7/15/2024

Funmilayo Michael

Student no: 12691429

Deciphering Big Data

Project Executive Summary

Contents:

[Introduction 2](#_Toc171974796)

[Project Overview 2](#_Toc171974797)

[Logical Design 2](#_Toc171974798)

[Proposed Database Model 2](#_Toc171974799)

[Data Items/Entities and Attributes 2](#_Toc171974800)

[**Table 1**. Database Entities and Attributes 2](#_Toc171974801)

[Primary Keys 3](#_Toc171974802)

[Foreign Keys 4](#_Toc171974803)

[Relationships and Associations 4](#_Toc171974804)

[Data Types and Formats 5](#_Toc171974805)

[Choosing a Database Management System (DBMS) 5](#_Toc171974806)

[Data Management Pipeline Process 5](#_Toc171974807)

[Data Capturing and Source 5](#_Toc171974808)

[Methods for Cleaning Data 6](#_Toc171974809)

[Normalization 6](#_Toc171974810)

[Critical Evaluation 6](#_Toc171974811)

[Data Wrangling 6](#_Toc171974812)

[Methodologies and Tools 6](#_Toc171974813)

[Legal and Compliance Requirements 7](#_Toc171974814)

[GDPR Compliance 7](#_Toc171974815)

[GDPR Implementation 7](#_Toc171974816)

[Conclusions and Recommendations 8](#_Toc171974817)

[Conclusion 8](#_Toc171974818)

[Priority Recommendations 8](#_Toc171974819)

[Graphics and Charts 9](#_Toc171974820)

[Figure 1: Entity-Relationship Diagram 10](#_Toc171974821)

[Figure 2: Data Flow Diagram 12](#_Toc171974822)

[References 13](#_Toc171974823)

# Introduction

ABC Electronics, a mid-sized retailer specializing in electronics and home appliances, has engaged our team of software consultants and developers to design and construct a logical database. ABC Electronics operates through two sales streams: physical and online stores. The objective of this project is to develop a comprehensive database system to enhance data management, expedite data retrieval and modification, and support optimal inventory utilization, real-time consumer engagement, and informed decision-making.

# Project Overview

The project involves designing a logical database architecture tailored to ABC Electronics' operations. This database will facilitate efficient data handling for physical and online sales processes. Key elements of the project include defining data items/entities, establishing relationships and associations, choosing an appropriate database management system (DBMS), and implementing a robust data management pipeline.

# Logical Design

## Proposed Database Model

We have chosen a relational database approach because it can handle complex queries and ensure data integrity. This model aligns with ABC Electronics' requirements, enabling seamless data Storage, management, retrieval, and update across interconnected processes.

## Data Items/Entities and Attributes

The primary entities and their attributes, based on the workflow, are detailed in the table below:

### **Table 1**. Database Entities and Attributes

|  |  |
| --- | --- |
| **Entity** | **Attributes** |
| **Customer** | Customer ID\*, First Name, Last Name, Email, Phone Number, Address, Join Date |
| **Product** | Product ID\*, Product Name, Category, Brand, Price, Stock Quantity, Supplier ID\*\* |
| **Order** | Order ID\*, Order Date, Customer ID\*\*, Employee ID\*\*, Total Amount |
| **Order Details** | Order Details ID\*, Order ID\*\*, Product ID\*\*, Quantity, Unit Price, Delivery Requested, Delivery ID\*\*, Completion Status, Completion Date |
| **Delivery** | Delivery ID\*, Delivery Date, Delivery Address, Delivery Postcode, Employee ID\*\*, Completion Status |
| **Stock Orders** | Stock Order ID\*, Stock Order Date, Product ID\*\*, Employee ID\*\*, Quantity, Total Amount |
| **Supplier** | Supplier ID\*, Supplier Name, Contact Name, Phone Number, Email |
| **Employee** | Employee ID\*, First Name, Last Name, Email, Phone Number, Position, Hire Date, Salary |
| **Payments** | Payment ID\*, Order ID\*\*, Stock Order ID\*\*, Employee ID\*\*, Payment Amount, Payment Date, Purpose |
| **Date** | Date ID\*, Date, Day, Month, Year, Quarter |

