Instructions to the Examiners:

1. **May the Examiners not look for exact words from the text book in the Answers.**
2. **May any valid example be accepted - example may or may not be from the text book**

1. Attempt *any two* of the following: **10**

a. State and justify the characteristics of a Data Warehouse with suitable examples.

– Subject oriented data
– Integrated data
– Time referenced data
– Nonvolatile data

b. Differentiate OLTP and OLAP.

OLTP Database	*Data Warehouse Database*
Designed for real-time business transactions and processes	Designed for analysis of business measures by subject area, category and attributes
Optimized for a common and known set of transactions, usually intensive nature; addition, updations and deletion of rows at a time, per table	Optimized for bulk loads and large complex, unpredictable queries that access many rows per table
Designed for validation of data during transactions, heavily influenced by business rules and database constraints and entity relationships	Designed to be loaded with consistent, valid data; uses very minimal validation routines but employs business empirical formulas for analytical purpose
Supports few concurrent users relative to the OLTP environment	Supports large user bases often distributed across geographies
Houses very minimal historical data	Houses a mix of most current information as well as historical data often regulated by the data purging and data retention strategies of the organization

c. Discuss the different types of facts with respect to measures stored in the fact table in a Data Warehouse.

- Additive - Measures that can be added across any dimension.
- Non Additive - Measures that cannot be added across any dimension.
- Semi Additive - Measures that can be added across some dimensions
- In the real world, it is possible to have a fact table that contains no measures or facts. These tables are called "factless fact tables", or "junction tables".

d. Why a dimension is called Slowly changing dimension?

Slowly Changing Dimension (SCD) refers to the fact that dimension values will change over time. Although this doesn't happen often, they will change and hence the "slowly" designation. For example, we might have an SKU assigned to a Super Ball made by the ACME Toy Manufacturing Company, which then gets bought out by the Big Toy Manufacturing Company. This causes the Brand that is stored in the dimension for that SKU to change. We have to specify how we want to handle the change. We will have the following three choices, which are related to the issue of whether or how we want to maintain a history of that change in the dimension:

Type 1: Do not keep a history. This means we basically do not care what the old value was and just change it.

Type 2: Store the complete change history. This means we definitely care about keeping that change along with any change that has ever taken place in the dimension.

Type 3: Store only the previous value. This means we only care about seeing what the previous value might have been, but don't care what it was before that.

(Any valid example may be accepted, not necessarily from the text book)

2. Attempt *any two* of the following: **10**

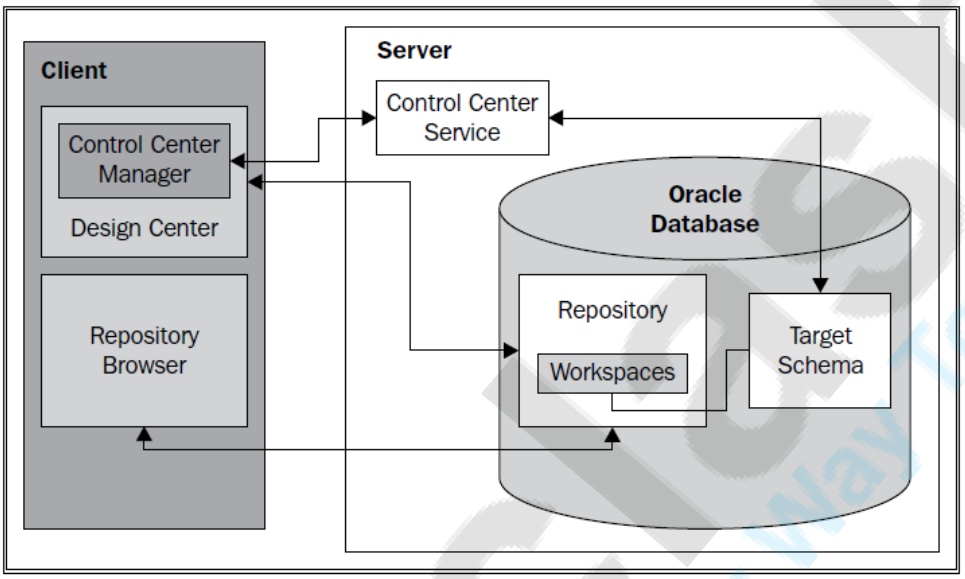
a. What is the relationship between OWBSYS and Oracle Warehouse Builder?

