BCS THE CHARTERED INSTITUTE FOR IT

BCS HIGHER EDUCATION QUALIFICATIONS
BCS Level 6 Professional Graduate Diploma in IT

ADVANCED DATABASE MANAGEMENT SYSTEMS

Monday 22nd September 2014 – Afternoon
Answer **any** THREE questions out of FIVE. All questions carry equal marks.
Time: THREE hours

Answer any Section A questions you attempt in Answer Book A
Answer any Section B questions you attempt in Answer Book B

The marks given in brackets are **indicative** of the weight given to each part of the question.

Calculators are **NOT** allowed in this examination.

Section A

Answer Section A questions in Answer Book A

A1

Examiners' general comments : This question was attempted by 31% of candidates and of those, 62% achieved pass level marks.

The increase in global communication through the internet has resulted in a rise in the deployment of databases connected by XML web services.
Describe how each of the following objectives are realised by the use of XML based web services technology

- The integration of database services over the WWW,
- Data can travel freely over the WWW,
- The means by which a DBMS consumes and generates XML data.
- Strong Data independence
- Loose coupling of physically distributed databases

Include in your answer sample code illustrating the role of the following key technologies:
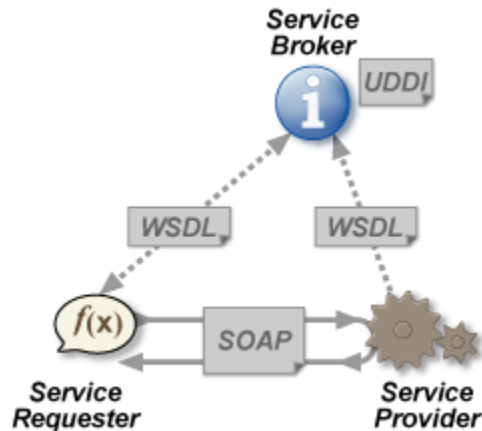The Simple Object Access Protocol (SOAP),
Web Services Description Language (WSDL),
Universal Description Discovery and Integration (UDDI)

(25 marks)

Answer Pointer
A general diagram of the above 3 components and how they interact is shown below:-

>> Data travels freely over the WWW because SOAP is an XML-based messaging protocol. It defines a set of rules for structuring messages that can be used for simple one-way messaging but is particularly useful for performing RPC-style (Remote Procedure Call) request-response dialogues. It is not tied to any particular transport protocol, though HTTP is popular, nor is it tied to any particular operating system or programming language so theoretically the clients and servers in these dialogues can be running on any platform and written in any language as long as they can formulate and understand SOAP messages.

>> The integration of database services over the WWW,

Here is an example. Imagine you have a very simple corporate database that holds a table specifying employee reference number, name and telephone number. You want to offer a service that enables other systems in your company to do a lookup on this data. The service should return a name and telephone number (a two element array of strings) for a given employee reference number (an integer). Here is a Java-style prototype for the service:

String[] getEmployeeDetails (int employeeNumber);

>> The means by which a DBMS consumes and generates XML data.
Central to the web services approach are XML SOAP based on XML and used for communication over the internet. WSDL again based on XML and is used to describe the web service. It acts as a layer of abstraction between the interface and the implementation. It is WSDL that provides loosely coupled service for future flexibility. UDDI as already mentioned above.

An example ius shown below of how data is exchanged by messaging.
SOAP is a protocol that provides this facility and in the example below enables the looking up of a stock price for a given stock name expressed in XML as a SOAP request

```
<?xml version = "1.0??">
<soap:envelope xmlns:soap = http://www.w3.org./2014/soap:envelope
Soap encoding style =  " http://ww.w3.org"
<soap:body xmlns= "http://example.name.space">
<m:getstockprice><m:stockname>Spanner
</m:stockname></m:getstockprice>
</soap:body></soap envelope>
```

>> Strong Data independence
>> Loose coupling of physically distributed databases
In recent years web services have become established as a way of connecting database applications running on different platforms and different DBMS at remote locations. It is a key technology that makes for better integration necessary because of the open-nature of the www. Therefore web services are based on open standards and focus on collaboration between users who see the world as a set of services with the database still playing the role of a backend server and hence highly abstracted and transparent. Unlike other web technologies, web services have no user interface and are thus not DBMS dependent.
Data is seen as consumable by convention, applications that are provided mechanisms in which data is transferred seamlessly using a text based format called XML. This format must adhere to standards and protocols such as w3c. It is important to emphasise the last point – all data sources (not necessarily relational dbms for that matter are disconnected until a service is called. This is UDDI a way of locating registered services and matching services to the data they consume.

