

Введение

Объемы данных, собираемых и анализируемых людьми, постоянно растут. Кроме того, постоянно повышаются требования к скорости чтения/записи этих данных. Сами данные обычно хранятся в структурированном виде, обладая рядом характеристик, по которым необходимо осуществлять их выбор. К решению этой проблемы можно подходить несколькими способами: во-первых, масштабировать систему хранения данных — добавление вычислительных мощностей должно снижать время обработки и поиска. Однако, в реальности такое масштабирование не является выгодным — оно стоит денег: сначала покупка оборудования, затем его поддержка.

Глава 1. Введение

1.1 Постановка задачи

Задачи хранения и обработки большого числа данных встречаются повсеместно. При этом хочется делать это эффективно и быстро.

Представим себе торговую площадку, каждый товар обладает рядом характеристик непрерывных или дискретных, как например на рисунках 1.1.1, 1.1.2, 1.1.3.




	<p>Ноутбук Lenovo V130 15</p> <p>5.0</p> <p>Покупателям нравится долгое время работы, небольшой вес</p> <p>Процессор: Core i3 Объем жесткого диска: 500 ГБ Диагональ экрана: 15.6 " Видеокарта: Intel HD Graphics 520 Вес: 1.8 кг</p> <p>28 человек купили этот товар</p>	<p>Есть акции</p> <p>от 16 270 Р</p> <p>251 предложение от 16 270 Р</p>	<p>Показать всё</p> <p>Тип</p> <p><input type="checkbox"/> ноутбук <input type="checkbox"/> ноутбук-планшет <input type="checkbox"/> трансформер</p> <p>Размер экрана</p> <p><input type="checkbox"/> 11"-11.9" <input type="checkbox"/> 12"-12.9" <input type="checkbox"/> 13"-13.9" <input type="checkbox"/> 14"-14.9" <input type="checkbox"/> 15"-15.9" <input type="checkbox"/> 17" и более <input type="checkbox"/> до 11"</p> <p>Разрешение экрана</p> <p><input type="checkbox"/> 1920x1080 <input type="checkbox"/> 1366x768 <input type="checkbox"/> 1600x900 <input type="checkbox"/> 3840x2160 <input type="checkbox"/> 2560x1600</p> <p>Показать всё</p> <p>Процессор</p> <p><input type="checkbox"/> Core i5 <input type="checkbox"/> Core i7 <input type="checkbox"/> Core i3 <input type="checkbox"/> Pentium <input type="checkbox"/> Celeron</p> <p>Показать всё</p>
<p>Выбор покупателей</p> 	<p>Ноутбук Xiaomi Mi Notebook Air 13.3" 2018</p> <p>5.0 7 отзывов</p> <p>Покупателям нравится экран, мощный процессор</p> <p>Объем жесткого диска: 256 ГБ Диагональ экрана: 13.3 " Видеокарта: NVIDIA GeForce MX150 Вес: 1.3 кг Оптический привод: DVD нет</p> <p>410 человек купили этот товар</p>	<p>от 54 400 Р</p> <p>30 предложений от 54 400 Р</p>	
<p>Выбор покупателей</p> 	<p>Ноутбук Xiaomi Mi Gaming Laptop</p> <p>5.0 3 отзыва</p> <p>Покупателям нравится мощный процессор, экран</p> <p>Диагональ экрана: 15.6 " Видеокарта: NVIDIA GeForce GTX 1060 Вес: 2.7 кг Оптический привод: DVD нет 4G LTE: нет</p> <p>398 человек купили этот товар</p>	<p>от 69 000 Р</p> <p>18 предложений от 69 000 Р</p>	

Рисунок 1.1.1 — Продажа электроники

Правильно организованная работа с данными увеличивает скорость выполнения запросов и позволяет экономить на оборудовании. Для этого стоит выбрать правильную структуру данных для хранения. Далее будут рассмотрены типы индексов, используемые в современных СУБД и проведено их сравнение.

Легковые автомобили

Все Новые С пробегом Сохранить поиск

Марка Модель Поколение +

Кузов Коробка Двигатель Привод Объем от, л до

Год от до Пробег от, км до Цена от, ₽ до

Все параметры Показать 52 532 предложения

Audi	2675	Ford	2956	LADA (BA3)	3780	Opel	1989
BMW	4061	Hyundai	2276	Mercedes-Benz	4293	Toyota	2056
Chevrolet	1910	Kia	2339	Nissan	2693	Volkswagen	3432

до 50 000 ₽ до 100 000 ₽ до 150 000 ₽ до 200 000 ₽ до 250 000 ₽ до 300 000 ₽ до 350 000 ₽ до 400 000 ₽ < >

Рисунок 1.1.2 — Продажа автомобилей

1.2 Математическая модель

Сформулируем математическую модель, а также характеристики, на которых будет сделан акцент в данной работе.

В базовом случае интересующий нас объект — это хранилище данных 1.2.1, обрабатывающее пользовательские запросы. Запросы могут быть нескольких типов: создание, чтение, обновление и удаление записей. В зависимости от типа запроса запрашивающая сторона (ей может быть пользователь или какая-то другая система) должна получить некоторый результат: данные, удовлетворяющие условию запроса, если это чтение, и «подтверждение» факта, что данный запрос успешно выполнен. Важна также корректность результатов наших запросов — если наш запрос содержит некоторые условия, то при его работе не должны возвращаться, изменяться и удаляться записи, не удовлетворяющие данному условию. Также мы не должны иметь побочных эффектов в виде появления записей, не создаваемых напрямую с помощью запросов или не предусмотренных внутренней задокументированной логикой работы данной базы данных.

Конечно, обычно СУБД предлагают более широкий список возможностей: поддержку групп и ролей, разграничение доступа, создание и использование

Квартира

Комнаты

Цена

+ Еще фильтры

Найти

Город

Метро

Район

Шоссе

Все

Новостройка

Вторичка

Сохранить мой поиск

Опубликовано

За месяц

Цена

За всю площадь

Общая площадь (м²)

От

До

Площадь кухни (м²)

От

До

Жилая площадь (м²)

От

До

Материал здания

☐ Кирпич
☐ Блоки
☐ Монолит
☐ Панель
☐ Дерево

Год постройки

С

По

Дома под снос

Неважно

Этаж

От

До

Не последний

Этажей в доме

От

До

До метро (по карте)

-

Продавец

Любой

Агент/агентство

Собственник

Показать только

С фотографиями

Добавить ключевое слово

Например: окна во двор, консьерж, название ЖК или адрес

Исключить ключевое слово

Например: окна на улицу, требует ремонта или балкона нет

Рисунок 1.1.3 — Продажа недвижимости



Рисунок 1.2.1 — Схематическое представление системы хранения данных

пользовательских функций, но это не будет рассмотрено, поскольку не имеет непосредственного отношения к рассматриваемой области.

Не имеет особой разницы форма запроса и ответа на него — предполагаем, что существует некоторый протокол, который и используется при общении с данной СУБД при этом на коммуникацию между системами тратится сравнимое и меньшее время, чем на выполнение запроса.

Получаемые результаты могут достаточно сильно зависеть от платформы, на которой будут запускаться тесты. Более подробное описание среды тестирования, платформы и самих тестов будет приведено в следующих главах.

1.3 Оценка полученных результатов

Оценка полученных результатов — результат проведения нескольких тестов. Для начала, считаем, что одна запись — вектор, который содержит числа и строки. При этом осуществлять поиск планируется не по всем значениям, а только по их части. Тестирование должно проводиться для нескольких размеров записей: 10/100 байт, 1/10/100 килобайт, 1 мегабайт. Количество полей, по которым осуществляется поиск стоит взять следующими: 1 (для сравнения со стандартными решениями), 2, 3 (гео-данные), 5, 20, 50. На первом этапе стоит проводить вставку 1/10/100 миллионов записей и замерять количество вставок за единицу времени, а также зависимость этого времени от количества уже вставленных данных. Следующий этап — чтение. Для вышеописанных конфигураций измеряется количество чтений за одну секунду. Финальный этап наиболее приближен к реальной жизни: должны выполняться запросы обоих типов в следующем соотношении — 80/20.

Глава 2. Анализ предметной области

2.1 Типы индексов

B-Tree

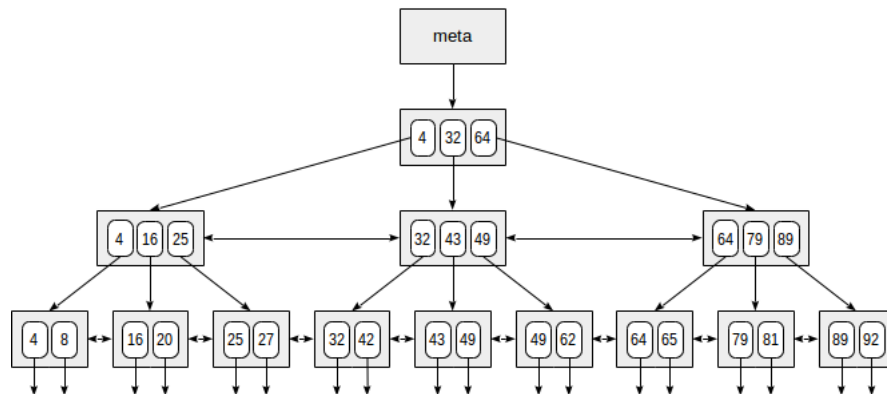


Рисунок 2.1.1 — Схематичный пример индекса по одному полю с целочисленными ключами

Для хранения данных в отсортированном виде используется B-Tree. Чтобы примерно представить себе работу следует вспомнить обычное бинарное дерево (поиск по нему имеет логарифмическую скорость поиска элемента). Однако в данном случае всё устроено сложнее: дерево сбалансировано и сильно ветвистое — каждый узел обычно имеет более двух потомков [2.1.1](#).

Элементы данного дерева отсортированы по возрастанию. Что позволяет эффективно выполнять поиск как отдельного значения, так и интервала значений. Ситуация ухудшается, когда необходимо делать поиск по нескольким измерениям. В этом случае мы можем ограничить лишь одно измерение, поиск по другим будет производиться полным перебором.

Тем не менее для большинства задач B-дерево всё-таки является хорошим вариантом. B-Tree можно назвать самым популярным индексом, используемым в большинстве современных СУБД как реляционных, так и нереляционных при этом абсолютно не важно, где именно хранятся данные — в памяти или на диске. Существует много модификаций B-Tree: B+Tree (используется в CouchDB, MongoDB), SB-Tree (OrientDB), B*-Tree.

