ETL

Cleaning and Conforming

- With the wisdom of hindsight from thousands of dimensional data warehouses, a set of ETL best practices have emerged.
- Careful consideration of these best practices has revealed 34 subsystems that are required in almost every dimensional data warehouse back room.
- The Kimball Group has organized these 34 subsystems of the ETL architecture into the following categories:
    - Three subsystems focus on extracting data from source systems.
    - Five subsystems deal with cleaning and conforming, including dimensional structures to monitor quality errors.
    - Thirteen subsystems deliver data as dimensional structures to the final BI layer, such as a subsystem to implement slowly changing dimension techniques.
    - Thirteen subsystems help manage the production ETL environment.

Subsystems that have been covered in the previous lectures:

- Subsystem 1: Data Profiling
- Subsystem 2: Change Data Capture System
- Subsystem 3: Extract System

Defining Data Quality

- Correct
  - The values and descriptions in data describe their associated objects truthfully and faithfully.
  - For example, the name of the city in which one of the authors currently live is called New Hope.
  - Therefore, accurate data about that home address needs to contain New Hope as the city name to be correct.
- Unambiguous.
  - The values and descriptions in data can be taken to have only one meaning.
  - For example, there are at least ten cities in the U.S. called New Hope, but there is only one city in Pennsylvania called New Hope.
  - Therefore, accurate data about an address in this city needs to contain New Hope as the city name and Pennsylvania as the state name to be unambiguous.

Defining Data Quality

- Consistent.
  - The values and descriptions in data use one constant notational convention to convey their meaning.
  - For example, the U.S. state Pennsylvania might be expressed in data as PA, Penn., or Pennsylvania.
  - To be consistent, accurate data about current home addresses should utilize just one convention (such as the full name Pennsylvania) for state names and stick to it.
- Complete.
  - There are two aspects of completeness.
  - The first is ensuring that the individual values and descriptions in data are defined (not null) for each instance.
  - The second aspect makes sure that the aggregate number of records is complete or makes sure that you didn’t somehow lose records altogether somewhere in your information flow.

Subsystem 4: Data Cleansing System

- Anomaly Detection
- Count the rows in a table while grouping on the column in question.
- The outliers in the result set are data anomalies and should be presented to the business owner with a strong recommendation that they be cleaned up in the source system.

```
select state, count(*)
 from order_detail
group by state
```

State	Count(*)
Rhode Island	1
Missisippi	2
New Yourk	5
Connecticut	7
New Mexico	43,844
Vermont	64,547
Mississippi	78,198
Utah	128,956
Wyoming	137,630
Missouri	148,953
Rhode Island	182,067
Minnesota	195,197
North Dakota	203,286
Michigan	241,245
Washington	274,528
Pennsylvania	287,289
Montana	337,128

Types of Enforcement

- It is useful to divide the various kinds of data-quality checks into four broad categories:
  - Column property enforcement
  - Structure enforcement
  - Data enforcement
  - Value enforcement

