QUESTION

9 Identify and describe three components of a data dictionary.

10 Select the most appropriate data type for the following information stored about flights:

Flight number (e.g. BA372)
Departure date
Departure time
Airport code (e.g. ACF)
Max number of passengers
Type (e.g. scheduled or charter)
Arrived?

11 Give reasons for the use of the text data type for storing a mobile phone number.

9.04 File and data management

File types

When data is saved it is stored in a file. Different software applications use data in different ways and so the way the data is stored differs between application types. For example, a database stores data in tables, whereas graphics software stores data about pixels.

Each file will typically include a header, which will be metadata (data about the file), then the main content will be stored followed by an end-of-file marker.

To a user, file types are usually identified by their extension. For example, Students.txt has an extension of txt which identifies it as a text file.

EXAMPLE

Examples of file types include:

Extension	File type	Purpose
.txt	Text	Stores plain text without any formatting. It is useful for transferring data between applications, but any formatting is lost.
.csv	Comma separated values	Stores structured data as plain text in rows with each column separated by commas. It is useful for transferring data between databases and spreadsheets or other applications which require data in a structured format.
.rtf	Rich text format	Stores text-based documents and includes the formatting (rich text). It is used to transfer data between different word processing or other text-based applications.
.docx	Microsoft Word XML document	Stores Microsoft's word processing documents in open XML format by saving all objects separately within a compressed file.
.pdf	Portable Document Format	Used to share read-only documents in a common format that can be accessed by any PDF reader software. It is commonly used for storing documents on the web as its contents can be indexed by search engines.
.odt	OpenDocument Text	An open-source file type for word processor documents that is used by open-source word processors and is not tied to one manufacturer.
.ods	OpenDocument Spreadsheet	An open-source file type for spreadsheets that is used by open-source spreadsheet software and is not tied to one manufacturer.
.odp	OpenDocument Presentation	An open-source file type for presentations that is used by open-source presentation software and is not tied to one manufacturer.
.html	Hypertext Markup Language	Stores web pages that can be opened by any web browser.
.xml	Extensible Markup Language	A data file that uses markup language to define objects and their attributes. They are used to transfer data between applications and can be read by a simple text editor.
.avi	Audio Video Interleave (video file)	Microsoft's method of storing video files with very little compression. File sizes are very big but no data is lost.

Extension	File type	Purpose
.mp4	Moving Pictures Experts Group (MPEG) Layer-4 (video file)	Audio and video are compressed and videos can be shared across the internet.
.wav	Waveform Audio File Format	Stores audio files as waveform data and enables different sampling rates and bit rates. This is the standard format for audio CDs but does not include compression so files are large.
.mp3	MPEG Layer-3 audio compression	Stores audio files in a compressed format approximately 10% the size of .wav files. Enables audio files to be shared across the internet.
.bmp	Bitmap image	Stores images as uncompressed raster images, storing each pixel individually. They are large files but can be accessed by any software.
.jpg	Joint Photographic Experts Group (compressed image)	Stores images as compressed raster images. It is used by most digital cameras and is a common format for web graphics but its use of lossy compression can mean some quality is lost.
.png	Portable Network Graphic	Stores images as compressed raster images and can include background transparency colours making it useful when images are required on different colour backgrounds.
.svg	Scalable Vector Graphics	Stores images as two-dimensional (2D) vector graphics. It is a standard format for using vector graphics on the web.
.exe	Executable program file	Stores program object code which enables the program to be executed by the computer.

Proprietary and open-source file formats

Proprietary file formats

Proprietary file formats are file types that have been developed by software manufacturers solely for use within their software. Using their own formats means that software manufacturers are free to develop software features that will store data in a way that is most suitable for the software and without waiting for a standard format to adapt to the software's needs. This enables software to improve and provide new features that otherwise would not be available.

