# Assignment

# *Deadline date:*

Your assignment must be uploaded/submitted before or on

### Rules for Marking

It should be clear that your assignment will not get any credit if:

* The assignment is submitted after due date
* The assignment is copied.
* The assignment is copied from net so write answer in your own wordings.

**Objective**

* To learn and understand Dimensional Modeling.
* To learn and understand Extract Transform Load.

***Assignment***

**Q.1 Explain the difference between Star schema, Star query and Multi-star schema. [15]**

The [**star schema**](http://www.stanford.edu/dept/itss/docs/oracle/10g/server.101/b10736/glossary.htm#i997131) is perhaps the simplest data warehouse schema. It is called a star schema because the entity-relationship diagram of this schema resembles a star, with points radiating from a central table. The center of the star consists of a large fact table and the points of the star are the dimension tables.

A [**star query**](http://www.stanford.edu/dept/itss/docs/oracle/10g/server.101/b10736/glossary.htm#i997128) is a join between a fact table and a number of dimension tables. Each dimension table is joined to the fact table using a primary key to foreign key join, but the dimension tables are not joined to each other. The optimizer recognizes star queries and generates efficient execution plans for them.

In a **multi-star schema**, the fact table has both a set of foreign keys, which reference dimension tables, and a primary key, which consists of one or more columns that provide a unique identifier for each row. The primary key and the foreign keys are not identical in a multi-star schema. This fact distinguishes a multi-star schema from a single-star schema.

**Q.2 In what situation Snowflake schemas are better to use than the star schema’s. What are the advantages of using Star schema? [4+6]**

Some situations are:

* If a dimension is very sparse (i.e. most of the possible values for the dimension have no data) and/or a dimension has a very long list of attributes which may be used in a query, the dimension table may occupy a significant proportion of the database and snow-flaking may be appropriate.
* A multidimensional view is sometimes added to an existing transactional database to aid reporting. In this case, the tables which describe the dimensions will already exist and will typically be normalized. A snowflake schema will hence be easier to implement.
* A snowflake schema can sometimes reflect the way in which users think about data. Users may prefer to generate queries using a star schema in some cases, although this may or may not be reflected in the underlying organization of the database.
* Some users may wish to submit queries to the database which, using conventional multidimensional reporting tools, cannot be expressed within a simple star schema. This is particularly common in data mining of customer databases, where a common requirement is to locate common factors between customers who bought products meeting complex criteria. Some snow flaking would typically be required to permit simple query tools such as Cognos Power-play to form such a query, especially if provision for these forms of query weren't anticipated when the data warehouse was first designed.

The main advantages of star schemas are that they,

* Provide a direct and intuitive mapping between the business entities being analyzed by end users and the schema design.
* Provide highly optimized performance for typical star queries.
* Are widely supported by a large number of business intelligence tools, which may anticipate or even require that the data warehouse schema contain dimension tables.

Star schemas are used for both simple data marts and very large data warehouses

**Q.3 what are the three main types of parallelisms** **as implemented in ETL applications? [15]**

There are **3 main types of parallelisms** as implemented in ETL applications:

**Data**: By splitting a single sequential file into smaller data files to provide parallel access.

**Pipeline**: Allowing the simultaneous running of several components on the same data stream. An example would be looking up a value on record 1 at the same time as adding together two fields on record 2.

**Component**: The simultaneous running of multiple processes on different data streams in the same job. Sorting one input file while performing reduplication on another file would be an example of component parallelism.

All three types of parallelism are usually combined in a single job.

***Note:***

* Assignment should be in your own wordings not copied from net, handouts or books.
* Your answer should be “to the point” and not more than 5 to 6 lines.
* Your assignment should be in doc format.