Chapter 11

Appliances and Big Data Warehouses

In This Chapter

- Defining the big data warehouse
- Creating a new model to support changing requirements
- The data warehouse big data hybrid system
- Evaluating deployment models for big data warehousing

The concept of the data warehouse originated almost 30 years ago. The data warehouse was intended to solve a big problem for customers that had many operational systems that were siloed. Increasingly, management wanted to replace inefficient decision-support systems with a more streamlined model. Companies wanted to be able to have a single architectural model that would make it much easier to make business decisions. This approach, whether in the form of a full data warehouse or a more limited data mart, has been the norm. However, with the advent of big data, the data warehouse concept is now changing so that it can be applied to new use cases. The traditional data warehouse will continue to survive and thrive because it is very useful in analyzing historical operational data for decision making. However, new types of data warehouses will be optimized for the big data world. In this chapter, we give you a perspective on how the data warehouse has evolved to support the characteristics of big data.

Integrating Big Data with the Traditional Data Warehouse

Unlike traditional operational database systems and applications, the data warehouse was used by business line and financial analysts to help make decisions about the direction of a business strategy. Data had to be gathered

from a variety of relational database sources and then ensured that the metadata was consistent, and that the data itself was clean and then well integrated. Bill Inmon, considered the father of the modern data warehouse, established a set of principles of the data warehouse, which included the following characteristics:

- It should be subject oriented.
- It should be organized so that related events are linked together.
- The information should be nonvolatile so that it cannot be inadvertently changed.
- Information in the warehouse should include all the applicable operational sources. The information should be stored in a way that has consistent definitions and the most up-to-date values.

Optimizing the data warehouse

Data warehouses have traditionally supported structured data and have been closely tied to the operational and transactional systems of the enterprise. These carefully constructed systems are now in the midst of significant changes as organizations try to expand and modify the data warehouse so that it can remain relevant in the new world of big data. While the worlds of big data and the data warehouse will intersect, they are unlikely to merge anytime soon. You can think of the traditional data warehouse as a system of record for business intelligence, much like a customer relationship management (CRM) system or an accounting system. These systems are highly structured and optimized for specific purposes. In addition, these systems of record tend to be highly centralized. Figure 11-1 shows a typical approach to data flows with warehouses and marts.

Differentiating big data structures from data warehouse data

Organizations will inevitably continue to use data warehouses to manage the type of structured and operational data that characterizes systems of record. These data warehouses will still provide business analysts with the capability to analyze key data, trends, and so on. However, the advent of big data is both challenging the role of the data warehouse and providing a complementary approach. You might want to think about the relationship between the data warehouse and big data as merging to become a hybrid structure. In this hybrid model, the highly structured optimized operational data remains in the tightly controlled data warehouse, while the data that is highly distributed and subject to change in real time is controlled by a Hadoop-based (or similar NoSQL) infrastructure.

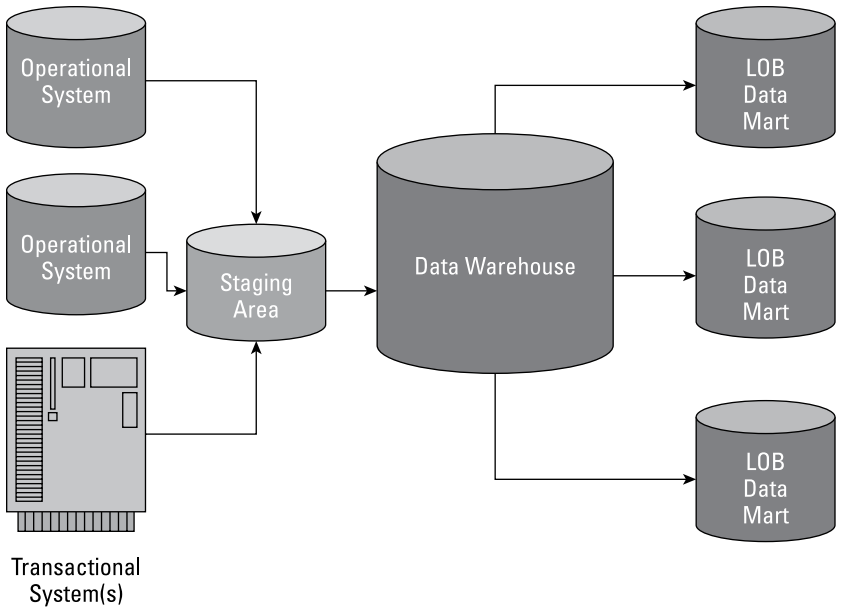

Figure 11-1: Data flows for traditional data warehouses and data marts.

