INTERNATIONAL JOURNAL OF RESEARCH IN COMMERCE, IT & MANAGEMENT

IJRCM

A Monthly Double-Blind Peer Reviewed Refereed Open Access International e-Journal - Included in the International Serial Directories

Indexed & Listed at:

Ulrich's Periodicals Directory ©, ProQuest, U.S.A., EBSCO Publishing, U.S.A., Index Copernicus Publishers Panel, Poland,

Open J-Gage, India [link of the same is duly available at Inflibnet of University Grants Commission (U.G.C.)]

as well as in Cabell's Directories of Publishing Opportunities, U.S.A.

Circulated all over the world & Google has verified that scholars of more than Hundred & Twenty One countries/territories are visiting our journal on regular basis.

Ground Floor, Building No. 1041-C-1, Devi Bhawan Bazar, JAGADHRI – 135 003, Yamunanagar, Haryana, INDIA

www.ijrcm.org.in

CONTENTS

CALL FOR MANUSCRIPTS

We invite unpublished novel, original, empirical and high quality research work pertaining to recent developments & practices in the area of Computer, Business, Finance, Marketing, Human Resource Management, General Management, Banking, Insurance, Corporate Governance and emerging paradigms in allied subjects like Accounting Education; Accounting Information Systems; Accounting Theory & Practice; Auditing; Behavioral Accounting; Behavioral Economics; Corporate Finance; Cost Accounting; Econometrics; Economic Development; Economic History; Financial Institutions & Markets; Financial Services; Fiscal Policy; Government & Non Profit Accounting; Industrial Organization; International Economics & Trade; International Finance; Macro Economics; Micro Economics; Monetary Policy; Portfolio & Security Analysis; Public Policy Economics; Real Estate; Regional Economics; Tax Accounting; Advertising & Promotion Management; Business Education; Business Information Systems (MIS); Business Law, Public Responsibility & Ethics; Communication; Direct Marketing; E-Commerce; Global Business; Health Care Administration; Labor Relations & Human Resource Management; Marketing Research; Marketing Theory & Applications; Non-Profit Organizations; Office Administration/Management; Operations Research/Statistics; Organizational Behavior & Theory; Organizational Development; Production/Operations; Public Administration; Purchasing/Materials Management; Retailing; Sales/Selling; Services; Small Business Entrepreneurship; Strategic Management Policy; Technology/Innovation; Tourism, Hospitality & Leisure; Transportation/Physical Distribution; Algorithms; Artificial Intelligence; Compilers & Translation; Computer Aided Design (CAD); Computer Aided Manufacturing; Computer Graphics; Computer Organization & Architecture; Database Structures & Systems; Digital Logic; Discrete Structures; Internet; Management Information Systems; Modeling & Simulation; Multimedia; Neural Systems/Neural Networks; Numerical Analysis/Scientific Computing; Object Oriented Programming; Operating Systems; Programming Languages; Robotics; Symbolic & Formal Logic and Web Design. The above mentioned tracks are only indicative, and not exhaustive.

Anybody can submit the soft copy of his/her manuscript **anytime** in M.S. Word format after preparing the same as per our submission guidelines duly available on our website under the heading guidelines for submission, at the email addresses: infoijrcm@gmail.com or info@ijrcm.org.in.

GUIDELINES FOR SUBMISSION OF MANUSCRIPT

1. **COVERING LETTER FOR SUBMISSION:**

DATED: _____

THE EDITOR
IJRCM

Subject: **SUBMISSION OF MANUSCRIPT IN THE AREA OF** _____.

(e.g. Finance/Marketing/HRM/General Management/Economics/Psychology/Law/Computer/IT/Engineering/Mathematics/other, please specify)

DEAR SIR/MADAM

Please find my submission of manuscript entitled '_____' for possible publication in your journals.

I hereby affirm that the contents of this manuscript are original. Furthermore, it has neither been published elsewhere in any language fully or partly, nor is it under review for publication elsewhere.

I affirm that all the author (s) have seen and agreed to the submitted version of the manuscript and their inclusion of name (s) as co-author (s).

Also, if my/our manuscript is accepted, I/We agree to comply with the formalities as given on the website of the journal & you are free to publish our contribution in any of your journals.

