

ПЛАТФОРМА ДАННЫХ СЕЛЕНА
РУКОВОДСТВО ПОЛЬЗОВАТЕЛЯ
РФ.DIS.00011-01 41

АННОТАЦИЯ

Документ РФ.DIS.00011-01 41 «Руководство пользователя» подготовлен на основе рекомендаций государственного стандарта ГОСТ Р 59795—2021 Информационные технологии. Комплекс стандартов на автоматизированные системы. Автоматизированные системы. Требования к содержанию документов.

Электронная версия документа хранится в составе пакета программной и эксплуатационной документации на изделие РФ.DIS.00011 «Платформа данных СЕЛЕНА» (далее по тексту Платформа данных СЕЛЕНА).

Ознакомление с документом «Руководство пользователя» персонала подразделения, принимающего участие в работе автоматизированной системы, производится под роспись с внесением соответствующей записи в журнал первичного инструктажа.

СОДЕРЖАНИЕ

Аннотация	2
Обозначения и сокращения	5
Предисловие	6
1 Обзор Платформы Данных	7
1.1 Обзор продукта	7
1.2 Архитектура и компоненты репликации данных	9
2 Модуль управления кластером Cluster Manager	13
2.1 Вход в систему	13
2.2 Управление виртуальными машинами	13
2.3 Добавления компонентов кластера	13
2.4 Удаление компонентов кластера	14
2.5 Мониторинг нагрузки кластера	14
2.5 Управление доступом к системе	15
2.6 Управление доступом к данным в системе	16
2.7 Управление конфигурацией кластера	17
2.8 Управление лицензией кластера	17
3 Управление расписание и задачами	18
3.1 Вход в систему	18
3.2 Просмотр общей статистики по выполняемым процессам	18
3.3 Создание подключений	18
3.4 Создание проектов	19
3.5 Создание и настройка процессов в проектах	19
3.6 Запуск и настройка расписания в процессах	20
3.7 Мониторинг и отладка выполняемых процессов	21
3.8 Мониторинг состояния сервиса	22
3.9 Создание и изменение пользователей	22
3.10 Создание и изменение групп пользователей	22
3.11 Настройка оповещений	23
4 Работа с модулем генеративного ИИ (AI-copilot)	24
4.1 Вход в систему	24
4.2 Выполнение запроса Text -> SQL	24
4.3 Выполнение запроса SQL -> Text	25
5 Подключение к системе из внешних инструментов	27

5.1	Подключение к платформе данных из Apache Superset	27
5.2	Подключение к платформе данных из Tableau Desktop	28
5.3	Подключение к платформе данных из DataGrip	29
5.4	Подключение к платформе данных из DBeaver.....	29
5.5	Подключение к платформе данных из Jupyter.....	30

ОБОЗНАЧЕНИЯ И СОКРАЩЕНИЯ

В документе РФ.DIS.00011.0141 «Руководство пользователя» используются следующие обозначения и сокращения, имеющие соответствующие значения:

ОБОЗНАЧЕНИЕ ИЛИ СОКРАЩЕНИЕ	ЗНАЧЕНИЕ
ВМ	Виртуальная машина
ГОСТ	Государственный стандарт
РФ	Российская Федерация
СУБД	Система управления базами данных
ПКМ	Правая кнопка мыши
УЗ	Учетная запись
AI	Искусственный интеллект
BI	Business intelligence. Система сбора, обработки и анализа данных
CPU	Центральный процессор
DDL	Data Definition Language (DDL) (язык описания данных)
DS	Data Science
ETL	Extract, Transform, Load
HDFS	Hadoop Distributed File System
JDBC	Java DataBase Connectivity
ML	Machine learning
ODBC	Open Database Connectivity
OLAP	Online analytical processing
MPP	Massively parallel processing
SIMD	Single instruction, multiple data
S3	Simple Storage Service
SQL	Structured Query Language — «язык структурированных запросов»
URI	Uniform Resource Identifier
URL	Uniform Resource Locator

Предисловие

Данное руководство пользователя описывает программный комплекс «Платформа хранения и обработки данных Селена» (в дальнейшем Селена), включая его архитектуру, варианты использования, принципы развертывания, источники и целевые объекты, а также основные концепции.

В руководстве также описывается, как использовать интерфейс панели управления для настройки, запуска, мониторинга и администрирования.

1 ОБЗОР ПЛАТФОРМЫ ДАННЫХ

1.1 Обзор продукта

Селена – сверхбыстрая платформа хранения и обработки данных нового поколения с массивно-параллельной обработкой (MPP), разработанная для упрощения и ускорения доступа к данным, быстрой аналитики в реальном времени. MPP и векторизованный механизм выполнения вычислений позволяют пользователям выбирать между различными схемами для разработки многомерных аналитических отчетов.

Платформа Селена предназначена для применения для следующих случаев:

- многомерная аналитика OLAP;
- построение отчетности любого уровня сложности;
- аналитика данных в реальном времени;
- high-concurrency аналитика;
- унифицированная аналитика.

Селена поддерживает самые разнообразные функции, обеспечивая надежную и быструю работу вашего корпоративного хранилища данных, что указано в списке ниже.

- а) MPP-фреймворк: платформа Селена использует фреймворк массивно-параллельной обработки. Один запрос разделяется на несколько физических вычислительных блоков, которые могут выполняться параллельно на нескольких машинах. Каждая машина имеет выделенные ресурсы CPU и памяти.
- б) Оптимизатор: он находит наиболее оптимальный план на основе собранной статистики ваших данных. Это ключ к лучшей в своем классе производительности запросов, особенно для мультитабличных запросов.
- в) Полностью векторизованный механизм выполнения: благодаря колоночному механизму хранения и полностью векторизованным операторам платформа Селена в полной мере использует современные многоядерные процессоры и инструкции SIMD для повышения производительности.
- г) Гибридное строково-столбцовое хранилище: оно обеспечивает более 10 000 запросов в секунду уже на 16-ядерных экземплярах вычислительных серверов

(нод кластера) за счет оптимизированных точечных запросов и первичного ключевого индексируемого ускорения в гибридном строково-столбцовом хранилище.

