Volume 4, Issue 12, December 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Lifecycle based Testing of Data Warehouse

Shraddha Saxena
JSS Academy of Technical Education, Noida
Uttar Pradesh Technical University, India

Ms. Sonali Mathur
JSS Academy of Technical Education, Noida
Uttar Pradesh Technical University, India

Abstract- Data warehouse testing (DWT) forms a very crucial stage in the DW development as all the important decisions are taken based on the information residing in Data warehouse (DW). As business information is available in diverse form but mostly in repositories of unstructured data. While developing the DW , immense amount of data is transformed, integrated , cleansed and structured into DW as a single set. These disparate changes could result into corrupt data or missing values. Also , testing is responsible for validation of validation of data.
Nevertheless, while most of phrases data warehouse development cycle have received deliberated amount of attention, work on testing the DW is still in the process. In this paper, we introduce a lifecycle based DWT and also implemented the same for a dataset using two different algorithm and comparision is made on these algorithm as well.

Keywords : Data Warehouse, Data Warehouse Testing, Lifecycle based Testing

I. INTRODUCTION

Data warehouse is a like a storehouse of data where data is present in many different forms. Data warehouse is subject oriented; integrated, time variant and non-volatile cluster of data on which various crucial decision making is performed. It is compound and associative data model that acquire the whole data of organization. The quality of real world data that is fed to DW is of major concern that comes from divergent sources. If any error or false value is present it may affect the analytical and strategic decision of the organization. Hence, testing is required for validating that DW meets the business and technical requirements.
Testing should focus on completeness of data, transformation of data, quality and performance of data. As agreed by most authors, the difference between testing DW and software system has many differences.

- DWT focuses on validating the data whereas software testing focuses upon the code.
- DW deals with tremendous amount of data.
- DWT is post placement activity whereas software testing is prior of placement of software.
- DWT is system triggered whereas software testing is user triggered.
- Test cases in DWT can be unlimited as different views of data whereas in software testing test cases are limited.
- DWT has a broader scope than software testing as it focus upon the correctness and usefulness of the data.

CHALLENGES IN DATA WAREHOUSE TESTING
As it is agreed that DW is totally different from other system, hence there are some challenges in testing the DW

- Lack of clarity on requirements.
- Lack of knowledge in testing tools
- Volume of data is huge
- Large amount of missing values
- Heterogeneous source of data
- Data present in different formats.

TESTING DATA WAREHOUSE
DWT is a process to figure out the quality of a DW and also it seeks for improvement by identifying the defects and problem. Testing implementation undergoes the wheel of unit testing, integration testing, system testing and usability testing. Also testing should focus on test cases generation and requirement gathering for the system.
The remainder of this paper is organized as follows: Section 2 briefly explains the survey of existing different techniques of DWT. Section 3 explains our proposed work. Section 4 explains advantages of our proposed lifecycle over the others. Section 5 is for Results obtained after implementing the proposed cycle. Finally, we will conclude our work in section 6.

II. RELATED WORK

Many trials has been made to address the DW testing process. Some companies published their white papers whereas researchers also published their work in this field.
In [1], author proposed a model based testing approach in which he lay stress on test cases generation. He build up the model in which he build some test cases and then test script is made and test cases run on that script.

In [2], author has proposed DWT framework which comprises of DW architecture analyser that studies the data received, test plan then splits into validation and verification phase. Validation manager involves the business experts and system experts where as verification manager involves system tester and DB administrator for assistance in process of Test Case preparation.

In [3], authors proposed object oriented framework which has two phases Requirement Level and Design Level. In requirement level , the requirement is gathered from different users and analysis is made. In Design Level, UML design is made by extracting the information from data gathered.

In[4], in this white paperauthor discuss the practices in DWT is mentioned which includes ETL process, Performance testing, user acceptance testing, report validation and also he discuss about the critical success factors like Risk based testing, data obfuscation, effective defect management, focus on automation.

In [5], author proposed the matrices which is categorized as WHERE, WHAT, WHEN categories will result in 3 dimensional matrix, the rows represent where dimension, the column reperesent the what dimension . He marked the performance in this matrix.

Table 1: DW testing matrices[5]

DW Testing Matrices

		Schema	Data	Operation
Backend	DS→ODS			
	ODS→DW			
Frontend	DW→DM			
	DM→UI			

In[6], author presented DW testing ttypes with respect to Dw development stages and focuses upon two high level aspects : underlying data and DW components. In underlying data he mentioned two points : data coverage and data complying with the transformational logic in accordance with hundred rules. Whereas in DW components he mentioned performance , scalability , component orchestration testing and regression testing.