Oracle configures its databases with most of the pre-installed schemas locked, and so users cannot access them. It is necessary to unlock them specifically, and assign our own passwords to them if we need to use them. One of them is the OWBSYS schema. This is the schema that the installation program automatically installs to support the Warehouse Builder. We will be making use of it when we start running OWB. So under Password Management we see the OWBSYS schema and click on the check box to uncheck it (indicating we want it unlocked) and then type in a password and confirm it.

b. i) Name and define the utility that has to be configured before creating an Oracle database.

The listener is the utility that runs constantly in the background on the database server, listening for client connection requests to the database and handling them. It can be installed either before or after the creation of a database……..

ii) Draw a neat diagram that illustrates the various components of OWB.



c. What is the significance of HS parameters in the heterogeneous service configuration file?
The file named initdg4odbc.ora is the default init file for using ODBC connections. This file contains the HS parameters that are needed for the Database Gateway for ODBC

HS_FDS_CONNECT_INFO = <odbc data_source_name>
HS_FDS_TRACE_LEVEL = <trace_level>

The HS_FDS_CONNECT_INFO line is where we specify the ODBC DSN. So replace the <odbc data_source_ name> string with the name of the Data Source, which is for example ACME_POS.

The HS_FDS_TRACE_LEVEL line is for setting a trace level for the connection. The trace level determines how much detail gets logged by the service and it is OK to set the default as 0 (zero).

d. Explain the term module with reference to design of a DW in a Design Center.

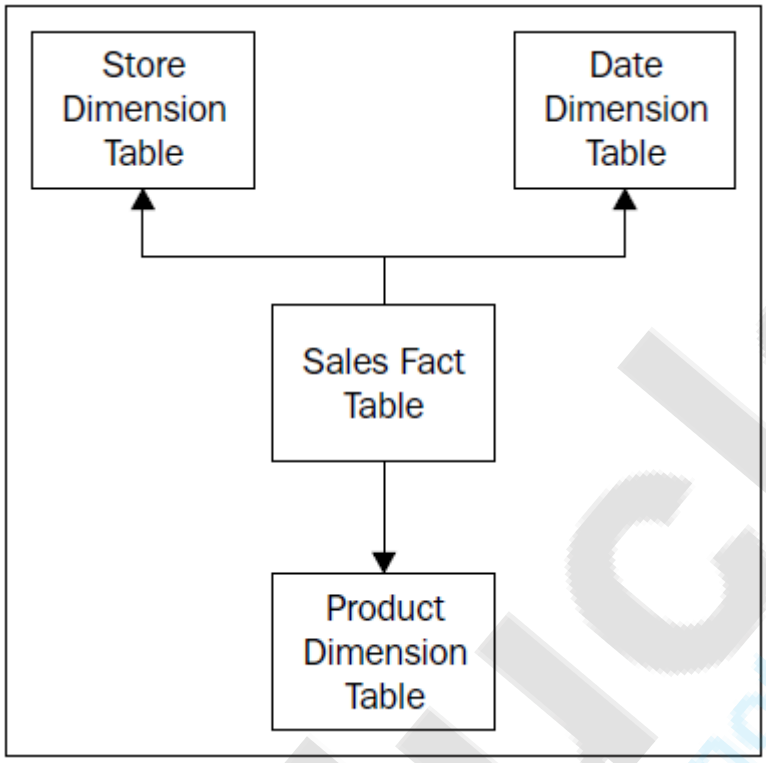
Creating a project is the first step. But before we can define or import a source data definition, we must create a module to hold it. A module is an object in the Design Center that acts as a storage location for the various definitions and helps us logically group them. There are Files modules that contain file definitions and Databases modules that contain the database definitions. These Databases modules are organized as Oracle modules and Non-Oracle modules. Those are the main modules we're going to be concerned with here.

We have to create an Oracle module for the ACME_WS_ORDERS database for the web site orders, and a non-Oracle module for the ACME_POS SQL Server database.

(Consider, if students write the steps to create these modules)

3. Attempt *any two* of the following: **10**

a. Explain the relational implementation of a dimensional model – Star Schema.



(students may explain a schema of this type)

b. Name and explain the objects that are relational and dimensional objects in an Oracle module.

