**ОПИСАНИЕ ФУНКЦИОНАЛЬНЫХ ХАРАКТЕРИСТИК**

**RT.DATALOADER**

2021

ТЕРМИНЫ И ОПРЕДЕЛЕНИЯ

| Термин | Определение |
| --- | --- |
| RT.DataLoader, Система | Система выгрузки и доставки данных, которая обеспечивает извлечение данных из систем-источников, их доставку и дальнейшую интеграцию данных в файловую систему HDFS ХД.  RT.DataLoader состоит из следующих компонентов:   * Модуль выгрузки; * Модуль доставки; * Оркестратор; * База матаданных. |
| Модуль | Обособленный компонент RT.DataLoader, выполняющий определённые функции, имеющий конкретную структуру и использующий конкретные технологии. |
| Модуль выгрузки | Модуль RT.DataLoader, предназначенный для извлечения данных из систем-источников в Область файлового обмена. |
| Модуль доставки | Модуль RT.DataLoader, предназначенный для доставки данных из Области файлового обмена в Интерфейсную область Системы. |
| БД метаданных, БД | База данных, которая содержит настроечные и управляющие процессами таблицы и данные. |
| Система – источник, СИ | Информационная система, интегрированная с Хранилищем данных посредством RT.DataLoader, являющаяся поставщиком данных для последующей обработки в Хранилище и формирования отчётности на их основе. |
| Хранилище данных, ХД | Информационно-аналитическая система, являющаяся приемником данных, доставляемых посредством RT.DataLoader и включающая в себя:   * нереляционное хранилище (Hive, HDFS, Hadoop) — программно-аппаратный комплекс и комплекс инструментов, предназначенный для загрузки, хранения и обработки «больших» данных; * реляционное хранилище (Greenplum) — хранилище данных на базе реляционной базы данных. |
| Область файлового обмена | Место размещения данных, выгруженных из систем-источников. |
| Интерфейсная область | Место размещения данных, доставленных из Области файлового обмена в файловую систему HDFS ХД. |
| Оркестровка | Единый процесс, который управляет последовательностью взаимодействия компонентов RT.DataLoader в соответствии с бизнес-логикой. |
| Разработчик | Организация, которая выполняет разработку задач (в том числе анализ требований, проектирование, приёмочные испытания) в процессе жизненного цикла. |
| Airflow | Инструмент, используемый для оркестровки процессов выгрузки и доставки данных в ХД. |

ОГЛАВЛЕНИЕ

[ТЕРМИНЫ И ОПРЕДЕЛЕНИЯ 2](#_Toc74056857)

[1 ФУНКЦИОНАЛЬНЫЕ ХАРАКТЕРИСТИКИ 5](#_Toc74056858)

[2 ПРИНЦИПЫ ФУНКЦИОНИРОВАНИЯ RT.DATALOADER 7](#_Toc74056859)

[2.1 Структура и функционирование RT.DataLoader 7](#_Toc74056860)

[2.2 Реализация потока данных 7](#_Toc74056861)

[3 БИЗНЕС-ПРОЦЕССЫ 8](#_Toc74056862)

ФУНКЦИОНАЛЬНЫЕ ХАРАКТЕРИСТИКИ

Инструмент RT.DataLoader предназначен для выгрузки данных из источников, поддерживающих соединение jdbc, и доставки в файловое хранилище HDFS.

Позволяет:

* выполнять подключения к реляционным БД;
* извлекать данные из таблиц и представлений по фильтру и целиком или по sql-запросу;
* приземлять данные в HDFS.

Даёт возможность быстро подключать к ХД новые источники данных, новые таблицы, корректировать атрибутивный состав.

Продукт автоматизирует процесс выгрузки, позволяет гибко настраивать расписание выгрузок, формировать очередь выгрузок, поддерживает регламентные выгрузки данных, позволяет балансировать нагрузку на сервере выгрузки. Даёт возможность значительно сократить привлечение ETL разработчиков к процессу подключения новых таблиц источников к ХД, сокращая стоимость расширения ХД.

**Модуль выгрузки** предназначен для извлечения данных из систем-источников в Область файлового обмена.

Основными функциями Модуля выгрузки являются:

мониторинг наличия данных в СИ;

формирование файловых пакетов данных;

выгрузка данных из систем-источников;

размещение выгруженных пакетов данных в Области файлового обмена;

логирование запусков процесса выгрузки;

логирование процесса формирования пакетов данных;

фиксирование количества выгруженных строк.

**Модуль доставки** предназначен для доставки пакетов данных из Области файлового обмена в Интерфейсную область.

Основными функциями Модуля доставки являются:

мониторинг наличия готовых к доставке данных в Области файлового обмена;

доставка пакета данных в Интерфейсную область;

очистка Области файлового обмена от доставленных ранее данных;

контроль качества данных на всех этапах процесса доставки;

логирование процесса доставки данных;

архивация успешно доставленных пакетов в специальной директории в Интерфейсной области.

**Оркестратор** предназначен для оркестрации процесса выгрузки и доставки данных.

Основными функциями оркестратора являются:

оркестрация процесса выгрузки, доставки данных из систем-источников;

управление нагрузкой на серверы выгрузки и доставки.

**База метаданных** предназначена для централизованного хранения метаданных, таких как:

* флаги синхронизации процессов выгрузки, доставки;
* настроечные и конфигурационные данные, логи;
* данные для автоматизации процессов выгрузки, доставки.

Основными функциями Базы метаданных являются:

хранение метаданных.

**RT.DataLoader обладает следующими особенностями:**

1. высокое быстродействие;
2. использование для больших объёмов данных;
3. работа с данными, полученными напрямую из систем-источников;
4. наличие области хранения выгруженных данных;
5. надёжность хранения данных, архивация данных;
6. техническая поддержка, обеспечиваемая Разработчиком;
7. протестированная совместимость компонентов.

ПРИНЦИПЫ ФУНКЦИОНИРОВАНИЯ RT.DATALOADER

Структура и функционирование RT.DataLoader

RT.DataLoader включает в себя следующие функциональные компоненты, предназначенные для решения соответствующих комплексов задач:

1. Модуль выгрузки;
2. Модуль доставки;
3. Оркестратор Airflow;
4. БД метаданных.

Реализация потока данных

Схема взаимодействия функциональных компонентов RT.DataLoader представлена на Рисунке 1.



Рисунок 1 – Схема взаимодействия функциональных компонентов RT.DataLoader

БИЗНЕС-ПРОЦЕССЫ

RT.DataLoader может выполнять автоматизацию следующих бизнес-процессов:

1. сбор и передача большого объёма данных из реляционных источников для последующей структуризации, аналитики и выявления необходимых компании данных;
2. архивация данных для компактного хранения данных;
3. оценка качества загруженных данных, для минимизации ошибок, обеспечивающая надёжность данных;
4. разделение процессов выгрузки и доставки данных для снижения нагрузки на источник при проблемах на стороне ХД;
5. управление и мониторинг процессами выгрузки и доставки данных;
6. управление нагрузкой на сервер выгрузки.