Data Warehousing

Dimensional Modeling

2022/2023

Ana Lucas

Dimensional Modeling

1. Parts of this presentation were taken from the backing material of the book

   Hoffer, J. A., Ramesh,V., Topi, H. (2019). *Modern Database Management*, (13th ed.). Pearson Higher Education.

2. And from the book:

   Kimball, R. & Ross, M. (2013). *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling* (3rd ed.). Indianapolis: John Wiley & Sons, Inc.

Architecture Choices (1/2)

Top Down

(Inmon)

- Enterprise data warehouse (3NF)
- Higher integration levels
- Logically centralized
- Larger project scope

Bottom Up

(Kimball, Dimensional Model)

- Independent data marts
- Some integration levels
- Logically decentralized
- Smaller project scope

ISEG EXECUTIVE EDUCATION

Architecture Choices (2/2)

Hybrid approach (2000) Mostly Top-Down

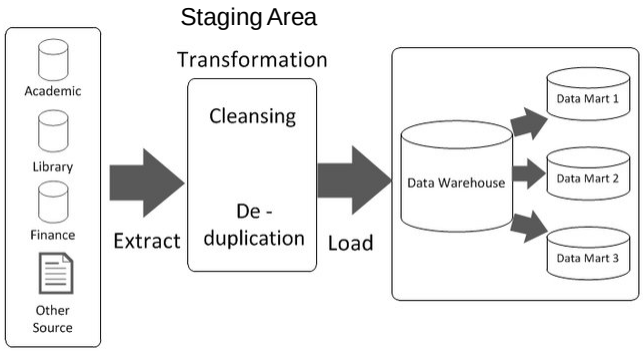
(Dan Linstedt, Data Vault 2.0)

Uses 3rd normal form (3NF) with some archetypes for the DW and Dimensional Model for the Data Marts.

"The Data Vault Model is a detail oriented, historical tracking and uniquely linked set of normalized tables that support one or more functional areas of business. It is a hybrid approach encompassing the best of breed between 3rd normal form (3NF) and star schema. The design is flexible, scalable, consistent and adaptable to the needs of the enterprise" - Dan Linstedt

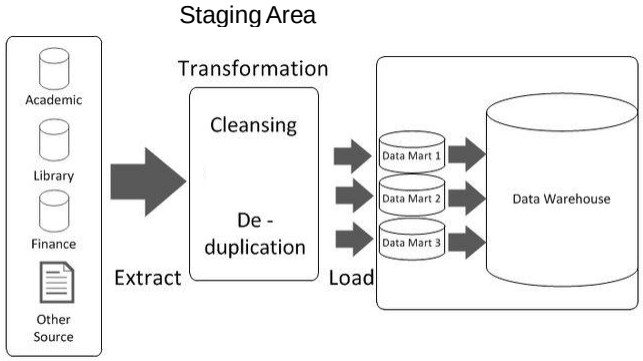
ISEG EXECUTIVE EDUCATION

The Inmon Data Warehouse



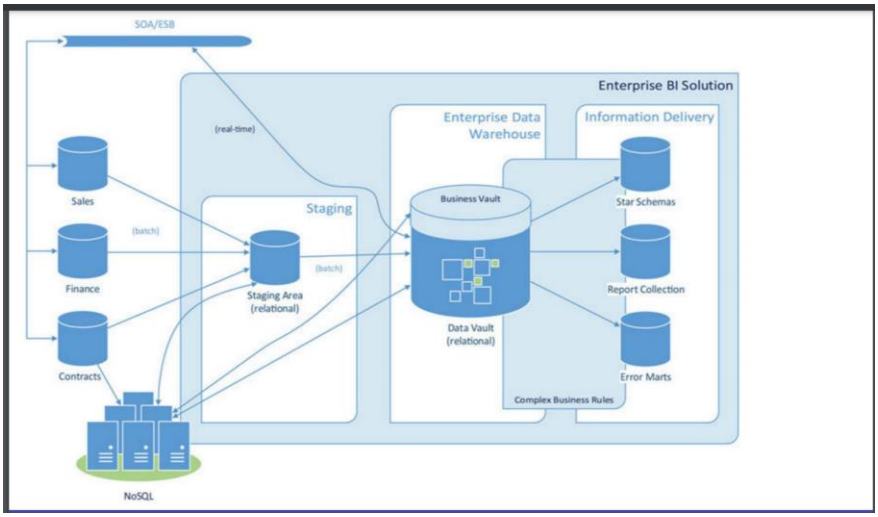
https://www.researchgate.net/publication/328434296_A_Holistic_View_of_Data_Warehousing_in_Education/figures?lo=1

The Kimball Data Warehouse



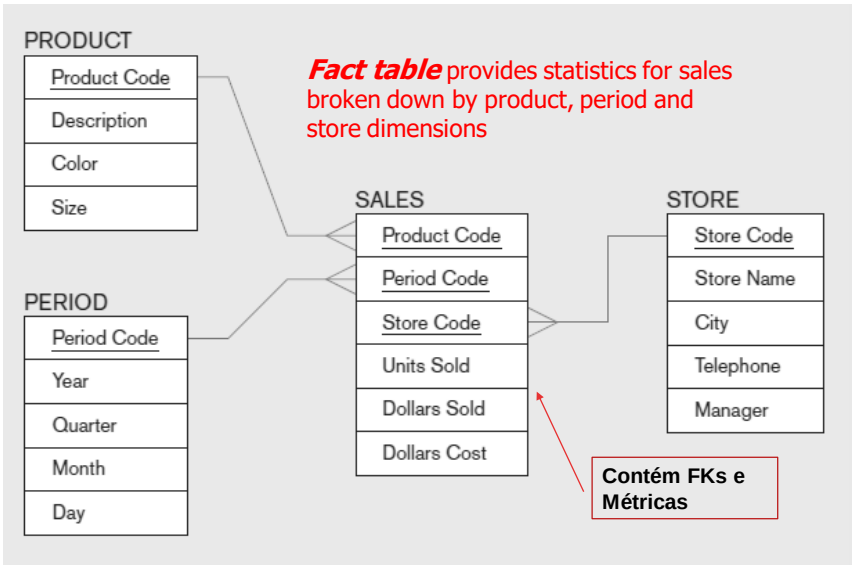
https://www.researchgate.net/publication/328434296_A_Holistic_View_of_Data_Warehousing_in_Education/figures?lo=1