NAME OF CORRESPONDING AUTHOR:
Designation:
Affiliation with full address, contact numbers & Pin Code:
Residential address with Pin Code:
Mobile Number (s):
Landline Number (s):
E-mail Address:
Alternate E-mail Address:

NOTES:

a) The whole manuscript is required to be in ***ONE MS WORD FILE*** only (pdf. version is liable to be rejected without any consideration), which will start from the covering letter, inside the manuscript.
b) The sender is required to mention the following in the **SUBJECT COLUMN** of the mail:
New Manuscript for Review in the area of (Finance/Marketing/HRM/General Management/Economics/Psychology/Law/Computer/IT/ Engineering/Mathematics/other, please specify)
c) There is no need to give any text in the body of mail, except the cases where the author wishes to give any specific message w.r.t. to the manuscript.
d) The total size of the file containing the manuscript is required to be below **500 KB**.
e) Abstract alone will not be considered for review, and the author is required to submit the complete manuscript in the first instance.
f) The journal gives acknowledgement w.r.t. the receipt of every email and in case of non-receipt of acknowledgment from the journal, w.r.t. the submission of manuscript, within two days of submission, the corresponding author is required to demand for the same by sending separate mail to the journal.

2. **MANUSCRIPT TITLE:** The title of the paper should be in a 12 point Calibri Font. It should be bold typed, centered and fully capitalised.

3. **AUTHOR NAME (S) & AFFILIATIONS:** The author (s) **full name**, **designation**, **affiliation** (s), **address**, **mobile/landline numbers**, and **email/alternate email address** should be in italic & 11-point Calibri Font. It must be centered underneath the title.

4. **ABSTRACT:** Abstract should be in fully italicized text, not exceeding 250 words. The abstract must be informative and explain the background, aims, methods, results & conclusion in a single para. Abbreviations must be mentioned in full.

5. **KEYWORDS:** Abstract must be followed by a list of keywords, subject to the maximum of five. These should be arranged in alphabetic order separated by commas and full stops at the end.

6. **MANUSCRIPT:** Manuscript must be in ***BRITISH ENGLISH*** prepared on a standard A4 size ***PORTRAIT SETTING PAPER***. It must be prepared on a single space and single column with 1" margin set for top, bottom, left and right. It should be typed in 8 point Calibri Font with page numbers at the bottom and centre of every page. It should be free from grammatical, spelling and punctuation errors and must be thoroughly edited.

7. **HEADINGS:** All the headings should be in a 10 point Calibri Font. These must be bold-faced, aligned left and fully capitalised. Leave a blank line before each heading.

8. **SUB-HEADINGS:** All the sub-headings should be in a 8 point Calibri Font. These must be bold-faced, aligned left and fully capitalised.

9. **MAIN TEXT:** The main text should follow the following sequence:

INTRODUCTION

REVIEW OF LITERATURE

NEED/IMPORTANCE OF THE STUDY

STATEMENT OF THE PROBLEM

OBJECTIVES

HYPOTHESES

RESEARCH METHODOLOGY

RESULTS & DISCUSSION

FINDINGS

RECOMMENDATIONS/SUGGESTIONS

CONCLUSIONS

SCOPE FOR FURTHER RESEARCH

ACKNOWLEDGMENTS

REFERENCES

APPENDIX/ANNEXURE

It should be in a 8 point Calibri Font, single spaced and justified. The manuscript should preferably not exceed ***5000 WORDS***.

10. **FIGURES &TABLES:** These should be simple, centered, separately numbered & self explained, and **titles must be above the table/figure**. **Sources of data should be mentioned below the table/figure**. It should be ensured that the tables/figures are referred to from the main text.

11. **EQUATIONS:** These should be consecutively numbered in parentheses, horizontally centered with equation number placed at the right.

12. **REFERENCES:** The list of all references should be alphabetically arranged. The author (s) should mention only the actually utilised references in the preparation of manuscript and they are supposed to follow **Harvard Style of Referencing**. The author (s) are supposed to follow the references as per the following:

- All works cited in the text (including sources for tables and figures) should be listed alphabetically.
- Use (**ed.**) for one editor, and (**ed.s)** for multiple editors.
- When listing two or more works by one author, use --- (20xx), such as after Kohl (1997), use --- (2001), etc, in chronologically ascending order.
- Indicate (opening and closing) page numbers for articles in journals and for chapters in books.
- The title of books and journals should be in italics. Double quotation marks are used for titles of journal articles, book chapters, dissertations, reports, working papers, unpublished material, etc.
- For titles in a language other than English, provide an English translation in parentheses.
- The location of endnotes within the text should be indicated by superscript numbers.

PLEASE USE THE FOLLOWING FOR STYLE AND PUNCTUATION IN REFERENCES:

BOOKS

- Bowersox, Donald J., Closs, David J., (1996), "Logistical Management." Tata McGraw, Hill, New Delhi.
- Hunker, H.L. and A.J. Wright (1963), "Factors of Industrial Location in Ohio" Ohio State University, Nigeria.

