

Prioritizing Technical Debt in Database Normalization Using Portfolio Theory and Data Quality Metrics

Mashel Albarak
 School of Computer Science
 University of Birmingham
 UK
 King Saud University
 KSA
 mxa657@cs.bham.ac.uk

Rami Bahsoon
 School of Computer Science
 University of Birmingham
 UK
 r.bahsoon@cs.bham.ac.uk

ABSTRACT

Database normalization is the one of main principles for designing relational databases. The benefits of normalization can be observed through improving data quality and performance, among the other qualities. We explore a new context of technical debt manifestation, which is linked to ill-normalized databases. This debt can have long-term impact causing systematic degradation of database qualities. Such degradation can be likened to accumulated interest on a debt. We claim that debts are likely to materialize for tables below the fourth normal form. Practically, achieving fourth normal form for all the tables in the database is a costly and idealistic exercise. Therefore, we propose a pragmatic approach to prioritize tables that should be normalized to the fourth normal form based on the metaphoric debt and interest of the ill-normalized tables, observed on data quality and performance. For data quality, tables are prioritized using the risk of data inconsistency metric. Unlike data quality, a suitable metric to estimate the impact of weakly or un-normalized tables on performance is not available. We estimate performance degradation and its costs using Input/Output (I/O) cost of the operations performed on the tables and we propose a model to estimate this cost for each table. We make use of Modern Portfolio Theory to prioritize tables that should be normalized based on the estimated I/O cost and the likely risk of cost accumulation in the future. To evaluate our methods, we use a case study from Microsoft, AdventureWorks. The results show that our methods can be effective in reducing normalization debt and improving the quality of the database.

CCS CONCEPTS

- Information systems → Relational database model; • Software and its engineering → Designing software;

KEYWORDS

Technical debt; Database design; Normalization

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

TechDebt '18, May 27–28, 2018, Gothenburg, Sweden

© 2018 Association for Computing Machinery.
 ACM ISBN 978-1-4503-5713-5/18/05...\$15.00
<https://doi.org/10.1145/3194164.3194170>

ACM Reference Format:

Mashel Albarak and Rami Bahsoon. 2018. Prioritizing Technical Debt in Database Normalization Using Portfolio Theory and Data Quality Metrics. In *Proceedings of TechDebt '18: International Conference on Technical Debt (TechDebt '18)*. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3194164.3194170>

1 INTRODUCTION

Information systems evolve in response to data growth, improving quality, and changes in users' requirements. The evolution process face many risks, such as cost overruns, schedule delays and increasing chances of failure; it is believed that 80% of the cost is spent on system's evolution [7]. Databases are the core of information systems; evolving databases schemas through refactoring is common practice for improving data qualities and performance, among other structural and behavioral qualities. Database normalization is one of the main principles for relational database design, invented by the Turing Award winner Ted Codd [10]. The concept of normalization was developed to organize data in "relations" or tables following specific rules to reduce data redundancy, and consequently, improve data consistency by decreasing anomalies. The benefits of normalization go beyond data quality, and can have ramifications on improving maintainability, scalability and performance [12, 20]. However, developers tend to overlook normalization due to time and expertise it requires, and instead turn to other strategies such as, creating more indexes or writing extra code fixes to achieve quick benefits. With the growth of data, ad-hoc fixes can become ineffective, calling for costly and inevitable future refactoring to address these pitfalls.

The technical debt metaphor was coined to describe and quantify issues arising from ill, inadequate or suboptimal practices in developing, maintaining and evolving systems, while compromising long-term qualities for short-term benefits, such as fast delivery and immediate savings in cost and effort [11]. The technical debt metaphor can be a useful tool for justifying the value of database normalization, through capturing the likely value of normalization (e.g. quality improvements) relative to the cost and effort of embarking on this exercise, or delaying it. Though technical debt research has extensively looked at code and architectural debts [5, 22], technical debt linked to database normalization, has not been explored [5]. In our previous work [4], we defined database design debt as:

"The immature or suboptimal database design decisions that lag behind the optimal/desirable ones, that

manifest themselves into future structural or behavioral problems, making changes inevitable and more expensive to carry out".

In [4], we developed a taxonomy that classified different types of debts, which relate to the conceptual, logical and physical design of the database.

In this study, we explore a specific type of database design debt that relates to the fundamentals of normalization theory. The underlying assumption of the theory is that the database tables should be normalized to the ideal normal form, a hypothetical fourth normal form, to achieve benefits [15]. While this assumption holds in theory, practically it fails due to the required time and expertise. Adhering to this assumption, we link database normalization debt to tables in the database that are below the fourth normal form.

Given the high cost of database normalization, prioritizing tables to be normalized is challenging. Studies have showed that the most critical technical debt items are the ones whose metaphoric interests grow fast over time [6]. Therefore, a method that would aid database developers and stakeholders to analyze and estimate the impact of normalization debts is highly needed to avoid negative and costly consequences on database qualities due to lack of normalization. Performance and data quality are among the most concerned qualities when it comes to database normalization [12, 31, 32]. In this study, we will analyze the impact of normalization debt on both data quality and performance and use the results to prioritize tables that should be normalized to the fourth normal form.

To measure the impact of normalization debt on data quality, we use the risk of data inconsistency metric of the International Standardization Organization [1]. Unlike data quality, a suitable metric to measure the impact of normalization debt on performance is not available. We analyze database performance, by looking at the Input and Output "I/O" cost of the operations performed on tables below fourth normal form. I/O cost is the number of disk pages read or written to execute the operation [18]. Weakly or un-normalized tables will acquire more pages due to big amount of data duplication they hold and big records size. Therefore, operations executed on those tables will increase the I/O cost, which in turn will affect performance. As part of our contribution, we propose a model to control the debt impact on the I/O cost. We build on Modern Portfolio Theory [24] to manage the debt by prioritizing tables that should be normalized based on the I/O cost and the risk of table's size growth.

