Australian Journal of Basic and Applied Sciences, 9(20) June 2015, Pages: 608-617

ISSN:1991-8178

Australian Journal of Basic and Applied Sciences

Journal home page: www.ajbasweb.com

A Survey of Data Warehouse and ETL Processes

[1]R. Malarvannan and [2]S. Rajalakshmi

[1]*Research Scholar, Dept. of Computer Science and Engineering, SCSVSM University, Kanchipuram, Tamilnadu, INDIA*
[2]*Professor. & Head, Dept. of Computer Science and Engineering, SCSVSM University. Enathur, Kanchipuram. Tamilnadu, INIDA*

ARTICLE INFO

Article history:
Received 6 March 2015
Accepted 25 May 2015
Published 29 June 2015

Keywords:
ETL, Data warehouse, ETL Modelling, ETL Maintenance

ABSTRACT

Extraction, Transformation and Loading is responsible for the extraction of data, their cleaning, conforming and loading into the target database. ETL is a Critical layer in data warehouse setting. It is widely recognized that building ETL processes is very expensive regarding time, money and effort depending upon the size of data. Here, firstly we review commercial ETL tools and prototypes coming from academic world. After that we review designing works in ETL field and modelling ETL maintenance issues. Here review works in connection with optimization and incremental ETL, then finally challenges and research opportunities around ETL

To Cite This Article: R. Malarvannan and S. Rajalakshmi., A Survey of Data Warehouse and ETL Processes. ***Aust. J. Basic & Appl. Sci.,*** *9(20): 608-617, 2015*

INTRODUCTION

ETL (Extract-Transform-Load), which is the process of extracting data from a variety of heterogeneous data sources, and transforming those extracted data into needed format, and then loading those data into the DW (Data Warehouse). ETL processes contain three parts: Extraction, Transformation and Loading, each of which has its own metadata.

a. Extraction:

Data extraction is the process of capturing data source, that is to say, reading the data from all kinds of original operation systems and cleansing the data, which is the premise of all the work. If there are no related mapping rules and metadata.

b. Transformation:

Data transformation is the process of transforming above data by some prearrange rules, and dealing with some redundant, ambiguous, incomplete and anti-rules data to realize a unity of data granularity and data format. If we want to finish the data transformation from the source data storage format to the target data storage format, we have to know the information about source data and target data, which are also metadata.

c. Loading:

Data loading is the process of importing above data to DW system by all or by planned increment. In order to load transformed data to the DW system, we also need metadata about mapping rules.

From the above three processes, here see metadata plays an important role in ETL, whose mishandling can lead to the ineffectiveness of ETL processes directly. ETL processes often fails through its triviality and fallibility. The architecture of ETL is shown as Figure.1. The phases of extract, transform and load were executed in one single process. Under the framework of conventional ETL, the ETL process is defined: for different data source, develop and compile program or script; retrieval records from database; after extract, exchange the data according to users' requirement; load the data to target data warehouse; and process the records piece by piece until the end of source database. The framework of ETL is simple and would be easily implemented under the conventional architecture, but the weakness is obvious: The efficiency and reliability of load is lame which makes the overall scenario weak and difficult.

Modelling And Design Of Etl:

ETL are areas with high added value labelled costly and risky. In addition, software engineering requires that any project is doomed to switch to maintenance mode. For these reasons, it is essential to overcome the ETL modelling phase with elegance in order to produce simple models and understandable. This method is spread over four steps:

1. Identification of sources

Corresponding Author: R. Malarvannan, Research Scholar, Dept. of Computer Science and Engineering, SCSVSM University, Kanchipuram, Tamilnadu, INDIA
E-mail: malarresearchscholar@gmail.com

2. Distinction between candidates' sources and active sources.
3. Attributes mapping.
4. Annotation of diagram (conceptual model) with execution constraints.

Meta-data models based on ETL Design the designer needs to:
1. Analyse the structure and sources.
2. Describe mapping rules between sources and targets. The based on meta-model, provides a graphical notation to meet this necessitate.

Reliability:

The probability that the ETL process will perform its intended operation during specified time period under given conditions. Any reason for not performing the intended operation is considered to be a failure.

Maintainability:

The ability of an ETL process to be operated at the design cost and with service level agreements.

Freshness:

The ability of the system to provide the desired latency in updating the data warehouse.

Recoverability:

The ability to restore an ETL process to the point at which a failure occurred within a specified time window.

Scalability:

The ability of an ETL process to handle higher volumes of data.

Availability:

The probability that the ETL process is operational during a specific time period. Flexibility: The ability to accommodate previously unknown, new or changing requirements. Robustness: The ability of an ETL process to continue operating well or with minimal harm. Affordability: The ability to maintain or scale the cost of an ETL process appropriately. Consistency: The extent to which the data populating the data warehouse is correct and complete.

Traceability:

The ability of an ETL process to track the lineage of data and data changes.

Auditability:

The ability of an ETL process to protect data privacy and security, and to provide data and business rule transparency.

