МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования

**«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г.ЧЕРНЫШЕВСКОГО»**

Кафедра математического и компьютерного моделирования

Реляционные и нереляционные базы данных

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 4 курса 411 группы

направление 01.03.02 — Прикладная математика и информатика

механико-математического факультета

Пыниной Юлии Дмитриевны

Научный руководитель
доцент, к.ф.-м.н., доцент _____ С.П. Шевырев

Зав. кафедрой
зав. каф., д.ф.-м.н., доцент _____ Ю.А. Блинков

Саратов 2025

Введение Современные информационные системы требуют эффективного хранения и обработки данных. Две основные технологии управления базами данных — SQL (реляционные) и NoSQL (нереляционные) — предлагают разные подходы к решению этих задач. В последние годы активно развиваются технологии интеллектуального анализа данных - OLAP и Data Mining. Этому способствует как потребность конечного пользователя в отображении результатов анализа данных, так и увеличивающаяся мощность и быстродействие компьютеров, способных обрабатывать массивы данных за секунды. При этом стоит отметить, что OLAP (OnLine Analytical Processing)оптимизирована для эффективной работы с данными, организованными в соответствии с распространенной многомерной моделью, применяемой в массивах данных. A Data Mining - это концепция, совокупность технологий, направленная на анализ данных, хранящихся как в хранилищах данных так и в реляционных моделях.

Актуальность данной работы заключается в том, как в настоящее время изучение возможностей SQL и NoSQL для анализа рынка энергетиков актуально как с теоретической (развитие методов обработки данных), так и с практической точки зрения (оптимизация бизнес-процессов). Особую значимость приобретает в условиях цифровизации экономики и роста объемов информации.

Цель данной бакалаврской работы является проведение сравнительного анализа SQL и NoSQL, а также продемонстрировать их применение для исследования рынка энергетических напитков в России с использованием OLAP (SQL) и Data Mining (NoSQL). Бакалаврская работа состоим из четырех разделов.

Бакалаврская работа состоим из четырех разделов.

В первом разделе приводится краткое описания основных понятий баз данных и описание технологии NoSQL и SQL. База данных - это совокупность связанных данных, организованных по определенным правилам, предусматривающим общие принципы описания, хранения и манипулирования, независимая от прикладных программ. База данных является информационной моделью предметной области. Обращение к базам данных осуществляется с помощью системы управления базами данных (СУБД). СУБД обеспечивает

поддержку создания баз данных, централизованного управления и организации доступа к ним различных пользователей. Основными конструктивными элементами модели хранения данных в БД является:

- сущность — любой различимый объект (объект, который мы можем отличить от другого), информацию о котором необходимо хранить в базе данных;

- атрибут — поименованная характеристика сущности;

- ключ — минимальный набор атрибутов, по значениям которых можно однозначно найти требуемый экземпляр сущности;

- связь — ассоциирование двух или более сущностей.

СУБД бывают двух видов: реалиционные и нереалиционные.

Нереалиционные СУБД или технология NoSQL — это термин, обозначающий ряд подходов, направленных на реализацию хранилищ баз данных, имеющих существенные отличия от моделей, используемых в традиционных реляционных СУБД с доступом к данным средствами языка SQL.

Технологии NoSQL появились при попытке решить проблему реалиционных баз данных, а именно горизонтальную масштабируемость, то есть невозможность наращивать производительность путем добавления новых вычислительных узлов к уже работающим. Поэтому большинству NoSQL-систем имеют распредел нную архитектуру, которая решает проблему горизонтальной масштабируемости и увеличивает надежность системы с помощью поддержания нескольких копий данных.

Технологии NoSQL не подразумевают внутренних связей, которые позволяют снять ограничения с формирования сущностей и допускают хранения данных в виде ключ-значение.

