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

**Руководство программиста**

Киров, 2025 г.АННОТАЦИЯ

Настоящий документ содержит основные положения и сведения, необходимые для сопровождения, модификации и дальнейшего развития программного комплекса «СОВД-Spark». Документ предназначен для программистов и дата-инженеров, обладающих квалификацией, достаточной для работы с используемым стеком технологий.

Документ разработан согласно требованиям ГОСТ Р 59795–2021 и ГОСТ Р 59853–2021.

**Содержание**

[1 Общие положения 4](#_Toc212653905)

[1.1 Наименование программы 4](#_Toc212653906)

[2 Назначение и условия применения 5](#_Toc212653907)

[2.1 Назначение программы 5](#_Toc212653908)

[2.2 Функции, выполняемые программой 5](#_Toc212653909)

[2.3 Сведения о технических и программных средствах, обеспечивающих выполнение программы 5](#_Toc212653910)

[2.3.1 Требования к составу и параметрам комплекса технических средств 5](#_Toc212653911)

[2.3.2 Требования к общесистемному программному обеспечению 6](#_Toc212653912)

[2.3.3 Требования к персоналу 7](#_Toc212653913)

[3 Характеристика программных средств программы 9](#_Toc212653914)

[3.1 Описание основных характеристик программы 9](#_Toc212653915)

[3.1.1 Временные характеристики программы 9](#_Toc212653916)

[3.1.2 Средства контроля правильности выполнения программы 9](#_Toc212653917)

[4 Обращение к программе 10](#_Toc212653918)

[5 Входные и выходные данные 11](#_Toc212653919)

[5.1 Характер и организация входных данных 11](#_Toc212653920)

[5.2 Характер и организация выходных данных 11](#_Toc212653921)

[6 Сообщения 12](#_Toc212653922)

[7 Аварийные ситуации 13](#_Toc212653923)

[7.1 Действия при сбоях в IT-инфраструктуре 13](#_Toc212653924)

[7.2 Действия при ошибках в коде или логике ETL 13](#_Toc212653925)

[7.3 Действия при обнаружении некорректных данных в источнике 13](#_Toc212653926)

# Общие положения

Настоящий документ содержит сведения, необходимые для работы программиста с программным обеспечением для анализа и обработки речевых сигналов.

## Наименование программы

Полное наименование системы: Программный комплекс для обновления аналитических витрин данных «СОВД-Spark».

# Назначение и условия применения

## Назначение программы

Основным назначением программного комплекса является автоматизация ETL-процессов (извлечение, преобразование, загрузка) в корпоративном хранилище данных компании «Энергосбыт».

## Функции, выполняемые программой

Перечень функций, реализуемых программой:

* извлечение данных из реляционных СУБД и файловых источников;
* распределенная обработка (очистка, трансформация, агрегация) данных с использованием Apache Spark;
* загрузка обработанных витрин в аналитическую СУБД ClickHouse;
* оркестрация и запуск конвейеров данных по расписанию с помощью Apache Airflow.

## Сведения о технических и программных средствах, обеспечивающих выполнение программы

### Требования к составу и параметрам комплекса технических средств

Средства аппаратного обеспечения, необходимые для функционирования программы, рассчитываются индивидуально под потребность решаемых задач, минимальные требования представлены в таблице.

Таблица 1 – Минимальные системные требования

|  |  |
| --- | --- |
| **Вид ресурса** | **Минимальные требования** |
| Тип ЭВМ | Серверная платформа x86-64 |
| Процессор | 8 vCPU |
| Оперативная память (ОЗУ) | Не менее 32 ГБ |
| Накопитель (HDD/SSD) | 500 ГБ для HDFS и системных файлов |
| Сетевое соединение | Ethernet 1 Гбит/с и выше |

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

### Требования к общесистемному программному обеспечению

Перечень и описание Системного ПО, обеспечивающего корректную работу приложения, представлены в таблице 2.

Таблица 2 – Сведения о программном обеспечении

|  |  |
| --- | --- |
| **Вид ресурса** | **Минимальные требования** |
| Операционная система | Серверная ОС семейства Linux |
| Интерпретатор Python | Версия не ниже 3.8 |
| Apache Hadoop | Версия 3.x (для YARN и HDFS) |
| Apache Spark | Версия 3.x |
| Apache Airflow | Версия 2.x |
| СУБД ClickHouse | Версия 21.x и выше |
| Библиотеки Python | PySpark, pandas, clickhouse-driver |
| Средства контроля версий | GitLab |

Средства, обеспечивающие функционирование Системы:

Сервер оркестрации Apache Airflow – реализует управляющую логику, запуск ETL-процессов по расписанию (DAGs). Взаимодействие с кластером происходит через YARN API.

Вычислительный кластер Apache Spark – выполняет ресурсоемкие вычисления. Задачи (jobs) написаны на Python с использованием библиотеки PySpark.

Распределенная файловая система HDFS – используется для хранения промежуточных данных (например, в формате Parquet) между этапами обработки.

