Яндекс

ML + ACS
Splunk DL – первое прикосновение

Наша проблема: вводная

- Здание со «свободным доступом»
- Дублирование пропусков
- Передача пропусков
- Утеря\кража пропусков

Наша проблема: splunk (попытка 1)

- Проблемы формирования большого набора для обучения
- Статистические классификаторы меняют id-классов

Исходные данные: факторы

- Дневной
- Недельный
- Квартальный

```
index=acs earliest=-90d@ latest=@d
| dedup _time, CARDNUM
| eval r_{date_wday}_{date_hour}_{DEVID} = 1
| eval r_{date_hour}_{DEVID} = 1
| eval r_{DEVID} = 1
| fillnull
| bin span=1d _time
| stats sum(r_*) AS r_* min(date_hour) AS min_date_hour max(date_hour) AS max_date_hour by CARDNUM, _time
```

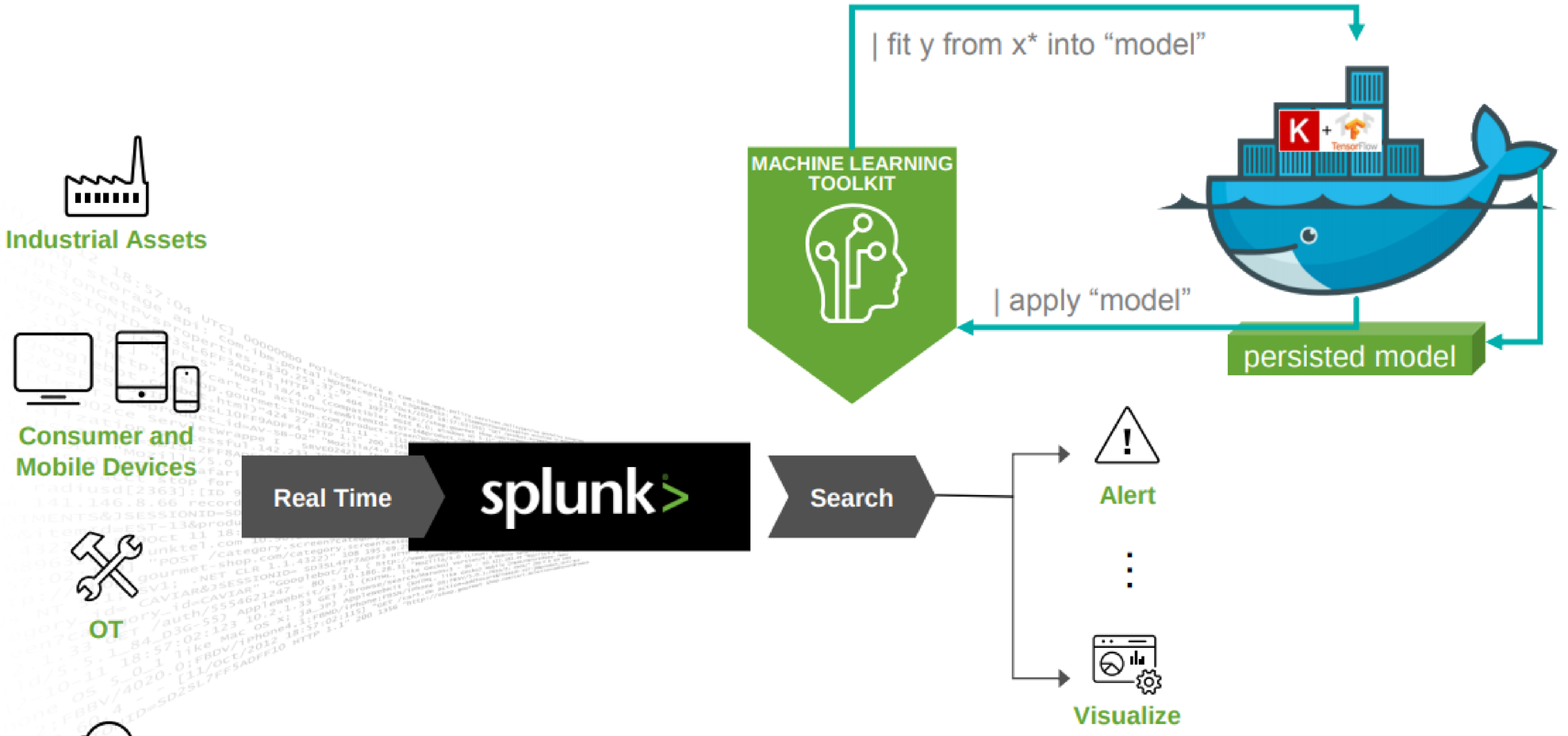
Исходные данные: данные

CARDNUM	_time	r_10	r_10_10	r_10_13	r_10_17	r_10_19	r_10_22	r_10_25	r_10_30	r_10_31	r_10_32	r_10_34	r_10_39	r_10_4
72c0e08631ef9a5c230fec2a2e8d97a3	2018-09-17	0	0	0	0	0	0	0	0	0	0	0	0	0
72c0e08631ef9a5c230fec2a2e8d97a3	2018-09-18	0	0	0	1	1	1	1	0	0	0	0	0	0
72c0e08631ef9a5c230fec2a2e8d97a3	2018-09-19	0	0	1	1	1	2	1	1	0	0	0	0	0
72c0e08631ef9a5c230fec2a2e8d97a3	2018-09-20	1	0	0	0	0	0	0	0	0	0	0	0	0
72c0e08631ef9a5c230fec2a2e8d97a3	2018-09-22	10	0	0	0	0	1	0	1	0	0	0	0	0
72c0e08631ef9a5c230fec2a2e8d97a3	2018-09-25	0	0	0	0	0	0	0	0	0	0	0	0	0
72c0e08631ef9a5c230fec2a2e8d97a3	2018-09-26	1	0	0	0	0	0	0	0	0	0	0	0	0
72c0e08631ef9a5c230fec2a2e8d97a3	2018-09-27	0	0	0	0	0	0	0	0	0	1	0	0	0
72c0e08631ef9a5c230fec2a2e8d97a3	2018-09-28	0	0	0	0	0	0	0	0	0	0	0	0	0
72c0e08631ef9a5c230fec2a2e8d97a3	2018-09-29	1	0	0	0	0	0	0	0	0	0	0	2	0
72c0e08631ef9a5c230fec2a2e8d97a3	2018-10-01	0	0	0	0	0	0	0	0	0	0	0	0	0
72c0e08631ef9a5c230fec2a2e8d97a3	2018-10-02	0	0	0	0	0	0	0	0	0	0	0	0	0
72c0e08631ef9a5c230fec2a2e8d97a3	2018-10-03	0	0	0	0	0	0	0	0	0	0	0	0	0
72c0e08631ef9a5c230fec2a2e8d97a3	2018-10-04	0	0	0	0	0	0	0	0	0	0	1	0	0
