# Task 1

## (1)

### (a)

Two new data quality problems that have emerged since the publication of the paper are:

1. Handling Unstructured and Semi-Structured Data: With the rise of big data, unstructured (e.g., text, images, videos) and semi-structured data (e.g., JSON, XML) have become more popular. Cleaning and standardizing these data types pose significant challenges, as traditional methods designed for structured data often fail to address issues like inconsistent formatting, embedded metadata, and contextual relevance.

2. Data Privacy and Compliance: As regulations like GDPR(General Data Protection Regulation) have been published, ensuring data quality now includes maintaining privacy and compliance. This introduces challenges in anonymizing personal data while preserving its utility. Inaccuracies or over-cleaning could lead to the loss of critical information, making it harder to ensure compliance without compromising data quality.

### (b)

Yes, many problems raised in the paper are still relevant today.

Those issues stems from the enduring nature of data quality challenges. Despite technological advancements, the **core problems** of data inconsistencies, duplicates, missing values, and integration across heterogeneous sources remain fundamental to ensuring data integrity. These issues are intrinsic to any data management process, whether dealing with traditional databases or modern big data environments.

The **ETL process**, central to the paper, is still crucial for transforming and loading data into warehouses or other storage systems. Even with the advent of real-time processing and advanced tools, ETL remains a vital step where data quality issues must be addressed.

Moreover, the **limitations of tools** and the need for **manual intervention**, as highlighted in the paper, continue to be a challenge. While automation has improved, complex data scenarios often require human expertise, especially when dealing with domain-specific data or ambiguous cases.

Additionally, the expansion of **data types** (e.g., unstructured and semi-structured data) and the growth in **data volume** have magnified these problems, making the paper’s insights even more pertinent today. Thus, the foundational issues and solutions discussed remain relevant as they underpin the challenges faced in modern data environments.