ADVANCE TECHNIQUES PROCEDURE AND CHALLENGES FOR DATA WAREHOUSE OR ETL TESTING

Praveen Kumar-1 & Dr. Akash Saxena-2

1- Research Scholar, Department Of Computer Science And Engineering, at Jayoti Vidyapeeth Women's Universiy, Jaipur, India.
e-mail-praveen.agrawal15@gmail.com

2- Associate Professor, Department of Computer Science, Compucom Institiute of Information Technology, Jaipure-mail-akash27saxena@gmail.com

ABSTRACT

Testing is an integral part of the design lifecycle of any software system. While there are a number of software testing tools in the market, none of them address all aspects of data warehouse (DW) testing.

This paper presents some of the instructions, technique for etl testing / data warehouse testing and focuses on procedure of Data Warehouse Test which addresses the challenges which we faced while performing data warehouse testing. The paper focuses on the difference between database and data warehouse testing which creates popular misunderstanding that database testing and data warehouse is similar while the fact is that both hold different direction in testing and highlights its key features. It concludes with the number of universal verifications that have to be carried out for any kind of data warehouse testing.

KEYWORDS: Data Warehouse Testing, Techniques, Procedure, Challenges, Database Testing.

I. INTRODUCTION

Test Data Management has always been a challenge and when it comes to DW testing then it is reasonably more challenging. This is very tough to get correct, clean and secure data considering the fact that data is coming from multiple sources. In simple words it is collection of data from different departments of a company or from different technologies to one warehouse. Extract from source, transform in a suitable format and then load into destination data warehouse.

Today let me take a moment and explain my testing fraternity about one of the much in demand and upcoming skills for my tester friends i.e. ETL testing (Extract, Transform, and Load). This article will present you with a complete idea about ETL testing and what we do to test ETL process.

It has been observed that Independent Verification and Validation is gaining huge market potential and many companies are now seeing this as prospective business gain. Customers have been offered different range of products in terms of service offerings, distributed in many areas based on technology, process and solutions. ETL or data warehouse is one of the offerings which are developing rapidly and successfully.

II. NEED OF DATA WHAREHOUSE TEST

A data warehouse provides an integrated view of an organization's data and delivers actionable, timely and reliable business information for decision making. Poor quality data in the data warehouse can lead to wrong decisions affecting the organization's performance. Hence, Data Warehouse Testing assumes a significant role in the long term dependability in decisions.

III. ORGANIZATIONAL NEED OF DATA WAREHOUSE TEST

Organizations with organized IT practices are looking forward to create a next level of

technology transformation. They are now trying to make themselves much more operational with easy-to-interoperate data. Having said that data is most important part of any organization, it may be everyday data or historical data. Data is backbone of any report and reports are the baseline on which all the vital management decisions are taken.

Most of the companies are taking a step forward for constructing their data warehouse to store and monitor real time data as well as historical data. Crafting an efficient data warehouse is not an easy job. Many organizations have distributed departments with different applications running on distributed technology. ETL tool is employed in order to make a flawless integration between different data sources from different departments. ETL tool will work as an integrator, extracting data from different sources; transforming it in preferred format based on the business transformation rules and loading it in cohesive DB known are Data Warehouse. Well planned, well defined and effective testing scope guarantees smooth conversion of the project to the production. A business gains the real buoyancy once the ETL processes are verified and validated by independent group of experts to make sure that data warehouse is concrete and robust.

IV.   DATA WAREHOUSE TESTING IS CATEGORIZED INTO FOUR DIVERSE ACTIVITIES IRRESPECTIVE OF TECHNOLOGY OR ETL TOOLS USED:

- **New Data Warehouse Testing** – New DW is built and verified from scratch. Data input is taken from customer requirements and different data sources and new data warehouse is build and verified with the help of ETL tools.
- **Migration Testing** – In this type of project customer will have an existing DW and ETL performing the job but they are looking to bag new tool in order to improve efficiency.
- **Change Request** – In this type of project new data is added from different sources to an existing DW. Also, there might be a condition where customer needs to change their existing business rule or they might integrate the new rule.

- **Report Testing** – Report are the end result of any Data Warehouse and the basic propose for which DW is build. Report must be tested by validating layout, data in the report and calculation.

V.   DATA WAREHOUSE or ETL TESTING TECHNIQUES:

1. Verify that data is transformed correctly according to various business requirements and rules.
2. Make sure that all projected data is loaded into the data warehouse without any data loss and truncation.
3.   Make sure that ETL application appropriately rejects, replaces with default values and reports invalid data.
4. Make sure that data is loaded in data warehouse within prescribed and expected time frames to confirm improved performance and scalability.

