

УДК [611.018.51+615.47]:612.086.2

ИСПОЛЬЗОВАНИЕ CLOUD РЕШЕНИЙ ПРИ РАБОТЕ С BIG DATA



Н.О. Дубовик
Инженер-программист
EPAM Systems

ИООО «ЭПАМ Системз», Республика Беларусь.
E-mail: dybovik.mikita@gmail.com.

Н. О. Дубовик

С отличием окончила Белорусский государственный университет информатики и радиоэлектроники. Магистрант БГУИР. Работает в EPAM в должности инженера-программиста.

Аннотация. С ростом количества обрабатываемых данных растет и необходимость в вычислительных мощностях для их обработки, а также возникает необходимость в использовании специальных инструментов или сервисов для повышения удобства и надёжности процесса анализа Big Data. Возможен вариант использования серверных мощностей компании, которая занимается данной работой, но такой способ не всегда удобен и выгоден. Cloud решения предоставляют максимально гибкую систему реагирования на изменения количества входящих данных, а также предоставляют встроенных в них сервисы для контроля за потоками данных.

Ключевые слова: Cloud, Big Data, интеграционная шина, Event Hubs, потоки данных.

Введение.

На текущий момент все больше и больше компаний переходят с on-premise [1] решений в Cloud [2], так как модель, при которой компании самой необходимо приобретать оборудование при увеличении количества запросов на сервера или из-за увеличивающегося количества поступающих данных, работает только до тех пор, пока можно спрогнозировать необходимое изменение мощностей. То есть пока сохраняется некий тренд на количество необходимого серверного оборудования. В ситуации, когда такой тренд невозможно спрогнозировать или прогноз оказался не верен из-за, например, изменения конъюнктуры рынка, что поспособствовало уменьшению внимания к приложению или некой системе, то компания будет терпеть убытки.

Для предотвращения такой проблемы Cloud сервисы предоставляют возможность гибко настраивать как количество машин, необходимых для обработки данных, так и их мощности, то есть физические характеристики, например количество оперативной памяти машины или версию процессора. Существует множество различных Cloud решений, как от небольших компаний, так и от IT-гигантов таких как Google, Amazon, Microsoft, Alibaba [3]. Возьмем облако Microsoft Azure [4] для рассмотрения того, как с помощью Cloud можно ускорить работу с Big Data, все остальные крупные облака-конкуренты имеет схожий функционал поэтому выбор Cloud не так важен, но акцентирование внимания на одном из них позволит взглянуть на решение с практической точки зрения.

Материалы и методы.

В ходе анализа эффективности использования Cloud решений при работе с Big Data использовалось финансовое приложение, в котором количество активных пользователей составляет один миллион пользователей, которые ежедневно формируют и отправляют финансовые отчетности. Приложение написано на языке C#, с использованием Angular для отображения данных клиентам. Сравнение производилось в отношении on-premise решения, которое предоставляет компания Tele2.

Для обработки такого большого количества данных использовались встроенные возможности

Azure по масштабированию, как по принципу scale-up [5], так и по системе scale-out [6], рисунок 1, а также использовалась интеграционная шина Service Bus [7] и специальный сервис отложенной доставки Big Data Event Hubs [8].

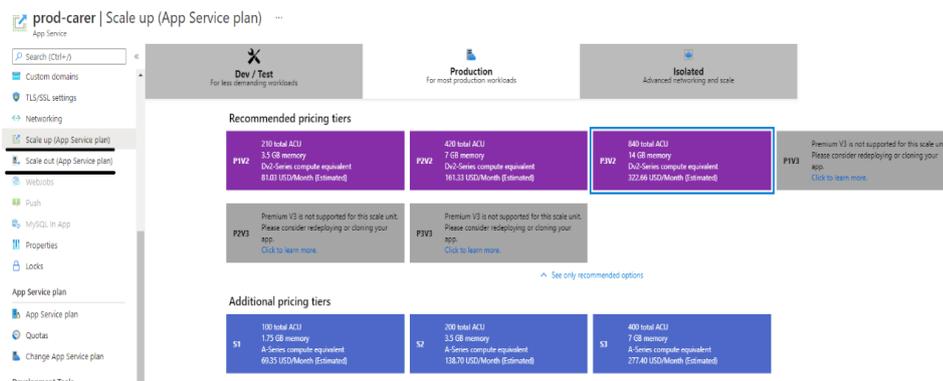


Рисунок 1. Использование встроенных возможностей Azure для обработки Big Data

Приложение работает по микросервисной архитектуре [9], с главным компонентом, который представляет из фасад для пользовательских запросов, соответственно этот компонент должен иметь максимальные пропускные возможности, чтобы справиться с потоками данных, но даже при максимально возможной комплектации машины ее может не хватать, поэтому в пиковые часы загрузки создаются дублирующие версии машины для перераспределения нагрузки. Все остальные сервисы настраиваются в соответствии с нагрузкой, регулируемой с помощью Application Insights [10]. В свою очередь, Service Bus позволяет связать все сервисы воедино и контролировать консистенцию данных.



