001

Практическая работа

Применение технологий искусственного интеллекта и машинного обучения для поиска угроз инорфмационной безопасности

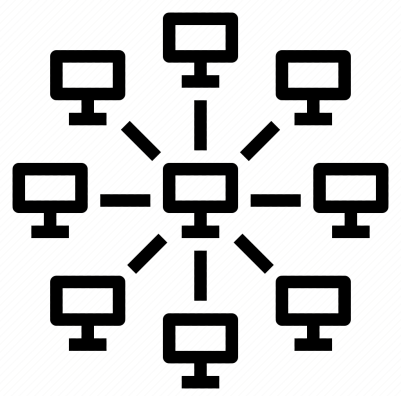
Использование технологии Yandex Query для анализа данных сетевой активности

Цель работы

1. Изучить возможности технологии Yandex Query для анализа структурированных наборов данных
2. Получить навыки построения аналитического пайплайна для анализа данных с помощью сервисов Yandex Cloud
3. Закрепить практические навыки использования SQL для анализа данных сетевой активности в сегментированной корпоративной сети

Общая ситуация

Вам стали доступны данные сетевой активности в корпоративной сети компании XYZ. Данные хранятся в Yandex Object Storage. Проведите разведочный анализ данных и ответьте на вопросы.



Задание

Используя сервис Yandex Query настроить доступ к данным, хранящимся в сервисе хранения данных Yandex Object Storage. При помощи соответствующих SQL запросов ответить на вопросы.

Ход работы

Для выполнения предложенного задания Вам необходимо последовательно проделать следующие шаги:

1. Проверить доступность данных в Yandex Object Storage

1. Проверьте доступность данных (файл yaqry_dataset.pqt) в *бакете* arrow-datasets S3 хранилища Yandex Object Storage. О принципах построения пути можно посмотреть здесь. Проверить можно просто перейдя по правильно сконструированному URL в браузере.

> **Что за хранилище S3**
>
> **S3** или **Simple Storage Service** – сервис, где хранятся данные большого объема. По сути, современный потомок протокола FTP, разработанный компанией Amazon. Может работать как по одноименному протоколу S3, так и по HTTPS. Подробнее смотрите по ссылке

> **Что за pqt файл**
>
> Parquet — это бинарный, колоночно-ориентированный формат хранения данных, со встроенным сжатием. Изначально создавался для экосистемы Hadoop.
>
> В языке R формат parquet поддерживается с помощью применения пакета *arrow*.

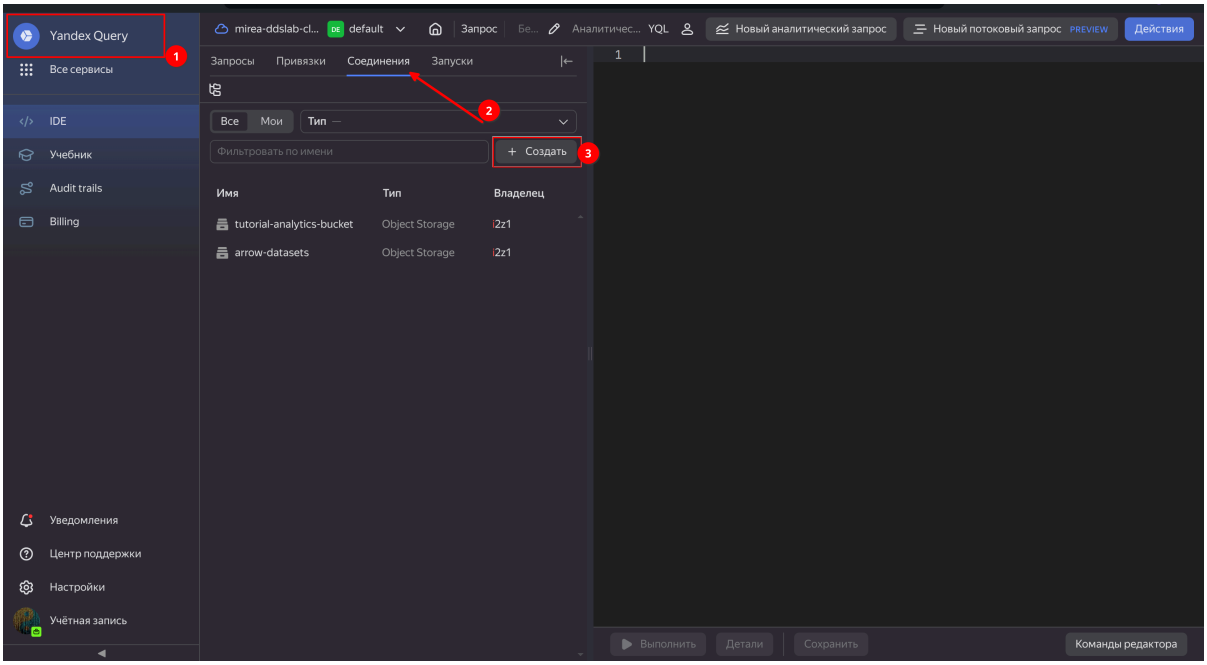
2. Подключить бакет как источник данных для Yandex Query

