ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Using Optimum ETL Tool to Handle Different Form of Data

Sandeep Kumar
M.Tech (CSE),
Dept. of Computer Science & Engineering,
Greater Noida Institute of Technology,
Greater Noida, India.

Dr. Yatin Agarwal
Associate Professor,
Dept. of Computer Science & Engineering,
Greater Noida Institute of Technology,
Greater Noida, India.

Abstract:

Suppose data are in different formats, is nothing but in different source systems and willing to read those data from different source systems and trying to load those data into the target systems. Here sources may be like- Flat File, XML or Database etc. and target may be Flat File, XML, Database or Data ware House etc. So to handle different form of data, in the market some tools are available. These tools are called ETL TOOLS. ETL (Extraction–transformation–loading) tools are pieces of software responsible for the extraction of data from several source systems, its cleansing, customization, reformatting, integration, and insertion into target systems or into a data warehouse. So there are many types of commercial ETL tools are available in the market that are used for data integration, transformation and loading into the target systems.

However the issue is all the ETL tools do not support all types of source systems data. So in this paper we are going to represent an optimum ETL tool. Informatica PowerCenter is a optimum and premium data integration solution available today. The reason it provides the best solution in large enterprises is because it is: database neutral and hence can communicate with any database. Informatica provides the market's leading data integration platform. It is robust, and can be used in both windows and UNIX based systems. It is high performing yet very simple for developing, maintaining and administering. It provides the best solution in large enterprises due to below reasons:

- ***It is database neutral and hence can communicate with any database***
- ***The most powerful data transformations tool. It converts one application's data to another's format***

Before loading the data into target systems some operations (cleansing, customization, reformatting or integration) are performed by Informatica Powercenter ETL tool on source data. In Informatica, transformation plays a vital role.

Transformations are nothing but objects which is used to creates, modified or passes data to the target systems. The main aim of transformation is to modify the source data as per business requirement or target system. So Informatica provides various transformations to perform specific functionalities.

INTRODUCTION OF INFORMATICA POWERCENTER

Informatica PowerCenter is a data integration ETL (Extract, transform and load) tool means it is used to read the data from various source systems and loading into target system. It is the most powerful and flagship product in the in the market now a days from Informatica Corporation which is used to extract data from multiple source systems and load it into the various target systems after applying business logic

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

and it also helps us to load data into a federal place such as a Data warehouse or an Operational Data Source. Informatica PowerCenter provides data integration software and services for various businesses, industries and government organizations, telecommunication, financial and insurance services. In another word we can say that ETL is tool which is responsible for the extraction of data from numerous sources and loading into target. Here sources may be like- Flat File, XML or Database etc. and target may be Flat File, XML, Database or Data ware House etc. Now before loading the data into target some operations (cleansing, customization, reformatting or integration) are performed by ETL tool on source data.

In informatica, transformation plays a vital role. Transformations are nothing but objects which is used to creates, modified or passes data to the target system. The main aim of transformation is to modify the source data as per business requirement or target system. So Informatica provides various transformations to perform specific functionalities. [1], [2], [7]

WHY INFORMATICA POSITIONED AS A LEADER IN 2017

Due to below points we will come know why Informatica is so popular in the market:

- 2017 Revenue : $1.06 billion, more than the combined revenue of Abinitio, datastage, SSIS, and other ETL tools
- 7-year Annual CAGR: 30%
- Partners : 450+
- Customers: Over 5,000
- Customers in 82 countries & direct Presence in 28 countries
- It is robust, and can be used in both windows and UNIX based systems
- It is high performing yet very simple for developing, maintaining and administering

With the help of above points we can conclude that there is great demand of Informatica powercenter across the globe. Here one thing came into the picture

WHY ETL

Now-a-days almost every company processes a vast set of data. The data may come from different sources systems and needs to be processed and ensure that to give insightful information for making business decisions. However the data which comes from different source systems has following challenges:

Challenges-Suppose there is large companies which is having huge data and such huge data may be in any format either available in multiple databases or in many unstructured files or in both. Requirement is like this we need to gather, combined, compared, and made to work as a seamless whole. However here the main issue is different databases don't communicate well.

