Database Vocabulary: Is Your Data Set a Dimension (LookUp) Table, a Fact Table or a Report?
Ronald J. Fehd, Centers for Disease Control and Prevention, Atlanta, GA, USA

ABSTRACT
This paper provides a review of database vocabulary and design issues. It reviews the categories of variables and tables in a relational database and offers tools to categorize variables in a data set and recode them so that the data set meets all the criteria of a relational database table.

This is another in the Journeymen's Tools series.

Audience intermediate users and programmers.

Keywords database design, dimension table, fact table, facts, foreign key, lookup table, normal forms, primary key, relational database

Information look-up tables.

In this paper This paper contains the following topics.

| Introduction | 2 |
| Concepts | 2 |
| Case Studies | 6 |
| SAShelp.Class | 6 |
| SAShelp.Shoes | 7 |
| Summary | 8 |
INTRODUCTION

Overview
This document reviews the following categories of database concepts.

- Cardinality Ratio
- categories of variables or columns
- categories of database tables

The source of the column and database table description is Kimball and Ross [3]. Kimball-Ross-DataWareHouseToolkit

Case studies of data sets from the the libref SAShelp are provided.
The purpose of this paper is to acquaint the reader with database concepts and provide examples of how these concepts may used to analyze the data structure and processing of their data sets.

CONCEPTS

Cardinality Ratio
Cardinality ratio is the number of levels of a variable divided by the number of observations. This ratio is used to differentiate variables into their respective categories. As a general rule cardinality ratio falls into four groups:

0 blank or missing
low foreign keys
high continuous integers or real numbers
1 primary keys

Nlevels from Proc Freq
The numerator of cardinality ratio is available from proc freq with the nlevels option.

The following code illustrates how to calculate cardinality ratio using Proc Freq:

```
PROC Freq data = sashelp.Class nlevels;
PROC Freq data = sashelp.Shoes nlevels;
```

The following table shows the number of variable levels for each variable in the SAShelp.Class and SAShelp.Shoes datasets.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Levels</th>
<th>Variable</th>
<th>Label</th>
<th>Levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>19</td>
<td>Region</td>
<td></td>
<td>10</td>
</tr>
<tr>
<td>Sex</td>
<td>2</td>
<td>Product</td>
<td></td>
<td>8</td>
</tr>
<tr>
<td>Age</td>
<td>6</td>
<td>Subsidiary</td>
<td></td>
<td>53</td>
</tr>
<tr>
<td>Height</td>
<td>17</td>
<td>Stores</td>
<td>Number of Stores</td>
<td>33</td>
</tr>
<tr>
<td>Weight</td>
<td>15</td>
<td>Sales</td>
<td>Total Sales</td>
<td>392</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Inventory</td>
<td>Total Inventory</td>
<td>395</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Returns</td>
<td>Total Returns</td>
<td>372</td>
</tr>
</tbody>
</table>
Categories of Variables

Data set variables or database columns are grouped into two categories.

- facts
- keys:

Facts may be character or numeric.

In a dimension table facts are character and contain text of information about the entity identified by the primary key.

In transaction tables facts are integers for quantity, and real numbers for measurements, prices, sums or other statistics.

Keys may be foreign keys or primary keys. Both foreign and primary keys are positive integers excluding zero. Foreign keys occur multiple times in transaction and snapshot tables. The cardinality ratio of foreign keys is low. A primary key is the unique row number of a dimension table; therefore the cardinality ratio of a primary key is one.

Reports are assembled by joining fact-table.key1 with dimension-table.key1.

Categories of Database Tables

Database tables are either of these categories:

1. Dimension tables
2. Fact tables

Dimension tables have a single primary key and contain text of information about the entity identified by the primary key.

The above statement is overly simplified for this discussion. It describes a dimension table in a Star Schema; a dimension table in a Snowflake Schema may contain foreign keys.

Fact tables have a composite key composed of a set of foreign keys from dimension tables; other columns contain event measurements.

Dimension or LookUp Tables

A dimension table contains information about an entity. Dimension tables are referred to as LookUp tables. In programs they are implemented as one-to-one formats.

Example:

```
Proc Format library = Library fmtlib;
value gender 1 = ‘female’
   2 = ‘male’;
```

As a data set or dimension table this would be:

```
obs GenderId GenderText
  1   1 female
  2   2 male
```
Categories of Fact Tables

1. Transactions

2. Snapshots:
   (a) Accumulating Snapshot
   (b) Periodic Snapshot

Transaction is a recording of an event
Snapshots are summarization of transactions
Accumulating Snapshot is a record of milestones achieved
Periodic Snapshot is a summarization of transactions for a time period

Transactions

A transaction is a recording of an event. One row is added to the table for each event. Rows are not updated. A transaction table has the following columns:

1. composite key: a set of foreign keys
   (a) date and/or time
   (b) location
   (c) vendor or seller
   (d) buyer or customer

2. facts
   (a) item
   (b) quantity
   (c) price per unit
   (d) purchase amount

An example is a line item on a purchase receipt.
A data set is said to be normalized when it meets the definition of a transaction table.

