

Свиридова И.В.

Преподаватель СПО Инжинирингового колледжа

НИУ «БелГУ» Россия, г. Белгород

Sviridova I.V.

Lecturer of STR of Engineering College

NRU "BelSU" Russia, Belgorod

Бабенко А.А.

Студент аспирантуры 2го года обучения

НИУ «БелГУ» Россия, г. Белгород

Babenko A.A.

2nd year postgraduate student

NRU "BelGU" Russia, Belgorod

Гончаров Д.В.,

Преподаватель СПО Инжинирингового колледжа

НИУ «БелГУ» Россия, г. Белгород

Goncharov D.V.

Lecturer of STR of Engineering College

NRU "BelSU" Russia, Belgorod

Бабенко А.А.

Студент аспирантуры 2го года обучения

НИУ «БелГУ» Россия, г. Белгород

Babenko A.A.

2nd year postgraduate student

NRU "BelGU" Russia, Belgorod

БОЛЬШИЕ ДАННЫЕ, ИХ ОБРАБОТКА И ПРИМЕРЫ ИХ

ИСПОЛЬЗОВАНИЯ

BIG DATA, THEIR PROCESSING AND EXAMPLES OF THEIR USE

Аннотация: В данной статье рассказывается о Больших данных, приводятся примеры их использования, рассказывается об одном из инструментов их сбора и обработки, приводятся способы анализа

Abstract: This article tells about Big Data, gives examples of their use, talks about one of the tools for collecting and processing them, and provides methods of analysis.

Ключевые слова: большие данные, система обработки сообщений, банковская система.

Keywords: big data, message processing system, banking system.

Сейчас многие говорят о Big Data (Больших данных), но далеко не все люди сталкивались с ними и понимают, как и нужно ли им вообще применять их в своей работе. В данной статье я познакомлю вас с терминологией и на нескольких примерах попытаюсь объяснить, нужно ли вам применять большие данные в вашей работе.

«Большие данные (BigData) — обозначение структурированных и неструктурированных данных огромных объёмов и значительного многообразия, эффективно обрабатываемых горизонтально масштабируемыми программными инструментами, появившимися в конце 2000-х годов и альтернативных традиционным системам управления базами данных и решениям класса BusinessIntelligence».

По мнению специалистов к Big Data можно отнести большинство потоков данных более 100Гб в день.

Представьте супермаркет, в котором все товары расставлены хаотично. Печенье рядом с туалетной бумагой, журналы рядом с выпечкой, алкогольные напитки напротив стеллажа с памперсами, на котором так же находятся фрукты, молоко и чипсы. Большие данные позволяют разобраться в этом хаосе расставляя все по местам и помогают найти необходимые продукты, проверить срок годности, стоимость, а так же узнать кто помимо вас покупает данный товар и чем он лучше или хуже своих аналогов.

Обработка объемных данных требуется для того, чтобы люди могли получить конкретные результаты для их дальнейшего эффективного применения.

Базовым принципом обработки больших данных является горизонтальная масштабируемость. При ней данные распределяются на нескольких вычислительных узлах, при этом обработка происходит без деградации производительности. Так же она позволяет при необходимости добавлять новые узлы, без приостановки работы текущих. Так же этот принцип подразумевает что вычислительных узлов может быть много и часть из них будет гарантированно выходить из строя. Методы работы с большими данными должны учитывать возможность сбоев и переживать их без особых проблем.

Данную технологию в России применяют несколько различных компаний, например «Сбер». Они используют Big Data для анализа и сравнения фотографий из базы, которые получают с камер на стойках, при идентификации клиентов. Благодаря этому количество случаев мошенничества уменьшилось в 10 раз.

Помимо «Сбера» большие данные используются многими другими Российскими банками («ВТБ24», «Альфа-банк», «Тинькофф») в скоринге, продажах, маркетинге, и т.д.

Но применение BigData не ограничивается сферой финансовых услуг, одним из случаев применения этой технологии может быть анализ движения транспортных средств на дорогах мегаполиса для прогнозирования возможного количества трафика в определенных районах. При помощи этих данных можно спрогнозировать места возможных пробок на дороге, и при установке маршрута в gps-навигаторе оптимизировать путь каждой из машин по отдельности, отталкиваясь от возможной загруженности дорог в будущем.