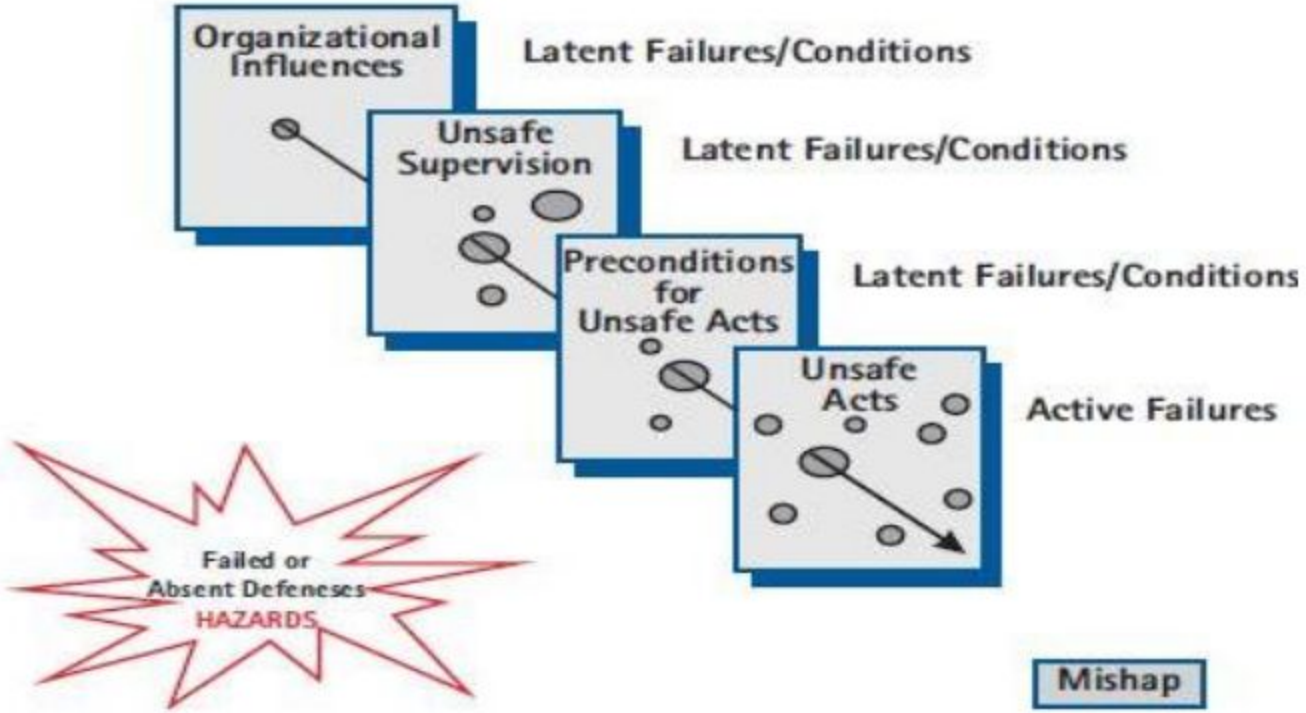
Column Property Enforcement

- Column property enforcement ensures that incoming data contains expected values from the providing system’s perspective.
- Column property enforcement checks include screens for:
  - Null values in required columns
  - Numeric values that fall outside of expected high and low ranges
  - Columns whose lengths are unexpectedly short or long
  - Columns that contain values outside of discrete valid value sets
  - Adherence to a required pattern or member of a set of patterns
  - Hits against a list of known wrong values where list of acceptable
  - values is too long
  - Spell-checker rejects

Based on the findings of these screens, the ETL job stream can choose to:

1. Pass the record with no errors
2. Pass the record, flagging offending column values
3. Reject the record
4. Stop the ETL job stream

- The general case is option two, passing records through the ETL stream and recording any validation errors encountered to the error event fact table to make these errors visible to the end user community and to avoid situations where data warehouse credibility is damaged. (Swiss cheese model)
- Data records that are so severely flawed that inclusion in the warehouse is either impossible or is damaging to warehouse credibility should be skipped completely (option three), the error event is duly noted in the error event fact table.
- And finally, data-validation errors that call into question the data integrity of the entire ETL batch should stop the batch process completely (option four), so that the data warehouse manager can investigate further.
- The screen dimension contains an exception action column that associates one of these three possible actions to each screen.


Organizational Influences
Latent Failures/Conditions
Unsafe Supervision
Latent Failures/Conditions
Preconditions for Unsafe Acts
Latent Failures/Conditions
Unsafe Acts
Active Failures
Failed or Absent Defeneses
HAZARDS
Mishap