Yandex Query – это облачное решение для анализа данных, в котором задачи организации хранения, обеспечения доступа и выполнения первичного анализа данных полностью берет на себя сервис-провайдер, то есть Yandex Cloud. Теоретически, весь анализ данных мы можем провестти с любого устройства, хоть со смартфона!

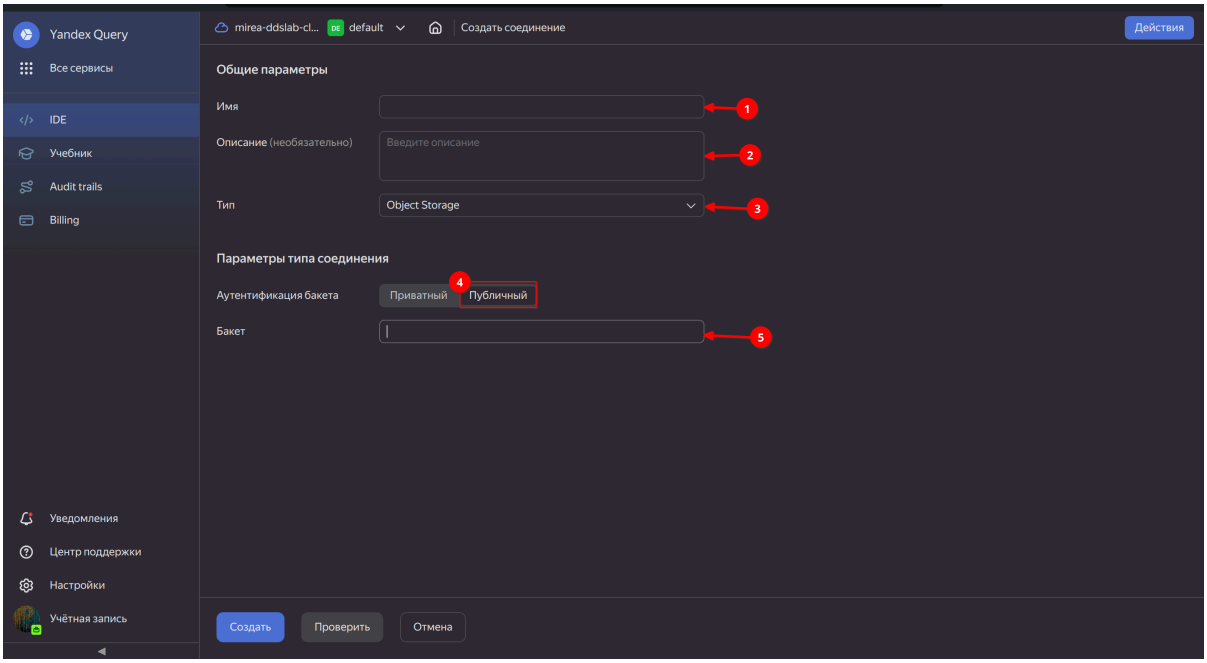
О Yandex Query можно посмотреть замечательный вебинар от его разработчиков.

