

Генератор анонимизированных баз данных

Воронежский Государственный Университет

Факультет Компьютерных Наук

Кафедра программирования и информационных технологий

Студент: Ячный М.А., специальность «Программная инженерия», 3 курс, 9 группа

Научный руководитель: старший преподаватель Самойлов Н.К.

Введение



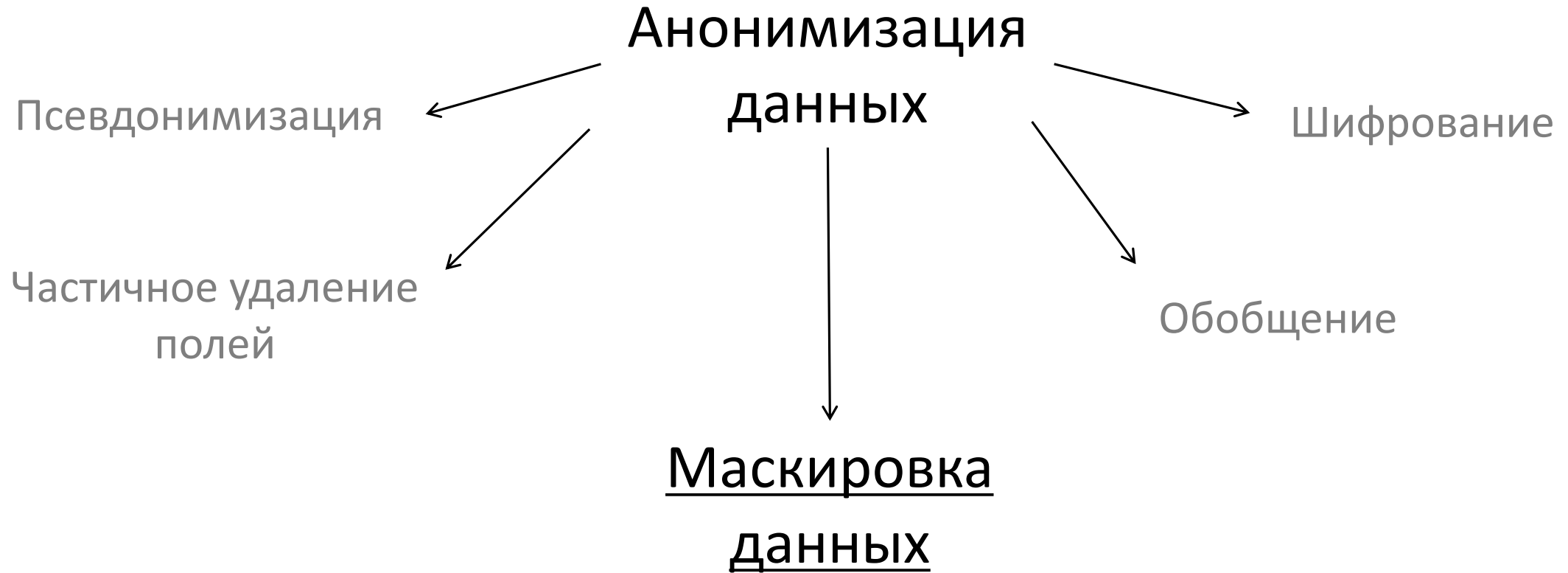
Постановка задачи

1. Система должна уметь работать со следующими типами данных: целые числа, числа с плавающей точкой, строки, даты, числовые интервалы, интервалы дат.
2. Система должна учитывать следующие типы ограничений целостности: первичный ключ, внешний ключ, уникальный ключ.
3. Система должна обеспечивать возможность использования в качестве типа данных первичного ключа целых чисел, UUID и ObjectID.
4. Система должна уметь работать с типами кардинальности связей между таблицами «Один к одному» и «Один ко многим».
5. Система должна обеспечивать возможность многопоточного (одновременного) заполнения нескольких таблиц

