**Система обновления аналитических витрин данных КХД «ЭнергосбыТ Плюс»  
с использованием Apache Spark**

**Руководство пользователя**

Киров, 2025 г.

**Содержание**

[1 Введение 3](#_Toc212587459)

[1.1 Область применения 3](#_Toc212587460)

[1.2 Краткое описание возможностей 3](#_Toc212587461)

[1.3 Уровень подготовки пользователя 3](#_Toc212587462)

[1.4 Перечень эксплуатационной документации, с которыми необходимо ознакомиться пользователю 4](#_Toc212587463)

[2 Назначение и условия применения 5](#_Toc212587464)

[2.1 Виды деятельности, функции, для автоматизации которых предназначено данное средство автоматизации 5](#_Toc212587465)

[2.2 Условия, при соблюдении которых обеспечивается применение средства автоматизации в соответствии с назначением 5](#_Toc212587466)

[3 Подготовка к работе 6](#_Toc212587467)

[3.1 Состав и содержание дистрибутивного носителя данных 6](#_Toc212587468)

[3.2 Порядок развертывания и настройки 6](#_Toc212587469)

[3.3 Порядок проверки работоспособности 6](#_Toc212587470)

[4 Описание операций 7](#_Toc212587471)

[4.1 Общее описание технологического процесса 7](#_Toc212587472)

[4.2 Описание экранных форм и интерфейса 7](#_Toc212587473)

[4.2.1 Главное окно (DAGs View) 7](#_Toc212587474)

[4.2.2 Окно процесса (Tree View) 8](#_Toc212587475)

[4.2.3 Окно логов (Log View) 8](#_Toc212587476)

[4.3 Выполнение операций 9](#_Toc212587477)

[4.3.1 Мониторинг выполнения ETL-процесса 9](#_Toc212587478)

[4.3.2 Ручной запуск ETL-процесса 9](#_Toc212587479)

[4.3.3 Диагностика сбоев 9](#_Toc212587480)

[4.3.4 Перезапуск процесса после сбоя 10](#_Toc212587481)

[5 Аварийные ситуации 11](#_Toc212587482)

[5.1 Действия при сбоях в IT-инфраструктуре 11](#_Toc212587483)

[5.2 Действия при ошибках в коде или логике ETL 11](#_Toc212587484)

[5.3 Действия при обнаружении некорректных данных в источнике 11](#_Toc212587485)

[6 Рекомендации по освоению 12](#_Toc212587486)

[7 Рекомендации по использованию 13](#_Toc212587487)

# Введение

Данный документ является руководством пользователя для системы обновления аналитических витрин данных «СОВД-Spark».

Данный документ разработан в соответствии с требованиями ГОСТ Р 59795–2021 «Программное обеспечение и системы в машиностроении. Документирование».

## Область применения

Настоящее руководство предназначено для технического персонала (дата-инженеров), ответственного за эксплуатацию, мониторинг и развитие системы обновления аналитических витрин данных «СОВД-Spark» (далее – Система) в компании «Энергосбыт».

## Краткое описание возможностей

Система «СОВД-Spark» предоставляет пользователю следующие возможности:

* автоматический запуск процесса обновления витрин по заданному расписанию;
* ручной (внеплановый) запуск процесса обновления;
* мониторинг статуса выполнения ETL-задач в режиме реального времени;
* доступ к подробным логам выполнения для диагностики проблем;
* перезапуск процесса с точки сбоя после устранения неисправностей.

## Уровень подготовки пользователя

Уровень подготовки пользователя – специалист. Пользователь должен обладать базовыми навыками администрирования в среде Linux и иметь представление о следующих технологиях: Apache Spark, Apache Airflow, SQL.

## Перечень эксплуатационной документации, с которыми необходимо ознакомиться пользователю

Пользователю необходимо ознакомиться со следующей документацией:

* Техническое задание на разработку Системы;
* Настоящее руководство пользователя.

# Назначение и условия применения

## Виды деятельности, функции, для автоматизации которых предназначено данное средство автоматизации

Система предназначена для автоматизации ETL-процессов, включая:

* извлечение больших объемов данных из источников КХД;
* очистка, преобразование и агрегация данных с использованием кластера Apache Spark;
* загрузка рассчитанных витрин данных в целевую аналитическую СУБД ClickHouse.