The Data Vault Data Warehouse

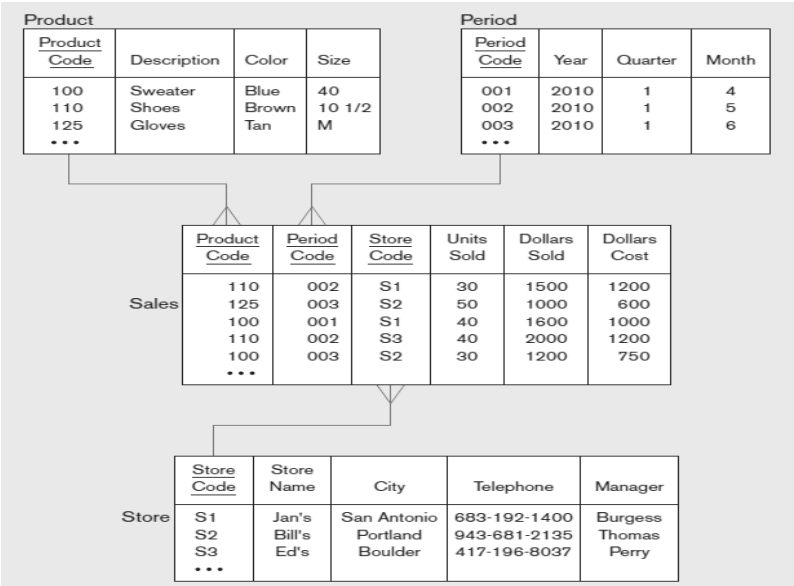


http://albertadataarchitecture.org/data/documents/Data-Integration-and-Warehousing-using-the-Data-Vault.pdf

Dimensional Model - Star schema example



Star schema with sample data



Product

Product Code	Description	Color	Size
100	Sweater	Blue	40
110	Shoes	Brown	10 1/2
125	Gloves	Tan	M
• • •			

Period

Period Code	Year	Quarter	Month
001	2010	1	4
002	2010	1	5
003	2010	1	6
• • •			

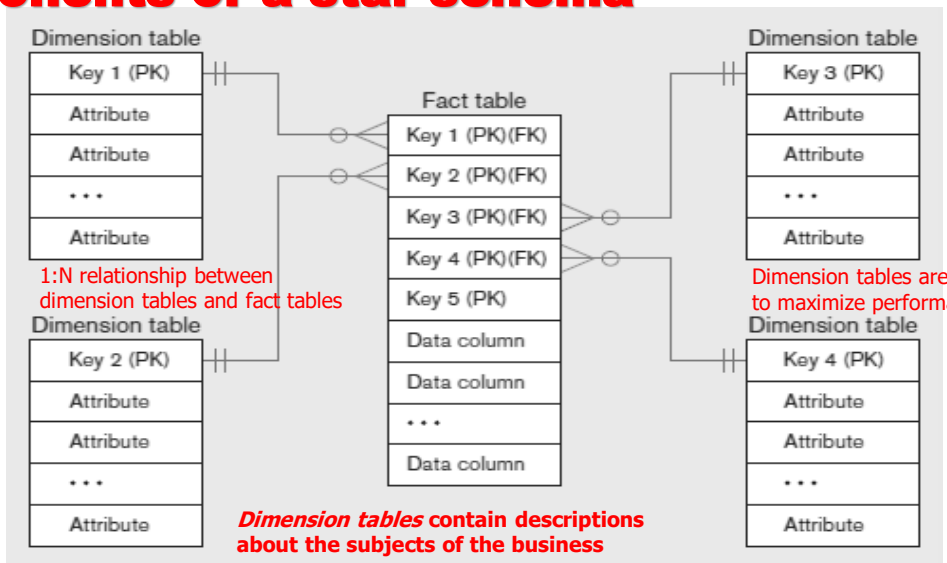
Sales

Product Code	Period Code	Store Code	Units Sold	Dollars Sold	Dollars Cost
110	002	S1	30	1500	1200
125	003	S2	50	1000	600
100	001	S1	40	1600	1000
110	002	S3	40	2000	1200
100	003	S2	30	1200	750
• • •					

Store

Store Code	Store Name	City	Telephone	Manager
S1	Jan's	San Antonio	683-192-1400	Burgess
S2	Bill's	Portland	943-681-2135	Thomas
S3	Ed's	Boulder	417-196-8037	Perry
• • •				

