恩門 BILLINMON.COM

Snapshots in the Data Warehouse

BY

W. H. Inmon

There are three types of modes that a data warehouse is loaded in:

- loads from archival data
- loads of data from existing systems
- loads of data into the warehouse on an ongoing basis.

The loading of data into the warehouse of archival data or from data residing in existing systems is of a "one time only" variety. In other words, if these types of loads of data are done at all, they are done only once. Because they are once in a lifetime activities, they can be very complex and time consuming and we can still cope with them. Almost anything done on a truly once in a lifetime basis is able to be survived, if not pleasant.

However, the third type of load into the data warehouse - the ongoing load of changes as they have occurred in the operational environment - is an entirely different matter. The management of these everyday, ongoing changes can consume an enormous amount of resources and can be very, very complex. It is this type of load that rivets the attention of the data architect. These ongoing loads of data are done in terms of "snapshots" that pass from the operational environment to the data warehouse environment.

SNAPSHOTS

Data in the data warehouse is stored in units of "snapshots". The records in the data warehouse are created as of some moment in time and are in effect a snapshot taken as of that moment in time. In this regard the data in the data warehouse is fundamentally different from the data in an operational data base environment. Data in an operational data base environment can be updated. Since data in the data warehouse environment is snapshot data it cannot be updated.

Figure 1 illustrates some simple forms of snapshot data found in the data warehouse.

One of the essential items of the snapshot taken is the element of time that applies to the snapshot. Said another way, each snapshot contains an element of time that further identifies the snapshot of data.

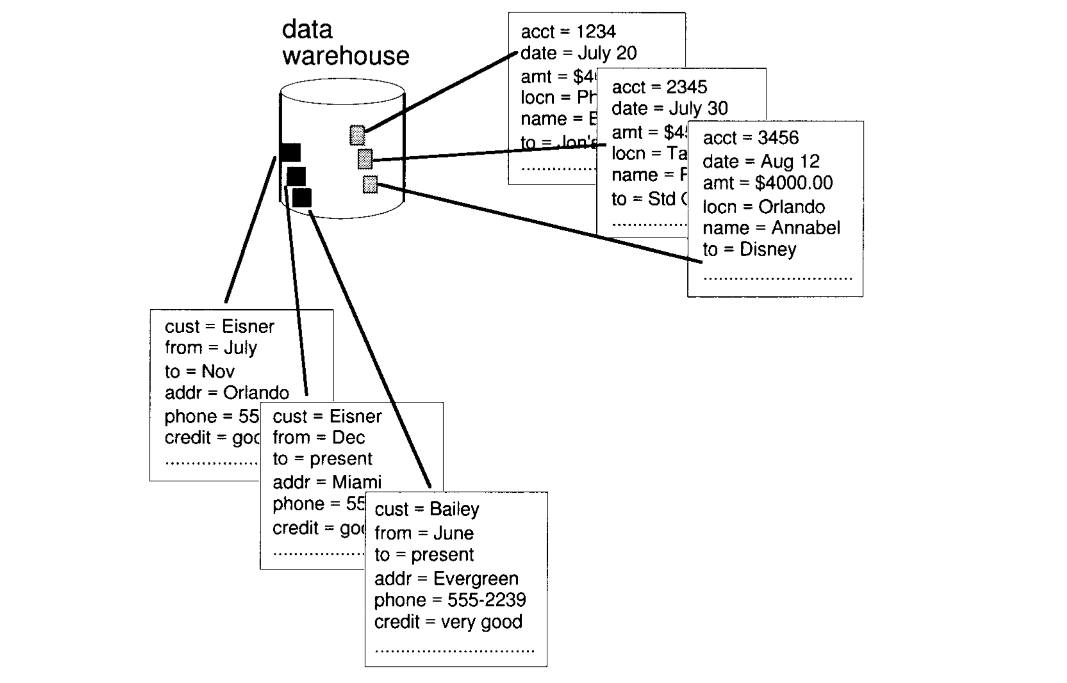


Figure 1: The data warehouse contains a series of snapshots of data.

There are many different forms of taking snapshots. This discussion will be on the different forms of snapshots and their associated design considerations.

EVENTS

The most basic consideration of a snapshot is that the snapshot has been taken as a result of an event. Figure 2 shows a snapshot being taken as a result of an event occurring.

