МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ИМЕНИ Н.Э. БАУМАНА (НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ)

«Метод автоматической нормализации для реляционных баз данных с использованием анализа функциональных зависимостей»

Студент: Зуев Тимофей Александрович

Группа: ИУ7-85Б

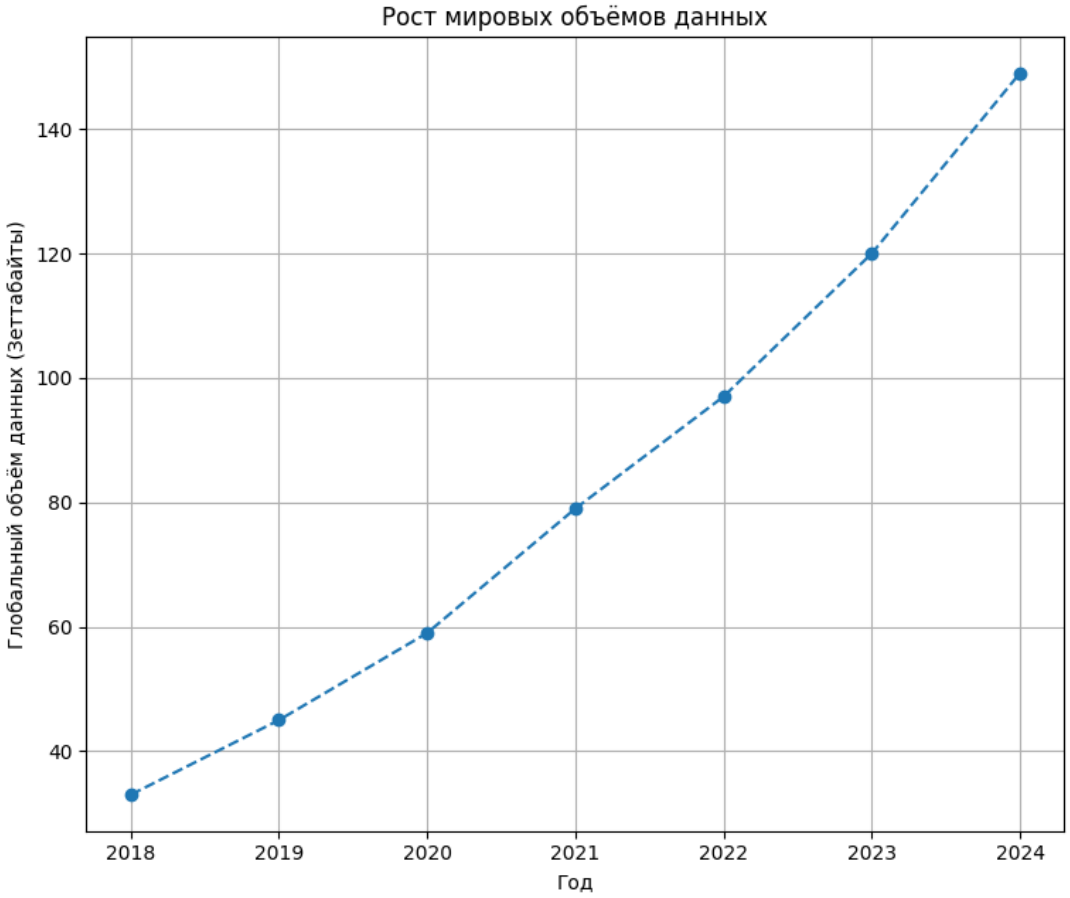
Руководитель: Исаев Андрей Львович

Актуальность

Нормальными формами называются уровни структурной организации отношений, определённые набором строгих требований к функциональным зависимостям, позволяющие устранить избыточность и аномалии обновления.

Нормализация отношений — важный процесс в проектировании баз данных, который преследует две основные цели:

1. Минимизация объема хранимых данных
2. Максимизация согласованности данных



Цель и задачи

Целью работы является разработка метода нормализации в реляционных базах данных с использованием анализа функциональных зависимостей.

Задачи:

- Проанализировать предметную область реляционных баз данных;
- Спроектировать метод автоматической нормализации в реляционных базах данных;
- Разработать спроектированный метод;
- Исследовать зависимость времени анализа и декомпозиции отношений от количества атрибутов и функциональных зависимостей.

Структура реляционных баз данных

Заголовком отношения называют множество
$U = \{A_1, A_2, \dots, A_n\}$, где A_i - имя i-ого атрибута.

Отношением называют пару вида $R(U) = (U, r)$, где

- $r = \{t_1, t_2, \dots, t_m\} \subseteq dom(A_1) \times dom(A_2) \times \dots \times dom(A_n)$, где
  - $t = (v_1, v_2, \dots, v_n), v_i \in dom(A_i)$, – кортеж из n значений
  - m — количество кортежей
- $U = \{A_1, A_2, \dots, A_n\}$ — заголовок отношения

Пусть существует отношение:
Экзамены (
Студент,
Дисциплина,
Преподаватель,
Оценка):

Студент	Дисциплина	Преподаватель	Оценка
Иванов И. И	Базы данных	Марьина Е. А.	5
Иванов И. И	Теория вероятностей	Земцова А. Ю.	4
Петров П. П	Базы данных	Марьина Е. А.	4
Смирнов А. А	Теория вероятностей	Земцова А. Ю	3

Функциональные зависимости

Пусть $X,\ Y \in U$ — непустые подмножества заголовка U. Говорят, что между множествами атрибутов X и Y существует **функциональная зависимость (ФЗ)**, когда:

$$\forall t_1, t_2 \in r\text{: если } t_1[X] = t_2[X], \text{то } t_1[Y] = t_2[Y],$$

где $t[X]$ — проекция кортежа t на подмножество атрибутов X.

