УДК 681.3
DOI 10.5281/zenodo.13999777

ВНЕДРЕНИЕ ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ ПО АНАЛИТИКЕ БОЛЬШИХ ДАННЫХ

IMPLEMENTATION OF BIG DATA ANALYTICS SOFTWARE

КОРТЕНКО ЛЮДМИЛА ВАСИЛЬЕВНА,
кандидат экономических наук, доцент,
Уральский государственный экономический университет.
СЫРОПЯТОВ МАКСИМ ВЯЧЕСЛАВОВИЧ,
заместитель директора по информационным технологиям,
Уральский государственный экономический университет,
Гимназия № 9, г. Екатеринбург.
ТАЙБОЛИН АЛЕКСАНДР НИКОЛАЕВИЧ,
бизнес-аналитик,
Уральский государственный экономический университет,
Сбербанк.

KORTENKO LYUDMILA VASILYEVNA,
Candidate of Economic Sciences, Associate Professor,
Ural State University of Economics.
SYROPYATOV MAXIM VYACHESLAVOVYCH,
Deputy Director for Information Technology,
Ural State University of Economics,
Gymnasium № 9.
TAYBOLIN ALEXANDER NIKOLAYEVICH,
Business analyst,
Ural State University of Economics,
Sberbank.

В статье рассмотрена ведущая массивно-параллельная СУБД с открытым кодом. Выделены преимущества описываемой системы управления базами данных при работе с большими объёмами данных. Проведен сравнительный анализ систем GreenPlum и PostgreSQL. В результате предложено внедрение новых систем управления базами данных (СУБД), способных справляться с поставленными массивными задачами. Сделан вывод, что проведенный анализ преимуществ и недостатков массивно-параллельной СУБД с открытым исходным кодом GreenPlum позволяет рекомендовать ее крупным организациям, обрабатывающим в повседневной деятельности нарастающие объемы данных или исходной информации.

The article discusses the leading massively parallel open source DBMS. The advantages of the described database management system when working with large amounts of data are highlighted. A comparative analysis of the GreenPlum and PostgreSQL systems has been conducted. As a result, the introduction of new database management systems (DBMS) capable of coping with massive tasks has been proposed. It is concluded that the analysis of the advantages and disadvantages of the massively parallel open source database GreenPlum allows us to recommend it to large organizations that process increasing amounts of data or source information in their daily activities.

Ключевые слова: *информационные технологии, массивно-параллельная СУБД, серверы, большие объемы данных, open source, big data.*

Key words: *information technology, massively parallel DBMS, servers, big data, open source.*

Информационные технологии работы с большими объемами данных развиваются стремительными темпами, применяемые IT-специалистами и пользователями информации (сотрудниками компаний) инструменты усложняются и оптимизируются. Развитие данного направления – затратный и трудоемкий процесс для широкого спектра IT-специалистов по написанию программ и инженерным решениям. При детальном анализе бизнес-процессов компаний авторами выявлено, что специалисты по информационным технологиям не только выявляют преимущества и недостатки возможных бизнес-решений, участвуют в определении вектора развития деятельности компаний, но и способствуют определению содержания их финансовой и экономической политик [1; 5].

В рамках повседневной работы крупной организации возможен переход на массивно-параллельную СУБД GreenPlum с открытым исходным кодом, совместимую с PostgreSQL и предназначенную для обработки больших объемов данных и выполнения сложных аналитических запросов с использованием технологии MPP (Massive Parallel Processing – от англ. *«вычислений с массовым параллелизмом»*). GreenPlum может масштабироваться до десятков терабайт, обеспечивать строгую консистентность данных, гарантируя их согласованность в различных узлах системы. Это делает ее идеальным выбором для задач, требующих высокой производительности и надежности при работе с большими объемами данных.

Архитектура GreenPlum (как и большей части других аналитических систем, таких как Citus, ClickHouse и прочие) построена на ядре PostgreSQL, ориентированном на эффективную и быструю работу с разнообразными аналитическими нагрузками [2; 3]. Кластеризация перечисленных выше систем неизбежна, поскольку одна машина не в состоянии масштабироваться вверх до бесконечности. Для обеспечения эффективной работы с данными различных систем в GreenPlum изменяются методы доступа (или «access methods») к ним.

Совместимость GreenPlum и PostgreSQL следует из создания первой на базе PostgreSQL. Сервис поддерживает реляционную СУБД PostgreSQL и может осуществлять функцию единой точки для сбора информации из различных реляционных систем для ее последующей обработки и качественной аналитики.

