# THE BCS PROFESSIONAL EXAMINATIONS
## Professional Graduate Diploma

**April 2007**

**EXAMINERS' REPORT**

## Advanced Database Management Systems

**General Comment**
The performance of this paper is similar to previous years. Candidates are advised not only to know the theoretic knowledge in books but the need to apply the knowledge in practice in real world situations.

**Question 1**

a) The definition of a relation in the entry for "Relational Database" in Wikipedia is "A *relation* is defined as a set of tuples that all have the same attributes. This is usually represented by a *table*, which is data organized in rows and columns."

Explain this definition. Your explanation should include a discussion on the difference between relations and tables. **(10 marks)**

b) SQL is not completely faithful to the relational model of data. Describe **three** ways in which SQL diverges from the mathematical theory of relations. Discuss the importance of each of these variations. **(15 marks)**

**Answer Pointers**

This question examines Section 1A the Relational Model of Data.

a) A tuple is a finite sequence. In the case of a relation, the tuple is a sequence of attribute value pairs. Each tuple will contain the same number of attribute:value pairs and the attributes will be the same for each tuple. There is no need for any special ordering of the attribute value pairs as the attribute name provides the context for the value. The values will be drawn from a domain which is a set that contains all the valid values for a given attribute. A relation is a set of such tuples. Consistent with the definition of set the order that these tuples appear is not significant but duplicate tuples are not permitted. Tables are constructed from rows and columns. These differ from relations in that ordering is used to associate a value with an attribute. In addition, rows in a table are ordered.

b) There are a number of possible answer to this question, three possible answers are:

i) SQL columns are ordered. This is not significant as long as the database management system does not rely on this sequencing to function.

ii) SQL allows duplicate rows. This can be significant as it means that objects stored in the database are distinguished by some principle external to the data that represents them.

iii) SQL supports aspects which are only related to database performance and not to data representation e.g. indexes. There are no problems caused by this unless these additions alter the results of queries.

**Examiners Guidance Notes**

This question was popular with candidates but unfortunately it did not attract many good answers.

Part a) invited candidates to present a formal definition of a relation and required them to comment on the fact that this definition is based on the concept of set. Many candidates included irrelevant facts in their answers, most notably a description of foreign key constraints. Some candidates recited Codd's twelve rules for relational database management systems. Whist some of these have relevance, the majority of the rules have little to do with the definition of a relation. Most candidates were unable to distinguish between a relation and a table.

In part b) many candidates were able to describe aspects of SQL that diverge from the mathematical relational model. Very few, however, attempted to discuss whether these aspects were significant and consequently lost marks.

## Question 2

a) Describe five relational algebra operations. For each relational algebra operation you list, illustrate the result of applying it to a relation. **(10 marks)**

b) Although it is theoretically possible to query a relational database using relational algebra, in practice the vast majority of users use a language based on relational calculus. Describe the main features of relational calculus and its relationship to relational algebra. **(15 marks)**

**Answer Pointers**

This question examines Section 1A the Relational Model of Data.

a) The first three Examples use the following tables:

A

| p | q | r |
|---|---|---|
| a1 | b1 | c1 |
| a2 | b1 | c2 |
| a3 | b2 | c3 |

B

| r | s | t |
|---|---|---|
| c1 | e1 | f1 |
| c2 | e2 | f2 |

Restrict identifies all the tuples which satisfy a given condition:
RESTRICT A where q = b1

| p | q | r |
|---|---|---|
| a1 | b1 | c1 |
| a2 | b1 | c2 |

Project selects the attributes
PROJECT A over q

| q |
|---|
| b1 |
| b2 |

A JOIN B is the set of all combination of tuples in A and B where the value of attribute r (the common attribute) matches.

| p | q | r | s | t |
|---|---|---|---|---|
| a1 | b1 | c1 | e1 | f1 |
| a2 | b1 | c2 | e2 | f2 |

The final two examples use the following tables:

A

| p | q | r |
|---|---|---|
| a1 | b1 | c1 |
| a2 | b1 | c2 |
| a3 | b2 | c3 |

B

| p | q | r |
|---|---|---|
| a1 | b1 | c1 |
| c2 | e2 | f2 |

A INTERSECTION B is the rows in both A and B

| p | q | r |
|---|---|---|
| a1 | b1 | c1 |

A DIFFERENCE B is the rows in A but not in B

| p | q | r |
|---|---|---|
| a3 | b2 | c3 |
| a2 | b1 | c2 |

b) Relational calculus is based on the mathematical predicate calculus. The execution of a relational calculus expression is a search for values of variables which make a given expression true. In tuple calculus the variables are tuple variables whereas in domain calculus the variables are domain variables. SQL is a tuple calculus language whereas Zloof's query by example is based on domain calculus.