Используемые технологии

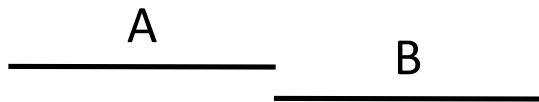


Способы анонимизации данных



Интервальная алгебра Аллена

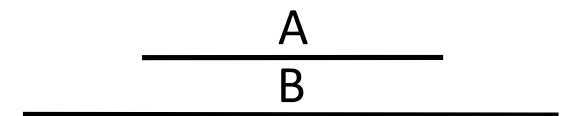
А встречается В
В встречен А



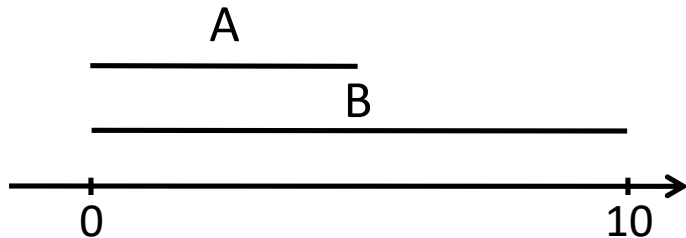
А перекрывает В
В перекрывается А



А во время В
В содержит А



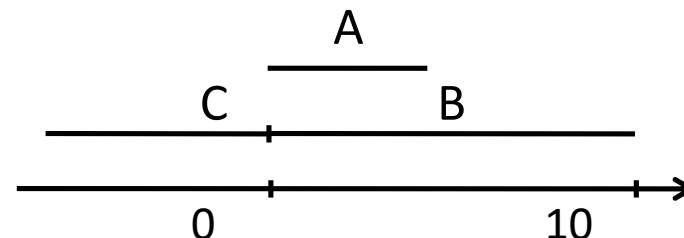
А начинает В



=

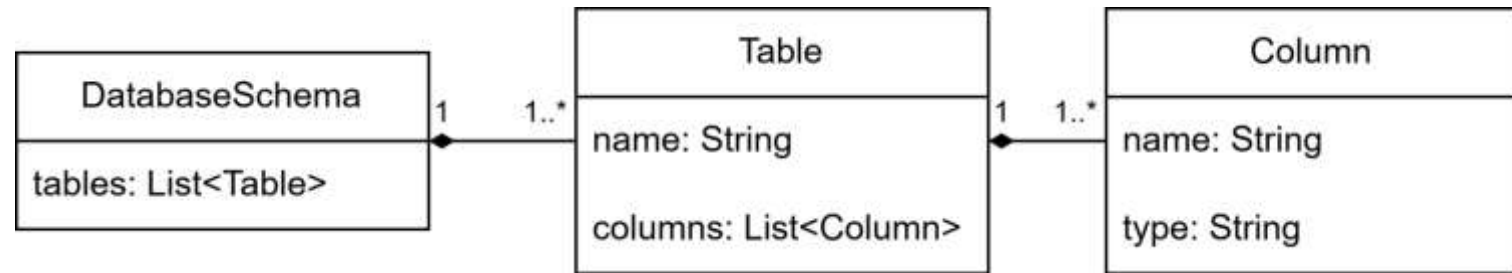
А встречен С
И

А во время В

