A Comparative Study of ETL Tools
Helstina S. Rose¹ Prof.Pankaj Raibagkar²
1,2S.I.E.S College

Abstract— In many organization’s valuable data is wasted as it lies around in different formats and in various resources. The consolidated data is stored in a complex system or the Central Repository called the Datawarehouse. The ETL process is the best way to handle such kind of data. ETL is a process of Extracting data from a source, Transforming the data by cleansing it and Loading it into a desired target database. A proper ETL tool should be selected to improve the quality of data.

Key words: Datawarehouse, ETL Tools, Complex Data, Pentaho Data Integration, SAP BODS

I. INTRODUCTION
Data should be loaded into a Datawarehouse on a daily basis so that it can serve the purpose of facilitating business analysis. For this, data from many operational systems have to be extracted and loaded into a warehouse. EXTRACT, TRANSFORM and LOAD also known as ETL is an important component of the Datawarehouse. Various ETL tools are available in the market helping the enterprises to select one tool based on their requirements and needs.

A. ETL Process and Concepts

![ETL Process Diagram]

Fig. 1: ETL Process

ETL tools are used to
- Extract data from various data sources
- Transform the data in proper format for querying and analysis
- Load it into the final target

In ETL tools, all the three phases execute in parallel since the data extraction takes time, so while the data is being pulled another transformation process executes, processing the already received data and prepares the data for loading and as soon as there is some data ready to be loaded into the target, the data loading starts without waiting for the completion of the previous phases. Extract, Transform and Load (ETL), is an important component of the Data Warehousing Architecture. Various ETL tools are available in the market. This makes it easy for an enterprise to select one based on its requirements. Over the time the tools have improved. The improvements include “data quality control, monitoring and cleansing, real-time and on-demand data integration”. Moreover ETL tools can be customized according to the functional requirements of an enterprise data warehouse.

II. ETL TOOL FEATURES

Some generic behavior of ETL tools are as follows:

A. Architecture

For evaluating an ETL tool with respect to architecture, aspects such as support for parallel processing, symmetric multiprocessing, clustering, load balancing and feasibility for grid computing should be considered. Also support for multi user management of ETL processes running on multiple machines is to be considered.

![Architecture Aspects Diagram]

Fig. 1: Architectural Support

B. Functionality

Two main aspects relating to functionality of an ETL tool are important i.e. the metadata support and the overall functionality provided by the tool. The main functionality focuses on whether the tool is data cleansing oriented or data transformation oriented, or it performs both equally. From this one gets a clear picture of what tool to select depending on the nature of data. On the other hand an ETL is also responsible of using metadata to map source data to
destination. Thus choosing an appropriate tool is very important.

C. Usability
The usability is one of the important factors of any tool. Thus points to consider are that the tool should be easy to use, understand and simple to get used to. The tool should have a well-balanced interface.

D. Reusability
The reusability depends on that the components of a data warehouse architecture, which is constructed using the ETL tool, must be reusable and can handle parameters. The tools should be capable of dividing the process into small building blocks, allow user to make user defined functions and allowing these functions to be used in the process flow.

E. Connectivity
The main aspects to consider is the native connections that the tool supports, the packages it can read metadata from, capability to graphically join tables, support for changed data capture principle, address cleansing ability etc.

F. Interoperability
Last but not the least the tool should be capable to run on a number of platforms and also on the different versions of a product.

III. OPEN SOURCE
Open source ETL tools are the latest and most used. They are an alternative to commercial packaged solutions as well as a productive and efficient option. They are designed to handle high performance and scalability requirements. The most important feature is that they are significantly less expensive than tools of commercial licence.
The 4 basic positives that companies adopt open source ETL tools:
- Costs are reduced and thus results in savings and has many other embedded components and features.
- A system that allows system integrators to deliver integration capabilities significantly faster and with a higher level quality.
- Using free ETL tools technology gives support to the initiatives of small enterprises
- Small companies likely support open source BI providers as they have less demanding needs for data integration.
### IV. SOME FAMOUS ETL TOOLS

Pentaho Data Integrator - Kettle: It is Java (Swing) application and library. Kettle is an interpreter of procedures written in XML format. Kettle provides a JavaScript engine to fine tune the data manipulation process. Kettle is also a good tool, with everything necessary to build even complex ETL procedures. Kettle provides a Java or JavaScript engine to take control of data processing. Kettle (PDI) is the default tool in Pentaho Business Intelligence Suite. The procedures can be also executed outside the Pentaho platform, provided that all the Kettle libraries and Java interpreter are installed. Pentaho Business Analytics is a business analytics software that can help users to discover, blend and access all sizes of data and take information-driven decisions. The system can help users to visualize data.

