ВЫБОР ОПТИМАЛЬНОЙ ОБЛАЧНОЙ ПЛАТФОРМЫ ДЛЯ ОРГАНИЗАЦИИ ХРАНИЛИЩА И ОБРАБОТКИ ДАННЫХ

Бессараб З. И.
Кафедра программного обеспечения информационных технологий, Белорусский государственный университет информатики и радиоэлектороники
Минск, Республика Беларусь
E-mail: zekker6@gmail.com

Данная статья рассматривает проблемы выбора провайдера облачных услуг предоставляющего возможность размещения платформы для организации хранилища и центра обработки данных.

Введение

В современных информационных системах все чаще используются механизмы анализа и принятия решений на основе использования исторических данных. Например, для определения предпочтений пользователей системы построения рекомендаций создают профиль пользователя путём сопоставления действий пользователя с аналогичными действиями, произведенными другими пользователями. Такие подходы анализа данных требуют возможности оперативного анализа и хранения больших объемов данных.

Согласно отчету об исследовании, проведенному компанией IDC [1], объем данных, созданных в следующие три года, превысит общее количество информации созданной за 30 предыдущих лет. По прогнозу об объеме прироста количества новых данных ожидается ежегодное увеличение общего объема данных не менее чем на 20% в год.

I. Проблемы использования собственных ЦОД

Интенсивность роста объема данных приводит к тому что при использовании наиболее дешевых вариантов развертывания систем хранения и анализа данных – развертывания в собственных центрах обработки данных – существенная часть работ, связанных с обеспечением бесперебойной работы системы, сводится к наращиванию мощности системы. Недостаточная скорость увеличения мощности системы может вызывать: ухудшение качества обработки данных, понижение скорости ответов системы, технические сбои во время работы системы.

Наиболее популярным подходом, позволяющим решить проблему скорости масштабирования, является использование облачных платформ. Использование публичные облачных платформ позволяет производить автоматическое масштабирование ресурсов необходимых как для хранения, так и для обработки информации. В связи с этим возникает следующая проблема: необходимо определить облачную платформу, которая будет удовлетворять требованиям к производительности, возможности масштабирования и стоимости решения до начала эксплуатации облачной системы.

II. Сравнение производительности OLAP систем

Для получения этой информации производится нагрузочное тестирование целевой платформы при помощи использования бенчмарков. Одним из наиболее известных подходов [2] к бенчмаркингу является применение бенчмарка TPC-DS предназначенного для тестирования аналитических систем анализа данных. Однако не существует единого стандарта для анализа данных, полученных в результате проведения тестирования системы с использованием данного бенчмарка. Большинство существующих исследований [3–5] рассматривают лишь два показателя: стоимость платформы и время обработки данных. Данные показатели не позволяют провести ранжирование по другим показателям, например, количеству обращений к хранилищу данных при выполнении запроса.

III. Алгоритм тестирования платформ

Для создания более универсальной системы, позволяющей уменьшить сложность принятия решения по переносу рабочей нагрузки в облачную систему, необходимо разработать программное средство, которое на основании данных о тестировании облачных систем и существующей системы позволит определить облачную платформу, подходящую к заданным входным параметрам. Для создания такого продукта необходимо описать условия для проведения тестирования, произвести тестирование, создать продукт для анализа полученных данных и ранжированию сравниваемых систем по различным характеристикам.

Предлагается производить тестирование следующим методом: с помощью бенчмарка TPC-DS создаются тестовые наборы с различными объемами данных(например, три выборки с размерами 1гб, 10гб, 100гб). Затем определяется количество симулируемых одновременно активных клиентов(например, от 10 до 50 с шагом в 10 клиентов). Для каждой пары из наборов исход-

ных данных проводится следующий тест: в течении часа поочередно запускаются запросы из тестовой выборки TPC-DS. При этом в результате работы каждого запроса должны быть получены как минимум следующие данные:

1. Объем тестовой выборки
2. Количество активных пользователей
3. Название запроса
4. Время выполнения, секунды
5. Затраченные ресурсы CPU, секунды
6. IO, кб

Благодаря полученным данным возможно построить модели нахождения зависимости между различными метриками полученными в результате работы, а также возможно вычислить функцию зависимости затраченного количества ресурсов от объема датасета и загруженности платформы. Это позволяет определять необходимое количество ресурсов для работы системы без проведения дополнительных тестов.

