# Chapter 11
# Mining Databases on the Web

## INTRODUCTION

While Chapters 9 and 10 provided an overview of Web data mining, this chapter discusses aspects of mining the databases on the Web. Essentially, we use the technologies discussed in Chapter 5 to describe the impact of Web mining.

As we have mentioned, there is a lot of data on the Web, some in databases, some in files or other data sources. The databases may be semistructured or they may be relational, object, or multimedia databases. These databases have to be mined so that useful information is extracted.

## CONCEPTS IN WEB DATABASE MINING

A simple illustration of Web database mining is shown in Exhibit 1. Note that we also discussed this figure in Chapters 3 and 5. The idea is that there are databases on the Web and these databases have to be mined to extract patterns and trends.

While we could use many of the data mining techniques to mine the Web databases, the challenge is to locate the databases on the Web. Furthermore, the databases may not be in the format that we need for mining the data. We may need mediators to mediate between the data miners and the databases on the Web. This is illustrated in Exhibit 2.

In Chapter 5, we discussed other aspects of Web database management, including processing queries and carrying out transactions, as well as metadata management, data warehousing, and data distribution. We discuss metadata mining as well as mining distributed databases later in this chapter. In the next section, we discuss Web database functions and data mining.

### Web Database Management Functions and Data Mining

We discussed Web data representation and Web database functions in Chapter 5. In this chapter, we examine the impact of data mining. As mentioned previously, data could be in relational, object, or semistructured

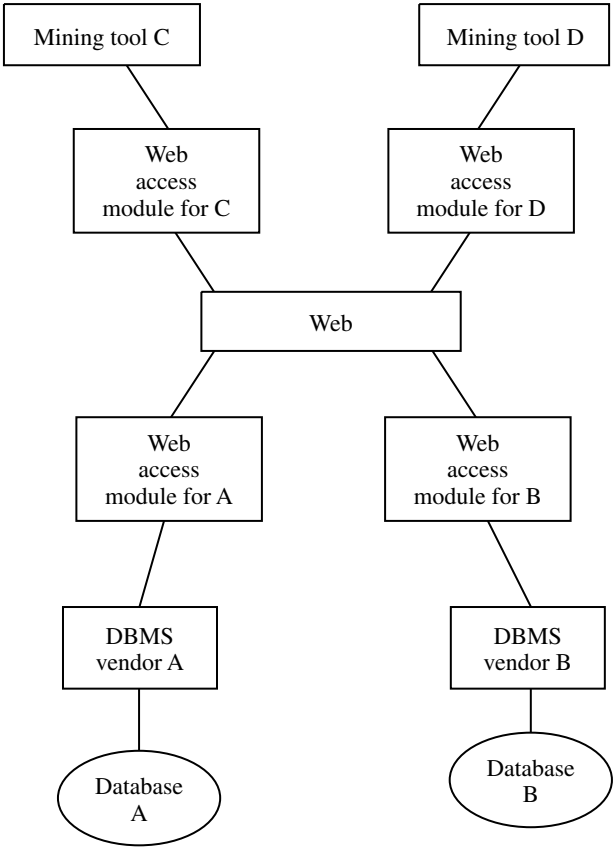


**Exhibit 1. Mining Databases on the Web**

databases. Mining semistructured databases is discussed in the next section. In the case of relational, object, and object-relational databases, we may apply the data mining tools directly on the database or we may extract key information from the data and then mine the extracted information. This is not different from what we have mentioned in our previous books (for example, see [THUR98]). However, because Web data may be coming from numerous sources, it may be incomplete or inconsistent. Therefore, we will have to reason under incompleteness and inaccuracy. An example of mining object-relational databases is illustrated in Exhibit 3.

In Chapter 5, we also discussed database functions, including query processing and transaction management. Query processing includes special optimization techniques and languages for the Web. We need to include data mining constructs into the languages as well as data mining techniques into the query optimization algorithms. Data mining can contribute in two ways for transaction management on the Web: mining Web transac-

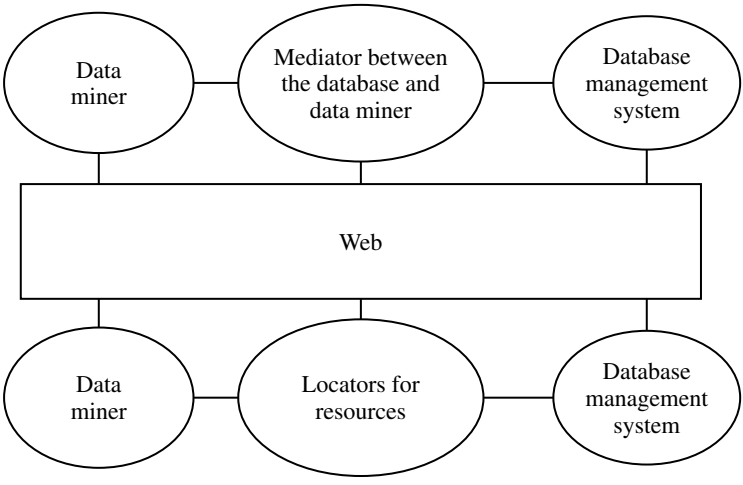


**Exhibit 2. Mediation and Location for Web Database Mining**

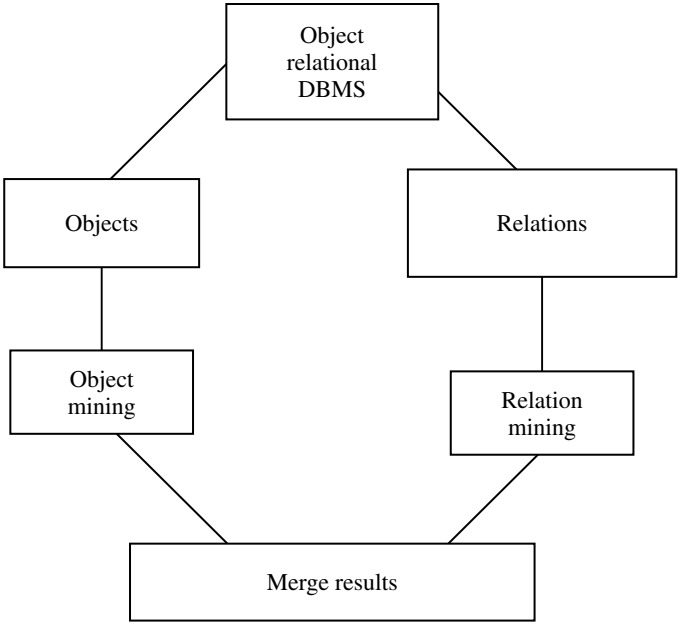


**Exhibit 3. Object-Relational Data Mining**

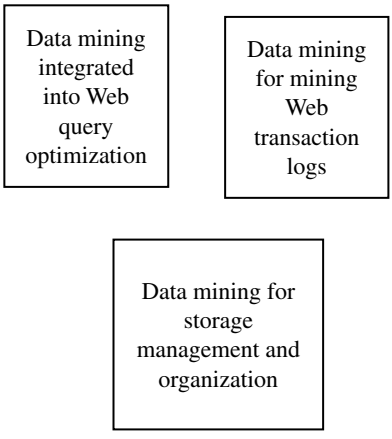


**Exhibit 4. Web Database Functions and Mining**

tion logs and mining when the transaction is being executed in real-time. There is little work on real-time data mining because data mining is usually carried out for analysis and it is difficult to build models in real-time. But we will see that for certain applications such as intrusion detection, we need to mine real-time databases. Building models in real-time will be a major challenge. Managing storage is also a function of Web database management. Here, data mining could help to determine the organization and structure of databases. We need more research in this area. Exhibit 4 illustrates some of the applications of data mining to Web database management functions.

