Общество с ограниченной ответственностью «Платформа Данных Селена»
123317, г.Москва, вн.тер.г. Муниципальный Округ Пресненский, ул. Литвина-Седого, д.2/13 К. В стр. 1, помещ. 1Ц
ИНН 9703195303, КПП 770301001, ОГРН 1247700725491

ФУНКЦИОНАЛЬНЫЕ ТРЕБОВАНИЯ
ПЛАТФОРМА ДАННЫХ СЕЛЕНА
РФ.DIS.00011–01 15

Общество с ограниченной ответственностью «Платформа Данных Селена»
123317, г.Москва, вн.тер.г. Муниципальный Округ Пресненский, ул. Литвина-Седого, д.2/13 К. В стр. 1, помещ. 1Ц
ИНН 9703195303, КПП 770301001, ОГРН 1247700725491

Общество с ограниченной ответственностью «Платформа Данных Селена»
123317, г.Москва, вн.тер.г. Муниципальный Округ Пресненский, ул. Литвина-Седого, д.2/13 К. В стр. 1, помещ. 1Ц
ИНН 9703195303, КПП 770301001, ОГРН 1247700725491

Аннотация

Документ РФ.DIS.00011-01 15 «Платформа Данных Селена» подготовлен на основе рекомендаций государственного стандарта ГОСТ 19.402—78 Единая система программной документации. Описание программы.

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

Ознакомление с документом «Платформа Данных Селена» персонала подразделения, принимающего участие в работе программного комплекса, производится под роспись с внесением соответствующей записи в журнал первичного инструктажа.

СОДЕРЖАНИЕ

Обозначения и сокращения

В документе РФ.DIS.00011-01 15 «Платформа Данных Селена» используются следующие обозначения и сокращения, имеющие соответствующие значения:

Обозначение или сокращение	Значение
ГОСТ	Государственный стандарт
РФ	Российская Федерация
СУБД	Система управления базами данных
УЗ	Учетная запись
AI	Искусственный интеллект
BI	Business intelligence. Система сбора, обработки и анализа данных
DAG	Directed acyclic graph. Ориентированный ациклический граф
ETL	Extract, Transform, Load
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	***S**tructured **Q**uery **L**anguage* — «язык структурированных запросов»

1 Общие сведения

1.1 Наименование программы

Наименование программы на русском языке: Платформа Данных Селена.
Наименование программы в латинской транслитерации: Selena Data Platform.
Краткое наименование программы на русском языке: Селена.
Краткое наименование программы в латинской транслитерации: Selena.

1.2 Описание программы

Платформа Данных Селена предназначена для построения сверхбыстрой платформы хранения и обработки данных с массивно-параллельной обработкой (MPP), разработанная для упрощения и ускорения доступа к данным, быстрой аналитики в реальном времени.

Платформа Данных Селена обеспечивает требуемым функционалом все сферы бизнеса:

- финансы и страхование;
- телекоммуникации;
- производственный сектор;
- государственный сектор;
- промышленность;
- торговля;
- и т.д.

1.3 Цели внедрения

Целями внедрения Платформы Данных Селена являются:

- построение централизованного хранилища данных;
- снижение затрат на инфраструктуру;
- Реализация аналитики в реальном времени;
- Ускорение аналитических запросов;
- Упрощение архитектуры хранилища данных.

1.4 Техническое описание

Платформа Данных Селена разработана на языках программирования:

- Java;
- C++;
- Kotlin;
- TypeScript.

Для функционирования Платформы Данных Селена требуется предварительно установленное программное обеспечение:

− пакет прикладных программ;

− базы данных источников информации;

− файловые ресурсы для размещения шаблонов документов;

− обозреватели для работы с веб интерфейсами Платформы Данных Селена;

- клиенты и консольные приложения для разработки запросов к данным на Платформе Данных Селена

− консольные приложения для настройки и обслуживания программы Платформа Данных Селена.

Компетенции обслуживающего персонала и пользователей Платформы Данных Селена должны обеспечивать бесперебойную работу всех модулей программного обеспечения в режиме реального времени с технологическими перерывами на техническое обслуживание оборудования и обновление программно-аппаратных средств.

1.5 Возможности приложения

Основные возможности:

- централизованное хранение данных
- обработка любых типов данных на любых языках;
- загрузка данных из различных источников;
- обработка и доставка потоковых данных в хранилище;
- возможность последовательного и параллельного выполнения процессов обработки;
- консолидация и нормализация данных;
- управление потоками данных.

2 Описание логической структуры

1.1 Логическая архитектура решения

Платформа Данных Селена основана на архитектуре открытых данных. Архитектура представлена на рисунке 1.