Сетевое взаимодействие – осуществляется по стандартным протоколам (JDBC для доступа к базам данных, HTTP для взаимодействия с API сервисов).

Взаимодействие всех компонентов Системы реализуется в распределенной архитектуре, представленной в Техническом проекте. Компоненты (Airflow, Spark, HDFS) требуют стабильного сетевого взаимодействия между собой для корректной работы.

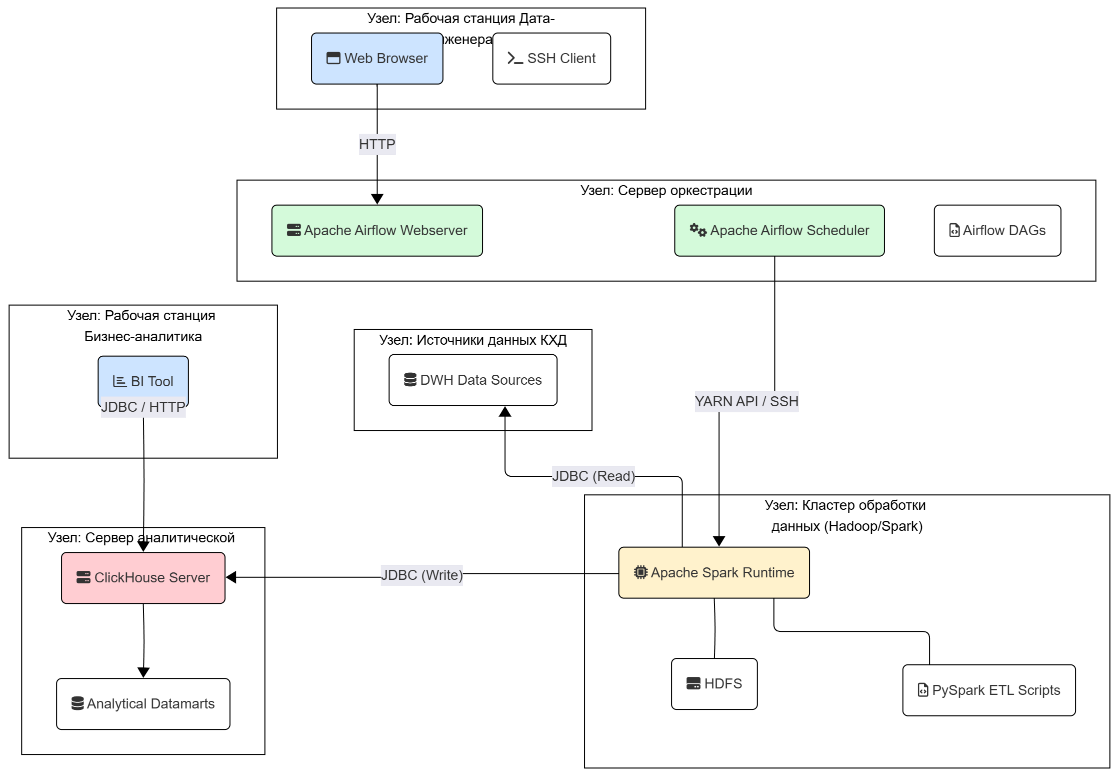


Рисунок 1 – Архитектурная схема функционирования системы

### Требования к персоналу

Обязательные навыки программиста/дата-инженера:

* Навыки работы в командной строке Linux;
* Знание стандартных утилит ОС (управление процессами, работа с файловой системой);
* Основы администрирования серверов на базе Linux.

Специальные навыки:

Python разработка:

* Глубокое знание Python 3.8+;
* Опыт работы с библиотекой PySpark для распределенной обработки данных;
* Навыки работы с Pandas для анализа и манипуляций с данными;
* Умение писать и оптимизировать Spark SQL запросы;
* Опыт разработки DAGs для Apache Airflow.

Базы данных:

* SQL язык (SELECT, INSERT, UPDATE, DELETE);
* Опыт проектирования реляционных БД и моделей данных (схемы "звезда", "снежинка");
* Понимание специфики колоночных СУБД, в частности ClickHouse (особенности синтаксиса, движки таблиц).

Смежные знания:

* Понимание принципов работы распределенных систем (экосистема Hadoop: HDFS, YARN);
* Основы проектирования ETL/ELT конвейеров и хранилищ данных;
* Знание форматов данных (Parquet, Avro, ORC).

Желательные навыки

* Опыт работы с Docker и контейнеризацией приложений;
* Знание принципов CI/CD и опыт работы с системами (Jenkins, GitLab CI);
* Опыт профилирования и оптимизации производительности Spark-приложений;
* Навыки работы с облачными платформами (AWS, Azure, Google Cloud);

# Характеристика программных средств программы

## Описание основных характеристик программы

Программный комплекс реализован как распределенная система, состоящая из:

* DAG-файлов для Apache Airflow, Python-скрипты, описывающие структуру, зависимости и расписание ETL-конвейеров.
* PySpark-приложений, основные модули на Python, содержащие бизнес-логику трансформации данных. Они запускаются на кластере Spark.
* Конфигурационных файлов, файлов в формате YAML или .conf, содержащие параметры подключения, пути к файлам и другие настройки, вынесенные из кода.

