УДК 004.8:658.8 https://doi.org/10.54251/2616-6429.2026.01.0018nu

Т.Д. Кузербаева[1*],С.Д. Куракбаева[1], М.М. Расулмухамедов[2], Ж.Д. Изтаев[1], Г.К. Тагай[1]
[1]магистрант, Южно-Казахстанский университет им. М. Ауэзова, Шымкент, Казахстан
[1]к.т.н. профессор, Южно-Казахстанскийуниверситетим.М.Ауэзова, Шымкент, Казахстан
[2]к. ф.-м.н., доцент, Ташкентский государственный транспортный университет, Ташкент, Узбекистан
[1]к.п.н., ассоциированный профессор, Южно-Казахстанский университет им. М. Ауэзова, Шымкент, Казахстан
[1]магистр, старший преподаватель Южно-Казахстанский университет им. М. Ауэзова, Шымкент
[*]**Автор для корреспонденции:** totykuzerbayeva@gmail.com

РАЗРАБОТКАСИСТЕМЫ СЕГМЕНТАЦИИ РЫНКА С ИСПОЛЬЗОВАНИЕМ BIGDATAДЛЯ ПОВЫШЕНИЯ ЭФФЕКТИВНОСТИ ПЕРСОНАЛИЗИРОВАННОГО МАРКЕТИНГА

Аннотация

В данном исследовании обоснован выбор стек технологий Big Data (Apache Spark, Parquet, Kafka, MLflow) для обеспечения масштабируемости и воспроизводимости. Предложена гибридная лямбда-архитектура, поддерживающая как пакетную, так и потоковую обработку событий. На синтетическом датасете, имитирующем поведение пользователей стриминговой платформы, проведены эксперименты по сегментации с использованием методов K-Means и DBSCAN. Построены комплексные признаки, включающие RFM-метрики, поведенческие и темпоральные характеристики. Качество кластеризации оценено с использованием метрик силуэта и Калински-Харабаса. Результатом является функционирующий прототип, который автоматизирует этапы от приема сырых событий до формирования интерпретируемых сегментов с рекомендациями по их использованию в персонализированных маркетинговых кампаниях. Практическая значимость работы заключается в демонстрации подхода, позволяющего снизить стоимость контакта и повысить CTR/CVR за счет более точного таргетирования.

Ключевые слова: Big Data, сегментация рынка, персонализированный маркетинг, кластеризация, RFM-анализ.

Введение. Цифровая трансформация маркетинга обусловила переход от массовых коммуникаций к гиперперсонализации, что требует глубокого понимания поведения каждого потребителя. В предыдущей работе [1] был проведен обзор особенностей, технологий и вызовов при построении интеллектуальных систем сегментации на основе Big Data. Определены ключевые проблемы: работа с высокоскоростными потоками данных (velocity), обеспечение конфиденциальности (veracity), масштабируемость алгоритмов (volume) и необходимость обработки разнородной информации (variety). Однако раннее наше исследование не затрагивало вопросы практической реализации.

Актуальность работы обусловлена потребностью бизнеса, в том числе малых и средних предприятий, стремящихся использовать технологии больших данных, в тиражируемых и экономичных архитектурных решениях. Целью исследования является создание воспроизводимого прототипа пайплайна интеллектуальной сегментации, от приема данных до генерации сегментов, с открытым кодом и четкой методологией, что составляет научную новизну. Практическая значимость заключается в снижении операционных затрат на маркетинг за счет автоматизации сегментации и повышении эффективности кампаний через точное таргетирование, что в перспективе ведет к росту конверсии (CVR) и кликабельности (CTR).

Постановкой задачи являетсяразработка архитектуры и реализация прототипа системы, которая на основе потоковых и накопленных пользовательских данных в масштабах,

характерных для крупной цифровой платформы (аналогично Netflix, Spotify), автоматически выделяет устойчивые поведенческие сегменты для последующей персонализации контента и рекламных коммуникаций.

В современной литературе по цифровому маркетингу сегментация рассматривается как ключевой механизм гиперперсонализации. Котлер, Картаджайя и Сетиауан в концепции Marketing 5.0 акцентируют переход к data-driven маркетингу, где ценность создается за счет точного понимания потребностей и контекста клиента на основе данных и ИИ [2]. Систематический обзор Gupta и соавт. фиксирует рост применения машинного обучения в маркетинге (в том числе кластеризации для сегментирования), однако подчеркивает дефицит воспроизводимых инженерных решений, связывающих алгоритмы с бизнес-метриками и процессами принятия решений. На уровне методов прикладные исследования предлагают расширения классических поведенческих моделей: например, Rungruang и соавт. показывают развитие RFM-подхода для устойчивой кластеризации клиентов, а Rodrigues и соавт. систематизируют NLP-ориентированное сегментирование на основе текстовых и событийных данных. [3]

С инженерной точки зрения, исследователи отдельно подчеркивают роль потоковой обработки и оперативного принятия решений. Jabbar, Akhtar и Dani рассматривают real-time обработку больших данных как основу «моментального» маркетинга и отмечают, что задержки обработки пользовательских событий снижают эффект таргетинга и персонализации. Для масштабируемой аналитики и построения распределенных пайплайнов часто используется Apache Spark; практическое руководство Damji и соавт. описывает подходы к обработке больших данных и построению ML-контура в распределенной среде [4-5].

