IST722: Class Exercise 8

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 Unit 8)

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 Data Quality? List three types of DQ rules with examples.

Ans – Data quality refers to the activities which ensure that the data in the data warehouse in complete and true, as the data warehouse will not be efficient or of any use if the data in it is not correct.

Three types of DQ are:
1. Incoming Data Validation - Rules for the data entered are checked as the staged data enters the DW. Example, marks of a student cannot be more than 100.
2. Cross Reference Validation – This rules to check the incoming data against the data already in the DW. Example, the number of people created account on a website compared to the running average of the last 2 weeks.
3. Data Warehouse Internal Validation – This checks the data warehouse data against itself, usually for aggregates. Example, the number of orders placed yearly matched the actual number of orders placed by year.

[2] What is the Data Cleaning requirement? Can you perform Analytics with Dirty Data?

Ans – Data cleaning is a process of identifying and correcting bad data. Data cleaning requirements can be:
a. Replacing nulls with defaults
b. Fixing case: MIKE to Mike
c. Formatting Data: 3154432911 to 315-443-2911
d. Regular Expressions: matching emails, ip addresses
e. Lookups on a business key
f. Fuzzy Matching: Do not to Don't
g. Rule-Based: [Bill, Will, William, Billy] to Bill

Analytics can be performed with bad or incorrect data, but there are major dangers and restrictions. Analyzing incorrect

data may result in inaccurate conclusions, wrong insights, and false forecasts. Errors, discrepancies, and missing values in data might jeopardize the accuracy and trustworthiness of the results.

[3] What is Master Data? How is this different from DW Dimensions?

Ans - Master Data serves as the primary entity, housing a centralized repository of the organization's crucial information. It encompasses a comprehensive collection of data from various entities, distinct from dimension data, which specifically focuses on data relevant to business processes. For instance, the employee entity is classified as master data, whereas the data concerning available employees in a particular process is treated as dimension data.

[4] How can a company benefit from Master Data Management? Does every company need an MDM strategy?

Ans - MDM initiatives will help us create conformed dimensions in our data warehouse, specifically when the dimensional data is sourced from multiple systems. Master Data Management (MDM) is a transformative solution that revolutionizes data practices for companies. By establishing data excellence, MDM optimizes processes, unifies business entities, and enables data-driven decision-making. Not every company needs an MDM strategy if the data is accurate, and the company individual takes their responsibilities seriously.

[5] Describe the composition and function of a data governance board.

Ans – Data governance refers to the complete management of an organization’s data. Data governance comprises of a Governing Body, a set of Defined Procedures and an Execution Plan to execute those procedures. The function of the data governance board is to:

- Devise a plan to share data to improve decision making or improve customer relations.
- Protect the data and maintain regulatory compliance.
- Maximize actionable insights from the data you have.
- Ensure data quality.

WORKS CITED:

Class slides and Professor Fudge slides.