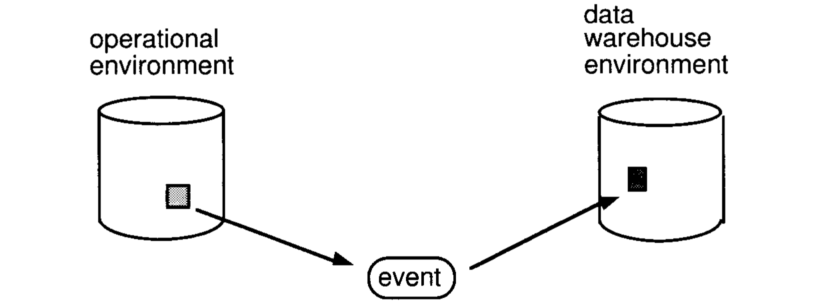


Figure 2: Data is moved into the data warehouse as a result of the occurrence of some event.

The event may be triggered by a wide variety of occurrences:

- an occurrence of a transaction,

- the periodic passage of time,
- a threshold having been reached,
- an audit,
- a special request, etc.

An example of these triggering events might be:

- a transaction occurring - a customer makes a purchase,
- periodic passage of time - the end of the month occurs,
- a threshold being reached - total orders exceed $1,000,000 for an account for a month,
- an audit - the inventory level is taken and recorded,
- a special request - management wants to know how many customers have made more than ten orders this year.

Almost any imaginable condition is capable of triggering a snapshot to be entered into the data warehouse. Once the event occurs the snapshot (or snapshots) is taken and the snapshot is loaded into the data warehouse.

On some occasions the date the snapshot is taken is entered as part of the record. On other occasions the date of the triggering event is entered. And on other occasions both the date of the snapshot and the date of the event are entered into the data warehouse. An example of each of these three conditions will be given.

- date of the snapshot - at the end of the month all accounts have their month ending balance captured. The event is the end of the month, and the month is stored as part of the data warehouse

- date of the activity - a loan request is processed by the bank and approved. The date of approval is stored in the data warehouse.

- both date of the activity and date of the snapshot - an insurance company receives payment for premiums. The date of premium receipt is stored in the data warehouse as well as the day the data is moved into the data warehouse is stored as part of the snapshot.

The first step in designing the data warehouse is to identify the events that will trigger an entry of data into the data warehouse.

Once the events are identified, the next step is to fully specify how the data warehouse snapshots will be managed. There are many types of snapshots that can go into the data warehouse, but they all can generally be classified into one of four types:

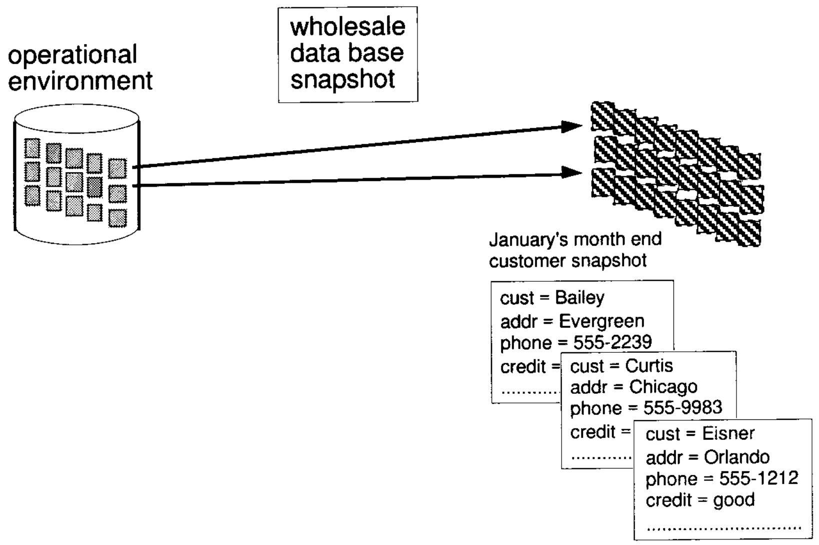
- wholesale data base snapshots,
- selected record snapshots,
- exceptional/special record snapshots, and

- cumulative snapshot records.

While these different classifications fit most environments, there are many different snapshot types that combine the types of records. Each of the basic types of snapshots will be discussed.

WHOLESALE DATA BASE SNAPSHOT

The simplest form of snapshot records in the data warehouse is that of wholesale data base snapshots. Figure 3 shows a common form of wholesale data base snapshots.


operational environment
wholesale data base snapshot
January's month end customer snapshot

cust = Bailey
addr = Evergreen
phone = 555-2239
credit =
....

cust = Curtis
addr = Chicago
phone = 555-9983
credit =
....

cust = Eisner
addr = Orlando
phone = 555-1212
credit = good
....

