

Solutions – Chapter 3

Critical Thinking Exercise

Vehicle Theft Database

Review Questions

1. Student responses will vary. Some examples of attributes (and approximate sizes) might include: theft number (4 bytes), date of theft (8 bytes), time of theft (6 bytes), place of theft (20 bytes), vehicle type (10 bytes), vehicle make (10 bytes), vehicle model (10 bytes), vehicle date (4 bytes), vehicle color (10 bytes), license plate number (8 bytes), and vehicle identification number (17 bytes).
2. Given the attributes listed above, I would assign theft number as the primary key.

Critical Thinking Questions

1. It might be helpful to include status information in the database. The database might store information as to whether the vehicle was found, if there is a known suspect, if an arrest has been made and if the case is under investigation or closed.
2. Student responses will vary. There could be problems with keeping the data updated, especially if the investigation information is stored as investigation information is constantly changing. You would need something committed to keeping this information current.

Critical Thinking Exercise

Cleansing Weather

Review Questions

1. Because data is coming from different sources and needs to be exchanged and integrated into different systems, it would make sense that an enterprise data model exists.
2. The domain of acceptable values for barometric pressure and humidity might not include negative numbers and temperature might not go below -90 or above 60 if measured in Celsius.

Critical Thinking Questions

1. Student responses will vary. There could be a number of reasons why the data might be incomplete or inaccurate including malfunctioning equipment and the location of equipment.
2. Student responses will vary. One way to correct the data might be to cross-check data from several sources in the same location to identify data that is inaccurate.

Critical Thinking Exercise

Database as a Service

Review Questions

1. With database as a service the database administration is handled by the service provider. This would increase performance as the service provider will have the technology require to complete the database processing as quickly as possible. You will not have to worry about investing large amounts of time or money into improving your technology as the data grows.
2. Student answers may vary, but the disadvantages would be similar to those disadvantages with cloud computing. The main issue might be security. If there is a problem with the service provider you might not have access to your data.

Critical Thinking Questions

1. Student responses will vary. Before deciding to use database as a service, you should look at cost, performance, scalability, management and maintenance.
2. Using database as a service will allow you to concentrate on your role in IT rather than constantly managing databases.

Critical Thinking Exercise

Walgreens Data Assimilation

Review Questions

1. Regulations that apply to the use and management of Walgreens and Rite Aide data are the California Senate Bill 1386, HIPPA, and Personal Information Protection and Electronic Documents Act.

2. It would make sense for Walgreens to appoint a data governance team. The team defines processes for how the data is to be stored, archived, backed up, and protected from cyberattacks, inadvertent destruction or disclosure, or theft. It also develops standards and procedures that define who is authorized to update, access, and use the data. The team also puts in place a set of controls and audit procedures to ensure ongoing compliance with organizational data policies and government regulations.

Critical Thinking Questions

1. Student responses will vary. An automated process would be much easier but does nothing to update or correct incorrect information. Also, errors in data could be introduced during the automated process that go undetected. A manual process might be more accurate. It also might have the benefit of updating outdated or incorrect customer data, but it relies on the customer to actually manually enter his data. Some customers may not want to do this.
2. Student answers will vary. Customer data could be entered incorrectly potentially resulting in serious issues such as incorrect insurance information or wrong mediation information.

Critical Thinking Exercise

Telefonica Brasil

Review Questions

1. Like Verizon, Telefonica Brasil could use clickstream data, chats, and even social media searches, to predict when a customer might switch to a new carrier.
2. The firm might elect to use a data warehouse and a NoSQL database.

Critical Thinking Questions

1. It is unlikely that a traditional database is used since the data will not be modeled using a simple two-dimensional tabular relation.
2. Student answers will vary. Like Verizon, the firm might want to use Hadoop.

Review Questions

1. A database is a well-designed, organized, and carefully managed collection of data. A database management system (DBMS) consists of a group of programs

used to access and manage a database as well as provide an interface between the database and its users and other application programs.