Архитектура кода PySpark-приложений. Код организован по модульному принципу:

* main.py: Точка входа в приложение, парсинг аргументов командной строки.
* jobs/: Директория с модулями для каждого ETL-процесса.
* readers/: Модули, отвечающие за чтение данных из различных источников.
* writers/: Модули, отвечающие за запись данных в целевые системы.
* transformers/: Модули с основной логикой трансформации данных.
* utils/: Вспомогательные функции (логирование, работа с датами и т. д.).

### Временные характеристики программы

Время инициализации Spark-сессии – до 30 секунд.

Время обработки данных зависит от объема данных и сложности трансформаций. Целевой показатель для суточной загрузки (<1ТБ) – не более 4 часов.

Время отклика Airflow UI стандартное для веб-приложений, не более 2 секунд на операцию.

### Средства контроля правильности выполнения программы

Для обеспечения корректности функционирования реализованы следующие механизмы:

* логирование, все компоненты системы (Airflow, Spark jobs) ведут подробные логи. Уровни логирования (INFO, WARN, ERROR) настраиваются в конфигурационных файлах;
* обработка исключений, в коде PySpark используются блоки try...except для перехвата и корректной обработки ошибок (например, недоступность источника), с последующим завершением задачи со статусом «failed»;
* Data Quality Checks, В ETL-процесс встроены проверки качества данных, например, проверка на NULL в ключевых полях, сверка контрольных сумм до и после обработки.

# Обращение к программе

Автоматический запуск осуществляется планировщиком Apache Airflow согласно расписанию, указанному в DAG-файле.

Ручной запуск инициируется через веб-интерфейс Apache Airflow (кнопка Trigger DAG) для целей отладки или внепланового пересчета.

Локальный запуск (для разработки) – PySpark-скрипты могут быть запущены локально с помощью команды spark-submit с передачей необходимых параметров через командную строку.

# Входные и выходные данные

## Характер и организация входных данных

Входные данные для системы «СОВД-Spark» представляют собой наборы данных из различных корпоративных систем, которые характеризуются следующими параметрами:

* Реляционные СУБД (Oracle, PostgreSQL): Данные извлекаются посредством JDBC-соединений с выполнением SQL-запросов;
* Форматы файлов: CSV, JSON, а также оптимизированные колоночные форматы, такие как Apache Parquet и Apache ORC;
* Структура данных: входные данные организованы в виде таблиц или наборов файлов, содержащих информацию о потреблении электроэнергии, клиентах, тарифах и платежах за определенный период;
* Объем данных: Система рассчитана на обработку больших объемов данных, от сотен гигабайт до нескольких терабайт за один цикл обновления;
* Параметры запуска: в качестве входного параметра ETL-процесс принимает дату или период, за который необходимо обработать данные. Этот параметр передается оркестратором Apache Airflow.

Данные проходят несколько этапов обработки внутри Spark: фильтрацию, очистку от аномалий и пропусков, обогащение из справочников и агрегацию.

## Характер и организация выходных данных

Выходные данные формируются после обработки входящих данных и содержат:

* Аналитические витрины данных:
  + Основной результат работы. Представляют собой таблицы в целевой колоночной СУБД ClickHouse;
  + Структура витрин спроектирована по схеме "Звезда" для обеспечения высокой скорости аналитических запросов;
  + Содержат агрегированные показатели (факты) и ссылки на измерения.
* Промежуточные данные: Наборы данных, сохраненные на HDFS в формате Apache Parquet.
* Логи выполнения, включающие информационные сообщения, предупреждения и детальные отчеты об ошибках с трассировкой стека.
* Отчеты о качестве данных, таблицы или файлы (CSV, JSON), содержащие результаты автоматических проверок качества данных.

# Сообщения

В программном комплексе «СОВД-Spark» сообщения генерируются различными компонентами (Apache Airflow, Apache Spark) и фиксируются в лог-файлах. Они обеспечивают информирование разработчика и инженера по эксплуатации о ключевых состояниях и событиях системы. Сообщения подразделяются на следующие виды:

* Информационные сообщения (INFO): уведомляют о штатном ходе выполнения операций. Отображаются в логах для трассировки процесса.
* Предупреждения (WARN): информируют о потенциальных проблемах, которые не привели к сбою, но требуют внимания.
* Ошибки (ERROR): отображаются при возникновении критических сбоев, которые приводят к остановке задачи.
* Аварийные сообщения (FATAL): сопровождают серьезные проблемы на уровне JVM или всего кластера, ведущие к невозможности продолжения работы.

Сообщения доступны для анализа через веб-интерфейс Apache Airflow. Формат сообщений стандартизирован и обычно включает: временную метку, уровень сообщения, имя модуля и текстовое описание. Сообщения об ошибках сопровождаются полной трассировкой стека.

# 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.