**Exercise 06:**

**1. Describe a data set and analysis scenario where you would choose to perform your analysis using only UNIX command line tools instead of a database, and explain why the command line is a good fit for the scenario.**

UNIX command line tools are extremely handier than databases when dealing with datasets of small size in simple csv and text, especially, or doing some initial exploration of data, like the “smell tests” before wrangling. Because of the UNIX has pipe lines and filters and pipelines can combine filters in many languages, it’s simpler to develop, maintain, and document files as well as to understand, remember and reuse. Also it can be combined with parallel to fit into repetitive workflows.

**2. Describe a data set and analysis scenario where you would choose to perform your analysis using a relational database instead of UNIX command line tools, and explain why the database is a good fit for the scenario.**

Relational database performs better than UNIX command line tools when working with large and complicated CSV files, like complex enterprises with inventory, business transactions, or when the work involves many simultaneous users and uses large volumes of data and constant changes, because of its integrity, atomicity, consistency And it supports concurrency of multiple users with differential access controls.

**3. Describe examples of Type 1 and Type 2 slow changes in a data warehouse.**

Type 1 slow changes in a data warehouse is overwriting the old value. This type is easy to maintain and is often use for data which changes are caused by processing corrections (e.g. removal special characters, correcting spelling errors).

Type 2 slow changes refers to creating a new additional record. It allows users to track historically significant attributes. The old records point to all history prior to the latest change, and the new record maintains the most current information. Like for 'effective date' and 'current indicator' columns are used in this method.

**4. Describe how SQL queries are processed within an RDBMS.**

SQL processing is the parsing, optimization, row source generation, execution of a SQL statement, and evaluation. Depending on the statement, the database may omit some of these stages. For parsing, it means separating the pieces of a SQL statement into a data structure that other routines can process. Optimization means generation of multiple execution plans. Row source generation means generation of query plan and it will produce a row source tree. Finally, during execution, the SQL engine executes each row source in the tree then query-execution engine takes a query-evaluation plan, executes that plan, and returns the answers to the query.

**5. How do dimensions and fact tables in a dimensional model relate to aspects of a business process?**

Since fact table holds the data to be analyzed, which are the facts that happens during the business process, and a dimension table stores data about the ways (dimensions) in which the data in the fact table can be classified and analyzed, like customer, time, products... Thus, the fact table consists of two types of columns. The foreign keys column allows joins with dimension tables, and the measures columns contain the data that is being analyzed.

**6. In general, is it usually more efficient to join relational tables first, and then filter the results, or to filter on individual relational tables first, and then join the results? Why?**

Filtering on individual relation tables first then join the results would be more efficient cause less duplicated data need to be processed after filtering which would speed the whole process up.

**7. Most relational database engines use row-oriented data storage techniques. Why might you choose to use column-oriented data storage instead?**

Since analytical workloads scan and write large amounts of data in a table. But the slowest thing to do in a database is read and write from disk. Also analytical workloads tend to have a lot of repeating values for each row (dimensions such as department name or product). So it’s would be better to choose to use column-oriented database when reading large amounts of repeating values.

**8. What happens during the ETL process?**

Since ETL is a process in data warehousing responsible for pulling data out of the source systems and placing it into a data warehouse. ETL involves the following tasks: extracting the data, transforming the data and loading the data.

**9. What is a “behavioral dimension”? Provide an example.**

A behavioral question is one that groups or filters facts based on the past behavior of members of a dimension. A very powerful analytic technique uses past behavior patterns to make sense of current. If we want to know if the customers who generate over $1 million in orders are receiving better discounts than those who generate $500,000 or less. It uses a fact, order dollars, as a dimension, providing a context for the study of discount dollars.

**10. What is required to have “tidy data”, according to Wickham?**

According to Wickham, Tidy data is a standard way of mapping the meaning of a dataset to its structure. A dataset is messy or tidy depending on how rows, columns and tables are matched up with observations, variables and types. In tidy data:

1. Each variable forms a column.

2. Each observation forms a row.

3. Each type of observational unit forms a table.

**11. Why are relational databases so successful? Provide 2-3 reasons.**

It offers better security by splitting data into tables, and certain tables can be made confidential. When a person logs on with their username and password, the system can then limit access only to those tables whose records they are authorized to view.

It caters for future requirements. By having data held in separate tables, it is simple to add records that are not yet needed but may be in the future.

**12. Why is it important to make research and data analyses reproducible?**

Reproducibility means ability to produce the same outputs from same inputs using same methods. It’s important as this helps to show that the experiment is valid and correct, because a number of different people did it and got the same results, and the results weren’t simply because of some mistake by one person.