DATA ANALYSIS AND ETL TOOLS IN BUSINESS INTELLIGENCE

Geetha.S
Computer Science Engineering Dept
SVKM'S Shri Bhagubhai Mafatlal Polytechnic Mumbai, India
geesubra@rediffmail.com

Keval Rajesh Dhanani
Computer Science Engineering Dept
SVKM'S Shri Bhagubhai Mafatlal Polytechnic Mumbai, India
dhanani.kev01@gmail.com

Param Pankaj Doshi
Computer Science Engineering Dept
SVKM'S Shri Bhagubhai Mafatlal Polytechnic Mumbai, India
paramdoshi1435@gmail.com

Abstract—Business intelligence (BI) is a collection of software and services to convert facts into actionable observations. These observations can impact strategic and tactical business decisions of an enterprise. BI can provide users with exhaustive intelligence about the state of the business with the assistance of tools which can access and scrutinize data sets and deliver pinpointing findings in reports and summaries etc. The technology has considerably advanced in the area of information systems through the evolvement of DSS (Decision Support Systems) to EIS (Executive Information Systems) to Business Intelligence systems. ETL (Extract, Transform, and Load) is a procedure of pulling out data from various data sources and processing them according to business calculations and transferring the reformed data into a data warehouse. ETL function lies at the core of Business Intelligence systems because of the in-depth analytics data it provides. Enterprises can gain past, current, and projecting views of real business data with ETL.
Keywords: Business Intelligence; ETL; DSS; EIS;

I. INTRODUCTION

A. Business Intelligence and Data Warehousing

Business intelligence and data warehousing have different objectives. BI is primarily dedicated for producing operational or strategic business insights like product positioning and product pricing, profitability, sales performance, forecasting, strategic directions on a broader level, while a data warehouse has its significance in storing all the company's data from heterogeneous sources in a single place. BI systems utilize data warehouse while data warehouse acts as a foundation for business intelligence.
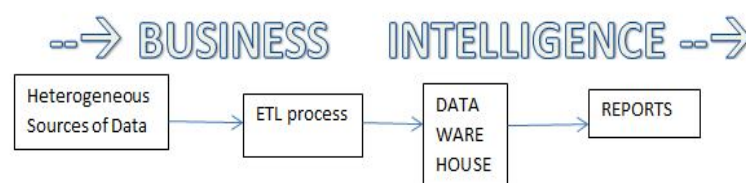

Fig. 1 Business Intelligence in Data Warehouse

II. DATA ANALYSIS AND BI

Data analysis in BI is used in to examine the data and generate some perceptions out of it. The process of data analysis consists of defining and studying the data, scrubbing and transforming the data so that it can give a meaningful outcome. The order followed in data analysis is data gathering, data scrubbing, and analysis of data and accurate interpretation of data so that anyone can understand what the data want to tell. Descriptive analysis, exploratory analysis, inferential analysis, predictive analyses are some beneficial forms of data analysis. It is a competitive advantage to make business decisions based on data. We need to link data analysis to business outcomes in order to show the impact of BI and data analysis in revenue-generating opportunities, increased operational efficiencies and improved customer's services. The actionable intelligence to be extracted from the raw data need to be decided during data analysis. In data analysis, investigation is performed on past dataset to understand what happened with the data so far. For example if we have 2GB customer purchase related data of past 2 years and we are trying to find what has materialized pertaining to purchase data, we are doing a past data analysis. Data analysis is necessary to understand the data. The BI solution can drill deep into the data, providing a combined view of goods, consumers and business data.

III. ETL AND BI

ETL refers to the extraction, transformation, and loading of data from one data source to another. In business intelligence, an ETL tool extracts data from one or more data-sources, transforms it and cleanses it to be optimized for reporting and analysis, and loads it into a data store or data warehouse.

