POINT:

1. ETL is used to populate the DWH
2. The tablename and mapping name will be present in technical documents (HLD, LLD, Technical specification etc)
3. The duration of running the mapping depends upon the data and the logic.
4. TOAD: Tool for Oracle Application Developer. It is a data access tool.In TOAD, F9 (to execute the query), F4 (table structure), Shift+delete(to delete the tables).
5. Transformation is nothing but a function.
6. Transformation will come after the qualifier and before target.
7. Negative cases use * instead of count(*).
8. Informatica does not support TRIM function, that the reason we use LTRIM and RTRIM for TRIM.

ETL TESTING LIFE CYCLE:

1. Requirement Understanding:
   BRD, Technical Specification (HLD and LLD), STM(Source and Target mapping doc)
2. Test Plan Creation (test Case Creation)
3. Test Data Creation – Real Time Data, provided by Client.
4. Test execution and reporting – Scheduling jobs

ETL TESTING

One of the Key elements contributing to the success of a Data Warehouse Solution is the ability of the Test team to Plan, Design and Execute a set of effective test that will help to identify multiple issues related to data inconsistency, Data quality failure in the Extract, Transform and Load (ETL) Testing.

- The primary focus of testing should be on the ETL process. This includes validating the data, loading of all required rows, correct execution of all transformations and successful completion of the cleaning operations.
- DWH projects largely concentrate on data and moving data from one stream to another stream to allow reporting.
- The sources of the data can be DB or flat file processed from other application.
-

OVERVIEW OF ETL OPERATIONS

1. **Extraction:** Extract is where data is pulled from source systems. (DB, flat files, XML files, Excel file etc)

2. **Transformation:-**Data transforms into new formats according to business rules. Apply business logic in this stage.
   Note: Transformations can be the most complex part of data warehousing.

3. **Loading:** Load is where data is located into the data warehouse.

TEST STRETEGY FOR ETL TESTING

There will be some standard tests for DWH that should be carried as part of testing for every DWH project. Strategies for testing ETL applications are identified as below:

1) ***DATA COMPLETENESS*** –Ensures that all expected data is loaded.
2) ***DATA TRANSFORMATION*** **–** Ensures that all data is transformed correctly according to business rules.
3) ***DATA QUALITY*** - Ensure that the ETL application correctly rejects, substitutes default values and reports invalid data.
4) ***INITIAL LOAD/FULL LOAD TESTING***
5) ***INCREMENTAL LOAD TESTING***
6) ***PRESENTATION LAYER (BI/REPORTING***) - Testing BI Reports in Data Warehouse testing.

7) ***INTEGRATION TESTING*** - Ensure that the ETL process function well with other Upstream and Downstream processes.
8) ***REGRESSION TESTING*** - This ensures the new data updates have not broken any existing functionalities.
9) ***UAT TESTING*** - Ensure the solutions meets user current expectations and anticipates their future expectations.

ABOUT INFORMATICA

START > PROGRAMS > INFORMATICA POWER CENTER 8.6.0
Current version in the market : 9.5

Informatics supports three clients/interfaces:

1) Power Center Designer
2) Power Center Workflow Manager
3) Power Center Workflow Monitor

Power Center Designer:

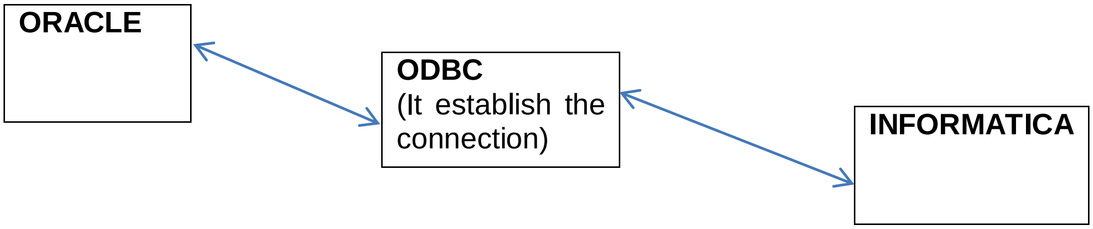
- In designer window, we (developer) will design/develop mappings (programs).
- Mapping: a mapping represents data flow from source to target.
- In designer window, we cannot execute the mapping, its only for designing.