Отдельный пласт работ связан с внедрением моделей и управлением жизненным циклом. Chen и соавт. описывают MLflow как платформу, поддерживающую воспроизводимость (tracking экспериментов), версионирование артефактов и регистрацию моделей, что снижает риски разрыва между исследовательским кодом и эксплуатацией; эти практики закреплены и в официальной документации MLflow [6-7]. Для оркестрации задач (ETL, обучение, переобучение и плановые пересчеты) широко применяется Apache Airflow. Практико-ориентированные исследования на русском языке также подчеркивают, что бизнес-ценность Big Data в маркетинге достигается только при наличии сквозной архитектуры интеграции данных, контроля качества и измерения эффекта кампаний [8].

Таким образом, существующие работы детально раскрывают отдельные аспекты (методы сегментации, потоковую аналитику, MLOps), но реже предоставляют целостный, воспроизводимый прототип «от данных до сегментов», ориентированный на масштабы крупной цифровой платформы и учитывающий конфиденциальность, интерпретируемость и эксплуатационные ограничения [9]. Научная новизна настоящей работы заключается в проектировании и реализации открытого end-to-end прототипа пайплайна интеллектуальной сегментации: прием потоковых и накопленных данных, их стандартизация и контроль качества, обезличивание идентификаторов на ранних этапах, формирование признаков и автоматическое выделение устойчивых поведенческих сегментов с последующей визуализацией результатов в веб-интерфейсе. Такой формат делает решение тиражируемым и пригодным для практического внедрения при ограниченных ресурсах, сохраняя ориентацию на масштабы, характерные для крупных цифровых платформ.

Методы исследования. В основе работы лежит комбинация инженерных и аналитических методов, направленных на создание воспроизводимого пайплайна сегментации. Для имитации реальных условий использован синтетический датасет, генерирующий поведенческие события пользователей VoD-платформы. Архитектурной основой выбрана гибридная лямбда-схема, обеспечивающая как пакетную, так и потоковую

обработку данных. Применены методы предобработки данных, включая дедупликацию, обработку пропусков и выбросов, а также масштабирование признаков. Для сегментации сравнивались алгоритмы кластеризации K-Means и DBSCAN, качество которых оценивалось с помощью внутренних метрик (силуэт, Калински–Харабас, Дэвис–Болдин). Визуализация и интерпретация сегментов выполнены с использованием веб-интерфейса, взаимодействующего с бэкендом через REST API. Все этапы пайплайна оркестрированы с помощью Apache Airflow, а эксперименты зафиксированы в MLflow для обеспечения воспроизводимости.

Результаты и обсуждение

Для эффективного хранения структурированных и неструктурированных данных используются различные решения, которые относятся к третьему этапу и включают в себя: NoSQL-базы данных (MongoDB, Cassandra, Redis) – позволяют хранить большие объёмы данных в удобном формате, обеспечивая высокую скорость чтения и записи; реляционные базы данных (PostgreSQL, MySQL, Microsoft SQL Server) – используются для хранения структурированных данных, если система требует строгих связей между таблицами; облачные хранилища (AWS S3, Google Cloud Storage, Azure Blob Storage) – масштабируемые и отказоустойчивые решения, обеспечивающие безопасное хранение данных и быстрый доступ из любой точки мира; Data Lake и Data Warehouse (Amazon Redshift, Google BigQuery, Snowflake) – используются для хранения больших массивов данных, подготовки их к аналитике и машинному обучению. Выбор хранилища зависит от типа данных, требуемой скорости обработки и масштаба проекта.

Для прототипирования используется синтетический датасет, который имитирует логи поведения пользователей на VoD-платформе (Video on Demand). Такой выбор позволяет избежать этических и юридических рисков, связанных с обработкой персональных данных, и одновременно обеспечивает полную воспроизводимость эксперимента: любой исследователь может повторить запуск, не сталкиваясь с ограничениями доступа к релевантным данным. Внутри набора данных моделируются ключевые сущности, характерные для цифровых платформ: поток событий пользователей (например, воспроизведение, пауза, поиск, клики по баннерам и покупка подписки), таблица транзакций с фактами платежей, профиль пользователя, который формируется на основе агрегирования активности за исторический период, а также справочник контента с метаданными видео (жанр, длительность, год релиза), необходимыми для интерпретации потребительских интересов [10].