Components of a star schema



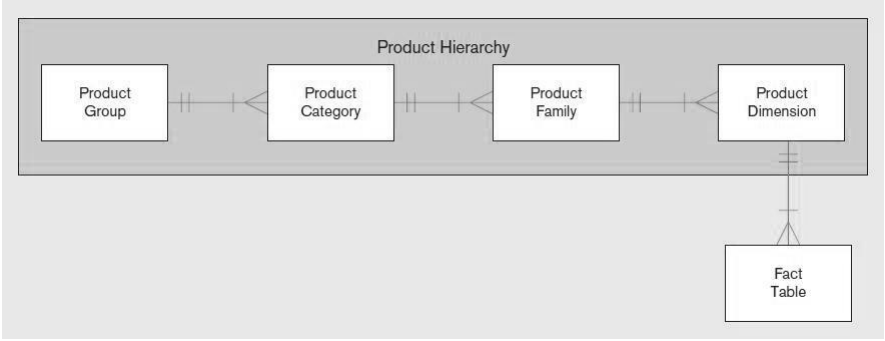
Excellent for ad-hoc queries, but bad for online transaction processing

Primary Keys of the Fact Tables

the primary key of the fact table is almost always defined as a subset of the foreign keys supplied by the dimensions

https://www.kimballgroup.com/2006/07/design-tip-81-fact-table-surrogate-key/

Dimension Product in 3NF - OLTP (Snowfake Schema – Do not use it)



Dimension Product with denormalized 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

Kimball, Ross (2013)

Overall Data Architecture for the Warehouse

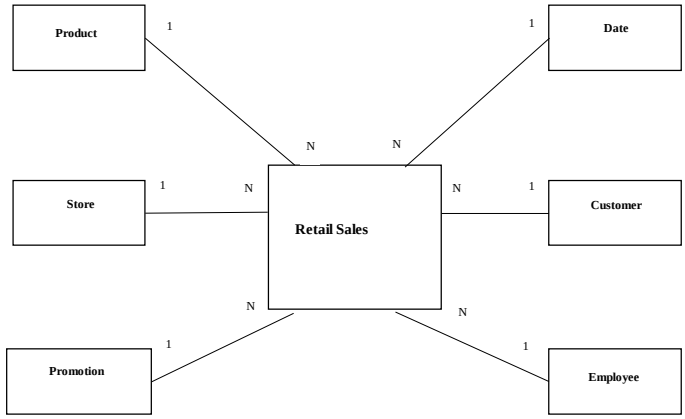
Sample enterprise data warehouse Bus Matrix for a retailer

	COMMON DIMENSIONS						
BUSINESS PROCESSES	Date	Product	Warehouse	Store	Promotion	Customer	Employee
Issue Purchase Orders	X	X	X				
Receive Warehouse Deliveries	X	X	X				X
Warehouse Inventory	X	X	X				
Receive Store Deliveries	X	X	X	X			X
Store Inventory	X	X		X			
Retail Sales	X	X		X	X	X	X
Retail Sales Forecast	X	X		X			
Retail Promotion Tracking	X	X		X	X		
Customer Returns	X	X		X	X	X	X
Returns to Vendor	X	X		X			X
Frequent Shopper Sign-Ups	X			X		X	X

Kimball, Ross (2013)

Overall Data Architecture for the Warehouse

Example: Business Process Retail Sales



Overall Data Architecture for the Warehouse

Detailed implementation of the bus matrix rows for procurement processes

Business Processes	Atomic Granularity	Metrics	Date	Product	Vendor	Contract Terms	Employee	Warehouse	Carrier
Purchase Requisitions	1 row per requisition line	Requisition Quantity & Dollars	X	X	X	X	X		
Purchase Orders	1 row per PO line	PO Quantity & Dollars	X	X	X	X	X	X	X
Shipping Notifications	1 row per shipping notice line	Shipped Quantity	X	X	X		X	X	X
Warehouse Receipts	1 row per receipt line	Received Quantity	X	X	X		X	X	X
Vendor Invoices	1 row per invoice line	Invoice Quantity & Dollars	X	X	X	X	X	X	
Vendor Payments	1 row per payment	Invoice, Discount & Net Payment Dollars	X	X	X	X		X	

Kimball, Ross (2013)

Overall Data Architecture for the Warehouse
Data Warehouse Bus Matrix

- **The bus matrix rows shouldn't correspond to the boxes on a corporate organization chart** representing functional groups.

- **The bus matrix shouldn't resemble a laundry list of requested reports. A single business process supports numerous analyses; the matrix row should reference the business process, not the derivative reports or analytics.**

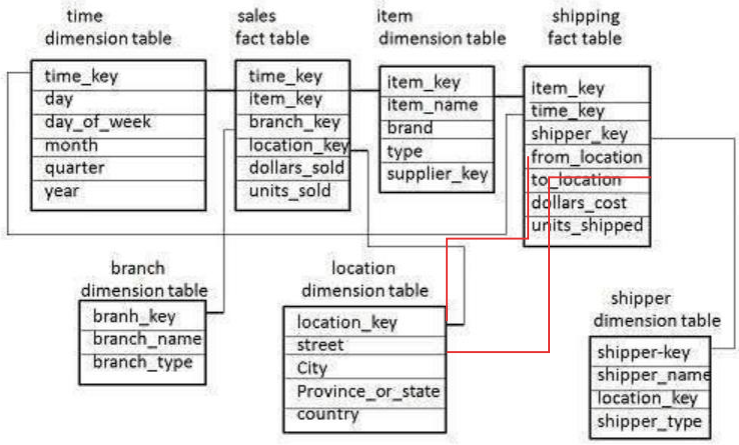
Kimball, Ross (2013)

Overall Data Architecture for the Warehouse
Relationship Between Business Processes and Fact Tables

