IMPROVED USE OF ETL TOOL FOR UPDATION AND CREATION OF DATA WAREHOUSE FROM DIFFERENT RDBMS

Agnas Vidya Michael
IT Department
Pillai College of Arts, Commerce and Science
Mumbai, India
agnas@mes.ac.in

Purnima Ahirao
IT Department
K J Somaiya College of Engineering
Mumbai, India
purnimaahirao@somaiya.edu

***Abstract*—ETL plays a very important role in any organization in the perspective of taking critical business decisions. The companies can use ETL tool to analyze their business data and answer complex business queries which Transactional databases cannot answer. ETL is a method of moving the data from various sources into a data warehouse. If the data source is changed, ETL facilitates updates on data warehouse automatically. Hence there is a need for well-designed and documented ETL system for the success of a Data Warehouse project. The Proposed paper shows the execution of ETL operation using Visual Studio 2017 with SQL Server and PowerBi with MS Excel. All the ETL operation steps executed using Visual Studio 2017 with SQL Server and PowerBi with MS Excel provides the improved results for all types of Complex Business Analytics and Queries on top of Data Warehouse. A comparison of these two technologies are carried out.**

Keywords: ETL, Data Warehouse, DBMS, Analytics, Execution

I. INTRODUCTION

ETL stands for Extract, Transform and Load. In this process, an ETL tool extracts the data from different RDBMS source systems then transforms the data like applying calculations, concatenations, etc. and then load the data into the Data Warehouse system. It may be obvious to think as, creating a Data warehouse is simply extracting data from multiple sources and loading into database of a Data warehouse. ETL process takes inputs from various stakeholders that includes developers, analysts, testers, top executives. In order to maintain the standard of this tool, data warehouse system needs to change with business changes. ETL also helps in providing facilities such as transformation, aggregation and calculation rules. Format issues between the data sources is taken care by ETL. A Well documented and designed system for the use and execution of ETL steps is a need for the Business enterprises. The proposed paper shows how the combination of Visual Studio 2017 with SQL Server and PowerBi with Excel acts as tools for ETL execution for Improved Productivity of Business through strong analytics and decision making. Following are some of the features of ETL:

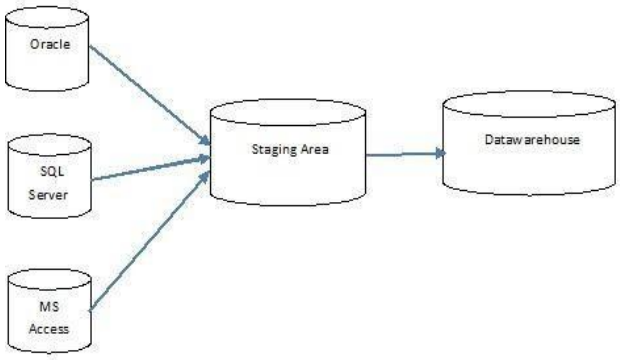
- Allow verification of data
- ETL process allows to perform data comparison between the source and the target system.
- Helps in performing complex transformations and requires the extra space to store the data.
- Helps to move data into a Data Warehouse by converting data into various formats and types so that one consistent system is adhered.
- One of the predefined processes for manipulating data from source into a target database.
- Offers deep historical context for the business.
- Helps in improving the productivity

II. RELATED WORK

Paper [1], mentions that every organization is generating data and information is gathered from that data. This information is useful in taking decisions. Business Intelligence (BI) allows to create reports required by every employee to see his own required content. Data are collected from various sources and then transformed into data warehouse. OLAP tools enable users to analyze data from multiple perspectives. There is a portal that allows to create static and dynamic reports and analysis. The BI software used are Cognos from IBM and some BI software from Oracle. Microsoft database was used for Cognos and Oracle database as used for Oracle solution. Data was extracted from information system Gemini. Data was stored in both the databases and both can be used as data warehouse. Paper [2], conducts a comparison of development platforms, mobile frameworks and data storage strategies. Future market trends are also discussed. It discusses the process of designing a mobile BI application prototype and then evaluates the mechanism of the 'App' and its functionalities. Finally, the paper also provides a guideline for developing Mobile BI applications. Paper [3], the value and use of Business Intelligence in an organization is clearly mentioned. The organizations can use Business intelligence and information technology to collect information and analyze it. Managers can make good decisions and additional cost savings and also improve performance and increase productivity by making use of BI. In paper [4], an overview of Business Intelligence, the key technology of Business Intelligence, as well as the establishment & application of Business Intelligence System in retail industry is explained. Business subject and dimension design, ETL tool design, Data Display middleware design and the main innovation are the key points of the system[4]. In paper [5], data is collected from different data sources such as GSMS and AIMS database, Excel and Access by meeting with different stakeholders. Data that is collected will be integrated and undergoes ETL

