Volume 5, Issue 7, July 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Integration of CWM and CAKE Architecture

Sidra Anam
M.Tech Scholar, CSE Department,
Pranveer Singh Institute of Technology,
Kanpur, Uttar Pradesh, India

Saurabh Gupta
Assistant Professor CSE Department,
Pranveer Singh Institute of Technology,
Kanpur, Uttar Pradesh, India

Abstract- The advancements in computing and communication over networks have resulted in the use of distributed environments. The large volume of data in such environment is distributed. The monitoring and analysis of such data need Distributed Mining technologies. Metadata captures all kinds of information necessary to analyze, design, build, use and interpret the data warehouse contents. To enable the interoperability between repositories and tool integration within data warehousing architectures, a standard for metadata representation and exchange is needed. This paper shows an approach which combines the Distributed Data Mining (DDM) using Multi-Agent System (MAS) technology with the architecture of CWM (Common Warehouse Metamodel). A data mining technique of "CAKE" (Classifying, Associating & Knowledge DiscovEry) is used. The model architecture is distributed, uses knowledge-driven mining technique and flexible enough to work on any data warehouse. The problem of heterogeneity i.e., use of different RDBMS at different distributed remote sites where Data warehouse resides, has been solved by making use of CWM. Integrating the architecture of CWM with CAKE architecture is the proposed solution.

Keywords: Distributed Data Mining, Multi-Agent System, Common Warehouse Metamodel, Classifying, Associating, Knowledge Discovery.

I. INTRODUCTION

In today's scenario, the increasing processing power and sophisticated technologies has increased the business needs of people they expect more from systems. The computer systems are not only used for storing data but also for providing information, forecasting business trends, analyzing the future patterns. All these needs have bought up the concept of Data Warehouse.

Extract, Transformation & Loading also known as ETL is a Data Warehouse pre-process that comprises of activities which are executed to transform the subject- oriented summarized data from Operational Systems to Data Warehouses. Many pre- processes are required to be performed in-order to remove the outliers, inconsistency and incompleteness of data. This is done to maintain the quality of data and its result.

- Data Cleaning: Removal of outliers and resolving inconsistency.
- Data Integration: Data from multiple sources are combined in a single standardized format.
- Data Transformation: Making data from multiple sources to a single aggregation form.
- Data Reduction: It is a selection process of required data to be available in data warehouse.
- Data Discretization: It deals with reduction of data, especially numerical data.

Data Warehouse deals with the storing of subjected data. Now what we require is the information in a manner that can help in accomplishing our objectives. The process of extracting and viewing of data in a Multidimensional Model is part of OLAP (Online Analytical Processing), which are set of tools used by knowledge workers or business domain experts to dig out what they exactly required. Data warehouses basically adopt three-tier architecture. The bottom tier is a warehouse database server, which is almost always a relational database system. The tools are used to fed the data into bottom tier from operational databases. The middle tier is the OLAP server that maps the operations on multidimensional data to standard relational operations. The top tier is a front-end client tool (web interface) which contains query, reporting and analysis tools. The use of intelligent agents in PArallel Data Mining Agents (PADMAs) makes this process of data warehousing fast and efficient. But data is still coming from multiple heterogeneous sources. Different organizations have different formats of storing their data. The syntax and semantics of organizations differs on the basis of their needs and usage of data. So, if we make the use of CWM then data warehousing can be done efficiently.

II. RELATED WORK

A term in Data Mining has been introduced known as "Distributed Data Mining (DDM)", it's an approach of performing Data Mining on Distributed Data Warehouses over different remote locations, which either contains the same data distributed over different locations or different data related to the same subject. There are several DDM approaches, which are developed using MAS that includes BODHI, PADMA and JAM, all these approaches deals with the centralized architecture. While another approach of DDM known as Papyrus deals with Peer-to-Peer (P2P) working style [5]. PArallel Data Mining Agents (PADMAs) is Multi-agent based architecture for Data Mining. It is a system that

makes use of Intelligent Data Mining Agents, which are responsible for accessing, analyzing and discovering the hidden patterns within the Data Warehouse. They all work together in conjunction with each other and share same repository or meta-data [6]. Data Mining techniques includes Classifying, Associating and Knowledge Discovery (CAKE), which are used to mine data on the basis of different defined rules and patterns, as shown in figure 1.