Structure Enforcement

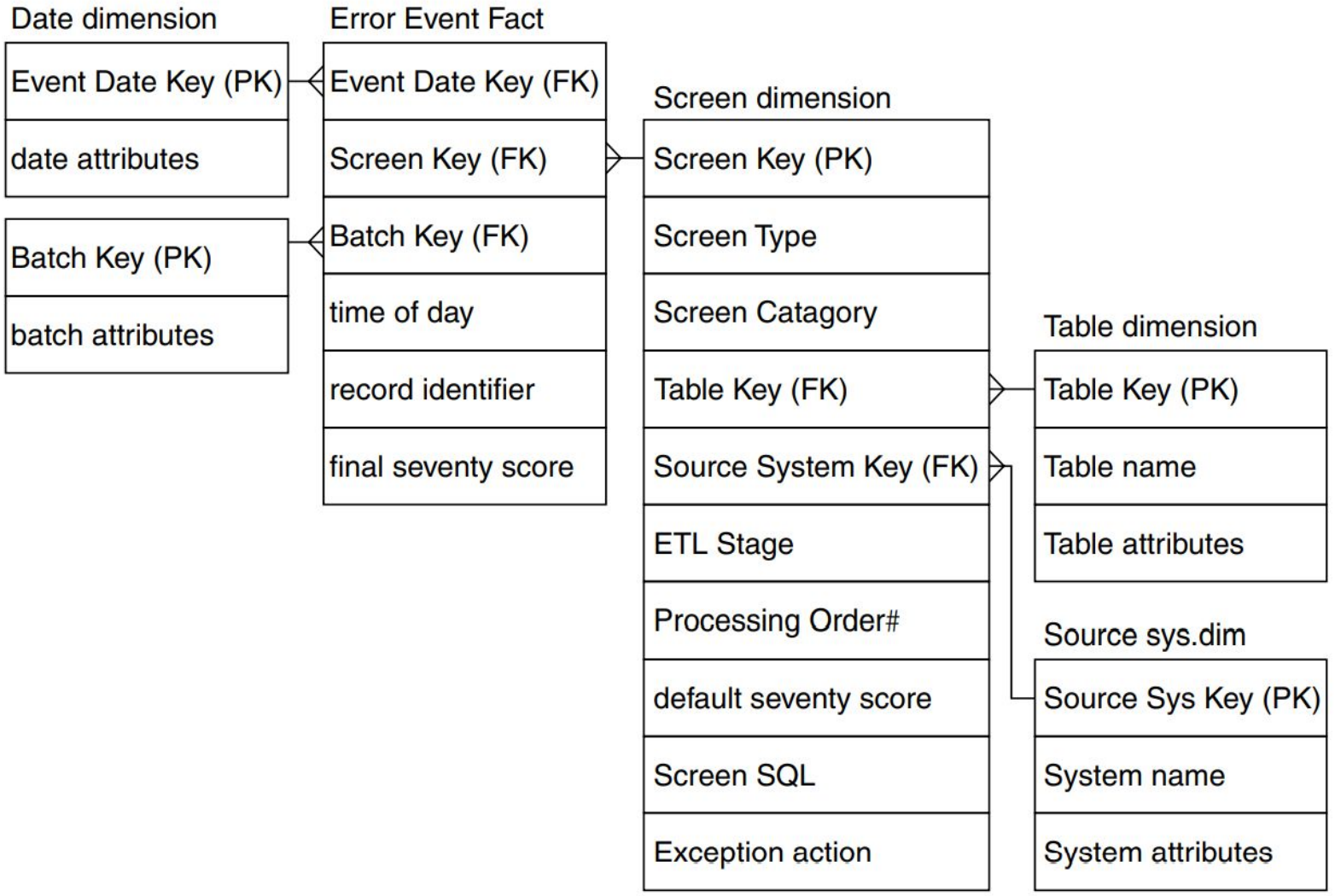
- Whereas column property enforcement focuses on individual fields, structure enforcement focuses on the relationship of columns to each other.
- We enforce structure by making sure that tables have proper primary and foreign keys and obey referential integrity.
- We check explicit and implicit hierarchies and relationships among groups of fields that, for example, constitute a valid postal mailing address.
- Structure enforcement also checks hierarchical parent-child relationships.

Data Value Rule Enforcement

- These range from simple business rules such as “if customer has preferred status, the overdraft limit is at least $1000” to more complex logical checks such as “a commercial customer cannot simultaneously be a limited partnership and a type C corporation”.
- They can also take the form of aggregate value business rules such as the physicians in this clinic are reporting a statistically improbable number of sprained elbows requiring MRIs.
- They can also provide a probabilistic warning that the data may be incorrect.

Subsystem 5 - Error Event Schema

- Each data-quality error or issue surfaced by the data-cleaning subsystem is captured as a row in the error event fact table.
- In other words, the grain of this fact table is each error instance of each data-quality check.
- Remember that a quality check is a screen.
- So, if you were to run ten separate screens against some set of data and each screen uncovered ten defective records, a total of 100 records would be written to the error event fact table.