Literature Review:

When changes happen, analyzing the impact of change is mandatory to avoid errors and mitigate the risk of breaking existent treatments. As a consequence, without a helpful tool and an effective approach for change management, the cost of maintenance task will be high. Particularly for ETL processes, previously judged expensive and costly. Using ETL terminology, above previous research efforts focus on the target unlike the proposal of which focuses onchanges in the sources. In these proposal dealing with change management in ETL are interesting and offer a solution to detect changes impact on ETL processes. However change incorporation is not addressed.

Stackowiak *et al*. (2007) defined Business intelligence as the process of taking large amounts of data, analyzing that data, and presenting a high-level set of reports that condense the essence of that data into the basis of business actions, enabling management to make fundamental daily business decisions.

Cui *et al*. (2007) view BI as way and method of improving business performance by providing powerful assists for executive decision maker to enable them to have actionable information at hand. BI tools are seen as technology that enables the efficiency of business operation by providing an increased value to the enterprise information and hence the way this information is utilized.

Zeng *et al*. (2006) define BI as "The process of collection, treatment and diffusion of information that has an objective, the reduction of uncertainty in the making of all strategic decisions." Experts describe Business intelligence as a "business management term used to describe applications and technologies which are used to gather, provide access to analyze data and information about an enterprise, in order to help them make better informed business decisions."

Tvrdikova (2007) describes the basic characteristic for BI tool is that it is ability to collect data from heterogeneous source, to possess advance analytical methods, and the ability to support multi user‟s demands.

Zeng *et al*. (2006) categorized BI technology based on the method of information delivery; reporting, statistical analysis, ad-hoc analysis and predicative analysis.

The concept of Business Intelligence (BI) is brought up by Gartner Group since 1996. It is defined as the application of a set of methodologies and technologies, such as J2EE, DOTNET, Web Services, XML, data warehouse, OLAP, Data Mining, representation technologies, etc, to improve enterprise operation effectiveness, support management/decision to achieve competitive advantages.

Golfarelli *et al*. (2004) defined BI that includes effective data warehouse and also a reactive component capable of monitoring the time critical operational processes to allow tactical and

operational decision-makers to tune their actions according to the company strategy.

Gangadharan and Swamy (2004) define BI as the result of in-depth analysis of detailed business data, including database and application technologies, as well as analysis practices. They widen the definition of BI as technically much broader tools, that includes potentially encompassing knowledge management, enterprise resource planning, decision support systems and data mining.

Berson et.al (2002); Curt Hall (1999) BI includes several software for Extraction, Transformation and Loading (ETL), data warehousing, database query and reporting, OLAP, data analysis, data mining and visualization.

Radhakrishna and Sreekanth , proposed a web based framework model for representing the extraction of data from one or more data sources and use transformation business logic and load the data within the data ware house. This is a good starting point for gathering information in the existing documentation for the system and also research for ETL phase in web based scenario modeling in distributed environment a provide the effective decision results for various organization. The models of the entire ETL process using UML because these structural and dynamic properties of an information system at the conceptual level are more natural than the naïve approaches. It is more flexible and is used to support trading corporation, banks, financial and human resource management system of an organization at various levels. The future direction of this paper includes analyzing multimedia information sources automating mechanisms for ETL process.

Owen Kaser *et al*., "The Lito Project data ware houses with Literature "describes to apply the business intelligence techniques of data warehousing and OLAP to the domain of text processing. A literary data ware – house is the conventional corpus but its data stored and organized in multidimensional stages, in order to promote efficient end user queries. This work improves the query engine, ETC process and the user interfaces. The extract, transform, load stage retains the information which are build by the data ware house. We believe the overall idea of applying OLAP to literary data is promising. The initial custom engine is slow for production use but until more optimization is attempted, its promise is unclear.

Lior sapir *et al*., This paper "A methodology for the design of a fuzzy data warehouse" a data ware house is a special database used for storing business oriented information for future analysis and decision making. In business scenario, where some of the data or the business attributes are fuzzy, it may be useful to construct a ware house that can support the analysis of fuzzy data and also outlined the Kimball's methodology for the design of a data warehouse can be extended to the construction of a fuzzy data ware house. A case study demonstrates the visibility of the methodology most commonly used methodology today is Kimball's . It describes the process of translating business data and process into a dimensional model. It has several advantages, such as users can make more intuitive and easy to understand queries in a natural language. Defining fuzzy dimensional allows the user to describe the facts with abstract human concepts which are actually more realistic. The fuzzy dimensional also allow more flexible and interesting to filtering of the facts. We have demonstrated that fuzzy measures used with fuzzy aggregation operators allow the user to understand his business and the data warehouse measures better.

