**ИНСТРУКЦИЯ**

**ПО УСТАНОВКЕ ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ**

**RT.DATALOADER**

2020

ТЕРМИНЫ И ОПРЕДЕЛЕНИЯ

| Термин | Определение |
| --- | --- |
| Инструмент доставки данных RT.DataLoader | Система выгрузки и доставки данных, которая обеспечивает извлечение данных из систем-источников, их доставку и дальнейшую интеграцию данных в файловую систему HDFS ХД. |
| 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. |
| 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, 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 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. |
| БД метаданных | База данных, которая содержит настроечные и управляющие процессами ETL таблицы и данные. |

ОГЛАВЛЕНИЕ

[ТЕРМИНЫ И ОПРЕДЕЛЕНИЯ 2](#_Toc66893199)

[1 Назначение системы и условия применения 4](#_Toc66893200)

[1.1 Назначение RT.DataLoader 4](#_Toc66893201)

[1.2 Условия применения 4](#_Toc66893202)

[2 Подготовка к установке 6](#_Toc66893203)

[3 Установка 7](#_Toc66893204)

[3.1 Установка модуля выгрузки 7](#_Toc66893205)

[3.2 Установка модуля доставки 7](#_Toc66893206)

[3.3 Установка Airflow 7](#_Toc66893207)

[3.4 Установка сервера БД 8](#_Toc66893208)

Назначение системы и условия применения

Назначение RT.DataLoader

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 атрибутный 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.

RT.DataLoader состоит из следующих модулей:

* Модуль выгрузки;
* Модуль доставки;
* Оркестратор;
* База матаданных.

Условия применения

Работа с инструментом RT.DataLoader доступна пользователям в соответствии с установленными правами доступа.

### Сервер Модуля выгрузки:

Минимальная конфигурация

* Центральный процессор: 1 ядро;
* Оперативная память: 200 Мб;
* Свободное дисковое пространство: 5 Гб;
* Доступ в интернет: скорость соединения 0,1 МБайт/с;
* Операционная система: Linux

Предварительную конфигурацию вычислительных ресурсов сервера для размещения модуля выгрузки данных можно оценить с помощью калькулятора:



Предустановленное программное обеспечение:

* Java;
1. Java версия "1.8.0\_144";
2. Java(TM) SE Runtime Environment (build 1.8.0\_144-b01);
3. Java HotSpot(TM) 64-Bit Server VM (build 25.144-b01, mixed mode).

### Сервер Модуля доставки, оркестратора и БД метаданных

* Центральный процессор: не менее 2 ядер;
* Оперативная память: не менее 8 Гб;
* Свободное дисковое пространство: не менее 70 Гб;
* Операционная система: Linux

Предустановленное программное обеспечение:

* Java;
1. Java версия "1.8.0\_144";
2. Java(TM) SE Runtime Environment (build 1.8.0\_144-b01);
3. Java HotSpot(TM) 64-Bit Server VM (build 25.144-b01, mixed mode).
* БД PosgreSQL версия 12.5 или выше;
* Airflow версия 2.0
* Python версия 3.6 – 3.8
* SVN система.

### Персональный компьютер пользователя:

* Центральный процессор: не менее 2 ГГц;
* Оперативная память: не менее 4 Гб;
* Предустановленное программное обеспечение:
	+ Браузер
	+ Текстовый редактор

Подготовка к установке

Установка компонентов системы RT.DataLoader производится на двух серверах:

1. Установка модуля выгрузки на сервер источника. Файл для установки находится в каталоге UNLOADING дистрибутива RTDL\_distribution.
2. Установка модуля доставки, оркестратора, БД метаданных на сервер RT.DataLoader. Файлы установки находятся в каталоге DELIVERY дистрибутива RTDL\_distribution.

Установка

Установка всех компонент системы происходит поэтапно.