To overcome above problem many organizations have implemented interfaces between these databases. So every pair of databases requires a unique interface. Suppose we change one database, many interfaces may have to be upgraded.

Hence data Integration technologies comes into picture to provide a solution of such type of problems and allow data from different databases and formats to communicate with each other. So there are different data integration technology is nothing but ETL tools are in the market. However Informatica powercenter uses the Extract, Transform & Load (ETL) is the most popular ETL tool to perform data integration. So we will discuss about ETL and how Informatica performs ETL to solve such business problems.

WHAT IS ETL

ETL is nothing but a data integration tool which is used to extracts, transforms, and then loads data into target systems.

FUNCTIONALITY OF ETL TOOLS

(i) Detection of appropriate information at the source side.
(ii) Extraction of this information.
(iii) Transportation of this information to the Data Staging Area (DSA).

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

(iv) Transformation (i.e. customization and integration) of the information coming from multiple sources into a common format.
(v) Loading the data to the target systems

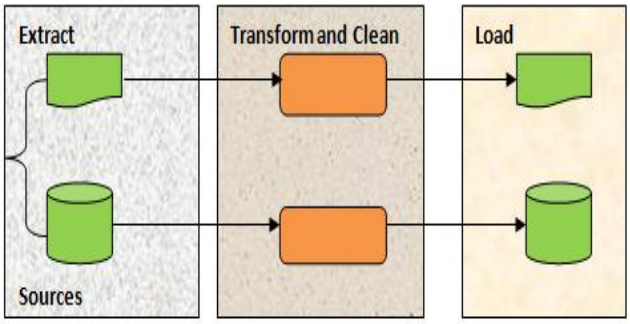

Fig: Shows Extraction-Transformation-Loading processes

There are three parts in the above figure. The first part has been extracted the data from source systems are files and relational databases. Once data have been extracted from the source systems then these data are transmitted to the data staging area (DSA). Staging area is a temporary storage area between the source of information and target systems.

Staging area is used to handle invalid data, replace missing value, eliminate duplicates and apply transformation as per the business requirement. This is the second part. Now coming to the third part i.e. target systems are files and relational databases or DWH. In DWH there are mainly two types of table fact and dimension tables. Fact tables stores the metrics or measurement of business process whereas dimension table stores textual or descriptive information about a business process [3], [4], [5], [8]

OBJECTIVE OF THESIS

Suppose we have data at source side in different format i.e. in different source systems and we are willing to transfer that data into target system so to do this work some tools are available in the market that tool is called ETL tool. ETL means Extract, Transfer and load the data in target system. In another word we can say that ETL tool is used to reading the data from different source system and loading into target system.

Now here the main goal of this thesis is to find out the optimum ETL tool to does this work nothing but reading the data from different source system and loading the data into target systems. Now a days in the market so many ETL tools are available however on the basis of study came to know **Informatica Powercenter** is the best and optimum ETL tool in the market. Informatica Powercenter is the most powerful ET tool and it plays a vital role to load data into target systems or to create a data ware house for business analysis. A Data Ware house stores historical data.

HOW INFORMATICA IS OPTIMUM ETL TOOL

As we know that there are many types of commercial ETL tools are available in the market that are used for data integration, transformation and loading into the target systems. However in this paper we are going to represent an optimum ETL tool. However Informatica powercenter is optimum ETL tool with respect to other ETL tools due to many reasons.