In Project Explorer window with target Oracle module expanded, we can see a number of objects that are available:
There are **objects that are relational** such as **Tables, Views, Materialized Views, and Sequences**. Also, there are **dimensional objects** such as **Cubes and Dimensions**.

c. Every dimension has four characteristics that have to be defined in OWB. What are they?

Every dimension, whether time or not, has four characteristics that have to be defined in OWB:

- **Levels**
- **Dimension Attributes**
- **Level Attributes**
- **Hierarchies**

d. Explain the tabs Name, storage, Attributes, Levels and Hierarchies in Editor Window of any object that is currently being edited.

(students should explain these tabs with Dimension as an object or Cube as an object)

4. Attempt *any two* of the following: 10

a. What is ETL? Explain with an example.

1. Extract the data from the source system by some method.
2. Load flat files using SQL Loader or via a direct database link. Then we have to transform that data with SQL or PL/SQL code in the database to match and fit it into the target structure.
3. Finally, we have to load it into the target structure.

(more detailed explanation needed here)

b. Explain the data flow operators

i) Aggregator

Aggregator—there are times when source data is at a finer level of detail than we need. So we need to sum the data up to a higher level, or apply some other aggregation type function such as an average function. This is the purpose of the Aggregator operator. This is implemented behind the scenes using an SQL group by clause with an aggregation SQL function applied to the amount(s) we want to aggregate.

ii) Joiner

Joiner—this operator will implement an SQL join on two or more input sets of data. A join takes records from one source and combines them with the records from another source using some combination of values that are common between the two. We will specify these common records as an attribute of the join. This is a convenient way to combine data from multiple input sources into one.

iii) Expression

Expression—this represents an SQL expression that can be applied to the output to produce the desired result. Any valid SQL code for an expression can be used, and we can reference input attributes to include them as well as functions.

c. What is set in the Keys tab in the Table Editor window in OWB?

d. Explain the Indexes and Partitions tab in the Table Editor.

An **index** can greatly facilitate rapid access to a particular record. It is generally useful for permanent tables that will be repeatedly accessed in a random manner by certain known columns of data in the table. It is not desirable to go through the effort of creating an index on a staging table, which will only be accessed for a short amount of time during a data load. Also, it is not really useful to create an index on a staging table that will be accessed sequentially to pull all the data rows at one time. An index is best used in situations where data is pulled randomly from large tables, but doesn't provide any benefit in speed if you have to pull every record of the table.

A **partition** is a way of breaking down the data stored in a table into subsets that are stored separately. This can greatly speed up data access for retrieving random records, as the database will know the partition that contains the record being searched for based on the partitioning scheme used. It can directly home in on a particular partition to fetch the record by completely ignoring all the other partitions that it knows won't contain the record.

5. Attempt *any two* of the following: **10**

a. Discuss any three transformation operators used in ETL processing.

TRIM()
UPPER()
SUBSTR()

b. What is the role of a LOOKUP operator in a mapping?

Key Lookup operators, as the name implies, are used for looking up information from other sources based on some key attribute(s) in a mapping.

(students should explain the complete role of the operator with an example)

c. Explain Full and Intermediate generation styles.

The generation style has two options we can choose from, Full or Intermediate. The Full option will display the code for all operators in the complete mapping for the operating mode selected. The Intermediate option allows us to investigate code for subsets of the full mapping option. It displays code at the attribute group level of an individual operator. If no attribute group is selected when we select the intermediate option in the drop-down menu, we'll immediately get a message in the Script tab saying the following:

Please select an attribute group.

When we click on an attribute group in any operator on the mapping, the Script window immediately displays the code for setting the values of that attribute group.

d. i) Validation will result in one of the three possibilities. What are they?

The validation will result in one of the following three possibilities:

1. The validation completes successfully with no warnings and/or errors as this one did.
2. The validation completes successfully, but with some non-fatal warnings.
3. The validation fails due to one or more errors having been found.

ii) Mention the five default operating mode of the mapping.

The three modes are as follows:

• Set-based
• Row-based
• Row-based (target only)

The following two additional options for operating modes are available, which are based on the previous three:

• Set-based fail over to row-based
• Set-based fail over to row-based (target only)

6. Attempt *any two* of the following: **10**

a. What are the different operations that can be performed on a snapshot of an object that is created?

