Dirt Bikes U.S.A. sells primarily through its distributors. It maintains a small customer database with the following data: customer name, address (street, city, state, zip code), telephone number, model purchased, date of purchase, and distributor:

These data are collected by its distributors when they make a sale and are then forwarded to Dirt Bikes. Dirt Bikes would like to be able to market more aggressively to its customers. The Marketing Department would like to be able to send customers e-mail notices of special racing events and of sales on parts. It would also like to learn more about customers’ interests and tastes: their ages, years of schooling, another sport in which they are interested, and whether they attend dirt bike racing events. Additionally, Dirt Bikes would like to know whether customers own more than one motorcycle. (Some Dirt Bikes customers own two or three motorcycles purchased from Dirt Bikes U.S.A. or other manufacturer.) If a motorcycle was purchased from Dirt Bikes, the company would like to know the date of purchase, model purchased, and distributor. If the customer owns a non–Dirt Bikes motorcycle, the company would like to know the manufacturer and model of the other motorcycle (or motorcycles) and the distributor from whom the customer purchased that motorcycle.

- Redesign Dirt Bikes’s customer database so that it can store and provide the information needed for marketing. You will need to develop a design for the new customer database and then implement that design using database software. Consider using multiple tables in your new design. Populate each new table with 10 records.
- Develop several reports that would be of great interest to Dirt Bikes’s marketing and sales department (for example, lists of repeat Dirt Bikes customers, Dirt Bikes customers who attend racing events, or the average ages and years of schooling of Dirt Bikes customers) and print them.
The solution file represents one of many alternative database designs that would satisfy Dirt Bikes’ requirements.

The design shown here consists of 4 tables:
1. Customer,
2. Distributor,
3. Purchase, and
4. Model.

Dirt Bikes’s old customer database was modified by breaking it down into these tables and adding new fields to these tables.

Data on both Dirt Bike customer purchases captured from distributors and customer purchases of non-Dirt Bike models are stored in the Purchase table.

The Customer table no longer contains purchase data but it does contain data on e-mail addresses, customer date of birth, years of education, additional sport of interest, and whether they attend dirt bike racing events.
Case Study: The Terrorist Watch List Database’s Troubles Continue

1. What concepts are illustrated in this case?

The first concept illustrated in the case focuses on consolidating multiple databases into a single cohesive one. That is a primary goal of the FBI’s Terrorist Screening Center (TSC). The organization is integrating at least 12 different databases; two years after the process began, 10 of the 12 have been processed. The remaining two databases are both fingerprint databases, and not technically watch lists.

Using data warehouses to serve all the agencies that need information is the second major concept. Agencies can receive a data mart, a subset of data, that pertains to its specific mission. For instance, airlines use data supplied by the TSA system in their NoFly and Selectee lists for prescreening passengers, while the U.S. Customs and Border Protection system uses the watch list data to help screen travelers entering the United States [presumably in transportation other than an airplane]. The State Department screens applicants for visas to enter the U.S. and U.S. residents applying for passports, while state and local law enforcement agencies use the FBI system to help with arrests, detentions, and other criminal justice activities.
Case Study: The Terrorist Watch List Database’s Troubles Continue

1. What concepts are illustrated in this case?

Managing data resources effectively and efficiently is the third major concept in this case. No information policy has been established to specify the rules for sharing, disseminating, acquiring, standardizing, classifying, and inventorying information. Data administration seems to be poor. Data governance that would help the organizations manage the availability, usability, integrity, and security of the data seems to be missing. It would help increase the privacy, security, data quality, and compliance with government regulations. Lastly, data quality audits and data cleansing are desperately needed to decrease the number of inconsistent record counts, duplicate records, and records that lacked data fields or had unclear sources for the data.
Case Study: The Terrorist Watch List Database’s Troubles Continue

2. Why was the consolidated terror watch list created? What are the benefits of the list?

The FBI’s Terrorist Screening Center, or TSC, was established to organize and standardize information about suspected terrorists between multiple government agencies into a single list to enhance communications between agencies. A database of suspected terrorists known as the terrorist watch list was born from these efforts in 2003 in response to criticisms that multiple agencies were maintaining separate lists and that these agencies lacked a consistent process to share relevant information concerning the individuals on each agency’s list.
Case Study: The Terrorist Watch List Database’s Troubles Continue

3. Describe some of the weaknesses of the watch list. What management, organization, and technology factors are responsible for these weaknesses?

**Management:** Policies for nomination and removal are not uniform between governmental departments. The size of the list is unmanageable – it has grown to over 750,000 records since its creation and is continuing to grow at a rate of 200,000 records each year since 2004. However, obvious non-terrorists are included on the list – a six-year old child and Senator Ted Kennedy. There is no simple or quick redress process for removal from the list. The watch list has drawn criticism because of its potential to promote racial profiling and discrimination.