Accumulating Snapshot

An Accumulating Snapshot is a set of milestones of activity during a specific time period. The value of each milestone is either missing or the date accomplished. Unlike the other fact tables, accumulating snapshot table rows are updated; this happens whenever an event is accomplished. An accumulating snapshot table has the following columns:

1. primary key of the entity
   (a) project
   (b) vendor or seller
   (c) buyer or customer

2. facts: milestones

An example is a student application tracking process where milestones include sending notices and receiving responses, as well as reminder notices of responses not received on time.
A Periodic Snapshot is a summarization of transactions during a specific time period, such as daily, monthly, or yearly. At the end of the time period, after the summarization process is completed one row is added to the table. Rows are not updated. A periodic snapshot table has the following columns:

1. composite key: a set of foreign keys indicating the granularity of the snapshot
   (a) time period: begin or end
   (b) entity: buyer, customer, seller or vendor
   (c) location of transaction collection: store, territory, city, county, state, region, etc.

2. summarization of transaction facts

Examples include:
- purchase receipt
- monthly financial statement from bank or credit card
- tax return

Kimball and Ross [3], Kimball-Ross-DataWareHouseToolkit, page 133, provide the following table comparing the types of fact tables.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Transaction Grain</th>
<th>Periodic Snapshot Grain</th>
<th>Accumulating Snapshot Grain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time period represented</td>
<td>Point in time</td>
<td>Regular, predictable intervals</td>
<td>Indeterminate time span, typically short-lived</td>
</tr>
<tr>
<td>Grain</td>
<td>One row per transaction event</td>
<td>One row per period</td>
<td>One row per life</td>
</tr>
<tr>
<td>Fact table loads</td>
<td>Insert</td>
<td>Insert</td>
<td>Insert and update</td>
</tr>
<tr>
<td>Fact row updates</td>
<td>Not revisited</td>
<td>Not revisited</td>
<td>Revisited whenever activity</td>
</tr>
<tr>
<td>Date dimension</td>
<td>Transaction date</td>
<td>End-of-period date</td>
<td>Multiple dates for standard milestones</td>
</tr>
<tr>
<td>Facts</td>
<td>Transaction activity</td>
<td>Performance for predefined time</td>
<td>Performance over finite lifetime</td>
</tr>
</tbody>
</table>
Agosta [1], Agosta-DataWarehousing defines a data warehouse fact as: “A customer buys a product at a certain location at a certain time.” The elements of this transaction definition statement are:

- actor
- verb
- object
- location
- time

This table compares the transaction definition statement for two common database transaction types.

<table>
<thead>
<tr>
<th>Description</th>
<th>Sales</th>
<th>Inventory</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actor</td>
<td>Customer</td>
<td>Clerk</td>
</tr>
<tr>
<td>Verb</td>
<td>purchased</td>
<td>counted</td>
</tr>
<tr>
<td>Object</td>
<td>Product</td>
<td>Product</td>
</tr>
<tr>
<td>Location</td>
<td>at Store</td>
<td>at Warehouse</td>
</tr>
<tr>
<td>Date-Time</td>
<td>on Date</td>
<td>on Date</td>
</tr>
<tr>
<td>Fact:</td>
<td>N(items)</td>
<td>N(Products)</td>
</tr>
<tr>
<td>Fact:</td>
<td>Price per unit</td>
<td></td>
</tr>
</tbody>
</table>

CASE STUDIES

Cardinality Ratio Report

The following reports are produced with version 2 of Fehd [2], sgf2008.003 which provides a cardinality ratio calculator for a data set.

SASHELPCLASS

Cardinality Ratio Report

The SASHelp.Class data set is provided with your installation.

The SmryEachVar Data Review Suite
Cardinality Ratio Report
Data: SASHelp.Class nobs: 19

<table>
<thead>
<tr>
<th>Var Num</th>
<th>Name</th>
<th>Type</th>
<th>Length</th>
<th>Label</th>
<th>Format</th>
<th>NLevels</th>
<th>Card Ratio</th>
<th>Card Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>Age</td>
<td>N</td>
<td>8</td>
<td></td>
<td></td>
<td>6</td>
<td>0.316</td>
<td>fkey?</td>
</tr>
<tr>
<td>4</td>
<td>Height</td>
<td>N</td>
<td>8</td>
<td></td>
<td></td>
<td>17</td>
<td>0.895</td>
<td>continuous</td>
</tr>
<tr>
<td>1</td>
<td>Name</td>
<td>C</td>
<td>8</td>
<td></td>
<td></td>
<td>19</td>
<td>1.000</td>
<td>pkey?</td>
</tr>
<tr>
<td>2</td>
<td>Sex</td>
<td>C</td>
<td>1</td>
<td></td>
<td></td>
<td>2</td>
<td>0.105</td>
<td>fkey?</td>
</tr>
<tr>
<td>5</td>
<td>Weight</td>
<td>N</td>
<td>8</td>
<td></td>
<td></td>
<td>15</td>
<td>0.789</td>
<td>continuous</td>
</tr>
</tbody>
</table>
Analysis

We can fill in the fact table description —actor, verb, object, location and time— with this sentence: “Staff measured students’ growth at unknown location, on unknown date.”