Date dimension
Event Date Key (PK)
date attributes
Batch Key (PK)
batch attributes
Error Event Fact
Event Date Key (FK)
Screen Key (FK)
Batch Key (FK)
time of day
record identifier
final seventy score
Screen dimension
Screen Key (PK)
Screen Type
Screen Catagory
Table Key (FK)
Source System Key (FK)
ETL Stage
Processing Order#
default seventy score
Screen SQL
Exception action
Table dimension
Table Key (PK)
Table name
Table attributes
Source sys.dim
Source Sys Key (PK)
System name
System attributes

Error Event Schema

- The **event date** is a standard dimension representing the calendar date.
- The time of day is represented in the fact table as the number of seconds since midnight, expressed as an integer.
- The **batch** dimension contains a record for each invocation of the overall batch process—and typically contains interesting timestamps, and numbers of records processed.
- The **screen** dimension table contains constant descriptive information about each data-quality check, or screen, applied.
- It is not a description of a specific run (that is what the fact table records) but rather is a description of what the screen does and where it is applied.
- One of its attributes, the default severity score, defines a severity value for each of the various types of errors it may encounter.
- These error-severity scores are used as the basis of the final severity score error event fact table. For example, the final severity score could be higher than the individual default scores if a large number had accumulated.

Attributes of the screen dimension

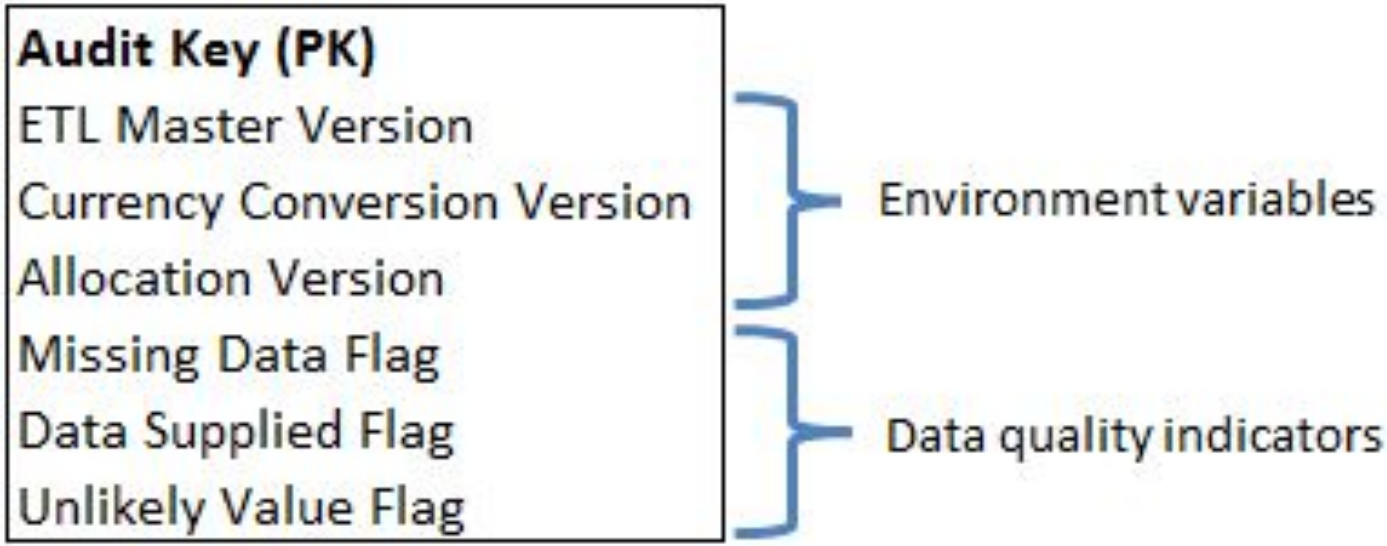
- The ETL Stage describes the stage in the overall ETL process in which the data-quality screen is applied.
- The Processing Order Number is a primitive scheduling/ dependency device, informing the overall ETL master process of the order in which to run the screens. Data-quality screens with the same processing-order number in the same ETL stage can be run in parallel.
- The Default Severity Score is used to define the error-severity score to be applied to each exception identified by the screen in advance of an overarching processing rule that could increase or decrease the final severity score as measured in the fact table.

Attributes of the screen dimension (2)

- The Exception Action attribute tells the overall ETL process whether it should pass the record, reject the record, or stop the overall ETL process upon discovery of error of this type.
- The Screen Type and Screen Category Name are used to group data-quality screens related by theme, such as Completeness or Validation or Out-of-Bounds.
- And finally, the SQL Statement captures the actual snippet of SQL or procedural SQL used to execute the data-quality check. If applicable, this SQL should return the set of unique identifiers for the rows that violate the data-quality screen so that this can be used to insert new records into the error event fact table.

