Question 1:

a) Discuss three disadvantages and three advantages of a distributed database management system (DDBMS) **(6 Marks)** (Slides 10-11 in DDBMS, 19:10 Rec, Info)

Distributed Database Management Systems have lots of advantages, but some disadvantages associated with them. One advantage is improved reliability. Since the database is distributed about different locations, if one computer fails, the rest are still available. **(1)** They also have improved availability, so since data may be replicated, a failed node does not make that data inaccessible. **(1)** They also have improved performance as they can utilise 'parallel computing' to vastly increase CPU and I/O performance. **(1)**

One dis-advantages of DDBMSs is that they are expensive to produce. Due to the increased complexity and additional hardware, procurement costs are increased. **(1)** Another dis-advantage is reduced security. Given that the database is distributed about a network, it is harder to secure the whole network opposed to a centralised system. **(1)** The final dis-advantage is increased complexity. Given the complex nature of DDBMSs and their transparency to the user, if not handled correctly the aforementioned advantages will suffer degradation and ultimately become a dis-advantage. **(1)**

b) Compare and contrast homogeneous and heterogeneous distributed database management systems. **(4 Marks)** (Slides 13-14 in DDBMS)

With Homogenous DDBMS, all sites must use the same DBMS product. However, with Heterogeneous DDBMS, sites may run different DBMS products, with possibly different underlying data models. **(1)** Homogeneous DDBMS are much easier to design and to manage compared to Heterogeneous DDBMS. **(1)** The Homogeneous DDBMS approach provides incremental growth and allows increased performance. **(1)** Heterogeneous DDBMS occurs when sites have implemented their own database and integration is considered later. **(1)**

c) Discuss the advantages of fragmentation within distributed databases **(4 Marks)** (Slides 26-27 and 31-32 in DDBMS)

Fragmentation in distributed databases has several advantages, some of which include; usage, applications work with views rather than entire relations, meaning the database administrations can provide users with only the data they require in views. **(1)** Another advantage is efficiency. Data is stored close to where it is most frequently used, meaning an increase in CPU and I/O performance. **(1)** A third advantage is parallelism. With fragments as a unit of distribution, transactions can be divided into several subqueries that operate on these fragments. **(1)** Finally, security. Data required by local applications is not stored and so is not available to unauthorised users. **(1)**

d) Compare and contrast strategies for the placement of data in distributed database design. **(8 Marks)** (Slides 28-30 in DDBMS)

Complete replication has the highest storage and communication costs, but also has the best read performance, the highest locality of reference, and the highest reliability and availability. **(2)** Selective replication has satisfactory performance, subject to good design, whereas centralised replication has unsatisfactory performance. **(2)** Centralised has the lowest locality of reference, whereas fragmented replication is high. **(2)** Selective replication has low reliability and availability per item, but is high for the overall system. **(2)**

e) Identify the role of transparency in distributed database design **(3 Marks)**

Distribution transparency allows the user to perceive the database as a single, logical entity. **(1)** There are different types of transparency, distribution, fragmentation, location but they all play the role of perception. **(1)** The user would not be aware that the database is distributed, or fragmented, or where the data is stored. **(1)**

Question 2:

a) Identify the operational reasons for implementing a OODBMS as opposed to a RDBMS. **(4 Marks)** (Slides 48-67 in DDBMS & Slides 14-16, 18-19 in OODBMS)

One reason is that RDBMSs have a homogenous data structure and therefore assume both horizonal and vertical homogeneity. **(1)** Another reason to choose OODBMS is RDBMSs have poor support for integrity and have enterprise constraints. **(1)** RDBMSs have limited operations, which cannot be extended like in an OODBMSs. **(1)** Finally, RDBMSs have difficulty handling recursive queries. **(1)**

b) Explain the concept of Pointer Swizzling and its role in achieving acceptable performance in object DBMS's **(8 Marks)** (Slides 110-112 in OODBMS)

Pointer Swizzling is the action of converting object identifiers (OIDs) into main memory pointers. **(1)** It aims to optimise access to objects, thus increasing performance. **(1)** Pointer Swizzling attempts to provide a more efficient strategy by storing memory pointers in the place of referenced OIDs, and vice versa, when the object is written back to disk. **(2)** One method is to hold lookup tables that map OIDs to memory pointers using hashing, for example. **(1)** Once objects have been read into cache, we want to record that the objects are now in memory to prevent them from being retrieved again. **(1)**

