ТЕРН ETL

Версия 1.0

Обзор функциональных возможностей

Листов 14

2023 год

АННОТАЦИЯ

В настоящем документе представлены общие сведения о программном продукте «ТЕРН ETL» (далее – Система), описаны функциональные возможности, приводится информация по их использованию.

Оглавление

ГЛОССАРИЙ

№	Термин	Пояснения
1	Система	Решение ТЕРН ETL
2	ETL	Extract, Transform, Load, трёхэтапный процесс управления данными, в дословном переводе значит «извлечение, преобразование, загрузка»
3	Профилирование	Процесс изучения данных, доступных в источнике информации
4	Тег, теги	Сигнатура юридически значимого документа, используемая экспертной системой для профилирования источников данных

ВАЖНО: Так будут выделены важные замечания.

1. Общие сведения

Терн ETL решение в области извлечения, преобразования и загрузки данных, разработанное кампанией ТЕРН (tern.ru).

Использование современных методов и подходов к обработке больших объемов данных, в том числе с использованием искусственного интеллекта, позволило создать решение, обладающее уникальными преимуществами перед уже существующими на рынке аналогами.

2. Назначение системы

2.1. Полное наименование продукта и его условное обозначение

Полное наименование системы: программный продукт «ТЕРН ETL».

Условное обозначение продукта: Система.

2.2. Назначение и область применения

«ТЕРН ETL» - решение в области извлечения, преобразования и загрузки данных и обеспечивает качественное построение и сопровождение хранилищ данных.

Система «ТЕРН ETL» предназначена для подразделений, отвечающих за формирование, ведение и использование корпоративных хранилищ данных, отвечающих за миграцию данных из одного источника в другой, обеспечивающих качество данных.

Система применяется для широкого круга задач по обработке, трансформации и миграции данных.

2.3. Задачи, решаемые системой

Задачами, решаемыми Системой, являются:

- Быстрое реагирование на события и на изменения условий и требований бизнеса.
- Поддержка непрерывного анализа неограниченного объема данных на сверхвысоких скоростях.

- Быстрая адаптация к изменениям форм и типов данных.
- Поддержка любых источников данных.
- Обеспечение высокой доступности, управление неоднородными данными и реализация новой потоковой парадигмы.
- Автоматическое профилирование источников любой сложности и из любой предметной области.
- Обеспечение защиты и конфиденциальности информации, предоставляемой для общего доступа.

К возможностям системы относятся:

- Корреляция и интеграция данных из любых источников, включая базы данных, веб-сервисы, файлов практически любых форматов (включая графические) и других приложений. Этот инструмент поможет объединить данные из различных источников, обработать и загрузить их в единое хранилище данных.
- Обеспечение точности и целостности данных с использованием таких технологий как валидация данных, автоматическая очистка данных, контроль целостности данных.
- Обработка больших объемов данных с использованием механизма распределенных вычислений, который позволяет быстро и эффективно обрабатывать большие объемы данных и уменьшает время их обработки.
- Мониторинг состояния источников данных в режиме реального времени, который облегчает поиск проблем в данных, автоматически проверяет обновления и обрабатывает ошибки, значительно сокращая время на обработку данных.

3. Функциональные возможности системы

Работа в системе «ТЕРН ETL» осуществляется через любой современный интернет-браузер (предпочтительным является Яндекс.Браузер).

3.1. Авторизация в системе

Авторизация в Системе происходит путем ввода логина и пароля, после чего пользователю становится доступен функционал Системы.

3.2. Соединения (коннекторы)

Работа с источниками данных начинается с создания для них соединений (коннекторов) – именованных множеств параметров, предназначенных для определения методов доступа к источникам данных.

С соединениями доступны следующие операции:

- Создание соединения (определение названия соединения, типа соединения, источника, типа соединения).
- Редактирование соединения (доступно редактирование всех ранее введенных полей).
- Удаление соединения (для удаления соединения необходимо дополнительно подтвердить данное действие).

ВАЖНО: Удаление соединения, к которому привязаны проекты и / или отображения невозможно.

Поддерживаемые типы источников для создания соединений:

- Реляционные базы данных, доступ к которым возможен через ODBC.
- Веб-сервисы, взаимодействующие через json или xml.
- Структурированные файлы – csv, xls(x) и т.д.
- Графические файлы – jpeg, gif и т.д., а также pdf.

ВАЖНО: В связи со спецификой подключения к различным СУБД требуются соответствующие драйвера, в данной версии системы поставляются драйвера только источников на базе PostgreSQL, Oracle, MS SQL.