**Organization:** Integrating 12 different databases into one is a difficult process – 2 databases still require integration. Balancing the knowledge of how someone is added to the list is difficult – information about the process for inclusion must be protected if the list is to be effective against terrorists. On the other hand, for innocent people who are unnecessarily inconvenienced, the inability to ascertain how they came to be on the list is upsetting. Criteria for inclusion on the list may be too minimal. The government agencies lack standard and consistent procedures for nominating individuals to the list, performing modifications to information, and relaying those changes to other governmental offices.
Case Study: The Terrorist Watch List Database’s Troubles Continue

3. Describe some of the weaknesses of the watch list. What management, organization, and technology factors are responsible for these weaknesses?

**Technology**: The poor quality of the database leads to inaccurate data, redundant data, and erroneous entries. The TSA needs to perform an intensive data quality audit and data cleansing to help match imperfect data in airline reservation systems with imperfect data on the watch lists. While government agencies have been able to synchronize their data into a single list, there is still more work to be done to integrate that list with those maintained by airlines, individual states, and other localities using more information to differentiate individuals.
Case Study: The Terrorist Watch List Database’s Troubles Continue

4. How effective is the system of watch lists described in this case study? Explain your answer.

It’s difficult to know for sure how effective the watch list system is because we don’t hear about the people that didn’t pass the checkpoints or the ones who didn’t try because they knew they were on the watch list. However, governmental reports assert that the list contains inaccuracies and that governmental departmental policies for nomination and removal from the lists are not uniform. There has also been public criticism because of the size of the list and that obvious non-terrorists are included on the list. Inconsistent record counts, duplicate records, and records that lacked data fields or had unclear sources for data cast doubt upon the effectiveness of the watch lists. Given the option between a list that tracks every potential terrorist at the cost of unnecessarily tracking some innocent people, and a list that fails to track many terrorists in an effort to avoid tracking innocent people, many choose the list that tracks every terrorist despite the drawbacks.
Case Study: The Terrorist Watch List Database’s Troubles Continue

5. If you were responsible for the management of the TSC watch list database, what steps would you take to correct some of these weaknesses?

Suggestions include performing data quality audits and using data cleansing software to correct many of the imperfections of the data. Information policies and data governance policies need to be developed to standardize the procedures for nomination and removal from the lists. The policies could also address the problem with inconsistent record counts, duplicate records, and records that lacked data fields or had unclear sources for their data. The TSA should develop consistent policies and methods to ensure that the government, not individual airlines, is responsible for matching travelers to watch lists.
Case Study: The Terrorist Watch List Database’s Troubles Continue

6. Do you believe that the watch list represents a significant threat to individuals’ privacy or Constitutional rights? Why or why not?

Most of you will probably answer that the watch list does represent threats to privacy and Constitutional rights. The TSA is developing a system called “Secure Flight” but it has been continually delayed due to privacy concerns regarding the sensitivity and safety of the data it would collect. Similar surveillance programs and watch lists, such as the NSA’s attempts to gather information about suspected terrorists, have drawn criticism for potential privacy violations. The watch list has drawn criticism because of its potential to promote racial profiling and discrimination.
1. What are the problems of managing data resources in a traditional file environment and how are they solved by a database management system?

1.1 List and describe each of the components in the data hierarchy.

The data hierarchy includes bits, bytes, fields, records, files, and databases. Data is organized in a hierarchy that starts with the bit, which is represented by either a 0 (off) or a 1 (on). Bits are grouped to form a byte that represents one character, number, or symbol. Bytes are grouped to form a field, such as a name or date, and related fields are grouped to form a record. Related records are collected to form files, and related files are organized into a database.

1.2 Define and explain the significance of entities, attributes, and key fields.

Entity is a person, place, thing, or event on which information is obtained.

Attribute is a piece of information describing a particular entity.

Key field is a field in a record that uniquely identifies instances of that unique record so that it can be retrieved, updated, or sorted. For example, a person’s name cannot be a key because there can be another person with the same name, whereas a social security number is unique. Also a product name may not be unique but a product number can be designed to be unique.
1.3 List and describe the problems of the traditional file environment.

Problems with the traditional file environment include:
1. data redundancy and confusion,
2. program-data dependence,
3. lack of flexibility,
4. poor security, and
5. lack of data sharing and availability.

**Data redundancy** is the presence of duplicate data in multiple data files. In this situation, confusion results because the data can have different meanings in different files.

**Program-data dependence** is the tight relationship between data stored in files and the specific programs required to update and maintain those files. This dependency is very inefficient, resulting in the need to make changes in many programs when a common piece of data, such as the zip code size, changes.

