CLOUD ALGEBRA FOR HANDLING UNSTRUCTURED DATA IN CLOUD DATABASE MANAGEMENT SYSTEM

Mansaf Alam

Department of Computer Science, Jamia Millia Islamia, New Delhi
mansaf_alam2002@yahoo.com

ABSTRACT

The handling of unstructured data in database management system is very difficult. The managing unstructured data like image, video textual data etc. are not easy task in database system. In this work a concept of cloud algebra introduced to handle unstructured data in CDBMS. The most popular concept, relational algebra is used in relational database management system. The relational algebra is helpful in query processing in SQL. Another concept, Object Algebra which is used in object oriented database management system for query processing. The object algebra is useful to manage the object in object oriented database management system. Now the concept of Cloud computing is introduced in new the era of computer technology. The concept of cloud is also introduced in the field of databases. The cloud database management system is newly introduced in the field of database technology to manage the cloud data. This work introduces the concept of cloud algebra for handling unstructured data the cloud database management system. This work proposed the concept of cloud algebra for query processing for unstructured data in the cloud. The data are spread over the internet as cloud. The creation of new cloud of unstructured data and setting the relationship among various cloud of unstructured data are facilitated by cloud algebra. The updating, deleting and retrieval of unstructured data in cloud are done by cloud algebra. The cloud algebra provides powerful computation while using the query processing in CDBMS for handling unstructured data.

KEYWORDS

Unstructured data, CDBMS, Cloud Algebra, cloud, Object algebra, Operator sets and relation

1. INTRODUCTION

The generally unstructured data access with object which is the concept of ORDBMS and OODBMS, objects, which consist of video, audio, images, textual data and even executable code such as Java applets, are becoming readily employed by desktop, network, and Internet applications[15]. The production of data are increasing exponentially, it is very difficult to manage the data easily in RDBMS and ODBMS. The concept of CDBMS is newly introduced in the field of database technology. Now the different natures of data are available on the internet as cloud. It is necessary to introduce the new mathematical technique for manipulation of data in cloud. The mathematical algebra is use to proposed the concept of cloud algebra. Some mathematical algebra is used in cloud algebra for powerful computation while performing the query processing. Reda Alhajj and M.Erol Akun have described in their paper that object algebra is as powerful as relational and nested relational algebra [1]. There are many query algebra have been developed for the OODBS on the basis of different data models, the association algebra [3] is one of them. A Method [2] is proposed for finding the relationship degree between two objects assuming in the relationships between classes in the fuzzy object oriented database model on the basis of constructing fuzzy association algebra based on a fuzzy object oriented data model [4].

DOI : 10.5121/ijccsa.2012.2603

The well known relational algebra is a query language for the relational model, this is pure theoretical concept used in RDBMS. The relational algebra is hidden from the user at the user interface level. Relational Algebra is more operational and very useful plan for representing execution. The Cloud Computing & Databases [5] in this, it is explained that the Cloud database using patterns, which are evolving. The business adoptions accelerate evolution of these technologies. In the beginning, the cloud databases provide the services of consumer applications. In the earlier these applications use a priority on read access, because the read and write access ration was very high. The purchase criteria of this technology was high-performance read access. To implement the cloud database management system there should be a set of rule that state how Cloud database management system will behave and operate. There for a set of rules are defined in this work for CDBMS. This work is done with the motivation of relational algebra and object algebra for object oriented database management system. The cloud algebra for unstructured data is proposed in this research work. The nature of unstructured data is complex. The cloud algebra developed in this research work will be beneficial in managing the unstructured data like textual, image, video data in cloud environment.