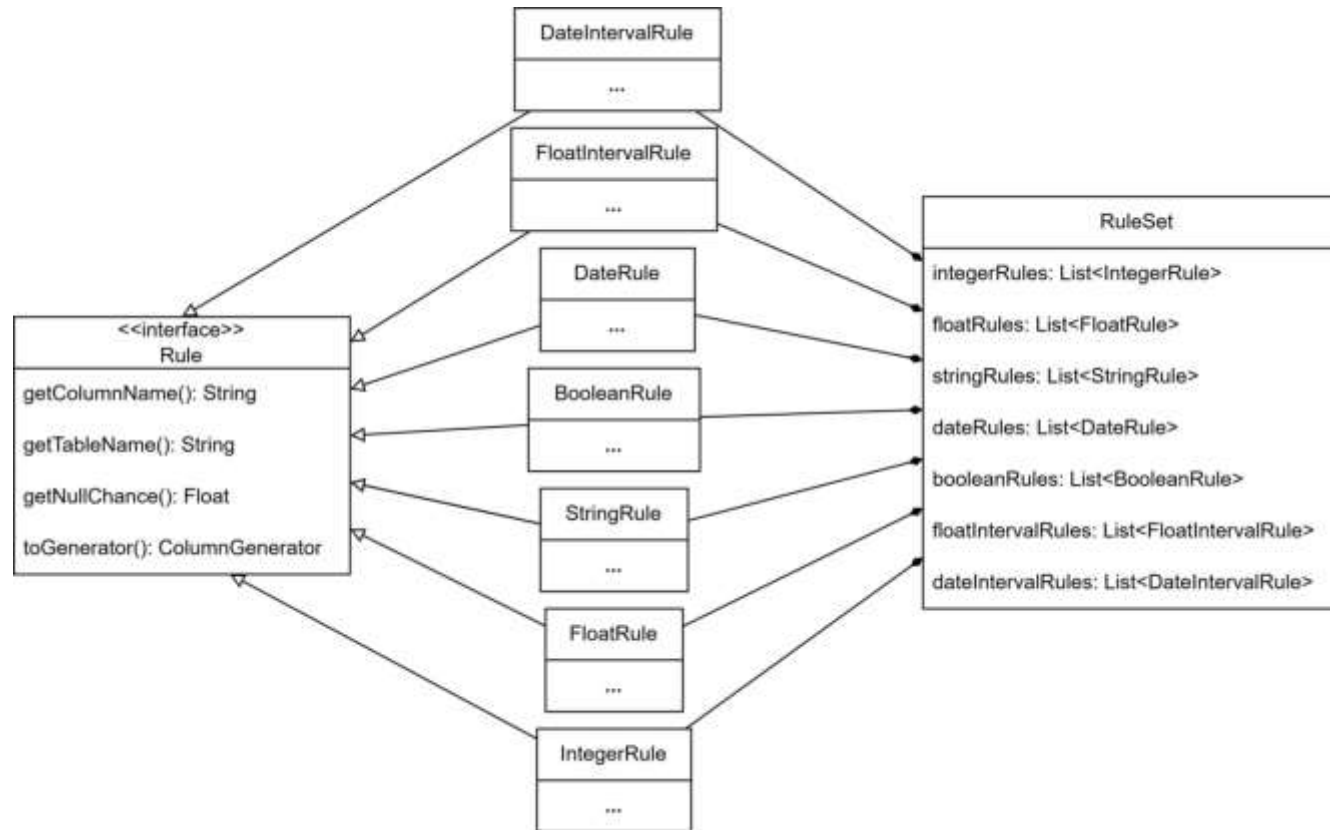


Описание схемы БД

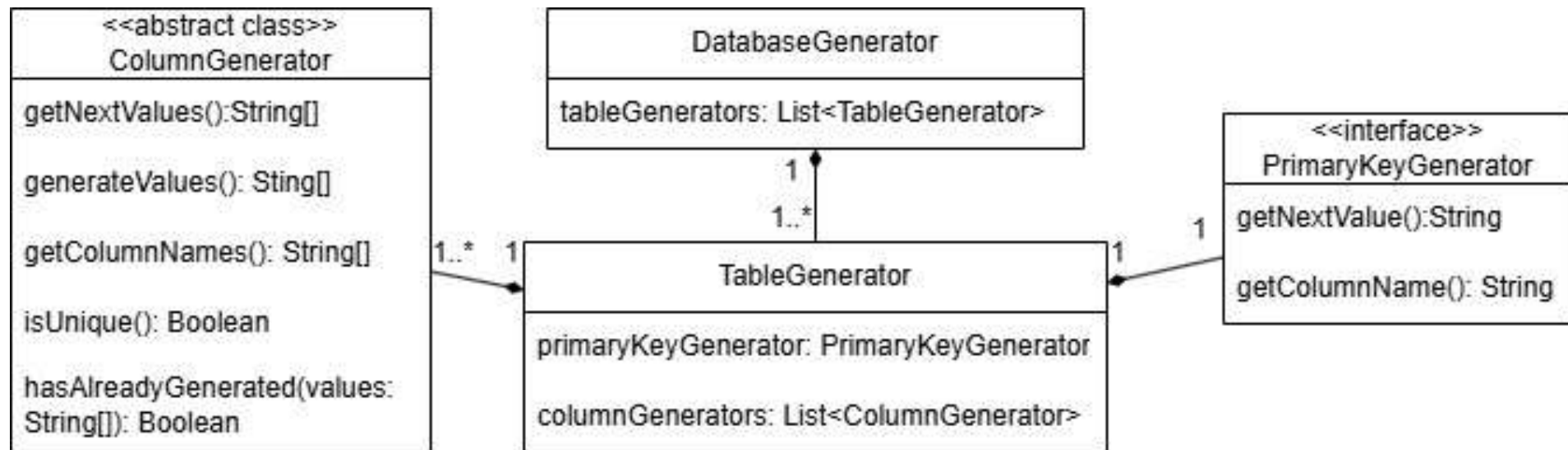
```
{  
  "name": "users",  
  "columns": [  
    {  
      "name": "id",  
      "type": "BIGINT"  
    },  
    {  
      "name": "name",  
      "type": "VARCHAR"  
    },  
    {  
      "name": "email",  
      "type": "VARCHAR"  
    },  
    {  
      "name": "age",  
      "type": "INTEGER"  
    }  
  ]  
},
```



