Lab 1. Database Environment

Noah Zhou

Sober is a startup company that is thinking about storing its data in Word documents, Excel files, and maybe some other files.

As a database designer at Sober, you are suggesting that Sober instead use a database management system (DBMS) to store their data. Advocate to Sober's top management whether to use a DBMS through the following questions:

Question 1A: What are the key differences between the file-based and database approaches to data management?

Through a file-based approach to data management, every independent application stores data into individual files. The data stored within these files are typically similar or even the same. This results in redundant data being stored across multiple files and used across multiple applications. There are multiple issues with this approach such as wasted storage, data inconsistency across applications, and a strong dependency between the application and the data file. A strong dependency can result in maintenance issues where a structural change in the data file requires changes in all the applications that rely on the data file. Another issue is concurrency control or simply put, simultaneous access to data without conflict. For example, if there are two completely independent applications using and updating the same file, there could be issues with inconsistent data if there are no established concurrency controls.

On the other hand, a database approach to data management uses a central database management system (DBMS). The purpose of the DBMS is to act as sort of an intermediary where the application now interfaces and requests data from the DBMS instead of directly sourcing data from a specific file. The DBMS typically stores two types of data, raw data and

metadata. Now instead of the applications storing the data definitions, the DBMS stores the definitions inside its catalog as metadata.

Question 1B:For each of the differences, come up with an example for Sober that would explain how file-based and database approaches could be used in Sober's case.

Sober could have separate Excel spreadsheets for customer information, customer orders, and customer invoices. The information spreadsheet contains information such as name, shipping address. The order and invoice spreadsheets both use the customer spreadsheet information spreadsheet to locate invoices and orders separately. There is a chance that if a customer has multiple orders, the invoice sheet might not update the running total correctly. In addition, if Sober wanted to add a new column to the customer information spreadsheet to see whether a customer is a new or returning customer in order to give discounts, this would have to be updated in both of the customer information spreadsheets that the order and invoice spreadsheets use On the other hand, through a database approach, the DBMS manages the data through the implementation of a customer information table that contains a unique customer identifier. In this case, both the order and invoice spreadsheets can be transferred to separate tables that refer to the unique customer identifier. This way, information can easily be accessed with a simple query instead of using a complex query to find information within three spreadsheets.

Question 2A: What are the advantages of using databases and database management systems? List all of them (mentioned in Chapter 1 of your textbook).

There are many advantages to using database and database management systems such as data independence, managing structured, semi-structured, and unstructured data, database

modeling, managing data redundancy, specifying integrity rules, specifying concurrency controls, data backup and recovery, data security, and data performance.

Data independence is the idea that changes in data definitions have little to no impact on the applications using that data.

Database modeling is the physical representation of data items together with their characteristics and relationships. Modeling can also include integrity rules and functions. Structured data is data where individual characteristics can be formally identified and specified such as name or address. Semi-structured data has a certain structure but is unstable and can easily change, such as people's profiles on a social media platform. Unstructured data is where items cannot be interpreted in a meaningful way even though individual items can be found within that data.

Managing data redundancy is where the DBMS is responsible for management of data redundancy through synchronization. If a local copy of data is updated, that change in data will be sent to all other copies in other locations.

Integrity rules can be used to enforce the correctness of the data. For example, a unique number that identifies a customer must be unique and must be an integer or names cannot contain any numbers.

Concurrency control is where typical read/write operations can be executed at the same time. Most systems must support the Atomicity, Consistency, Isolation, and Durability (ACID) properties.

Data backup and recovery is typically used when there is a loss of data due to software or hardware instability or even network instability.

Data security is where users have specified access such as being able to read and write or just read. Access is typically managed through assigned logins and passwords.

DBMS's also have performance utilities which address the three main key performance indicators which are response time, throughput rate, and space utilization. Utilities can include optimizing storage space, index tuning for faster query execution, or optimizing buffer management.

Question 2B: Outline whether each of the listed advantage may or may not be applicable to Sober.

Data independence most likely will be applicable due to customer's information constantly changing. Database modeling is definitely applicable to Sober, especially when deciding whether to continue with the file-based approach or use a database approach. Managing data redundancy goes hand in hand with data independence where changes to the information must be seen in other locations that use the same data. Integrity rules are also applicable where a customer's birthdate must be entered and must follow a specific format. Concurrency controls might not be used within Sober. Data backup and recovery is definitely applicable because losing a customer's information is unwanted. Data security is definitely important where customers can access and change information with their own personal logins which blocks malicious actors from accessing such information. Performance utilities might not be used in the beginning but might be required once Sober grows bigger in size.