RESEARCH ARTICLE OPEN ACCESS

A Survey of ETL Tools

Mr. Nilesh Mali[1], Mr.SachinBojewar[2]

[1](Department of Computer Engineering, University of Mumbai, ARMIET, Shahapur,Thane, India)
[2](Department of Information Technology, University of Mumbai, VIT, Wadala, India)

---*************************-----------------------------------

Abstract:

In most of the organizations valuable data is wasted because of its different formats in various resources. Data warehouses (DWs) are central repositories of integrated data from different disparate sources with an objective to support in decision making process. The Extraction-Transformation-Loading (ETL) processes are the key components of DWs, so the selection of ETL tool is complex and important issue in DWs. ETL is a process of extraction of data, their transformation to desire state by cleaning it, loading it to a target database. This paper first focus the ETL process briefly then discuses analysis of some of the ETL tools based on the generalized criteria for selection of better tool for the improvement of business growth.

Keywords — **Data warehouses, ETL Process, ETL tools, enterprise systems, Business Intelligence.**

---*************************-----------------------------------

I. INTRODUCTION

Data Warehouse (DW) defined by Inmon [1] as "collection of integrated, subject-oriented databases designated to support the decision making process" aims to improve decision process by supplying unique access to several sources. Data warehouses are central repositories of integrated current and historical data from one or more different disparate sources. It mainly contains historical data derived from transaction data and the current data from other sources. It usually characterized by collection of integrated; subject oriented, on-volatile and time variant databases. The heart of DWs is the Extraction-Transformation-Loading (ETL) process. ETL is a process which is used to extract data from various sources, transform that data to desired state by cleaning it and loading it to a target database. The result of this is used to create reports and analyze it. ETL consume up to 70% of resources [3], [5], [4], and [2]. Interestingly [2] reports and analyses a set of studies proving this fact. ETL is responsible to maintain accuracy and correctness of data.

This paper first focus the ETL process briefly then discuses analysis of some of the ETL tools based on the generalized criteria for selection of better tool.ETL tools are very important for evaluation of Business Intelligence. This includes your results are only as accurate as the input you feed it. Selection of right ETL tool is a fundamental step in achieving your strategic goals.

The research categories of the framework revealed by Nils Schmidt, Mario Rosa, Rick Garcia, Efrain Molina, Ricardo Reyna and John Gonzale[6]. The main purpose to develop a criteriaframework tocompare the ETL tools against each other.

II. ETL PROCESS

ETL (Extract, Transform and Load) is a process in data warehousing responsible for pulling data out of different source systems and placing it into a data warehouse.

Basically ETL involves the following tasks:

A. Extraction

Extracting the data from different source systems is converted into one consolidated data warehouse format which is ready for transformation purpose.

B. Transformation

Transforming the datamay involve the following tasks:

- Applying new business rules (so-called derivations, e.g., calculating new dimensions and measures),
- Cleaning the data (e.g., mapping NULL to 0 or "Male" to "M" and "Female" to "F" etc.),
- Filtering the data (e.g., selecting only specific columns to load),
- Splitting a column into multiple columns and merging multiple columns into a column,
- Joining together data from different sources (e.g., lookup, merge),
- Transposing rows into columns and vice versa.
- Applying simple or complex data verification and validation (e.g., if the first 4 columns in a row are empty then reject the row from processing) [7].

C. Loading

Loading the data into a data mart or data warehouse or data repository other reporting applications that houses data [7], [8].

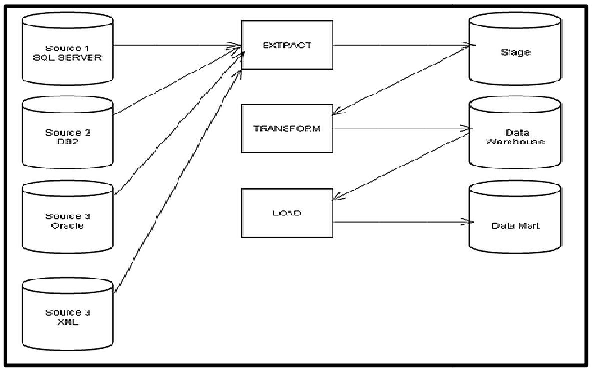

Fig. 1 ETL Workflow

III. ETL Tools

The ETL (Extract, transform, load) tools were used to simplify the data management by reducing the absorbed effort.These are designed to save time and money by eliminating the need of 'hand-coding' when a new data warehouse is developed [9]. Depending on the needs of customers there are many types of tools and you have to select appropriate one for you. Most of them are relatively quite expensive, some may be too complex to handle. The most important aspect to start with defining business requirements is selection of right ETL tool. The working of the ETL tools is based on ETL (Extract, transform, load) process.

