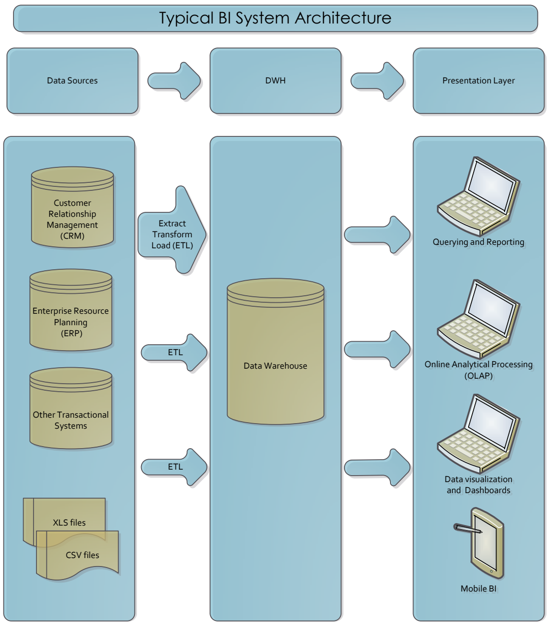
ETL in a nutshell

BY LEO KOZHUSHNIK

Figure 1 - Typical BI System architecture



Introduction

ETL, simply standing for Extract, Transform, Load, is a critical component in overall integration architectures and data transformation. Bouman, Casters and Doongen (2010) define ETL as a set of processes for getting the data from source systems into a Data Warehouse. Furthermore, both ETL & Data Warehouse are the core architectural components of the BI Systems.

BI systems go beyond just a set of querying and reporting tools used in an organization. BI Systems are in integrated set of tools, technologies, and software that are used to source data from disparate data sources, transform, and make it commonly available to the end users (see Figure 1). Typical BI system will include:

- Extract, Transform and Load (ETL) components to source the data from source systems and load into a data warehouse.
- Data warehouse that provide a central repository of data, transformed and aggregated, ready for analysis.
- Querying and Reporting tools that enable users to create reports.
- Online Analytical Processing tools (OLAP) allowing users to perform analysis.

More advanced BI System may also include:

- Data visualisation and Dashboards which provides users with an easy to read, high level, visual interface.
- Mobile BI enabling users' business insight on the move

ETL Defined

The main ETL steps involved in extracting data from source systems and bringing it into the data warehouse are grouped into 3x main sections: Extract, Transform & Load.

Bouman, Casters and Doongen (2010) describe **Extract** as a step that involves all the processing that is required to connect to various data sources, extract the data from these data sources, and makes the data available to the subsequent processing steps.

The main aim of the extract step is to retrieve all the required data from different sources, including database systems and applications, essentially like-for-like, into one area also known as the Staging area.

Once the data is extracted into the Staging area, the data can be cleaned. Cleaning is considered to be one of the most important steps in the ETL process as it ensures the quality of the data in the data warehouse before transformation is applied. This is achieved by applying basic cleaning rules incl. character validations, null removal, validation of addresses and post codes, etc.

Transformation involves any function applied to the extracted data between the extraction of the sources and loading data into targets (Bouman, Casters & Doongen 2010, p. 5).

Once the data is cleaned and/or cleansed, the transformation can be applied in the form of conversion, generating aggregates, and surrogate keys, sorting and deriving calculated measures. Complex business rules can also be applied at this step.

Post transformation, Bouman, Casters and Doongen (2010) describe **Load** as a process that involves all the processing required to load the data into the target system.

When the data is transformed, it can be loaded to the target system, which in this case is a target Data Warehouse. The load process can be performed Nightly, Weekly or Monthly.

Difference between ETL and ELT

Silvers (2008) describe earlier ETL applications as 3x separate platforms:

- Source system platform
- ETL platform
- Data Warehouse platform

Much like ETL, the ELT performs all the same steps/functions but in slightly different order and on a different platform/data warehouse. Silvers (2008) explains that the main difference with ELT is that operational data is extracted and loaded directly to staging tables on the data warehouse platform, instead of a separate ETL platform, then the transformation takes place.
In turn, the Staging area, where the data can be cleaned is occurring in the data warehouse database. The advantages to this are that the data warehouse platform is much more powerful and able to process much quicker and without disruption to operational system(s). The staging area can also be persisted where history is kept of the operational data and referred back to on needed basis.