Client-side Java handles Garbage Collection automatically, so there needs to be a mechanism in place that clears object references from memory on the database server side. Pointer swizzling can actually be a detriment to performance as if it not efficiently storing memory pointer in reference to OIDs, it can bloat the memory. It is finite.

c) Discuss three disadvantages and three advantages of an object-oriented database system (OODBMS) **(6 Marks)** (Slide 158-159 in OODBMS)

One advantage of OODBMSs is they have improved performance, benchmarks have shown that they have up to a 30x performance increase over RDBMS. **(1)** Another advantage is extensibility. They allow new data types to be built from existing types. This allows inheritance which reduces redundancy. **(1)** A third advantage is enriched modelling capabilities. Objects encapsulate state and behaviour, which is a more natural and realistic representation of real world objects. **(1)**

One dis-advantage is the lack of support for views. Considering that views provide many advantages such as data independence and security, this is negative. **(1)** Another dis advantage is the lack of support for security. OODBMSs currently do not provide adequate security mechanisms. Users cannot grant access rights on individual objects of classes. **(1)** Finally, a third dis-advantage is a lack of support for a Universal Data Model. There is no universally agreed data model for an OODBMS, so most models lack a theoretical foundation. **(1)**

d) The Object-Oriented Database Manifesto documents thirteen mandatory features for an OODMS. Identify seven of these features and analyse their impact on Database Design. **(7 Marks)** (Slide 100-102 in OODBMS, Info)

- Complex object must be supported **(1)** Non-Pojo
- Object identity must be supported **(1)**

- Encapsulation must be supported **(1)** Security
- Types or Classes must be supported **(1)**
- Types or Classes must be able to inherit from their ancestors **(1)**
- Dynamic Binding must be supported **(1)**
- The DBMS must support concurrent users **(1)**

CIETIDC (DICE-TIC)

Question 3:

a) What features of OLAP would support a national supermarkets chain business Analytics? **(6 Marks)** (Slides 10 or 11 in OLAP)

One feature is multi-dimensional views of data. These views provide a basis for analytical processing through flexible access to corporate data. They are also a core requirement of building a realistic business model. **(2)** Another feature beneficial to business analytics is OLAPs support for complex calculations. These computational methods are required for sales forecasting, which uses trend algorithms such as moving averages and percentage growth. **(2)** A third feature of OLAP is time intelligence. This can provide supermarket chains the ability to produce year-to-date and period-over-period comparisons of data, which is useful for sales analysis. **(2)**

b) Analyse the statement, "OLAP is just an extended set of grouping functions". **(6 Marks)** (Slides 53-67 in OLAP)

Aggregation is a fundamental part of OLAP. To improve aggregation capabilities, the SQL standard provides extensions to the GROUP BY clause such as the ROLLUP and CUBE functions. **(2)** ROLLUP supports calculations using aggregations such as SUM, COUNT, MAX, MIN and AVG at increasing levels of aggregation. **(2)** CUBE is similar to ROLLUP, enabling a single statement to calculate all possible combinations of aggregations. CUBE can generate the information needed in cross-tabulation reports with a single query. **(2)**

c) Describe the architecture, characteristics, and issues associated with each of the following categories of OLAP tools **(8 Marks)** (Slides 32-49 in OLAP)

MOLAP (Multi-Dimensional OLAP) uses specialised data structures and multi-dimensional database management systems (MDDBMSs) to organise, navigate and analyse data. **(1)** One issue with MOLAP is its products require a different set of skills and tools to build and maintain the database which increases the cost and complexity of support. **(1)** ROLAP (Relational OLAP) uses a metadata layer to support RDBMS. This avoids the need to create a static multi-dimensional data structure. It also facilitates the creation of multi-dimensional views of the two-dimensional relation. **(1)** One issue with ROLAP is complex queries require multiple passes through the relational data which creates performance problems during processing. **(1)** HOLAP (Hybrid OLAP) provides limited analysis capability, either directly against an RDBMS product, or by using an intermediate MOLAP server. **(1)** One issue is the HOLAP is the architecture results in a significant quantity of redundant data and may cause problems for networks that support many users. **(1)** DOLAP (Desktop OLAP) stores the OLAP data in client-based files and supports multi-dimensional processing using a client multi-dimensional engine. **(1)** One issue is it requires relatively small extracts of data to be held on the client machines. They may be distributed in advance, or created on demand. **(1)**

d) Discuss the relationship between Data Warehousing and OLAP. **(5 Marks)** (Slides 4-5 in OLAP and Slides 41-44 in Data Mining)