- д) Кэш данных: встроенная в платформу Селена структура кэширования на основе памяти и дискового пространства специально разработана для минимизации накладных расходов ввода-вывода при извлечении данных из внешнего хранилища для ускорения выполнения запросов.
- е) Аналитика в реальном времени: от потоковой передачи до сбора данных с богатым набором коннекторов возможно загружать данные в платформу Селена в реальном времени для получения самых свежих инсайтов.
- ж) Таблица первичного ключа обеспечивает непревзойденную производительность запросов с обновлениями, вставками и удалениями в реальном времени. Индекс первичного ключа позволяет эффективно разрешать изменения данных во время приема данных, оптимизируя производительность чтения, поддерживая актуальность данных на уровне менее десяти секунд от времени изменения данных.
- з) Синхронное материализованное представление может постепенно обновляться при приеме данных и выполнять переписывание запросов при их исполнении.
- и) Общая архитектура данных - Селена разделяет слои хранения и вычисления посредством сохранения данных в удаленном объектном хранилище, таком как S3 или HDFS.
- к) Единый каталог метаданных: с помощью одной команды при использовании встроенного каталога метаданных платформа Селена позволяет вам легко подключаться и напрямую запрашивать самые свежие данные, хранящиеся во всех озерах и других источниках данных.
- л) Асинхронное материализованное представление разработано для ускорения медленных запросов без каких-либо дополнительных внешних инструментов обработки. Платформа Селена обеспечивает возможность перезаписи запросов и позволяет создавать асинхронное материализованное представление в любое время без необходимости вручную изменять SQL-запрос.

- м) Унификация доступа к данным: платформа Селена поддерживает синтаксис ANSI SQL, протокол MySQL и обеспечивает поддержку диалекта Trino/Presto. Платформа совместима с широким спектром клиентского ПО и инструментов BI и аналитики.

1.2 Архитектура и компоненты репликации данных

Платформа Данных Селена включает следующие компоненты:

- а) Мастер сервер
- б) Вычислительный сервер
- в) Панель управления кластером (Cluster Manager)
- г) Модуль загрузки данных
- д) Метакаталог данных
- е) Мониторинг
- ж) Модуль среды разработки
- з) Модуль хранения данных

На рис. 1 показана архитектура Платформы данных Селена.