It is inevitable that operational and structured data will have to interact in the world of big data, where the information sources have not (necessarily) been cleansed or profiled. Increasingly, organizations are understanding that they have a business requirement to be able to combine traditional data warehouses with their historical business data sources with less structured and vetted big data sources. A hybrid approach supporting traditional and big data sources can help to accomplish these business goals.

Examining a hybrid process case study

Imagine that you are in charge of data management for an online travel site. Your company offers a wide range of services, including air travel, cruises, hotels, resorts, and more. The company offers these services in many different ways. For example, a public website is available that includes reviews of various trips, hotels, and so on. This website has relationships with various related companies that offer services such as trip insurance and local tour services. Specialized sites exist for different countries. In addition, a corporate travel service is customized for large companies. Needless to say, this travel company has to manage a huge volume of data and be able to present it differently depending on who is interacting with it. A data warehouse is used by the company to track its transactions and operational data. However, the data warehouse does not keep track of web traffic. Therefore, the company used web analytics solutions to capture customer interactions.

For example, what did the customer click on? What offers were made available to different customers, and which ones did they select? Was price the most important factor? Did customers like to be able to design their own travel packages, or were they more likely to purchase predesigned tours? Were some locations attracting more customers while other geographies were less popular? Which partners were attracting the most revenue?

While much of this data could be incredibly valuable for those planning for the future, it was not practical for the company to store all or most of this data in the data warehouse. As a result, most of this data was thrown away after it was examined. Soon the company realized that it would be valuable to keep as much of this data as possible to understand the changes and nuances of the business.

The information management team decided that rather than building a customized data warehouse to store this data, it would leverage a Hadoop distributed computing approach based on commodity servers. Now the company is able to keep all the data from the web interactions. This data is now stored across a vast array of servers running Hadoop and MapReduce. Leveraging tools such as Flume and Sqoop, the team is able to move data into and out of Hadoop and push it into a relational model so that it can be queried with familiar SQL tools.

Now the company is able to change its business offerings quickly when it is apparent that a demographic group of customers wants certain new services. The company can also predict changes in airfare that will impact how packages are priced. Some of this data remains in the Hadoop framework environment and is updated in near real time. Other data elements are cleansed and then are moved into the data warehouse so that the data is used to compare the historical information about customers and partners to the new data. The existing warehouse provides the context for the business while the Hadoop environment tracks what is happening on a minute-to-minute basis. The combination of the system-of-record approach with the data warehouse with the dynamic big data system provides a tremendous opportunity for the company to continue to evolve its business based on analyzing the massive amount of data generated by its web environments. Figure 11-2 depicts an example approach to hybridizing traditional and big data warehousing.