В-дерево было предложено ещё в 1970 году для эффективного поиска среди файлов [1], с тех пор появилось большое количество эффективных и компактных реализаций этой структуры. Такую популярность данная структура получила благодаря своей работе с памятью. Если мы хотим прочитать какое-либо значение, то в память/кэш помещается весь блок данных. Что существенно ускоряет скорость чтения, если кроме этого потребуются и соседние значения. Однако это является и недостатком этой структуры — если требуется записать новое или перезаписать существующее значение, будет обновлен весь блок. Такие «паразитные» чтения в литературе, посвящённой хранению на диске, называется *read amplification*, а «паразитные» записи — *write amplification*. Формально, *amplification factor*, то есть коэффициент умножения, вычисляется как отношение размера фактически прочитанных (или записанных) данных к реально необходимому (или изменённому) размеру. В случае В-дерева порядок этого коэффициента — десятки и сотни.

Hash

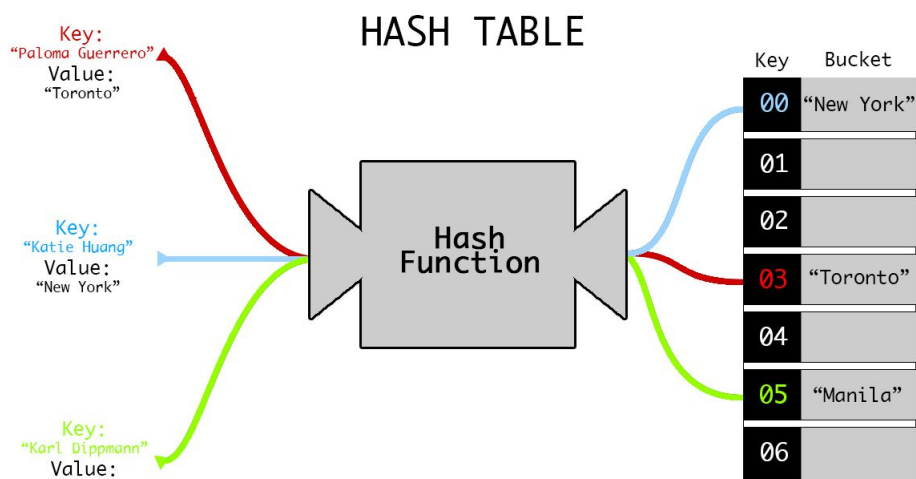


Рисунок 2.1.2 — Схематический пример организации работы hash-индекса

Hash-индекс работает не с индексируемыми ключами, а с их хэшами. Идея хэширования состоит в том, чтобы значению любого типа данных сопоставить некоторую битовую последовательность фиксированной длины. Функцию, осуществляющую такое преобразование, называют хэш-функцией. Вычисленное

значение указывает на некоторую область, хранящую нужную запись. Доступ к этой записи может быть получен за константное время — $O(1)$ (рис. 2.1.2).

Ключевым отличием от, например, B-Tree является отсутствие возможности поиска в интервале (близкие значения обычно имеют различные хэши). Кроме того отсутствует возможность выборки по префиксу, поскольку хэш вычисляется от полного ключа.

Существует класс задач, для которых важна скорость доступа к данным по простому ключу. Это привело к появлению так называемых key-value баз данных, например, Redis, Riak, memcached.

LSM-Tree

Ещё одним типом дерева, наравне с B-Tree, предназначенным для хранения данных является LSM-Tree — *Log-structured merge-tree* [2]. В отличие от B-Tree, которое можно использовать как для хранения в памяти, так и на диске, LSM-Tree предназначено для хранения данных именно на диске. Разделение на части, хранящиеся в памяти и на диске заложено в саму архитектуру данной структуры. Все операции вставки делаются в L0 (уровень, хранящийся в оперативной памяти), как только место там заканчивается, данные начинают сбрасываться на диск (рис. 2.1.3).

Ещё одним ключевым отличием является то, что в узлах дерева хранятся не сами данные, а операции с ними (рис. 2.1.4).

LSM-деревья работают быстрее для частых вставок и редких чтений, в отличие от B-деревьев, иными словами write amplification LSM-деревьев меньше, но read amplification выше. LSM-деревья стали довольно популярны в последнее время, это связано с тем, что стоимость внешней памяти стала меньше, при этом популярность приобретают SSD-диски, обладающие более высокой скоростью чтения по сравнению с устаревающими HDD-дисками.

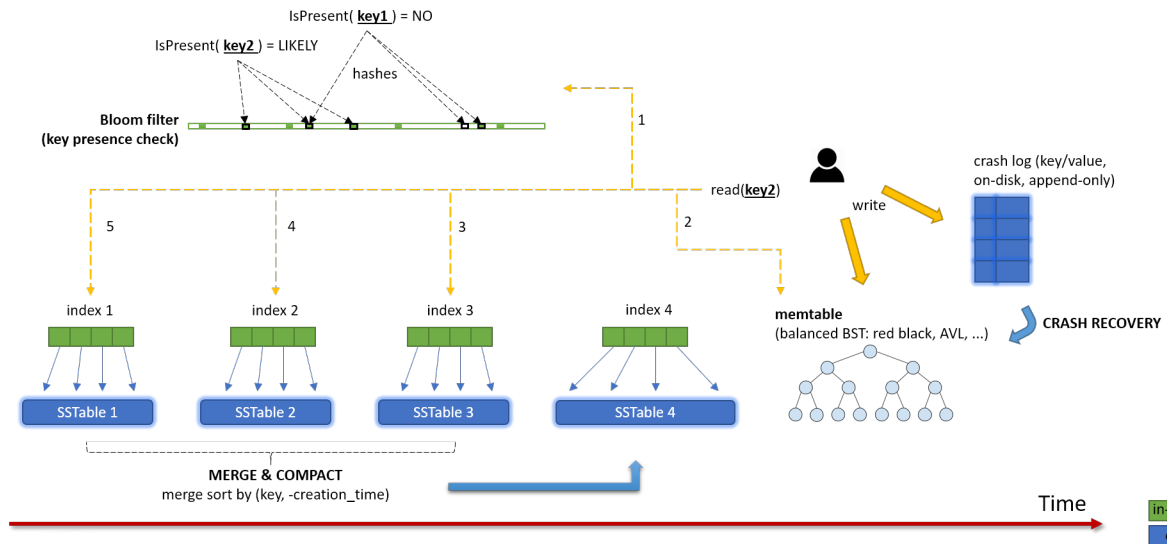


Рисунок 2.1.3 — Схематическое представление LSM-Tree

key	lsn	Op code	Value
1	176	REPLACE	2018-05-07 15:00:01
1	53	INSERT	2017-12-31 23:59:01
2	174	REPLACE	2018-05-06 00:00:00
3	175	REPLACE	2018-05-07 09:04:19
3	9	REPLACE	2017-01-01 19:25:43
3	7	INSERT	2017-01-01 19:22:16
4	173	DELETE	
4	168	INSERT	2018-05-05 07:40:01

Рисунок 2.1.4 — Хранение операций над данными, а не самих данных

Inverted index

Индекс, используемая для полнотекстового поиска. Содержит список всех уникальных слов и ссылки на документы, в которых эти слова встретились.

Полнотекстовые запросы выполняют лингвистический поиск в текстовых данных путем обработки слов и фраз в соответствии с правилами конкретного языка: разбиение на слова, отсекание окончаний, выбор однокоренных слов и т.д. Отдельными задачами при полнотекстовом поиске являются ранжирование результатов запроса и исключение ненужных слов.

Реализации полнотекстового поиска варьируются в различных СУБД. Инвертированный индекс используется в Microsoft SQL Server, MySQL, OrientDB и поисковом движке Elasticsearch.

Обобщением данного типа индекса является *GIN* (Generalized Inverted Index), реализованный в PostgreSQL. Кроме полнотекстового поиска, является подходящим для индексирования массивов и JSON. Обобщенным он называется, потому что операция над индексируемым объектом задается отдельно в отличие от, например, B-Tree, где все операции сравнения уже заданы. В качестве операции могут использоваться такие как «содержит», «пересекается», «содержится».

Количество текстовой информации, окружающей нас огромно: новости, книги, письма и т.д. Для индексирования содержимого этот тип индекса является подходящим. Однако работа с текстом не входит в поставленную задачу.

Пространственные индексы

Большинство современных СУБД имеют типы, предназначенные для работы с пространственными типами данных: точки, прямые, окружности и другие геометрические объекты. Для данных объектов используются свои стратегии индексирования.

Известными решениями является использование пространственной сетки (spatial grid), дерева квадрантов (quadtree) и R-Tree.

Данные индексы используются графовыми базами данных (Neo4j, AllegroGrath), однако существуют специальные дополнения и расширения для известных СУБД, но предназначенные для обработки исключительно пространственной информации — PostGIS, Oracle Spatial, GeoAPI в Redis.

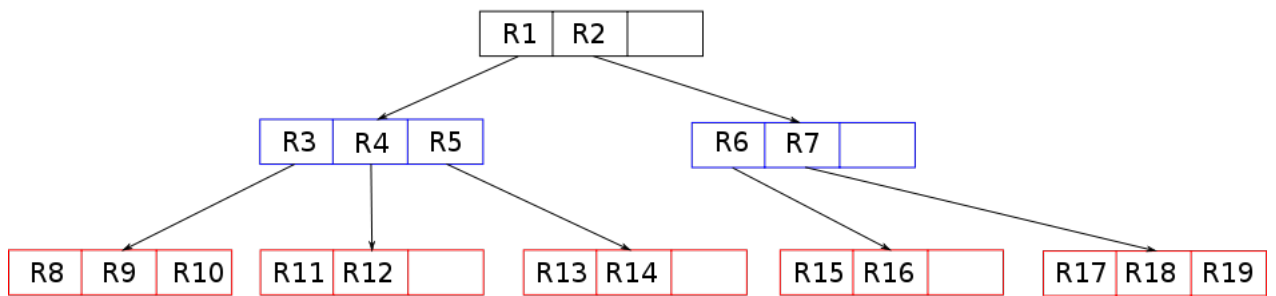
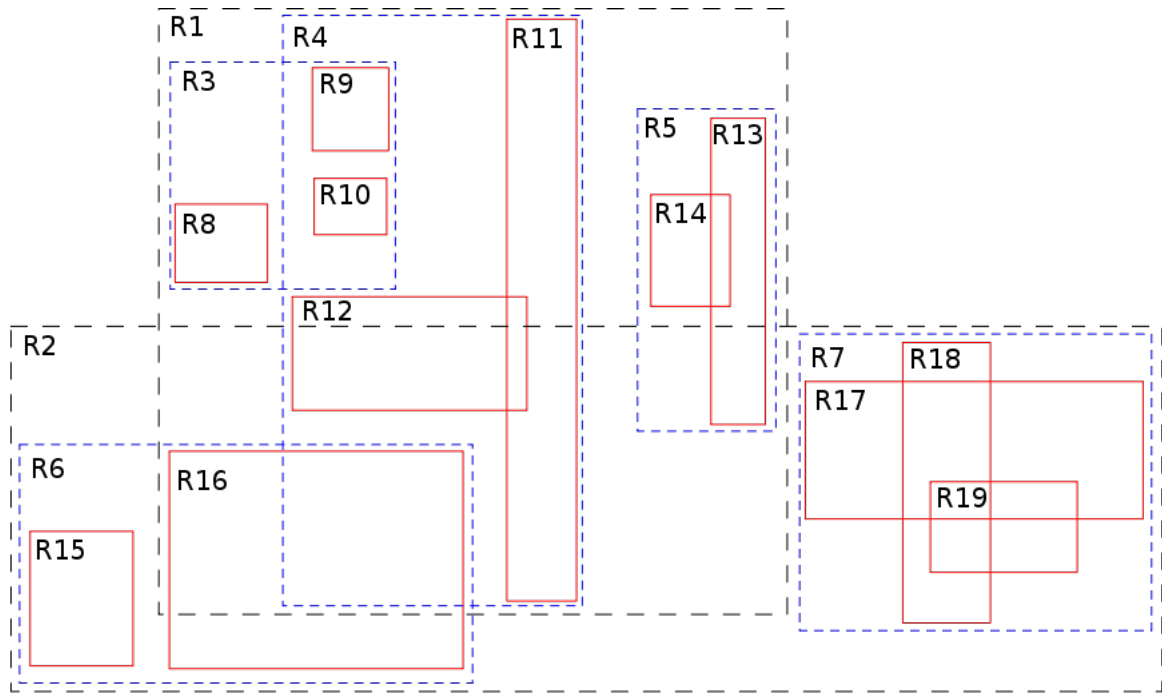