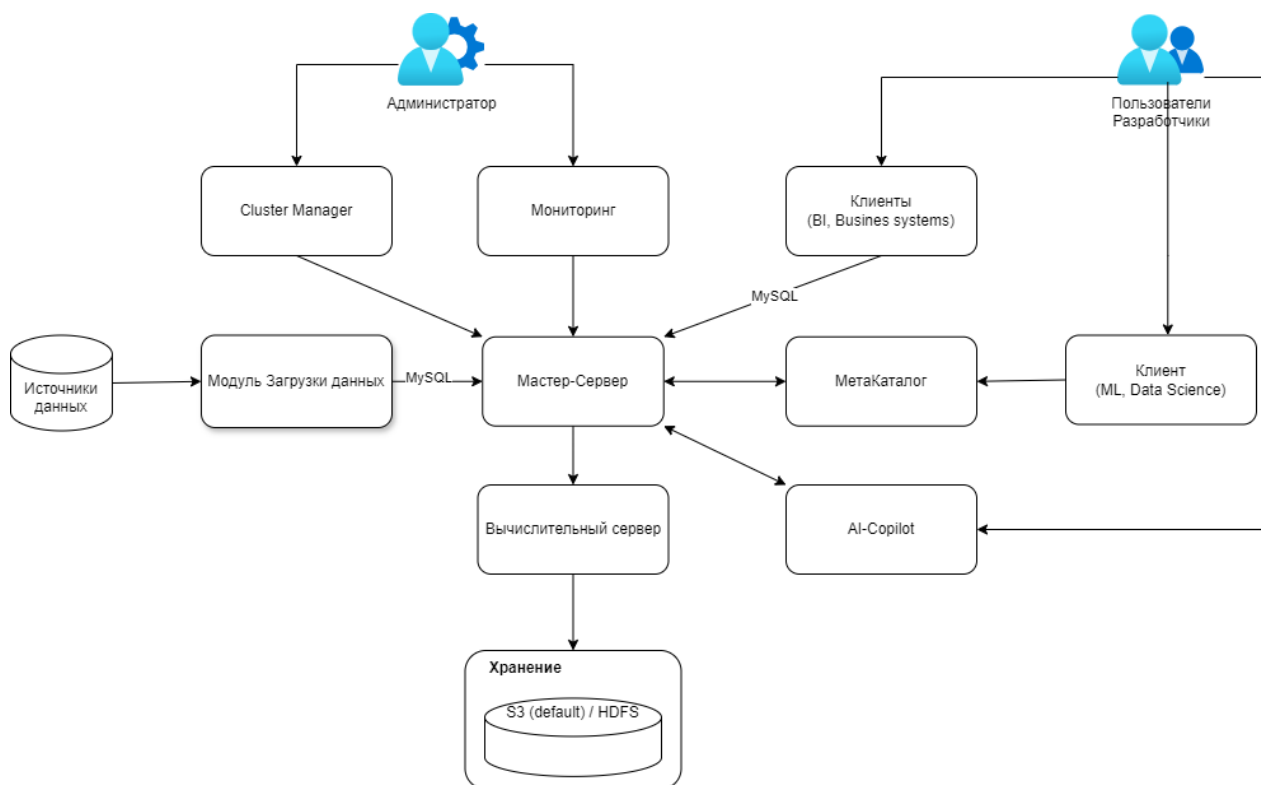


Рисунок 1 - Архитектура и компоненты Платформы Данных Селена

Мастер сервер

Мастер сервер отвечают за управление метаданными, управление клиентскими соединениями, планирование запросов и диспетчеризацию запросов. Каждый Мастер сервер хранит и поддерживает полную копию метаданных в своей памяти, что гарантирует неизбирательное обслуживание среди Мастер серверов. Мастер сервер может выполнять одну из ролей - лидер, последователь и наблюдатель.

Лидер считывает и записывает метаданные. Лидер обновляет метаданные, а затем синхронизирует изменения метаданных с последователями и наблюдателями. Записи данных считаются успешными только после того, как изменения метаданных синхронизированы с более чем половиной последователей.

Последователи синхронизируют и воспроизводят журналы с лидера для обновления метаданных.

Вычислительный сервер

Основной задачей Вычислительных серверов является выполнение SQL запросов. Мастер сервер разбирают каждый SQL-запрос на логический план выполнения в соответствии с семантикой запроса, а затем преобразуют логический план в физические планы выполнения, которые могут быть выполнены на Вычислительных серверах. Вычислительные сервера, хранящие данные назначения, выполняют запрос. Это устраняет необходимость в передаче и копировании данных, достигая высокой производительности запроса.

Панель управления кластером (Cluster Manager)

Cluster Manager - основной инструмент управления кластером, обеспечивает администраторам все необходимые функции по созданию и настройке кластера. Обеспечивает функции добавления и удаления нод, назначение типов и развертывание приложений на новых нодах. Помимо этого, выполняет функции безопасности, обеспечивая контроль доступа как к самим интерфейсам, используемым в платформе компонентов, так и к данным, расположенным в хранилище.

Модуль загрузки данных

Модуль загрузки данных обеспечивает бесперебойную загрузку данных из различных систем источников данных в целевое хранилище. Компонент обеспечивает возможность извлекать данные из самых разнообразных систем включая самые популярные реляционные и No-SQL базы данных, шины данных и т.д. Кроме задач по загрузке данных, компонент реализует функции визуального формирования потоков данных, и настройки расписания запусков, а также выполнять функции отладки и мониторинга выполнения процессов работы с данными.

Метакаталог данных

Метакаталог данных содержит информацию о метаданных, хранящихся в объектном хранилище, и обеспечивает быстрое формирование запросов на данные в объектном хранилище. Каталог метаданных в архитектуре платформы Селена обеспечивает единое представление о хранящихся в объектном хранилище данных.

Мониторинг

Мониторинг в системе обеспечивается выделенным модулем, который собирает всю необходимую информацию о работе кластера, состоянии основных компонентов и вычислительных мощностях. Пользователям предоставляется целый ряд пред настроенных мониторов отображая информацию от загрузки ЦПУ на нодах мастер серверов, заканчивая производительностью и количеством выполняемых запросов на вычислительных узлах кластера.

Модуль среды разработки

Модуль среды разработки позволяет различным бизнес-пользователям быстро и просто получить данные. Для разработчиков и аналитиков обеспечивается возможность создания ad-hoc запросов и формирование блокнотов. А для бизнес-пользователей для работы с данным компонентом нет необходимости уметь писать sql-запросы. Все обращения в систему формируются используя семантику, далее они автоматически преобразуются в SQL код, который в свою очередь и выполняется, возвращая ожидаемый результат пользователю.

Модуль хранения данных

В качестве компонента, отвечающего за непосредственное хранение данных, могут выступать два решения для хранения: объектное хранилище S3 и HDFS. Платформа хранения и обработки данных Селена поставляется вместе с S3 на базе Minio, но при этом может быть использовано любое клиентское S3 Compatible решение, а также HDFS для реализации хранения данных. Платформа обеспечивает возможность записывать и хранить данные в различных форматах, в зависимости от использования внешнего метакаталлога данные будут храниться в открытом формате, либо в нативном.

2 МОДУЛЬ УПРАВЛЕНИЯ КЛАСТЕРОМ CLUSTER MANAGER

2.1 Вход в систему

- а) Открыть страницу мастер сервера
- б) Ввести логин и пароль

2.2 Управление виртуальными машинами

Для добавления виртуальной машины в кластер, необходимо:

- а) Авторизоваться в системе под УЗ администратора
- б) Нажать в левой части экрана «виртуальные машины»
- в) В открывшемся окне нажать «+ новая виртуальная машина»
- г) В открывшейся оснастке указать:
 - название ВМ;
 - публичный IP адрес;
 - внутренний IP адрес;
 - пользователь;
 - тип аутентификации пароль или ssh-ключ;
 - если выбран пароль, указать пароль;
 - если выбран ssh ключ нажать choose file и указать путь к файлу с ssh ключом.
- д) Для проверки соединения нажать “проверить доступ ssh”
- е) Нажать «ОК»


2.3 Добавления компонентов кластера

Для добавления компонентов в кластер, необходимо:

- а) Авторизоваться в системе под УЗ администратора
- б) Нажать в левой части экрана «управление кластером»
- в) В кластере есть 4 основные роли у компонентов, это - Мастер сервер, Вычислительный сервер, Хранилище и Метакаталог. Для добавления роли