**Lack of flexibility** refers to the fact that it is very difficult to create new reports from data when needed. Ad-hoc reports are impossible to generate; a new report could require several weeks of work by more than one programmer and the creation of intermediate files to combine data from disparate files.

**Poor security** results from the lack of control over data.

**Data sharing** is virtually impossible because it is distributed in so many different files around the organization.
A database is a collection of data organized to service many applications efficiently by storing and managing data so that they appear to be in one location. It also minimizes redundant data. A database management system (DBMS) is special software that permits an organization to centralize data, manage them efficiently, and provide access to the stored data by application programs.

A DBMS can:
1. reduce the complexity of the information systems environment,
2. reduce data redundancy and inconsistency,
3. eliminate data confusion,
4. create program-data independence,
5. reduce program development and maintenance costs,
6. enhance flexibility,
7. enable the ad hoc retrieval of information,
8. improve access and availability of information, and
9. allow for the centralized management of data, their use, and security.
Review Questions

2. What are the major capabilities of DBMS and why is a relational DBMS so powerful?

2.1 Name and briefly describe the capabilities of a DBMS.

A DBMS includes capabilities and tools for organizing, managing, and accessing the data in the database. The principal capabilities of a DBMS include:
1. data definition language,
2. data dictionary, and
3. data manipulation language.

The **data definition language** specifies the structure and content of the database.
The **data dictionary** is an automated or manual file that stores information about the data in the database, including names, definitions, formats, and descriptions of data elements.
The **data manipulation language**, such as SQL, is a specialized language for accessing and manipulating the data in the database.
2.2 Define a relational DBMS and explain how it organizes data.

The relational database is the primary method for organizing and maintaining data in information systems. It organizes data in two-dimensional tables with rows and columns called relations. Each table contains data about an entity and its attributes. Each row represents a record and each column represents an attribute or field. Each table also contains a key field to uniquely identify each record for retrieval or manipulation.

2.3 List and describe the three operations of a relational DBMS.

In a relational database, three basic operations are used to develop useful sets of data: Select, project, and join.

Select operation creates a subset consisting of all records in the file that meet stated criteria. In other words, select creates a subset of rows that meet certain criteria.

Join operation combines relational tables to provide the user with more information that is available in individual tables.

Project operation creates a subset consisting of columns in a table, permitting the user to create new tables that contain only the information required.
3. What are some important database design principles?

3.1 Define and describe normalization and referential integrity and explain how they contribute to a well-designed relational database.

Normalization is the process of creating small stable data structures from complex groups of data when designing a relational database. Normalization streamlines relational database design by removing redundant data such as repeating data groups. A well-designed relational database will be organized around the information needs of the business and will probably be in some normalized form. A database that is not normalized will have problems with insertion, deletion, and modification.

Referential integrity rules ensure that relationships between coupled tables remain consistent. When one table has a foreign key that points to another table, you may not add a record to the table with the foreign key unless there is a corresponding record in the linked table.
3.2 Define and describe an entity-relationship diagram and explain its role in database design.

Relational databases organize data into two-dimensional tables (called relations) with columns and rows. Each table contains data on an entity and its attributes. An entity-relationship diagram graphically depicts the relationship between entities (tables) in a relational database. A well-designed relational database will not have many-to-many relationships, and all attributes for a specific entity will only apply to that entity. Entity-relationship diagrams help formulate a data model that will serve the business well. The diagrams also help ensure data are accurate, complete, and easy to retrieve.
4. What are the principal tools and technologies for accessing information from databases to improve business performance and decision making?

4.1 Define a data warehouse, explaining how it works and how it benefits organizations.

A data warehouse is a database with archival, querying, and data exploration tools (i.e., statistical tools) and is used for storing historical and current data of potential interest to managers throughout the organization and from external sources (e.g., competitor sales or market share). The data originate in many of the operational areas and are copied into the data warehouse as often as needed. The data in the warehouse are organized according to company-wide standards so that they can be used for management analysis and decision making. Data warehouses support looking at the data of the organization through many views or directions. The data warehouse makes the data available to anyone to access as needed, but it cannot be altered. A data warehouse system also provides a range of ad hoc and standardized query tools, analytical tools, and graphical reporting facilities. The data warehouse system allows managers to look at products by customer, by year, or by salesperson--essentially different slices of the data.
4.2 Define business intelligence and explain how it is related to database technology.

Powerful tools are available to analyze and access information that has been captured and organized in data warehouses and data marts. These tools enable users to analyze the data to see new patterns, relationships, and insights that are useful for guiding decision making. These tools for consolidating, analyzing, and providing access to vast amounts of data to help users make better business decisions are often referred to as business intelligence. Principal tools for business intelligence include software for database query and reporting tools for multidimensional data analysis and data mining.
4.3 Describe the capabilities of online analytical processing (OLAP).