CONTRIBUTIONS TO BOOKS

- Sharma T., Kwatra, G. (2008) Effectiveness of Social Advertising: A Study of Selected Campaigns, Corporate Social Responsibility, Edited by David Crowther & Nicholas Capaldi, Ashgate Research Companion to Corporate Social Responsibility, Chapter 15, pp 287-303.

JOURNAL AND OTHER ARTICLES

- Schemenner, R.W., Huber, J.C. and Cook, R.L. (1987), "Geographic Differences and the Location of New Manufacturing Facilities," Journal of Urban Economics, Vol. 21, No. 1, pp. 83-104.

CONFERENCE PAPERS

- Garg, Sambhav (2011): "Business Ethics" Paper presented at the Annual International Conference for the All India Management Association, New Delhi, India, 19–22 June.

UNPUBLISHED DISSERTATIONS AND THESES

- Kumar S. (2011): "Customer Value: A Comparative Study of Rural and Urban Customers," Thesis, Kurukshetra University, Kurukshetra.

ONLINE RESOURCES

- Always indicate the date that the source was accessed, as online resources are frequently updated or removed.

WEBSITE

- Garg, Bhavet (2011): Towards a New Natural Gas Policy, Political Weekly, Viewed on January 01, 2012 http://epw.in/user/viewabstract.jsp

DATA WAREHOUSING AND TESTING

VENKATESH RAMASAMY
SENIOR TEST ANALYST
COGNIZANT TECHNOLOGY SOLUTIONS
SIRUSERI

ABINAYA MURUGANANDHAN
MBA STUDENT
ANNA UNIVERSITY
TIRUCHIRAPPALLI

ABSTRACT

Over the last 20 years, with the advancement in the computing technology, the fall in the computer hardware and change in the nature of business – the value of information has been raised dramatically. The need of making good decisions on the basis of large amount of data, which has the property of diversification among the different units of organization, has risen to a level not comparable to any phase throughout the history of Information Technology. The indispensable requirement to store enormous amount of data lead to the analytic systems which in turn gave birth to the idea of Data Warehousing. In this paper, we discuss about the data warehouse design, implementation and its testing activities, also we classify them in terms of what is tested, how it is tested, and we explain how the test approach can be effectively designed.

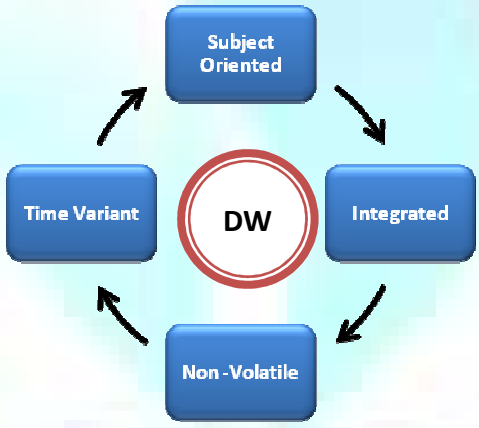
KEYWORDS

Data Warehouse (DW), Data Mining (DM), Query Processing, Testing.

INTRODUCTION

The Data Warehouse (DW) integrates data from multiple heterogeneous information sources and transforms them into a multidimensional representation for making decisions more effective. The responsibility of a data warehouse is to provide robust data management, scalability, high performance query processing and integration with organization data servers. "Data Warehousing is a Subject-Oriented, Integrated, Non-volatile, and Time variant collection of data in support of Management Decisions", - W.H. Inmon.

FIG. 01 – DATA WAREHOUSING INTRODUCTION



SUBJECT-ORIENTED

In data warehousing the prime objective of storing data is to facilitate decision process of a company, and within any company data naturally concentrates around subject area. This leads to the gathering of information around these subjects rather than around the applications or processes.

INTEGRATED

Though the data in the warehouses is scattered around different tables, databases or even servers, the data is integrated consistently in the values of variables, naming conventions and physical data definitions.

NON-VOLATILE

Being the snapshot of operational data on a given specific time, the data in the data warehouses should not be changed or updated, once it's loaded from organizational system. As the snapshot shows operational data at some moment of time and one expects data warehouse to reflect accurate values of that time frame. There exists only two operations – the time based loading of data, accessing the loaded data.

TIME VARIANT

The time based archival of data from operational systems to data warehouse, makes value of data, in the data warehouses, being function of time. As the data warehouse gives accurate picture of operational data for some given time and the change in the data in warehouse is based on time based change in operational data, data in the data warehouse is called, "time-variant".