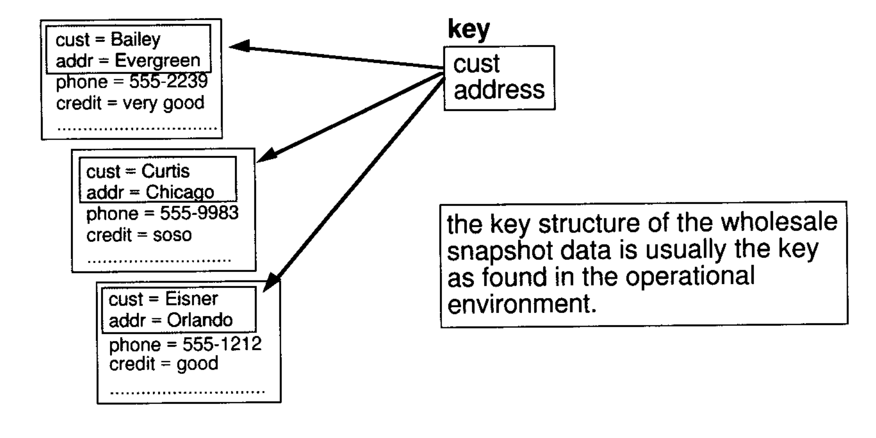
Figure 3: The simplest form of snapshot — the database is captured in its entirety as of one moment in time.

In Figure 3 at the end of every month the customer file is read in the operational environment and passed into the data warehouse. Then, throughout the month when a DSS analyst has to access customer data for informational purposes, the data warehouse holds the customer data. The snapshot taken here may or may not be a perfect image of the operational data. If the operational customer file contains fields of data or records of data that is only useful for the operational environment, then that data will be filtered out as the data passes into the data warehouse environment.

Advantages - the advantage of the wholesale data base approach is that it is simple to execute. Very little design and very little complex programming are required.

Disadvantages - the disadvantages of this approach are many. The wholesale data base approach applies only to small files. You would never think of copying over large files on a wholesale basis. A second disadvantage is that the wholesale data base - once having been captured - ages very quickly. Once the snapshot is taken, changes made to the data after the snapshot is made are not reflected in the data base. Correspondingly, because data ages quickly, the technique is not good for volatile files where data is changing constantly.

The technique of making wholesale data base snapshots implies that the key of the operational environment will be carried to the data warehouse environment, as shown in Figure 4.



the key structure of the wholesale snapshot data is usually the key as found in the operational environment.

Figure 4: It is implied that the moment the snapshot was taken is part of the key structure as well.

In Figure 4 the key of the data warehouse is the same as the key of the operational environment. It is assumed that the date the snapshot was made is part of the key structure as well (at least it is implicit in the key structure.)

SELECTED RECORD SNAPSHOTS

The most common form of snapshots in the data warehouse environment is that of selected record snapshots. Selected record snapshots are taken as the result of an event occurring. The records that will be passed to the data warehouse environment are selected based on some criteria contained within the record. Any data not being used for DSS processing is purged as data passes from the operational environment to the data warehouse environment. Figure 5 illustrates selected record snapshots.

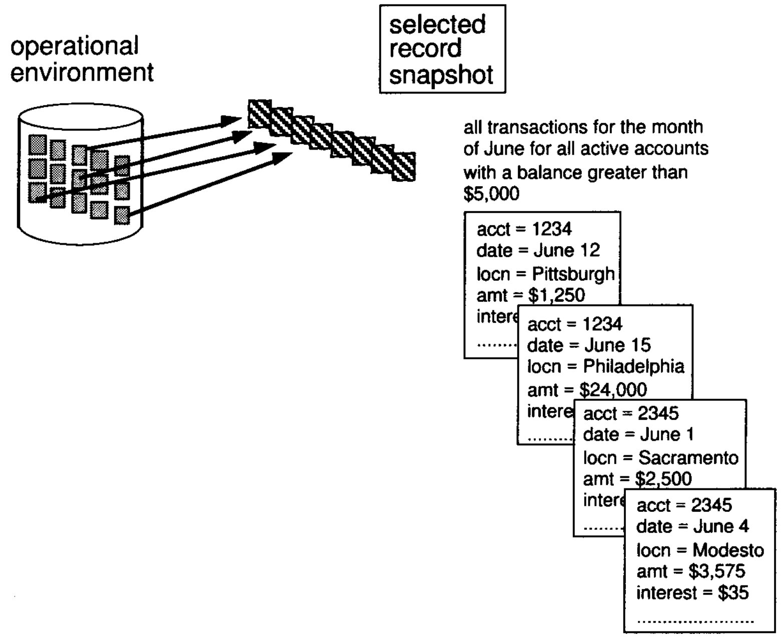


Figure 5: Selected records are identified and shipped over to the data warehouse environment.

