Data Mining Techniques in E-Commerce

Abdugofur Temirov[1], **Ren Dongxiao**[2]

School of Science, Zhejiang University of Science and Technology, Hangzhou 310023, Zhejiang, China

Abstract: ***This paper discusses the important role of business based on Data-Mining knowledge development to detection the relation of Data-Mining and E-commerce. Moreover, some applications, benefits and challenges in this case. E-commerce processes and Data-Mining tools have revolutionized many companies. Data-Mining is a form of business intelligence and data analysis. It is the process of analyzing data to draw useful conclusions or predictions from it. It is technique frequently adopted by large-scale E-commerce businesses to aid with marketing and product development.***

Keywords: Data-Mining (DM), E-commerce.

1. Introduction

With the development of economic globalization and trade liberalization, the emerging technology of computer networks has gradually penetrated into various aspects of everyone's life, and the e-commerce industry is generated as a platform. As a kind of new business model, e-commerce has changed people's opinion of the traditional commerce and trade, the business philosophy and the method of payments, and it has injected fresh blood into today's business community and brought a revolutionary technical impact to the traditional business model. Data mining as a kind of technology about data analyzing and finishing is in urgent need in e-commerce. It will processes and analysis a large amounts of information on the Internet for enterprises in the normal e-commerce trade more effectively, and it will provide business model ,marketing strategy and decision-making enterprises in the future as a more accurate technical and information support [5].

Data mining, extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were time-consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations. Most companies already collect and refine massive quantities of data. Data mining techniques can be implemented rapidly on existing software and hardware platforms to enhance the value of existing information resources, and can be integrated with new products and systems as they are brought on-line.

Data must be organized by data base tools and data warehouses, and then it needs an instrument for knowledge discovery. Data mining can be defined as the art of extracting non-obvious, useful information from large databases. This emerging field brings a set of powerful techniques which are relevance for companies to focus their efforts in taking advantage of their data [8].Data mining tools generate new information for decision makers from very large databases. The various mechanisms of this generation include abstractions, aggregations, summarizations and characterizations of data These forms are the result of applying sophisticated modeling techniques from the diverse fields of statistics, artificial intelligence, and database management and computer graphics [7].

E-commerce has changed the face of most business functions in competitive enterprises. Internet technologies have seamlessly automated interface processes between customers and retailers, retailers and distributors, distributors and factories, and factories and their myriad suppliers. In general, e-commerce and e-business have enabled on-line transactions. Also, generating large-scale real-time data has never been easy. With data pertaining to various views of business transactions being readily available, it is only apposite to seek the services of data mining to make (business) sense out of these data sets [4].

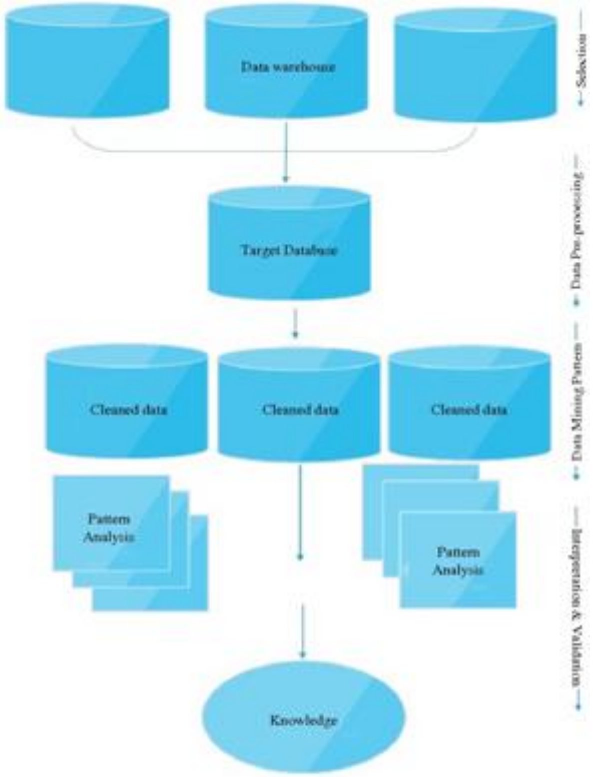
The success of a DM exercise is driven to a very large extent by the following factors:

