FACTFILE:
GCSE
DIGITAL TECHNOLOGY
Unit 1 – DIGITAL DATA

Rewarding Learning

Fact File 8: Database Applications (2)

Learning Outcomes

Students should be able to:

- extract data from a database structure using simple query structures and using the following logical operators: <, >, =, <=, >=, AND, OR and BETWEEN;
- demonstrate understanding of big data, referring to volume, velocity and variety;
- demonstrate understanding of the need for data analytics to interpret big data.

Content:

- An Example Database;
- Database Queries;
- Big Data;
- Interpreting Big Data.

An Example Database

A music database has been designed to hold data about musicians and the albums they played on. The database has the following structure.

Album(Album_Title, Record_label, Release_Date) Artist(Artist_Name, Birth_Year, Main_Instrument)
PlaysOn(Album_Title*, Artist_Name*)
Album_Name -> Album
Artist_Name -> Artist

Note that the primary key of the PlaysOn table is a composite primary key – i.e. it consists of two fields. Each of these fields is, individually, also a foreign key linking to another table.

The database is populated below with a small sample of data found at this website:

- http://www.allmusic.com/
  - http://www.allmusic.com/album/steamin-mw0000191715
  - http://www.allmusic.com/album/lush-life-mw0000187971
  - http://www.allmusic.com/album/red-alone-mw0000220452

Album

Album_Title	Record_Label	Release_Date
Steamin with the Miles Davis Quintet	Universal	1961
Relaxin with the Miles Davis Quintet	Prestige	1958
Cookin with the Miles Davis Quintet	Prestige	1957
Lush Life	Fantasy	1958
Red Alone	Universal	1960

Artist

Artist_Name	Main_Instrument	Birth_Year
Miles Davis	Trumpet	1926
Red Garland	Piano	1923
John Coltrane	Saxophone	1926
Philly Joe Jones	Drums	1923
Paul Chambers	Bass	1935
Earl May	Bass	1927
Art Taylor	Drums	1929

PlaysOn

Album_Title	Artist_Name
Steamin with the Miles Davis Quintet	Miles Davis
Steamin with the Miles Davis Quintet	Red Garland
Steamin with the Miles Davis Quintet	John Coltrane
Steamin with the Miles Davis Quintet	Philly Joe Jones
Steamin with the Miles Davis Quintet	Paul Chambers
Lush Life	John Coltrane
Lush Life	Earl May
Lush Life	Art Taylor
Red Alone	Red Garland

Database Queries

The following queries illustrate a range of features that are available in the query language, SQL.

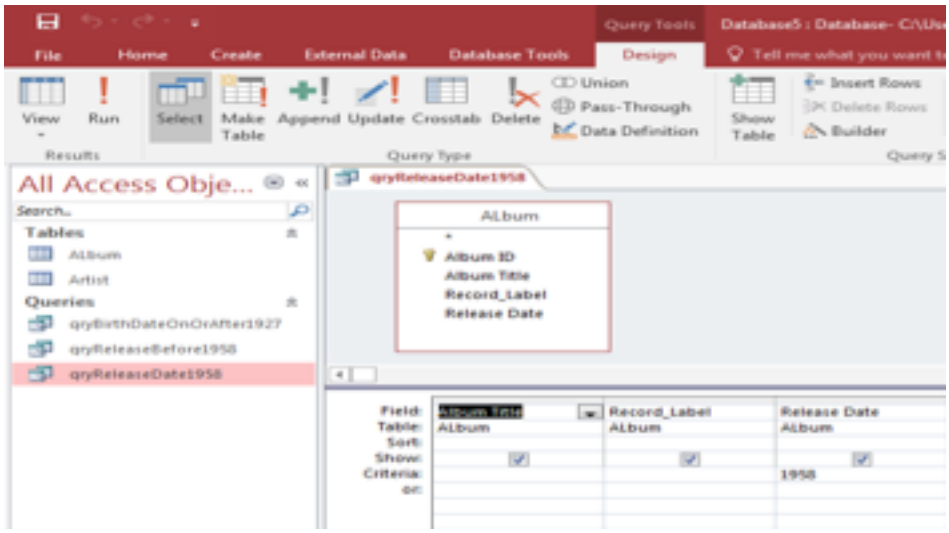
Query 1: Which albums were released in 1958?

```
SELECT Album_Title
FROM Album
WHERE Release_Date = 1958
```

Query 1 shows how the '=' operator is used in the where clause to return only those records with a release date field equal to 1958.

