УДК 004.75

ОСОБЕННОСТИ ОБРАБОТКИ БОЛЬШИХ ДАННЫХ В РЕАЛЬНОМ ВРЕМЕНИ В ОБЛАКЕ AWS С ИСПОЛЬЗОВАНИЕМ СЕРВИСА AWS KINESIS

К.О. Климов
Магистрант БГУИР, специалист по сопровождению ПО EPAM Systems

Г. А. Пискун
Заместитель декана факультета компьютерного проектирования по научной работе, кандидат технических наук, доцент

Д.В. Лихачевский
Декан факультета компьютерного проектирования, кандидат технических наук, доцент

В.Ф. Алексеев
Доцент кафедры проектирования информационных компьютерных систем, кандидат технических наук, доцент

В.В. Шаталова
Заместитель декана факультета компьютерного проектирования по учебно-методической работе, кандидат технических наук, доцент

Белорусского государственного университета информатики и радиоэлектроники, Республика Беларусь.
ИООО «ЭПАМ СИСТЕМЗ», Республика Беларусь.
E-mail: klimovkostya5@gmail.com.

К. О. Климов
Окончил Белорусский государственный университет информатики и радиоэлектроники. Магистрант БГУИР. Работает в EPAM Systems в должности специалиста по сопровождению ПО. Проводит научные исследования систем Интернета-вещей.

Г. А. Пискун
Заместитель декана факультета компьютерного проектирования по научной работе, канд. техн. наук, доцент.

В. Ф. Алексеев
Доцент кафедры проектирования информационных компьютерных систем, кандидат технических наук, доцент.

В. В. Шаталова
Заместитель декана факультета компьютерного проектирования по учебно-методической работе, канд. техн. наук, доцент.

Д. В. Лихачевский
Декан факультета компьютерного проектирования, кандидат технических наук, доцент.

Аннотация. Объем данных, генерируемый различными источниками, растет с каждым годом. К 2020 году объем хранимых данных увеличился до 59 зеттабайт, к 2025 этот объем увеличится в 3 раза [1]. По этой причине аналитика всех этих данных позволяет выявлять ценную информацию, благодаря чему бизнес может определять тенденции и прогнозировать показатели с высокой точностью. Максимальную выгоду можно получить, анализируя

эти данные в режиме реального времени и реагируя незамедлительно. Чтобы достигнуть таких целей, требуются огромные вычислительные ресурсы для обработки всех поступающих данных. Использование облачных вычислений значительно упростило аналитику такого объема данных.

Ключевые слова: Amazon Web Services, AWS, AWS Kinesis, аналитика больших данных, обработка данных.

Введение.

В быстроизменяющемся мире незамедлительная реакция на критические события может сэкономить десятки тысяч долларов или спасти чью-то жизнь. Огромное количество данных генерируется разными источниками ежесекундно. Анализируя эти данные, можно предсказывать проблемы и реагировать на них в реальном времени, избегая и предупреждая серьёзные последствия. До недавнего времени еженедельные и ежемесячные аналитические отчеты полностью покрывали запросы бизнеса.

Однако в последнее время произошел сдвиг в сторону аналитики в реальном времени, где данные обрабатываются по мере их поступления. Эти изменения обусловлены несколькими факторами: расширение доступа к источникам данных и повышение вычислительной мощности в результате использования облачных вычислений [2].

Методы и алгоритмы обработки данных в облаке AWS.

Существует огромное количество примеров, в которых бизнес может извлечь выгоду, используя аналитику данных в реальном времени:

– отслеживание поведения клиентов на веб-сайтах;
– выявление трендов в интернет-магазинах по мере продажи товаров;
– мониторинг данных о дорожном движении для улучшения маршрутов или формирования предупреждений о пробках.

Для осуществления аналитики в реальном времени AWS предлагает решение на базе AWS Kinesis Data Analytics для создания SQL запросов, которые непрерывно считывают и обрабатывают данные в режиме, близком к реальному времени [3].

В качестве примера будем отслеживать движение на автомагистралях Бельгии. Данные предоставляются с помощью открытого API Фламандского института дорожного движения, который предоставляет загруженность трафиком на автомагистралях. Визуализация ниже показывает состояние трафика на отдельных участках (рисунок 1).