На следующем этапе реализуется предобработка и контроль качества данных, потому что даже синтетические логи должны имитировать реалистичные искаженияданных, с которыми сталкивается бизнес. На этом шаге выполняется дедупликация событий, чтобы убрать повторные записи и снизить искажения при расчёте метрик поведения. Пропущенные значения обрабатываются по типу признаков: категориальные поля заполняются нейтральным значением вроде unknown, а числовые — устойчивыми оценками, например медианой (в практических кейсах это снижает влияние единичных аномалий). Для финансовых показателей дополнительно применяется обработка выбросов, поскольку расходы пользователей часто имеют тяжелые хвосты; в прототипе используется подход на основе межквартильного размаха (IQR), который хорошо работает без предположений о нормальности распределения.

Затем выполняется построение признаков (feature engineering) для сегментации. Для каждого обезличенного пользователя за фиксированный исторический интервал (например, 90 дней) формируется вектор признаков, который отражает как классическую маркетинговую модель RFM, так и поведенческие особенности потребления контента. Помимо Recency, Frequency и Monetary, добавляются характеристики сессий и взаимодействий с платформой: средняя длительность сессии, доля завершённых просмотров, разнообразие интересов через

число уникальных жанров, частота использования поиска и другие паттерны вовлечённости. Отдельно учитываются временные закономерности активности — например, склонность к просмотрам утром или вечером и распределение активности по дням недели. Чтобы признаки были сопоставимы и корректно работали в алгоритмах кластеризации, они масштабируются с помощью стандартной нормализации (StandardScaler)[11].

В прототипе сравниваются два подхода к сегментации. Первый — K-Means, широко используемый центроидный метод, который эффективен при достаточно “компактных” кластерах и хорошо масштабируется на больших данных, но требует заранее задать число кластеров k [10]. Второй — DBSCAN, плотностный метод, который способен находить кластеры произвольной формы и выделять шумовые точки, не требуя явного задания количества кластеров, однако он чувствителен к настройке параметров плотности (eps и min_samples) [11]. В качестве перспективного направления рассматривается применение автоэнкодера для нелинейного снижения размерности перед кластеризацией, что может повысить качество сегментов при сложных, нелинейных зависимостях в поведении пользователей.

Качество полученных сегментов оценивается с помощью внутренних метрик кластеризации, которые позволяют измерить компактность и разделимость групп в признаковом пространстве [12]. Для этого применяются Silhouette Score, индекс Калински–Харабаса и индекс Дэвиса–Болдина, поскольку они дают взаимодополняющие оценки структуры кластеров. При этом подчёркивается, что в прикладном маркетинге кластеризация должна подтверждаться не только математически, но и бизнес-эффектом, поэтому дальнейшая работа предполагает проверку полезности сегментов через A/B-эксперименты и измерение uplift по CTR или CVR при персонализации предложений для выделенных групп.

Таблица 1 - Обоснование выбора инструментария

Задача	Инструмент	Обоснование выбора
Язык программирования	Python 3.9+	Де-факто стандарт для DataScience, богатая экосистема библиотек.
Обработка данных (блок)	Pandas, NumPy	Эталонные библиотеки для анализа в памяти на одном узле. Используются для отладки и прототипирования логики.
Обработка BigData	ApacheSpark 3.x (PySpark)	Распределенная вычислительная платформа для обработки больших объемов данных. MLlib предоставляет масштабируемые реализации алгоритмов машинного обучения.
Хранилище (DataLake)	Parquet на S3-совместимом хранилище (MinIO)	Колоночный формат Parquet обеспечивает эффективное сжатие и выборку столбцов. S3-протокол – стандарт для облачных объектных хранилищ. MinIO используется для локального прототипирования.
Потоковый ingestion	ApacheKafka	Высокопроизводительная распределенная потоковая платформа. Используется для приема событий в реальном времени.

Продолжение таблицы 1

Оркестрацияпайплайпнов	ApacheAirflow	Позволяет описывать, планировать и мониторить сложные пайплайны данных в виде направленных ациклических графов (DAG).
Трекинг ML-экспериментов	MLflow	Платформа для управления жизненным циклом машинного обучения: логирование параметров, кода, метрик и артефактов (моделей).
Контейнеризация	Docker	Обеспечивает воспроизводимость среды выполнения, изолируя все зависимости.

В работе предложена лямбда-архитектура, которая объединяет пакетную (batch) и потоковую (stream) обработку, чтобы одновременно сохранять точность на исторических данных и обеспечивать приемлемую задержку при обновлении сегментов. На входе (ingestion) события из источников — логов приложений и транзакционных баз — поступают в Apache Kafka (топик raw-events), что позволяет устойчиво принимать поток данных и отделить источники от последующей обработки. Для пакетной загрузки исторических данных используется прямое размещение в объектном хранилище, чтобы можно было быстро сформироватьретроспективу и не перегружать потоковый контур.

