IIT School of Applied Technology

ILLINOIS INSTITUTE OF TECHNOLOGY

information technology & management

526 Data Warehousing

January 27, 2016

Week 2 Presentation

Week 02 Topic: Dimensional Modeling Fundamental Concepts

- We will discuss
  - Business-driven goals of DW/BI
  - Dimensional modeling core concepts and vocabulary
  - Kimball DW/BI architecture
  - Alternative DW/BI architectures

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

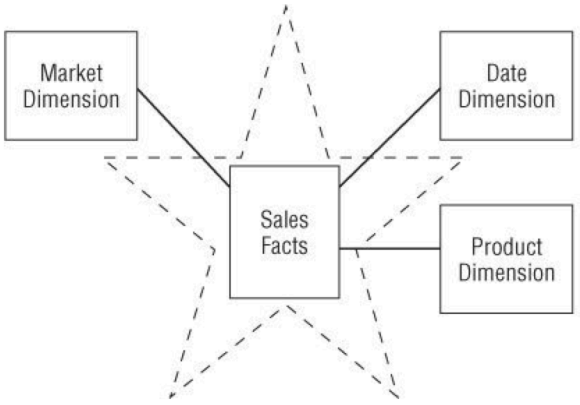
- Make information accessible and understandable to the business users
- Present information consistently in a timely manner
- Ensure DW/BI deliverables accepted by business community to support their decision making

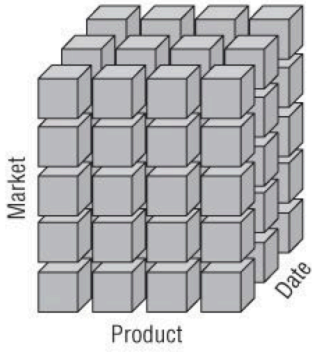
DW/BI Deliverables

- Dimensional models
  - Relational star schemas
    - Fact tables for measurements
    - Dimension tables for descriptors
  - Multidimensional online analytical processing (OLAP) cubes
  - Goals:
    - Easy to use
    - Fast queries
- Business intelligence applications
  - Reporting tools, analytic tools, etc.

Star Schema Versus 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.



