UNIVERSITY PARTNER

UNIVERSITY OF WOLVERHAMPTON

HERALD COLLEGE KATHMANDU

Big Data (6CS030)

Internal Assignment - 1

Student Name	: Nayan Raj Khanal
Student Number	: 2227486
Tutor	: Mr. Sujan Shrestha
Word Count	: 768
Submission Date	: 7th March, 2024

1. Introduction

Data analysis is a crucial task for today's large corporations to run their businesses and to perform that task they are building data warehouses. But only recently that a problem has surfaced i.e., the data they are feeding their warehouses with are often 'dirty'. The paper "A Taxonomy of Dirty Data" broadly defines dirty data as missing data, wrong data, and non-standard representations of the same data (Kim, et al., 2003). This report aims to summarize the key points regarding the data quality problems categorized in the two papers and to identify the top three data quality issues.

2. Literature Review

In the paper "Data Cleaning: Problems and Current Approaches" data cleaning or scrubbing is defined as identifying and eliminating errors and inconsistencies in data to enhance its quality (Rahm & Do, 2000). Misspellings, missing information, and invalid data entries are common data quality problems that are in turn common in data warehouses, federated database systems, or global web-based information systems. Thus, the data-cleaning process is crucial due to the presence of redundant data in various formats to ensure accurate data for decision-making processes. Before the data is sent to the data warehouses it goes through the ETL (extraction, transformation, loading) process but due to the complexity and volume of the data manual intervention and low-level programming are often necessary thus calling for an effective data cleaning approach which is supported by tools to minimize manual efforts. The paper divides data-cleaning problems into two parts; Single-Source Problems and Multi-Source Problems. When there is no clear structure data does not need to follow strict rules leading to single-source problems. Examples include misspellings, duplicates, uniqueness, etc. Multi-Source Problems are just aggravated Single-Source Problems, i.e., the data are collected from multiple places, like different databases or files and because each place might organize information differently it gets even more complicated. The main problem is overlapping data which needs to be solved by either choosing the

correct one or merging them. Examples include naming conflicts, structural conflicts, inconsistent timings, etc. (Rahm & Do, 2000)

The paper "A Taxonomy of Dirty Data" highlights that large corporations have recognized the value of data warehouses bring to the table; being able to condense data from various sources into one centralized repository for analysis enhances competition. Various industries have their own data warehousing alongside numerous software products allowing them to effectively utilize data for their respective industry. But for analyzing the data effectively the data fitted should be correct. "Dirty" data, which may contain missing, incorrect, or non-standard representations drastically reduces the effectiveness of data analysis applications. Dirty data can originate from numerous sources ranging from human data entry errors to bugs in data processing systems. The paper's taxonomy of data is based on the fact that dirty data manifests itself in three different ways: missing data, not missing but wrong data, and not missing and not wrong but unusable. Missing data refers to information that is absent or incomplete within a dataset. Examples are incomplete records, non-responses in surveys, data corruption, etc. Not missing but wrong data are instances where data is present but incorrect or inconsistent. Examples are abbreviations or aliases, wrongful data entry, etc. Not missing not wrong data but unusable data includes data that is present, and correct, but still unusable due to various reasons. Reasons can be incompatibility with other systems or datasets, integration of two or more databases, etc. (Kim, et al., 2003)

3. Analysis

Both papers discuss data quality problems in relational database systems and agree that managing transactions and keeping data accurate is important. However, the paper "A Taxonomy of Dirty Data" talks more about problems with data such as different data representations and data unusability across databases. Also, while both papers agree that current database technology has limits again the paper "A Taxonomy of Dirty Data" talks more about these limits and how to solve them.

4. Conclusion

- **Different Representations of Data:**
  There are numerous ways to represent data which makes it confusing. Data can be encoded in different format, represented in different way or measured in different units which makes it hard to compare and understand the data which in turn makes it harder to analyze.
- **Wrong Data Due to Integrity Constraint Violations:**
  In current relational database systems complex data types face issues with data quality as integrity constraints are not enforced properly leading to inconsistent and inaccurate data.
- **Unusable Data from Differences across Databases:**
  When integrating and analyzing data from multiple databases several challenges arise due to differences in data representation, standards, and schemas which renders the database system to be less effective.

5. References

Kim, W. et al., 2003. A Taxonomy of Dirty Data. *Data Mining and Knowledge Discovery,* 7(1), pp. 81-99.

Rahm, E. & Do, H. H., 2000. Data Cleaning: Problems and Current Approaches. *IEEE Data Eng. Bull.,* Volume 23, pp. 3-13.