Рисунок 2. Финансовые затраты на различные среды выполнения приложения

Для анализа данных пользователей и выявления аномалий используется Event Hubs, который аккумулирует в себе поступающие данные и потом отправляет их в Cognitive Service [11], способный сделать отчет о данных.

Результаты.

Первичные итерации показали, что конечная развернутая prod like инфраструктура в Azure будет иметь меньшую стоимость, чем аренда серверов Tele2, но суммарная стоимость всех сред, необходимых для разработки приложения, будет дороже, рисунок 2.

Таким образом, финансовая выгода от использования Cloud напрямую зависит от того, как быстро произойдет переход с on-premise решения, а также от того насколько активен процесс разработки новых усовершенствований для приложения.

С точки зрения пользовательского опыта [12], за счет распределенной системы, а также наличия большого количества дата-центров Azure время обработки большинства пользовательских запросов уменьшилось с 1 секунды до 3 миллисекунд, что позитивно сказалось на пользовательском мнении.

Для разработчиков при использовании Cloud решений необходимо иметь базовую подготовку и знания о том, какие сервисы и инструменты предлагаются, а также их стоимость, чтобы

контролировать расходы.

Заключение.

Можно заключить, что использование Cloud решений превосходят on-premise решения при работе с Big Data, так как они позволяют добиться большей эффективности при меньших расходах, а также улучшить пользовательский опыт при работе с приложениями. Тем не менее, финансовые затраты в процессе переноса инфраструктуры могут превысить изначальные затраты на инфраструктуру on-premise, что необходимо учитывать при планировании бюджета.

Еще одной проблемой могут стать законы о пользовательских данных. В некоторых странах необходимо хранить определенный набор данных о действиях пользователей на территории этих стран, из-за этого может возникнуть необходимость в дублировании данных в базы других облаков, у которых есть дата-центры на территории этих стран, что в свою очередь скажется на финансовых затратах.

Список литературы

- [1] Techland // What is On Premises [Электронный ресурс]. – Режим доступа: <https://www.techslang.com/definition/what-is-on-premises/> Дата доступа: 06.04.2021.
- [2] Citrix // What is a Cloud Service [Электронный ресурс]. – Режим доступа: <https://www.citrix.com/glossary/what-is-a-cloud-service> Дата доступа: 06.04.2021.
- [3] ZDnet // Top cloud providers in 2021 [Электронный ресурс]. – Режим доступа: <https://www.zdnet.com/article/the-top-cloud-providers-of-2021-aws-microsoft-azure-google-cloud-hybrid-saas/> Дата доступа: 07.04.2021.
- [4] Microsoft // Microsoft Azure [Электронный ресурс]. – Режим доступа: <https://azure.microsoft.com/en-us/> Дата доступа: 07.04.2021.
- [5] Microsoft // Scale up an app in Azure App Service [Электронный ресурс]. – Режим доступа: <https://docs.microsoft.com/en-us/azure/app-service/manage-scale-up> Дата доступа: 07.04.2021.
- [6] Microsoft // Scale out to multiple instances [Электронный ресурс]. – Режим доступа: <https://docs.microsoft.com/en-us/azure/azure-monitor/autoscale/autoscale-get-started?toc=/azure/app-service/toc.json> Дата доступа: 08.04.2021.
- [7] Microsoft // Service Bus queues, topics, and subscriptions [Электронный ресурс]. – Режим доступа: <https://docs.microsoft.com/en-us/azure/service-bus-messaging/service-bus-queues-topics-subscriptions> Дата доступа: 08.04.2021.
- [8] Microsoft // Event Hubs [Электронный ресурс]. – Режим доступа: <https://azure.microsoft.com/en-us/services/event-hubs/> Дата доступа: 08.04.2021.
- [9] Microservices // What are microservices [Электронный ресурс]. – Режим доступа: <https://microservices.io/> Дата доступа: 09.04.2021.
- [10] Microsoft // What is Application Insights [Электронный ресурс]. – Режим доступа: <https://docs.microsoft.com/en-us/azure/azure-monitor/app/app-insights-overview> Дата доступа: 09.04.2021.
- [11] Microsoft // Azure Cognitive Services [Электронный ресурс]. – Режим доступа: <https://azure.microsoft.com/en-us/services/cognitive-services/> Дата доступа: 09.04.2021.
- [12] Interaction Design Foundation // User Experience (UX) Design [Электронный ресурс]. – Режим доступа: <https://www.interaction-design.org/literature/topics/ux-design> Дата доступа: 09.04.2021.

USING CLOUD SOLUTIONS DURING WORKING WITH BIG DATA

M.A. DUBOVİK

Software engineer epam systems

EPAM Systems, Republic of Belarus

E-mail: dybovik.mikita@gmail.com

Abstract. As the amount of data being processed increases, so does the need for computing power to process it, as well as the need to use special tools or services to improve the convenience and reliability of the Big Data interaction and analysis process. It is possible to use the server capacity of the company that is engaged in this work, but this method is not always convenient and profitable. Cloud solutions provide the most flexible system for responding to changes in the amount of incoming data, as well as provide built-in services for monitoring data flows.

Keywords: Cloud, Big Data, Integration bus, Event Hubs, data streams.