Если реалиционные СУБД базируются на требованиях ACID к транзакционной системе: атомарности, согласованности, изолированности, надежности, то NoSQL ориентируется на наборе свойств BASE:

- базовая доступность, т.е каждый запрос гарантированно завершается (успешно или безуспешно);

- гибкое состояние, т.е состояние системы может изменяться со временем, даже без ввода новых данных, для достижения согласования данных;

- согласованность в конечном счете, т,е данные могут быть некоторое время рассогласованы, но приходят к согласованию через некоторое время.

Технологии NoSQL присуще:

- Применение различных типов хранилищ.

- Возможность разработки базы данных без задания схемы.

- Использование многопроцессорности.

- Линейная масштабируемость (добавление процессоров увеличивает производительность).

- Инновационность: не только SQLї открывает много возможностей для хранения и обработки данных.

- Сокращение времени разработки.

- Скорость: даже при небольшом количестве данных конечные пользователи могут оценить снижение времени отклика системы с сотен миллисекунд до миллисекунд.

Технология NoSQL обладает нереляционными моделями данными, которые значительно проще, чем модели классической реляционной модели, и свои способы для осуществления запросов, которые существенно отличаются от традиционных баз данных.

Основываясь на модели данных и подходах к распределенности и репликации можно выделить основные классы технологии NoSQL:

1. Системы ключ-значение (Key-Value Stores).

В данном типе NoSQL-систем хранятся как структурированные, так и неструктурированные данные и доступ к ним осуществляется при помощи уникального ключа, который является единственным, без поддержки вторичных ключей и индексов. Так же здесь может поддерживаться некоторая структура данных, позволяющая менять отдельные поля объекта, но не позволяющая строить по ним запросы. Работа с данными обычно осуществляется с помощью простых операций вставки, удаления и поиска по ключу. В этом отношении системы ключ-значение похожи на популярную распределtнную систему кэширования в оперативной памяти Memcached, но предоставляют постоянное хранение данных и ряд дополнительных возможностей.

2. Документно-ориентированые (Document Stores).

Документо-ориентированные базы данных в отличие от систем класса ключ-значениеї у документоориентированные СУБД предоставляют намного больше возможностей. Еще одним отличием документо-ориентированные

СУБД от систем типа ключ-значение является то, что документные СУБД могут запрашивать коллекции документов на основании нескольких ограничений на атрибуты, могут осуществлять агрегатные запросы, сортировку результатов, поддерживают индексы на полях документов и т.д. Также документноориентированные системы поддерживают поиск по полям документов, индексы, часто допускаются вложенные документы и массивы, при этом схемы данных, которая предопределена заранее, нет. Единицей хранения данных в таких системах является документ - некоторый объект, обладающий произвольным набором атрибутов (полей), который может быть представлен, например, в JSON.

3. Системы хранилищ семейств колонок (Extensible Record Stores / Wide Column Stores / Column Families).

Системы хранилищ семейств колонок основной идей такого класса систем NoSQL является хранение данных не по строкам, как это принято в реляционных СУБД, а по колонкам. Это означает, что с точки зрения пользователя данные представлены как обычно в виде таблиц, но физически эти таблицы являются совокупностью колонок, каждая из которых по сути представляет собой таблицу из одного поля. Физически эти данные хранятся последовательно друг другу. Такая структура хранения данных означает, что при выполнении запроса на чтение, в котором фигурируют только 2 поля из 20 полей таблицы, реально будут прочитаны только 2 колонки. Это означает что нагрузка на канал ввода/вывода будет приблизительно в 10 раз меньше чем при выполнении такого же запроса в реляционной СУБД.

4. Графовые базы данных.

Графовая база данных - это база данных, предназначенная для хранения данных в виде графа. Т.е. это хранения данных по вершинам и ребрам. По определению, графовая база данных - это любая структура для хранения, где взаимосвязанные элементы соединены без применения индекса. Смежные по структуре элементы доступны при разыменовании физического показателя. Существует несколько типов графов, которые могут хранить: от однотипного ненаправленного графа до гиперграфа, включая свойственные подграфы.

