On Using Distributed Extended XQuery for Web Data Sources as Services*

Muhammad Intizar Ali[1], Reinhard Pichler[1], Hong-Linh Truong[2], and Schahram Dustdar[2]

[1] Database and Artificial Intelligence Group, Vienna University of Technology
{intizar,pichler}@dbai.tuwien.ac.at
[2] Distributed Systems Group, Vienna University of Technology
{truong,dustdar}@infosys.tuwien.ac.at

Abstract. DeXIN (**D**istributed **e**xtended **X**Query for data **IN**tegration) integrates multiple, heterogeneous, highly distributed and rapidly changing web data sources in different formats, e.g. XML, RDF and relational data. DeXIN is a RESTful data integration web service which integrates heterogeneous distributed data sources, including data services (DaaS – data as a service). At the heart of DeXIN is an XQuery extension that allows users/applications to execute a single query against distributed, heterogeneous web data sources or data services. In this system demo, we show how DeXIN can provide an optimized, distributed and parallel query processing and data integration at the same time.

1 Introduction

In recent years, there has been an enormous boost in Semantic Web technologies and Web services. Web applications thus have to deal with huge amounts of data which are normally scattered over various data sources using various languages. Hence, these applications are facing two major challenges, namely (i) how to integrate *heterogeneous* data and (ii) how to deal with *rapidly growing* and continuously changing *distributed data sources*.

The concept of providing data as a service (DaaS) [1] enables applications to expose data sources as Web services that can be consumed by Web clients within a corporate network and across the internet. In this paper, we demonstrate DeXIN, a RESTful Web Service for data integration of heterogeneous and distributed data sources. DeXIN receives a user query in extended XQuery syntax as presented in [2], this extension enables DeXIN to execute a single query in XQuery language, which can contain multiple sub-queries of SPARQL or SQL. Data sources supported by DeXIN can be Web services or databases which provide a query interface based on Web services and XML wrapping facility for results. Supporting Web services based on databases is important as current database management systems increasingly provide REST/SOAP APIs for querying hosted data. The integrated results of all the data sources are presented in XML. Currently available heterogeneous data integration approaches normally work

* This work was supported by the Vienna Science and Technology Fund (WWTF), project ICT08-032.

either (i) by transforming the data sources into common format [3, 4] or (ii) by query rewriting [5]. In contrast, DeXIN executes distributed parallel queries towards native data sources, without fetching all the data sources into one centralized place or transforming data sources. These features make DeXIN a powerful tool for data integration in a highly distributed, peer to peer, heterogeneous and rapidly changing Web environment, providing the user with a uniform access to this data.

2 Overview of DeXIN

An architectural overview of DeXIN is depicted in Figure 1. The main task of DeXIN is to provide an integrated access to different distributed, heterogeneous, autonomous data sources. DeXIN provides a single entry point to access different data sources by

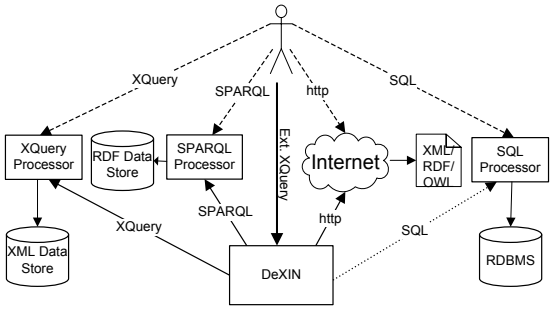

Fig. 1. Architectural overview of DeXIN framework

using our extension of XQuery [2]. The DeXIN service can be utilized by many web applications which require an integrated access to heterogeneous web data sources. Distributed, parallel query execution and avoiding data transformation make it a strong tool for data integration and optimized query execution in distributed and peer to peer networks. Normally, the user would have to query each of these data sources separately. With the support of DeXIN, he/she has a single entry point to access all these data sources. In total, the user thus issues *a single query* (in our extended XQuery language) and receives *a single result*. All the tedious work of decomposition, connection establishment, document retrieval, query execution, etc. is done behind the scene by DeXIN.

3 Distributed Extended Query for Data Integration

3.1 DeXIN : Data Integration Web Service

DeXIN is a RESTful data service which takes a single query in extended XQuery syntax as input, decomposes the query into sub-queries, executes each sub-query independently on its appropriate distributed data source at remote locations and outputs the integrated results from all data sources in XML format. Consider an example of a web

application which needs to provide the integrated access to the distributed and heterogeneous Web data sources dynamically. Typical data integration approaches e.g. warehousing, mediation or ontology based, do not provide the desired results because they require some prior knowledge about the data sources. DeXIN can ideally serve such applications because it provides integrated access to heterogeneous distributed web data sources dynamically. Figure 2 shows the user interface of the DeXIN service. The user can write a query in extended XQuery format and gets the accumulated results of all the data sources. Currently, DeXIN supports two types of sub-queries inside XQuery namely, (i) SPARQL for RDF, OWL and (ii) SQL for relational data.