A BP gives rise to one or more fact tables, usually one

Overall Data Architecture for the Warehouse

Fact Constellation with Conformed Dimensions



https://www.researchgate.net/publication/340546587_COMPARATIVE_STUDY_ON_DATA_WAREHOUSE_TABLES_AND_SCHEMA-AN_OVERVIEW/figures?lo=1

Overall Data Architecture for the Warehouse

Data Warehouse Bus Matrix

Creating the **data warehouse bus matrix** is one of the most important up-front **deliverables of a data warehouse implementation**. It is a hybrid resource that is **part technical design tool**, **part project management tool**, and **part communication tool**

Kimball, Ross (2013)

Four-Step Dimensional Design Process

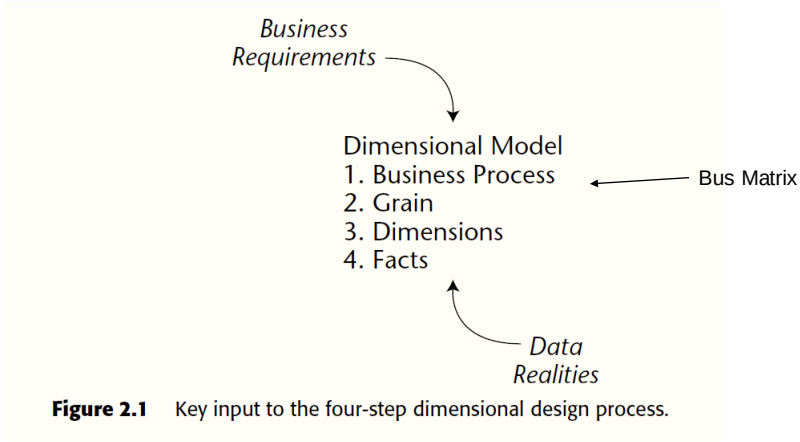


Figure 2.1 Key input to the four-step dimensional design process.

Kimball, Ross (2013)

Step 1: Select the Business Process

A business process is **a low-level activity performed by an organization**, such as taking orders, invoicing, receiving payments, handling service calls, registering students, performing a medical procedure, or processing claims.

Business Processes Characteristics

- Business processes are typically **supported by an operational system**, such as the billing or purchasing system
- Business processes **generate or capture key performance metrics**

Kimball, Ross (2013)

Step 2: Declare the Grain

Declaring the grain means **specifying exactly what an individual fact table row represents**

Example grain declarations include:

- One row per scan of an individual product on a customer's sales transaction
- One row per individual boarding pass scanned at an airport gate
- One row per purchase order (PO) line
- One row per daily snapshot of the inventory levels for each item in a warehouse

Kimball, Ross (2013)

Grain of the Fact Table

Granularity of Fact Table – what level of detail do you want?

Transactional grain – finest level
Aggregated grain – more summarized
Finer grains → better ***market basket analysis*** capability
Finer grain →more dimension tables, more rows in fact table

Grain Example

Sales fact table grain

Coarse: customer postal codes (1,000), product category (100), store (200), week (52)

Fine: individual customer (200,000), individual product (2,000), store (200), day (365)

Impact

Higher storage requirements for fine grain

More reporting flexibility for fine grain

Step 3: Identify the Dimensions

Dimensions fall out of the question, **"How do business people describe the data resulting from the business process measurement events?"**

Examples:

- Date
- Product
- Customer
- Employee
- Facility
- Supplier

Kimball, Ross (2013)

Dimensions

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
...

The dimension tables contain the **textual context associated with a business process measurement event**.

Dimension attributes serve as the primary source of query constraints, groupings and report labels

Kimball, Ross (2013)

Date Dimensions
Time-of-Day as a Dimension

Date Dimension
Date Key (PK)
Date
Full Date Description
Day of Week
Day Number in Calendar Month
Day Number in Calendar Year
Day Number in Fiscal Month
Day Number in Fiscal Year
Last Day in Month Indicator
Calendar Week Ending Date
Calendar Week Number in Year
Calendar Month Name
Calendar Month Number in Year
Calendar Year-Month (YYYY-MM)
Calendar Quarter
Calendar Year-Quarter
Calendar Year
Fiscal Week
Fiscal Week Number in Year
Fiscal Month
Fiscal Month Number in Year
Fiscal Year-Month
Fiscal Quarter
Fiscal Year-Quarter
Fiscal Half Year
Fiscal Year
Holiday Indicator
Weekday Indicator
SQL Date Stamp
...

Although date and time are blended in an operational date/time stamp, **time-of-day is typically separated from the date dimension to avoid a row count explosion in the date dimension.**

Because the date dimension is likely the most frequently constrained dimension in a schema, it should be kept as small and manageable as possible.

Kimball, Ross (2013)

Promotion Dimension

Promotion Dimension
Promotion Key (PK)
Promotion Code
Promotion Name
Price Reduction Type
Promotion Media Type
Ad Type
Display Type
Coupon Type
Ad Media Name
Display Provider
Promotion Cost
Promotion Begin Date
Promotion End Date
...

The various promotion dimensions are usually highly correlated.

Another way of modeling the promotions is **to separate the four causal mechanisms (price reductions, ads, displays, and coupons) into separate dimensions** rather than combining them into one dimension.

Kimball, Ross (2013)

Conformed Dimensions

Conformed dimensions have consistent dimension keys, consistent attribute column names, consistent attribute definitions, and consistent attribute values (which translates into consistent report labels and groupings).

