Database Concept 1

Quitoles, Sheena Patrice B.

BSIT 2A

1. **Define each of the following terms:**
    a) Data - refers to any information that can be stored and processed by a computer.
    b) Field - is a specific piece of data within a larger set of information.
    c) Record - is a collection of related fields that describe a specific entity or object.
    d) File - is a collection of related data that is stored on a computer's hard drive or other storage device.
2. **What is data redundancy, and which characteristics of the file system can lead to it?**

    Data redundancy occurs when the same data is stored in multiple places, which can lead to inconsistencies and errors. Characteristics of a file system that can contribute to data redundancy include the lack of data normalization, the absence of a centralized data storage and management system, and the use of multiple, independent files to store related data.

3. **What is data independence, and why is it lacking in file systems?**

    Data independence refers to the ability to change the underlying structure or organization of data without affecting the way that data is used or accessed. File systems lack data independence because they are designed around a specific file structure, which can make it difficult to make changes to the data without disrupting the way that it is stored and accessed.

4. **What is a DBMS, and what are its functions?**

    A DBMS (Database Management System) is a software system that is used to create, maintain, and manage databases. Its functions include data storage, retrieval, and manipulation, as well as the ability to enforce data integrity and security, and support for data backup and recovery.

5. **What is structural independence, and why is it important?**

    Structural independence refers to the ability to modify the underlying structure of a database without affecting the applications that use the data. It is important because it allows for greater flexibility and adaptability in database design and can help to reduce the costs and risks associated with database maintenance and upgrades.

6. **Explain the difference between data and information.**

   Data refers to raw facts and figures, while information is data that has been processed and organized in a meaningful way to support decision-making and other activities.

7. **What is the role of a DBMS, and what are its advantages? What are its disadvantages?**

   The role of a DBMS is to provide a centralized, secure, and efficient way to store, manage, and retrieve data. Its advantages include improved data organization and accessibility, enhanced data security and integrity, and support for data sharing and collaboration. Its disadvantages include high implementation and maintenance costs, as well as the potential for performance and scalability issues.

8. **List and describe the different types of databases.**

   The different types of databases include relational databases, NoSQL databases, object-oriented databases, and others. Relational databases are the most common type, and they use tables to organize data into rows and columns.

9. **What are the main components of a database system?**

   The main components of a database system include the database itself, the DBMS software, the users who access and interact with the data, and the hardware and network infrastructure that supports the system.

10. **What are metadata?**

    Metadata is data that describes other data, such as information about the structure, organization, and properties of a database.

11. **Explain why database design is important.**

    Database design is important because it determines the overall structure and organization of the data, which can have a significant impact on its usability, accessibility, and accuracy.

12. **What are the potential costs of implementing a database system?**

    The potential costs of implementing a database system include the cost of hardware and software, as well as the cost of designing and implementing the system, training users, and maintaining and upgrading the system over time.

13. **Use examples to compare and contrast unstructured and structured data. Which type is more prevalent in a typical business environment?**

Unstructured data refers to information that does not have a defined data model or format. This can include things like emails, social media posts, images, audio files, and video files. Unstructured data is typically more difficult to analyze because it does not have a consistent format or structure.

Structured data, on the other hand, has a defined data model and format. Examples of structured data include sales data, inventory data, and customer information. Structured data is easier to analyze and can be easily organized and queried.

In a typical business environment, structured data is more prevalent because it is easier to manage and analyze. However, unstructured data is becoming increasingly important as businesses collect more data from sources like social media and customer feedback.

For example, a company may have structured data in a database containing customer information such as name, address, and purchase history. However, they may also have unstructured data in the form of customer feedback on social media, which can be more difficult to analyze and organize.

14. What are some basic database functions that a spreadsheet cannot perform?

Basic database functions that a spreadsheet cannot perform include data normalization, enforcing referential integrity, handling concurrent data access, and providing a centralized and secure data storage and management system.

15. What common problems does a collection of spreadsheets created by end users share with the typical file system?

A collection of spreadsheets created by end users shares common problems with the typical file system, including the lack of data normalization, the potential for data redundancy and inconsistency, and the difficulty of managing and analyzing data across multiple files.

16. Explain the significance of the loss of direct, hands-on access to business data that end users experienced with the advent of computerized data repositories.

The advent of computerized data repositories, such as databases, brought about a significant shift in the way that end users access business data. Prior to the use of computerized data repositories, end users typically had direct, hands-on access to the data they needed. They could physically access files and documents, and they could easily manipulate data using tools like spreadsheets.

With the introduction of computerized data repositories, the ability of end users to directly access and manipulate data was greatly reduced. Instead, access to data was controlled by the database management system (DBMS), which required end users to interact with data through specific interfaces and query

languages. This change meant that end users had to rely on the DBMS to provide them with the data they needed, rather than being able to directly access and manipulate data as they had in the past.