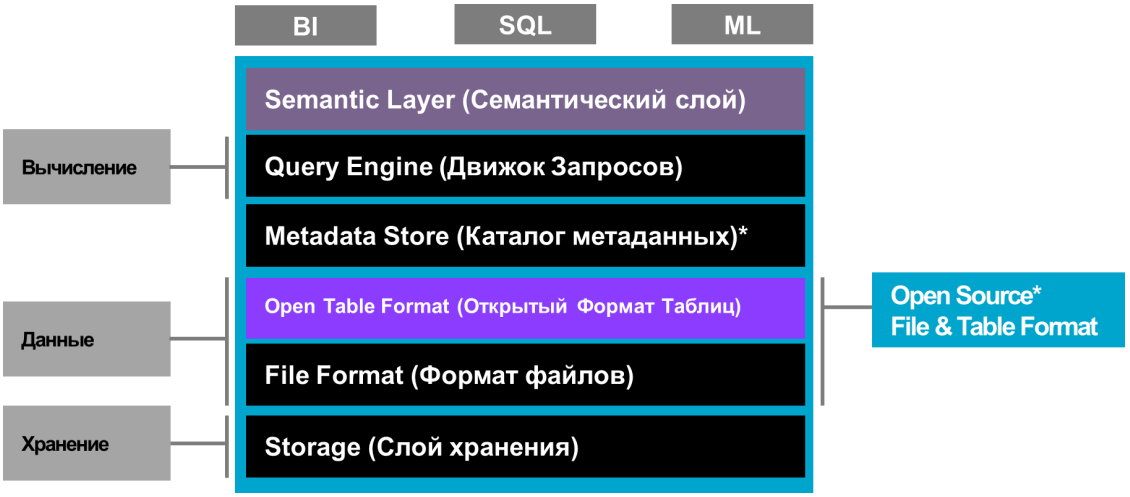

Рисунок 1. Архитектура открытых данных

Архитектура открытых данных используется при построении хранилищ данных Data Lakehouse. Это архитектурный подход, объединяющий лучшие практики, взятые из архитектур Data Warehouse и Data Lake, предоставляющий лучшие возможности хранения и обработки данных.

Платформа данных Селена строится на принципах и подходах, полностью соответствующих архитектуре Data Lakehouse, таких как:

- использование открытых форматов таблиц, обеспечивающий независимость единого слоя данных;
- использование каталога данных, который хранит в себе информацию о местонахождения данных, и являющийся обязательным компонентом платформы;
- использование независимого решения для хранения данных. В Селене используется S3, который является технологией для удобного и надежного хранения большого объема данных;

- использование движка запросов, для выполнения задач, связанных с обработкой данных и вычислениями;

Соблюдение этих подходов при построении хранилища данных обеспечивают следующие положительные преимущества:

- универсальность. По сравнению с блочными и файловыми хранилищами решение S3 позволяет хранить абсолютно любые данные;
- независимость и масштабируемость всех слоев архитектуры;
- доступность данных для сторонних систем. Все объекты хранилища располагаются в плоском адресном пространстве, без иерархии, упрощая доступ и работу с файлами;
- высокую надежность, данные хранятся одновременно в разных узлах;
- экономичность, версионность и многое другое.

1.2 Компоненты Платформы Данных Селена

Платформа данных Селена реализована с использованием следующих компонентов:

- мастер сервер, является частью ядра системы, обеспечивает управление метаданными, управление клиентскими соединениями, планирование запросов и диспетчеризацию запросов;
- вычислительный сервер, является частью ядра системы, обеспечивает выполнение SQL запросов и вычисление данных;
- панель управления кластером (Cluster Manager) обеспечивает администраторам все необходимые функции по созданию и настройке кластера. Обеспечивает функции добавления и удаления нод, назначение типов и развертывание приложений на новый 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 и управления данными обеспечивает загрузку данных из различных источников в целевое хранилище, а также реализует функции визуального формирования потоков данных, и настройки расписания;

- метакаталог данных содержит информацию о метаданных, хранящихся в объектном хранилище, и обеспечивает быстрое формирование запросов на данные в объектном хранилище;
- модуль мониторинга обеспечивает сбор информации о работе кластера, состоянии основных компонентов и вычислительных мощностях;
- модуль генеративного ИИ обеспечивает пользователям быстрый и простой доступ к данным;
- модуль хранения данных обеспечивается непосредственное хранение данных.

3 Функциональные требования

Платформа Данных Селена представляет собой новый подход к хранению и работе с данными. Селена предлагает широкий набор функций для быстрой аналитики данных в реальном времени в больших масштабах.