Рисунок 2.1.5 — Пример организации хранения данных в R-Tree

R-Tree

Эта структура данных разбивает многомерное пространство на множество иерархически вложенных и, возможно, пересекающихся, прямоугольников и гиперкубов (для многомерных структур) [2.1.5](#).

Подходит для поиска объектов в 2-3-мерном пространстве. Идея лежащая в основе индекса — группировка объектов в зависимости от расстояния друг до друга. Это ускоряет поиск, однако происходит потеря точности, и возвращенный результат может не быть абсолютно точным.

Данный тип индекса поддерживается некоторыми движками СУБД MariaDB (SPATIAL INDEX), PostgreSQL (RTREE), Oracle.

Существует несколько модификаций R-Tree: R+-Tree, R*-Tree. Обобщением R-Tree является X-Tree, который позволяет индексировать данные произвольных размерностей.

Другое обобщение *GiST* (*The Generalized Search Tree*) — обобщенное дерево поиска. Реализовано в PostgreSQL и подобно GIN поддерживает индексирование произвольной информации (геоданные, тексты, изображения и т.д.) с использованием операций «принадлежит», «содержит», «совпадает», «соответствует».

Z-order curve (Кривая Мортон)

Индекс используемый для хранения многомерных данных в одномерной структуре. К каждому значению применяется преобразование Мортон, заключающееся в чередовании двоичных цифр координатных значений, полученный результат называется Z-последовательностью (Z-order curve). Для обработки одномерных данных используется обычное B-дерево.

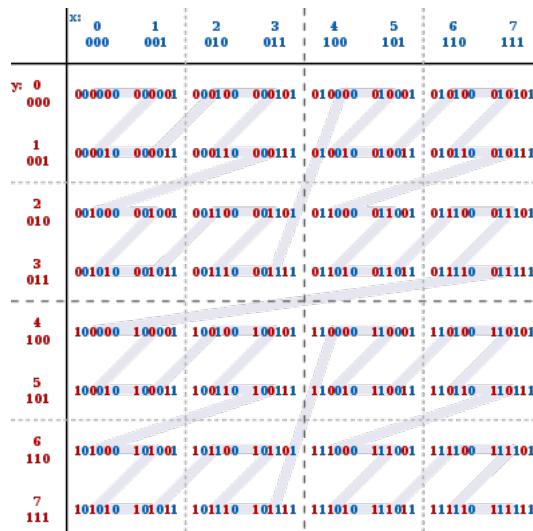


Рисунок 2.1.6 — Построение Z-последовательности

Позволяет эффективно производить поиск по интервалам значений, однако часть возвращаемого результата может и не находиться в указанном интервале (рис. 2.1.7), поэтому при запросе приходится применять дополнительные механизмы для фильтрации данных. Это накладывает некоторые

ограничения на целесообразность применения данного индекса. Запросы должны быть.

- **Часто задаваемыми.** Распространена практика, когда часть параметров запроса не задается, а остается открытой, однако в данном случае это может серьезно влиять на производительность.
- **«Избирательным».** Границы, устанавливаемые при запросе должны исключать большие объемы данных. Для неравномерно распределенных логических значений пространство поиска может быть сильно увеличено.

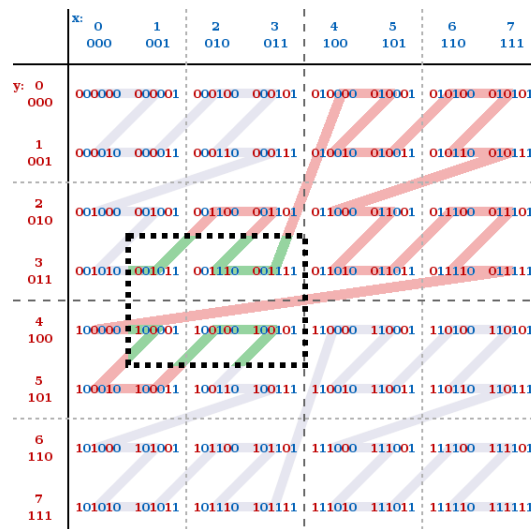


Рисунок 2.1.7 — Поиск значений в интервале

При реализации Z-адрес рассматривается исключительно как битовая последовательность. Это значит, что единственное ограничение на тип ключа — возможность упорядочивания при переходе к двоичному представлению. В итоге доступные типы ограничиваются не только целыми числами, но и числами с плавающей точкой, строками, временными метками...

Данный тип индексирования используется в TransBase[3], Accumulo, HBase [4], DynamoDB[5; 6].

Стоит отметить, что данное преобразование не является единственным для отображения многомерных данных в одномерные. Могут использоваться кривые Гильберта или Пеано. Однако Z-последовательность гораздо проще для вычисления.

Индексы с использованием машинного обучения

Можно выделить несколько подходов, которые могут быть использованы для поиска информации и выделения закономерностей в больших массивах данных — Latent Semantic Indexing (LSI) и Hidden Markov Model (HMM). Данные варианты хоть и являются интересными и полезными в некоторых сферах, но примеров их использования в каких-либо СУБД нет.

2.2 Используемые индексы в различных СУБД

СУБД	Индексы
PostgreSQL	B-Tree, R-Tree, Hash, GiST, SP-GiST, GIN, RUM, BRIN, Bloom
MySQL/MariaDB	B-Tree, Hash, R-Tree, Inverted Index
Oracle	B-Tree, B-Tree-cluster, Hash-cluster, Reverse key, Bitmap
MongoDB	B-Tree, Geohash, Text index, Hash
OrientDB	SB-Tree, Hash, Lucene Fulltext, Lucene Spatial
MemSQL	SkipList, Hash, Columnstore

2.3 Выбор платформы для разработки

В качестве платформы, для которой будет реализован индекс на основе кривой Мортонa была выбрана СУБД Tarantool — СУБД с открытым исходным кодом и активно разрабатываемая в настоящее время.

Предлагаемое решение достаточно просто может быть встроено в существующую кодовую базу — около 400 тысяч строк на языке C и 200 тысяч строк на языке Lua.

Основные особенности СУБД Tarantool:

- Удовлетворение принципу ACID

- Наличие двух движков для хранения данных на диске (vinyl) и в оперативной памяти (memtx)
- Хранение данных в формате MessagePack
- Обработка транзакций в одном потоке
- Не только база данных, но и сервер приложений

Реализовываемый индекс будет работать с движком Memtx — все данные будут храниться в оперативной памяти. Часть уже существующих решений может быть переиспользована, например, B⁺-Tree (BPS, в терминологии Tarantool).

В СУБД Tarantool достаточно необычная система типов. При работе как с сервером приложений пользователь использует язык программирования Lua 5.1 (LuaJIT 2.1 beta 2) и его типы: "string", "number" (double-precision floating-point), "boolean", "table" (ассоциативный массив), "nil" (отсутствие значения), а так же пользовательские типы данных, представленные типами "userdata" и "cdata".

Уровнем ниже находится MessagePack, обладающий большим количеством типов, более близким к представленным в других языках программирования — unsigned int, signed int, float, double, ... С этим уровнем также взаимодействуют коннекторы, позволяющие внешним системам работать с данными.

Ещё ниже следует система типов Tarantool, обладающая типами которые могут сочетать в себе сразу несколько MessagePack типов, например, "number—double" для нецелых чисел, signed integer для отрицательных целых чисел и unsigned integer для положительных чисел.

При реализации необходимо будет решить несколько инженерных задач:

- Разработка пользовательского интерфейса для работы с индексом
- Встраивание решения в существующую кодовую базу
- Адаптация решения под существующую систему типов

Элементарной единицей является кортеж (tuple), аналог строки в реляционных базах данных. Однако в отличие от реляционных БД кортеж может иметь произвольную длину и содержать произвольные типы. Строгая типизация требуется только для индексируемых полей.

Глава 3. Реализация

Реализацию индекса можно декомпозировать на несколько частей:

- Битовый массив (bit array)
- Логика z-order curve
- Встраивание в движок БД
- Lua-frontend

3.1 Bit array

Как описано в предыдущих пунктах, индексируемое значение получается в результате перемешивания битов указанных индексируемых полей. Данную структуру будем называть битовый массив. При реализации прототипа было решено найти и использовать готовую реализацию на языке C. Реализация была найдена — <https://github.com/noporpoise/BitArray>, она удовлетворяла необходимым функциональным потребностям, однако имела достаточно небольшое покрытие тестами, содержала баги, которые пришлось исправить — <https://github.com/noporpoise/BitArray/pull/15> и была достаточно неоптимальной с учетом специфики решаемой задачи.

С учетом того, что для хранения и представления большинства типов используется 64 бита, размер массива всегда будет кратен 64 элементам. А именно $N * 64$, где N — количество частей в индексе. Приведенная выше реализация содержала достаточно большое число проверок и занимала больше памяти из-за того, что являлось массивом общего назначения с возможностью динамического изменения размера. Для достижения максимальной эффективности пришлось реализовать свой битовый массив постоянного размера и более оптимально работающий с памятью (для выделения памяти использовалось 2 системных вызова, один — выделение структуры со служебными полями, другой сам массив). В полученной реализации создание массива — одна аллокация. Служебное поле "количество элементов в массиве" также потеряло смысл и было удалено поскольку вычисляется как $N * 64$.

