

Testing of Extract Load and Transform (ETL) in Assorted Dimensions and Perspectives



Nanasaheb Mahadev Halgare

Abstract: In day today technical field, we are working on data science. It is the field that increasing rapidly, data science is similar to data mining but if we need to perform data mining then it is necessary to have data warehouse. And if we are interested to create data warehouse then we need to perform Extract Load and Testing (ETL). ETL implies Extraction of data from various sources, Transform that extracted data into proper format and finally load the data into data house. The integration of data science with the ETL is quite prominent and required so that the higher degree of performance can be attained. In addition, the performance elevation is very important to have the testing with more accuracy.

Keywords: Extract Load Testing, ETL testing, Test case, Bugs.etc

I. INTRODUCTION

Before starting of testing of ETL we need to know the basics of things like Business Intelligence, Data warehouse, ETL. As we know that the data extraction, loading and transformation is very important in any organization with the more elevation towards the performance.

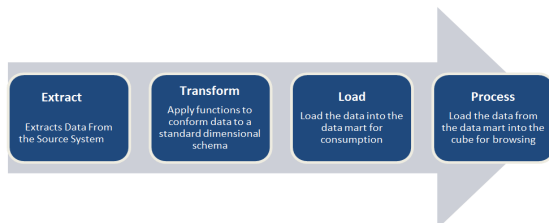


Figure 1.1 ETL Process and Key Perspectives

II. EXTRACTION TRANSFORMATION AND LOADING TESTING

The extraction of data is hugely important to have the higher performance and to have the key aspects with the elimination of dilemma of data processing with the loading and testing.

ETL TESTING PROCESS

There are various testing process similar to that of other testing process ETL testing process is there. It has different phases like shown in following figure 3.1.

Revised Manuscript Received on December 30, 2019.

* Correspondence Author

Dr. Nanasaheb M.Halgare*, Department of Information Technology, M. S. Bidve Engineering College Latur (Maharashtra), India. Email:halgare@gmail.com

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>)

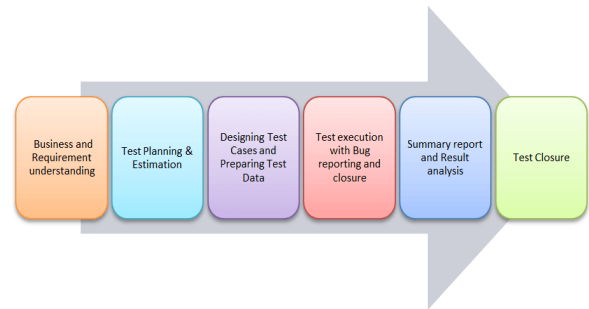


Figure 3.1: Process with ETL Analytics

Stages of ETL testing integrates the following key points and aspects

1. Analyzing the data sources as well as requirements
2. Gaining and extraction of data
3. Applying the dataset to the transformation
4. Key framing perspectives
5. Framing with the documents

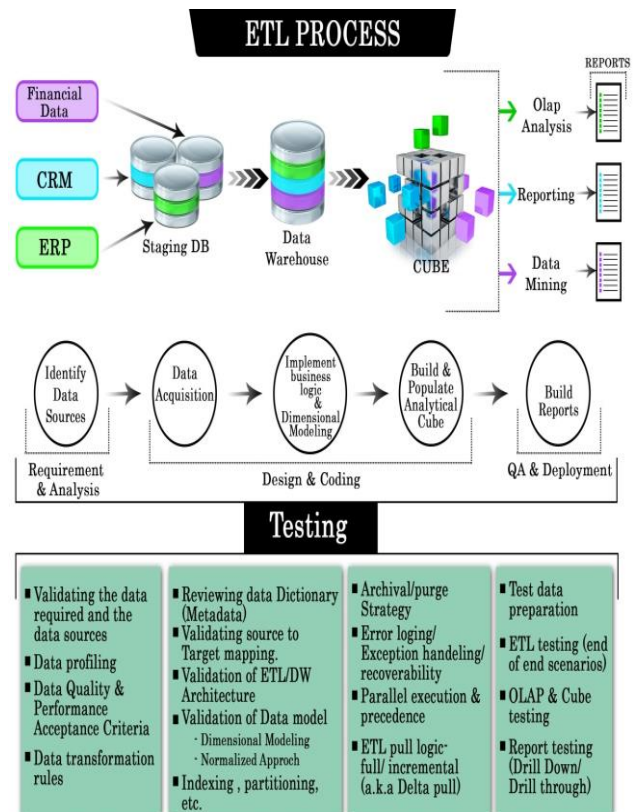


Figure 3.2: Key Perspectives

III. TYPES OF ETL TESTING

A. Production Validation Testing:

1. **Testing Process:** It is simply “Table balancing” or “production reconciliation” this type of testing is used to work provided to cavernous applications. To support business decision, whatever the data is available in management systems has to be in the correct sequence. Informatica Data Validation alternative provides the ETL testing automation and management abilities to ensure that the integration segments.

2. **Validation Testing**

3. **Upgradation to the Testing:**

4. **Metadata Testing:** This testing is used to test data type audit, data length audit and index/constraint audit.

