Data Quality and Integration

Group-2

COURSE PRESENTER

(DR. Ebtisam Alharbi)

Members:

SUBMITTED BY:	STUDENT ID:
Layan Adel Babkur	444002368
Jana Abulraouf Al-Lihyani	444001382
Reham Faisal Alsubhi	444003014
Manar Ali Al-Subhi	444003523

DEPARTMENT OF (INFORMATION SCIENCE-DATA SCIENCE)

COLLEGE OF COMPUTER AND INFORMATION SYSTEMS

UMM AL-QURA UNIVERSITY

Table of Contents

1. Introduction

The era of big data has its importance in data quality and data integration from different sources. This project aims to address two major problems: DQI (Data Quality and Integration) will be one of the key challenges that will help to deal with the challenges of increasing volume, variety, and speed of data in modern database management and analytics. The project will analyze ways and methods for data integrity protection, and unification of several data sets, which will enable organizations to have tools and frameworks that could help them overcome complexities in data management. The primary goal is to facilitate organizations not only detect and resolve errors, but also to unify data records and implement the data governance policies which in the long run will enable the organization to fully realize the potential of their data assets for data-driven and strategic decisions.

2. Comprehensive Description

Data Quality and Integration

Data quality and integration (DQI) covers the entire data acquisition approach, which includes all related processing, activities, and techniques applied for ensuring the rightness, consistency, and availability of data from heterogeneous sources. Data usage here emerges as a key step that organizations may use when they are looking to make strategic decision.

Data Quality involves:

- *Accuracy:* The data modeling process involves getting the data to accurately represent the real-world objects that it models as one of the most significant aspects is the data modeling.
- *Consistency:* To ensure uniformity across datasets
- *Completeness*: To ensure that all the necessary data is available
- *Timeliness*: To ensure that the data is up to date
- *Validity:* To ensure the data compliance with the format and constraint

Data Integration involves:

- *Data Cleansing:* Identifying and correcting errors and inconsistencies.
- *Data Matching:* Merging records that correspond to the same entity across different datasets.
- *Data Transformation:* Converting data into common formats or structures for integration.
- *Data Consolidation:* Integrating data from multiple sources to generate a single view of the issue.
- Data Governance: Providing policies and procedures on how to handle the quality and integration of data to be done in an effective manner.

The project will be working on these aspects by developing a complete framework that combines different tools and techniques to achieve high levels of data quality as well as integration. Such architecture will be built to handle the issues that come with modern big data ecosystems that are characterized by huge volumes of heterogeneous data that is being generated in great speed. The framework will consist of data cleansing that can be automated, advanced data matching, scalable data transformation, real-time data integration, and complete data governance.

3. Background

Historical Context and Evolution

For many years, data quality and integration were the two most critical elements of database management. It was however with the emergence of big data and complex IT environments that their role became more dominant. Conversely, ensuring data quality was traditionally a laborious process and very error-prone, while integration efforts were often done in isolation and inefficiently.
The emergence of the internet, social media and IoT has exponentially increased the amount, type and speed of data which makes maintaining the quality of data and data integration more challenging. The conventional ways of data management did not have the capabilities to handle the large amounts of data generated by these novel sources.

The Importance of Data Quality and Integration

Data quality and integration are critical for several reasons:

1.Decision-Making: Top-notch and integrated data is a base for making right decisions for the business.

2. Operational Efficiency: For instance, the right data can enhance operational activities and decrease the expenses related to data inaccuracies.

3. Regulatory Compliance: Numerous sectors are subject to regulations demanding for keeping up with the information that is up to date and consistent.

4. Customer Satisfaction: Precise and up-to-date data improve customer experience by ensuring that customers are provided with correct and instant updates.

The development of DQI has been motivated by the need to deal with these challenges, which in turn, has led to the introduction of new methodologies and technologies whose purpose is to improve data quality and to ensure the smooth interoperability between different sources of data.

4. Literature Review

Early Research and Foundational Works

The first stage of data quality and integration research was centered on the introduction of basic data cleansing techniques as well as methodologies for data matching and deduplication. For example, Rahm and Do (2000) gave an extensive description of the existing issues and solutions in data cleaning and indicated the role of data quality in database systems [Rahm & Do, 2000]. Batini and Scannapieco (2006) expanded the dimensions and metrics of data quality, presenting a comprehensive structure for data quality assessment and improvement [Batini & Scannapieco, 2006].

Data Quality

Data Cleansing and Error Detection

Data cleansing, commonly known as data scrubbing, is a process of identifying and rectifying errors and inconsistencies in data aiming at improving its integrity. Hernández and Stolfo (1998) brought up the topic of real-world data cleansing and merge/purge problem, pinpointing how critical it is to have data cleaning techniques that work [Hernández & Stolfo, 1998]. Data cleansing techniques include:

- *Validation*: Making sure the data's integrity being in the predefined rules and formats.
- *Standardization:* Data transformation into a universal format or structure.
- *Enrichment:* Data improvement through adding an additional information or improving precision.