In[7], author introduced data warehouse testing activities framed within a DW development methodology introduced. They stated that following components needs to be tested : conceptual schema, Logical Schema, ETL procedures, Database and frontend.

Table 2: DW components vs testing types[7]

DW components Vs testing types

	Conceptual schema	Logical schema	ETL procedures	Database	Front-end
Functional	✓	✓	✓		✓
Usability	✓	✓			✓
Performance		✓	✓	✓	✓
Stress			✓	✓	✓
Recovery			✓	✓	
Security			✓	✓	✓
Regression	✓	✓	✓	✓	✓
	Analysis & design		Implementation		

In [8], authors explained why data warehousing is different by mentioning about the user tiggerred vs. system triggered approach, volume of test data, test data preparation. He also explains different testing approaches like extraction testing, Transformation Testing, Loading Testing, End user testing and User browser testing and stress and volume testing.

III. PROPOSED WORK

The testing activities in data warehousing projects begin in the requirement gathering phase and carried out in an iterative manner. In data warehousing testing, every component of the project needs to be tested. Following is our proposed lifecycle for data warehouse testing :

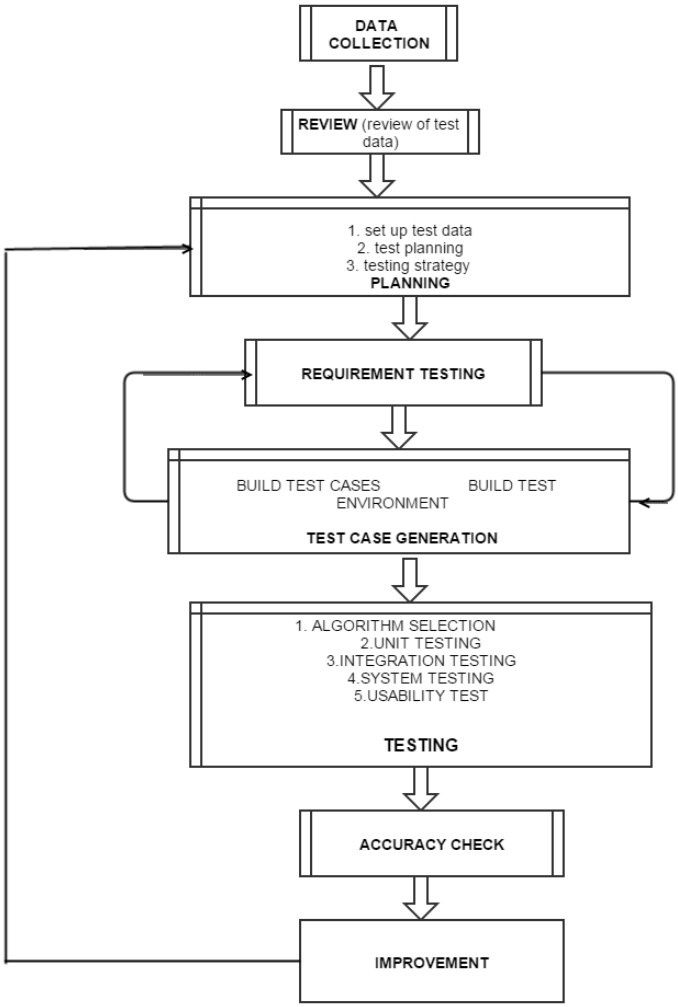


Figure 1: Proposed Lifecycle of Data Warehouse Testing

We are briefly described the above phases as follows:

1. ***DATA COLLCETION PHASE :***
   All the data is collected from different sources like sql server ,access sheets.

2. ***REVIEW***
   The data collected is reviewed for any correction present in the data. All data which is to be tested should be included in the data gathering.

3. ***PLANNING***
   In this phase, test data is prepared for testing . in this phase how testing is to be carried out is planned. Decision for selecting of which algorithm is made here.

4. ***REQUIREMENT TESTING***
   In this emphasis is given defining business rules and requirement stated should be complete, clear , consistent and understandable. Interface review must be done to check the usability of the system. Tools must be decided in this phase on which testing must be performed.

5. ***TEST CASE GENERATION***
   Different test cases is generated by using different combinations and according tools are set to operate.