- *Availability of data with rich descriptions*: This means that unless the relations captured in the database are of high degree, extracting hidden patterns and relationships among the various attributes will not make any practical sense.
- *Availability of a large volume of data*: This is mostly mandated for statistical significance of the *rules* to hold. Absence of say, at least a hundred thousand transactions will most likely reduce the usefulness of the rules generated from the transactional database.
- *Reliability of the data available*: Although a given terabyte database may have hundreds of attributes per relation, the DM algorithms run on this dataset may be rendered defunct if the data itself was generated by manual and error prone means and wrong default values were set.
- *Ease of quantification of the return on investment (ROI) in DM*: Although the earlier three factors may be favorable, unless a strong business case can be easily made, investments in the next level DM efforts may not be possible. In other words, the utility of the DM exercise needs to be quantified the domain of application.
- *Ease of interfacing with legacy systems*: It is commonplace to find large organizations run on several legacy systems that generate huge volumes of data. A DM

exercise which is usually preceded by other exercises like extract, transformation and loading (ETL), data filtering etc, should not add more overheads to system integration.

Data-Mining in E-Commerce

Data mining in E-commerce is a vital way of repositioning the e-commerce company for supporting the enterprise with the required information concerning the business. Recently, most companies adopt e-commerce and being in possession of big data in their data repositories. The only way to get the most out of this data is to mine it to increase decision making or to enable business intelligence. In e-commerce data mining there are three important processes that data must pass before turning into knowledge or application.



Data Mining in E-commerce

In general data mining process iterates from the following five basic steps:

1) *Data selection:* This step is all about identifying the kind of data to be mined, the goals for it and the necessary tool to enable the process. At the end of it the right input attributes and output information in order to represent the task are chosen.
2) *Data transformation:* This step is all about organizing the data based on the requirements by removing noise, converting one type of data to another, normalizing the data if there is need to, and also defining the strategy to handle the missing data.
3) *Data mining step per sec:* Having mined the transformed data using any of the techniques to extract pattern of interest, the miner can also make data mining method by performing the proceeding steps correctly.
4) *Result interpretation and validation:* For better understanding of data and it synthesized knowledge together with its validity span, the robustness is check by data mining application test. The information retrieved can also be evaluated by comparing it with the earlier expertise in the application domain.
5) *Incorporation of the discovered knowledge:* This has to do with presenting the result of discovered knowledge to decision maker so that it is possible to compare or check/resolve for conflict with an earlier extracted knowledge where a new discovered pattern can be applied [16].

The technique that is used to perform these feats in data mining is called modeling. Modeling is simply the act of building a model in one situation where it is known the answer and then applying it to another situation that is unknown. This act of model building is thus something that people have been doing for a long time, certainly before the advent of computers or data mining technology. What happens on computers, however, is not much different than the way people build models. Computers are loaded up with lots of information about a variety of situations where an answer is known and then the data mining software on the computer must run through that data and distil the characteristics of the data that should go into the model.

The goal in prospecting is to make some calculated guesses about the information in the lower right hand quadrant based on the model that it builds going from Customer General Information to Customer Proprietary Information.

Data-Mining for prospecting

	Customers	Prospects
General information (e.g. demographic data)	Known	Known
Proprietary information (e.g. Customer transactions)	Known	Target

Test marketing is an excellent source of data for this kind of modeling. Mining the results of a test market representing a broad but relatively small sample of prospects can provide a foundation for identifying good prospects in the overall market.

Data-Mining for predictions

	Yesterday	Today	Tomorrow
Static information and current plans (e.g. demographic data, marketing plans)	Known	Known	Known
Dynamic information (e.g. Customer transactions)	Known	Known	Target

Techniques used in Data Mining

The most commonly used techniques in data mining are:

1) *Artificial neural networks:* Non-linear predictive models that learn through training and resemble biological neural networks in structure.
2) *Decision trees:* Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi
3) *Genetic algorithms:* Optimization techniques that use processes such as genetic combination, mutation, and natural selection in a design based on the concepts of evolution.
4) *Nearest neighbor method:* A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical

dataset. Sometimes called the k-nearest neighbor technique.

