# 6.2

# BIG DATA/EXISTING SYSTEMS INTERFACE

One of the challenges of information systems is determining how they all fit together. In particular, how does Big Data fit with the existing systems environment? There is no question that Big Data brings new opportunities for information and decision making to the organization. And there is no question that Big Data has great promise. But Big Data is not a replacement for the existing systems environment. In fact Big Data accomplishes one task and the existing systems environment accomplishes another task. They are (or should be) complementary to each other.

So exactly how does Big Data need to interface with and interact with the existing systems environment?

## The Big Data/Existing Systems Interface

Figure 6.2.1 shows the recommended way in which Big Data and existing systems interface with each other and the overall system flow between Big Data and the existing systems environment. Each of the interfaces will be discussed in detail.

Raw Big Data is divided into two distinct sections (see the "great divide"). There is repetitive raw Big Data and nonrepetitive raw Big Data. Repetitive raw Big Data is handled entirely differently than nonrepetitive raw Big Data.

## The Repetitive Raw Big Data/Existing Systems Interface

The interface from repetitive raw Big Data to the existing systems environment in some ways is the simplest interface. In many ways this interface is like a "distillation" process. The mass of data found in raw repetitive Big Data is winnowed down – distilled – into the few records that are of interest.

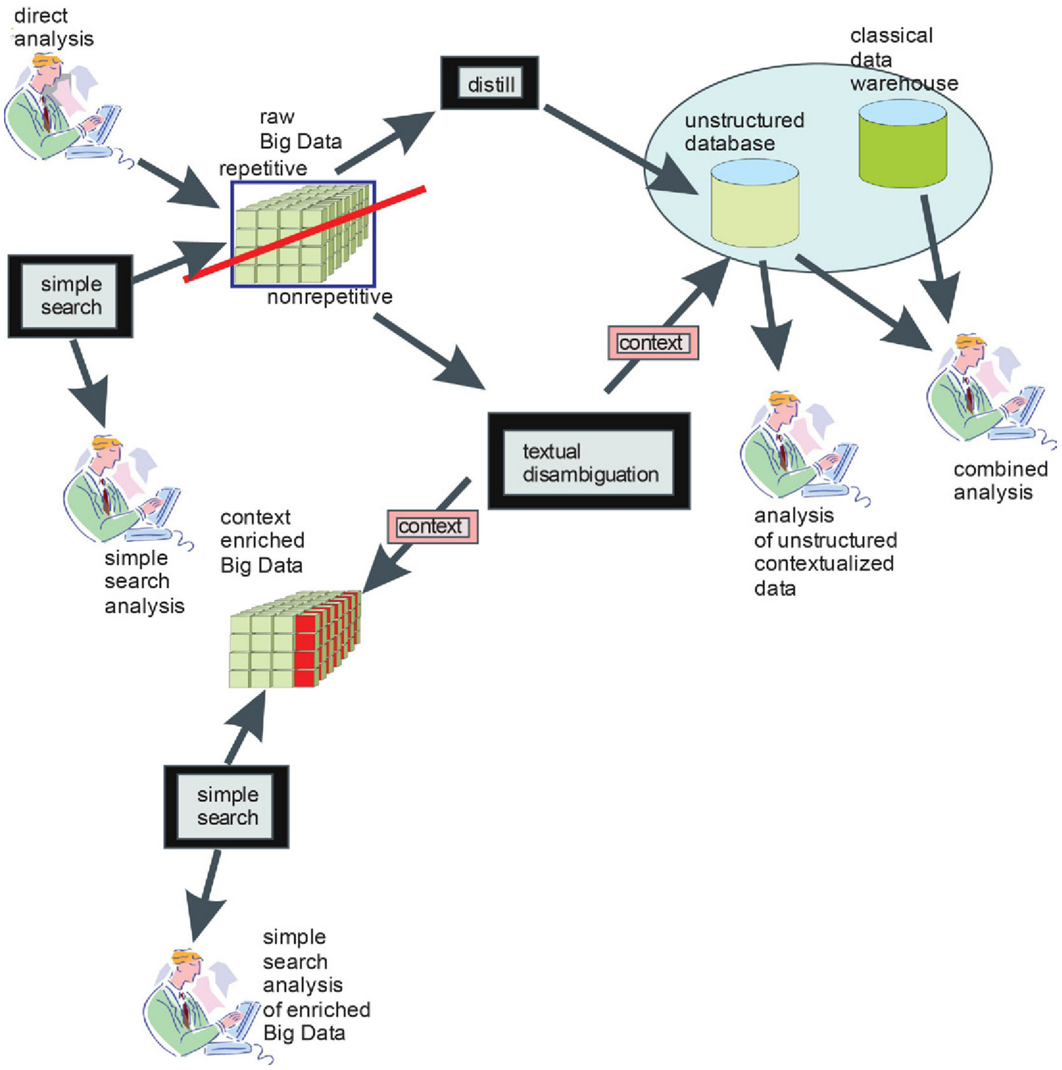

**Figure 6.2.1**

The repetitive raw Big Data is processed by parsing each record. And when the records that are of interest are located, the records of interest are then edited and passed to the existing systems environment. In such a fashion the data that is of interest is distilled from the mass of records found in the raw repetitive Big Data environment. One assumption made by this interface is that the vast majority of records found in the repetitive component of raw Big Data will not be passed to the existing systems environment. The assumption is that only a few records of interest are to be found.

In order to explain this assumption, consider the following cases of manufacturing, telephone calls, log tape analysis, and metering.

