cedrusdata

Описание функциональных характеристик CedrusData Catalog

ООО «Кверифай Лабс»

ОГРН 1217800163790

ИНН 7811766769

КПП 781101001

Введение

CedrusData Catalog — это система управления метаданными для современных аналитических платформ.

Преимущества CedrusData Catalog: - Поддержка протокола Iceberg REST Catalog (https://iceberg.apache.org/concepts/catalog/#decoupling-using-the-rest-catalog) - Поддержка популярных аналитических систем обработки данных: CedrusData, Trino, Apache Spark, Apache Flink и др. - Поддержка файловых систем S3 и HDFS - Расширенные возможности мониторинга - Возможность приобретения коммерческой версии с технической поддержкой

Данный документ содержит описание функциональных характеристик CedrusData Catalog.

1. Сценарии использования

CedrusData Catalog реализует протокол Iceberg REST Catalog и позволяет управлять таблицами Apache Iceberg.

Типичными сценариями применения CedrusData Catalog являются:

- Создание, изменение и удаление данных Apache Iceberg в озерах данных на основе S3 или HDFS
- Предоставление аналитическим SQL-движкам доступа к данным Apache Iceberg в озерах данных на основе S3 или HDFS

2. Работа с данными Apache Iceberg

2.1. Конфигурация подключения к озеру данных

CedrusData Catalog поддерживает работу с озерами данных на основе S3 или HDFS. Для подключения к озеру данных, пользователь вызывает команду `file-system create`. Аргументами команды являются наборы пар ключ-значение, описывающих конфигурацию подключения к файловой системе:

- S3: URL сервиса S3, статические или динамические ключи доступа, режим формирования путей (virtual-hosted-style или path-style), и др.
- HDFS: Набор XML-файлов конфигурации (`core-site.xml`, `hdfs-site.xml`, и др).

После создания подключения к озеру данных пользователь может проверить работоспособность подключения с помощью команды `file-system check`.

2.2. Создание каталога Apache Iceberg

Данные Apache Iceberg организованы в таблицы. Таблицы организованы в логические схемы, называемые namespace. Namespace организованы в логические каталоги, называемые catalog.

После создания подключения в озеру данных необходимо создать каталог, который будет хранить в себе таблицы и схемы. Для создания каталога пользователь вызывает

команду `iceberg catalog create`, куда передает уникальный идентификатор файловой системы озера данных, а также относительный путь внутри файловой системы по которому будет происходить создание новых объектов.

2.3. Работа с данными Apache Iceberg

После создания каталога, пользователь получает возможность подключиться к CedrusData Catalog по протоколу Apache Iceberg REST и начать создавать, изменять или удалять данные Apache Iceberg.

Для подключения к каталогу пользователь может воспользоваться официальной библиотекой Apache Iceberg или использовать одну из популярных аналитических систем, поддерживающих протокол Apache Iceberg. Например, пользователь может настроить Apache Spark на работу с CedrusData Catalog, передав в качестве параметров конфигурации URL CedrusData Catalog, уникальный идентификатор каталога и ключ доступа.

3. Развертывание CedrusData Catalog

CedrusData Catalog может быть развернут из архива или с использованием Docker-образа.

CedrusData Catalog может быть развернуть как on-premises, так и в облачном окружении.

При развертывании CedrusData Catalog в облаке пользователь может использовать Docker-образ CedrusData Catalog, а также интегрировать его с другими облачными технологиями, такими как Kubernetes, Terraform, Helm.