Informatica PowerCenter is a premium data integration solution available today. The reason it provides the **best** solution in large enterprises is because it is: database neutral and hence can communicate with any database. It is the most powerful data transformations **tool**. Informatica provides the market's leading data integration platform. Give customers what they need: a data integration solution that can integrate more data, from more systems, in less time than any other. Tested on nearly 500,000 combinations of platforms and applications, our data integration platform interoperates with the broadest possible range of disparate standards, systems, and applications. This unbiased and universal view makes us unique in today's market. It also makes us the ideal strategic partner for companies looking to solve data integration issues of any size. Because data integration is sole focus, Tool out-performs and out-innovate offerings from other providers—especially in cases where integration is only one of several lines of business. Informatica was founded on innovation. Data integration platform was architected from the ground

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

up to meet the most challenging data integration requirements. From the start, they approached data integration differently. They separated the logic from the execution environment. To some, that may sound like a deep-dive technical detail, but it is the architectural approach that has made all the difference in how our platform performs, and how rapidly is able to respond to customers' evolving requirements. They have continued to build upon this innovation and are known, and respected, for consistent track record of delivering solid technology though regular releases.

Performance:- It's consistently at the top of the list when it comes to evaluating data integration technologies. Improving performance is also consistently at the top of the list for our customers. Today, Informatica provides the most high-performance, scalable, adaptive solution in the industry—with support for any latency, whether batch or real time. With key capabilities like adaptive load balancing, parallel processing, interference handling, memory management, 64-bit processing support, and a variety of methods available to optimize performance for your infrastructure (e.g. Dynamic Partitioning, Workflow on Grid, Session on Grid, Pushdown Optimization) Informatica delivers the superior performance you can count on to support your most demanding data integration needs.

Near 100% -- "Go Live" success rate
95% -- Rate of renewal, significantly higher than the industry average of 86%

It support all types of source systems and used maximum now a day's in the market as per business requirement and this tool is very appropriate to fill the business logic. Informatica powercenter tool plays a vital role to extract data from various sources and loading it into target systems. It is a very popular and effective ETL tool with its high availability as well as being fully scalable and high-performing. Powercenter provides the ground work for all major data integration projects throughout the enterprise. Powercenter provides reliable solutions to the IT management, global IT teams, developers and business analysts as it delivers not only data that can be trusted and guarantees to meet analytical and operational requirements of the business, but also offers support to various data integration projects and collaboration between the business and IT across the globe. Informatica PowerCenter enables access to almost any data sources from one platform. Informatica PowerCenter is capable of managing the broadest range of data integration initiatives as a single platform. Informatica offers a rich set of features like operations at row level on data, integration of data from multiple structured, semi-structured or unstructured systems, scheduling of data operation. It also has the feature of metadata, so the information about the process and data operations are also preserved. This ETL tool makes it possible to simplify the development of data warehouses and data marts.

COMPARISON OF DIFFERENT TYPES OF ETL TOOL

1. INFORMATICAPOWERCENTER:-Informatica PowerCenter is a very popular data integration ETL tool with its high availability as well as being fully scalable and high-performing. PowerCenter provides the groundwork for all major data integration projects and initiatives throughout the enterprise. [5], [6]

Basically Informatica PowerCenter consists of 3 main components.

a) Informatica PowerCenter Client Tools:-These are the development tools which are used to create workspace/Folder or to change user's password etc.

b) Informatica PowerCenter Repository:-It is a relational database that is used to store the metadata like Source definition, Target definition, Transformation used

c) Informatica PowerCenter Server:-It is responsible for execution of the data and loading of these data into target system.

Advantages:

- frozen technology

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

- Easy to learn
- Consistent to track the record
- ability to address real-time data integration schemes
- Informatica is highly specialized in ETL and Data Integration.
- focus on B2B data exchange
- It is robust, and can be used in both windows and UNIX based system
- It is high performing yet very simple for developing, maintaining and administering

Disadvantage:

- Limited experience in the field.