Album_Title
Relaxin with the Miles Davis Quintet
Lush Life

Using QBE the = operator is implied when the year is typed into the criteria.



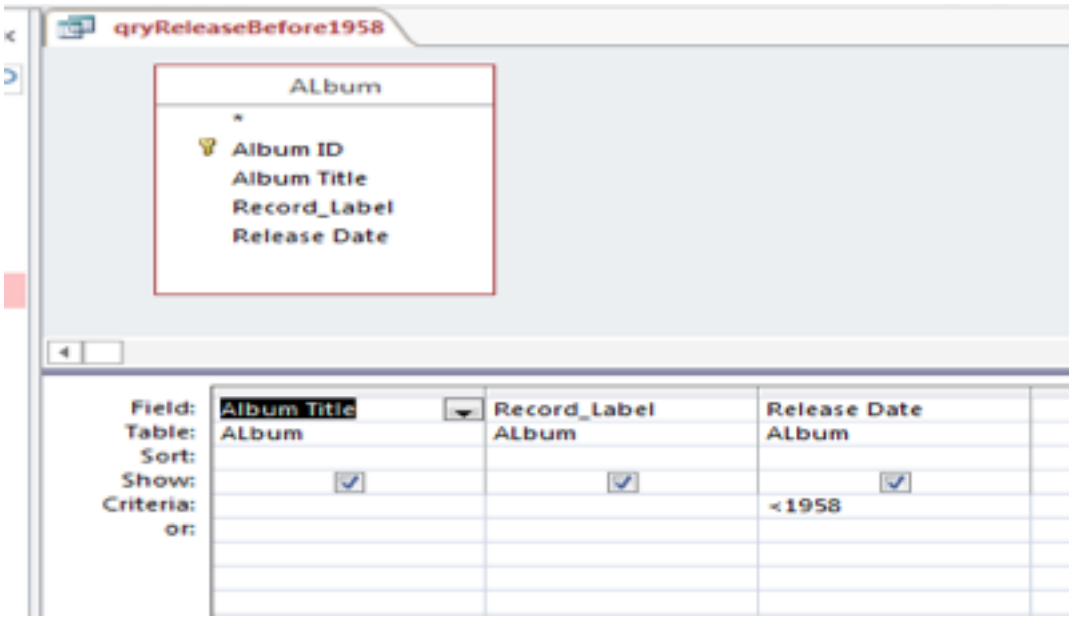
Query 2: Which albums were released earlier than 1958?

```
SELECT Album_Title
FROM Album
WHERE Release_Date < 1958
```

Album_Title
Cookin with the Miles Davis Quintet

Query 2 shows how the '<' operator is used to return only those records with a release date field less than 1958.

QBE



Query 3: Which albums were released later than 1958?

```
SELECT Album_Title
FROM Album
WHERE Release_Date > 1958
```

Album_Title
Steamin with the Miles Davis Quintet
Red Alone

Query 3 shows how '>' operator is used to return only those records with a release date field greater than 1958.

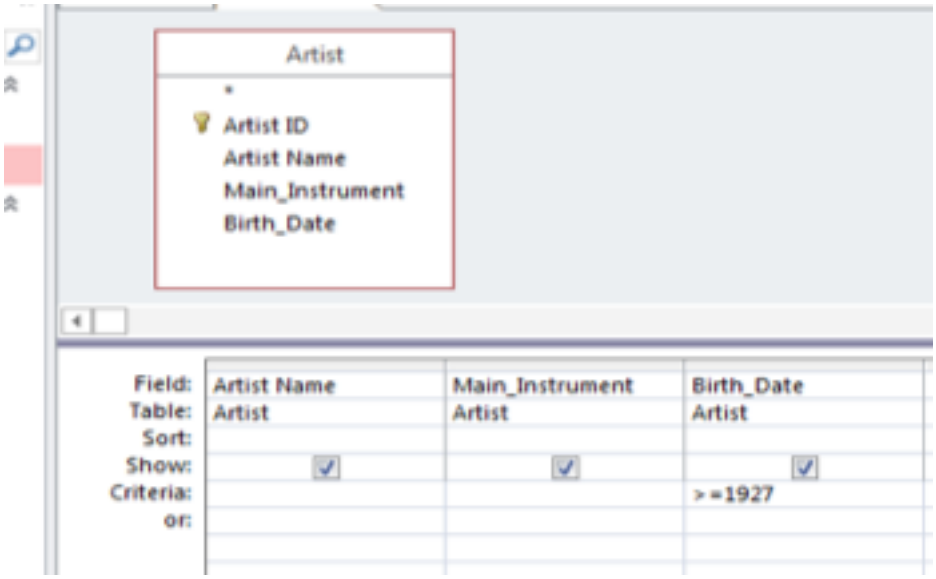
Query 4: Which artists were born in 1927 or later? Give full artist details.

```
SELECT *
FROM Artist
WHERE Birth_Year >= 1927
```