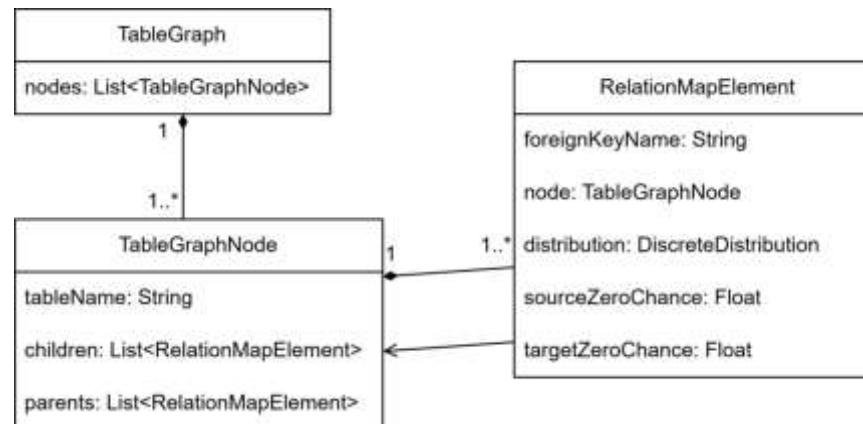
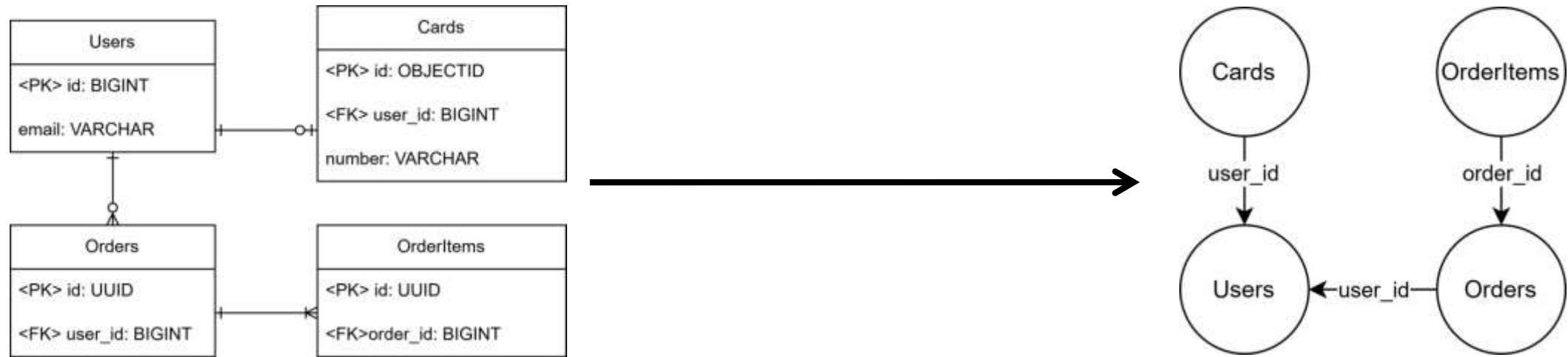
Описание правил генерации