Most of the organizations run their businesses on the basis of collection of data for strategic decision- making. To get a competitive edge the organization should have the ability to review historical trends and monitor real-time functional data. Hence, the concept of data warehousing in the sense, once the data is extracted from the operational system it can be validated, reformatted, reorganized, summarized, restructured and supplemented with data from other source. The resulting data warehouse become the best source of information for the decision making process. Data Warehousing technology has grown much in scale and reputation in the past few years, as evidenced by Gartner Group survey of Fortune 500 IT managers who found that 90% of all organizations had planned to implement Data Warehouses by 2011.

NEED OF DATA WAREHOUSING

BUSINESS PROCESSING AND DECISION MAKING

In large organizations good decision-making process becomes very critical because of its diversified data which is available across different platforms. The best decisions are made when all the relevant data available are taken into consideration. The best possible source for that data is a well-designed Data Warehouse System.

DATA MINING AND DATA TRANSFORMATION

A common problem that exists in many organizations is the inability to quickly retrieve and combine operational data about the same entity that exists in multiple systems. At a very simple level, data reporting quality are concerned with collecting accurate data, transforming it into a common presentation format that is convenient for business partners. This sounds simple enough but there are many complicating factors that must be considered.

DATA SECURITY

With the provision of being multi-user database system, the primary focus is on providing controls for the security of database. The controls include unauthorized access to database as well as individual schema objects, assessment of environment parameters such as disk usage and system resource usage. It needs a set of privileges and a user is restricted by the grants given by these privileges. Data Warehousing provides a multi-level secure database management and mandatory access control to monitor the user actions on the Database.

DATA VOLUME

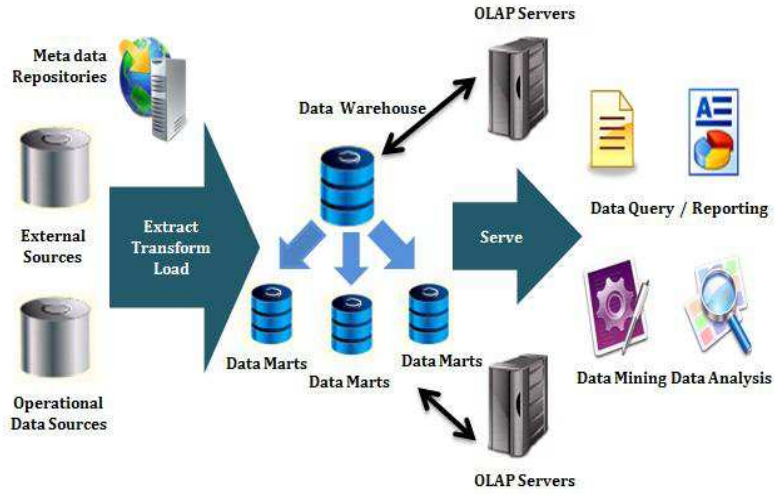
In any organization, it is frequently necessary to load a very large volume of data into the databases in a short amount of time, thereby requiring parallel processing of different data. The initial data loads are usually the most voluminous, so the organizations need to have a relatively long load window to accomplish this task.

Even though additional hardware and software are needed, the presence of a data warehouse is obligatory to handle the huge volume of data.

DATA WAREHOUSING ARCHITECTURE

Many organizations want to implement an integrated enterprise warehouse that collects information about all subjects (e.g., customers, products, sales, assets, personnel) spanning the whole organization. However, building an enterprise warehouse is a long and complex process, requiring extensive business modeling, and may take many years to succeed. The below architectural design explains about the implementation of data warehousing in organization. It includes tools for extracting data from multiple operational databases and external sources; for cleaning, transforming and integrating this data; for loading data into the data warehouse and for periodically refreshing the warehouse to reflect updates at the sources and to purge data from the warehouse, perhaps onto slower archival storage. In addition to the main warehouse, there may be several departmental data marts.

FIG.2: DATA WAREHOUSING ARCHITECTURE



Note: Data marts are departmental subsets focused on selected subjects (e.g., a marketing data mart may include customer, product, and sales information). These data marts enable faster roll out, since they do not require enterprise-wide consensus, but they may lead to complex integration problems in the long run, if a complete business model is not developed.]

Data in the warehouse and data marts is stored and managed by one or more warehouse servers, which present multidimensional views of data to a variety of front end tools: query tools, report writers, analysis tools, and data mining tools. Finally, there is a repository for storing and managing metadata, and tools for monitoring and administering the warehousing system.

The warehouse may be distributed for load balancing, scalability, and higher availability. In such a distributed architecture, the metadata repository is usually replicated with each fragment of the warehouse, and the entire warehouse is administered centrally. An alternative architecture, implemented for expediency when it may be too expensive to construct a single logically integrated enterprise warehouse, is a federation of warehouses or data marts, each with its own repository and decentralized administration.

