IST722: Class Exercise 7

This is an individual assignment.

Before you begin, please make sure you've read and understand 1) our class honor code, 2) course policies on late work and 3) participation policies as posted on the syllabus. "I didn't know" is not an excuse.

You should cite your sources in a standard format like MPA or APA and include a list of works cited.

Your Name:	Bhavya Shah
Your Email:	bhshah@syr.edu

Instructions (Refer Units 6 & 7)

Answer each of the following questions as concisely as possible. More is not necessarily better. Please justify your answer by citing your sources from the assigned readings from our textbooks, our class lectures, or online if directed to do so. Be sure to cite in text and include a list of works cited. Place your answer below each question. When you're finished, print out this document and bring it to class as part of your participation grade.

Questions

[1] How can you identify changes to a business entity when there is no natural key?

Ans - When there is no natural key in a data warehouse, changes to a business entity can be identified using a combination of surrogate keys and effective date/timestamps. Surrogate keys are unique identifiers generated by the system for each record. By comparing these surrogate keys and timestamps, you can track changes and determine the most recent version of the entity.

[2] What are CET and LSET? How are they used in data warehousing?

Ans - CET, known as the Current Extraction Timestamp, represents the current timestamp of the data warehouse. LSET, standing for the Last Successful Extraction Timestamp, indicates the time of the last successful data extraction. These timestamps play a crucial role in incremental data extraction, focusing solely on the modified data. By comparing the timestamp of a newly added record with LSET, data extraction occurs if it is greater. Similarly, extraction occurs when the last updated timestamp surpasses LSET.

[3] Briefly explain late-arriving dimensions and late-arriving facts, and how you would manage them.

Ans - Late-arriving dimensions occur when a dimension record is not available when the associated fact record is loaded. This can happen for a variety of reasons, such as a delay in the data source, a problem with the ETL process, or a change to the dimension schema.

Late-arriving facts occur when a fact record is loaded before the associated dimension records. This can happen for the same reasons as late-arriving dimensions, or it can happen if the fact data is received in real time and the dimension data is not.

One common approach to manage them is to create placeholder dimension records for late-arriving dimensions. These placeholder records contain the natural key of the dimension record, but they have null values for all other attributes.

[4] Explain what is meant by the surrogate key pipeline in your own words. Keep this brief.

Ans - The surrogate key pipeline is a process in data warehouse ETL that replaces the natural keys in the dimension tables with surrogate keys. Surrogate keys are unique integers that are assigned to each record in a dimension table. They are used to ensure the referential integrity of the data warehouse and to make it easier to join the fact and dimension tables. The surrogate key pipeline typically consists of the following steps:

1. Extract the natural keys.
2. Query a lookup table to find the corresponding surrogate key.
3. Replace the natural keys with the surrogate keys.
4. Load the dimension tables into the data warehouse.

[5] What is the purpose of the lookup transformation? How many attributes must match for the lookup to succeed?

Ans - The primary objective of the lookup transformation is to update the fact table by replacing its natural keys with corresponding surrogate keys from dimension tables. By using surrogate keys, the fact table can maintain accurate references to dimension data despite any changes that might occur.
For the lookup to succeed, all the natural keys from the dimension tables in the fact table must match with the surrogate keys. This ensures proper data integration and consistency between the fact and dimension tables in the data warehouse.

WORKS CITED:

Class slides and Professor Fudge’s videos.