Kimball, Ross (2013)

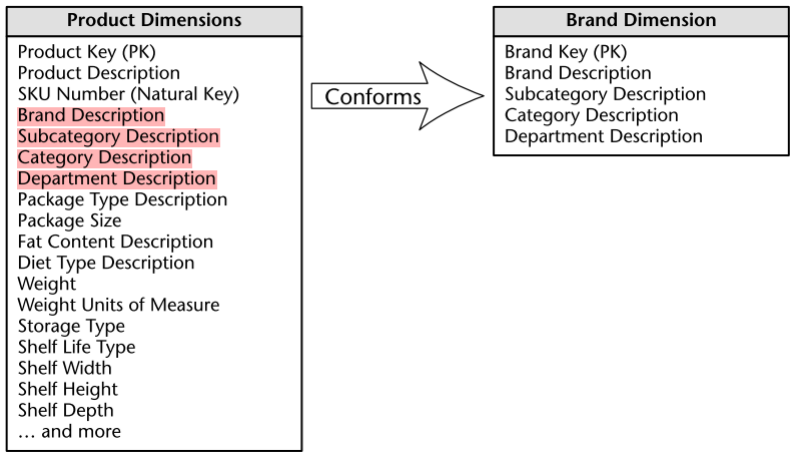
Conformed Dimensions

One of the key objectives of the **data governance function** is to reach agreement on **data definitions, labels, and domain values so that everyone is speaking the same language.**

Defining a **conformed dimension** requires **organizational consensus**

Kimball, Ross (2013)

Conforming roll-up dimension subsets



Roll-up dimensions conform to the base-level atomic dimension if they are a strict subset of that atomic dimension

Kimball, Ross (2013)

Alternatives for identifying roll-up dimensions on the bus matrix

	Date
Issue Purchase Orders	X
Receive Deliveries	X
Inventory	X
Retail Sales	X
Retail Sales Forecast	X Month

OR

	Date	
	Day	Month
Issue Purchase Orders	X	
Receive Deliveries	X	
Inventory	X	
Retail Sales	X	
Retail Sales Forecast		X

Kimball, Ross (2013)

Degenerate Dimensions

Translates into

Retail Sales Facts
Date Key (FK)
Product Key (FK)
Store Key (FK)
Promotion Key (FK)
Customer Key (FK)
Clerk Key (FK)
Transaction #
Sales Dollars
Sales Units

Degenerate Dimension

Kimball, Ross (2013)

Degenerate Dimensions

Operational control numbers such as order numbers, invoice numbers and POS transaction numbers usually give rise to empty dimensions and are represented as degenerate dimensions (that is, dimension keys without corresponding dimension tables) in fact tables where the grain of the table is a line item in the document

Kimball, Ross (2013)

Surrogate Dimension Keys

Dimension table keys should be ***surrogate*** (non-intelligent and non-business related), because:

- **Integrate multiple source systems, with different natural keys**
- **Surrogate keys are simpler and shorter and the Joins with the Fact Tables are more performant**

Slowly Changing Dimensions (SCD)

How to maintain knowledge of the past
Kimball's approaches:

Type 1: just replace old data with new (lose historical data)

Type 2: create a new dimension table row each time one dimension attribute changes, with all dimension characteristics at the time of change. **Most common approach**

Type 3: for each changing attribute, create a current value field and several old-valued fields (multivalued)

Example of Type 2 SCD Product dimension table

Original row in Product dimension:

Product Key	SKU (NK)	Product Description	Department Name	...	Row Effective Date	Row Expiration Date	Current Row Indicator
12345	ABC922-Z	IntelliKidz	Education	...	2012-01-01	9999-12-31	Current

Rows in Product dimension following department reassignment:

Product Key	SKU (NK)	Product Description	Department Name	...	Row Effective Date	Row Expiration Date	Current Row Indicator
12345	ABC922-Z	IntelliKidz	Education	...	2012-01-01	2013-01-31	Expired
25984	ABC922-Z	IntelliKidz	Strategy	...	2013-02-01	9999-12-31	Current

Kimball, Ross (2013)

Slowly Changing Dimensions – Type 2

When a new row is created for a dimension member, a **new primary surrogate key is assigned** and used as a foreign key in all fact tables from the moment of the update until a subsequent change creates a new dimension key and updated dimension row

A minimum of **three additional columns** should be added to the dimension row with type 2 changes: **1) row effective date or date/time stamp; 2) row expiration date or date/time stamp; and 3) current row indicator**

Kimball, Ross (2013)

Normalizing Dimension Tables

Hierarchies

Sometimes a dimension forms a natural, **fixed depth hierarchy**

Design options

- Include all information for each level in a single denormalized table – **the good option**
- Normalize the dimension into a nested set of 1:M table relationships – **do not use it**

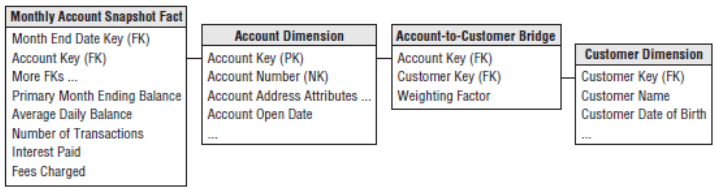
Multivalued Dimensions

Facts qualified by a set of values for the same dimension

Normalization involves creating a table for an associative entity between dimensions – **Bridge Table**

Multivalued dimension

Account-to-customer bridge table with weighting factor



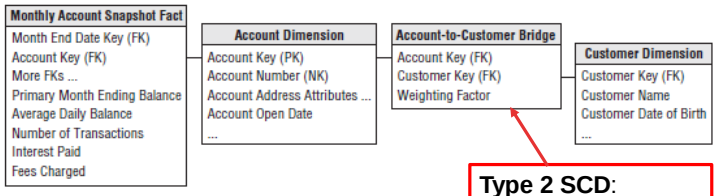
How to solve the problem of changing the customers of an account?

