Министерство образования Республики Беларусь
Учреждение образования
Белорусский государственный университет
информатики и радиоэлектроники

УДК 004.75

Бессараб
Захар Игоревич

Выбор оптимальной облачной платформы для организации хранилища данных Data Warehouse

АВТОРЕФЕРАТ
на соискание академической степени
магистра
по специальности 1-40 80 05 – Программная инженерия

Научный руководитель И. И. Пилецкий,
к.ф.-м.н., доцент

Минск 2021

ВВЕДЕНИЕ

Наиболее частым видом развертывания систем ранее был вариант развертывания с использованием собственной инфрастркутуры. Сюда входят варианты при развертывании в собственных помещениях, либо с использованием услуг коммерческих центров обработки данных.

Данный подход требовал больших затрат на содержание и планирование инфраструктуры, особенно для компаний относящихся к малому и среднему бизнесу. В результате увеличения спроса на услуги по управлению инфраструктурой начали появляться провайдеры облачных услуг, такие как Amazon Web Services (запущен в 2006 году), Google Cloud Platform (запущен в 2008 году), Microsoft Azure (запущен в 2010 году).

После того как решения поставщиков облачных услуг стали достаточно надежными и популярными повысился уровень доверия к таким решениям, после чего многие компании начали рассматривать возможность использования облачной инфраструктуры в качестве замены собственной. При этом первоочередной проблемой была необходимость выбора какого-либо провайдера облачных услуг, который будет удовлетворять критериям необходимым компании.

Проблема выбора поставщика облачных услуг связана с сложностью и комплекстностью мер для оценки преимуществ того или иного провайдера услуг. Для того чтобы провести всесторонний анализ каждой из предлагаемых систем необходимо затратить немалое количество ресурсов(человеческих и материальных), так же зачастую тестирование производится применимо к требованиям конкретной компании, что не позволяет переиспользовать эти результаты для других сравнений.

Для оценки экономической целесообразности важно учитывать различные виды затрат на содержание серверной инфраструктуры, такие как:

– стоимость электричества;
– стоимость обслуживания и установки систем охлаждения;
– содержание персонала для обслуживания системы;
– поддержание сетевой инфраструктуры;
– стоимость оборудования;
– резервирование систем (при необходимости обеспечения высокой доступности).

В то время как при использовании облачных услуг нет необходимости во всем вышеперечисленном так как эта задача делегируется поставщику облачных услуг. Для компании арендующей ресурсы необходимо лишь оплачивать используемые ресурсы. Это приводит к тому что одним из мотивирующих факторов является экономия денежных средств компании.

Другим немаловажным фактором является возможность экономии человеческих ресурсов. Это достигается благодаря тому что при использовании

облачных услуг управление инфраструктурой зачастую сводится к использованию API или графического интерфейса управления. Возможность экономии человеческих ресурсов так же означает что возможно выделить больше ресурсов для достижения других целей и задач компании, например увеличить скорость и качество основной разработки программных средств.

Это вызвало интерес к исследованию решений позволяющих произвести сравнение производительности различных облачных платформ в общем случае. Было принято решение сконцентрировать направление исследование на системах Data Warehouse в виду актуальности проблем обработки большого количества данных в текущее время.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Цель и задачи исследования

Целью диссертационной работы является проведение анализа применяемых моделей и алгоритмов, которые используются для оптимизации при выборе различных облачных платформ для использования систем Data Warehouse. На основе проведенного анализа предложить решение позволяющее унифицировать существующие модели для проведения тестирования различных систем.

Объектом исследования являются существующие методологии, алгоритмы, модели, используемые для сравнения производительности различных систем.

Предметом исследования являются методологии, алгоритмы, модели используемые для сравнения производительности облачных систем при использовании Data Warehouse.

Основной *гипотезой*, положенной в основу диссертационной работы, является возможность унификации процесса тестирования различных облачных систем Data Warehouse для получение достоверных результатов позволяющих сравнить их производительность.

Апробация результатов диссертации

Материалы, положенные в основу работы, докладывались и обсуждались на Международной конференции «ИТС-2020 БГУИР» (Минск, Беларусь, 2020), Международной научно-практической конференции «Big Data Minsk» (Минск, Беларусь, 2020), а также в Республиканской научно-практической конференции «Вычислительные методы, модели и образовательные технологии» (Брест, Беларусь, 2020).

Опубликованность результатов диссертации

По теме диссертации опубликовано 3 печатных работы в сборниках материалов международных научных конференций. Из них 1 работа в сборнике трудов и материалов международной конференции ИТС-2020 БГУИР, 1 работа в сборнике материалов Международной очной научно-практической конференции «Big Data Minsk 2020» и 1 работа в сборнике материалов IX Республиканской научно-практической конференции «Вычислительные методы, модели и образовательные технологии».

