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

УДК 004.65:681.51

Федосенко
Егор Константинович

МЕТОДЫ АВТОМАТИЗАЦИИ ПРОЦЕССОВ И ИНФРАСТРУКТУРА ДЛЯ УПРАВЛЕНИЯ БОЛЬШИМИ ДАННЫМИ

АВТОРЕФЕРАТ

На соискание степени магистра технических наук
по специальности 1-23 80 08 – Психология труда, инженерная психология, эргономика

Научный руководитель
Е.А. Криштопова, кандидат
технических наук, доцент

Минск 2020

ВВЕДЕНИЕ

Инфраструктура для управления большими данными обеспечивает хранение и обработку информации с устройств пользователей для дальнейшего анализа и принятия правильных решений в контексте многих задач.

Большие данные – это термин, который описывает большой объем данных, как структурированных, так и неструктурированных, которые ежедневно наполняют бизнес. Но важен не объем данных, а то, что организации делают с данными. Большие данные могут быть проанализированы для понимания, которые приводят к лучшим стратегическим и деловым решениям.

В наши дни компании используют большие данные, чтобы превзойти своих конкурентов. В большинстве отраслей существующие лидеры рынка и новые участники будут использовать стратегии, основанные на проанализированных данных, чтобы конкурировать и вводить новшества. Большие данные помогают организациям оценивать ситуацию на рынке анализируя поведение пользователей. Эти организации имеют достаточную информацию о продуктах и услугах, покупателях и поставщиках, предпочтениях потребителей, которая может быть зафиксирована и проанализирована.

Со стороны внедрения функционала больших данных появляется огромное количество технических проблем. Одной из таких проблем является выбор технологий, который бы подходил компании. Должны учитываться многие критерии, такие как надежность хранения и скорость обработки данных. В последнее время, в результате стремительно набирающих популярность облачных разработок, данный запрос успешно удовлетворяют такие сервисы, как Amazon Web Services.

После этапа выбора технологий требуется грамотное проектирование инфраструктуры. При проектировании инфраструктуры должно быть учтено как удобство взаимодействия с интерфейсом для управления процессами, так и возможность автоматизации. Цена ошибки человека-оператора должна быть минимальной, а основные концепции, такие как объемность, скорость и разнообразие, не должны быть нарушены.

Целью работы является разработка системы взаимодействия с большими данными в облачном сервисе Amazon Web Services.

Для реализации цели необходимо решить следующие задачи:

– выбрать подходящие сервисы, с помощью которых можно построить корректную систему обработки больших данных; соблюдая основные принципы данной концепции;

– построить инфраструктуру, которая при взаимодействии разных сервисов использует все их преимущества;

– разработать алгоритмы взаимодействия пользователей с системой;

– автоматизировать процессы сбора и хранения данных.

Таким образом, система для управления большими данными должна быть гибкой, решать задачи компании быстро и надежно, без ущерба качества. Риски, связанные с неработоспособностью системы в результате ошибки человека-оператора, должны быть минимизированы, путем внедрения методов автоматизации.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Ключевые слова:

Amazon Web Services, S3, Lambda, Apache Hive, EMR, Cloud Watch, Step Functions, SNS.

Цель работы – разработка эргономичной инфраструктуры взаимодействия пользователя с большими данными в облачном сервисе Amazon Web Services.

Задачи работы:

1. Проанализировать методы автоматизации процессов и инфраструктуры для управления большими данными.

2. *Разработать инфраструктуры для управления большими данными*

3. Провести т*естирование интерфейса взаимодействия пользователя с инфраструктурой для управления большими данными и устранить выявленные по итогам тестирования недостатки.*

Разработка данной системы должна обеспечить возможность удобного и надежного способа сбора и хранения данных.

Объектом исследования является инфраструктура для управления большими данными.

Предметом исследования является взаимодействие пользователя с инфраструктурой для управления большими данными.

Проведен сравнительный анализ существующих методов и подходов взаимодействий с большими данными внутри локальных и облачных систем. Были выявлены ключевые недостатки и проблемы локального сбора и хранения больших данных. В результате анализа было принято решение использовать облачные сервисы для разработки и автоматизации таких систем.

Для построения инфраструктуры был выбран следующий комплекс сервисов:

− AWS S3 как сервис для хранения информации;

− Serverless лямбда-функции для взаимодействия с сервисами AWS;

− EMR (Elastic Map Reduce) как сервис обработки данных при помощи кластеров;

− Cloud Watch как сервис автоматизации вызовов Lambda функций;

− Step Functions как сервис построения последовательности вызова лямбда-функций;

− SNS (Simple Notification Service как сервис для нотификации пользователей.