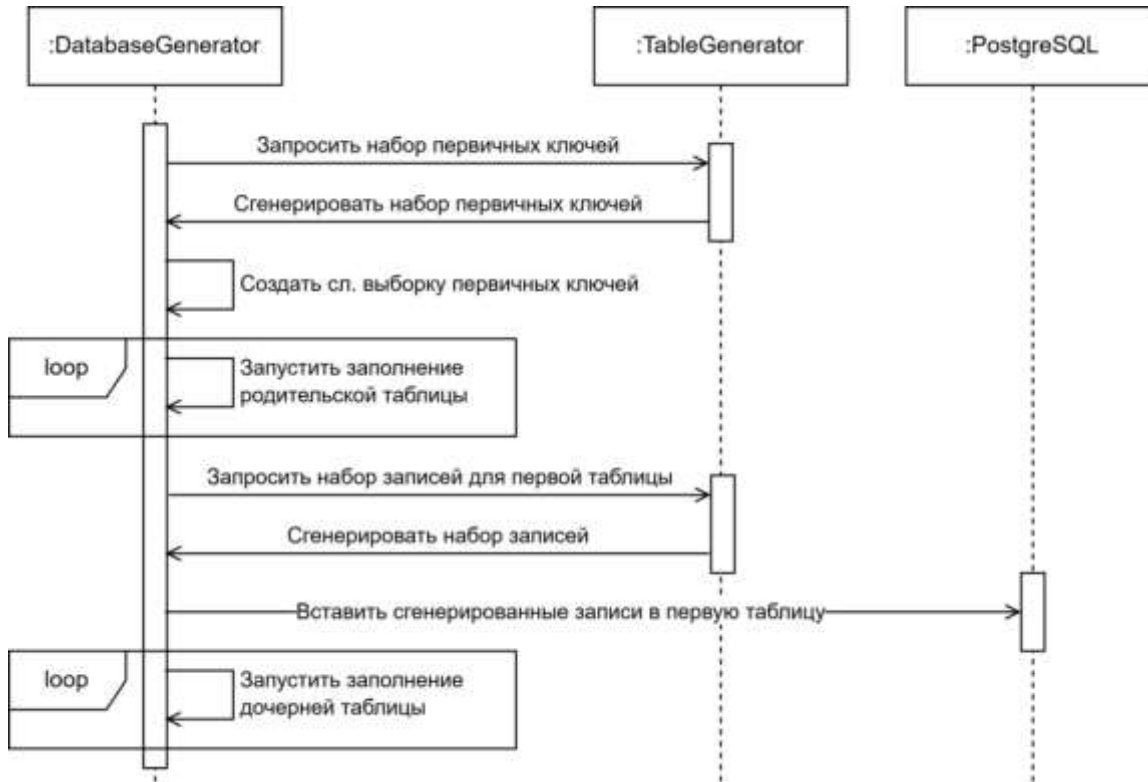
Генераторы таблиц и колонок



Граф отношения таблиц



Заполнение первой таблицы



1. Создание набора первичных ключей



2. Создание случайной выборки из них



3. Запуск заполнения родительских таблиц



4. Заполнение текущей таблицы



5. Запуск заполнения дочерних таблиц



Заполнение родительских таблиц

Дочерняя таблица

[Первичный ключ]
590c7fed-4d25-492a-9891		
d9066e12-f4b7-4db2-972e		
d932a1f8-6126-41b6-879c		
...		

[Первичный ключ]	...	[Внешний ключ]
1		590c7fed-4d25-492a-9891
2		d9066e12-f4b7-4db2-972e
3		d932a1f8-6126-41b6-879c
...		...

Родительская таблица

[Первичный ключ]	...	[Внешний ключ]
1		590c7fed-4d25-492a-9891
2		d9066e12-f4b7-4db2-972e
3		d932a1f8-6126-41b6-879c
...		...

Родительская таблица

Заполнение дочерних таблиц

1.

[Первичный ключ]	...	[Внешний ключ]
1		null
2		null
3		null
...		...

[Первичный ключ]
590c7fed-4d25-492a-9891		
d9066e12-f4b7-4db2-972e		
d932a1f8-6126-41b6-879c		
...		

2.

[Первичный ключ]	...	[Внешний ключ]
1		590c7fed-4d25-492a-9891
2		d9066e12-f4b7-4db2-972e
3		d932a1f8-6126-41b6-879c
...		...