## Условия, при соблюдении которых обеспечивается применение средства автоматизации в соответствии с назначением

Для стабильной работы приложения персональный компьютер должен соответствовать минимальным системным требованиям (см. таб. 1).

Таблица 1 – Минимальные системные требования

|  |  |
| --- | --- |
| Процессор | не менее 8 vCPU |
| ОЗУ | не менее 32 ГБ |
| Операционная система | Серверная ОС семейства Linux |
| Монитор | С поддержкой разрешения 1024x768 и выше |
| Устройства ввода | Мышь и клавиатура |
| Свободное место на диске | 500 ГБ минимум |

Приложение должно использоваться в стабильных условиях окружающей среды без воздействия экстремальных температур и влажности.

# Подготовка к работе

## Состав и содержание дистрибутивного носителя данных

Дистрибутив Системы поставляется в виде репозитория GitLab и содержит:

* исполняемые скрипты PySpark, реализующие логику ETL;
* конфигурационные файлы;
* файл определения DAG для Apache Airflow.

## Порядок развертывания и настройки

Установка приложения:

1. Клонировать репозиторий Системы на сервер Apache Airflow;
2. Разместить исполняемые скрипты PySpark в рабочем каталоге, доступном для  
   кластера Spark;
3. Скопировать файл определения DAG в директорию dags Apache Airflow;
4. Через веб-интерфейс Apache Airflow настроить подключения к источникам данных СУБД ClickHouse.

## Порядок проверки работоспособности

Для проверки корректного развертывания и работоспособности Системы выполнить следующие действия:

* запустить веб-интерфейс Apache Airflow;
* проверить, что в списке DAG появился новый процесс и напротив него отсутствует ошибка импорта;
* включить DAG, переведя переключатель Off в положение On;
* выполнить ручной запуск (см. п. 4.3.2) на тестовом наборе данных и убедиться в его успешном завершении.

# Описание операций

Данный раздел настоящего руководства пользователя содержит описание выполняемых функций, задач и процедур программой.

## Общее описание технологического процесса

Типичный процесс работы Системы включает следующие этапы:

* Apache Airflow запускает главный DAG по расписанию или вручную;
* задачи Spark подключаются к системам-источникам и считывают данные за прошедшие сутки;
* данные очищаются, обогащаются и агрегируются в памяти кластера Spark;
* рассчитанные витрины данных записываются в целевые таблицы в ClickHouse;
* Airflow фиксирует успешное или неуспешное выполнение процесса.

## Описание экранных форм и интерфейса

Управление и мониторинг Системы осуществляется через веб-интерфейс Apache Airflow.

### Главное окно (DAGs View)

При запуске веб-интерфейса открывается главное окно, представленное на рисунке 4.2.1.

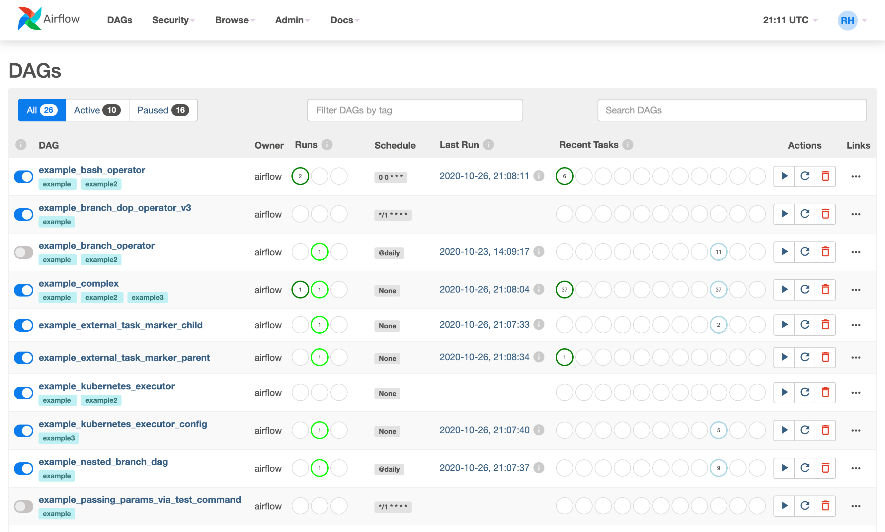


Рисунок 4.2.1 – Главное окно (DAGs View)

DAGs View отображает список всех доступных ETL-процессов (DAG), их статус, расписание и историю запусков.

