ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

Volume 3, Special Issue 3, March 2014

2014 International Conference on Innovations in Engineering and Technology (ICIET'14)

On 21st& 22nd March Organized by

K.L.N. College of Engineering, Madurai, Tamil Nadu, India

Improving the Demand Multidimensional Data Integration Including Formal Data Models and Re-Write Rules for Optimization Using Cloud Warehouse

Prof.S.Saravanan, Dr.V.Venkatachalam

Ph.D.Scholar, Dept of CSE, M.Kumarasamy College of Engineering, Karur, India.

Principal, The Kavery Engineering College, Mecheri, Salem, India.

Abstract : Cloud intelligence is a collection of technologies emerging from the migration of business intelligence and analytics technologies to a cloud computing environment combined with exploiting the massive range of new intelligence opportunities opened up by cloud computing. Cloud computing introduces several trends which require traditional business intelligence techniques to be re-thought, including agility, the ability to assemble resources, e.g., data sources, on-demand, and virtualization, e.g., that data are provided as a service over the web rather than stored in local databases.

This paper focuses on the combination of data source agility and data-as-a-service virtualization and its use for cloud intelligence. After presenting the novel vision of the Cloud Warehouse, the paper goes on to present a comprehensive semantic foundation for on-demand multidimensional data integration, including for- mal data models, a range of query operators and re-write rules for optimization. This semantic foundation provides a sound formal basis for on-demand multidimensional data integration, which is a cornerstone of cloud intelligence.

Keywords: Data integration · Multi-dimensional databases · OLAP · Federated databases · Cloud intelligence

I.INTRODUCTION

Cloud intelligence is a collection of technologies emerging from the migration of business intelligence and analytics technologies to a cloud computing environment combined with exploiting the massive range of new intelligence opportunities opened up by cloud computing. With cloud computing, we have to handle a paradigm shift in the computing infrastructure. Here, existing concepts such as storage, processes, computation, etc., are *virtualized* to all be parts of "the cloud." Thus, technologies have to become device and location independent, e.g., different parts of the data will all reside "in the cloud," but can be dispersed, and even migrating, between many physical locations and devices. We thus have to start thinking about data management as a service, which in turn entails that analytics and intelligence also become services. Data-as-a-service will often be delivered using existing web-based technologies such as XML-based data formats.

II.AGILITY

Cloud intelligence also entails *agility*, the ability to assemble the necessary resources on demand, not only in terms of computing power, but also in terms of data sources. Here, the situation will often be that a given organization has one or more existing core data sources, which must then be integrated on demand with additional data sources delivered through data services.

III.MEASURED VALUES

OLAP systems enable powerful analysis of large amounts of summary data commonly drawn from a number of different transactional databases. OLAP data are often organized in multidimensional *cubes* containing *measured values* that are characterized by a number of hierarchical *dimensions*. The multidimensional approach offers a number of advantages over traditional types of DBMS, including automatic aggregation, visual querying, and good query performance due to the use of pre-aggregation .

In a cloud intelligence setting, the we need to integrate new data sources *on demand* into the cloud warehouse. This means that *virtual*, rather than physical, integration of data is the way to go. This paper focuses on the combination of data source agility and data-as-a-service virtualization and how this can be used for cloud intelligence. The paper first presents the novel concept of a *cloud warehouse* that integrates the multitude of data sources found in the cloud to form a virtual warehouse "in the cloud."

The data model for the cloud warehouse is based on multidimensional cube concepts, since this has proven to be superior for data analysis tasks. The paper then goes on to present a comprehensive semantic foundation for on-demand multidimensional data integration, including a range of query operators and re-write rules for optimization. The integration paradigm is similar to the "pay-as-you-go" data integration principles known from data-spaces. Concretely, we assume that the external data-as-a-service are delivered in XML format, and that the existing core data sources take the form of multidimensional data cubes. However, the presented principles also generalize to a much wider range of data types, and this paper can thus be seen as a first step toward a comprehensive semantic foundation for cloud intelligence.

The presented integration foundation allows external XML data to be used as "virtual" dimensions, enabling three specific uses of XML data. First, OLAP query results may be "decorated" with XML data. Second, external XML data may be used for selection. Third, OLAP data may be grouped by external XML data when aggregation is performed. Special care is taken to ensure that the possibly irregular structure of the XML data does not cause problems w.r.t. correct aggregation of data. A flexible linking mechanism is devised to associate cube data with parts of XML documents.

IV.EXTENDED MULTIDIMENSIONAL

We make no assumptions about the existence of Document Type Definitions (DTDs) or XML Schemas . To demonstrate the capabilities of the approach, we present a data model and a multi-schema query language, *XML-Extended Multidmensional SQL* (SQL$_{XM}$), based on SQL and XPath. SQL and XPath are chosen for their simplicity, wide-spread use, and compact syntax. The foundation of the approach is a novel multidimensional algebra operator, the *decoration operator*, which allows external data to be integrated *on demand* into OLAP cubes as new dimensions, i.e., the cube is "decorated" with the new dimensions which can subsequently be used just as the regular dimensions. We formally specify the semantics of the decoration operator, giving three different possible interpretations if dimension hi-eparchies are *irregular*.