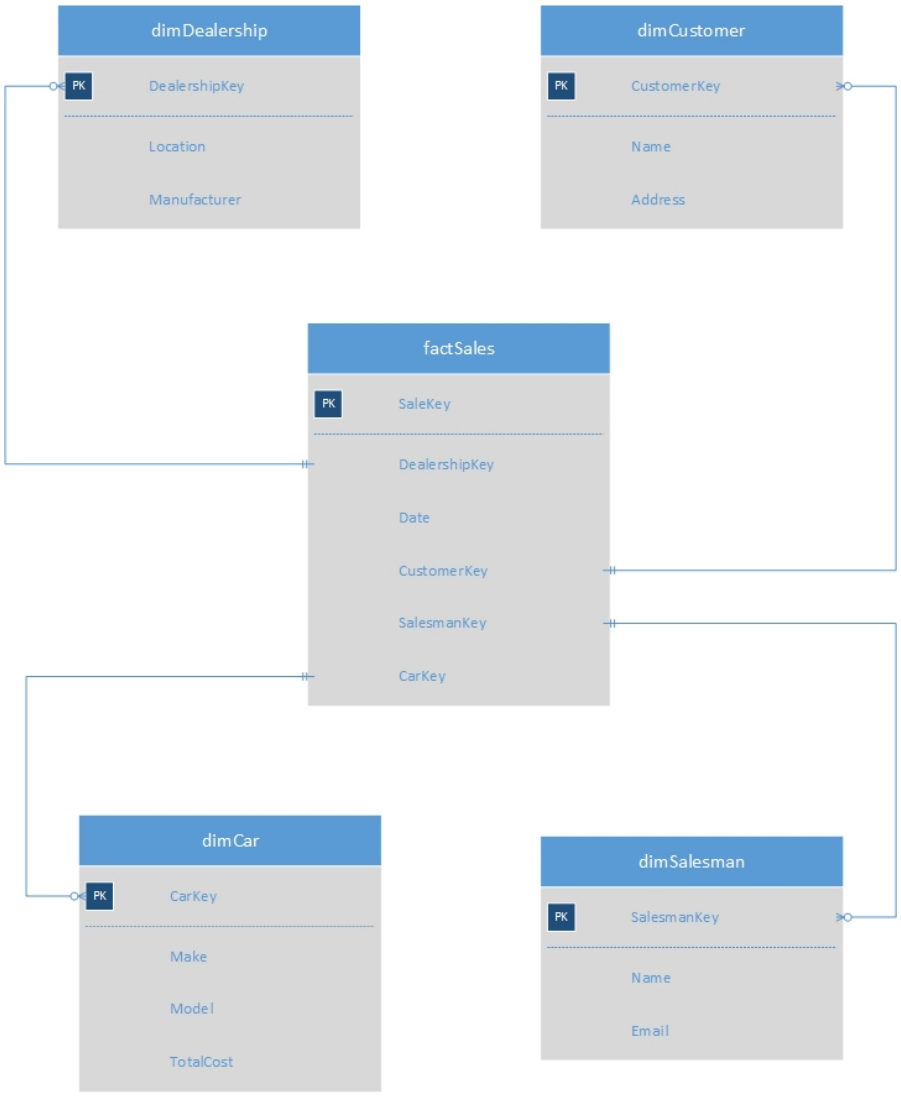
OLAP and Data Mining differ in what they offer the user. Because of this they are complementary technologies. **(1)** An environment that includes a data warehouse (or more commonly, one or more data marts) together with tools such as OLAP or data mining, are collectively referred to as Business Intelligence (BI) technologies. **(2)** OLAP and Datamining are synonymous, but they can exist independently. For example, a company may just want to improve its operational efficiency, so just OLAP could be used. **(1)** Furthermore, It is advantageous to mine data from multiple sources to discover as many interrelationships as possible. Data warehouses contain data from a number of sources. **(1)**

