**1.** (a) What is a Data Warehouse and how can it benefit organizations? Discuss how data warehouse data differs from operational data.

**(5 marks)**

1. Describe the stages that data, on its way to the data warehouse, goes through in the ETL process.

**(10 marks)**

1. A data warehouse is required to accumulate information on the rental market in London. It will hold information on landlords and tenants (names), houses (no of rooms, type of housing (e.g. flat, house, etc)), location (e.g. Chelsea, Camden, etc), the date of tenancy agreement, and the rental price.. Create a star schema for the data warehouse to accommodate these requirements. The schema should indicate the dimensions and fact table with some reasonable attributes, the primary and foreign keys and relationships.

**(10 marks)**

**2** a) Define the following :

* + - Subject-Oriented
    - Integrated
    - Time-varant
    - Non-Volatile  **(6 marks)**

b) Describe the stages that data, on its way to the data warehouse goes through in the ETL Process. Why it this stage necessary. **(10 marks)**

c) Consider the following records from two different operational systems

|  |  |
| --- | --- |
| ProductCode: 3 057483 273523 | Product Code: 3057483273523 |
| Model Name: 208 | Model Name: Peugeot 208 |
| Price: £15,000 | Price: 20,000 (euros) |
| Supplier: A D Wicks | Supplier: Wicks |
| Address: 32, Wick Street, Kew | Add Line 1: 32 Wick St |
|  | Add Line 2: Kew |
| Postcode: TW9 3AB | Postcode: TW93AB |
| Telephone: 020 8546 5342 | Telephone: +4402085465342 |
| Product Code: P023456 |  |

Discuss the problems that arise from these records that will need to be resolved in the integration of the data from the two systems for a data warehouse system. **(9 marks)**

**Total 25 marks**

**3** a) In the context of data warehousing, what is the difference between operational and informational processing?

**(5 marks)**

1. Discuss how online transaction processing systems (OLTP) differ from data warehousing systems.

**(10 marks)**

1. Describe the stages that data, on its way to the data warehouse, goes through in the ETL process. **(10 marks)**

**Total 25 marks**

**SOLUTIONS**

**QUESTION 1**

1. What is a Data Warehouse and how can it benefit organizations? Explain how data warehouse data differs from operational data.

**[5 marks]**

A data warehouse is a database with archival, querying and data exploring tools (i.e., statistical tools) for storing historical and current data of potential interest to managers throughout the organization and from external sources (e.g. competitor sales or market share). The data originate in many of the operational areas and are copied into the data warehouse as often as needed. Data warehouses support looking at the data of the organization through many views or directions. For example a data warehouse might allow managers to look at products by customer by year by salesperson -- different slices of the data. Normal operational databases do not permit such different views.

b) Describe the stages that data, on its way to the data warehouse, goes through in the ETL process.***(10 marks)***

The ETL process goes through the stages of capture, scrub or data cleansing, transform and load.

The capture stage takes data from a data source either in its entirety or a selected subset of that data.

Frequently source data contains inaccuracies that are tolerant within the source system but not in the data warehouse. The scrub or data cleansing stage uses pattern recognition and AI techniques to remove the errors in the data, for example, misspellings, incorrect field usage, etc, to improve the data quality.

Data from different sources can need to be reconciled in terms of representation, for example, one data source can hold an address in single field whilst another holds the address in multiple fields. The transform stage reconciles data formats. This can be done at field level (one field into many and vice versa) and at the record level (selection, join and aggregation).

The final stage is load phase where the transformed data is placed in the data warehouse and indexed.

(c ) A data warehouse is required to accumulate information on the rental market in London. It will hold information on landlords and tenants (names), houses (no of rooms, type of housing (e.g. flat, house, etc), location (e.g. Chelsea, Camden, etc), the date of tenancy agreement, and the rental pric.. Create a star schema for the data warehouse to accommodate these requirements. The schema should indicate the dimensions and fact table with some reasonable attributes, the primary and foreign keys and relationships.

(10 marks)

One mark for each of the dimensions with primary keys identified.

One mark for the relationships with the fact table which should be 1:n

Two marks for the fact table with primary and foreign keys identified.