Хранилище организовано как Data Lake с разделением зон данных. Потоковые события через Spark Structured Streaming записываются в “горячую” зону (например, в формате Parquet), а пакетные данные поступают в сырой слой. После очистки и обогащения данные перемещаются в подготовленный слой (curated), который становится основной витриной для построения признаков. Такое разнесение слоев позволяет управлять качеством данных и уменьшать вероятность того, что в моделирование попадут невалидные или неоднородные записи.

Далее работает слой обработки и feature engineering. По расписанию (через Airflow DAG) запускается Spark-задача, которая агрегирует события из подготовленного слоя за выбранный период и вычисляет вектор признаков для каждого пользователя. Полученные признаки сохраняются в feature store, реализованный в виде таблиц внутри того же Data Lake, что упрощает воспроизводимость и снижает стоимость хранения. На следующем этапе запускается моделирование: периодически (например, раз в неделю или месяц) выполняется кластеризация (K-Means/DBSCAN) над актуальной матрицей признаков, а гиперпараметры и метрики качества логируются в MLflow. Обученная модель и сопоставление user_anon_id → segment_id сохраняются как артефакты, чтобы можно было восстановить точную версию сегментации и сравнивать результаты между итерациями.

На слое сервинга предполагается API-сервис (например, на FastAPI), который загружает актуальную модель и маппинг сегментов и предоставляет интерфейс для получения сегмента пользователя по его идентификатору. Результаты сегментации также выгружаются в хранилище аналитики (например, ClickHouse) как атрибут segment_id в профилях пользователей, чтобы аналитики и маркетинговые команды могли строить отчеты и проверять эффект персонализации. Отдельно предусмотрен мониторинг: отслеживаются показатели качества данных (полнота, аномалии), возможный дрейф признаков и прикладные метрики, такие как размеры сегментов и их активность во времени, что важно для контроля стабильности сегментации.

Вопрос безопасности учитывается на ранних этапах пайплайна. Уже на ingestion выполняется обезличивание — исходный user_id преобразуется в user_anon_id (например, посредством хеширования), а доступ к данным на уровне хранилища ограничивается ролями (IAM/ACL). Внутри решения используются отдельные технические учетные записи для

оркестратора, обработки и сервинга, чтобы минимизировать риски несанкционированного доступа и разделить полномочия по принципу наименьших привилегий.

Для экспериментальной проверки используется синтетически сгенерированный датасет, имитирующий поведение 50 000 пользователей и около 5 млн событий за 90 дней. Генерация данных закладывает реалистичные закономерности: сессионность, предпочтения по жанрам и связь активности с монетизацией. Эксперименты выполнялись в локальном окружении: Spark в standalone-режиме (драйвер и два воркера по 4 ГБ RAM), Kafka версии 3.x и Python 3.9, а также библиотеки pyspark 3.3.0, pandas и scikit-learn. После предобработки и построения признаков (12 признаков) была проведена кластеризация: для K-Means число кластеров выбрано равным 5 на основе метода “локтя” и метрики силуэта, а DBSCAN настроен с параметрами eps = 0.5 и min_samples = 10.

На рисунке 1 представлена главная страница интерфейса, где пользователь может выполнить следующие операции: загрузить CSV-файл с исходными данными, выбрать алгоритм кластеризации (K-Means или DBSCAN), настроить гиперпараметры выбранного алгоритма (количество кластеров для K-Means или параметры eps и min_samples для DBSCAN), а также инициировать процесс сегментации нажатием кнопки запуска. Интерфейс взаимодействует с бэкенд-сервисом через REST API, отправляя задачу на обработку в распределенную среду.