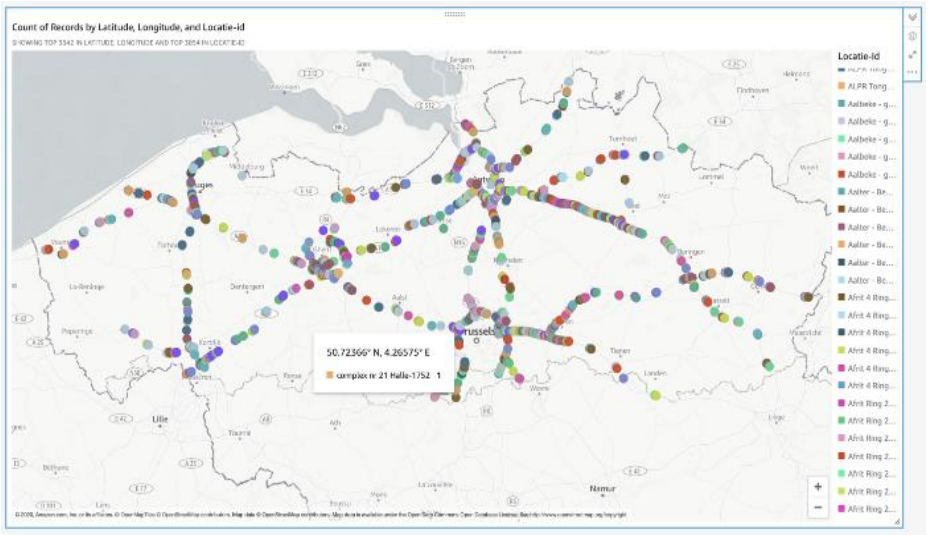

Рисунок 1. Визуализация данных API

Нашей задачей является установление для нескольких критических мест во Фландрии следующих показателей:

– отправка оповещения при возникновении пробки;

– анализ того, возникает или исчезает пробка;
– получение статуса текущей ситуации в режиме реального времени.

Данные о трафике публикуются в виде *.xml файла, который содержит информацию для всех 4500 мест измерения во Фландрии. Эти данные будут преобразованны в формат JSON и разделены на события измерения в каждом местоположении. Эта предварительная обработка достигается с использованием функций AWS Lambda. Затем измерения по каждому местоположению передаются потоком в Firehose.

Данные из Firehose используется в качестве входных данных для Kinesis Data Analytics, который будет предоставлять информацию в реальном времени [4]. Далее результаты аналитики в реальном времени посредствам Kinesis Data Analytics отправляются в Kinesis Data Stream, который может запускать AWS Lambda для создания оповещений о новых пробках или сохранять в DynamoDB (рисунок 2).

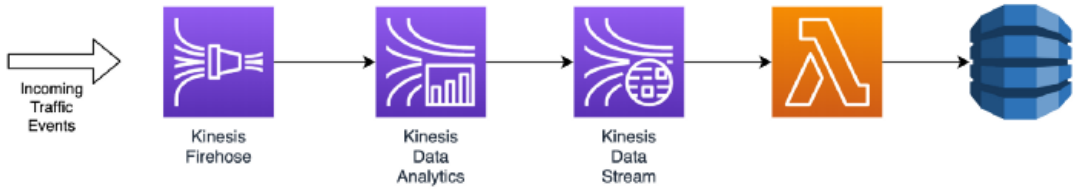

Рисунок 2. Поток обработки данных

Данные, поступающие в Firehose, содержат:
– отметку времени, указывающую, когда было проведено измерение;
– уникальный идентификатор места измерения;
– информацию о состоянии измерительного датчика;
– информацию о скорости движения транспортных средств определённых классов. Эти классы представляют тип транспортного средства, например, грузовой автомобиль, легковой, автобус.

Рассмотрим подробнее Kinesis Firehose, который является источником данных. Kinesis Firehose – это сервис, работающий в реальном времени, который может загружать данные в хранилище данных или сервис аналитики и может масштабироваться автоматически в зависимости от поступающих запросов. Следует учитывать, что данные обрабатываются в режиме близком к реальному времени: данные поступают с минимальными интервалами в 60 секунд или 1MiB.

Важно отметить, что существует два варианта потоковой передачи данных в сервисе Kinesis: Kinesis Firehose и Kinesis Data Stream [5]. В текущей архитектуре используется Kinesis Firehose. Firehose позволяет обрабатывать 5000 операций записи в секунду. Firehose имеет дополнительное преимущество, заключающееся в том, что он может сохранять исходные данные в S3, позволяя построить хранилище данных для последующей пакетной обработки.

Для аналитики потока данных в реальном времени используется Kinesis Data Analytics.

