Data Management - LaTeX Report

February 2019

Emilia Szynkowska
30186773
eas1g18@soton.ac.uk

1 Scripts

1.1 Basic File Processing with Unix

Code for countreviews.sh

```
#shebang - shows the code is bash
#!/bin/bash

location=$1
echo "Path:" $location

#iterates through all the files in the reviews folder
for file in $location/*; do {

#prints the file name
#finds all occurences of the pattern <Author>
#-c counts the number of occurences
filename=$(basename $file .dat)
count=$(grep -c '<Author>' $file)
echo "$filename $count"

} done | sort -nr -k2

#sorts in descending order of reviews
#-n means sort by numerical value
#-r means reverse
#-k2 means sort by the second column
```

1.2 Data Analysis with Unix

Code for countreviews.sh

```
#!/bin/bash

location=$1
echo "Path: $location"

#iterates over each file in the reviews folder
for file in $location/*; do {

    #prints the filename
    #finds all values for the overall rating
    #finds the average by dividing the total over the number
       of values
    #rounds to 2 decimal places
    filename=$(basename $file .dat)
     count=0; total=0;
     a=$(grep -E '<Overall>' $file | sed 's/<Overall>//' |
        awk '//{count++; total+=$0} END {print total/count}')
     average=$(printf "%0.2f" $a)
     echo "$filename $average"

} done | sort -nr -k2

#sorts the averages in descending order
#-n means sort by numerical value
#-r means reverse order
#-k2 means sort by the second column
```

2 Discussion

2.1 Structured and Unstructured Data

Structured Data: Structured data is data that has been organized into a formatted repository, typically a database, so that its elements can be made addressable for more effective processing and analysis.
(Source: Techtarget)
Unstructured Data: Unstructured data is information that either does not have a predefined data model or is not organized in a predefined manner.
(Source: Wikipedia)

Examples of structured database systems include SQL and noSQL, DB2, Oracle, and Teradata. Structured databases follow a **relational model**, a concept which was invented by Edgar Codd in the 1970s. In a relational model, data is stored in a table which contains attributes. These are arranged in rows and columns.

Attribute: property which defines a relation
Schema: representation of a group of data e.g. Student(Name, Age, ID)
Tuple: row in a table
Cardinality: number of tuples in a table
Column: set of values for a particular attribute

- Structured data is machine-readable. Machines are much faster than humans at processing large amounts of data. The fastest algorithms can reach best-case time complexities of log(n)

- Data in relational models can be queried, for example XML data is queried with XQuery. This is essential for analysing data to find patterns and trends. Data analysis is used to understand customers and improve a company's services

- Databases provide a better framework for security and privacy. Methods of data protection include encryption, firewalls/proxies, passwords, system/object privileges, and auditing:
  Encryption: data is encoded. Public-key encryption uses two different keys at once - to decode an encrypted message, a computer uses the public key, provided by the original computer, and its own private key
  Firewalls and proxies: a proxy server is an intermediate between a server and computer which allows data to be filtered before reaching its destination, whereas a firewall is a security system which controls network traffic. Firewalls and proxy servers allow harmful files to be recognised and removed
  Auditing: changes in data can be monitored. Users can track access to the database and changes to objects inside it

- Databases can store huge amount of information. The human brain is estimated to have a maximum capacity between 1 terabyte and 2.5 petabytes of data, and is highly susceptible to memory leakage. However, if servers or other forms of computer memory are connected together, they can store an almost infinite amount of non-volatile data

- Digital storage takes less space than writing on paper and is more eco-friendly. Electronic data is also easily changed and copied

Unstructured data is written in the form of raw text, such as Word documents, emails, images, audio, and video. This type of data cannot be queried and must be searched for the appropriate information.

- Unstructured data is human-readable

- Unstructured data contains a large amount of useful information which can be analysed for marketing and advertising. The majority of data on the internet is unstructured:
  Social media: data is automatically created whenever a user mentions a company or product in a social media post
  Media: images and videos are unstructured but contain a range of useful data
  Customer-generated content: online reviews, comments, emails, and phone calls contain feedback which a company can use

- However, unstructured data is not compact and takes up a large amount of storage. Not all unstructured data is useful

- Analysis of unstructured data is slow, as the system must search for particular patterns. This is in contrast to structured databases, in which data can be extracted immediately by looking at its attributes

In conclusion, structured databases are significantly better than unstructured data as they allow computers to filter, edit, and refine large amounts of data with increased accuracy and efficiency.

2.2 The Review Ranking System

This system uses .dat text files containing tags followed by the associated data, e.g. <Overall>3. Each review contains the Author, Content, Date, Overall Rating, Value, Rooms, Location, Cleanliness, Service, and other information.

This data is presented as raw unordered text. It cannot be classed as SGML because it does not use closing tags, and does not use attributes.