2. RELATED WORKS

The relational algebra is most popular concept. These concepts are widely used in relational database management system (RDBMS). The relational algebra is explained in fundamental of database systems [6] that the basic set of operations for the relational data model and the sequences of relational operation. Object algebra [1] is developed for object oriented databases management system. The concepts of mathematical algebra are used in object algebra. A query algebra for program databases [7] this approach is for modelling of source code as many sorted algebra. X. Sean Wang [8] introduced temporal algebras on a temporal database model that incorporate multiple temporal types and also introduced the notion of data-domain independence. Array algebra for Database application [9] provides a complete and algebraic programming environment for database application system where large numbers with most structured data are involved. The pattern oriented relational algebra [10] is uses for characterization and rewrite quires for performance optimization. Object Algebra for the ODMG Standard [11] present object algebra based on formal model of object oriented database management system. The currently research work on unstructured data management[16] present an architecture for making MDM text-aware and showcase its implementation as IBM Info-Sphere MDM Extension for Unstructured Text Correlation, an add-on to IBM InfoSphere Master Data Management Standard Edition. They have also highlighted that how MDM benefits from additional evidence found in documents when doing entity resolution and relationship discovery. Managing Unstructured Data [18] discuss that the unstructured data are going mainstream in various organizations, they expect that the same evolution to happen in managing unstructured data objects. They have also discussed that the recognition of unstructured data as a vital corporate asset and has forced leading IT company to utilize enterprise content management systems (ECMS) for managing all non-structured data objects. This type of systems allow for searching storing, securing , retrieving and administering content from a centralized base in the same method that a DBMS allows administering structured data.

3. CLOUD DATABASE

The cloud database is run on cloud computing platform such as Amazon EC2, GoGrid and Rackspace. Using virtual machine image, users can run database on cloud independently. The user can purchase access to a database service. This database service is maintained by cloud database service provider. Providing Database as a Service [12] shows that “database as a service” providing its customer’s seamless mechanisms to create, store, and access their databases at the host site. The data are spread over the internet at various database servers, the group of data is known as cloud database. The databases are available in various data center connected via internet. Any user required a particular database, they can use database as service provide by cloud database provider. The architecture of Cloud Databases [17] is designed for very high and consistent performance, along with container-based virtualization instead of dedicated storage network, high performance SAN storage, and a traditional hardware virtualization.

4. OPERATOR SET AND RELATION

The cloud algebra [19] used various mathematical operator set to manipulate the cloud data in cloud database management system. There are four mathematical operator sets are Union, Intersection, deference and Cartesian product that can be used in cloud database management system manipulation of data. Let there are two data centers A and B in the cloud. Both the data centers are joined with the internet as shown in the Figure 1. Two data centers connected via internet.

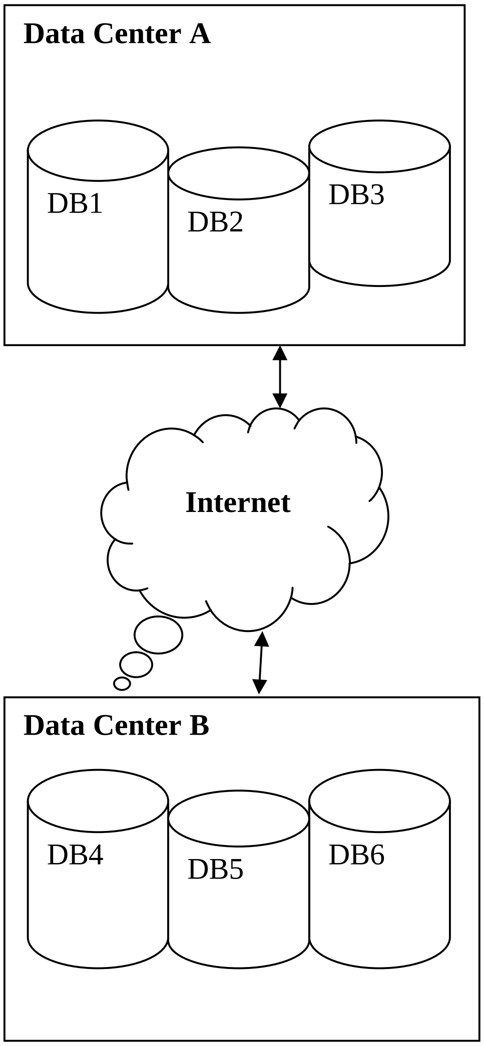

Figure 1. Two data centers connected via internet

In the above figure DB stand for database, the center A have three database namely BD1, BD2 and DB3 respectively and data centre B have again three databases DB4,DB5 and DB6 respectively.

The database of data centre A is defined mathematically as follows:
DB1={Product, Supplier}
DB2={Customer, Product}
DB3={Supplier, Customer}

Similarly database for data centre B are defined as follows:

DB4={Product, supplier, customer}
DB5={Project, Client} and
DB6={Client, supplier, product}.
The databases are in two clouds namely cloud1 and cloud2. The cloud1 is group of database DB1,DB2 and DB3. The cloud2 is a group of databases DB3, DB4 and DB5. The cloud1 and cloud2 are mathematically defined as follows:
Cloud1={DB1,DB2,DB3} and Cloud={DB4,DB5,DB6}.