5) *Rule induction:* The extraction of useful if-then rules from data based on statistical significance.

Many of these technologies have been in use for more than a decade in specialized analysis tools that work with relatively small volumes of data. These capabilities are now evolving to integrate directly with industry-standard data warehouse and OLAP platforms.

Benefits of Data Mining in E-Commerce

Application of data mining in e-commerce refers to possible areas in the field of e-commerce where data mining can be utilized for the purpose of enhancements in business. It is well known while visiting an online store for shopping, users normally leave behind certain facts that companies can store in their database. These facts represent unstructured or structured data that can be mined to provide a competitive advantage to the company. The following areas are where data mining can be applied in the field of e-commerce for the benefits of companies:

Customer Profiling

This is also known as customer-oriented strategy in e-commerce. This allows companies to use business intelligence through the mining of customer's data to plan their business activities and operations as well as develop new research on products or services for prosperous e-commerce. Classifying the customers of great purchasing potentially from the visiting data can help companies to lessen the sales cost. Companies can use users' browsing data to identify whether they purposefully shopping or just browsing or buying something they are familiar with or something new. This helps companies to plan and improve their infrastructure.

Personalization of Service

Personalization is the act to provide contents and services geared to individuals on the basis of information of their needs and behavior. Data mining research related to personalization has focused mostly on recommender systems and related subjects such as collaborative filtering. Recommender systems have been explored intensively in the data mining community. These systems can be divided into three groups: Content-based, social data mining and collaborative filtering. These systems are cultured and learned from explicit or implicit feedback of users and are usually represented as the user profile. Social data mining, in considering the source of data that are created by the group of individuals as part of their daily activities, can be important source of important information for companies. Contrarily, personalization can be achieved by the aid of collaborative filtering, where users are matched with particular interest and in the same vein the preferences of these users to make recommendations.

Basket Analysis

Every shopper's basket has a story to tell and Market Basket Analysis (MBA) is a common retail, analytic and business intelligence tool that helps retailers to know their customers better. There are different ways to get the best out of market basket analysis and these include:

- Identification of product affinities: tracking not so apparent product affinities and leveraging on them is the real challenge in retail. Wal-Mart customers purchasing Barbie dolls shows an affinity towards one of three candy bars, obscure connection such as this can be discovered with an advanced market basket analytics for planning more effective marketing efforts.
- Cross-sell and up-sell campaigns: these shows the products purchased together, so customers who purchase the printer can be persuaded to pick up high quality paper or premium cartridges.
- Pangrams and product combos: are used for better inventory control based on product affinities, developing combo offers and design effective user friendly pangrams in focusing on products that sells together.
- Shoppers profile; in analyzing market basket with the aid of data mining over time to get a glimpse of who your shoppers really are, gaining insight to their ages, income range, buying habits, likes and dislikes, purchase preferences, levering this and giving the customer experience .

Sales Forecasting

Sales forecasting involves the aspect of the time an individual customer spend to buy an item and in this process trying to predict if the customer will buy again. This type of analysis can be used to determine a strategy of planned obsolescence or figure out complimentary products to sell. In sales forecasting, cash flow can be projected into three which include the pessimistic, optimistic and the realistic. This helps to have a plan on the adequate amount of capital available to endure the worst possible scenario that is if sales do not go actually as planned.

Merchandise Planning

Merchandise planning is useful for both online and offline retail companies. In the case of online business, merchandise planning will help to determine stocking options and the inventory warehousing, while in the case of offline companies, business that are looking to boost by adding stores can assess the required amount of merchandise they will be adequately needing by having a foresight at the exact layout of the current store. Using the right approach to merchandise planning will definitely lead to answers on what to do with:

- Pricing: the aspect of database mining will help determining the suited best price of products or services in the processes of revealing customer sensitivity.
- Deciding on products; data mining provides e-commerce businesses with the aspect of which products customers actually desire, which includes the aspect of intelligence on competitor's merchandise.
- Balancing of stocks; in mining the retail database, it helps determine the right and specific amount of stocks needed not too much and not too less, throughout the business year and also during the buying seasons.