The significance of this change is that it had both advantages and disadvantages. On the one hand, the use of computerized data repositories made it possible to store and manage vast amounts of data in a more efficient and organized way. It also provided greater control over who could access and modify data, which helped to improve security and data integrity.

On the other hand, the loss of direct, hands-on access to data meant that end users had to learn new tools and techniques for working with data. It also created new dependencies on the DBMS and other IT professionals who were responsible for managing the database environment. Finally, it made it more difficult for end users to manipulate data quickly and easily, since they had to rely on more complex and structured query languages and interfaces.

FIGURE P1.1 The file structure for Problems 1–4

PROJECT_CODE	PROJECT_MANAGER	MANAGER_PHONE	MANAGER_ADDRESS	PROJECT_BID_PRICE
21-5Z	Holly B. Parker	904-338-3416	3334 Lee Rd., Gainesville, FL 37123	16833460.00
25-2D	Jane D. Grant	615-898-9909	216 Clark Blvd., Nashville, TN 36362	12500000.00
25-5A	George F. Dorts	615-227-1245	124 River Dr., Franklin, TN 29165	32512420.00
25-9T	Holly B. Parker	904-338-3416	3334 Lee Rd., Gainesville, FL 37123	21563234.00
27-4Q	George F. Dorts	615-227-1245	124 River Dr., Franklin, TN 29165	10314545.00
29-2D	Holly B. Parker	904-338-3416	3334 Lee Rd., Gainesville, FL 37123	25559999.00
31-7P	William K. Moor	904-445-2719	216 Morton Rd., Stetson, FL 30155	56850000.00

1. **How many records does the file contain? How many fields are there per record?**

   The file contains seven records (21-5Z through 31-7P) and each of the records is composed of five fields (PROJECT_CODE through PROJECT_BID_PRICE).

2. **What problem would you encounter if you wanted to produce a listing by city? How would you solve this problem by altering the file structure?**

   The city names are contained within the MANAGER_ADDRESS attribute and decomposing this character (string) field at the application level is cumbersome at best. (Queries become much more difficult to write and take longer to execute when internal string searches must be conducted.) If the ability to produce city listings is important, it is best to store the city name as a separate attribute.

3. **If you wanted to produce a listing of the file contents by last name, area code, city, state, or zip code, how would you alter the file structure?**

If we wanted a listing by LAST_NAME, AREA_CODE, CITY, STATE or ZIP_CODE then we can add those fields to the file structure then loop through the records and do comparison using field name = value and list those records that match. The file structure needs to be changed to the following:

PROJECT_ CODE	PROJECT_ MANAGER	MANAGER_ PHONE	MANAGER_ ADDRESS	PROJECT_ BID_PRICE	LAST_NAME	AREA_CODE	CITY	STATE	ZIP

4. **What data redundancies do you detect? How could those redundancies lead to anomalies?**

Note that the manager named Holly B. Parker occurs three times, indicating that she manages three projects coded 21-5Z, 25-9T, and 29-2D, respectively. (The occurrences indicate that there is a 1:M relationship between PROJECT and MANAGER: each project is managed by only one manager but, apparently, a manager may manage more than one project.) Ms. Parker's phone number and address also occur three times. If Ms. Parker moves and/or changes her phone number, these changes must be made more than once, and they must all be made correctly... without missing a single occurrence. If any occurrence is missed during the change, the data are "different" for the same person. After some time, it may become difficult to determine what the correct data is. In addition, multiple occurrences invite misspellings and digit transpositions, thus producing the same anomalies. The same problems exist for the multiple occurrences of George F. Dorts.

FIGURE P1.5 The file structure for Problems 5–8

PROJ_NUM	PROJ_NAME	EMP_NUM	EMP_NAME	JOB_CODE	JOB_CHG_HOUR	PROJ_HOURS	EMP_PHONE
1	Hurricane	101	John D. Newson	EE	85.00	13.3	653-234-3245
1	Hurricane	105	David F. Schwann	CT	60.00	16.2	653-234-1123
1	Hurricane	110	Anne R. Ramoras	CT	60.00	14.3	615-233-5568
2	Coast	101	John D. Newson	EE	85.00	19.8	653-234-3254
2	Coast	108	June H. Settlemeir	EE	85.00	17.5	905-554-7812
3	Satellite	110	Anne R. Ramoras	CT	62.00	11.6	615-233-5568
3	Satellite	105	David F. Schwann	CT	26.00	23.4	653-234-1123
3	Satellite	123	Mary D. Chen	EE	85.00	19.1	615-233-5432
3	Satellite	112	Allecia R. Smith	BE	85.00	20.7	615-678-6879

5. **Identify and discuss the serious data redundancy problems exhibited by the file structure shown in Figure P1.5.**

Given the file's poor structure, the stage is set for multiple anomalies. For example, if the charge for JOB_CODE = EE changes from $85.00 to $90.00, that change must be made twice. Also, if employee June H. Sattlemeier is deleted from the file, you also lose information about the existence of her JOB_CODE = EE, its hourly charge of $85.00, and the PROJ_HOURS = 17.5. The loss of the PROJ_HOURS value will ultimately mean that the Coast project costs are not being charged properly, thus causing a loss of PROJ_HOURS*JOB_CHG_HOUR = 17.5 x $85.00 = $1,487.50 to the company.