Функциональную зависимость X от Y обозначают как $X \to Y$.

Пусть существует отношение Экзамены (Студент, Дисциплина, Преподаватель, оценка):

Студент	Дисциплина	Преподаватель	Оценка
Иванов И. И	Базы данных	Марьина Е. А.	5
Иванов И. И	Теория вероятностей	Земцова А. Ю.	4
Петров П. П	Базы данных	Марьина Е. А.	4
Смирнов А. А	Теория вероятностей	Земцова А. Ю	3

Тогда в нем выполняются ФЗ:

- $\{\text{Студент},\ \text{Дисциплина}\} \to \{\text{Оценка}\}$
- $\{\text{Дисциплина}\} \to \{\text{Преподаватель}\}$

Проблема избыточности данных

Аномалией называют любое нарушение целостности или согласованности базы данных, при котором операции изменения содержания отношений приводят к избыточному хранению или потере информации.

Аномалия обновления

	Студент	Дисциплина	Преподаватель	Оценка
1	Иванов И. И	Базы данных	Марьина Е. А.	5
2	Иванов И. И	Теория вероятностей	Земцова А. Ю.	4
3	Петров П. П	Базы данных	Марьина Е. А.	4

Обновить на "Карпова З. А"

Если обновить значение поля «Преподаватель» только в строке 3, возникнет рассогласованность – два преподавателя на один предмет.

Аномалия вставки

	Студент	Дисциплина	Преподаватель	Оценка
1	Иванов И. И	Базы данных	Марьина Е. А.	5
2	Иванов И. И	Теория вероятностей	Земцова А. Ю.	4
3	Петров П. П	Базы данных	Марьина Е. А.	4

Добавить дисциплину "Анализ алгоритмов"

Нельзя добавить дисциплину без сдающих ее студентов

Аномалия удаления

	Студент	Дисциплина	Преподаватель	Оценка
1	Иванов И. И	Теория вероятностей	Земцова А. Ю.	4
2	Петров П. П	Базы данных	Марьина Е. А.	4
3	Иванов И. И	Базы данных	Марьина Е. А.	5

Удалить информацию о результатах по "Базам данных"

Если удалить все сдачи студентов по «Базам данных», потеряется связь с преподавателем по этой дисциплине.

Решение для всех трех случаев – **декомпозиция** отношения на два новых: Курс(Дисциплина, Преподаватель) и Экзамен(Студент, Дисциплина, Оценка).

Декомпозиция без потерь

Декомпозицией отношения $R(U)$ называют разложение его схемы и множества кортежей на два или более подотношений $R_1(U_1),\ R_2(U_2), \ldots ,\ R_k(U_k)$, таких что $U = U_1 \cup\ U_2 \cup\ \ldots \cup\ U_k$.

Разложение R на R_1 и R_2 называется **декомпозицией без потерь**, если при естественном соединении R_1 и R_2 восстанавливается исходное отношение R без появления лишних или утраченных кортежей. В ином случае, разложение называется **декомпозицией с потерями**.

Декомпозиция без потерь

Студент	Дисциплина	Оценка
Иванов И. И	Базы данных	5
Иванов И. И	Теория вероятностей	4
Петров П. П	Базы данных	4
Смирнов А. А	Теория вероятностей	3

Дисциплина	Преподаватель
Базы данных	Марьина Е. А.
Теория вероятностей	Земцова А. Ю.

Естественное соединение $R_1 \bowtie R_2$ по атрибуту «Дисциплина» состоит из 4-х исходных кортежей.

$$\boldsymbol{R_1 \bowtie R_2 = R}$$

Декомпозиция с потерями

Студент	Преподаватель
Иванов И. И	Марьина Е. А.
Иванов И. И	Земцова А. Ю.
Петров П. П	Марьина Е. А.
Смирнов А. А	Земцова А. Ю

Дисциплина	Оценка
Базы данных	5
Теория вероятностей	4
Базы данных	4
Теория вероятностей	3

Естественное соединение $R_1 \bowtie R_2$ дает декартово произведение из 16 кортежей, из которых только 4 – исходные.

$$\boldsymbol{R_1 \bowtie R_2 \neq R}$$

Нормальные формы

- **Первая нормальная форма (1НФ):** каждое значение атрибута неделимо (атомарно).
- **Вторая нормальная форма (2НФ):** отношение находится в 1НФ и каждый неключевой атрибут полностью зависит от всего составного ключа, а не от его части.
- **Третья нормальная форма (3НФ):** отношение в 2НФ и каждый неключевой атрибут зависит непосредственно от ключа, без транзитивных зависимостей через другие неключевые атрибуты.
- **Нормальная форма Бойса–Кодда (BCNF):** для любой функциональной зависимости $X \to Y$ в отношении X является суперключом.
- **Четвёртая нормальная форма (4НФ):** отношение в BCNF и не содержит ненулевых многозначных зависимостей.

Студент	Курсы_и_оценки	Руководитель	Факультет
Иванов И. И	Базы данных:5, Экономика:4	Марьина Е. А.	ИУ
Петров П. П	Математика:5	Марьина Е. А.	ФН
Смирнов А. А	Базы данных:4, Философия:3	Земцова А. Ю	ИУ

Приведем к 1НФ, сделав все записи атомарными:

Студент	Курс	Оценка	Руководитель	Факультет
Иванов И. И	Базы данных	5	Марьина Е. А.	ИУ
Иванов И. И	Экономика	4	Марьина Е. А.	ИУ
Петров П. П	Математика	5	Марьина Е. А.	ФН
Смирнов А. А	Базы данных	4	Земцова А. Ю	ИУ
Смирнов А. А	Философия	3	Земцова А. Ю	ИУ