Во втором разделе описывается технологии OLAP И Data Mining .

OLAP (Online Analytical Processing) - представляет собой технологию многомерного анализа данных, которая позволяет пользователям получать быстрые ответы на сложные аналитические запросы.

Основные характеристики OLAP-систем:

Многомерные структуры данных:

- Использование OLAP-кубов для хранения и обработки данных;
- Возможность работы с данными по различным измерениям;
- Поддержка иерархической структуры данных.

Операции анализа данных:

- Срез (Slice) - извлечение данных по одному измерению;
- Врез (Dice) - извлечение данных по нескольким измерениям;
- Ротация (Pivot) - изменение ориентации измерений;
- Агрегация (Roll-up) - объединение данных на более высоком уровне;
- Детализация (Drill-down) - переход к более детальным данным.

Типы OLAP-систем:

- ROLAP (Relational OLAP) - работа с реляционными базами данных;
- MOLAP (Multidimensional OLAP) - использование многомерных баз данных;
- HOLAP (Hybrid OLAP) - комбинированный подход;
- DOLAP (Desktop OLAP) - настольные решения.

Ключевые преимущества:

- Высокая скорость обработки запросов;
- Интуитивно понятный интерфейс;
- Возможность создания сложных отчетов;
- Визуализация данных;
- Поддержка временных рядов.

В контексте анализа рынка энергетических напитков OLAP-технологии позволяют проводить анализ продаж по различным временным периодам, выявлять региональные особенности потребления, анализировать сезонные колебания спроса, оценивать эффективность маркетинговых кампаний, прогнозировать объемы продаж, создавать детализированные отчеты по различным параметрам.

OLAP-системы обеспечивают быстрый доступ к агрегированным данным и позволяют пользователям самостоятельно проводить анализ, не прибегая к помощи IT-специалистов, что существенно повышает эффективность принятия управленческих решений.

Технологии Data Mining, представляет собой процесс обнаружения полезных закономерностей или знаний из больших объемов данных. Это совокупность статистических методов и алгоритмов для анализа данных и выявления в них значимых закономерностей.

Основные направления Data Mining:

Классификация:

- Отнесение объектов к определенным классам;

- Алгоритмы: деревья решений, метод k-ближайших соседей, нейронные сети;

- Применение: прогнозирование покупательского поведения, оценка рисков.

Визуализация данных:

- Графическое представление результатов анализа;

- Инструменты: диаграммы, графики, тепловые карты;

- Применение: представление результатов анализа, принятие решений.

В контексте анализа рынка энергетических напитков Data Mining позволяет выявлять группы потребителей со схожими предпочтениями, определять закономерности в покупательском поведении, прогнозировать спрос на различные виды напитков,выявлять сезонные колебания потребления, анализировать эффективность рекламных кампаний, определять оптимальные ценовые категории, выявлять необычные паттерны потребления.

Процесс Data Mining включает следующие этапы:

- Подготовка данных;

- Выбор методов анализа;

- Построение моделей;

- Оценка качества моделей;

- Внедрение и использование.

Data Mining позволяет извлекать ценную информацию из больших массивов данных, что помогает принимать более обоснованные управленческие решения и разрабатывать эффективные стратегии развития бизнеса.

В третьем разделе представлена информация о сравнении SQL и NoSQL баз данных.

Структура данных для SQL:

- Строгая реляционная модель;
- Таблицы с предопределенной схемой;
- Жесткие связи между данными;
- Типизированные поля.

Структура данных для NoSQL:

- Гибкая схема данных;
- Хранение в виде документов, графов или пар "ключ-значение";
- Отсутствие жестких связей;
- Возможность хранения разнородных данных.

Производительность и масштабируемость для SQL:

- Вертикальное масштабирование;
- Высокая производительность для структурированных данных;
- Ограничения при обработке больших объемов.