As an example of selected record snapshots, the data architect selects all transactions which have occurred in the month of June for all active accounts with a month ending balance of greater than $5,000. The selection program reads through the operational file and upon encountering a record that meets the qualifications, moves the record to the data warehouse.

Advantages - the advantages of the selected record approach are that only a subset of operational records have to be considered for input into the data warehouse environment. There is no need to eliminate a file for consideration because the file is large. (Of course it is assumed that the operational file being scanned can be selectively scanned.) This technique is one of the most popular techniques because it applies to operational files of many sizes. The only type of file that the technique does not apply to is that of the file with MANY, MANY records where every change cannot be trapped.

A huge advantage of this approach is that it may be run not against the operational data base directly, but may be able to be run against the log and transaction files that have been created while the operational data base was being operated on.

Disadvantages - the searching of the operational file can become surprisingly complex. In addition, if care is not taken, huge amounts of data can appear in the data warehouse. Another disadvantage is that maintenance of the interface can become a burden since this technique is much more closely entwined with the data warehouse data than the wholesale data base approach.

The key structure of the selected record approach is such that the key is taken from the operational environment, as shown in Figure 6.

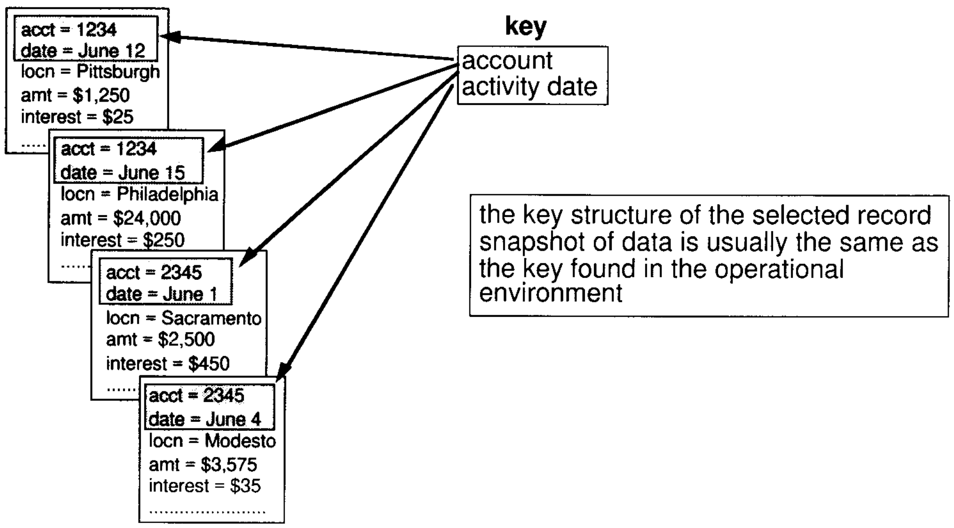


Figure 6: In almost all cases an element of time is included as part of the key structure.

In Figure 6 account and activity date are shown as the key taken from the operational environment. Note that in using this technique, sometimes TWO time stamps are created. For example, suppose the snapshots are taken at the end of the month - June 30. The selection is on (at least partially) the month in which the activity occurred. It is sometimes useful to store BOTH the transaction date and the snapshot selection date in the data warehouse.

EXCEPTIONAL / SPECIAL RECORD SNAPSHOT

The exceptional/special record snapshot technique applies where there are so many records in the operational environment that only selected records can be trapped and sent to the data warehouse environment. Unlike the previously discussed technique of trapping all changes to the operational data base environment, this technique traps only selected records. Figure 7 shows the employment of the exceptional/special record technique of selecting records for the data warehouse environment.