3.2 Searching Available Data Services

Many service providers have started to expose their data as a service by implementing the Resource Oriented Architecture (ROA). Some Database Management Systems also provide access to their data using Query Language + REST/SOAP. DeXIN can communicate with Data Service directories to find out the appropriate data service. The user can initiate a keyword search for the required data service, and all the available data services are listed by DeXIN to help the user to select an appropriate data service.

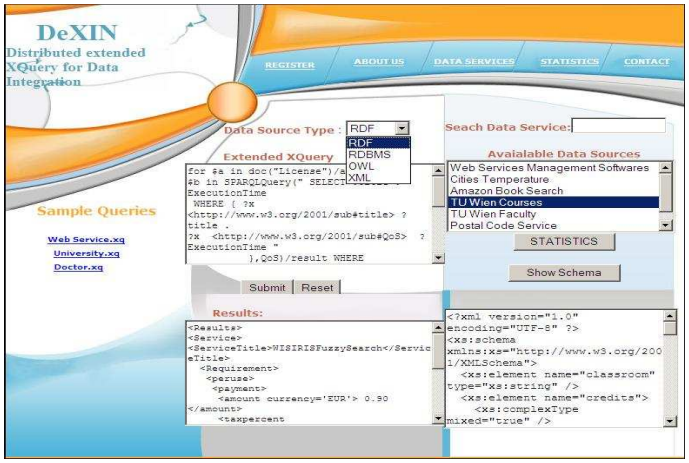

Fig. 2. User Interface of DeXIN

3.3 Registration of Data Sources

The registration of data sources at DeXIN is not mandatory because DeXIN can interact with any data service at runtime, but providing some metadata about data sources by following the registration procedure makes querying simpler from the user's perspective. Different Data Sources (e.g. RDF,XML,OWL or RDBMS) can register at DeXIN to benefit from the integration facilities provided by DeXIN. Each data source must provide a unique name, should have one of the DeXIN supported data types, querying interface with connectivity facility and XML converter for query results. Data providers can provide additional information about schema, user privileges, license and legal issues to facilitate the users to interact with their data sources effectively. It is worth

mentioning here that utilizing the concept of Data as a Service greatly eases the process of registration, because it uses the standard HTTP protocol to interact with the data sources and XML for data transfer.

3.4 Data Source Statistics and Schema Information

DeXIN stores some metadata and statistics about registered data sources, which are helpful for the selection of the best available service from the user's perspective. The user can select any data service from the list of available data sources shown by DeXIN (see top right of Figure2) and can see its statistics which are either stored in DeXIN or retrieved by DeXIN after communicating with the Service Management System.

If the data service provider provides some schema information about data sources, the user can click on "Show Schema", to see the schema information, which is helpful for designing queries for that particular data source.

3.5 Query Execution

Once the user submits a query to DeXIN in extended XQuery format, DeXIN (i) decomposes the query into multiple sub-queries for distributed, heterogeneous data sources (ii) connects with the data sources mentioned in the query (iii) dispatches queries to their particular data source at remote locations (iv) displays integrated results of all the sub-queries into XML format.

4 Conclusion

In this demo, we present DeXIN, a web based system to integrate data by executing distributed XQuery over Heterogeneous Data Sources. We demonstrate typical use cases of heterogeneous data integration which show that DeXIN is a simple but powerful tool to integrate rapidly changing heterogeneous data sources dynamically. DeXIN can be utilized by many applications where the data sources are unknown at design time, and it eases the integration process from the user's perspective by not requiring prior knowledge of data sources.

References

1. Fujun Zhu, Mark Turner, Ioannis A. Kotsiopoulos, Keith H. Bennett, Michelle Russell, David Budgen, Pearl Brereton, John Keane, Paul J. Layzell, Michael Rigby, and Jie Xu. Dynamic data integration using web services. In *ICWS*, pages 262–269, 2004.
2. Muhammad Intizar Ali, Reinhard Pichler, Hong-Linh Truong, and Schahram Dustdar. DeXIN: An extensible framework for distributed XQuery over heterogeneous data sources. In *Proc. ICEIS 2009*, pages 172–183, 2009.
3. Fabien Gandon. GRDDL Use Cases: Scenarios of extracting RDF data from XML documents, April 2007. W3C Proposed Recommendation.
4. Sven Groppe, Jinghua Groppe, Volker Linnemann, Dirk Kukulenz, Nils Hoeller, and Christoph Reinke. Embedding SPARQL into XQuery/XSLT. In *Proc. SAC 2008*, pages 2271–2278, 2008.
5. Waseem Akhtar, Jacek Kopecký, Thomas Krennwallner, and Axel Polleres. XSPARQL: Traveling between the XML and RDF Worlds - and Avoiding the XSLT Pilgrimage. In *Proc. ESWC 2008*, pages 432–447, 2008.