ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol. 3 , No.1, Pages : 533– 536 (2014)
Special Issue of ICETETS 2014 - Held on 24-25 February, 2014 in Malla Reddy Institute of Engineering and Technology, Secunderabad– 14, AP, India

DATA CENTRIC KNOWLEDGE MANAGEMENT SYSTEM

R. SRAVANTHI [1], A. KALYANI [2]

[1]*Student, Malla Reddy Institute of Engineering and Technology, sravanthi.rachuri@gmail.com*
[2]*Asst. professor, Malla Reddy Institute of Engineering and Technology, @gmail.com*

Abstract-

The purpose of the Data Centric Knowledge Management System (DCKMS) is to centralize knowledge generated by employees working within and across functional areas, and to organize that knowledge such that it can be easily accessed, searched, browsed, navigated, and curated.

DCKMS is a web based application which allows employees of a company to share their knowledge with others in the company. Also it allows them to search for knowledge assets when in need. It provides a facility for the employees to register themselves as 'experts' as well as search for other 'experts' incase of any problem/requirement in their project. It is a one stop shop for finding solutions for your problems.

Index Terms: **Datamart, Data warehouse, Clustering, Information management, Data port, Data Centric.**

I.INTRODUCTION:

This is every employee needs some help at some point of time. To solve some issues or bugs or problems employees has to depend upon many sources like internet. This is very difficult and time consuming task. Also accurate solution may not be available. Data Centric Knowledge Management System is a perfect solution to Overcome the above mentioned problems. It provides a facility to share your knowledge by submitting various knowledge assets and to search for assets when in need. It allows users to search documents based on keywords as well as name of the author, topic, category etc.

This application allows users to register themselves as experts in their favorite areas. Also allows users to find and contact experts in order to seek help from them. This application provides end to end solution to maintain shared knowledge assets in a company. It allows K-Team and Experts to evaluate the documents submitted by various employees before publishing them. Also based on this rating various awards are being awarded to employees.

This application maintains the entire data in a centralized and secured database server to maintain consistency in report generation and allows users to access from any location. This is an online application that allows multi-user access of system and to track or manage the data simultaneously. Various roles and authentications have been provided and access to various areas in the tool is restricted according to the role given to users.

This system design is modularized into various categories. This system has enriched UI so that a novice user did not feel any operational difficulties. This system mainly concentrated in designing various reports requested by the users as well as higher with export to excel options.

1. Datamart:

A data mart is a subject-oriented archive that stores data and uses the retrieved set of information to assist and support the requirements involved within a particular business function or department. Data marts exist within single organizational data warehouse repository.Data marts improve end-user response time by allowing users to have access to the specific type of data they need to view most often by providing the data in a way that supports the collective view of a group of users.

Designing:

The design step is first in the data mart process. This step covers all of the tasks from initiating the request for a data mart through gathering information about the requirements, and developing the logical and physical design of the

ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol. 3 , No.1, Pages : 533– 536 (2014)
Special Issue of ICETETS 2014 - Held on 24-25 February, 2014 in Malla Reddy Institute of Engineering and Technology, Secunderabad– 14, AP, India

Constructing

This step includes creating the physical database and the logical structures associated with the data mart to provide fast and efficient access to the data.

Populating

The populating step covers all of the tasks related to getting the data from the source, cleaning it up, modifying it to the right format and level of detail, and moving it into the data mart. More formally stated.

Accessing:

The accessing step involves putting the data to use: querying the data, analyzing it, creating reports, charts, and graphs, and publishing these. Typically, the end user uses a graphical front-end tool to submit queries to the database and display the results of the queries.