#### A. Pentaho Business Analytics Specifications
- **Name:** Pentaho Business Analyst
- **Company:** Pentaho
- **Address:** Pentaho, Hazeltine National Drive, Orlando, FL
- **Software Type:** Standalone, On Premise, Cloud, Saas
- **Typical Customers:** Small business, Medium business, Large business
- **Business Area:** Cross Business Area
- **Platforms:** Desktop, Cloud

#### V. FEATURES
- It has improved system performance monitoring such as minor bug fixes to the PDI-specific portions of Pentaho.
- Data Profiling Perspective includes Data Cleaner. Analyze tables and columns in preparation for ETL.
- The PDI marketplace makes it possible to share and download new plugins.
- The Data Services and Kettle JDBC Driver enable you to deliver data from multiple data sources.
- Database connections can be used with all jobs. It enables commits rollbacks on a job level. Prior it was only possible with transformations.

#### A. Pentaho Business Analytics Review

1) **Overview**

Pentaho Business Analytics is a business analytics software that can help users to discover, blend and access all sizes of data and take information-driven decisions. The system can help users to visualize data in a variety of multiple dimensions, starting from basic reports to predictive modeling. It can also help users to minimize their dependence on IT.

Pentaho, the company behind this business analytics system, has been providing business analytics solutions since 2004. The company is headquartered in Orlando, Florida and has served over 1,200 enterprise customers including Kiva, Swissport and Marketo. The company is popular for providing a cost-effective suite for data visualization, access, analysis, integration and mining.

#### B. Pentaho Business Analytics Average Rating

The average rating of Pentaho Business Analytics is 5 stars. The rating is based on 69 aggregated online reviews.

### C. Pentaho Business Analytics Typical Customers

Pentaho Business Analytics is perfect for small and growing business companies. Industries providing healthcare, retail and financial services and government can use this software. Marketo, Telefonica and Lufthansa are some of the clients of Pentaho Business Analytics.

#### D. Pentaho Business Analytics 5 Pros
- Pentaho Business Analytics supports visualizing and analyzing various types of data across multiple dimensions
- Pentaho Business Analytics offers true mobile experience for users as it which can be used from anywhere
- Pentaho Business Analytics provides 24x7 technical live support to its users
- The interactive visual analysis feature can give users immediate access to any data
- The business analytic software offers a 30-day free trial version for users

#### E. Pentaho Business Analytics 2 Cons
- Pentaho Business Analytics includes a Weka data-mining module that is less capable and advanced compared to other similar tools
- Pentaho Business Analytics’ documentation is weak comparing to other similar tools and can be difficult to use for some users

#### F. Commercial ETL Tool: Sap Business Objects

SAP Business Objects Data Services is an Extract, Transform and Load (ETL) tool used to move and manipulate data between source and target. SAP Data Services provides a data management platform that can support various initiatives including business intelligence, data migration, application integration and many more specific applications. SAP Data Services are the executable components within the application that can be deployed in either a batch or real time based architecture.

SAP BusinessObjects is SAP’s BI solution. By having immediate access to the latest data affecting their business, the company can benefit from it. Business Objects Data Integrator is an ETL tool for data integration, mainly focus on data quality features. It also uses data repository for storing the created jobs and projects. BODI is commonly used for building data marts, data warehouses, etc. It provides a lot of options in data manipulation like • 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.