**Facts:** variables number four and five, height and weight, are facts; this is confirmed by their high cardinality ratio and that they are measurements. Note that units of measurement are missing in their labels.

**Primary Key** variable number one, Name, is the primary key with cardinality ratio of one.

**Foreign Key** variable number two, Sex or gender, is a foreign key because of its low cardinality ratio. Note that this variable is a constant attribute of the person, so it could be moved to a dimension table.

**Age** variable number three is a time interval calculated as the difference between birthdate and data-collection date, both of which are missing from this data.

Summary

**Table type:** report, this judgement is based on age which is calculated.

**Guesstimate:** derived from join of:

- **Dimension table:** Student-Id, name, gender, date-of-birth
- **Fact table:** Student measurements: date, Student-Id, height, weight

SASHELP.SHOES

Cardinality Ratio Report

The SAShelp.Shoes data set is provided with your installation.
The SmryEachVar Data Review Suite
Cardinality Ratio Report
Data: SAShelp.Shoes nobs: 395

<table>
<thead>
<tr>
<th>Var Num</th>
<th>Name</th>
<th>Type</th>
<th>Length</th>
<th>Label</th>
<th>Format</th>
<th>NLevels</th>
<th>Card Ratio</th>
<th>Card Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>Inventory</td>
<td>N</td>
<td>8</td>
<td>Total Inventory</td>
<td>DOLLAR12.</td>
<td>395</td>
<td>1.000</td>
<td>pkey?</td>
</tr>
<tr>
<td>2</td>
<td>Product</td>
<td>C</td>
<td>14</td>
<td></td>
<td></td>
<td>8</td>
<td>0.020</td>
<td>fkey?</td>
</tr>
<tr>
<td>1</td>
<td>Region</td>
<td>C</td>
<td>25</td>
<td></td>
<td></td>
<td>10</td>
<td>0.025</td>
<td>fkey?</td>
</tr>
<tr>
<td>7</td>
<td>Returns</td>
<td>N</td>
<td>8</td>
<td>Total Returns</td>
<td>DOLLAR12.</td>
<td>372</td>
<td>0.942</td>
<td>continuous</td>
</tr>
<tr>
<td>5</td>
<td>Sales</td>
<td>N</td>
<td>8</td>
<td>Total Sales</td>
<td>DOLLAR12.</td>
<td>392</td>
<td>0.992</td>
<td>continuous</td>
</tr>
<tr>
<td>4</td>
<td>Stores</td>
<td>N</td>
<td>8</td>
<td>Number of Stores</td>
<td></td>
<td>33</td>
<td>0.084</td>
<td>fkey?</td>
</tr>
<tr>
<td>3</td>
<td>Subsidiary</td>
<td>C</td>
<td>12</td>
<td></td>
<td></td>
<td>53</td>
<td>0.134</td>
<td>fkey?</td>
</tr>
</tbody>
</table>

Analysis

Based on the variable labels, which contain the word *Total*, this is a periodic snapshot. Reviewing the values in each of the variables with low cardinality, we find that Region contains names of areas of continents, Subsidiary contains city names, and Product contains types of shoes.

We can fill in the fact table description —actor, verb, object, location and time— with this sentence: “(Number of) Stores sold types of shoes (Product) in location (city = Subsidiary), during unknown time period.”
As a check to our assumption that Region, Subsidiary and Product are the set of foreign keys which describe the granularity of this periodic snapshot table, we can do a proc freq cross-tabulation of the three variables.

```sas
1 PROC Freq data = sashelp.Shoes
2   nlevels;
3   tables Region
4   * Subsidiary
5   * Product
6   / list missing noprint
7   out = Work.Freq;
```

NOTE: There were 395 observations read from the data set SASHELP.SHOES.
NOTE: The data set WORK.FREQ has 394 observations and 5 variables.

```sas
8 PROC Print data = Work.Freq
9   (where = (Count ge 2));
```

NOTE: There were 1 observations read from the data set WORK.FREQ.
WHERE Count>=2;

We expect the data set, as a periodic snapshot, to be unique on its set of foreign keys; this is not the case.
Remember: it is an example data set!

---

**SUMMARY**

**Conclusion**

By gaining knowledge of the vocabulary of database design programmers and users can more easily describe their input, processing and output.

**Further Reading**

**Database Design Theory** Edgar F. Codd describes the basic rules of relational database design in Codd’s 12 rules.

**REFERENCES**


---

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. © indicates USA registration.

Author: Ronald Fehd
Centers for Disease Control
4770 Buford Hwy NE
Atlanta GA 30341-3724

mailto:RJF2@cdc.gov