Examiners comment: Generally answered reasonably well. However, this was not that apopular question being attempted by only a third of candidates. For many years this topic area has been assessed, yet this topic area still causes problems due to lack of knowledge. This topic area covers emerging technology associated with advanced database technology encompassing web/distributed database technology and non-relational eg semi-structured data based on XML. Most candidates produced simplistic answers covering only XML and XML query languages. The example code was rarely supplied so that many answers lacked a degree of quality and understanding. Some good points were made about CORBA and its differences with web services. Also candidates should be aware that technologies like JSON should be considered as alternatives to XML based services. Candidates who excelled in this question produced clear examples and good illustrative diagrams of web service architecture showing the role of databases in this loosely coupled distributed system architecture. The 5 bullet points were intended to focus the candidates' answers and to help structure their answers accordingly.

A2

a) The sample ERD (Figure A2 below) describes a scenario in which there are many shipping companies that own a number of ships that travel from one port to another. A ship follows a prescribed route consisting of stages which consists of a number of ports that are visited on the route starting from and finishing at a ships home port. A ship journeys over one route but a route may be used by many ships. A route is also covered by many shipping companies but a company only travels over one route.

Figure A2 ERD of a shipping company
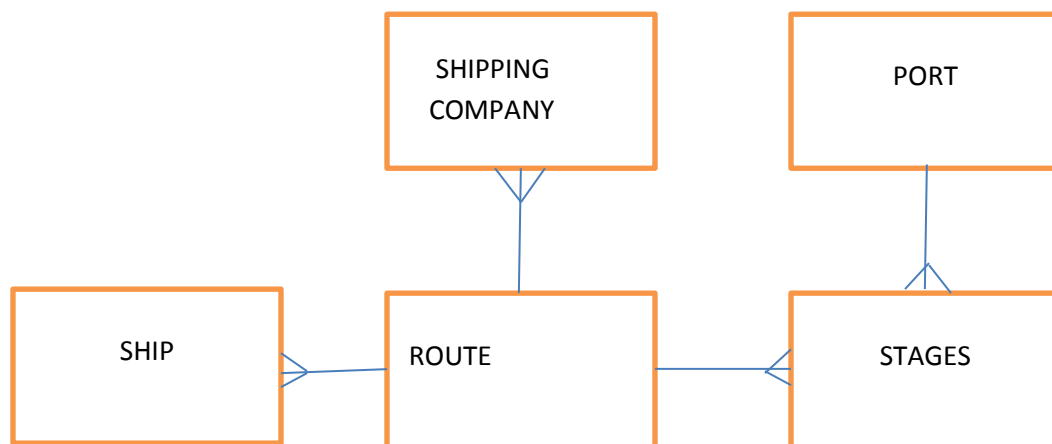
With reference to the ER model Figure A2

(i)   Explain the problem known as *fan trap* . Identify a potential *fan trap* in the above ER model and show how this is resolved.

Answer Pointer

There is a fan trap which occurs when a model represents a relationship between entity types, but the pathway between certain entity occurrences is ambiguous. It occurs when 1:m relationships fan out from a single entity.   Thus a ship follows one route but a route is covered by many ships so the owning company of the ship cannot be determined. This can be resolved by reorganising the relationships thus:

ship→ Company→ route

(6 marks)

(ii)   With aid of an example explain the problems caused by a *chasm trap* and explain how this might be resolved.

Answer Pointer

A chasm trap occurs when a model suggests the existence of a relationship between entity types, but the relationship pathway does not exist for all entity occurrences. It occurs where there is a relationship with partial participation, which forms part of the pathway between entities that are related.

A suitable example:-

Each company has a number of ships. Each of their ships has a number of allocated staff. If this is the case, then a relationship (1:M) could be drawn between company and ship and another 1:M relationship from ship to staff.