Data warehouses support multidimensional data analysis, also known as online analytical processing (OLAP), which enables users to view the same data in different ways using multiple dimensions. Each aspect of information represents a different dimension.

OLAP represents relationships among data as a multidimensional structure, which can be visualized as cubes of data and cubes within cubes of data, enabling more sophisticated data analysis. OLAP enables users to obtain online answers to ad hoc questions in a fairly rapid amount of time, even when the data are stored in very large databases. Online analytical processing and data mining enable the manipulation and analysis of large volumes of data from many perspectives, for example, sales by item, by department, by store, by region, in order to find patterns in the data. Such patterns are difficult to find with normal database methods, which is why a data warehouse and data mining are usually parts of OLAP.
4.4 Define data mining, describing how it differs from OLAP and the types of information it provides.

Data mining provides insights into corporate data that cannot be obtained with OLAP by finding hidden patterns and relationships in large databases and inferring rules from them to predict future behavior. The patterns and rules are used to guide decision making and forecast the effect of those decisions. The types of information obtained from data mining include associations, sequences, classifications, clusters, and forecasts.
4.5 Explain how text mining and Web mining differ from conventional data mining.

Conventional data mining focuses on data that has been structured in databases and files. Text mining concentrates on finding patterns and trends in unstructured data contained in text files. The data may be in email, memos, call center transcripts, survey responses, legal cases, patent descriptions, and service reports. Text mining tools extract key elements from large unstructured data sets, discover patterns and relationships, and summarize the information.

Web mining helps businesses understand customer behavior, evaluate the effectiveness of a particular Web site, or quantify the success of a marketing campaign. Web mining looks for patterns in data through:
1. Web content mining: extracting knowledge from the content of Web pages
2. Web structure mining: examining data related to the structure of a particular Web site
3. Web usage mining: examining user interaction data recorded by a Web server whenever requests for a Web site’s resources are received
4.6 Describe how users can access information from a company’s internal databases through the Web.

Conventional databases can be linked via middleware to the Web or a Web interface to facilitate user access to an organization’s internal data. Web browser software on a client PC is used to access a corporate Web site over the Internet. The Web browser software requests data from the organization’s database, using HTML commands to communicate with the Web server. Because many back-end databases cannot interpret commands written in HTML, the Web server passes these requests for data to special middleware software that then translates HTML commands into SQL so that they can be processed by the DBMS working with the database. The DBMS receives the SQL requests and provides the required data. The middleware transfers information from the organization’s internal database back to the Web server for delivery in the form of a Web page to the user. The software working between the Web server and the DBMS can be an application server, a custom program, or a series of software scripts.
5. Why are information policy, data administration, and data quality assurance essential for managing the firm’s data resources?

5.1 Describe the roles of information policy and data administration in information management.

An **information policy** specifies the organization’s rules for sharing, disseminating, acquiring, standardizing, classifying, and inventorying information. Information policy lays out specific procedures and accountabilities, identifying which users and organizational units can share information, where information can be distributed, and who is responsible for updating and maintaining the information.

**Data administration** is responsible for the specific policies and procedures through which data can be managed as an organizational resource. These responsibilities include developing information policy, planning for data, overseeing logical database design and data dictionary development, and monitoring how information systems specialists and end-user groups use data.

In large corporations, a formal data administration function is responsible for information policy, as well as for data planning, data dictionary development, and monitoring data usage in the firm.
5.2 Explain why data quality audits and data cleansing are essential.

Data that is inaccurate, incomplete, or inconsistent create serious operational and financial problems for businesses because they may create inaccuracies in product pricing, customer accounts, and inventory data, and lead to inaccurate decisions about the actions that should be taken by the firm. Firms must take special steps to make sure they have a high level of data quality. These include using enterprise-wide data standards, databases designed to minimize inconsistent and redundant data, data quality audits, and data cleansing software.

A data quality audit is a structured survey of the accuracy and level of completeness of the data in an information system. Data quality audits can be performed by surveying entire data files, surveying samples from data files, or surveying end users for their perceptions of data quality.

Data cleansing consists of activities for detecting and correcting data in a database that are incorrect, incomplete, improperly formatted, or redundant. Data cleansing not only corrects data but also enforces consistency among different sets of data that originated in separate information systems.
END
Exam Examples

1) Internet advertising is growing at approximately 10 percent a year.
Answer: TRUE

2) The organization's rules for sharing, disseminating, acquiring, standardizing, classifying, and inventorying information is called a(n)
A) information policy.
B) data definition file.
C) data quality audit.
D) data governance policy.
Answer: A