A manufacturer makes a product. The quality of the product is quite high. On the average only 1 out of 10,000 products is defective. However, the defective products are still a bother. All the product manufacturing information is stored in Big Data. But only the information about the defective products is brought to the existing systems environment for further analysis. In this case, based on a percentage basis, very little data is brought to the existing systems environment.

Millions of telephone calls are made on a daily basis. But of those millions of telephone calls, only a handful – maybe three or four – are of interest for call record details. Only the phone calls that are of interest are brought from the Big Data environment to the existing systems environment

For log tape analysis, a log tape of transactions is created. In a day, tens of thousands of log tape entries are created. But only a few hundred entries on the log tape are of interest. Those few hundred log tape entries that are of interest are the only entries that find their way back into the existing systems environment for further analysis.

An organization collects metering data. The vast majority of the metering activity is normal and not of particular interest. But on a few days of the year certain metering data reacts in an unexpected manner. Only those readings that have reacted abnormally are brought to the existing systems environment for further analysis.

And there are many more examples of repetitive raw Big Data being examined for exceptional data.

As a rule when data goes from the Big Data environment to the existing systems environment, it is convenient to place the data in a data warehouse. However, if there is a need, data can be sent elsewhere in the existing systems environment.

## Exception-Based Data

Once the data in the raw repetitive Big Data environment is selected (which is usually chosen on an exception basis and is then moved to the existing systems environment) the exception-based data can undergo all sorts of analysis, such as:

- Pattern analysis. Why are the records that have been chosen exceptional? Is there a pattern of activity external to the records that match with the behavior of the records?

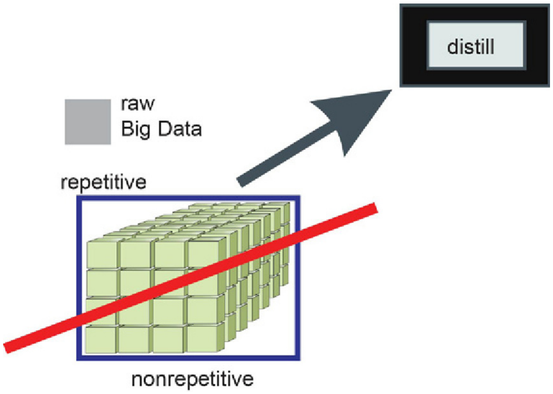

**Figure 6.2.2**

- Comparative analysis. Is the number of exceptional records increasing? Decreasing? What other events are happening concurrent to the collection of the exceptional records?
- Growth and analysis of exceptional records over time. Over time what is happening to the exceptional records that have been collected from Big Data?

And there are *many* more ways to analyze the data that has been collected.

Figure 6.2.2 shows the interface from Big Data to the existing systems environment.

