003

Практическая работа

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

Анализ данных сетевого трафика при помощи библиотеки Arrow

Цель работы

1. Изучить возможности технологии Apache Arrow для обработки и анализ больших данных
2. Получить навыки применения Arrow совместно с языком программирования R
3. Получить навыки анализа метаинфомации о сетевом трафике
4. Получить навыки применения облачных технологий хранения, подготовки и анализа данных: Yandex Object Storage, Rstudio Server.

Общая ситуация

Вы – специалист по информационной безопасности компании “СуперМегатек”. Вы, являясь специалистом Threat Hunting, часто используете информацию о сетевом трафике для обнаружения подозрительной и вредоносной активности. Помогите защитить Вашу компанию от международной хакерской группировки AnonMasons.

У Вас есть данные сетевой активности в корпоративной сети компании “СуперМегатек”. Данные хранятся в Yandex Object Storage.

⚠ Данные

Данные хранятся в бакете

`arrow-datasets/tm_data.pqt`

О принципах построения пути к датасету в облаке можно посмотреть здесь

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.

Ваши ресурсы

Вычисления

У Вас есть доступ к облачному серверу RStudio Server, с подготовленным рабочим окружением и установленной библиотекой arrow. Доступ к нему осуществляется при помощи любого современного браузера. Однако, для обеспечения дополнительной безопасности подключение осуществляется через сетевой туннель.

Подключение осуществляется через ssh-туннель (local port forwarding) к сереверу на интефейс 127.0.0.1:8787. Для авторизации используйте ключ (смотрите в группе Telegram).

SSH-туннель

О возможностях SSH по созданию сетевых туннелей можно ознакомиться здесь и из документации.

В общем виде, подключение осуществляется при помощи команды (Linux, MacosX)

```
ssh -i <путь к ключу> -L 8787:127.0.0.1:8787 user<ВашНомер>@62.84.123.211>
```

После установления соединения с удаленным сервером и появления консоли, перейдите с помощью Вашего браузера по адресу `http://127.0.0.1:8787` – появится интерфейс Rstudio Server.

Обязательно смените пароль!!!

Пароль по умолчанию в Вашем аккаунте на удаленном сервере – `12345678`.

Его нужно сменить. Для этого после установления туннеля в консоли выполните команду `passwd`, введите текущий пароль, а затем введите новый и подтвердите его.

Этот пароль Вам нужен для авторизации в Rstudio Server.

Rstudio Server – это полноценная IDE (Integrated Development Environment) для разработки на языках R и Python, доступ к которой осуществляется с помощью браузера.

Данные

Стуктура датасета

- Описание полей датасета: timestamp,src,dst,port,bytes
- IP адреса внутренней сети начинаются с 12-14

- Все остальные IP адреса относятся к внешним узлам

Задание

Используя язык программирования R, библиотеку arrow и облачную IDE Rstudio Server, развернутую в Yandex Cloud, выполнить задания и составить отчет.

Задание 1: Надите утечку данных из Вашей сети

Важнейшие документы с результатами нашей исследовательской деятельности в области создания вакцин скачиваются в виде больших заархивированных дампов. Один из хостов в нашей сети используется для пересылки этой информации – он пересылает гораздо больше информации на внешние ресурсы в Интернете, чем остальные компьютеры нашей сети. Определите его IP-адрес.

Задание 2: Надите утечку данных 2

Другой атакующий установил автоматическую задачу в системном планировщике `cron` для экспорта содержимого внутренней wiki системы. Эта система генерирует большое количество трафика в нерабочие часы, больше чем остальные хосты. Определите IP этой системы. Известно, что ее IP адрес отличается от нарушителя из предыдущей задачи.

Задание 3: Надите утечку данных 3

Еще один нарушитель собирает содержимое электронной почты и отправляет в Интернет используя порт, который обычно используется для другого типа трафика. Атакующий пересылает большое количество информации используя этот порт, которое нехарактерно для других хостов, использующих этот номер порта.

Определите IP этой системы. Известно, что ее IP адрес отличается от нарушителей из предыдущих задач.

Ход работы: рекомендации

Импорт данных

Для получения данных можно использовать функцию `read_parquet` или `open_dataset` пакета `arrow`

```
library(arrow)
df <- arrow::open_dataset()
```

Обработка данных

Для обработки данных можно использовать синтаксис, подобный синтаксису dplyr. Например:

```
sw <- read_parquet(file_path, as_data_frame = FALSE)

result <- sw %>%
  filter(homeworld == "Tatooine") %>%
  rename(height_cm = height, mass_kg = mass) %>%
  mutate(height_in = height_cm / 2.54,
         mass_lbs = mass_kg * 2.2046) %>%
  arrange(desc(birth_year)) %>%
  select(name, height_in, mass_lbs)
```

Данной конструкцией создается схема вычислений. Чтобы ее запустить (выполнить) необходим отдельный этап:

Для dplyr:

```
result %>% compute()
```

Для всего остального R:

```
result %>% collect()
```

Подробнее о нем можно узнать в https://arrow.apache.org/docs/r/articles/data_wrangling.html.

> **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