Handling Missing Data in ETL - Theoretical Answers

Q1. What are the most common reasons for missing data in ETL pipelines?

Missing data in ETL pipelines occurs due to several reasons:

1. Data Entry Errors – Users may skip optional fields or enter incorrect values.
2. System Integration Issues – Different source systems may not contain all required fields.
3. Data Transmission Failures – Network/API failures may cause partial data loads.
4. Schema Changes – Newly added columns may not have historical data.
5. Data Corruption – File damage or format mismatches may result in NULL values.
6. Business Logic Rules – Some fields are conditionally applicable.
7. Improper Data Transformation – Incorrect joins or filters may introduce NULL values.

Q2. Why is blindly deleting rows with missing values considered a bad practice in ETL?

Blindly deleting rows with missing values can lead to loss of valuable information, biased analysis, and inaccurate insights. If a large portion of data is removed, the dataset may no longer represent the real population. Instead of deleting, ETL developers should analyze the cause and choose an appropriate handling method.

Q3. Difference between Listwise Deletion and Column Deletion

Listwise Deletion: Removes entire rows where at least one value is missing. Appropriate when missing data is very small and random.

Column Deletion: Removes an entire column if most of its values are missing. Appropriate when a column has very high missing percentage and is not critical for analysis.

Q4. Why is median imputation preferred over mean imputation for skewed data such as income?

Median imputation is preferred for skewed data because the median is not affected by extreme outliers. Income data is usually right-skewed, where a few high-income values can distort the mean. Using the median provides a more accurate central value.

Q5. What is forward fill and in what type of dataset is it most useful?

Forward fill is a method where missing values are replaced with the previous valid value. It is most useful in time-series or sequential datasets such as sales data, stock prices, or sensor readings where previous values logically continue forward.

Q6. Why should flagging missing values be done before imputation in an ETL workflow?

Flagging missing values before imputation preserves information about which records originally had missing data. This helps in analysis, model training, and identifying patterns in missingness, which may provide important business insights.

Q7. How can missing income data provide business insights?

If income is missing for many customers, it may indicate privacy concerns, incomplete data collection processes, or specific customer segments (such as students or unemployed individuals). Analyzing this missingness can help improve data collection strategies and target marketing efforts more effectively.

(8, 9 , 10 QUESTIONS ARE THERE IN THE LINK BELOW 👇👇)

EXCEL