DATA STAGE ACCELERATOR

Mrs. Dhanamma Jagli
Assistant Professor,
Department of MCA,V.E.S. Institute Of Technology,
University of Mumbai, India.
dhana1210@yahoo.com

Mrs. Sangita Oswal
Assistant Professor,
Department of MCA,V.E.S. Institute Of Technology,
University of Mumbai, India.
dhana1210@yahoo.com

ABSTRACT
Data Stage is an application which connects to data sources and targets and processes the data as they move through the application. In DataStage there are many jobs which are similar in nature only the source and the target requirements have to be changed every time one has to repetitively make the jobs, change the settings and then run it.In this paper we propose a data stage tools where developer can even develop these ETL jobs at home and come next day and execute them in batch which save the development time. It only requires the user to provide the key requirements like server details, login credentials and schema definition and then generate the required mapping and execute the job and save design time.
.

KEYWORDS
Data mining, Clustering Algorithms

INTRODUCTION

LITERATURE REVIEW

1) *Introduction to* Data Warehousing

A data warehouse is a place where data is stored for archival, analysis and security purposes. Usually a data warehouse is either a single computer or many computers (servers) tied together to create one giant computer system. Data can consist of raw data or formatted data. It can be on various types of topics including organization's sales, salaries, operational data, summaries of data including reports, copies of data, human resource data, inventory data, external data to provide simulations and analysis, etc.

Warehouse ETL (Extraction, Transformation and Loading of data) is an essential part of data warehousing where the data warehousing professional populate data warehouse with information from production databases. Data warehousing professionals work with business analysts and make changes to warehouse ETL in order to maintain consistent and accurate reporting on warehouse table structures. Besides being a store house for large amount of data, they must possess systems in place that make it easy to access the data and use it in day to day operations. A data warehouse is sometimes said to be a major role player in a decision support system (DSS). DSS is a technique used by organizations to come up with facts, trends or relationships that can help them make effective decisions or create effective strategies to accomplish their organizational goals. Data warehousing is combining data from multiple and usually varied sources into one comprehensive and easily manipulated database. Common accessing systems of data warehousing include queries, analysis and reporting. Because data warehousing creates one database in the end, the number of sources can be anything we want it to be, provided that the system can handle the volume, of course. The final result, however, is homogeneous data, which can be more easily manipulated. Data warehousing is commonly used by companies to analyze trends over time. In other words, companies may very well use data warehousing to view day-to-day operations, but its primary function is facilitating strategic planning resulting from long-term data overviews. From such overviews, business models, forecasts, and other reports and projections can be made. Routinely, because the data stored in data warehouses is intended to provide more overview-like reporting, the data is read-only. If we want to update the data stored via data warehousing, we'll need to build a new query when we're done. This is not to say that data warehousing involves data that is never updated. On the contrary, the data stored in data warehouses is updated all the time. It's the reporting and the analysis that take more of a long-term view. Data warehousing is not the be-all and end-all for storing all of a company's data. Rather, data warehousing is used to house the necessary data for specific analysis. More comprehensive data storage requires different capacities that are more static and less easily manipulated than those used for data warehousing. Data warehousing is typically used by larger companies analyzing larger sets of data for enterprise purposes. Smaller companies wishing to analyze just one subject, for example, usually access data marts, which are much more specific and targeted in their storage and reporting. Data warehousing often includes smaller amounts of data grouped into data marts. In this way, a larger company might have at its disposal both data warehousing and data marts, allowing users to choose the source and functionality depending on current needs.