Relational algebra and calculus are equivalent in their expressive power.

Relational algebra provides a collection of explicit operations - join, union, projection, etc.

The operations are used to tell the system how to build some desired relation in terms of other relations.

The calculus provides a notation for formulating the definition of a desired relation in terms of other relations.

Relational Algebra is procedural; it is more like a programming language;

Relational Calculus is nonprocedural it is more close to a natural language.

Codd provided a reduction algorithm that can transform any arbitrarily complex calculus expression into a set of relational algebra statements. Using this a database implementer can create an RDBMS by creating software that can carry out relational algebra operations (there are only a small number of these). The reduction algorithm can be used to translate any SQL statement to a sequence of these operations.

**Examiner’s Guidance Notes**

This was the most popular question on the paper. It received a number of good answers, however, candidates tended to perform better in part a) rather than part b).

Part a) was designed to remind students about the relational algebra they would have encountered in the diploma paper. Since this is basic knowledge it was marked fairly strictly. Worryingly, it appears that a significant number of candidates have yet to master this elementary material.

Part b) explored the relevance of relational algebra to actual database implementations. A good answer would discuss the essential characteristics of algebra and calculus and point out their equivalence. From this candidates could argue that whilst users tended to use calculus at an internal level the database executed algebra. The were a number of well informed answers to this part but also some answers which described aspects of relational databases such as query optimisation which were not relevant to the question.

**Question 3**

In a paper entitled “The Object-oriented Database System Manifesto”, Atkinson et al state: “An object-oriented database system must satisfy two criteria: it should be a DBMS, and it should be an object-oriented system, i.e., to the extent possible, it should be consistent with the current crop of object-oriented programming languages.” Discuss the implications of this statement.

**(25 marks)**

**Answer Pointers**

This question examines Section 1c Emerging Database Management System Technologies

Candidates were not expected to have read the paper cited, however, if they had been exposed to the aims and objectives of object oriented systems they should be able to identify those aspects of databases which should be implemented by an OODBMS as well as the aspects of OO programming the system should support. From the database point of view an OODBMS will support persistence, secondary storage management, concurrency, recovery and an ad hoc query facility. Additionally such a database may support distribution, design transactions and versions. From the programming language point of view an OODBMS will support complex objects, object identity, encapsulation, types or classes, inheritance, overriding combined with late binding, extensibility, computational completeness. It may also support multiple inheritance, type checking and inferencing though these are not essential. Full marks will be awarded for a discussion of both database and programming language capabilities expected with an explanation of the issues that can be associated with the provision of the characteristic.

**Examiner’s Guidance Notes**

This question was not answered particularly well by the majority of the candidates who attempted it. Most answers were able to identify important features of object-oriented programming languages but failed to say how they would be realised in a database management system. Very few answers pointed out the aspects of databases that should be supported by an OODBMS. A number of answers simply discussed SQL support for objects, which was too narrow to attract all the marks available. Other answers chose to discuss Object-Relational systems, which were not the subject of the question.

**Question 4**

a) In the context of query optimisation explain the following terms and describe the effect each will have on the query optimiser's decision as to whether or not to use an index. Give suitable examples from the sample tables to illustrate your answers.