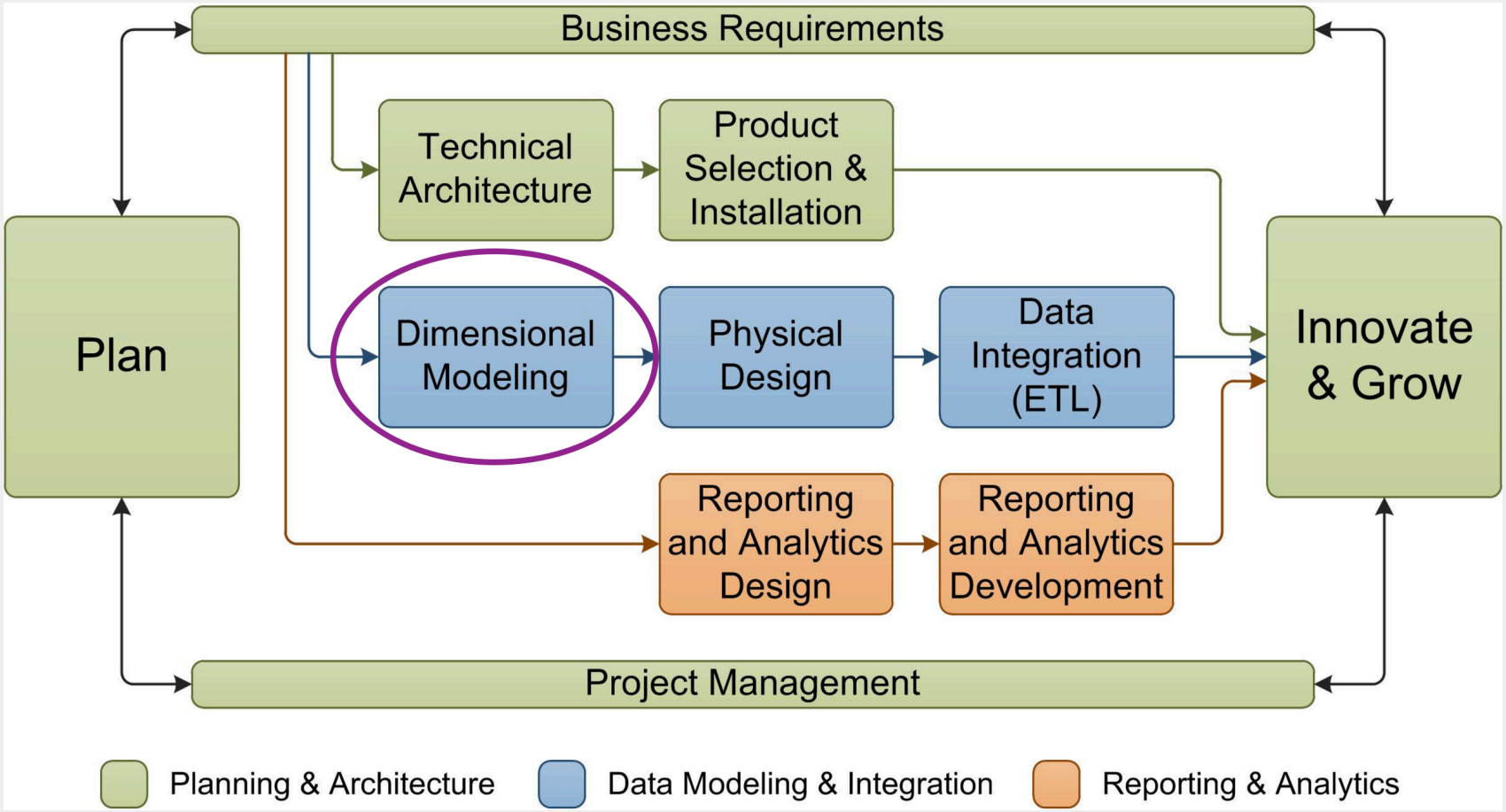


Star Schema Versus OLAP Cubes

- 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

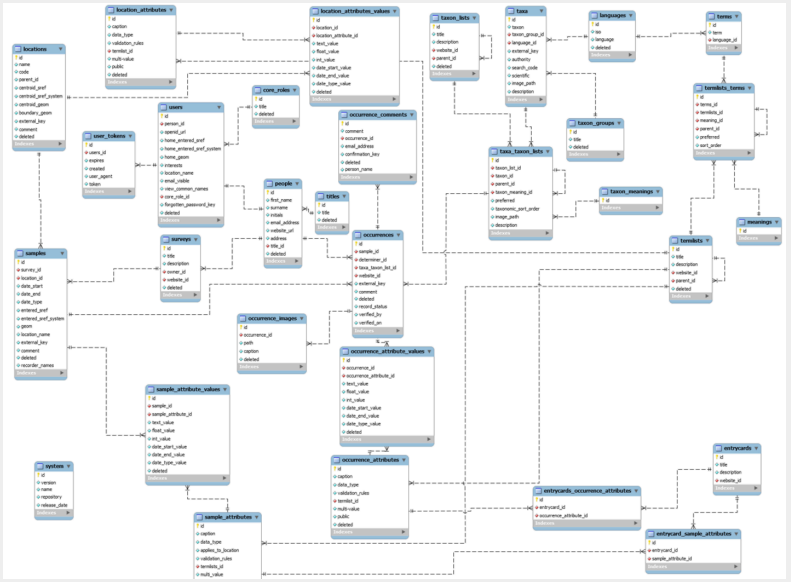
- Kimball focuses on the star schema rather than the OLAP cubes
- Star schema usually has 15 dimensions
- OLAP usually has 8-10 dimensions

