**What is a data warehouse, and why is it used?**

A data warehouse is a repository of data. The pieces of information stored are relevant to each other and support the decision making tree of some corporation or entity. It can incorporate multiple data sources to store all the data connected to the subject. Tipically it is composed by archive or historical data that can be analised. A data warehouse is supported on a database system.

**What are the basic stages of a data warehouse?**

The first stage to build a data warehouse is the initial data introduction, tipically this can be achived by copying some operational database. This is called and offline operational database. Then, we have to feed new sets of data to the newest created data warehouse. Therefore, this database is updated with large sets of data in a regular time basis (week, month). With this step, we’ve successfully built a offline data warehouse.

To achived a Real-time data warehouse you have to insert the operational data in real time. When this is integrated with the application, reporting on the data, it’s called a Integrated data warehouse.

**What is OLAP and OLTP, and which are their main differences?**

OLAP performs the analysis on the data, reporting the information. The focus on these kind of systems is the reading of data, thus using the SELECT database statement. OLTP manages the transaction system that collects the data. Actions like INSERT, UPDATE or DELETE are the focus here.

**What is a fact table?**

The fact table is a concrete measure that is tipically stored as numeric values, they have the core business information.

In detail, the fact table contains two different kinds of information. The foreign keys to the related dimension tables, providing joining relationships, and the measure columns which represent the added data.

**And a dimension table?**

Dimension tables describe the quantified data on the fact tables, giving context on its fields. They contain descriptive attributes which provide more information related to the fact table.

Fact tables have foreign keys to the dimension ones and the relation is one to many.

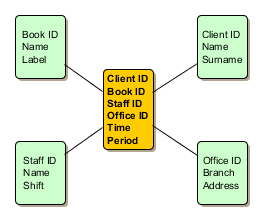
**Describe the star schema.**

In the star schema we have a centralized fact table and multiple dimensions linked to it. These dimensions are only related to the fact table, so the only link they have is to that specific table.

The fact table relates to the dimensions having their primary keys as foreign keys, and other extra attributes relevant to the data warehouse.

Therefore this kind of schema is denormalized and  better for simple querys, which are usually faster.

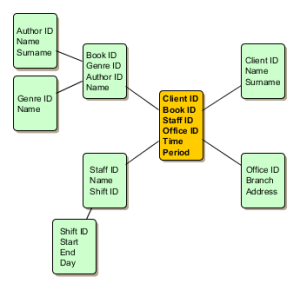
The next diagram represents a simple star schema implementation.

[](http://alldata-tprojects.rhcloud.com/wp-content/uploads/2015/01/starschema.png)

**Describe the snowflake schema.**

The Snoflake schema have links and relationships between dimensions, becoming a normalized organization of tables: fact and dimensions. This type of schema is usually more complex because each dimension can be composed of many other dimensions.

This kind of organization is explain in the next schema.

[](http://alldata-tprojects.rhcloud.com/wp-content/uploads/2015/01/snowflakeschema.png)

**What is a OLAP cube?**

A OLAP data cube is a representation of data in multiple dimensions, using facts and dimensions. It is characterized by the combination of information according to it’s relationship.

It can consist in a colection of 0 to many dimensions, representing specific data. There are five basic operation to perform on these kind of data cubes:

* Slicing
* Dicing
* Roll-Up
* Drill-Up and Drill-Down
* Pivoting

**Explain the slicing operation.**

The slicing operation on a OLAP Cube establishes a single value for one of the dimensions of the cube, selecting all the data that corresponds to the selected value.

So, by executing a slice on the cube we get all the selected dimension and fact information for the specific value assigned.

**Explain the dicing operation**

Dicing on OLAP Cubes consists on choosing an interval of values for some of the dimensions representing in the cube, and selecting the data that corresponds to those intervals.

This operation creates a subset of the cube which contains the data between the intervals.

**Explain the roll up operation.**

The roll-up operation performs some computing rules on the data of a OLAP cube specific dimension, returning the computed information to the end user.

These applied rules can be defined and summarize the information on that specific dimension.

**Explain the drill-up/drill-down operation**

These operations allow the exploration of information between the levels of data presented on dimensions and facts on the data warehouse.

It can select summarized information or the details that compose that data aggregation.

**Explain the pivoting operation**

Pivoting allows the rotation of the cube on its dimensions providing the user a different point of view of the explored data.

The cube can be rotated on every face.

**Explain the concept of data mart.**

Data mart is a specific group of data linked to a subject, which is part of a specific data warehouse. Therefore, a data warehouse have multiple data marts.

Basically a data mart is a small data warehouse with condensed information about a specific subject and it’s relationships. Usually each data mart is related to a department, business unit or something that can function individually within a data warehouse.