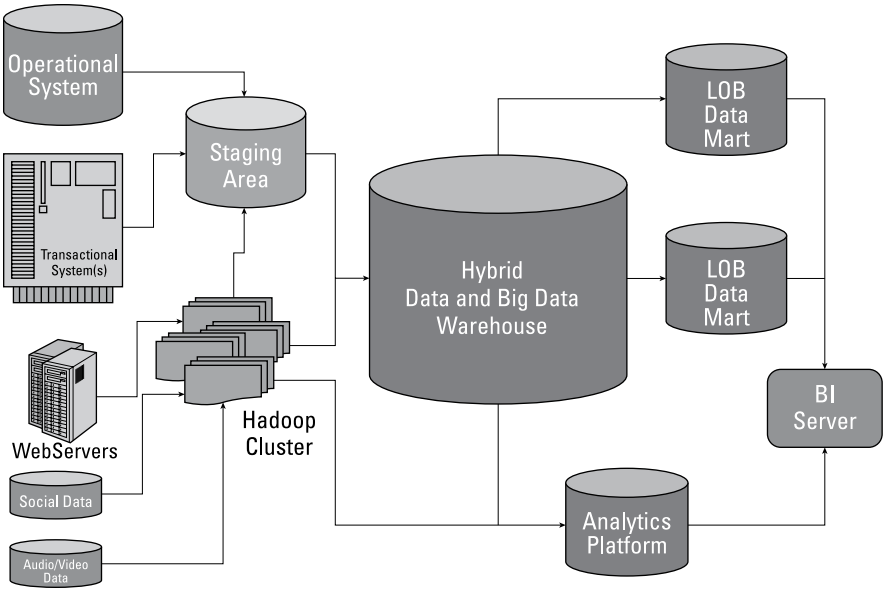

Figure 11-2: Data flows for big data warehouses and data marts.

Big Data Analysis and the Data Warehouse

From the preceding examples, you find value in bringing the capabilities of the data warehouse and the big data environment together. You need to create a hybrid environment where big data can work hand in hand with the data warehouse. First it is important to recognize that the data warehouse as it is designed today will not change in the short term. Therefore, it is more pragmatic to use the data warehouse for what it has been designed to do — provide a well-vetted version of the truth about a topic that the business wants to analyze. The warehouse might include information about a particular company's product line, its customers, its suppliers, and the details of a year's worth of transactions. The information managed in the data warehouse or a departmental data mart has been carefully constructed so that metadata is accurate. With the growth of new web-based information, it is practical and often necessary to analyze this massive amount of data in context with historical data. This is where the hybrid model comes in.

Certain aspects of marrying the data warehouse with big data can be relatively easy. For example, many of the big data sources come from sources that include their own well-designed metadata. Complex e-commerce sites include well-defined data elements (customer, price, and so on). Therefore, when conducting analysis between the warehouse and the big data source, the information management organization is working with two data sets with carefully designed metadata models that have to be rationalized.

Of course, in some situations, the information sources lack explicit metadata. Before an analyst can combine the historical transactional data with the less structured big data, work has to be done. Typically, initial analysis of petabytes of data will reveal interesting patterns that can help predict subtle changes in business or potential solutions to a patient's diagnosis. The initial analysis can be completed leveraging tools like MapReduce with the Hadoop distributed file system framework. At this point, you can begin to understand whether it is able to help evaluate the problem being addressed. In the process of analysis, it is just as important to eliminate unnecessary data as it is to identify data relevant to the business context. When this phase is complete, the remaining data needs to be transformed so that metadata definitions are precise. In this way, when the big data is combined with traditional, historical data from the warehouse, the results will be accurate and meaningful.

The integration lynchpin

To make the process we describe practical requires a well-defined data integration strategy. We cover the issue of data integration in detail in Chapter 15. While data integration is a critical element of managing big data, it is equally important when creating a hybrid analysis with the data warehouse. In fact, the process of extracting data and transforming it in a hybrid environment is very similar to how this process is executed within a traditional data warehouse. In the data warehouse, data is extracted from traditional source systems such as CRM or ERP systems. It is critical that elements from these various systems be correctly matched.

Rethinking extraction, transformation, and loading

In the data warehouse, you often find a combination of relational database tables, flat files, and nonrelational sources. A well-constructed data warehouse will be architected so that the data is converted into a common

format, allowing queries to be processed accurately and consistently. The extracted files must be transformed to match the business rules and processes of the subject area that the data warehouse is designed to analyze. For example, it is common to have the concept of a purchase price as a calculated field in a data warehouse because it will be used in many of the queries used by management. Processes may exist within the warehouse to validate that the calculations are accurate based on business rules. While these ideas are foundational to the data warehouse, it is also a key principle of marrying the warehouse to big data. In other words, the data has to be extracted from the big data sources so that these sources can safely work together and produce meaningful results. In addition, the sources have to be transformed so that they are helpful in analyzing the relationship between the historical data and the more dynamic and real-time data that comes from big data sources.