The designing and rolling out of a data warehouse is a complex process, consisting of the following activities,

- Define the architecture, do capacity planning, and select the storage servers, database and OLAP servers, and tools.
- Integrate the servers, storage, and client tools.
- Design the warehouse schema and views.
- Define the physical warehouse organization, data placement, partitioning, and access methods.
- Connect the sources using gateways, ODBC drivers, or other wrappers.
- Design and implement scripts for data extraction, cleaning, transformation, load, and refresh.
- Populate the repository with the schema and view definitions, scripts, and other metadata.
- Design and implement end-user applications.
- Roll out the warehouse and applications.

FACTORS INFLUENCING THE DATA WAREHOUSING DESIGN & IMPLEMENTATION

When designing a data warehouse, organizations employ either a top-down or bottom-up development approach. In the top-down approach, an enterprise data warehouse is built and information processes are created business area by business area with underlying dependent data marts pulled out from the enterprise data warehouse contents. In the bottom-up approach, independent data marts with information processes are created with the view of integrating them into an enterprise data warehouse in the future. The six important factors that potentially influence the design and implementation of data warehousing,

INFORMATION INTERDEPENDENCE BETWEEN ORGANIZATIONAL UNITS

There is a high level of information interdependence when the work of one organizational unit is dependent upon information from one or more other organizational units. In this situation, the ability to share consistent, integrated information is important.

UPPER MANAGEMENT'S INFORMATION NEEDS

In order to carry out their job responsibilities, senior management often requires information from lower organizational levels. It may need to monitor progress on meeting company goals, drill down into areas of interest, aggregate lower-level data, and be confident that the company is in compliance with regulations.

NATURE OF END USERS TASKS

Some users perform non-routine tasks. Structured queries and reports are insufficient for their needs. They have to analyze data in novel ways. These users require an architecture that provides enterprise-wide data that can be analyzed "on the fly" in creative ways.

COMPATIBILITY WITH EXISTING SYSTEMS

There are many benefits to implementing IT solutions that are compatible with the existing computing environment. Consequently, the design of data warehouse architecture is likely to be impacted by the systems and technologies that are already in place. This may include compatibility with source systems, metadata integration, data access tools, and technology vendors.

TECHNICAL ISSUES

A variety of technical considerations can affect the design of architecture – the ability to integrate metadata; scalability in terms of the number of users, volume of data, and query performance; the ability to maintain historical data; and the ability to adapt to technical changes, such as in source systems.

CONSTRAINTS ON RESOURCES

The data warehouse architecture requires more resources to develop and operate. As a result, the availability of IT personnel, business unit personnel, and monetary resources can impact the design of the architecture.

DATA ANALYSIS TECHNIQUES

A data warehouse is built to provide an easy to access source of high quality data. It is typically the need to perform analysis and decision making through the use of that source of data. There are several techniques for data analysis that are in common use today. They are query and reporting, multidimensional analysis, and data mining. They are used to formulate and display query results, to analyze data content by viewing it from different perspectives, and to discover patterns and clustering attributes in the data that will provide further insight into the data content. The techniques of data analysis can impact the type of data model selected and it's content. It also seems clear that, in most implementations, multiple types of data models might be used to best satisfy the varying requirements.

QUERYING AND REPORTING

Query and reporting analysis is the process of posing a question to be answered, retrieving relevant data from the data warehouse, transforming it into the appropriate context, and displaying it in a readable format. It is driven by analysts who must pose those questions to receive an answer. Traditionally, queries have dealt with two dimensions, or two factors, at a time. Query definition is the process of taking a business question or hypothesis and translating it into a query format that can be used by a particular decision support tool. When the query is executed, the tool generates the appropriate language commands to access and retrieve the requested data, which is returned in what is typically called an answer set. The data analyst then performs the required calculations and manipulations on the answer set to achieve the desired results. Those results are then formatted to fit into a display or report template that has been selected for ease of understanding by the end user. The report is delivered to the end user on the desired output medium, which could be printed on paper, visualized on a computer display device, or presented audibly.

MULTI DIMENSIONAL ANALYSIS

Multi-dimensional analysis has become a popular way to extend the capabilities of query and reporting. That is, rather than submitting multiple queries, data is structured to enable fast and easy access to answers to the questions that are typically asked. For example, the data would be structured to include answers to the question, "How much of each of our products was sold on a particular day, by a particular sales person, in a particular store?" Each separate part of that query is called a dimension. By pre calculating answers to each sub-query within the larger context, many answers can be readily available because the results are not recalculated with each query; they are simply accessed and displayed. By having the results to the above query, one would automatically have the answer to any of the sub-queries. Multidimensional analysis enables users to look at a large number of interdependent factors involved in a business problem and to view the data in complex relationships. End users are interested in exploring the data at different levels of detail, which is determined dynamically. The complex relationships can be analyzed through an iterative process that includes drilling down to lower levels of detail or rolling up to higher levels of summarization and aggregation.