We also provide a comprehensive set of algebraic re-write rules. These specify how the decoration operator interacts with the other operators in the cube algebra, selection and generalized projection. It is shown how the rules can be used for effective processing of on-demand multidimensional data integration queries. We also present a number of additional effective optimization techniques for the approach, including so-called "in lining" and optimizations over data-as-a-service interfaces.

As almost all data sources can be efficiently wrapped in XML format , the approach also allows external data from most kinds of data source in the cloud, e.g., also relational, object-relational, and object databases to be integrated on demand.

V.THE CLOUD WAREHOUSE

In this section, Cloud-based, scalable approach to the integration of different types of data for analysis purposes, and then explain how on-demand multidimensional data integration fits into the picture.

Data warehouses (DWs) have become very successful in many enterprises, by al- lowing the storage and analysis of large amounts of structured business data. DWs are mostly based on a so-called "multidimensional" data model, where important business events, e.g., sales, are modeled as so-called facts, characterized by a number of hierarchical dimensions, e.g., time and products, with associated numerical measures, e.g., sales price.

The multidimensional model is unique in providing a framework that is both intuitive and efficient, allowing data to be viewed and analyzed at the desired level

of detail with excellent performance. Traditional data warehouses have worked very well for traditional, so-called structured data. However, the ongoing revolution of cloud computing means that traditional DWs are increasingly only solving a small part of the real integration and analysis needs of most enterprises.

There is a multitude of different types of data found "in the cloud," including structured, relational data, multidimensional data in DWs, text data in documents, emails, and web pages, and semi-structured/XML data such as electronic catalogs. With the ongoing developments within mobile, pervasive and ubiquitous computing, the cloud will also contain large quantities of geo-related data, as well as data from a large amount of sensors. Finally, many analytical models of data have been developed through data mining.

VI.CLOUD INTELLIGENCE

To achieve true *cloud intelligence*, all these types of data/models must be integrated and analyzed in a coherent fashion, which is not possible with current solutions. Instead, applications must develop ad hoc solutions for integration and analysis, typically for each pair of data types, e.g., relational and text. This obviously is both expensive and error-prone. Privacy protection is, although important, often ignored or given low priority, given the problems with doing the integration and analysis in the first place.

Our overall vision to overcome this is to develop a breakthrough set of technologies that enables a "cloud warehouse". The base for the cloud warehouse will a data model that uses concepts from multidimensional and semi-structured data models, with additional support for handling geo-related data (geo models, etc.), sensor data (high speed data streams, missing or incorrect values, etc.), semi-structured and un-structured data (enabling analysis across structured, semi-structured, and unstructured data), and imperfect (imprecise, uncertain, etc.) data. Support for privacy management will also be built into the framework.

This will enable the creation of a Cloud Warehouse (CW) that provides the same benefits to all the described data types as is currently available in traditional DWs for structured data only. The Cloud Warehouse enables the integration and analysis of all types of data using the developed data model and query language. As a distinguishing feature, the Cloud Warehouse is protected by an all-encompassing "shield" that provides integrated privacy management. All queries to the CW must pass through, and be approved by, the shield, thus ensuring that privacy is not violated.

The overall idea in the vision is to repeat the "data warehouse success" when integrating all the different types of data found in the cloud for analysis purposes. Basically, this means that data of a particular type should only need to be "integrated" once, and the results of the integration should be put into a common, "harmonized" data store that can accommodate all these types of data (or their derivations) and still support data analysis tasks very well.

The concept of the Cloud Warehouse (CW) is illustrated in Fig. 1. Starting from the center of the figure, we see that the CW has the form of the cube, meaning that it is based on multidimensional modeling principles. Inside the CW, we see that the content has different "shades." Intuitively, this means that data in the CW are "not just black and white." More concretely, this refers to the fact that all data in the CW have a built-in notion of "perfection," i.e., precision and certainty of the data. Thus, data may be very precise and totally certain (like ordinary DW data) or quite imprecise and rather uncertain, e.g., due to sampling errors or due to the fact that the data come from an analytical model rather than a traditional data source.