5. **Data Completeness Testing:** To justify that all the normal

IV. HOW TO CREATE ETL TEST CASE

Extraction Transformation loading testing is a approach which is used to enforced variety of tools and databases in information management companies.

Table 4.1: Etl Test Scenarios And Test Cases

Test Scenario	Test Cases
Mapping doc validation	<ol style="list-style-type: none"> Dynamic document extraction and key formatting for ETL Key Processing Data Analytics Cavernous Processing
Validation	<ol style="list-style-type: none"> ETL integration validation process Validation analytics
Constraint Validation	Setup of objectives and constraints
Data consistency issues	Consistency analytics
Completeness Issues	<ol style="list-style-type: none"> Assure that all normal data is weighted into destination table. Analyze document calculation between origin and destination. Certify for any denied documents. Certify data cannot be deleted in the column of destination tables. Certify boundary value inquiry.
Correctness Issues	<ol style="list-style-type: none"> Misspelled or inaccurately data is recorded.
Transformation	Transformation
Data Quality	<ol style="list-style-type: none"> Number audit: Need to number certify and verify it Date Audit Data audit.
Null Validate	Validation with the null values

Duplicate Audit	<ol style="list-style-type: none"> Necessity to approve the uncommon key, primary key and any other attribute must uncommon as per the business needs are having any corresponding rows.
Data Cleaness	Unwanted attribute should be removed aforetime loading into the open space.
Date Validation	<p>Codes of dates are passed on in divers fields in ETL advancement</p> <ol style="list-style-type: none"> To detect the row generation date. To detect operative records as per ETL.
Complete Data Validation	<ol style="list-style-type: none"> To approve the entire data set in origin and destination table less a query in a better explanation.
CompleteData Validation	<ol style="list-style-type: none"> To approve the entire data set in origin and destination table less a query in a better explanation. We work out on origin less destination and destination less origin. If less query return any code those should be studied as discord tuples. Covet to duplicate tuples amid origin and destination applying converge statement.

V. TYPES OF ETL BUGS

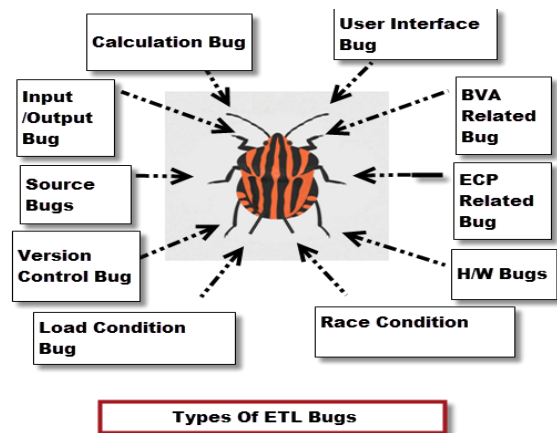


Figure 7.1: Bugs Analytics

VI. KEY ETL DIMENSIONS AS TESTING DOMAIN

- Performance of Data Processing
- Testing Elevations
- Testing Evaluations
- Key Perspectives with Testing in ETL
- ETL with the diversified Aspects
- ETL in addition to cavernous perspectives
- ETL in addition to the major data processing applications

- ETL in addition to the existing approaches for performance elevations

Responsibilities with Testing

- Test components of ETL data warehouse
- Execute backend data-driven test
- ETL tool based verifications
- Generate, layout and run test cases, verify plans and verify equipments.
- find the difficulties and implement clarification for capability problems
- Accept needs and layout blueprint.

VII. ETL ADMINISTRATION EXAMINATION AND CONFORM

The ETL key based administration and confirmation is required so that the overall performance and effectiveness can be elevated in addition to the existing aspects. The ETL based testing is quite prominent and needs and elevated segments so that the enormous aspects can be analyzed to the more accuracy levels.

VIII. KEY AUTOMATION ASPECTS WITH THE ETL

The key automation involves the presentation and association of fast integration for the ETL to have the performance aware integration. In addition, the processing is quite rigorous for the analytics patterns

IX. ENORMOUS DIMENSIONS WITH ETL PERSPECTIVES

The associated dimensions involves the performance factors with following points

- Efficiency
- Effectiveness
- Accuracy
- Cost Factor
- Cavernous Analytics

X. PROPOSED METHODOLOGY

The proposed methodology is associated with the integration of high performance tools and scripts to get the performance aware outcomes on the assorted parameters. The dimensions to include the analytics include the effectiveness, accuracy and related parameters using which the data processing the associated key factors are analyzed. The usage patterns of Python scripts and associated libraries are done so that the overall performance and evaluations can be done with higher degree of accuracy.

XI. RESULTS AND OUTCOMES

Table 11.1: Effectiveness Analysis

Execution Attempt	Effectiveness
1	99
2	98
3	97
4	95
5	96

6	97
---	----



Figure 11.1: Outcomes on Effectiveness

The tabular as well as the graphical results depict evaluations on assorted parameters with elevated performance.

