DATA LAKE OR DATA WAREHOUSE?
DATA CLEANING OR DATA WRANGLING?
HOW TO ENSURE THE QUALITY OF YOUR DATA?

Anastasija Nikiforova
Assistant Professor of Information Systems, Faculty of Science and Technology, Institute of Computer Science, Chair of Software Engineering, University of Tartu
European Open Science Cloud (EOSC) Task Force “FAIR metrics and data quality”

Data Science Seminar: When, why and How? The important of Business Intelligence
9.11.2022, Tartu, Estonia
BIO

PhD in Computer Science – Data Processing Systems and Data Networking

Research interests include but are not limited to data management with a focus on data quality, open government data, Smart City, Society 5.0, sustainable development, IoT, HCI, digitization.

Most recent experience:
✔ Assistant professor at the University of Tartu, Faculty of Science and Technology, Institute of Computer Science, Chair of Software Engineering
✔ European Open Science Cloud Task Force “FAIR Metrics and Data Quality”
✔ associate member of the Latvian Open Technology Association.
✔ expert of the COST – European Cooperation in Science & Technology
✔ visiting researcher at the Delft University of Tehnology, Faculty Technology Policy and Management
✔ assistant professor at the Faculty of Computing, University of Latvia
✔ researcher in the Innovation Laboratory, Faculty of Computing, University of Latvia
✔ IT-expert at the Latvian Biomedical Research and Study Centre, BBMRI-ERIC LV National Node
✔ advisor for the Institute for Social and Political Studies, University of Latvia
BEFORE WE START…

Data Quality Management
- Towards automating data quality specification by extracting data quality requirements from data features
- Data Deduplication (A Record Linkage-Based Data Deduplication Framework)
- Data Quality-aware Software Development
- User-oriented data object-driven approach to data quality assessment → DQMBT (Data Quality Model-Based Testing of IS)
- Data Lake and Data Wrangling for Ensuring Data Quality in CRIS

Integrating artificial intelligence technologies into customer service: Improving the Actionability of Customer Feedback Analysis Using Machine Learning
BEFORE WE START...

Data Quality Management

✔ Towards automating data quality specification by extracting data quality requirements from data features
✔ Data Deduplication (A Record Linkage-Based Data Deduplication Framework)

✔ Data Quality-aware Software Development
✔ User-oriented data object-driven approach to data quality assessment ⇒ DQMBT (Data Quality Model-Based Testing of IS)
✔ Data Lake and Data Wrangling for Ensuring Data Quality in CRIS ➔


Integrating artificial intelligence technologies into customer service: Improving the Actionability of Customer Feedback Analysis Using Machine Learning
BACKGROUND

Today, billions of data sources continuously generate, collect, process, and exchange data. With the rapid increase in the number of devices and IS, the amount and variety of data are increasing.

There is a need to integrate ever-increasing volumes of data, regardless of the source, format or amount, where the data quality, flexibility and scalability in connecting and processing different data sources are crucial.

an effective mechanism should be employed to ensure faster value creation from these data
DATA QUALITY

Why data quality? Again? and still?
Among other “nuances”, data quality is use-case dependent and dynamic (as well as relative) in nature!

*** “absolute data quality” == a level of data quality at which the data would satisfy all possible use cases - is not possible to achieve, but this is the objective to be pursued
DATA REPOSITORY
DATA WAREHOUSE

DATA LAKE?

Maybe even something more?
DATA WAREHOUSE vs DATA LAKE

Data is processed and organized into a single schema before being put into the warehouse.

The analysis is done on the cleansed data in the warehouse.

Raw and unstructured data goes into a data lake.

Data is selected and organized as and when needed.

**DATA WAREHOUSE vs DATA LAKE**

- **Schema on Write**
  - Data is processed and organized into a single schema before being put into the warehouse.

- **Schema on Read**
  - Raw and unstructured data goes into a data lake.

- The analysis is done on the cleansed data in the warehouse.
- Data is selected and organized as and when needed.

---

"single source of truth"

---

**DATA LAKE** vs **DATA WAREHOUSE**

**Raw**
Data Lakes contain unstructured, semi-structured and structured data with minimal processing. It can be used to contain unconventional data such as log and sensor data.

