**Документация, содержащая описание функциональных характеристик модуля программного обеспечения модуль «Инфраструктура для сервиса»**

**г. Москва 2025**

[Введение](#_sc406tk5ia27) 3

[1. Описание программного обеспечения](#_gi26yb8goyhg) 3

[2. Начало работы с программным обеспечением](#_dvt6yjxigt7) 3

[Для эффективной работы с программным обеспечением необходимо выполнить настройку инфраструктуры, интеграцию с источниками данных, а также запустить соответствующие модули для обработки и хранения данных. В случае возникновения вопросов или проблем, обратитесь к документации по конкретным модулям или к технической поддержке.](#_tx2q5kxj3fjq) 3

[2.1 Установка и настройка](#_20lvckza067) 3

[2.2 Запуск модуля](#_fi0l2mou803l) 3

[3. Рабочие процессы](#_2n1wtlnfancu) 4

[3.1 Загрузка данных](#_vg1x3532r7d7) 4

[3.2 Обработка данных](#_jgik21rg0ygd) 5

[3.3 Хранение данных](#_3jiowy73yzjh) 5

[3.4 Визуализация и бизнес-аналитика](#_ulvrmuqiebeh) 5

[4. Обслуживание системы](#_kzwano5jknag) 5

##

## **Введение**

Данный документ предназначен для пользователей, ответственных за эксплуатацию программного обеспечения (ПО), которое предоставляет сервисы для хранения исторических данных, интеграции с источниками данных и их обработки. Основное внимание уделено процессу начала работы в системе, запуску модулей, а также использованию функционала для обработки данных.

## **1. Описание программного обеспечения**

ПО предназначено для обработки и хранения исторических данных, а также интеграции с аналитическими системами и источниками данных. Среди ключевых возможностей — автоматическая и периодическая загрузка данных, хранение и обработка больших объемов информации, создание аналитических витрин данных, а также интеграция с инструментами для бизнес-анализа и отчетности.

Основные функциональные требования к ПО:

* Автоматическая загрузка данных из аналитических систем (Яндекс.Метрика, АппМетрика).
* Периодическая загрузка данных о пользовательском поведении.
* Долгосрочное хранение данных в объектном хранилище на базе Яндекс.Облака.
* Обработка данных с использованием SQL и Apache Spark.
* Интеграция с BI-инструментами для визуализации данных.
* Обработка видео и использование мультимодальных нейросетей.

## **2. Начало работы с программным обеспечением**

### Для эффективной работы с программным обеспечением необходимо выполнить настройку инфраструктуры, интеграцию с источниками данных, а также запустить соответствующие модули для обработки и хранения данных. В случае возникновения вопросов или проблем, обратитесь к документации по конкретным модулям или к технической поддержке.

### 2.1 Установка и настройка

Проверьте соответствие системных требований (среда выполнения Java, версии Python, доступ к облачным ресурсам).

Настройте доступ к облачной инфраструктуре (например, Яндекс.Облако), указав ключи API для подключения к сервисам хранения и обработки данных.

Настройте коннекторы для источников данных (например, Яндекс.Метрика) с использованием соответствующих API. Для настройки автоматической загрузки данных следуйте руководству по интеграции с каждой системой.

### 2.2 Запуск модуля

1. Запуск основной службы:

Для запуска программного обеспечения, откройте командную строку и выполните команду для старта основного сервиса:

shell
$ ./start\_service.sh

Убедитесь, что все службы запустились без ошибок.

1. Запуск модулей загрузки данных:

Для запуска модуля автоматической загрузки данных выполните следующую команду:

shell
$ ./start\_data\_loader.sh

Этот модуль будет автоматически загружать данные из аналитических систем (Яндекс.Метрика, АппМетрика) и внутренних сервисов Яппи.

1. Запуск модуля обработки данных:

Для обработки больших объемов данных используйте команду для старта инфраструктуры Apache Spark:
shell
$ ./start\_spark\_processing.sh

Этот процесс позволит использовать распределённую обработку данных с использованием SQL.

1. Запуск видеопроцессинга:

Для параллельной обработки видео и использования мультимодальных нейросетей используйте команду:
shell
$ ./start\_video\_processing.sh

## **3. Рабочие процессы**

### 3.1 Загрузка данных

**Автоматическая загрузка:** Каждый день осуществляется автоматическая загрузка данных о поведении пользователей и активности из интегрированных систем (например, Яндекс.Метрика).

**Периодическая загрузка:** Раз в неделю выполняется загрузка данных о просмотрах, лайках, подписках и других событиях внутри приложения Yappy.

**Модели загрузки данных:** Все данные, загруженные в систему, проходят через предварительную обработку и нормализацию.

### 3.2 Обработка данных

1. **Обработка через Apache Spark:**

Использование SQL-запросов для обработки больших данных. Все запросы к данным обрабатываются с использованием распределённой системы для повышения производительности.

1. **Мультимодальные нейросети:**

Для обработки видео данных используется интеграция с нейросетями, такими как XClip и ViCLIP.

### 3.3 Хранение данных

Все данные, включая видео и пользовательские данные, сохраняются в объектном хранилище на базе Яндекс.Облака. Реализовано георезервирование, что обеспечивает надежность и доступность данных в разных регионах.

### 3.4 Визуализация и бизнес-аналитика

Для визуализации данных используется Tableau, что позволяет создавать интерактивные отчеты и графики, помогая в принятии бизнес-решений.

## **4. Обслуживание системы**

**Мониторинг работы системы.** Используйте встроенные инструменты для мониторинга состояния серверов и загрузки данных.

Для отслеживания ошибок и производительности системы доступны логи, которые можно просматривать через административную панель.

**Обновления и поддержка.** Программное обеспечение регулярно обновляется. Следите за выпусками и установите последние версии для улучшения функционала и безопасности.

**Резервное копирование данных.** Все данные регулярно резервируются в облаке для обеспечения их сохранности.