Principles of Dataspaces

Seminar From Databases to Dataspaces
Summer Term 2007

Monika Podolecheva

University of Konstanz

Department of Computer and Information Science
Tutor: Prof. M. Scholl, Alexander Holupirek

Abstract. This seminar paper introduces a concept in the area of data management called dataspaces. The goal of this concept is to offer a way to handle multiple data sources with different models as answer to the rapidly increasing demand of working with such a data. Management challenges like providing search/query capability, integrity constraints, naming conventions, recovery, and access control arise permanently, in organizations, government agencies, libraries, on ones PC or other personal devices. Often there is a problem with locating the data and discovering the relationships between them. Managing data in a principal way will offer significant benefits to the organizations. The paper presents an abstraction for a dataspace management system as supposed by [1] and [2]. Specific technical challenges in realizing it are outlined and some applications are described.

1 Introduction and Motivation

In many real life situations the users are often faced with multiple data sources and the arising management challenges across these heterogeneous collections. General problem is how to locate all the relevant data and how to discover relationships between them. Having the data, they cannot be fitted into a conventional relational DBMS and uniformly managed. Instead, we have different data models, providing different search and query capability. The idea of dataspace is to provide base functionality over all data sources. On the administration side, these challenges include integrity constraints and naming conventions across a collection, providing availability, recovery and access control. If advanced functionalities such as data mining or certain guaranties in term of consistency, durability of updates etc are required, the data from the different sources must be integrated. Despite the difficulties, the benefits of managing data in a principal way are significant for organizations - hence the motivation for the development of Dataspace Management System (DSMS) - a system for managing heterogeneous data sources.
A dataspace should contain all information relevant to a particular organization regardless of its format and location, and model a rich collection of relationships

between them. The elements of a dataspace are a set of participants and a set of relationships. The participants are individual data sources (RDB, XML, web services, text). They contain description about the kind of data included, the format used for its storage and the querying mechanisms allowed. Further, they involve information about the data location and the relations to other participants as well as if the data is replicated and if the participant is added by a user or automatically generated by the system. The relationships denote the relation in which the participants are: for example, if the participant A is a view or a replica of another participant B, if A and B are created independently, but reflect the same physical system, or less specific, if two databases came from the same source at the same time. How this data is controlled is briefly discussed in the next sections, starting with all the requirements posted to the DSMS, its components that have to face these requirements, their architecture and the challenges arising during its realization.

2 Requirements on the Dataspace

Catalog A dataspace must hold information about its elements. The component Catalog is responsible for storing a detailed description of all participants included into the dataspace. It must contain basic information, such as owner, creation date, type of the participant and semantic information about the data in the participant. The Catalog must also incorporate the schema of the source, the query answering capabilities, accuracy, access and privacy information. The user should be able to browse the catalog to get more information about specific data sources.

Information Retrieval Main tasks performed by the users are querying and searching. These are handled by the Information retrieval component. Key challenge here is the requirement to support querying and searching for all participants regardless to their data model. This implies intelligent methods for interpreting and translating queries into various languages. Further requirements concerning the system are formulated as follows:

- The user should be able to refine the queries
- Support of transition between keyword querying, browsing and structured querying
- Ranking of the query answers
- Support of a wide spectrum of meta-data queries, such as: including the source of an answer or how it was delivered or computed and querying the source and degree of uncertainty in the answers
- Support of filtering and aggregation, also called monitoring of the dataspace

Information Extraction Information has to be extracted from different data collections, such as RDB, XML databases as well as from the World Wide Web, which has become a huge data container and thus storage of knowledge. Post-processing of data obtained from the web is a challenging issue and requires information extracting techniques from unstructured, semi-structured and structured web documents. Structured documents allow easier access and integration due to the rich semantically information included in the data representation. Dealing with semi-structured web pages, the extracted relevant information has to be transformed into structural information and saved local for further processing. Most problematic are the non-structured web documents. Text mining techniques can be used to map the parsed documents into groups organized with the help of ontologies, which allow a keyword search.