Managing:

This step involves managing the data mart over its lifetime. In this step, you perform management tasks such as the following:

1. Managing the growth of the data
2. Optimizing the system for better performance
3. Ensuring the availability of data even with system failures

2. Data Warehouse

A data warehouse is a relational database that is designed for query and analysis rather than for transaction processing. It usually contains historical data derived from transaction data, but it can include data from other sources. It separates analysis workload from transaction workload and enables an organization to consolidate data from several sources. In addition to a relational database, a data warehouse environment includes an extraction, transportation, transformation, and loading (ETL) solution, an online analytical processing (OLAP) engine, client analysis tools, and other applications that manage the process of gathering data and delivering it to business users

Subject Oriented:

Data warehouses are designed to help you analyze data. For example, to learn more about your company's sales data, you can build a warehouse that concentrates on sales. Using this warehouse, you can answer questions like "Who was our best customer for this item last year?" This ability to define a data warehouse by subject matter, sales in this case, makes the data warehouse subject oriented.

Integrated:

Integration is closely related to subject orientation. Data warehouses must put data from disparate sources into a consistent format. They must resolve such problems as naming conflicts and inconsistencies among units of measure. When they achieve this, they are said to be integrated.

Nonvolatile:

Nonvolatile means that, once entered into the warehouse, data should not change. This is logical because the purpose of a warehouse is to enable you to analyze what has occurred.

Time Variant

In order to discover trends in business, analysts need large amounts of data. This is very much in contrast to online transaction processing (OLTP) systems, where performance requirements demand that historical data be moved to an archive. A data warehouse's focus on change over time is what is meant by the term time variant.

ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol. 3 , No.1, Pages : 533– 536 (2014)
Special Issue of ICETETS 2014 - Held on 24-25 February, 2014 in Malla Reddy Institute of Engineering and Technology, Secunderabad– 14, AP, India

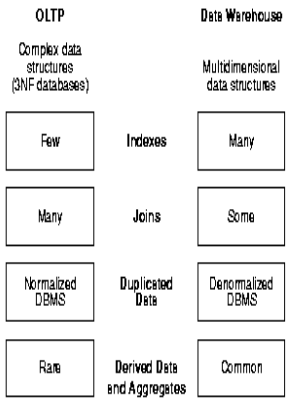

Figure 1-1 Contrasting OLTP and Data Warehousing Environments

3. Clustering:

Clustering is the task of grouping a set of objects in such a way that objects in the same group (called a **cluster**) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical, data analysis, used in many fields, including machine learning, pattern reorganization, image process, information retrieval. Cluster analysis itself is not one specific algorithm, but the general task to be solved. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with small distances among the cluster members, dense areas of the data space, intervals or particular statistical distribution. Clustering can therefore be formulated as a multi_objective problem. The appropriate clustering algorithm and parameter settings (including values such as the distance function to use, a density threshold or the number of expected clusters) depend on the individual data set and intended use of the results. Cluster analysis as such is not an automatic task, but an iterative process of knowledge discovery or interactive multi-objective optimization that involves trial and failure.

Centroid-based clustering:

In centroid-based clustering, clusters are represented by a central vector, which may not necessarily be a member of the data set. When the number of clusters is fixed. K-means clustering gives a formal definition as an optimization problem: find the cluster centers and assign the objects to the nearest cluster center, such that the squared distances from the cluster are minimized.

Distribution-based clustering:

The clustering model most closely related to statistics is based on distribution model. Clusters can then easily be defined as objects belonging most likely to the same distribution. A nice property of this approach is that this closely resembles the way artificial data sets are generated: by sampling random objects from a distribution. while the theoretical foundation of these methods is excellent, they suffer from one key problem known as overriding, unless constraints are put on the model complexity. A more complex model will usually always be able to explain the data better, which makes choosing the appropriate model complexity inherently difficult.

One prominent method is known as Gaussian mixture models. Here, the data set is usually modeled with a fixed (to avoid over fitting) number that are initialized randomly and whose parameters are iteratively optimized to fit better to the data set. This will converge to a local optimum, so multiple runs may produce different results. In order to obtain a hard clustering, objects are often then assigned to the Gaussian distribution they most likely belong to; for soft clustering's, this is not necessary. Distribution-based clustering is a semantically strong method, as it not only provides you with clusters, but also produces complex models for the clusters that can also capture correlation and dependence of attributes. However, using these algorithms puts an extra burden on the user: to choose appropriate data models to optimize, and for many real data sets, there may be no mathematical model available the algorithm is able to optimize (e.g. assuming Gaussian distributions is a rather strong assumption on the data).