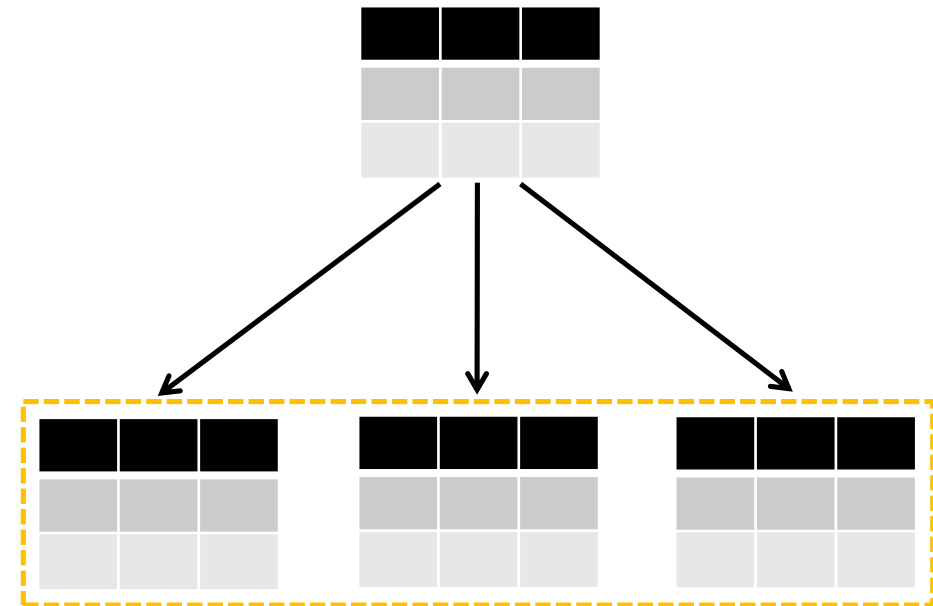
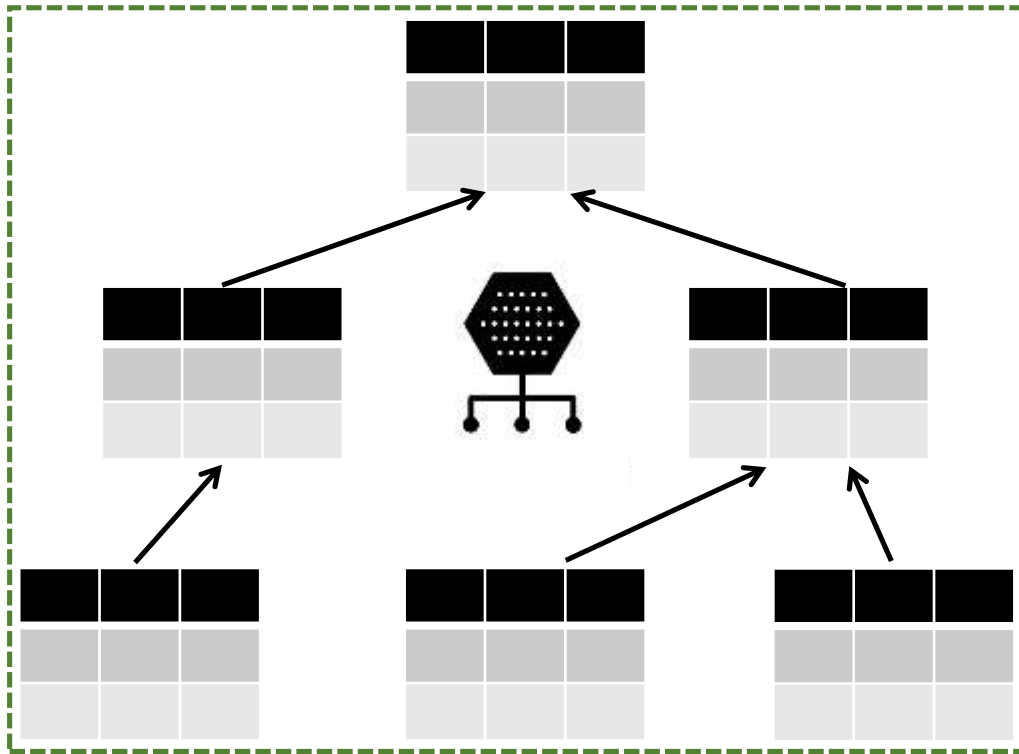
UPDATE

[Первичный ключ]
590c7fed-4d25-492a-9891		
d9066e12-f4b7-4db2-972e		
d932a1f8-6126-41b6-879c		
...		