Data Management Extension Some of the data underlying models (e. g. simple file server, set of web documents) do not provide or provide only limited management features such as backup, recovery, and replication. Task of a DSMS should be to offer features for enhancing the low level participants. Responsible for this task is the Data Management Extension component. Additionally, this component supports ”value-added” information that is held by the DSMS but not initially present in the participants, such as translation tables for coded values, classification and rating of documents etc. This information must be able to span the participants. An example is storing connections between presentations, papers and programs, all related to one and the same project.

Manager The system should provide an interface between the user and the dataspace participants. The manager is the central component that is responsible for the interaction with the user. It manages the user authentication and right assignment.

Discovery Component As mentioned above, a main step in organizing a dataspace is to locate the participants and to discover relationships between them. The task of the Discovery Component covers these requirements. Initially, it locates participants in a dataspace, usually performing a classification with regard to participant type and content. Then it is responsible for the semi-automatical creation of relationships, for the improvement and maintenance of already existing relationships. This involves finding out which pairs of participants are likely to be related to each other, and then proposing relationships (schema mappings, replicas, etc.) that are afterwards manually verified and refined. Finally, it is important that the discovery component monitors the content of the dataspace to propose additional relationships over time.

Local Store and Index This component manages the caching of search and query results so that certain queries can be answered without accessing the actual data. It creates efficiently queryable associations between data objects in different participants and improves the access to data sources with limited access patterns. The Index has to be highly adaptive to heterogeneous environments.

It takes as input any token appearing in the dataspace and returns the locations at which the token appears and the role of each occurrence (string in a text file, element in file path, or tag in XML file). Thus, it spreads information across all participants involved. Another important aspect is the robust handling of multiple references to real-world objects (e. g., different ways to refer to a company or person). Caching increases availability of data stored in participants that may not be reliable and reduces the query load on participants that cannot allow ad-hoc external queries.

Replication Storage This component allows the replication of the data in order to increase access performance and thus to ensure availability and recovery support.

Application Repository This component allows user to share data analysis tools, domain specific models, and evaluations.

3 Architecture of a Dataspace Management System

An overview of the architecture is given in Fig. 1. It contains three participants having different data models: a RDB, a XML, and a file repository. Initially each participant is registered in the Catalog and thus becomes available as data source. Responsible for the registration is the Manager as well as for generating relationships between them, for allowing the user to refine these relationships,

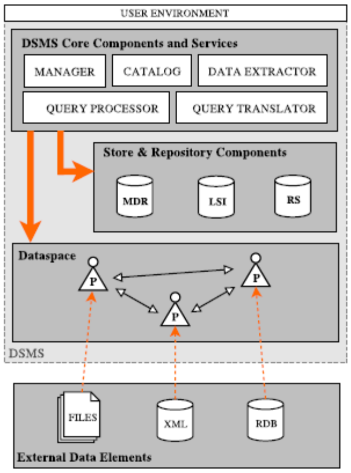

Fig. 1. DSMS Architecture

for user authentication and assignment of access rights. Here one can take the additional Discovery Component as part of the Manager. The DSMS does not assume complete control over the data. Instead, it allows the data to be managed by the participant systems, but provides a set of services to adjust probably missing management features. There are three internal data containers, namely the Metadata Repository (MDR), the Replication Storage (RS), and the Local Storage and Indexing (LSI) which allow users to transfer data between the components and local machines. The information extracted (by the Data Extractor) is locally stored and indexed. The Query Processor faces the problem of global query on multiple databases at multiple sites. The Query Interpreter translates the query into languages supported by the participants.