\*Primary Key, \*\*Foreign Key

## Primary Keys

To maintain relationships and data integrity within a database, primary keys are essential. A primary key serves as a unique identifier for each record, ensuring that every entry in the database is distinct. Examples of primary keys include the Customer ID in the Customer table and the Product ID in the Product table, which ensure the uniqueness of each customer and product, respectively. Similarly, each order in the Order table is uniquely identified by its Order ID, and every entry in the Order Details table is made unique by its Order Details ID. The Supplier ID in the Supplier table guarantees the unique identification of each supplier, while the Employee ID in the Employee table does the same for each employee. Additionally, the Date ID in the Date table ensures the uniqueness of every date entry, facilitating accurate data tracking. PK represents the primary key in the diagram below:

A screenshot of a computer

Description automatically generated

(lucid.app, n.d.)

## Foreign Keys

Foreign keys are used to establish connections between tables, ensuring referential integrity within the database. For example, the Customer ID in the Order table is a foreign key that references the Customer ID in the Customer table, ensuring that every order is linked to an existing customer. Similarly, in the Order Details table, the Order ID and Product ID act as foreign keys, referencing the Order and Product tables respectively. This ensures that each order detail item is associated with a valid order and product.

To maintain the relationship between products and their suppliers, the Supplier ID in the Product table is a foreign key that references the Supplier ID in the Supplier table. Additionally, the Employee ID in the Order table is a foreign key that references the Employee ID in the Employee table, ensuring that every order is associated with an active employee.

Finally, the Date ID in the Order table functions as a foreign key that references the Date ID in the Date table, linking each order to a specific date. This comprehensive use of foreign keys maintains the integrity and consistency of the data across the database.

FK represents the foreign key in the diagram below:

A screenshot of a computer

Description automatically generated

(lucid.app, n.d.)

# Relationships and Associations

The entities are interrelated primarily through one-to-many (1:M) relationships, with some one-to-one (1:1) relationships:

Customer (1) → (M) Order

Product (1) → (M) Order Details, Stock Orders

Order (1) → (M) Order Details

Order (1) → (1) Payments

Order Details (1) → (M) Delivery

Employee (1) → (M) Order, Payments, Delivery, Stock Orders

Supplier (1) → (M) Product

Stock Orders (1) → (1) Payments

Date (1) → (M) Order, Order Details, Customer, Employee, Delivery, Stock Orders, Payments

# Data Types and Formats

The database will utilize various data types, including Integer, String, Decimal, Boolean, and Date, ensuring appropriate formatting for each field.

# Choosing a Database Management System (DBMS)

We recommend using MySQL for ABC Electronics' database needs, supported by the following justifications:

1. Scalability: MySQL can grow to accommodate increasing data volumes as ABC Electronics expands.

2. Performance: MySQL provides excellent performance for read-intensive workloads, beneficial for retail operations.

3. Security: MySQL includes robust security measures to protect sensitive customer and transactional data.

4. Cost-Effectiveness: As an open-source DBMS, MySQL is cost-effective and supported by a large community.

# Data Management Pipeline Process

## Data Capturing and Source

Primary data sources include online orders, inventory management, point-of-sale systems, existing operational databases, and third-party vendors. ETL (Extraction, Transformation, and Loading) procedures will integrate these data sources into the database, minimizing errors by using structured input methods like drop-down boxes and postcode lookups.

## Methods for Cleaning Data

To ensure data accuracy and quality, the following data cleansing methodologies will be implemented:

1. Data Validation: Ensuring data adheres to specified formats and standards.

2. Managing Missing Values: Applying imputation or deletion strategies based on context.

3. Eliminating Duplicates: Maintaining data integrity by identifying and removing duplicate records.

4. Standardization and Normalization: Computing new values from existing data and removing outliers for uniform data formats.

# Normalization

Normalization will be applied to prevent data anomalies and ensure data integrity, adhering to the following standard forms:

1. First Normal Form (1NF): Ensuring each row is unique and columns contain atomic values.
2. Second Normal Form (2NF): Ensuring every non-key attribute is fully functionally dependent on the primary key.
3. Third Normal Form (3NF): Eliminating transitive dependencies to ensure all non-key attributes depend solely on the primary key.