### Data Sharing vs. Data Mining on the Web

As we have stressed, one of the challenges in data mining is to get the data ready for mining. This means that organizations have to be willing to share the data. In many cases, data is private and may be sensitive. Therefore, organizations and agencies may not be willing to share all of the data. As we have stated, with bad data one cannot have good data mining results even if using excellent data mining tools. So the question is, how can we share data so that we can mine? Essentially, we need to carry out federated data mining. This aspect will be discussed in Part III. We will see that federated data mining may have applications in counter-terrorism.

We discuss federated data mining in a later section. We discuss some preliminaries in this section. For example, in a federated environment, organizations form a federation with the objective to share data and still have autonomy. That is, we need federated data management practices for data sharing and mining. There are various ways to carry out federated data mining. In one approach, we can export certain data and schema to

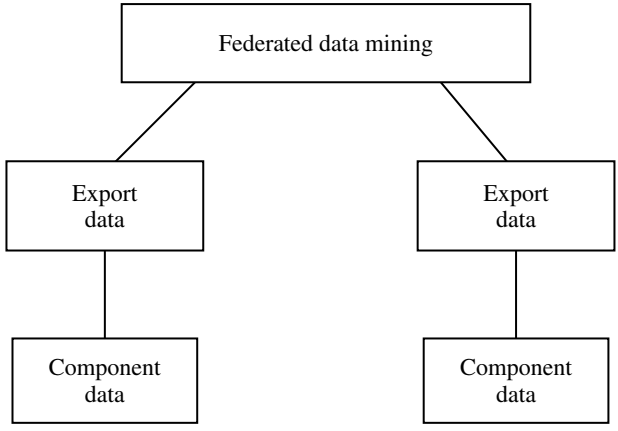


**Exhibit 5. Federated Data Mining: Approach I**

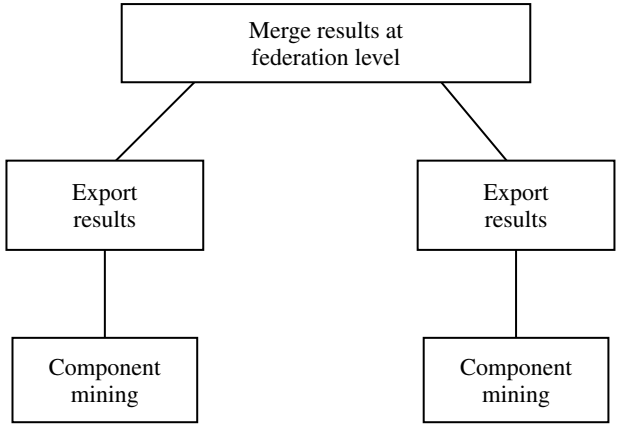


**Exhibit 6. Federated Data Mining: Approach II**

the federation and then carry out mining, as illustrated in Exhibit 5. In another approach, we carry out mining at the component level and then put the pieces together at the federation level. This latter approach is illustrated in Exhibit 6. We address distributed, heterogeneous, legacy, and federated database mining in a later section.

## MINING SEMISTRUCTURED DATABASES

Chapter 5 discussed semistructured databases. We elaborated on semistructured databases as well as XML databases in [THUR02]. Essentially, we use the terms *semistructured databases* and *XML databases* interchange-

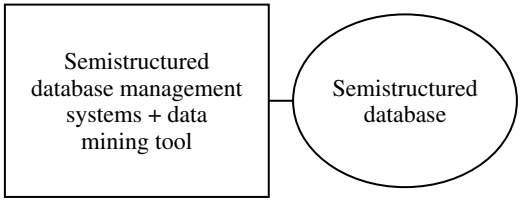


**Exhibit 7. Tight Integration between Data Miner and DBMS**

ably, although semistructured databases are much broader and include RDF document databases.

In our example of object-relational database mining, we gave some idea of how to mine objects and relations and merge the results. With semistructured database management, there are two approaches to managing the semistructured documents. One is to develop a database management system to manage the semistructured documents, known as the tight coupling approach. The other approach is to build an interface, for example, over relational databases to manage the semistructured documents, known as the loose coupling approach. We discussed these approaches in [THUR02].

In the case of data mining, there are various approaches. In one approach, we can extend the semistructured database management system with a data miner to mine the documents (see Exhibit 7), or we can build an interface to the semistructured database management system (see Exhibit 8). The former is the tight coupling between the data miner and the database management system (DBMS), and the latter is the loose coupling between the data miner and the DBMS. In the loose coupling approach to semistructured database management, we can mine the relations and the semistructured documents and then integrate the results (see Exhibit 9). Another approach is to extract structure from the semi-

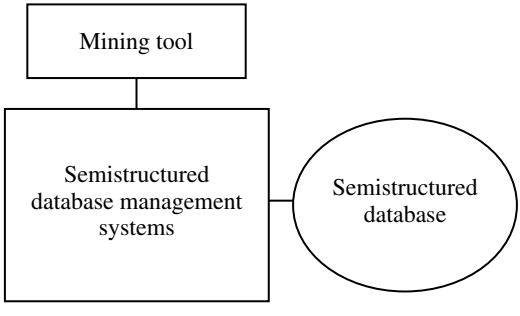


**Exhibit 8. Loose Integration between Data Miner and DBMS**

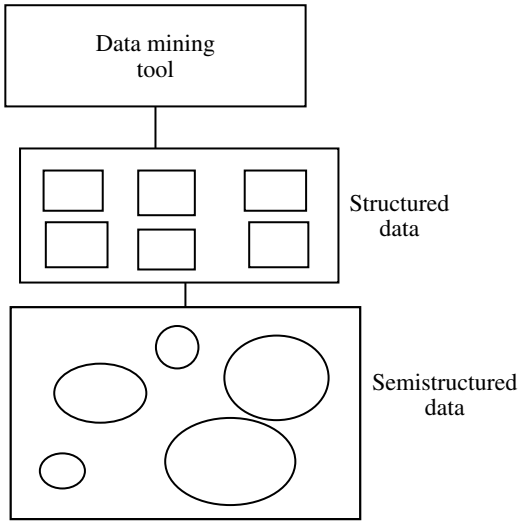


**Exhibit 9. Extract Structure and then Mine**

structured documents and then mine the structured documents (Exhibit 10). Note that there are various permutations and combinations of these approaches. For example, even in the loose coupling approach to semistructured database management, we can build a data miner as an interface to the database management system (Exhibit 11). The point we are making is that there are two ways to manage the semistructured databases: the loose coupling approach and the tight coupling approach. Also, in the case of data mining, we can have a tight coupling or a loose coupling with the data miner.

With respect to data warehousing, there are two aspects. One is that XML documents, semistructured databases, relational databases, and other data sources have to be integrated into a warehouse. Much of the work until now has been in integrating relational databases into a warehouse, also based on a relational model. When the databases are XML documents as well as semistructured databases, the question is, how do we integrate them into a warehouse? What is an appropriate model for a warehouse? Because there are now mappings between SQL and XML, can we still have an SQL-based model for the warehouse? The second aspect is representing the warehouse as a collection of XML documents. In this case, for example, we need mappings between the data sources based on relational and object models, to XML data models. We also need to develop techniques for accessing, querying, and indexing the warehouse. Both aspects of data warehousing are illustrated in Exhibits 12 and 13.