Kimball Lifecycle Approach



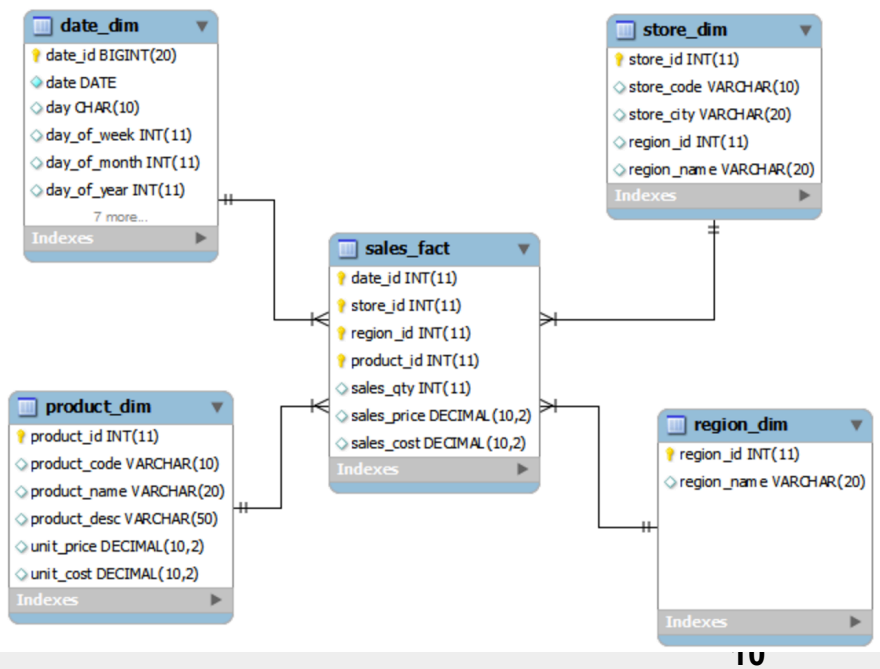
Dimensional Modeling Fundamental Concepts: Make it Simple

- 3NF: Immensely useful in operational processing
- However, it's too complicated for users to use. It looks like a subway map
- Unpredictable queries cause performance problems



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 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, Time
  - Performance: sales volume, profit
  - → Make a simple modeling
- Simple modeling is important
  - Resist temptation to over-engineer

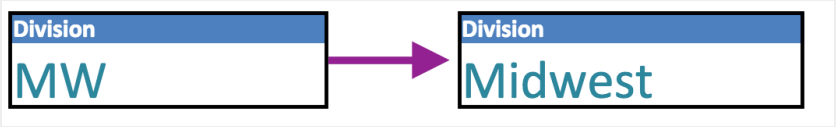
Dimensions – Physical Table Elements

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

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
...

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

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**
(Snowflaking: normalizing dimensions)

- Operational natural business key as attribute, not primary key

ITMD - 526

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

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

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	2004-12-22	NULL

Fact Tables

- Contains the measurements or facts about a business process
- Are thin and deep
- Usually is:
  - Business transaction
  - Business Event
- The grain of a fact table is the level of the data recorded

Fact Table – Grain

- The grain of a fact table: what a fact table record exactly represents
- Examples:
  - an individual line item on a customer's retail sales ticket as measured by a scanner device
  - an individual transaction against an insurance policy
  - a line item on a bill received from a doctor
  - an individual boarding pass used by someone making an airplane flight

Fact Tables – Physical Table Elements

- Contains the following elements
  - Primary key – surrogate key (optional)
  - Foreign keys to dimensions
  - Degenerate dimensions
    - Transaction indicators or flags
  - Measure or metrics
    - Transaction amounts