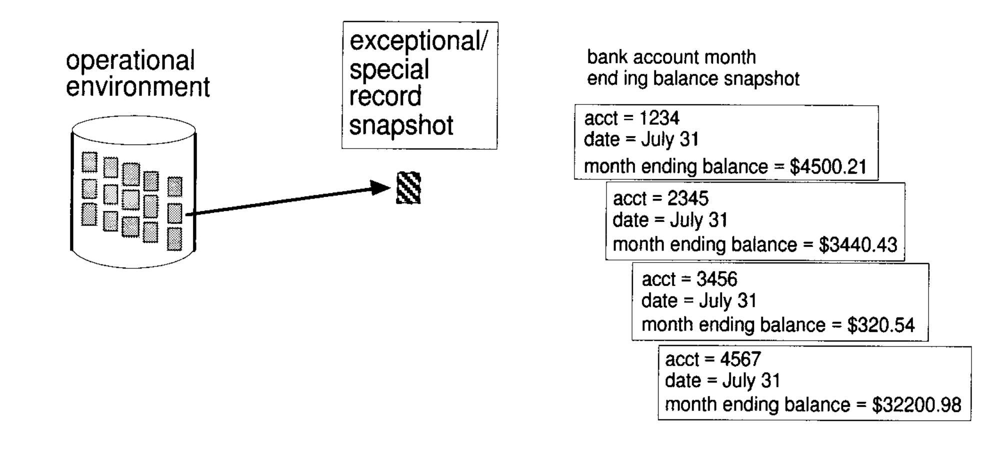

operational environment
exceptional/ special record snapshot
bank account month end ing balance snapshot
acct = 1234
date = July 31
month ending balance = $4500.21
acct = 2345
date = July 31
month ending balance = $3440.43
acct = 3456
date = July 31
month ending balance = $320.54
acct = 4567
date = July 31
month ending balance = $32200.98

Figure 7: Exceptional/special records are identified and shipped over to the data warehouse environment.

In Figure 7, at the end of the month, each account is queried and the balance of the account at the end of the month is transferred to the data warehouse environment. Note that one account may have had no activities during the month and another account may have had 200 activities during the month. Both accounts will show up as exactly one record in the data warehouse environment. No continuity of activity is assumed using this technique.

The passage of time - day end, week end, month end, etc. - are all common ways of triggering an exceptional snapshot. But the periodic passage of time is hardly the only way that snapshots are triggered. For example, when an account first goes active, or when an account becomes overdrawn, or when an account has activity over a certain threshold are all common ways that exceptional/special snapshot records are triggered.

When an exceptional/special record is triggered, there can be any amount of data that is captured as part of the snapshot. What data goes into the snapshot depends on the imagination of the data architect and/or the DSS analyst using the data warehouse environment.

Advantages - useful for large operational stores of data or where the data going into the operational environment changes quickly. Usually the programming required for the snapshots is not very complex. The primary advantage is that exceptional snapshots of data do not require much space.

Disadvantages - the primary disadvantage of exceptional snapshots of data is that they do not form a continuous record of data. Instead they make a statement about data as of only a single instant in time. For some kinds of processing the lack of a continuous record is a serious drawback. (Conversely, for other kinds of processing, lack of continuity is no problem.)

The key structure of the exceptional/special record is taken from the operational environment and in some cases blended with the moment in time when the snapshot was made. Figure 8 shows the key structure for exceptional/special records of data.

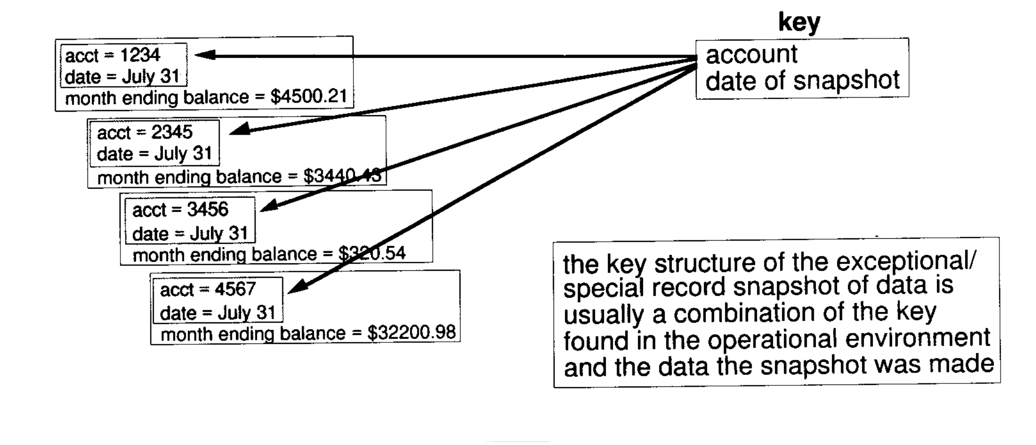

key
account
date of snapshot
acct = 1234
date = July 31
month ending balance = $4500.21
acct = 2345
date = July 31
month ending balance = $3440.43
acct = 3456
date = July 31
month ending balance = $320.54
acct = 4567
date = July 31
month ending balance = $32200.98
the key structure of the exceptional/ special record snapshot of data is usually a combination of the key found in the operational environment and the data the snapshot was made

Figure 8: Often the element of time has to be added to the key structure as data passes into the data warehouse in this type of structuring of data.

CUMULATIVE SNAPSHOT RECORDS

Cumulative snapshot records are created as a result of gathering related operational records together and summarizing or otherwise calculating the data. Figure 9 shows how a data warehouse record is created from the accumulation of multiple operational records.