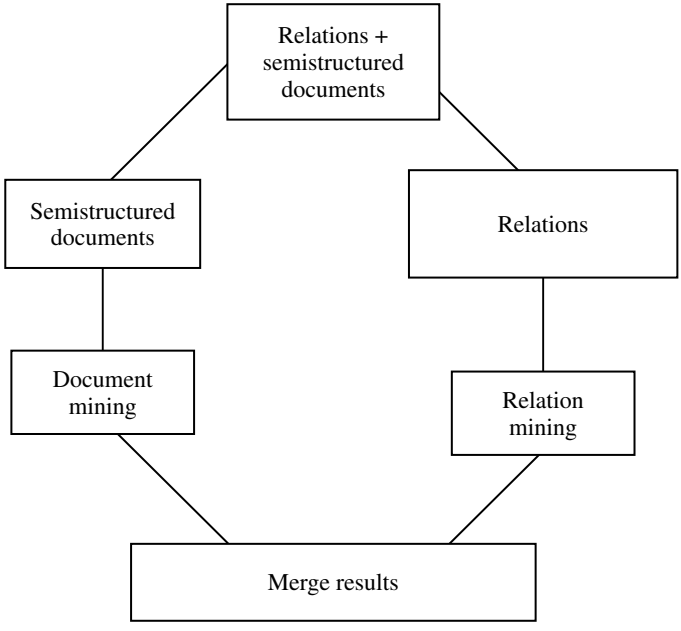


**Exhibit 10. Mining and then Merging**

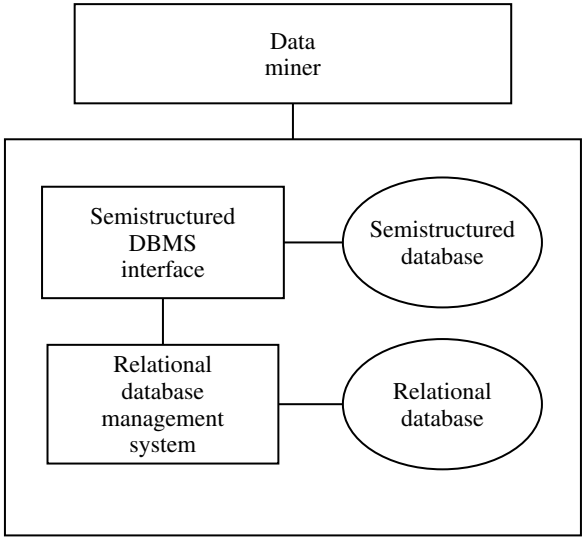


**Exhibit 11. Data Miner as an Interface to a Loose Coupling Semistructured DBMS**

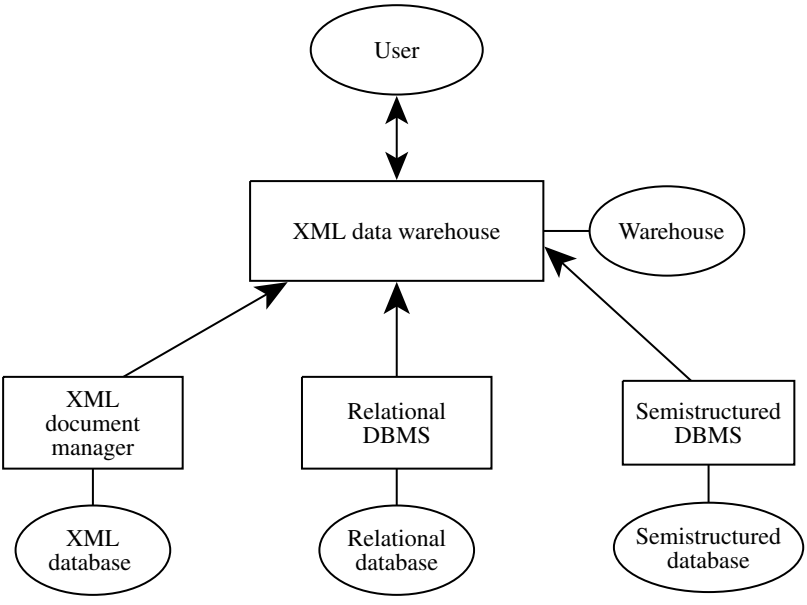


**Exhibit 12. XML Data Warehouse: Approach I**

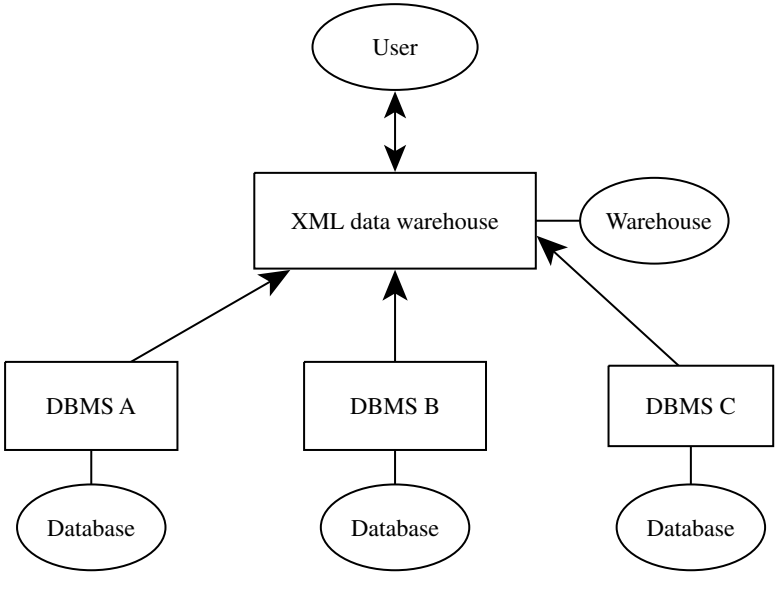


**Exhibit 13. XML Data Warehouse: Approach II**

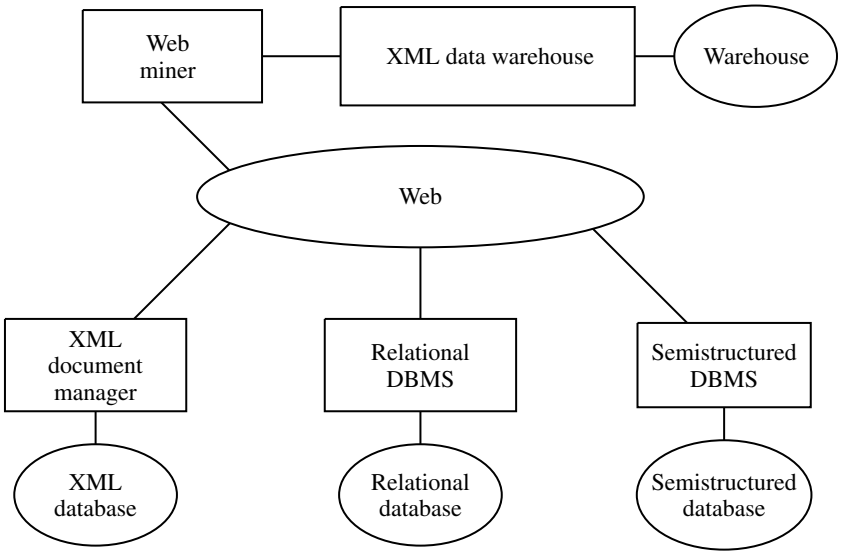


**Exhibit 14. Mining Web Warehouses**

Mining XML documents is receiving more attention recently. There are two aspects here. One is to mine the documents to extract useful information such as patterns and trends. For example, XML documents may be mined for business intelligence. One could also mine these documents for customer relationship management. The other aspect is to mine the links in an XML document and extract some information from these links. There is still much to be done here. Exhibit 14 illustrates mining the XML-based warehouses shown in Exhibits 12 and 13.

## METADATA AND WEB MINING

As discussed previously, metadata by itself is becoming a key technology for various tasks such as data management, data warehousing, Web searching, multimedia information processing, and data mining. Because metadata has been so closely aligned with databases in the past, we have included a discussion of the impact of metadata technology on Web data mining in this book.