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

Fact Table - Types of Measures

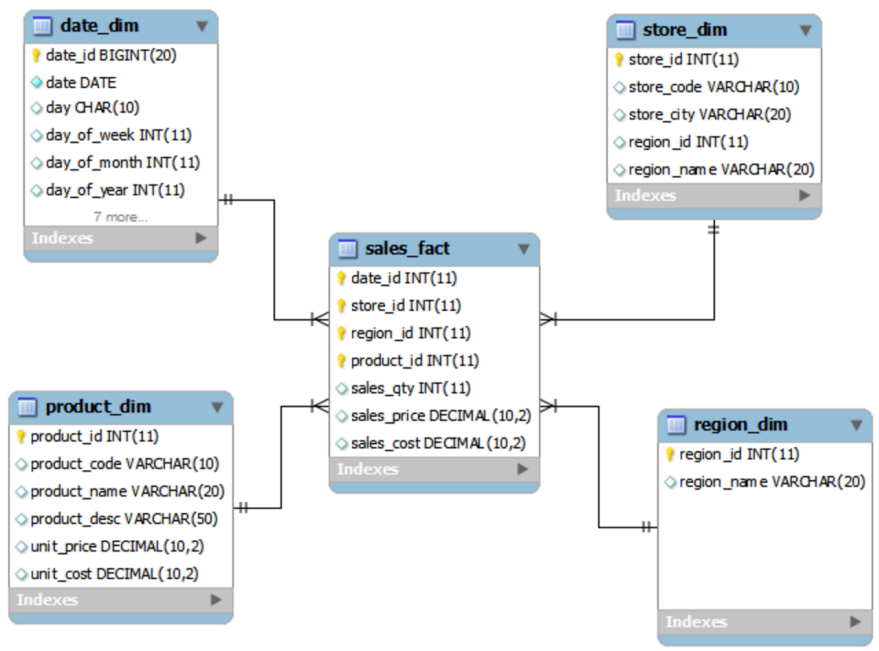
- Additive facts - Measures that can be added across any dimensions.
  - Amounts
- Non additive facts - Measures that cannot be added across any dimension.
  - Rates
- Semi additive facts - Measures that can be added across some dimensions.
  - Balances

Fact Tables - Types of Tables

- **Transactional** - A transactional table is the most basic and fundamental. The grain associated with a transactional fact table is usually specified as "one row per line in a transaction".
- **Periodic snapshots** - The periodic snapshot, as the name implies, takes a "picture of the moment", where the moment could be any defined period of time.
- **Accumulating snapshots** - This type of fact table is used to show the activity of a process that has a well-defined beginning and end, e.g., the processing of an order. An order moves through specific steps until it is fully processed. As steps towards fulfilling the order are completed, the associated row in the fact table is updated.