### Окно процесса (Tree View)

При выборе конкретного ETL-процесса открывается окно процесса, представленное на рисунке 4.2.2.

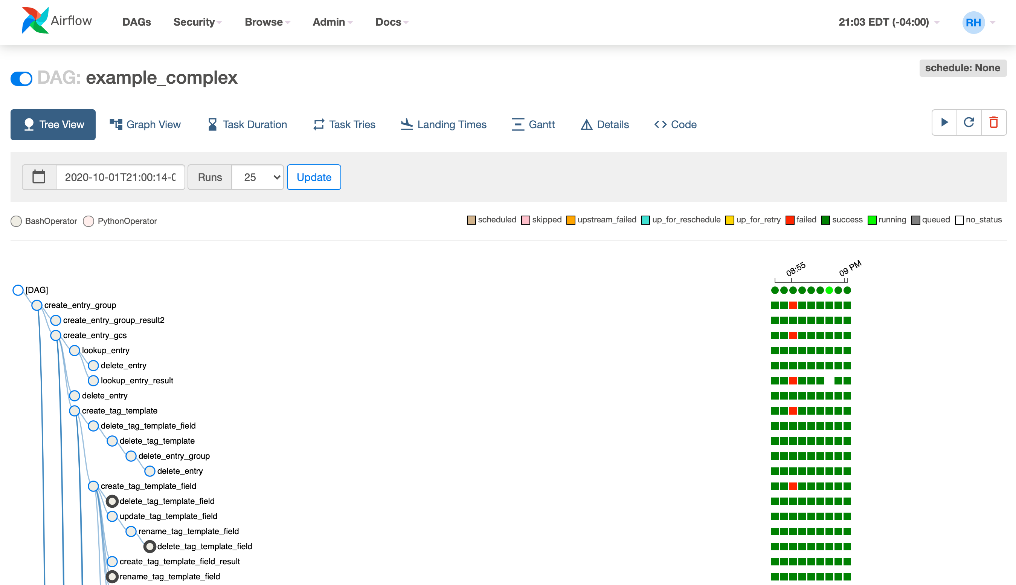


Рисунок 4.2.2 – Окно процесса (Tree View)

Представляет конкретный запуск ETL-процесса в виде дерева задач, где цветом обозначен статус каждой задачи (зеленый - успех, красный – сбой).

### Окно логов (Log View)

При выборе конкретной задачи ETL-процесса открывается окно логов выполнения данной задачи, представленное на рисунке 4.2.3.

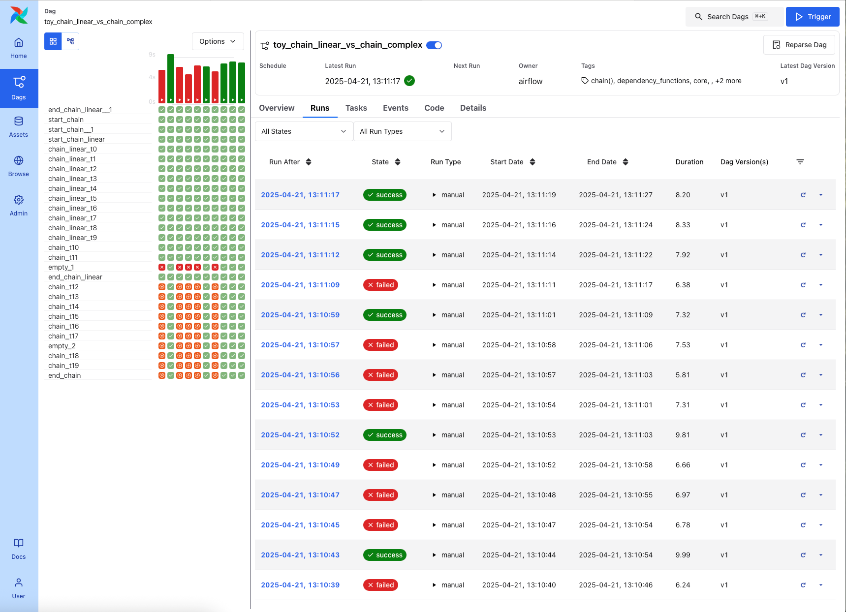


Рисунок 4.2.3 – Окно логов (Log View)

Данное текстовое окно содержит подробную информацию о ходе выполнения отдельной задачи, включая сообщения Spark и трассировку ошибок.