Metadata plays an important role in data mining. It could guide the data mining process. That is, the data mining tool could consult the metadatabase and determine the types of queries to pose to the DBMS. Metadata may be updated during the mining process. For example, historical information as well as statistics may be collected during the mining process, and the metadata has to reflect the changes in the environment. The role of metadata in guiding the data mining process is illustrated in Exhibit 15.

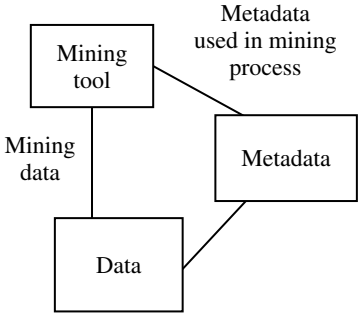


**Exhibit 15. Metadata Used in Data Mining**

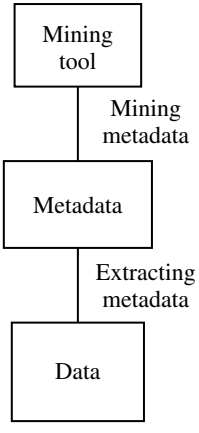


**Exhibit 16. Metadata Mining**

Extracting metadata from the data and then mining the metadata is illustrated in Exhibit 16.

There has been much discussion on the role of metadata for data mining [META96]. There are many challenges here. For example, when is it better to mine the metadata? What are the techniques for metadata mining? How does one structure the metadata to facilitate data mining? Researchers are working on addressing these questions.

Closely associated with the metadata notion is that of a repository. A repository is a database that stores possibly all the metadata, the mappings between various data sources when integrating heterogeneous data sources, the information needed to handle semantic heterogeneity such as "ship X and submarine Y are the same entity," the enforced policies and procedures, as well as information on data quality. So the data mining tool

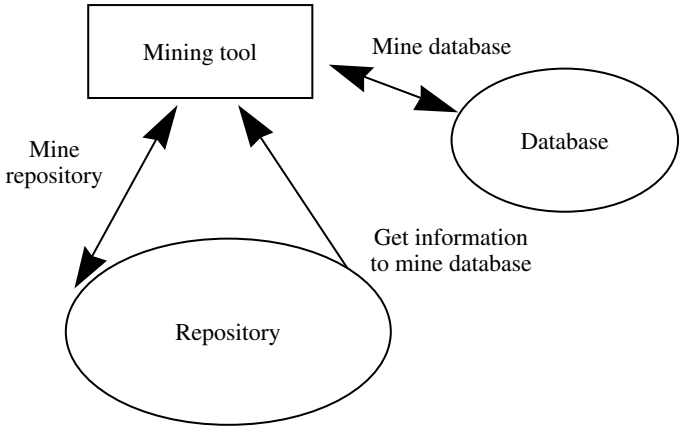


**Exhibit 17. Repository and Mining**

may consult the repository to carry out the mining. On the other hand, the repository itself may be mined. Both scenarios are illustrated in Exhibit 17.

Metadata plays an important role in various types of mining. For example, in the case of mining multimedia data metadata may be extracted from the multimedia databases and then used to mine the data. For example, as illustrated in Exhibit 18, the metadata may help in extracting the key entities from the text. These entities may be mined using commercial data mining tools. Note that in the case of textual data, metadata may include information such as the type of document, the number of paragraphs, and other

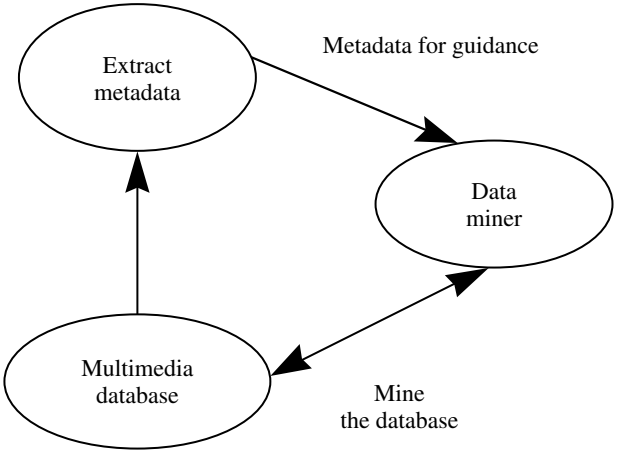


**Exhibit 18. Metadata for Multimedia Mining**

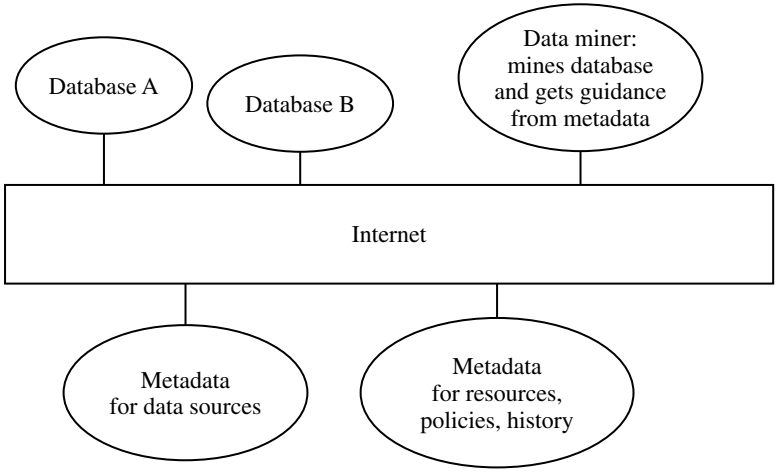


**Exhibit 19. Metadata for Web Mining**

information describing the document, but not the contents of the document itself.

Metadata is also critical in the case of Web mining, which is the main focus of this book. Because there is so much information and data on the Web, mining this data directly could become quite challenging. Therefore, we may need to extract metadata from the data, and then either mine this metadata or use this metadata to guide in the mining process. This is illustrated in Exhibit 19. Note that languages such as XML, which we will briefly discuss in the next section, will play a role in describing metadata for Web documents.

In Part III, we will address privacy issues for data mining. Policies and procedures will be a key issue for determining the extent to which we want to protect the privacy of individuals. These policies and procedures can be regarded as part of the metadata. Therefore, such metadata will have to guide the process of data mining so that privacy issues are not compromised through mining.

In almost every aspect of mining, metadata plays a crucial role. Even in the case of data warehousing, which we have regarded as a preliminary step to mining, it is important to collect metadata at various stages. For example, in the case of a data warehouse, data from multiple sources has to be integrated. Metadata will guide the transformation process from layer to layer in building the warehouse (see the discussion in [THUR97]). Metadata will also help in administering the data warehouse. Also, metadata is used in extracting answers to the various queries posed.

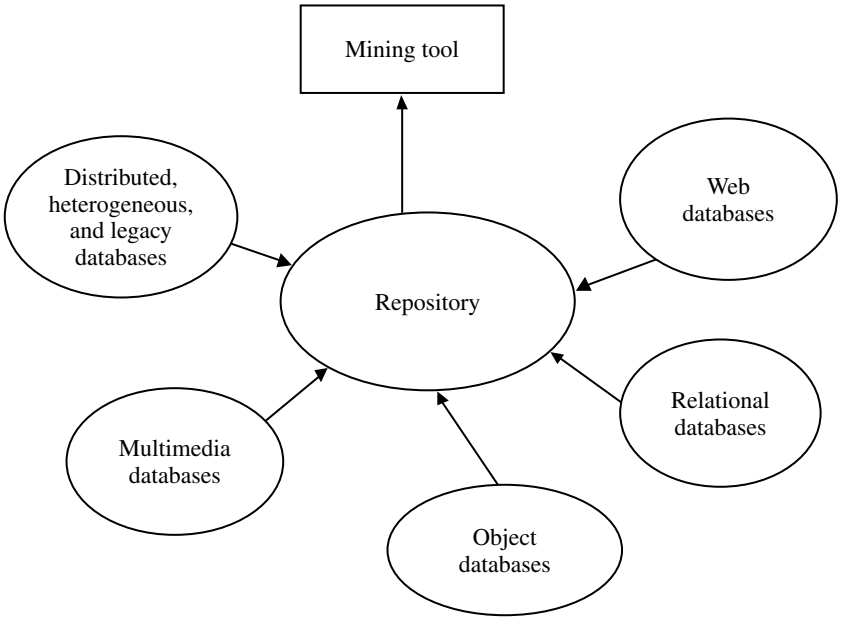


**Exhibit 20. Metadata as the Central Repository for Mining**

Because metadata is key to all kinds of databases including relational, object, multimedia, distributed, heterogeneous, legacy, and Web databases, one could envisage building a metadata repository that contains metadata from the different kinds of databases and then mining the metadata to extract patterns. This approach is illustrated in Exhibit 20 and could be an alternative if the data in the databases is difficult to mine directly.