Установка модуля выгрузки

Для того чтобы установить модуль необходимо выполнить следующие действия:

1. Сохранить каталог UNLOADING из дистрибутива RTDL\_distribution в желаемую директорию;

В каталоге содержатся файлы приложения и Makefile для установки Модуля выгрузки

1. Содержимое дистрибутива:

target\rtcLoad\bin\rtlc-load.bat

target\rtcLoad\bin\run.bat

target\rtcLoad\bin\run.sh

target\rtcLoad\bin\run\_cache.sh

target\rtcLoad\bin\run\_microsoftsql.sh

target\rtcLoad\bin\run\_mysql.sh

target\rtcLoad\bin\run\_oracle.sh

target\rtcLoad\bin\run\_oracle\_new.sh

target\rtcLoad\bin\run\_postgres.sh

target\rtcLoad\bin\test\_run.sh

target\rtcLoad\config\application.properties

target\rtcLoad\config\logback-spring-debug.xml

target\rtcLoad\config\logback-spring.xml

target\rtcLoad\config\pull.yaml

target\rtcLoad\config\syslog.properties

target\rtcLoad\jdbc\readme.txt

target\rtcLoad\key\key

target\rtcLoad\lib\activation-1.1.jar

target\rtcLoad\lib\antlr-2.7.7.jar

target\rtcLoad\lib\antlr-3.2.jar

target\rtcLoad\lib\antlr-runtime-3.2.jar

target\rtcLoad\lib\aopalliance-1.0.jar

target\rtcLoad\lib\apache-log4j-extras-1.2.17.jar

target\rtcLoad\lib\aspectjrt-1.8.7.jar

target\rtcLoad\lib\aspectjweaver-1.8.7.jar

target\rtcLoad\lib\audience-annotations-0.5.0.jar

target\rtcLoad\lib\avro-1.4.0-cassandra-1.jar

target\rtcLoad\lib\avro-1.8.1.jar

target\rtcLoad\lib\bonecp-0.8.0.RELEASE.jar

target\rtcLoad\lib\cassandra-all-0.8.1.jar

target\rtcLoad\lib\cassandra-thrift-0.8.1.jar

target\rtcLoad\lib\com.ibm.jbatch-tck-spi-1.0.jar

target\rtcLoad\lib\common-config-5.3.0.jar

target\rtcLoad\lib\common-utils-5.3.0.jar

target\rtcLoad\lib\commons-cli-1.3.1.jar

target\rtcLoad\lib\commons-codec-1.9.jar

target\rtcLoad\lib\commons-collections-3.2.1.jar

target\rtcLoad\lib\commons-compiler-2.7.8.jar

target\rtcLoad\lib\commons-compress-1.10.jar

target\rtcLoad\lib\commons-csv-1.4.jar

target\rtcLoad\lib\commons-lang-2.4.jar

target\rtcLoad\lib\commons-logging-1.2.jar

target\rtcLoad\lib\commons-testing-1.0.1.jar

target\rtcLoad\lib\commons-testing-core-1.0.1.jar

target\rtcLoad\lib\concurrentlinkedhashmap-lru-1.1.jar

target\rtcLoad\lib\guava-15.0.jar

target\rtcLoad\lib\hamcrest-core-1.3.jar

target\rtcLoad\lib\hamcrest-library-1.3.jar

target\rtcLoad\lib\high-scale-lib-1.1.2.jar

target\rtcLoad\lib\hsqldb-2.3.3.jar

target\rtcLoad\lib\httpclient-4.3.6.jar

target\rtcLoad\lib\httpcore-4.3.3.jar

target\rtcLoad\lib\jackson-annotations-2.11.3.jar

target\rtcLoad\lib\jackson-core-2.11.3.jar

target\rtcLoad\lib\jackson-core-asl-1.4.0.jar

target\rtcLoad\lib\jackson-databind-2.11.3.jar

target\rtcLoad\lib\jackson-mapper-asl-1.4.0.jar

target\rtcLoad\lib\jamm-0.2.2.jar

target\rtcLoad\lib\janino-2.7.8.jar

target\rtcLoad\lib\JavaEWAH-1.1.6.jar

target\rtcLoad\lib\javafaker-0.6.jar

target\rtcLoad\lib\javax.batch-api-1.0.jar

target\rtcLoad\lib\javax.mail-1.5.4.jar

target\rtcLoad\lib\jcl-over-slf4j-1.7.12.jar

target\rtcLoad\lib\jettison-1.2.jar

target\rtcLoad\lib\jetty-6.1.22.jar

target\rtcLoad\lib\jetty-util-6.1.22.jar

target\rtcLoad\lib\jline-0.9.94.jar

target\rtcLoad\lib\joda-time-2.9.1.jar