Subsystem 6 : Audit Dimension Assembler

- The audit dimension, like all dimensions, provides the context for a particular **fact** row.
- Thus when a fact row is created, the environment variables are fetched from a small table containing the version numbers in effect for specific ranges of time.
- The data quality indicators are fetched from the error event fact table that records data quality errors encountered along the ETL pipeline.


Audit Key (PK)
ETL Master Version
Currency Conversion Version
Allocation Version
Missing Data Flag
Data Supplied Flag
Unlikely Value Flag
Environment variables
Data quality indicators

- The environment variables in the above figure are version numbers that change only occasionally.
- The ETL master version number is a single identifier, similar to a software version number, that refers to the complete ETL configuration in use when the particular fact row was created.
- The currency conversion version is another version number that identifies a specific set of foreign currency conversion business rules in effect when the fact table row was created.
- The allocation version is a number that identifies a set of business rules for allocating costs when calculating profitability.

- The data quality indicators are flags that show whether some particular condition was encountered for the specific fact row.
- If the fact row contained missing or corrupt data (perhaps replaced by null) then the missing data flag would be set to true.
- If missing or corrupt data was filled in with an estimator, then the data supplied flag would be true.
- If the fact row contained anomalously high or low values, then the unlikely value flag would be true.
- Note that this simple audit dimension does not provide a precise description of the data quality problem, rather it only provides a warning that the business user should tread cautiously.

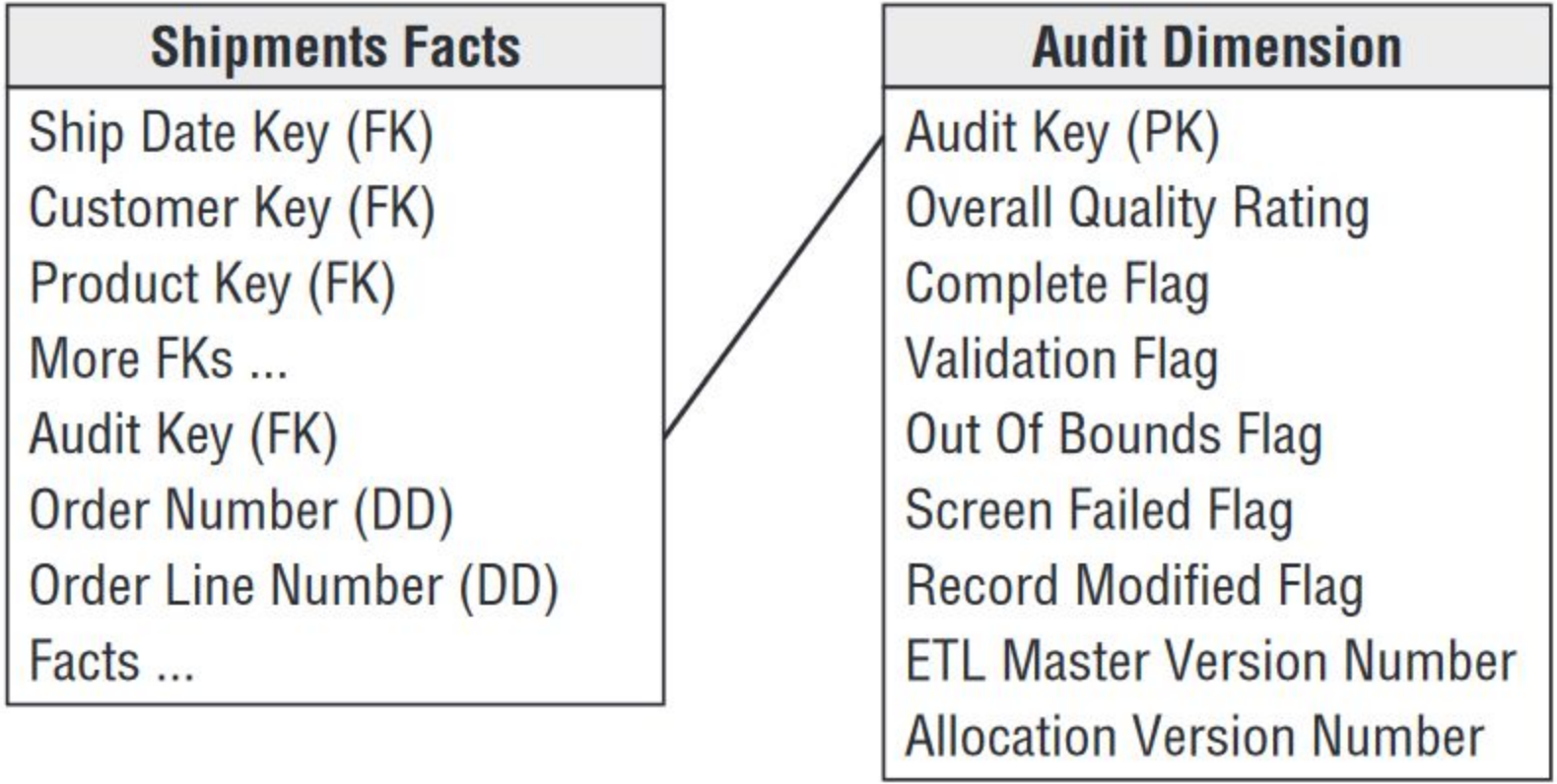
- An audit dimension can help your business users.
- Here’s a before and after portion of a simple tracking report using an out of bounds indicator with values “Abnormal” and “OK” that provides a useful warning that a large percentage of the Axon West data contains unlikely values.
- The instrumented report is created just by dragging the out of bounds indicator into the query.
- Business users get to know why the data has been flagged and get sufficient information for making business decisions.