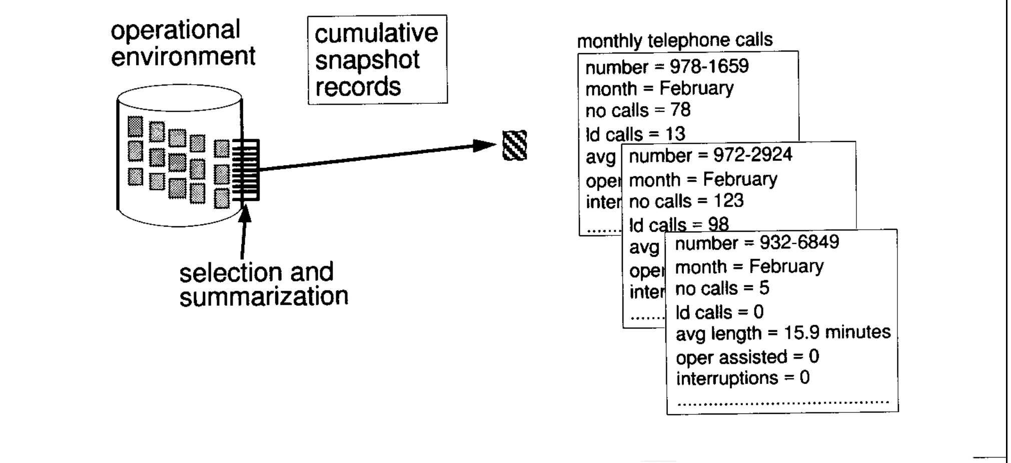


Figure 9: Cumulative snapshot records are created as a result of record selection and summarization as data warehouse records are created.

In Figure 9 monthly phone call records are accumulated by phone number and stored in the data warehouse. It is much more efficient to store and process a few records in the data warehouse than it is to have to store detailed records and calculate those records repeatedly. The degree of condensation of data that can be achieved is remarkable. Depending upon the specifics, it is not unusual to achieve a savings of three to four orders of magnitude. The savings show up in the disk storage consumed and the CPU resources needed.

Of course, as with every case of summarization of data (or the changing of the "granularity of data") whenever details are summarized, some amount of functionality is lost. Functions that require a detailed level of data are not able to be done when summarization occurs.

The issues that revolve around creating cumulative snapshots of data are:

- what time period should be used for summarization - a day? a week? a month?
- what fields should the summarized record contain?
- what algorithms should be used?

Advantages - great compaction of data.

Disadvantages - loss of functionality when gross levels of detail are required; complexity of processing; complexity of design; the need to sequence input data so that related input records physically reside next to each other.

The key of the records in the cumulative snapshot record technique is a combination of operational record data and the unit of time over which the summarization is made. Figure 10 shows the key structure common to the technique.

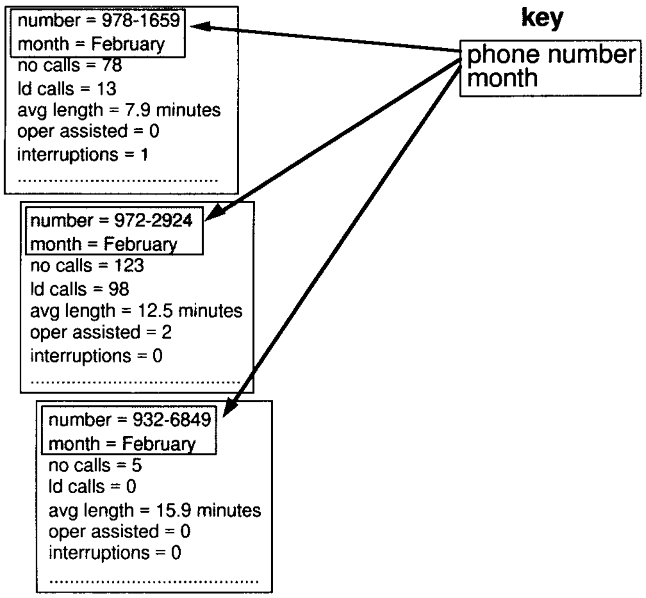


Figure 10: The key structure of the cumulative record snapshot is taken from the key of the operational environment coupled with the unit of time covered by the snapshot.