Artist_Name	Main_Instrument	Birth_Year
Paul Chambers	Bass	1935
Earl May	Bass	1927
Art Taylor	Drums	1929

Query 4 shows how '>=' is used to return only those records with a birth year field greater than or equal to 1927. It is also possible to use '<=' to mean less than or equal to.

QBE



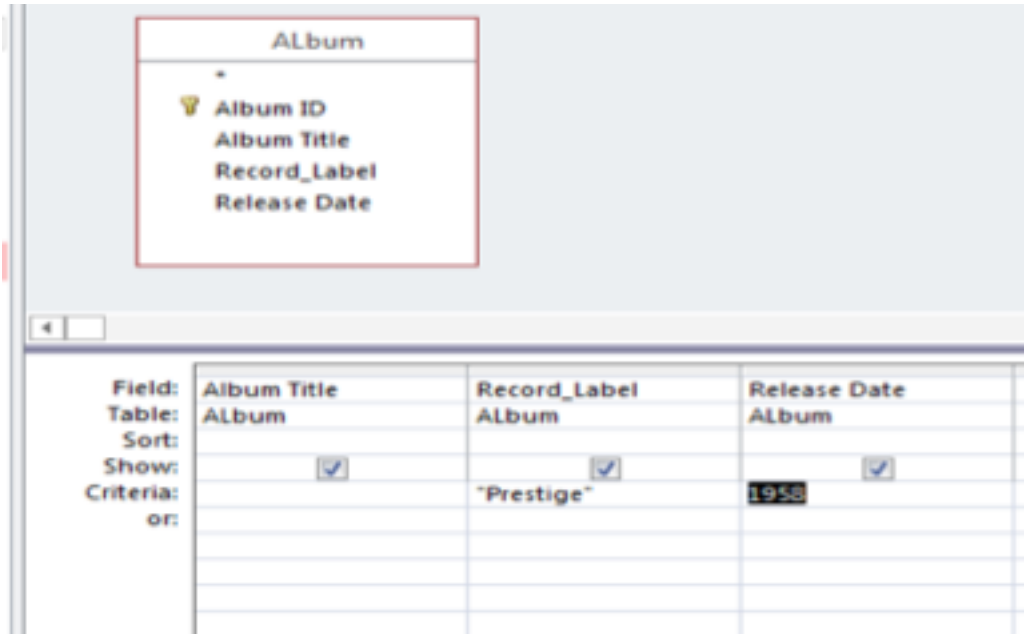
Query 5: Which albums were released on the Prestige label in 1958? Give full album details.

```
SELECT *
FROM Album
WHERE Record_Label = 'Prestige' AND Release_Date = 1958
```

Album_Title	Record_Label	Release_Date
Relaxin with the Miles Davis Quintet	Prestige	1958

Query 5 shows how the 'AND' operator is used to return only those records that simultaneously satisfy two different constraints. It is also possible to combine more than two constraints using this operator.

With QBE the 'AND' operator is implied as the criteria is entered into separate fields