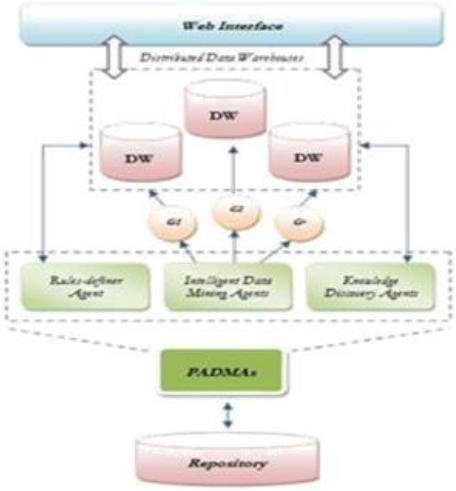

Figure 1: CAKE (Architecture) [6]

III. PROPOSED ARCHITECTURE

The architecture of CAKE was based on 4-tier. Distributed Data Mining (DDM) is implemented using PArallel Data Mining Agents (PADMAs) using centralized meta-data, which contains all the rules of Classification and Association along with the data structure details and Web Interface is provided to the users for viewing the mined results. There could be any Reporting or Analytical tool that can be used to satisfy the needs of the end-user. Three Multi-purpose agents are used which are described as follows:

1.1 Rule-definer Agents. These are used to define the Meta-data of Data Warehouse on the basis of rules that are to be defined by the users. These rules are then used by the Intelligent Data Mining Agents for the purpose of Data Mining and by Knowledge Discovery Agents for extracting the knowledge out from the defined patterns.

1.2 Intelligent Data mining Agents. These groups of agents works together to mine the data and compute the desired result.

1.3 Knowledge-Discovery Agents. These agents are used to determine the final computed results along with the explanation on the computed data. These decisions are taken on the basis of defined requirement in the repository.

But this architecture was based on centralized repository data and supporting distributed data mining approach. The end user queries are now been processed faster and efficiently. But many problems are identified, some are addressed, some are still needed to be resolved. Few of them are as follows:

- Heterogeneous: Using of different RDBMS at different distributed remote sites where Data warehouse resides.
- Complexity: More architecture redesigning will be required to cater complex mining needs.

IV. SOLUTION TO THE PROBLEM

The problem of heterogeneity can be solved by using CWM (Common Warehouse Metamodel). Data warehousing and business intelligence often involve the use of a variety of tools and products, each with its own definition and format for metadata. Creating, sharing and managing the metadata for these tools and products is time consuming and error prone [1]. The solution was given by CWM. It is a successful framework which provides a standard language for defining the structure and semantics of metadata in a formal way, a standard interchange mechanism for sharing metadata defined in the standard language and a standard common specification that defines the structure and semantics of shared metadata in data warehousing and business intelligence. CWM is a complete specification of the syntax and semantics needed to export/import shared warehouse metadata and the common warehouse metamodel. It is used for following purposes:

- Interchange of all warehouse metadata including both technical metadata and business metadata.
- Interchange of metadata that describes all warehouse data elements including data sources, transformations and data targets.
- Interchange of metadata that describes all warehouse processing elements including scheduling, status.
- Interchange of metadata that describes informational data and the use of major types of informational data models (such as multidimensional and hierarchical classification) for representing informational data.
- Interchange of metadata that describes operational data and the use of major types of operational data models (such as relational, object-oriented and hierarchical) for representing operational data [1].

