

## Keys in Relational Databases: Theory and Practice. Part 4. Key Problem #1. Primary Keys Which Also Describe.

Dr. Tom Johnston  
MindfulData.com

Many enterprises are undertaking primary key reengineering projects for their core business tables. These tables keep information on the enterprise's suppliers, partners, customers, orders, locations, employees, and so forth. The purpose of these projects is to replace the *natural keys* of these tables with system-generated *surrogate keys*.

### *Natural Keys, Intelligent Keys and Surrogate Keys.*

A *natural key* is a primary keys at least one of whose columns contains business data.

**Figure 17. Definition: “Natural Key”.**

*Business data* is data whose values convey useful information to a business user.

**Figure 18. Definition: “Business Data”.**

To clarify these definitions, let's look at a few different examples. Consider, for example, a Customer table.

<u>cust-nbr</u>	other-data
C1	.....

C2	.....
----	-------

**Figure 19. Sample Data: a Primary Key.**

Is cust-nbr in this sample table a natural key, according to this definition? The answer is: it depends. In this example, the values shown are “C1” and “C2”. These values are probably not meaningful to a business user. They would be meaningful if she knew exactly who customer C1 and C2 were. But is she likely to recognize each of her company’s ten-thousand customers from customer numbers like these? Clearly, she is not.

So in this case, the values for cust-nbr are probably system-generated values, with a new one generated for each new customer added to the table. But being system-generated isn’t what prevents data from being business data. The reason cust-nbr isn’t business data is that it doesn’t convey any useful information to the business user. Not being business data, it is not a natural key. It is, instead, a *surrogate key* – a concept we will examine more closely in a moment.

Now let’s look at another Customer table.

<u>cust-nbr</u>	other-data
DNV-1723	.....
ATL-0456	.....

**Figure 20. Sample Data: Another Primary Key.**

Is cust-nbr in this sample table a natural key, according to this definition? Again, it depends. Let’s suppose that all cust-nbr values we find in the table have this two-part structure, separated by a hyphen. This suggests that one or both of these pieces of this single column have their own individual meaning. Let us suppose that looking at a large number of cust-nbr values, we find that the alphabetic piece seems to be based on about a hundred or so distinct values, and that the numeric piece looks like an added-on sequence

number. Let's suppose, further, that SMEs confirm this impression, as does the code which generates these cust-nbrs.

Knowing this, plus a little bit about the alphabetic values, a business user would know that customer ATL-0456 is a customer headquartered in Atlanta. She would also know (but probably doesn't care) that this is the 456<sup>th</sup> Atlanta-headquartered customer added to the Customer table. So in this example, cust-nbr is meaningful. It is business data. And so it is a natural key.

It is also an intelligent key.

An *intelligent key* is a primary key that contains at least one column with the following property. Although each instance of the column is a single atomic data element to the DBMS, there is a pattern of data inside each such instance in which one or more parts of that data are separately meaningful to the business user.

**Figure 21. Definition: "Intelligent Key".**

Since the first part of cust-nbr, in this table, is meaningful to the business user – telling her in which city the customer is headquartered – but there is one additional internal component to the column, this cust-nbr is an intelligent key.

But many primary keys contain more than one column. So let's look again at our Invoice Line Item table, but this time to illustrate a primary key, not (as we did earlier) to illustrate a non-primary key semantic key.

<u>invc-nbr</u>	<u>line-nbr</u>	prod-code	qty	unit-price
031256	3	4X18 WM Rolls	24	\$225.00

171804	3	4X18 WM Rolls	24	\$225.00
--------	---	------------------	----	----------

**Figure 22. Sample Data: a Third Primary Key.**

In this table, *invc-nbr* and *line-nbr* make up the primary key. *invc-nbr* is a foreign key back to an Invoice Header table. *Line-nbr* keeps the lines of each invoice distinct, and like *invc-nbr*, appears on the printed invoice.

Both components of this primary key are business data. In understanding why this is so, we will better understand the definition of “business data” as “data whose values convey useful information to a business user”. First of all, consider *invc-nbr*, and assume that it is a system-generated sequence number. If it’s system-generated, how can it be business data?

Notice that I said that *invc-nbr* “appears on the printed invoice”. By doing so, this system-generated number has, as it were, escaped the confines of the relational database in which it plays the roles of primary key and foreign key. It has escaped into the rest of the company. Appearing on printed invoices, customers will use it when they call to question a charge on an invoice. Company personnel will refer to the invoice by means of its *invc-nbr* value when discussing the invoice over the phone and in emails to one another.

This also explains why *line-nbr* is business data. The “0456” in “ATL-0456” was not business data because it didn’t convey any useful information about the customer. But the “3” in *line-nbr* *is* business data because it is used throughout the company, and even outside it, to designate the third line on an invoice. For example, a customer might call up and complain “I should have received a 10% discount on line 3 of invoice 171804.”

Being business data, this primary key is also a natural key. Since none of its component columns seem to contain semantically distinct components, however, it is *not* an intelligent key.

As a final example, consider this primary key.

<u>cust-nbr</u>	other-data
19930ABFCC876237	.....
337BD8C8FF83265A	.....

**Figure 23. Sample Data: a Fourth Primary Key.**

These cust-nbr values are obviously system-generated. But just as obviously, they will never be used as business data. Business users are just not going to refer to customers with numbers like these, and they aren't going to expose such numbers to the customers themselves. So this cust-nbr is *not* business data. Therefore, it is neither a natural key, in general, nor an intelligent key in particular. It is, instead, a surrogate key.

A *surrogate key* is a primary key none of whose component columns, in whole or in part, are business data.

**Figure 24. Definition: "Surrogate Key".**

Let's look, now, at the cost of making changes to natural keys. That cost is usually very high, and always for the same two reasons:

1. When the natural key primary keys of a table are changed, all their foreign key instances must be changed, also.
2. Often, these changes cannot be done in real-time, as part of a single atomic transaction.

**Figure 25. Cost Components for Changing Natural Keys.**