Power Center Workflow Manager

- In this we will run/execute the workflow (mapping), which has been created in designer window.
- In this window, we cannot design the mapping, we will only be able to run the mapping.

Power Center Workflow Monitor

- We can monitor/check the status of the workflow (mapping)
- We can also execute the workflow but the workflow should have executed at least once from the workflow Manager.



How to run Workflow

1. Select the required workflow from workflow window.
2. Select the session or workflow and then right click on it.
3. It will display a popup window with multiple window with different options, from there you click on START TASK option.

How to see the status of the Workflow?

1. Go to the workflow MONITOR, select the respective session.
2. MONITOR the status column for respective session.

Informatica session status in workflow window:

1. RUNNING (EXECUTING/IN-PROGRESS)
2. SUCCEEDED (COMPLETED/PASSED)
3. FAILED
4. STOP
5. ABORT

How to get the log file?

Form workflow MONITOR WINDOW select required session than right click on "GET SESSION LOG" option.

How to see the status of the running session (how many records Pass/Fail)?

From workflow MONITOR WINDOW, select the required session, click on "get run properties" option.
It will display multiple properties from those properties select the "SOURCE/TARGET STATICES"

Can we run the workflow in the monitor window?

Right click > Rerun > Restart task

STM (SOURCE TARGET MAPPING DOCUMENT)

It is the heart of DWH. This document is for testing and development team. This is basically developed by developer in consultation with Architect/Moduler, the documents is then review by Architect/modular and ready for use.

STM-SOURCE & TARGET MAPPING DOCUMENT						
S.TABLE	**S.COLM**	**S.DATATYPE**	**BUSSINESS LOGIC**	**T.TABLE**	**T.COLM**	**T.DATATYPE**
EMP	EMPNO	NUMBER(5)	INTCAP()	EMP_DIM	ID	NUMBER(10)
EMP	ENAME	VARCHAR2(10)		EMP_DIM	EMPNAME	VARCHAR2(15)
EMP	SAL	NUMBER(5)		EMP_DIM	BASIC	NUMBER(25)
EMP	JOB	VARCHAR2(10)		EMP_DIM	DEPT	VARCHAR2(30)

Note:

1) If the business logic is null/blank it means straight pull
2) If the Target data type is empty it mean it can have same data type and size as of source.

How to cross verify the mapping are correct?

- ⇨ Through STM doc

What are the pre-requisite for testing:

1) Source and target table exists.
2) Source table should have data.
3) Target table should be empty
4) Mapping should be there.

DATA COMPLETENESS TEST (DCT)

It is designed to verify that all the expected data loads into the data warehouse. This includes running detailed tests to verify that all records, all fields and the full contents of each field are loaded.

TEST CASE IN DWH

Write the test case to "verify the source and target record count".

Step#	Description	SQL	Expected
1	Get Source Record Count	select count(*) from emp;	The query should be executed successfully and should display source record count. Note: note down the record count
2	Get Source Target Count	select count(*) from dept;	The query should be executed successfully and should display source record count. Note: note down the record count
3	Compare Source and Target Count		Both the count should be same

Write the test case to "verify the duplicate records in the target DB 'EMP_DIM'"

Step#	Description	SQL	Expected
1	check the duplicate records in the 'emp_dim' target table	select emp, count(*) from emp_dimgourp by empno having count(*)>1;	The query should be executed successfully and should return zero rows.

Write the test case to "verify source (emp) and target (emp_dim) table column mapping".

Step#	Description	SQL	Expected
1	Compare source and target column mapping	select empno, ename, sal from oltp.emp Minus select empno, ename, sal from emp_dim;	The query should be executed successfully and should return zero rows.

DATA TRANSFORM TEST (DTT)

It is a process of converting the data, cleansing the data into a required business format. Validating that data is transformed correctly based on business rules. Business rules can be the most complex part of testing in ETL application with significant transformation logic.