1. Restore
2. Delete
3. Convert to signature
4. Export
5. Compare

b. What happens if, let's say for example, a table definition is updated after we have defined it and created a mapping or mappings that included it?

Discuss Synchronizing Objects……

c. What is the difference between ROLAP and MOLAP?

MOLAP (Multidimensional Online Analytical Processing)

The MOLAP storage mode causes the aggregations of the partition and a copy of its source data to be stored in a multidimensional structure in Analysis Services when the partition

is processed. This MOLAP structure is highly optimized to maximize query performance. The storage location can be on the computer where the partition is defined or on another computer running Analysis Services. Because a copy of the source data resides in the multidimensional structure, queries can be resolved without accessing the partition's source data. Query response times can be decreased substantially by using aggregations. The data in the partition's MOLAP structure is only as current as the most recent processing of the partition.

ROLAP (Relational Online Analytical Processing)

The ROLAP storage mode causes the aggregations of the partition to be stored in indexed views in the relational database that was specified in the partition's data source. Unlike the MOLAP storage mode, ROLAP does not cause a copy of the source data to be stored in the Analysis Services data folders. Instead, when results cannot be derived from the query cache, the indexed views in the data source is accessed to answer queries. Query response is generally slower with ROLAP storage than with the MOLAP or HOLAP storage modes. Processing time is also typically slower with ROLAP. However, ROLAP enables users to view data in real time and can save storage space when you are working with large datasets that are infrequently queried, such as purely historical data.

d. Illustrate data explosion with reference to data storage in Data Warehouse.

Data explosion is the phenomenon that occurs in multidimensional models where the derived or calculated values significantly exceed the base values. There are three main factors that contribute to data explosion as listed below:

- Sparsely populated base data
- Many dimensions in a model
- A high number of calculated levels in each dimension

	YEAR
A	10
C	20
D	8
F	15
Total	53

Figure 1

	YEAR	Q1	Q2	Q3	Q4
A	10	10	0	0	0
B	20	0	20	0	0
C	8	0	0	8	0
D	15	0	0	0	15
A+B	30	10	20	0	0
C+D	23	0	0	8	15
A+B+C+D	53	10	20	8	15

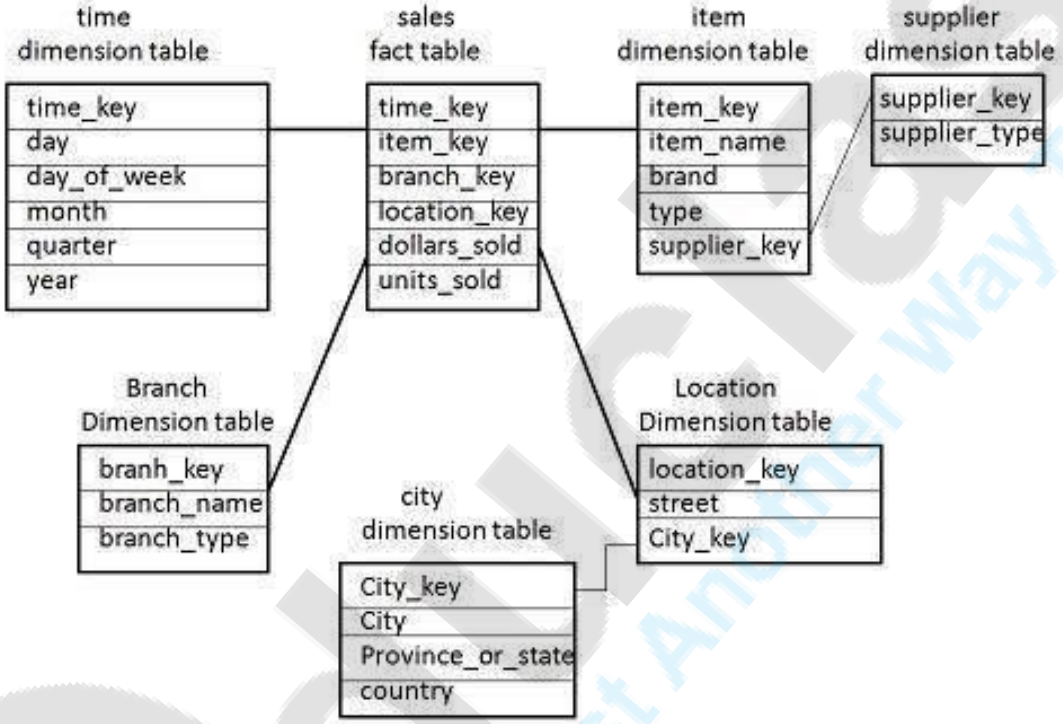
Figure 2

As demonstrated in Figures 1 & 2, we begin with our original 100% dense table. Suppose we wanted to find combination of sales figures for products A+B, C+D, and A+B+C+D. Suppose also that we wanted to figure this information over 4 quarters of the year. Notice how these have drastically added to the amount of data being stored in comparison to the original.

