DATA WAREHOUSE TESTING APPROACHES

Dr. J. Hamed, Sahar Ahmed Al-Badani

Department Information Technology, Mansoura University, Egypt

ABSTRACT

Data Warehouse is a collection of large amount of data which is used by the management for making strategic decisions. The data in a data warehouse is gathered from heterogeneous sources and then populated and queried for carrying out the analysis. The data warehouse design must support the queries for which it is being used for. The design is often an iterative process and must be modified a number of times before any model can be stabilized. The design life cycle of any product includes various stages wherein, testing being the most important one. Data warehouse design has received considerable attention whereas data warehouse testing is being explored now by various researchers. This paper discusses about various categories of testing activities being carried out in a data warehouse at different levels.

Keywords

Data Warehouse, Data warehouse testing, Software Testing

1. INTRODUCTION

Data warehouse is collection of data from heterogeneous sources thus data analysis poses as a great challenge due to large volume and complexity of data. Testing is an essential part in the life cycle of a data warehouse. Data warehouse testing is carried out to check the quality of data. In data warehouse testing data is collected from heterogeneous sources therefore, the data exists in different forms and formats. Users that work on the data warehouse need to be ensured that the data has been collected and integrated properly from various sources and then transformed correctly in specific formats in order to remove inconsistencies and then stored in desired formats according to business requirements or design specifications [1]. Therefore data warehouse testing is carried out to remove the inconsistencies that occur due to data being collected in different formats from different sources which shall there by affect the decision making process. Data in the data warehouse is used for strategic decision making by higher managements of an organization. By carrying out data warehouse testing, data in data warehouse is transformed in the desired form and then analysed for strategic decision making purpose and management of resources. Various researches have been done on other phases of data warehouse but very little work has been carried out on testing of data warehouse.

As we all know that the data warehouse is a repository of any organizations historical data therefore, we may say that data is integrated in the data warehouse from various sources and various formats. Testing of all the integrated data is very exhaustive, therefore, selective data is filtered and extracted. Then it is transformed and loaded to carry out the data warehouse testing. For example, we take into account any banking industry, data warehouse testing helps in answering many business questions about geographic variations in choosing of banks, plan of their performance, choice of the customers, types of the customers, market share, profit analysis, etc. Thus if the data is very huge and critical it becomes challenging to integrate and test the data. In this competitive era, any organization should be able to review its historical data and monitor the real time functional data. If any bug occurs in later stage of testing it may lead to huge financial losses. Therefore, data warehouse testing is done in the early stages of integration of data thereby producing good quality data that is dependable for planning and making strategic or business or organization decisions.

Testing is an activity which is carried out with the objective to find errors in the given data [2]. To carry out the testing activity test plan and test cases need to be developed. To carry out successful testing of a data warehouse the following points needs to be taken care of:

- **Completeness of Data** – Whole data must be loaded to carry out Data warehouse testing. The data loaded must also be correct and validated for testing.

- **Transformation of Data** – The data loaded must comply with the requirements and design specifications and must be arranged according to business protocols defined.

- **Quality of Data –** The data loaded needs to be checked for quality as it shall be used for decision making purpose. The methods by which quality of data can be assessed are data correction, data substitution, data rejection and data notification. All these methods can be carried out without making any changes in the existing data.

- **Scalability of Data –**The data loaded should be checked for scalability. Scalability of a data warehouse is directly represented by the amount of data being queried and the number of concurrent users simultaneously running the queries.

- **Performance of Data –** To carry out successful testing, the data warehouse should be capable of handling large volume of data.

- **Risk Analysis –** During data warehouse testing, the risks involved should be taken into account and what would be the impact of these risks on the decision making process.

The layout of the paper is organized as follows; section 1 gives a brief introduction about data warehouse testing and the processes that should be carried out for a successful data warehouse testing. In section 2 all the work related to different types of testing approaches has been discussed. In section 3 comparison between software testing and data warehouse testing is done. Section 4 describes the various categories in which the testing approaches have been classified. Finally our work has concluded in section 5.

2. RELATED WORK

Data warehouse testing is the area which is being explored by the researchers now due to the need of the hour to test databases having enormous data and to take out the relevant data which can be used by the organizations for decision-making process. Some of the researchers have tried to juxtapose various testing approaches. Various consulting firms and IT companies are also lending a hand to promote research in data warehouse testing.

In this section we shall be briefly describing the various testing approaches that have been proposed by various researchers. We have compiled all the testing approaches in our research paper which will help the researchers to reconnoitre various areas of data warehouse testing for a better and improved understanding of its different phases that can be carried out in the data warehouse. The various testing approaches are described in the following section.