4 Application

Ecological data analysis The ecological research community is trying to integrate several ecological data collected and managed by 27 organizations spread across Europe, each of them with its own structure, semantic, and metadata description. The goal is to share data on biodiversity, ecology and socio-economy in order to extend their understanding and co-operation. The data sources include data from different interest domains which are represented as participants within the Eco-Dataspace. On demand, the participants could be dynamically integrated into different sub-Dataspaces performing different data exploration processes, e. g. geostatistics and building prediction models. Within the Application Repository the community will share specialized models and evaluations that are very expensive, difficult to develop and, as well as quite resource consuming and thus are distributed among multiple institutions and limited accessible. An integration of the results of these specialized applications into the Dataspace will make them ”globally” accessible and available for ecological data analysis.

Environmental observation and forecasting Consider a scientific research group monitoring a coastal ecosystem through weather stations, shore-and buoy-mounted sensors and remote imagery. Consider, they run atmospheric and fluid-dynamics models for simulating past, current and future conditions. For these simulations they often need to import data and model outputs from other groups, such as river flows and ocean circulation forecast. On the other side, the results of the simulations and the observation serve as input for programs generating a wide range of data products (such as comparison plots between observed and simulated data and images of surface-temperature distributions, etc.). The user is interested in basic file attributes such as time period covered, geographic region, physical variables (temperature, wind speed), kind of data product (graph, plot, animation), etc. Information about the version, simulation time step, used datasets etc. is also required for an appropriate analysis and comparison. Local copies or additional indices for fast search are neccesarry. Obtaining a huge amount of results, the overview of the entire set may get lost. This is a typicall example for the need of DSMS. Following dataspace requirements are illustrated here (and adequate for the former example): a dataspace-wide catalog, support for data lineage and creating collections and indexes over entities that span more than one participating source.

Personal Information Management (PIM) PIM should provide easy access and manipulation of all the personal information stored on a desktop, with possible extension to mobile devices, personal information on the Web, or even information accessed during a person's lifetime. The idea is to extend the desktop search (now limited to keyword queries) using the associations between the disparate items on the desktop, e. g. "Compute the aggregate balance of my bank accounts". Additional, we would like to query about sources, e. g. "Find all the experiments run by student X". Following principles of dataspaces come here in play: a PIM tool must enable the access to all data, to ensure integration of the multiple sources data and to return best-effort results. For realization of a PIM systems see iMeMex [8] and SEMEX [9].

5 Research Challenges

Challenges in terms of Data Models, Querying and Answers : Because of the multiple data models that should be supported by a DSMS, methods for interpreting queries in various languages must be found. Specialized search and query interfaces and an intuitive visualization of the results should be provided to the user, the difference between search and query should remain hidden for him. There are some specifics regarding the query answers that can be outlined in the following challenges:

- Development of intuitive semantics for answering a query that takes into consideration a sequence of earlier queries leading up to it (allows refinement of previous queries)
- Development of formal model of information gathering tasks that include a sequence of lower-level operations on a dataspace (allows handling or other actions like browsing, creating content etc)
- Development of algorithms that will rank the data sources according to how likely they are to contain the answer[1].

The next point is to shift the attention away from the limited semantic mappings developing techniques for obtaining the best-effort answers, such as: application of several approximate or uncertain mappings and comparison of the answers obtained by them; application of keyword search techniques for obtaining some data or some constants that can be used in instantiating mappings; examination of previous queries and answers obtained from data sources in the dataspace and ensuring mapping between the data sources; having queries that span multiple data sources, finding out how the sources are related (e.g. join attributes should

[1] For works related to key-word searching, browsing in RDB and ranking: see Hristidis et al. [4], Dbxplorer [5], DISCOVER [6], BANKS [7]

provide some hint of common domains). An intuitive challenge concerns the accuracy of the query answers, and extend of completeness and precision and latency required by the user.

- Development of formal model for approximate semantic mapping and for measuring the accuracy of answers obtained with them.
- Definition of metrics for comparing the quality of answers and answer sets over dataspaces, and efficient query processing techniques.

Challenges in terms of Uncertainty, Inconsistency and Lineage : The data in a dataspace will be uncertain and often inconsistent. The uncertainty will increase due to the best-effort query answering. Answers can be different, depending on the level of latency and completeness required. Hence, the authors see as a crucial point that a DSSP must be able to introspect on the answers presented to the users, and specify the assumptions and lineage underlying them. This section describes some challenges concerning dataspace introspection with respect to uncertainty, inconsistency and lineage and especially the way they affect each other.

