УДК 004.65

КЛЮЧЕВЫЕ АСПЕКТЫ ВЫБОРА СУБД ДЛЯ АНАЛИТИКИ БОЛЬШИХ ДАННЫХ

Н.В. Павлович
Магистрант кафедры проектирования информационно-компьютерных систем БГУИР
mpaulovich00@gmail.com

И.Н. Тонкович
Доцент кафедры проектирования информационно-компьютерных систем БГУИР, кандидат технических наук, доцент
intonkovich@gmail.com

В.Ф. Алексеев
Доцент кафедры проектирования информационно-компьютерных систем БГУИР, кандидат технических наук, доцент
alexvikt.minsk@gmail.com

Н.В. Павлович

Окончил БГУИР (2021 г.), в настоящее время является магистрантом этого университета. Проводит научные исследования в области распределенной обработки больших массивов данных.

И.Н. Тонкович

Окончила Белорусский государственный университет, факультет прикладной математики. Основная область научных интересов связана с применением инновационных подходов в системе высшего образования, разработкой методов и алгоритмов построения информационно-компьютерных систем, организацией учебного и научно-исследовательского процессов в техническом университете.

В.Ф. Алексеев

Окончил Минский радиотехнический институт. Область научных интересов связана с разработкой методов и алгоритмов построения информационно-компьютерных систем, организацией учебного и научно-исследовательского процессов в техническом университете.

Аннотация. В данной работе представлены ключевые аспекты выбора различных типов СУБД для аналитики больших данных. Рассмотрены критерии и факторы, влияющие на выбор СУБД. Выявлены особенности использования реляционных, нереляционных, графовых, колоночных и In-memory СУБД для аналитики Big Data.

Ключевые слова: аналитика, большие данные, СУБД для аналитики, реляционные, нереляционные, графовые, колоночные, In-memory.

Введение.

Аналитика больших данных становится одним из ключевых направлений в современном бизнесе. По оценкам ResearchAndMarkets мировой рынок аналитики больших данных будет увеличиваться ежегодно в среднем на 11,9% с 2020 по 2028 годы. Такой рост проектов по внедрению систем бизнес-аналитики Big Data приведет к серьезному спросу на специализированные СУБД для обработки и анализа больших объемов данных. Выбор СУБД для аналитики Big Data является критически важной задачей, для решения которой необходимо поддерживать быстрые итерации для правильного анализа данных, учитывать способность обрабатывать структурированные и неструктурированные наборы данных, возможность масштабирования, поддержки параллельной обработки и распределенных вычислений, возможность работы с различными форматами данных, а также способность обеспечивать безопасность и защиту данных.

В данном исследовании рассмотрены ключевые аспекты выбора СУБД для аналитики Big Data.

Критерии и факторы, влияющие на выбор СУБД для аналитики Big Data.

Выделим ряд общих критериев, которые следует учитывать при выборе СУБД для аналитики Big Data [1]:

1 Размер данных. СУБД должна быть способна обрабатывать большие объемы данных, а также поддерживать их хранение.

2 Скорость обработки данных. СУБД должна быть способна обрабатывать данные быстро и эффективно.

3 Сложность запросов. В зависимости от сложности запросов и аналитических операций необходимо выбирать СУБД с соответствующими функциональными возможностями.

4 Надежность и доступность. СУБД должна обеспечивать высокую надежность и доступность, чтобы обеспечить постоянный доступ к данным.

5 Стоимость внедрения и использования. Стоимость СУБД также является важным фактором при выборе, поскольку одни СУБД могут быть дороже, чем другие.

6 Совместимость с существующими инструментами. СУБД должна быть совместима с существующими инструментами аналитики и инфраструктурой.

7 Поддержка и развитие. СУБД должна иметь поддержку и обновления, а также продолжать развиваться и адаптироваться к новым технологиям.

Кроме того, при выборе СУБД для аналитики Big Data также важно учитывать ее тип (реляционная, NoSQL, NewSQL и т.д.), способность к параллельной обработке, возможность масштабирования и другие факторы.

Существуют различные типы СУБД, которые могут быть использованы для анализа Big Data. Каждый тип СУБД имеет свои особенности и применяется в зависимости от характеристик проекта и требуемых функциональных возможностей, типа данных и характера анализа. Более того, аналитика больших данных может быть проблематичной, поскольку она часто включает сбор и хранение различных типов неструктурированных данных или используется для обработки непрерывного потока данных в режиме реального времени.

