IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 3, Ver. I (May-Jun. 2016), PP 43-47
www.iosrjournals.org

A Survey on Different Levels of Risks during Different Phases in Data Warehouse

Prangyan Mohapatra[1], Nachiketa Tarasia[2], Ananta Chandra Das[3]
[1,3]*M.Tech, School of Computer Engineering, KIIT University*
[2]*Associate professor, School of computer Engineering, KIIT University*

Abstract: *The term Data Warehouse represents huge collection of historical data which are subject-oriented, non-volatile, integrated, and time-variant and such data is required for the business needs [1]. Data warehouses and on-line analytical processing (OLAP) tools have become essential elements of decision support systems. Traditionally, data warehouses are refreshed periodically (for example, nightly) by extracting, transforming, cleaning and consolidating data from several operational data sources. The data in the warehouse is then used to periodically generate reports, or to rebuild multidimensional (data cube) views of the data for on-line querying and analysis. Increasingly, business intelligence applications in telecommunications, electronic commerce, and other industries, that are characterized by very high data volumes and data flow rates, and that require continuous analysis and mining of the data. For such applications, rather different data warehousing and on-line analysis architectures are required [2]. Although Data warehousing is a part of Business Intelligence and the motto of implementing a data warehouse is for business strategic needs, however our scope is to understand the various components that is involved in data warehousing, their purpose of work and scope of each components. Through this study we will be working on the methodologies of data cleansing techniques and improvement in such area [3].*

Keywords *: Data Warehouse, Business Intelligence(BI), OLAP, OLTP, ETL*

I. Literature Survey

1.1 Need of Data warehouse

Business Intelligence refers to a set of methods and techniques that are used by organizations for tactical and strategic decision making. It leverages technologies that focus on counts, statistics and business objectives to improve business performance.

A Data Warehouse (DW) is simply a consolidation of data from a variety of sources that is designed to support strategic and tactical decision making. Its main purpose is to provide a coherent picture of the business at a point in time. Using various Data Warehousing toolsets, users are able to run online queries and 'mine" their data.Many successful companies have been investing large sums of money in business intelligence and data warehousing tools and technologies. They believe that up-to-date, accurate and integrated information about their supply chain, products and customers are critical for their very survival.

1.2Why do us Implement Business Intelligence

Larry Ellison of Oracle has said of Strategic Business Intelligence that the best run businesses run better with business intelligence. Without BI, a company runs the risk of making critical decisions based on insufficient or inaccurate information. Making decisions based on "gut feel" will not get the job done!
It thus helps you to achieve the following :

- Quickly Identify and Respond to Business Trends.
- Empowered Staff Using Timely, Meaningful Information and Trend Reports
- Easily Create In-Depth Financial, Operations, Customer, and Vendor Reports
- Efficiently View, Manipulate, Analyze, and Distribute Reports Using Many Familiar Third-Party Tools
- Extract Up-to-the-Minute High-Level Summaries, Account Groupings, or Detail Transactions
- Consolidate Data from Multiple Companies, Divisions, and Databases
- Minimize Manual and Repetitive Work

1.3Components of Data Warehouse

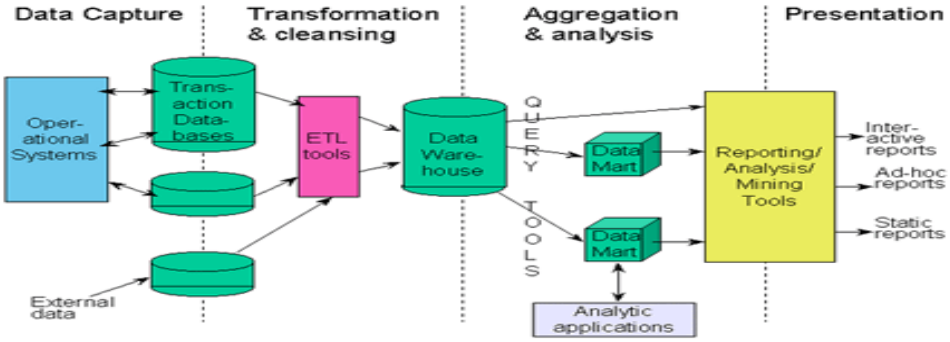

Figure 2.3(a): Components of Data Warehouse

1.3.1 Data Capture

Gathering data is concerned with collecting or accessing data which can then be used to inform decision making. Gathering data can come in many formats and basically refers to the automated measurement and collection of performance data. For example, these can come from transactional systems that keep logs of past transactions, point-of-sale systems, web site software, production systems that measure and track quality, flat files that used old system of storing data. A major challenge of gathering data is making sure that the relevant data is collected in the right way at the right time. If the data quality is not controlled at the data gathering stage then it can jeopardize the entire BI efforts that might follow – always remember the old adage - garbage in garbage out.