Kimball, Ross (2013)

Multivalued dimension

Account-to-customer bridge table with weighting factor

How to solve the problem of changing the customers of an account?

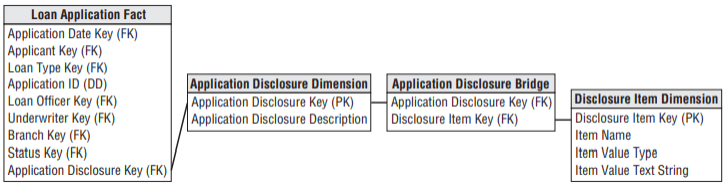


Type 2 SCD:
1. Row effective date or date/time stamp;
2. Row expiration date or date/time stamp;
3. Current row indicator

Kimball, Ross (2013)

Multivalued dimension

Bridge table design for multiple disclosure Items (Clásulas de um contrato)

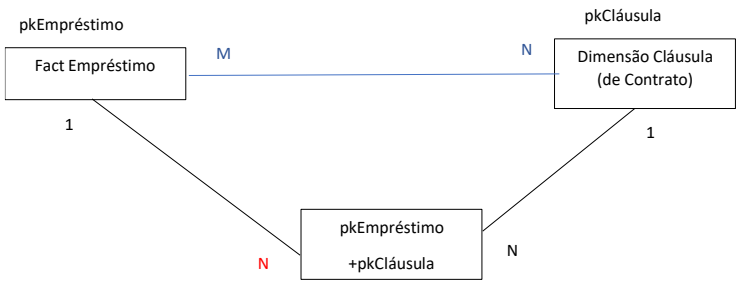


A disclosure statement (cláusula de um contrato) is a document explaining the rules of a financial transaction in plain, nontechnical language.

Kimball, Ross (2013)

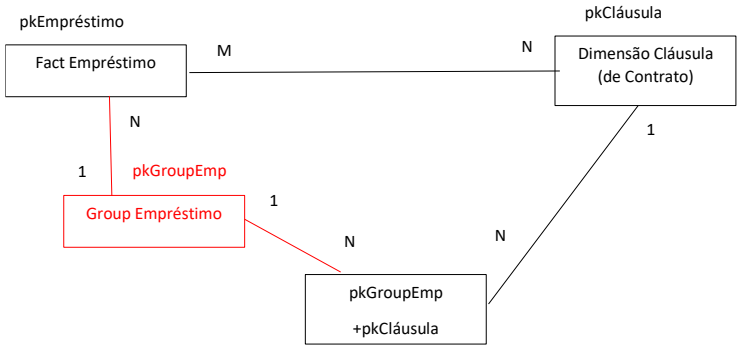
Multivalued dimension

Bridge table design for multiple disclosure Items (Clásulas de um contrato)



Multivalued dimension

Bridge table design for multiple disclosure Items (Cláusulas de um contrato)



Ver Design Tip #142 Building Bridges - Kimball Group

Step 4: Identify the Facts

Facts are determined by answering the question, **"What is the process measuring?"**

- **All candidate facts in a design must be true to the grain defined in step 2**
- **Facts that clearly belong to a different grain must be in a separate fact table**
- Typical facts are **numeric additive figures**, such as quantity ordered or dollar cost amount

Kimball, Ross (2013)

Fact Table
Measure Aggregation Properties

Additive

Summarized by addition across all dimensions

Common measures such as sales, cost, and profit

Semi-Additive

Summarized by addition in some but not all dimensions such as time

Periodic measurements such as bank account balances and inventory levels

Non-Additive

Cannot be summarized by addition through any dimension

Historical facts such as unit price for a sale

The most useful facts in a fact table are numeric and additive

Fact Table
Types of Fact Tables

- **Transaction Fact Tables**

Date	Amount
2/1/2014	$3,000
2/4/2014	($200)
2/9/2014	$1,000

http://valuabledata.blogspot.pt/2014/03/transaction-fact-tables.html

- **Periodic Snapshot Fact Tables**

Date	Amount
2/1/2014	$3,000
2/4/2014	$2,800
2/9/2014	$3,800

http://valuabledata.blogspot.pt/2014/03/periodic-snapshot-fact-tables.html

Fact Table
Types of Fact Tables

- **Accumulating Snapshot Fact Tables**

Date Ticket Opened	Date Ticket Assigned	Date Solution Provided To Customer	Date Customer Accepted Solution	Date Ticket Closed	Ticket Number
3/1/2014	3/3/2014	3/4/2014	3/5/2014	3/6/2014	10012