Dimensional Star Schema

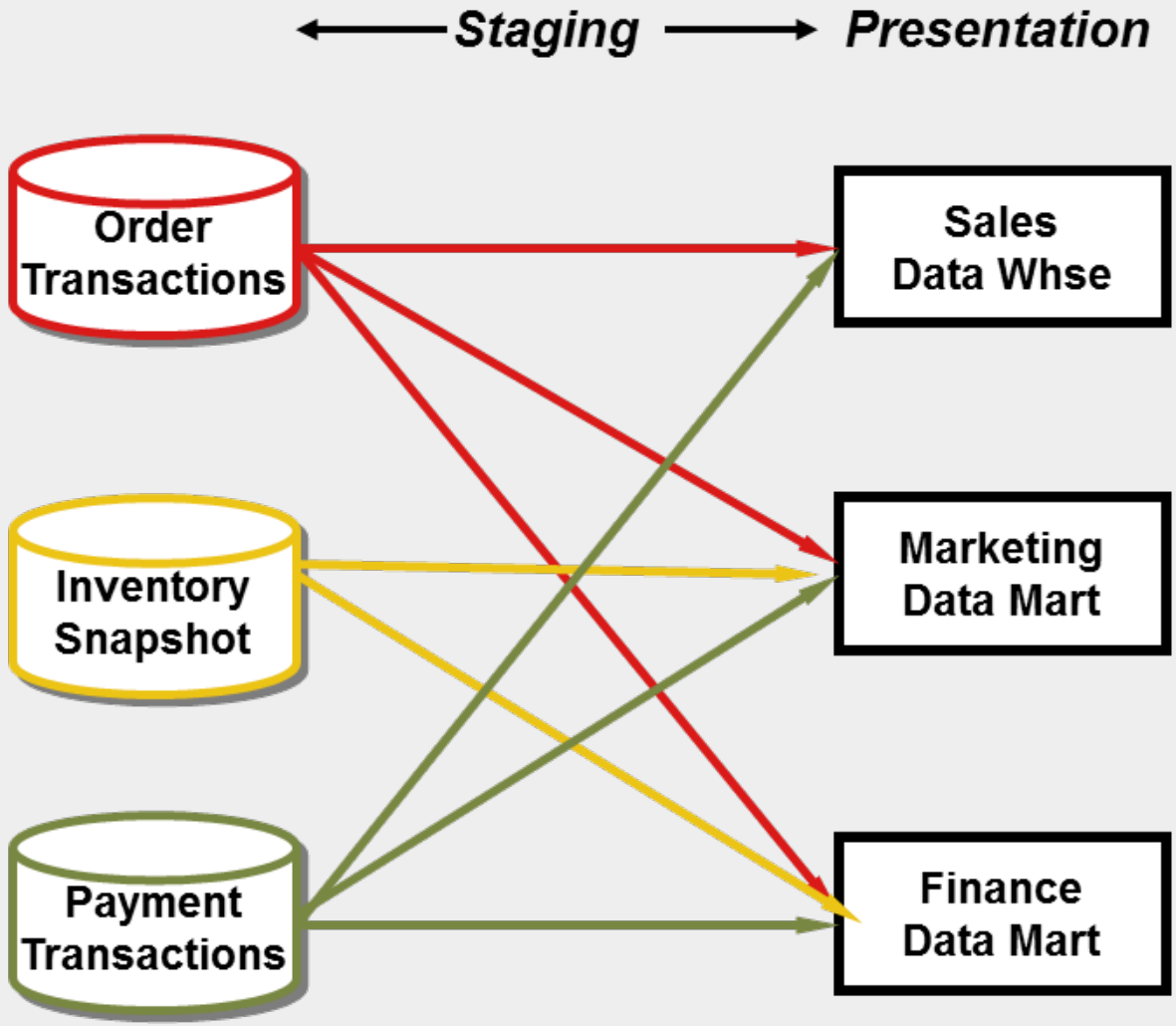
- Atomic fact table per business process event (grain)
- Benefits:
  - Easy to understand
  - Better query performance from fewer joins
  - Resilience to change via extensibility (adding attributes, dimensions)



Dimensional Star Schema: Start Simple, Finish Simple

"A model that starts simple has a chance of remaining simple at the end of the design. A model that starts complicated surely will be overly complicated at the end, resulting in slow query performance and business user rejection." - Kimball