Kinesis Data Analytics – это сервис анализа потоковых данных в реальном времени с помощью SQL или интегрированных Java-приложений. Для создания аналитических запросов будет использован язык SQL. В настройках сервиса укажем источник данных – Kinesis Firehose (рисунок 3).

Для подготовки данных будет использоваться функция AWS Lambda, которая будет получать пакеты событий и преобразовать их один за одним. Рассмотрим SQL запрос для обработки получаемых данных (листинг 1).

Рисунок 3. Выбор источника данных в AWS Kinesis

Листинг 1. SQL запрос для обработки получаемых данных.

```
CREATE OR REPLACE STREAM «INCOMING_STREAM» (
«uniqueId» INTEGER,
«speed» INTEGER,
«bezettingsgraad» INTEGER,
«recordTimestamp» TIMESTAMP);
CREATE OR REPLACE PUMP «INCOMING_STREAM_PUMP» AS INSERT INTO «INCOMING_STREAM» SELECT STREAM

«unieke_id»,
«voertuigsnelheid_rekenkundig_klasse2»,
TO_TIMESTAMP (CAST (»tijd_waarneming» AS BIGINT) * 1000) AS «recordTimestamp».
FROM «SOURCE_SQL_STREAM_001»;
```

Запрос создаёт промежуточный поток под названием INCOMING_STREAM, далее создаётся PUMP для пересылки данных в промежуточный поток и затем определяется запрос, который отделит данные из промежуточного потока.

Далее рассмотрим создание запроса-окна, который может агрегировать данные за определённый момент времени. Существует два вида оконных запросов: временные и строковые и 3 их типа: пошаговые, поворотные и скользящие.

Скользящие запросы – это непрерывно агрегирующий запрос с использованием фиксированного интервала времени или количества строк. Поворотный запрос – непрерывный агрегирующий запрос, использующий определённые временные окна, которые открываются и закрываются через регулярные интервалы. Пошаговые окна используются в тех случаях, когда связанные записи не попадают в одно и то же ограниченное по времени окно.

В архитектуре приложения используется запрос со скользящим окном, чтобы узнать среднюю скорость на участке за последние x минут. Ниже приведён пример запроса (листинг 2).

Листинг 2. SQL запрос для поиска средней скорости за 0, 2 и 10 минут.

```
CREATE OR REPLACE PUMP «STREAM_PUMP_SPEED» AS INSERT INTO «SPEED_SQL_STREAM» SELECT STREAM

«uniqueId»,
AVG (»speed») over W0,
AVG (»speed») over W2,
AVG (»speed») over W10.
FROM «INCOMING_STREAM» WINDOW
W0 AS (PARTITION BY «uniqueId».
RANGE INTERVAL '0' MINUTE PRECEDING),
W2 AS (PARTITION BY «uniqueId».
RANGE INTERVAL '2' MINUTE PRECEDING),
```

W10 AS (PARTITION BY «uniqueId».
RANGE INTERVAL '10' MINUTE PRECEDING);

Для обогащения результатов добавим мета-данные, которые будут содержать идентификатор места, где были измерены данные. Названия самих мест в данные не включаются. Чтобы добавить эти данные воспользуемся следующим SQL запросом (листинг 3).

Листинг 3 .SQL запрос для поиска обогащения результатов мета-данными.
CREATE OR REPLACE PUMP «YOUR_IN_APPLICATION_STREAM» AS
INSERT INTO «YOUR_IN_APPLICATION_PUMP» (»uniqueId», «currentSpeed», , «location»).
SELECT STREAM
«sdi».»uniqueId»,
«sdi».»currentSpeed»,
...,
«ml«."locatie»,
FROM «SPEED_DIFF_INDICATOR_SQL_STREAM» AS «sdi» LEFT JOIN «measurementLocations» as «ml«.
ON «sdi».»uniqueId» = «ml«.«id«;

Этот запрос построчно добавит мета-данные о каждой точке измерений. После обработки данных мы имеем поток подготовленных данных, которые далее можем передавать в другие приложения и сервисы для последующей работы с ними.

На приведённой ниже схеме архитектуры видно (рисунок 4), что поток данных Kinesis, который получает результаты аналитики, связан с функцией AWS Lamba. Такая интеграция позволяет отправлять предупреждения на основе данных, получаемых функцией из потока Kinesis. Если она замечает, что флаг пробки изменил свое значение с True на False, отправляется сообщение в мессенджер, чтобы уведомить клиентов о том, что появилась пробка в определённой точке.