Between SOURCE and TARGET: - Test should make sure that the data type of each column of each table is as per the column mapping document. If no specific details are mentioned in STM document about the tables schema etc,then test should make sure on the below :

- ⇨ The data type of the source column and destination column are same.
- ⇨ The destination column length is equal to or greater that the source column length.

- ⇨ Validation should be done that all data specified get extracted.
- ⇨ Test should include the check to see that the transformation and cleansing process are working correctly. The following types of data transformation activities take place in staging.
  1. Data Cleansing
  2. Data Merging
  3. Data Scrubbing
  4. Data Aggregation

Data Cleansing

It is a process of changing inconsistence, inaccurate data and removing unwanted data.
The ultimate goal of data cleaning is to improve the organization confidence in their data.
List types of data error that needs to be addressed such as:

- ✓ Making first character as Capital letter.
- ✓ Decoding data
- ✓ Rounding the decimal data
- ✓ Missing data—Removing records which contains NULL's
- ✓ Data that contains unwanted junk such as an apostrophic or a comma or extra spaces.
- ✓ Telephone numbers in the wrong format.

Data Merging

It is process of integrating the data from multiple operational sources into a single output pipeline.

Data Scrubbing

It is process of deriving new attributes to meet their warehouse requirement.
Example: - Fact Table Column "SOLD_QTY" ,"SALES_AMOUNT"

Data Aggregation

It is process of calculating the summarized from details data.The following aggregation functions can be used to calculate the aggregates.
SUM (), COUNT (), AVG (), MAX (), MIN ()

Following are the list of transformations available in INFORMATICA.

1) Aggregator transformation
2) Expression transformation
3) Filter transformation
4) Joiner transformation
5) Router transformation
6) Sequence generator transformation
7) Sorter transformation
8) Look up transformation
9) Update strategic transformation

What is the difference between aggregator and expression transformation?

Aggregator transformation is applied on multiple rows whereas expression transformation is applied on single rows.

Aggregator transformation

This transformation is useful to perform calculations such as average, sum etc (mainly to perform calculations on multiple rows or groups.

Create a table in the source but cannot see the values in the target table?

- ⇨ The Commit command would have not done.

ROUTER TRANSFORMATION

This is similar to filter transformation. The only difference is:
- filter transformation drop the data that do not meet the condition whereas router transformation has an option to capture the data that do not meet the condition. It is useful to test multiple condition.

What is the approach of transformation in router transformation?

- ⇨ Based on the criteria, it will check and load the records.

JOINER TRANSFORMATION

- Belongs to flat file testing
- It is used to join two sources coming from two different locations or from same location.Example:

1) To join a flat file and relational source (table) and then load into target.
2) To join two flat files and then load into target

How to upload flat file:

1) Create the structure of the table.
   Create table joiner (id number (3), name varchar2 (10), locvarchar2 (10));
2) TOAD > DATABASE > IMPORT > IMPORT TABLE DATA > SELECT THE TABLE "JOINER" > NEXT > browse the file > NEXT > select the delimiter "tab" > NEXT" > select the first row as "2" (so that the column name should not be imported) > Preview > NEXT > Execute

How to insert the data in excel without separator?

1) Select 'NAME|LOC' from dual;
2) Copy paste the value to the excel
3) In excel, Data > text to column
4) Select delimiter radio button > NEXT > Other : | > NEXT > in the data preview, Shift+click on the grid > FINISH

TOAD:
1. BOOKMARK →SERVICES→C:\ROOT FOLDER
2. UTILITES→COMPARE FILES **(RED→ISSUE;BLACK→NO ISSUE; BLUE → ANY SPACES)**
NOTE: For COLUMN NAME IF RED ALSO NO ISSUES

FILTER TRANSFORMATION

FILTER Transformation can be used to filter rows in a mapping that meet the condition.
Example: Load only selected data from Source to Target.

LOOKUP TRANSFORMATION

It is used to lookup data in database table. Lookup definition can be imported either from source or target tables. Lookup can be either on source/target.
Example: load the required data into target from the lookup tables using source or target data.