The contributions of this paper can be summarized as follows:

- Explore technical debt related to database normalization: Define and estimate the impact of normalization debt, where we specifically look at the ramifications of the normalization decisions on performance and data qualities.
- In large-scale systems, designers are often faced with the challenge of limited resources when it comes to normalization and managing the debt. To address this challenge, our approach prioritize tables that should be normalized based on their impact on data quality and performance. For performance, we utilize portfolio theory to prioritize tables that should be normalized, taking into account: debt impact on operations performance and its accumulation over time.

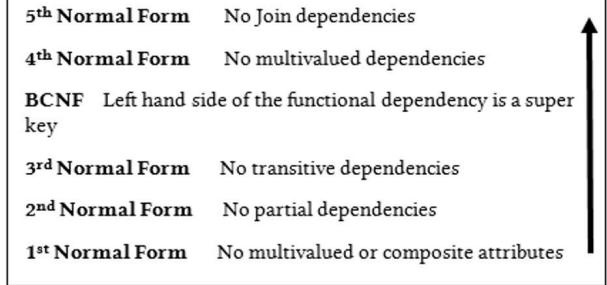


Figure 1: Normal forms hierarchy

The contribution provides a basis for software engineers and database developers to understand normalization, not only from the technical perspective but also from its connection to debt. Coining normalization with debt hopes to provide a systematic procedure that can eliminate unnecessary normalization efforts, justify the essential ones, and/or prioritize investments in normalization, when resources and budget are limited.

The methods are evaluated using a case study from Microsoft, AdventureWorks database [2], and StoreFront web application built on top of the database [3]. The database has a total of 64 tables, each populated with large amount of data to support e-commerce transactions of a hypothetical cycles retail company.

2 BACKGROUND AND MOTIVATION

In this section, we will summarize the key concept of our work. The normalization process was first introduced by Codd in 1970 [10], as a process of organizing the data in relations or "tables". The main goal of normalization is to reduce data redundancy, which is accomplished by decomposing a table into several tables after examining the dependencies between attributes. Benefits of normalization was discussed and proved in the literature [12, 15]. Examples of such benefits include: improving data quality as it reduces redundancy; minimizing update anomalies and facilitating maintenance. Figure 1 illustrates the normal forms hierarchy. Higher level of normal form indicates a better design [15], since higher levels reduce more redundant data. The main condition to go higher in the hierarchy is based on the constraint between two sets of attributes in a table, which is referred to as dependency relationship.

2.1 Benefits of Normalization: Data Quality and Performance

Improving data quality is one of the main advantages of normalization [12, 15]. This improvement is linked to decreasing the amount of data duplication as we move higher in the normalization hierarchy. In the presence of data duplication, in poorly or un-normalized tables, there is always a risk of updating only some occurrences of the data, which will affect the data consistency. Data quality is a crucial requirement in all information systems, as the success of any system relies on the reliability of the data retrieved from the system. The trustworthiness of business processes depends highly on the quality of the exchanged data.

Benefits of normalization go beyond data quality and can be observed on other aspects of the system such as performance. Performance has always been a controversial subject when it comes to normalizing databases. Some may argue that normalizing the database involves decomposing a single table into more tables, henceforth, data retrieval can be less efficient since it requires joining more tables as opposed to retrieving the data from a single table. Indeed, de-normalization was discussed as the process of "downgrading" table's design to lower normal forms and have limited number of big tables to avoid joining tables when retrieving the required data [12]. Advocates of de-normalization argue that the Database Management System (DBMS) stores each table physically to a file that maintains the records contiguously, and therefore retrieving data from more than one table will require a lot of I/O. However, this argument might not be correct: even though there will be more tables after normalization, joining the tables will be faster and more efficient because the sets will be smaller and the queries will be less complicated compared to the de-normalized design [12]. Moreover, weakly or un-normalized tables will be stored in a large number of files as opposed to the normalized design due to big amount of data redundancy, and consequently, increased records size and increased I/O cost [12, 31]. Adding to this, not all DBMS store each table in a dedicated physical file. Therefore, several tables might be stored in a single file due to reduced table size after normalization, which means less I/O and improved performance, as proven in [12, 31, 32].

Despite the controversy about normalization and performance, several arguments in favor of normalization is presented by C. J. Date [12], an expert who was involved with Codd in the relational database theory:

- Some of the de-normalization strategies to improve performance proposed in the literature are not "de-normalizing" since they do not increase data redundancy. In fact, as Date states, some of them are considered to be good normalized relational databases.
- There is no theoretical evidence that de-normalizing tables will improve performance. Therefore it's application dependent and may work for some applications. Nevertheless, this does not imply that a highly normalized database will not perform better.

In this study, we view these deteriorations in performance and data quality as debt that can be rooted in inadequate normalization of database tables. The debt can accumulate overtime with data growth and increased data duplication. It is evident that the short-term savings from not normalizing the table can have long-term consequences which may call for inevitable costly maintenance, fixes and/or replacement of the database. Therefore, we motivate rethinking database normalization from the debt perspective linked to performance and data quality issues. Rethinking databases normalization from the technical debt angle can lead to software that has the potential to better evolve and cope with changes and continuous data growth.

2.2 Cost of Normalization

As mentioned previously the process of normalizing the database is relatively complex due to expertise and resources required. According to [29], decomposing a single table for normalization involves several refactoring tasks:

- Modifications to the database schema: create the new tables after decomposition; create triggers to ensure that the data in new tables will be synchronized and finally updating all stored views, procedures and functions that access the original table.
- Migrating data to new table/s: a strategy should be set to migrate the data from the old weakly or un-normalized table to the new tables.
- Modifications to the accessed application/s, which involve introducing the new tables' meta data and refactor the presentation layer accordingly.