**Large**
Data Lakes contain vast amounts of data in the order of petabytes. Since the data can be in any form or size, large amounts of unstructured data can be stored indefinitely and can be transformed when in use only.

**Undefined**
Data in data lakes can be used for a wide variety of applications, such as Machine Learning, Streaming analytics, and AI.

**DATA LAKE**

**DATA WAREHOUSE**

**Refined**
Data Warehouses contain highly structured data that is cleaned, pre-processed and refined. This data is stored for very specific use cases such as BI.

**Smaller**
Data Warehouses contain less data in the order of terabytes. In order to maintain data cleanliness and health of the warehouse, Data must be processed before ingestion and periodic purging of data is necessary.

**Relational**
Data Warehouses contain historic and relational data, such as transaction systems, operations etc.
**DATA LAKE**

- **Raw**
  Data Lakes contain unstructured, semi-structured and structured data with minimal processing. It can be used to contain unconventional data such as log and sensor data.

- **Large**
  Data Lakes contain vast amounts of data in the order of petabytes. Since the data can be in any form or size, large amounts of unstructured data can be stored indefinitely and can be transformed when in use only.

- **Undefined**
  Data in data lakes can be used for a wide variety of applications, such as Machine Learning, Streaming analytics, and AI.

---

**DATA WAREHOUSE**

- **Refined**
  Data Warehouses contain highly structured data that is cleaned, pre-processed and refined. This data is stored for very specific use cases such as BI.

- **Smaller**
  Data Warehouses contain less data in the order of terabytes. In order to maintain data cleanliness and health of the warehouse, Data must be processed before ingestion and periodic purging of data is necessary.

- **Relational**
  Data Warehouses contain historic and relational data, such as transaction systems, operations etc.
Raw
Data Lakes contain unstructured, semi structured and structured data with minimal processing. It can be used to contain unconventional data such as log and sensor data.

Large
Data Lakes contain vast amounts of data in the order of petabytes. Since the data can be in any form or size, large amounts of unstructured data can be stored indefinitely and can be transformed when in use only.

Undefined
Data in data lakes can be used for a wide variety of applications, such as Machine Learning, Streaming analytics, and AI.

Implementing a Data Lake or Data Warehouse Architecture for Business Intelligence? | by Lan Chu | Towards Data Science
Data Swamp

- No metadata
- Broken metadata management
- No data governance
- Broken ingestion process

Data Lake

- Metadata
- Information is in rows and columns
- Easily ordered and processed with data mining tools
- Has data context
- Contains a data set for running analytics
- Has directories and sub-directories
So how to get its benefits?
DATA LAKE

Data scientist

Data engineer
DATA LAKE & DATA WAREHOUSE

DATA LAKEHOUSE
Data lakehouse is seen as a combination of data warehousing workloads & data lake economics.
DATA LAKE FOR BUSINESS INTELLIGENCE

BUSINESS DATA LAKE
DIRTY DATA YOU HAVE
CLEAN IT UP YOU MUST

Or how to avoid GIGO*?

*“garbage in, garbage out”
DATA CLEANING or DATA WRANGLING?
DATA WRANGLING VERSUS DATA CLEANING

DATA CLEANING

Process of detecting and removing corrupted or inaccurate records from a record set, table or database

Data cleansing is another name for data cleaning

Visit www.PEDIAA.com

DATA WRANGLING

Process of transforming and mapping data from one raw data form into another form with the intent of making it more appropriate and valuable for various tasks

Data munging is another name for data wrangling

a process of iterative data exploration and transformation that enables their further analysis by making them (1) usable, (2) credible and (3) useful
The nature of data lake allows to store a variety of data within the memory

**BUT**

there is a need to clean up dirty data and enrich them in a pre-processing process, where data wrangling is found to be suitable for these purposes.