Data Quality Dimensions and Metrics

Batini and Scannapieco (2006) pointed out the most important aspects of data quality, where accuracy, consistency, completeness, timeliness, and validity were the main ones [Batini & Scannapieco, 2006]. These measures give the basis of assessing whether data is appropriate and possible improvements to it. Data quality is measured by the metrics for each dimension which track the improvements over time.

Data Integration

Data Matching and Deduplication

Data matching refers to the activity of joining up the records that are referring to the same person in different datasets. This is the main element in data integration as it is the source of duplicates and inconsistencies as well as errors in integrated data. Winkler described methods of record linkage and deduplication, and he also emphasized the importance of using suitable methods of matching [Winkler, 1999].

Data Transformation and Consolidation

Transformation of data deals with the conversion of data into a common layout or pattern to be integrated. In this phase, data integration is the most important thing in the process of combining data from various data sources. In their paper, Stonebraker and Hellerstein (2001) specifically dealt with the e-business data architecture, but they were particularly interested in the data transformation techniques [Stonebraker & Hellerstein, 2001].

Data Governance

Data governance means to formulate the rules and procedures of data quality and data integration properly. These activities include assigning roles and responsibilities, creating quality standards for data, and implementing procedures for data monitoring and improvement. Data governance is a must to be ensured that data quality and data integration objectives are in harmony with the organizational goals and regulatory demands.

Recent Advances and Emerging Trends

Machine learning and artificial intelligence have been the main areas of DQI research which can be applied to automate and improve DQI activities. This is a case study of Hentschel et al. (2016) on cloud-based database systems and the lessons learned from industry and academia, where machine learning technology was used to improve data quality and integration [Hentschel et al. , 2016].

5. Current Technology

ETL Tools

ETL (Extract, Transform, Load) tools are the basis of data integration which greatly facilitates automating the process of data extraction from different sources, data transformation into a target format and loading the data into a target database or data warehouse. The widely used ETL tools are mentioned as: Talend, Informatica, and Apache NiFi.

Data Virtualization

Virtualization data platforms allow the users to get the information immediately without having to call for physical consolidation from multiple sources. The combination of data from diverse systems into a single view and the improvement of the data availability and usability are the advantages that come along with this and the data becomes more accessible and useful.

Machine Learning and AI

Artificial intelligence and machine learning have now become common in the last couple of years as automation and improvement of the data quality and integration process. Technologies can be used in data matching, error identification and anomaly detection that all help to improve the DQI process efficiency and accuracy.

Anomaly Detection

Besides, machine learning techniques can be utilized for anomaly detection and identifying and correcting data errors. By means of historical data, organizations can train models to spot patterns and outliers that might become a threat to data quality, and they can thus take preventive measures before the problem actually occurs.

6. References

1. Rahm, E., & Do, H. H. (2000). Data Cleaning: Problems and Current Approaches. *IEEE Data Engineering Bulletin, 23(4), 3-13*.
2. Batini, C., & Scannapieco, M. (2006). *Data Quality: Concepts, Methodologies and Techniques*. Springer.
3. Hernández, M. A., & Stolfo, S. J. (1998). Real-world Data is Dirty: Data Cleansing and The Merge/Purge Problem. *Data Mining and Knowledge Discovery, 2(1), 9-37*.
4. Stonebraker, M., & Hellerstein, J. M. (2001). Content Integration for E-business. *Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data, 552-560*.
5. Winkler, W. E. (1999). The State of Record Linkage and Current Research Problems. *Statistics of Income Division, Internal Revenue Service Publication R99/04*.
6. Hentschel, M., Jacobsen, H. A., & Stolze, K. (2016). Cloud-Based Database Systems: What We Learned from Industry and Academia. *Proceedings of the 2016 International Conference on Management of Data (SIGMOD '16)*.
7. White, T. (2015). *Hadoop: The Definitive Guide*. O'Reilly Media.
8. Zaharia, M., Chowdhury, M., Franklin, M. J., Shenker, S., & Stoica, I. (2010). Spark: Cluster Computing with Working Sets. *Proceedings of the 2nd USENIX Conference on Hot Topics in Cloud Computing*.
9. Dean, J., & Ghemawat, S. (2008). MapReduce: Simplified Data Processing on Large Clusters. *Communications of the ACM, 51(1), 107-113*.
10. Bertino, E., & Sandhu, R. (2005). Database Security—Concepts, Approaches, and Challenges. *IEEE Transactions on Dependable and Secure Computing, 2(1), 2-19*.