Paper CC-04

Database Vocabulary: Is Your Data Set a Dimension (LookUp) Table, a Fact Table or a Report?

Ronald J. Fehd, SAS-L peer, Atlanta, GA, USA

Abstract

Description : 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 the criteria of a relational database table.

Purpose : 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.

Audience : intermediate users and programmers.

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

Contents

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 [5, Kimball-Ross-DataWareHouseToolkit]. The description of a data warehouse fact is from Agosta [1, Agosta-DataWarehousing]. Case studies of data sets from the libref sashelp are provided.

Concepts

Cardinality Ratio

Cardinality ratio is the number of levels of a variable divided by the number of observations of the data set. 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 from proc freq with the nlevels option.

proc-freq-nlevels.sas

```
1 PROC Freq data = sashelp.class
2                  nlevels;
3 PROC Freq data = sashelp.shoes
4                  nlevels;
```

sashelp.class

Number of Variable Levels

Variable	Levels
Name	19
Sex	2
Age	6
Height	17
Weight	15

sashelp.shoes

Number of Variable Levels

Variable	Label	Levels
Region		10
Product		8
Subsidiary		53
Stores	Number of Stores	33
Sales	Total Sales	392
Inventory	Total Inventory	395
Returns	Total Returns	372

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 a fact-table.key with the dimension-table.key.

Categories of Database Tables

Database tables are either of these categories:

- dimension tables, also known as lookup
- 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

Overview

Database tables are organized in these broad categories:

- transactions
- snapshots:
  - accumulating snapshot
  - periodic snapshot

transaction : is a recording of an event

snapshots : are summarization of transactions

accumulating : is a record of milestones achieved

periodic : 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 these columns:

- composite key: a set of foreign keys
  - date and/or time
  - location
  - vendor or seller
  - buyer or customer
- facts
  - item
  - quantity
  - price per unit
  - 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.

Categories of Snapshots or Reports

Overview

Reports are referred to as Snapshots and are either of these categories:

- accumulating
- periodic

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:

- primary key of the entity
  - project
  - vendor or seller
  - buyer or customer
- 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.

Periodic Snapshot

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

Continued on next page.

Comparison

Kimball and Ross [5, Kimball-Ross-DataWareHouseToolkit], page 133, provide the following table comparing the types of fact tables.

Fact Table Type Comparison			
Characteristic	Transaction Grain	Periodic Snapshot Grain	Accumulating Snapshot Grain
Time period represented	Point in time	Regular, predictable intervals	Indeterminate time span, typically short-lived
Grain	One row per transaction event	One row per period	One row per life
Fact table loads	Insert	Insert	Insert and update
Fact row updates	Not revisited	Not revisited	Revisited whenever activiity
Date dimension	Transaction date	End-of-period date	Multiple dates for standard milestones
Facts	Transaction activity	Performance for predefined time	Performance over finite lifetime

Transaction Categories

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.

description	sales	inventory
actor	customer	clerk
verb	purchased	counted
object	product	product
location	at store	at warehouse in row, bin
date-time	on date at time	on date
fact:	n(items)	n(products)
fact:	price per unit	

Case Studies

Cardinality Ratio Report

The following reports are produced with version 2 of Fehd [4, sgf2008.003] which provides a cardinality ratio calculator for a data set.

SAShelp.Class

Cardinality Ratio Report

The data set sashelp.class is provided with your installation.

The SmryEachVar Data Review Suite
Cardinality Ratio Report
Data: sashelp.class nobs: 19

Var Num	Name	Type	Length	Label	Format	NLevels	Card Ratio	Card Note
3	Age	N	8			6	0.316	fkey?
4	Height	N	8			17	0.895	continuous
1	Name	C	8			19	1.000	pkey?
2	Sex	C	1			2	0.105	fkey?
5	Weight	N	8			15	0.789	continuous

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 data set sashelp.shoes is provided with your installation.

The SmryEachVar Data Review Suite
Cardinality Ratio Report
Data: sashelp.shoes nobs: 395

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

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."

Data Review SAShelp.Shoes

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.

```
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.
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

programs : for this paper are in Fehd [2, sco.Cardinality-Ratio]; see also, in these conference proceedings:

! $\rightarrow$ Data Review Information: N-Levels or Cardinality Ratio

database theory : Edgar F. Codd describes the basic rules of relational database design in Codd's 12 rules.

SmryEachVar : Fehd [3, sco.SmryEachVar] contains the suite of programs used to calculate Cardinality Ratio.

References

[1] Lou Agosta. *The Essential Guide to Data Warehousing*. Prentice-Hall Inc., Upper Saddle River, NJ, 2000. URL `http://www.pearsonhighered.com/academic/product/0,3110,013085087X,00.html`. 19 chap., 454 pp., glossary: 25 pp., references: 4 pp., index: 15 pp.

[2] Editor R.J. Fehd. Cardinality-ratio. In *sasCommunity.org*, 2008. URL `http://www.sascommunity.org/wiki/Cardinality_Ratio`. topics: definition and programs.

[3] Editor R.J. Fehd. SmryEachVar: A data-review suite for each variable in all data sets in a libref. In *sasCommunity.org*, 2008. URL `http://www.sascommunity.org/wiki/SmryEachVar_A_Data_Review_Suite`. list processing using parameterized includes.

[4] Ronald J. Fehd. SmryEachVar: A data-review routine for all data sets in a libref. In *Proceedings of the SAS Global Forum Annual Conference*, 2008. URL `http://www2.sas.com/proceedings/forum2008/003-2008.pdf`. Applications Development, 24 pp.; topics: data review; info: utilities to repair missing elements in data structure.

[5] Ralph Kimball and Margy Ross. *The Data Warehouse Toolkit, The Complete Guide to Dimensional Modeling, Second Edition*. John Wiley & Sons, Inc., New York, 2002. URL `http://www.kimballgroup.com/html/booksDWT2.html`. subtitle: Complete Guide to Dimensional Modeling; 17 chap., 387 pp., glossary: 29 pp., index: 18 pp.

Contact Information: **Ronald J. Fehd** `mailto:Ron.Fehd.macro.maven@gmail.com`
`http://www.sascommunity.org/wiki/Ronald_J._Fehd`

About the author:

education:	B.S. Computer Science, U/Hawaii,	1986
	SAS User Group conference attendee since	1989
	SAS-L reader since	1994
experience:	programmer: 25+ years	
	data manager using SAS:	17+ years
	statistical software help desk:	7+ years
	author: 30+ SUG papers	
	sasCommunity.org: 300+ pages	
SAS-L:	author: 6,000+ messages to SAS-L since	1997
	Most Valuable SAS-L contributor:	2001, 2003

Trademarks

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.

Other brand and product names are trademarks of their respective companies.