В рамках разработки была реализована автоматизированная инфраструктура для управления большими данными. Также был разработан интерфейс, который позволяет осуществлять последовательные и параллельные вызовы шагов обработки и записи данных.

Результаты магистерской диссертации представлены на 53-й и 56-й научно-технических конференциях аспирантов, магистрантов и студентов БГУИР.

Глава 1

Целью диссертации является разработка системы управления большими данными при помощи облачных сервисов AWS

Рассмотренные сервисы для работы с большими данными позволяют осуществлять все эти операции частично, или по отдельности. Не все из этих систем могут предоставить гибкий доступ к функционалу, быстрый переход между компонентами системы, низкую стоимость готового программного продукта. Одним из недостатков в рассмотренных системах была выявлена невозможность автоматизации рутинных операций пользователя внутри системы, таких как вызовы последовательных запросов. Помимо этого, из-за ряда факторов, влияющих на автоматизацию, большое количество времени системы тратят свои ресурсы не оптимально.

Облачные сервисы имеют избыточный функционал, пытаясь охватить все виды деятельности в компании и, тем самым, устраняя необходимость в использовании других приложений, вытесняя конкурентов. С одной стороны, это является преимуществом, т.к. сотрудникам не надо переключаться между разными приложениями, с разным дизайном и структурой, что может на несколько минут снизить производительность, а с другой стороны, недостаток заключается в том, что для клиента может быть неприемлемым использование нового программного продукта для общения и получения информации. К тому же, не всем компаниям в работе необходим широкий набор функциональных возможностей.

Таким образом, разрабатываемая система должна обладать следующими свойствами:

– универсальность: возможность использования в разных сферах сотрудниками, имеющими разный уровень компьютерных знаний и навыков;

– простота и удобство интерфейса: обеспечение легкости в его изучении и в использовании;

–хорошая навигация по сайту: работа с системой не должна вызывать у пользователя сложностей в поиске необходимых директив (элементов интерфейса) для управления процессом решения поставленной задачи;

–привлекательность: дизайн приложения не должен отталкивать пользователей;

–доступность: доступ информации с любого устройства при наличии подключения к сети Интернет.

Разрабатываемая система должна позволять решать следующие задачи:

– хранение данных;

– обработка больших объемов данных

– оперативное оповещение о форс-мажорных ситуациях;

– предоставление гибкой системы доступа: передача прав на управление системой другим пользователям;

– быстрый поиск по уже сгенерированным отчетам;

– доступность к отчетам на различных устройствах;

– быстрый доступ к последней полученной информации;

– возможность сохранения данных в архив для последующего анализа и получения статистики за длительный период времени.

Также облачные сервисы должны быть совместимы для корректной работы внутри системы. Это следует учесть при проектировании и разработке инфраструктуры для управления большими данными.

Автоматизация системы должна быть выполнена по следующим направлениям:

– автоматизация последовательных задач пользователя по управлению большими данными;

– оповещение пользователя о результатах работы системы, путем отправления отчетности по средствам электронной почты.

Глава 2

В результате проектирования и разработки системы была получена система для хранения и обработки больших данных.

Данная система спроектирована и выполнена при помощи использования следующих ресурсов:

− AWS S3 как сервис для хранения информации;

− Serverless лямбда-функции для взаимодействия с сервисами AWS;

− EMR (Elastic Map Reduce) как сервис обработки данных при помощи кластеров;

− Cloud Watch как сервис автоматизации вызовов Lambda функций;

− Step Functions как сервис построения последовательности вызова лямбда-функций;

− SNS (Simple Notification Service как сервис для нотификации пользователей.

Система обработки и хранения больших данных имеет ряд преимуществ относительно не облачных систем, таких как:

–универсальность: возможность использования в разных сферах сотрудниками, имеющими разный уровень компьютерных знаний и навыков;

–простота и удобство интерфейса: обеспечение легкости в его изучении и в использовании;

–хорошая навигация по сайту: работа с системой не должна вызывать у пользователя сложностей в поиске необходимых директив (элементов интерфейса) для управления процессом решения поставленной задачи;

–доступность: доступ информации с любого устройства при наличии подключения к сети Интернет.

Ключевым преимуществом использования облачных сервисов, таких как Amazon Web Services, является эффективность использования вычислительных ресурсов. Поскольку данная система является динамической, то время, при котором вычислительные машины находятся в режиме ожидания сводится к минимуму.

Благодаря облачному сервису S3 была осуществлена система надежного хранения данных. Данный факт важен, поскольку проблема потери данных является очень частой в локальных системах управления большими данными.

Лямбда-функции внутри облачных сервисов AWS позволяют совмещать использование различными сервисами, при помощи библиотеки Boto. Данное преимущество помогает осуществлять вызовы внутри различных сервисов, используя одну лямбда-функцию, что облегчает работу пользователя.