**Which are the reasons to create a Data Mart?**

There are various reasons that lead to a the creation of a data mart. The most important ones are:

* Create a data specific environment, providing easy access to it
* Easy to create
* Data is more relevant to users having only the essential information
* Lower cost than creating a whole data warehouse

**What does Normalization mean?**

Normalization is the process in which tables and fields are organized in a database in order to reduce the redundancy of stored data. Therefore many relationships between tables are defined, providing a better organized database system.

The key benefits of normalization are:

* Low database data redundancy
* Searching and indexing is faster
* Fewer null values since data is well distributed
* Cleaner and easier to mantain

**What is a ETL process?**

A ETL process consists on getting data from different sources and converting it to enter in a specific data warehouse.

Theses processes transform and normalize the data, providing a common base for all sources to integrate with a data warehouse.

**What is aggregation?**

Aggregation is the representation of a set of data, joined by some aggregation function.

This functions may be simple or complex depending of the purpose of the selected aggregation data. A simple function is the sum of every value.

**Explain what is partitioning.**

Partitioning is the process of dividing all data warehouse elements into smaller and distinct sets of data, keeping the relationships between the elements.

The benefits of partitioning are:

* Easy management
* Better performance
* Availability
* Easier backup and recovery

**What is the difference between metadata and data dictionary?**

A data dictionary has all the definitions of a database, the tables and fields, rows, number of rows, and that kind of information.

Metadata describes some kind of information with addicional and important data which is complementary.

**1) What is ETL?**

In datawarehousing architechture, ETL is an important component, which manages the data for any business process. ETL stands for **Extract, Transform** and **Load**.  Extract does the process of reading data from a database.  Transform does the converting of data into a format that could be appropriate for reporting and analysis. While, load does the process of writing the data into the target database.

**3) Mention what are the types of data warehouse applications and what is the difference between data mining and data warehousing?**

The types of data warehouse applications are

* Info Processing
* Analytical Processing
* Data Mining

Data mining can be define as the process of extracting hidden predictive information from large databases and interpret the data while data warehousing may make use of a data mine for analytical processing of the data in a faster way. Data warehousing is the process of aggregating data from multiple sources into one common repository

**5) What is fact? What are the types of facts?**

It is a central component of a multi-dimensional model which contains the measures to be analysed.  Facts are related to dimensions.

Types of facts are

* Additive Facts
* Semi-additive Facts
* Non-additive Facts

**7) Explain what is tracing level and what are the types?**

Tracing level is the amount of data stored in the log files.  Tracing level can be classified in two Normal and Verbose. Normal level explains the tracing level in a detailed manner while verbose explains the tracing levels at each and every row.

**19)  Explain what is the difference between OLAP tools and ETL tools ?**

The difference between ETL and OLAP tool is that

ETL tool is meant for the extraction of data from the legacy systems and load into specified data base with some process of cleansing data.

Example: Data stage, Informatica etc.

While OLAP is meant for reporting purpose in OLAP data available in multi-directional model.

Example: Business Objects, Cognos etc.

**22) Explain what staging area is and what is the purpose of a staging area?**

Data staging is an area where you hold the data temporary on data warehouse server.  Data staging includes following steps

* Source data extraction and data transformation ( restructuring )
* Data transformation (data cleansing, value transformation )
* Surrogate key assignments

**23) What is Bus Schema?**

For the various business process to identify the common dimensions, BUS schema is used.  It comes with a conformed dimensions along with a standardized definition of information

**24) Explain what is data purging?**

Data purging is a process of deleting data from data warehouse. It deletes junk data’s like rows with null values or extra spaces.

**25) Explain what are Schema Objects?**

Schema objects are the logical structure that directly refer to the databases data.  Schema objects includes tables, views, sequence synonyms, indexes, clusters, functions packages and database links

* **Complete backup** - It backs up the entire database at the same time. This backup includes all the database files, control files, and journal files.
* **Partial backup** - As the name suggests, it does not create a complete backup of the database. Partial backup is very useful in large databases because they allow a strategy whereby various parts of the database are backed up in a round-robin fashion on a day-to-day basis, so that the whole database is backed up effectively once a week.
* **Cold backup** - Cold backup is taken while the database is completely shut down. In multi-instance environment, all the instances should be shut down.
* **Hot backup** - Hot backup is taken when the database engine is up and running. The requirements of hot backup varies from RDBMS to RDBMS.
* **Online backup** - It is quite similar to hot backup.

**24. What is the difference between round-robin partitioning and Hash partitioning?**

In round-robin partitioning, data is evenly distributed among all the partitions so the number of rows in each partition is relatively same. Hash partitioning is when the server uses a hash function in order to create partition keys to group the data.

**26. What is lookup transformation and when is it used?**

Lookup transformation allows you to access data from relational tables which are not defined in mapping documents. It allows you to update slowly changing dimension tables to determine whether the records already exist in the target or not.