2. ORACLE WAREHOUSE BUILDER (OWB):- Oracle Warehouse Builder (OWB) is an Oracle's ETL tool that allow graphical environment to build, manage and maintain data integration processes in a custom Business Intelligence application. Oracle Warehouse Builder allows creation of dimensional, relational and metadata models, and also star schema data warehouse architectures. Oracle Warehouse Builder supports Oracle Database and flat files as a target system.

Advantage:

- Tight connection to all Oracle data warehousing applications.
- Tendency to integrate all tools into one application and one environment.

Disadvantage:

- customers are mostly confused in the functionality and the future is uncertain
- Focus on ETL solutions only.
- Tools are used mostly for batch-oriented work, transformation rather than real-time processes or federation data delivery.

3. SAS ETL:-It is an ETL tool which combines data warehousing and intelligence applications for traditional business. The SAS Enterprise ETL Server also provides the concept of multithreaded and multiprocessing data extraction to speed up the data transfer and related operations. The SAS platform is also able to reduce duplicate or inaccurate data with the help of drag and drop interface, not required of programming or SQL (Structured Query Language) for managing data.

Advantage:

- Very flexible to collect data through number of sources.
- Can work on many operating systems.
- experienced company, great support and most of all very powerful data integration tool with lots of multi-management features
- great support for the business-class companies as well for those medium and minor ones

Disadvantage:

- Misplaced sales force, company is not well-known.
- High cost

4 IBM INFOSPHEREDATASTAGE:-Info Sphere Data Stage is a powerful data integration tool. It uses a graphical notation to construct data integration solutions. The IBM Info Sphere Data Stage is capable of integrating data on demand across multiple and high volumes of data sources and target applications using a high performance parallel framework. It provides powerful scalable platform for easy and flexible integration of all types of data, including big data at rest (Hardtop-based) or in motion (stream-based), on distributed and mainframe. It is available in various versions such as the Server Edition, the Enterprise Edition, and the MVS Edition. The Enterprise Edition introduces parallel processing architecture and parallel jobs. The Server Edition mainly representing the Server Jobs.

It has three levels of Parallelism which are:

- Pipeline Parallelism
- Data Parallelism
- Component Parallelism

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Advantage:

- Flexibility and strongest tool on the market.
- high level of satisfaction from clients and a variety of initiatives

Disadvantage:

- difficult to learn
- Long and time consuming implementation.
- became very heavy (lots of GBs) with version 8.x and requires a lot of processing power

5. SQL SERVER INTEGRATION SERVICES (SSIS):-SQL Server Integration Services (SSIS) is a tool that is used to perform ETL operations; i.e. extract, transform and load data. It is used to perform a large volume of data migration, transformation and integration. The first version of SSIS was released with SQL Server 2005. SSIS is a replacement for Data Transformation Services (DTS) which was available with SQL Server 7.0 and SQL Server 2000.

With help of this tool we can move data from source system to target system very fast. However it can retrieve data from any source and perform various transformations on the data; e.g. convert from one type to another, convert to uppercase or lowercase, perform calculations, etc. and loading data into target.

Advantage:

- standardized data integration
- simplicity speed of implementation
- Comparatively low cost, excellent support and distribution.

Disadvantage:

- Problems in non-Windows environments. Takes over all Microsoft Windows limitations.
- unclear vision and strategy

INFORMATICA POWERCENTER TRANSFORMATIONS

The name transformations itself describes that it is used to transfer the source data into target system according to the business requirements. In Informatica, transformations help us to transform the source data according to the requirements of target system and also ensuring the quality of the data being loaded into target. In another word we can say that Informatica transformations are the objects which creates, read, modify or pass data to the defined target structures like tables, files, or any other targets required. A transformation is mainly used to represent a set of rules, which define the data flow and how the data is loaded into the targets and transformation is also used to link the source and target objects.