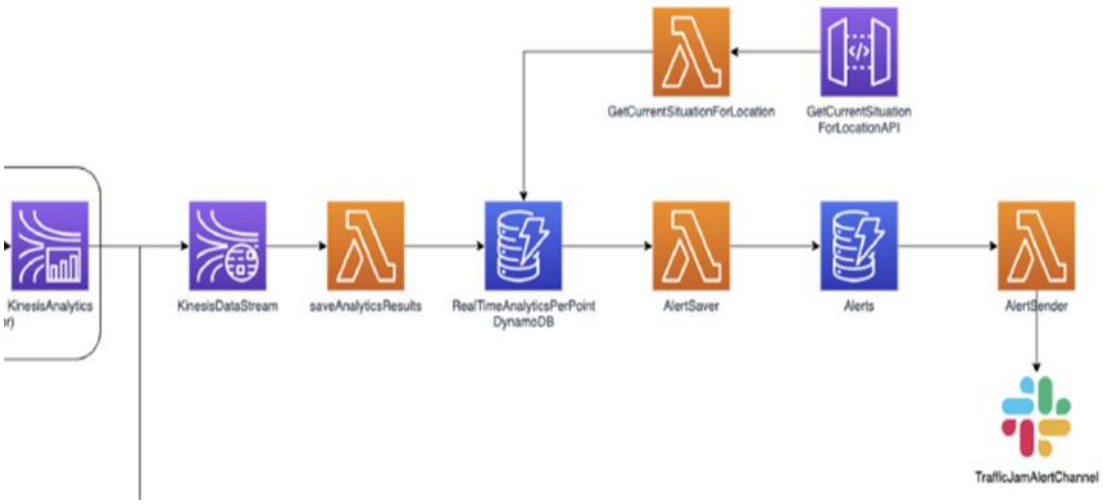

Рисунок 4. Алгоритм обработки поступающих данных

Заключение.

Использование сервисов Amazon Web Services позволяет создать приложение для обработки большего объёма данных в кратчайшие сроки. Преимущество такого подхода заключается в том, что все сервисы поддерживают автоматическое масштабирование в зависимости от нагрузки.

Обработка и анализ данных в режиме реального времени позволяет бизнесу значительно

ускорить реакцию на возникающие события, что позволяет им моментально реагировать на происходящие события и избегать нежелательных событий.

Список литературы

[1] Volume of data/information created [Электронный ресурс]. – Режим доступа: https://www.statista.com/statistics/871513/worldwide-data-created/

[2] AWS Kinesis [Электронный ресурс]. – Режим доступа: https://aws.amazon.com/ru/kinesis/

[3] AWS Kinesis Data Analytics [Электронный ресурс]. – Режим доступа: https://aws.amazon.com/ru/kinesis/data-analytics/

[4] Amazon Kinesis Data Streams [Электронный ресурс]. – Режим доступа: https://aws.amazon.com/ru/kinesis/data-streams/

[5] Amazon Kinesis Data Firehose [Электронный ресурс]. – Режим доступа: https://aws.amazon.com/ru/kinesis/data-firehose

REAL-TIME BIG DATA ANALYTICS IN THE AWS CLOUD USING THE AWS KINESIS SERVICE

K.O. KLIMOV
BSUIR Master,
DevOps Engineer
EPAM Systems

G. A. PISKUN
Deputy Dean of the Faculty of Computer Design for Scientific Work, Candidate of Engineering of Sciences, Associate Professor

D.M. LIKHACHEVSKY
Dean of the Faculty of Computer Design, candidate of Technical Sciences, Associate Professor

V.F. ALEKSEEV
Associate Professor, Department of Information Computer Systems Design, Candidate of Technical sciences, Associate Professor

V.V. SHATALOVA
Deputy Dean of the Faculty of Computer Design for Educational and Methodological Work, Associate Professor

Department of Information and Computer Systems Design
Faculty of Computer Engineering
Belarusian State University of computer science and Radio Electronics, Republic of Belarus
EPAM Systems, Republic of Belarus
E-mail: klimovkostya5@gmail.com

Abstract. The amount of data generated by different sources grows every year. By 2020, the amount of stored data will increase to 40-44 zettabytes, by 2025 this amount will be 10 times bigger. For this reason, the analytics of all this data can help to identify valuable information, so that the business can determine trends and predict indicators with high accuracy. The maximum benefit can be obtained by analyzing this data in real time and reacting immediately. To achieve such goals, huge computational resources are required to process all incoming data. Cloud computing has greatly simplified the analytics of this amount of data.

Keywords: Amazon Web Services, AWS, AWS Kinesis, Big Data Analytics, Big Data processing.