Одним из инструментов для хранения и больших данных являются системы обработки сообщений, такие как ApacheKafka.

Кафка позволяет централизовать сбор, передачу и обработку большого количества сообщений в непрерывных информационных потоках, а также хранить эти большие данные, не волнуясь о рисках их потери и производительности системы.

Яркий пример использования ApacheKafka – непрерывная передача информации со smart-периферии (конечных устройств) в IoT-платформу, когда данные не только передаются, но и обрабатываются множеством клиентов, которые называются подписчиками (consumers). В роли подписчиков выступают приложения и программные сервисы. Здесь имеют место отложенные вычисления, когда подписчиков меньше, чем сообщений от издателей – источников данных (producer). Сообщения (messages) записываются по разделам (partition) темы (topic) и хранятся в течении заданного периода. Подписчики сами опрашивают Kafka на предмет наличия новых сообщений, и указывают, какие записи им нужно прочесть, увеличивая или уменьшая смещение к нужной записи. Записанные события могут переигрываться или обрабатываться повторно.

На рисунке 1 показан принцип действия распределенной BigData системы обработки сообщений ApacheKafka.

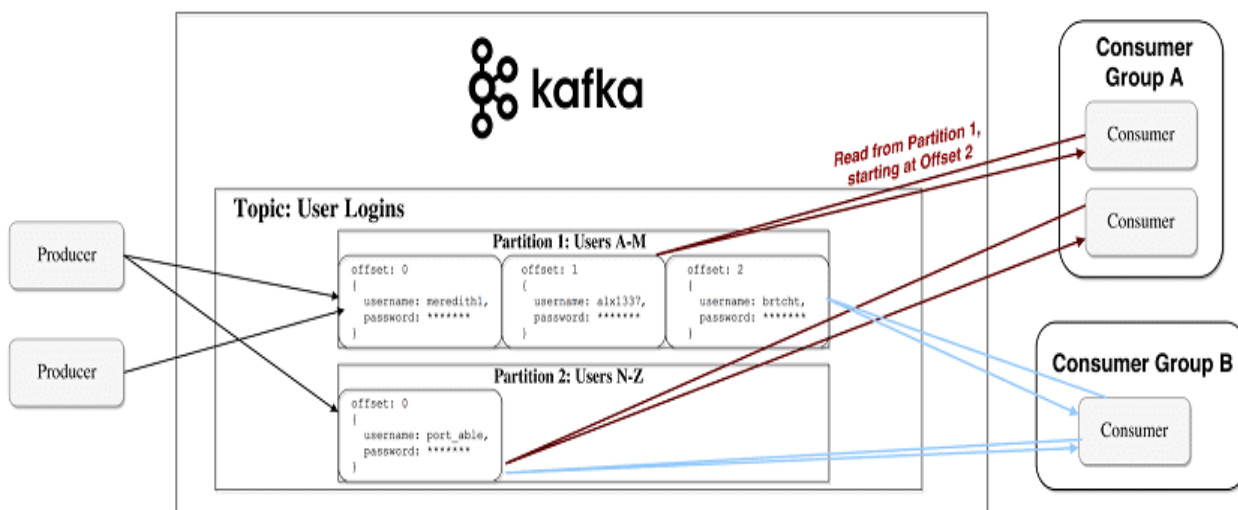


Рисунок 1 - Принцип действия распределенной BigData системы обработки сообщений ApacheKafka

Использованные источники:

1. Big Data [Большие данные] [Электронный ресурс]– Электрон. текстовые дан. – Киев: [б.и.], 2000. – Режим доступа: <https://www.it.ua/ru/knowledge-base/technology-innovation/big-data-bolshie-dannye>, свободный (дата обращения: 08.11.2020).

2. Хабр. Big Data от А до Я. Часть 1: Принципы работы с большими данными, парадигма MapReduce [Электронный ресурс] / Хабр. – Электрон. текстовые дан. – Москва: [б.и.], 2000. – Режим доступа: <https://habr.com/ru/company/dca/blog/267361/>, свободный (дата обращения: 08.11.2020).