Two marks for a reasonable set of attributes.

**Diagram below :- substitute sale with rental, seller with landlord and buyer with tenant**



**QUESTION 2**

*4a) A data warehouse is ‘a subject-oriented, integrated, time-variant and non-volatile collection of data …’ (Bill Inmon 1993). Explain what the terms ‘subject-oriented’, integrated’, ‘time-variant’ and ‘non-volatile’ mean in this context.* ***(6 marks)***

**Subject-oriented:** deals with subjects rather than systems, e.g. customers, suppliers, orders, etc, rather than order processing.

**Integrated:** integrates data from a variety of sources including systems that are not computer based.

**Time-variant:** data is not updated. It is simply added to. Data is stored in versions

**Non-volatile:** all data is stored permanently.

*b) Describe the ETL process and why it is necessary* ***(10 marks)***

The ETL process stands for extract, transform and load. The extract phase extracts data from the various sources. Source data is not always correct, some parts may be missing, some data may be incorrect. Whilst this is tolerable in operational systems, the data must be correct in the data warehouse. So the next step is to clean the data to remove any inconsistencies. Because data for the data warehouse comes from different sources there is a need to reconcile the data. In transformation the data is converted into a common format. The load phase is where the data is loaded into the data warehouse.

The ETL process is necessary because a data warehouse integrates data from a variety of sources and hence there is a need to reconcile the data to create a data warehouse that is consistent and correct.

*c) Discuss the problems that arise from these records that will need to be resolved in the integration of the data from the two systems for a data warehouse system.*

***(9 marks)***

One bar code has spaces the other does not. Does the bar code need spaces or not?

One product name has the manufacturer’s name included the other does not.

One price is in pounds, the other in dollars. Which is significant and how will the currency be denoted?

Do we need to store the supplier’s initials?

Does the address need to be split up into separate lines?

Does the postcode need a space in the middle?

What should be the format for the telephone number – with or without the international dialling code?

What is the significance of the product code attribute?

**QUESTION 3**

*a) In the context of data warehousing, what is the difference between operational and informational processing?* ***(5 marks)***

Operational processing captures, stores and processes data to manage the day to day operations of an organisation. Informational processing is the analysis of summarised data or other forms of information to support the decision making processes within an organisation.

*b) Discuss how on-line transaction processing (OLTP) systems differ from data warehousing systems* ***(10 marks)***

OLTP systems are designed to maximise transaction processing capacity while data warehouses are designed to support ad hoc query processing. Organisations will normally have a number of OLTP systems for business processes such as sales, invoicing, accounts, etc, but, in contrast, an organisation will normally have a single data warehouse. The following table compares the two types of system:

|  |  |
| --- | --- |
| OLTP Systems | Dare Warehousing Systems |
| Holds current data  Stores detailed data  Data is dynamic  Repetitive processing  High level of transaction throughput  Predictable pattern of usage  Transaction driven  Application oriented  Supports day to day decisions  Serves large number of operational users | Holds historical data  Stores detailed, lightly summaries and highly summarised data  Data is largely static  Ad hoc, unstructured and heuristic processing  Medium to low level of transaction throughput  Analysis driven  Subject oriented  Supports strategic decisions  Serves relatively low number of managerial users |

Answers are not expected to take all these factors into account but should demonstrate the difference between the two types of system.

**(c)** *Describe the stages that data, on its way to the data warehouse, goes through in the ETL process.* ***(10 marks)***

The ETL process goes through the stages of capture, scrub or data cleansing, transform and load.

The capture stage takes data from a data source either in its entirety or a selected subset of that data.

Frequently source data contains inaccuracies that are tolerant within the source system but not in the data warehouse. The scrub or data cleansing stage uses pattern recognition and AI techniques to remove the errors in the data, for example, misspellings, incorrect field usage, etc, to improve the data quality.

Data from different sources can need to be reconciled in terms of representation, for example, one data source can hold an address in single field whilst another holds the address in multiple fields. The transform stage reconciles data formats. This can be done at field level (one field into many and vice versa) and at the record level (selection, join and aggregation).

The final stage is load phase where the transformed data is placed in the data warehouse and indexed.