Следующий шаг в оптимизации — «векторизация». Большинство современных процессоров поддерживает так называемые векторные инструкции, служащие для обработки массивов данных (SIMD — Single Instruction Multiple Data). По словам разработчиков использование таких инструкций позволяет получать прирост производительности до 30%. О том, что какой-то цикл может быть векторизован можно с помощью директивы `#pragma simd`, однако это не дает гарантий, что цикл будет векторизован, компилятор может и проигнорировать данную директиву, к тому же большинство современных компиляторов могут автоматически находить векторизуемые циклы. Посмотреть за тем, векторизуется цикл или нет, можно с помощью специальных опций компилятора.

Циклы для операций AND и OR были векторизованы для компиляторов GCC v7.4.0 и Apple Clang v11.0.0.

Bit interleaving

Поскольку одной из ключевых операций является перемешивание битов, эта функциональность была добавлена для битового массива.

Вычисление производится не тривиальным образом, а с помощью так называемых lookup-таблиц. Таблица ставит в соответствие любому числу от 0 до 255 значение, вычисляемое по правилу $\sum i = 0^7(k_i * 2^{i*n})$, где n - число массивов, которые должны быть перемешаны, k_i - значение i -ого бита. Для каждой части в зависимости от её номера m хранится сдвинутая на m разрядов копия такой lookup-таблицы.

Соответственно вычисление z -адреса происходит за $8 * n$ обращений, сдвигов и применения логической операции OR, где n - число массивов. Стоит обратить внимание, что lookup-таблицы вычислены для октетов, а не больших размеров. Использование для $n = 16/32/64$ значений затруднено, поскольку расходы памяти на хранение этой таблицы растут экспоненциально как 2^n .

3.2 Z-order curve

Изначально информация о существовании индекса на основе кривой z-порядка была получена из статей Amazon [5; 6], данные статьи являются скорее инструкцией для пользователей, как создать индекс на основе кривой Мортон для удовлетворения пользовательских потребностей с минимальными накладными расходами (z-order curve не является встроенной функциональностью DynamoDB).

Основные идеи, которые можно было бы почерпнуть из статьи — алгоритмы работы с кривой z-порядка и работа с различными типами данных - необходимо найти функцию, которая для каждого типа данных возвращает битовую строку. Для различных значений битовые строки должны быть лексикографически упорядочены. Например, рассмотрим знаковые целые числа -1 и 2 . Для них выполняется следующее неравенство $-1 < 2$. Однако в бинарном представлении -1 — это 11111111 , а 2 — 00000010 . Лексикографический порядок нарушен. Для восстановления предлагается инвертировать старший бит, тогда значения для -1 — 01111111 и 10000010 для 2 . Другие типы будут рассмотрены в следующей главе.

После вычисления z-адреса для каждого проиндексированного значения и размещения его в B-дереве мы получаем структуру, из которой хотелось бы выбирать значения, соответствующие нашим запросам. Самое первое, что может определить пользователь - нижнюю и верхнюю границы поиска (lower и upper bound). Рассмотрим пример для случая двух измерений. Мы хотим сделать выборку в прямоугольнике $[x_{min}; x_{max}]$ и $[y_{min}; y_{max}]$ - границами будут $z_address(x_{min}, y_{min})$ и $z_address(x_{max}, y_{max})$. В интервал между двумя этими значениями может попадать достаточно большое число значений, не принадлежащих заданному прямоугольнику. Перед тем как объяснять алгоритм итерации по данной структуре следует ввести 2 ключевые функции - $is_relevant(lower_bound, upper_bound, z_address)$, которая проверяет $z_address$ на принадлежность заданному прямоугольнику, и $next_jump_in(lower_bound, upper_bound, z_address)$, возвращающая для z-адреса за пределами прямоугольника первое по порядку значение, попадающее в прямоугольник. С учетом вышесказанного итерация выглядит следующим образом. Мы начинаем с некоторого $z_address$ значения (большого

или равного *lower_bound*), проверяем его принадлежность прямоугольнику с помощью *is_relevant*, если функция вернула *true*, то это значение возвращается пользователю, иначе с помощью *next_jump_in* переходим к следующему подходящему значению, пока не выйдем за границу *upper_bound*.

Описание двух, используемых выше алгоритмов, можно найти в [3; 7; 8] С некоторыми изменениями и модификациями они были реализованы в рамках дипломной работы. Например, функция *is_relevant* может рассматриваться как часть функции *next_jump_in* и практически без изменений извлечена в отдельную функцию. Однако такой наивный подход неоптимален, функцию можно упростить, добавить дополнительные оптимизации и применить эвристики, способные ускорить данные функции. Во-первых, там, где можно было отказаться от массивов в пользу битовых масок, это было сделано. Как было сказано, ключ — это массив 64-битных чисел, однако если индексируемые числа небольшие, то большая часть битов будет нулевой. Приведенные алгоритмы проверки и поиска работают за $O(N)$, где N — длина ключа. При этом если биты с одинаковым порядковым номером *zvalue*, *lower bound* и *upper bound* могут быть безболезненно пропущены, это же справедливо и для более крупных единиц, например, байтов. Данная эвристика для размерности 3 и чисел из интервала $[0; 255]$ дала выигрыш 15 — 20%.

Отдельно стоит отметить более рациональный подход к используемым типам данных. Использование типов как можно меньшей размерности, скажем, *uint8_t* вместо *uint64_t* дало существенный прирост производительности, сократив время работы функций практически в 2 раза.

3.3 Tarantool index

В терминах ООП — индекс это класс, обладающий некоторым набором функций и реализующий некоторую структуру данных, хранящую информацию о проиндексированных кортежах. Как было описано в предыдущих частях, вычисленные *z*-адреса сохраняются в *B*-дереве. В СУБД Tarantool уже была готовая реализация $B+^*$ -tree, которую было решено переиспользовать.

СУБД Tarantool имеет собственную систему типов. Для индекса было выбрано несколько поддерживаемых типов - *unsigned* (*unsigned integer*), *integer*

(signed integer), number (double-precision floating-point number) и string (только префиксный поиск по первым 64 битам). Это достаточно сильно отличает данный индекс от уже существующего R-Tree, предназначенного для работы с числами с плавающей точкой.

Как было сказано, для корректной работы нужно научиться преобразовывать значения определенного типа к некоторым битовым лексикографически упорядоченным словам. Все указанные типы используют 64 бита для хранения, соответствующие им значения решено было хранить как unsigned integer размером 64 бита.

Для unsigned integer преобразование является тривиальным, поскольку в битовые представления и так лексикографически упорядочены.

У знаковых целых чисел (signed integers) старший бит является меткой знака. То есть с точки зрения бинарного представления любое отрицательное число больше любого положительного. Получить значение, удовлетворяющее нужным критериям можно с помощью инверсии старшего бита.

Поиск по строкам, как было сказано, возможен только префиксный. Первые 8 байт любой строки сохраняются как unsigned integer число, в случае если строка короче, она дополняется нулями в конце. Если используется строка в кодировке ASCII, то значения будут уже лексикографически отсортированы. Кодировка UTF-8 обратно совместима с ASCII, поэтому также может использоваться. Такой подход исключает поддержку collation'ов. Но всё-таки поддержку данного типа было решено оставить, поскольку 8 байт может быть вполне достаточно для многих пользовательских сценариев. Кроме того, данный поиск можно использовать как первичный, и при необходимости пользователь сам может выполнить дополнительную фильтрацию.

Наиболее сложный для рассмотрения вариант — числа с плавающей точкой двойной точности (double). Для того, чтобы разобраться с этим случаем следует обратиться к спецификации IEEE 754, говорящего что “если два числа с плавающей запятой одного и того же формата упорядочены (скажем, $x < y$), то они упорядочиваются таким же образом, когда их биты интерпретируются как целые числа со знаками (sign-magnitude integers)”. Нужное нам преобразование имеет следующий вид: для положительных чисел инвертируется старший бит, для отрицательных инвертируются все биты.

Отдельно стоит отметить *null*-значения. Они не поддерживаются в силу того, что невозможно задать битовое значение, которое бы соответствовало

бы отсутствию любого значения. Однако в пользовательском интерфейсе можно будет использовать такие значения для задания минимально возможного и максимально возможного значения ключа. Но это будет рассмотрено в следующих частях.

В реляционных базах данных индексы могут быть уникальными (индекс может содержать один экземпляр определенного значения) и неуникальными. Для Tarantool это тоже справедливо. Предполагается, что для каждого спейса (аналог реляционной таблицы) определен как минимум один индекс. Самый первый индекс должен быть уникальным и называется первичным. Именно в качестве первичного ключа предлагается использовать z-order curve по задумке авторов статьи Amazon. Однако в нашем случае было решено поступить иначе, запретить делать индекс на основе кривой z-порядка уникальным.

Любой индекс предполагает, что элементы внутри него располагаются детерминировано при построении при запуске, вставке очередного элемента и т.д. При появлении неуникальных значений встает вопрос о том, как сортировать эти значения. В Tarantool принято сортировать неуникальные вторичные ключи путем дописывания первичного ключа. Полученный ключ является уникальным и обычно не возникает проблем с сортировкой таких значений. Рассмотрим подробнее, как эта проблема была решена. Начнем с того, как это делается для B-Tree индекса. Вспомним, что кортеж (tuple) - это массив значений, закодированных в формат MessagePack. Для каждого индекса определены 2 значения *key_def* - поля, входящие в индекс, и *cmp_def* — *key_def*, расширенный первичным ключом, а значит уникальный. Это не банальное дописывание первичного ключа в конец, это добавление полей, отсутствующих в *key_def*, т.е. каждое поле может встретиться только один раз в *key_def* и *cmp_def*. *key_def* виден пользователю, однако в самом дереве хранится значение, соответствующее расширенному ключу. Сравнение двух кортежей производится с учетом полей, перечисленных в *cmp_def*.

Таким образом работу с индексом можно описывать как независимые 2 части — чтение и запись. Алгоритм записи следующий:

- Извлечь из полученного кортежа его z-адрес;
- Вставить в дерево элемент — структуру, состоящую из z-адреса и указателя на кортеж.

Алгоритм чтения:

- Извлечь z-адрес из ключа поиска;

- Получить итератор на наименьший элемент с таким ключом;
- Последовательно итерироваться по дереву, проверяя каждый элемент на принадлежность к региону поиска;
- В случае выхода за границу сделать прыжок для возврата в регион поиска;
- Прекратить поиск, если очередной элемент больше, чем *upper_bound*.