Производительность и масштабируемость для NoSQL:

- Горизонтальное масштабирование;
- Высокая производительность при работе с большими данными;
- Линейное увеличение производительности при добавлении узлов.

Гибкость и удобство использования для SQL:

- Мощный язык запросов;
- Сложные аналитические запросы;
- Хорошая поддержка для сложных связей.

Гибкость и удобство использования для NoSQL:

- Простой язык запросов;
- Легкость внесения изменений в структуру;
- Хорошая работа с неструктурированными данными.

Области применения для SQL:

- Финансовые системы;

- Системы учета;

- Приложения с жесткими связями между данными;

- Структурированные данные.

Области применения для NoSQL:

- Социальные сети;

- Системы с большими данными;

- Приложения с гибкой структурой;

- Хранение медиаконтента;

- Мобильные приложения.

Ключевые преимущества: SQL-базы - (надежность, целостность данных, сложные запросы, хорошая документация)

NoSQL-базы - (масштабируемость, гибкость, высокая производительность, экономия ресурсов)

Выбор между SQL и NoSQL зависит от конкретных требований проекта, объема данных, необходимости в масштабировании и сложности бизнес-логики. Часто эффективным решением становится комбинирование обеих технологий для достижения оптимального результа.

В четвертом разделе представлено практическое исследование рынка энергетических напитков.

Сравнительный анализ SQL и NoSQL, При анализе данных об употреблении энергетических напитков в России были получены следующие результаты: SQL-система показала высокую эффективность в (анализе продаж по регионам, отслеживании сезонных колебаний спроса, выявлении корреляции между ценами и объемами продаж и ормировании отчетности), а NoSQL-система продемонстрировала преимущества в (обработке неструктурированных данных о потребительском поведении, анализе отзывов и комментариев пользователей, обработке больших объемов данных в реальном времени, масштабировании при росте объема данных).

На основе проведенного исследования выявлены:

- Региональные особенности потребления энергетических напитков;

- Сезонные колебания спроса;

- Демографические характеристики основных потребителей;

- Предпочтения по брендам и вкусам;

- Тенденции развития рынка.

Для анализа рынка энергетических напитков рекомендуется:

- Использовать SQL - системы для структурированного анализа продаж и финансовой отчетности;

- Применять NoSQL-системы для обработки неструктурированных данных и анализа потребительского поведения;

- Комбинировать технологии для получения наиболее полной картины рынка;

- Использовать OLAP для многомерного анализа данных.

- Применять методы Data Mining для выявления скрытых закономерностей

Практическая значимость работы заключается в возможности применения разработанных методик и полученных результатов для оптимизации бизнес-процессов в сфере анализа рынка энергетических напитков.

Основное предназначения хранилищ данных - это представление пользователям информации для статистического анализа и принятия управленческих решений. Они должны обеспечивать высокую скорость добычи данных, а также достоверность и полноту информации. Data Mining - это ключевой компонент анализа данных, хранящихся в хранилищах данных. Это не конкретная технология, а совокупность технологий, направленная на анализ данных, хранящихся как в хранилищах данных, так и в реляционных моделях.

В рамках данной бакалаврской работы были рассмотрены технологии OLAP и Data Mining: понятия, их классификации, некоторые виды OLAP-систем, преимущества и недостатки всех способов анализа. А также были показаны способы применения технологий добычи данных.

В данной работе были рассмотрены преимущества и недостатки SQL и NoSQL систем при анализе больших данных, разработана комплексная методика анализа рынка с использованием OLAP и Data Mining, продемонстрирована эффективность применения современных технологий баз данных для решения бизнес-задач а также, сформированы рекомендации по оптимизации процессов обработки данных в сфере анализа рынка энергетиков.

Практическая значимость работы заключается в возможности применения разработанных методик и полученных результатов для оптимизации бизнес-процессов в сфере анализа рынка энергетических напитков.