A. ETL tools comparison criteria

- mode of connectivity or adapter support
- mode of data transformation and delivery support
- data modeling and metadata support
- architecture design, development and data governance support
- debugging facility and execution or runtime platform support
- additional services and requirements for vendors
- customers usability support
- cost of hardware or software , installation, OS, Support
- functionality, flexibility and performance support
- infrastructure support

B. Details of ETL Tools

Most popular used commercial and freeware (open-sources) ETL Tools are listed below.

Commercial ETL Tools

A variety of commercial ETL tools used for data integration, transformation, some of them are listed below

1) ***1) IBM InfosphereDataStage:*** IBMInfoSphereDataStage uses the features of high performance parallel framework and graphical notation to integrate data across multiple systems. It provides powerful scalable platform for easy and flexible integration of all types of data, including big data at rest (Hadoop-based) or in motion (stream-based), on distributed and mainframe platforms [11]. It manages workload and business rules by optimization of hardware. 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. The MVS Edition related with mainframe jobs.

2) InformaticaPowerCenter:InformaticaPowerCenter is an enterprise data integration platform working as a unit for B2B Data Exchange, Cloud Data Integration, data migration, Complex Event Processing, Data Masking, Data Quality, Data Replication and synchronization, Data Virtualization, Master Data Management, Ultra Messaging, etc.

Basically InformaticaPowerCenter consists of 3 main components.

- InformaticaPowerCenter Client Tools: These are the development tools, installed at developer end to enable the mapping process.
- InformaticaPowerCenter Repository: Repository stores all the metadata for your application so it is the heart of Informatica tools. It act as a data inventory.
- InformaticaPowerCenter Server: Server is responsible for execution of all the data and loading of these data into target system.

3) Oracle Warehouse Builder (OWB):Oracle Warehouse Builder (OWB) is an Oracle’s ETL tool that enables 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 [9].Oracle Warehouse Builder supports Oracle Database (releases 8i, 9i and newer) and flat files for target database. It provides data quality, data auditing and full lifecycle management of data and metadata of target database.

4) Oracle Data Integrator (ODI):Oracle Data Integrator (ODI) is an ETL based software application used for data transformation and merging or data integration from high-volume, high-performance load, to event-driven, to SOA-enabled data services processes by adding parallelism. The important architecture component of ODI is repository which is collection of all metadata and is accessed by client-server mode or thin client mode. Oracle Data Integrator works as well in the staging and transforming area as the support for other Oracle software[9].

5) SAS ETL Studio:This ETL tool is developed by **SAS Enterprise** (USA) which offers an integrated ETL platform. SAS is one of the market leaders which combine data warehousing and intelligence applications for traditional business process. It provides the facility of multithreaded and multiprocessing data extraction to speed up the data transfer and related operations. SAS helps to reduce duplicate or inaccurate data by providing drag and drop interface, not necessary of programming or SQL (Structured Query Language) for managing data [9]. SAS Data Integration Studio enables users to quickly build and edit data integration, to automatically capture and manage standardized metadata from any source, and to easily display, visualize, and understand enterprise metadata and your data integration processes [11],[12].

6) Business Objects Data Integrator (BODI): Business Objects Data Integrator is an ETL tool for data integration, mainly focus on data quality features. It also uses data repository for storing of created jobs and projects. BODI is commonly used for building data marts, data warehouses, etc.

It provides a lot of options in data manipulation likes

- Data unification: makes possible quick and trouble-free or easily updating and creating the universes objects.

- Data auditing: data integrity gets verified, especially when the data is read, processed and loaded successfully.
- Data cleansing: maintaining the quality and standard of data.

7)SQL Server Integration Services (SSIS):SQL Server Integration Services (SSIS) is a fast and flexible component of the Microsoft SQL Server database software that can be used to perform a wide range of data migration, data transformation and data integration tasks. It can be used to maintain and update multidimensional cube data of SQL Server database for solving complex business problems. SSIS was first released with Microsoft SQL Server in 2005, later replaced Data Transformation Services, which had been a feature of SQL Server from the Version 7.0. It is only available in the editions likes "Standard", "Enterprise" and "Business Intelligence". Using SSIS any type of data can be moves quickly from a variety of source types to a variety of destination types. SSIS includes a rich set of built-in transformations likes Aggregation, Audit, Cache Transform, Data Conversion, Data Mining Query, Dimension Processing, Export & Import Column, Fuzzy Grouping, pivot, row sampling, term extraction, etc. It also provides programmable object model that allows developers to create, store, and load packages for execution. Basically SSIS is used to handle large enterprises, as it requires a Microsoft Server operating system so it requires high operating system and support cost.