Выделяют следующие типы СУБД для аналитики Big Data:

1 Реляционные СУБД (RDBMS) – это классические СУБД, которые используются для хранения и управления структурированными данными. Они могут использоваться для анализа больших данных, но часто не обладают достаточной масштабируемостью для обработки Big Data в режиме реального времени.

2 NoSQL СУБД – используются для хранения и обработки неструктурированных данных, таких как данные социальных сетей, IoT-данные и т.д. NoSQL СУБД более гибкие и масштабируемые, чем реляционные СУБД, что делает их популярным выбором для аналитики Big Data.

3 Графовые СУБД – как правило, используются для анализа связей между объектами в данных, таких как социальные сети, транспортные системы и т.д. Графовые СУБД могут быстро находить связи между данными, что делает их очень полезными для аналитики Big Data.

4 Колоночные СУБД – используются для хранения и обработки больших объемов структурированных данных. Колоночные СУБД могут обрабатывать большие объемы данных очень быстро и обладают высокой степенью сжатия данных, что делает их эффективными инструментами для аналитики Big Data.

5 In-memory СУБД – они используются для обработки данных в оперативной памяти компьютера, что обеспечивает высокую скорость обработки данных. In-memory СУБД обычно используются для анализа больших объемов данных в режиме реального времени.

Каждый тип СУБД имеет свои преимущества и недостатки, и выбор конкретного типа СУБД для аналитики Big Data зависит от требований проекта и доступных ресурсов.

Особенности использования реляционных СУБД для аналитики Big Data.

Реляционные СУБД, такие как MySQL, Oracle и Microsoft SQL Server, долгое время были основным инструментом для работы с данными в больших компаниях. Они хорошо подходят

для обработки структурированных данных и отлично работают с небольшими и средними объемами информации.

Однако, с появлением Big Data, объемы и сложность обрабатываемых данных значительно возросли. Реляционные СУБД не всегда эффективно работают с такими объемами информации, и могут стать узким местом при анализе данных.

Тем не менее, реляционные СУБД все еще могут использоваться для аналитики Big Data, особенно в тех случаях, когда данные имеют ясную структуру и формат.

Для увеличения производительности и эффективности работы реляционных СУБД с Big Data, используются различные инструменты, такие как партиционирование, шардирование, репликация данных и т.д. Также в реляционных СУБД могут быть добавлены функции аналитики, такие как OLAP и Data Warehousing.

Однако, в целом, для работы с Big Data лучше использовать специализированные СУБД, которые оптимизированы для работы с большими объемами информации и могут обрабатывать неструктурированные данные.

С точки зрения практического внедрения или использования выделяют следующие наиболее популярные реляционные СУБД для аналитики Big Data:

1 Oracle Database – это одна из самых популярных реляционных СУБД в мире. Она обладает широким спектром возможностей для анализа данных и имеет хорошую масштабируемость для обработки больших объемов данных.

2 Microsoft SQL Server – это реляционная СУБД, разработанная Microsoft. Она имеет множество функций и возможностей для анализа данных, включая встроенный анализ данных и машинное обучение.

3 IBM Db2 – это еще одна популярная реляционная СУБД, которая используется для анализа Big Data. Она предлагает высокую производительность и масштабируемость, а также широкий набор функций и инструментов для анализа данных.

4 PostgreSQL – это бесплатная и открытая реляционная СУБД, которая может быть использована для анализа Big Data. Она имеет множество расширений и плагинов, которые могут быть использованы для расширения ее функциональности.

5 MySQL – это популярная бесплатная реляционная СУБД. Хотя она может иметь некоторые ограничения для обработки больших объемов данных, ее можно использовать для анализа небольших или средних объемов данных.

Это лишь некоторые из множества реляционных СУБД, которые могут быть использованы для аналитики Big Data. Выбор конкретной СУБД будет зависеть от конкретных потребностей и требований организации.

Особенности использование нереляционных СУБД для аналитики Big Data.

Нереляционные СУБД, также известные как NoSQL базы данных, стали популярны в последние годы благодаря их способности обрабатывать большие объемы неструктурированных данных с высокой скоростью и масштабируемостью. Они используют различные модели данных и алгоритмы хранения, которые отличаются от традиционных реляционных баз данных.