3.4 Lua frontend

Приводится описание того, как конечный пользователь должен выполнять запрос к базе данных с целью сохранения, поиска, удаления и модификации какого-либо элемента.

```

-- Создание space "test" - аналога таблицы в реляционных БД
local space = box.schema.space.create('test',
  { engine = 'memtx' })
-- Создание первичного ключа
5 local pk = space:create_index('pk',
  { type = 'tree', parts = {{1, 'unsigned'}}})
-- Создание индекса на основе z-order curve
-- Индексируемые поля 2 типа "unsigned" и 3 типа "integer"
local sk = space:create_index('secondary',
10 { type = 'zcurve', parts = {{2, 'unsigned'}, {3, 'integer'}}})
-- Вставка кортежа {1, 2, 3} в space
space:replace({1, 2, 3})
-- Прибавление ко второму полю кортежа единицы
space:update({1}, {{'+', 2, 1}})
15 -- Удаление кортежа
space:delete({1})
-- Выборка по всему индексу
sk:select()
sk:select({}, {iterator = 'ALL'})
20 -- Выборка в интервале от [0; 2] и [3; 5]
sk:select({0, 2, 3, 5})
sk:select({0, 2, 3, 5}, { iterator = 'EQ' })
sk:select({0, 2, 3, 5}, { iterator = 'GE' })
-- Выборка значений {5; 6}
25 sk:select({5, 6})
sk:select({5, 6}, {iterator = 'EQ'})

```

```

30 | -- Выборка значений [5; +inf] и [6; +inf]
    | sk:select({5, 6}, {iterator = 'GE'})
    | -- Выборка значений [-inf; 5] и [-inf; +inf]
    | sk:select({box.NULL, 5, box.NULL, box.NULL}, {iterator = 'GE'})
    | -- Удаление индекса
    | sk:drop()

```

Как видно, индекс поддерживает 3 итератора «ALL» — выборка всех данных, «EQ» - выборка данных с указанным ключом и «GE» - выборка данных, больше либо равных указанному ключу.

«box.NULL» — специальный символ, указывающий на отсутствие значения. Если он стоит на нечетном месте, то эквивалентен $-\infty$, на чётном — $+\infty$.

Глава 4. Результаты

Было проведено несколько тестов:

- Сравнение с R-Tree — поиск точек внутри гиперкуба размерности от 1 до 20;
- Сравнение с B-Tree — поиск точек внутри гиперкуба размерности от 1 до 20;
- Сравнение с B-Tree — префиксный поиск.

В каждом из тестов отдельно интересуют значения времени, потраченного на вставку тестового датасета, время выполнения запросов и объем потребляемой памяти.

Также будет проверено 3 варианта распределения значений в пространстве. В первом случае запрос будет делаться за данными, распределенными равномерно. Т.е. координаты точек — случайно распределенные числа в некотором интервале $[V_{min}; V_{max}]$. Во втором случае запрос должен будет вернуть пустое множество — данных удовлетворяющих запросу нет. И в последнем случае запрос будет высокоселективным, данные будут сконцентрированы вблизи одной точки, большой вклад во время выполнения запроса начнет давать не только время поиска нужных данных, но и время десериализации результатов перед возвратом клиенту.

4.1 Сравнение с R-Tree. Поиск точек внутри гиперкуба

Описание среды и условий проведения измерений

Для каждого из тестов создавались 2 спейса, содержащих 2 индекса — первичный B-Tree и вторичный R-tree или Z-order curve. Последовательно в каждый из спейсов вставлялись приготовленные кортежи, измерялось время вставки и память, которую занимает индекс. Стоит отметить, что для ускорения вставки и исключения факторов, способных её замедлить, были отключены

WAL (Write-ahead logs) — запись всех операций в специальный журнал, предотвращающий потерю данных в случае перезагрузки или остановки СУБД. После чего производилось N_q запросов, и измерялось время, за которое они все были исполнены.

В первом случае равномерного распределения генерировалось 10^6 значений в интервале от 0 до 10^5 . Область запроса — интервал $[0.35 \times v_{max}; 0.75 \times v_{max}]$ по каждой из размерностей, v_{max} — максимально сгенерированное случайное значение. Для усреднения результатов каждый из запросов повторялся 10 раз.

В случае высокоселективной выборки случайным образом для размерности D генерировались значения v_i в диапазоне от 0 до $V = 2.5 \times 10^8$ при этом большая часть значений лежит около 0, ими заполнялись 2 кортежа — для R-Tree (с массивом: $\{id, \{v_1, v_2, \dots, v_D\}\}$) и для Z-Order curve (без массива: $\{id, v_1, v_2, \dots, v_D\}$). id - первичный ключ, число типа unsigned. Всего $N_t = 10^6$ значений.

Далее генерировалось 2 числа Q_{min} и Q_{max} ($Q_{min} < Q_{max}$) для ограничения области в пространстве $[Q_{min}; Q_{max}]$ в каждой размерности $N_q = 10^3$ пар.

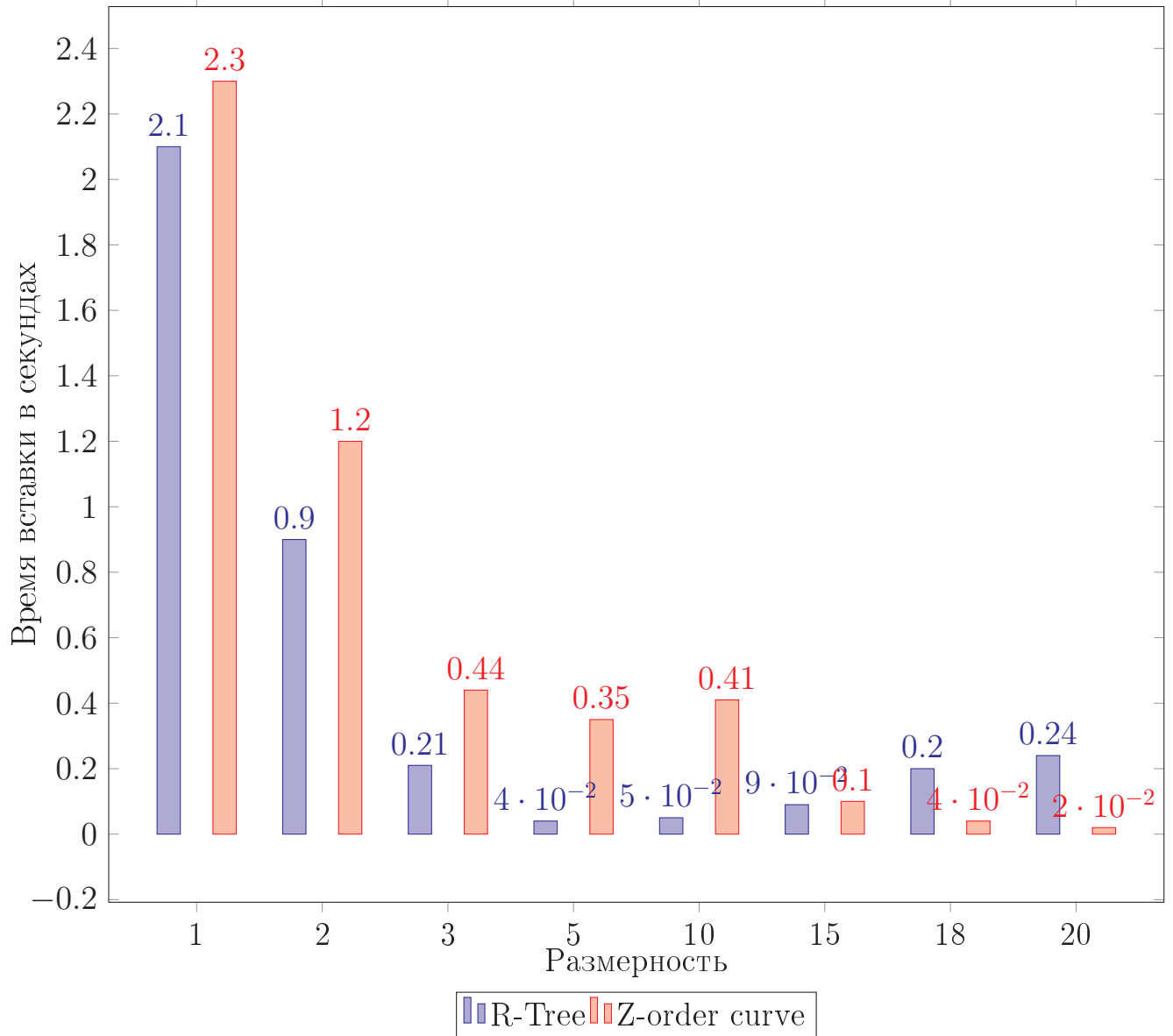
Результаты

Дадим интерпретацию полученным результатам. Во-первых, как видно из рис. 4.1 скорость вставки в Z-order curve индекс выше, чем в R-Tree. Данный тренд сохраняется и при увеличении размерности. На практике однако результат не особо релевантен, поскольку, как было отмечено выше, тест проводился с выключенным WAL.

С точки зрения расхода памяти оптимальнее Z-order curve. Причем на больших размерностях практически в три раза. Это уже можно считать релевантным положительным результатом, поскольку объем оперативной памяти обычно ограничен. С другой стороны, её стоимость постоянно снижается.