6. ***TESTING***
   Different testing techniques are mentioned :
   - Algorithm selection : selection of algorithm from different set of algorithms
   - Unit Testing : In this white box testing is performed. The developer loads the data from data source, data modules is tested individually.
   - Integration testing : the process of combining and testing the components togerher one by one to check their integrity issues and to make sure they perform well working together.

- System Testing: this testing executes at developer site to make sure the system performs well. All the components run well together.
- Useablity testing : this is a black box testing. This checks for fulfillment of the requirements and ensure the validation of data.

7. ACCURACY CHECK

Data is tested against the accuracy of the data sets and performance is judged on the basis of different data mining attributes like precision, accuracy and time taken.

8. IMPROVEMENT

If any suggestion aur data is added then it must go through the whole cycle.

IV. ADVANTAGES OF PROPOSED TESTING LIFECYCLE FOR DATA WAREHOUSE

1. Test planning is done prior to the test cases development.
2. Strategy is prepared as all the test data should be covered in test cases so that defects can be recognized at the early stages.
3. Stress on Requirement testing is given as it is important for the need of development of the system and tools is studied for testing the data.
4. Testing the data on the basis of accuracy is the major part as in data mining there are many algorithms to choose , the best algorithm according to the need is chosen.
5. Since cycle is iterative in nature so it become easy for the testers to change the system according to the requirements or add on any strategy without any hindrance.

V. RESULTS

Testing of data warehouse is successful.
Testing of data is done using the proposed lifecycle.
Firstly, all data is collected and data is extracted. All the data is transformed according to the business rules, which is then loaded into the desired tool and algorithm is loaded and results are captured. Data is validated using this process.
Secondly, GUI (Graphical Interface) is made in which data sets is loaded and queried and result is displayed so as to ensure whether the validated data is producing correct results.
Thirdly, comparision is made using different algorithms.
In this scenario, we have used two algorithms

- Naiye Bayes and
- SVM (Support Vector Machine)

Naïve Bayes:
Briefly explaining naïve bayes classifier , In simple terms, a naive Bayes classifier assumes that the value of a particular feature is unrelated to the presence or absence of any other feature, given the class variable. For example, a fruit may be considered to be an apple if it is red, round, and about 3" in diameter. A naive Bayes classifier considers each of these features to contribute independently to the probability that this fruit is an apple, regardless of the presence or absence of the other features.

Support Vector Machine
In machine learning, support vector machines (SVMs) are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.
Here are some graphs used for interpretation