## The Nonrepetitive Raw Big Data/Existing Systems Interface

The interface from the nonrepetitive raw Big Data environment is one that is very different from the repetitive raw Big Data interface. The first major difference is in the percentage of data that is collected. Whereas in the repetitive raw Big Data interface only a small percentage of the data is selected, in the nonrepetitive raw Big Data interface the majority of the data is selected. This is because there is business value in the majority of the data found in the nonrepetitive raw Big Data environment whereas there is little business value in the majority of the repetitive Big Data environment. But there are other major differences as well.

The second major difference in the environments is in terms of context. In the repetitive raw Big Data environment, context is usually obvious and easy to find. In the nonrepetitive raw Big Data environment context is not obvious at all and is not easy to find. It

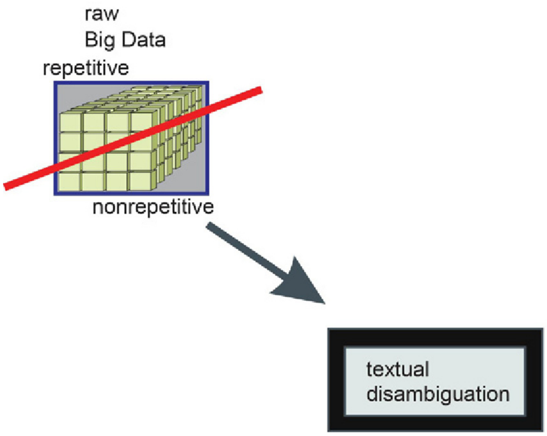

Figure 6.2.3

is noted that context is in fact there in the nonrepetitive Big Data environment; it just is not easy to find and is anything but obvious.

In order to find context, the technology of textual disambiguation is needed. Textual disambiguation reads the nonrepetitive data in Big Data and derives context from the data. (See the chapter on textual disambiguation and taxonomies for a more complete discussion of deriving context from nonrepetitive raw Big Data.)

While *most* of the nonrepetitive raw Big Data is useful, some percentage of data is not useful and is edited out by the process of textual disambiguation. Once the context is derived, the output can then be sent to either of the existing systems environments.

Figure 6.2.3 shows the interface from nonrepetitive raw Big Data to textual disambiguation.

## Into the Existing Systems Environment

Once data has come from nonrepetitive raw Big Data and has passed through textual disambiguation, the data can be passed to the existing systems environment.

As the data is passed through textual disambiguation, it is greatly simplified. Context is inferred and each unit of text that passes the filtering process is turned into a flat file record. The flat file record is reminiscent of a standard relational record. There is a key and dependent data, as is found in a relational format.

The output can be sent to a load utility so that the output data can be placed in whatever DBMS is desired. Typical output DBMSs include Oracle, Teradata, UDB/DB2, and SQL Server.

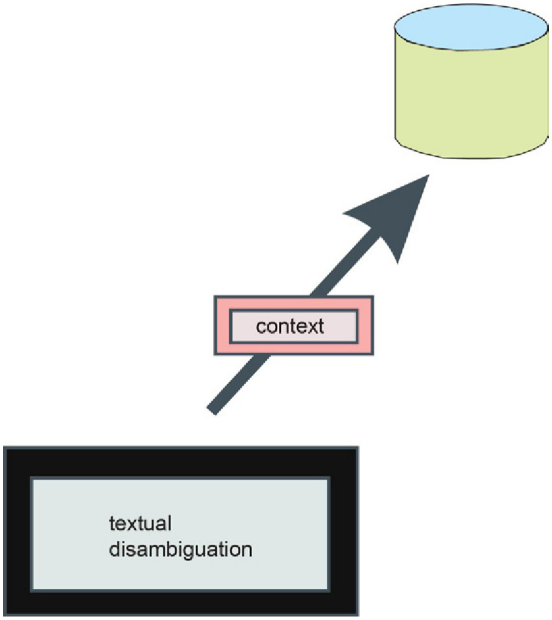

**Figure 6.2.4**

Figure 6.2.4 shows the movement of data into the existing systems environment in the form of a standard DBMS.