http://valuabledata.blogspot.pt/2014/03/accumulating-snapshot-fact-tables.html

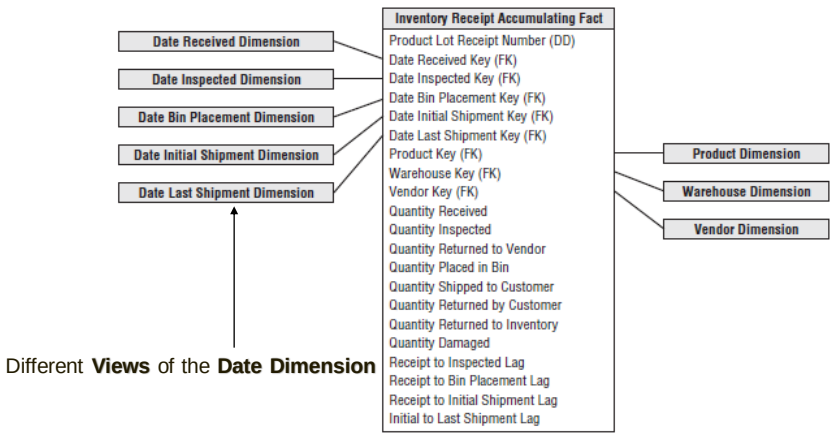
FACT_CLAIM_PROCESSING	
P * CLAIM_KEY	NUMBER
P * CUSTOMER_KEY	NUMBER
P * POLICY_KEY	NUMBER
* CLAIM_DATE	DATE
INVESTIGATION_DATE	DATE
REVIEW_DATE	DATE
DESCISION_DATE	DATE
PAYMENT_DATE	DATE

OR

FACT_CLAIM_PROCESSING	
P * CLAIM_KEY	NUMBER
P * CUSTOMER_KEY	NUMBER
P * POLICY_KEY	NUMBER
* CLAIM_DATE	DATE
INVESTIGATION_DATE	DATE
DAYS_TO_INVESTIGATION	NUMBER
REVIEW_DATE	DATE
DAYS_TO_REVIEW	NUMBER
DESCISION_DATE	DATE
DAYS_TO_DESCISION	NUMBER
PAYMENT_DATE	DATE
DAYS_TO_PAYMENT	NUMBER

http://www.nuwavesolutions.com/accumulating-snapshot-fact-tables/

Fact Table
Star Schema of Accumulating Snapshot Fact Table



Different **Views** of the **Date Dimension**

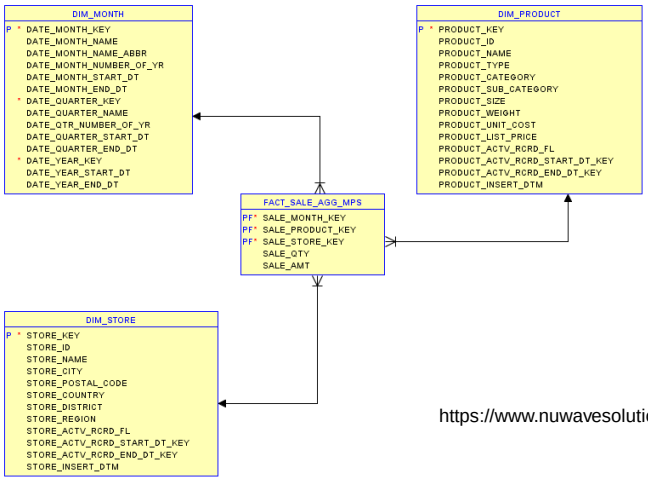
Fact Table
Fact Table Type Comparison

	Transaction	Periodic Snapshot	Accumulating Snapshot
Periodicity	Discrete transaction point in time	Recurring snapshots at regular, predictable intervals	Indeterminate time span for evolving pipeline/workflow
Grain	1 row per transaction or transaction line	1 row per snapshot period plus other dimensions	1 row per pipeline occurrence
Date dimension(s)	Transaction date	Snapshot date	Multiple dates for pipeline's key milestones
Facts	Transaction performance	Cumulative performance for time interval	Performance for pipeline occurrence
Fact table updates	No updates, unless error correction	No updates, unless error correction	Updated whenever pipeline activity occurs

Kimball, Ross (2013)

Fact Table
Aggregate Fact Tables

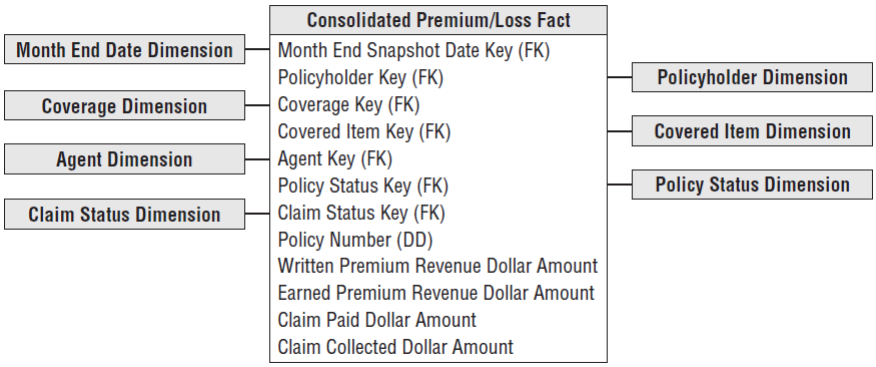
Aggregate fact tables **are simple numeric rollups of atomic fact table data built solely to accelerate query performance**



https://www.nuwavesolutions.com/aggregates/

Fact Table
Consolidated Fact Tables

It is often convenient to combine facts from multiple processes together into a single *consolidated fact table* if they can be expressed at the same grain



Policy/claim consolidated fact table

Kimball, Ross (2013)

