Utilization of central Repository in Data Warehouse

Abhijeet Umare Shoheb Shaikh,
Department of MCA G.H.Raisoni
College of Engineering Nagpur,Maharashtra (India)
abijeetumare77@gmail.com , shaikh.shoheb13@gmail.com

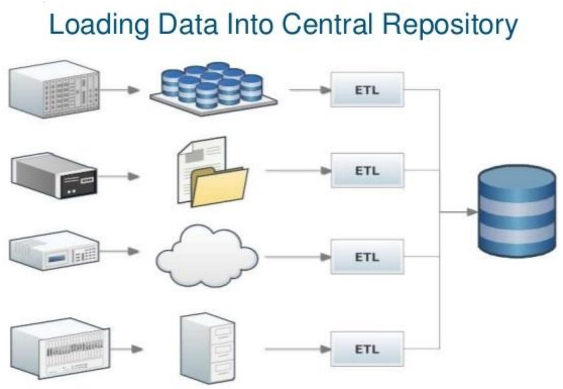
Abstract – The data warehouse is a collection of database. It very important to handle such huge amount of data from internal treats and other such problems. In order to achieve this, the central repository in data warehouse is utilized. This paper puts a light on the utilization of central repository in data warehouse. Centralized storage protects data, increases speed, convenience and efficiency. File sharing allows for quick and easy access to important data from almost anywhere in the world. Relative mobility and control of data can move a company workflow to a higher level of effectiveness. The idea of cost-effectiveness can also relate to power supplies and peripheral equipment. It is more efficient to supply a central server than to power several individual machines. The same thing applies to any peripheral equipment. It is much easier to equip a cluster than every individual machine.

Introduction-

Generically refers to a central place where data is stored and maintained. A repository can be a place where multiple databases or files are located for distribution over a network, or a repository can be a location that is directly accessible to the user without having to travel across a network. In a CASE development system, a database of information about the software, including data elements, processes, inputs, outputs and interrelationships. A CASE system uses a repository to identify objects and rules for reuse. The best practice is have local repository for each developer in their local machine and use central repository for sharing the objects i.e. jobs, data flows etc.

In computing, a data warehouse, also known as an enterprise data warehouse, is a system used for reporting anddata analysis. Data warehouse are central repositories of integrated data from one or more disparate sources. They store current and historical data and are used for creating trending reports for senior management reporting such as annual and quarterly comparisons.

The data stored in the warehouse is uploaded from the operational systems such as marketing, sales, etc. The data may pass through an operational data store for additional operations before it is used in the DW for reporting.



Data repository is a somewhat general term used to refer to a destination designated for data storage. However, many IT experts use the term more specifically to refer to a particular kind of setup within an overall IT structure, such as a group of databases, where an enterprise or organization has chosen to keep various kinds of data. Some experts refer to a data repository as a partitioning of data, where partitioned data types are stored together. It is also commonly called data warehousing.

Data Repository is a logical (and sometimes physical) partitioning of data where multiple databases which apply to specific applications or sets of applications reside. For example, several databases (revenues, expenses) which support financial applications could reside in a single financial Data Repository.

A database warehouse is one large Data Repository of all business related information including all historical data of the business organization implementing the data warehouse. Data warehousing is a complex process of building a data repository in the form of a relational database so that the company can support web or text mining in order to leverage data and transform or aggregate them into useful information.

In all cases, organizations use data warehousing to gain a competitive advantage, support for decision making processes through comprehensive data analysis.

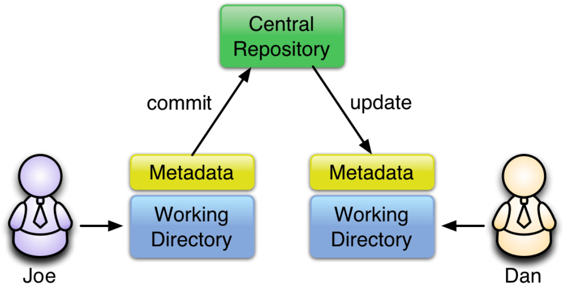
Some of the key components of data warehousing are Decision Support Systems (DSS) and Data Mining (DM).