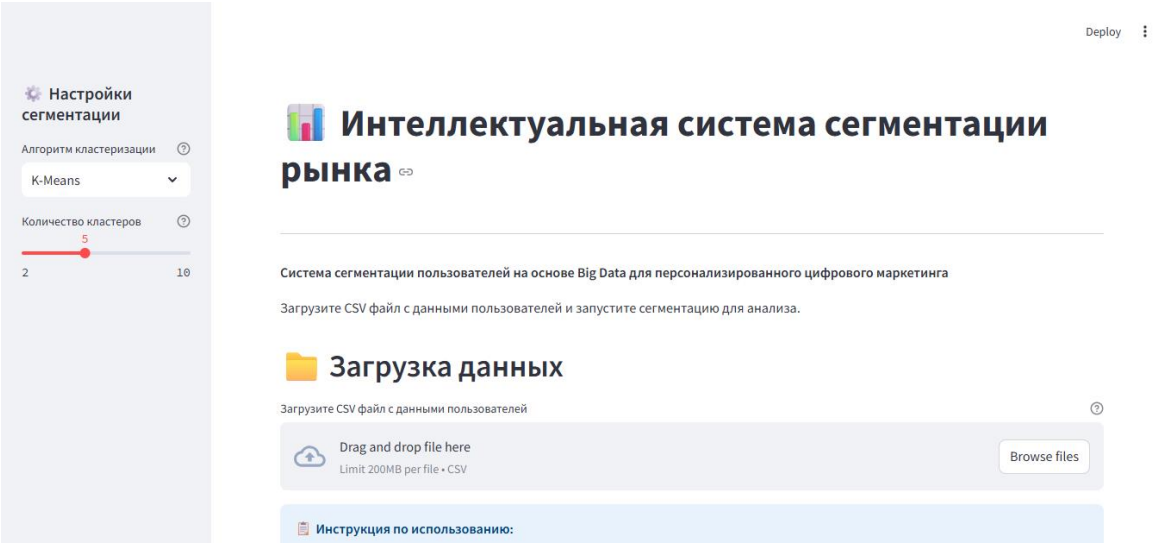

Рисунок 1 - Главная страница прототипа системы сегментации рынка (загрузка данных и параметры сегментации).

Для обеспечения корректной работы системы и предотвращения ошибок при загрузке данных, интерфейс включает раздел технической документации. На рисунке 2 представлен блок инструкций, который содержит требования к формату входных данных, включая список обязательных полей (например, user_id, timestamp, event_type, amount), описание типов данных и примеры заполнения CSV-файла. Этот раздел предназначен для самостоятельной подготовки данных пользователями и упрощения интеграции с системой.

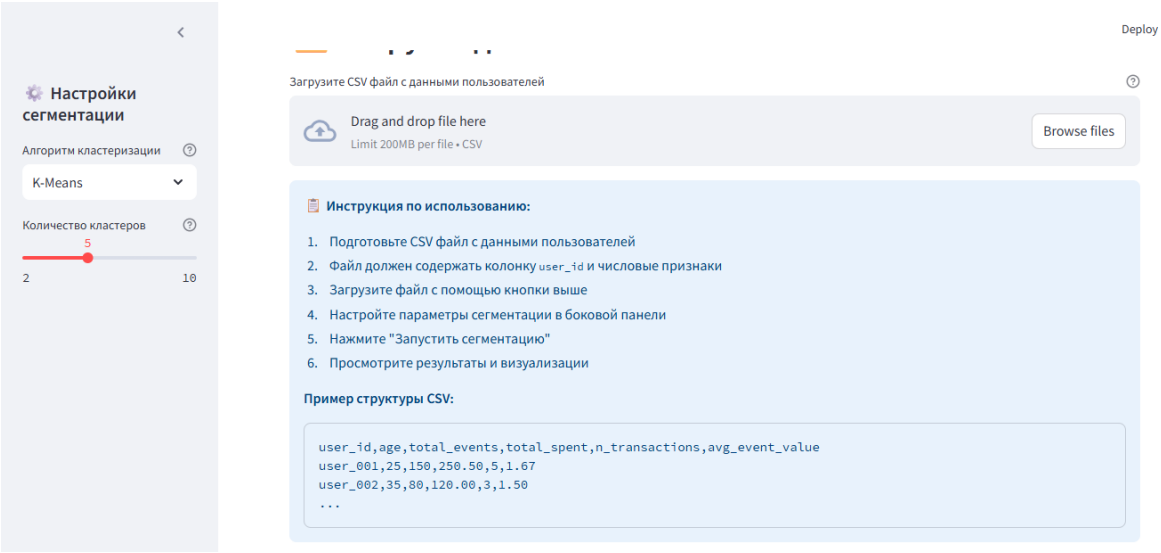

Рисунок 2 - Блок инструкций и пример структуры входного CSV-файла в пользовательском интерфейсе прототипа.

Таблица 2 - Архитектурные компоненты и этапы работы системы интеллектуальной сегментации

Этап/Компонент	Назначение в системе	Ключевые процессы
Сбор данных (Ingestion)	Прием данных из внешних источников	Потоковый прием через Kafka, пакетная загрузка в S3, обезличивание идентификаторов
Хранение (DataLake)	Организация структурированного хранилища	Зонирование (сырые, очищенные, признаки), формат Parquet, управление доступом
Обработка (ETL/FeatureEngineering)	Подготовка данных для анализа	Очистка, дедупликация, агрегация, расчет признаков (RFM, поведенческие, темпоральные)
Моделирование (ML Pipeline)	Построение и обучение моделей	Кластеризация (K-Means/DBSCAN), валидация, логирование экспериментов в MLflow

