ETL Testing

1.What is Data?

Data is collection of raw material (in unorganized format).

2.What is Data Ware House(DWH) used for?

Data ware house used for gather analytical operation. Avoid slowdown in transaction response times. Minimize the CPU cost.

3. Why Data Ware House (DWH)?

* Adaptation and implementation
* Finance and Budget allocation
* Sizeable ROI (Return on Investment)
* Technology Maturity

4. Explain what is Data Ware House (DWH)?

Bill Inmon is 1990 started DWH main goal of DWH are:

* Subject Oriented: related to the single subject
* Integrated: gather data from multiple places, multiple formats and multiple system.
* Non-Volatile: feature Historical data
* Time Variant: information should be available with prescribed time.

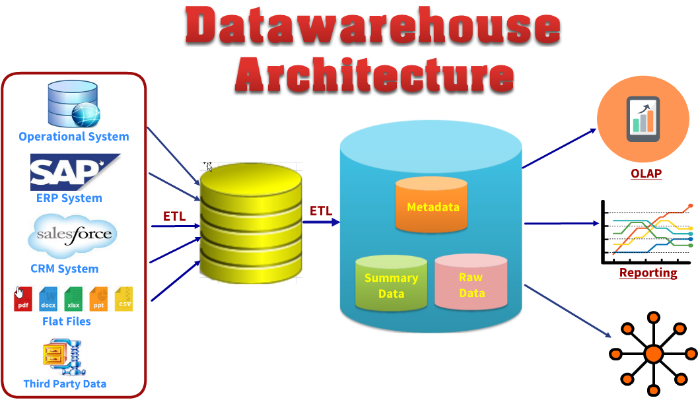
5. Key Features of Data Ware House(DWH)

* Structured data
* Better response time
* Available Historical Data
* Option for Ad-Hoc queries

6. Define Data Mart?

Data Mart is an implementation of Data Ware House which is designed small and more tightly restricted data which service single department of part of a organization. In other words, Data mart is the subset of DWH.

7. Data Ware House Architecture?



7.1 Single Department Data Mart (Federated)

* Logically consolidate but store in physical database
* Local Data Mart stores only relevant date of the department.

7.2 Multi- tiered Data Ware House

* Is distributed data approach.
* Many sources have to integrated to the central warehouse first. That means gather all the data to Central ware house and distributed according to the requirement using Data Mart.

8. Component of a Data ware house Architecture.

* Sources
* Data Extraction/ Transformation/ Load (ETL) Tool.
* Data ware House maintenance and administration tools.
* Data modelling tool or interface to external data model
* Ware house databases
* End-user data access and analysis tools.

9. Importance of a staging Area in Data ware House

* Complete depend on what kind of source is coming from
* Staging is a place or phase architecture where you hold temporary data on data ware house server. We need staging area to perform data cleansing and merging before loading the data into ware house
* Combine all the format to one format.

10. Describe how to structure the data in your Data ware House (DWH)

[Data Modeling]

It is the process that produce abstract date models for one or more database components of the data ware house.

Data modelling is a technique aimed at optimizing way to use in the Enterprise. It begins with identification of main data group i.e. Sales, product… for the Enterprise data ware house.

11. Why tester need to know data modelling?

* Function and technical aspects of database design
* Completeness in the design
* Understanding (BRD) scope of DB test
* DB test Execution 🡪 tables need to test
* Validation

12. Data Modelling Technique

1. ER Model (Entity- Relation Modeling)

* The ER Modeling technique used to illuminate the microscopic relationships among the data elements
* The highest art form of ER Modeling is to remove all the redundancy in the data.
* Traditional Modeling Technique
* Technique of choice for OLTP (Online Transaction Processing)
* Suited for cooperate date ware house (DWH)

2. Dimensional Modelling

* Analyzing business measures in the specific business context
* Helps visualize very abstract business question
* End users can easily understand and navigate the data structure.

13. Types of Schemas on Dimension model.

1. Star Schema
2. Snowflake Schema
3. Galaxy Schema / Fact Constellation

14. Difference Between Star Schema and Snow Flake Schema

|  |  |  |
| --- | --- | --- |
|  | Star Schema | Snowflake Schema |
| 1. Ease of maintenance / Change | Has redundant date. Hence less easy to change | No redundancy, so it is easy to maintain and change |
| 1. Easy to Use | Lower query complexity and easy to understand | More complex queries.  Less easy to understand |
| 1. Query Performance | Less number of foreign keys and short query execution time | More foreign keys. Longer query execution time (Slower) |
| 1. Types of Data Ware House | Good for Data Marts with simple relationship (1:1 or 1: many) | Good for Date Ware house core to simplify  Complex relationship (many: many) |
| 1. Joins | Fewer joins | Higher number of joins |
| 1. Dimension table | Contains only single dimension table for each dimension. | More than one-dimension table for each dimension. |
| 1. When to use | When dimension table contain less number of rows | When dimension table is relatively big in size. |
| 1. Normalization/ De-Normalization | Both Dimension and fact table are in Denormalized form | Dimension tables are in Normalized form but Fact table is in Denormalized form |
| 1. Data Mode | Top Down approach | Bottom up Approach |

Data Integration and ETL

1. What is Data Integration?

Data integration is to integrate the data from various sources into Data ware house or its equivalent system for single unified view for various purpose.