### VI. SAP BUSINESS OBJECTS BI SPECIFICATIONS

- **Name:** SAP Business Objects BI
- **Company:** SAP
- **Address:** Walldorf, Germany
- **Software Type:** Standalone, On Premise, Cloud
- **Typical Customers:** Start up, Small business, Medium business, Large business
- **Business Area:** Cross Business Area
Platforms: Desktop, Cloud

**A. Features of SAP BODS**

- SAP BODS combines industry leading data quality and integration into one platform.
- One of the fundamental capabilities of Data Services is extracting, transforming, and loading (ETL) data from heterogeneous sources into a target database or data warehouse.
- It supports multi users
- It has comprehensive administrative tools and reporting tool
- It provides high performance parallel transformations
- SAP BODS is very flexible with web-service based application
- It allows scripting language with rich sets of functions

SAP Business Objects BI allows each user to create reports, which they can save on their own folders or share with other users. They can also specify users that can access certain reports. Dashboards can be customized based on user and company needs.

Reports saved on the application or the local database can be accessed easily and merged for convenient analysis and reporting. SAP Business Objects BI comes with a scheduling feature to ensure that communication is sent to the right people at the right time. Team collaboration and communication can be handled easily without involving another application as it comes with an inbox for each user and a comment box for each report. Decisions can be made easily with the most up to date information and rules. Instant access to data also allows to quickly provide the right answers to clients.

**B. SAP Business Objects BI Review**

1) **SAP Business Objects BI Overview**

SAP Business Objects is SAP’s BI solution. By having immediate access to the latest data affecting their business, businesses can make sound decisions for the benefit of the company. This is what SAP Business Objects BI is all about: it allows users to create, access, use and share information within the company.

SAP Business Objects BI was developed by SAP. The company is a world leader when it comes to enterprise application solutions. Over 251,000 clients rely on SAP for their business operations.

2) **SAP Business Objects BI Average Rating**

This software product is not yet rated

3) **SAP Business Objects BI Typical Customers**

SAP Business Objects BI is ideal for small to large enterprises in which access to information is crucial for business operation. It can be used in various industries, including financial services, manufacturing and distribution, government, non-profits and more. Some of the customers that use this solution are Citrix, Feeding America, Plan International and Keen Footwear.

4) **SAP Business Objects BI 5 Pros**

- SAP Business Objects BI allows a company to achieve its goals by immediate access to real time critical operational data
- Reports can be easily analyzed and compared, allowing users to determine if they are on the right track
- Users can easily work with each other, ensuring fast issue resolution and arriving to the best decision through teamwork
- SAP Business Objects BI can be accessed via mobile devices, so users are always updated on business critical data
- SAP Business Objects BI supports various languages, making it practical for companies with global operations

VII. BODS BENEFITS

- SAP BODI provides tools for data mining and quality.
- View dashboards of DI execution to see at a glance the status and performance of the Job execution for one or more repositories over a given time period.
- Analyze end-to-end impact for DI tables and columns, and BO objects such as universes, business views and reports.
- Evaluate the reliability of the target data based on the validation rules created in the Batch Jobs in order to quickly review, assess and identify potential inconsistencies or errors in source data.
- View, analyze and print graphical representations of all objects as depicted in DI Designer, including their relationships and properties.

**A. Advantages**

- SAP Business Objects created a firm company determined to stir the market
- Good data modeling and data-management support
- SAP Business Objects provides tools for data mining and quality: profiling due to many acquisitions of other companies.
- Quick learning curve and ease of use

**B. Disadvantages**

- Uncertain future. Controversy over deciding which method of delivering data integration to use: SAP BW or BODS.
- Business Objects Data Integrator also known as Data Services may not be seen as a stand-alone capable application to some organizations.

VIII. CONCLUSION AND FUTURE SCOPE

Security of data is a major challenge. This research proves that there is improvement in data extraction speed by using flat file with security measures. This is achieved by converting the database table into flat files and then extract them. The flat file needs less storage space and its extraction is faster than that of a database table. As security is provided to the flat file, the hacker cannot use the data of the file during the extraction process. This work can be extended further for the later building process of the data warehouse for transformation and loading process of ETL.
REFERENCES


[3] https://www.google.co.in/search?q=images+of+the+etl+process&espv=2&biw=1366&bih=667&tbm=isch&tbo=u&source=univ&sa=X&ved=0ahUKEwiKre6js9DNAhXMdRfKjRBqQQ7Ahg