Data warehouse testing has been categorised into various phases as discussed in [8] . They are:

1. **Requirements Testing** – Emphasis had been given on defining business rules and requirements stated should be complete, unambiguous, clear, consistent and understandable.
2. **Unit Testing** – It is a white box testing. The developer loads the data from the data source, transforms the data according to the business rules and rejects the data which do not comply as per the defined requirements.
3. **Integration Testing** – The developer checks the initial data loaded and also the incremental data that keeps on loading due to modifications, updations and transformations. In this phase the quality of the system is tested from the lowest level. Thus this phase has again been divided in the following sub-phases:
    - Understanding of Requirements
    - Development of Test Plan and Test Design
    - Preparation of Test Case
    - Execution of Test Case

4. **Acceptance Testing** – It is the last phase of testing that specifies that there are no errors and the data warehouse system is fully functional and it should be accepted by the customer.

It has been discussed in [7] that Data warehouse testing should have both White box testing (tester has access to source code) and Black box testing (tester does not have access to source code). He has proposed the usage of V-Model to be used in Data Warehouse Testing. Various stages of data warehouse testing given are:

1. **Unit Testing** – The data is loaded from the data source into the warehouse and tested individually.

2. **Integration Testing** – The transformed data is gathered together into one data module.

3. **System Testing** – The resultant system as a whole is checked for errors and modifications.

4. **Acceptance Testing** – The system is checked whether it is performing as per the user requirements and user is satisfied with the system.

The author has also suggested that incremental approach is the best approach which should be used for testing because of huge complexity and large size of Data warehouse. He has laid emphasis on starting testing of data early during the development cycle of data warehouse as removing defects late in the development lifecycle may be costlier to handle.

Researchers in [10] have related the different phases of designing of data mart to data warehouse testing. The methodological framework as elaborated by the authors includes eight phases: Requirement analysis, Analysis and Reconciliation, Conceptual design, Workload refinement,

Logical design, Data staging design, Physical design and Implementation. The authors have briefly summarized the components that need to be tested first, thereby emphasizing on the importance of testing the design quality of the data warehouse. The various components listed are: Conceptual schema, Logical Schema, ETL procedures, Database and Front-end. On the basis of the listed components, seven testing activities have been elaborately defined by the authors. They are:

1. **Functional Test** – It checks that the data is loaded into the system as per the business rules.

2. **Usability Test** – The user interacts with the system to evaluate its ease of use and intelligibility.

3. **Performance Test** – The effectiveness of the system is tested under given workload environment.

4. **Stress Test** – It evaluates the performance of system under heavy workload environment.

5. **Recovery Test** – It evaluates how the system recuperates from crashes, failures, etc.

6. **Security Test** – The data in the system is secure under the given working conditions.

7. **Regression Test** – It verifies that the system is functioning properly after certain data updation and data manipulation.

[11] throws some light on testing the design quality along with testing the data quality of a data warehouse system. In regards to the above concern the items to be tested summarized by the author are: Multidimensional schema, ETL procedures, Physical schema and Front-end. To test these items, seven testing activities described by the authors are: Functional Test, Usability Test, Performance Test, Stress Test, Recovery Test, Security Test and Maintainability Test. Working on the concept of a prototype model, the testing activities summarized for the four items described are:

1. **Multidimensional Schema** – The various testing activities carried out for this item are: Workload Test, Hierarchy Test, Conformity Test, Usability Test, Nomenclature check, Performance Test, Early loading Test, Security Test and Maintainability Test.

2. **ETL Procedures** – The various testing activities involved for this item are: Code Test, Integrity Test,

   Integration Test, Administrability Test, Performance/Stress Test, Recovery Test, Security Test and Maintainability Test.

3. **Physical Schema** – The various testing activities carried out for this item are: Performance/Stress Test, Recovery Test and Security Test.

4. **Front-end** – The various testing activities involved are: Balancing Test, Usability Test, Performance/Stress Test and Security Test.

Through this research paper, the authors have been able to evaluate that the maximum testing time is consumed by the ETL activities and procedures as compared to other development phases.

It has been emphasized in [12] the importance of maintaining integrity of data in a data warehouse. If the data is not reliable then the reports that would be generated will not display correct result. Therefore, the author has divided the standard testing approaches into two categories in order to preserve the integrity of data. The approaches are:

1. Manual Sampling – The testing activities covered under this approach are:

   - End-to-End Testing – It checks that data is properly loaded into the systems from which data warehouse will extract data for nreport generation.