виртуальной машине, необходимо нажать кнопку



- г) В открывшемся окне, указать роль и выбрать свободную ВМ, на которой требуется развернуть соответствующую роль. Нажать «Сохранить». После этого будет запущена установка программного обеспечения.
- д) В разделе, для которого добавляется новая ВМ, появится новая строка, с указанием:
 - Состояния
 - IP адреса
 - Версии установленного приложения
 - И прочей дополнительной информацией относительно роли
- е) Для проверки журнала установки необходимо нажать кнопку просмотра файла журналирования  напротив требуемой ноды кластера

2.4 Удаление компонентов кластера

Для добавления компонентов в кластер, необходимо:

- а) Авторизоваться в системе под УЗ администратора
- б) Нажать в левой части экрана «управление кластером»
- в) Напротив ноды которую требуется убрать их кластера нажать 

2.5 Мониторинг нагрузки кластера

Для просмотра состояния и статистики по нагрузке на кластер платформы данных необходимо:

- а) Авторизоваться в системе под УЗ администратора
- б) Нажать в левой части экрана «мониторинг»
- в) В открывшемся окне в левой части экрана нажать «Dashboards»
- г) В открывшемся окне нажать Selen Overview
- д) Будет открыт набор пред настроенных панелей, с большим количеством разнообразных метрик по состоянию системы.

2.5 Управление доступом к системе

Для задания дефолтного пароля для новых пользователей необходимо:

- а) Авторизоваться в системе под УЗ администратора
- б) Нажать в левой части экрана «Пользователи»
- в) В появившемся окне выбрать Создание и удаление пользователей
- г) Выбрать вкладку «организация»
- д) Выбрать в списке необходимую организацию
- е) Напротив поля default password задать пароль , по умолчанию для всех новых УЗ.

Для создания пользователя необходимо:

- ж) Авторизоваться в системе под УЗ администратора
- з) Нажать в левой части экрана «Пользователи»
- и) В появившемся окне выбрать Создание и удаление пользователей
- к) В открывшемся окне нажать кнопку «добавить пользователя»/«adduser»
- л) В появившейся оснастке указать:
 - Имя
 - Отображаемое имя
 - Пароль
 - Роль
 - Задать привелегии

Для редактирования пользователя необходимо:

- а) Авторизоваться в системе под УЗ администратора
- б) Нажать в левой части экрана «Пользователи»
- в) В появившемся окне выбрать Создание и удаление пользователей
- г) Нажать кнопку «редактировать» напротив необходимого пользователя
- д) Изменить требуемые параметры
- е) Нажать «Сохранить»

Для удаления пользователя необходимо:

- а) Авторизоваться в системе под УЗ администратора
- б) Нажать в левой части экрана «Пользователи»
- в) В появившемся окне выбрать Создание и удаление пользователей
- г) Напротив УЗ, которую требуется удалить, нажать «удалить»
- д) Подтвердить свой выбор

2.6 Управление доступом к данным в системе

Для создания роли пользователя необходимо:


- а) Авторизоваться в системе под УЗ администратора
- б) Нажать в левой части экрана «Роли»
- в) В появившемся окне нажать «Новая роль»
- г) Задать имя роли, нажать «сохранить»

Для изменения у роли пользователей полномочий необходимо:

- а) Авторизоваться в системе под УЗ администратора
- б) Нажать в левой части экрана «Роли»
- в) Напротив требуемой роли нажать кнопку конфигурации  , и затем Просмотр
- г) Для добавления, в появившейся оснастке нажать кнопку добавления полномочий 
- д) Выбрать привелегию, действие, объект , идентификатор. Нажать «Сохранить»
- е) Для добавления, в появившейся оснастке нажать кнопку добавления полномочий 
- ж) Выбрать привелегию, которую требуется удалить. Нажать «Сохранить»

2.7 Управление конфигурацией кластера

Для создания роли пользователя необходимо:

- а) Авторизоваться в системе под УЗ администратора
- б) Нажать в левой части экрана «Конфигурация»
- в) Напротив необходимого параметра, нажать кнопку изменения 
- г) В открывшейся оснастке задать новое значение параметра, нажать «сохранить»

2.8 Управление лицензией кластера

Для создания роли пользователя необходимо:

- а) Авторизоваться в системе под УЗ администратора
- б) Нажать в левой части экрана «Лицензия»
- в) Напротив необходимого параметра, нажать кнопку «Обновить Лицензию»
- г) В открывшейся оснастке указать путь к файлу лицензии, нажать «сохранить»
- д) На обновившемся экране лицензии, убедиться, что конфигурация обновилась корректно.

3 УПРАВЛЕНИЕ РАСПИСАНИЕ И ЗАДАЧАМИ

3.1 Вход в систему

- а) Открыть страницу системы оркестрации и управления потоками данных
- б) Ввести логин и пароль
- в) Нажать «Войти»

3.2 Просмотр общей статистики по выполняемым процессам

- а) Авторизоваться в системе
- б) В левой части экрана нажать «Дашборд»
- в) В появившемся окне будет отображена информация по выполняемым в системе процессам и задачам.
- г) Для того, чтобы отфильтровать процессы и задачи по временному диапазону необходимо задать соответствующие даты начала и конца в соответствующем полях над выводимой информацией
- д) Для того, чтобы отфильтровать процессы с не интересующим вас состоянием необходимо в центральной части нажать по показателю статуса, который следует убрать или добавить