## MINING DISTRIBUTED, HETEROGENEOUS, LEGACY, AND FEDERATED DATABASES ON THE WEB

In [THUR97], we placed much emphasis on heterogeneous database integration and interoperability. Many applications require the integration of multiple data sources and databases. These data sources may need to be mined to uncover patterns. Furthermore, interesting patterns may be found across the multiple databases. Mining heterogeneous and distributed data sources is a subject that has received little attention.

In the case of distributed databases, one approach is to have the data mining tool as part of the distributed processor where each distributed processor (DP) has a mining component also, as illustrated in Exhibit 21. This way, each data mining component could mine the data in the local database and the DP could combine all the results. This will be quite chal-

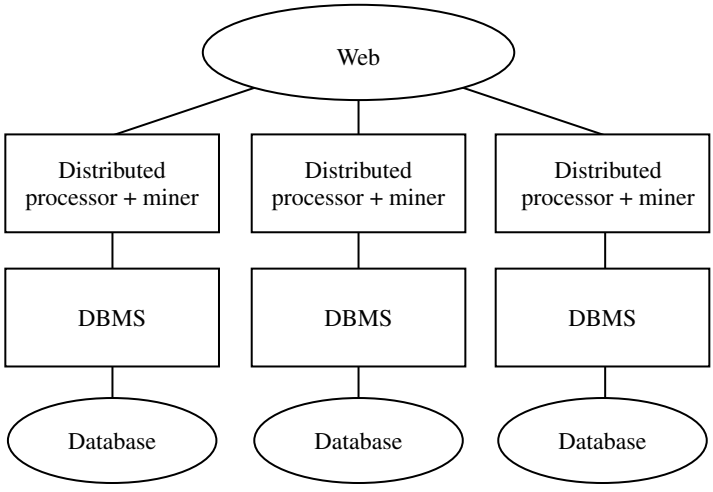


**Exhibit 21. Distributed Processing and Mining**

lenging, as the relationships between the various fragments of the relations or objects have to be maintained in order to mine effectively. Also, the data mining tool could be embedded into the query optimizer of the DQP (distributed query processor). Essentially, with this approach the DP has one additional module, a distributed data miner (DDM), as shown in Exhibit 22.

We illustrate distributed data mining with an example shown in Exhibit 23. Each DDM mines data from a specific database. These databases contain information on projects, employees, and travel. The DDMs can mine and get the following information: John and James travel together to London on project XXX at least 10 times a year. Mary joins them at least four times a year.

An alternative approach is to implement the data mining tool on top of the distributed system. As far as the mining tool is concerned, the data-

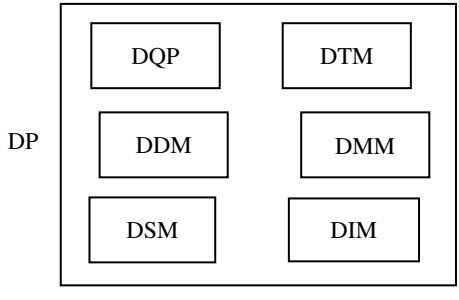


**Exhibit 22. Modules of DP for Data Mining**

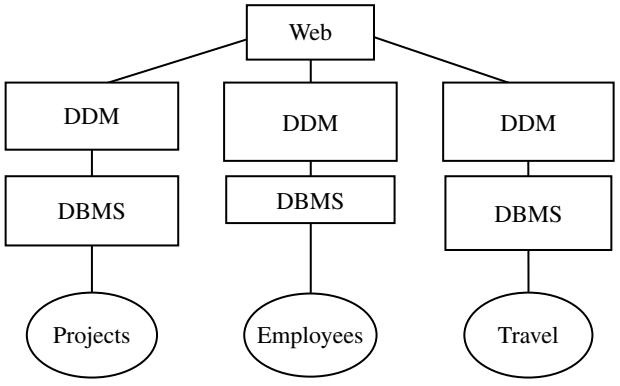


**Exhibit 23. Example Distributed Data Mining**

base is one monolithic entity. The data in this database has to be mined and useful patterns have to be extracted as illustrated in Exhibit 24.

In the case of heterogeneous data sources, we can either integrate the data and then apply data mining tools as shown in Exhibit 25, or apply data mining tools to the various data sources and then integrate the results, as shown in Exhibit 26. Note that if we integrate the databases first, then integration methods for interoperating heterogeneous databases are different from those for providing an integrated view in a distributed database. Some of these issues are discussed in [THUR97]. Furthermore, for each data mining query, one may need first to send that same query to the various data sources, get the results, and integrate the results, as shown in