- Row count Testing – To avoid any loss of data, all rows of data are counted after the ETL process to ensure that all the data is properly loaded.

- Field size Testing – It checks that the data warehouse field should be bigger than the data field for the data being loaded, as if it is not checked will lead to data truncation.

- Sampling – The sample used for testing must be a good representation of whole data.

2. Reporting Tool – The testing activities covered under this approach are:

   - Report Testing – The reports are checked to see that the data displayed in the reports are correct and can be used for decision-making purpose.

The researchers in [13] have laid emphasis on the testing environment. The authors have described the importance of gathering requirements to test a data warehouse. On the basis of the requirements, test plan and test cases are created. The authors have described the various phases of data warehouse testing as:

1. Business Understanding
   a. High Level Test Approach
   b. Test Estimation
   c. Review Business Specification
   d. Attend Business Specification and Technical Specification walkthroughs
2. Test plan creation, review and walkthrough
3. Test case creation, review and walkthrough
4. Test Bed & Environment setup
5. Receiving test data file from the developers
6. Test predictions creation, review (Setting up the expected results)
7. Test case execution and (regression testing if required)
   a. Comparing the predictions with the actual results by testing the business rules in the test environment.
   b. Displaying the compare result in the separate worksheet.
8. Deployment
   a. Validating the business rule in the production environment.

The most important testing activities given by the author are:

- Data validation
- Regression Testing
- Oneshot/Retrospective Testing
- Prospective Testing
- View Testing
- Sampling

The authors discusses about the significance of data validation during the testing process in [14]. They have defined two approaches for validation of data during testing. They are:

1. Approach I – One should follow the data from the source to the target warehouse.

2. Approach II – One should follow the data from the source through the ETL process and then into the Target warehouse.

Various types of testing that can be carried out in the data warehouse have been discussed in detail in [5]. Also it has been specified that good data warehouse testing shall lead to good quality assurance in a data warehouse.

The various types of data warehouse testing described are:

1. Extraction Testing – It is carried out to check whether the required fields can be extracted by the data.

2. Transformation Testing – It transforms the data as per the expected logic.

3. Loading Testing – The transformed data is loaded in the data warehouse and checked.

4. End User Browsing Testing – It checks for the scheduled reports whether they are complete or accurate or not.

5. Ad-hoc Query Testing – The queries are created as per the expected functionalities and executed in a selected time frame.

6. Down Stream Flow Testing – It checks the updation of data in the data marts.

7. One Time Population Testing – It is one time testing of ETL applications and the checks that the data warehouse reports should match with the production reports.

8. End-to-End Integrated Testing – The end to end flow of data from the source to the target data warehouse should be complete and accurate.

9. Stress and Volume Testing – In this type of testing the robustness and capacity of the system is checked.

10. Parallel Testing – It is done when the data warehouse is executed on production data and the reports of the actual warehouse reports should be in sync with the reports of the production data.

11. Security Framework Testing – It checks for all features of security framework.

Thus we see that various testing approaches have been considered and carried out by various researchers. While in the execution of different testing approaches, sequence of implementation of these approaches have not been discussed and risk analysis approach has not been considered. These issues are addressed in this paper. The first step towards deciding the sequence of the testing approaches is to analyze the difference between software testing and data warehouse testing.

3. DIFFERENCE BETWEEN SOFTWARE TESTING AND DATA WAREHOUSE TESTING

Various authors in their recent researches have stated that there is a lot of difference between testing a software system and testing a data warehouse. Software Testing is an essential part of software development life cycle. It does not assure and error free software but plays a very important role. An error detected in the latter half of development would prove to be more costly as rectification is difficult as compared to if error is detected in the early stages of development. Data Warehouse Testing on the other hand focuses on delivering high quality data which shall be useful for making critical strategic decisions. Thus data cleansing, data verification, data validation and data formatting all play an important role in data warehouse testing. Some of the differences between software testing and data warehouse testing are listed below:

- Software testing is carried out prior to deployment of the software whereas data warehouse testing is a post deployment activity.

- Software testing focuses on testing of each use case which contains various test cases [3] whereas data ware house testing is focused on querying the test data loaded by the ETL process.

- Software testing is source code specific whereas data warehouse testing is content specific.

- Software testing can be carried out easily by testing the test cases generated on the other hand queries are triggered to carry out data warehouse testing.

- Tester bridges the gap between informal specifications and formal verification in software testing [3] whereas it is not possible in data warehouse testing due to large volume of database.

- In software testing it is possible to test the data storage and retrieval without having to go through the business logic and vice-versa [4] but this is not possible in data warehouse testing as all the tests are focused on business logic and data content.