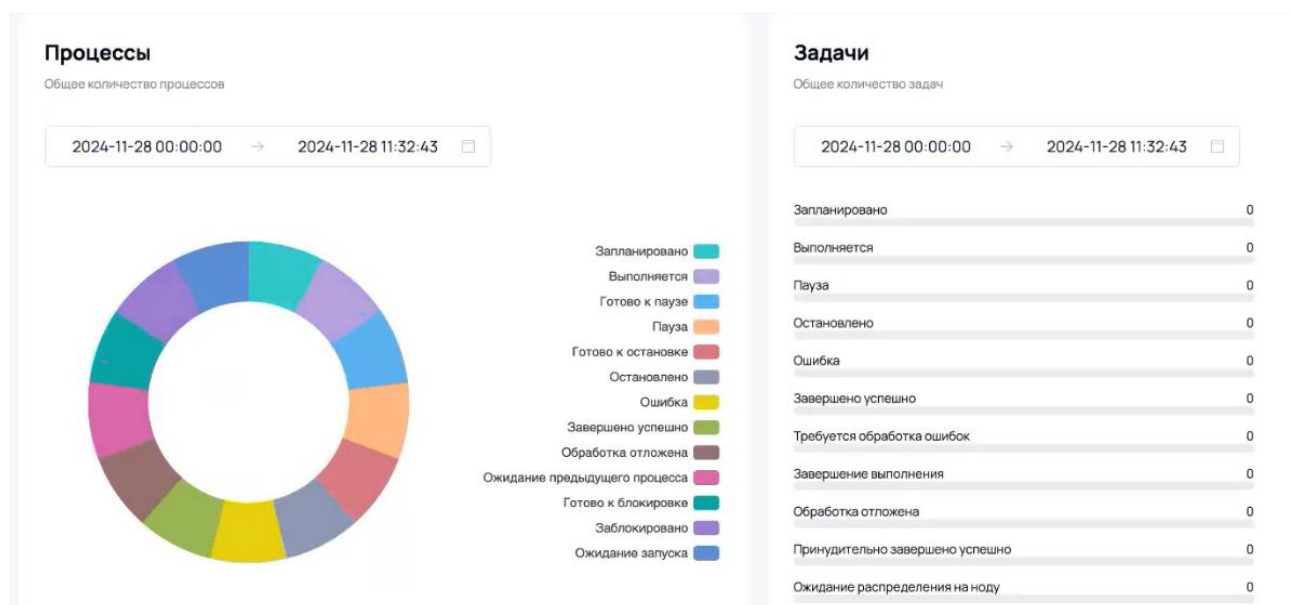


Рисунок 2

3.3 Создание подключений

- а) Авторизоваться в системе

- б) В левой части экрана нажать «Подключения»
- в) В появившемся окне нажать кнопку «новое подключение»
- г) Выбрать из списка подключений необходимую СУБД или источник данных
- д) Указать требуемые настройки в зависимости от системы
- е) Нажать «тест соединения» для проверки корректности введенной информации
- ж) Нажмите «сохранить»
- з) Если все указано корректно, то схема данных взятая из добавленного источника автоматически отобразиться в настройках при создании процессов

3.4 Создание проектов



- а) Авторизоваться в системе
- б) В левой части экрана нажать «Проекты»
- в) В появившемся окне нажать кнопку «новый проект»
- г) Указать название, владельца и описание, нажать «подтвердить»
- д) Открыть проект нажав по нему
- е) В открывшемся окне, на вкладке Информация собрана общая информация о работе данного проекта, включая все запускаемые процессы и задачи в нем. Настройка фильтрация по времени или состоянию задач осуществляется согласно разделу 3.2 пунктам 4 и 5.

3.5 Создание и настройка процессов в проектах

- а) Авторизоваться в системе
- б) В левой части экрана нажать «Проекты»
- в) Открыть необходимый проект нажав по нему
- г) Нажать вкладку «процессы»
- д) Нажать «новый процесс»
- е) В открывшемся окне используя функции drag and drop. Перетащите из левой части экрана задачу, на область для проектирования.
- ж) При помещении задачи на область для проектирования автоматически откроется всплывающее окно с настройками. В нем необходимо задать:
 - Название задачи
 - Выполняемые код, запрос или действие в зависимости от выбранной задачи.

- При необходимости развернуть раздел «дополнительные настройки» и указать необходимые дополнительные параметры
- з) Для соединения задач в еденный поток, нажмите ЛКМ на «кружке» одной из них и перетащите появившуюся стрелку к другой задаче
- и) Для редактирования ранее добавленной задачи, нажмите на задачу правой кнопкой мыши, в контекстном меню нажмите «редактировать», измените настройки, нажмите «сохранить».
- к) Для копирования ранее добавленной задачи, нажмите на задачу правой кнопкой мыши, в контекстном меню нажмите «копировать», дубликат задачи автоматически создастся рядом с копируемой задачей.
- л) Для удаления ранее добавленной задачи, нажмите на задачу правой кнопкой мыши, в контекстном меню нажмите «удалить»
- м) По окончании создания потока действий нажмите «сохранить» в правом нижнем углу области проектирования.