DATA MINING

Data mining is a relatively new data analysis technique. It is very different from query and reporting and multidimensional analysis in that is uses a discovery technique. That is, we do not ask a particular question on the data but rather use specific algorithms that analyze the data and report what they have discovered. Unlike query and reporting and multidimensional analysis where the user has to create and execute queries based on hypotheses, data mining searches for answers to questions that may have not been previously asked. This discovery could take the form of finding significance in relationships between certain data elements, a clustering together of specific data elements, or other patterns in the usage of specific sets of data elements. After finding these patterns, the algorithms can infer rules. These rules can then be used to generate a model that can predict a desired behavior, identify relationships among the data, discover patterns, and group clusters of records with similar attributes. Data mining is most typically used for statistical data analysis and knowledge discovery. Statistical data analysis detects unusual patterns in data and applies statistical and mathematical modeling techniques to explain the patterns. The models are then used to forecast and predict. Knowledge discovery extracts implicit, previously unknown information from the data. This often results in uncovering unknown business facts. Data mining can help discover new insights about the business by giving us answers to questions we might never have thought to ask.

TOOLS USAGE IN DATA WAREHOUSING

Creating and managing a data warehousing system is a complex computing problem because of its integration with different organizational data sources. Many different classes of tools are available to facilitate different aspects of the design and implementation process. Development tools are used to design and edit schemas, views, scripts, rules, queries, and reports. Planning and analysis tools are used for what-if scenarios such as understanding the impact of schema changes or refresh rates, and for doing capacity planning.

Data Warehouse management tools (e.g., HP Intelligent Warehouse Advisor, IBM Data Hub, Prism Warehouse Manager) are used for monitoring a warehouse, reporting statistics and making suggestions to the administrator: usage of partitions and summary tables, query execution times, types and frequencies of drill downs or rollups, which users or groups request which data, peak and average workloads over time, exception reporting, detecting runaway queries, and other quality of service metrics. System and network management tools (e.g., HP OpenView, IBM NetView, and Tivoli) are used to measure traffic between clients and servers, between warehouse servers and operational databases, and so on. Finally, only recently have workflow management tools been considered for managing the extract-scrub-transform-load-refresh process. The steps of the process can invoke appropriate scripts stored in the repository, and can be launched periodically, on demand, or when specified events occur. The workflow engine ensures successful completion of the process, persistently records the success or failure of each step, and provides failure recovery with partial roll back, retry, or roll forward. Data quality tools are available to enhance the quality of the data at several stages in the process of developing a data warehouse. Data Cleansing tools like SAP Business Objects Data Services XI, can be useful in automating many of the activities that are involved in cleansing the data- parsing, standardizing, correction, matching, transformation and house holding.

DATA WAREHOUSE TESTING

Testing is an essential part of the design life-cycle of any software product. Needless to say, testing is especially critical to success in data warehousing projects because users need to trust in the quality of the information they access. Data warehouse testing may have the same principles/fundamentals of a general testing project, but testing data warehousing projects involve significant programming work as there are limited front end screens but mostly back end processes that work on data sets. The attitude and methodology required for data warehousing testing are not same as those required for normal testing. The skill sets required includes resources with a strong aptitude and with technology specialization and programming language.

The difference between testing data warehouse systems and generic software systems or even transactional systems depends on several aspects,

- Software testing is predominantly focused on program code, while data warehouse testing is directed at data and information. As a matter of fact, the key to data warehouse testing is to know the data and what the answers to user queries are supposed to be.
- Differently from generic software systems, data warehouse testing involves a huge data volume, which significantly impacts performance and productivity.
- Data warehouse testing has a broader scope than software testing because it focuses on the correctness and usefulness of the information delivered to users. In fact, data validation is one of the main goals of data warehouse testing.
- Though a generic software system may have a large number of different use scenarios, the valid combinations of those scenarios are limited. On the other hand, data warehouse systems are aimed at supporting any views of data, so the possible combinations are virtually unlimited and cannot be fully tested.
- While most testing activities are carried out before deployment in generic software systems, data warehouse testing activities still go on after system release.
- Typical software development projects are self-contained. Data warehousing projects never really come to an end; it is very difficult to anticipate future requirements for the decision-making process, so only a few requirements can be stated from the beginning. Besides, it is almost impossible to predict all the possible types of errors that will be encountered in real operational data. For this reason, regression testing is inherently involved.