Структура и объем диссертации

Диссертация состоит из введения, общей характеристики работы, трех глав, заключения, списка использованных источников, списка публикаций автора и приложений. В первой главе представлен анализ предметной области DWH, выявлены основные существующие проблемы в рамках тематики исследования, показаны направления их решения. Вторая глава посвящена описанию и анализу современных облачных платформ, предоставляемых услуг и преимуществ предоставляемых ими. В третьей главе предложен алгоритм

проведения тестирования облачной платформы DWH, описан последующий результат процесса анализа полученных данных на основании тестирования.

КРАТКОЕ СОДЕРЖАНИЕ

Во введении рассмотрено современное положение и роль облачных систем в организации инфраструктуры различных решений, а также приведена проблема, которая приводит к необходимости исследования.

В первой главе проведен анализ систем Data Warehouse. Расмотрены основные направления решения задач возникающих при обработке данных, приведены примеры использования облачных технологий для улучшения качества работы DWH систем.

Во второй главе рассмотрены облачные технологии которые используются для современного построения инфрастркутуры. Рассмотрены различные виды облачных систем, методы управления, предоставления услуг. Проведен обзор облачных систем DWH предлагаемых на рынке.

В третьей главе предложена методика для произведения тестирования различных систем DWH. Выделены критерии для сравнения, описано проведение тестирования и процесс получения данных по результатам тестирования. Проведен анализ полученных данных. Подведены итоги сравнения систем участвовавших в сравнении.

ЗАКЛЮЧЕНИЕ

В ходе исследования были рассмотрены аспекты систем DWH которые важны для понимания методики измерения производительности и масштабируемости систем. Были рассмотрены облачные технологии которые позволяют повысить производительность систем DWH с помощью применения подходов реализуемых в облачной инфраструктуре.

Был предложен алгоритм для проведения тестирования различных платформ, описан процесс получения данных необходимых для анализа, произведено тестирование для платформ Amazon Redshift, Teradata On Premise и Teradata Vantage. Полученный результат тестирования был проанализирован с точки зрения поиска наиболее производительной системы, так же было проведено сравнение производительности систем одного поставщика работающих в облачной инфраструктуре и без неё.

Полученные результаты позволяют сделать вывод о масштабируемости рассмотренных платформ и влиянии на неё различных факторов. Так, лучшую масштабируемость по результатам исследований имеет платформа Amazon Redshift, что подробно рассмотрено в данной работе. Результаты для платформы Teradata On Premises оказались хуже, чем для Vantage, что обусловлено использованием технологий от одного поставщика – компании Teradata – и работой Vantage в облаке AWS.

Важным аспектом продолжения данной работы является продолжение анализа полученных данных, а так же их обогащение новой информацией. Обогащение данных информацией о стоимости каждой из тестируемых платформ позволит произвести разработку модели для предсказания необходимой конфигурации облачной платформы для удовлетворения целей бизнеса при минимальной стоимости. Также существует возможность продолжения анализа данных с целью поиска закономерностей в различных показателях работы систем, например, поиске триггеров используемых платформами в качестве сигнала к началу масштабирования.

Другим направлением развития данной работы является получение большего количества данных для уже протестированных систем, а так же проведение аналогичного тестирования на других облачных платформах для получения более точной модели предсказания производительности.

СПИСОК ПУБЛИКАЦИЙ СОИСКАТЕЛЯ

[1] Бессараб, З. И. Выбор оптимальной облачной платформы для организации хранилища и обработки данных / Бессараб З. И, // Компьютерные системы и сети: материалы международной научной конференции аспирантов, магистрантов и студентов, Минск, 18 ноября 2020 г. / Белорусский государственный университет информатики и радиоэлектроники. – Минск, 2020. – С. 151 – 152.

[2] Бессараб, З. И. Оценка масштабируемости облачных хранилищ данных / В.Н. Козуб, З.И. Бессараб, Е.Г. Гусаковская // Материалы шестой международной научно-практической конференции «BIG DATA and Advanced Analytics. BIG DATA и анализ высокого уровня», Минск, 20-21 мая 2020 г. / Белорусский государственный университет информатики и радиоэлектроники. – Минск, 2020. – С. 439 – 447.

[3] Бессараб, З. И. Проблемы миграции в облачное окружение для систем хранения и обработки данных / Бессараб З.И. // Материалы 9 Республиканской научно-практической конференции «Вычислительные методы, модели и образовательные технологии», Минск, 22 октября 2020 г. / Брестский Государственный Университет Имени А.С. Пушкина. – Брест, 2020.