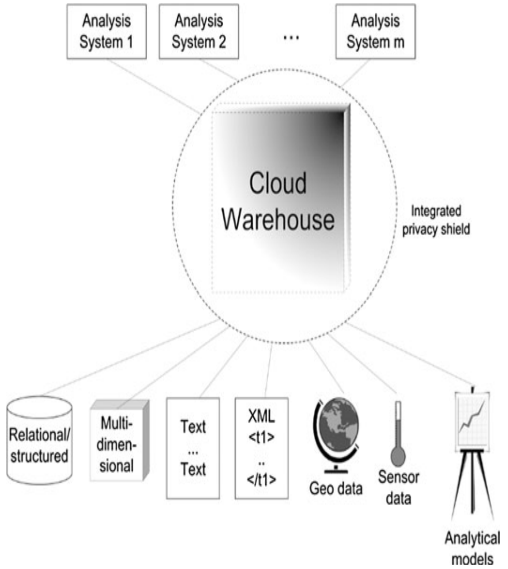

Fig. 1 The Cloud Warehouse

At the bottom of the figure, sources with different types of data is connected to the CW through only ONE "connection" per data type, e.g., one for text, one for geo- data, etc. This means that the difficult task of integrating a particular type of data can (mostly) be handled once-and-for-all, by figuring out how this particular type of data should be mapped into the CW data model and developing algorithms and tools for doing this.

Similarly, the analysis systems only have one "connection" each to the CW. This means that they can take advantage of all the functionality and types of data available in the CW, and they are also relieved of the very difficult tasks of performing the integration of different types of data themselves (as the systems in the previous section had to do).

Finally, the CW is surrounded by an "integrated privacy shield.” This shield makes sure that data privacy is not violated in the CW. This can be done in two situations. First, when data from the sources is coming in, the shield may analyze the data and find that certain modifications (aggregation, swapping, randomization, . . .) needs to be performed on the data *before* it is placed in the CW. Second, when data are re- quested from an analysis system, the CW may decide to perform modifications to the query or the query result in order to protect privacy.

In summary, the CW approach means that the “complexity” of the integration of all the different types of data for n types of data and m analysis systems drops to $n + m$ (from $n * m$). The “hard” tasks such as integrating a new type of data or protecting privacy are generally handled **only once, by the CW rather than in the** analysis systems, meaning great relief for the development of the analysis systems.

The basis for the CW will be a novel kind of data model. This model should encompass the best of several worlds. First, it should support multidimensional modeling concepts, as these have proven superior for analysis purposes. Second, it should support the flexibility and generality found in semi-structured data models. At the same time, the model should be capable of supporting a much wider range of data.

Specifically, support will be added for handling geo-related data (geo models, etc.), sensor data (high speed data streams, missing or incorrect values, etc.), semi-structured and unstructured data (enabling analysis across structured, semi- structured, and unstructured data), and imperfect (imprecise, uncertain, etc.) data. As mentioned above, support for privacy management will also be built into the model.

In this context, there are almost endless possibilities for research. It is relevant to explore query languages, query processing/optimization techniques, data integration techniques, and techniques for integrating databases, sensors, and analytical/predictive models of data.

The creation of the Cloud Warehouse will provide the same benefits to all the de- scribed data types as is currently available in traditional DWs for *structured data only*. The CW enables the integration and analysis of all types of data using the developed data model and query language. A distinguishing feature of the CW is the protection by an all-encompassing “shield” that provides integrated privacy management. All queries to the CW must pass through, and be approved by, the shield, thus ensuring that privacy is not violated.

The concrete integration approach presented in the following can be seen as a first step toward providing a semantic foundation for the on-demand multidimensional data integration needed for realizing the cloud warehouse vision.

VII.CONCLUSION

Motivated by the challenges posed by cloud intelligence and cloud computing in general, this paper focused on the combination of data source agility and data- as-a-service virtualization and its use for cloud intelligence. After introducing the novel vision of the Cloud Warehouse, the paper presented a comprehensive semantic foundation for on-demand multidimensional data integration, including formal data models, a range of query operators and re-write rules for optimization. This semantic foundation provided a sound formal basis for on-demand multidimensional data integration which is a cornerstone of cloud intelligence. Specifically, the paper presented an approach for the federation of OLAP and XML data, given in terms of a formal data model and algebraic query language.

To demonstrate the approach we introduced a federated query language, SQL$XM$ , incorporating the XML query language XPath into a subset of SQL adapted to multidimensional querying. SQL$XM$ al- lows XML data to be used directly in an OLAP query to *decorate* multidimensional cubes with external XML data, and to *group* and *select* cube data based on XML data values. The incorporation of XML data in cubes was made such that semantic problems were avoided, e.g., when aggregation was performed on the resulting cube no double-counting of data could occur.

REFERENCES

1. Abadi DJ (2009) Data management in the cloud: limitations and opportunities. IEEE Data Eng Bull 32(1):3–12. Special issue on cloud data management

2 Body M, Miquel M, Bédard Y, Tchounikine A (2003) Handling evolutions in multidimensional struc- tures. In: Proceedings of ICDE, pp 581–591

3. Agrawal R, Gupta A, Sarawagi S (1997) Modeling multidimensional databases. In: Proceedings of the thirteenth international conference on data engineering, pp 232–243

4. Beyer KS, Ercegovac V, Krishnamurthy R, Raghavan S, Rao J, Reiss F, Shekita EJ, Simmen DE, Tata S, Vaithyanathan S, Zhu H (2009) Towards a scalable enterprise content analytics platform. IEEE Data Eng Bull 32(1):28–35. Special issue on cloud data management