1) Types of Data Warehouses

With improvements in technology, as well as innovations in using data warehousing techniques, data warehouses have changed from Offline Operational Databases to include an Online Integrated data warehouse. Offline Operational Data Warehouses are data warehouses where data is usually copied and pasted from real time data networks into an offline system where it can be used. It is usually the simplest and less technical type of data warehouse. Offline Data Warehouses are data warehouses that are updated frequently, daily, weekly or monthly and that data is then stored in an integrated structure, where others can access it and perform reporting. Real Time Data Warehouses are data warehouses where it is updated each moment with the influx of new data. For instance, a Real Time Data Warehouse might incorporate data from a Point of Sales system and is updated with each sale that is made. Integrated Data Warehouses are data warehouses that can be used for other systems to access them for operational systems. Some Integrated Data Warehouses are used by other data warehouses, allowing them to access them to process reports, as well as look up current data.

2) Advantages & Disadvantages

The number one reason why you should implement a data warehouse is so that employees or end users can access the data warehouse and use the data for reports, analysis and decision making. Using the data in a warehouse can help you locate trends, focus on relationships and help you understand more about the environment that your business operates in. Data warehouses also increase the consistency of the data and allow it to be checked over and over to determine how relevant it is. Because most data warehouses are integrated, you can pull data from many different areas of your business, for instance human resources, finance, IT, accounting, etc.While there are plenty of reasons why you should have a data warehouse, it should be noted that there are a few negatives of having a data warehouse including the fact that it is time consuming to create and to keep operating. You might also have a problem with current systems being incompatible with your data. It is also important to consider future equipment and software upgrades; these may also need to be compatible with you data. Finally, security might be a huge concern, especially if your data is accessible over an open network such as the internet. You do not want your data to be viewed by your competitor or worse hacked and destroyed.

2. OVERVIEW OF THE SYSTEM

Many tools for data warehousing ETL process are used in the past decade. One such tool is the Ascentail DataStage. In DataStage there are many jobs which are similar in nature.At least 10-15% of the jobs which are performed frequently are somewhat same. So there are repeated jobs which are executed and the time and effort taken to develop these jobs is also increased. In DataStage some of the jobs are same, that is, the source and the target requirements have to be changed every time and this is very tedious. In short, one has to repetitively make the jobs, change the settings and then run it. The process is very time consuming since when many people are trying to access the same DataStage server to execute the jobs. So they have to access the same ETL environment which takes a lot of time increasing the waiting time for the job. The ETL developers can even develop these ETL jobs at home and come next day and execute them in batch. This saves almost 80% of the development time at the office, which can be utilized to effectively plan for the ETL process. DataStage Accelerator only requires the user to provide the key requirements like server details, login credentials and schema definition. DataStage Accelerator will then generate the required mapping and execute the job. This again saves 85% of design time.Datastage Accelerator reduces the complexities associated with the ETL process using Data Stage.

3) NEED FOR AN AUTOMATED SYSTEM

In the present time, many tools for data warehousing ETL process are used. One such tool is the Ascent ail DataStage. In DataStage there are many jobs which are similar in nature. At least 10-15% of the jobs which are performed frequently are somewhat same. So there are repeated jobs which are executed and the time and effort taken to develop these jobs is also increased. In DataStage some of the jobs are same, that is, the source and the target requirements have to be changed every time and this is very tedious. In short, one has to

repetitively make the jobs, change the settings and then run it. The process is very time consuming since when many people are trying to access the same DataStage server to execute the jobs. So they have to access the same ETL environment which takes a lot of time increasing the waiting time for the job. The ETL developers can even develop these ETL jobs at home and come next day and execute them in batch. This saves almost 80% of the development time at the office, which can be utilized to effectively plan for the ETL process.DataStage Accelerator only requires the user to provide the key requirements like server details, login credentials and schema definition. DataStage Accelerator will then generate the required mapping and execute the job. This again saves 85% of design time.Datastage Accelerator reduces the complexities associated with the ETL process using Data Stage.

PROPOSED SYSTEM

Datastage

DataStage is a ETL tool set for designing, developing, and running applications that populate one or more tables in a data warehouse or data mart. DataStage integrates data across multiple and high volume of data source and target applications. It integrates data on demand with a high performance parallel framework, extended metadata management, and enterprise connectivity. - Supports the collection, integration and transformation of large volumes of data, with data structures ranging from simple to highly complex.