Ключ отношения – (Студент, Курс). Это отношение нарушает 2НФ, так как есть частичная зависимость: {Руководитель, Факультет} зависит только от «Студент», а не от всего ключа. Приведем к 2НФ:

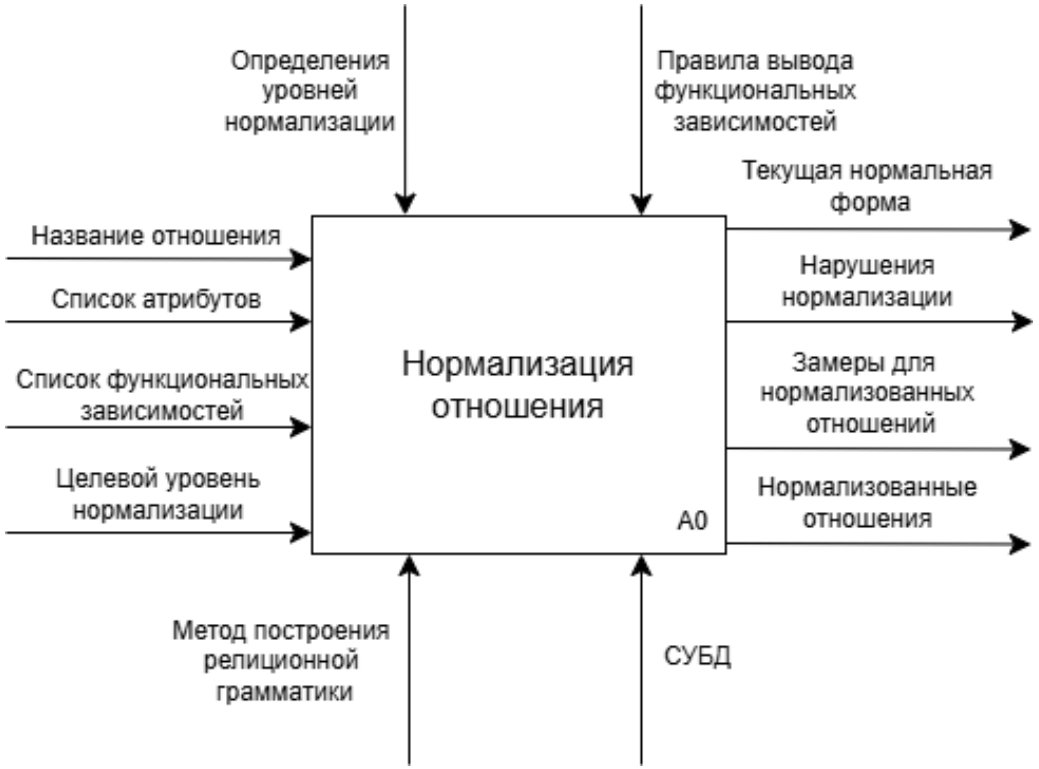
Студент	Курс	О-а
Иванов И. И	Базы данных	5
Иванов И. И	Экономика	4
Петров П. П	Математика	5
Смирнов А. А	Базы данных	4
Смирнов А. А	Философия	3

Студент	Р-ель	Ф-т
Иванов И. И	Марьина Е. А.	ИУ
Иванов И. И	Марьина Е. А.	ИУ
Петров П. П	Марьина Е. А.	ФН
Смирнов А. А	Земцова А. Ю	ИУ
Смирнов А. А	Земцова А. Ю	ИУ

Сравнение существующих методов нормализации отношений

Решение	Метод	Максимальная НФ	Графический интерфейс	Явное задание ФЗ
Micro	Графовый анализ	НФБК	+	+
RDBNorma	Графовый анализ	3НФ	+	+
JMathNorm	Грамматика для реляционных операций	НФБК	+	+
Genetic algorithm for decomposing relational databases	Эволюционная оптимизация декомпозиции	3НФ	-	-
Анализатор И. А. Зорина	Графовый анализ	3НФ	+	+

Формализованная постановка задачи



Автоматическую нормализацию можно формализовать как задачу оптимизации по двум критериям:

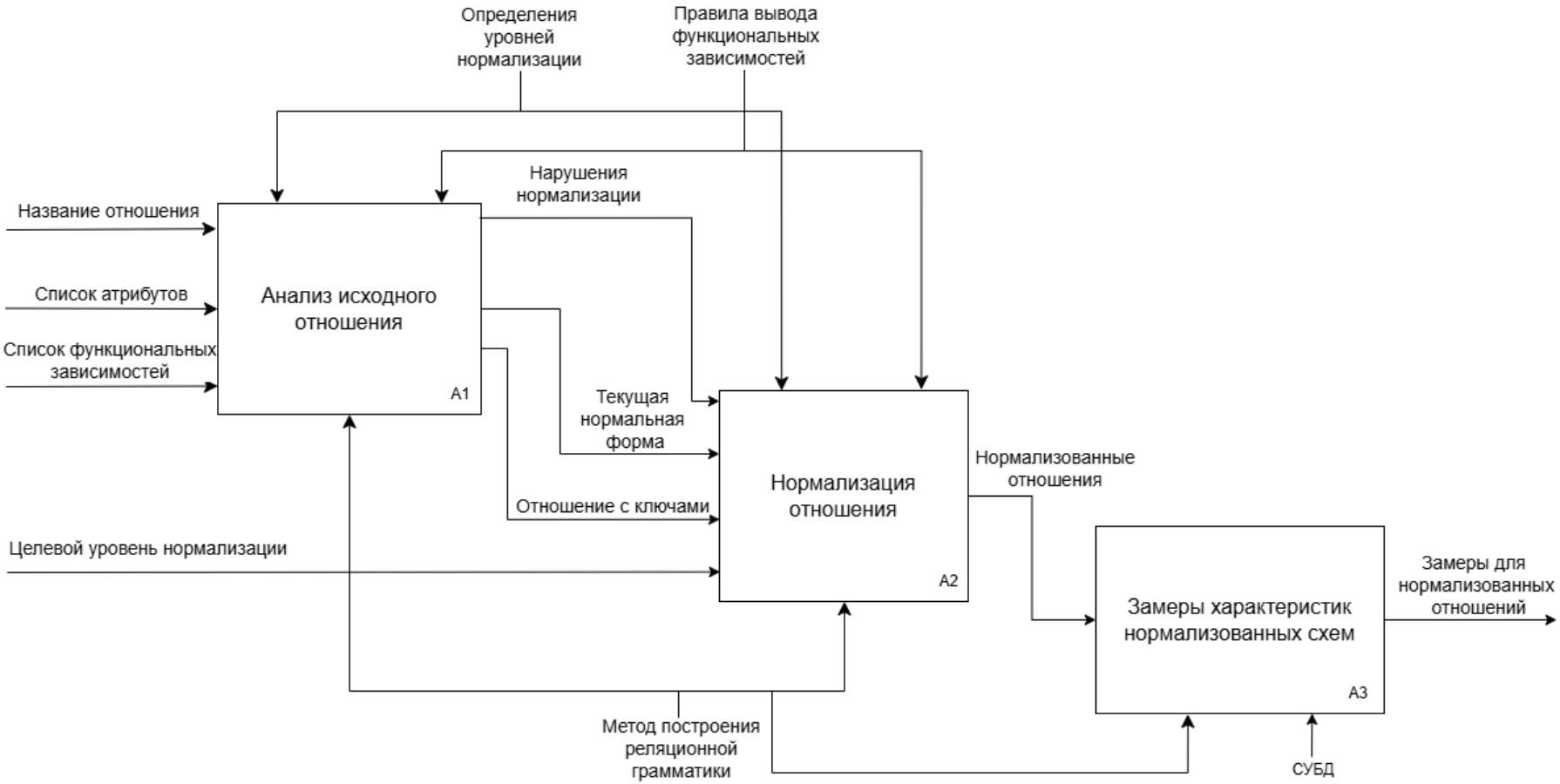
$$V = \sum_{i=1}^{k} |R_i| \rightarrow min\ ;$$

$$A = Count\_anomalies(R_1, \dots, R_k) \rightarrow min,$$

где:

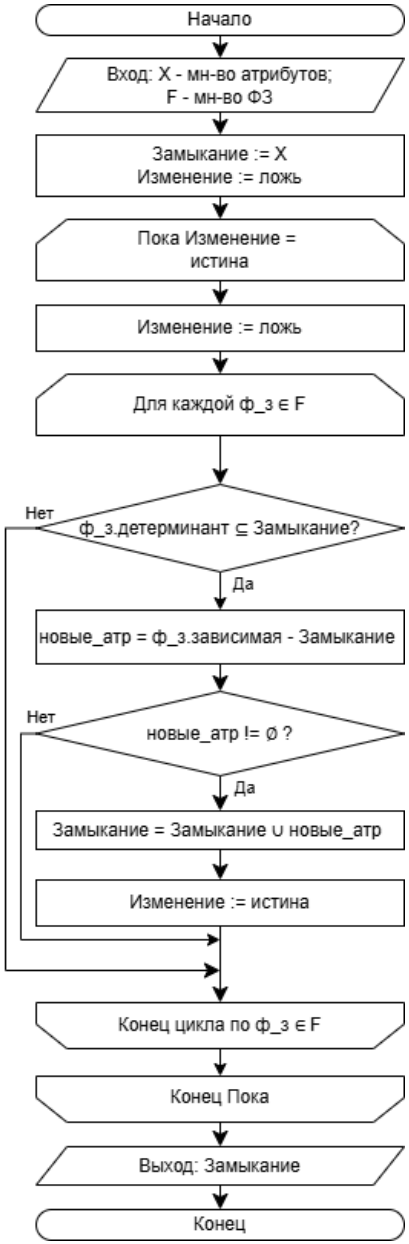
- k – количество декомпозированных отношений
- $Count\_anomailes(R_1, \dots, R_k)$ – функция, вычисляющая количество аномалий всех типов.