Various purpose might be Report generating and Analysis, letter/ Monthly statement generation.

* What happens

Discovery, Cleansing, monitoring, transformation and delivery of data form a variety of sources.

* Why?

with the increase in volume, variety and Velocity of data integration has become crucial step for may enterprises to be successful

* How does it help?
* Cleanse data and monitor data quality
* Understand your information and help in collaboration between Business and IT
* Transform data in any style and deliver it to any system.

1. What is ETL?

ETL stands for Extract, Transform and Load.

This is a general term used for moving data from one system to another or one database to another.

In Data Ware Housing term, ETL is part of Data integration where data is moved from transactional system to Data ware House or Data mart or operational Data store.

1. Data Transformation

* Data Merging: Join or Union
* Data Cleansing: Changing the inconsistency and inaccuracy, removing unwanted records.
* Data Scrubbing: amending and remove database, making more accurate and consistent
* Data Aggregation: Calculating the summary. (SUM, MAX, MIN)

Difference Between ETL (Extract Transform Load) and ELT (Extract Load Transform)

|  |  |  |
| --- | --- | --- |
|  | ETL | ELT |
| Flow | Data pipeline are used.  Transformation to the data one record at a time.  Intermediate data result is Stored in memory | Data loaded into destination Server  Set base process  Transformation and Lookups with in SQL. |
| Advantage | Complex transformation intermediate result in memory is faster than persisting to disk | The power of relational data base system can be utilized for very large dataset. |
| Disadvantage | Large data set could be overwhelmed the memory.  Updates are more efficient using set-based processing | Load on RDBMS  More disk activity. |
|  |  |  |

Difference between UI or GUI Testing and ETL Testing

UI (User Interface) or GUI (Graphical User Interface)

* Test the component as we see it.
* Open to visual, web page, online forms

DB testing(Database)

* Goes in the world of RDBMS (RelationalDatabase Management System) where we look at the tables, Storage, attributes and functionality.
* Primarily testable item which are in the backend.

**Types of DB Testing**

1. Structural: The structural data testing should involve the validation of all those element inside that are used primarily for storage of data and which are not allowed to manipulated by end user.
2. Functional: Combination of structural setup to that of the business required document.
   * 1. Mandatory Check
     2. Datatype
     3. Attribute naming convention
     4. Derived Attribute
     5. Overall process flow
     6. Data integrity and Data Consistency
     7. User restrictions
3. Schema Testing:

* Entity / Attribute Testing
* Views - Extract from different table
* Trigger
* Store procedure

1. Non-Functional: Splits into various categories as business requirement

* Stress
* Load Performance testing

Performance testing:

* Risk identification
* Max threshold - maximum capacity system can work

Load testing: to test the optimum response under huge number of uses from requirement.

# **Database Testing VS ETL Testing**

|  |  |  |
| --- | --- | --- |
| **Function** | **DB Testing** | **ETL Testing** |
| Primary Goal | Data validation and Integration | Data Extraction, Transform and Load for BI Reporting |
| Applicable System | Transactional system where business flow occurs. | System containing historical data and not in business flow environment |
| Common Tools | QTP, Selenium | Data Stage, Informatica, Pentaho, Talend etc. |
| Business need | It is used to integrate data from multiple application. | It is used for analytical reporting and information and forecasting. |
| Modeling | ER method | Multidimensional |
| Database type | It is normally used in OLTP (Online Transaction Process) system | It is applied in OLAP (Online Analytical Process) system |
| Data type | Normalized data with more join | Denormalized data with less joins. More indexes and aggregation. |

# Defect Management

After finding the Defect:

* Writing the Bug Summary
* Using appropriate language
* Steps to reproduce the defect
* Include Test Data
* Include Screenshot
* Severity/ Priority
* Logs
* Other information

## **Possible Reasons for Defects**

* Insufficient/ Incorrect Requirements
* Assumption / Mis-understanding
* Wrong decision during design
* Good Programmer with wrong Goal
* Complexity: 1. Business Logic 2. Technology
* Time Constraints: 1. Too much time 2. Too less time
* Frequently Changing Business Requirement.

## Issue/ Defect Management Process

* Discrepancy Reporting
* Communication the discrepancy
* Escalation
* Generating Metrices

Defect Categories

|  |  |
| --- | --- |
| **Defect Status** | **Description** |
| New | Issue get new status when the bug is opened by QA and assign to appropriate party |
| Open | The project Manager changes the status to open and assign the issue to the appropriate resources for resolution. |
| Fixed | The developer changes the status to fixed when developer fixes the issue in Development environment |
| Retest | The issue is ready to retest in Test Environment and assigned to the QA team. |
| Validated | The issue is Validate by QA Team |
| Reopen | The QA team finds that issue has not been resolved and reassign the issue back. |
| Closed | The QA team finds the issue to be resolved. |
| BA Review | The issue requires clarification in the business specification |
| Defer | The issue will be fixed in future release (PM approval required) |
| Not Reproducible | The issue cannot be reproduced more information required. |
| Duplicate | The issue already exists in QC |
| Training | The issue is related to training and is not issue with either application or environment. |

# Different ETL Testing Categories

* Meta data Testing
* Data completeness Testing
* Data quality Testing
* Data Transformation Testing
* Reference Data Testing
* Incremental Data Testing
* ETL Regression Testing
* ETL Integration Testing
* ETL Performance Testing