Data Integration

Data integration involves combining data from several disparate sources, which are stored using various technologies and provide a unified view of the data. Data integration becomes increasingly important in cases of merging systems of two companies or consolidating applications within one company to provide a unified view of the company's data assets. The later initiative is often called a data warehouse.

Probably the most well known implementation of data integration is building an enterprise's data warehouse. The benefit of a data warehouse enables a business to perform analyses based on the data in the data warehouse. This would not be possible to do on the data available only in the source system. The reason is that the source systems may not contain corresponding data, even though the data are identically named, they may refer to different entities.

Data Integration Areas

Data integration is a term covering several distinct sub-areas such as:

- Data warehousing
- Data migration
- Enterprise application/information integration
- Master data management

This article concentrates on the process of data integration. More detailed information about the above areas can be found in related articles.

Challenges of Data Integration

At first glance, the biggest challenge is the technical implementation of integrating data from disparate often incompatible sources. However, a much bigger challenge lies in the entirety of data integration. It has to include the following phases:

Design

- The data integration initiative within a company must be an initiative of business, not IT. There should be a champion who understands the data assets of the enterprise and will be able to lead the discussion about the long-term data integration initiative in order to make it consistent, successful and benefitial.
- Analysis of the requirements (BRS), i.e. why is the data integration being done, what are the objectives and deliverables. From what systems will the data be sourced? Is all the data available to fulfill the requirements? What are the business rules? What is the support model and SLA?
- Analysis of the source systems, i.e. what are the options of extracting the data from the systems (update notification, incremental extracts, full extracts), what is the required/available frequency of the extracts? What is the quality of the data? Are the required data fields populated properly and consistently? Is the documentation available? What are

the data volumes being processed? Who is the system owner?
- Any other non-functional requirements such as data processing window, system response time, estimated number of (concurrent) users, data security policy, backup policy.
- What is the support model for the new system? What are the SLA requirements?
- And last but not least, who will be the owner of the system and what is the funding of the maintenance and upgrade expenses?
- The results of the above steps need to be documented in form of SRS document, confirmed and signed-off by all parties which will be participating in the data integration project.

Implementation

Based on the BRS and SRS, a feasibility study should be performed to select the tools to implement the data integration system. Small companies and enterprises which are starting with data warehousing are faced with making a decision about the set of tools they will need to implement the solution. The larger enterprise or the enterprises which already have started other projects of data integration are in an easier position as they already have experience and can extend the existing system and exploit the existing knowledge to implement the system more effectively. There are cases, however, when using a new, better suited platform or technology makes a system more effective compared to staying with existing company standards. For example, finding a more suitable tool which provides better scaling for future growth/expansion, a solution that lowers the implementation/support cost, lowering the license costs, migrating the system to a new/modern platform, etc.

Testing

Along with the implementation, the proper testing is a must to ensure that the unified data are correct, complete and up-to-date.

Both technical IT and business needs to participate in the testing to ensure that the results are as expected/required. Therefore, the testing should incorporate at least Performance Stress test (PST), Technical Acceptance Testing (TAT) and User Acceptance Testing (UAT) PST, TAT (Technical Acceptance Testing), UAT (User Acceptance Testing).

Data Integration Techniques

There are several organizational levels on which the integration can be performed. As we go down the level of automated integration increases.

Manual Integration or Common User Interface - users operate with all the relevant information accessing all the source systems or web page interface. No unified view of the data exists.

Application Based Integration - requires the particular applications to implement all the integration efforts. This approach is manageable only in case of very limited number of applications.

Middleware Data Integration - transfers the integration logic from particular applications to a new middleware layer. Although the integration logic is not implemented in the applications anymore, there is still a need for the applications to partially participate in the data integration.

Uniform Data Access or Virtual Integration - leaves data in the source systems and defines a set of

views to provide and access the unified view to the customer across whole enterprise. For example, when a user accesses the customer information, the particular details of the customer are transparently acquired from the respective system. The main benefits of the virtual integration are nearly zero latency of the data updates propagation from the source system to the consolidated view, no need for separate store for the consolidated data. However, the drawbacks include limited possibility of data's history and version management, limitation to apply the method only to 'similar' data sources (e.g. same type of database) and the fact that the access to the user data generates extra load on the source systems which may not have been designed to accommodate.

Common Data Storage or Physical Data Integration - usually means creating a new system which keeps a copy of the data from the source systems to store and manage it independently of the original system. The most well know example of this approach is called Data Warehouse (DW). The benefits comprise data version management, combining data from very different sources (mainframes, databases, flat files, etc.). The physical integration, however, requires a separate system to handle the vast volumes of data.