Daniel Fasel demonstrates the users a fuzzy data ware house approach to support the fuzzy analysis of the customer performance measurement. The potential of the fuzzy data ware house approach is illustrated using a concrete example of a customer performance measurement of a hearing instrument manufacture. A few for combining fuzzy concepts with the hierarchies of the data ware house have been proposed. A method of summary can be guaranteed using this approach and the data ware house concepts retained flexibility. Using a fuzzy approach in data ware house concepts improves information quality for the company. It provides broader possibilities to create indicators for customer performance measurement as in the example given of a hearing instrument manufacturer. The proposed approach does not include fuzzy linguistic concepts directly in to the hierarchical structure of dimension or into fact tables of the data ware house model and also explains how the fuzzy concepts can be aggregated over dimensions without having to redefined the fuzzy sets in every degree of granularity. Visualization should provide easily understand the results for fuzzy queries in the fuzzy data ware house

Christ Sophie *et al*., focus that in the field of human resources there is a growing trend towards moving from activity based functions to a more strategic, business oriented role. The data mart defines on the HR information needs is the best solution to meet the objectives. The main purpose of this paper is to explain how the SAS system can be used in top of SAS R/3 HR, and obtained real business benefits on a very short time. It is also based on the practical experience at the Belgain Gas and electricity provider. The structure of this paper first explained the business of the short comings and discussed the business objectives for the data mart. Finally this paper explains the project approach and focuses on the specific attention points when building a data mart. It provides end to end solution and the data management facilities possible to deliver quick result to the end-users.

Kari Richardon and Eric Rossland describes the hands-on work shop will give users a basic tour through the functionality of SAS ETL studio health to build a small data mart. The participants in this

workshop will use SAS ETL studio to define necessary library definitions also source and target table definitions. Participants will create a process flow diagram using a simple transformation and load the target table. In the last step, participants will create 2 reports using target table. Finally, this hands-on workshop provides an overview of SAS ETL studio and how it can be used to create a data mart.

D. Ashok Kumar and M.C. Loraine explained modern electronic health records are designed to capture and render vast quantities of clinical data during the health care prone. Utilization of data analysis and data mining methods medicine and health care is sparse. Medical data is one of the heavily and categorical type data. A Dichotomous variable is type of categorical which is binary with categorical zero and one. Binary data are the simplest form of data used for medical database in which close ended questions can be used. It is very efficient based on computations efficiency and memory capacity to represent categorical type data. Data mining technique called clustering is involved here for capacity to represent categorical type data. Data mining techniques called clustering is involved here for dichotomous medical data due to its high dimensional and data scarcity. Usually the binary data clustering is done by using 0 and 1 as numerical value. The clustering is performed after transforming the binary data into real by wiener transformation. The proposed algorithm in this paper can be usable for large medical and health binary data bases for determining the correction are the health disorders and symptoms observed.

S. Vikram Phaneendra & E. Madhusudhan Reddy et.al. Illustrated that in olden days the data was less and easily handled by RDBMS but recently it is difficult to handle huge data through RDBMS tools, which is preferred as "big data". In this they told that big data differs from other data in 5 dimensions such as volume, velocity, variety, value and complexity. They illustrated the hadoop architecture consisting of name node, data node, edge node, HDFS to handle big data systems. Hadoop architecture handle large data sets, scalable algorithm does log management application of big data can be found out in financial, retail industry, health-care, mobility, insurance. The authors also focused on the challenges that need to be faced by enterprises when handling big data: - data privacy, search analysis, etc.

Kiran kumara Reddi & Dnvsl Indira et.al. Enhanced us with the knowledge that Big Data is combination of structured , semi-structured ,unstructured homogenous and heterogeneous data .The author suggested to use nice model to handle transfer of huge amount of data over the network .Under this model, these transfers are relegated to low demand periods where there is ample ,idle bandwidth available . This bandwidth can then be repurposed for big data transmission without impacting other users in system. The Nice model uses a store –andforward approach by utilizing staging servers. The model is able to accommodate differences in time zones and variations in bandwidth. They suggested that new algorithms are required to transfer big data and to solve issues like security, compression, routing algorithms.

Jimmy Lin et.al. used Hadoop which is currently the large –scale data analysis " hammer" of choice, but there exists classes of algorithms that aren't " nails" in the sense that they are not particularly amenable to the MapReduce programming model . He focuses on the simple solution to find alternative non-iterative algorithms that solves the same problem. The standard MapReduce is well known and described in many places .Each iteration of the pagerank corresponds to the MapReduce job. The author suggested iterative graph, gradient descent & EM iteration which is typically implemented as Hadoop job with driven set up iteration &Check for convergences. The author suggests that if all you have is a hammer, throw away everything that's not a nail.

Wei Fan & Albert Bifet et.al. Introduced Big Data Mining as the capability of extracting Useful information from these large datasets or streams of data that due to its Volume, variability and velocity it was not possible before to do it. The author also started that there are certain controversy about Big Data. There certain tools for processes. Big Data as such hadoop, strom, apache S4. Specific tools for big graph mining were PEGASUS & Graph. There are certain Challenges that need to death with as such compression, visualization etc.

Albert Bifet et.al. Stated that streaming data analysis in real time is becoming the fastest and most efficient way to obtain useful knowledge, allowing organizations to react quickly when problem appear or detect to improve performance. Huge amount of data is created everyday termed as " big data". The tools used for mining big data are apache hadoop, apache big, cascading, scribe, storm, apache hbase, apache mahout, MOA, R, etc. Thus, he instructed that our ability to handle many exabytes of data mainly dependent on existence of rich variety dataset, technique, software framework.