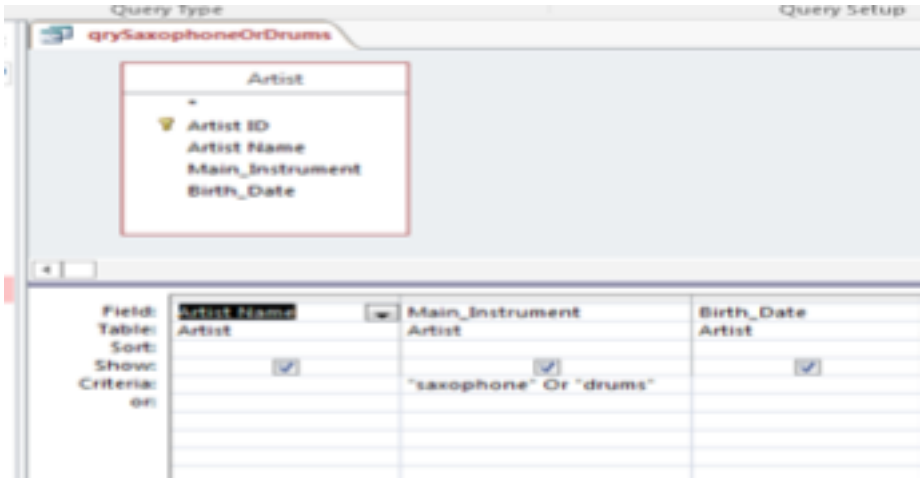
Query 6: Which artists play either saxophone or drums? Give full artist details.

```
SELECT *
FROM Artist
WHERE Main_Instrument = 'saxophone' OR Main_Instrument = 'drums'
```

Artist_Name	Main_Instrument	Birth_Year
John Coltrane	Saxophone	1926
Philly Joe Jones	Drums	1923
Art Taylor	Drums	1929

Query 6 shows how the 'OR' operator is used to return only those records that satisfy at least one of two alternative constraints. It is also possible to combine more than two alternative constraints using this operator.

Using QBE the 'OR' can be seen to the left of screen below criteria, you can type in 2 separate rows or write as can be seen below.



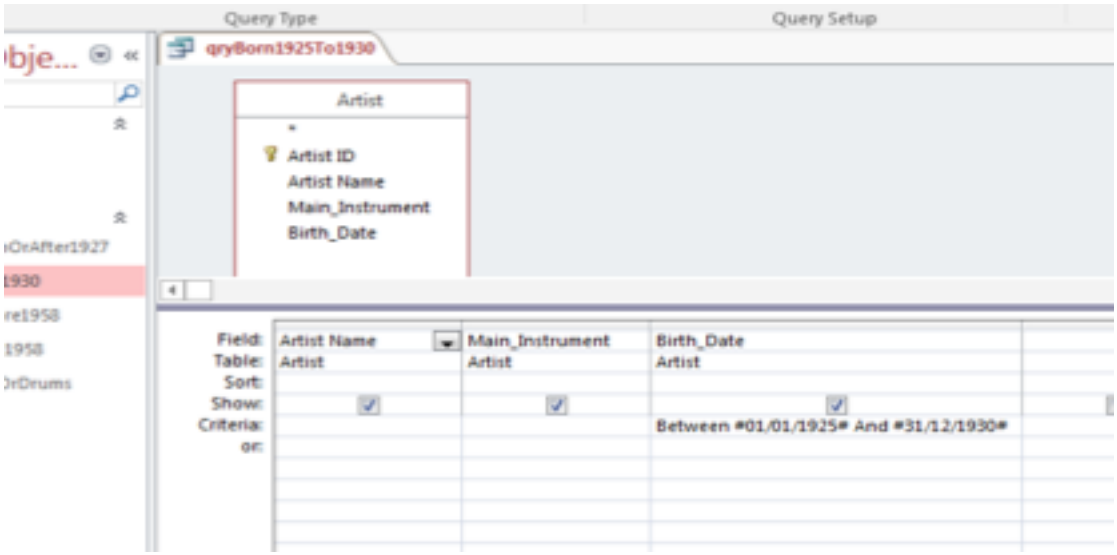
Query 7: Which artists were born after 1925 but before 1930? Give full artist details.

```
SELECT *
FROM Artist
WHERE Birth_Year BETWEEN 1925 AND 1930
```

Artist_Name	Main_Instrument	Birth_Year
Miles Davis	Trumpet	1926
John Coltrane	Saxophone	1926
Earl May	Bass	1927
Art Taylor	Drums	1929

Query 7 shows how the 'BETWEEN' operator can be used to return only those records with a birth year field that lies between two specified values.

There are occasions when you have to type the operators using QBE as in this example below.



Query 8: What albums did John Coltrane play on, and what are the corresponding record labels?

```
SELECT Album.Album_Title, Record_Label F
ROM Album, PlaysOn
WHERE Artist_Name = 'John Coltrane' AND PlaysOn.Album_Title = Album.Album_Title
```

Album_Title	Record_Label
Steamin with the Miles Davis Quintet	Universal
Lush Life	Fantasy

Query 8 is a little more complicated than the previous ones so a more detailed explanation may help:

- No single table contains all of the information needed to answer the query. The albums that John Coltrane played on can be found on the PlaysOn table, while the corresponding record labels can only only be found on the Album table. In order to make use of information from both tables, they must both be listed in the FROM clause of the query.
- Both the named tables have a field named Album_Title, so it is necessary to distinguish between using the dotted notation: **Album.Album_Title** means the Album_Title field from the Album table, while **PlaysOn. Album_Title** means the Album_Title field from the PlaysOn table.