Additionally, testing the database and the applications before deployment can be expensive and time consuming. Consequently, the aim should not be normalizing all the tables, rather than prioritize tables that has the most negative effect to be normalized. In this study, we aim to formulate the normalization problem as a technical debt. We start from the intuitive assumptions that tables, which are weakly/not normalized to the deemed ideal form, can potentially carry debt. To manage the debt, we adhere to the logical practice in paying the debt with the highest negative impact first. We use the impact information of the debt item to inform normalization decisions, constrained to time, budget and quality objectives.

2.3 Database Refactoring

Database refactoring, or some may refer to as schema evolution, has been largely studied over the past years [9, 26, 29]. Schema evolution is basically altering the schema of the database to adapt it to a certain required change[29]. Attempts have been made to analyze the schema evolution from earlier versions [9], and tools have been developed to automate the evolution process [26]. However, schema evolution literature has focused on limited database refactoring tactics, such as adding or renaming a column, changing the data type, etc. Evolving the schema for the purpose of normalizing the database to create value and avoid technical debt has not been explored. We posit that normalization is an exercise that should create a value. By creating a value, it is ensured that the system will sustain and be much more usable and maintainable.

The process of normalizing a database can be very costly since it involves decomposing weakly normalized tables into several tables following specific rules, and updating the applications that use the database. Researchers have provided several algorithms to automate database normalization [13, 14]. Their studies aimed to produce up to 3rd normal form or BCNF tables automatically. However, these studies looked at the schema in isolation from applications using the database. The consideration of the applications is important to better estimate the cost of normalization taking into account refactoring, configuration and data migration tasks. Since this process can be costly, our study aims to provide a method to prioritize tables to be normalized to improve the design and avoid negative consequences.

Table 1: Debt Tables Below 4th Normal Form

Table Name	Normal Form
Product	Un-normalized
Employee	1 st normal form
EmployeePayHistory	1 st normal form
ProductProductPhoto	1 st normal form
WorkOrder	2 nd normal form

3 NORMALIZATION DEBT DEFINITION

There is no general agreement on the definition of technical debt in the literature [21]. However, the common ground of existing definitions is that the debt can be attributed to poor and suboptimal engineering decisions that may carry immediate benefits, but are not well geared for long-term benefits. Database normalization had been proven to reduce data redundancy, improve data quality and performance as the table is further normalized to higher normal forms [12]. Consequently, tables below the 4th normal form can be subjected to debts as they potentially lag behind the optimal, where debt can be observed on data consistency and performance degradation as the database grows [12, 31]. To address this phenomenon of normalization and technical debt, we have chosen 4th normal form as our target normal form. Although most practical database tables are in 3rd (rarely they reach BCNF) normal form [15], 4th normal form is considered to be the optimal target since more duplicate data is removed. Additionally, 4th normal form is based on multi valued dependency which is common to address in real world [34]. Moreover, higher normal form of the 5th level is based on join dependencies that rarely arise in practice, which makes it more of a theoretical value [12, 15].

Following similar ethos of database design debt definition [4] , we view that database normalization debt is likely to materialize for any table in the database that is below the 4th normal form.

Using available data in the AdventureWorks data dictionary [2], we have identified 23 debt tables below 4th normal form. Due to space limitation, we will analyze five tables presented in Table 1:

4 NORMALIZATION DEBT IMPACT

Data quality and performance are among the qualities that are affected by database normalization or its absence. Normalization technical debt can be observed on these qualities. As it can be difficult to eliminate the debt, we hope to manage the debt by prioritizing debt tables below the 4th normal form that are believed to have the highest negative effect on them. Henceforth, the first step is to estimate the impact of each debt table on those qualities, where the estimation will be the main driver of prioritizing debts to be paid.

In technical debt area, the debt interest is a crucial factor for managing debts. In general, the interest of technical debt is the cost paid overtime by not resolving the debt [17]. Unlike finance, interest on technical debt has been acknowledged to be difficult to elicit and measure for software engineering problems [6]. The interest can span various dimensions; it can be observed on technical, economic, and/or sociotechnical dimensions, etc. One important aspect of

Table 2: Risk of Data Inconsistency for the Debt Tables

Table Name	Risk of Data Inconsistency
Product	17
Employee	8.213
EmployeePayHistory	1.034
ProductProductPhoto	0.978
WorkOrder	2.910

technical debt interest is that it is context dependent, meaning the same detected debt can have more or few interest in the future depending on the project and its circumstances . Researchers have coined interest with implication on qualities [35]. The analogy is applicable to the case of normalization as striving for the ideal 4th normal form will reduce data duplication substantially and consequently, decrease the rate at which the current debt tables negatively affect data quality and performance.

4.1 Data Quality

Data quality is considered as one of the main advantages of normalization. In this study, data quality is represented by its consistency. In weakly or un-normalized tables that store big amount of data duplication, there is always a possibility to change or update some occurrences of the data leaving the same duplicated data un-updated. Therefore, the risk of data inconsistency will be higher. This risk is decreased by normalization as the amount of data duplication is decreased. Thus, we view the risk of data inconsistency as an impact incurred due to weakly normalized tables. This Impact can be quantified using the International Standardization Organization (ISO) metric [1]. According to ISO, The risk of data inconsistency is proportional to the number of duplicate values in the table and it is decreased if the table further normalized to higher normal forms. This risk can be measured using the following formula:

$$X = A \div B$$

With $\binom{n}{k}$ sets of k attributes for a table with n attributes (k=1..n), where $A = \sum k \sum iDi$, Di = number of duplicate values found in set i of k attributes, and $B = \text{number of rows} \times \text{number of columns}$.

Rows refer to the data stored in the table and columns refer to the attributes of that table. For X, lower is better. Using this metric, developers can prioritize tables to be normalized by calculating the risk of data inconsistency for all the tables in the database and identify tables with higher risks to be normalized.