ETL importance

From consolidation data from disparate data sources perspective, abovementioned ETL's importance is highlighted as a critical component in overall integration architectures and data transformation. Without ETL, there would be limited ability to extract data from a single or multiple source systems, at least, not systematically.
ETL allows for

- Consolidation of Data from disparate source systems and applications
- Performing routinely Data Validation
- Performing routinely Data Cleansing to ensure Data Quality
- Data Governance

ETL vendors

Many ETL tools exist within the marketplace. The Passionned Group (2015) that coordinates the ETL Tools & Data Integration Survey 2015, suggests that the ETL tools comparison is a 100% vendor independent study and reveals all the product details of all the major ETL tools in the market. Figure 2 outlines some of the current leaders in the market place.

Figure 2 – ETL market place leaders

ETL Tool	*Description*
SAP Business Objects Data Services	SAP Data Services delivers a single enterprise-class solution for data integration, data quality, data profiling, and text data processing that allows businesses to integrate, transform, improve, and deliver trusted data to critical business processes (SAP 2015).
IBM Infosphere Information Server	IBM InfoSphere Information Server is a market-leading data integration platform which includes a family of products that enable businesses to understand, cleanse, monitor, transform, and deliver data, as well as to collaborate to bridge the gap between business and IT (IBM 2015).
SQL Server Integration Services (SSIS)	Microsoft Integration Services is a platform for building enterprise-level data integration and data transformations solutions that can be used to solve complex business problems by copying or downloading files, sending e-mail messages in response to events, updating data warehouses, cleaning and mining data, and managing SQL Server objects and data (Microsoft TechNet 2015).
Oracle Warehouse Builder (OWB)	Oracle Warehouse Builder is a single, comprehensive tool for all aspects of data integration. Warehouse Builder leverages Oracle Database to transform data into high-quality information. It provides data quality, data auditing, fully integrated relational and dimensional modelling, and full lifecycle management of data and metadata (Oracle Help Center 2015).
Pentaho Data Integration	Pentaho data integration prepares and blends data to create a complete picture of that drives actionable insights. The complete data integration platform delivers accurate, "analytics ready" data to end users from any source. With visual tools to eliminate coding and complexity, Pentaho puts big data and all data sources at the fingertips of business and IT users alike (Pentaho 2015).

How transformation (T) plays its role in the ETL/ELT and its importance with an example from a sector

Bouman, Casters and Doongen (2010) suggest that the core value of ETL is in its Transform capabilities. The role of the T is multidimensional and covers many more functions than simply Transforming the data. Data Quality & Data Validation are part of transforming data and include Cleansing & Conforming.

Sigma Pharmaceuticals – industry sector example

Sigma Pharmaceuticals is a leading full line wholesale and distribution business to pharmacy and is also the owner of two of Australia's best known pharmacy retail brands (Sigma, 2015).

With ageing systems, Sigma identified a need for a development of new Data Warehouse to deliver a data set available for analytical capability that enables for the achievement of company's planned strategic initiatives.

Feasibility study for the project was conducted and found that access to good & reliable data was ranked as foremost priority, particularly:

- Enforcing data quality
- Efficient and Scalable
- Enabled Audit & Compliance reporting

This agrees with the findings by the QuerySurge (2015), who found that, based on information from analyst firm Gartner, bad data has been found in every database and data warehouse studied and is estimated to cost firms on average $8.2 million annually.

This is critical in the pharmaceutical industry where data needs to be accurate to meet critical audits and compliance reporting, and assurances must be made to warrant that the data warehouse is not populated with bad data.

This highlights paramount importance of and reliance on the Transformation component, particularly in addressing data quality issues.