Метод автоматической нормализации

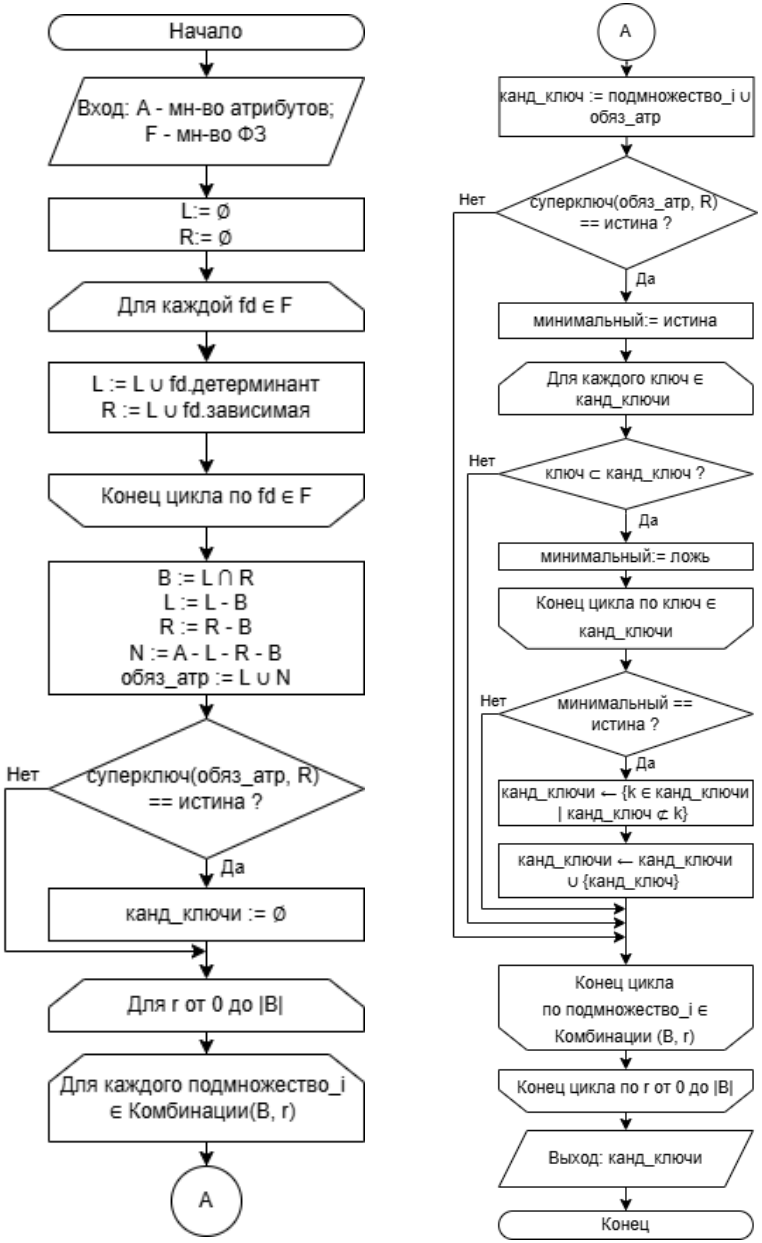


Анализ исходного отношения

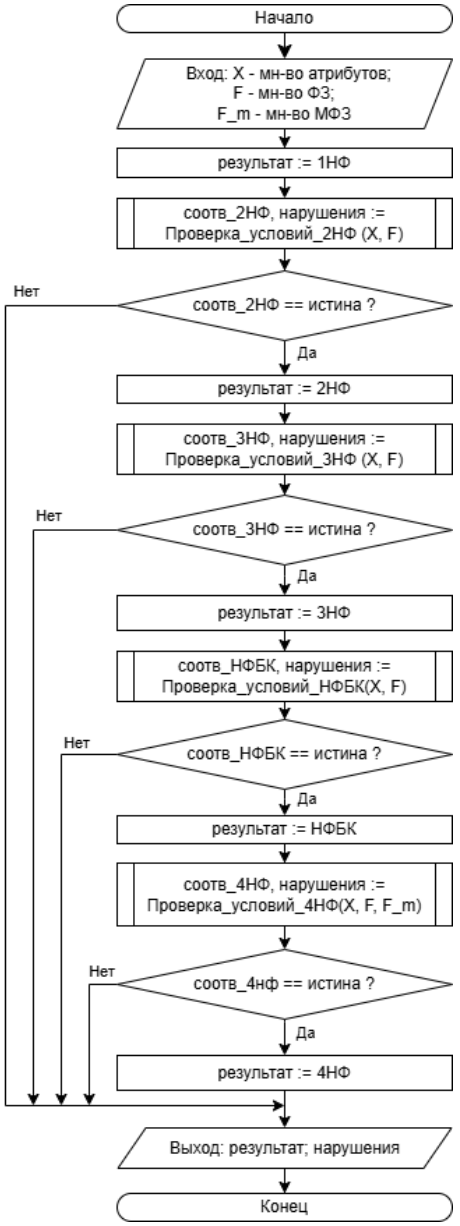