4.1 Union Operator

The Union operator is applied on cloud1 and cloud2 as follows:

Cloud1∪ Cloud2= {DB1, DB2, DB3, DB4, DB5, DB6}
The union operator will help to get all databases from data center and Data Center B.

4.2 Intersection Operator

The situation in which user want to get the data available in both the data centers then the Intersection operator can be applied to fulfil the requirement of users. The intersection operator can be applied on Cloud1 and Cloud2 as follows:
Cloud1 ∩ Cloud2={φ},This result shows that a common database is not available in both data centre A and B.

4.3 Cartesian product

There is a situation in which user wants to access the data from both the data center A and B. The users can first join the both centre with the help of Cartesian product, the desired database can be accessed. The Cartesian product of data centre A and B are as follows:

Cloud1XCloud2= {(DB1, DB4),
(DB1, DB5),
(DB1, DB6),
(DB2, DB4),
(DB2, DB5),
(DB2, DB6),
(DB3, DB4),
(DB3, DB5),
(DB3, DB6}

The user can access DB1 from data centre A and Database DB4, DB5 and DB6 from data Centre B or DB2 from data centre A and DB4, DB5 and DB6 from data Centre B or DB3 and DB4, DB5 and DB6 from data Centre B. Now consider DB1 from data center A and DB4 from data center B, then DB1xDB4 is defined as follows:

DB1= {Product, Supplier} and DB4= {Product, supplier, customer}.
DB1xDB4= {(product, product), (Product, supplier), (Product, supplier),(product, customer),(Supplier, Product), (supplier, supplier),(supplier, customer)} now use the intersection operator consecutive two elements of set DB1xBD4 to common element as follows: DB1xDB4={(product, product) ∩ (Product, supplier) ∩ (Product, supplier) ∩ (product, customer) ∩ (Supplier, Product) ∩ (supplier, supplier) ∩ (supplier, customer) this result will be helpful in manipulation the data in different databases in different data centers.

4.4 Difference Operator

This operator will be helpful to find the difference between the databases from two different data centers. This operator can be used in cloud mathematically as follows:

Cloud1 - Cloud2= {D1, DB2, DB3}. The database DB4, DB5 and DB6 are not available in data centre A. This operator set is not applicable in the case of two different data centre because database member of one data centre can't be member of another data centre.

4.5 Select Operator

Select operator is used to select the structured and unstructured data from different databases available in different data centers. Consider d is data may be structured and unstructured in database DB1 is available in cloud A. The select operation is defined as follows:

Select (d, DB1, A, c) = {d|(d in DB1)∩c∩A} where first DB1 will be search in cloud A the data will be search in Database DB1 with the certain condition c. When the given condition is satisfy the select return the result to the user who is searching for the data. Select operation [13] is used in A Query Algebra for Object-Oriented Databases. Select can be used as follows:

Select d1, d2, d3----------Dn from (DB1 in A, DB2 in B , ... , DBn in A, B) where c(DB1, DB2---------DBn.

In the above example d1, d2, d3, -----------------------dn are the data, and DB1, DB2-----------------------------------.DBn are databases available in data centre A and B. The c is the condition on which data will be retrieved.

5. Cloud Algebra for Handling Unstructured Data

In this section, it is discuss that how cloud algebra handle the unstructured data. The structured of unstructured data is not constrained by schema. Let us consider a graph with the set of node N, the element of N are a, b, c and d, it is represented in set as N={a, b, c, d}. The edge of this graph E is a set of edges, e1, e2, e3 and e4, it can be represented as E= {e1, e2, e3, e4}. The graph with this specification is given below:

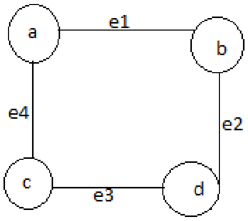

Figure 2. Unstructured graphical data

The above unstructured data diagram is represented in matrix form as follows:

$$N= \begin{array}{ccccc} & a & b & c & d \\ a & 0 & 1 & 1 & 0 \\ b & 1 & 0 & 0 & 1 \\ c & 1 & 0 & 0 & 1 \\ d & 0 & 1 & 1 & 0 \end{array}$$

$$N^{|} = \begin{array}{ccccc} & a & b & c & d \\ a & 1 & 0 & 0 & 1 \\ b & 0 & 1 & 1 & 0 \\ c & 0 & 1 & 1 & 0 \\ d & 1 & 0 & 0 & 1 \end{array}$$