2. An attribute is a characteristic of an entity. For example, employee number, last name, first name, hire date, and department number are attributes for an employee. The specific value of an attribute, called a data item, can be found in the fields of the record describing an entity. The domain for a particular attribute indicates what values can be placed in each column of the relational table.
3. The database approach to data management is an approach to data management where multiple information systems share a pool of related data.
4. Data archiving refers to how long data must be stored. When operating a database it's important to consider the storage needs for older data that is still important to the organization and needed for future reference.
5. Entity-relationship (ER) diagrams use basic graphical symbols to show the organization of and relationships between data. In other words, ER diagrams show data items in tables (entities) and the ways they are related. ER diagrams help ensure that the relationships among the data entities in a database are correctly structured so that any application programs developed are consistent with business operations and user needs.
6. Basic data manipulations include selecting, projecting, joining, and linking.
7. Data scrubbing (data cleaning or data cleansing) is the process of detecting and then correcting or deleting incomplete, incorrect, inaccurate, or irrelevant records that reside in a database.
8. DaaS is similar to Software as a Service (SaaS). A SaaS system is one in which the software is stored on a service provider's servers and is accessed by the client company over a network. With DaaS, the database is stored on a service provider's servers and accessed by the client over a network, typically the Internet, with the database administration handled by the service provider.

Amazon Relational Database Service (Amazon RDS) is a Database as a Service that enables organizations to set up and operate their choice of a MySQL, Microsoft SQL, Oracle, or PostgreSQL relational database in the cloud. The service automatically backs up the database and stores those backups based on a user-defined retention period.

TinyCo is a mobile gaming firm whose games Tiny Monsters, Tiny Village, and Tiny Zoo Friends can be found at the Amazon, Google Play, and iTunes app stores. The company employs Amazon Web Services (AWS) to enable it to

- support the rapid growth in the number of its users without having to devote constant time and effort to organize and configure its information systems infrastructure. This arrangement has allowed the company to focus its resources on developing and marketing its new games. TinyCo application data is stored in the Amazon Relational Database Service (Amazon RDS) for MySQL.
9. Hadoop is an open-source software framework with several software modules that provide a means for storing and processing extremely large data sets. Hadoop has two primary components: a data processing component (a Java-based system called MapReduce, which is discussed in the next section) and a distributed file system (Hadoop Distributed File System, HDFS) for data storage.
 10. Schemas are used to describe the entire database, its record types, and its relationships to the DBMS. Schemas are entered into the computer via a data definition language, which describes the data and relationships in a specific database.
 11. Concurrency control is a method of dealing with a situation in which two or more users or applications need to access the same record at the same time. Without proper database control, data updates might be incorrect, resulting in inaccurate records.
 12. An in-memory database (IMDB) is a database management system that stores the entire database in random access memory (RAM). This approach provides access to data at rates much faster than storing data on some form of secondary storage (e.g., a hard drive or flash drive) as is done with traditional database management systems. IMDBs enable the analysis of big data and other challenging data-processing applications, and they have become feasible because of the increase in RAM capacities and a corresponding decrease in RAM costs.
 13. Projecting involves eliminating columns in a table while joining involves combining two or more tables.
 14. Big data is the term used to describe data collections that are so enormous (terabytes or more) and complex (from sensor data to social media data) that traditional data management software, hardware, and analysis processes are incapable of dealing with them. There are three characteristics associated with big data – volume, velocity, and variety.
 15. A data warehouse is a database that holds business information from many sources in the enterprise, covering all aspects of the company's processes, products, and customers. Data warehouses allow managers to drill down to get greater detail or roll up to take detailed data and generate aggregate or summary

reports. The primary purpose is to relate information in innovative ways and help managers and executives make better decisions.

16. A data lake (also called an enterprise data hub) takes a “store everything” approach to big data, saving all the data in its raw and unaltered form, while a traditional data warehouse is created by extracting (and discarding some data in the process), transforming (modifying), and loading incoming data for predetermined and specific analyses and applications.
17. An in-memory database (IMDB) provides fast access to data because the entire database is stored in random access memory (RAM).

Discussion Questions

1. Student responses may vary. One major concern is error correction and loss of information. The most challenging problem within data cleansing remains the correction of values to remove duplicates and invalid entries. However, this is usually done through the deletion of data which leads to loss of information. This loss can be costly if there is a large amount of deleted data.
2. One data cleansing solution is to identify and correct the data by crosschecking it against a validated data set. For example, street number, street name, city, state, and zip code entries in an organization’s database may be cross-checked against the United States Postal Zip Code database. Data cleansing may also involve standardization of data, such as the conversion of various possible abbreviations (St., St, st., st) to one standard name (Street).
3. The ACID properties are atomicity, consistency, isolation, durability. These properties guarantee database transactions are processed reliably and ensure the integrity of data in the database. Basically, these principles mean that data is broken down to atomic values—that is, values that have no component parts—such as employee_ID, last_name, first_name, address_line_1, address_line_2, and city. The data in these atomic values remains consistent across the database. The data is isolated from other transactions until the current transaction is finished, and it is durable in the sense that the data should never be lost.