72c0e08631ef9a5c230fec2a2e8d97a3	2018-10-05	0	0	0	0	0	0	0	0	0	0	0	1	0

Результаты попытки 1

- Точность: 60-90% (целевая > 95%)
- Ежедневное обучение модели
- Срабатывания для истории

Splunk > MLTK > Dockerized Deep Learning



https://conf.splunk.com/files/2018/recordings/exciting-to-be-announced-fn1478.mp4
https://clck.ru/EhWJC

Splunk MLTK: минусы

- Определенная версия CUDA
- Текущая версия скриптов не предполагает оффлайн установки
- Формат обмена данных со спланком не CSV (numpy.ndarray)
- Модель создается руками

Splunk MLTK: создание модели

```
# ----------------------------------------------------------------------------
# WRAPPER functions for notebook stages to execute dynamic code
# ----------------------------------------------------------------------------
# wrapper function to create a model x -> y with parameters p
def create_model(X,Y,param,model_filename):
    K.clear_session()
    input_shape = int(X.shape[1])
    print("FIT build model with input shape " + str(X.shape))
    print(Y)
    activation_function='relu' #'tanh'
    model = Sequential()
    model.add(layers.Dense(int(input_shape), activation=activation_function, input_dim=int(input_shape) ))
    #for l in range(0,2):
    #    model.add(layers.Dense(int(X.shape[1]*3), activation='relu'))
    model.add(layers.Dense(int(input_shape), activation=activation_function))
    model.add(layers.Dense(int(Y.shape[1]), activation=activation_function))
    model.compile(loss='categorical_crossentropy', optimizer='adam', metrics=['accuracy'])

    return model

# wrapper function to fit a model x -> y with parameters p
def create_model_fit(model,X,Y,param):
    returns = {}
    model_epochs = 10
```

Splunk MLTK: плюсы

- Контейнеризация
- Интеграция в pipeline
- Бесшовная обработка данных из Splunk
- ...

Спасибо за внимание!

Гоц Игорь

gots@yandex-team.ru