Market Segmentation

Customer segmentation is one of the best uses of data mining. From the lots of data gotten, it can be broken down into different and meaningful segments like income, age, gender, occupation of customers, and this can be used when either the companies are running email marketing campaigns or SEO strategies. The aspect of market segmentation can also help a company identify its own competitors. This provided information alone can help the retail company identify that the periodic respondents are usually not the only ones pointing the same customer money as the present company is. Segmenting the database of a retail company will improve the conversion rates as the company can focus there promotion on a close-fitted and highly wanted market. This also helps the retail company to understand the competitors that are involved in each and every segment in the process permitting the customization of products that will actually satisfy the target audience in a generic way.

2. Challenges of Data Mining in E-Commerce

Besides the benefits data mining provides challenges for e-commerce companies, which are as follows:

Spider Identification
As it is commonly known main aim of data mining is to convert data into useful knowledge. Main source of data for e-commerce companies is web pages. Therefore, it is critical for e-commerce companies to understand how search engines work to follow how quickly things happen, how they happen and when changes will show up in the search engines. Spiders are software programs that are sent out by the search engine to find new information. These spiders can also be called as bots or crawlers. It is a software program that search engine uses to request pages and download them, it comes as a surprise to some people, however what the search engine does is they use a link of an existing website to find a new website and request a copy of that page to download it to their server. This is what the search engines use to run the ranking algorithm against and that is what shows up in the search engine result page. Therefore, the challenge here is that the search engines need to download a correct copy of the website. E-commerce website needs to be readable and seeable and the algorithm is applied to the search engines database. Tools are needed to have the mechanisms to enable them automatically remove unwanted data that will be transformed to information in order for data mining algorithm to provide reliable and sensible output.

Data Transformations
In this case data transformation poses a challenge for data mining tools. Today, the data needed to transform can only be gotten from two different sources, one of which an active and operational system for the data warehouse to be built and secondly it should include some activities that involves assigning new columns, binning data and also aggregating the data as well. In the first process, it is needed to be modified infrequently that is only when there is a change in the site and lastly the set of the transformed data gives a significantly great challenge in the data mining process.

Scalability of Data Mining Algorithms
With yahoo which has over 1.2 billion page views in a day with the presence of large amount of data, scalability arises with significant issues;

- Due to the large amount of data size gathered from the website at a reasonable time, the data mining algorithmcan handle or process it as much as it’s needed especially because of the scale nonlinearly.
- The models that are generated tend to be too complicated for individuals to understand how it is interpreted.

Make Data Mining Models Comprehensible to Business Users
The results of data mining should be clearly understood by business users, from the merchandisers who are incharge of decision making to the creative designers that design the sites to marketers to spend advertising money.

Support Slowly Changing Dimensions
The demographic aspect of visitors change, in that they may get married, there is an increase in salaries or income, the rapid growth of their children, needs which are the bases on which it is modeled changes. Thus, the products attributes also change, in terms of new choices may be available, the design and the way the products or service is packaged and also the increase or degrade of quality. These attribute that change over time are often known as “Slowly Changing Dimensions”. In this case the main challenge here is to keep track of those changes and in the same vein providing support for the identified change in the analysis.

Make Data Transformation and Model Building Accessible to Business Users
Having the ability to provide definite answers to questions by individual business users, this requires the aspects of data transformations but with the technical understanding of the tools used in the analysis. Many commercials report designers and also online analytical processing (OLAP) tools are basically hard to understand by business users. In this case, two preferred solutions are (I) provision of templates, for the expected questions and (ii) provision of the experts via consultation or even a service organization. This mentioned challenge basically is to find a way to enrich the business users to as to be able to analyze the information themselves without and hiccups.

3. Trends That Effect Data Mining

In this section, it describes five external trends which promise to have a fundamental impact on data mining.

Data Trends
Perhaps the most fundamental external trend is the explosion of digital data during the past two decades. During this period, the amount of data probably has grown between six to ten orders of magnitude. Much of this data is accessible via networks. On the other hand, during this same period the number of scientists, engineers, and other analysts available to analyze this data has remained relatively constant. For example, the number of new Ph.D.’s in statistics graduating each year has remained relatively constant during this period. Only one conclusion is possible: either most of the data is destined to be write-only, or techniques, such as data mining, must be developed, which can automate, in part, the analysis of this data, filter irrelevant information, and extract meaningful knowledge.