Bernice Purcell et.al. Started that Big Data is comprised of large data sets that can't be handle by traditional systems. Big data includes structured data, semi-structured and unstructured data. The data storage technique used for big data includes multiple clustered network attached storage (NAS) and object based storage. The Hadoop architecture is used to process unstructured and semi-structured using map reduce to locate all relevant data then select only the data directly answering the query. The advent of Big Data has posed opportunities as well challenges to business.

Sameer Agarwal et.al. Presents a BlinkDB, a approximate query engine for running interactive

SQL queries on large volume of data which is massively parallel. BlinkDB uses two key ideas: (1) an adaptive optimization framework that builds and maintains a set of multi-dimensional stratified samples from original data over time, and (2) A dynamic sample selection strategy that selects an appropriately sized sample based on a query's accuracy or response time requirements.

Yingyi Bu et.al. Used a new technique called as HaLoop which is modified version of Hadoop MapReduce Framework, as Map Reduce lacks built-in-support for iterative programs HaLoop allows iterative applications to be assembled from existing Hadoop programs without modification, and significantly improves their efficiency by providing interiteration caching mechanisms and a loop-aware scheduler to exploit these caches. He presents the design, implementation, and evaluation of HaLoop, a novel parallel and distributed system that supports large-scale iterative data analysis applications. HaLoop is built on top of Hadoop and extends it with a new programming model and several important optimizations that include (1) a loop-aware task scheduler, (2) loop-invariant data caching, and (3) caching for efficient fix point verification.

Shadi Ibrahim et.al. Project says presence of partitioning skew1 causes a huge amount of data transfer during the shuffle phase and leads to significant unfairness on the reduce input among different data nodes In this paper, author develop a novel algorithm named LEEN for locality aware and fairnessaware key partitioning in MapReduce. LEEN embraces an asynchronous map and reduce scheme. Author has integrated LEEN into Hadoop. His experiments demonstrate that LEEN can efficiently achieve higher locality and reduce the amount of shuffled data. More importantly, LEEN guarantees fair distribution of the reduce inputs. As a result, LEEN achieves a performance improvement of up to 45% on different workloads. To tackle all this he presents a present a technique for Handling Partitioning Skew in MapReduce using LEEN.

Kenn Slagter et.al. Proposes an improved partitioning algorithm that improves load balancing and memory consumption. This is done via an improved sampling algorithm and partitioner. To evaluate the proposed algorithm, its performance was compared against a state of the art partitioning mechanism employed by Tera Sort as the performance of MapReduce strongly depends on how evenly it distributes this workload. This can be a challenge, especially in the advent of data skew. In MapReduce, workload distribution depends on the algorithm that partitions the data. One way to avoid problems inherent from data skew is to use data sampling. How evenly the partitioner distributes the data depends on how large and representative the sample is and on how well the samples are analyzed by the partitioning mechanism. He uses an improved partitioning mechanism for optimizing massive data analysis using MapReduce for evenly distribution of workload.

Ahmed Eldawy et.al. presents the first full-fledged MapReduce framework with native support for spatial data that is spatial data Spatial Hadoop pushes its spatial constructs in all layers of Hadoop, namely, language, storage, MapReduce and operations layers. In the language layer, a simple high level language is provided to simplify spatial data analysis for nontechnical users. In the storage layer, a two-layered spatial index structure is provided where the global index partitions data across nodes while the local index organizes data in each node. This structure is used to build a grid index, an R-tree or an R+- tree. Spatial-Hadoop is a comprehensive extension to Hadoop that pushes spatial data inside the core functionality of Hadoop. Spatial Hadoop runs existing Hadoop programs as is, yet, it achieves order(s) of magnitude better performance than Hadoop when dealing with spatial data. SpatialHadoop employs a simple spatial high level language, a two-level spatial index structure, basic spatial components built inside the MapReduce layer, and three basic spatial operations: range queries, k-NN queries, and spatial join. Author presents an efficient MapReduce framework for Spatial Data.

Jeffrey Dean et.al. Implementation of MapReduce runs on a large cluster of commodity machines and is highly scalable: a typical MapReduce computation processes many terabytes of data on thousands of machines. Programmers and the system easy to use: hundreds of MapReduce programs have been implemented and upwards of one thousand MapReduce jobs are executed on Google's clusters every day. Programs written in this functional style are automatically parallelized and executed on a large cluster of commodity machines. The runtime system takes care of the details of partitioning the input data, scheduling the program's execution across a set of machines, handling machine failures, and managing the required inter-machine Communication. This allows programmers without any experience with parallel and distributed systems to easily utilize the resources of a large distributed system. Author proposes Simplified Data Processing on Large Clusters.

Chris Jermaine et.al. Proposes an Online Aggregation for Large-Scale Computing. Given the potential for OLA to be newly relevant, and given the current interest on very largescale, data-oriented computing, in this paper we consider the problem of providing OLA in a shared-nothing environment. While we concentrate on implementing OLA on top of a MapReduce engine, many of author's most basic project contributions are not specific to MapReduce, and should apply broadly. Consider how online aggregation can be built into a MapReduce system for large-scale data processing. Given the MapReduce paradigm's close relationship with cloud

computing (in that one might expect a large fraction of MapReduce jobs to be run in the cloud), online aggregation is a very attractive technology. Since large-scale cloud computations are typically pay-as-you-go, a user can monitor the accuracy obtained in an online fashion, and then save money by killing the computation early once sufficient accuracy has been obtained.