**27. What is a surrogate key in a database?**

A Surrogate key is something having sequence-generated numbers with no meaning, and just to identify the row uniquely. It is not visible to users or application. It is also called as Candidate key.

**28. What is the difference between surrogate key and primary key?**

A Surrogate key has sequence-generated numbers with no meaning. It is meant to identify the rows uniquely.

A Primary key is used to identify the rows uniquely. It is visible to users and can be changed as per requirement.

**29. If there are thousands of records in the source system, how do you ensure that all the records are loaded to the target in a timely manner?**

In such cases, you can apply the checksum method. You can start by checking the number of records in the source and the target systems. Select the sums and compare the information.

**30. What do you understand by Threshold value validation Testing? Explain with an example.**

In this testing, a tester validates the range of data. All the threshold values in the target system are to be checked to ensure they are as per the expected result.

Example − Age attribute shouldn’t have a value greater than 100. In Date column DD/MM/YY, month field shouldn’t have a value greater than 12.

**31. Write an SQL statement to perform Duplicate Data check Testing.**

Select Cust\_Id, Cust\_NAME, Quantity, COUNT (\*)  
FROM Customer GROUP BY Cust\_Id, Cust\_NAME, Quantity HAVING COUNT (\*) >1;

**32. How does duplicate data appear in a target system?**

When no primary key is defined, then duplicate values may appear.

Data duplication may also arise due to incorrect mapping, and manual errors while transferring data from source to target system.

**What is Regression testing?**

Regression testing is when we make changes to data transformation and aggregation rules to add a new functionality and help the tester to find new errors. The bugs that appear in data which comes in Regression testing are called Regression.

**Using SSIS ( SQL Server Integration Service) what are the possible ways to update table?**

To update table using SSIS the possible ways are:

Use a SQL command  
Use a staging table  
Use Cache  
Use the Script Task  
Use full database name for updating if MSSQL is used

**15. In case you have non-OLEDB (Object Linking and Embedding Database) source for the lookup what would you do?**

In case if you have non-OLEBD source for the lookup then you have to use Cache to load data and use it as source

**16. In what case do you use dynamic cache and static cache in connected and unconnected transformations?**

Dynamic cache is used when you have to update master table and slowly changing dimensions (SCD) type 1  
For flat files Static cache is used

**17. Name the three approaches that can be followed for system integration.**

The three approaches are − top-down, bottom-up, and hybrid.

**18. What are the different ETL Testing categories as per their function?**

ETL testing can be divided into the following categories based on their function −

Source to Target Count Testing − It involves matching of count of records in source and target system.

Source to Target Data Testing − It involves data validation between source and target system. It also involves data integration and threshold value check and Duplicate data check in target system.

Data Mapping or Transformation Testing − It confirms the mapping of objects in source and target system. It also involves checking functionality of data in target system.

End-User Testing − It involves generating reports for end users to verify if data in reports are as per expectation. It involves finding deviation in reports and cross check the data in target system for report validation.

Retesting − It involves fixing the bugs and defects in data in target system and running the reports again for data validation.

System Integration Testing − It involves testing all the individual systems, and later combine the result to find if there is any deviation.

**19. Explain the key challenges that you face while performing ETL Testing.**

Data loss during the ETL process.

Incorrect, incomplete or duplicate data.

DW system contains historical data so data volume is too large and really complex to perform ETL testing in target system.

ETL testers are normally not provided with access to see job schedules in ETL tool. They hardly have access on BI Reporting tools to see final layout of reports and data inside the reports.

Tough to generate and build test cases as data volume is too high and complex.

ETL testers normally doesn’t have an idea of end user report requirements and business flow of the information.

ETL testing involves various complex SQL concepts for data validation in target system.

Sometimes testers are not provided with source to target mapping information.

Unstable testing environment results delay in development and testing the process.

**20. What are your responsibilities as an ETL Tester?**

The key responsibilities of an ETL tester include −

Verifying the tables in the source system − Count check, Data type check, keys are not missing, duplicate data.

Applying the transformation logic before loading the data: Data threshold validation, surrogate ky check, etc.

Data Loading from the Staging area to the target system: Aggregate values and calculated measures, key fields are not missing, Count Check in target table, BI report validation, etc.

Testing of ETL tool and its components, Test cases − Create, design and execute test plans, test cases, Test ETL tool and its function, Test DW system, etc.

**21. What do you understand by the term ‘transformation’?**

A transformation is a set of rules which generates, modifies, or passes data. Transformation can be of two types − Active and Passive.

**22. What do you understand by Active and Passive Transformations?**

In an active transformation, the number of rows that is created as output can be changed once a transformation has occurred. This does not happen during a passive transformation. The information passes through the same number given to it as input.