Question 4:

a) Design a data warehouse structure for a national car dealer chain to provide business decision makers with the important data they need. The company records information about its dealerships, customers and cars sold in its database. For each dealership it records location and manufacturer. For each car it records make, model and total cost. For customer it records name and address. For each sale the dealership, date, salesman and car are recorded. Using the four-step dimensional modelling process design a star schema for the data warehouse. **(9 Marks)** (Slides 31-40 in Data Warehousing)

Four-Step Dimensional Modelling Process:

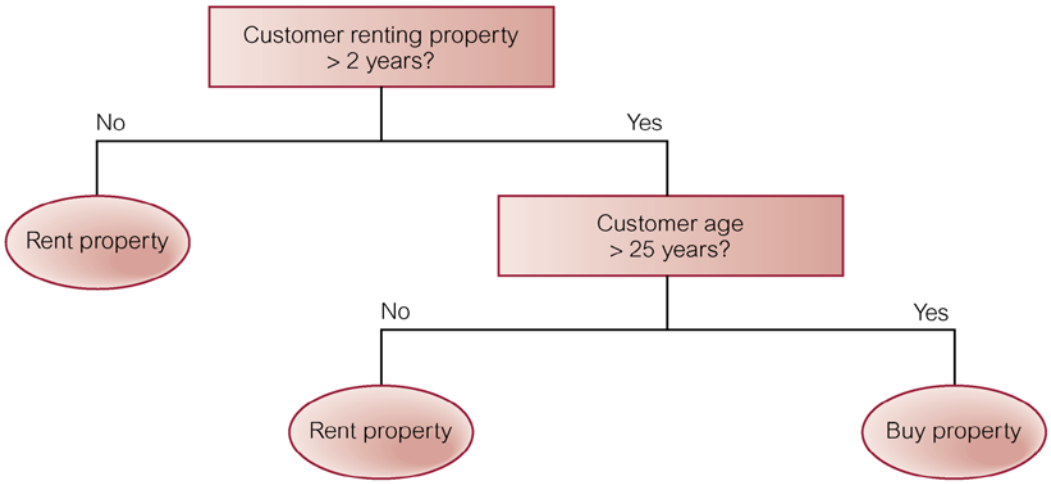
- **Step 1:** Select the Business Process
- **Step 2:** Declare the Grain
- **Step 3:** Identify the Dimensions
- **Step 4:** Identify the Facts



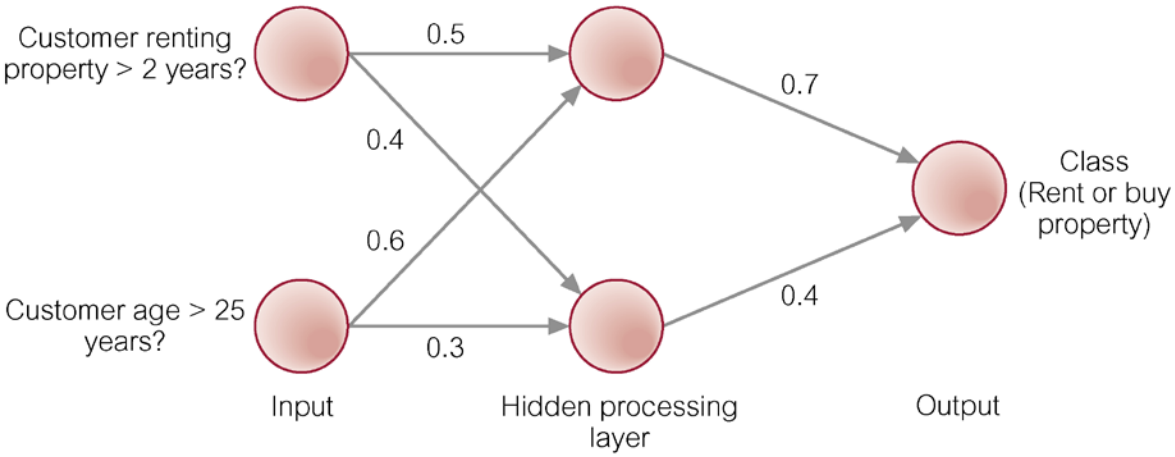
b) In data mining terms, explain how Classification and Value Prediction are used as Predictive Modelling methods. **(8 Marks)** (Slides 15-24 in Data Mining)