- Software testing is carried out at the user interface and at business logic whereas data warehouse testing is only focussed on business logic.

- Software testing is user triggered as the input is given by the user and individual transactions are processed, on the other hand data warehouse testing is system triggered due to Extraction, Transformation and Loading process[5].

- Instant or overnight result can be obtained by execution of the transactions in software testing whereas transactions are processed at the back end in data warehouse testing and may take a long time to be processed.

- Software defects found later in development lifecycle increases the development cost whereas in a data warehouse the focus is on correctness of data on the basis of which critical business decisions shall be taken.

The common factor that needs to be considered in software testing and data warehouse testing is that both focus on the Requirement analysis phase. If the requirements defined are clear, understandable and consistent then it will lead to a successful software testing and data warehouse testing.

Thus we conclude that software testing and data warehouse testing are very different. In the next section we have classified the different testing approaches under various levels. All these levels are executed while carrying out data warehouse testing only the user needs to decide which type of testing has to be carried out at the specific level.

4. CATEGORIZATION OF TESTING APPROACHES

As seen from the data collected and analysed from various research papers, categorization of the testing approaches for a data warehouse can be done in various levels. At the application level the user interacts with the system and the testing approaches that can be applied are requirement testing, unit testing, integration testing, system testing and acceptance testing. Based on this approach the data warehouse testing approaches can be classified under five levels. The user can choose any one type of testing approach under each level for testing a data warehouse. The five levels on which testing approaches for a data warehouse can be classified are listed as follows:

1. Query level – When the data has been extracted and transformed in the data warehouse view testing, sample testing, report testing and adhoc query testing can be carried out to check the execution of queries at various levels and whether the user is getting the required output or not.

2. Functional level – To check the functionality of the data warehouse, verification and validation of the programs can be checked with the help of correctness testing, reliability testing, usability testing, integrity testing and downstream testing can be done.

3. Quality Assurance level – The quality of a data warehouse can be tested by carrying out Stress testing, security testing, regression testing, performance testing and volume testing.

4. Engineering level – The data warehouse can be assessed on the basis of efficiency testing, testability, documentation and structural testing.

5. Adaptability level – Adaptability and feasibility of the

   data warehouse to adjust in different environments can be checked by its flexibility, reusability, maintainability and recoverability.

The various categories of testing approaches have been classified in five levels as shown in Table 1. Each level has its impact and importance in testing a data warehouse. Execution of all the levels of testing may vary based on different user requirements but testing at all the levels is done by choosing any one type of testing approach at each level.

At the application level the data warehouse to be tested is subjected to requirement analysis through requirement testing and then based on the problem, the database of the data warehouse is formatted and any of the given testing – unit testing, integration testing , system testing and acceptance testing can be carried out. For example we have a bank data warehouse. It would contain data with regards to customer personal details, customer account number, customer transaction details, customer account details, customer ATM details and so on. Certain branch of the bank in a specific region has very few customers enrolled which is a matter of concern to the bank. In order to expand that branch in the specific region, a study needs to be carried out to find out its causes and thereby solutions can be planned by the management for increasing their clientele. Thus requirement testing can be carried out to study the business requirements and find out the reasons why people do not prefer this bank over other banks in their region. After the requirements have been studied carefully, based on the causes unit testing, integration testing, system testing or acceptance testing can be carried out.

The first level is the query level, where the queries related to the specific problems are executed in the database and results are analyzed. For example the requirement analysis reveal that the interest rate of the bank is less than the other banks and it is the prime reason for less clientele. Thus execution of queries for different problems may yield different solutions. Based on the database being queried any one type of testing can be performed like view testing, sample testing, report testing and ad-hoc query testing and reports related to queries can be prepared.

The second level is the functional level where the functionality of the programs are tested by taking any one of the testing approaches like correctness testing, reliability testing, usability testing, integrity testing and downstream testing based as per the given problem. The reports generated from these testing approaches shall guide the management to take different measures to expand their bank clientele.

Table 1. Levels of Testing Approaches

Query level	**Functional level**	**Quality Assurance level**	**Engineering level**	**Adaptability level**
View Testing	Correctness Testing	Stress Testing	Efficiency Testing	Flexibility
Sample Testing	Reliability Testing	Security Testing	Testability	Reusability
Report Testing	Usability Testing	Regression testing	Documentation Testing	Maintainability
Ad-hoc Query Testing	Integrity Testing	Performance Testing	Structure Testing	Recoverability
	Downstream Testing	Volume Testing		

The third level is the quality assurance level which assesses then quality of the data warehouse and works toward improving it. Good quality data warehouse testing shall result in correct results that can be used for making strategic decisions by the management. Any one of the given testing stress testing, security testing, regression testing, performance testing or volume testing can be carried out at this level to increase the quality of data warehouse testing for generating effective results. For example by carrying out performance testing at this level we concluded that by increasing the interest rate of the bank, customers will definitely enrol themselves. But if it does not happen then it is concluded that the quality of the data warehouse tested was not good as accurate results were not found. Then any other type of testing approach can be carried out at this level.