Incidentally, note that the file contains different JOB_CHG_HOUR values for the same CT job code, thus illustrating the effect of changes in the hourly charge rate over time. The file structure appears to represent transactions that charge project hours to each project. However, the structure of this file makes it difficult to avoid update anomalies and it is not possible to determine whether a charge change is accurately reflected in each record. Ideally, a change in the hourly charge rate would be made in only one place and this change would then be passed on to the transaction based on the hourly charge. Such a structural change would ensure the historical accuracy of the transactions.

You might want to emphasize that the recommended changes require a lot of work in a file system.

6. **Looking at the EMP_NAME and EMP_PHONE contents in Figure P1.5, what change(s) would you recommend?**

A good recommendation would be to make the data more atomic. That is, break up the data components whenever possible. For example, separate the EMP_NAME into its components EMP_FNAME, EMP_INITIAL, and EMP_LNAME. This change will make it much easier to organize employee data through the employee's name component. Similarly, the EMP_PHONE data should be decomposed into EMP_AREACODE and EMP_PHONE. For example, breaking up the phone number 653-234-3245 into the area code 653 and the phone number 234-3245 will make it much easier to organize the phone numbers by area code. (If you want to print an employee phone directory, the more atomic employee name data will make the job much easier.

7. **Identify the various data sources in the file you examined in Problem 5.**

JOB_CODE, PROJ_ HOURS, JOB_CHG_HOUR

8. **Given your answer to Problem 7, what new files should you create to help eliminate the data redundancies found in the file shown in Figure P1.5?**

The data sources are probably the PROJECT, EMPLOYEE, JOB, and CHARGE. The PROJECT file should contain project characteristics such as the project name, the project manager/coordinator, the project budget, and so on. The EMPLOYEE file might contain the employee names, phone number, address, and so on. The JOB file would contain the billing charge per hour for each of the job types – a database designer, an applications developer, and an accountant would generate different billing charges per hour. The CHARGE file would be used to keep track of the number of hours by job type that will be billed for each employee who worked on the project.

FIGURE P1.9 The file structure for Problems 9–10

BUILDING_CODE	ROOM_CODE	TEACHER_LNAME	TEACHER_FNAME	TEACHER_INITIAL	DAYS_TIME
KOM	204E	Williston	Horace	G	MWF 8:00-8:50
KOM	123	Cordoza	Maria	L	MWF 8:00-8:50
LDB	504	Patroski	Donald	J	TTh 1:00-2:15
KOM	34	Hawkins	Anne	W	MWF 10:00-10:50
JKP	225B	Risell	James		TTh 9:00-10:15
LDB	301	Robertson	Jeanette	P	TTh 9:00-10:15
KOM	204E	Cordoza	Maria	I	MWF 9:00-9:50
LDB	504	Williston	Horace	G	TTh 1:00-2:15
KOM	34	Cordoza	Maria	L	MWF 11:00-11:50
LDB	504	Patroski	Donald	J	MWF 2:00-2:50

9. Identify and discuss the serious data redundancy problems exhibited by the file structure shown in Figure P1.9.(The file is meant to be used as a teacher class assignment schedule. One of the many problems with data redundancy is the likely occurrence of data inconsistencies—two different initials have been entered for the teacher named Maria Cordoza.)

The data redundancy problems exhibited by the file are:

- The **TEACHER_INITIAL** column holds two different initials for the teacher, **Maria Cordoza**, that are **'I'** and **'L'**.

  This would lead to inconsistency and inaccuracy of data regarding the identification of the teacher **Maria Cordoza**. Both the initials would retrieve only partial and incomplete details of her classes.

- The **DAYS_TIME** column holds two classes, one with **Horace Willingston** and the other with **Maria Cordoza**, on the same days and timings **"MWF 8:00-8:50"**.

  This would lead inconsistency of data regarding the lecture on **"MWF 8:00-8:50".**

10. Given the file structure shown in Figure P1.9, what problem(s) might you encounter if building KOM were deleted?

Deletion of building **KOM** would lead to the below mentioned consequences:

- Details of the teacher **Maria Cordoza** would be deleted from the database.
- Details of classes on **"MWF"** would be deleted from the database.

This would lead to a lack of data integrity in the database as the data would no longer be accurate and consistent with the real-life scenario.