Informatica PowerCenter provides multiple transformations, each transformation provides functionality. Informatica Transformations can be mainly divided into two categories. First is based on the change in the overall no of rows between the source and target and second is based on the connectivity (Linking in mapping) of the transformations with each other. [9] ,[10], [11], [12], [13], [14], [15]

TYPES OF TRANSFORMATIONS BASED ON THE CHANGE IN NO OF ROWS

- Active
- Passive

Active Transformation:

An active transformation can change the number of rows that pass through it from source to target i.e. it eliminates rows that do not meet the condition in transformation. For example:-Filter transformation, Aggregator, Filter, Joiner, Normalizer, etc.

Passive Transformation:

A passive transformation does not change the number of rows that pass through it. In passive transformation the number of input and output remain same. For example:-Expression Transformation

TYPES OF TRANSFORMATIONS BASED ON CONNECTIVITY

- Connected

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

- Unconnected

Connected Transformation:

If the transformation is connected to other transformations or directly to target table in the mapping, is called connected transformation. The connected transformations are elected when for every input row, transformation is called or expected to return a value. For Example:-Source qualifier transformation can be connected to filter transformation to filter records.

Unconnected Transformation:

An unconnected transformation is not connected to other transformations in the mapping. It is called within another transformation, and returns a value to that transformation. For Example: -Lookup, and Stored Procedure

DIFFERENT TYPES OF TRANSFORMATIONS IN POWERCENTER

Transformation	Type	Description
Expression	Passive/Connected	Calculates a value.
Filter	Active/ Connected	removes rows that do not meet the filter condition
Source Qualifier	Active/ Connected	Use to filter, sort, join if source is table
Sorter	Active/ Connected	Used to sort the data based on sort key
Router	Active Connected	Routes data into multiple target based on multiple filter conditions.
Sequence Generator	Passive Connected	to generate auto sequence numbers
Aggregator	Active Connected	Performs aggregate calculations.
Joiner	Active Connected	Joins data from different databases or flat file systems.
Normalizer	Active Connected	Used in the pipeline to normalize data from relational or flat file sources.
Look Up	Active Connected or Passive Connected or Active Unconnected or Passive Unconnected	Lookup and return data from a flat file, relational table, view, or synonym.
Rank	Active Connected	Limits records to a top or bottom range.
Union	Active Connected	Merges data from different databases or flat file systems.
SQL	Active Connected or Passive Connected	Executes SQL queries against a database.

Table. Short descriptions about transformations

REFERENCES:

[1] Himanshu Agrawal, Girish Chafle, Sunil Goyal, Sumit Mittal, and Sougata Mukherjea. An Enhanced Extract-Transform-Load System for Migrating Data in Telecom Billing. In Proceedings of the International Conference on Data Engineering (ICDE), Canc´un, M´exico, 2008.

[2] Alexander Albrecht and Felix Naumann. Managing ETL processes. In Proceedings of the VLDB International Workshop on New Trends in Information Integration (NTII), Auckland, NZ, 2008.

[3] P. Vassiliadis and A. Simitsis, “EXTRACTION, TRANSFORMATION, [4] ETL Tools information, http://etltools. info/en/bi/etl_process.htm accessed on September 12, 2014

[5] https://www.etltool.com/list-of-etl-tools/

[6] https://adeptia.com/products/etl_vendor_comparison.html

[7] "Informatica Tutorial"

[8] "Chapter 1: Service Oriented Architecture (SOA)". msdn.microsoft.com. Retrieved 2016-09-21.

[9] https://www.youtube.com/watch?v=yw8uJdUc42o

[10] https://www.youtube.com/watch?v=aPBaDpw1S2w

[11] https://www.youtube.com/watch?v=JGJlvbKnEpI

[12] https://www.youtube.com/watch?v=NVkKxJS7Y2g

[13] https://www.youtube.com/watch?v=22gNUDhPUTk

[14] https://www.youtube.com/watch?v=a4TvwOKSbWQ

[15]https://www.youtube.com/watch?v=4nNmXXGUc8E