We have calculated the risk of data inconsistency for the tables in AdventureWorks database, the results are shown in Table 2.

Based on the results from Table 2, the logical approach would be to normalize table *Product* to the 4th normal form as it has the highest risk of data inconsistency. It is also observable from the results that the risk of data inconsistency is independent from the normal form of the debt table. As shown, table *WorkOrder* has a higher risk of data inconsistency than tables *EmployeePayHistory* and *ProductProductPhoto* even though they are in a weaker normal form, which denotes that the normal form alone is not sufficient to make the right decision on which table to normalize.

4.2 Performance

Enhanced operations performance is one of the advantages of database normalization, as the database will store smaller sets of data in tables due to the decreased amount of data duplication. Operations performed on the database include update; insert; delete data stored in the tables, in addition to data retrieval operations. Each of these operations incur a cost on the number of disk pages read or written, which is referred to as the Input/Output "I/O" cost [18]. The impact of normalization debt can be observed through the I/O cost incurred by the operations performed on the debt table in its current normal form. Due to the huge amount of data redundancy in debt tables below 4th normal form, tables will require more pages to be stored in, which will affect the performance of the operations executed on those tables. Meaning, the more data stored in the debt table, the more disk pages it requires and henceforth, more time to go through the pages to execute the operation. Therefore, normalization is the sensible solution that will not only improve performance, as it has been proven in several studies [12, 31, 32], but it will also improve the overall design of the database to adapt easily to future changes.

Unlike data consistency, a metric to quantify the I/O cost of all the operations performed on debt tables is not available. In the following section we propose a model to estimate this cost and suggest an approach, which is grounded on portfolio approach to manage it effectively.

4.2.1 Performance Estimate: The Impact of Normalization Debt on the I/O cost.

I/O cost of each operation can be elicited from the database management system via query analyzer tools. To quantify the total I/O cost for each debt table, we use the execution rate of the operation and the I/O cost of operations. Let n_u , n_i , n_d , n_s be the number of update, insert, delete and select operations on the debt table R respectively. Λ_x^u , Λ_x^i , Λ_x^d , Λ_x^s represent the execution rates of the x^{th} update, x^{th} insert, x^{th} delete and x^{th} select respectively. Finally, the I/O costs of the x^{th} update, x^{th} insert, x^{th} delete and x^{th} select are denoted by C_x^u , C_x^i , C_x^d , C_x^s respectively. It is important to note that operations on a specific single table are considered (excluding the join operations when calculating the impact). The reason for this is because we want to examine the cost incurred by a specific debt table due to its weakly normalized design. Moreover, after the decision is made whether to pay or keep the debt, the join operations are still needed to retrieve the required data. Therefore, the I/O cost of a debt table can be calculated as follows:

$$I/O\text{Cost} = \sum_{x=1}^{n_u} C_x^u \Lambda_x^u + \sum_{x=1}^{n_i} C_x^i \Lambda_x^i + \sum_{x=1}^{n_d} C_x^d \Lambda_x^d + \sum_{x=1}^{n_s} C_x^s \Lambda_x^s$$

Let us consider a table named Staff as an example. This table stores staffs' information. Assuming table Staff is in the 2nd normal form, where the ideal is the 4nd normal form and henceforth, considered as a debt table. Suppose that there are 2 select statements to retrieve data from this table, 1 update statement to update some information, 1 insert statement and no delete statements. Values of the other variables are assumed as in Table 3

Referring to Table 3, the I/O cost of table Staff's operations would be:

Table 3: Values of Table Staff I/O Cost Variables

Variable	Value
n_u	1
n_i	1
n_s	2
C_1^u	2 I/Os
Λ_1^u	100/month
C_1^i	1 I/O
Λ_1^i	50/month
C_1^s	5 I/Os
Λ_1^s	200/month
C_2^s	3 I/Os
Λ_2^s	500/month

$(2 \times 100) + (1 \times 50) + (5 \times 200) + (3 \times 500) = 2750$ Estimated average I/O cost monthly.

4.2.2 Technical Debt Impact Accumulation.

Technical debt grows with the accumulated interests on the debt [17]. When likely interest accumulation is overlooked, it can lead to ill-justified decisions regarding debt payment. For example, suppose that that a debt is identified in a software artifact and the developer demands that it should be fixed before the next release. As a result, the developer may end up wasting effort, time and cost for fixing something that may not have severe or noticeable impact on the system or business in the future. In normalization debt context, similar situation is resembled in I/O cost accumulation. I/O cost changes based on the debt tables' growth rate. If the table is likely to grow faster than other tables, the I/O cost for the operations executed on that table will accumulate faster than others. This due to the fact increasing table size implies more disk pages to store the table and therefore, more I/O cost. Tables' growth rate is a crucial measure to prioritize tables needed to be normalized. If the table is not likely to grow or its growing rate is less than other tables, a strategic decision would be to keep the debt and defer its payment. Table growth rate can be elicited from the database monitoring system. The growth rate of a table can be viewed as analogous to interest risk or interest probability. Interest probability captures the uncertainty of interest growth in the future [23]. Debt tables which experience high growth rate in data can be deemed to have higher interest rate. Consequently, these tables are likely to accumulate interest faster.