ВАЖНО: При работе с источниками, представленными графическими файлами, а так же pdf, используется собственная разработка на базе искусственного интеллекта, которая распознает и выявляет табличные формы, содержащиеся в источниках и помещает распознанную информацию во внутреннюю базу данных. В текущей версии системы распознавание возможно для русского и английского языков с вероятностью порядка 97%.

3.3.Проекты.

Проекты являются основной сущностью Системы, обеспечивающие выполнение свойственных Системе задач. Проекты содержат набор отображений для ETL-операций. Проекты хранятся во внутреннем репозитории Системы, именуемым Библиотекой.

Над проектами доступны следующие операции:

- Создание проекта (определение названия проекта).
- Просмотр проекта (просмотр доступен как для всего списка существующих в Системе проектов, так и для отображений ETL-операций в рамках какого-то одного проекта, видимость доступных для просмотра объектов определяется исходя из настроек прав доступа).
- Редактирование / модификация проекта (Редактирование / модификация отображений ETL-операций в рамках проекта доступно исходя из настроек прав доступа).
- Удаление проекта (для удаления проекта необходимо дополнительное подтверждение).

Важно: Соединения настраиваются из проектов. К каждому соединению может быть подключено несколько проектов.

3.4.Аналитика.

В Системе представлена экспертная система, построенная на сплаве искусственного интеллекта и сигнатурного анализа, решающая задачи профилирования (определения типов данных, хранящихся в таблицах-справочниках) источников.

Экспертная система функционирует на основании загружаемой в нее отраслевой модели «золотой записи» и использует в своей работе теги, которые являются отображениями юридически значимых документов.

Результаты профилирования источника с помощью экспертной системы могут быть просто соотнесены с целевой архитектурой хранилища данных, и на их основе могут быть построены необходимые отображения ETL-операций. Результаты каждого запуска аналитики для источника сохраняются во внутреннем репозитории Системы и могут в дальнейшем, использоваться в сравнительном анализе.

ВАЖНО: Обучение экспертной системы проходит во внутреннем контуре компании ТЕРН и заказчикам поставляются уже обученные копии экспертной системы.

ВАЖНО: Экспертная система не занимается сбором, обработкой и хранением персональных данных.

ВАЖНО: В представленной версии Системы, экспертная система может распознавать следующий набор тегов:

- BCARD' – номер банковской карты.
- 'INNFL' – ИНН физлица.
- 'INNUL' – ИНН юрлица.
- 'OGRN' – ОГРН.
- 'OGRNIP' – ОГРН индивидуального предпринимателя.
- 'SNILS' – СНИЛС.
- 'ORG' – организация.

- 'PER' – персона.
- 'LOC' – локация, адрес.

3.5.Получатели.

При создании получателя (СУБД или файла, являющегося финальной стадией ETL процесса) необходимо указать наименование подключения, выбрать тип подключения и необходимые отображения ETL-операций.

С получателями доступны следующие операции:

- Создание получателя (определение наименования, типа подключения).
- Просмотр получателя (просмотр доступен как для всего списка существующих в Системе получателей, так и для отображений ETL-операций для какого-то одного получателя, видимость доступных для просмотра объектов определяется исходя из настроек прав доступа).
- Редактирование / модификация получателя (Редактирование / модификация параметров подключения и отображений ETL-операций для конкретного получателя доступно исходя из настроек прав доступа).
- Удаление получателя (для удаления получателя необходимо дополнительное подтверждение).

ВАЖНО: Система предоставляет возможность создания и настройки новых таблиц в получателе посредством встроенных возможностей интерфейса.

3.6.Отображения.

Отображения это одна из основных сущностей Системы, описывающая непосредственно инструкции и действия, которые необходимо выполнить с данными в части их обработки и преобразования.

С отображениями возможны следующие действия:

- Создание отображения. Создание нового отображения.
- Просмотр и выбор отображения. Просмотр и выбор доступных отображений в соответствии с настроенными правами доступа.
- Редактирование отображения. Внесение изменений в сущностные атрибуты отображения.
- Удаление отображения. Действие требует дополнительного подтверждения.
- Создание таблицы квалификатора в отображении. Таблица квалификатора – копия таблицы источника с внутренними типами данных для полей, которые будут участвовать в процессе трансформации.
- Добавление таблицы квалификатора. При добавлении таблиц источника добавляется таблица квалификатора.

Важно: В одном отображении могут использоваться несколько таблиц источника и таблицы из разных источников.