- Development of formalisms that capture uncertainty about common forms of inconsistency in dataspaces (e.g. observe where the values come from and how they were created)
- Development of formalisms for representing and reasoning about external lineage again in term of the uncertainty.
- Development of a general technique to extend any uncertainty formalisms with lineage, and study the representational and computational advantages of doing so (combines the uncertainty and the lineage)

Dataspace discovery and reusing human attention : Large enterprises often do not know which data sources they have. That leads to the problem of locating the participants in the dataspace. For discovering relationships between them, clustering and other data mining algorithms must be applied. A good approach is to reuse human attention when possible for automatically creation of relationships. Consider the following cases: a user have done annotations, have temporary made collection of data for a particular task or written queries on the data. All this information can be exploited by machine learning techniques to gain knowledge about the relationships between the dataspaces. Methods for analyzing users' activities when interacting with a dataspace and creation of additional meaningful relationships in a dataspace or other enhancements to the dataspace are needed. The following challenges result from the nature of the human interaction:

- Development of techniques that examine collections of queries over data sources and their results to build new mappings between disparate data sources.
- Development of algorithms for grouping actions on a dataspace into tasks.

- Development of facilities for explicit enhancement of dataspace information that give high return on the investment of human attention.
- Development of a formal framework for learning from human attention in dataspaces

Challenges in terms of dataspace storage and indexing : Indexing is challenging due to the heterogeneity of the data. It should uniformly index all possible data regardless of their data model. Further, the index needs to consider multiple ways of referring to the same realworld object. A problem will be to keep the index up to date, especially for participants that do not have mechanisms to notify of updates.

6 Conclusions

The dataspaces were presented as a solution for the uniform management of disparate data sources. Requirements on the DSMS components were discussed to give a more concrete idea of the suggested system. Besides the benefits, there are many open points that have to be considered. Some of the challenges are already under research, others are still pure abstraction. Seamless querying of heterogeneous data is one of the main problems. Further, the data are not under full control of the system, which leads to a lack of limited ACID guarantees. Well functioning DSMS are needed and will be very valuable in many contexts, however, it will take a lot of time and research to overcome the challenges faced with heterogeneity.

References

1. Halevy, M. Franklin and D. Maier: Principles of Dataspace Systems, 2006 ACM1595933182/06/0006
2. Ibrahim Elsayed and Peter Brezany, A Min Tjoa: Towards Realization of Dataspaces Proceedings of the 17th International Conference on Database and Expert Systems Applications (DEXA'06) 0-7695-2641-1/06, 2006 IEEE
3. M. Franklin, A. Halevy and D. Maier: From Databases to Dataspaces: A New Abstraction for information Management, 2005 SIGMOD Record, Vol. 34, No. 4, Dec. 2005
4. V. Hristidis,L. Gravano,Y. Papakonstantinou: Efficient IR-Style Keyword Search over Relational Databases VLDB, 2003
5. S. Agrawal, S. Chaudhuri, and G. Das: Dbxplorer: A system for keyword-based search over relational databases, 2002 Proceedings of ICDE, 2002.
6. V. Hristidis, Y. Papakonstantinou: DISCOVER: Keyword search in relational databases, VLDB 2002
7. G. Bhalotia, C. Nakhey, A. Hulgeri, S. Chakrabarti, and S. Sudarshanz: Keyword Searching and Browsing in Databases using BANKS Proceedings of ICDE, 2002
8. Jens-Peter Dittrich: iMeMex: A Platform for Personal Dataspace Management 2nd Invitational Workshop for Personal Information Management, 2006
9. Yuhan Cai, Xin Luna Dong, Alon Halevy, Jing Michelle Liu, and Jayant Madhavan: Personal Information Management with Semex SIGMOD 2005 June