TABLE 1. SUMMARY OF THE SURVEY

Sr, No.	Tool	Advantage	Disadvantage
1	IBM InfosphereDataStage	• flexibility and strongest tool on the market, • provides high level of satisfaction for the clients	• difficult to learn • long and time consuming implementation • requires large amount of memory and processing power
2	InformaticaPowerCenter	• consistent to track the record, easy learning, ability to address real-time data integration schemes • focus on B2B data exchange	• diminishing the value of technologies diminished by several partnerships • In the field experience is limited.
3	Oracle Warehouse Builder (OWB)	• strong, powerful data integration • tight connectivity with respective application • all the tools are integrated in one application of one environment	• focus on ETL solutions only • mostly used for batch-oriented work, • customers are mostly confused
4	Oracle Data Integrator (ODI)	• strong connection to all Oracle data warehousing applications, • all the tools are integrated in one application of one environment	• focus on ETL solutions only • mostly used for batch-oriented work, • Using this future is uncertain
5	SAS ETL Studio	• experienced company, great support and most of all very powerful data integration tool with lots of multi-management features • can work on many operating systems and gather data through number of sources – very flexible	• misplaced sales force, not well recognized organization. • Future is Uncertain. • Cost is high

		• great support for the business-class companies as well for those medium and minor ones	
6	Business Objects Data Integrator (BODI)	• SAP integration •Better data modeling and data-management support; •provides SAP tool for data mining • Quick learning and ease to use	• different companies uses different SAP Business Objects. •uncertain future. • not supported as a stand-alone capable application of few organization.
7	Microsoft SQL Server Integration Services(SSIS)	• integrates data with standard • Ease speed of implementation details. • low cost, excellent support and distribution	• Window limitation, complexity increases • unclear strategy and vision

Freeware (open-sources) ETL Tools

A variety of open source ETL tools used for data integration, transformation, some of these listed below:

1) Pentaho Data Integration (Kettle):Kettle has been recently acquired by the Pentaho group so it is also called as Pentaho Data Integration. It is a leading open source ETL application or tool in the market. Basically it is composed of four elements, **ETTL** (**extraction** from different sources, **transport** of data, **transformation** of data, **loading** of data into data warehouse). Kettle can supports deployment on single node computer as well as on a cloud, or cluster. It can load and process big data sources in familiar ways [10] by providing flexibility and security. Kettle can support various data sources and databases like oracle, MySQL, AS/400, MS Access, MS SQL Server, Sybase, IBM DB2, dBase, Informix, Hypersonic, any database using ODBC on Windows, etc [8]. By purchasing the Enterprise Edition would provide the services like recovery and backup. Pentaho is easy to use as it being open source software allows administrators to utilize its new drag and drop environment.

The main components of Pentaho Data Integration are:

Spoon : Related with the transformation (reading, validating, refining and writing) of data.

Pan : Used to run data transformations designed in Spoon.
Chef : Used to automate the database update process
Kitchen :Used to help and to execute the jobs in a batch mode.
Carte : a web server which allows remote monitoring of the running Pentaho Data integration ETL processes through a web browser [8].

2) Talend Open Studio:Talend Open Studio is a Data Integration platform that enables designing of data integration processes and their monitoring and operates as a code generator, producing data-transformation scripts and underlying programs in Java. It consists of metadata repository which provides the data (definitions and configuration related data for each job) to all the components of Talend Open Studio. It is commonly used for data migration, synchronization or replication of databases. It also used to improve the quality of big data. It easy to operate does not require extra technical skills.

3) Clover ETL: Clover is a Commercial Open Source ETL tool considered for data transformation and integration, cleanse, and distribute data into applications, databases, and data warehouses. The Clover ETL tool is based on Java so it is independent and resource- efficient. It can be used as standalone as well as a command- line application or server application or just like Java library it can be even embedded in other applications [9].Clover ETL is operated on any type of operating system like windows, Linux, HP-UX, etc. It can be both used on low-cost PC as on high-end multi processors servers [9]. Graph or transformation dataflow are used to represent data flow in Clover ETL. Edges of graph can represent the data flow from one component to another.

Clover ETL components

- CloverETLEngine : the core for running data transformation graphs

- CloverETLDesigner: a commercial visual data integration tool used to design and execute transformation graphs

- Clover ETL Server :An enterprise automation and data integration monitoring platform having the features such as workflows, scheduling, monitoring, user management, or real-time ETL abilities.