- Настройка полей таблицы квалификатора. Можно указать какие поля источника требуются в процессе ETL, какие будут загружены в систему и использованы.
- Отображение связей в таблице квалификатора. Можно переключаться между разными способами отображения и присваивания связей – по полям и по таблицам в целом.
- Удаление таблицы квалификатора.
- Добавление/удаление таблицы получателя в / из отображения. Добавление таблицы получателя является обязательным действием при создании отображения.
- Настройка связей таблиц в отображении. Соединение связями таблиц источника и получателя. Связи можно создавать как между таблицами в целом, так и между конкретными полями таблиц.

- Удаление связей между таблицами. Лишние/ошибочные связи могут быть удалены. Удаление возможно как единичной связи, так и выбранной группы связей. При удалении таблицы, все входящие и исходящие связи также удаляются.

В отображения также можно добавить следующие операции трансформации данных:

- Фильтр. Оператор фильтрации служит для выбора из множества исходных данных только тех строк, которые удовлетворяют заданным пользователем фильтрам, для дальнейшей обработки или для записи в таблицу.
- Объединение строк. Оператор объединения строк служит для объединения данных (строк) из нескольких таблиц в одну таблицу.
- Объединение полей. Оператор объединения полей (joiner) служит для объединения данных из нескольких таблиц в общие строки итоговой таблицы. Например, для добавления в итоговую таблицу дополнительных полей-деталей из дополнительной таблицы по внешнему ключу.
- Агрегация. Оператор агрегации служит для получения таблицы с данными, сгруппированными по указанным полям с возможностью агрегации значений других полей в рамках сформированных групп.
- Маршрутизация (Роутинг). Оператор роутинга служит для разделения данных таблицы источника на несколько таблиц-получателей по заданным условиям фильтрации.

3.7.Рабочие процессы.

Рабочие процессы это одна из основных сущностей Системы, обеспечивающая выполнение инструкций и действий над данными в части их обработки и преобразования.

С рабочими процессами возможны следующие действия:

- Создание рабочего процесса. Создание рабочего процесса доступно только для полностью сформированных отображений (указаны таблицы источник и приемник, сформированы связи между ними, добавлены необходимые операции над данными).
- Просмотр рабочего процесса. Доступен просмотр как перечня существующих рабочих процессов, так и сущностных атрибутов выбранного процесса. Видимость определяется настройкой прав доступа к объектам.
- Удаление рабочего процесса. Удаление рабочего процесса не удаляет отображения задействованные в нем.
- Загрузка и запуск рабочего процесса. Для исполнения, рабочий процесс должен быть загружен в операционную среду и ему должна быть дана команда на активацию.
- Настройка свойств запуска и активации рабочего процесса.

ВАЖНО: По умолчанию, все рабочие процессы имеют опцию активации по требованию. Для изменения этой опции для активации по расписанию или непрерывно, необходимо внести соответствующие изменения в настройки процесса. Также в Системе доступны расширенные настройки запуска рабочего процесса по расписанию.

3.8.Мониторинг.

Каждый рабочий процесс представляет собой последовательность выполняемых операций – таких как, чтение из базы данных, запись в базу данных, а также трансформации, производимые с данными.

Система предоставляет возможности мониторинга рабочих процессов, в рамках которого рабочий процесс можно развернуть и посмотреть список операций, из которых он состоит, а также их статусы.

Мониторинг позволяет получить информацию по следующим параметрам рабочего процесса:

- Запуск рабочего процесса. Наименование запущенного рабочего процесса.
- Время начала. Формат: время начала выполнения рабочего процесса в формате год-месяц-день часы:минуты:доли секунд.
- Время завершения. Формат: время завершения выполнения рабочего процесса в формате год-месяц-день часы:минуты:доли секунд.
- Статус. Статус рабочего процесса может быть в диапазоне – Wait (ожидание запуска), Running (исполняется), Success (завершено без ошибок), Fail (завершено с ошибками).
- Подробности рабочего процесса. Для процесса в целом и для каждой операции в отдельности отображается информация о статусе, времени и длительности выполнения процесса или операции, а в случае операции с таблицами - о количестве успешных строк.
- Подробности прогресса задач. Для процесса в целом и для каждой операции в отдельности отображается визуальная временная диаграмма исполнения процессов.
- Просмотр логов. Для каждой отдельной операции можно просмотреть логи выполнения операции.

ВАЖНО: Информация в окне мониторинга обновляется раз в 10 секунд. Параметр не может быть изменен в текущей версии Системы.