Алгоритм поиска замыкания



Алгоритм поиска кандидатных ключей



Общий алгоритм анализа



Декомпозиция отношения

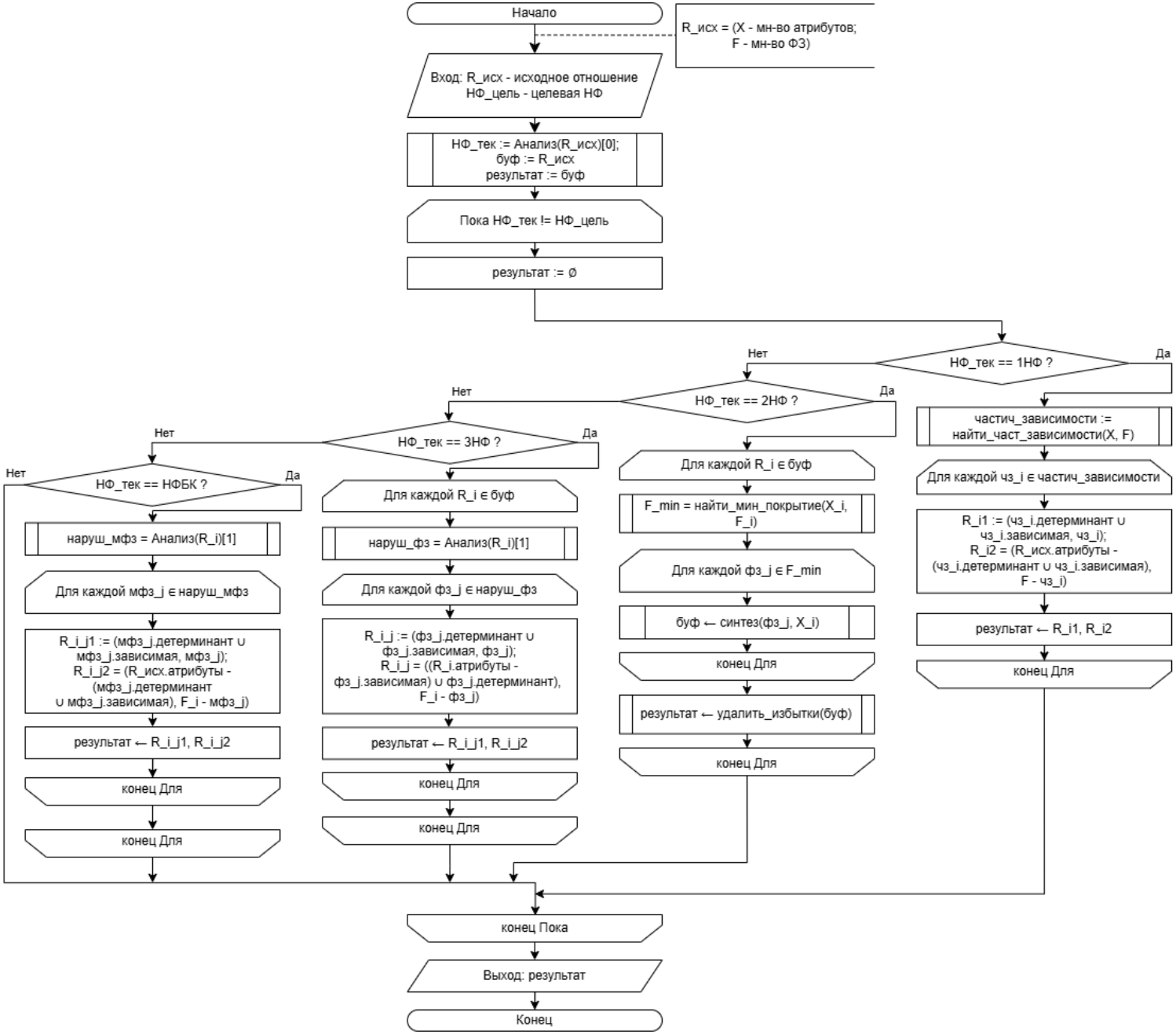