i) Selectivity **(4 marks)**

ii) Density **(4 marks)**

b) Explain, for a stated DBMS, the indexing strategy that you would recommend for the Customer table in **Figure 1** below. The table has a primary key called *custId* and indexes are required to support the following queries:

i) SELECT custLName, custFName
FROM Customer
ORDER BY custLName, custFName **(6 marks)**

ii) SELECT custLName, custFName
FROM Customer
ORDER BY custId; **(6 marks)**

c) Explain how the DBMS query processor would retrieve the rows returned by each of the queries in *b)*above. As part of each explanation draw a suitable diagram.
**(5 marks)**

| **custId** | **custEmail** | **custPassword** | **custTitle** | **custFName** | **custLName** |
|---|---|---|---|---|---|
| 1234 | fredjones@hotmail.com | freddie806 | Mr | Fred | Jones |
| 1235 | annekrakowski@demon.net | kraKKers | Ms | Anne | Krakowski |
| 1236 | jones9004@aberdovey.com | Blaenau | Mrs | Myfanwy | Jones |
| 1237 | gerhartshcroeder@hotmail.com | Chancell0R | Herr | Gerhart | Shroeder |
| 1238 | thesmiths@marske.com | seaSide567 | Miss | Jane | Smith |

**Figure1: Sample Extract of a Customers Table containing Millions of Rows**

**Answer Pointers**

**Part a) i and ii)**
**Selectivity** is derived from the percentage of rows in a table that are accessed or returned by a query. In high selectivity, the search criteria limit the number of rows returned to a low percentage of the total possible. One row returned is the highest selectivity that can be achieved.

Using an index will optimise a query with high selectivity. In low selectivity, the search criteria return a high percentage of rows in a table. Using an index will not optimise a query with low selectivity.

***2 + 2 marks for any suitable example.*****Density**

**Density** is related to density, but is the average percentage of duplicate rows in an index. An index with a large number of duplicates, eg an index on last name, has high density. A composite index, eg on last name and first name, will be much less dense. A unique index, for example the nationalidentityNumber, has low density. However, density relates to specific data elements and can vary accordingly. For example data elements of an index on last name are very dense for popular last names such as Smith, whereas a rare last name such as Otachataka is not likely to be very dense.

Because data is not distributed evenly, the query optimiser might use an index or not. It might perform a table scan to retrieve the last name Smith and use an index to access the last name Otachataka.
The higher the number of rows returned by a query, the less likely it is that the query optimiser
will use an index, unless the data can be retrieved JUST from the index and therefore the data pages do not need to be read at all.

***2 + 2 marks for any suitable example.***

Part b.

i.
```
SELECT custLName, custFName
FROM Customer
ORDER BY custLName, custFName;
```

Here all the data required is in the non-clustered index, which is also in the order required, so the database query optimiser will use an index scan, ie it will start at the first item in the non-clustered index and work through the index in order to the end, retrieving the required data.

3 marks for explanation, 3 for diagram

ii.
```
SELECT custLName, custFName
FROM Customer
ORDER BY custId;
```

Here all the data required is in the non-clustered index, but the order required is the same as the order of the clustered index, so database query optimiser will not use the non-clustered index at all. It will simply use an index scan of the clustered index, ie it will start at the first item and work through the index in order to the end, retrieving the required data.

3 marks for explanation, 3 for diagram

Part c)

Candidates are expected to draw the expression tree of relational algebra operations that the query optimiser would build to schedule the order of processing each data access operation.
The key relational algebra operations are PROJECT and SELECT.
Firstly for the above queries a SELECT operation would be the first relational algebra task in the expression tree to filter the WHERE clause followed by the PROJECT operation to specify the target output (Columns) of the result set.
To get full marks candidates would be expected to reproduce the output from a graphical showplan of a query processor of an actual server DBMS. Finally knowledge of how the ORDER BY clause is interpreted by an actual DBMS graphical showplan – the optimiser will implement a sort/merge operation, please note this is not a Relational Algebra operation and might be implemented differently across a range of database servers.
5 marks for diagram + explanation

**Examiner's Guidance Notes**

Overall very disappointing answers for the entire question, answers were mostly shallow and superficial for this level.

The key to getting high marks in this question is to understand the rules that a query optimiser uses to execute a SQL query efficiently given indexes and rules of relational algebra (to translate SQL statements).

It is a concern for the examiner that candidates do not seem to appreciate the reasons why database queries (usually) execute so efficiently. Perhaps candidates have not been encouraged to examine the output from a query optimiser (text or the graphical showplan) which informs SQL programmers of how their SQL statements are executed and perform.

**Question 5**

a) The convergence of database and Internet technologies has meant that major database vendors are required to support software components that handle distributed data in their flagship products.

Two important software components in this convergence are *Web Services* and *Data Replication*.

*i)* What is a Web Service?