Перед проведением анализа нам надо связать Yandex Query с хранилищем наших данных. В нашем случае это S3 Object Storage от Yandex Cloud.

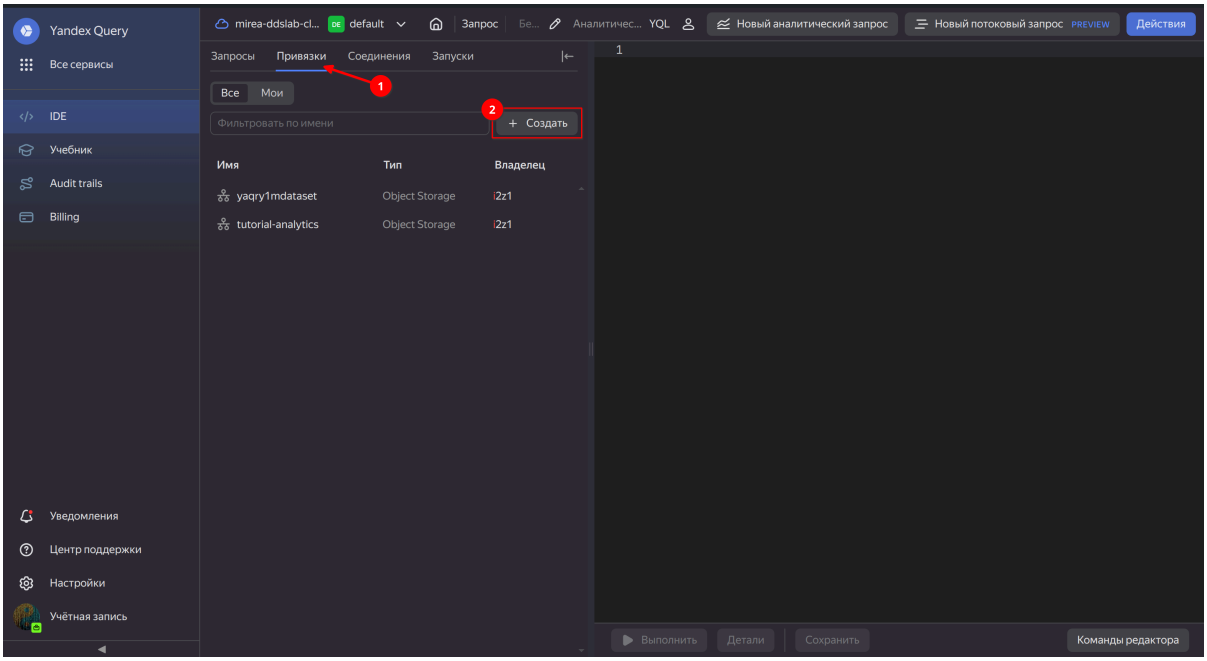
1. Создать соединение для бакета в S3 хранилище