4.2.3 Managing I/O cost of normalization debt (Portfolio approach).

4.2.3.1 Modern Portfolio Theory.

Modern Portfolio Theory (MPT) was developed by the Nobel Prize winner Markowitz [24]. The aim of this theory is to develop a systematic procedure to support decision making process of selecting capital of a portfolio consisting of various investment assets. The assets may include stocks, bonds, real estate, and other financial products on the market that can produce a return through investment. The objective of the portfolio theory is to select the combination of assets using a formal mathematical procedure that can maximize the return while minimizing the risk associated with

every asset. Portfolio management involves determining the types that should be invested or divested and how much should be invested in each asset. This process draws on similarity with the normalization debt management process, where developers can make decisions about prioritizing investments in normalization, based on which technical debt items should be paid, ignored, or can further wait. With the involvement of uncertainty, assets expected return and variance of the return are used to evaluate the portfolio performance. The expected return of a portfolio is presented by the following equation as the weighted sum of the expected return of the assets in the portfolio. The weight of an asset represents the proportion from the capital invested in this asset.

$$E(R) = \sum_{i=1}^n X_i E(R_i)$$

Where $E(R)$ is the expected return of a portfolio of n assets, X_i is the weight of asset i . The weight determines the proportion of the money that should be invested in asset i . $E(R_i)$ is the expected return of that asset. This expected return of the portfolio is constrained by the following equation, which denotes the sum weight of all the assets in the portfolio should equal to one:

$$\sum_{i=1}^n X_i = 1$$

To measure the risk of a financial portfolio, the variance and standard deviation is used. The risk of a portfolio depends on:

- The volatility of each asset's return which is estimated based on the observation of its return over time.
- The weight invested in each asset
- The correlation between assets return which is estimated by the observation of assets overtime.

The risk of a portfolio R is calculated using the following equation [24]:

$$R = \sqrt{\sum_{i=1}^n X_i^2 \sigma_i^2 + \sum_{i=1}^n \sum_{j=2}^n X_i X_j p_{ij} \sigma_i \sigma_j}$$

Where X_i is the weight of an asset i , σ is the variance of this asset return and P_{ij} is the correlation between asset i and asset j .

4.2.3.2 Modelling the problem (Portfolio Based Approach to Manage the I/O cost of Normalization Debt).

Our approach aims to help database developers make strategic decisions about refactoring the database for normalization. Let us consider an existing database system with tens of tables, all or most of which are below the 4th normal form. Previous research and the classical theory of normalization would encourage normalizing all tables to the ideal normal form [12]; the exercise would require a lot of time and resources, where a complete refactoring of the whole system might be an alternative cost saving option. Our approach acknowledges the fact that time, resources, and budget is often a constraint that prevent exhaustive and unjustified normalization. Our approach selects and prioritizes the tables to be normalized by constructing a portfolio of multiple debt tables that has the highest priority for normalization. The objective is to minimize the negative impact of the debt tables on performance taking into consideration

the likely growth rate of the table size and henceforth, the risk of interest accumulation.

We view a database of debt tables below the 4th normal form as a market of assets. To fit in portfolio management, each debt table is treated as an asset. For each table, we need to determine whether it is better to normalize that table to the 4th normal form (pay the debt) or keep the table in its current normal form (defer the payment). To decide on this, we need to determine what the expected return of each debt table is. In the case of normalization debt, the expected return of the debt table resembles the estimated performance impact of the table. Tables with the lowest estimated impact are deemed to carry higher expected return. In other words, If the estimated I/O of table's A operations is less than estimated I/O of table's B operations, then table A expected return would be higher than B ; B will then have a higher priority for normalization due to high I/O. We balance the expected return with the risk. In portfolio management, this risk is represented by the variance of the return. For the debt tables, this risk is represented by the tables' growth rate. Tables with the highest growth rate are considered to be risky assets, their likely interest and so the debt will grow faster than other tables of low growth rate.

In order to apply the portfolio theory to normalization debt, few considerations need to be taken into account:

- The expected return of the debt table is equal to $1 \div$ I/O cost.
- The risk of each debt table is equal to the table growth rate for each debt table. This information can be elicited from the database management system by monitoring the table's growth.
- We set the correlation between the debt tables to zero for several reasons: First, the I/O costs of the debt tables are independent. Meaning, the I/O cost of the operations executed on a debt table has no effect on the I/O cost of the operations executed on another debt table. Moreover, the growth rate for each table, which affects the I/O cost, is unique and independent from each other. Lastly, each debt table design is independent from other debt tables, as the decision to keep the debt or normalize the table have no effect on the design and the data of the other debt tables.

Taking into account these considerations, we can apply the portfolio theory, where the database developer is investing in tables' normalization. The database developer needs to build a diversified portfolio of multiple debt tables. Multiple debt tables in the database represent the assets. For each asset i , it has its own risk R_i and I/O cost C_i . Based on these values the developer then can prioritize tables to be normalized. The expected return of debt tables portfolio E_p , built by prioritizing debt tables from the database of m debt tables can be calculated as in the following equation:

$$E_p = \sum_{i=1}^m w_i \frac{1}{C_i}$$

With one constraint represented in the following equation:

$$\sum_{i=1}^m w_i = 1$$

Where w_i represents the resulted weight of each debt table. This weight will resemble the priority of each table for normalization as explained in the process steps.

The risk of table growth rate for debt table i is represented by R_i . The global risk of the portfolio R_p is calculated as the following:

$$R_p = \sqrt{\sum_{i=1}^m w_i^2 R_i^2}$$

Process Steps:

- (1) Identify debt tables: tables below 4th normal form should be identified. If this is not already documented, it would require knowledge of the functional dependencies and rules which can be elicited from the requirements and business analysts.
- (2) Determine the debt tables' growth rate from the database monitor. This step will simplify the method to examine only debt tables with high growth rate.
- (3) Consider only the tables with high growth rates to calculate I/O costs of their operations.
- (4) For each debt table of high growth rate, list all queries, update, insert and delete operations, execution rate for each operation and their I/O costs
- (5) Calculate the I/O cost of each table's operations as explained in section 4.2.1
- (6) Determine the values of the portfolio model variables (expected return= $1/I/O$ cost) and (risk= table growth rate)
- (7) Run the model on the available data to produce the optimal portfolio of the debt tables. The portfolio model will provide the highest weights to those tables with low I/O cost and low table growth rate. Therefore, debt table that has the lowest weight implies the highest priority table that should be normalized.
- (8) Use the results to justify the decisions to stakeholders.