Особенности использования нереляционных СУБД для аналитики Big Data:

1 Высокая масштабируемость. Нереляционные СУБД могут обрабатывать большие объемы данных и масштабироваться на несколько серверов, обеспечивая горизонтальное масштабирование.

2 Гибкость в модели данных. Нереляционные СУБД не требуют определения жесткой схемы базы данных, что облегчает добавление и изменение данных.

3 Более быстрый доступ к данным. Нереляционные СУБД могут обеспечить быстрый доступ к данным при использовании определенных алгоритмов хранения, таких как хранение в оперативной памяти или индексирование.

4 Поддержка неструктурированных данных. Нереляционные СУБД могут обрабатывать данные различных форматов, таких как документы, графы и временные ряды.

Нереляционные СУБД делятся на несколько категорий в зависимости от используемой модели данных, таких как ключ-значение, документоориентированные, столбцовые, графовые и др. Каждая категория подходит для конкретного типа данных и типичных запросов в аналитике Big Data.

Наиболее известными представителями нереляционных СУБД для аналитики Big Data являются [2]:

1 MongoDB – документоориентированная СУБД, которая хранит данные в формате JSON-подобных документов. Она хорошо подходит для работы с большими объемами неструктурированных данных.

2 Apache HBase – СУБД, работающая на основе Hadoop и HDFS. HBase подходит для хранения больших объемов данных с высокой производительностью в режиме реального времени.

3 Apache Couchbase – распределенная NoSQL СУБД, которая поддерживает хранение данных в оперативной памяти и на диске. Она может обрабатывать большие объемы структурированных и неструктурированных данных в режиме реального времени.

4 Amazon DynamoDB – управляемая NoSQL СУБД, которая хранит данные в виде ключ-значение. Она обеспечивает высокую доступность, производительность и масштабируемость при работе с большими объемами данных.

Использование графовых и колоночных СУБД для аналитики Big Data.

Графовые СУБД – это тип нереляционных СУБД, которые используют графовую модель данных для хранения и обработки информации. Графы представляют собой набор вершин (узлов), соединенных ребрами (связями), которые могут иметь различные свойства и атрибуты.

В аналитике Big Data графовые СУБД используются для анализа связей и зависимостей между данными, таких как социальные сети, транспортные сети, сети связи и т.д.

Одной из наиболее известных графовых СУБД для аналитики Big Data является Neo4j. Она представляет собой гибридную СУБД, которая сочетает в себе возможности графовой модели данных и языка запросов Cypher с традиционными реляционными возможностями, такими как ACID-транзакции и индексирование.

Графовые СУБД обычно имеют высокую производительность и масштабируемость для анализа больших объемов данных. Однако, они могут быть менее удобными для анализа данных, которые не являются связанными или зависимыми.

Колоночные СУБД являются специальным типом нереляционных СУБД, предназначенных для работы с большими объемами данных. Они используются для хранения и анализа структурированных данных, которые обычно представлены в виде больших таблиц. В колоночных СУБД данные хранятся в виде колонок, в отличие от реляционных СУБД, где они хранятся в виде строк.

Колоночные СУБД используются для аналитики Big Data, так как они обеспечивают быстрый доступ к большим объемам данных, позволяют проводить анализы в режиме реального времени и поддерживают параллельную обработку данных. Кроме того, они обладают высокой степенью сжатия данных, что позволяет существенно снизить объем хранимых данных.

Наиболее популярными решениями колоночных СУБД для аналитики Big Data являются:

1 Vertica – колоночная СУБД, разработанная для аналитической обработки больших объемов данных. Она поддерживает распределенную архитектуру и реализует функциональность параллельной обработки запросов, что позволяет обеспечить быстродействие и масштабируемость.

2 Amazon Redshift – облачная колоночная СУБД, разработанная для аналитической обработки больших объемов данных. Она использует распределенную архитектуру и высокопроизводительное хранилище данных, что позволяет обеспечить высокую скорость выполнения запросов и масштабируемость.

3 Google BigQuery – облачная колоночная СУБД, которая предоставляет масштабируемое хранилище данных и высокопроизводительный механизм выполнения запросов. Она использует технологии облачных вычислений, что позволяет обеспечить высокую доступность и масштабируемость.

4 Apache Cassandra – распределенная колоночная СУБД, разработанная для обработки больших объемов данных с высокой доступностью и масштабируемостью. Она использует распределенную архитектуру и реализует механизм репликации данных, что обеспечивает высокую отказоустойчивость.