Density-based clustering:

In density-based clustering, clusters are defined as areas of higher density than the remainder of the data set. Objects in these sparse areas - that are required to separate clusters - are usually considered to be noise and border points. the most popular density based clustering method is DBSCAN In contrast to many newer methods, it features a well-defined cluster model called "density-reach ability". Similar to linkage based clustering, it is based on connecting points within certain distance thresholds. However, it only connects points that satisfy a density criterion, in the original variant defined as a minimum number of other objects within this radius. A cluster consists of all density-connected objects (which can form a cluster of an arbitrary shape, in contrast to many other methods) plus all objects that are within these objects' range. Another interesting property of DBSCAN is that its complexity is fairly low - it requires a linear number of range

ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol. 3 , No.1, Pages : 533– 536 (2014)
Special Issue of ICETETS 2014 - Held on 24-25 February, 2014 in Malla Reddy Institute of Engineering and Technology, Secunderabad– 14, AP, India

queries on the database - and that it will discover essentially the same results in each run, therefore there is no need to run it multiple times. OPTICS is a generalization of DBSCAN that removes the need to choose an appropriate value for the range parameterε, and produces a hierarchical result related to that of linkage clustering. DeLi-Clu, Density-Link-Clustering combines ideas from single linkage clustering and OPTICS, eliminating the εparameter entirely and offering performance improvements over OPTICS by using an R-tree index.

EXISTING SYSTEM:

Here in the existing system, the company maintains all the knowledge based documents in a separate system which will be accessible for all employees through LAN and they can post their new documents into this and access the earlier documents. Searching for related documents based on author, technology etc is a time taking process. Managing the documents category wise and restrict them not to be accessible based on the user type becomes complicated. This system doesn't restrict unnecessary documents to be posted.

Disadvantage:

Difficulty in maintaining security levels and searching for required documents.

PR0POSED SYSTEM:

The proposed system is fully computerized, which removes all the drawbacks of existing system. In the proposed system, it allows different employees of the company to upload their knowledge document into this system which will be verified by next level users to avoid unnecessary documents. Also it allows them to search for knowledge assets very easily when in need. It provides a facility for the employees to register themselves as 'experts' as well as search for other 'experts' incase of any problem/requirement in their project. It provides a facility for the evaluator to rate the documents posted by the employees.

Advantage:

It provides a facility a to share knowledge documents across the company

Conclusion and future work:

The new system, Data Centric Knowledge Management System has been implemented to cater the needs of company employees in sharing different knowledge assets effectively with role based access. The present system has been integrated with the already existing. The database was put into the My SQL server. This was connected by JDBC. The database is accessible through Intranet on any location. This system has been found to meet the requirements of the users and departments and also very satisfactory. The database system must provide for the safety of the information stored, despite system crashes or attempts at unauthorized access.

REFERENCES:

Core java volume-II Advanced features 7th edition by *Cay S.Horstmann* and *Gary Cornell* (Pearson education).

Java Servlet Programming by O'relly publishers
Java Complete Reference 5th edition by *Herbert Schildt* (Tata McGraw Hill).

Algorithm and applications in java 3rd edition by *Satraj Sahni* (Tata McGraw Hill).

Classical Data Structures by *Samantha* (Pearson education).

Java Server Programming 2.0 with complete J2EE concepts included (après).

Software Engineering practice and principles 6th edition by *Roger Pressmen* (Tata McGraw Hill).
Java How to program 5th edition *Deitel and Deitel* (Prentice Hall of India).

Internet & World Wide Web How to program 3rd edition by *Deitel & Deitel and Goldberg* (Pearson education).

Web enabled commercial application development using Java 2.0 by *Ivan Bayross* (Prentice Hall of India).

Data base System Concepts 4th edition by *Silbershatz, Korth, and Sudharshan* (Tata McGraw Hill).

Fundamentals of Data base systems 4th edition by Ramez *Elmasri and Shamkant B.Navathe* (Pearson education).