But suppose some of the staff of the company dare not allocated to ships. (In other words, the relationship of staff to ship is optional (partial participation)). For these there would be no link available to their company because the link is established only through a ship. Hence the company that employs these staff cannot be identified.

The missing relationship which is called `employs' between the Staff and the Company entities needs to be added.

Care is also needed when removing relationships which are considered to be redundant.

(7 marks)

Examiner's comment: This part of the question was answered by less than half of candidates and those answers showed a lack of knowledge of extended ER modelling features encompassing connection traps. Connection traps are an important check in the modelling process to confirm if the relationships in the model conform when mapping to Tables.

b) With reference to the scenario below show how Object Oriented modelling concepts can be used to model Aggregation and Inheritance and explain how these constructs are translated to database tables/relations.

(12 marks)

A ship carries a cargo and a crew. Crew are personnel with different responsibilities – maintenance, navigation, catering and loading. Loading crew are contractors that are not employed by the shipping company. Maintenance crew are assigned properties that reflect their speciality. A transport job involves transporting cargo consisting of different products that are stored in containers. Each container may hold one or many products. There may be none, one or many containers transported on a ship between visits to ports. Containers can be of a standard size or a large size but some containers are of a special type of load, these being either Refrigerated or Hazardous. Containers are either loaded or unloaded during the visit of a ship at a port.

Answer Pointer
Ship carries Crew and these are generalised by a Class of the same name. Crew are specialised into Maintenance, Navigation and Catering crew. Loading crew have a specific relationship as they are not employed by the ship company and have a responsibility of loading operations. Within each of these sub classes are assigned properties such as qualifications, or some specialised aspect of their role.
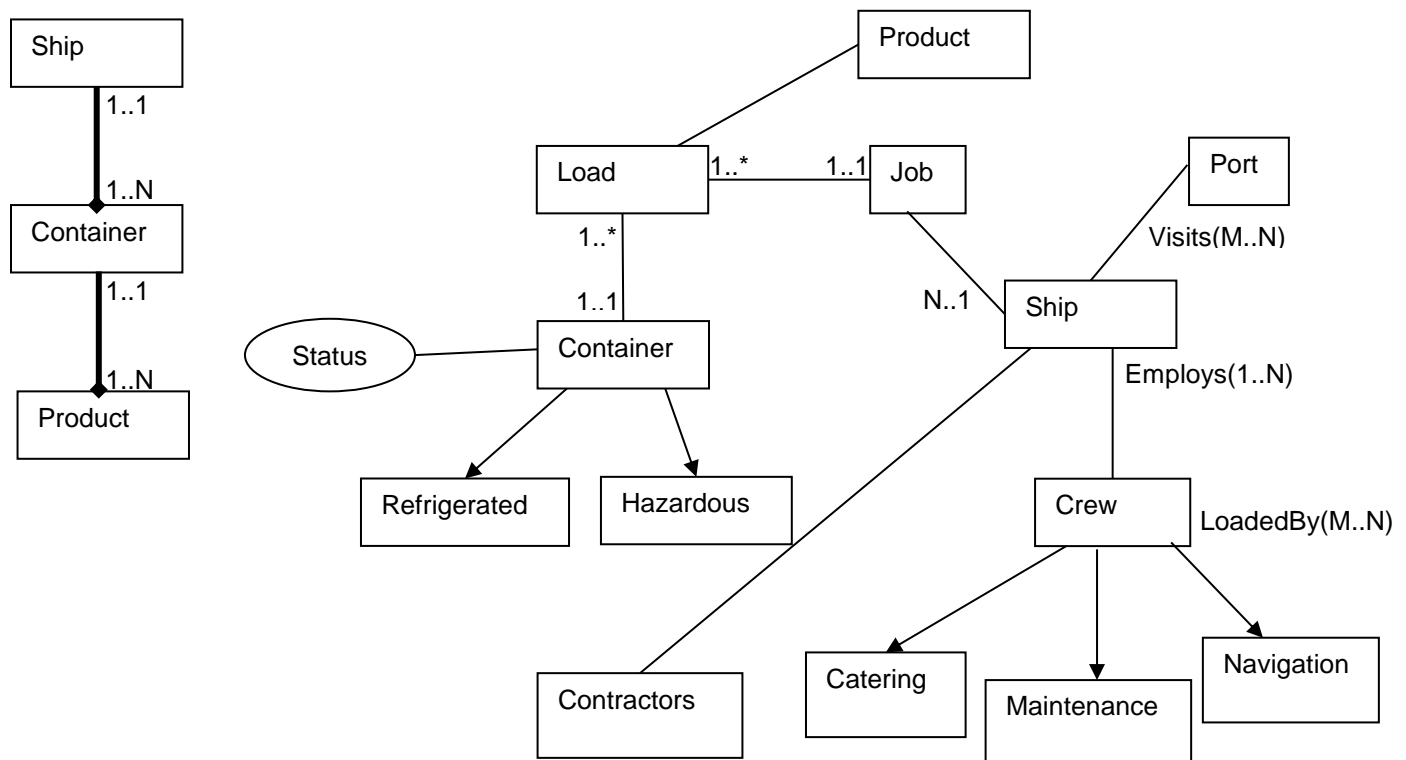
Mapping to tables

