**Документация, содержащая описание функциональных характеристик модуля программного обеспечения**

**Модуль «BI аналитика для стриминговых сервисов»**

**г. Москва 2025**

**Содержание**

[1. Введение 3](#_heading=h.r8e07vu1cxor)

[2. Назначение модуля 3](#_heading=h.tnbbzl639ime)

[3. Функциональные возможности 3](#_heading=h.yimv833x4tph)

[4. Архитектура и принцип работы 3](#_heading=h.aeej687w2juc)

[5. Используемые языки программирования и фреймворки 5](#_heading=h.jp5kc83w4j3)

### 1. Введение

Модуль «BI аналитика для стриминговых сервисов» предназначен для продуктовой аналитики видео платформ, таких как Rutube. Он предоставляет инструменты для анализа данных о пользователях, контенте и их взаимодействиях, что позволяет улучшать рекомендации, поиск и пользовательский опыт.

Модуль интегрируется с другими системами платформы данных, поисковым и рекомендательным модулями, а также поддерживает гибкость работы с данными через подготовленные дашборды и Adhoc-запросы с использованием SQL и Big Data технологий.

###

### 2. Назначение модуля

Модуль предназначен для анализа ключевых метрик видео платформ и взаимодействий пользователей. Его основные задачи:

* Мониторинг ключевых метрик (время просмотра, CTR, retention rate).
* Анализ пользовательского поведения.
* Оптимизация поиска и рекомендаций.
* Оценка эффективности контента.
* Поддержка A/B тестирования.
* Интерактивные дашборды и Adhoc-запросы.

###

### 3. Функциональные возможности

**Визуализация**

Интерактивные дашборды (DAU, MAU, время просмотра).

A/B тестирование.

Автоматические отчеты.

Adhoc аналитика (детализация данных с SQL-запросами).

**Аналитика поиска**

Анализ популярных запросов.

Оценка релевантности результатов.

Анализ нулевых результатов поиска.

**Аналитика рекомендаций**

Оценка CTR, времени просмотра и retention rate.

Персонализация рекомендаций.

**Пользовательская аналитика**

Поведенческая аналитика (пути пользователей, частота возвращений).

Сегментация пользователей (холодные/горячие, демографические данные).

**Аналитика контента**

Определение популярных видео, категорий и трендов.

Оценка влияния метаданных на вовлеченность.

###

### 4. Архитектура и принцип работы



*рис. 1 - диаграмма концептуальной архитектуры модуля «BI аналитика для стриминговых сервисов»*

Модуль интегрирован с платформой данных и получает информацию из внешних сервисов, включая:

* Авторизацию пользователей.
* A/B тестирование.
* Логи плеера.
* Данные о просмотрах и взаимодействиях.
* Поисковые запросы и результаты.
* Рекомендательные алгоритмы.

**Взаимодействие с другими модулями и сервисами.**

Платформа данных - универсальный модуль, отвечающий за сбор, хранение и обработку данных.

Рекомендательная система - источник данных о рекомендациях для анализа их эффективности.

Поисковый модуль - источник данных о пользовательских запросах и результатах поиска.

Внешние сервисы:

Авторизация Rutube

A/B тестирование

Бэкенд сервисы Rutube

Логи плеера

AppMetrica

Яндекс.Метрика

**Уровень данных**

Данные о контенте - название видео, описание, длительность, время публикации.

Данные об авторах/каналах - название, описание, дата создания.

Данные о подписках - информация о подписках пользователей на авторов и каналы.

Данные о взаимодействиях - просмотры, досмотры, перемотки.

Логи работы - данные о поисковых запросах, ответах и рекомендациях.

### 5. Используемые языки программирования и фреймворки

Базовые технологии

**Язык программирования**: Python

Инфраструктурные программные средства

Базовая ОС: Ubuntu

Система виртуализации: Виртуализация в Yandex Cloud

Контейнеризация:

* Docker
* Kubernetes (Yandex Managed Service for Kubernetes®)

Облачное хранилище: Yandex Cloud (Yandex Object Storage)

BI и визуализация: Apache Superset

Обработка данных:

* Dagster (open source версия, Docker)
* Apache Spark (часть Yandex Data Processing)

Используемые библиотеки: Polars, Pandas, deltalake, PyArrow, Dagster, PySpark

#

Основными техническими средствами, обеспечивающими функционирование модуля BI аналитики, являются:

ClickHouse - Высокопроизводительная колоночная СУБД для adhoc аналитики по агрегированным витринам данных.

SuperSet - Инструмент для визуализации и анализа данных.

Dataproc (Apache Spark) - Масштабируемая система для распределенной обработки больших объемов данных.

DataSphere: Интерактивный ноутбук для adhoc анализа данных.