target\rtcLoad\lib\jsch-0.1.53.jar

target\rtcLoad\lib\json-simple-1.1.jar

target\rtcLoad\lib\jsr250-api-1.0.jar

target\rtcLoad\lib\jsr305-3.0.2.jar

target\rtcLoad\lib\jul-to-slf4j-1.7.12.jar

target\rtcLoad\lib\junit-4.12.jar

target\rtcLoad\lib\kafka-avro-serializer-5.3.0.jar

target\rtcLoad\lib\kafka-clients-2.6.0.jar

target\rtcLoad\lib\kafka-json-serializer-4.1.1.jar

target\rtcLoad\lib\kafka-log4-appender-0.0.2.jar

target\rtcLoad\lib\kafka-schema-registry-client-5.3.0.jar

target\rtcLoad\lib\libthrift-0.6.1.jar

target\rtcLoad\lib\log4j-1.2.17.jar

target\rtcLoad\lib\log4j-over-slf4j-1.7.12.jar

target\rtcLoad\lib\logback-classic-1.1.3.jar

target\rtcLoad\lib\logback-core-1.1.3.jar

target\rtcLoad\lib\lz4-java-1.7.1.jar

target\rtcLoad\lib\mockito-core-1.10.19.jar

target\rtcLoad\lib\netty-3.10.6.Final.jar

target\rtcLoad\lib\objenesis-2.1.jar

target\rtcLoad\lib\org.eclipse.jgit-4.6.0.201612231935-r.jar

target\rtcLoad\lib\paranamer-2.7.jar

target\rtcLoad\lib\rtcLoad-3.0.1.jar

target\rtcLoad\lib\servlet-api-2.5-20081211.jar

target\rtcLoad\lib\servlet-api-2.5.jar

target\rtcLoad\lib\slf4j-api-1.7.7.jar

target\rtcLoad\lib\slf4j-simple-1.6.1.jar

target\rtcLoad\lib\snakeyaml-1.14.jar

target\rtcLoad\lib\snappy-java-1.1.7.3.jar

target\rtcLoad\lib\spotbugs-annotations-3.1.9.jar

target\rtcLoad\lib\spring-aop-4.1.8.RELEASE.jar

target\rtcLoad\lib\spring-batch-core-3.0.5.RELEASE.jar

target\rtcLoad\lib\spring-batch-infrastructure-3.0.5.RELEASE.jar

target\rtcLoad\lib\spring-beans-4.1.8.RELEASE.jar

target\rtcLoad\lib\spring-boot-1.2.7.RELEASE.jar

target\rtcLoad\lib\spring-boot-autoconfigure-1.2.7.RELEASE.jar

target\rtcLoad\lib\spring-boot-starter-1.2.7.RELEASE.jar

target\rtcLoad\lib\spring-boot-starter-aop-1.2.7.RELEASE.jar

target\rtcLoad\lib\spring-boot-starter-batch-1.2.7.RELEASE.jar

target\rtcLoad\lib\spring-boot-starter-jdbc-1.2.7.RELEASE.jar

target\rtcLoad\lib\spring-boot-starter-logging-1.2.7.RELEASE.jar

target\rtcLoad\lib\spring-boot-starter-mail-1.2.7.RELEASE.jar

target\rtcLoad\lib\spring-boot-starter-test-1.2.7.RELEASE.jar

target\rtcLoad\lib\spring-context-4.1.8.RELEASE.jar

target\rtcLoad\lib\spring-context-support-4.1.8.RELEASE.jar

target\rtcLoad\lib\spring-core-4.1.8.RELEASE.jar

target\rtcLoad\lib\spring-expression-4.1.8.RELEASE.jar

target\rtcLoad\lib\spring-jdbc-4.1.8.RELEASE.jar

target\rtcLoad\lib\spring-retry-1.1.0.RELEASE.jar

target\rtcLoad\lib\spring-test-4.1.8.RELEASE.jar

target\rtcLoad\lib\spring-tx-4.1.8.RELEASE.jar

target\rtcLoad\lib\stringtemplate-3.2.jar

target\rtcLoad\lib\svnClientAadapter-1.0.jar

target\rtcLoad\lib\svnjavahl-1.0.jar

target\rtcLoad\lib\tomcat-jdbc-8.0.28.jar

target\rtcLoad\lib\tomcat-juli-8.0.28.jar

target\rtcLoad\lib\velocity-1.7.jar

target\rtcLoad\lib\xmlpull-1.1.3.1.jar

target\rtcLoad\lib\xpp3\_min-1.1.4c.jar

target\rtcLoad\lib\xstream-1.4.7.jar

target\rtcLoad\lib\xz-1.6.jar

target\rtcLoad\lib\zkclient-0.10.jar

target\rtcLoad\lib\zookeeper-3.4.14.jar

target\rtcLoad\lib\zstd-jni-1.4.4-7.jar

1. Запустить Makefile;

make install

Каталог установки по умолчанию /opt/rtc/rtcload

Дождаться "install success"