Businesses depend on the ETL process for a combined data view that can enable improved business judgments. High-level Data Mapping of ETL allows mapping data for specific applications. ETL also guarantees the quality of data in the warehouse through standardization and eliminating duplicates. The modern-day ETL tools perform automatic & Faster Batch Data Processing. Using ETL and data integration, enterprises can obtain the best data view across multiple sources. SSIS, SSAS and SSRS are some of the tools used in MSBI (Microsoft Business Intelligence) for ETL, Analysis and reporting services. We will briefly analyze the features each of these tools in the following part.
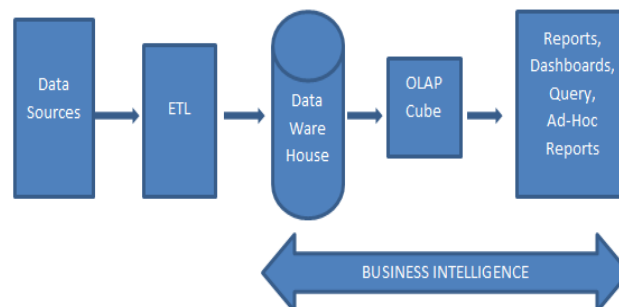

Fig. 2 BI and ETL Process

A. SQL Server Integration Services(SSIS)

SSIS is a platform for building enterprise-level integration and conversions solutions for data. Integration Services can extract and transform data from hetrogeneous data sources such as XML data files, flat files, and relational data sources, and then load the data into one or more destinations.
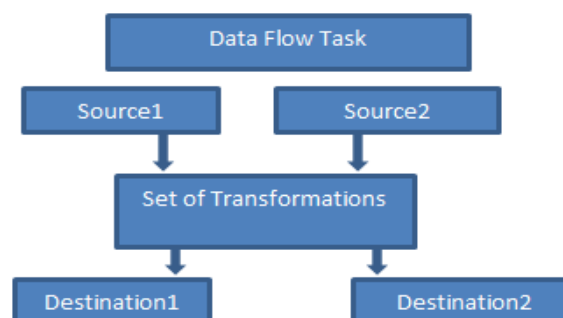

Fig. 3 Integration Services Data Flow

_____

These Services includes a rich set of incorporated tasks and transformations, graphical tools for constructing packages.It also provides   Integration Services Catalog database,which can help in storing, running and managing packages. The graphical Integration Services tools can create solutions without writing a single line of code. SSIS Solutions can be developed in SQL Server Business Intelligence Development Studio (BIDS), a visual development tool based on Microsoft Visual Studio. SSIS files are structured into packages, projects and solutions. The package is at the bottom of the hierarchy and covers the tasks required to achieve the actual ETL operations. Each package is part of a project is saved as a .dtsx file and one or more packages can be included in a project. Solution is at the top of the hierarchy and project is a part of a solution and we can include one or more projects within a solution. A Conditional Split Transformation scenario in SSIS is presented here. Conditional Split Transformation in SSIS checks the given condition and based on the condition result the output will send to the appropriate destination path. It has one input and can have many outputs. In the current scenarion we are splitting employees into retired and non-retired employees based on their age(age>=60 or age<60) and storing in two different data store destinations.
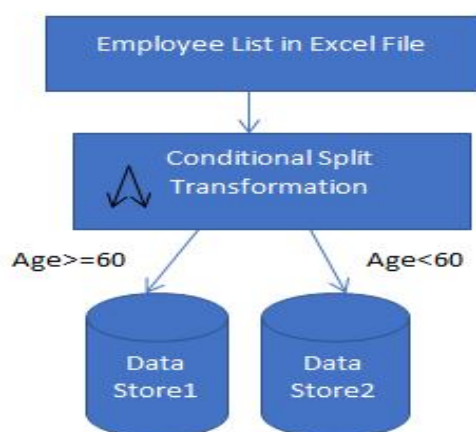

Fig. 4 Conditional Split Workflow

Excel File: List of all employees

EmpID	EmpName	EmpAge	EmpSalary
1	Dhruv	50	50000
2	Meet	62	65000
3	Mohit	63	53000
4	Yash	59	52000
5	Jay	55	60000