Продолжение таблицы 2

Оркестрация (Orchestration)	Управление выполнением пайплайнов	Планирование задач в Airflow, мониторинг выполнения, обработка ошибок
Сервинг (Serving)	Предоставление результатов для внешних систем	REST API (FastAPI), выгрузка сегментов в аналитические хранилища (ClickHouse)
Мониторинг (Monitoring)	Контроль работы системы и качества данных	Отслеживание метрик данных, дрейфа признаков, производительности API

С точки зрения реализации, система разрабатывалась как набор слабосвязанных модулей, взаимодействующих через очереди данных (Kafka) и общее хранилище (Data Lake). Основной пайплайн реализован в виде направленного ациклического графа (DAG) в Apache Airflow, который координирует выполнение Spark-задач для обработки данных, обучения моделей и обновления артефактов. Для обеспечения воспроизводимости все этапы, включая инженерию признаков и обучение моделей, логируются в MLflow с фиксацией параметров, кода и метрик. Веб-интерфейс (описанный на рисунках 1-2) представляет собой отдельное клиентское приложение, которое взаимодействует с бэкендом через REST API, инкапсулируя сложность распределенных вычислений и предоставляя пользователю простой способ управления процессом сегментации.

Выводы. В работе разработан и апробирован воспроизводимый прототип пайплайна интеллектуальной сегментации пользователей для цифровой платформы класса VoD, реализующий полный цикл — от генерации синтетических логов до выдачи сегментов в виде артефактов и пригодных для дальнейшей персонализации данных. Использование синтетического датасета, имитирующего поведение десятков тысяч пользователей и миллионы событий за исторический период, позволило одновременно обеспечить юридическую и этическую корректность исследования и зафиксировать повторяемость результатов при одинаковых настройках окружения и параметрах моделей.

Показано, что комбинирование RFM-признаков с поведенческими и темпоральными характеристиками формирует более содержательные и интерпретируемые сегменты, чем использование только транзакционных показателей. В рамках прототипа выполнено сравнение кластеризации K-Means и DBSCAN; подобранные параметры демонстрируют, что центроидные методы удобны для масштабирования и регулярного пересчёта сегментов, а плотностные — полезны для выявления “шума” и нетипичного поведения, хотя требуют более аккуратной настройки. Предложена методика оценки качества сегментации, которая не ограничивается внутренними метриками (Silhouette, Calinski–Harabasz, Davies–Bouldin), а формирует основу для перехода к прикладной валидации через бизнес-эффект — uplift по CTR/CVR в A/B-экспериментах, что особенно важно для маркетинговых систем, где “красивые кластеры” не всегда равны коммерческой ценности.

Обоснован выбор лямбда-архитектуры как компромисса между точностью и задержкой: потоковый контур обеспечивает устойчивый приём событий и актуализацию данных, а пакетный — стабильный пересчёт признаков и сегментов на исторической выборке. Применение связки Kafka + SparkStructuredStreaming + DataLake (Parquet и зонирование данных) + оркестрация (Airflow) + логирование экспериментов (MLflow) обеспечивает масштабируемость, отказоустойчивость и прозрачность экспериментов, а также создаёт

основу для переноса решения из локального окружения в распределённый кластер без изменения методологии. Дополнительно подтверждена практическая применимость прототипа через пользовательский веб-интерфейс, позволяющий загружать данные, задавать параметры сегментации и получать результаты в форме, удобной для аналитического использования.

Перспективы дальнейшей работы связаны с переходом от синтетических данных к реальным потокам при строгом соблюдении требований конфиденциальности (обезличивание, контроль доступа, регламенты обработки), а также с постановкой и проведением A/B-тестов для проверки бизнес-гипотез и измерения uplift. Отдельными направлениями являются развитие механизма динамического обновления сегментов на потоковых событиях (в сторону Kappa-подхода/онлайн-обновлений), исследование методов снижения размерности (включая автоэнкодеры) для повышения качества кластеризации на высокоразмерных признаковых пространствах и внедрение мониторинга дрейфа признаков/сегментов для поддержания стабильности сегментации во времени. Работа представляет собой законченное IT-решение в области инженерии данных и MLOps, демонстрирующее подход к построению масштабируемого, воспроизводимого и безопасного пайплайна для интеллектуальной сегментации больших данных. Предложенная архитектура и технологический стек могут быть адаптированы для решения аналогичных задач в других предметных областях.