# Critical Evaluation

## Data Wrangling

Data wrangling involves preparing and transforming raw data into a usable format. Major challenges include managing diverse data sources, addressing inconsistent or missing data, and maintaining data integrity. Utilizing tools such as SQL queries, Tableau, and Python's Pandas module can mitigate these challenges.

## Methodologies and Tools

For this project, the following methodologies and tools were adopted:

1. ETL Tools: Talend or Apache NiFi for data extraction, transformation, and loading.
2. Data Cleaning: Python libraries (Pandas, NumPy) for data manipulation and cleaning.
3. Database Management: MySQL for efficient data storage and retrieval.
4. Data Visualization: Tableau for presenting data insights visually.

# Legal and Compliance Requirements

## GDPR Compliance

Ensuring compliance with the General Data Protection Regulation (GDPR) is critical in database management systems. This involves implementing strong data protection measures, securing explicit consent for data collection, and guaranteeing customers' rights to access, rectify, and erase their data. Although GDPR primarily aims to safeguard European citizens' privacy, its enforcement poses challenges, especially within Big Data systems. These systems process large volumes and varieties of data, making it difficult to track data throughout its intricate lifecycle—from collection and ingestion to storage and analytics.

Between 2016 and 2021, research efforts were directed toward creating security tools for GDPR compliance. These tools, however, tend to be application-specific or only partially meet the regulation's requirements. To assess the scope of these tools, identify deficiencies, and set necessary metrics for comparison, we propose a GDPR compliance framework. This framework delineates essential components for implementing the regulation by aligning GDPR requirements with IT design needs. Employing this framework, we analyse significant GDPR solutions within the Big Data sector and offer guidelines for verifying and implementing GDPR in these systems. This strategy aims to streamline GDPR compliance and bolster data protection in complex Big Data environments. (Rhahla, Allegue and Abdellatif, 2021)

## GDPR Implementation

Bańka, Soczyński, and Wasiak (2022) outline practical methods for implementing GDPR compliance, focusing on enhancing the security of personal data processing. Key recommendations include:

1. Legal and Technical Integration: Combining legal frameworks with technical measures such as encryption, pseudonymization, and anonymization.

2. Certification and Good Practices: Utilizing national and European certification mechanisms and adopting sector-specific codes of good practice.

3. Continuous Improvement: Regularly updating procedures and conducting system recovery tests to ensure ongoing compliance.

# Conclusions and Recommendations

## Conclusion

The proposed logical database design for ABC Electronics aims to significantly enhance the retailer's data management capabilities. By adopting a relational database model, we ensure robust data integrity and efficient handling of complex queries, which are essential for managing both physical and online sales operations. The chosen entities and their relationships reflect the core aspects of ABC Electronics' business processes, promoting seamless data integration and real-time decision-making. With MySQL as the selected DBMS, the database will benefit from scalability, performance, security, and cost-effectiveness. Additionally, rigorous data management, cleansing, and GDPR compliance practices will safeguard data quality and privacy, fostering customer trust and operational efficiency.

## Priority Recommendations

1. Implement and Test the Relational Database Design:

* Develop the database schema based on the defined entities, attributes, and relationships.
* Conduct thorough testing to ensure data integrity, query performance, and the correct implementation of primary and foreign keys.

2. Adopt MySQL for Database Management:

* Leverage MySQL’s capabilities to handle growing data volumes and ensure high performance for both read and write operations.
* Utilize MySQL’s security features to protect sensitive customer and transactional data.

3. Enhance Data Quality Through Rigorous Data Management:

* Implement ETL processes to integrate data from various sources into the database.
* Apply data validation, standardization, and normalization techniques to maintain data accuracy and consistency.

4. Ensure GDPR Compliance:

* Develop and integrate GDPR-compliant data protection measures, including encryption, pseudonymization, and anonymization.
* Establish procedures for obtaining explicit consent for data collection and ensuring customers' rights to access, rectify, and erase their data.