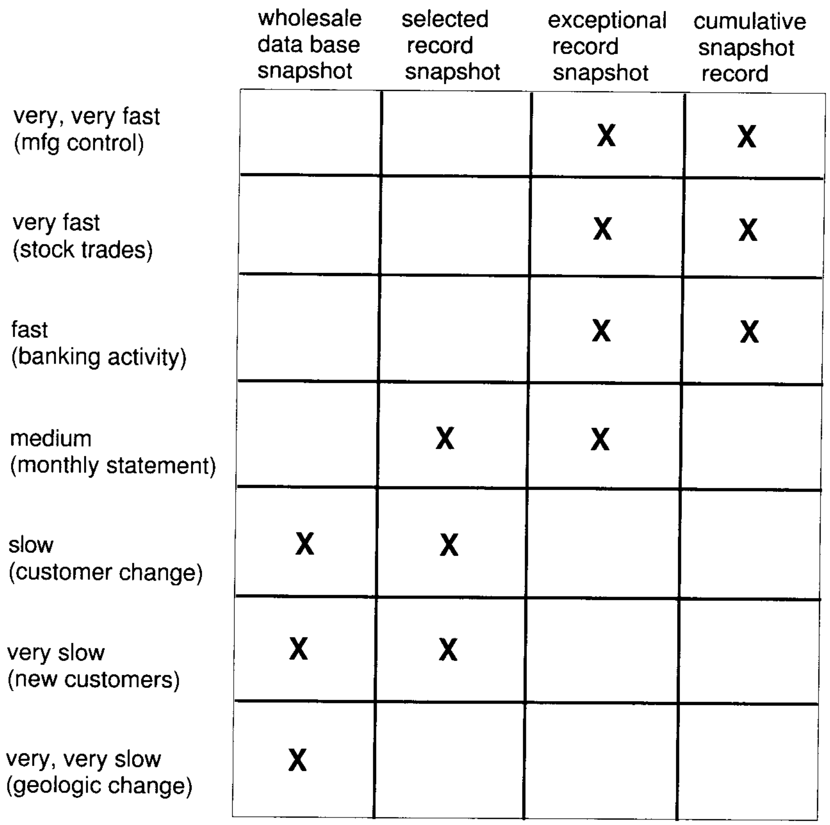
RATES OF CHANGE

Which technique of creating snapshots is appropriate is related to the amount of data to be found in the operational environment. And the amount of data found in the operational environment is related to the rate of change of variables. There is then - syllogistically - a relationship between the appropriate technique for creating data warehouse snapshots and the rate of change of variables.

In order to fully explore this relationship, one must begin with a discussion of a rate of change. Some variables change very fast and other variables change very slowly. The following are some variables and their associated rate of change:

- very, very fast change - manufacturing control, where analog measurements may be taken as much as 100 times per second.
- very fast change - stock trades, which may occur for a given stock at a rate of 100 trades per minute (or collectively, for many stocks, at a much accelerated rate.)
- fast change - banking activity - where for a given account there may be as many as three or four activities per hour (or collectively, for many accounts, at a much accelerated rate.)
- moderate change - monthly statementing of an account, where for a given account there may be one activity (i.e., statementing) which gathers up several transactions and analyses their effects on the account for the month,
- slow change - customer changes to an account - a change in address, a change in phone, a change in employer, etc, where any given customer may have two or three changes per year
- very slow change - the addition of new customers - where a customer walks in the door and opens a new account.
- very, very slow change - geologic change - where a well is drilled and the change in pools, fields, formations is noted.

These then are some representative types of changes that may be measured by their rate, which is quite different. The different rates of change can be compared to the different techniques of creating snapshots to create the following chart, shown in Figure 11.



	wholesale data base snapshot	selected record snapshot	exceptional record snapshot	cumulative snapshot record
very, very fast (mfg control)			X	X
very fast (stock trades)			X	X
fast (banking activity)			X	X
medium (monthly statement)		X	X	
slow (customer change)	X	X		
very slow (new customers)	X	X		
very, very slow (geologic change)	X			

Figure 11: A chart showing the appropriate type of snapshot based on the volatility of data.

Figure 11 shows that depending on the rate of change of variables, there is a different technique for creating data warehouse snapshots that is applicable. As a rule, the faster the rate of change, the fewer and more cumulative snapshots are applicable. Conversely, as a rule, the slower the rate of change, the more wholesale snapshots are applicable.

CONTENT AND STRUCTURE

While the timing and volume of the snapshots of operational data are extremely important, there is another dimension. (NOTE: This dimension has been covered in depth in the Tech Topic on data models and the data warehouse and will be covered in only a summary fashion here.) That dimension is the relationship of the data model to the snapshot. Figure 12 shows the dynamics of the relationship.