Based on this, to provide a scalable solution, Microsoft's SQL Server Integration Services (SSIS) product was used as an ETL solution. A series of packages were developed that followed the ELT pattern. The data was extracted from the source system and into the Staging area of the data warehouse. Transformation tasks/routines were set up for each source table to be cleaned using a series of transformations to:

- Eliminate quality issues like NULL values
- Trimming blank characters
- Validating incoming addresses as spelling mistakes and fields left blank in the source system were very common
- Hashing banking details to comply with privacy laws

Once cleaned, data was loaded into dimensional model with additional Transformation tasks/routines to

- Populate dimensions and fact tables
- Creating aggregated summations of General Ledger amounts, Sales and Purchase Invoices
- Aggregate Inventory quantities

Transformation (T) component has enabled delivery of efficient and scalable solution that ensured good data enabled Audit & Compliance reporting, and easily accessible for necessary analysis, interpretation and trend identification (Lohrey 2015).

Conclusion

ETL is a critical component in overall integration architectures and data transformation. ETL provides a vehicle for connecting data from disparate source systems that may include various databases, flat files, mainframe systems, xml, etc. Ability cleanse and conform data in ways of filtering, reformatting, sorting, joining, merging, and aggregating provides for good data quality and governance principles that ensure that end users are presented with accurate data in a timely manner.

Biography

Leo Kozhushnik is a Principal Consultant at Expert Skills IT Professional Services who are a focused Business Intelligence & Data Warehouse Technology Professional Services provider. Leo has over 15 years of IT industry experience spread across various industry sectors with the main focus on relational database management, data warehouse design, development and implementation, development and deployment of BI solutions. Leo holds a Bachelor of Information Management (Business Systems) from Monash University, Post Graduate Certificate of Business Management (Project Management) from Swinburne University, Microsoft Technology Associate (MTA) and a TDWI Melbourne Chapter Member.

Leo is currently undertaking Masters of Business Analytics at Deakin University with expected completion in 2018.

References

- Bouman, R, Casters, M & Donger, J 2010, Pentaho Kettle Solutions: Building Open Source ETL Solutions with Pentaho Data Integration, John Wiley & Sons, retrieved 10 April 2015, Google eBook.
- IBM 2015, IBM InfoSphere Information Server, IBM, retrieved 20 April 2015, <http://www-01.ibm.com/software/data/integration/info_server/>.
- Lohrey, J 2015,The Need for Data Warehousing for Pharmaceutical Companies, Houston Chronicle, retrieved 21 April 2015, < http://smallbusiness.chron.com/need-data-warehousing-pharmaceutical-companies-80768.html>.
- Microsoft TechNet 2015, SQL Server Integration Services, Microsoft TechNet, retrieved 20 April 2015, < https://technet.microsoft.com/en-us/library/ms141026.aspx>.
- Oracle Help Center 2015, Warehouse Builder User's Guide, Oracle Help Center, retrieved 20 April 2015, <http://docs.oracle.com/cd/B28359_01/owb.111/b31278/concept_overview.htm#WBDOD10100>..
- Passionned Group 2015, ETL Tools Comparison, retrieved 13 April 2015, <http://www.etltool.com/etl-tools-comparison/>.
- Pentaho 2015, Pentaho Data Integration, Pentaho, retrieved 19 April 2015, < http://www.pentaho.com/product/data-integration>.
- QuerySurge 2015, Data Warehousing in the Pharmaceutical Industry, QuerySurge, retrieved 21 April 2015, <http://www.querysurge.com/solutions/pharmaceutical-industry>.
- Rahman, N, Marz, J & Akhter, S 2012, 'An ETL Metadata Model for Data Warehousing',Journal Of Computing & Information Technology, 20, 2, pp. 95-111, Applied Science & Technology Source, EBSCOhost, viewed 9 April 2015.
- SAP 2015, SAP Data Services 4.2, SAP, retrieved 19 April 2015, <http://help.sap.com/bods>
- Sigma 2015, Welcome to Sigma Pharmaceuticals Limited, Sigma, retrieved 21 April 2015, <http://www.sigmaco.com.au/>.
- Silvers, F 2008, Building and Maintaining a Data Warehouse, CRC Press, retrieved 10 April 2015, Google eBook.