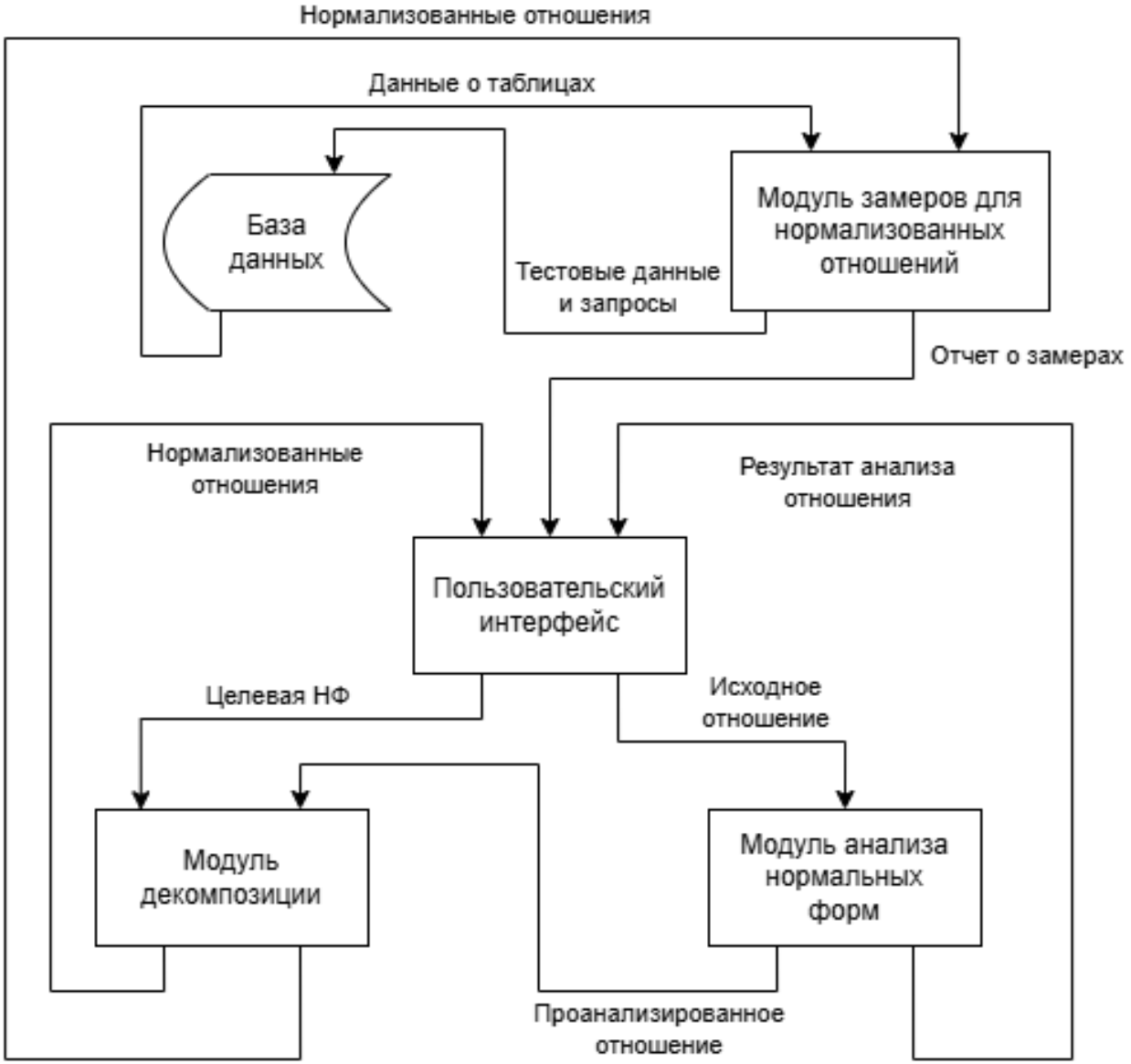
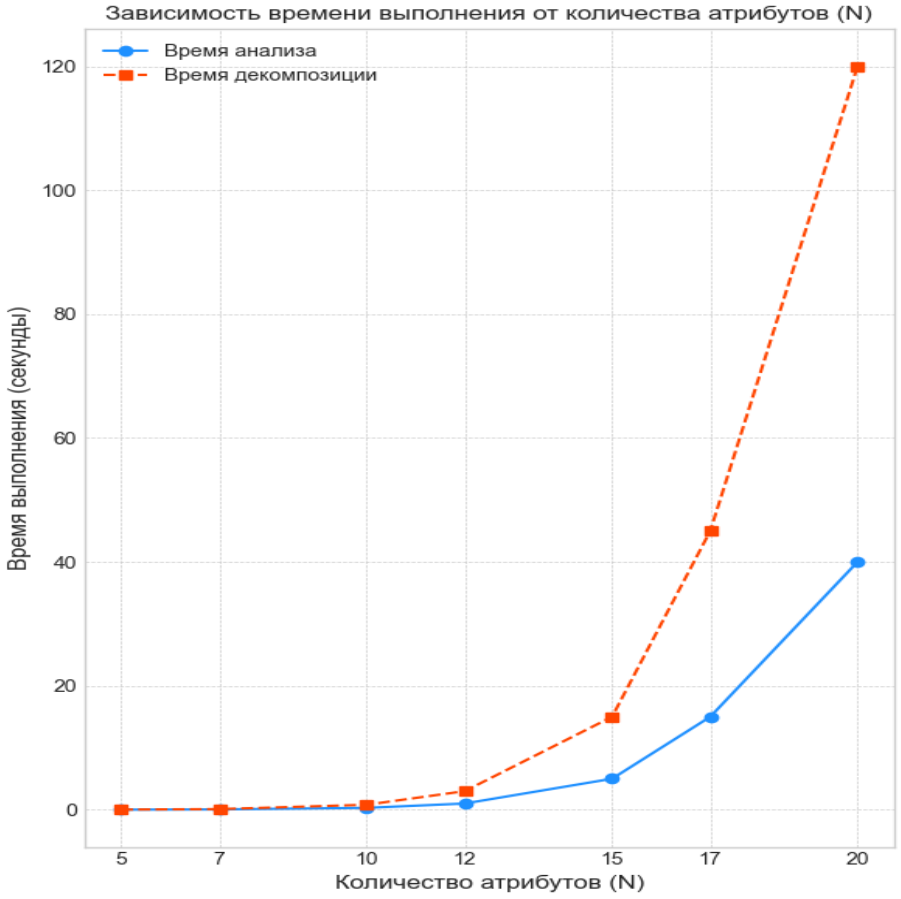
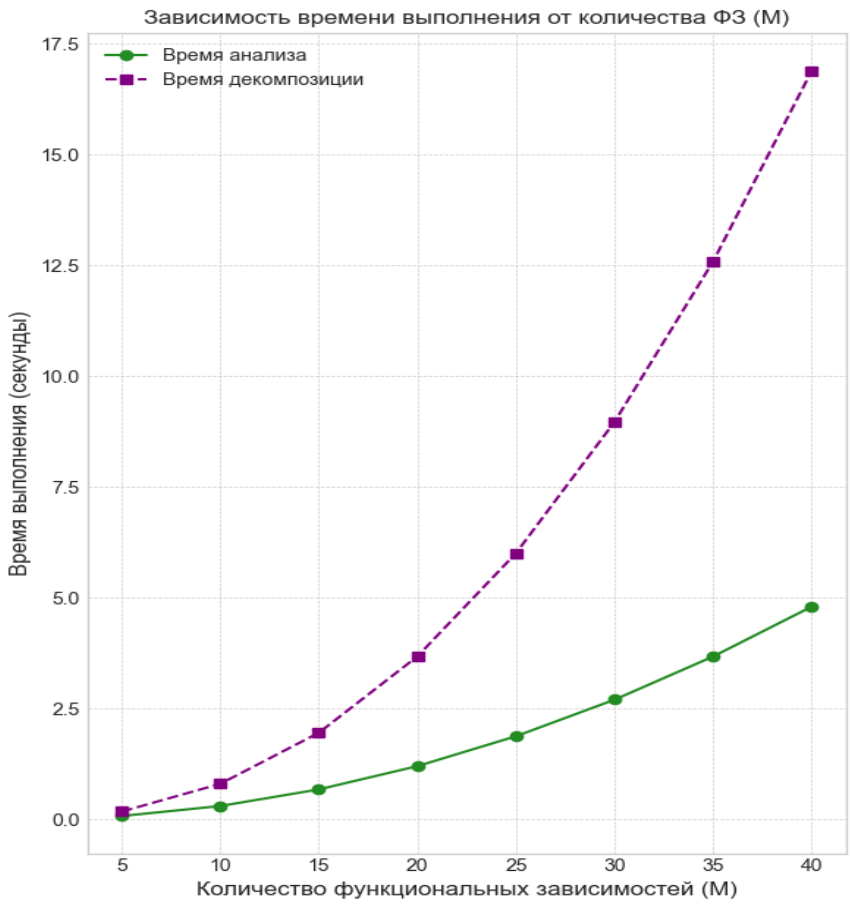