1.3.2 Transformation & Cleansing

The next component is modelling the data. Here we take the data that has been captured and in this process we inspect the current data, transform the current raw data to the present format available in the warehouse or model it in order to gain new insights that will support our business decision making. Data analysis comes in many different formats and approaches, both quantitative and qualitative. Analysis techniques includes the use of statistical tools, data mining approaches as well as visual analytics or even analysis of unstructured data such as text or pictures. Cleansing includes conversion of the raw data to the data formats present in the warehouse.

1.3.3 Aggregation & Analysis

The next component is analyzing the data. After the data is modelled and cleansed and stored into the warehouse, generally the proposed BI model is published into the portal where it is accessed by reporting & OLAP tools. Thus from the specific model designed by the modeler, various dimensional reports are developed by thorough analysis and different mathematical operations are performed to get various results.

2.3.4 Presentation

The next component is presenting the data to the business users. The reporting tools and OLAP tools have beautiful enhancing capabilities to present the designed data to the business users. The data can be visualized by standard reports, bar, graphs, pie-charts etc.

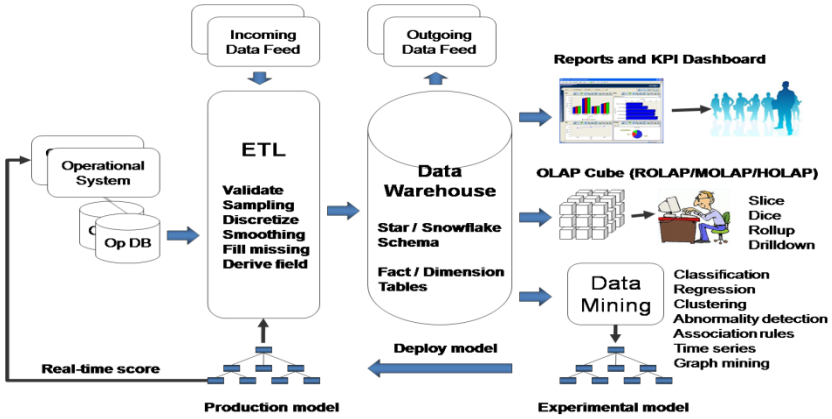

Figure 2.4(a): Flow Diagram of Data warehouse

2.5 Working Mechanism of Data warehouse

The below is the working mechanism of Data warehousing.

Step-1

At first the business users have to give the requirement on what the analysis has to be done or what is the prime requirement for the top level management for the growth their business.

Step-2

Now based on the requirement of the business, a model has to be prepared on which the existing system is running.

Step-3

Based on the proposed model, data is gathered from various sources, the sources may be live database, flat files, production servers etc. The data collected is thoroughly verified in the first stage for the correctness.

Step-4

After the data is collected from various sources, then the data is cleaned removing duplicity, cleansed to fit the standard formats in the warehouse, transformed to the suitable model that has been proposed for the business needs.Here the techniques of ETL is applied to fetch the data from different sources and on applying various techniques is stored in the data warehouse.

Step-5

The modelled data that is now stored in the warehouse is fetched by different BI tools. Where the data undergoes different mathematical operation and the result is generated for the business users. Data generated form the BI tools may be represented in various dimensions.

Step-6

After the operations are completed in the BI tools, the data is presented to the Business users using enhanced capabilities provided by the tool. The data is beautified and presented by pictorial means or by simple reports based on the business users need.

II. Motivation

Business growth purely depends on the data. The more is the data present, the best will the business grow. Due to day by day growth of business, huge amount of transactional data occurs and to analyze the transactional data, data warehouse is required. Now a days every business has its own warehouse where they store historical data and analyses them. We motivate ourselves to analyses the existing techniques and to implement newly developed techniques such that the business gets huge benefits.

III. Problem Definition

The thesis problem definitionis to understand every components, its function, implementation, and benefits with respect to the data warehouse. In our Scope, we will be studying techniques involved in data cleansing and improvement to be suggested in this area.

3.1 Data Capture

Type of Scope	Types of risk involved	Level of risk	Existing Methods to reduce risk
Flat Files	Error from the Flat Files	Low	Manual Testing
	Corrupt Files	High	File Recovery System
	Permission Issue	medium	Grant Permission
	File format Issue	High	Reloading the File
OLTP	Server Overload	medium	Waiting for the previous process to complete
	Missing rows	High	Back Tracking
	Overwriting on previous records or duplicacy	Low	Removing if duplicate
	Format not matching	medium	Handled by the ETL tool
Production Server	Duplicate records	Low	Removing if duplicate
	Server Overload	medium	Waiting for the previous process to complete

Table 4.1(a): Data Capture

3.2 Transformation & Cleansing

Type of Scope	Types of risk involved	Level of risk	Existing Methods to reduce risk
Extract	Data sources not available	high	Creating new connection
	Format not matching	medium	ETL tools handles it
	poor connectivity	medium	wait for proper connection
Transformation	Format not matching	medium	ETL tools handles it
	Improper group by clause	medium	Manual Testing
	cross joins	high	Manual testing
	look up failures	medium	ETL tools handles it
	date time format mismatch	high	manual overriding conversion
	aggregation not working	medium	ETL tools handles it
Load	Target not found	high	Creating new connection
	table lock	high	removing lock
	table not found	high	Creating new table columns
	table columns not matching	high	Creating new tables