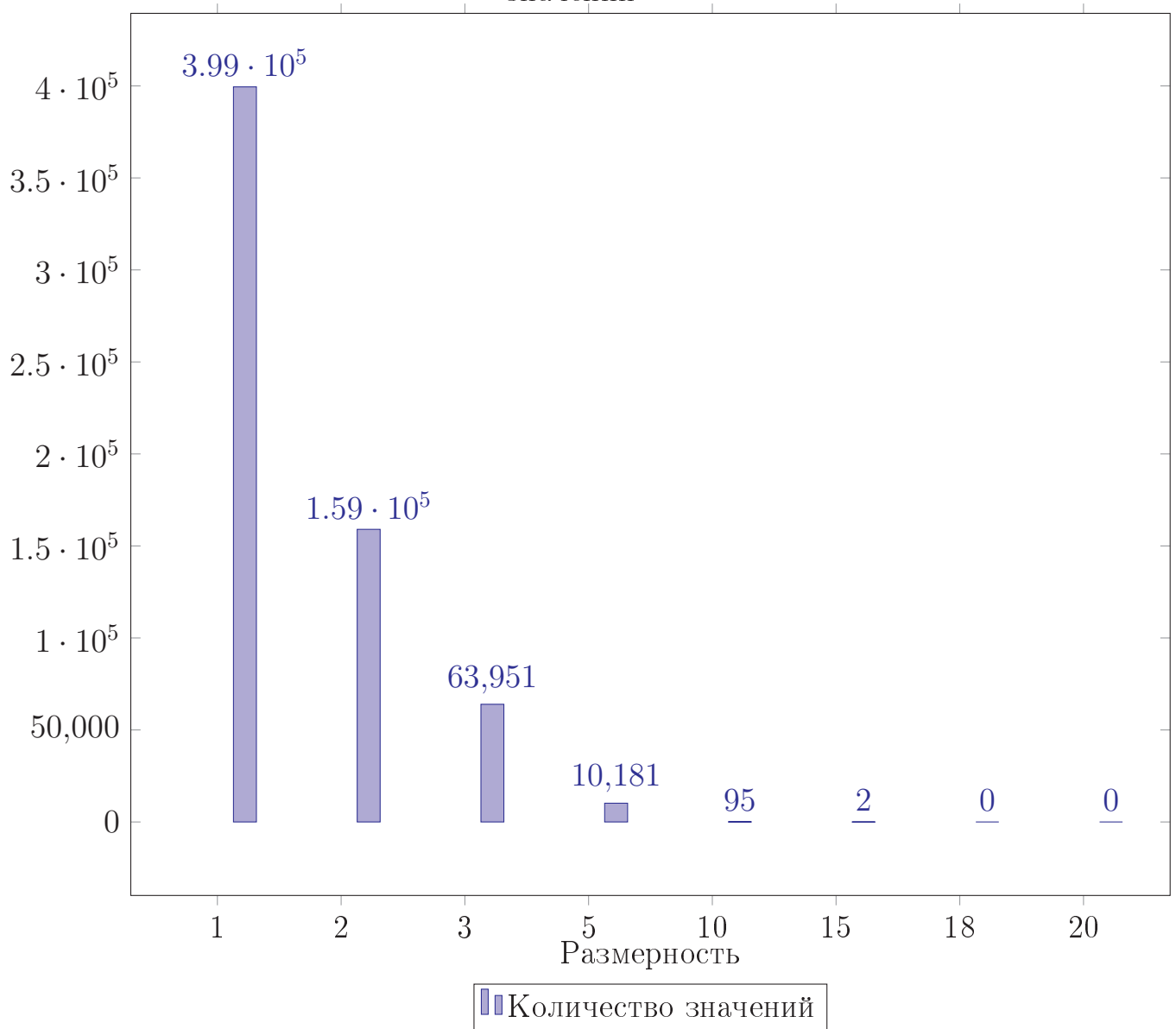
Последний график, показывающий время выполнения запросов на чтение наиболее интересен. Как видно примерно до размерности 5 время уменьшается, после чего для R-Tree остается практически постоянным, а для Z-order curve начинает драматически возрастать. Дело в том, что при одинаковом количестве

Рисунок 4.1.1 — Сравнение скорости вставки в R-Tree и Z-order curve индексы



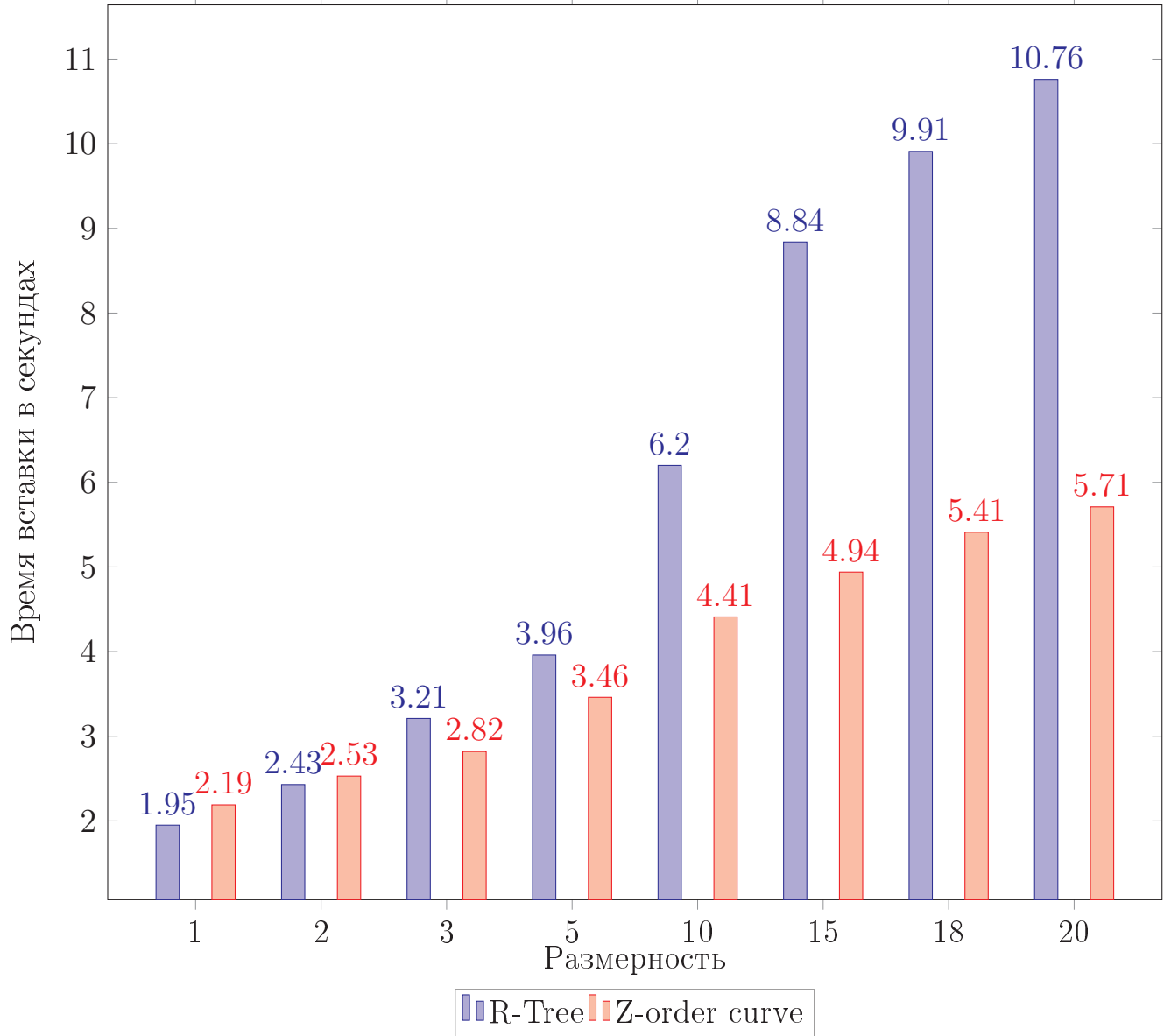
точек с увеличением размерности начинает падать селективность наших запросов. Больше размерностей — меньше вероятность того, что координаты точки попадут в интервал между границами запроса. Для R-Tree низкая селективность не критична, поэтому в конце время выполнения запросов практически одинаково. Совершенно иначе обстоит дело с Z-order Curve. Как только итератор натывается на точку, лежащую вне границ поиска, вычисляется Z-адрес следующего вхождения в указанную область. После этого совершается прыжок — спуск по дереву в эту точку. Если данной точки не существует, то мы попадем в следующую по порядку. Данная точка снова может лежать вне области поиска, тогда ситуация будет повторяться вновь и вновь. Это обоснование правой части графика. Слева при высокой селективности большая часть времени может тратиться не на сам поиск, а на то, чтобы упаковать данные и вернуть

Рисунок 4.1.2 — Селективность запросов в случае равномерно распределенных значений



их клиенту. При работе с внешними коннекторами — это пересылка по сети, при доступе из Lua-приложения это дополнительная работа по сериализации/-десериализации Lua-объектов перед возвратом пользователю. Также, если при спуске по R-дереву понять принадлежность точки/области гиперкубу можно за количество операций, зависящих от размерности, то при работе с Z-адресами алгоритмы линейно зависят от длины этого адреса — размерность, умноженная на достаточно большую константу — 64.

Рисунок 4.1.3 — Сравнение скорости вставки в R-Tree и Z-order curve индексы

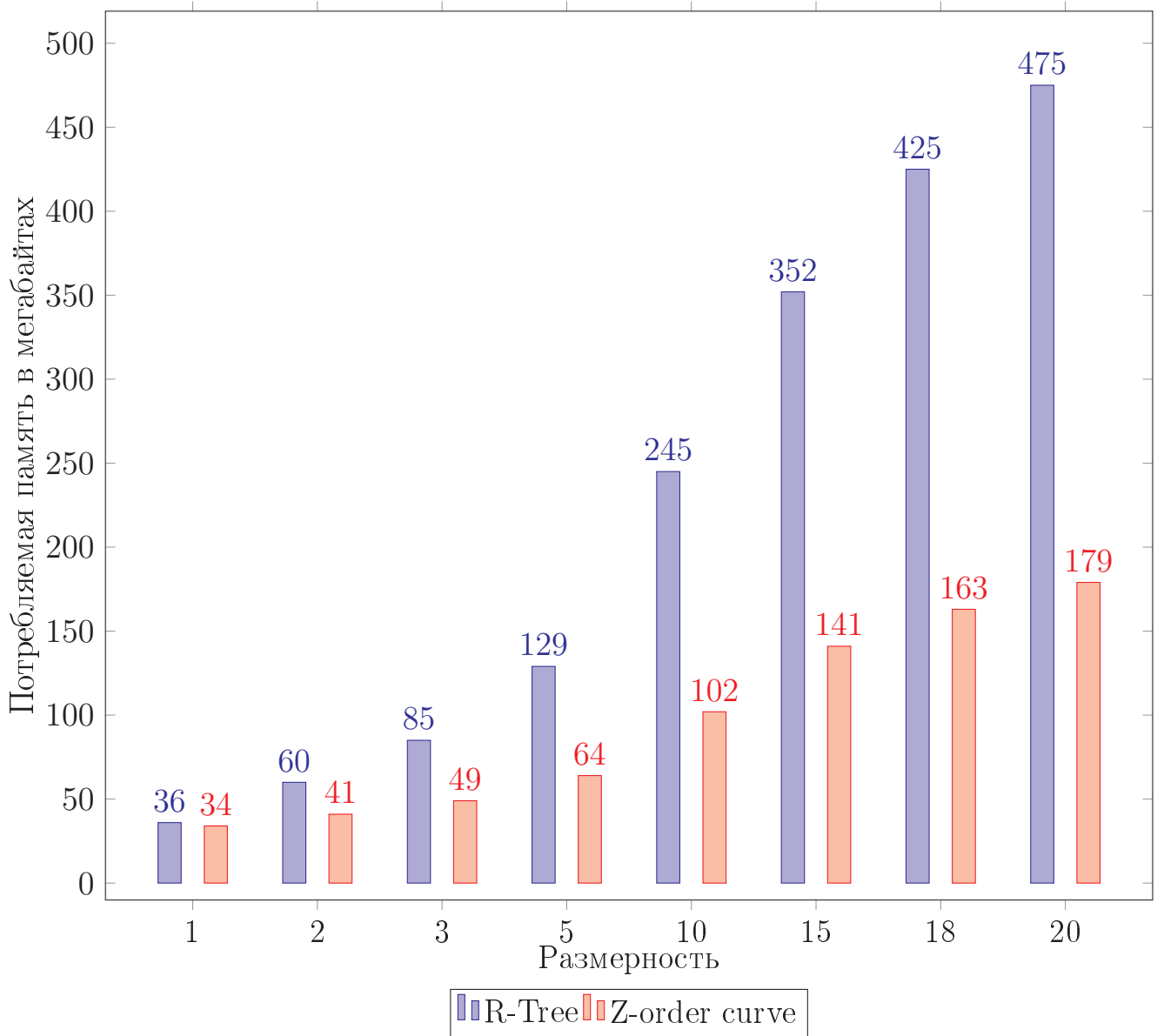


4.2 Сравнение с B-Tree. Поиск точек внутри гиперкуба

Описание среды и условий проведения измерений

В первом случае генерируется 10^6 значений равномерно распределенных в интервале $[0; 10^5]$. Область запроса — интервал $[0.35 \times v_{max}; 0.75 \times v_{max}]$ по каждой из размерностей, v_{max} — максимально сгенерированное случайное значение. Для усреднения результатов каждый из запросов повторялся 10 раз.