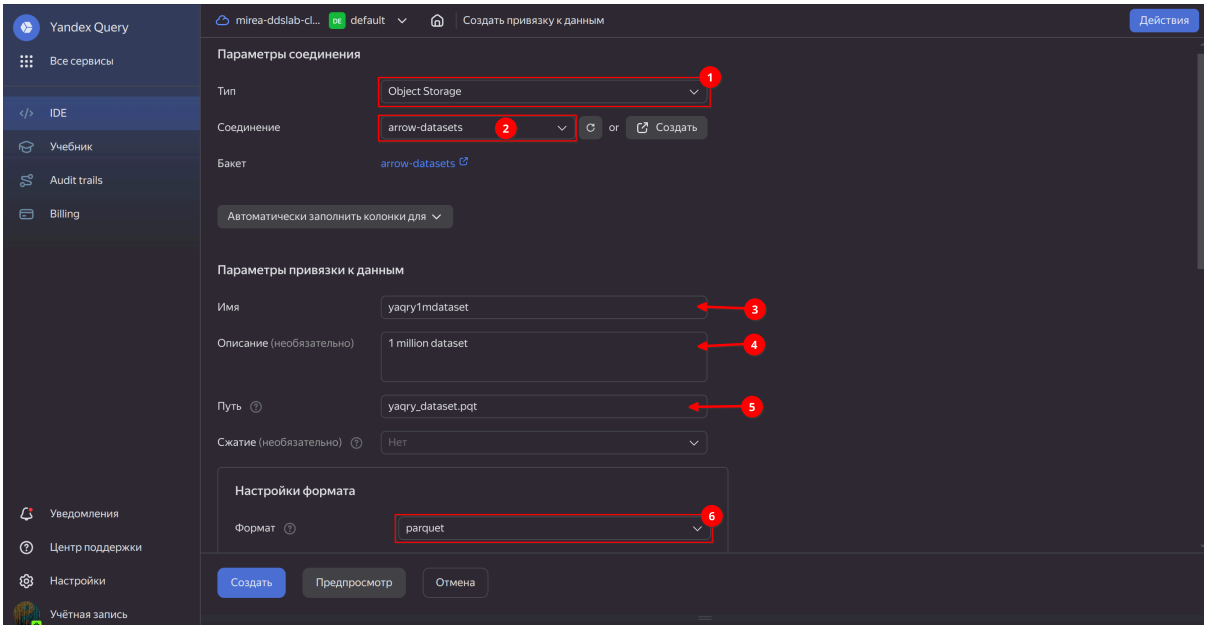
2. Заполняем поля с учетом допустимых символов, выбираем тип аутентификации (*все же помнят что это такое!*) – публичный. Вводим имя бакета в соответствующее поле и сохраняем.



3. Теперь, после создания соединения, укажем какой объект использовать в качестве источника данных. Для этого нужно сделать привязку данных.



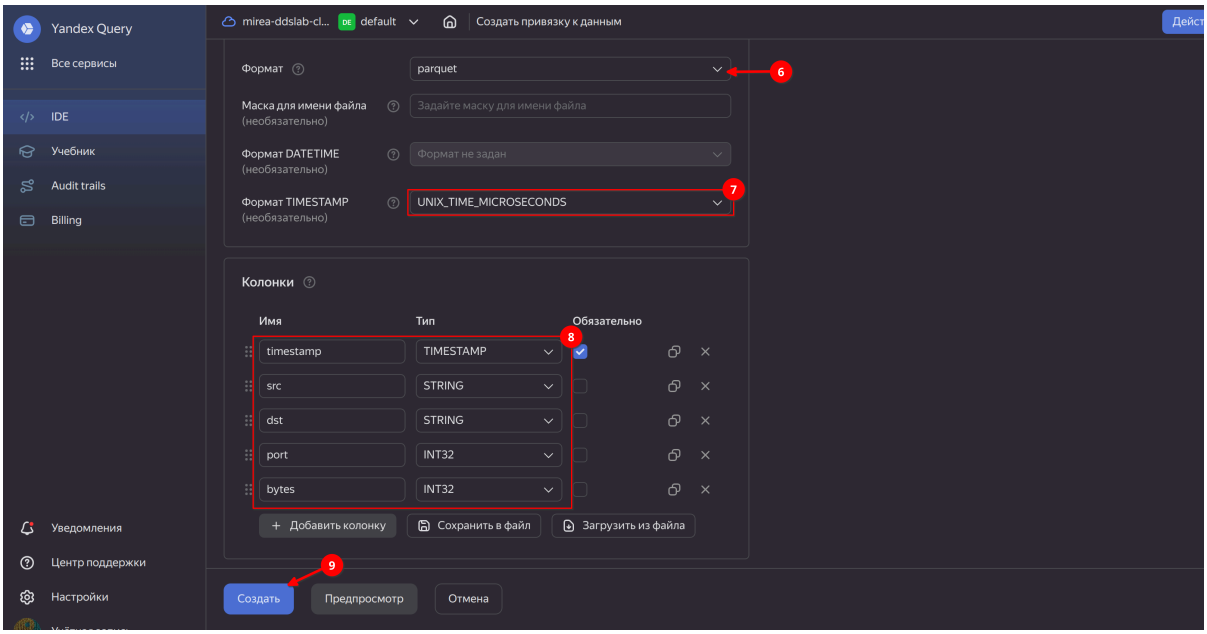
4. Начинаем самый ответственный этап – настройку привязки данных!