When ignoring data explosion, one of the major consequences is massive databases that are hundreds or thousands of times larger than is necessary. Also it requires expensive hardware to process and accommodate exploded data, which leads to expensive cost. Load or calculation times take much longer, which results a poor performance. The hidden cost may incur due to the failing to provide timely and relevant enterprise system. Therefore, the data explosion is the single most limiting factor in many large OLAP implementations

7. Attempt *any three* of the following: **15**

a. Explain Snow Flake Schema with an example.



(students may explain a schema of this type)

b. The first step required in making use of Oracle Heterogeneous Services used to access a Non-Oracle database is to create an ODBC connection. How is this connection established?

The first step that is required in making use of Oracle Heterogeneous Services to access a non-Oracle database using the generic connectivity agent is to create an ODBC connection. We do that by setting up a system DSN (Data Source Name). A DSN is the name you give to an ODBC connection. An ODBC connection defines which driver to use and other physical connection details that are needed to make a connection to a database. On Microsoft Windows, we configure DSNs in the ODBC Data Source Administrator.

(after this students may explain in short the steps for configuring DSN)

c Define and explain the term staging with respect to ETL processing.

Staging is the process of copying the source data temporarily into a table(s) in our target database. Here we can perform any transformations that are required before loading the source data into the final target tables. The source data could actually be copied to a table in another database that we create just for this purpose. This process involves saving data to storage at any step along the way to the final target structure, and it can incorporate a number of intermediate staging steps. The source and target designations will be affected during the intermediate steps of staging…….staging area…….

(students may explain the same with an example – to be considered accordingly if definition is not written)

d Differentiate Surrogate identifier and Business identifier with suitable examples.

Each dimension is represented in the cube attributes by a **surrogate identifier**, which is the primary key for the dimension, and the business identifier(s) defined for the dimension. The **business identifiers** that we specified when we designed our dimensions will identify the dimension record for this cube record and the surrogate identifier will be used as the foreign key to actually link to the appropriate dimension record in the database. Let's take a look at these attributes for the dimensions.

For example for the PRODUCT dimension, we have seen three attributes earlier that are product related—PRODUCT_NAME, PRODUCT_SKU, and PRODUCT. If we were to open our PRODUCT dimension in the Data Object Editor to view its attributes, we wouldn't see any attributes with exactly these names. The Warehouse Builder has provided us with attributes that represent the corresponding attributes from the dimension, but with a slightly different naming scheme. The PRODUCT_NAME and PRODUCT_SKU attributes correspond to the NAME and SKU attributes from the dimension that are the business identifiers we defined. The PRODUCT attribute corresponds to the ID attribute that was created automatically for us as the surrogate identifier for the dimension.

e The process of building the DW from the model created in the Warehouse Builder involves Deploying and Executing. Discuss the same.

The process of building the data warehouse from our model in the Warehouse Builder involves the following four steps:

- **Validating**: Checking objects and mappings for errors
- **Generating**: Creating the code that will be executed to create the objects and run the mappings
- **Deploying:** Creating the physical objects in the database from the logical objects we designed in the Warehouse Builder

- **Executing:** Executing the logic that is found in the deployed mappings for mappings and transformations

(students should explain deploying and executing in detail)

f Explain the change management related tool - the Metadata Loader.

With this feature we can export anything from an entire project down to a single data object or mapping. It will save it to a file that can be saved for backup or used to transport metadata definitions to another repository for loading, even if that repository is on a platform with a different operating system. Some other possible uses for the export and import of metadata are to quickly make copies of workspace objects in multiple workspaces for multiple developers, or to migrate to a newer product version. We would need to export from the old product version, upgrade the Warehouse Builder, and then import back to the new workspace version.