Схема программного обеспечения



Зависимость временных характеристик метода от количества атрибутов отношения и от количества функциональных зависимостей





Время выполнения анализа и декомпозиции отношений растет как при росте количества атрибутов, так и при росте количества функциональных зависимостей. При этом при росте количества атрибутов время выполнения растет быстрее.

Постановка исследования

Схема отношения «Сотрудники_компании»:

$$U = \left\{ \begin{array}{c} \text{КодСотрудника}, \\ \text{ИмяСотрудника}, \\ \text{Отдел}, \\ \text{НачальникОтдела}, \\ \text{Код проекта} \\ \text{НазваниеПроекта} \\ \text{БюджетПроекта} \end{array} \right\}.$$

Функциональные зависимости:

1. $\{\text{КодСотрудника}\} \rightarrow \{\text{ИмяСотрудника}, \text{Отдел}\}$;
2. $\{\text{Отдел}\} \rightarrow \{\text{НачальникОтдела}\}$.
3. $\{\text{КодПроекта}\} \rightarrow \{\text{Бюджет}, \text{НазваниеПроекта}\}$.

Исходные характеристики:

- Исходная нормальная форма – 1НФ;
- Количество строк – 3000;
- Объем исходного отношения – 0.469 МБ,

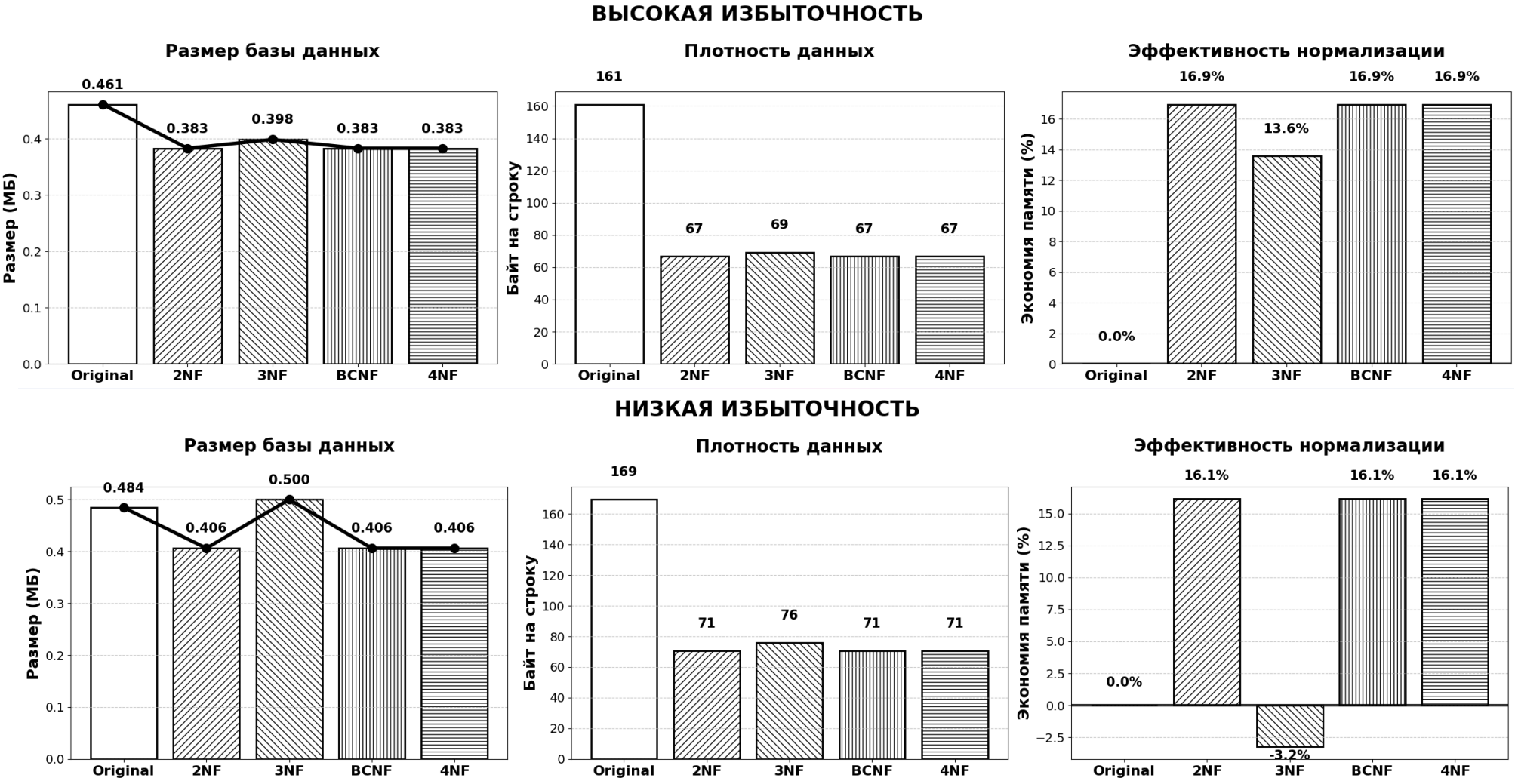
Заполнение с высокой избыточностью:

- Уникальных отделов: 5
- Уникальных проектов: 10
- Уникальных начальников: 5
- Уникальных имен сотрудников: 200

Заполнение с низкой избыточностью:

- Уникальных отделов: 30
- Уникальных проектов: 28
- Уникальных начальников: 30
- Уникальных имен сотрудников: 500

Исследование зависимости занимаемой таблицами памяти от уровня нормализации



Вывод: для данного отношения нормализация дает существенное уменьшение используемой памяти, причем для более высокой избыточности данных эффективность нормализации выше

Заключение

В ходе выполнения выпускной квалификационной работы была достигнута цель разработки метода автоматической нормализации в реляционных базах данных, а также были успешно выполнены все поставленные задачи:

- Проанализирована предметная область реляционных баз данных;
- Спроектирован метод автоматической нормализации в реляционных базах данных;
- Разработан спроектированный метод;
- Исследована зависимость времени анализа и декомпозиции отношений от количества атрибутов и функциональных зависимостей.

Направления дальнейшего развития для разработанного метода могут включать:

- Реализацию обратной композиции отношений;
- Вычисление характеристик нормализованных отношений для различных типов запросов.