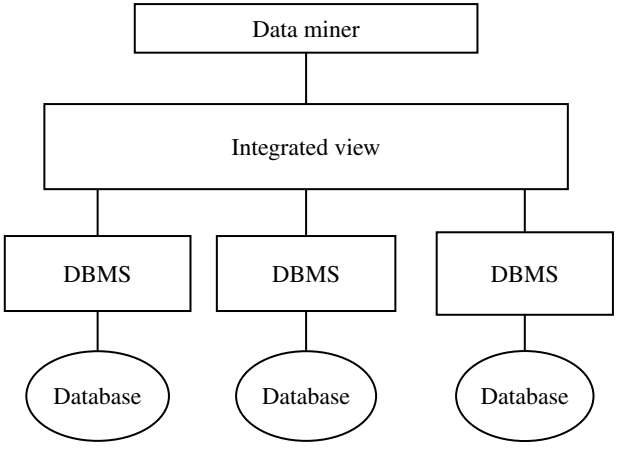


**Exhibit 24. Data Mining Hosted on a Distributed Database**

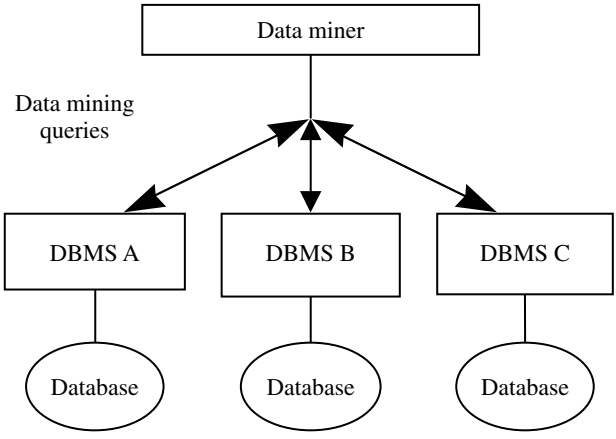


**Exhibit 25. Data Mining on Heterogeneous Data Sources**

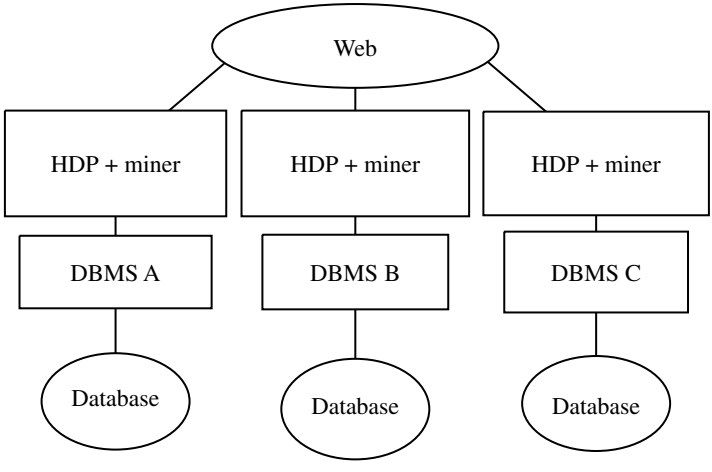


**Exhibit 26. Mining and then Integration**

Exhibit 25. If the data is not integrated, then a data miner may need to be integrated with the heterogeneous distributed processor (HDP), as illustrated in Exhibit 26. If each data source is to have its own data miner, then each data miner is acting independently. We are not sending the same query to the different data sources as each data miner will determine how to operate on its data. The challenge here is to integrate the results of the various mining tools applied to the individual data sources so that patterns may be found across data sources.

If we integrate the data sources and then apply the data mining tools, the question is, do we develop a data warehouse and mine the warehouse, or

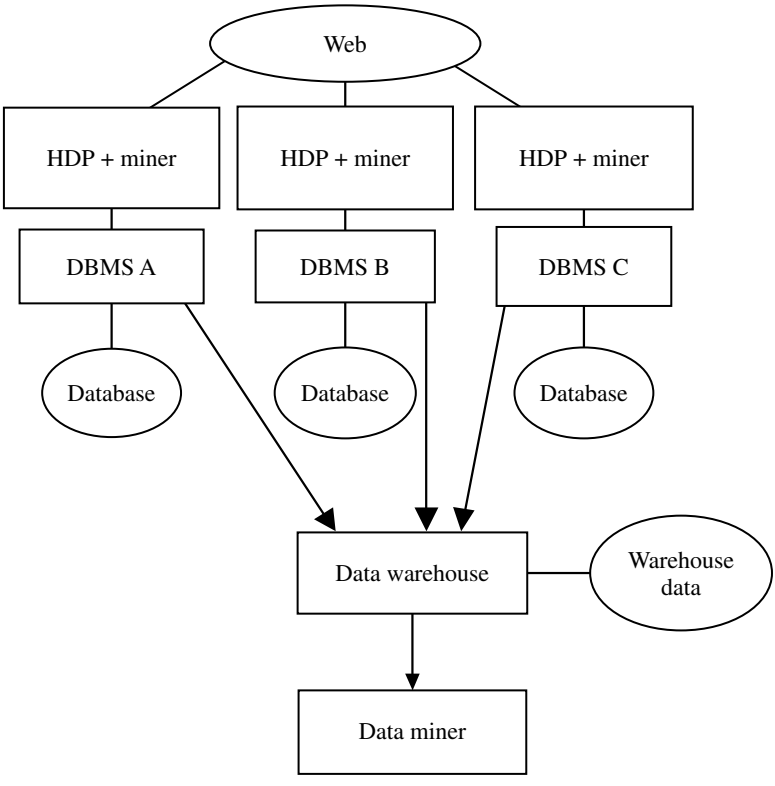


**Exhibit 27. Mining, Interoperability, and Warehousing**

do we mine with interoperating database systems? Note that in the case of a warehouse approach, not all of the data in the heterogeneous data sources is brought into the warehouse. Only decision support data is brought into the warehouse. If interoperability is used together with warehousing, then the data miner could augment both the HDP and the warehouse, as illustrated in Exhibit 27.

One could also use more sophisticated tools such as agents to mine heterogeneous data sources, as illustrated in Exhibit 28 where an integration agent integrates the results of all the mining agents. The integration agent may give feedback to the mining agents so that the mining agents may pose further queries to the data sources and obtain interesting information. There is two-way communication between the integration agent and the mining agents. Another alternative is to have no integration agent, but have instead the various mining agents collaborate with each other and discover interesting patterns across the various data sources. This is illustrated in Exhibit 29. This latter approach is also called collaborative data

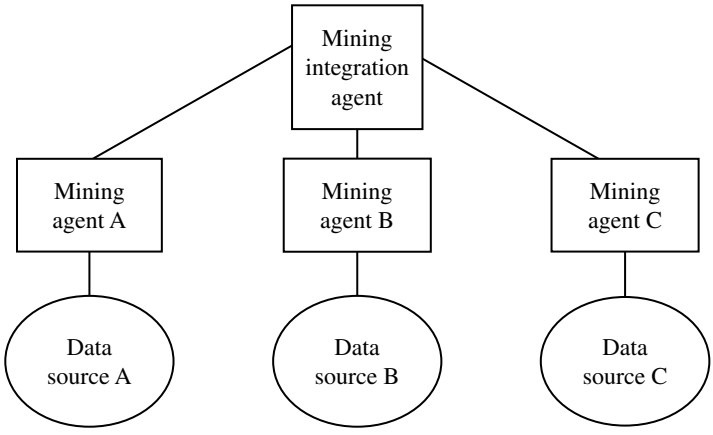


**Exhibit 28. Integrating Data Mining Agents**

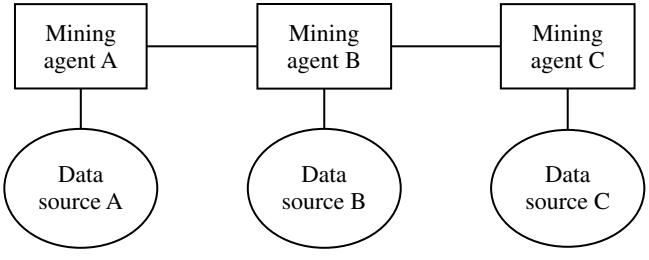


**Exhibit 29. Collaboration among Mining Agents**

mining. In this approach, collaborative computing, data mining, and heterogeneous database integration technologies have to work together.

The specific approach to mining heterogeneous data sources, whether to use an integration agent or have the mining agents collaborate, is yet to be determined. One may need both approaches or there may be yet another approach. Note also that heterogeneity may exist with respect to data models, data types, and languages. This could pose additional challenges to the data mining process. There is much research to be done in this area.

Another scenario for collaborative data mining is illustrated in Exhibit 30. Here, two teams at different sites use collaboration and mining tools to mine the shared database. One could also use mediators to mine heterogeneous data sources. Exhibit 31 illustrates an example where we assume that general purpose data miners and mediators are placed between the data miners and the data sources. We also use a mediator to integrate the results from the different data miners.