Сервер-сегментами и мастером или главным экземпляром GreenPlum являются PostgreSQL-инстансы или резервные мастера. Это обеспечивает возможность интеграции с любым ПО, поддерживающим PostgreSQL. С целью подключения используются драйверы Postgres или драйверы самого GreenPlum, особенно, если есть необходимость использовать специфичный функционал для решения задачи. Это свойство важно, если организация планирует перемещение с уже существующей базы данных Postgres в облачные решения [6].

Общим свойством для GreenPlum и PostgreSQL является ориентация на возможность качественной работы с большими данными в результате того, что у PostgreSQL отсутствуют ограничения на максимальный размер базы данных или количество индексов, содержащихся в самой таблице. Как следствие, данное преимущество технологии обеспечивает возможность её применения в проектах Big Data [7]. Другими унаследованными чертами GreenPlum от PostgreSQL являются: высокая производительность, открытый исходный код, поддержка JSON (формата текстовых данных, используемого для обмена данными в веб- и мобильных приложениях), высокая ёмкость таблиц и полей, масштабируемость, надёжность, транзакционность.

PostgreSQL предназначен в том числе для OLTP-кейсов с использованием небольших баз данных, поскольку обеспечивает подключение к большому количеству систем обработки

транзакций в режиме реального времени. Использование MPP-технологий с архитектурой систем вычислений с массовым параллелизмом в OLAP-сценариях (характеризующихся частыми запросами на чтение данных большого количества строк и редкими запросами на добавление данных), является допустимым, но не лучшим вариантом, так как отсутствует функция сжатия данных, автоматического партиционирования, колоночного хранилища и распараллеливания запросов.

Одним из преимуществ GreenPlum является возможность параллельной обработки данных, что позволяет сервису выполнять запросы по большим наборам данных в несколько раз быстрее, чем аналогичные системы без MPP-архитектуры. Это является ее ключевым отличием от PostgreSQL, которая при обработке запросов использует только один многоядерный сервер. MPP-архитектура обеспечивает возможность работы GreenPlum без разделения ресурсов, разделяя только сетевую инфраструктуру. В конструкции сервиса есть мастер-хост (в нем располагается инстанс Postgres, к которому обращаются пользователи) и сегментные хосты, на которых расположены инстансы Postgres, включая и мастер-хост.

Взаимодействие между мастер-хостом и сегментными хостами осуществляется через внутреннюю сеть. Поэтому GreenPlum считается аналитической базой данных и его использование для OLTP-систем (характеризующихся непрерывной записью и чтением транзакций в реальном времени) не рекомендуется. Из-за отсутствия разделения ресурсов происходят сетевые задержки, генерируются расходы на обработку запросов на мастер-хосте и на каждом из узлов, в частности. Для запросов с небольшим временем исполнения это критически важно.

Выбор свободной системы управления базами данных из PostgreSQL и GreenPlum зависит от характера решаемой задачи.

GreenPlum оптимизирован под хранение и аналитику больших наборов данных, при этом в транзакционной среде его эффективность снижается. Серверная часть веб-приложений создает рабочую нагрузку OLTP. GreenPlum не в состоянии обеспечить больше 500-600 TPS, так как это распределённая система с большими расходами на обработку транзакций.

Таблица 1. Отличия GreenPlum и PostgreSQL

GreenPlum	PostgreSQL
Архитектура	
реализует массивно-параллельную обработку ресурсов без их разделения.	осуществляет стандартную клиент-серверную обработку ресурсов.
Сценарии использования	
обеспечивает обширную OLAP-аналитику больших данных.	предназначен для небольшого размера баз данных с OLTP-кейсами.
Перенос базы данных	
включает два планировщика запросов: – по умолчанию используется GPORCA, оптимизированный под определённые операции (например, запросы к секционированным таблицам); – для запросов с большим количеством соединений повторяет PostgreSQL.	осуществляется в графическом интерфейсе pgAdmin либо созданием, переносом и восстановлением резервной копии базы данных между серверами в зависимости от размера базы данных, допустимого времени простоя и требуемого уровня сохранения целостности транзакций и данных.

Для OLTP или баз данных до 10 ТБ более подходит для использования PostgreSQL, но она может обеспечить обработку только одного хоста без возможности разбиения на разде-

лы, осуществления сжатия и хранения столбцов. А вот GreenPlum уже способен обрабатывать данные параллельно в кластере, что обеспечивает его успешное использование при масштабном анализе данных в OLAP-сценариях, т.к. он основан на многомерном анализе данных и позволяет проводить анализ данных в различных измерениях.