Список литературы

1. Кузербаева Т.Д., Куракбаева С.Д., Изтаев Ж.Д., Тагай Г.К. Обзор особенностей разработки интеллектуальной системы сегментации рынка на основе Big Data для цифрового маркетинга // Вестник науки Южного Казахстана. – 2025. – С. 170-171.
2. Котлер Ф., Картаджайя Х., Сетиауан И. Marketing 5.0: Technology for Humanity. Hoboken, NJ: John Wiley & Sons, 2021. 224 p. ISBN 978-1-119-66851-0.
3. Gupta, S., et al. Machine Learning in Marketing: A Literature Review and Research Agenda // Journal of Marketing Research. – 2020. – Vol. 57(1). – P. 28–47.
4. Захариа М., Хендари Р.М., Конарлик Дж., Венкатараман С., У. Джозеф А., Гходаси А., Ходи З., Хинд М. Апаче Спарк: унифицированный движок для обработки больших данных // Коммьюникейшнс оф зе ЭйСиЭм (Communications of the ACM). – 2016. – Т. 59, № 11. – С. 56–65.
5. Damji J. S., Wenig B., Das T., Lee D. Learning Spark: Lightning-Fast Data Analytics. 2nd ed. O'Reilly Media, 2020. 400 p. ISBN 978-1492050018.
6. Крепс Дж., Наркхеде Н., Рао Дж. Кафка: распределенная система обмена сообщениями для обработки журналов // Труды конференции НетДБ (Proceedings of the NetDB). – 2011.
7. MLflow: A Platform for the Machine Learning Lifecycle. [Электронный ресурс]. – Official Documentation, 2023. URL: https://mlflow.org/docs/latest/index.html
8. Apache Airflow Documentation. [Электронный ресурс]. – URL: https://airflow.apache.org/docs/ Доступно
9. Marz, N., Warren, J. Big Data: Principles and best practices of scalable real-time data systems. – Manning Publications, 2015.
10. Хан Дж., Камбер М., Пей Дж. Добыча данных: концепции и техники (Data Mining: Concepts and Techniques). – Морган Кауфманн (Morgan Kaufmann), 2011.
11. Мамбетсапаев К.А., Турапова Ш.З. Методы кластеризации для анализа потребительского поведения в цифровом маркетинге // Институт "International school of finance, technology and science", 2023. doi: 10.5281/zenodo.14900124.

References

1. Kuzerbaeva T.D., Kurakbaeva S.D., Iztaev Zh.D., Tagai G.K. Obzor osobennostei razrabotki intellektualnoi sistemy segmentatsii rynka na osnove Big Data dlya tsifrovogo marketinga // Vestnik nauki Yuzhnogo Kazakhstana. – 2025. – S. 170-171.
2. Kotler F., Kartadzhaiya Kh., Setiauan I. Marketing 5.0: Technology for Humanity. Hoboken, NJ: John Wiley & Sons, 2021. 224 p. ISBN 978-1-119-66851-0.
3. Gupta, S., et al. Machine Learning in Marketing: A Literature Review and Research Agenda // Journal of Marketing Research. – 2020. – Vol. 57(1). – P. 28–47.
4. Damji J. S., Wenig B., Das T., Lee D. Learning Spark: Lightning-Fast Data Analytics. 2nd ed. O'Reilly Media, 2020. 400 p. ISBN 978-1492050018.
5. MLflow: A Platform for the Machine Learning Lifecycle. [Elektronnyi resurs]. – Official Documentation, 2023. URL: https://mlflow.org/docs/latest/index.html
6. Apache Airflow Documentation. [Elektronnyi resurs]. – URL: https://airflow.apache.org/docs/ Dostupno
7. Marz, N., Warren, J. Big Data: Principles and best practices of scalable real-time data systems. – Manning Publications, 2015.
8. Zakharia M., Khendari R.M., Konarlik Dzh., Venkataraman S., U. Dzhozef A., Gkhodasi A., Khodi Z., Khind M. Apache Spark: unifitsirovannyi dvizhok dlya obrabotki bolshikh dannykh // Kommyunikeishns of ze EiSiEm (Communications of the ACM). – 2016. – T. 59, № 11. – S. 56–65.
9. Kreps Dzh., Narkkhede N., Rao Dzh. Kafka: raspredelennaya sistema obmena soobshcheniyami dlya obrabotki zhurnalov // Trudy konferentsii NetDB (Proceedings of the NetDB). – 2011.
10. Khan Dzh., Kamber M., Pei Dzh. Dobycha dannykh: kontseptsii i tekhniki (Data Mining: Concepts and Techniques). – Morgan Kaufmann (Morgan Kaufmann), 2011.
11. Mambetsapaev K.A., Turapova Sh.Z. Metody klasterizatsii dlya analiza potrebitelskogo povedeniya v tsifrovom marketinge // Institut "International school of finance, technology and science", 2023. doi: 10.5281/zenodo.14900124.