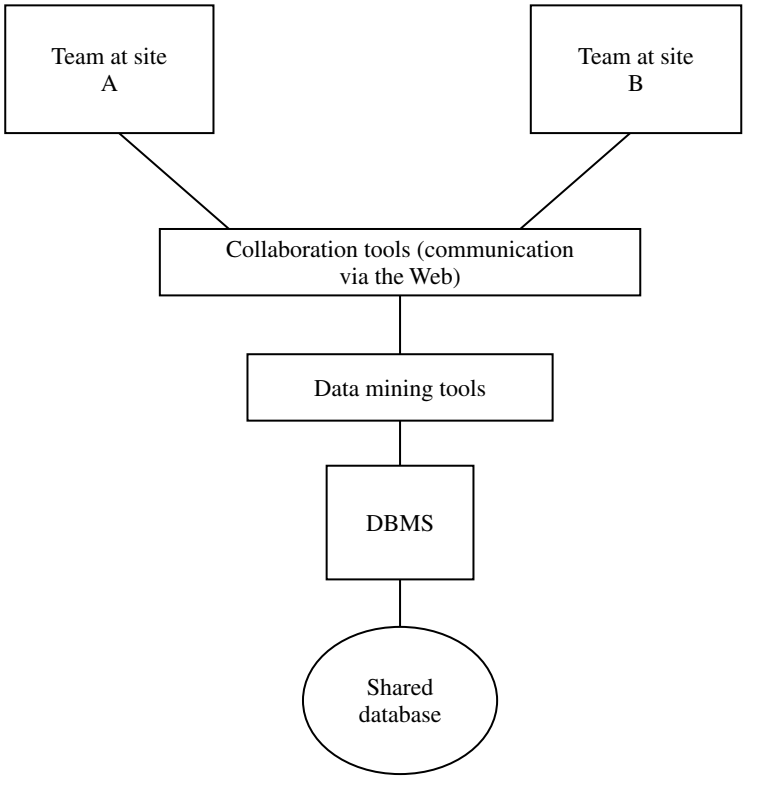


**Exhibit 30. Teams Conducting Mining on Shared Database**

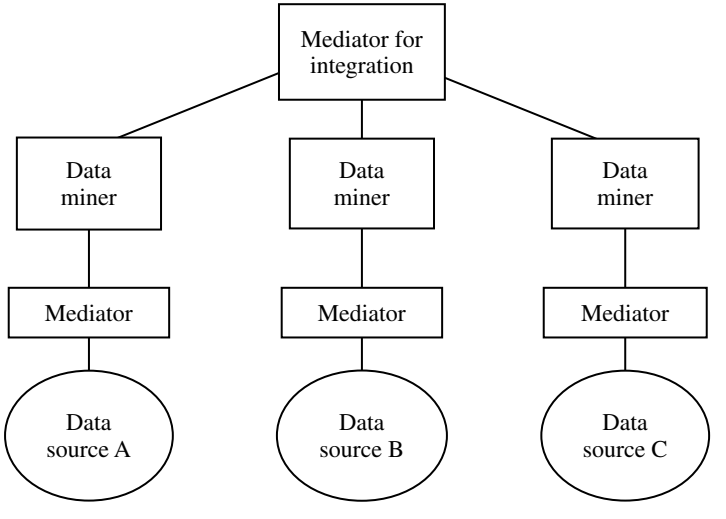


**Exhibit 31. Mediator for Integration**

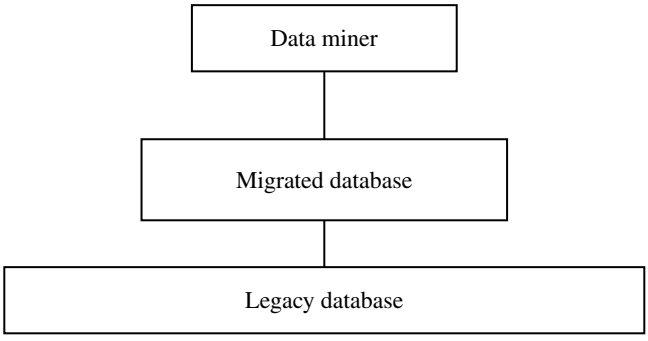


**Exhibit 32. Migration and then Mining**

Next, let us focus on legacy databases. One of the challenges here is how the legacy databases are to be mined. Can we rely on the data in these databases? Is it worth organizing and formatting this data, especially if it has to be migrated to newer systems? Is it worth developing tools to mine the legacy databases? How easy is it to integrate the legacy databases to form a data warehouse? There are some options. One is to migrate the legacy databases to new systems and mine the data in the new systems (see Exhibit 32). Another approach is to integrate legacy databases and form a data warehouse based on new architectures and technologies, and then mine the data in the warehouse (see Exhibit 33). In general, it is not a good idea to directly mine legacy data, as this data could soon be migrated, or it could be incomplete, uncertain, and therefore expensive to mine. Note that

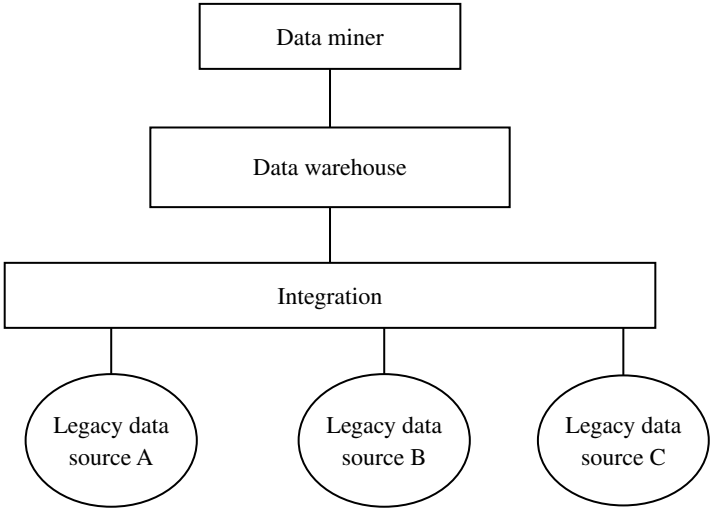


**Exhibit 33. Mining Legacy Databases**

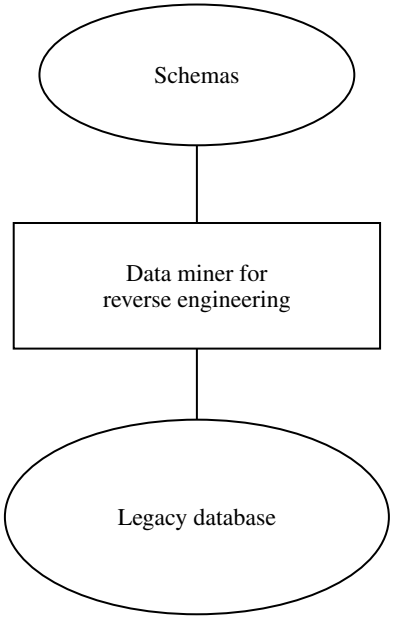


**Exhibit 34. Extract Schemas from Legacy Databases**

mining could also be used to reverse engineer and extract schemas from the legacy databases (see Exhibit 34 and the discussions in [THUR98]).

Finally, we will examine federated architectures and data mining. We call this federated data mining. Here, the data miners at the local sites need some autonomy and also need to share information with the foreign sites. As we have stated in our previous book [THUR97], Sheth and Larson [SHET90] came up with this very interesting schema architecture for federated databases. We adapted this architecture for security policies. Now we need to adapt it for data mining. Exhibit 35 illustrates our preliminary ideas on federated data mining. Note that we also discussed some approaches when we discussed data sharing and data mining.