Tyson Condie et.al. propose a modified MapReduce architecture in which intermediate data is pipelined between operators, while preserving the programming interfaces and fault tolerance models of other MapReduce frameworks. To validate this design, author developed the Hadoop Online Prototype (HOP), a pipelining version of Hadoop. Pipelining provides several important advantages to a MapReduce framework, but also raises new design challenges. To simplify fault tolerance, the output of each MapReduce task and job is materialized to disk before it is consumed. In this demonstration, we describe a modified MapReduce architecture that allows data to be pipelined between operators. This extends the MapReduce programming model beyond batch processing, and can reduce completion times and improve system utilization for batch jobs as well. We demonstrate a modified version of the Hadoop MapReduce framework that supports online aggregation, which allows users to see "early returns" from a job as it is being computed. Our Hadoop Online Prototype (HOP) also supports continuous queries, which enable MapReduce programs to be written for applications such as event monitoring and stream processing.

Kyong-Ha Lee Hyunsik Choi et.al. Proposes a prominent parallel data processing tool MapReduce survey intends to assist the database and open source communities in understanding various technical aspects of the MapReduce framework. In this survey, we characterize the MapReduce framework and discuss its inherent pros and cons. We then introduce its optimization strategies reported in the recent literature. author also discuss the open issues and challenges raised on parallel data analysis with MapReduce. Chen He Ying Lu David Swanson et.al develops a new MapReduce scheduling technique to enhance map task's data locality. He has integrated this technique into Hadoop default FIFO scheduler and Hadoop fair scheduler. To evaluate his technique, he compares not only MapReduce scheduling algorithms with and without his technique but also with an existing data locality enhancement technique (i.e., the delay algorithm developed by Facebook). Experimental results show that his technique often leads to the highest data locality rate and the lowest response time for map tasks. Furthermore, unlike the delay algorithm, it does not require an intricate parameter tuning process.

Jonathan Paul Olmsted et.al. Derive the necessary results to apply variation Bayesian inference to the ideal point model. This deterministic, approximate solution is shown to produce comparable results to those from standard estimation strategies. However, unlike these other estimation approaches, solving for the (approximate) posterior distribution is rapid and easily scales to 'big data'. Inferences from the variation Bayesian approach to ideal point estimation are shown to be equivalent to standard approaches on modestly-sized roll call matrices from recent sessions of the US Congress. Then, the ability of variation inference to scale to big data is demonstrated and contrasted with the performance of standard approaches.

Jonathan Stuart Ward et.al. did a survey of Big data definition, Anecdotally big data is predominantly associated with two ideas: data storage and data analysis. Despite the sudden Interest in big data, these concepts are far from new and have long lineages. This, therefore, raises the question as to how big data is notably different from conventional data processing techniques. For rudimentary insight as to the answer to this question one need look no further than the term big data. \Big" implies significance, complexity and challenge. Unfortunately the term\big" also invites quantification and therein lies the difficulty in furnishing a definition. The lack of a consistent definition introduces ambiguity and hampers discourse relating to big data. This short paper attempts to collate the various definitions which have gained some degree of traction and to furnish a clear and concise definition of an otherwise ambiguous term.

Albert Bifet et.al. Discuss the current and future trends of mining evolving data streams, and the challenges that the field will have to overcome during the next years. Data stream real time analytics are needed to manage the data currently generated, at an ever increasing rate, from such applications as: sensor networks, measurements in network monitoring and traffic management, log records or click-streams in web exploring, manufacturing processes, call detail records, email, blogging, twitter posts and others. In fact, all data generated can be considered as streaming data or as a snapshot of streaming data, since it is obtained from an interval of time. Streaming data analysis in real time is becoming the fastest and most efficient way to obtain useful knowledge from what is happening now, allowing organizations to react quickly when problems appear or to detect new trends helping to improve their performance. Evolving data streams are contributing to the growth of data created over the last few years. We are creating the same quantity of data every two days, as we created from the dawn of time up until 2003. Evolving data streams methods are becoming a low-cost, green methodology for real time online prediction and analysis.

Mrigank Mridul, Akashdeep Khajuria, Snehasish Dutta, Kumar N. et.al did the analysis of big data he stated that Data is generated through many sources

like business processes, transactions, social networking sites, web servers, etc. and remains in structured as well as unstructured form . Today's business applications are having enterprise features like large scale, data-intensive, web-oriented and accessed from diverse devices including mobile devices. Processing or analyzing the huge amount of data or extracting meaningful information is a challenging task. The term "Big data" is used for large data sets whose size is beyond the ability of commonly used software tools to capture, manage, and process the data within a tolerable elapsed time. Big data sizes are a constantly moving target currently ranging from a few dozen terabytes to many peta bytes of data in a single data set. Difficulties include capture, storage, search, sharing, analytics and visualizing. Typical examples of big data found in current scenario includes web logs, RFID generated data, sensor networks, satellite and geo-spatial data, social data from social networks, Internet text and documents, Internet search indexing, call detail records, astronomy, atmospheric science, genomics, biogeochemical, biological, and other complex and/or interdisciplinary scientific project, military. Surveillance, medical records, photography archives, video archives, and large-scale ecommerce.

