cedrusdata

Инструкция по эксплуатации 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

CedrusData может быть установлена из архива или из Docker-образа. Краткая инструкция по установке доступна в документации:

- Установка из архива: https://docs.cedrusdata.ru/catalog/latest/deployment/archive.html
- Установка из Docker-образа: https://docs.cedrusdata.ru/catalog/latest/deployment/docker.html

Данный документ представляет собой руководство по эксплуатации кластера CedrusData Catalog, установленного из архива.

1. Основные понятия

Основной задачей CedrusData Catalog является хранение и обработка метаданных таблиц и виртуальных представлений Apache Iceberg. Для этого CedrusData Catalog определяет следующие объекты, организованные в иерархию:

1. CedrusData Catalog — объект верхнего уровня, экземпляр CedrusData Catalog
2. Principal — принципал CedrusData Catalog. Представляет собой пользователя или роль. Выполнение каждой команды CedrusData Catalog происходит от имени определенного принципала
3. File System — подключение к распределенной файловой системе озера данных (S3 или HDFS)
4. Iceberg Catalog — логический каталог Apache Iceberg, который хранит схемы Apache Iceberg Логический каталог имеет ассоциированную с ним файловую систему
5. Iceberg Namespace — логическая схема Apache Iceberg, которая хранит таблицы и виртуальные представления Apache Iceberg
6. Iceberg Table — таблица Apache Iceberg
7. Iceberg View — виртуальное представление Apache Iceberg

Рис. 1: Иерархия объектов CedrusData Catalog

2. Запуск CedrusData Catalog

Для запуска CedrusData Catalog из архива скачайте архив:

```
wget https://downloads.cedrusdata.ru/releases/cedrusdata-catalog-server-458-2.tar.gz
```

Распакуйте архив:

```
tar -xf cedrusdata-catalog-server-458-2.tar.gz
```

Запустите CedrusData Catalog. Параметр `--data-dir` указывает на рабочую директорию каталога:

```
cedrusdata-catalog-server-458-2/bin/launcher start --data-dir data
```

3. Настройка СУБД

CedrusData Catalog хранит состояние объектов в персистентной СУБД. Поддерживаются два типа СУБД: SQLite и PostgreSQL.

По умолчанию CedrusData использует SQLite, файлы которой сохраняются внутри рабочей директории, переданной параметром `--data-dir`.

Для использования PostgreSQL необходимо задать JDBC URL, имя пользователя и необязательный пароль пользователя в файле конфигурации `etc/config.properties`. Например:

```
store.type=postgresql
store.postgresql.jdbc-url=jdbc:postgresql://example.net:5432/database
store.postgresql.username=root
store.postgresql.password=secret
```

После изменения конфигурации необходимо перезапустить CedrusData Catalog:

```
cedrusdata-catalog-server-458-2/bin/launcher restart --data-dir data
```

Подробное описание конфигурации подключения к СУБД приведено в документации https://docs.cedrusdata.ru/catalog/latest/config.html#config-store.

4. Настройка файловой системы

Файлы таблиц Apache Iceberg обычно хранятся в распределенной файловой системе. CedrusData Catalog поддерживает работу с распределенными файловыми системами S3 и HDFS.

Перед началом работы с объектами Apache Iceberg необходимо сконфигурировать подключение к файловой системе. Для этого можно воспользоваться утилитой командной строки `catalog`, которая находится в директории `cedrusdata-catalog-server-458-2/bin`.

Например, для создания подключения к файловой системе S3 на основе MinIO, которая доступна по адресу `example-minio:9000` и имеет статические access key и secret key со значениями `accesskey` и `secretkey` (соответственно), следует выполнить следующую команду:

```
catalog file-system create \
  --file-system-name minio \
  --type s3 \
```

```
  -p endpoint=http://example-minio:9000 \
  -p access-key=accesskey \
  -p secret-key=secretkey
```

Детальные инструкции по настройке файловых систем приведены в документации https://docs.cedrusdata.ru/catalog/latest/file-system.html.

5. Создание логического каталога

CedrusData Catalog позволяет создать один или несколько логических каталогов. Каждый каталог представляет собой логический контейнер, который хранит схемы (namespace) Apache Iceberg, и работает с конкретной файловой системой. Например, вы можете иметь каталог `sales`, который работает с S3-совместимой файловой системой MinIO, и другой каталог `hr`, который работает с файловой системой HDFS.

Создать каталог можно с помощью утилиты командной строки `catalog`.

Например, для создания логического каталога `sales`, который работает с созданной ранее файловой системой `minio`, и хранит данные по относительному адресу `s3://sales`, необходимо выполнить следующую команду:

```
catalog iceberg catalog create \
  --catalog-name sales \
  --file-system-name minio \
  --file-system-location s3://sales
```

Детальные инструкции по работе с логическими каталогами приведены в документации https://docs.cedrusdata.ru/catalog/latest/iceberg.html.

6. Подключение из аналитических движков

Основной задачей CedrusData Catalog является предоставление доступа к объектам Apache Iceberg различным аналитическим движкам. Например, CedrusData, Trino, Apache Spark.

Аналитические движки взаимодействую с CedrusData Catalog по протоколу Apache Iceberg REST. Данный протокол позволяет движку взаимодействовать с CedrusData Catalog согласно публично доступному утвержденному стандарту, развитием которого занимается сообщество Apache Iceberg. Таким образом, аналитические движки могут взаимодействовать с CedrusData Catalog, не имея знаний о деталях реализации и настройки CedrusData Catalog.

Каждый аналитический движок предоставляет собственные инструкции по настройке подключения к каталогу по протоколу Apache Iceberg REST. Детали настройки могут меняться между версиями аналитических движков. Документация CedrusData Catalog предоставляет пошаговые инструкции по подключению к CedrusData Catalog из некоторых популярных движков. Данные инструкции позволяют пользователю создать тестовое подключение и проверить его работоспособность на локальном компьютере.

1. Подключение из CedrusData:
   https://docs.cedrusdata.ru/catalog/latest/engines/engines-cedrusdata.html
2. Подключение из Trino:
   https://docs.cedrusdata.ru/catalog/latest/engines/engines-trino.html

3. Подключение из Apache Spark:
   https://docs.cedrusdata.ru/catalog/latest/engines/engines-spark.html