Inheritance (equivalent to specialisation and generalisation) represents weak entities (dependent entities become specialised tables containing a posted ID from the strong entity type). So for example ContainerID (the strong entity ID) would be posted in Refrigerated and Hazardous weak entity types.

A form of aggregation are intra relationships between the same entity i.e. a product is itself formed from a product (a minor part of a major product)

MajorProductID	SubProductID
5	3
5	2
5	11
11	22
11	23

A3

a) Consider the following table:

employees (<u>empID</u>, name, salary)

The table is stored on a disk file consisting of 40 blocks and the primary key index is a B-Tree with 3 levels and 20 leaf nodes.

For each of the following queries, state how the query is to be executed (e.g., full table scan, full index scan, etc.) and calculate the associated cost (in number of blocks):

(i)   SELECT empID FROM employees;

(3 marks)

(ii)   SELECT name FROM employees WHERE empID = 120;

(3 marks)

(iii)   SELECT * FROM employees WHERE salary > 15000;

(3 marks)

Answer Pointers

(i) No need to access the table on disk as all the data requested in the query is contained in the index. Therefore a full scan of the index will cost: number of leaf nodes = 20 blocks. (3 marks)

(ii) An index unique scan will be performed to locate empID value, and then ROWID will be used to read a data block from the table. Cost = Number of levels + 1 = 3 + 1 = 4 blocks. (3 marks).

(iii) Given that there is no index on salary, a full table scan is performed. Cost = 40 blocks. (3 marks).

Examiner's comment: Many candidates seem to understand the basic working of a B-Tree index and how it links to the data file, however, there seems to be some misunderstanding of when (and how) a query can benefit from such an index.

b) The SGA (System Global Area) is a shared memory area in an Oracle database instance. Describe its role in achieving good query performance.

(4 marks)

Answer Pointers

The SGA is a shared memory area between the multiple users of the database. Therefore, all users' queries (and the data resulting from those queries) is temporarily saved into the SGA to avoid repeated, time-consuming access to disk. (4 marks)

Examiner's comment: Very few candidates answered this question correctly. It is important that candidates familiarise themselves with modern DBMS architectures and the specific functions of their components.

c) Security is a major concern in database systems.

    (i)   Describe four of the main threats to database security.

(4 marks)

    (ii)   Using examples, discuss how Triggers can contribute to the security of a database, before and after a security breach.

(4 marks)

    (iii)   Describe two mechanisms for guarding against SQL injection.

(4 marks)

Answer Pointers

(i) Three of any of the following (or any other sensible answer) + brief description: (1 mark each)
- **Theft and fraud**
- **Loss of confidentiality**
- **Loss of privacy**
- **Loss of integrity**
- **Loss of availability**

(ii) Triggers can be used to (for example): (2 marks each)
- **Before: Preventing access/change to a table outside normal working hours,**
- **After: Finding out who connected to the database and what they did (event logging)**

(iii) Any sensible two measures (with brief description), for example: (2 marks each):
- **Input validation**
- **The use of bind variables in SQL statements**
- **Displaying appropriate error messages**
- **Granting the least privileges to database users**
- **Changing database default passwords**

Examiner's Comment: Almost all candidates gave sensible answers to part (i). However, not all were able to provide a relevant example for the use of triggers in securing databases. Also, candidates need to get more familiar with ways of defending against SQL injection which remains one of the top threats to database security.