Etl Tools - General Information:

ETL tools are designed to save time and money by eliminating the need of 'hand-coding' when a new data warehouse is developed. They are also used to facilitate the work of the database administrators who connect different branches of databases as well as integrate or change the existing databases.

* The main purpose of the ETL tool is: extraction of the data from legacy sources (usually heterogenous)
* data transformation (data optimized for transaction --> data optimized for analysis)
* synchronization and cleansing of the data
* loading the data into data warehouse.

There are several requirements that must be had by ETL tools in order to deliver an optimal value to users, supporting a full range of possible scenarios. Those are:

- data delivery and transformation capabilities
- data and metadata modelling capabilities
- data source and target support
- data governance capability
- runtime platform capabilities
- operations and administration capabilities
- service-enablements capability.

Etl Tools Comparison Criteria:

The research presented in this article is based on Gartner's data integration magic quadrant, forrester researches and our professional experience. The etltools.org portal is not affiliated with any of the companies listed below in the comparison.

The research inclusion and exclusion criteria are as follows:

- range and mode of connectivity/adapter support
- data transformation and delivery modes support
- metadata and data modelling support
- design, development and data governance support
- runtime platform support
- enablement of service and three additional requirements for vendors:
- support of customers in not less than two major geographic regions
- have customer implementations at cross departmental and multiproject level.

Etl Tools Comparison:

The information provided below lists major strengths and weaknesses of the most popular ETL vendors.

IBM (INFORMATION SERVER INFOSPHERE PLATFORM)

* Advantages: strongest vision on the market, flexibility
* progress towards common metadata platform
* high level of satisfaction from clients and a variety of initiatives
* Disadvantages: difficult learning curve
* long implementation cycles
* became very heavy (lots of GBs) with version 8.x and requires a lot of processing power

Informatica Powercenter:

* Advantages: most substantial size and resources on the market of data integration tools vendors
* consistent track record, solid technology, straightforward learning curve, ability to address real-time data integration schemes
* Informatica is highly specialized in ETL and Data Integration and focuses on those topics, not on BI as a whole
* focus on B2B data exchange
* Disadvantages: several partnerships diminishing the value of technologies
* limited experience in the field.

Microsoft (Sql Server Integration Services):

* Advantages: broad documentation and support, best practices to data warehouses
* ease and speed of implementation
* standardized data integration
* real-time, message-based capabilities
* relatively low cost - excellent support and distribution model
* Disadvantages: problems in non-Windows environments. Takes over all Microsoft Windows limitations.
* unclear vision and strategy

Oracle (Owb And Odi):

* Advantages: based on Oracle Warehouse Builder and Oracle Data Integrator – two very powerful tools;
* tight connection to all Oracle data warehousing applications;
* tendency to integrate all tools into one application and one environment.
* Disadvantages: focus on ETL solutions, rather than in an open context of data management;
* tools are used mostly for batch-oriented work, transformation rather than real-time processes or federation data delivery;
* long-awaited bond between OWB and ODI brought only promises - customers confused in the functionality area and the future is uncertain

Sap Business Objects (Data Integrator / Data Services):

* Advantages: integration with SAP
* 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
* Disadvantages: SAP Business Objects is seen as two different companies
* Uncertain future. Controversy over deciding which method of delivering data integration to use (SAP BW or BODI).
* Business Objects Data Integrator (Data Services) may not be seen as a stand-alone capable application to some organizations.

SAS:

* Advantages: 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
* great support for the business-class companies as well for those medium and minor ones
* Disadvantages: misplaced sales force, company is not well recognized
* SAS has to extend influences to reach non-BI community
* Costly

Sun Microsystems:

* Advantages: Data integration tools are a part of huge Java Composite Application Platform Suite - very flexible with ongoing development of the products
* 'Single-view' services draw together data from variety of sources; small set of vendors with a strong vision
* Disadvantages: relative weakness in bulk data movement
* limited mindshare in the market
* support and services rated below adequate

Sybase:

* Advantages: assembled a range of capabilities to be able to address a mulitude of data delivery styles
* size and global presence of Sybase create opportunities in the market
* pragmatic near-term strategy - better of current market demand
* broad partnerships with other data quality and data integration tools vendors
* Disadvantages: falls behind market leaders and large vendors
* gaps in many aspects of data management

Syncsort:

* Advantages: functionality; well-known brand on the market (40 years experience); loyal customer and experience base;
* easy implementation, strong performance, targeted functionality and lower costs
* Disadvantages: struggle with gaining mind share in the market
* lack of support for other than ETL delivery styles
* unsatisfactory with lack of capability of professional services

Tibco Software:

* Advantages: message-oriented application integration; capabilities based on common SOA structures;
* support for federated views; easy implementation, support and performance
* Disadvantages: scarce references from customers; not widely enough recognised for data integration competencies
* lacking in data quality capabilities.