Hardware Trends
Data mining requires numerically and statistically intensive computations on large data sets. The increasing memory and processing speed of workstations enables the mining of data

sets using current algorithms and techniques that were too large to be mined just a few years ago. In addition, the commoditization of high performance computing through SMP workstations and high performance workstation clusters enables attacking data mining problems that were accessible using only the largest supercomputers of a few years ago.

Network Trends
The next generation internet (NGI) will connect sites at OC-3 (155 MBits/sec) speeds and higher. This is over 100 times faster than the connectivity provided by current networks. With this type of connectivity, it becomes possible to correlate distributed data sets using current algorithms and techniques. In addition, new protocols, algorithms, and languages are being developed to facilitate distributed data mining using current and next generation networks.

Scientific Computing Trends
As mentioned above, scientists and engineers today view simulation as a third mode of science. Data mining and knowledge discovery serves an important role linking the three modes of science: theory, experiment and simulation, especially for those cases in which the experiment or simulation results in large data sets.

Business Trends
Today businesses must be more profitable, react quicker, and offer higher quality services than ever before, and do it all using fewer people and at lower cost. With these types of expectations and constraints, data mining becomes a fundamental technology, enabling businesses to more accurately predict opportunities and risks generated by their customers and their customers' transactions.

Common DM Tools

Weka: To have accurate data mining result require the right tool for the dataset which are mining. Weka however, gives the ability to put into reality the learning methods algorithms. The tool has lots of benefits as it's include all the standard data mining procedures like data pre-processing, clustering, association, classification, regression and also attribute selection. It has both the Java and non-Java version together with visualization application, and the tool is free to users to customize it to their own specification.

NLTK: It is mainly for language processing task with pool of different language processing tools together with machine learning, data mining and sentiment analysis, data scrapping and different language processing tasks.NLTK tool require a user to install the tool on their systems and have access to the full package. It is built in pythonand a user can build application on top and can play around with the tool to their own specification. All thethree mentioned tools above are open source.

Spider Miner: A data mining tool that does not require a user to write a code, written in Java programming language. Part of SpiderMiner tool capability is that, it provides a thorough analytics via template-based frameworks. It is very flexible tool and user friendly offered as a service, and apart from data mining function, the tool can visualize, predict, data pre-processing, deployment statistical modeling and of course evaluation functions. In the tool there learning schemes, algorithms and models from WEKA and R script which makes the tool to be more powerful. All the three mentioned tools above are open source [15].

Data- Mining Applications
A wide range of companies have deployed successful applications of data mining. While early adopters of this technology have tended to be in information-intensive industries such as financial services and direct mail marketing, the technology is applicable to any company looking to leverage a large data warehouse to better manage their customer relationships. Two critical factors for success with data mining are: a large, well-integrated data warehouse and a well-defined understanding of the business process within which data mining is to be applied.

- A pharmaceutical company can analyze its recent sales force activity and their results to improve targeting of high-value physicians and determine which marketing activities will have the greatest impact in the next few months. The data needs to include competitor market activity as well as information about the local health care systems. The results can be distributed to the sales force via a wide-area network that enables the representatives to review the recommendations from the perspective of the key attributes in the decision process. The ongoing, dynamic analysis of the data warehouse allows best practices from throughout the organization to be applied in specific sales situations.
- A credit card company can leverage its vast warehouse of customer transaction data to identify customers most likely to be interested in a new credit product. Using a small test mailing, the attributes of customers with an affinity for the product can be identified. Recent projects have indicated more than a 20- fold decrease in costs for targeted mailing campaigns over conventional approaches.
- A diversified transportation company with a large direct sales force can apply data mining to identify the best prospects for its services. Using data mining to analyze its own customer experience, this company can build a unique segmentation identifying the attributes of high-value prospects. Applying this segmentation to a general business database such as those provided by Dun & Bradstreet can yield a prioritized list of prospects by region.
- A large consumer package goods company can apply data mining to improve its sales process to retailers. Data from consumer panels, shipments, and competitor activity can be applied to understand the reasons for brand and store switching. Through this analysis, the manufacturer can select promotional strategies that best reach their target customer segments.