## The "Context-Enriched" Big Data Environment

The other route that data can take after it passes through textual disambiguation is that the output of data can be placed back into Big Data. There may be several reasons for wanting to send output back into Big Data. Some of the reasons include:

- The volume of data. There may be a lot of output from textual disambiguation. The sheer volume of data may dictate that the output data be placed back into the Big Data environment.
- The nature of the data. In some cases the output data may have a natural fit with the other data placed in the Big Data environment. Future analytical processing may be greatly enhanced by placing output data back into Big Data.

In any case, after data passes through textual disambiguation and is placed back into Big Data, it enters Big Data in a very different state. When data has passed through textual disambiguation and is placed back into the Big Data environment, it is placed into the environment with the context of the data clearly identified and prominently a data part of the data in Big Data.

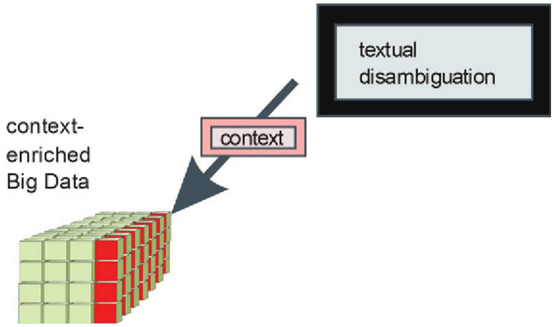

Figure 6.2.5

By placing the output of textual disambiguation back into Big Data, there now is a section of Big Data that can be called the "context-enriched" section of Big Data. From a structural standpoint, the context-enriched component of Big Data looks to be very similar to repetitive raw Big Data. The only difference is that the context-enriched component of Big Data has context open and obvious and attached to the base data in this component of Big Data.

Figure 6.2.5 shows that output data from textual disambiguation can be placed back into Big Data. Another perspective of the Big Data environment is shown in Figure 6.2.6.

Figure 6.2.6 shows that there is the division of Big Data into the repetitive and nonrepetitive sections. However, in the repetitive section, it is seen that when context-enriched Big Data is added to the Big Data environment, context-enriched data simply becomes another type of repetitive data. Stated differently, there are two types of repetitive data in Big Data – simple repetitive data and context-enriched repetitive data.

This division becomes important when doing analytical processing. Repetitive data is analyzed in a completely different fashion than context-enriched repetitive data.

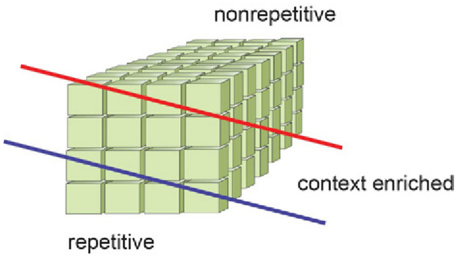

Figure 6.2.6

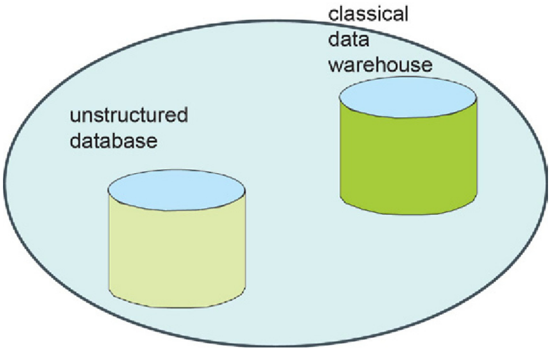

Figure 6.2.7

## Analyzing Structured Data/Unstructured Data Together

The final interface of interest in the Big Data environment is that data that has come from Big Data either through the distillation process or textual disambiguation. The data that arrives here can be placed into a standard DBMS.

Figure 6.2.7 shows the database that has been created from unstructured data being placed in the same environment as the classical data warehouse. Of course the data in the classical data warehouse has been created from structured data entirely.

Figure 6.2.7 shows that data whose origin is quite different can be placed in the same analytical environment. The DBMS may be Oracle or Teradata. The operating system may be Windows or Linux. In any case doing analytical processing against the two databases is as easy as doing a relational join. In such a manner it is easy and natural to do analytical processing against data from the two different environments. This means that structured data and unstructured data can be used together analytically. By combining these two types of data together, entirely new vistas of analytical processing open up.