Data volumes in data warehouse could grow at an exponential rate so there should be a way to handle this tremendous growth. With respect to storage requirements, the critical needs that need to be seriously considered in a data warehouse are high availability, high data volume, high performance and scalability, simplification and usability and easy management.

Partitioning of data into a logical or in some cases physical Data Repository could greatly help meet the requirement in relation to dealing with the exponential growth of data volumes in the data warehouse. If all the data in the data warehouse were not partitioned into several Data Repositories, then there will be profound disadvantage in terms of performance and efficiency.For one, if the central server fails, the system would come to a halt. This is because data is just located in one monolithic system, and when the hardware fails, there is no sort back up. It may take some time to get the server up, depending on the nature of the problem. But in a business company, even a few minutes of business stoppage can already translate into thousands of potential dollars lost from the business.

When Data Repository is employed in the data warehouse, the load can be distributed across many databases or even across many servers. For instance, instead of having one computer handle the database related to customers, several databases could be handling the different aspects of customers.

In a very large company such as a company that has several branches around the country, instead of having all the customers in one database, several databases may be handling different branch customer database in a data repository. Or as earlier mentioned, several company departmental database may be broken down into various Data Repository such as one data repository supporting several databases (revenues, expenses) which support financial applications (A/R, A/P) could reside in a single financial Data Repository.



Data Repository offers easier and faster access due to the fact that related information are, to some degree, lumped or clustered together. For instance, in the example with financial Data Repository, anybody from the financial department or any other data use wanting information related to financials will not have to dig through the entire volume of the data in the data warehouse.

For database administrators, employing Data Repository means a lot easier way to maintain the data warehouse system because of the compartmentalized nature. When there is problem within the system, it may be easy to trace the cause of the problem without having to use a top down approach for the whole data warehouse. In most companies, one database manager or administrator is usually assigned to one data repository to ensure data reliability for the whole system.

Central Identities Data Repository (CIDR) is a government agency in India that stores andmanages data for the country's Aadhaar project. CIDR, which is regulated by the Unique Identification Authority of India (UIDAI), is responsible for verifying the authenticity of documents submitted by an individual and that an applicant is actually the person he or she claims to be.CIDR is tasked with ensuring that the information contained in the Aadhaar cards is unique to each individual and kept updated and relevant. While certain types of information such as birth date and gender will remain unchanged, other demographic details may undergo changes with time. The agency monitors these changes periodically with the assistance of a network of registrars who oversee the initial enrollment process for the issuance of UIDs and subsequent change requests.

Data warehouses have had staying power because the concept of a central data repository—fed by dozens or hundreds of databases, applications, and other source systems—continues to be the best, most efficient way for companies to get an enterprise-wide view of their customers, supply chains, sales, and operations.

For this reason, businesses that have data warehouses are upgrading and augmenting them with technologies such as Hadoop and in-memory processing, which help with “big data” workloads that are 10 times or 100 times or 1,000 times bigger than before. Meanwhile, businesses that have relied on piecemeal data-analysis solutions in the past are now establishing data warehouses to get a more complete picture of the enterprise. For more on that, see my recent article, “Healthcare’s Next Innovation? The Answer Is In The Data.”

The losses are staggering, and only represent incidents currently known to the U.K. government. Although there are a number of culprits contributing to the number of data breaches, I believe the primary cause may be that enterprises store personal information in many databases, resulting in the need to continually transfer information between departments and companies. From a conceptual perspective, if a single database existed with all personal and company information, then far less personal data would need to be stored on computers or passed between departments and organisations; only public references to individuals or companies involved would need to be shared.

However, in such a scenario it would be crucial to ensure that access to this central database is strictly controlled toprevent data loss.Building a central repository using digital signature technology. The first step in enabling such a central data repository is to create a network of distributed databases that each contain the virtual representations of each entity (i.e. individual, company, department, social group, etc.) and are associated with a particular country or region. Each virtual representation will have a unique identifier (similar to a URL for a website) and the following structure:

Each virtual representation will be associated with information, ranging from simple data such as name, address, date of birth, etc., to more complex structured documents, such as medical or employment records. Each piece of information is protected by a set of access control rules. The rules will determine which privileges, such as the ability to read, update or delete information, are granted to an authenticated third party attempting to gain access.