Table 4.2(a): Transformation & Cleansing

3.3 Aggregation & Analysis

Type of Scope	Types of risk involved	Level of risk	Existing Methods to reduce risk
Report Model	Modelling not done as per the Business requirement	high	Manual modelling done
	Errors while creating the model	high	Manual testing
	Connection failure	medium	Check the connection
	Unable to fetch the database tables	high	Check the table
	Permission not granted	medium	Grant permission
	Report not working	high	Test the report or recreate it
Cube Model	Modelling not done as per the Business requirement	high	Manual modelling done
	Errors while creating the model	high	Manual testing
	Connection failure	medium	Check the connection
	Unable to fetch the database tables	high	Check the table
	Permission not granted	medium	Grant permission
	Cube Report not working	high	Test the report or recreate it
	Dimensions are not inter related	high	Check the dimensions properly
	Aggregation not working	medium	re- aggregate
	Fact table data mismatch	high	Check for valid data. Manual testing
	Report model not working	high	Check the report model

Table 4.3(a): Aggregation & Analysis

3.4 Presentation

Type of Scope	Types of risk involved	Level of risk	Existing Methods to reduce risk
Reports	Wrong Report	high	connect to the proper report model
	Permission issue	medium	Grant permission
	Data not matching	high	Check the data
	Irrelevant data	high	Check the data
	Data format not matching	medium	Check the data format
	Pictoral Data is wrong	high	Check the data

Table 4.4(a): Presentation

3.5 Data Alliance rule

Data alliance rules are based on mathematical
association rules [8] and are defined as follows:
Let F= {f1, f2------- fn) be a set of fields, where each field Fi ⊆DM. DM is a collection of data marts suchthat DM= {DM1, DM2, ------- DMm}.

K= {k1, k2----------kk} is set of k scores, where each Kk⊆ F.
Also the score set has integer as the numerical domain D (K € D) owing relationships defined in D: = equal, ≠ not equal, ≥ greater or equal.
Then, K1, K2 => K1 op K2, where op ⊆{≠, =, ≥}, is an alliance rules if,

1) K1 and K2 occur together in at least s% of the n records, where s is the support of the rule.

2) In c% of the records, L=> K1 op K2 op T where K1, K2 are reference table score and data warehouse table score respectively and op $\{=, \geq\}$.

Confidence of threshold value is 50%. If 50% letters do not match the word is bound to be an outlier value. The rule 2 is applied for q-grams matching. The qgrams [1] are the substrings of a length q of a given string.The data marts taken into consideration over here are the sub sections or are the multiple sources from where the data is collected and amalgamated to construct a complete data warehouse.

IV. Work Progress Status

Current work progress status includes study of different components in data warehouse. We will be working on the mechanism of data cleansing and the improvements that can be suggested.
The above Alliance rules is used to trace duplicate rows and remove duplicity. As well as based on this rule above various date format conversion is done and the data is cleansed.
But it is now applied to all the cases of data cleansing. We will propose methods to Cleanse data even if there is some space between the letters to maximize joining and retrieving accurate data.

V. Conclusion

Data Warehousing is not a new phenomenon. All large organizations already have data warehouses, but they are just not managing them. Over the next few years, the growth of data warehousing is going to be enormous with new products and technologies coming out frequently. In order to get the most out of this period, it is going to be important that data warehouse planners and developers have a clear idea of what they are looking for and then choose strategies and methods that will provide them with performance today and flexibility for tomorrow.

References

[1] Arora, R.; Pahwa, P.; Bansal, S.”Alliance Rules for Data Warehouse Cleansing”.2009 International Conference on Signal Processing Systems15-17 May 2009. Pages: 743– 747.
[2] Prasad, K.H.; Faruquie, T.A.; Joshi, S.; Chaturvedi, S.; Subramaniam, L.V.; Mohania, M.”Data Cleansing Techniques for Large Enterprise Datasets”. SRII Global Conference (SRII), 2011 Annual March 29 2011-April 2 2011.Page 135-144.
[3] Savitri, F.N.; Laksmiwati, H.” Study of localized data cleansing process for ETL performance improvement in independent datamart”. Electrical Engineering and Informatics (ICEEI), 2011 International Conference on 17-19 July 2011. Pages: 1 – 6
[4] Xingquan Zhu; Peng Zhang; Xindong Wu; Dan He; Zhang, C.; Yong Shi.” Cleansing Noisy Data Streams”. Data Mining, 2008. ICDM '08. Eighth IEEE International Conference on 15-19 Dec. 2008. Pages: 1139 - 1144
[5] Das, A.C.; Mohanty, S.N.; Pani, S.K.;"A Comparative Study on Data analytics and Big Data Analytics", IJCSITR, VOL 4, Issue 1, 2016, Pages: 67-75.