The goal is to convert complex data types and data formats into structured data without programming efforts ➔ users should be able to prepare and transform their data without the need of using the ETL tools or familiarity and use of programming languages, where these transformations should be automatically suggested after reading the data based on machine learning algorithms that greatly speeds up this process.
DATA LAKE + DATA WRANGLING
= DATA QUALITY IN IS

[an asset, not a silver bullet]
The data wrangling process to prepare research information and integrate it into CRIS

Depending on the IS and the desired or required target quality*, individual steps should be carried out several times ➔ data wrangling is a continuous process that repeats itself repeatedly at regular intervals.
<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Select data</td>
<td>The required data records are identified in different data sources. When selecting data, a record is evaluated by its value if there is added value, the availability and terms of use of the data and subsequent data from this data source are checked.</td>
</tr>
<tr>
<td>Structure</td>
<td>In most cases, there is little or no structure in the data, change the structure of the data for easier accessibility.</td>
</tr>
<tr>
<td>Clean</td>
<td>Almost every dataset contains some outliers that can skew the analysis results, the data are extensively cleaned for better analysis (processing of null values, removing duplicates and special characters, and standardization of the formatting to improve data consistency).</td>
</tr>
<tr>
<td>Enrich</td>
<td>The data needs to be enriched - an inventory of the data set and a strategy for improving it by adding additional data should be carried out. The data set is enriched with various metadata:  ✔ Schematic metadata provide basic information about the processing and ingestion of data, the data wrangler analyzes / parses data records according to an existing schema.  ✔ Conversation metadata are exchanged between accessing instances with the idea to document information obtained during the processing or analysis of these data for subsequent users. The recognized peculiarities/ features of a data set can be saved.</td>
</tr>
<tr>
<td><em>Data lake</em></td>
<td>The physical transfer of data in the data lake. Although data are prepared using metadata, the record is not pre-processed.  The goal is to avoid a data swamp, estimate the value of the data and decide on their lifespan depending on the data quality and its interconnectedness with the rest of the DB.  Analyzes are not performed directly in the data lake, but only on the relevant data.  To be able to use the data, the requester needs the appropriate access rights. Data Wrangler performs data extraction, however, general viewing and exploration of the data should be possible directly in the data lake.</td>
</tr>
<tr>
<td><em>Data governance</em></td>
<td>The contents of the data lake, technologies and hardware used are subject to change, an audit is required to take care of the care and maintenance of the data lake. The main principles / guidelines and measures that regulates data maintenance coordinating all processes in the data lake and responsibilities are defined.</td>
</tr>
<tr>
<td>Validate</td>
<td>the data are checked one more time before they are integrated into the target CRIS to identify problems with the data quality and consistency of the data, or to confirm that the transformation has been successful. Verify that the values of the attribute are correct and conform to the syntactic and distribution constraints, thus ensuring high data quality AND document every change so that older versions can be restored, or history of changes can be viewed. If new data are generated during data analysis in CRIS, it can be re-included in Data Lake**  **New data go through the data wrangling process, starting with the step 2 of data validating and structuring the data.</td>
</tr>
</tbody>
</table>

At the end of this process, research information can be used by analytical applications and protected from unauthorized access by access control.
USE-CASE

- ✔ Data formatting
- ✔ Correction of incorrect data (e.g. address data)
USE-CASE

- Normalization and standardization (e.g. phone numbers, titles, etc.)
- Structuring (e.g. separation of names into titles, first and last names, etc.)
- Identification and cleaning of duplicates
USE-CASE: TRIFACTA FOR DATA WRANGLING
CONCLUSIONS

✔ As the volume of research information and data sources increases, the prerequisite for data to be complete, findable, comprehensively accessible, interoperable, reusable (compliant with FAIR principles), but also securely stored, structured, and networked in order to be useful remain critical but at the same time become more difficult to fulfill ➔ data wrangling can be seen a valuable asset in ensuring this.

✔ The goal is to counteract the growing number of data silos that isolate data from different areas of the organization. Once successfully implemented, data can be retrieved, managed and made available and accessible to everyone within the entity.

✔ A data lake and data wrangling can be implemented to improve and simplify IT infrastructure and architecture, governance and compliance. They provide valuable support for predictive analytics and self-service analysis by making it easier and faster to access large amount of data from multiple sources.

✔ The proper organization of the data lake makes it easier to find the data the user needs. Managing the data that have already been pre-processed results in an increased efficiency and cost saving, as preparing data for their further use is the most resource-consuming part of data analysis.

✔ By providing pre-processed data, users with limited or no experience in data preparation (low level of data literacy) can be supported and analyzes can be carried out faster and more accurately.
THANK YOU FOR ATTENTION!

QUESTIONS?

For more information, see ResearchGate, anastasijanikiforova.com

For questions or any queries, contact me via Nikiforova.Anastasija@gmail.com,