«OTRi.DI»

Описание функциональных характеристик

ОТР Интеграция	Проект:	«OTRi.DI»
	Документ:	Описание функциональных характеристик
	Дата:	16.07.2024
	Версия:	1.0

СОДЕРЖАНИЕ

1 Термины и сокращения

Термины и сокращения	Определения
API	Программный интерфейс, то есть описание способов взаимодействия одной компьютерной программы с другими
AVRO	Ориентированная на строки среда для удаленного вызова процедур и сериализации данных, разработанная в рамках проекта Apache Hadoop. Он использует JSON для определения типов данных и протоколов и сериализует данные в компактном двоичном формате
CSV	Текстовый формат, предназначенный для представления табличных данных. Строка таблицы соответствует строке текста, которая содержит одно или несколько полей, разделенных запятыми
Greenplum	
gzip	Утилита сжатия и восстановления файлов, использующая алгоритм Deflate. Применяется в основном в UNIX-системах, в ряде которых является стандартом де-факто для сжатия данных
JSON	Текстовый формат обмена данными, основанный на JavaScript
HTML	Стандартизированный язык гипертекстовой разметки документов для просмотра веб-страниц в браузере
HTTP-заголовок	Строки в HTTP-сообщении, содержащие разделённую двоеточием пару имя-значение. Формат заголовков соответствует общему формату заголовков текстовых сетевых сообщений ARPA (см. RFC 822). Заголовки должны отделяться от тела сообщения хотя бы одной пустой строкой
HTTPS	(аббр. от англ. HyperText Transfer Protocol Secure) – расширение протокола HTTP для поддержки шифрования в целях повышения безопасности
MS SQL Server	Система управления реляционными базами данных, разработанная корпорацией Microsoft
POP3	(англ. Post Office Protocol Version 3 – протокол почтового отделения, версия 3) – стандартный интернет-протокол прикладного уровня, используемый клиентами электронной почты для получения почты с удалённого сервера по TCP-соединению
PostgreSQL	Свободная объектно-реляционная система управления базами данных (СУБД)
REST	(от англ. REpresentational State Transfer – «передача репрезентативного состояния» или «передача „самоописываемого“ состояния») – архитектурный стиль взаимодействия компонентов распределённого приложения в сети
SQL	Декларативный язык программирования, применяемый для создания, модификации и управления данными в реляционной базе данных, управляемой соответствующей системой управления базами данных
SOAP	(от англ. Simple Object Access Protocol – протокол доступа к объектам) – протокол обмена структурированными сообщениями в распределённой вычислительной среде
SFTP	(англ. Secure File Transfer Protocol) – протокол прикладного уровня передачи файлов, работающий поверх безопасного канала. Предназначен для копирования и выполнения других операций с файлами поверх надёжного и безопасного соединения
SSL	(англ. Secure Sockets Layer – уровень защищённых сокетов) – криптографический протокол, который подразумевает более безопасную связь

Термины и сокращения	Определения
XLS	Файлы с расширением XLS представляют собой формат двоичных файлов Excel
XML	(англ. eXtensible Markup Language) – «расширяемый язык разметки». Спецификация XML описывает XML-документы и частично описывает поведение XML-процессоров (программ, читающих XML-документы и обеспечивающих доступ к их содержимому)
ZIP	Формат архивации файлов и сжатия данных без потерь. Архив ZIP может содержать один или несколько файлов и каталогов, которые могут быть сжаты разными алгоритмами
База данных	Сущность, обеспечивающая хранение и применение параметров доступа (адрес, порт, логин, пароль пользователя и др. настройки) к различным источникам данных (базам данных, слою представления Oracle BI)
ПО	Программное обеспечение
ПО «OTRi.DI»	Программное обеспечение для подключения к источникам данных, извлечения и трансформации данных и их последующей загрузки в целевые БД
Процесс	Ориентированный граф, шаги которого выполняют высокоуровневые задачи: от получения файлов с веб-сервера до запуска других трансформаций. Связи графа процесса определяют порядок выполнения и условия запуска шагов
СУБД	Комплекс программно-языковых средств, позволяющих создать базы данных и управлять данными
Трансформация	Ориентированный граф, состоящий из шагов (блоков, выполняющих определенные действия с данными) и настроенных между ними связей. Эти связи задают порядок передачи данных между шагами. При старте трансформации шаги выполняются последовательно в зависимости от указанных связей, первыми запускаются шаги, не имеющие входящих связей

2 Функциональные характеристики

ПО «OTRi.DI» обеспечивает выполнение функций, перечисленных в разделах 2.1–2.4 настоящего документа.