3.6 Запуск и настройка расписания в процессах

- а) Авторизоваться в системе
- б) В левой части экрана нажать «Проекты»
- в) Открыть необходимый проект нажав по нему
- г) Нажать вкладку «процессы»
- д) Для единоразового запуска процесса, напротив требуемого процесса нажать кнопку редактирования , в контекстном меню нажать «запустить»
- е) Для настройки запуска задачи по расписанию, напротив требуемого процесса нажать кнопку редактирования , в контекстном меню нажать «перевести в онлайн», подтвердить свой выбор нажав «подтвердить».
- ж) В появившемся окне задать настройки выполнения по расписанию:
 - Время начала и окончания
 - расписание выполнения
 - действие при ошибке
 - стратегии уведомления
 - приоритет процесса

- группу выполняемых серверов
 - группу пользователей
- з) Нажать «Подтвердить»



3.7 Мониторинг и отладка выполняемых процессов

- а) Авторизоваться в системе
- б) В левой части экрана нажать «Проекты»
- в) Открыть необходимый проект нажав по нему
- г) Нажать вкладку «инстансы процессов»
- д) В открывшемся окне будет отображена таблица со всеми запусками процессов в данном проекте.
- е) По каждому запуску фиксируется следующая информация:
 - Название процесса
 - Тип запуска
 - Статус
 - Расписание
 - Время начала и окончания
 - Продолжительность
 - Повторы
- ж) При нажатии на интересующий нас процесс откроется область проектирования. Нажав на требуемую задачу, откроется контекстное меню, в котором можно выбрать одно из следующих действий для отладки процесса
 - Редактирование
 - Копирование
 - Удаление
 - Просмотр лога
 - Очистка кэша
 - Запуск
 - Запуск в обратном порядке
 - Запуск в прямом порядке

3.8 Мониторинг состояния сервиса


- а) Авторизоваться в системе
- б) В левой части экрана нажать «мониторинг»
- в) В открывшемся окне нажать на название Ноды сервиса
- г) В открывшемся окне будет указана сводная информация о работе сервера и создаваемой нагрузке

3.9 Создание и изменение пользователей

- а) Авторизоваться в системе
- б) В левой части экрана раскрыть «управление» нажать «Безопасность»
- в) В открывшемся окне в вкладке пользователи нажать на кнопку «создать пользователя»
- г) В открывшейся оснастке задать:
 - Имя
 - Пароль
 - Группа
 - Email
 - Телефон
 - Статус
- д) Нажать «подтвердить»
- е) Для изменения напротив требуемой УЗ нажать кнопку конфигурирования  , в контекстном меню выбрать «редактирование», после внесения изменений нажать подтвердить.
- ж) Для удаления требуемой УЗ нажать кнопку конфигурирования  , в контекстном меню выбрать «удалить», подтвердить свой выбор.

3.10 Создание и изменение групп пользователей

- а) Авторизоваться в системе
- б) В левой части экрана раскрыть «управление» нажать «Безопасность»
- в) В открывшемся окне нажать вкладку группы пользователей

- г) Нажать на кнопку «создать группу»
- д) В появившейся оснастке задать имя и описание, нажать подтвердить
- е) Напротив созданной группы нажать кнопку конфигурирования , в контекстном меню выбрать «редактирование», указать пользователей, входящих в группу, нажать «подтвердить»

3.11 Настройка оповещений

- а) Авторизоваться в системе
- б) В левой части экрана раскрыть «управление» нажать «Оповещение»
- в) В открывшемся окне нажать «создать оповещение»
- г) В открывшейся оснастке задать:
 - название оповещение
 - Глобальное оповещение
 - Тип предупреждения
 - Используемый плагин
- д) Нажать сохранить

4 РАБОТА С МОДУЛЕМ СРЕДЫ РАЗРАБОТКИ

4.1 Вход в систему

Для перехода в среду разработки необходимо:

- а) Авторизоваться в системе под УЗ Пользователя
- б) Нажать в левой части экрана «Данные»

4.2 Выполнение запросов Ad-Hoc

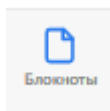
- а) Нажать в верхней части экрана Консоль
- б) В средней части экрана указать необходимые SQL-запросы.

- в) Для выполнения нажать ctrl+enter или кнопку



4.3 Создание блокнотов

- а) Нажать справа кнопку блокноты



- б) В появившейся оснастке нажать
- в) В появившейся оснастке выбрать «Создать блокнот»
- г) В блокноте можно задать:



- Имя блокнота введя его в соответствующее поле
- Добавить дополнительное текстовое поле, нажав кнопку
- Добавить SQL-запрос, нажав кнопку
- Добавить диаграмму связанную с запросом, нажав кнопку

+ Текст

+ Запрос

+ Диаграмма

- д) Для выполнения одного запроса нажмите кнопку



над ним

- е) Для выполнения всех запросов, нажмите кнопку
страницы






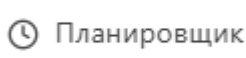
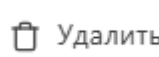
▶ Запустить все

вверху

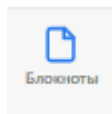

- ж) Для настройки диаграммы нажмите над ней кнопку 

4.3 Дополнительные действия с блокнотами

Для выполнения дополнительных действий с блокнотами необходимо:

- а) Нажать кнопку  в верхней части экрана
- б) Для клонирования в выпадающем меню нажать 
- в) Для экспорта в PDF в выпадающем меню нажать 
- г) Для просмотра списков в выпадающем меню нажать 
- д) Для работы с шаблонными переменными в выпадающем меню нажать 
- е) Для создания расписания в выпадающем меню нажать 
- ж) Для удаления в выпадающем меню нажать 

4.4 Создание блокнотов с помощью ИИ

- а) Нажать справа кнопку блокноты 
- б) В появившейся панели нажать 
- в) В появившейся панели выбрать «Создать блокнот с помощью AI»
- г) В открывшейся панели указать:
- Каталог, с которым предстоит работать
 - БД, по которой будет происходить формирование блокнота
 - Описание блокнота

д) Нажать кнопку

Создать

5 ПОДКЛЮЧЕНИЕ К СИСТЕМЕ ИЗ ВНЕШНИХ ИНСТРУМЕНТОВ

5.1 Подключение к платформе данных из Apache Superset

- а) на сервере Apache Superset установите Python client StarRocks, выполнив:

pip install starrocks

- б) В Apache Superset перейдите в раздел подключения БД в качестве источника данных
- в) В разделе поддерживаемых Баз Данных (supported databases) из выпадающего списка выберите StarRocks