EXAMPLE

Some examples of proprietary file formats include:

Extension	Software / file type	Manufacturer
.docx	Word processor	Microsoft Word
.wpd	Word processor	Corel Word Perfect
.msg	Email message	Microsoft Outlook
.ra	Audio / video streaming	Real Networks
.MOV	Movie	Apple
.psd	Graphics	Adobe Photoshop
.ai	Graphics	Adobe Illustrator
.accdb	Database	Microsoft Access

Open-source file formats

Open-source file formats are file types that have been developed for the purpose of being used by any proprietary software or open-source software. They are free from copyright, patents and trademarks, and their structure is known publicly. They are usually maintained by an international standards organisation or a public interest group. Their main advantage is that the files can be shared between users of different software. However, they can hold back development of open-source software because new features will require the file format standard to be updated.

EXAMPLE

Some examples of open-source file formats include:

File type	Type of data	Standards organisation
JPG	Compressed raster graphics	Developed by the Joint Photographic Experts Group (JPEG) and standardised by the International Organization for Standardization (ISO)

File type	Type of data	Standards organisation
PNG	Compressed raster graphics with transparency support	ISO
ePub	E-book	International Digital Publishing Forum
XML	Extensible Markup Language	World Wide Web Consortium (W3C)
MPEG	Compressed video	Developed by the Moving Picture Experts Group (MPEG) and standardised by the ISO.

Generic file formats

Generic file formats enable data to be transferred between software. Data can be exported from software to a generic file format and generic file formats can be imported into software. They store the essential data but will not include any formatting.

The two main file formats used within databases are CSV and TXT. These were described earlier in this chapter in the section about importing data.

Indexed sequential access

Many years ago, data was often stored on tape, which required records to be written one after another onto the tape. This was known as storing the data serially. To access the data, all the records would need to be read from the first onwards until the required record was found or until the end of the file was reached. It could take a very long time to read through a whole table of data and so indexed sequential files were developed.

Indexed sequential files still store records one after each other but they are sorted into an order based upon a field. For example, data about customers might be sorted into surname order or customer ID order. Sequential files are particularly useful when data is being batch processed such as when gas bills are being generated and the master customer file will be processed in order of customer ID and any transaction files will also be processed in order of customer ID.

EXAMPLE

Here is an example of part of a master customer file showing the customers, the date the current meter reading was taken, the previous meter reading (amount of gas used) and the current meter reading:

Customer ID	Surname	Date of reading	Previous reading	Current reading
10	Black	12/1/16	32721	34872
11	Brown	15/12/15	02717	03281
12	White	8/1/16	47270	48572
13	Green	8/1/16	21827	23593

Here is an example of part of a transaction file that will be processed to update the master customer file with new gas meter readings:

Customer ID	Date of reading	Meter reading
11	12/3/16	03692
13	12/3/16	23997

This is what the master customer file will look like once the transaction file has been processed:

Customer ID	Surname	Date of reading	Previous reading	Current reading
10	Black	12/1/16	32721	34872
11	Brown	**12/3/16**	**03281**	**03692**
12	White	8/1/16	47270	48572
13	Green	**12/3/16**	**23593**	**23997**

However, when reading the data, it was still necessary to read the whole file serially from the beginning because there was no way of knowing where each record was stored. Indexed sequential files are stored in exactly the same way as sequential files but the file also has an index based on the field used to sort the file. A field with an index is known as a secondary key. The index file stores each secondary key value and the address in storage (e.g. tape or disk) where the first record containing that value is stored.

The index is small enough to store in main memory and so all that needs to be done to find a record is to search the index, find the location in storage and then read the records from that point until the record is found.

EXAMPLE

If Employee ID is the secondary key, then an index will exist with Employee ID as one column and the storage address as the other column. Rather than storing every single Employee ID, the index may store every tenth Employee ID for example.

Employee ID	Storage address
0001	A8FB2DC3
0011	9AEB08E3
0021	8C4DDDF5

Direct file access

The use of indexed sequential file access still requires some serial access of data and there are problems with trying to maintain a file in a sequential order as new records are added and old records deleted.

With direct file access, records are stored in a random order. There is no sequence. When storing a file, a hashing algorithm (calculation) is performed on the key field to determine the storage address where the record should be stored. Then when the record is searched for, the same hashing algorithm will be performed on the key field to determine where the record can be found. The computer system can then directly access that record without having to read through other records.