*ii)* What is Data Replication? **(8 marks)**

b) Refer to the scenario on the next page. This describes the requirements of a mobile internet application.

Explain how web services and data replication could be used in combination to support the application requirements described. Use diagrams and examples to illustrate your answer wherever possible and state any assumptions you make. **(10 marks)**

c) In order for web services and data replication to work various software standards (such as SOAP and XML) have to be built into software products. Describe ONE of these standards and discuss why this particular standard is important. Use examples from the scenario to assist in your answer. **(7 marks)**

**Scenario: Requirements of a Mobile Internet Application (for use in Question 5)**

The ‘GP Out-of Hours Service’ is a service run by a local health authority whereby patients can request medical treatment outside the normal visiting times of a GP practice. GPs (also known as Doctors) are on call to visit patients usually at their home. A GP practice may have many GPs on call at the same time and the practice area may cover large rural areas.

A patient’s medical records (including diagnosis and treatment history) are located at the national central database called SPINE. Specific image-based medical data such as X-ray photographs are stored in the database located at the hospital that carried out the test (for example an X-ray), this database is called TRUST. In order to monitor the location of GPs and the status of the GP (for example they may be free, attending a call etc) another database called CENTROL records the geographical location of every doctor who is on call.

Each GP has a driver who drives the GP to the patient’s address. The GP/driver can at any time receive instructions to visit other patients who need treatment. GPs carry with them a hand-held device called a PDA that has a small database capable of storing a small amount of patient medical data. GP’s use the local PDA database to input data following a visit. (For example the GP may record the patient’s diagnosis and treatment). It is important that the GP has an up-to-date medical history of the patient and that any drugs dispensed do not conflict with existing medications. Medical data should be removed from the PDA database after the GP has finished his calls.

**Brief specification of the PDA**

The PDA is a hand-held computer that runs a mini version of the Windows operating system. It has 128K Ram and has a removable 1GB memory card. The PDA has a built-in mobile phone, supports wireless internet connections using GPRS. The PDA can also connect to a location tracking device called a GPS receiver. The PDA transmits its location data to the CENTROL database every 2 minutes.

**Answer Pointers**

Part a)

Web services – a web technology that supports loosely coupled interaction between distributed components. 1 mark Key points to get across include a description of its basic principles

**1 mark** – message oriented approach using

**1 mark** – open architecture and WSDL

**1 mark** – SOAP – based on exchange of XML formatted data

The convergence of database and Internet technologies has meant that major Database vendors are required to support software components that handle distributed data in their flagship products.

Replication – supports physical distribution of data copied on distant servers. **1 mark** The basic concepts are

- **1 mark** Maintenance and synchronisation
- **1 mark** Types **-** merge / transactional and snapshot approaches
- **1 mark** Networks mainly LAN based (e.g. named pipes) with limited support for HTTP

b)

Fairly open ended, marks were awarded for a system diagram of the components that made up the distributed system. **5 marks.**

An understanding that both web services and replication can co-exist and indeed complement each other by using web synchronisation over HTTP. This means candidates will have to explain the security implications and the role played by IIS/web servers/ data sets , MTS i.e. features that cross the web services – replication areas of responsibility. **5 marks**

c)

This might seem like bookwork but there must be some description to explain how a particular standard would be applied to the discourse/scenario. For example the examiner would expect an example of a SOAP message, in the example it would be represented as an XML encoded envelope. This would be used to hold the following standardised information (W3C) within a well defined set of rules for example

- the contents of a message and how to process it,
- A set of encoding rules for expressing instances of application-defined datatypes,
- a convention for representing remote procedure calls and responses.

**3 * 2 marks** each above + **1 mark** stating that W3C impose the standard

**Examiner's Guidance Notes**

Again many candidates had difficulty applying knowledge to a discourse but given the poor knowledge of web services this did not matter too much. Candidates should look at the 2006 paper where knowledge of web services was asked before.

Many candidates expressed in their answers a misunderstanding of the technical aspects of web services or simply relayed knowledge from an application viewpoint devoid of any relevance to the scenario provided.

Knowledge of the older technology (database replication) was generally good but it seems that many candidates are not reading up on the fast moving distributed web technologies that are driving Web 2.0 and now Web 3.0 futures. With poor knowledge of the former meant many candidates could not reflect on the association between data replication and its evolution and coexistence with web delivery systems.