Fact Table
Variations of the Star Schema

Factless Facts Tables

No nonkey data, but foreign keys for associated dimensions

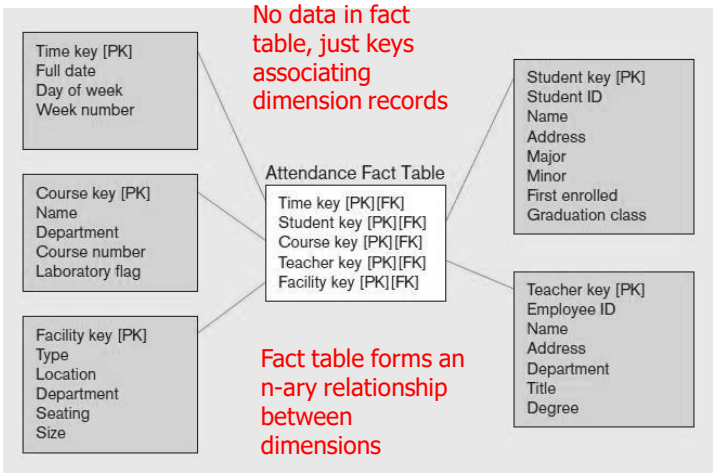
Used for:

- Tracking events
- Inventory coverage

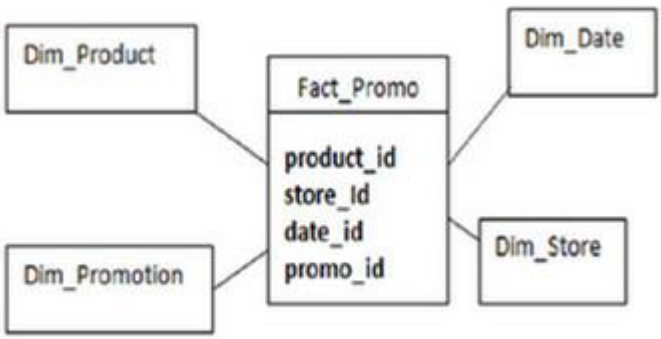
Snowflake Schema

A refinement of star schema where some dimensional hierarchy is further splitting (normalized) into a set of smaller dimension tables, forming a shape similar to snowflake

Fact Table
Factless Fact Table

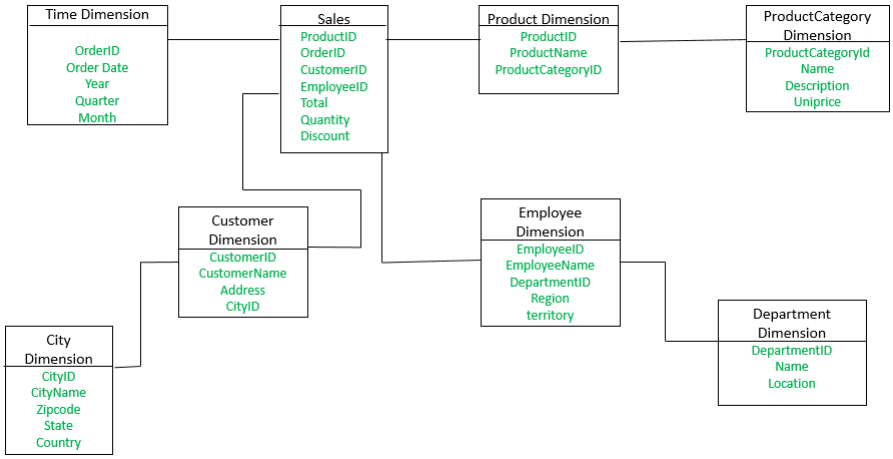


Fact Table
Factless Fact Table



https://informaticareference.files.wordpress.com/2012/09/2.jpg

Snowflake schema
Just Say NO!



https://dev.to/pedrojmfidalgopt/star-schema-vs-snowflake-schema-and-why-you-should-care-40fh

57

10 Essential Rules for Dimensional Modeling

Rule #1: Load detailed atomic data into dimensional structures.

Rule #2: Structure dimensional models around business processes.

Rule #3: Ensure that every fact table has an associated date dimension table.

Rule #4: Ensure that all facts in a single fact table are at the same grain or level of detail.

Rule #5: Resolve many-to-many relationships in fact tables – Bridge Tables.

Rule #6: Resolve many-to-one relationships in dimension tables. Dimension denormalization is the name of the game in dimensional modeling.

Rule #7: Store report labels and filter domain values in dimension tables.

Rule #8: Make certain that dimension tables use a surrogate key.

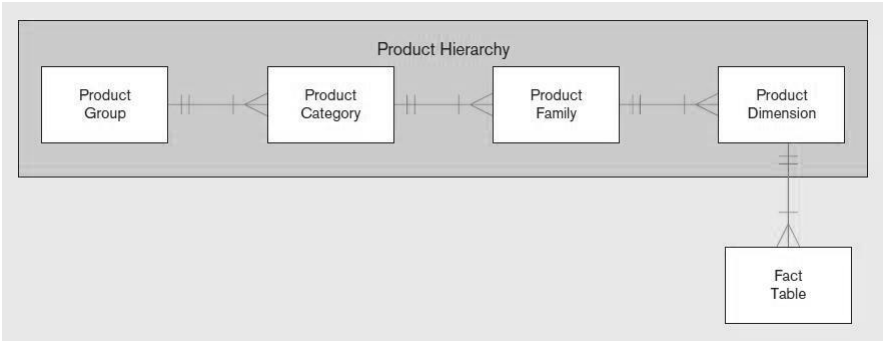
Rule #9: Create conformed dimensions to integrate data across the enterprise.

Rule #10: Continuously balance requirements and realities to deliver a DW/BI solution that's accepted by business users and that supports their decision-making.

https://www.kimballgroup.com/2009/05/the-10-essential-rules-of-dimensional-modeling/

58

Rule #6
(Snowfake Schema – Do not use it)



60

Rule #6 -Dimension Product with denormalized 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

Kimball, Ross (2013)

ana.lucas@iseg.ulisboa.pt

www.isegexecutive.education