Eti:

* Advantages: proven and mature code-generating architecture
* one of the earliest vendors on the data integration market; support for SOA service-oriented deployments;
* successfully deals with large data volumes and a high degree of complexity, extension of the range of data platforms and data sources;
* customers' positive responses to ETI technology
* Disadvantages: relatively slow growth of customer base
* rather not attractive and inventive technology.

Iway Software:

* Advantages: offers physical data movement and delivery; support of wide range of adapters and access to numerous sources;

* well integrated, standard tools;
* reasonable ease of implementation effort
* Disadvantages: gaps in specific capabilities
* relatively costly - not competitive versus market leaders

Pervasive Software:

* Advantages: many customers, years of experience, solid applications and support;
* good use of metadata
* upgrade from older versions into newer is straightforward.
* Disadvantages: inconsistency in defining the target for their applications;
* no federation capability;
* limited presence due to poor marketing.

Open Text:

* Advantages Simplicity of use in less-structured sources
* Easy licensing for business solutions
* cooperates with a wide range of sources and targets
* increasingly high functionality
* Disadvantages: limited federation, replication and data quality support; rare upgrades due to its simplicity;
* weak real-time support due to use third party solutions and other database utilities.

Pitney Bowes Software:

* Advantages: Data Flow concentrates on data integrity and quality;
* supports mainly ETL patterns; can be used for other purposes too;
* ease of use, fast implementation, specific ETL functionality.
* Disadvantages: rare competition with other major companies, repeated rebranding trigger suspicions among customers.
* narrow vision of possibilities even though Data Flow comes with variety of applications.
* weak support, unexperienced service.

Conclusion:

As the role of enterprises becomes increasing real-time such as real-time Business Intelligence will be increasing important to such companies. In the traditional ETL approach, the most current data is not available. With the increase demands by businesses for real-time Business Intelligence and Predictive Analytics, there is a need to build ETL tools, which provide real-time data into Data Warehousing. Not every analysis task warrants real-time analysis. The trade-off between the overhead of providing real-time Business Intelligence and Data Warehousing, and the need for such an analysis calls for serious research and thought. Otherwise, or the resulting system may have forbidden costs associated with it (Agrawal, 2009). The underlying technology components and custom solutions are excessively expensive. The importance, complexity and criticality of such an environment make real-time BI and DW a significant topic of research and practice; therefore, these issues need to be addressed in the future by both the industry and the academia (Vassiliadis, Simitsis, 2008).

REFERENCE

Zhang, X.F., W.W. Sun, W. Wang, 2006. Generating Incremental ETL Processes Automatically. Computer and Computational Sciences, 516-521.

Zhang Zhongping and Zhao Ruizhen, 2006. Design of architecture for ETL based on metedata-driven, Computer Applications and Software, 26: 61-63

Sun Wei and Zhang Zhongneng, 2005. ―ETL Architecture Research Ⅱ. Micro-computer Application, 21(3): 13-15.

Zhao Xiaofei and Huang Zhiqiu, 2006. ―A Formal Framework for Reasoning on Metadata Based on CWM, Ⅱ The 25th International Conference on Conceptual Modeling, 371-384.

Herzog, T.N. and F. Scheuren and W.E. Winkler, 2007. Data Quality and Record Linkage Techniques.Springer, Heidelberg.

Gomes, P. and J. Farinha and M.J. Trigueiros, 2007. A Data Quality Metamodel Extension to CWM. In 4th Asia Pacific Conference on Conceptual Modelling Proceedings, pages, 17–26. APCCM.

Labio, W., J. Yang, Y. Cui, H. Garcia-Molina and J. Widom, 2000. ―Performance Issues in Incremental Warehouse Maintenance‖, International Conference on Very Large Data Bases (VLDB).

Yang, J. and J. Widom, 2001. ―Temporal View Self-Maintenance‖, 7th Int. Conf. Extending Database Technology (EDBT).

Lujan-Mora, S. and J. Trujillo, 2005. Physical modeling of data warehouses using UML. In I. Song and K. Davis, editors, Proceedings of the 7th ACM International Workshop on Data Warehousing and OLAP,DOLAP'04, pages 48{57, Washington, D.C., USA, ACM Press.

Giga Information Group, 2002. Market Overview Update:ETL. Technical Report RPA-032002-00021.

Adzic, J., V. Fiore, 2003. Data Warehouse Population Platform,in: Proceedings of the Fifth International Workshop on the Design and Management of Data Warehouses (DMDW‟03), Berlin, Germany.

Yang, J., 2001. ―Temporal Data Warehousing‖, Ph.D. Thesis, Dept. Computer Science, Stanford University.

Yang, J. and J. Widom, 2001. ―Incremental Computation and Maintenance of Temporal

Aggregates‖, 17th Intern. Conference on Data Engineering (ICDE).

Bouzeghoub, M., F. Fabret and M. Matulovic, 1999. ―Modeling Data Warehouse Refreshment Process as a Workflow Application‖, Intern. Workshop on Design and Management of Data Warehouses (DMDW)

Theodoratus, D. and M. Bouzeghoub, 1999. ―Data Currency Quality Factors in Data Warehouse Design‖, International Workshop on the Design and Management of Data Warehouses (DMDW)

Inmon, W., D. Strauss and G. Neushloss, 2007. "DW 2.0 The Architecture for the next generation of data warehousing",Morgan Kaufman.