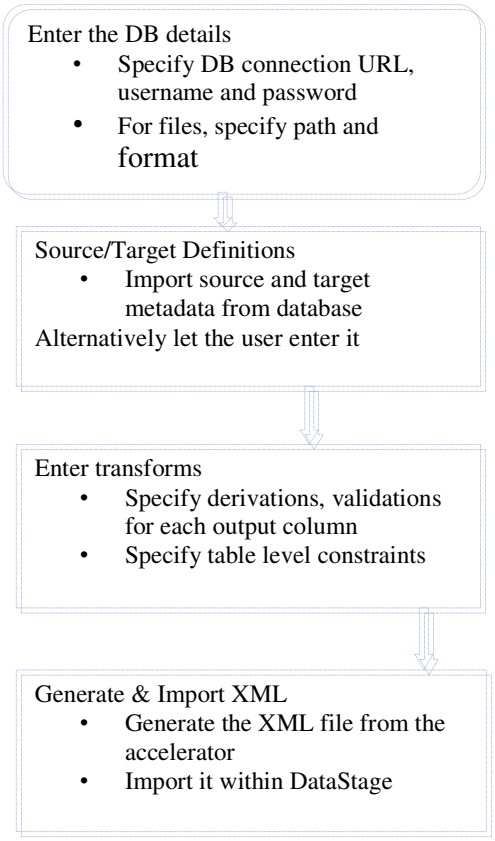
- Offers scalable platform that enables companies to solve large-scale business problems through high-performance processing of massive data volumes - Supports real-time data integration

- Enables developers to maximize speed, flexibility and effectiveness in building, deploying, updating and managing their data integration infrastructure - Completes connectivity between any data source and any applications Client Components DataStage has four client components which are installed on any PC running Windows 95, Windows 2000, or Windows NT 4.0 with Service Pack 4 or later:

1) DataStage Designer: A design interface used to create DataStage applications (known as jobs). Each job specifies the data sources, the Transforms required, and the destination of the data. Jobs are compiled to create executables that are scheduled by the Director and run by the Server

2) DataStage Director: A user interface used to validate, schedule, run, and monitor DataStage jobs

3) DataStage Manager: A user interface used to view and edit the contents of the Repository



Workflow

There are three server components which are installed on a server:

1) Repository: A central store that contains all the information required to build a data mart or data warehouse

2) DataStage Server: Runs executable jobs that extract, transform, and load data into a data warehouse.

3) DataStage Package Installer: A user interface used to install packaged DataStage jobs and plug-ins

There are three basic types of DataStage job:

1) Server jobs: These are compiled and run on the DataStage server. A
Server job will connect to databases on other machines as necessary, extract data, process it, and then write the data to the target data warehouse
2) Parallel jobs: These are available only if you have Enterprise Edition installed. Parallel jobs are compiled and run on a DataStage UNIX
Server, and can be run in parallel on SMP, MPP, and cluster systems
3) Mainframe jobs: These are available only if you have Enterprise MVS Edition installed. A mainframe job is compiled and run on the mainframe. Data extracted by such jobs is then loaded into the data warehouse

CONCLUSION

The proposed Data stage is an effort to simplify the work of an ETL Developer. It allows developers to focus on data rather than focusing on how to access and modify it. It allows a developer to build fast and accurate mappings which are essential to increase the speed of any Data warehousing project. It works with or without Data Stage Server to create mappings without the need of a client. It cuts the development time and accelerates the developments of mappings in Data Stage.

ACKNOWLEDGEMENTS

The authors would like to thank everyone, just everyone!

REFERENCES

[1] **Data loading and mapping using staging DBMS in the grid-Ejaz Ahmed**

[2] ETL Process Modeling Conceptual for Data Warehouses: A Systematic Mapping Study- Mufioz, L.

[3] Study of localized data cleansing process for ETL performance improvement in independent datamart- Savitri, F.N.