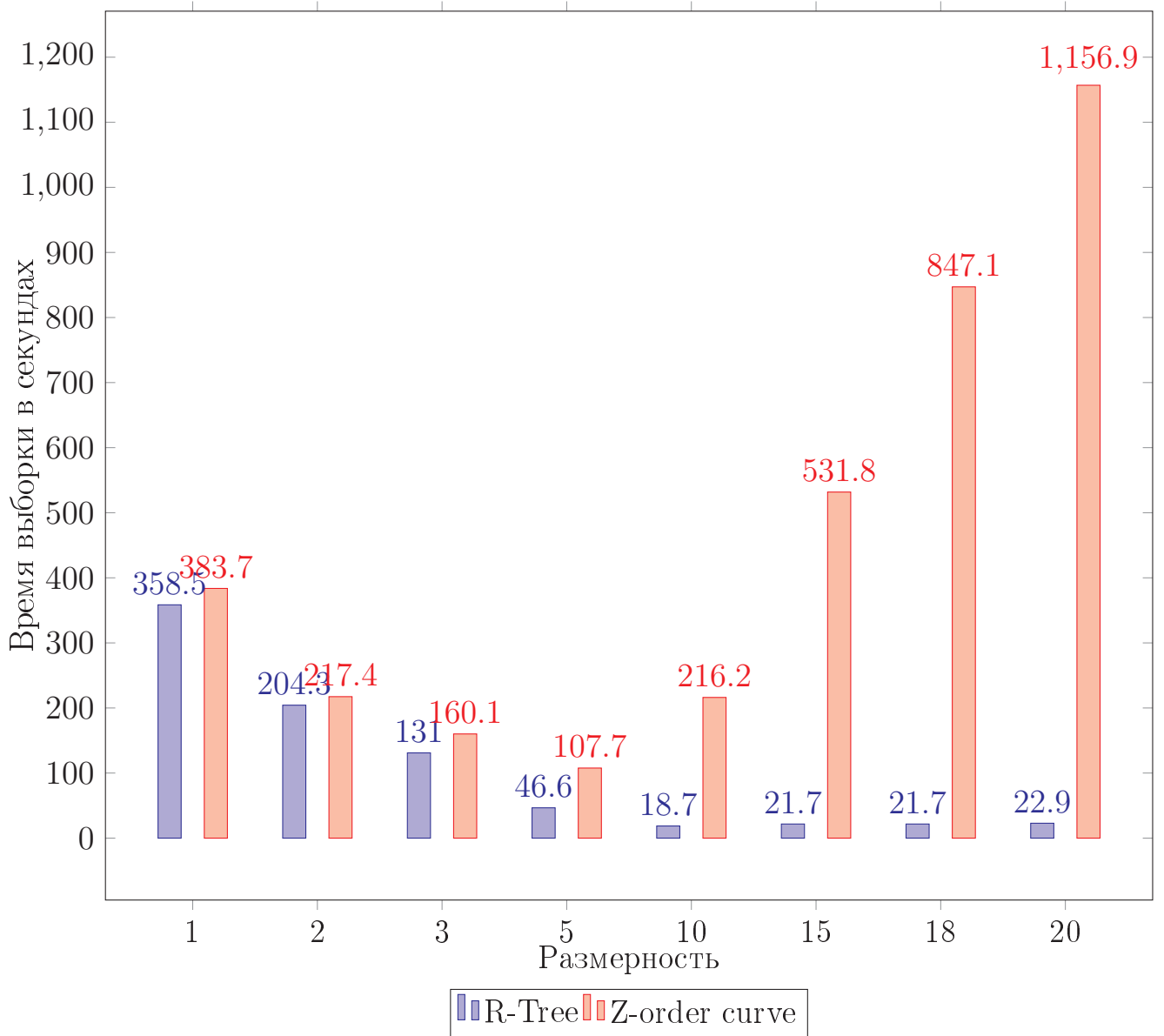
Для высокоселективных запросов, как и при сравнении с R-Tree, генерировалось $N = 10^6$ значений. В том же диапазоне от 0 до 2.5×10^8 , однако большая



из часть сконцентрирована в районе 0. Значения вставлялись в один спейс, в котором были созданы три индекса — первичный уникальный ключ, B-Tree индекс по первому полю и Z-order curve по второму и всем последующим полям. Далее выбирались точки из гиперкуба в интервале $[1; v_{max}/4]$, где v_{max} — самое большое сгенерированное значение. Чтобы избежать случайных выбросов и усреднить результаты, каждый запрос был сделан 10 раз.

Результаты и интерпретация

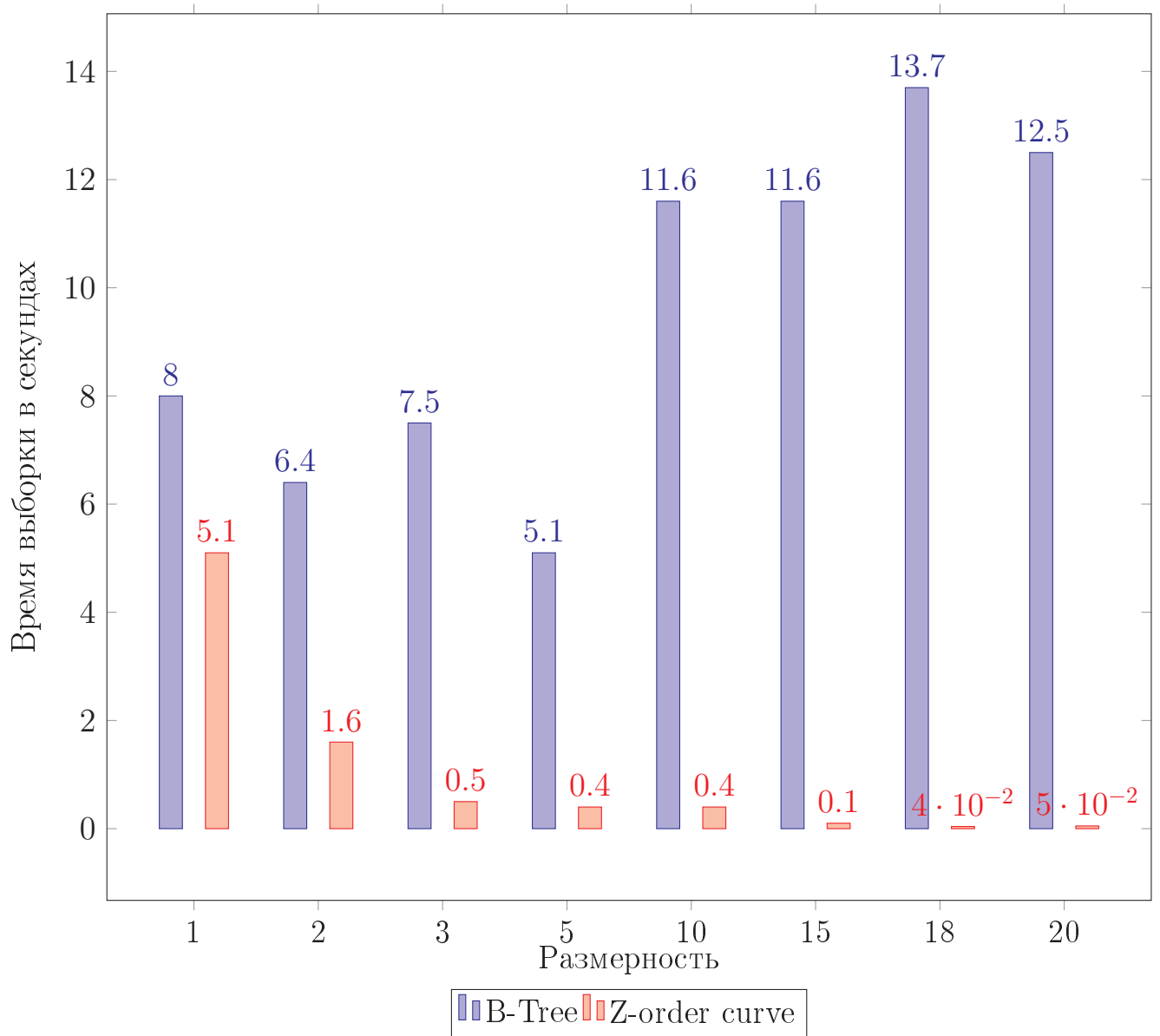
Результаты с нормальным распределением приведены на рисунке 4.2. Как и ожидалось, Z-order curve показал себя быстрее, чем B-Tree. Время,



которое тратилось на сканирование значений, не входящих в интервал поиска, значительно превышало время, которое затрачивалось на вычисление $next_intersection_point$ в случае выхода за границу поиска и прыжка к этой точке. При этом заметим, что результаты, полученные для больших размерностей не являются особо показательными, поскольку значений, удовлетворяющих критериям запроса было либо слишком мало, либо совсем не было.