Now, the idea is to integrate the existing architecture of CWM with the CAKE architecture. This can be done by combining the two architectures. Classification, Association and Knowledge Discovery using PADMAs is made more efficient by introducing the CWM. The multi-agents will execute the user queries faster and CWM act as a standard

interchange interface for exchanging data from different platforms (figure 2). The CWM will provide all the necessary interfaces, languages and specifications for storing the data in a standard format so that data can be used by the business analysts for analysis of all data and identifying the future trends. These data will help the companies or organizations to increase their profits and overall sales. The processes of classification, association and knowledge discovery of rules are done by the multi-agent system. The queries of end user from the web interface will be now based on the Standards as defined by the CWM. The data from different sources need to follow the standards and semantics defined by the CWM. All the analysis data is combined and then fed to the intelligent agents. The agents perform their respective tasks like the Rule Definer Agent will define the rules of metadata, the Intelligent Data Mining Agent will apply the standard data mining techniques to the rules which are defined by Rule Definer Agent and finally the Knowledge Discovery Agent will extract the result for end user queries. The proposed architecture is as follows:

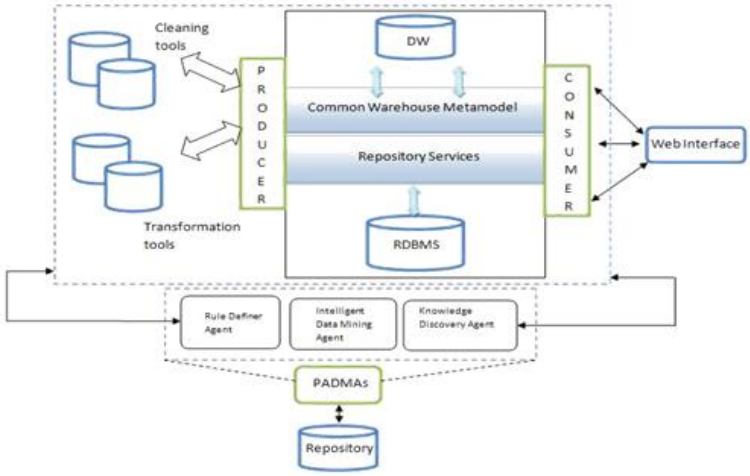

Figure 2: Integrated architecture of CAKE and CWM

V. CONCLUSION AND FUTURE SCOPE

The proposed architecture can be efficiently used for the purpose of data mining even we have the industries or organizations which have different formats of storing their data. The heterogeneity of data can now be removed by implementing Common Warehouse Metamodel (CWM) with the existing architecture of CAKE. The use of Multi-Agent System will answer the user queries faster. The CWM will provide all the necessary interfaces, languages and specifications for storing the data in a standard format so that data can be used by the business analysts for analysis of all data and identifying the future trends. The future work can be based upon reducing the complexity of the system and mining the data faster.

REFERENCES

[1] Dr. Daniel T. Chang, "Common Warehouse Metamodel (CWM), UML and XML", IBM Database Technology Institute, Meta Data Conference, March 19-23, 2000.

[2] OMG CWMI RFP Web page http://www.omg.org/techprocess/meetings/schedule/CWMI_RFP.html CWM Specification (ad/2000-01-01).

[3] CWM Forum Web site http://www.cwmforum.org/

[4] Chris Clifton, "Privacy Preserving Distributed Data Mining", Department of Computer Sciences, November 9, 2001.

[5] PArallel Data Mining Agents (PADMA) http://www-fp.mcs.anl.gov/ccst/research/reports_pre1998/algorithm_development/padma/kargupta.html

[6] Danish Khan , "CAKE – Classifying, Associating & Knowledge DiscovEry An Approach for Distributed Data Mining (DDM) Using PArallel Data Mining Agents (PADMAs)", 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology.

[7] OMG CWMI RFP Web page http://www.omg.org/techprocess/meetings/schedule/CWMI_RFP.html CWM Specification Volume 2. XML, IDL and DTD(ad/2000-01-02).

[8] OMG CWMI RFP Web page http://www.omg.org/techprocess/meetings/schedule/CWMI_RFP.html CWM Specification Volume 3. Extensions (CWMX) (ad/2000-01-03).

[9] OMG CWMI RFP Web page http://www.omg.org/techprocess/meetings/schedule/CWMI_RFP.html CWM Specification Volume 4. Extensions XML, IDL and DTD (ad/2000-01-11).

