IST722: Unit 07 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] A fellow student reported:

> *I am stuck in my ETL. After troubleshooting I found that it's the DayofYear column not matching in type. Here's the specific error message: "Cannot map columns of different types. Column 'DayOfYear' is of type 'System.Int32' and column 'YearDay' is of type 'System.Int16'." I need help.*

What would be your advice to your fellow student?

Ans. I would suggest to typecast the datatype Dayofyear from int to tinyint and would make the change in the stage or warehouse stage as schema is flexible.

[2] When you find a data type or length mismatch, where should you fix the issue as best practice: source, stage, dw?

Ans. If we find a data type or length mismatch the best practice is to fix the issue in the stage and DW .We cannot make change in source and thus it is untouched.

[3] What is Upsert? When is Loading via Upsert used?

Ans. Upsert essentially means - To insert rows into a database table if they do not already exist or update them if they do. SCD type 1 and SCD type 2 use the Upsert functionality when data is needed to be loaded into the fact tables. Once the data is loaded into the fact tables, if there is a match in the data, the data is updated. If there is no match, a new row is inserted while also retaining the previous row.

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

Ans. Data warehouses commonly use a surrogate key to uniquely identify an entity. A surrogate is not generated by the user but by the system. A primary difference between a primary key and surrogate key in few databases is that PK uniquely identifies a record while a SK uniquely identifies an entity.

e.g.An employee may be recruited before the year 2000 while another employee with the same name may be recruited after the year 2000. Here, the primary key will uniquely identify the record while the surrogate key will be generated by the system (say a serial number) since the SK is NOT derived from the data.

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

Ans. The Lookup Transformation in Informatica is very useful to look up data present in Flat Files, Relational tables, and Views. Lookup is very useful transformation SSIS component it performs lookup operation by connecting input value with data-table or table dataset columns. It compares source data with existing table dataset and filters matching ones and un-matching ones.

The primary purpose of Lookup transformation is to replace natural keys with surrogate keys in a fact table to keep the table update. These natural keys present in the fact table must match with the surrogate keys in order for the lookup transformation to be successful.

WORKS CITED:

Thursday Lecture discussions

Professor Humayun Explainations

Professor fudge videos

https://www.careerride.com/Data-warehousing-surrogate-key.aspx