In lookup table what type of queries do you write?

- ⇨ Joins

SLOWLY CHANGING DIMENSIONS (SCD)

SCD are also known as CDC (Change Data Capture)

MOD-It is a function used to generate a checksum values and it is faster.

TYPE-1: THE NEW RECORD REPLACES THE ORIGINAL RECORD. NO TRACE OF THE OLD RECORD EXISTS. IT'S LIKE OVERRIDING THE VALUES.

How to ensure the values/data is loaded to the target table:
Check the value in the EMP_CHK_SUM table, example, if the value 3333 is not getting displayed mean the data is not loaded to the target table.

SORTER TRANSFORMATION

It is an active and connected transformation used to sort the data. The data can be sorted in ascending or descending order by specifying the sort key. A port can be defined as sort key.

ACTIVE & PASSIVE TRANSFORMATION

Active Transformation

An Active transformation can change the number of rows that pass through it from source to target. It eliminates rows that do not meet the condition in transformation.
eg: filter transformation passes the rows from the source to target that which meet the filter condition.

Active Transformations list

- Source Qualifier Transformation
- Aggregator Transformation
- Filter Transformation
- Joiner Transformation
- Router Transformation

Passive Transformation

A Passive transformation does not change the number of rows that pass through it i.e., it passes all rows through the transformation.
eg: an expression transformation that performs the calculation on the data and passes all the rows from source to target.

Passive Transformations list -

- Expression Transformation
- Sequence Generate Transformation\

Tell me one mapping in which you have used multiple transformation and in the order?

- IN SCD's
- Source ➔ Sequence ➔ Filter ➔ Expresssion ➔Aggregator ➔ router
- Source ➔ Sequence ➔ Filter ➔ joiner ➔ Expresssion

DATA QUALITY:

- We say we have achieved the quality when we successfully fulfilled customer's requirements.
- Since in Data Warehouse Testing, the test execution moves around the data, so it is important to achieve the degree of excellence for the data and for that we do the data validation for both the data extracted from the source and then getting loaded into the target table.

DATA QUALITY is defined as "how the ETL system handles data rejection, substitution, correction". To ensure success in testing data quality, includes as many data scenarios as possible.
Note: Typically, data quality roles are defined during design.

Data Quality Rules:

- ✓ Reject the record if a certain decimal fields has nonnumeric data.
- ✓ Substitute Dept Name if a Dept No is given.
- ✓ Substitute -1 or 0 if source data has null or empty or spaces.
- ✓ Validate and correct the STATE field if necessary based on the ZIP Code.
- ✓ Compare product code values in lookup table and if there is no match load anyway but report to users.

Depending on the data quality rules of the application begin tested, scenario to test might include null key values, duplicate records in target data and involved data types in fields (Example Alphabetic characters in a decimal field)

Review the detailed test scenarios with business users and technical designers to ensure that all are on the same page.

Data quality rules applied to the data will usually be invisible to the users once the application is in production ; users will only see what's loaded to the database .For this reason, it is important to ensure that what is done with invalid data is reported to the business users.

SYSTEM INTEGRATION TESTING (SIT)

- It shows how the application fits into the overall flow of all UPSTREAM (DWH) & DOWNSTREAM (DATAMART) system.
- When creating integration test scenarios, consider how overall process can break and focus on touch points between applications rather that with in one application.
- Consider how process failures of each step would be handled and how data would be handled and how data would be recovered if necessary.
- Most issues during integration testing are either data related to or resulting from false assumptions about the design of other applications.
- Therefore, it is important to integration testing with production-like data. Real production data is ideal, but depending on the contents of the data, there could be a privacy or security concern that requires certain fields to be randomized before using it in a test environment.
- Run the overall process from end to end in the same order and with the dependencies as in production.
- Testing carried out by the test teams to make sure Source data corresponds to the final data as appeared in reporting.
- Included in this testing, should be data in the dimension and FACT tables in the DWH as well as the data presented through the presentation layer (BI Reports).