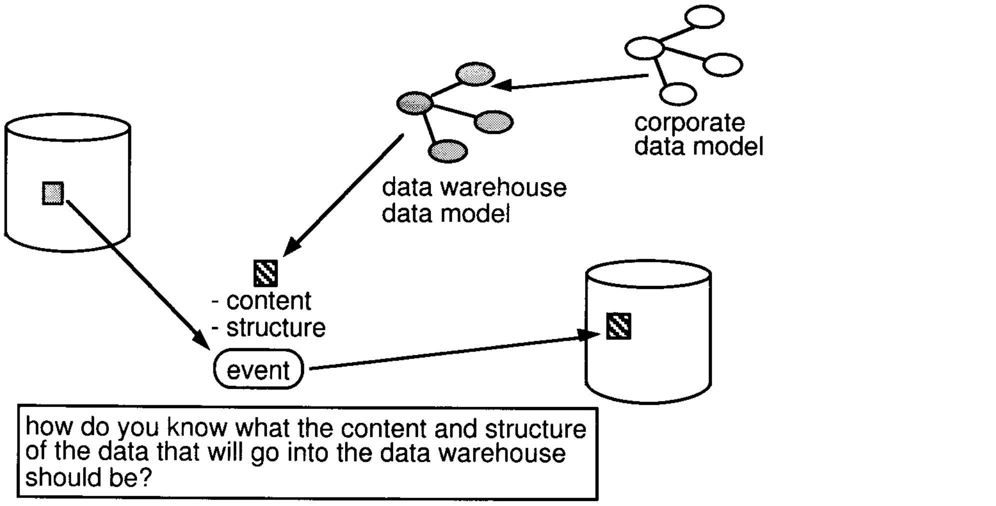

corporate
data model
data warehouse
data model
- content
- structure
event
how do you know what the content and structure of the data that will go into the data warehouse should be?

Figure 12: The data warehouse data model is the most important guid to the design of the data in the warehouse.

Figure 12 shows that the starting point for the design of the structure and contents of the snapshot is the corporate data model. The corporate data model is transformed into the data model for the data warehouse. After the data model for the data warehouse is created, the next step is the physical design of the snapshot itself.

Not all data found in the operational environment is placed in the data warehouse. Only that data that has a good chance of being used in the data warehouse is passed on. Figure 13 shows the filtering of data that occurs as data passes into the data warehouse.

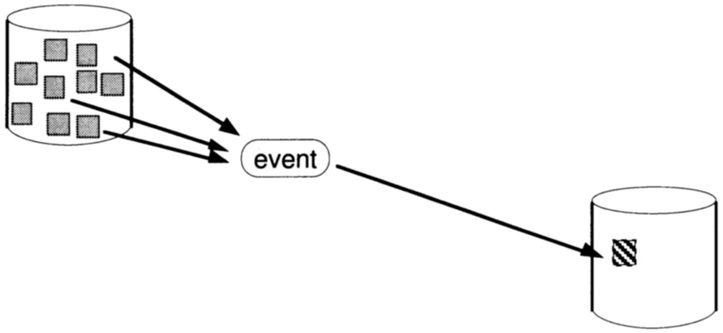


Figure 13: Both data attributes and records of data are filtered as data passes from the operational environment to the data warehouse.

Another important design issue is that of managing the relationships of data that are found in the operational environment (which has been discussed at length in a Tech Topic.) Figure 14 shows that only the artifact of the relationship is captured at the moment of taking the snapshot.

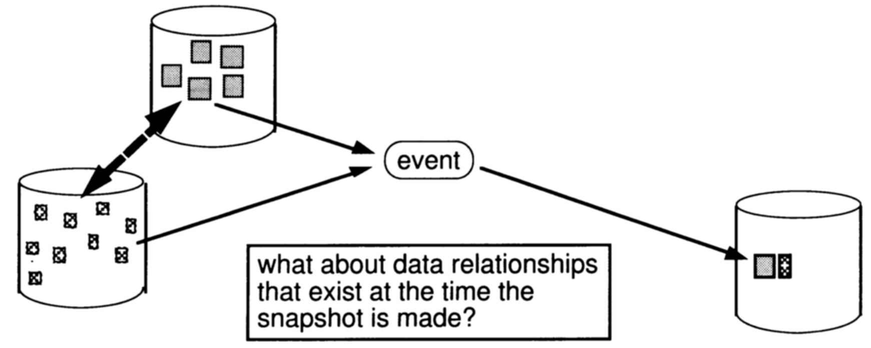

event
what about data relationships that exist at the time the snapshot is made?

Figure 14: Artifacts of the relationship are stored in the data warehouse?

SUMMARY

This discussion has addresses the issues surrounding the creation of snapshot data to go into the data warehouse. The data warehouse is populated by snapshots of data. The snapshots are created by the occurrence of an event. The event can be a wide variety of activities.

The snapshot is usually of one of four varieties:

- a wholesale snapshot of data,
- a selection of a group of records,
- a selection of exceptional or special records, or
- the creation of a cumulative record.