Loading information in the big data model will be different than what you would expect in a traditional data warehousing model. In the data warehouse, after data has been codified, it is never changed. A typical data warehouse will provide the business with a snapshot of data based on the need to analyze a particular business issue that requires monitoring, such as inventory or sales quotas. Loading information can be dramatically different with big data. The distributed structure of big data will often lead organizations to first load data into a series of nodes and then perform the extraction and transformation. When creating a hybrid of the traditional data warehouse and the big data environment, the distributed nature of the big data environment can dramatically change the capability of organizations to analyze huge volumes of data in context with the business.

Changing the Role of the Data Warehouse

It is useful to think about the similarities and differences between the way data is managed in the traditional data warehouse and when the warehouse is combined with big data.

Similarities between the two data management methods include

- Requirements for common data definitions
- Requirements to extract and transform key data sources
- The need to conform to required business processes and rules

Differences between the traditional data warehouse and big data include

- The distributed computing model of big data will be essential to allowing the hybrid model to be operational.
- The big data analysis will be the primary focus of the efforts, while the traditional data warehouse will be used to add historical and transactional business context.

Changing Deployment Models in the Big Data Era

With the advent of big data, the deployment models for managing data are changing. The traditional data warehouse is typically implemented on a single, large system within the data center. The costs of this model have led organizations to optimize these warehouses and limit the scope and size of the data being managed. However, when organizations want to leverage the massive amount of information generated by big data sources, the limitations of the traditional models no longer work. Therefore, the data warehouse appliance has become a practical method of creating an optimized environment to support the transition to new information management.

The appliance model

When companies need to combine their data warehouse structure with big data, the appliance model can be one answer to the problem of scaling. Typically, the appliance is an integrated system that incorporates hardware (typically in a rack) that is optimized for data storage and management. Because they are self-contained, appliances can be relatively easy and quick to implement, as well as offer lower costs to operation and maintain, Therefore, the system will be preloaded with a relational database, the Hadoop framework, MapReduce, and many of the tools that help ingest and organize data from a variety of sources. It also incorporates analytical engines and tools to simplify the process of analyzing data from multiple sources. The appliance is therefore a single-purpose system that typically includes interfaces to make it easier to connect to an existing data warehouse.

The cloud model

The cloud is becoming a compelling platform to manage big data and can be used in a hybrid environment with on-premises environments. Some of the new innovations in loading and transferring data are already changing the potential viability of the cloud as a big data warehousing platform. For example, Aspera, a company that specializes in fast data transferring between networks, is partnering with Amazon.com to offer cloud data management services. Other vendors such as FileCatalyst and Data Expedition are also focused on this market. In essence, this technology category leverages the network and optimizes it for the purpose of moving files with reduced latency. As this problem of latency in data transfer continues to evolve, it will be the norm to store big data systems in the cloud that can interact with a data warehouse that is also cloud based or a warehouse that sits in the data center.

Examining the Future of Data Warehouses

The data warehouse market has indeed begun to change and evolve with the advent of big data. In the past, it was simply not economical for companies to store the massive amount of data from a large number of systems of record. The lack of cost-effective and practical distributed computing architectures meant that a data warehouse had to be designed so that it could be optimized to operate on a single unified system. Therefore, data warehouses were purpose-built to address a single topic. In addition, the warehouse had to be carefully vetted so that data was precisely defined and managed. This approach has made data warehouses accurate and useful for the business to query these data sources. However, this same level of control and precision has made it difficult to provide the business with an environment that can leverage much more dynamic big data sources. The data warehouse will evolve slowly.

Data warehouses and data marts will continue to be optimized for business analysis. However, a new generation of offerings will combine historical and highly structured data stores with different stages of big data stores. First, big data stores will provide the capability to analyze huge volumes of data in near real time. Second, a big data store will take the results of an analysis and provide a mechanism to match the metadata of the big data analysis to the requirements of the data warehouse.