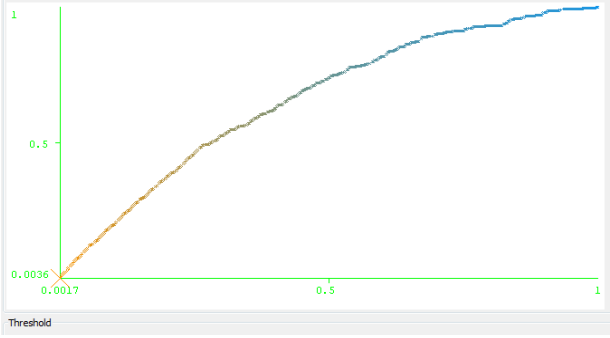


Figure 2 : threshold graph for naive bayes

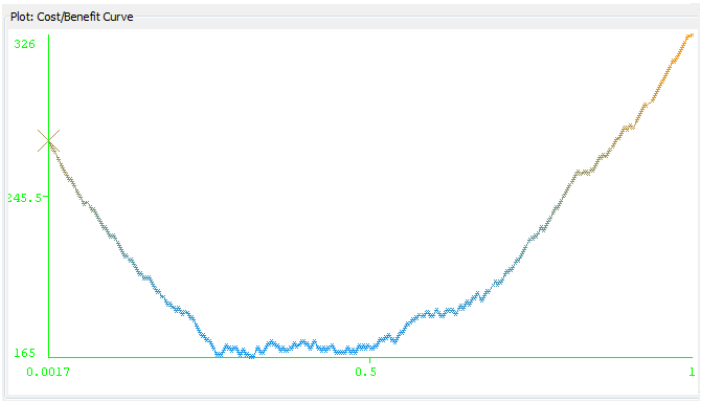


Figure 3: cost Benefit graph for Naïve Bayes

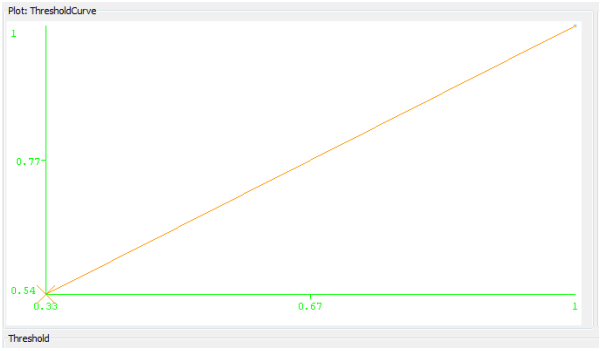


Figure 4 : threshold graph for SVM (SMO)

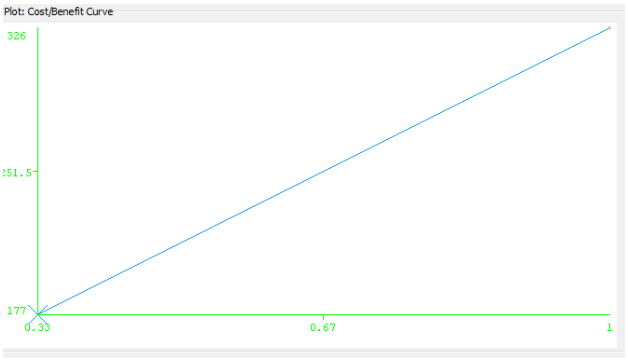
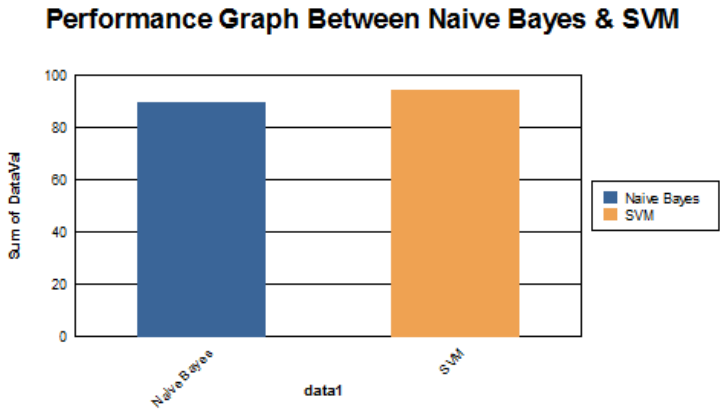


Figure 5: Cost Benefit Graph for SVM (SMO)

From the analysis we concluded that SVM (SMO) gives better accuracy than Naïve Bayes.



VI. CONCLUSION

From our work, we have concluded that data warehouse testing is a crucial phase of any data warehouse development process. Any bug found in later stage can effect the analysis of data. Hence it become very important to validate the data first and data is presented in GUI , can help the data analytic to do analysis at accurate rate and in simplified manner. This paper explain the iterative DWT life cycle. It also discusses DWT and there various stages and comparision is done between two algorithm using this data warehouse life cycle testing.

VII. FUTURE WORK ROADMAP

In future , we plan to conduct the studies on various data gathering techniques and would work on data gathering and will focus more on building a data warehouse with more improved testing strategies.

ACKNOWLEGMENT

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:

I would like to thank Ms. Sonali Mathur for reading my research paper and providing valuable advices and for reproofing the paper.

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.

REFERENCES

[1] Kuldeep Deshpande," Model Based testing of Datawarehouse", International Journal of Computer Science Issues (IJCSI), vol 10, Issue 2, March 2013

[2] Naveen ElGamal ," Data Warehouse Testing", EDBT/ICDT, March 2013

[3] Rajini Jindal and Shweta Taneja,"Comparitive Study Of Data Warehouse Design Approaches: A survey" , International Jornal of Database Management System (IJDMS), vol February 2012

[4] Syntel , " Proven Testing Techniques In Large Data Warehousing Projects : A white paper " , Syntel 2012

[5] Naveen ElGamal, Ali Ei Bastawissy and Galal Edeen, " Towards A Data warehouse esting Framework", Ninth International Conference On ICT ad Knowledge Engineering, IEEE 2011

[6] Manoj Philip Mathen, " Data Warehouse Testing : a white paper" , Infosys, March 2010

[7] 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.

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

[9] Weka tutorials : http://sentimentmining.net/weka/

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

[11] 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

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

[13] www.wikipedia.com