Hierarchical database management systems

The hierarchical database model was created in the 1960s and is not commonly used today. The model relies upon a tree structure where each parent branch is the one side of a relationship and each child branch is the many side of a relationship. The tree structure can only deal with one-to-many relationships and can only work in one direction. Hierarchical databases are only suitable for models which have a strict hierarchy.

One such hierarchy is the file system used within computer systems. The file system may look something like this:

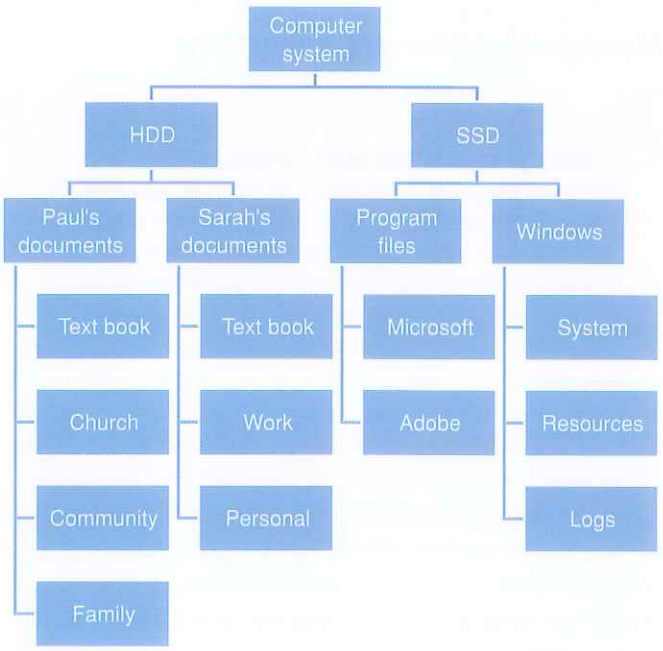

Figure 9.41 - Folder structure.

Each disk contains folders and there may be further subfolders within each folder. Each subfolder has only one folder at the level above it. To find the data, the user browses through the system, selects the disk the data is stored on, then selects the folder, then selects the next subfolder until eventually the file is found.

This same process is used when searching for data within a hierarchical database. This means that data at the top of the tree is very fast to access.

EXAMPLE

A bank could store data about customers and the accounts they hold:

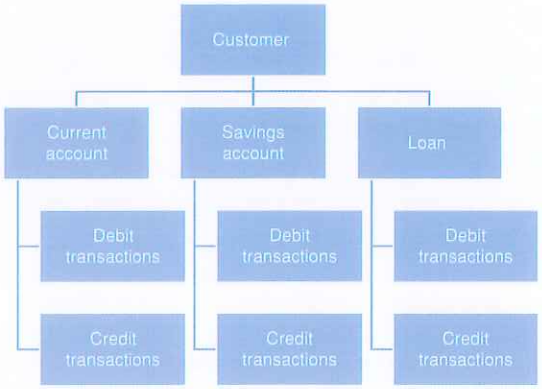

Figure 9.42 - Hierarchical bank.

Management information systems

Remember

A **management information system** (MIS) provides summary information to managers to enable them to make decisions. The MIS will collate data from a database and present it in the form of reports and charts. These reports and charts can be produced within the database system itself or they may be part of an additional piece of software that is used to analyse the data.

The additional software is likely to collate data from more than one database and interconnect the data from those databases to produce reports that analyse all the data together. When additional software is used to collate data from more than one database, it is often referred to as an executive information system (EIS).

A MIS has the following essential features:

- data is collated from databases and other sources
- data is interconnected from different sources
- data is analysed to provide the data that is required by management
- summary reports and charts are produced for managers that will help with decision making.

The reports and charts are created by people, but once they are created they can be reused as the data changes within the data sources. It's important that the reports and charts provide information that managers need.

Using MISs

TIP

Information from a MIS is used by managers to make decisions. Managers can examine the summary information and then decide upon actions to take. Reports are provided at regular times and it's also possible for managers to request ad hoc reports if they need additional information.