Big Data

The term *Big Data* has been used to refer to a range of problems and technologies that relate to the management of very large data sets. The amount of data that we currently generate is enormous, and is increasing each year.

> *"IBM say that 'every day, we create 2.5 quintillion bytes of data – so much that 90 per cent of the data in the world today has been created in the last two years alone.'"*
> http://www.bcs.org/upload/pdf/big-data.pdf

Sources of this data include:

- text, images and videos uploaded to social media sites;
- medical records;
- fitness monitors;
- financial markets – share price data, currencies, etc.;
- web server logs;
- mobile phone records;
- ecommerce transactions;
- device logs (internet of things).

There is no generally agreed definition of big data, and it is clearly not just about the size of the data sets. There is, however, widespread recognition that big data problems are characterised by what have become known as the *three Vs*: volume, velocity and variety.

Volume

It is difficult to put a size limit in what counts as *big* data; indeed what was considered big yesterday may not be considered big tomorrow. It is clear, however, that big data problems involve data sets that are so large and complex that traditional tools, such as relational database management systems, are not able to deal with them effectively.

Velocity

Big data problems frequently have to deal with real time flows of data. Events happen in the real world and the data generated is real time, and is often time sensitive. For example:

- It has been estimated that, in September 2016, there were almost 891 thousand transactions per day on the London Stock Exchange (Statista, 2016). Each transaction results in money moving between accounts, as well as shareholder registers being updated.
  https://www.statista.com/statistics/325326/uk-lse-average-daily-trades/
- It has been estimated that there are between 4.5 and 9 million security closed circuit television cameras (CCTV) in the UK. While there is debate about the desirability of all this surveillance, it is hard to ignore the potential to enhance security and combat crime.
  http://www.bbc.co.uk/news/uk-30978995

Variety

Much of the data being produced is unstructured – this is not data that can easily be put into relational database tables. It is video, audio, and image, as well as text and numeric data.

The effective management of the volume, velocity and variety of all this data is an important challenge for the big data industry.

Interpreting Big Data

The management of big data is, however, only one part of the problem. There is no value in managing data unless we can interpret it and extract useful information. This is the role of *data analytics*.

Data analytics applies algorithms to raw data in order to spot patterns, relationships and trends. The algorithms may include statistical analysis as well as artificial intelligence techniques.

Applications

Potential applications of big data and analytics include:

- Healthcare.
  - https://www.newscientist.com/article/dn28340-big-data-better-health/
  - http://www.bbc.co.uk/news/health-38055509

- Crime prevention combat crime.
  - http://www.bbc.co.uk/guides/zqsg9qt
  - http://graymattersystems.com/big-data-crime-patterns/
  - http://www.bbc.co.uk/news/uk-30978995

- Energy Management.
  - http://www.bbc.co.uk/news/business-35722324

You can find more applications of big data and data analytics here:

- http://www.bbc.co.uk/news/business-29147254

Exercises

1. Write SQL queries to answer the following questions.

   a. Which artists were born in 1927 or earlier?

   b. On which albums did Red Garland play?

   c. Which albums were released in the 1950s?

2. Write a query that is equivalent to query 7 but without using the BETWEEN operator.

3. Repopulate the database structure used in the SQL examples above with data that relates to your own musical preferences. In other words, take the existing table structures and populate them with your own data. You might find it useful to consult the AllMusic website: http://www.allmusic.com/

Resources

W3Schools provides excellent resources on many aspects of web development. This link takes you to an interactive SQL processor that will enable you to try out SQL queries on an existing database.

- http://www.w3schools.com/sql/trysql.asp?filename=trysql_select_all

phpMyAdmin is a widely used, free, web-based database management system. This link takes you to an instance of phpMyadmin where you can create your own database and run SQL queries against it. The queries in this Fact File were tested using phpMyAdmin and mySQL.

- https://demo.phpmyadmin.net/master-config/

You can find a collection of BBC stories involving big data, here:

- http://www.bbc.co.uk/news/business-29147254