Apart from these 4 main ETL testing methods other testing methods like integration testing and user acceptance testing is also carried out to make sure everything is smooth and reliable.

VI.   PROCEDURE OF DATA WAREHOUSE TESTING

Similar to any other testing that lies under Independent Verification and Validation, ETL also go through the same phase.

- Business and requirement understanding
- Validating
- Test Estimation
- Test planning based on the inputs from test estimation and business requirement
- Designing test cases and test scenarios from all the available inputs
- Once all the test cases are ready and are approved, testing team proceed to perform pre-execution check and test data preparation for testing
- Lastly execution is performed till exit criteria are met

- Upon successful completion summary report is prepared and closure process is done.

- It is necessary to define test strategy which should be mutually accepted by stakeholders before starting actual testing. A well defined test strategy will make sure that correct approach has been followed meeting the testing aspiration. ETL testing might require writing SQL statements extensively by testing team or may be tailoring the SQL provided by development team. In any case testing team must be aware of the results they are trying to get using those SQL statements.

VII. DIFFERENCE BETWEEN DATABASE AND DATA WAREHOUSE TESTING

There is a popular misunderstanding that database testing and data warehouse is similar while the fact is that both hold different direction in testing.

- Database testing is done using smaller scale of data normally with OLTP (Online transaction processing) type of databases while data warehouse testing is done with large volume with data involving OLAP (online analytical processing) databases.

- In database testing normally data is consistently injected from uniform sources while in data warehouse testing most of the data comes from different kind of data sources which are sequentially inconsistent.

- We generally perform only CRUD (Create, read, update and delete) operation in database testing while in data warehouse testing we use read-only (Select) operation.

- Normalized databases are used in DB testing while demoralized DB is used in data warehouse testing.

VIII. VALIDATION IN DATA WAREHOUSE TESTING

There are number of universal verifications that have to be carried out for any kind of data warehouse testing. Below is the list of objects that are treated as essential for validation in ETL testing:

- Verify that all DB fields and field data is loaded without any truncation

- Verify that for rejected data proper error logs are generated with all details

- Verify data checksum for record count match

- Verify that data transformation from source to destination works as expected

- Verify NULL value fields

- Verify data integrity

- Verify that duplicate data is not loaded

- Verify that expected data is added in target system

IX. CHALLENGES OF DATA WAREHOUSE TESTING

ETL testing is quite different from conventional testing. There are many challenges we faced while performing data warehouse testing. Here is the list of few ETL testing challenges I experienced on my project:

- *Huge volumes of data*: While transactional systems involve testing individual records, data warehouse testing involves huge volumes of data from heterogeneous sources, which significantly impact performance and productivity.

- *High cost of data quality:* A data warehouse is primarily used for taking tactical and strategic business decisions. Every defect that slips into production transforms into high costs for the organization when you are looking at huge volumes that can amplify a small error

- *Large scope of testing:* A transactional system may have a limited number of valid combinations of use cases. Since a data warehouse is aimed at supporting any number of views of data, the number of possible combinations of use cases is virtually unlimited.

- *Substantial business knowledge:* Transactional system testing focuses on program code, while data warehouse testing is directed at data. A good understanding of the data is required to ensure figures provided to the business user's queries are correct.

Data is important for businesses to make the critical business decisions. ETL testing plays a significant role validating and ensuring that the business information is exact, consistent and reliable. Also, it minimizes hazard of data loss in production.

X. CONCLUSION

- Data warehouses present an incorporated vision of an organization's data as well as distribute actionable, appropriate along with dependable production information for decision building.

- Test Data Management has constantly been a challenge and when it comes to Data warehouse testing; subsequently it is extra challenging. It is compilation of data from diverse departments of a company or from diverse technologies to one warehouse.

- Extract from resource, transform in an appropriate layout and then load into intention data warehouse.

- This paper covers various instructions, technique for ETL testing or data warehouse testing and also focused on procedure of Data Warehouse which help organizations at the stage of data warehouse testing.

- It concludes with the amount of entire verifications that have to be carried out for any type of data warehouse testing. This paper presented you with an entire plan concerning data warehouse testing along with what we execute to test data warehouse procedure.

- Deprived excellence data in the data warehouse is able to direct toward incorrect decisions distressing the organization's performance. Therefore, Data Warehouse Testing presupposes a major task in the long term constancy in decisions.

- ETL or data warehouse is one of the contributions which are rising speedily moreover fruitfully.