The fourth level is the Engineering level which is concerned with the physical implementation of data warehouse testing. Efficiency testing, testability, documentation testing or structure testing can be carried out at this level. Based on the user problem any one of the testing approaches can be chosen. For example by increasing the interest rate of the bank still more customers are not opting for it then by carrying out documentation testing more reasons can be found out for less clientele.

The fifth level is the Adaptability level where flexibility, reusability, maintainability or recoverability of various data warehouse testing approaches are carried out. For example a new module like internet banking is to be added in the bank structure then adaptability level testing can be carried out to check the compatibility of old data warehouse with the new data warehouse.

Thus we conclude, that at all the five levels the testing approaches are executed in the data warehouse testing. Based on the problem formulation any one of the testing approach is selected at each level and reports are generated which can be used by the management for strategic decision making purpose.

5. CONCLUSION

Enough research has already been carried out to study the different testing approaches that can be carried out in data warehouse testing. We have categorised all the testing approaches into five levels. In data warehouse testing, testing at all the five levels is always performed, only the testing approach taken at various levels may be different. Any one testing approach is chosen at each level and it also primarily depends upon the problem for which the data warehouse testing in being carried out. With the help of a bank architecture we have shown what is the importance of these approaches at various levels and how will they be useful for the higher management for planning and making strategic business decisions.

Basically we have tried to address the generic categories for a data warehouse testing framework. Each of the testing approaches chosen at each level is problem oriented and the focus is on improving quality of data. The reports generated help in gaining the user's faith and simplifying problems and ensure trust of the user from the tested data warehouse.

6. ACKNOWLEDGMENTS

Our thanks to all the experts who have contributed towards in this field. We would also like to dedicate our acknowledgment of gratitude towards the following significant advisors and contributors:

First and foremost, I would like to thank Dr. S.L Gupta for his most valuable support and encouragement. He kindly read my paper and offered invaluable detailed advices on grammar, organization, and the theme of the paper.

Second, I would like to thank Dr. Payal Pahwa for reading my research paper and providing valuable advices and for reproofing the paper..

Finally, I sincerely thank to my parents, family, and friends, who provide the advice and financial support. The product of this research paper would not be possible without all of them

7. REFERENCES

[1] Pooniah, P., 2001, "Data Warehousing Fundamentals – A Comprehensive Guide for IT Professionals", John Wiley & Sons, Inc.

[2] Myers, Glenford J., 1979, "The art of software testing", New York: Wiley, ISBN: 0471043281, xi, 177 p.

[3] Sneed M. Harry, 2006, "Testing a Data Warehouse – an Industrial Challenge", in proceedings of the Testing:

Academic & Industrial Conference on Practice and Research Techniques, IEEE Computer, p. 203-210

[4] Sneed M. Harry, 2006, "Reengineering for Testability", at workshop on Software Reengineering.

[5] Executive-MiH, "Data Warehouse Testing is different".

[6] Theobald J., 2007, "http://www.information management.com.

[7] Breue T., 2010, "Turning data into Dollars", XLNT Consulting.

[8] Brahmkshatriya K., 2007, "Data Warehouse Testing", in www.stickyminds.com.

[9] Mathen M. P., 2010, "Data Warehouse Testing", in www.infosys.com.

[10] Golfarelli M. and Rizzi S., 2009, "A Comprehensive Approach to Data Warehouse Testing", in ACM 12th international workshop on Data Warehousing and OLAP (DOLAP'09), Hong Kong, China.

[11] Golfarelli M. and Rizzi S., 2011, "Data warehouse testing: A prototype-based metholdology", on Information and Software Technology Journal, vol. 53, pp. 1183-1198.

[12] Ceres E., 2011, "Data Warehouse Testing: Why QA projects need Automation", white paper published for Real-Time Technology Solutions.

[13] Mookerjea A. and Malisetty P., 2008, "Data Warehouse/ETL Testing: Best Practices", www.pureconferences.com.

[14] Arbuckle S. and Cooper R., 2002, "How to thoroughly Test a Data Warehouse", in Software Testing Analysis and Review (STAREAST), Orlando, Florida.