

IST722: Unit 06 Participation Questions

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:	Akshay Bhala
Your Email:	abhala@syr.edu

Instructions

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] What is the difference between ETL and ELT?

ETL stands for Extract, Transform and Load. It is the process of retrieving data from OLTP sources, transforming it and placing it into the data warehouse. It is moving the data around the data warehouse and takes 70% of the data warehouse effort. ELT stands for Extract, Load and Transform and the only difference is the location where the transformational logic takes place (target data source). It is used mainly in MPP data warehouse as MPP can handle high workloads effectively.

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

When there is no natural key, you can identify changes to a business entity using CRC/Hash which allows to perform compare for assigning a cryptographic hash to data and comparing it to incoming data to detect changes.

[3] Explain snapshotting and its purpose in data warehousing.

A snapshot in a data warehouse can be used to track activities. For example, every time an employee attempts to change his address, the data warehouse can be alerted for a snapshot. This means that each snap shot is taken when some event is fired. A snapshot has three components – Time when event occurred, A key to identify the snap shot, Data that relates to the key.

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

CET is Current Extraction Timestamp which tells you the current timestamp of the data warehouse, LSET is Last Successful Extraction Timestamp which tells when the last data extractions successfully happened. These are used for incremental data extraction – i.e. when you want to get only that data that has been modified or added to the OLTPs. So basically, you would check the timestamp when a record was added to the OLTP, if this is greater than LSET, that means a new record has been added, so you would extract that data; or when Last updated timestamp is greater than LSET.

[5] What are 3 ways for matching data that are character strings?

The three ways if matching data that are character strings are:

1. **Exact match**- all the characters in the string must be exactly the same. Example: Date and Time match
2. **Fuzzy logic**- Generates similarity score and cutoff to determine if there is a match in strings (example- 0.5 for 50% match). Example: misspell words
3. **Rule-based logic**- It decides based on the context. Example- book=novel

WORKS CITED:

Professor Humayun Explanation in lecture

Thursday Lecture discussions

<https://www.careerride.com/Data-warehousing-what-is-snapshot.aspx#:~:text=from%20the%20catalogue.,What%20is%20snapshot%20with%20reference%20to%20data%20warehouse%3F,when%20some%20event%20is%20fired.>