The graph of matrix $N^{|}$ which is compliment of given graphs represented in matrix N is given below:

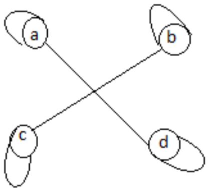

Figure 3. Compliment of Figure 2

$$NxN'= \begin{array}{ccccc} & a & b & c & d \\ a & 0 & 1 & 1 & 0 \\ b & 1 & 0 & 0 & 1 \\ c & 1 & 0 & 0 & 1 \\ d & 0 & 1 & 1 & 0 \end{array} \quad x \quad \begin{array}{ccccc} & a & b & c & d \\ a & 1 & 0 & 0 & 1 \\ b & 0 & 1 & 1 & 0 \\ c & 0 & 1 & 1 & 0 \\ d & 1 & 0 & 0 & 1 \end{array} \quad = \begin{array}{ccccc} & a & b & c & d \\ a & 0 & 0 & 0 & 0 \\ b & 0 & 0 & 0 & 0 \\ c & 0 & 0 & 0 & 0 \\ d & 0 & 0 & 0 & 0 \end{array}$$

$$N+N'= \begin{array}{ccccc} & a & b & c & d \\ a & 0 & 1 & 1 & 0 \\ b & 1 & 0 & 0 & 1 \\ c & 1 & 0 & 0 & 1 \\ d & 0 & 1 & 1 & 0 \end{array} + \begin{array}{ccccc} & a & b & c & d \\ a & 1 & 0 & 0 & 1 \\ b & 0 & 1 & 1 & 0 \\ c & 0 & 1 & 1 & 0 \\ d & 1 & 0 & 0 & 1 \end{array} = \begin{array}{ccccc} & a & b & c & d \\ a & 1 & 1 & 1 & 1 \\ b & 1 & 1 & 1 & 1 \\ c & 1 & 1 & 1 & 1 \\ d & 1 & 1 & 1 & 1 \end{array}$$

The above computation show that $N * N^{|} = e$, where * is define as set of operator {X, +}. If the operator * is treated as x operator then $NxN^{|} = 0$ and the operator * is treated as + then $NxN^{|} = 1$. The statement $NxN^{|} = 0$ means the result of applying x operator on a graph and its compliment give the result in a graph represented in matrix form which all elements are 0. The statement $NxN^{|} = 1$ means applying the + operator on a graph which is unstructured data and its compliment the resultant will be a graph represented in matrix form which all elements will be 1. The different unstructured data are available in different data centres. The cloud algebra will be used there to handle the unstructured data available in data centres. The graph is an example of unstructured data. The handling of unstructured data is illustrated with the help of example here. This cloud algebra can be used to handle the unstructured data available in data centre as well as on remote server. This can work at database as a service[20] for providing the services to the users by service provider. This can also be used in private cloud, Public cloud, External cloud, hybrid cloud and community cloud.

4. CONCLUSION

In this work mathematical algebra is applied on unstructured data in cloud database management system (CDBMS). The union operator is used to combine the different data center for searching a particular unstructured data item from the different data centre where the unstructured data is stored. The intersection operator is used to select the unstructured data item available in different data centers and Cartesian product is used to join different data center for manipulate the unstructured data item in deferent data centers. The concept of cloud algebra is proposed in this research work for handling the unstructured data, it is done with the motivation of relational algebra and object algebra for object oriented database. The various mathematical operators are used to handling the unstructured data. This cloud algebra for handling the unstructured data will work at databases as a service.

ACKNOWLEDGEMENTS

I thank almighty blessing me and giving wisdom and strength throughout my work. I would like thank all staff members of DCS as well as my family and friends for their kind cooperation.

REFERENCES

[1] Reda Alhajj and M. Erol Arkun, An object algebra for object oriented database Systems, ACM SIGMIS Database, Volume 24 Issue 3, Aug. 1993, ACM New York, NY, USA.