It is important to mention that this process should be executed iteratively before each release for both the I/O cost and growth rate of debt tables vary during the system's life.

4.2.4 Case Study.

We considered the AdventureWorks database, designed by Microsoft [2] and StoreFront web application built on top of the database [3]. This database supports standard e-commerce transactions for a fictitious bicycle manufacturer. The database has a total of 64 tables, each filled with thousands of fictitious data. The data dictionary is available with a fair description of the tables and the attributes, which will facilitate the process of identifying the normal form for each table. To better understand our approach, the following example demonstrates how the steps performed to manage the impact of normalization debt on performance.

Step 1: Identify debt tables: 23 tables have been identified below 4th normal and considered to be debt tables.

Step 2&3: Growth rate can be monitored or retrospectively captured from the database monitoring system. For simplicity, we assume that this procedure has identified 5 tables presented in Table 4 to have the highest growth rate among the debt tables.

Step 4: A list for each table has been constructed similar to Table 5 with all the select, insert, update delete statements, their assumed

Table 4: Debt Tables with the Highest Growth Rate

Table Name	Growth Rate/monthly
Product	0.2
Employee	0.1
EmployeePayHistory	0.1
ProductProductPhoto	0.2
WorkOrder	0.5

Table 5: Table Product List of Operations

Select	State-ments	Execution Rate per month	I/O Cost per ex-ecution
1		3000	2
2		5000	2
3		4000	1
Update	State-ments	Execution Rate per month	I/O Cost per ex-ecution
1		2000	2
2		3000	1
Insert	State-ment	Execution Rate per month	I/O Cost per ex-ecution
1		100	1

Table 6: I/O Cost of the Debt Tables

Table Name	I/O Cost
Product	27100
Employee	15000
EmployeePayHistory	12000
ProductProductPhoto	20000
WorkOrder	4000

execution rates and I/O cost per execution. Table 5 demonstrates a list of table *Product* operations.

Step 5: Table 6 shows the calculated I/O cost of the operations performed on each debt table as detailed in section 4.2.1

Step 6: Table 7 presents the values of the portfolio model variables (expected return and risk); the expected return values presented in the table are rounded values.

Step 7: After running the portfolio model on the available data, the weights presented in Table 8 were determined for each debt table.

From Table 8 we can determine that table *Product* has the highest priority to normalize since it got the lowest weight. Although this table's growing rate is considered to be small when compared to table *WorkOrder*, the I/O cost incurred by the operations executed on table *Product* is the highest, meaning the number of operations performed on that table and the frequencies of those operations on a monthly basis are high. Henceforth, this table will have more I/O cost compared to the other tables; this I/O cost will accumulate

Table 7: Portfolio Model Variables

Table Name	Expected Return=1÷I/O Cost	Risk=Growth Rate/monthly
Product	0.003	0.2
Employee	0.006	0.1
EmployeePayHistory	0.008	0.1
ProductProductPhoto	0.005	0.2
WorkOrder	0.025	0.5

Table 8: Debt Tables Weights

Table Name	Weight
Product	4.4
Employee	35.29
EmployeePayHistory	47.06
ProductProductPhoto	7.3
WorkOrder	5.8

faster than other tables despite the relatively smaller table growth rate. On the other hand, table *WorkOrder* has the second priority for normalization despite its lowest I/O cost. This is because it has the fastest growth rate among the tables, which will accelerate I/O cost accumulation in the future and affect performance. As seen, the difference of the weights between table *Product* and table *WorkOrder* is relatively small, which indicates that both tables are semi-equally important to normalize, considering time and budget constraints. Tables *Employee* and *EmployeePayHistory* are the tables with the least priority to be normalized since both their I/O costs and growth rates are the smallest among tables. However, if time and budget permits only one table to normalize, table *Product* would be the correct choice since it also has the highest risk of data inconsistency. Therefore, by normalizing this table, both performance and data quality will be improved.

5 EVALUATION

One objective of the case study is to investigate how the described methods can be applied by database developers to reason about normalization decisions; improve data quality and performance in an attempt to manage normalization debts and their accumulated interests, while minimizing the cost of database refactoring. The evaluation contrasts our debt-aware approach to the ad-hoc one. The debt metric provides insights on the significance of the impact of weakly or un-normalized tables and their accumulation leading to deterioration in the quality. We have claimed the 4th normal form as the target normal form for the "debt-friendly" table design. However, practically, achieving this target for all tables, as conventional approaches of database normalization suggest, is idealistic and costly process due to constraints on expertise, time and budget. Therefore, our approach facilitates the decision making process for normalization debt management, regarding which table has a higher priority to normalize, taking into consideration the effect of the debt table and likely future quality degradations overtime. In

Table 9: Difference in effort between Conventional approach and Debt-aware approach of Normalization

Approach	Number of tables to normalize
Conventional approach	5
Prioritize based on risk of data inconsistency	1
Prioritize based on I/O cost	2

particular, our debt-aware approach provides insights on whether it is beneficial to go for the 4th normal form based on likely benefits relative to improving the data quality and performance due to this exercise. Table 9 presents how our prioritization method has followed a systematic and informed procedure in significantly reducing normalization effort while improving qualities (rather than ad-hocly targeting all the tables that are below 4th normal form).

Table 9 shows that following the conventional approach, which encourage normalizing all the tables to 4th normal form is more costly; time consuming and ad-hoc than the debt-aware approach for two scenarios: The first scenario prioritizes debt-tables based on their impact on data quality, measured using the ISO metric Risk of Data Inconsistency [1], where it suggests to only normalize one table, table *Product*, based on the biggest amount of data duplication it holds. The second scenario considers the performance impact of each debt table, and the likely accumulation of this impact in the future. By utilizing the portfolio approach, two tables are suggested for normalization, *Product* and *WorkOrder*. Moreover, depending on available resources, developers can include more tables to normalize and justify their decisions based on the debt tables' impact on data quality and performance. In summary, our debt-aware approach has provided more systematic and informed cost-effective decision than the conventional approach for normalization taking into consideration debts linked to performance and data quality.