5 ClickHouse – колоночная СУБД, разработанная для обработки больших объемов данных с высокой производительностью. Она использует распределенную архитектуру и механизмы параллельной обработки запросов, что позволяет обеспечить высокую скорость выполнения запросов и масштабируемость.

Использование In-memory СУБД для аналитики Big Data.

In-memory СУБД (IMDB) представляют собой базы данных, в которых данные хранятся в оперативной памяти компьютера. Такой подход позволяет быстро обрабатывать большие объемы данных, так как операции чтения и записи осуществляются непосредственно в памяти компьютера, а не на диске. Использование IMDB для аналитики Big Data позволяет существенно ускорить обработку данных и повысить скорость принятия решений.

IMDB часто используются в комбинации с другими СУБД, например, с колоночными или нереляционными СУБД, для обработки больших объемов данных. Кроме того, некоторые производители СУБД предлагают гибридные решения, которые объединяют в себе функциональность различных типов СУБД, включая IMDB.

Использование IMDB для аналитики Big Data может быть особенно полезным в случаях, когда требуется быстрый доступ к большим объемам данных, например, для аналитики рынка, финансового анализа или обработки больших объемов транзакций [3].

Топовые In-memory СУБД, которые можно использовать для аналитики Big Data:

1 SAP HANA – высокопроизводительная In-memory СУБД, которая может использоваться для анализа больших объемов данных в режиме реального времени.

2 Oracle TimesTen – In-memory СУБД, предназначенная для обработки транзакций и анализа данных в режиме реального времени.

3 IBM SolidDB – In-memory СУБД с открытым исходным кодом, которая обеспечивает быстрый доступ к данным и возможность анализа больших объемов данных.

4 MemSQL – In-memory СУБД с поддержкой SQL, которая может использоваться для анализа больших объемов данных в режиме реального времени.

5 Apache Ignite – распределенная In-memory СУБД, которая может использоваться для анализа больших объемов данных, распределенных по нескольким узлам.

6 Aerospike – In-memory СУБД, которая может использоваться для анализа больших объемов данных в режиме реального времени и обеспечивает высокую доступность и масштабируемость.

7 Altibase – In-memory СУБД, которая может использоваться для анализа больших объемов данных в режиме реального времени и предоставляет поддержку SQL и NoSQL.

Заключение.

В заключении можно отметить, что выбор СУБД для аналитики Big Data является важным и ответственным шагом. Необходимо учитывать ряд критериев, таких как тип данных, производительность, масштабируемость и прочее. Для аналитики Big Data могут использоваться различные типы СУБД, такие как реляционные, нереляционные, гибридные, графовые, колоночные и In-memory СУБД. Каждый тип имеет свои преимущества и недостатки, и выбор должен основываться на конкретных потребностях бизнеса.

Список литературы

[1] Выбор базы данных для аналитики [Электронный ресурс]. – Режим доступа: https://habr.com/ru/post/487498/.

[2] NoSQL [Электронный ресурс]. – Режим доступа: https://www.bigdataschool.ru/wiki/nosql.

[3] Использование In-memory в Big Data [Электронный ресурс]. – Режим доступа: https://www.bigdataschool.ru/blog/tarantool-use-cases-and-advantages.html.

KEY ASPECTS OF CHOOSING A DBMS FOR ANALYTICS BIG DATA

N.V. Paulovich
Master student of the Department of Information and Computer Systems Design BSUIR

I.N. Tonkavich
Associate Professor, Department of Information Computer Systems Design, PhD of Chehnical sciences, Associate Professor

V.F. Alekseev
Associate Professor, Department of Information Computer Systems Design, PhD of Technical sciences, Associate Professor

Department of Information and Computer Systems Design
Faculty of Computer Engineering
Belarusian State University of computer science and Radio Electronics, Republic of Belarus
E-mail: mpaulovich00@gmail.com, intonkovich@gmail.com, alexvikt.minsk@gmail.com

Abstract. This paper presents the key aspects of choosing different types of DBMS for big data analytics. Criteria and factors influencing the choice of DBMS are considered.

The features of using relational, non-relational, graph, column and In-memory DBMS for Big Data analytics are revealed.

Keywords: analytics, big data, DBMS for analytics, relational, non-relational, graph, column, In-memory.