IV. Результаты тестирования

Данные полученные в результате данного тестирования можно использовать для выбора оптимальной облачной платформы по различным критериям сравнения. На рисунке 1 показан график относительного взвешенного среднего времени отклика трёх платформ для набора данных размером 10 ГБ.

График показывает относительное, а не абсолютное время отклика, так как в данном случае мы исследовали степень влияния увеличения количества конкурирующих пользователей платформы на время отклика и не сравнивали значения абсолютного времени отклика среди рассмотренных платформ.

По показателю времени выполнения запроса в данном объеме данных видно что платформа Vantage показывает наименьшее время отклика.

Заключение

Проведя тестирование с использованием предложенного алгоритма тестирования были получены практические результаты в виде результатов тестирования. Эти данные могут быть использованы для дальнейшего исследования показателей производительности различных платформ, сравнения архитектурных подходов платформ позволяющих показывать лучшие показатели производительности.

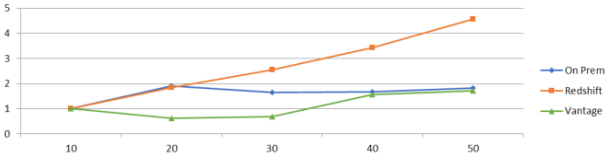

Рис. 1 – Зависимость времени отклика от количества активных пользователей

1. IDC's Global DataSphere Forecast Shows Continued Steady Growth in the Creation and Consumption of Data [Электронный ресурс]. – 2020. – Режим доступа: https://www.idc.com/getdoc.jsp?containerId=prUS46286020 – Дата доступа: 02.10.2020.
2. Meikel Poess, Raghunath Othayoth Nambiar, and David Walrath. 2007. Why you should run TPC-DS: a workload analysis. In Proceedings of the 33rd international conference on Very large data bases (VLDB '07). VLDB Endowment, 1138–1149.
3. EDW performance comparison | Grid Dynamics Blog [Электронный ресурс]. – 2020. – Режим доступа: https://blog.griddynamics.com/edw-performance-comparison/ – Дата доступа: 05.10.2020.
4. Evaluating modern data warehousing platforms with a performance per-dollar approach | West Monroe [Электронный ресурс]. – 2020. – Режим доступа: https://www.westmonroepartners.com/perspectives/point-of-view/evaluating-modern-data-warehousing-platforms-with-a-performance-per-dollar-approach – Дата доступа: 05.10.2020.
5. Cloud Data Warehouse Performance Testing | Gigaom [Электронный ресурс]. – 2020. – Режим доступа: https://gigaom.com/report/cloud-data-warehouse-performance-testing/ – Дата доступа: 05.10.2020.

Таблица 1 – Пример аггрегированных данных собранных в ходе выполнения тестирования

Объем тестовой выборки	Количество активных пользователей	Название запроса	Время выполнения, секунды	Количество выполнений	Затраченные ресурсы CPU, секунды	IO, кб
1	10	Query1	0.478	5767	2315	457549
1	10	Query2	0.795	4322	4.114	416.167
1	20	Query1	1.226	3799	2.410	102.335
1	20	Query2	1.964	3323	4.340	149.938
10	10	Query1	3.399	1165	18.108	297.692
10	10	Query2	5.032	1163	28.345	2380.264
10	20	Query1	7.497	1573	18.248	503.289
10	20	Query2	12.314	1570	28.914	670.933
100	10	Query1	34.907	123	178.049	71470.4
100	10	Query2	67.531	67	323.946	68077
100	20	Query1	58.548	107	176.095	14903.579
100	20	Query2	122.581	55	322.459	19861.106