Though our approach has looked at two essential qualities in database normalization, performance and data quality, the approach can be extended and be applicable to reason about the likely debts and interests relative to other structural or behavioral qualities such as maintenance, availability among the others. The extension requires identifying appropriate metrics for the analysis, however. The inclusion of other qualities has the promise to provide database designers and developers with more comprehensive approach for prioritization and debt management.

6 RELATED WORK

Technical debt metaphor was first introduced by Ward Cunningham in 1992 [11]. Back then he described coding debt as a trade-off between short term goals (i.e. shipping the application early to meet the market) and applying the optimal coding practices for long term goals. Technical debt has captured the attention of many researchers over the past years. The attempts have exceeded code level debt to encompass other aspects of the system such as architectural debt [23], requirements debt [16], testing debt [28], among

other types of technical debt [17, 22]. As our work is close to debt prioritization, authors in [35] utilized prioritization to manage code level debts in software design. The authors estimated the impact of God classes on software maintainability and correctness. They prioritized classes that should be refactored based on their impact on those qualities. Portfolio Theory was proposed by researchers to manage technical debt [19, 30]. In [19] the authors viewed each debt item as an asset, and they utilized Portfolio Theory to construct a portfolio of debt items that should be kept based on debt principal, which they defined as the effort required to remove the debt, and debt interest which is the extra work needed if the debt is not removed. Portfolio Theory was also proposed to manage requirements compliance debt in [27]. The authors viewed compliance management as an investment activity that needs decisions to be made about the right compliance goals under uncertainty. They identified an optimal portfolio of obstacles that needed to be resolved, and address value-driven requirements based on their economics and risks. Despite the vast contributions on technical debt in software, database design debts received very little attention. In [33], the authors studied technical debt related to referential integrity constraints in databases schema. They viewed missing foreign keys in the tables as a debt that will affect data quality. They proposed an iterative process to reduce the debt, which involves measuring the debt by making use of missing foreign keys detection algorithms; define modification activities that should be done in the development site and finally, package all modifications to the deployment sites. They illustrated their approach using OSCAR electronic medical record system.

Technical debt interest is among the essential elements needed to manage the debt effectively [8, 25]. Ampatzoglou et al. [6] presented a framework to manage the interest of technical debt. Their framework involved interest definition; classification; evolution and finally, interest management theory, based on Liquidity Preference Theory.

7 CONCLUSION

We have explored a new context of technical debt, which is linked to database normalization. Normalizing the database is essential to improve data quality and performance. However, practically developers tend to overlook normalization process due to time and expertise it requires, and instead, turn to other quick procedures, such as creating more indexes or writing extra code fixes to save time and effort. With data growth and systems' evolution, normalizing the database becomes essential to avoid long-term effects on data quality and performance. We have asserted that normalization debt is likely to be materialized for tables below the 4th normal form and we discussed the validity of this assertion.

Conventional approaches for normalization encourage normalizing all the tables to achieve benefits; the exercise is often judged in isolation of the cost, long-term value and technical debt avoidance. Furthermore, it is impractical to embark on exhaustive normalization of all tables because of costs and uncertain benefits. In this study, we proposed an approach to manage the debt by prioritizing tables that should be normalized. The prioritization is based on the impact of the weakly or un-normalized tables on data quality and performance. To improve data quality, the prioritization is based

on a metric provided by the ISO [1] to measure the risk of data inconsistency for each debt table below the 4th normal form. To enhance performance, a model was proposed to estimate the I/O cost of the operations performed on each weakly or un-normalized table. These costs tend to increase with the growth of the tables. To manage this debt effectively and avoid accumulation of I/O cost, Portfolio Theory was utilized to prioritize tables that should be normalized based on the I/O costs of the operations performed on the tables and the risks of future cost accumulation (e.g., interest on the debt). The techniques was applied to AdventureWorks database from Microsoft. The results show that rethinking conventional database normalization from the debt angle can provide more systematic guidance and informed decisions to improve the database quality, while reducing the cost and effort that is linked with unnecessary normalization.