Table 1: Retired Employees

EmpID	EmpName	EmpAge	EmpSalary
2	Meet	62	65000
3	Mohit	63	53000

Table 2: Non-Retired Employees

EmpID	EmpName	EmpAge	EmpSalary
1	Dhruv	50	50000
4	Yash	59	52000
5	Jay	55	60000

Fig. 5 Sample Conditional Split

B. SQL Server Reporting Services(SSRS)
SSRS is a reporting software which produces reports with tables in the form of data, graph, images, and charts. It is part of Microsoft SQL Server Services suite.
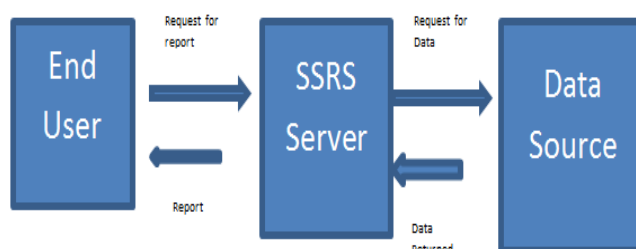

Fig. 6 Workflow of SSRS

_____

SSRS offers faster processes of reports on both relational and multidimensional data. SSRS can retrieve data from OLEDB, and ODBC DB connections. The key components of SSRS are Report Builder, Report Designer, Report Manager, Report Server, and Data sources. SSRS can create reports from various data sources with rich data visualization like charts, maps, spark lines etc and these reports can be observed through web browsers. SSRS reports can be exported in various formats like Excel, PDF, word etc. SSRS is a full-featured report engine. Reports can be generated against any data source that has an accomplished code provider like OLEDB, or ODBC data source. We can effortlessly retrieve data from SQL Server, Oracle, Analysis Services, Access, or Essbase, and many other databases.
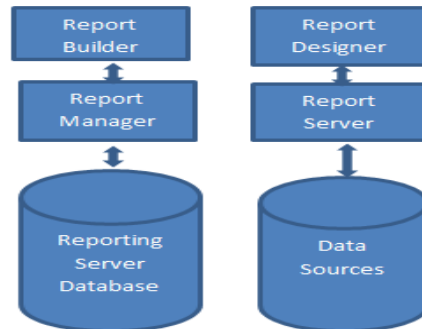
Fig. 7 Components of SSRS

A sample SSRS report generated on student details along with total marks obtained is shown in figure 8. In this report, the students are grouped with respect to their country. The Country column is grouped using parent grouping. The sum function is used to get the total subject marks of each student country wise.

Country	Student_Name	City	Marks
[Country]	[StudentName]	[City]	[Sum(Marks)]

Country	Student_Name	City	Marks
United States	Alex	New York	455
	Ana	Boston	854
India	Ramesh	Mumbai	965
	Suresh	Delhi	541
Sri Lanka	Mathews	Colombo	659
	Dinesh	Galle	789

Fig. 8 Sample Student Report

C. SQL Server Analysis Services(SSAS)

SSAS is the technology from the Microsoft Business Intelligence stack to develop Online Analytical Processing (OLAP) solutions.
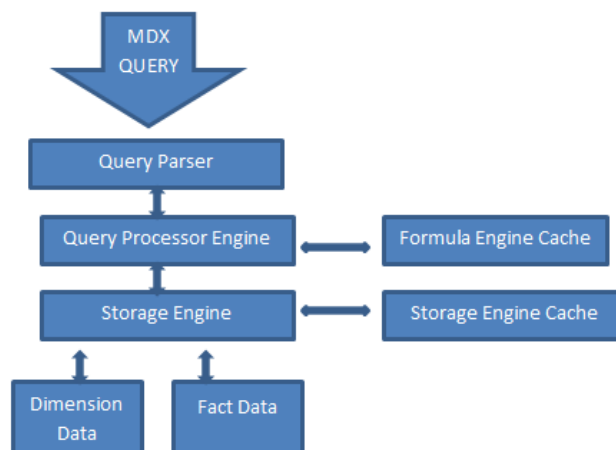
Fig. 9 Internal Architecture of SSAS

We can use SSAS to create cubes using data from data marts or data warehouse for deeper and faster data analysis. Cubes are multi-dimensional data sources which have dimensions and facts as its core elements. Dimensions can be assumed as tables and facts can be assumed quantifiable features. These features are generally warehoused in a pre-aggregated standard format and users can examine enormous volumes of data and slice this data by dimensions very effortlessly. Multi-dimensional expression (MDX) is the query language used to query a cube. The examples of dimensions can be product or geography or time or customer, and examples of facts can be orders or sales. A typical scenario can be sales analysis in Europe Region during the past 5 years. Here the data can be assumed as a pivot table where region is the column-axis and years is the row axis, and sales can be seen as the values.