Key focus areas need to be considered before data warehouse testing,

- Loss of data during the ETL process.
- Inaccurate and incomplete data coming from disparate source systems.
- Data duplication appearing in source feeds.
- Differing definitions of operational data from disparate systems.
- Authentication and security concerns.
- Multiple country integration issues, business definition and source systems disparity.
- Non-availability of comprehensive test bed.
- Imbalances in performance across the different environment like Development, UAT and Production.
- Data Quality Issues.

Data Warehouse Testing attempts to plug each of the above gaps by ensuring that every stage of the source to target movement/transformation of data is tested and is working fine.

CHALLENGES IN DATA WAREHOUSE TESTING

- Voluminous data, from heterogeneous sources.
- Data Quality not assured at source.
- Transaction-level traceability will be difficult to attain in a Data Warehouse.
- Difficult to estimate. Only volume might be available. No accurate picture of the quality of the underlying data.
- Business Knowledge. Organization-wide Enterprise data knowledge may not be feasible.
- 100% Data verification will not be feasible. In such cases, the extraction, transformation and loading components will be thoroughly tested to ensure all types of data behaves as expected, within each of these modules.
- Very High Cost of Quality. This is because any defect slippage will translate into significantly high costs for the organization.
- The Heterogeneous sources of data will be updated asynchronously. Temporal Inconsistency is part and parcel of a Data warehouse implementation.

DATA WAREHOUSE TEST APPROACH

Like for most generic software systems, different types of tests can be devised for data warehouse system. The peculiar characteristics of data warehouse testing and the complexity of data warehouse projects ask for a deep revision and contextualization of these test types, aimed in particular at emphasizing the relationships between testing activities.

Since the correctness of a system can only be measured with reference to a set of requirements, a successful testing begins with the gathering and documentation of end-user requirements. Since most end-users requirements are about data analysis and data quality, it is inevitable that data warehouse testing primarily focuses on the ETL process on the one hand (this is sometimes called back- end testing), on reporting and OLAP on the other (front-end testing). While back-end testing aims at ensuring that data loaded into the data warehouse are consistent with the source data, front-end testing aims at verifying that data are correctly navigated and aggregated in the available reports.

The below mentioned testing types are tightly related to the software quality factors like correctness, usability, efficiency, reliability, integrity, flexibility and also best fit the characteristics of data warehouse systems:

UNIT TESTING

It has been the task of the developer. This is white-box testing to ensure the module or component is coded as per agreed upon design specifications. The developer should focus on the following:

- That all inbound and outbound directory structures are created properly with appropriate permissions and sufficient disk space. All tables used during the ETL3 are present with necessary privileges.
- The ETL routines give expected results:

i. All transformation logics work as designed from source till target
ii. Boundary conditions are satisfied– e.g. check for date fields with leap year dates
iii. Surrogate keys have been generated properly
iv. NULL values have been populated where expected
v. Rejects have occurred where expected and log for rejects is created with sufficient details
vi. Error recovery methods
vii. Auditing is done properly

- That the data loaded into the target is complete:

i. All source data that is expected to get loaded into target actually get loaded– compare counts between source and target and use data profiling tools
ii. All fields are loaded with full contents– i.e. no data field is truncated while transforming
iii. No duplicates are loaded
iv. Aggregations take place in the target properly
v. Data integrity constraints are properly taken care of

SYSTEM TESTING

It verifies that the item is compliant with its specified business requirements. We test for the functionality of the application and mostly it is black-box. The major challenge here is preparation of test data. An intelligently designed input dataset can bring out the flaws in the application more quickly.

The QA team must test for:

- Data completeness – match source to target counts.
- Data aggregations – match aggregated data against staging tables and/or ODS4.
- Granularity of data is as per specifications.
- Error logs and audit tables are generated and populated properly.
- Notifications to IT and/or business are generated in proper format.

INTEGRATION TESTING

It is used to ensure that the application developed works from an end-to-end perspective. We must consider the compatibility of the Data Warehouse application with upstream and downstream flows. We need to ensure for data integrity across the flow. It is a combined responsibility and participation of experts from all related applications is a must in order to avoid misinterpretation of results.

PERFORMANCE TESTING

The Data Warehouse must necessarily go through another phase called performance testing. It checks that the item performance is satisfactory under typical workload conditions with huge volume of data. Any Data Warehousing application is designed to be scalable and robust. Therefore, when it goes into production environment, it should not cause performance problems. We must ensure that the load window is met even under such volumes. This phase should involve DBA team, and ETL expert and others who can review and validate the code for optimization.