2.2.1 Advantages and Disadvantages of the Review Ranking System

Advantages:

- Human-readable, easy to understand
- Each review is separated by line breaks
- Each element is on a new line
- Tags are provided to separate data elements
- Each review uses the same tags, these are consistent across the file
- Files have appropriate names, e.g. hotel_85003.dat

Disadvantages:

- Is not written in a database language such as SQL
- Cannot be queried to search for particular data
- Cannot be ordered easily by particular tags or values
- Must be searched using text parsing commands such as grep and sed, this is highly inefficient
- Sorting the data is slow and takes up time
- It is difficult to collect and present the data due to its unorganised format

2.2.2 Authenticating the authors of reviews

One issue with the review system is authentication of users. The system must filter out comments which contain spam, harassment or bad language, and must also check the comment has been made by a valid user.

The most obvious way to do this is to require users to register on the hotel website. Each user will have to create an account and choose a username and password. To further increase security, the password should be at least 8 characters long and contain a number, special character, or capital letter. The database should not store the actual passwords; instead it should use a salted hashing or encryption.

Hashing is a technique where an algorithm is used to generate a hash value, which is used to map data to a location in a hash table. Salted hashing is an extra step where an additional value known as a 'salt' is added to the end of a password before it is hashed, which alters the resulting hash value.

Encryption is when a type of algorithm called a cipher is used to encrypt and decrypt data. In public-key encryption, public and private keys are used; one key encrypts and one key decrypts. This means you must know both keys in order to read the data, and is very secure.

Lastly, CAPTCHAs (Completely Automated Public Turing tests to tell Computers and Humans Apart) are an easy to make sure users are human and to prevent automated attacks because they limit the number of password entry attempts. A CAPTCHA will usually involve recognising distorted letters or images.

To check the content of a review, the computer can parse over the text to perfrom simple operations like finding particular words or phrases. Alternatively for more thorough checking the comment can be saved and read by a moderator before it is posted/published.

2.2.3 Improvements on the Review Ranking System

To improve this system, the company can use a markup language or database to store the data.

One way of doing this is to use XML (eXtensible Markup Language). This type of markup language is a subset of SGML (Standard Generalised Markup Language) and is more ordered than plain text. Elements in XML must contain entities, elements, and attributes. An entity is a character used to write text, such as a letter, number, or symbol. An element represents structure or desired behaviour; it has an opening tag, content, and a closing tag. An attribute is a property of an element which controls its behaviour and allows it to be located within a database. XML is useful because XQuery can be used to query the data. XQuery can search for nodes, subnodes, locations, and content within data. There are many predefined commands available which can be used to find information quickly.

JSON (JavaScript Object Notation) is a good way to store and transport data over the internet. .json files contain data in attribute-value pairs, separated by commas. Data is also separated into objects and arrays. It is a good idea to use JSON because it can be parsed instantly and converted to JavaScript using the JSON.parse() command. This means the relevant information can easily be found from files and extracted.

SQL is designed for storing data in a RDBMS (Relational Database Management System). SQL uses four sublanguages: data querying language (DQL), definition (DDL), control (DCL), and management (DML). These offer a wide range of operations such as insertion, deletion, updating, schema creation and modification, and data access control.

Elements of SQL:
Clauses: components of statements and queries
Expressions: return scalar values or tables containing data
Predicates: variables which are evaluated to Boolean values
Queries: retrieve data based on specific criteria
Statements: control schemata, data, program flow, connections, sessions, diagnostics

The most efficient and secure way to store the data about the hotels would be to use a Rational Database Management System. The most popular systems include MySQL, Oracle, IBM DB2, and Microsoft Access. RDBMSs allow:

- **Multi-user access:** this prevents clashes when data is being changed by multiple users and helps to keep data up-to-date

- **Authorisation and privilege control:** the administrator can restrict or grant access to certain data, meaning that sensitive data is protected
- **Network access:** RDBMSs provide access to the database through a server daemon. This software receives and authenticates requests from clients to read and edit the database, and the database can be accessed from any computer
- **Maintenance:** maintenance utilities allow administrators to maintain, test, repair, and back up databases. These functions can also be automated
- **Querying:** SQL is a predefined language containing many commands to sort and analyse data

2.3 Data Storage Issues

Flat file systems can be inefficient because the files are not linked, meaning the time complexity of searching is $\Omega(n)$, where n is the number of files or elements. As well as this, the data can be difficult to share and is incompatible with multi-user systems. Compared to a relational database, it is relatively complicated to edit data as it must be added manually and with the correct syntax. Another issue with flat-file data is that it does not offer good security. Flat files have operating system level security, but do not offer access control over particular nodes or parts of data.

These issues can cause problems for a company:

- Hardware tends to be more expensive as more storage is needed
- More time is needed for search and sort algorithms to process big data
- Software costs can increase
- The data is prone to errors and redundancy
- Poor security means that it is easier for unauthorised users to access data