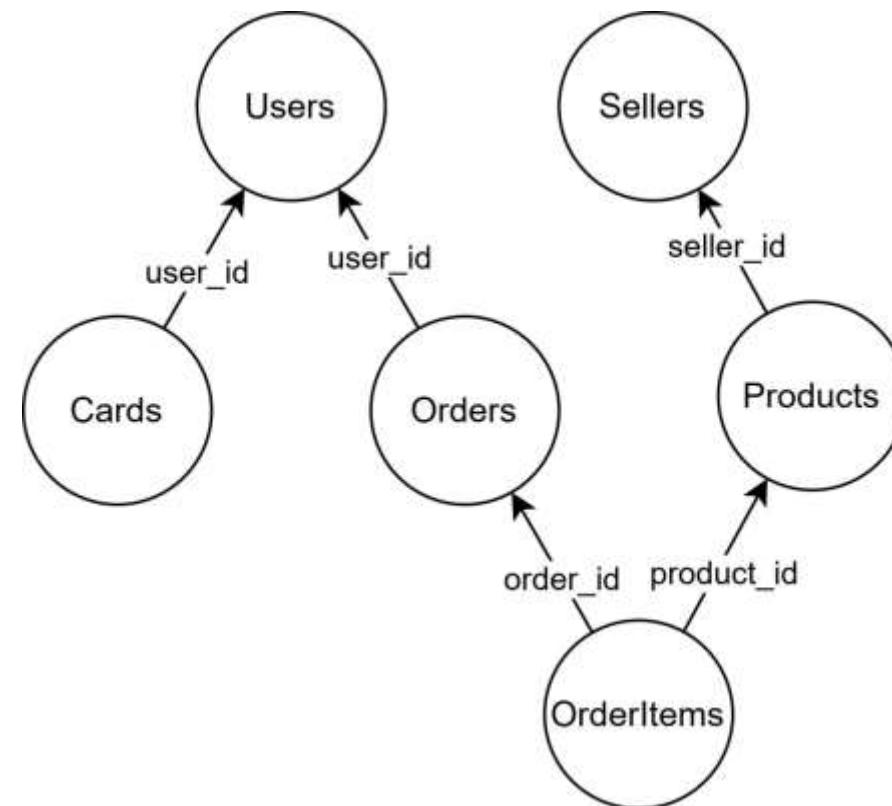
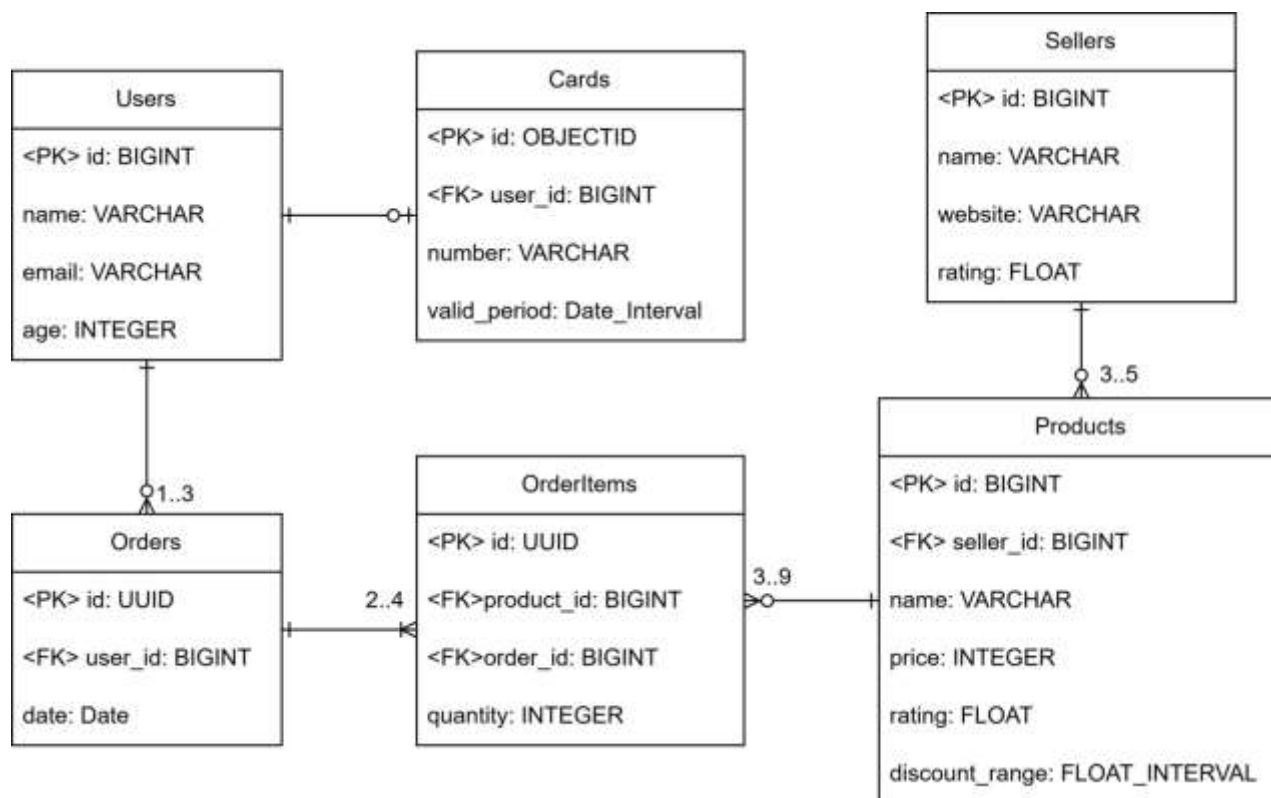
Родительская таблица