REGRESSION TESTING

Data Warehouse application is not a one-time solution. Possibly it is the best example of an incremental design where requirements are enhanced and refined quite often based on business needs and feedbacks. In such a situation it is very critical to test that the existing functionalities of a DW application are not messed up whenever an enhancement is made to it. Generally this is done by running all functional tests for existing code whenever a new piece of code is introduced. However, a better strategy could be to preserve earlier test input data and result sets and running the same again. Now the new results could be compared against the older ones to ensure proper functionality.

RECOVERY TESTING

It checks how well an item is able to recover from crashes, hardware failures and other similar problems. The recovery test of ETL checks for robustness by simulating faults in one or more components and evaluating the system response. For example, you can cut off the power supply while an ETL process is in progress or you can set a database offline while an OLAP session is in progress to check for restore policies effectiveness. It enables testers to verify the Data Warehouse behavior after critical errors such as power leaks during update, network fault, and hard disk failures.

SECURITY TESTING

It mainly concerns the possible adoption of some cryptography technique to protect data and the correct definition of user profiles and database access grants. It checks for user profiles to be properly set up. It is also check for single-sign-on policies to be set up properly after switching between different analysis applications.

TESTING PHASES

- Business Understanding Document
  - a. High Level Test Approach & Test Estimation
  - b. Review Business & Technical Specification
- Test Plan Creation, Review and Walkthrough
- Test Case Creation, Review and Walkthrough
- Test Bed & Environment Setup
- Test Data Creation
- Test Predictions Creation, Review (Setting up the expected results)
- Test Case Execution
- Deployment
- Test Summary & Deliverables

CONCLUSION

Now-a-days every organization recognizes the significant advantages and values that data warehousing can provide for data analysis and decision making with complex operational systems. The implementation of dedicated data warehousing system ensures robust data management, faster query processing and dynamic report generations at any point of time with utmost quality. The well defined test approach & strategy for data warehousing system ensures that the data from Data Warehouse is uniform, accurate, and consistent and thereby serves as a "single version of truth" for the enterprise.

REFERENCES

- A. Mookerjea and P. Malisetty, (2008), "Best Practices in Data Warehouse Testing", In proceedings of Test 2008 conference, Delhi.
- Barbara Mento & Brendan Rapple, (2003), "Data Mining and Data Warehousing", DC: Association of Research Libraries, Washington.
- Chuck Ballard, Dirk Herreman, Don Schau, (1998), "Data Modeling Techniques for Data Warehousing".
- Hugh J. Watson, Thilini Ariyachandra, (2005), "Data Warehouse Architectures: Factors in the Selection Decision and the Success of the Architectures", College of Business University of Cincinnati, Ohio.
- Inmon, W.H., (1996), "Building the Data Warehouse", 4th Edition, John Wiley, New York.
- Kimball, R. (1996), "The Data Warehouse Toolkit", 2nd Edition, John Wiley, New York.
- Devlin, Barry, (1996), "Data Warehouse: From Architecture To Implementation", 1st Edition, Addison-Wesley Professional.
- Matteo Golfarelli & Stefano Rizzi, (2009), "A Comprehensive Approach to Data Warehouse Testing", In proceedings of DOLAP'09, Hong Kong, China.
- Surajit Chaudhuri & Umeshwar Dayal, (1997), "An Overview of Data Warehousing and OLAP Technology", Appears in ACM Sigmod Record.
- Vibhor Raman Srivastava, (2010), Testing Trends in Data Warehouse, In proceedings of Step-Auto 2010 conference, Pune.

REQUEST FOR FEEDBACK

Dear Readers

At the very outset, International Journal of Research in Commerce, IT and Management (IJRCM) acknowledges & appreciates your efforts in showing interest in our present issue under your kind perusal.

I would like to request you to supply your critical comments and suggestions about the material published in this issue as well as on the journal as a whole, on our E-mails i.e. **infoijrcm@gmail.com** or **info@ijrcm.org.in** for further improvements in the interest of research.

If you have any queries please feel free to contact us on our E-mail **infoijrcm@gmail.com**.

I am sure that your feedback and deliberations would make future issues better – a result of our joint effort.

Looking forward an appropriate consideration.

With sincere regards

Thanking you profoundly

Academically yours

Sd/-

Co-ordinator

ABOUT THE JOURNAL

In this age of Commerce, Economics, Computer, I.T. & Management and cut throat competition, a group of intellectuals felt the need to have some platform, where young and budding managers and academicians could express their views and discuss the problems among their peers. This journal was conceived with this noble intention in view. This journal has been introduced to give an opportunity for expressing refined and innovative ideas in this field. It is our humble endeavour to provide a springboard to the upcoming specialists and give a chance to know about the latest in the sphere of research and knowledge. We have taken a small step and we hope that with the active co-operation of like-minded scholars, we shall be able to serve the society with our humble efforts.

Our Other Journals