## ARCHITECTURES AND WEB DATA MINING

There are several dimensions to Web data management architectures. We discussed many of them in Chapter 5. In this section, we focus on some of the data mining aspects. First of all, the three-tier architecture for Web data mining that we have discussed in our books (see [THUR00] and [THUR01]) is illustrated in Exhibit 36. In this architecture, we have the Web server which includes the data miner in the middle tier. Note that parts of the data miner could also reside in the client and the server DBMS. That is, there are various combinations for this architecture. Exhibit 37 illustrates

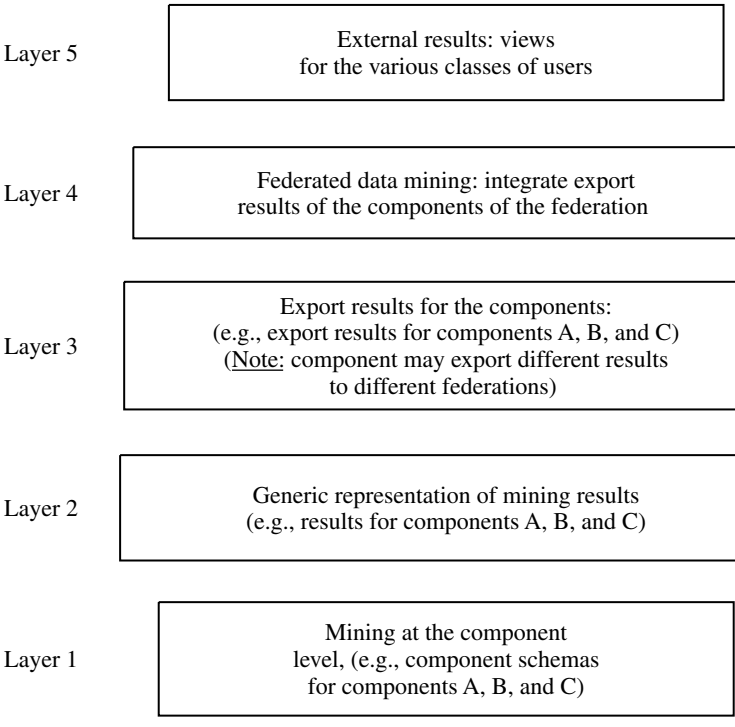


**Exhibit 35. Five-Level Architecture for Federated Data Mining**

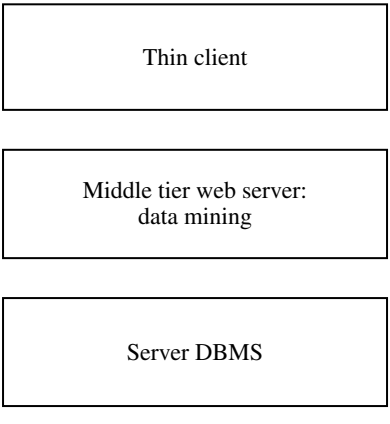


**Exhibit 36. Three-Tier Architecture for Data Miner**

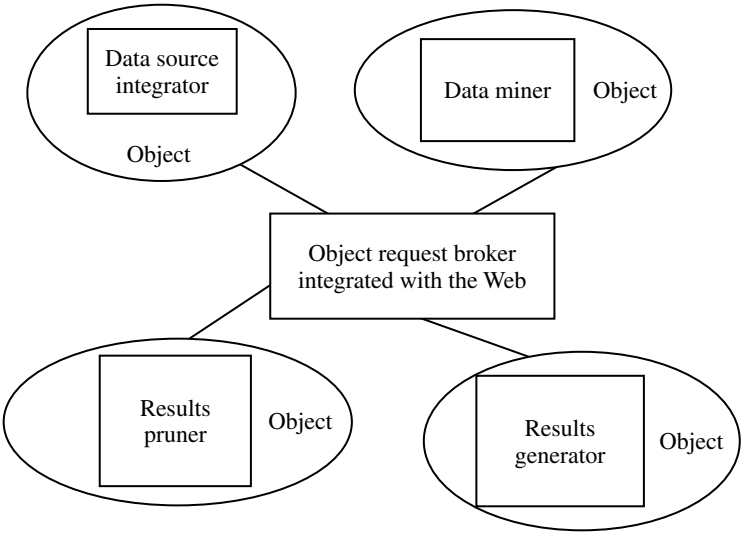


**Exhibit 37. Encapsulating Data Mining Modules as Objects**

the use of objects for data mining. Here, various data mining components are encapsulated as objects and an object request broker (ORB) is used to integrate the objects via the Web.

In Chapter 4, we also discussed various push/pull architectures. The question is, where does the data miner reside? In the case of pulling the data and pushing it to the consumer, we could have a data miner to carry out selective pulling as well as selective pushing of the data, as illustrated in Exhibit 38. Essentially, data mining is used as a Web service. We will discuss this further in Chapter 13. Another example of data mining from an architectural perspective is the application server/data server architecture discussed in Chapter 4. We can place a data miner to mine data at the server level as well as the application level, as illustrated in Exhibit 39. Note that there are several dimensions to architectures, and we have discussed just a few in this section.

## SUMMARY

This chapter has discussed various aspects of mining Web databases. Essentially, we examined the concepts in Chapter 5 and discussed the impact of data mining. First we discussed issues on mining databases such as integrating data mining into Web query optimization. Then we addressed mining semistructured databases. Metadata mining was discussed next. This was followed by a broad overview of mining distributed and heterogeneous databases. We also discussed approaches to

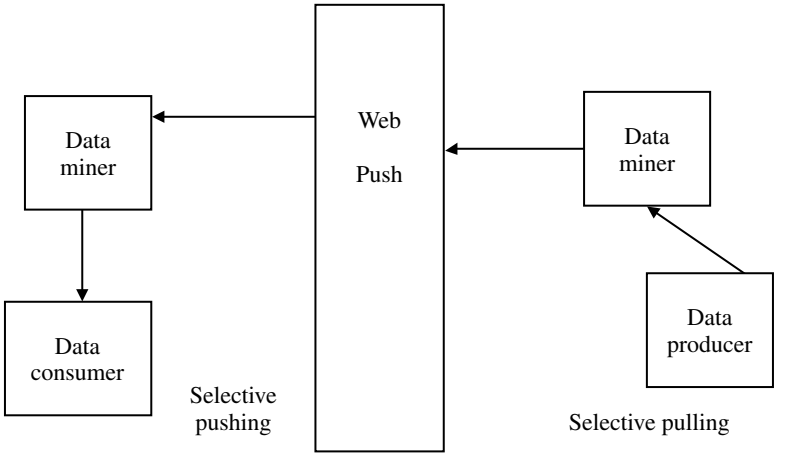


**Exhibit 38. Push/Pull and Data Mining: An Example**

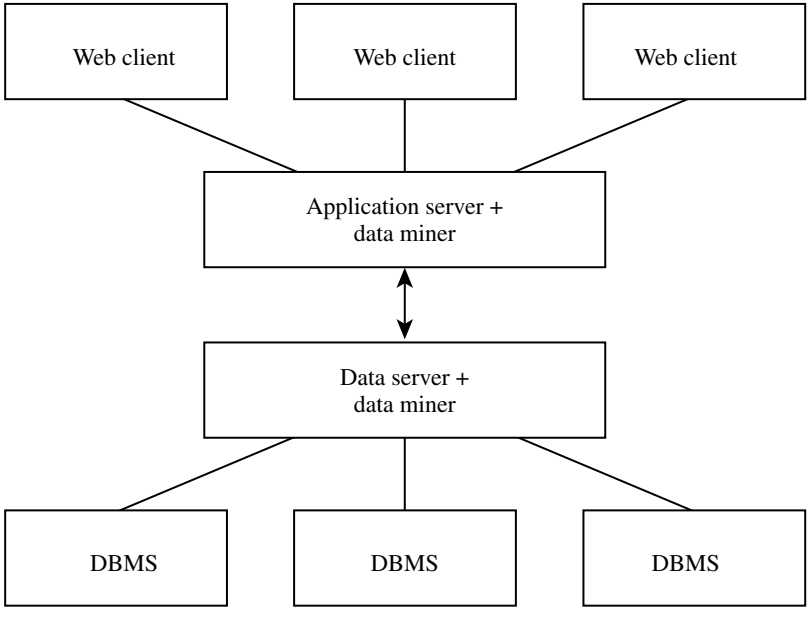


**Exhibit 39. Data Mining, Data Servers, and Application Servers**

federated data mining. Finally, we discussed architectural aspects of Web data mining.

As we have stressed in this book, our goal is to give the essential information at a high level. We have not given the details of algorithms for query optimization and metadata mining. For details, we refer the reader to the

various articles in journals and conference proceedings (see Appendix D). Our goal is to provide enough detail so that the reader can start thinking about applying the technologies to critical applications such as counter-terrorism.