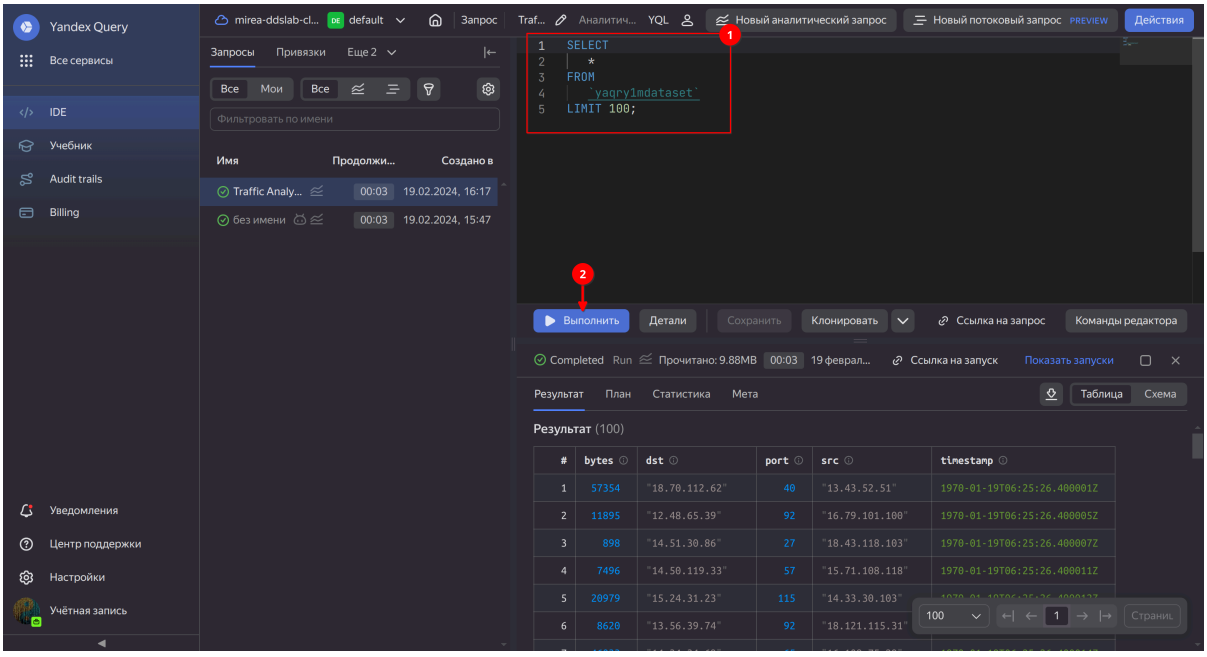
> **Обратите внимание!**
>
> Критический этап – описание состава и формата входных данных. Любая ошибка на данном этапе не позволит выполнить дальнейший анализ данных!

> **но Вам – подскажу :)**

```
SCHEMA=(
timestamp TIMESTAMP NOT NULL,
src STRING,
dst STRING,
port INT32,
bytes INT32
)
```



Если настройки сделаны правильно, то можно попробовать сделать аналитический запрос и посмотреть результат



Если запрос показал не пустую таблицу – тогда Вы на верном пути!

Анализ

Решите следующие задания:

1. Известно, что IP адреса внутренней сети начинаются с октетов, принадлежащих интервалу [12-14]. Определите количество хостов внутренней сети, представленных в датасете.
2. Определите суммарный объем исходящего трафика
3. Определите суммарный объем входящего трафика

> **Tip**
>
> Дополнительные материалы можно найти в Telegram https://t.me/datadrivencybersec

Отчет

Для оформления отчета используйте следующие материалы:

1. https://i2z1.ddslab.ru/posts/lab_recommendations/
2. https://i2z1.quarto.pub/checklab/criteria.html
3. https://github.com/i2z1/Report_template

Сайт курса

https://i2z1.ddslab.ru/MLCTH