Independent Data Mart Architecture

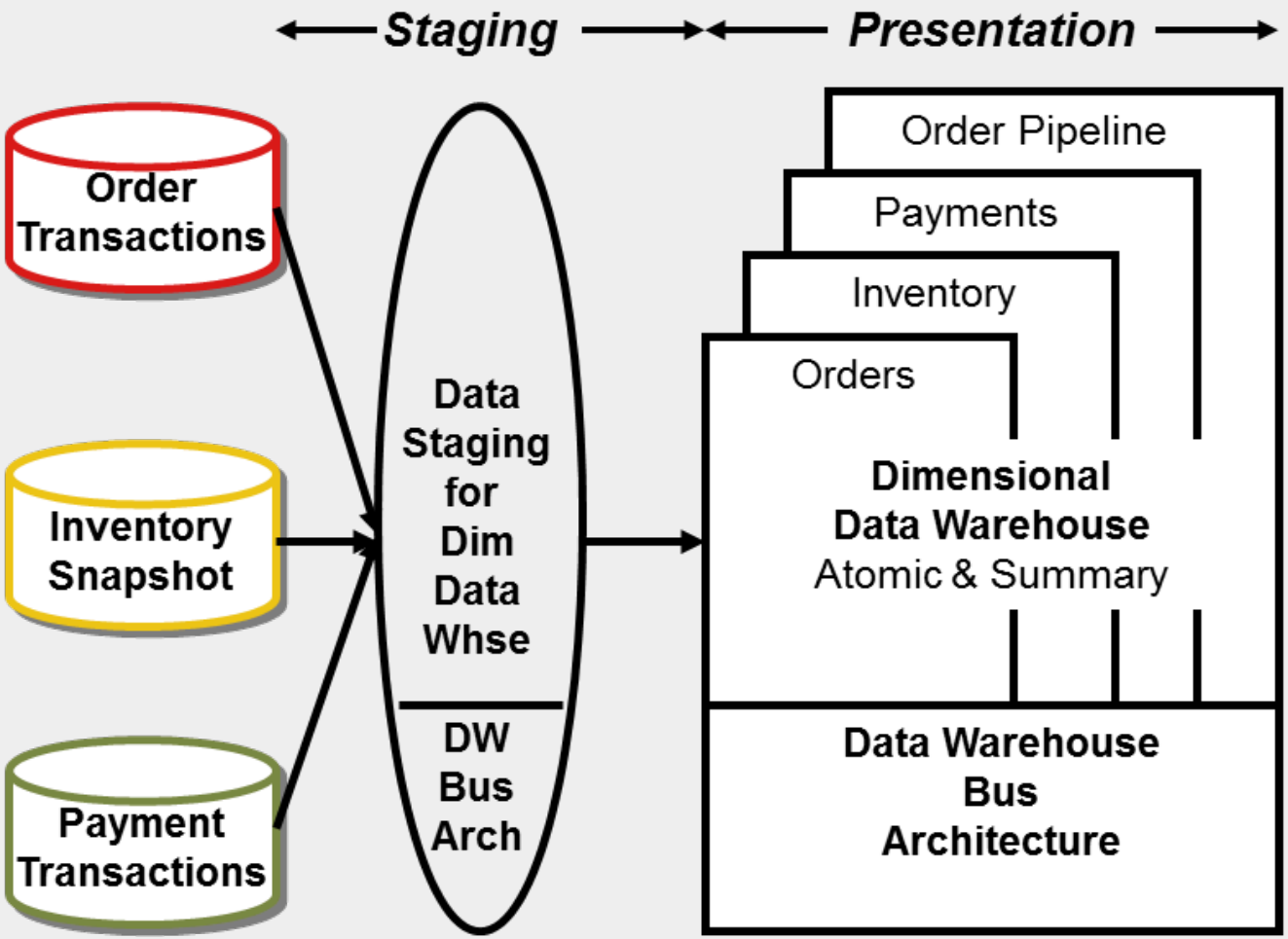


Source: http://www.kimballgroup.com/2004/03/differences-of-opinion/

Independent Data Mart Architecture

- ETL System
  - Multiple independent extracts from the same source data
    - Inefficient
    - Potentially different transformation business rules applied to created each data mart
- Advantages: Shorter time to delivery
- Disadvantages:
  - Inconsistency across data marts for the same source data
  - Undue burden on the source system
- This approach is NOT recommended