Автоматизация системы выполнена по следующим направлениям:

– автоматизация последовательных задач пользователя по управлению большими данными;

– оповещение пользователя о результатах работы системы, путем отправления отчетности по средствам электронной почты.

Были рассмотрены пошаговые функции как сервис для построения последовательностей вызова внутри AWS. Сервис AWS Step functions является эффективным инструментом для автоматизации вызовов лямбда-функций. Это позволяет при наличии большого количества ресурсов минимизировать вероятность ошибки, которую может сделать человек-оператор.

В контексте управления большими данными технология пошаговых функций широко применяется, поскольку она помогает экономить время человека-оператора, которое тратится на рутинные вызовы внутри системы.

Из-за проблемы большого времени отработки кластеров могут тратится часы на мониторинг процессов внутри системы человеком оператором. Данную проблему помог решить принцип уведомления пользователя.

Внутри облачных сервисов AWS были рассмотрены функциональности сервисов Simple Email Service (SES) и Simple Notification Service (SNS). Оба сервиса успешно выполняют свои задачи, однако SNS является более гибким инструментом и позволяет выполнять нотификацию не только на электронную почту, но и на любые устройства, имеющие доступ к интернету.

В результате автоматизации процессов были решены следующие задачи:

– уменьшено время мониторинга пользователем системы управления большими данными;

– уменьшено количество ручных вызовов пользователем;

– увеличена детальность отчетности о состоянии системы для человека-оператора.

Глава 3

В данной главе были выполнено юзабилити-тестирование системы хранения и обработки больших данных. Юзабилити-тестирование позволило сделать программный продукт более удобным в использовании, тем самым не только повышая эффективность работы конечных пользователей и бизнес-процессов в целом, но и улучшая впечатление от взаимодействия с системой.

Поскольку система для хранения и обработки больших данных имеет направленность на узкий круг специалистов, которые имеют компетенции для изменения и детальной конфигурации сервисов, то наиболее оптимальным методом для юзабилити тестирования был выбран экспертный подход.

Для тестирования были использованы эвристики Якоба Нильсена и знания экспертов, полученные в результате опыта работы с облачными сервисами AWS.

В результате тестирования были найдены следующие проблемы:

– отсутствует возможность выполнять отдельные задачи внутри системы с пропуском всех остальных

– недостаточно информации о состоянии пошаговой функции при нотификации пользователя

– если требуется запуск системы для обработки данных, то это может быть тяжело реализуемо для пользователя, поскольку требуется ожидание выполнения каждого кластера

– в случае, когда задача, выполненная в параллели пошаговой функции, завершается с ошибкой, вся система завершает работу

Данные проблемы были решены путем улучшения инфраструктуры, а именно: добавлением лямбда-функций и изменением конфигурации пошаговой функции.

После устранения выявленных недостатков было произведено повторное юзабилити тестирование, которое показало, что данные изменения помогли улучшить степень удовлетворенности взаимодействия пользователя и системы.

ЗАКЛЮЧЕНИЕ

Анализ статических локальных сервисов для хранения и обработки показал, что не все из этих систем могут предоставить гибкий доступ к функционалу, быстрый переход между компонентами системы, низкую стоимость готового программного продукта. Одним из недостатков в рассмотренных системах была выявлена невозможность автоматизации рутинных операций пользователя внутри системы, таких как вызовы последовательных запросов. Помимо этого, из-за ряда факторов, влияющих на автоматизацию, большое количество времени системы тратят свои ресурсы не оптимально.

Наличие данных недостатков обусловило необходимость разработки инфраструктуры для управления большими данными используя облачные сервисы. В процессе проектирования данной системы были выбраны сервисы, которые поддерживают концепцию динамического выделения ресурсов.

Динамическая инфраструктура была разработана при помощи использования следующих сервисов Amazon Web Services: EC2, S3, EMR. Для автоматизации системы использовались сервисы: Step functions, Lambda, SNS. В качестве языка для написания лямбда-функций был использован Python и библиотека boto3 для взаимодействия с ресурсами AWS.

После разработки данной инфраструктуры было произведено тестирование интерфейса взаимодействия пользователя с инфраструктурой для управления большими данными. В результате тестирования были выявлены ошибки, которые были позже исправлены путем расширения функциональности инфраструктуры, а именно добавления лямбда-функций и пошаговых функций. После исправления ошибок системы было выполнено

дополнительное юзабилити-тестирование, которое показало хорошие результаты.

Таким образом была спроектирована, разработана и протестирована эргономичная инфраструктура взаимодействия пользователя с большими данными в облачном сервисе Amazon Web Services.

Результаты магистерской диссертации представлены на 53-й и 56-й научно-технических конференциях аспирантов, магистрантов и студентов БГУИР.