Т.Д. Кузербаева[1]*, С.Д. Куракбаева[1], М.М. Расулмухамедов[2], Ж.Д. Изтаев[1], Г.К. Тагай[1]

[1] магистрант, tksway17@gmail.com, М. Әуезов атындағы ОҚУ, Шымкент, Қазақстан

[1] т.ғ.к., профессор, sevam@mail.ru, М. Әуезов атындағы ОҚУ, Шымкент, Қазақстан

[2] ф.-м.ғ.к., доцент, prof.rasulmukhamedov@gmail.com, Ташкент мемлекеттік көлік университеті, Ташкент, Өзбекстан

[1] п.ғ.к., қауымдастырылған профессор, zhalgasbek71@mail.ru, М. Әуезов атындағы ОҚУ, Шымкент, Қазақстан

[1] магистр, аға оқытушы, zhadi.t@mail.ru, М. Әуезов атындағы ОҚУ, Шымкент, Қазақстан

ПЕРСОНАЛДАНДЫРЫЛҒАН МАРКЕТИНГТІҢ ТИІМДІЛІГІН АРТТЫРУ ҮШІН BIG DATA НЕГІЗІНДЕ НАРЫҚТЫ СЕГМЕНТТЕУ ЖҮЙЕСІН ӘЗІРЛЕУ

Түйін

Бұл мақала персонализацияланған маркетинг үшін Big Data-ға негізделген тұтынушыларды сегменттеу жүйесінің қайталанатын прототипін жобалау мен іске асыруды ұсынады. Жүйені әзірлеу мәселелеріне алдыңғы шолу жұмысын жалғастыра отырып, бұл зерттеу тәжірибелік инженерлік аспектілерге баса назар аударады. Big Data стекін (Apache Spark, Kafka, Parquet, MLflow) таңдау негізделіп, партиялық және ағындық өңдеуді қолдайтын гибридті лямбда-сәулет ұсынылған. Мінездемелерді инженерлеу RFM-метрикаларын, мінез-құлықтық және уақыттық үлгілерді қамтиды. Ағындық платформадағы пайдаланушы әрекетін имитациялайтын синтетикалық деректер жинағында жүргізілген эксперименттерде K-Means және DBSCAN кластерлеу алгоритмдері салыстырылды, нәтижесінде силуэт және Калински-Харабас метрикалары бойынша K-Means түсіндірімді сегменттерді берді. Прототип деректерді қабылдау мен сақтаудан модельдеу және сегменттерді қызмет етуге дейінгі толық конвейерді қамтамасыз етеді. Практикалық маңыздылығы тұтынушыны тарту құнын

төмендетуге және CTR мен CVR-дың әлеуетті өсуіне әкелетін нысаналаудың дәлдігін арттыруға бағытталғас масштабируемі тәсілді көрсетуде. Қайталануды қамтамасыз ету үшін код пен әдіснама толық ашып көрсетілген.

Кілттік сөздер: Big Data, нарықты сегменттеу, персонализацияланған маркетинг, кластерлеу, RFM-талдау

T.D. Kuzerbayeva[1]*, S.D. Kurakbayeva[1], M.M. Rasulmukhamedov[2], Zh.D. Iztayev[1], G.K. Tagay[1]

[1]Master's student, tksway17@gmail.com, M. Auezov South Kazakhstan University, Shymkent, Kazakhstan
[1]Candidate of Technical Sciences, Professor, sevam@mail.ru, M. Auezov South Kazakhstan University, Shymkent, Kazakhstan
[2]Candidate of Physical and Mathematical Sciences, Associate Professor, prof.rasulmukhamedov@gmail.com, Tashkent State Transport University, Tashkent, Uzbekistan
[1]Candidate of Pedagogical Sciences, Associate Professor, zhalgasbek71@mail.ru, M. Auezov South Kazakhstan University, Shymkent, Kazakhstan
[1]Master of Science, Senior Lecturer, zhadi.t@mail.ru, M. Auezov South Kazakhstan University, Shymkent, Kazakhstan

DEVELOPMENT OF A BIG DATA-BASED MARKET SEGMENTATION SYSTEM TO IMPROVE THE EFFECTIVENESS OF PERSONALIZED MARKETING

Abstract

The article presents the design and initial implementation of a reproducible prototype for a Big Data-driven customer segmentation system aimed at personalized marketing. Building upon a prior review of system development challenges, this work focuses on the practical engineering aspects. It justifies the selection of the Big Data stack (Apache Spark, Kafka, Parquet, MLflow) and proposes a hybrid lambda architecture supporting both batch and stream processing. Feature engineering incorporates RFM metrics, behavioral, and temporal patterns. Experiments on a synthetic dataset simulating user behavior on a streaming platform compare K-Means and DBSCAN clustering algorithms, with K-Means yielding more interpretable segments based on silhouette and Calinski-Harabasz scores. The prototype provides a complete pipeline from data ingestion and storage to modeling and segment serving. The practical significance lies in demonstrating a scalable approach to reduce customer acquisition cost and improve targeting precision, leading to potential increases in CTR and CVR. The code and methodology are fully disclosed to ensure reproducibility.

Keywords: Big Data, market segmentation, personalized marketing, clustering, RFM analysis