Classification:

Classification is used in Predictive Modelling in order to establish a specific predetermined class for each record in a database, from a finite set of possible class values. **(1)** There are two specialisations of classification. Tree Induction and Neural Induction. **(1)**



(2)



Value Prediction:

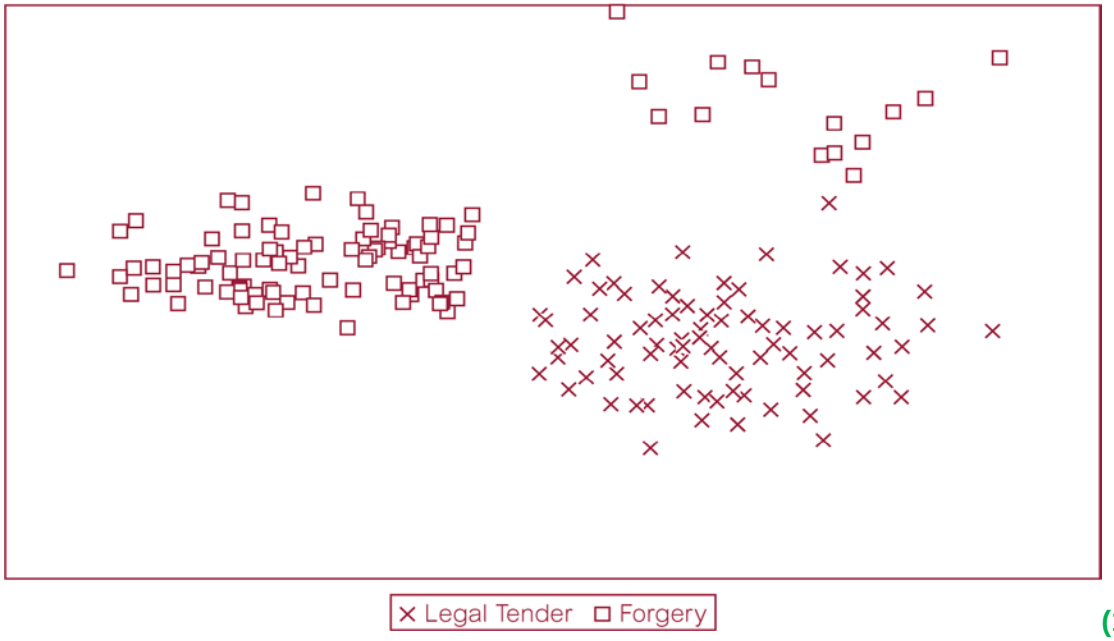
Value Prediction is used to estimate a continuous numeric value that is associated with a database record. **(1)** It uses the traditional statistical techniques of linear regression and non-linear regression. **(1)** Linear Regression attempts to fit a straight line through a plot of the data which represents the best average of all observations at that point in the plot. **(1)** One issue with this is the technique only works well with linear data and is sensitive to outliers. **(1)**

c) Identify two methods used for Descriptive Modelling and explain how that can be used to derive useful information from a data set. **(8 Marks)** (Slide 14, 25-28, 33-34 in Data Mining)

Database Segmentation:

Database Segmentation aims is to partition a database into an unknown number of segments, or clusters, of similar records. **(1)** It uses unsupervised learning to discover

homogeneous sub-populations in a database to improve the accuracy of the profiles. (1) Applications of database segmentation include customer profiling, direct marketing, and cross selling. (1) An example of database segmentation using a scatterplot:



(1)

Deviation Detection:

Deviation Detection is often a source of true discovery because it identifies outliers, which express deviation from some previously known expectation and norm. (1) It can be performed using statistics and visualization techniques or as a by-product of data mining. (1) Applications include fraud detection in the use of credit cards and insurance claims, quality control, and defects tracing. (1)