[2] Doan Van Ban, Ho Cam Ha and Vu Duc Quang, Querying Fuzzy Object-Oriented Data Based On Fuzzy Association Algebra, 2011 Third International Conference on Knowledge and Systems Engineering, IEEE,14-17 Oct. 2011, pp: 40 – 47, Hanoi.

[3] Stanley Y.W.Su, Mingsen Guo, Herman Lam, Association Algebra: A Mathematical Foundation for Object- oriented Databases, IEEE,Tran. On Knowledge and Data Engineering, 5 (5), 1993, 775 – 798.

[4] Selee Na; Seog Park, A fuzzy association algebra based on a fuzzy object oriented data model, Computer Software and Applications Conference(COMPSAC 96), IEEE, 21-23 Aug 1996, pp: 276 – 281, Seoul, South Korea.

[5] Mike Hogan, Cloud Computing & Databases: How databases can meet the demands of cloud computing, ScaleDB Inc., November 14, 2008.

[6] Elmasri and Navathe, Fundamental of database systems: The Relational Algebra and Relational Calculus, sixth edition, Pearson Education, Inc. Publishing as Pearson Addison-Wesley, 2011.

[7] Santanu Paul and Atul Prakash,A query algebra for program databases, IEEE transaction on software engineering, vol.22,no.3,march 1996.

[8] X. Sean Wang, Algebraic Query Languages on Temporal Databases with Multiple Time Granularities, Technical report ISSE-TR-94-107,Revised April 1995.

[9] Levent Orman, An Array Algebra for Database Applications, Journal of Management Information Systems/spring, vol 1, no 4, Page [44] of 44-56, 1985.

[10] Arnaud Giacometti, Patrick Marcel and Arnaud Soulet, A Relational view of Pattern Discovery, International Conference(DASFAA2011),pp:153-167,LNCS, Springer- Verlag, Brlin Heidelberg, 2011.

[11] Alexandre Zamulin, An Object Algebra for the ODMG Standard, ADBIS2002, LNCS 2435, pp. 291–304, Springer-Verlag Berlin Heidelberg , 2002.

[12] Hakan Hacıg¨um¨us, Bala Iyer, and Sharad Mehrotra, Providing Database as a Service, University of California. Irvine, CA 92697, USA.

[13] Gail M. Shaw, et. al., A Query Algebra for Object-Oriented Databases, Department of Computer Science, Brown University, Providence, Rhode Island-02912,Cs-89-19,1989.

[14] Mansaf Alam, Six Layers Architecture Model for Object Oriented Database, International Journal of Computer Science Issues(IJCSI, Republic of Mauritius, 2011.

[15] Alexander P. Pons and Hassan Aljifri, Handling Unstructured Data Type in DB2 and Oracle, Communications of the International Information Management Association, Volume 3, Issue 2, pp. 117-128, 2003, USA.

[16] Karin Murthy, Prasad M Deshpande, Atreyee Dey, Ramanujam Halasipuram, Mukesh Mohania, Deepak P, Jennifer Reed, Scott Schumacher, Exploiting Evidence from Unstructured Data to Enhance Master Data Management, Proceedings of the VLDB Endowment, Vol. 5, No. 12, 2012, Istanbul, Turkey.

[17] Rockspace(The open Cloud Company), The high performance MySQL database on the cloud (Deliver faster applications on the first relational database service built on OpenStack), http://www.rackspace.com/cloud/public/databases/.

[18] Larissa Moss, Managing Unstructured Data, Volume 3, Issue 9 - October 2009.

[19] Mansaf Alam, Cloud algebra for cloud database management system, cciet2012, Coimbatore, published by ACM- ACM International Conference Proceeding Series (ICPS), October, 2012.

[20] Carlo Curino, Evan P. C. Jones, Raluca Ada Popa, Nirmesh Malviya, Eugene Wu, Sam Madden, Hari Balakrishnan, Nickolai Zeldovich, Relational Cloud: A Database-as-a-Service for the Cloud, 5th Biennial Conference on Innovative Data Systems Research, Asilomar, CA, January 2011.

Author

Dr. Mansaf Alam has been working as Asstt. Professor in the Department of Computer Science, Faculty of Natural Sciences, Jamia Millia Islamia, New Delhi-110025. He has published several research articles in reputed International Journals and Proceedings of reputed International conferences published by IEEE, Springer, Elsevier Science, and ACM.