Установка ПО на сервер RT.DataLoader

Для того чтобы установить ПО на серевер RT.DataLoader необходимо сохранить каталог DELIVERY из дистрибутива RTDL\_distribution в желаемую директорию сервера.

Дистрибутив содержит:

* Набор файлов для установки модуля выгрузки в каталоге RTDL\_J\_JDDM;
* Файлы для установки оркестратора RTDL\_AIRFLOW;
* Скрипты для создания БД метаданных RTDL\_DB.

### Модуль доставки

Для того чтобы установить модуль необходимо выполнить следующие действия:

1. Сохранить каталог из дистрибутива RTDL\_distribution/ DELIVERY/ RTDL\_J\_JDDM в желаемую директорию;

В каталоге содержатся файлы приложения и Makefile для установки Модуля доставки

1. Запустить Makefile;

make install

Каталог установки по умолчанию /opt/rtc/rtcload

Дождаться "install success"

### Установка Airflow

Для проведения оркестрации процессов используется инструмент Airflow.

Этапы установки Airflow:

1. Установить Airflow

AIRFLOW\_VERSION=2.0.1

PYTHON\_VERSION="$(python --version | cut -d " " -f 2 | cut -d "." -f 1-2)"

CONSTRAINT\_URL="https://raw.githubusercontent.com/apache/airflow/constraints-${AIRFLOW\_VERSION}/constraints-${PYTHON\_VERSION}.txt"

# For example: https://raw.githubusercontent.com/apache/airflow/constraints-2.0.1/constraints-3.6.txt

pip install "apache-airflow==${AIRFLOW\_VERSION}" --constraint "${CONSTRAINT\_URL}"

1. Создать БД

CREATE DATABASE airflow\_db;

CREATE USER airflow\_user WITH PASSWORD 'airflow\_user';

GRANT ALL PRIVILEGES ON DATABASE airflow\_db TO airflow\_user;

1. Инициализировать БД

airflow db init

1. Распаковать дистрибутив RTDL\_AIRFLOW.zip в желаемую директорию;

Дистрибутив содержит:

./ddm\_airflow/ddm\_check\_dag.py

./ddm\_airflow/ddm\_clear\_archive\_dag.py

./ddm\_airflow/ddm\_clear\_pkg\_dag.py

./ddm\_airflow/ddm\_replication\_dag.py

./ddm\_airflow/ddm\_run\_dag.py

./ddm\_airflow/ddm\_scripts/ddm\_lib.py

./ddm\_airflow/ddm\_scripts/ddm\_pkg\_clear.py

./ddm\_airflow/ddm\_scripts/ddm\_pkg\_ctl.py

./ddm\_airflow/ddm\_scripts/ddm\_pkg\_lib.py

./ddm\_airflow/ddm\_scripts/ddm\_pkg\_rep.py

./ddm\_airflow/ddm\_scripts/ddm\_queries.py

./export\_package\_airflow/ddm\_manual\_export\_dag.py

./export\_package\_airflow/export\_package\_dag\_pattern.py

./export\_package\_airflow/scripts/export\_lib.py

./export\_package\_airflow/scripts/export\_manual\_lib.py

./export\_package\_airflow/scripts/export\_package\_lib.py

Дополнительная информация по установке Airflow: <https://airflow.apache.org/docs/apache-airflow/stable/installation.html>

### Установка БД метаданных RT.DataLoader

Необходимая версия PostgreSQL 12.5 устанавливается для различных версий ОС. Пройдя по ссылке <https://www.postgresql.org/download/>, выберите необходимую ОС и загрузите установочный файл.

В БД PostgrSQL выполнить скрипты создания объектов из каталогов RTDL и RTDL\_READ дистрибутива RT.DataLoader под пользователем RTDL и RTDL\_READ соответственно.