Table 11.2: Accuracy Analysis

Execution Attempt	Accuracy
1	96
2	97
3	99
4	96
5	97
6	99

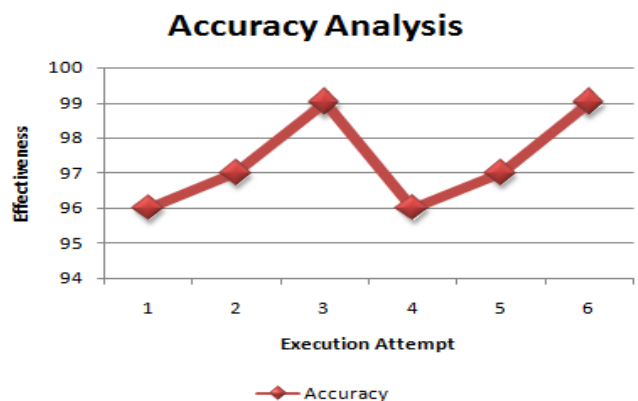


Figure 11.2: Outcomes Plot on Accuracy Parameter

XII. CONCLUSION AND SCOPE OF FUTURE WORK

The presented work is can be elevated to the effectiveness with integration of soft computing and metaheuristic approaches so that the overall performance can be improved. Following are few approaches which can be integrated for the soft computing and nature inspired techniques



- Social Cognitive Optimization (SCO)
- Bacteria Chemotaxis (BC) Algorithm
- Intelligent Water Drops algorithm, or the IWD algorithm
- Gases Brownian Motion Optimization
- Vortex Search Algorithm
- Bird Mating Optimizer
- MBO: Marriage in Honey Bees Optimization
- Krill Herd
- Harmony Search (HS)
- Grey Wolf Optimizer
- Cultural Algorithms (CA)
- Optics Inspired Optimization (OIO)
- Seed Based Plant Propagation Algorithm
- League Championship Algorithm (LCA)
- Firefly Algorithm (FA)
- Memetic Algorithm (MA)
- Particle Swarm Optimization (PSO)
- Simulated annealing (SA)
- Group Search Optimizer
- Parliamentary optimization algorithm POA
- Golden Ball
- Central Force Optimization
- Animal Migration Optimization (AMO) Algorithm
- Artificial Chemical Process Algorithm
- Self-propelled Particles
- The Raven Roosting Optimization Algorithm
- Forest Optimization Algorithm
- Bumble Bees Mating Optimization (BBMO) Algorithm
- Elephant herding optimization (EHO)
- Artificial Algae Algorithm (AAA)
- Strawberry Algorithm
- Boids
- Plant Propagation Algorithm
- Water Wave Optimization
- Bees Algorithm
- Artificial immune systems (AIS)
- Genetic Algorithm (GA)
- Invasive Weed Optimization (IWO)
- Bull optimization algorithm
- Bat Algorithm
- Social Spider Optimization (SSO)
- Shuffled Frog Leaping Algorithm (SFLA)
- Spider Monkey Optimization (SMO) algorithm
- Bacterial Colony Optimization
- Bacterial Foraging Optimization
- Ant Colony Optimization Algorithm (ACO)
- Cuttlefish Algorithm
- Chemical Reaction Algorithm
- Spiral Dynamic Algorithm (SDA)
- Bumble bees mating optimization BBMO
- Bacterial Evolutionary Algorithm (BEA)
- Artificial Chemical Reaction Optimization Algorithm
- Altruism Algorithm

- Flower pollination algorithm (FPA)

REFERENCES

1. O. Benjelloun, H. Garcia-Molina, D. Menestrina, Q. Su, S. E. Whang, J. Widom, "Swoosh: A generic approach to entity resolution", *The VLDB Journal*, vol. 18, no. 1, pp. 255-276, Jan. 2009.
2. E. Rogstad, L. Briand, R. Dalberg, M. Rynning, E. Arisholm, "Industrial experiences with automated regression testing of a legacy database application", *Software Maintenance (ICSM) 2011 27th IEEE International Conference on*, pp. 362-371, Sept 2011.
3. "Cost-effective strategies for the regression testing of database applications: Case study and lessons learned", *Journal of Systems and Software*, vol. 113, pp. 257-274, 2016..
4. W. J. Labio, H. Garcia-Molina, "Comparing very large database snapshots", *Tech. Rep.*, 1995.
5. S. Zhang, D. Jalali, J. Wuttke, K. Muşlu, W. Lam, M. D. Ernst, D. Notkin, "Empirically revisiting the test independence assumption", *Proceedings of the 2014 International Symposium on Software Testing and Analysis*, pp. 385-396, 2014.

AUTHORS PROFILE



Dr. Nanasaheb M. Halgare, Education Qualification Bachelor of Engineering (Information Technology), Master in Computer Science and Engineering, and PhD in Computer Science and Engineering. Having Experience of 15 years in teaching field, in which 8 years worked as a Head of Department of Computer Engineering in SND college of Engineering yeola Maharashtra and 5 years in VDF school of Engineering and Technology Latur Maharashtra. Area of interest Data mining and Data warehousing. Working as a Asst .Professor in M.S. Bidve Engineering College Latur, Maharashtra **INDIA**. Member of ijeat.org. Member of International Association of Engineers (IAENG).