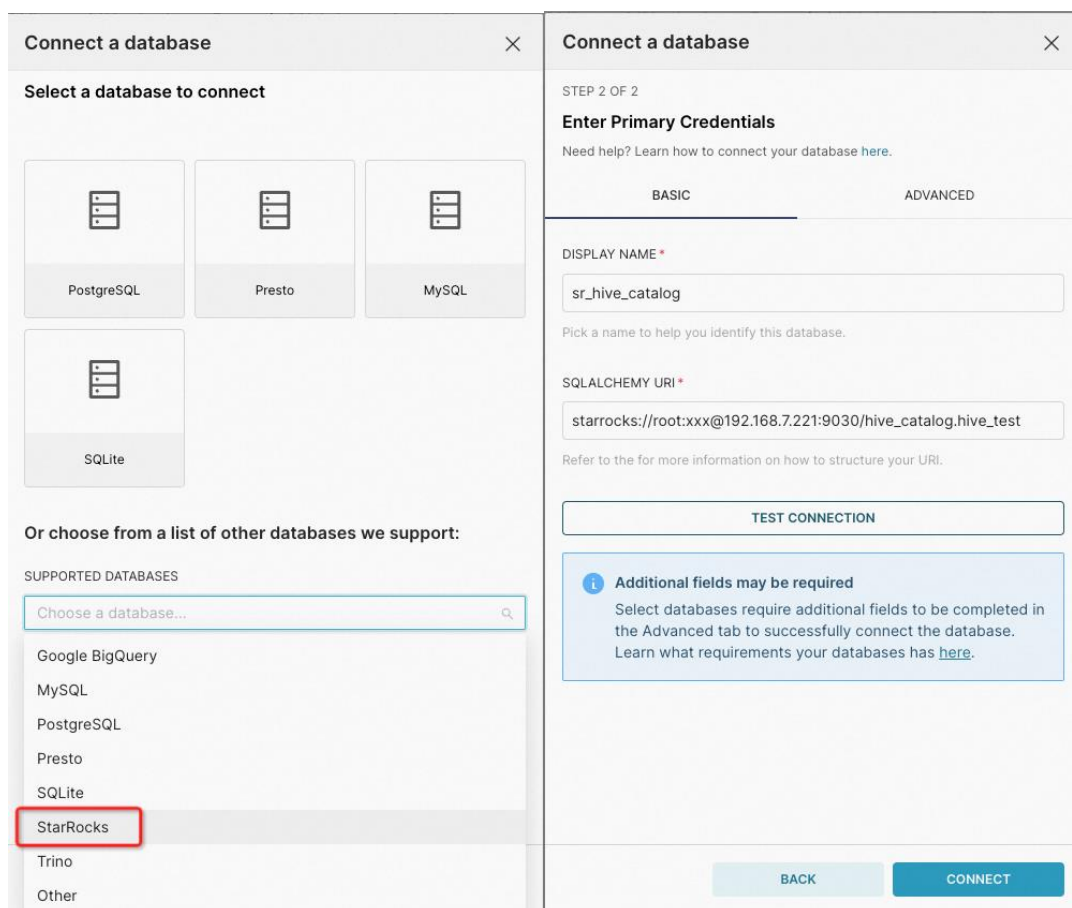


Рисунок 3

- г) На Этапе настройки подключения укажите:
- Отображаемое имя в поле Display Name

- Путь подключения в поле SQLAlchemy URI:

starrocks://<User>:<Password>@<Host>:<Port>/<Catalog>.<Database>

- User – имя пользователя технической УЗ
- Password – пароль от УЗ
- Host – hostname или ip адрес мастер сервера
- Port – порт мастер сервера
- Catalog – имя целевого каталога, поддерживаются внутренние и внешние каталоги
- Database – имя целевой База Данных, поддерживаются внутренние и внешние базы данных

5.2 Подключение к платформе данных из Tableau Desktop

Tableau Desktop поддерживает запросы и визуализацию как внутренних, так и внешних данных в Селене.

- Создайте базу данных в Tableau
- Выберите Other Databases (JDBC) в качестве источника данных.
- Для Dialect выберите MySQL.
- Для URL введите URL в формате MySQL URI, как показано ниже:

jdbc:mysql://<Host>:<Port>/<Catalog>.<Databases>

Параметры в URL описаны следующим образом:

- Host: hostname или IP-адрес хоста мастер сервера вашего кластера.
 - Port: порт хоста мастер сервера , например, 9030.
 - Catalog: целевой каталог в вашем кластере. Поддерживаются как внутренние, так и внешние каталоги.
 - Database: целевая база данных в вашем кластере. Поддерживаются как внутренние, так и внешние базы данных.
- Настройте имя пользователя и пароль.

5.3 Подключение к платформе данных из DataGrip

DataGrip поддерживает запросы как внутренних, так и внешних данных.

Создайте источник данных в DataGrip. Обратите внимание, что в качестве источника данных необходимо выбрать MySQL.

Параметры, которые вам необходимо настроить, описаны ниже:

- Хост: hostname или IP-адрес хоста мастер сервера вашего кластера.
- Порт: порт хоста мастер сервера, например, 9030.
- Аутентификация: метод аутентификации, который вы хотите использовать. Выберите Имя пользователя и пароль.
- Пользователь: имя пользователя, которое используется для входа в ваш кластер, например, admin.
- Пароль: пароль, который используется для входа в ваш кластер.
- База данных: источник данных, к которому вы хотите получить доступ в вашем кластере. Значение этого параметра имеет формат `<catalog_name>.<database_name>`.