Section B
Answer Section B questions in Answer Book B

B4

Examiners' general comment: A very popular question attempted by almost all candidates, the vast majority achieving pass level marks.

a) How does a *data warehouse* differ in function, content and structure from an *OLTP database*?

(5 Marks)

Answer pointers

- **An OLTP database is designed to support low-level, day-to-day transactional data with a view to satisfying the core business activities by junior to mid-level staff. By contrast, a data warehouse is all about long-term strategic decision making by senior management.**
- **The data warehouse is commonly defined by four attributes: *subject-oriented* (customer, product etc.), *integrated* (uniform data drawn from multiple sources), *time-variant* (historical data means a series of snapshots of data over a time period) and *non-volatile* (data is always added, often aggregated and usually not deleted or modified).**

- **An OLTP database will, for efficiency reasons, not hold historical data and is application/business-oriented (customer invoicing etc.)**
- **It is the difference between a low-level routine transaction-focus and a high-level ad-hoc, analysis focus.**

Examiner's Comment: Universally well answered – it was perfectly clear that just about every student had a good grasp of these fundamental issues. Consequently, most students scored the full 5 marks.

b) For each of the following items, explain what the term means, the underlying concepts involved, any associated benefits or limitations, typical applications and features. You should support your discussion with suitable diagrams and/or examples.

   (i)   OLAP, its varieties, the role of aggregation and the impact on SQL

(5 Marks)

   (ii)   Data Mining, associated algorithms/techniques and data cleaning

(5 Marks)

   (iii)   Multi-Dimensional Data, dimensional analysis, roll-up and pivoting

(5 Marks)

Answer pointers

Marked holistically OLAP (On-Line Analytical Processing) is the dynamic synthesis, analysis and consolidation of large volumes of aggregated multi-dimensional data (or similar). Key features include aggregation, time-series analysis, drill-down queries, highly complex 'what if' and 'why' type queries (as well as the more traditional OLTP-type 'what', 'when' and 'who' queries) across multiple views (dimensions) of the data over extended time periods. Applications could include budgeting and financial modelling, sales analysis & forecasting, market research etc. The better students may go onto discuss Codd's twelve rules for OLAP tools, the different categories of OLAP – MOLAP, ROLAP and HOLAP (multi-dimensional, relational and hybrid respectively) including a brief discussion of the key features of each. Finally, a few comments on OLAP extensions to SQL could be included such as the ROLLUP and CUBE functions (within the GROUP BY clause) and RANK, DENSE_RANK functions and 'windowing' calculations on aggregates.

Data Mining – The process of identifying patterns & trends and thus extracting previously unknown, actionable, valid and valuable information from large data sets for commercial or other advantage (or similar). Typical applications could be identifying customer buying patterns, detecting patterns of fraudulent credit card use, analyzing insurance claims/frauds etc. Techniques include *predictive modelling* (using supervised learning with training/testing stages), *link analysis* (based on establishing associations or affinities between different data items) and *deviation detection* (using statistical or data visualization techniques to identify 'outliers' in the data that defy 'normal' or

expected behaviour – thus leading to new, unexpected knowledge). The better students could also cover the key issues around DM tools such as data preparation (data cleaning, handling missing data etc.), selection of data mining algorithms (how it relates to the training/testing stages and how it handles 'noise' in the data), and scalability & performance (use of parallel processing and other optimization issues).

Multi-Dimensional Data begins with the traditional two-dimensional relational table before moving onto three-dimensional data cubes and, logically, four dimensions and more (for example dimensions could be time, location, product, staff) with each cell in the 'cube' representing sales figures. To maintain acceptable query performance as the number of dimensions rises, 'rolled-up' or pre-computed (consolidated or aggregated) data may be stored that in turn may be drilled down for finer detail as needed (for example, time could be aggregated to years but drilled down to quarters, months, weeks and even days' worth of sales). Data may also be 'sliced & diced' (pivoting). The role of dense/sparse data in different cells and the role of data compression could be covered.