process before it is stored in Data Warehouse. BI technology is used for aggregation and analysis of data from different sources. Finally, result analysis and data analysis of the integrated data is carried out. This paper highlights the importance of data collection, data integration and ETL in business intelligence framework and its significance for better education management. In paper [6], the author expounds the decision supporting system in light with the data warehouse technology, and analyzes profoundly the systematic structure and disadvantages of traditional decision supporting system. On the basis, the author of this paper discusses the importance of applying data excavation technology to the decision supporting system with examples that can provide new measures to cultivating the decision supporting system with better performance. In paper [7], architecture for data warehouse systems is proposed. The architecture involves client/server properties and deductive database features. It can provide good properties of generality, flexibility, efficiency, scalability, and intelligence. It includes detailed information flow in warehouse systems and functions of warehouse components. Big data, not only let the database known for efficient and secure storage embarrassed, but also allow the data warehouse that provide the convenience for business intelligence and decision-making lose their role. Paper [8] discusses about Big Data, data warehouse and DSS. It focuses on the research of data warehouse in big data, features and technical process of big data, analysis of the data warehouse' characteristics and application. A comparison of traditional data warehousing and data warehousing in big data is mentioned in this paper. Paper [9], gives an account of the emerging technologies in business intelligence and data warehousing that significantly help with improving the performance data warehouses and business intelligence tools. It explains parallel processing dbms, in-memory analytics in Business analytics and dbms data warehousing, BI with cloud computing, BI in mobile.

III. ETL PROCESS



ETL is a process that consists of 3steps is a

Fig. 1. ETL Process

A. *Extraction*

In this step, the data of specific interest is extracted from different database sources like Oracle, SQL Server, MS Access, flat files and applications. No spam is allowed in this process. Most of the time it becomes difficult to extract data of specific interest. So, more data than required are extracted and then later relevant data are identified. Staging area is used in this step as most of the source systems are available for extraction for a limited period of time which is less than the total data-load time. Our data are extracted from the data sources and stored in the staging area before the time slot ends. Staging area allows us to extract data from multiple data sources and also to perform joining of two tables from different databases. Depending on the source system's capabilities few transformation can also be carried out in this step. The data extraction time slot is different for different time zones and operational hours. The size of the extracted data can vary from hundreds to gigabytes depending upon the business.

B. *Transformation*

Data that is extracted from data sources are in raw form. It needs to be cleansed, mapped and then transformed. Transformation is the process of cleaning and aggregation of data that are required to prepare data for analysis. Following are some of the basic transformations:

- Mapping of NULL values to 0 or "Male" to "M" and "Female" to "F," etc.
- Identifying and removing duplicate records.
- Conversion of Units of Measurements like Date Time Conversion, currency conversions, numerical conversions, etc.
- Selecting only certain rows or columns for loading.
- Splitting a single column into multiple columns.
- Key restructuring: Establishing key relationships across tables.

C. *Loading*

- It is the last step of the ETL process. After the data is extracted from multiple sources and transformed into a standard format, it is then loaded into a storage system, such as a cloud data warehouse.
- In a typical data warehouse, large volumes of data are loaded in a very short period of time. During load failure, recovery mechanisms should be arranged to restart from the point of failure without any loss of data integrity.

There are many Data Warehousing tools available in the market. Here are some of the most prominent one:

1) CloverETL:

It is a commercial open source software. CloverETL is used to cleanse, standardize, transform, and distribute data to applications, database, and warehouses. It can be used standalone, as a command-line application, a server application, or even embedded in other applications. CloverETL has been used on the Windows platform and also on Linux, HP-UX, AIX, AS/400, Solaris and OSX. CloverETL Components are CloverETL Engine, CloverETL Designer, and CloverETL Server. CloverETL should allow users to combine, transform, and move data from any source.

2) KETL:

Your business never stops generating data. The ability to pull apart, focus and analyse this flow of information is the

answer to making better decisions about your supply chain, your product, and your customers. KETL's is designed to assist in the development and deployment of data integration efforts which require ETL and scheduling

3) Jaspersoft:

Jaspersoft data integration software extracts, transforms, and loads (ETL) data from different sources into a data warehouse or data mart for reporting and analysis purposes. You can leverage and combine several disparate relational or non-relational data sources.

4) MarkLogic:

MarkLogic is a data warehousing solution which makes data integration easier and faster using an array of enterprise features. It can query different types of data like documents, relationships, and metadata.

5) Oracle:

Oracle is the industry-leading database. It offers a wide range of choice of Data Warehouse solutions for both on-premises and in the cloud. It helps to optimize customer experiences by increasing operational efficiency.

6) Amazon RedShift:

Amazon Redshift is Data Warehouse tool. It is a simple and cost-effective tool to analyze all types of data using standard SQL and existing BI tools. It also allows running complex queries against petabytes of structured data.

IV. ETL USING TOOLS

A. ETL performed using Visual Studio 2 017 and SQL Server

Source Data is extracted from Excel file. This data is then transformed (conversion of the column values from lowercase to uppercase) and loaded into target data warehouse (SQL server database). Following figure shows the design view of ETL process performed using Visual Studio.