Дочерняя таблица

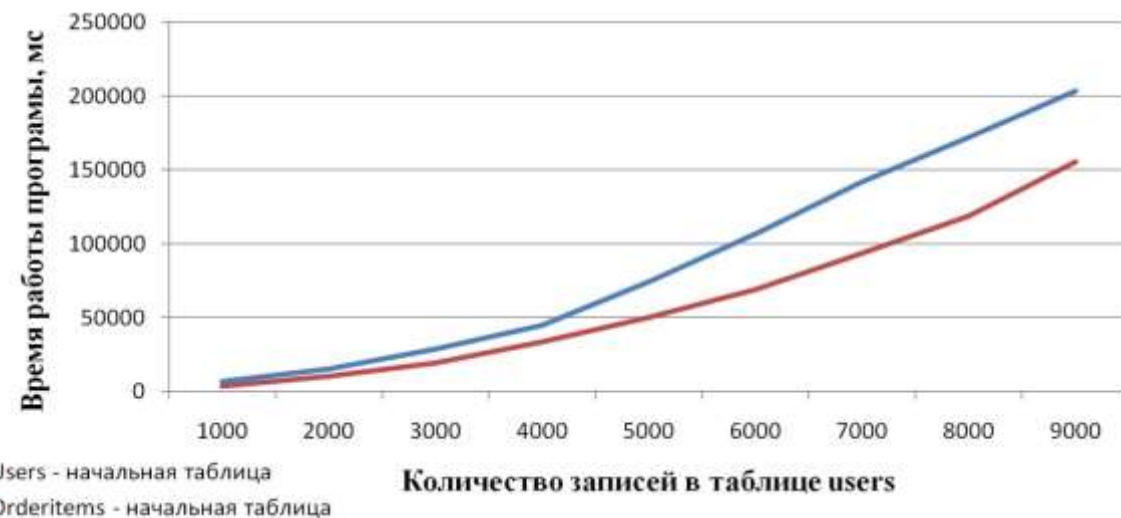
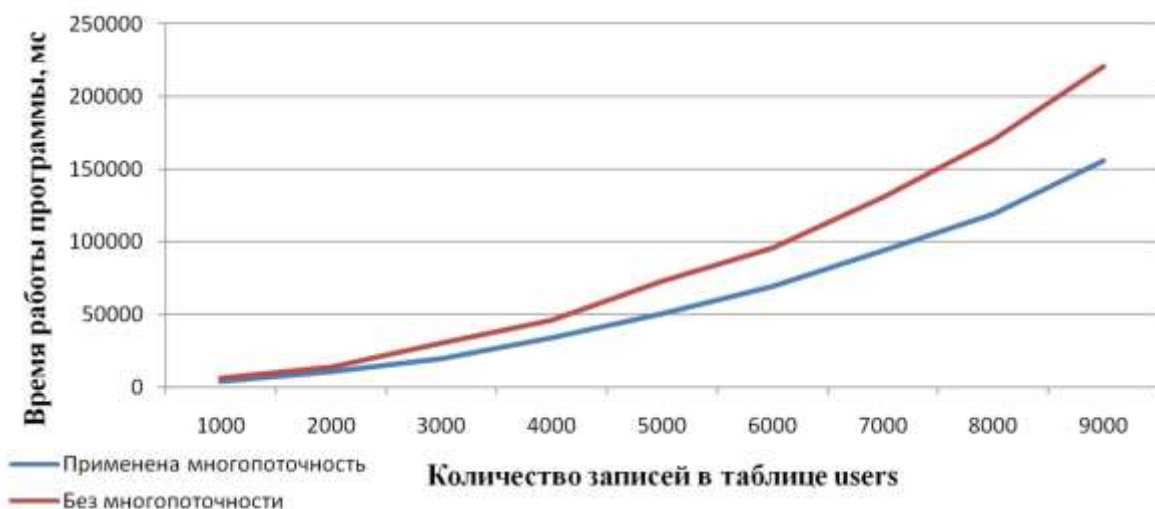
Применение многопоточности



Тестирование



Измерение производительности



Центральный процессор Intel Core i7-8700: 6 ядер, частота 3.2ГГц;

Оперативная память типа DDR4: частота 2.4МГц, объём 32Гб;

SSD Kingston sa400s37240g: скорость записи до 350Мб/с.

Заключение

В ходе выполнения работы были решены следующие задачи:

- Реализовано консольное приложение для генерации анонимизированных локальных баз данных
- Приложение умеет работать с основными типами данных
- Приложение умеет работать с основными ограничениями целостности
- Реализован метод генерации интервальных значений с помощью интервальной алгебры Аллена
- Обеспечена возможность многопоточного заполнения таблиц

Генератор анонимизированных баз данных

Воронежский Государственный Университет

Факультет Компьютерных Наук

Кафедра программирования и информационных технологий

Студент: Ячный М.А., специальность «Программная инженерия», 3 курс, 9 группа

Научный руководитель: старший преподаватель Самойлов Н.К.