5. Utilize Data Wrangling and Visualization Tools:

* Use SQL queries, Python libraries (Pandas, NumPy), and Tableau for data manipulation, cleaning, and visualization.
* Regularly transform raw data into actionable insights to support informed decision-making.

6. Continuous Improvement and Monitoring:

* Regularly review and update data management processes to adapt to changing business needs and regulatory requirements.
* Conduct routine system recovery tests and compliance audits to ensure ongoing adherence to GDPR and other relevant regulations.

By following these priority recommendations, ABC Electronics can establish a robust and efficient database system that enhances operational effectiveness, supports strategic decision-making, and ensures data privacy and security.

# Graphics and Charts

To visualize the database structure and relationships, the following diagrams are provided:

1. Entity-Relationship Diagram (ERD): Illustrates the entities, attributes, and relationships in the database.

2. Data Flow Diagram (DFD): Shows the flow of data through the system, from data capture to storage and retrieval.

## Figure 1: Entity-Relationship Diagram

A screenshot of a computer

Description automatically generated

(lucid.app, n.d.)

A close-up of a graph

Description automatically generated

(lucid.app, n.d.)

A screenshot of a computer

Description automatically generated

(lucid.app, n.d.)

## Figure 2: Data Flow Diagram

A diagram of a process

Description automatically generated

These visual aids help in understanding the database structure and the flow of data within the system, ensuring clarity and effective communication of the database design.

# References

Kameswari, P., Ramesh, V., Ishwarappa Bhavikatti, V., Santhoshi Gondesi, G., et al. 2024. Analyzing the role of big data and its effects on the retail industry. *Web  Intelligence*. 22 (2), pp.1-19. DOI: https://doi.org/10.3233/WEB-230027

Hossain, M.A., Akter, S., Yanamandram, V. and Strong, C. 2024. Navigating the  platform economy: Crafting a customer analytics capability instrument. *Journal of  Business Research*, 170. DOI: https://doi.org/10.1016/j.jbusres.2023.114260

Cuzzocrea, A., Song, I.-Y., & Davis, K. C. 2011. Analytics over large-scale multidimensional data: the big  data revolution! *Proceedings of the ACM 14th International Workshop on Data Warehousing and OLAP*,  101-104. DOI: https://doi.org/10.1145/2064676.2064695

Panda, M., & Patra, M. R. 2015. Data wrangling: Techniques and challenges of big  data. *International Journal of Applied Research*, 1(10), 997-1001. DOI:  https://doi.org/10.5441/002/edbt.2016.44

Taipalus, T. 2024. Database management system performance comparisons: A  systematic literature review. *Journal of Systems and Software*, 208. DOI:  https://doi.org/10.1016/j.jss.2023.111872

Mirayala, N.K. 2024. Evolving trends in open-source RDBMS: Performance,  scalability and security insights. *International Journal of Science and Technology*,  12(2). DOI: https://doi.org/10.21275/SR24126224648

Ma, X. and Wang, Z. 2024. Computer security technology in E-commerce platform  business model construction. *Heliyon*, 10(7), e28571. DOI: https://doi.org/10.1016/j.heliyon.2024.e28571

Rhahla, M., Allegue, S. and Abdellatif, T. (2021). Guidelines for GDPR compliance in Big Data systems. *Journal of Information Security and Applications*, 61, p.102896. doi: <https://doi.org/10.1016/j.jisa.2021.102896>

‌Bańka, M., Soczyński, T. and Wasiak, D. (2022). Practical Methods of Implementation for the Indispensable Mechanism of GDPR Compliance. *Wroclaw Review of Law, Administration & Economics*, 0(0). doi: <https://doi.org/10.2478/wrlae-2021-0013>

‌ lucid.app. (n.d.). *Lucid visual collaboration suite: Log in*. [online] Available at: https://lucid.app/lucidchart/7829a207-f5d0-46f6-94b0-a69cdda9f61a/edit?beaconFlowId=8BF742544AB8653F&invitationId=inv\_ddd0c119-7f57-417d-9e91-9f2c860cbb62&page=0\_0# [Accessed 15 Jul. 2024].

‌