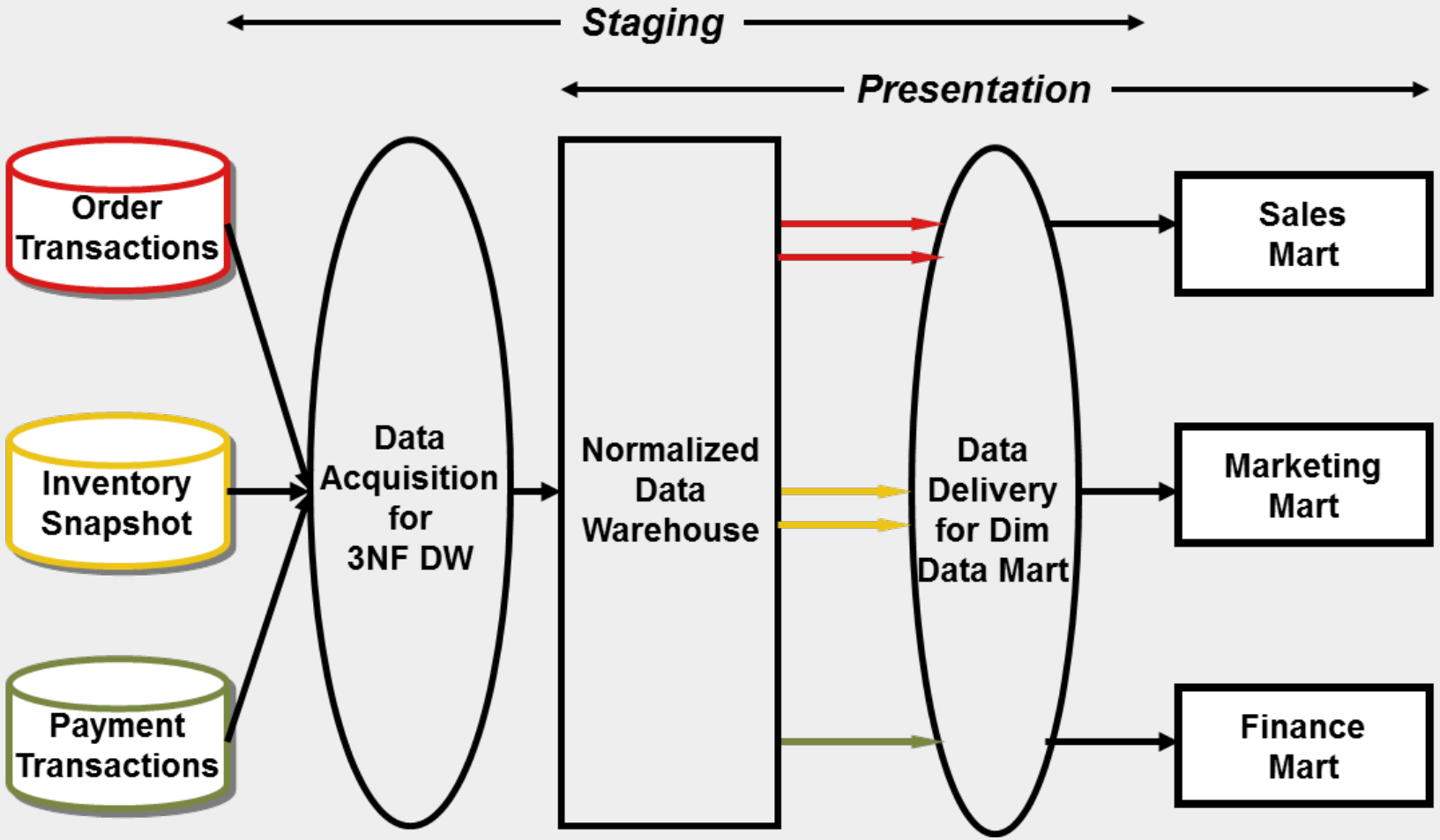
Kimball Architecture



Kimball Architecture

- ETL System
  - "Back room" and is off limit to the business users
  - No user query Support
  - Design Goals
    - Staging throughput, Integrity and Consistency
- Presentation Server
  - Stores dimensional models (star or OLAP cube)
  - Dimensional models typically contain atomic details in star schema; in addition, summary data may be stored in OLAP for query performance