2.1 Мониторинг и настройка уведомлений

ПО «OTRi.DI» позволяет отслеживать и анализировать в реальном времени продолжительность выполнения как трансформаций и процессов в целом, так и отдельных составляющих их шагов. Для нахождения проблемных мест в трансформациях и процессах собираются метрики. Основные действия системы записываются в лог, глубину детализации которого можно настраивать.

Мониторинг работоспособности приложения происходит с использованием стандартных средств, таких как инструменты логирования, сбора и визуализации метрик, а также отправки уведомлений по условиям.

Для обеспечения контроля зависимостей в ПО «OTRi.DI» предусмотрено регистрирование изменений структуры данных в источниках, настройка сценария обработки таких событий; отслеживание доступности внешних источников для всех типов ресурсов и индикация статусов.

Функционал уведомлений включает в себя следующие возможности:

– получение уведомлений о недоступности внешних ресурсов;

– получение отчетов о выполнении трансформаций и процессов на электронную почту;

– настройка уведомлений при выполнении различных условий (выход метрик за граничные значения, появление данных в источнике и др.).

2.2 Загрузка и трансформация данных из источников

ПО «OTRi.DI» располагает различными методами интеграции с источниками: файлообменные протоколы, API, MQ и прямое подключение к БД и др. Для интеграции посредством API используются протоколы SOAP и REST.

Алгоритмы загрузки данных можно описывать в текстовом формате, поддерживающем чтение и хранение в системе контроля версий.

Файлы размером до 100 Гб могут обрабатываться в поточном режиме. Также в ПО предусмотрена параллельная обработка данных в несколько потоков для более эффективного использования ресурсов.

Настройка подключения к ресурсам производится в отдельной конфигурации для использования в трансформациях и процессах только логических названий ресурсов, без прописывания параметров подключения.

В ПО «OTRi.DI» предусмотрена поддержка обработки ошибок загрузки и трансформации данных, настройки повторных запусков и интервала между ними, а также логирование ошибок.

ПО «OTRi.DI» позволяет работать со следующими видами источников:

– файлы (данные могут читаться из файловых ресурсов следующих типов: локальные и сетевые диски, FTP/FTPS, SFTP. Также предусмотрена возможность работы с файловыми директориями);

– веб-ресурсы:

а) данные могут быть получены посредством веб-запросов;

б) базовая (basic) аутентификация происходит с использованием защищенных протоколов передачи данных. Поддерживается веб-запрос аутентификации с сохранением cookie для последующих запросов к этому веб-ресурсу и управление набором HTTP-заголовков;

в) предусмотрено использование SSL-сертификатов и отправка запросов по HTTPS;

– получение писем (электронная почта) по протоколу POP3 с возможностью чтения тела письма, а также вложений;

– базы данных:

а) выборка данных может быть получена в результате выполнения SQL-запросов из баз данных следующих типов СУБД: PostgreSQL, Greenplum, Oracle, MySQL, MS SQL Server и др.;

б) данные, полученные из внешних источников, могут быть объединены с данными в БД с возможностями оптимизации, настраиваемым перечнем полей индексов и с указанием правил по изменению атрибутов новых, существующих и отсутствующих записей в полученных данных;

в) алгоритмы загрузки могут быть расширены за счёт хранимых процедур для обработки данных, параметризованных хранимых процедур, а также хранимых процедур и функций из внешних баз данных в качестве источников данных.

ПО «OTRi.DI» позволяет работать со следующими форматами данных:

– CSV с возможностями настройки параметров файла, таких как кодировка, символ-разделитель, символ цитирования, столбец с номером строки, строка-заголовок с названиями столбцов, список столбцов с типами данных, форматами и др.;

– JSON и XML с настраиваемым списком столбцов результата (с указанием типов данных, форматов и др.) и маппингом (сопоставлением) по путям в структуре, в т.ч. с возможностью чтения вложенных коллекций с поддержкой нескольких уровней вложенности;

– XLS/XLSX с поддержкой нескольких листов, указанием начальной строки, столбца, строки-заголовка, списка столбцов с типами данных, форматами и др.;

– AVRO с настраиваемым списком столбцов, путей, типов данных, форматов и др.;

– YAML с настраиваемым списком столбцов, типов данных, форматов и др.;

– ZIP-архивы с одним и более файлами поддерживаемых форматов, в т.ч. защищенные паролем, поддержка чтения сжатых GZIP-файлов.