`catalog_name`: имя целевого каталога в вашем кластере. Поддерживаются как внутренние, так и внешние каталоги.

`database_name`: имя целевой базы данных в вашем кластере. Поддерживаются как внутренние, так и внешние базы данных.

5.4 Подключение к платформе данных из DBeaver

DBeaver — это клиентское программное обеспечение SQL и инструмент администрирования баз данных.

Чтобы подключиться к базе данных, выполните следующие действия:

- а) Запустите DBeaver.
- б) Щелкните значок плюса (+) в верхнем левом углу окна DBeaver или выберите База данных > Новое подключение к базе данных в строке меню, чтобы получить доступ к помощнику.
- в) Выберите драйвер MySQL.

- г) На этапе выбора базы данных вам будет представлен список доступных драйверов. Нажмите Analytical на левой панели, чтобы быстро найти драйвер MySQL. Затем дважды щелкните значок MySQL.
- д) Настройте подключение к базе данных.
- е) На этапе «Параметры подключения» перейдите на вкладку «Главное» и настройте следующие основные параметры подключения:
 - Хост сервера: hostname или IP-адрес хоста мастер сервера вашего кластера.
 - Порт: порт хоста мастер сервера, например, 9030.
 - База данных: целевая база данных в вашем кластере. Поддерживаются как внутренние, так и внешние базы данных
 - Имя пользователя: имя пользователя, которое используется для входа в ваш кластер, например, admin.
 - Пароль: пароль, который используется для входа в ваш кластер.
- ж) Вы также можете просматривать и редактировать свойства драйвера MySQL на вкладке «Свойства драйвера», если это необходимо. Чтобы изменить определенное свойство, щелкните строку в столбце «Значение» для этого свойства.
- з) Протестируйте подключение к базе данных.
- и) Нажмите «Проверить подключение», чтобы проверить точность параметров подключения. Появится диалоговое окно с информацией о драйвере MySQL. Нажмите «ОК» в диалоговом окне, чтобы подтвердить информацию. После успешной настройки параметров подключения нажмите «Готово», чтобы завершить процесс.
- к) После установки подключения вы можете просмотреть его в дереве подключений к базе данных слева, и DBeaver сможет эффективно подключиться к базе данных.

5.5 Подключение к платформе данных из Jupyter

Вы можете использовать JupySQL поверх Jupyter для выполнения запросов к Селена.

После загрузки данных в кластер вы можете запрашивать и визуализировать их с помощью построения графиков SQL.

Перед началом работы вам необходимо установить локально следующее программное обеспечение:

- JupySQL: *pip install jupysql*
- Jupyterlab: *pip install jupyterlab*
- SKlearn Evaluation: *pip install sklearn-evaluation*
- Python
- pymysql: *pip install pymysql*

После выполнения вышеуказанных требований вы можете открыть Jupyter lab, просто вызвав `jupyterlab` — это откроет интерфейс блокнота. Если Jupyter lab уже запущен в блокноте, вы можете просто запустить ячейку ниже, чтобы получить зависимости.

```
%pip install --quiet jupysql sklearn-evaluation pymysql
```

Для использования обновленных пакетов может потребоваться перезапуск ядра.

```
import pandas as pd
from sklearn_evaluation import plot
# Import JupySQL Jupyter extension to create SQL cells.
%load_ext sql
%config SqlMagic.autocommit=False
```

Вам нужно будет настроить строку подключения в соответствии с типом экземпляра, к которому вы пытаетесь подключиться (url, пользователь и пароль). В примере ниже используется локальный экземпляр.

Подключение к Селена через JupySQL

В этом примере используется экземпляр Docker, который отражает данные в строке подключения.

Пользователь `root` используется для подключения к локальному экземпляру Селена, создание базы данных и проверки того, что данные действительно могут быть прочитаны и записаны в таблицу.

```
%sql mysql+pymysql://root:@localhost:9030
```

Создать JupySQL БД:

```
%sql CREATE DATABASE jupysql;
```

```
%sql USE jupysql;
```

Создать таблицу:

```
%%sql
```

```
CREATE TABLE tbl(c1 int, c2 int) distributed by hash(c1) properties  
("replication_num" = "1");
```

```
INSERT INTO tbl VALUES (1, 10), (2, 20), (3, 30);
```

```
SELECT * FROM tbl;
```

Сохранение и загрузка запросов

Теперь, после создания базы данных, вы можете записать в нее некоторые образцы данных и запросить их.

JupySQL позволяет разбивать запросы на несколько ячеек, упрощая процесс создания больших запросов.

Вы можете писать сложные запросы, сохранять их и выполнять при необходимости, аналогично CTE в SQL.

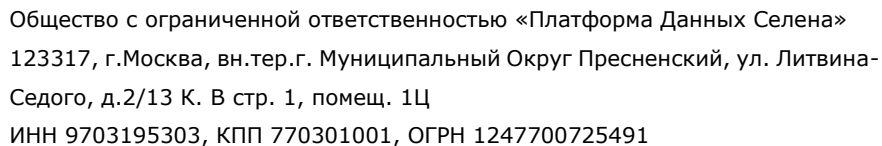
```
%%sql --save initialize-table --no-execute
```

```
CREATE TABLE tbl(c1 int, c2 int) distributed by hash(c1) properties ("replication_num" =  
"1");
```

```
INSERT INTO tbl VALUES (1, 1), (2, 2), (3, 3);
```

```
SELECT * FROM tbl;
```

Обратить внимание, что используется `--with;`, это извлечет ранее сохраненные запросы и добавит их (используя CTE). Затем мы сохраняем запрос в `track_fav`.

04.12.2024 | 33