4) Jasper ETL:Jasper ETL is one of the easiest open sources ETL tool considered for data integration, cleansing, transformation and movement on the market [13]. In this case it is said to reduce both the costs of ownership and the complexity of an IT infrastructure services. It consists of aggregation or integration of large data from different data sources. While transforming the data it maintains both consistency and accuracy of data. Finally loads the data into data warehouse in optimize way. This tool is easily affordable to anyone. It is easy to manage and proven superior performance than many commercial ETL tools. Is can be used to any type of business, it may be small or complex.

IV. CONCLUSIONS

ETL tools are designed and used to save time and cost when a new data mart or data warehouse is developed. ETL tools are basically at the foundation of Business Intelligence for migration or transformation of Data from one format to another or Data mobility then you might have to employ these ETL tools to enable your business process. They help to extract the data from different heterogeneous database, to transform the data into a unified standard format by cleansing and applying various processes and finally load it into data mart or data warehouse. From our survey we have studied different commercial ETL software tools; we find that Microsoft SQL Server Integration Services (SSIS) are mostly satisfied the needs of large organizations, as it can handle the large database. In case of freeware or open sources ETL tools, Pentaho Data Integration (Kettle) is mostly used for small enterprises, as it limits the speed and having limited debugging facility. This survey mostly helps us for selection of best ETL tool, but ultimately, the decision will depends on your organization and factors considered for selection of best ETL tool.

ACKNOWLEDGEMENTS

I would like to express my deepest appreciation to all those who contributed in some way to the work described in this paper. My sincere thanks to my project guide **Prof. SachinBojewar** for giving me intellectual freedom of work and guiding me on proper time. I would also like to thanks head of computer department and to the principal of AlamuriRatnamala Institute of engineering & technology, Shahapur, Thane for extending his support for our work.

REFERENCES

[1] W. Inmon D. Strauss and G.Neushloss, “DW 2.0 The Architecture for the next generation of data warehousing”, Morgan Kaufman, 2007.

[2] A.Simitisis, P. Vassiliadis, S.Skiadopoulos and T.Sellis“DataWarehouse Refreshment”, Data Warehouses and OLAP: Concepts,Architectures and Solutions, IRM Press, 2007, pp 111-134.

[3] R. Kimball and J. Caserta. “The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data”, Wiley Publishing, Inc, 2004.

[4] A. Kabiri, F. Wadjinny and D. Chiadmi, "Towards a Framework for Conceptual Modelling of ETL Processes", Proceedings of The first international conference on Innovative Computing Technology (INCT 2011), Communications in Computer and Information Science Volume 241, pp 146-160.

[5] P. Vassiliadis and A. Simitsis, “EXTRACTION, TRANSFORMATION,

ANDLOADING*“,http://www.cs.uoi.gr/~pvassil/publications/2009_DB_encyclopedia/Extract-Transform-Load.pdf*

[6] Nils Schmidt, Mario Rosa, Rick Garcia, Efrain Molina, Ricardo Reyna and John Gonzale, “ETL TOOL EVALUATION- A Criteria Framework “

[7] Data Warehouse, *http://datawarehouse4u.*info accessed on September 10, 2014

[8] ETL Tools information, *http://etl-tools.info/en/bi/etl_process.htm* accessed on September 12, 2014

[9] ETL Tools, *http://www.etltools.net* accessed on September 12, 2014

[10] Pentaho data integration, http://www.pentaho.com/product/data-integration accessed on September 14, 2014

[11] IBM InfosphereDataStage, *www.ibm.com/software/products/en/ibminfodata* accessed on September 14, 2014

[12] SAS ETL studio *,http://support.sas.com/software/products/etls/* accessed on September 14, 2014

[13] Jasper ETL tool *,http://community.jaspersoft.com*accessed on September 14, 2014

AUTHORS

Mr. NileshMali is currently a graduate student pursuing masters in Computer Engineering at ARMIE, Thane, University of Mumbai, India. He has received hisB.E in Computer Engineering from University of Pune. He has 3 year of past experience in teaching. His areas of interest are distributed database, data mining, Information Retrival, etc.

Mr. SachinBojewar is Associate professor in Information Technology Department, at VIT, Wadala, University of Mumbai, India. He has completed his B.E and M.E in computer Engineering. Currently He is pursuing his PhD in Software Engineering. He is having 25 years of experience in teaching. His areas of interest are Object Oriented Design, Software Engineering, Management Information System, Human Computer Interaction, Programming Languages, Management Information System, Software Architecture, etc.