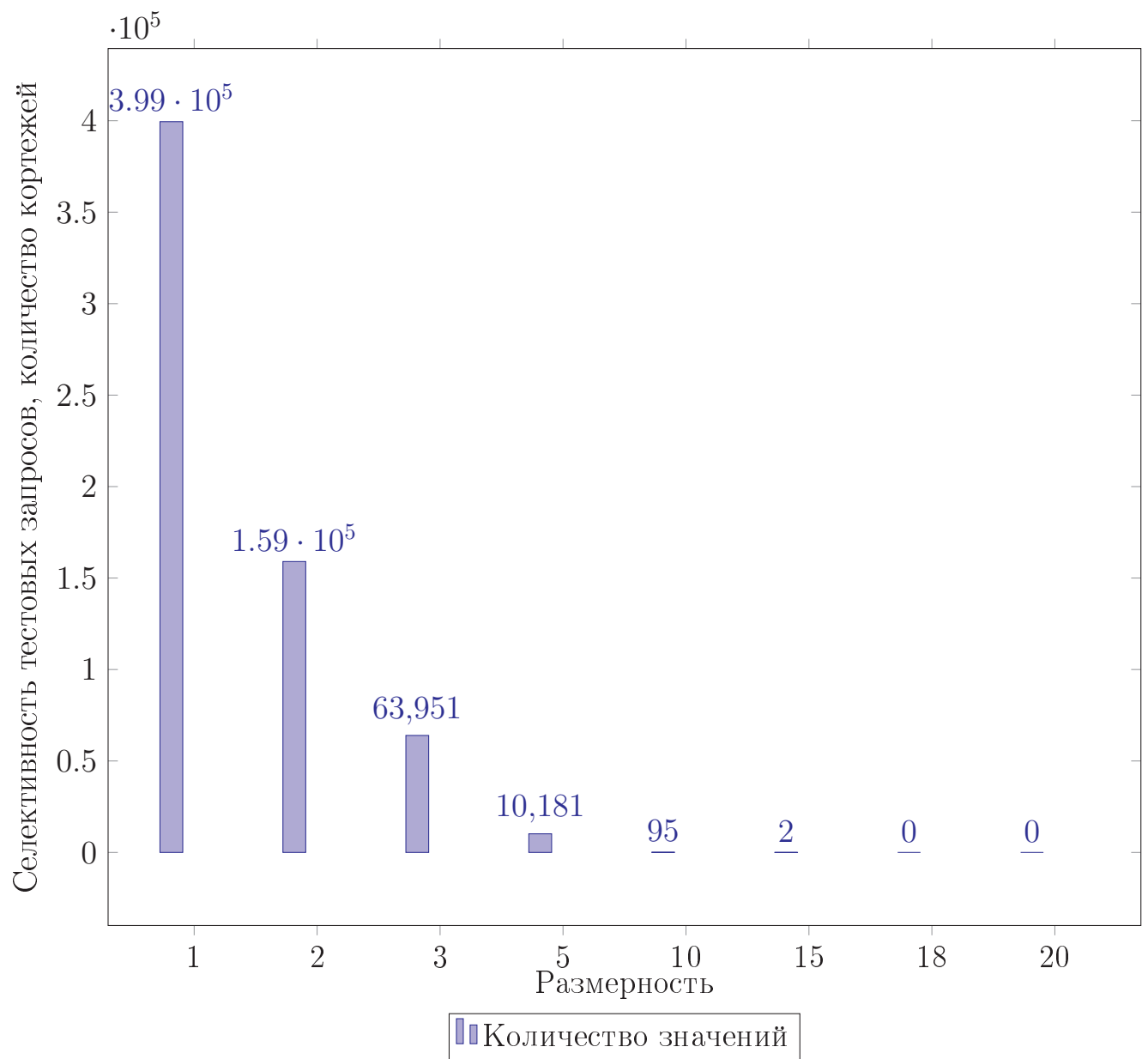
Результаты приведены на рисунке 4.2

В графиках присутствуют выбросы, но общая тенденция видна — при высокоселективных запросах использование Z-order curve оправдано. Однако с увеличением размерности и уменьшением селективности сканирование начинает быть эффективнее, чем частые спуски по дереву для пропуска интервала, не входящего в область запроса.



4.3 Сравнение с B-Tree. Префиксный поиск

Поиск производился по спейсу, заполненную одинаковыми строками. Выполнялся запрос с итератором «EQ» — equal. B-дерево оказалось медленнее примерно на 15%. Это связано с тем, что существующая реализация не использует дополнительной памяти для хранения проиндексированных значений. Поэтому каждая итерация — это декодинг message pack (см. рис 4.3.1, [9]) строки и затем уже сравнение. В противоположность Z-order curve индекс хранит дополнительно 8 байт на каждое проиндексированное значение. Декодирования на каждой итерации не происходит, что и является причиной подобного прироста производительности.



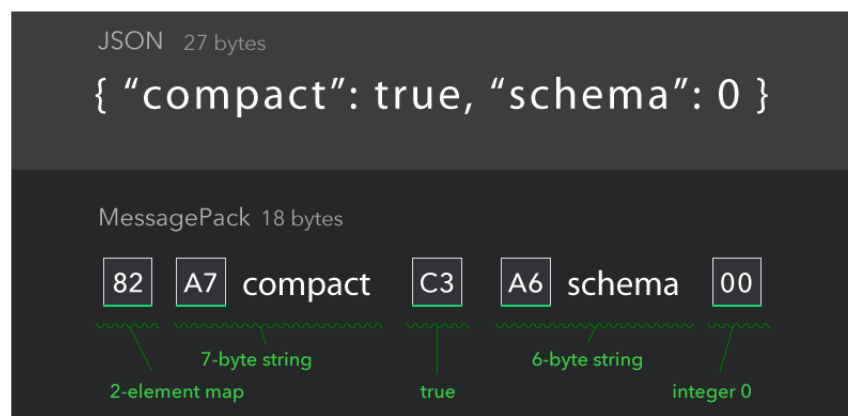
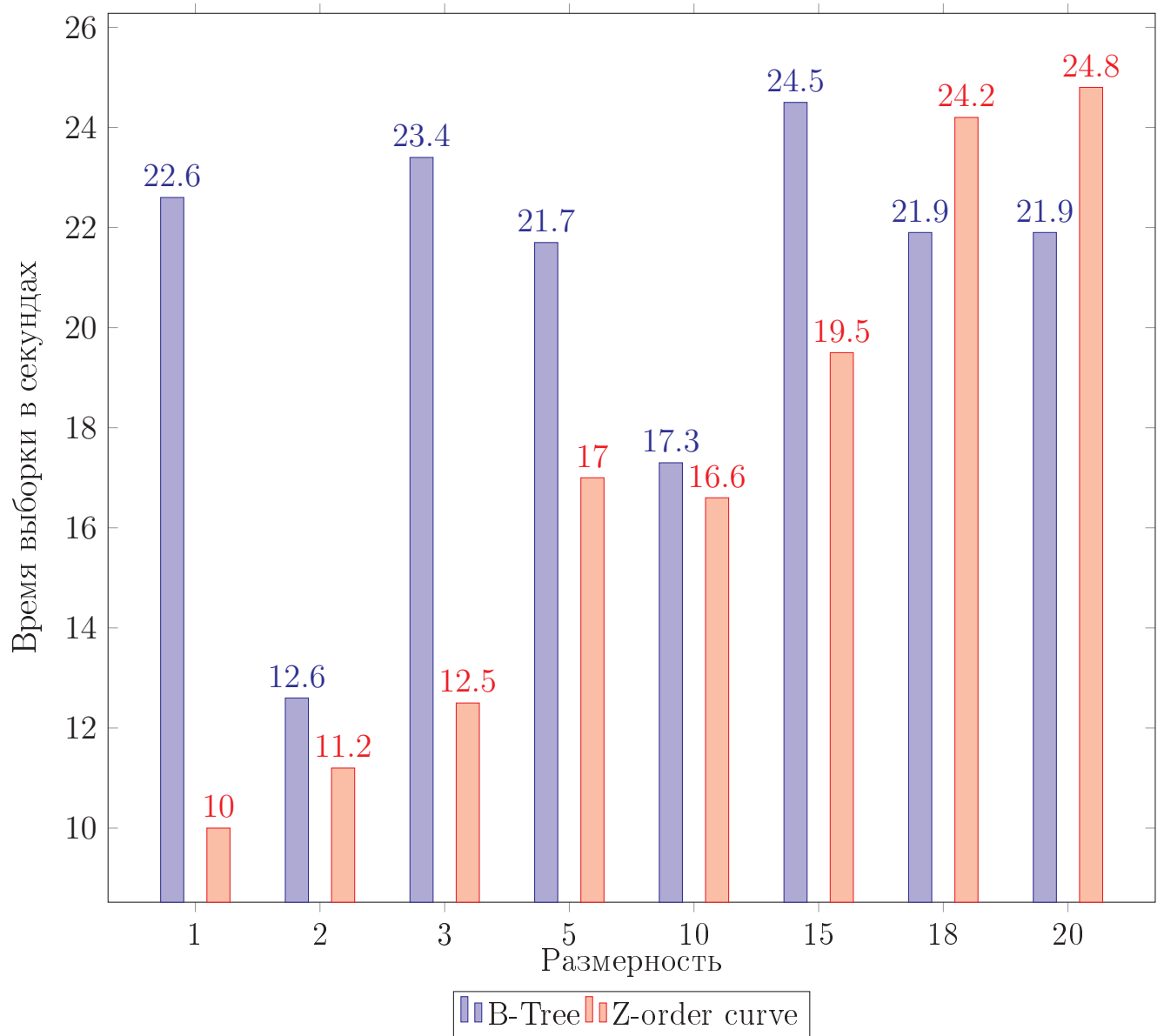


Рисунок 4.3.1 — Сравнение хранения в JSON-формате (plain text). И message pack (binary)

Список литературы

1. *Bayer, R.* Organization and maintenance of large ordered indexes / R. Bayer, E. McCreight // Software pioneers. — Springer, 2002. — С. 245—262.
2. The log-structured merge-tree (LSM-tree) / P. O’Neil [и др.] // Acta Informatica. — 1996. — Т. 33, № 4. — С. 351—385.
3. Integrating the UB-tree into a database system kernel. / F. Ramsak [и др.] // VLDB. T. 2000. — 2000. — С. 263—272.
4. MD-HBase: A scalable multi-dimensional data infrastructure for location aware services / S. Nishimura [и др.] // Mobile Data Management (MDM), 2011 12th IEEE International Conference on. Т. 1. — IEEE. 2011. — С. 7—16.
5. *Slayton, Z.* Z-Order Indexing for Multifaceted Queries in Amazon DynamoDB: Part 1 / Z. Slayton. — 2017. — URL: <https://aws.amazon.com/ru/blogs/database/z-order-indexing-for-multifaceted-queries-in-amazon-dynamodb-part-1/> (дата обр. 07.10.2018).
6. *Slayton, Z.* Z-order indexing for multifaceted queries in Amazon DynamoDB: Part 2 / Z. Slayton. — 2018. — URL: <https://aws.amazon.com/ru/blogs/database/z-order-indexing-for-multifaceted-queries-in-amazon-dynamodb-part-2/> (дата обр. 07.10.2018).
7. *Widhopf-Fenk, R. J.* Advanced Concepts and Applications of the UB-tree : дис. ... канд. / Widhopf-Fenk Robert Josef. — Technische Universität München, 2005.
8. *Prukl, A.* A relational approach to indexing / A. Prukl. — 2007.
9. *Furuhashi, S.* MessagePack / S. Furuhashi // URL: <https://msgpack.org>. — 2013.