Similarly, relationships between one virtual representation and another (e.g. individual X "works for" company Y) are protected by rules that govern who can read, update and delete them.

Third, 'create' rules are used to automatically determine if authenticated users can associate new information or relationships and accompanying access control rules with the virtual representation. If a suitable 'create' rule is not found that can automatically approve the association of the information/relationship, then the request could be submitted to a manual approval process, where the entity that owns the virtual representation will be informed that another user wishes to associate new information or a new relationship with his or her virtual representation, allowing the owner to approve or deny the request.

Each entity in the database would have a digital signature as a means to authenticate itself when accessing the central repository or any system. The use of digital signatures, for instance, would ensure that customers or end users could create, access and update their information securely and govern subsequent access by others to such data or relationships.Having the ability for one entity to associate information with another entity, and specify rules that govern subsequent access to that information, provides a replacement mechanism for the multitude of existing proprietary databases.

Let's consider two more specific, illustrative examples:

1) For an organization with a website, rather than having a local database recording users' profile data, it would associate any additional 'website-specific' information with the user's virtual representation in the central data repository. For example, when the user accesses the Acme Corp. website, its digital signature will identify the user without him or her having to log into the site specifically. Acme's system would then access the central repository to retrieve the information associated with the user's virtual representation. The information will be protected by access control rules that ensure it can only be read by Acme and updated by the user.

2) Medical information can be associated with an individual's virtual representation. Unlike the website example, the access control rules associated with medical information would prevent the individual from reading, updating or deleting his or her own records. (Although information is associated with an individual's virtual representation, it does not necessarily imply that he or she has the right to view or change the information.)

Considering access rules based on digital signatures the problem with using the 'pure' digital signature approach is that the access control mechanism, which governs access to individuals' protected information in the central data repository, can only be based on information contained in the digital signature of the requester, such as his or her identity.

This approach is sufficient if the potential group of entities accessing information is small, as in the website example above. In this example, only the user and Acme will have the right to access information, and therefore the access control rules have to authenticate both parties.However, for the medical information example, the rules governing who may access and update information are more complicated. It is not a case of identifying the requester based on identity. Instead, it may be necessary to distinguish whether the requester is a doctor or medical assistant. This type of access control rule could not be implemented based on digital signature identity.

Another problem with using digital signatures as the source of information in access control rules isthat the information is static; in other words, it only gets updated when the certificate is renewed. Access control rules may need to be based upon the most up-to-date information possible, e.g. whether a doctor currently has a license to practice.So it's clear that digital signatures alone would not be sufficient to facilitate a secure central repository for sensitive data. The approach is not scalable to enable access by wide-ranging groups or entities and certainly not in situations where access needs to be based on dynamic information about an entity, as opposed to someone's actual identity. However, such a secure central repository is possible with in-depth access control rules. In my next tip, I will discuss the creation and implementation of these rules for virtual representations.

CONCLUSION

Data storage centralization, centralized management, and consolidation are challenges that all companies face in order to store and share data. Data storage centralization provides safety and gives users the ability to stream and download files provided by network users. The concept of centralized data management opens a variety of possibilities which are essential for any company in today's competitive and demanding market. What does this mean for the storage industry and for the evolution of storage software? Primarily, an expanded range of possibilities is changing the standard from an ordinary file server to a complex and multifunctional device.

REFERENCES

Books

Data Warehousing Fundamentals: A Comprehensive
By Paulraj Ponniah (Author)

Introduction to Building the Data
Warehouse Paperback – 2005
By I.B.M. (Author)

Website

http://www.learn.geekinterview.com/data-warehouse/dw-basics/what-is-data-repository.html

http://encyclopedia2.thefreedictionary.com/Data+repository

http://www.forbes.com/sites/oracle/2014/03/10/the-top-10-trends-in-data-warehousing/

http://searchbusinessintelligence.techtarget.in/definition/Central-Identities-Data-Repository-CIDR

http://en.wikipedia.org/wiki/Data_warehouse

http://en.wikipedia.org/wiki/Data_warehouse

http://www.webopedia.com/TERM/R/repository.html