Отличия GreenPlum и PostgreSQL по аспектам архитектуры, сценариям использования и переноса базы данных представлены авторами в табл. 1.

Принципиальными достоинствами GreenPlum, отличающими ее от PostgreSQL являются:

– в GreenPlum реализована опция, дифференцирующая потребление серверных ресурсов и устанавливающая ограничения для пользователя, группы, сеанса, запроса в отношении ресурсов вычислительных машин (оперативной памяти, дискового пространства, ресурсов контейнера и прочее);

– GreenPlum включает несколько методов сжатия, например, в версии «6x» поддерживаются ZSTD, ZLIB и RLE. Для каждого из методов есть возможность выбирать уровень сжатия от 1 до 9 [6];

– высокая скорость обмена данными между процессорами;

– универсальность и простота при обслуживании;

– относительно невысокая цена самого продукта.

Недостатками GreenPlum можно считать:

– необходимость использования специальной техники программирования для реализации обмена сообщениями между процессорами;

– ограниченность доступного каждому процессору объёма локального банка памяти;

– высокую стоимость ПО для массово-параллельных систем с раздельной памятью вследствие наличия представленных архитектурных недостатков, требующих большие усилия для того, чтобы максимально эффективно обеспечить использование имеющихся системных ресурсов.

Облачные технологии и Greenplum. Сравнение с Hadoop.

MPP-системы на основе PostgreSQL применяются широко, в том числе в облачных технологиях. Можно отметить Greenplum, Arenadata DB, ApsaraDB AnalyticDB for PostgreSQL. В Greenplum реализована возможность правки кода продукта как развиваемого open-source-решения [7].

В сравнении с Hadoop, Greenplum отличает поддержка ACID-транзакции. При этом использование ACID-механизмов Greenplum требует аккуратного подхода при обновлении данных.

Таким образом, в целом Greenplum – это реляционная СУБД с архитектурой MPP без разделения ресурсов. Система предназначена для хранения и обработки больших объемов данных методом их распределения и обработки запросов на нескольких серверах. Она отлично подходит для построения корпоративных хранилищ данных, решения аналитических задач, задач машинного обучения и искусственного интеллекта. Так как в основе Greenplum лежит PostgreSQL, то можно сказать, что одна СУБД Greenplum функционируют как множество PostgreSQL, что, несомненно, делает данное СУБД одним из лучших инструментов для работы с большими объемами данных, даже при наличии некоторых перечисленных выше недостатков.

Проведенный анализ преимуществ и недостатков массивно-параллельной СУБД с открытым исходным кодом GreenPlum позволяет рекомендовать ее крупным организациям, обрабатывающим в повседневной деятельности нарастающие объемы данных или исходной информации. В общей тенденции можно отметить, что объем информации, хранящейся на серверах, увеличивается и это делает обработку больших данных все более важной для

крупных федеральных компаний. Для ускорения извлечения этих данных из различных источников, например, с серверов следует учитывать, что новые системы управления базами данных (СУБД) такие как GreenPlum внедряются, так как ни один из серверов в единственном числе не может обеспечить стабильную и бесперебойную работу крупной организации при наличии большого объема данных, и каждая компания, выбирая инструмент для работы с большими объемами данных, опирается на свой опыт и доступные технологические решения, исходя из текущих условий, перспектив развития и требуемого масштабирования.

СПИСОК ЛИТЕРАТУРЫ

1. Гумирова М.Н. Сравнительный анализ систем бизнес-аналитики компаний в целях их устойчивого развития. Томск, 2021. URL: https://vital.lib.tsu.ru/vital/access/manager/Repository/vital:14-036.
2. Кислицын Е.В. Исследование рынка программных продуктов в России // Мир экономики и управления. 2019. № 2. С. 49-64.
3. Кузнецов С.Д. MapReduce: внутри, снаружи или сбоку от параллельных СУБД? // Труды Института системного программирования РАН. 2010. Т. 19. С. 35-70.
4. Панов М.А. Перспективы роста национальной и региональной экономики в условиях кризиса // Вестник ЧелГУ. 2022. № 6 (464). С. 94-105.
5. Просто о больших данных: книга / Дж. Гурвиц, А. Ньюджент, Ф. Халпер, М. Кауфман. М.: Альпина Паблишер, 2016. 58 с.
6. Связь GreenPlum и PostgreSQL. URL: https://habr.com/ru/companies/slurm/articles/682248 (дата обращения 05.09.2024).
7. СУБД Greenplum для Big Data и машинного обучения. URL: https://elibrary.ru (дата обращения: 05.09.2024).