REFERENCES

- [1] 2015. ISO/IEC 25024:2015 Systems and software engineering – Systems and software Quality Requirements and Evaluation SQuaRE- Measurement of Data Quality. (2015).
- [2] 2018. AdventureWorks Sample Database. (2018). [https://technet.microsoft.com/en-us/library/ms124501\(v=sql.100\).aspx](https://technet.microsoft.com/en-us/library/ms124501(v=sql.100).aspx)
- [3] 2018. StoreFront Application. (2018). [https://technet.microsoft.com/en-us/library/ms160795\(v=sql.90\).aspx](https://technet.microsoft.com/en-us/library/ms160795(v=sql.90).aspx)
- [4] Mashel Al-Barak and Rama Bahsoon. 2016. Database design debts through examining schema evolution. In *Managing Technical Debt (MTD), 2016 IEEE 8th International Workshop on*. IEEE, 17–23.
- [5] Nicoll SR Alves, Thiago S Mendes, Manoel G de Mendonça, Rodrigo O Spínola, Forrest Shull, and Carolyn Seaman. 2016. Identification and management of technical debt: A systematic mapping study. *Information and Software Technology* 70 (2016), 100–121.
- [6] Areti Ampatzoglou, Apostolos Ampatzoglou, Paris Avgeriou, and Alexander Chatzigeorgiou. 2015. Establishing a framework for managing interest in technical debt. In *5th International Symposium on Business Modeling and Software Design, BMSD*. Citeseer.
- [7] Aybuke Arum and Claes Wohlin. 2007. A value-based approach in requirements engineering: explaining some of the fundamental concepts. In *International Working Conference on Requirements Engineering: Foundation for Software Quality*. Springer, 109–115.
- [8] Sofia Charalampidou, Apostolos Ampatzoglou, Alexander Chatzigeorgiou, and Paris Avgeriou. 2017. Assessing code smell interest probability: a case study. In *Proceedings of the XP2017 Scientific Workshops*. ACM, 5.
- [9] Anthony Cleve, Maxime Gobert, Loup Meurice, Jerome Maes, and Jens Weber. 2015. Understanding database schema evolution: A case study. *Science of Computer Programming* 97 (2015), 113–121.
- [10] Edgar F Codd. 1970. A relational model of data for large shared data banks. *Commun. ACM* 13, 6 (1970), 377–387.
- [11] Ward Cunningham. 1993. The WyCash portfolio management system. *ACM SIGPLAN OOPS Messenger* 4, 2 (1993), 29–30.
- [12] Chris Date. 2012. *Database design and relational theory normal forms and all that jazz*. O'Reilly Media, Inc.
- [13] Mousa Demba. 2013. Algorithm for relational database Normalization up to 3NF. *International Journal of Database Management Systems* 5, 3 (2013), 39.
- [14] YV Dongare, PS Dhabe, and SV Deshmukh. 2011. RDBNorma:-A semi-automated tool for relational database schema normalization up to third normal form. *arXiv preprint arXiv:1103.0633* (2011).
- [15] R Elmasri and S Navathe. 2011. Database systems: models, languages, design, and application programming.(2011). (2011).
- [16] Neil A Ernst. 2012. On the role of requirements in understanding and managing technical debt. In *Proceedings of the Third International Workshop on Managing Technical Debt*. IEEE Press, 61–64.
- [17] Carlos Fernández-Sánchez, Juan Garbajosa, Agustín Yagüe, and Jennifer Perez. 2017. Identification and analysis of the elements required to manage technical debt by means of a systematic mapping study. *Journal of Systems and Software* 124 (2017), 22–38.
- [18] Viviana E Ferragine, Jorge H Doorn, and Laura C Rivero. 2009. *Handbook of Research on Innovations in Database Technologies and Applications: Current and Future Trends*. Information Science Reference Hershey, PA.
- [19] Yuepu Guo and Carolyn Seaman. 2011. A portfolio approach to technical debt management. In *Proceedings of the 2nd Workshop on Managing Technical Debt*. ACM, 31–34.

- [20] Davendranath G Jha. 2013. *Computer Concepts and Management Information Systems*. PHI Learning Pvt. Ltd.
- [21] Philippe Kruchten, Robert L Nord, and Ipek Ozkaya. 2012. Technical debt: From metaphor to theory and practice. *Ieee software* 29, 6 (2012), 18–21.
- [22] Zengyang Li, Paris Avgeriou, and Peng Liang. 2015. A systematic mapping study on technical debt and its management. *Journal of Systems and Software* 101 (2015), 193–220.
- [23] Zengyang Li, Peng Liang, and Paris Avgeriou. 2014. Architectural debt management in value-oriented architecting. In *Economics-Driven Software Architecture*. Elsevier, 183–204.
- [24] Harry Markowitz. 1952. Portfolio selection. *The journal of finance* 7, 1 (1952), 77–91.
- [25] Antonio Martini and Jan Bosch. 2016. An empirically developed method to aid decisions on architectural technical debt refactoring: AnaConDebt. In *Software Engineering Companion (ICSE-C), IEEE/ACM International Conference on*. IEEE, 31–40.
- [26] Loup Meurice and Anthony Cleve. 2014. Dahlia: A visual analyzer of database schema evolution. In *Software Maintenance, Reengineering and Reverse Engineering (CSMR-WCRE), 2014 Software Evolution Week-IEEE Conference on*. IEEE, 464–468.
- [27] Bendra Ojameruaye and Rami Bahsoon. 2014. Systematic elaboration of compliance requirements using compliance debt and portfolio theory. In *International Working Conference on Requirements Engineering: Foundation for Software Quality*. Springer, 152–167.
- [28] Ken Pugh. 2010. The Risks of Acceptance Test Debt. *Cutter Business Technology Journal* (2010).
- [29] Pramod J Sadalage and Scott W Ambler. 2006. Refactoring databases: evolutionary database design. (2006).
- [30] Carolyn Seaman, Yuepu Guo, Clemente Izurieta, Yuanfang Cai, Nico Zazworka, Forrest Shull, and Antonio Vetrò. 2012. Using technical debt data in decision making: Potential decision approaches. In *Proceedings of the Third International Workshop on Managing Technical Debt*. IEEE Press, 43–48.
- [31] Marion G Sobol, Albert Kagan, and Hirohisa Shimura. 1996. Performance criteria for relational databases in different normal forms. *Journal of Systems and Software* 34, 1 (1996), 31–42.
- [32] Jurgita Tonkunaite, Lina Nemuraitė, and Bronius Paradauskas. 2006. Model driven development of data warehouses. In *Databases and Information Systems, 2006 7th International Baltic Conference on*. IEEE, 106–113.
- [33] Jens H Weber, Anthony Cleve, Loup Meurice, and Francisco Javier Bermudez Ruiz. 2014. Managing technical debt in database schemas of critical software. In *Managing Technical Debt (MTD), 2014 Sixth International Workshop on*. IEEE, 43–46.
- [34] Margaret S Wu. 1992. The practical need for fourth normal form. In *ACM SIGCSE Bulletin*, Vol. 24. ACM, 19–23.
- [35] Nico Zazworka, Carolyn Seaman, and Forrest Shull. 2011. Prioritizing design debt investment opportunities. In *Proceedings of the 2nd Workshop on Managing Technical Debt*. ACM, 39–42.