Simitisis, A., P. Vassiliadis, S. Skiadopoulos and T. Sellis, 2007. "DataWarehouse Refreshment", Data Warehouses and OLAP:Concepts, Architectures and Solutions, IRM Press, 111-134.

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

Kabiri, A., F. Wadjinny and D. Chiadmi, 2011. "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, 241: 146-160.

Vassiliadis, P. and A. Simitsis, 2009. "EXTRACTION, TRANSFORMATION, AND LOADING",

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

Adzic, J., V. Fiore and L. Sisto, 2007. "Extraction, Transformation, and Loading Processes", Data Warehouses and OLAP:Concepts, Architectures and Solutions, IRM Press, 88-110.

Eckerson, W. and C. White, 2003. "Evaluating ETL and Data Integration Platforms", TDWI REPORT SERIES,101communications LLC.

IBM InfoSphereDataStage, http://www-01.ibm.com/software/data/infosphere/datastage/

Vikram Phaneendra, S., E. Madhusudhan Reddy, 2013. "Big Data- solutions for RDBMS problems- A survey" In 12th IEEE/IFIP Network Operations & Management Symposium (NOMS 2010) (Osaka, Japan, Apr 19.

Kiran kumara Reddi, Dnvsl Indira, 2013. "Different Technique to Transfer Big Data : survey" IEEE Transactions on, 52(8): 2348-2355.

Jimmy Lin, 2013. "MapReduce Is Good Enough?" The control project. IEEE Computer, 32.

Umasri., M.L., 2014. Shyamalagowri.D ,Suresh Kumar.S "Mining Big Data:- Current status and forecast to the future", 4(1): 2277-128X.

Albert Bifet, 2012. "Mining Big Data In Real Time" Informatica, 37: 15–20.

Bernice Purcell, 2013. "The emergence of "big data" technology and analytics" Journal of Technology Research.

Sameer Agarwal, Barzan MozafariX, Aurojit Panda, Henry Milner, Samuel MaddenX, Ion Stoica 2013."BlinkDB: Queries with Bounded Errors and Bounded Response Times on Very Large Data" Copyright © ì ACM 978-1-4503-1994 2/13/04

Yingyi Bu _ Bill Howe _ Magdalena Balazinska _ Michael D. Ernst, 2010. "The HaLoop Approach to Large-Scale Iterative Data Analysis" VLDB paper "HaLoop: Efficient Iterative Data Processing on Large Clusters.

Shadi Ibrahim, Hai Jin, Lu Lu, 2008. "Handling Partitioning Skew in MapReduce using LEEN" ACM, 51: 107–113.

Kenn Slagter, Ching-Hsien Hsu, 2013. "An improved partitioning mechanism for optimizing massive data analysis using MapReduce" Published online: © Springer Science+Business Media New York.

Ahmed Eldawy, Mohamed F. Mokbel, 2013. "A Demonstration of SpatialHadoop:An Efficient MapReduce Framework for Spatial Data" Proceedings of the VLDB Endowment, 6. 12 Copyright VLDB Endowment 21508097/13/10.

Jeffrey Dean and Sanjay Ghemawat, 2010. "MapReduce: Simplified Data Processing on Large Clusters" OSDI.

Niketan Pansare1, Vinayak Borkar2, Chris Jermaine1, Tyson Condie, 2011. "Online Aggregation for Large MapReduce Jobs" August 29September 3, 2011, Seattle, WA Copyright VLDB Endowment, ACM.

Tyson Condie, Neil Conway, Peter Alvaro, Joseph M. Hellerstein, 2010. "Online Aggregation and Continuous Query support in MapReduce" SIGMOD'10, 6–11, Indianapolis, Indiana, USA. Copyright ACM 978-1-4503-0032- 2/10/06.

Jonathan Paul Olmsted, 2014. "Scaling at Scale: Ideal Point Estimation with 'Big-Data" Princeton Institute for Computational Science and Engineering.

Jonathan Stuart Ward and Adam Barker, 2012. "Undefined By Data: A Survey of Big Data Definitions" Stamford, CT: Gartner.

Balaji Palanisamy, 2010. Member, IEEE, Aameek Singh, Member, IEEE Ling Liu, Senior Member, IEEE" Cost-effective Resource Provisioning for MapReduce in a Cloud"gartner report, 25.

Mrigank Mridul, Akashdeep Khajuria, Snehasish Dutta, N. Kumar, 2014. " Analysis of Bidgata using Apache Hadoop and Map Reduce" 4(5): 27.

Kyong-Ha Lee Hyunsik Choi, 2011. "Parallel Data Processing with MapReduce: A Survey" SIGMOD Record, 40-4.

Chen He Ying Lu David Swanson, 2010. "Matchmaking: A New MapReduce Scheduling" in 10th IEEE International Conference on Computer and Information Technology (CIT'10), 2736–2743.