Normal Report:

Product	Ship From	Qty Shipped	Revenue
Axon	East	1438	$235,000
Axon	West	2249	$480,000

Instrumented Report (add Out of Bounds Indicator to SELECT):

Product	Ship From	Out of Bounds Indicator	Qty Shipped	Revenue
Axon	East	Abnormal	14	$2,350
Axon	East	OK	1424	$232,650
Axon	West	Abnormal	675	$144,000
Axon	West	OK	1574	$336,000


Shipments Facts
Ship Date Key (FK)
Customer Key (FK)
Product Key (FK)
More FKs ...
Audit Key (FK)
Order Number (DD)
Order Line Number (DD)
Facts ...
Audit Dimension
Audit Key (PK)
Overall Quality Rating
Complete Flag
Validation Flag
Out Of Bounds Flag
Screen Failed Flag
Record Modified Flag
ETL Master Version Number
Allocation Version Number

Subsystem 7: Deduplication System

- Often dimensions are derived from several sources.
- This is a common situation for organizations that have many customer-facing source systems that create and manage separate customer master tables.
- Customer information may need to be merged from several lines of business and outside sources.
- Sometimes, the data can be matched through identical values in some key column.
- However, even when a definitive match occurs, other columns in the data might contradict one another, requiring a decision on which data should survive.

- Unfortunately, there is seldom a universal column that makes the merge operation easy.
- Sometimes, the only clues available are the similarity of several columns.
- The different sets of data being integrated and the existing dimension table data may need to be evaluated on different fields to attempt a match.
- Sometimes, a match may be based on fuzzy criteria, such as names and addresses that may nearly match except for minor spelling differences.

- Survivorship is the process of combining a set of matched records into a unified image that combines the highest quality columns from the matched records into a conformed row.
- Survivorship involves establishing clear business rules that define the priority sequence for column values from all possible source systems to enable the creation of a single row with the best-survived attributes.
- If the dimensional design is fed from multiple systems, you must maintain separate columns with back references, such as natural keys, to all participating source systems used to construct the row.

Subsystem 8: Conforming System

- Conforming consists of all the steps required to align the content of some or all the columns in a dimension with columns in similar or identical dimensions in other parts of the data warehouse.
- For instance, in a large organization you may have fact tables capturing invoices and customer service calls that both utilize the customer dimension.
- It is highly likely the source systems for invoices and customer service have separate customer databases.
- It is likely there will be little guaranteed consistency between the two sources of customer information.
- The data from these two customer sources needs to be conformed to make some or all the columns describing customer share the same domains.

- For two dimensions to be conformed, they must share at least one common attribute with the same name and same contents.
- You can start with a single conformed attribute such as Customer Category and systematically add this column in a nondisruptive way to customer dimensions in each of the customer-facing processes.
- As you augment each customer-facing process, you expand the list of processes that are integrated and can participate in drill-across queries.
- You can also incrementally grow the list of conformed attributes, such as city, state, and country.

Deduplicating and survivorship processing for conformed dimension process