Examiner's comment: Again, very well answered by the bulk of candidates. Most supplied well-presented, clearly annotated diagrams (such as data cubes and star/snowflake schemas) and went into extensive detail on each sub-topic. This was a strong question for most students.

c) Explain the term 'ETL' with respect to data warehousing. Highlight common techniques, problems or issues in each stage of the use of ETL.

(5 Marks)

Answer pointers

Marked holistically but covering…

- *E = Extract -* **Copy data from operational and other external sources like OLTP databases, spreadsheets, websites etc**

- *T = Transform -* **Select/reject, clean, re-format, order, (dis)aggregate data into a consistent structure and regular format. Issues could include what to do with missing (NULL) entries, standardizing on one format (for example 'M/F', 'Male/Female' for Sex), how many fields to use for the Address data? or application of business rules.**

- *L = Load -* **Copy the selected and cleaned data into the data warehouse at regular intervals**

Specific examples gained bonus marks.

Examiner's comment: Yet again, strong responses across the board. Nearly all attempts achieved maximum marks.

B5

(a) In your own words, define the following database terms…

    (i)   Transaction

(2 Marks)

    (ii)   ACID properties

(2 Marks)

    (iii)   Isolation levels

(2 Marks)

    (iv)   Locking levels

(2 Marks)

You should support your discussion with suitable examples and/or diagrams.

Answer pointers

A transaction is a logical unit of work (comprising one or more operations) that transforms a database from one consistent state to another consistent state (or very similar). Students should make it clear that a transaction is a concept commencing with a DML operation and concluding with either a rollback or commit and is not the same as an individual SQL statement or query. Students should then move onto ACID as specifying atomicity, consistency, isolation and durability - each aspect must be named and explained in detail. Bonus marks for a clear worked example or well annotated diagram. A basic definition of isolation and locking levels will suffice but correct terminology should be used – such as X-locks and S-locks etc.

(a) Using your own examples and/or suitable diagrams, explain the following *transaction processing* concepts and their role in enforcing data integrity and consistency…

    (i)   COMMIT

(2 Marks)

    (ii)   ROLLBACK

(2 Marks)

    (iii)   SAVEPOINT

(2 Marks)

Answer pointers

- **COMMIT as a saving operation (of the complete transaction).**
- **ROLLBACK to undo the whole transaction. Or…**
- **SAVEPOINTS as transaction partitioning concepts whereby a large transaction can be sub-divided down into smaller units using embedded labels and how the DBMS can ROLLBACK to a named SAVEPOINT rather than undoing the whole transaction.**

Examiner's comment: Most candidates submitted extensive, quality responses that easily gathered the full 6 marks.

(c) Using your own examples and/or suitable diagrams, explain how the TWO-PHASE COMMIT protocol may be needed in a distributed database environment and why the above concepts in part (b) may be inadequate to ensure data integrity and consistency in such a distributed environment. You are specifically advised to address the following concepts:

- The two stages of the TWO-PHASE COMMIT protocol
- Global and local transactions
- Inter-site communication during the TWO-PHASE COMMIT protocol
- The impact of partial success and ACID rules

(11 Marks)

Answer pointers

Marked holistically but students should describe the need for the two-phase commit in a distributed database environment and how it constitutes the 'voting' phase (the 'are you ready to commit' message) and the 'decision' phase (the 'commit now' message). Students should also cover the concepts of *global transactions*, *localized sub-transactions*, *transaction coordinator site* (the transaction initiator site), *participant sites*, the need to pass *messages* between sites, the use of *timeouts* to avoid unnecessary blocks or delays, the possibility of a participant site issuing a *local abort* – thus forcing a *global abort* to be issued and the need for unanimous *local commits* before a *global commit* is circulated – all ensuring the *data integrity* and *ACID rules* are satisfied.

Examiner's comment: Less well done that parts (a) or (b) but most candidates did a reasonable job – clearly addressing the distinction between local and global transactions, the need for a two-part coordinated COMMIT process to ensure ACID-compliance etc. The only disappointing aspect was that relatively few candidates supplied a diagram to supplement their narrative. However, only a few poor attempts were submitted.