Simplified Hub-and-Spoke Corp Info Factory (CIF) Architecture

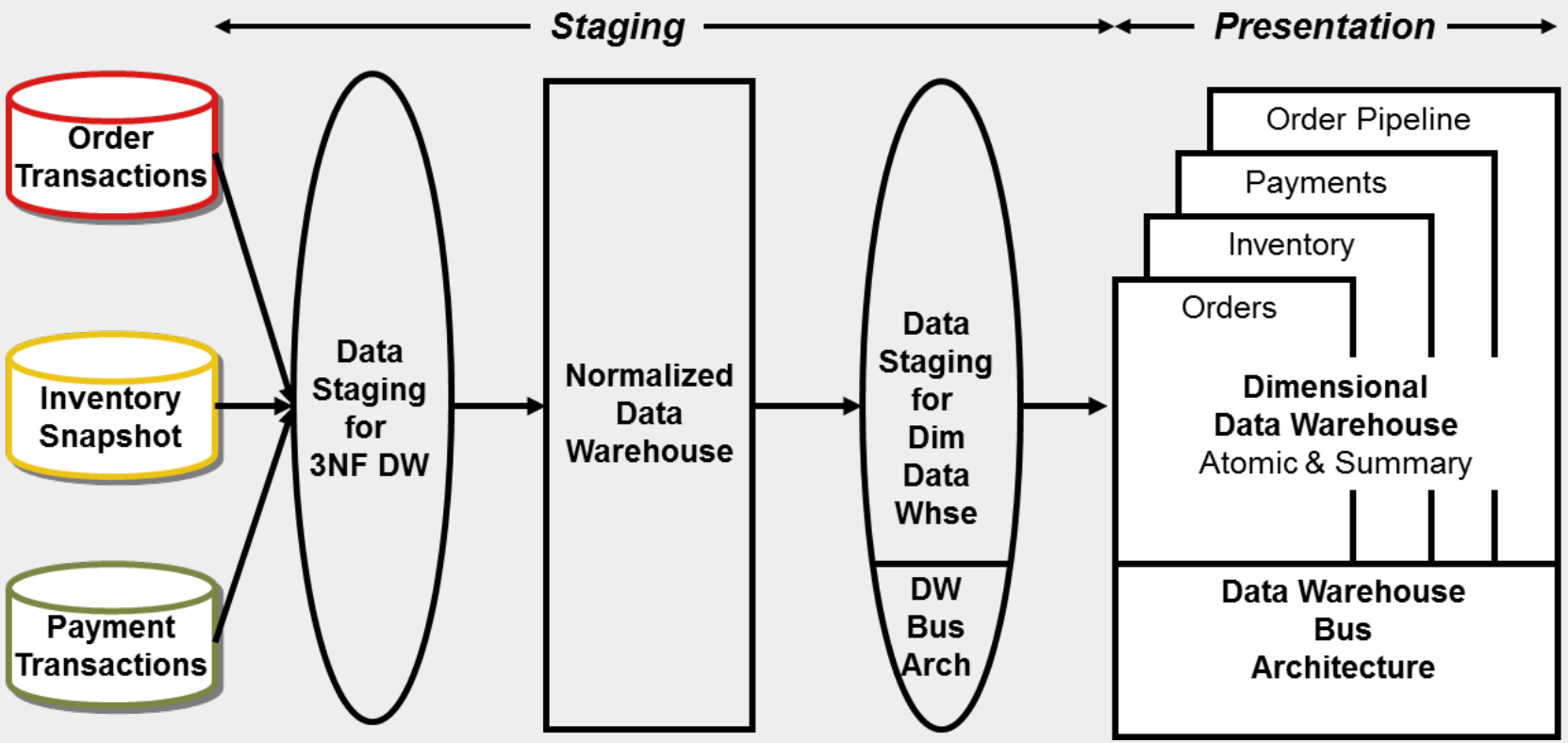


Source: http://www.kimballgroup.com/2004/03/differences-of-opinion/

Simplified Hub-and-Spoke Corp Info Factory (CIF) Architecture

- Similarities to Kimball Architecture
  - Deliver single version of the truth
  - Dimensional presentation in dimensional summary data
- Unique characteristics in CIF
  - The first ETL to EDW (normalized tables) and the second ETL to the dimensional structure
  - Users are given access to the EDW
  - Dimensional structures are typically for summary data only
  - The EDW and dimensional data can be out of sync depending on the ETL's timing

Hybrid Architecture



Source: http://www.kimballgroup.com/2004/03/differences-of-opinion/

Hybrid Architecture

- A normalized data warehouse from the CIF plus dimensional data warehouse of atomic detail
- Comes with high incremental costs and time lags
- It would work if there is already a normalized data warehouse built (e.g. SAP ERP system)

4-Step Dimensional Design Process

- Identity the Business Process
  - Source of "measurements"
- Identity the Grain
  - What does 1 row in fact table represent/mean?
  - Lowest atomic grain delivers most flexibilty
- Identity the Dimensions
  - Descriptive context, true to the grain
- Identify the Facts
  - Numeric additive measurements, true to the grain

Week 02 Topic: Dimensional Modeling Fundamental Concepts

Questions?