4. Conclusion

In this paper we discussed some common tools and techniques used in Data-Mining and also some applications of these tools to Data-Mining in E-commerce. E-commerce is using technology to improve business processes in last decades. There are several major Data-Mining techniques have been developing and using in Data-Mining projects recently In E-commerce. Data-Mining is useful for discovering patterns and relationships in data to help make better decisions. Data-Mining helps in developing smarter marketing campaigns and to predict customer loyalty. Data mining for e-commerce companies should no longer be privilege but requirement in order to survive and remain relevant in the competitive environment. On one hand, data mining offers number of benefits to E-commerce companies and allows them to do merchandise planning, analyze customers' purchasing behaviors and forecast their sales which in turn would place them over other companies and generate more revenue. On the other hand, there are certain challenges of data mining in the field of E-commerce such as spider identification, data transformation, scalability of data mining algorithms, making data mining model comprehensible to business users, support slow changing dimensions and making data transformation and model building accessible to business users. To sum up, the special issue highlights that although E-commerce systems are an ideal application for Data-Mining. However, there still is much research needed in order to improve E-commerce marketing.

5. Acknowledgment

The authors would like to thank the editors and anonymous reviewers for their insightful comments and constructive suggestions. This work is partially supported by the Education Department Scientific Research Funds of Zhejiang province (Grants Y201839622).

References

[1] Pankaj Gupta, Bharat Bhushan, " Data Mining Techniques in E-Business", International Journal of Advances in Engineering Sciences, Jan, 2012.

[2] M.S. Chen, J.S. Park, and P.S. Yu, "Efficient Data Mining for Path Traversal Patterns," IEEE Trans. Knowledge and Engg., Vol.10, n0.2, pp.209-221, Apr. 1998.

[3] Agarwal R, Srikant R. "Fast algorithms for mining association rules", 1994.

[4] N.R. SrinivasaRaghavan, " Data mining in e-commerce: A survey", April/June 2005.

[5] GuHongjiu, " Data Mining in the Application of E-Commerce Website", China.

[6] Ahmad Tasnim Siddiqui, Sultan Aljahdali, " Web Mining Techniques in E-Commerce Applications", International Journal of Computer Applications, May 2013.

[7] Hamid Rastegari, Mohd Noor Md. Sap, " Data Mining and E-Commerce: Methods, Applications and Chalenges", JuranalTeknogiMaklumat, December 2008.

[8] P.L. Carbone, " Expanding the Meaning of and applications for data mining", proceedings of the IEEE International Conference on Systems, Man and Cybernetics, vol.3, pp.1872-1873, 2000.

[9] J.-1. Jeng and Y. Drissi, "PENS: A predictive event notification system for eCommerceenvironment," Proceedings - IEEE Computer Society's International Computer Software and ApplicationsConference. pp. 93-98, 2000.

[10] X. Z. Zhang, "Building personalized recommendation system in E-Commerce usingassociation rule-based mining and classification," in Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, ICMLC 2007, 2007,pp.4113-4118.

[11] B. Mobasher, "Web usage mining and personalization," Practical Handbook of Internet Computing, 2004.

[12] S. Vallamkonduand L. Gruenwald, "Integrating purchase patterns and traversal patterns to predict http requests in e-commerce sites," IEEE Int. Conf. on e-commerce, pp. 256-263, 2003.

[13] R. Kohavi, "Lessons and Challenges from Mining Retail E-Commerce Data," 2004.

[14] S. Krishnaswamy, A. Zaslavsky, and S. W. Loke, "An architecture to support distributed data mining services in e-commerce environments," in Advanced Issues of E-Commerce and Web-Based Information Systems, 2000. WECWIS 2000. Second International Workshop on,2000, pp. 239-246.

[15] Mustapha Ismail, Mohammed Mansur Ibrahim, Zayyan Mahmoud Sanusi, Muesser Nat, "Data Mining in Electronic Commerce: Benefit and Chalanges" , Scientific Research Publishing, December 2015.

[16] Ralph, K. and Caserta, J. (2011) The Data Warehouse ETL Toolkit: Practical Techniques for Extraction, Cleaning, Conforming and Delivering Data. Wiley Publishing Inc., USA.