SQL databases rely upon concurrency control by locking database records to ensure that other transactions do not modify the database until the first transaction succeeds or fails. As a result, 100 percent ACID-compliant SQL databases can suffer from slow performance.

4. A NoSQL database is a way to store and retrieve data that is modeled using some means other than the simple two-dimensional tabular relations used in relational

databases.

The four main categories of NoSQL databases are:

- Key-value NoSQL databases—Similar to SQL databases, but have only two columns (“key” and “value”), with more complex information sometimes stored within the “value” columns.
 - Document NoSQL databases—Used to store, retrieve, and manage document-oriented information, such as social media posts and multimedia, also known as semi-structured data.
 - Graph NoSQL databases—Used to understand the relationships among events, people, transactions, locations, and sensor readings and are well-suited for analyzing interconnections such as when extracting data from social media.
 - Column NoSQL databases—Store data in columns, rather than in rows, and are able to deliver fast response times for large volumes of data.
5. Student responses may vary. One could argue that the Gramm-Leach-Bliley Act or the Health Insurance Portability and Accountability Act (HIPAA) has the most impact on safeguarding personal information.
 6. The ETL (extract, transform, load) process takes data from a variety of sources, edits and transforms it into the form to be used in the data warehouse, and then loads this data into the warehouse. This process is essential in ensuring the quality of the data in the data warehouse.
 7. Student responses will vary. Students should research the data collection policies of the IRS, their state’s Bureau of Motor Vehicles and Equifax. They should identify any concerns they may have over the collection or data storage policies.
 8. Student responses will vary. The following steps should be taken:
 - Determine the lifetime of information.
 - Select appropriate encryption technologies.
 - Set policies and procedures.
 - Install encryption technology.
 - Create an audit trail.
 - Train users.
 9. Student responses will vary. In the article, Weatherhead University Professor Gary King states, “The big data revolution is that now we can do something with the data.” The article goes on to mention that King explains that the revolution lies in improved statistical and computational methods, not in the exponential growth of storage or even computational capacity.

Problem-Solving Exercises

1. Students should develop a simple data model for student information.
2. Students should use a graphics program to develop an entity relationship diagram for a database application for a movie-streaming subscription service.
3. Students should use a database management system to build the data-entry screen discussed in question 2.
4. Students should modify the data-entry screen built in step 3.

Team Activities

1. Students should form teams and identify what other student attributes they might want to consider and how to collect this data. They should also create an ER diagram.
2. The teams should discuss the actions they would take to ensure that the student reporting needs and data privacy concerns of the students are fully identified.
3. Students should interview business managers from three different businesses that have implemented a database.

Web Exercises

1. Students should search a Web search engine to find information on the NSA PRISM program.
2. Students should search a Web search engine to find an organization that is struggling with big data growth.
3. Students should research and report on three different estimates of the rate at which the amount of data is growing.

Career Exercises

1. Students should research the roles of database administrator and data steward. A database administrator (DBA) is a skilled and trained IS professional who directs all activities related to an organization's database, including providing security from intruders. A data steward is an individual responsible for the management of critical data elements, including identifying and acquiring new data sources;

creating and maintaining consistent reference data and master data definitions;
and analyzing data for quality and reconciling data issues.

2. Students should discuss how they would use big data to do a better job at work.

Case Studies

Case One: WholeWorldBand: Digital Recording Studio

Critical Thinking Questions

1. Student responses will vary. The main data challenge the company faces might be supporting the uploading, streaming and storing of large audio and video files. Cloud computing can certainly help with this as the host company can configure and scale the infrastructure as needed.
2. The company likely employs a mix of both SQL and NoSQL. Managing the subscriber information likely requires a relational model.
3. Student responses will vary. Students should read and research the Terms of Use to identify how the company protects copyrighted material.

Case Two: Mercy's Big Data Project Aims to Boost Operations

Critical Thinking Questions

1. Redundancy is particular important in healthcare to ensure patients receive proper care at the appropriate time.
2. In a healthcare setting large amounts of data are being collection daily (volume and velocity). This data is being collected in a variety of ways from individuals (doctors and nurses) and machines (for example heart monitors).
3. Student responses will vary. With a large amount of time-sensitive data, it is very important to have the organization's data needs clearly defined.