2.1 Функции, реализуемые вычислительным движком

В части ядра Платформы Данных Селена обеспечивает следующие функциональные возможности:

- для расчетов и вычислений используется технология массовой параллельной обработки (MPP). Один запрос разбивается на несколько физических вычислительных блоков, которые могут выполняться параллельно на нескольких машинах;
- полностью векторизованный механизм выполнения обеспечивает более эффективное использование вычислительной мощности процессора, поскольку этот механизм организует и обрабатывает данные в виде столбцов. Векторизованный механизм выполнения в полной мере использует инструкции SIMD. Этот механизм может выполнять больше операций с данными с помощью меньшего количества инструкций;
- оптимизатор задач, обеспечивающий нахождение самого оптимального плана на основе собранной статистики данных. Автоматическое преобразование исходного запроса пользователей в альтернативную формулировку, с целью достижения того же результата за меньшее время и\или количества операций. Оптимизатор может определить и удалить ненужные части запроса, которые не влияют на конечный результат;
- автоматизированный режим работы с материализованными представлениями. Селена автоматически обновляет данные в соответствии с изменениями в базовой таблице без необходимости дополнительных операций по обслуживанию. Кроме того, выбор материализованных представлений также осуществляется автоматически;

- Селена может работать в качестве вычислительного механизма для анализа данных, хранящихся в озерах данных, таких как Apache Hive, Apache Iceberg, Apache Hudi и Delta Lake. Эта функция позволяет пользователям беспрепятственно запрашивать внешние источники данных, устраняя необходимость в переносе данных. Таким образом, пользователи могут анализировать данные из разных систем, таких как HDFS и Amazon S3, в различных форматах файлов, таких как Parquet, ORC, CSV и т. д.;
- Селена поддерживает различные отрытые форматов хранения данных Iceberg, Hudi, Data lakes и т.д.;
- Селена совместима с протоколом MySQL;
- Селена обеспечивает возможность разделения ресурсов для хранения данных и вычислительной среды;
- Селена обеспечивает расширенные возможности, связанные с кэшированием. Обеспечивая кэширование запросов, данных, результатов, планов запросов;
- Селена в полной мере использует современные многоядерные процессоры и инструкции SIMD для повышения производительности ;
- Селена поддерживает работу с гибридным строково-столбцовым форматов хранения.

2.2 Функции, реализуемые модулем управления кластером (Cluster Manager)

В части модуля управления кластером Платформа Данных Селена обеспечивает следующие функциональные возможности:

- Управление пользователями, группами и ролями, а также доступом к интерфейсам системы;
- Добавление и удаление хостов в кластер;
- Управление кластером, включая установку и удаление приложений на хостах в кластере;
- Управление доступом и привилегиями к объектам в хранилище;

- Управление конфигурационными параметрами кластера;
- Управление лицензией.

2.3 Функции, реализуемые модулем загрузки и управления данными

В части модуля загрузки и управления данными Платформа Данных Селена обеспечивает следующие функциональные возможности:

- позволяет визуализировать рабочие задачи, рабочие процессы и все процедуры обработки данных, в том числе в формате ориентированный ациклический граф (DAG);
- обеспечивает планирование задач в визуальных рабочих процессах;
- поддерживает взаимосвязи при оркестрации операций с данными для различных приложений, работающих с большими данными;
- поддерживает запланированное и ручное планирование на основе выражений cron;
- поддерживает запуск рабочего процесса, запуск выполнения с текущего узла, возобновление отказоустойчивого рабочего процесса, возобновление приостановленного процесса, запуск выполнения с отказавшего узла, дополнение, синхронизация, повторный запуск, пауза, остановка, возобновление ожидающего потока;
- поддерживает приоритезацию экземпляров процессов и задач;
- обеспечивает отслеживание статуса выполнения процессов и задач и осуществляет оповещение по электронной почте, Telegram, Http, Slack в зависимости от сработавшего тригера;
- объединяет задачи в режиме потоковой передачи ориентированного ациклического графа;
- позволяет отслеживать состояние выполнения задач в режиме реального времени;
- в рамках проектирования задачи или процесса поддерживает различные типы задач Shell, MR, Python, Spark, SQL и т. д.;
- в составе имеет коннекторы для озер данных и хранилищ данных, включая коннекторы Delta Lake, Hive, Hudi и Iceberg;

- в составе имеет коннекторы для систем управления реляционными базами данных, включая коннекторы MySQL, PostgreSQL, Oracle и SQL Server;
- в составе имеет коннекторы для множества других систем, включая коннекторы Cassandra, ClickHouse, OpenSearch, Pinot, Prometheus, SingleStore и Snowflake;
- в составе имеет ряд других полезных коннекторов, таких как коннекторы JMX, System и TPC-H.