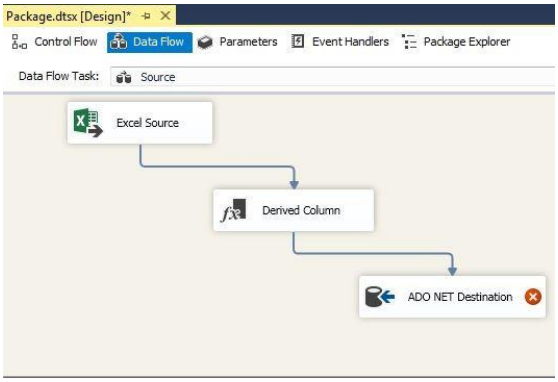


Fig. 2. ETL process flow from Excel Source to SQL Server (ADO NET) using Visual Studio

B. ETL performed using PowerBi and S QL Server

Source Data is extracted from Excel file. This data is then transformed (conversion of the column values from lowercase to uppercase) and loaded into target data warehouse (SQL server database). Following figure shows the design view of ETL process performed using PowerBi Editor.

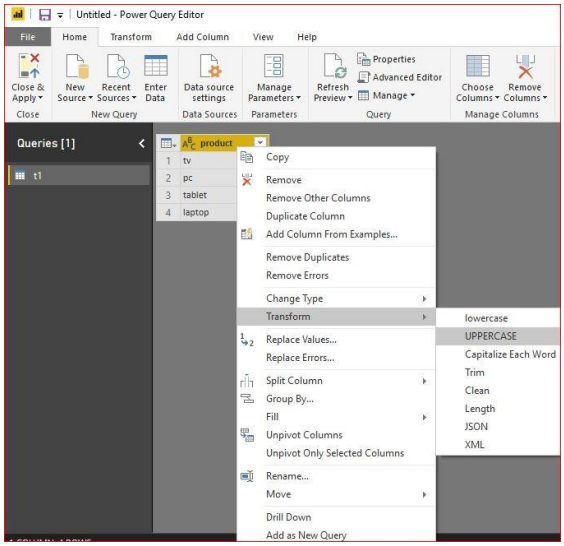


Fig. 3. Perform ETL Transformation from Excel Source to SQL Server (ADO NET) using PowerBi

C. Comparison of ETL using Visual Studio and PowerBi

The following table, table1 shows a comparison of ETL supported using Visual Studio and PowerBi based on the supported operating system, database connectivity and Big Data support provided by the softwares.

Table1: Comparison of ETL using Visual Studio and PowerBi

Tool	Supported Operating Systems	Database and Storage Connectivity	Big Data	Indemnification / Warranty
Microsoft Visual Studio	Windows 7 SP1, Windows 8, Windows Server2008 R2 SP1, Windows Server	SQL Server Database, Access Database, SQL Server Analysis Services	Apache Hadoop® and NoSQL Components	Yes

PowerBi	Windows 7/Windo wsServer 2008 R2, or later.	SQL Server Database, Access Database, SQL Server Analysis Services Database, Oracle Database, IBM DB2 Database, Amazon Redshift etc.	Apache Hadoop ® and NoSQL Compo nents	Yes

CONCLUSION

ETL acts as an important tool for decision making process. The different software tools used such as Visual Studio 2017, Power BI with SQL Server for ETL processing proves useful for Business organizations in handling complex data queries and their analysis. The analysis can further help Enterprises to improve their business productivity and increase the overall performance of the organization.

REFERENCES

[1] M.Miškuf, I.Zolotová , "Application of Business Intelligence solutions on Manufacturing data", IEEE 13th International Symposium on Applied Machine Intelligence and Informatics, January 22-24, 2015

[2] Sathyanath Lappasi Ramamoorthy, Jagdev Bhogal, "Developing a Mobile BusinessIntelligence Application", 2014 Eighth International Conference on Complex, Intelligent and Software Intensive Systems, pp. 392-397

[3] Fereydoon Azma , Mohammad Ali Mostafapour, "Business intelligence as a key strategy for development organizations", Procedia Technology (2012) 102 – 106

[4] Tong Gang and Cui Kai Song Bei, "The Research & Application of Business Intelligence System in Retail Industry", Proceedings of the IEEE International Conference on Automation and Logistics Qingdao, China September 2008, pp. 87-91

[5] Nur Alia Hamizah Mohamad Rodzi, Mohd Shahizan Othman, Lizawati Mi Yusuf, "Significance of Data Integration and ETL in Business Intelligence Framework for Higher Education", 2015 International Conference on Science in Information Technology (ICSITech),pp. 181-186

[6] Guo-xiang Liu,Zhi-heng Qi, "The Application of Data Warehouse in Decision Support System", 2012, pp. 1373-1376

[7] Jixue Liu Millist Vincent, "An architecture for DataWarehouse Systems", 1998, pp.107-110

[8] Hai-fei QIN, Zhi-ming QIAN, "On the Research of Data Warehouse in Big Data", 2015 International Conference on Network and Information Systems for Computers,pp.354-357

[9] Nayem Rahman, Fahad Aldhaban, Shameem Akhter, "Emerging Technologies in Business Intelligence", 2013 Proceedings of PICMET '13: Technology Management for Emerging Technologies, pp.52-57