EXAMPLE

Managers within a large second-hand car dealership need to be able to monitor sales. They need to be able to identify trends in sales for different makes and models of cars at different times of the year. This will enable them to identify which cars are selling the most and which are making the most profit. They can then decide which second-hand cars they want to acquire to sell.

Marketing managers can analyse how effective a marketing campaign was by comparing sales figures during an advertising campaign with sales figures outside the advertising campaign. This will help them to decide whether to run similar campaigns in the future.

QUESTIONS

12 Explain why generic file types are needed.

13 Describe the steps involved to find a file using indexed sequential access.

14 Explain why direct access is used for databases in preference to indexed sequential access.

15 Describe two features of management information systems (MISs).

9.05 Summary

A database contains structured data in tables which consist of records and fields. Data in fields has a type assigned to it, such as text, alphanumeric, integer/decimal, date/time or Boolean.

A flat file is a single table and has no relationships. Relationships connect entities (tables) together and can be one-to-one or one-to-many. Hierarchical

databases are based on a tree structure to deal with one-to-many relationships. Relationships are depicted in an entity relationship diagram (ERD). A primary key is a unique identifier for a record, a compound key is a primary key consisting of more than one field and a foreign key relates to a primary key in another table.

Referential integrity ensures that data exists in a related table. Validation rules can be used to ensure that data is sensible and allowed. Verification is the process of checking data has been transferred correctly.

Simple queries use one criterion to search for data and complex queries use two or more criteria. Summary queries can be used to find statistical information from a database. Static parameters are used in queries when the value of the parameter does not change and dynamic parameters are used when the user is likely to want to change the value each time the query is run.

Indexed sequential access involves the use of an index to determine where to start searching a file for a record. Direct file access involves using a hashing algorithm to find the location of a record in a file.

Normalisation is the process of structuring data within a database and is measured using normal forms. A data dictionary, known as metadata, describes the structure of the data held within the database.

Data entry forms are used for inputting data into a database. Data can be imported into a database from another data source or exported so it can be used in other software. Different software applications require different data types in order to store data.

Proprietary file formats are developed by manufacturers for their own software and open-source formats are developed for use by any software.

A management information system (MIS) provides summary information to managers to enable them to make decisions.

Review questions

A website accepts donations for charities. Each donor may make several donations to one or more charities. This information is stored in a relational database.

1a Identify three tables that should be used within the database. [3]

1b Describe two relationships that would be used within the database. [2]

1c Explain how referential integrity is important to this database. [2]

An apartment complex stores data about its customers, their bookings and the rooms they are staying in. The entity relationship diagram (ERD) is shown below:

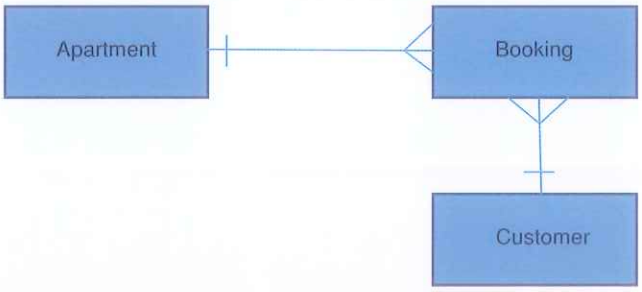

Figure 9.43 - Entity relationship diagram.

2a Identify two foreign key fields that should be used within the database. [2]

2b Select the most appropriate data type for each of the fields below in the apartment table: [3]

(i) Telephone Number

(ii) Swimming Pool

(iii) Bedrooms

2c Describe how a dynamic parameter query could be used to produce a list of customers that have stayed in an apartment during a specified time period. [4]

2d Explain why this query would be a complex query. [2]

2e Identify and describe three items of a data dictionary that could be used in this database. [6]

Students in a college belong to tutor groups. Each tutor group has one tutor. The students are able to borrow books from the college library.

4 Normalise the unnormalised data below to 3NF. Show each table, its attributes and its primary keys. [4]

STUDENT
Name
Address
Telephone
Tutor Group
Tutor Name
    Book ID
    Title
    Due Date

5 Describe the difference between proprietary and open-source file formats. [2]