IV. PENTAHO DATA INTEGRATION(PDI)

Using Pentaho ETL developers can create data management jobs with a user-friendly graphical creator, and without writing a single line of code. PDI uses a repository which enables remote ETL execution. The PDI components for implementing ETL processes in Pentaho are given below:

1. Spoon- ETL developers use it to create transformations and jobs.
2. Pan- executes transformations modelled in Spoon.
3. Kitchen- executes jobs designed in Spoon.
4. Carte- a simple web server used for running and monitoring data integration tasks.

The following scenario illustrates a basic PDI transformation. Here the source is a flat file of sales data which contain customer sales records with PIN codes and missing PIN codes. Through the PDI transformation we can resolve the missing PIN codes problem using a lookup file.
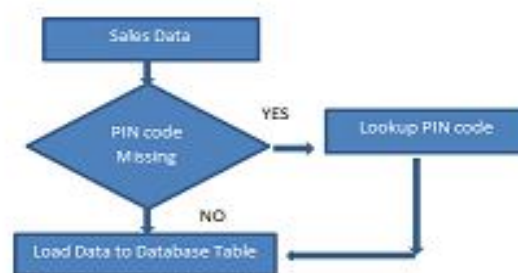

Fig. 10 Flowchart of PDI Transformation

Fig. 11 Task Steps in PDI Transformation

V. CONCLUSION

With ETL, business leaders can make data-driven business decisions. There are many ETL software solutions available to today's businesses - from enterprise level tools to open-source integration tools. ETL tools are designed and used to reduce time and cost when a new data mart or data warehouse is developed. Large organizations use SSIS, because it can handle the huge database. Small Enterprises use Pentaho, because of its limited speed and debugging facility. Advanced data analytics requires a modern approach to data integration for incorporating data from databases, streaming services, files, or other sources and choosing the right ETL toolset is critical for this task.

REFERENCES

1. Paulraj Ponniah, "DATA WAREHOUSING FUNDAMENTALS A Comprehensive Guide for IT Professionals" A Wiley-Interscience Publication New York, 2001.
2. Panos Vassiliadis, "A Survey of Extract–Transform–Load Technology", International Journal of Data Warehousing & Mining, 5(3), 1-27, July-September 2009.
3. Sonali Vyas and Pragya Vaishnav, "A comparative study of various ETL process and their testing techniques in data warehouse", Journal of Statistics and Management Systems, 20:4, 753-763, 2017.
4. Mohammad Atwah Al-ma'aitah. "The Role of Business Intelligence Tools in Decision Making Process", International Journal of Computer Applications 73(13):24-32, July 2013.
5. Nitin Anand, "Application of ETL Tools in Business Intelligence", International Journal of Scientific and Research Publications, Volume 2, Issue 11, November 2012, ISSN 2250-3153.
6. Ranjith Katragadda, Sreenivas Sremath Tirumala and David Nandigam, "ETL tools for Data Warehousing: An empirical study of Open Source Talend Studio versus Microsoft SSIS", WCCAIS 2015.
7. M. M. Al-Debei, "Data Warehouse as a Backbone for Business Intelligence: Issues and Challenges," European Journal of Economics, Finance and Administrative Sciences, vol.33, 2011.
8. Vassiliadis, Panos, "A Survey of Extract-Transform-Load Technology". International Journal of Data Warehousing and Mining. 5. 1-27, 2009.
9. Ranjan, J."Business Intelligence: Concepts, Components, Techniques and Benefits", Journal of Theoretical and Applied Information Technology, Vol 9. No 1, pp 060 - 070 ,2009
10. Microsoft, "Ssis technical report," Microsoft, SSIS. Retrieved from microsoft.com, Tech. Rep., [Online]. Available: https://docs.microsoft.com/en-us/sql/integration-services/sql-server-integration-services