Преобразование данных в «OTRi.DI» включает в себя:

– фильтрацию, а также первичное преобразование данных при получении из источника;

– выполнение базовой трансформации данных перед вставкой в базу: очистку значений по регулярным выражениям, конвертацию значений разных числовых форматов и схожих;

– генерация таблицы замыканий с использованием отношений родитель-потомок;

– объединение нескольких полей в одно новое поле;

– выполнение строковых операций (удаление пробелов, добавление символов, замена одной подстроки на другую, изменение регистра, обрезка и др.);

– преобразование строк в столбцы;

– разделение поля на несколько полей;

– сортировка строк;

– удаление повторяющихся строк;

– XSL преобразование;

– изменение кодировки и др.

Алгоритмы загрузки могут быть расширены программными обработками, реализованными на одном из языков программирования высокого уровня. В алгоритмах загрузки могут использоваться команды операционной системы с возможностью передачи атрибутов данных как параметров.

2.3 Оркестрация выполнения загрузки и рекалькуляция

Для оркестрации выполнения загрузки пользователь может:

– настраивать обработку ошибок;

– выполнять трансформации и процессы по расписанию и вызову API, а также вручную, в т.ч. запускать выполнение отдельных пользовательских задач и сценариев;

– настраивать параллельную обработку данных с ограничением по количеству параллельно выполняемых задач;

– настраивать зависимости между разными источниками данных для составления очередности выполнения задач в очереди.

Статус задач в очереди может отображаться в виде таблицы или графика. Состояние задачи в очереди отображается в режиме реального времени.

Рекалькуляция (пересчёт) совершается в следующих режимах: полный пересчет, пересчет по перечню и диапазону значений с возможностью комбинировать по нескольким условиям. Пересчёт может запускаться вручную по указанным параметрам.

2.4 Дистрибуция данных

Алгоритмы дистрибуции могут быть описаны в текстовом формате, поддерживающем чтение и хранение в системе контроля версий. В алгоритмах дистрибуции данных могут использоваться логические названия ресурсов без указания параметров подключения.

ПО «OTRi.DI» предлагает следующие способы дистрибуции данных:

– файлы и файловые системы (данные могут записываться на следующие файловые ресурсы: локальные и сетевые диски, FTP/FTPS, SFTP);

– веб-ресурсы (веб-ресурсы, используемые для записи данных, должны поддерживать передачу данных с помощью веб-запросов. В них должна быть предусмотрена поддержка basic-аутентификации с использованием защищенных протоколов передачи данных, управление набором HTTP-заголовков, возможность использования веб-запросов аутентификации с сохранением cookie);

– электронная почта:

OTRi.DI позволяет отправлять вложения по электронной почте. Для формирования тела письма используются шаблоны (на основе набора табличных данных).

– базы данных:

Данные можно вставлять в таблицы внешних баз данных и автоматически создавать временные таблицы с конфигурируемым перечнем индексов. Алгоритмы дистрибуции могут

расширяться параметризованными хранимыми процедурами для обработки элементов данных.

Данные могут быть записаны в следующих форматах:

– CSV с возможностями настройки параметров файла, таких как кодировка, символ-разделитель, символ цитирования, столбец с номером строки, строка-заголовок с названиями столбцов, список столбцов с типами данных, форматами и др.;

– JSON и XML с настраиваемым списком полей результата (с указанием типов данных, форматов и др.) и маппингом (сопоставлением) по путям в структуре, в т.ч. с возможностью чтения вложенных коллекций с поддержкой нескольких уровней вложенности;

– XLS/XSLX с настройками, защитой листа, использованием файла-шаблона при создании файла, указанием начальной строки, строки-заголовка, списка столбцов с типами данных, форматами и др.;

– AVRO с настраиваемым списком полей, путей, типов данных, форматов, файлом-схемой и др.;

– выгрузки в ZIP-архивах, в т.ч. защищенных паролем.

3 Список изменений

Версия	Дата	Внесенные изменения	Исполнитель