2.4 Функции, реализуемые метакаталог данных

В части метакаталога Платформа Данных Селена обеспечивает следующие функциональные возможности:

- реализация спецификации каталога REST для Apache Iceberg;
- обеспечивает подключение к любой системе управления доступом с провайдером OpenID;
- обеспечивает транслирование событий изменений во внешние системы;
- обеспечивает возможность запрета изменений в таблицах;
- бесшовная интеграция с S3;
- обеспечивает масштабирование по горизонтали и обновление без простоев;

2.5 Функции, реализуемые модулем мониторинга

В части модуля мониторинга Платформа Данных Селена обеспечивает следующие функциональные возможности:

- многомерная модель данных с данными временных рядов, идентифицированными по названию показателя и парам «ключ-значение»;
- отсутствие зависимости от распределенного хранилища;
- сбор временных рядов происходит с помощью модели извлечения по протоколу HTTP;
- передача временных рядов поддерживается через промежуточный шлюз;

- обеспечивает возможность обнаружения объектов мониторинга в автоматическом режиме или статической конфигурации;
- поддерживает несколько режимов построения графиков и информационных панелей;
- поддерживает возможность отображения график, таблицу, тепловую карту, статус параметра, состояние сервиса и свободный текст на отображаемой панели;
- поддерживает возможность расширения функций визуализации по средствам загрузки дополнительных плагинов;
- позволяет менять формат отображаемой панели на свое усмотрение;
- позволяет шаблонировать настроенные панели,

2.6 Функции, реализуемые модуль генеративного ИИ

В части модуля генеративного ИИ Платформа Данных Селена обеспечивает следующие функциональные возможности:

- упрощение работы с данными с помощью ИИ;
- обеспечивает поддержку построения больших языковых моделей (LLM);
- обеспечивает семантический поиск данных;
- обеспечивает обработку и преобразование из естественного языка в SQL запрос и обратно.

2.7 Функции, реализуемые модуль хранения данных

В части модуля хранения данных Платформа Данных Селена обеспечивает следующие функциональные возможности:

- записывает данные и метаданные вместе как объекты, устраняя необходимость в базе данных метаданных;
- выполняет функции (код стирания, проверку на битовый сбой, шифрование) как встроенные;

- обеспечивает высокую параллельной обработки благодаря набору распределенных серверов;
- защищает данные с помощью встроенного в каждый объект «стирающего кода». Стирающее кодирование позволяет выполнять восстановление на уровне объекта и может восстанавливать несколько объектов независимо друг от друга;
- обеспечить бесперебойное чтение и запись при наличии в развёртывании только ((N/2)+1) рабочих дисков;
- Обеспечивает защиту от Bitrot и гарантирует, что он никогда не будет считывать повреждённые данные. Модуль хранения выявляет и исправляет повреждённые объекты на лету;
- Обеспечивает целостность от начала до конца за счёт вычисления хэша при чтении и его проверки при записи из приложения, по сети и в память/на диск;
- поддерживает схемы шифрования на стороне сервера для защиты данных, где бы они ни находились;
- обеспечивает конфиденциальность, целостность и подлинность с незначительными потерями производительности. Шифрование на стороне сервера и на стороне клиента поддерживается с помощью AES-256-GCM, ChaCha20-Poly1305 и AES-CBC;
- Зашифрованные объекты защищены от несанкционированного доступа с помощью шифрования на стороне сервера AEAD;
- Обеспечивает совместимость со широко используемыми решениями для управления ключами (например, HashiCorp Vault);
- использует систему управления ключами (KMS) для поддержки SSE-S3;
- Обеспечивает возможность отключения всех API, которые потенциально могут изменять данные и метаданные объектов;
- Обеспечивает непрерывная репликация;

Лист регистрации изменений

Изм.	*Номера листов (страниц)*				*Всего листов (страниц) в документе*	*Номер документа*	*Входящий номер сопроводительного документа и дата*	*Подпись*	*Дата*
	Изменённых	*заменённых*	*новых*	*аннулированных*					

Общество с ограниченной ответственностью «Платформа Данных Селена»
123317, г.Москва, вн.тер.г. Муниципальный Округ Пресненский, ул. Литвина-Седого, д.2/13 К. В стр. 1, помещ. 1Ц
ИНН 9703195303, КПП 770301001, ОГРН 1247700725491