REFERENCES

[1] Burnstein, I.; Suwanassart, T.; Carlson, R., "Developing a Testing Maturity Model for software test process evaluation and improvement," Test Conference, 1996. Proceedings., International , vol., no., pp.581,589, 20-25 Oct 1996 doi: 10.1109/TEST.1996.557106

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

[3] Mustafa, K.M.; Al-Qutaish, R.E.; Muhairat, M.I., "Classification of Software Testing Tools Based on the Software Testing Methods," Computer and Electrical Engineering, 2009. ICCEE '09. Second International Conference on , vol.1, no., pp.229,233, 28-30 Dec. 2009 doi: 10.1109/ICCEE.2009.9.

[4] Holmes, A.; Kellogg, M., "Automating functional tests using Selenium," Agile Conference, 2006 , vol., no., pp.6 pp.,275, 23-28 July 2006, doi: 10.1109/AGILE.2006.19.

[5] Wissink, T.; Amaro, C., "Successful Test Automation for Software Maintenance," Software Maintenance, 2006. ICSM '06. 22nd IEEE International Conference on , vol., no., pp.265,266, 24-27 Sept. 2006 doi: 10.1109/ICSM.2006.63.

[6] Zhen Li.;Yong Hu Sun.;, "Use Selenium Grid to enhance testing of web applications", "IBM Techinical Library",07 June 2011.

[7] Sneed, H.M., "Testing a Datawarehouse - An Industrial Challenge," Testing: Academic and Industrial Conference - Practice And Research Techniques, 2006. TAIC PART 2006. Proceedings , vol., no., pp.203,210, 29-31 Aug. 2006 doi: 10.1109/TAIC-PART.2006.27.

[8] Kuhn, D.R.; Reilly, M.J., "An investigation of the applicability of design of experiments to software testing," Software Engineering Workshop, 2002. Proceedings. 27th Annual NASA Goddard/IEEE , vol., no., pp.91,95, 5-6 Dec. 2002 doi: 10.1109/SEW.2002.1199454.

[9] Caniupan, M.; Placencia, A., "Data Warehouse Fixer: Fixing Inconsistencies in Data Warehouses," Computer Science Society (SCCC), 2011 30th International Conference of the Chilean , vol., no., pp.28,32, 9-11 Nov. 2011 doi: 10.1109/SCCC.2011.5.

[10] Ramachandran, M., "Testing software components using boundary value analysis," Euromicro Conference, 2003. Proceedings. 29th , vol., no., pp.94,98, 1-6 Sept. 2003 doi: 10.1109/EURMIC.2003.1231572.

[11] Reid, S.C., "An empirical analysis of equivalence partitioning, boundary value analysis and random testing," Software Metrics Symposium, 1997. Proceedings., Fourth International , vol., no., pp.64,73, 5-7 Nov 1997, doi: 10.1109/METRIC.1997.637166.

[12] Min Chen; Xuedong Gao; HuiFei Li, "An efficient parallel FP-Growth algorithm," Cyber-Enabled Distributed Computing and Knowledge Discovery, 2009. CyberC '09. International Conference on , vol., no., pp.283,286, 10-11 Oct. 2009, doi: 10.1109/CYBERC.2009.5342148.

[13] Maity, S.; Nayak, A., "Improved test generation algorithms for pair-wise testing," Software Reliability Engineering, 2005. ISSRE 2005. 16th IEEE International Symposium on , vol., no., pp.10 pp.,244, 1-1 Nov. 2005, doi: 10.1109/ISSRE.2005.23.

[14] Glicker, S.; Hosch, F., "A design approach for a distributed test automation system," Applied Computing, 1990., Proceedings of the 1990 Symposium on , vol., no., pp.9,11, 5-6 Apr 1990, doi: 10.1109/SOAC.1990.82132.

[15] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

[16] AMELIA II: A Program for Missing Data. James Honaker, Gary King, and Matthew Blackwell. Version 1.7.2. June 8, 2013. (internet) http://cran.r-project.org/.

[17] Ying Wah Teh; Abu Bakar Zaitun; Lee, S.P., "Data mining using classification techniques in query processing strategies," Computer Systems and Applications, ACS/IEEE International Conference on. 2001 , vol., no., pp.200,202, 2001 doi: 10.1109/AICCSA.2001.933977.

[18] Feature Extraction, Construction and Selection: A Data Mining Perspective edited by Huan Liu, Hiroshi Motoda, Kluwer Academic Publishers, 2001.

[19] C Alexander. A primer on the orthogonal GARCH Model, 2000, URL http://www.icmacenter.rdg.ac.uk/pdf/orthogonal.pdf (internet) http://testng.org/.

[20] Unmesh Gundecha, "Distributed Testing with Selenium Grid" , Packt Publishing, November 2012.