding a new row to the dimension to reflect a changed dimension.

This is slightly more complex and requires adding some additional attributes to your dimension.

For example:

Surrogate Key (pk)	CusID (nk)	Customer Name	Address	City	State	Valid From	Valid To	Current Row
1	12345|NA	Dominic Sellitto	123 Fake St.	Buffalo	NY	1/1/21	N/A	Current

Becomes:

Surrogate Key (pk)	CusID (nk)	Customer Name	Address	City	State	Valid From	Valid To	Current Row
1	12345|NA	Dominic Sellitto	123 Fake St.	Buffalo	NY	1/1/21	4/1/21	Expired
2	12345|NA	Dominic Sellitto	123 Cool St.	Lancaster	NY	4/2/21	N/A	Current

Reference: Kimball, Ralph, and Margy Ross. *The Data Warehouse Toolkit : The Definitive Guide to Dimensional Modeling*, John Wiley & Sons, Incorporated, 2013,
https://www.nuwavesolutions.com/slowly-changing-dimensions/

Changing dimensional attributes

Type 3 changing dimensional attributes involve adding a new attribute to the dimension to reflect a changed dimension.

This method is not preferred, as it creates conflicting elements in a single record and is less scalable over time.

For example:

Surrogate Key (pk)	CusID (nk)	Customer Name	Address	City	State
1	12345|NA	Dominic Sellitto	123 Fake St.	Buffalo	NY

Becomes:

Surrogate Key (pk)	CusID (nk)	Customer Name	Old Address	Old City	Old State	New Addr	New City	New State
1	12345|NA	Dominic Sellitto	123 Fake St.	Buffalo	NY	123 Cool St.	Lancaster	NY

Reference: Kimball, Ralph, and Margy Ross. *The Data Warehouse Toolkit : The Definitive Guide to Dimensional Modeling*, John Wiley & Sons, Incorporated, 2013,
https://www.nuwavesolutions.com/slowly-changing-dimensions/

Changing dimensional attributes

Type 4, 5, 6,and 7 changing dimensions become increasingly complex and less used. We'll talk more about these later in the semester.

Reference: Kimball, Ralph, and Margy Ross. *The Data Warehouse Toolkit : The Definitive Guide to Dimensional Modeling*, John Wiley & Sons, Incorporated, 2013,
https://www.nuwavesolutions.com/slowly-changing-dimensions/

Common Dimensions

There are several common dimensions you'll repeatedly see across warehouse design initiatives:

- Date/time dimension
- Product
- Customer

Reference: Kimball, Ralph, and Margy Ross. *The Data Warehouse Toolkit : The Definitive Guide to Dimensional Modeling*, John Wiley & Sons, Incorporated, 2013,
https://www.nuwavesolutions.com/slowly-changing-dimensions/

Demo

Let’s take a look at some common dimensions…

Reference: Kimball, Ralph, and Margy Ross. *The Data Warehouse Toolkit : The Definitive Guide to Dimensional Modeling*, John Wiley & Sons, Incorporated, 2013,
https://www.nuwavesolutions.com/slowly-changing-dimensions/