## Выполнение операций

### Мониторинг выполнения ETL-процесса

Подготовительные действия:

* Находиться в главном окне Apache Airflow.

Выполнение операции:

* Найти в списке нужный DAG;
* Нажать на его название, чтобы перейти к истории запусков;
* Выбрать интересующий запуск и перейти в Tree View для детального просмотра статуса задач.

Ожидаемый результат:

На экране отображается актуальный статус выполнения всех задач процесса.

### Ручной запуск ETL-процесса

Подготовительные действия:

* Находиться в главном окне Apache Airflow. DAG должен быть включен.

Выполнение операции:

* Нажать кнопку «Trigger DAG» напротив нужного DAG.

Ожидаемый результат:

Система инициирует новый запуск ETL-процесса, который появляется в истории запусков со статусом «running».

### Диагностика сбоев

Подготовительные действия:

* В Tree View одна из задач отображается красным цветом (статус «failed»).

Выполнение операции:

* Нажать на квадрат, соответствующий сбойной задаче;
* В появившемся меню выбрать «Log»;
* Прокрутить лог в самый низ для поиска сообщения об ошибке (ERROR, Traceback).

Ожидаемый результат:

На экране отображается лог выполнения задачи с детальным описанием причины сбоя.

### Перезапуск процесса после сбоя

Подготовительные действия:

* Причина сбоя, выявленная на шаге 4.3.3, устранена.

Выполнение операции:

* Нажать на квадрат сбойной задачи;
* В появившемся меню выбрать «Clear»;
* В окне подтверждения нажать «OK».

Ожидаемый результат:

Статус задачи сброшен. Планировщик Airflow автоматически поставит ее в очередь на повторное выполнение.

# Аварийные ситуации

В данном разделе настоящего руководства пользователя содержится информация о действиях пользователя в случаях возникновения аварийных ситуаций.

## Действия при сбоях в IT-инфраструктуре

Возможные ошибки:

* Недоступен веб-интерфейс Airflow;
* Задачи не запускаются;
* В логах ошибки подключения.

Действия:

* Обратиться к системному администратору для проверки работоспособности серверов, сети, кластера Spark и баз данных. После восстановления инфраструктуры выполните перезапуск (см. п. 4.3.4).

## Действия при ошибках в коде или логике ETL

Возможные ошибки:

* Процесс падает с ошибкой Spark, которую не удается устранить перезапуском.

Действия:

* Проанализировать код PySpark-скрипта, на который указывает лог;
* Внести исправления в логику;
* Развернуть обновленную версию кода (см. п. 3.2);
* Выполнить перезапуск задачи.

## Действия при обнаружении некорректных данных в источнике

Возможные ошибки:

* Процесс завершился успешно, но контрольная проверка витрины выявила аномалии, например, отрицательные суммы, дубликаты.

Действия:

* Сообщить администратору КХД о проблеме с данными в источнике;
* После исправления данных в источнике выполнить ручной запуск DAG для перерасчета витрины за соответствующий период.

# Рекомендации по освоению

Для успешной работы с приложением пользователю рекомендуется:

Начинать с ручных запусков – перед активацией расписания выполните несколько ручных запусков на тестовых данных, чтобы убедиться в корректности настроек.

Изучить интерфейс Airflow – ознакомьтесь со всеми представлениями (Graph, Gantt, Code) для полного понимания структуры и зависимостей процесса.

Использовать тестовую среду – все изменения в коде и конфигурации сначала проверяйте в тестовой среде, прежде чем применять в продуктивной.

# Рекомендации по использованию

Весь код ETL-процессов и DAG-файлы рекомендуется хранить в системе контроля версий.

Периодически отслеживайте потребление ресурсов (CPU, память) заданиями Spark через его собственный веб-интерфейс (Spark UI) для своевременной оптимизации.

Любые изменения в логике трансформации данных фиксируйте во внутренней документации для обеспечения прозрачности процесса.