Десятая независимая научно-практическая конференция «Разработка ПО 2014»

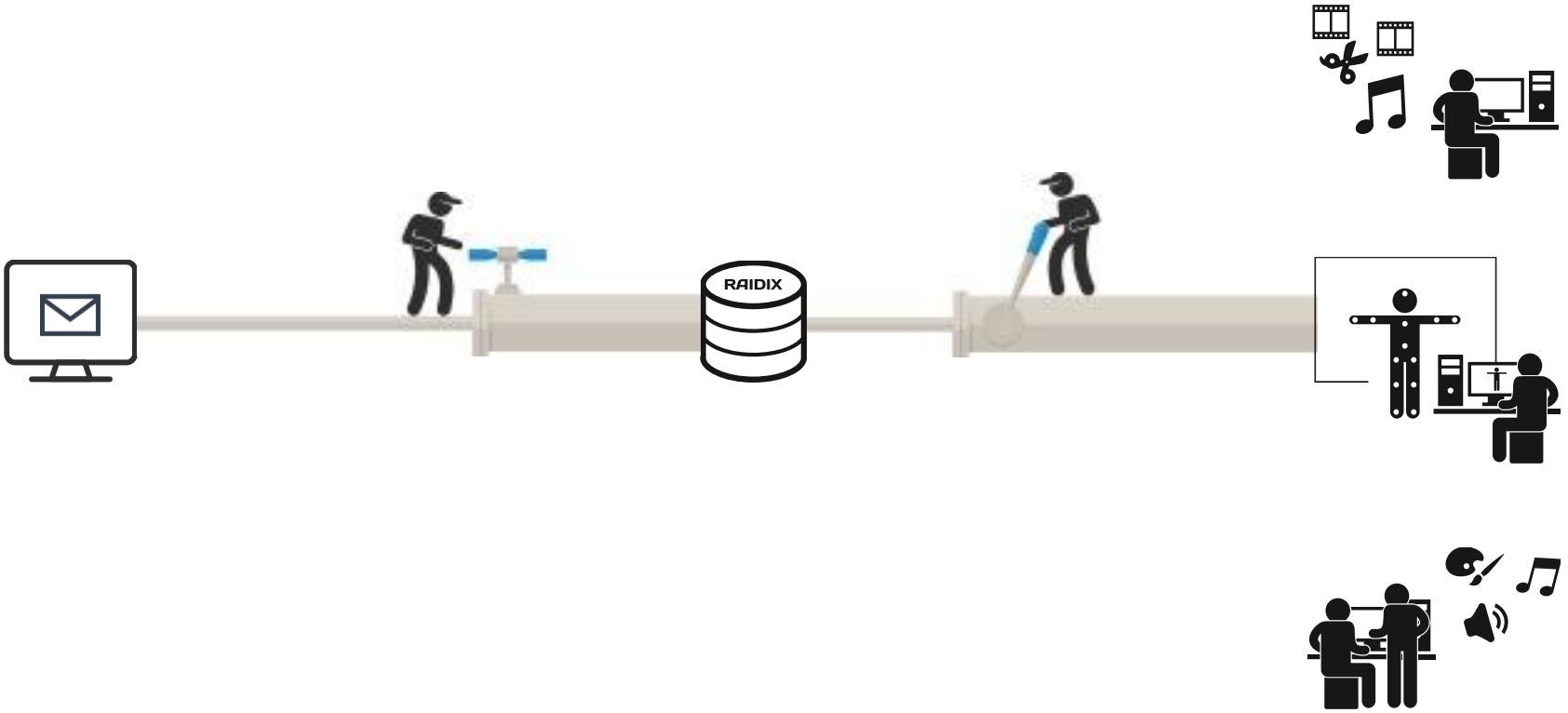
23 – 25 октября, Москва

Отделяем зерна от плевел в случайном лесу

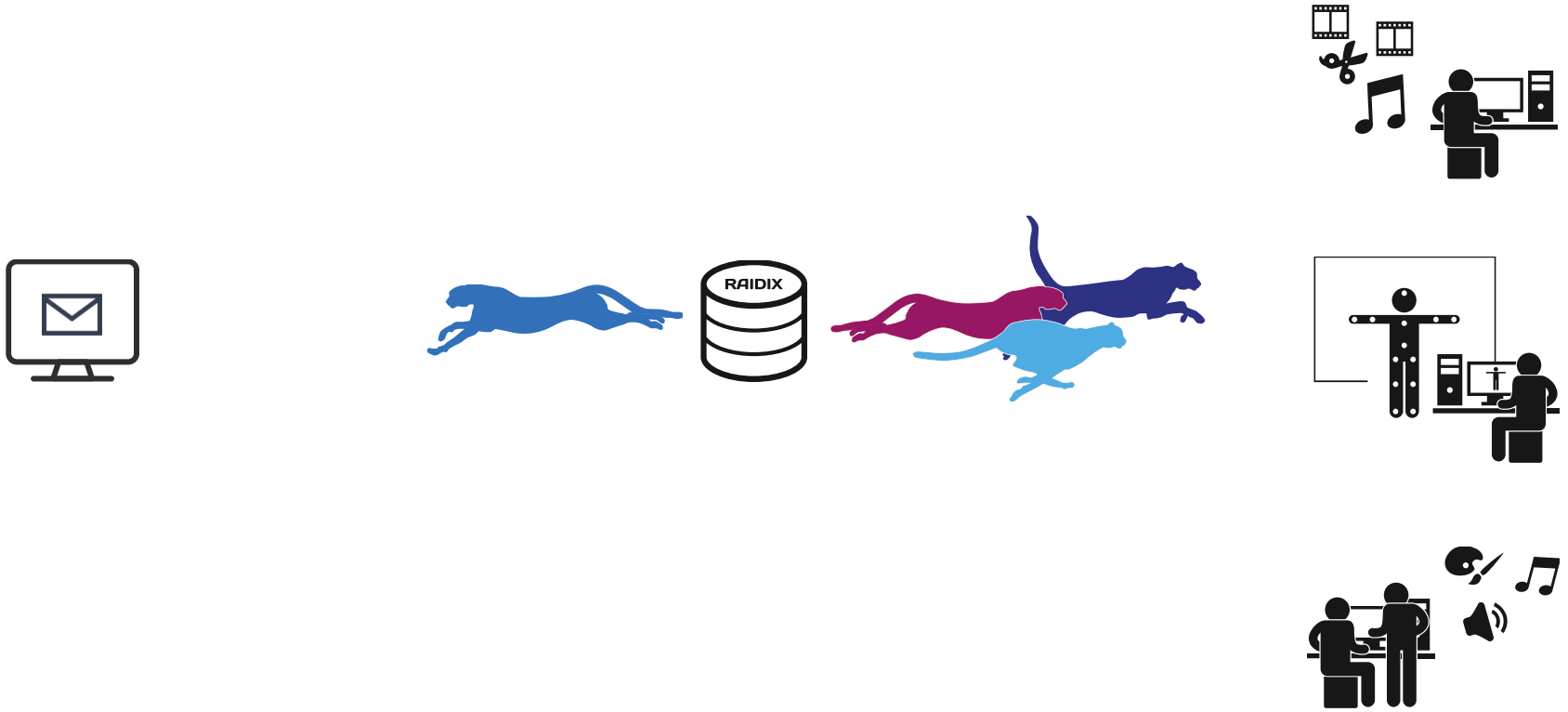
Анализ и классификация мультимедийного потока I/O запросов к системе хранения данных

Светлана Лазарева

Раньше: Приоритет для критически важных приложений выставлялся вручную

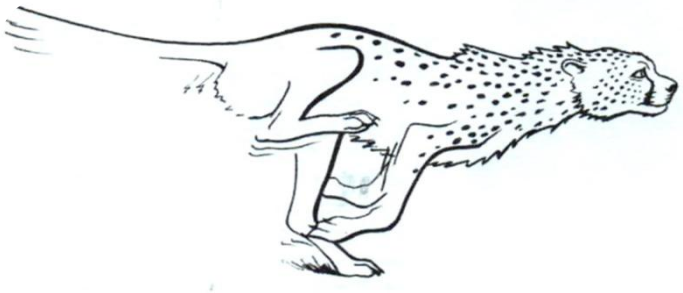


Сейчас: Автоматическое выставление приоритета критически важным приложениям



Требования заказчика

Высокая точность идентификации, порядка **99.9%**



Идентификация
за **20 секунд**

Идентификация только на основе **I/O запросов** (на стороне клиента нельзя устанавливать дополнительные программы для сбора информации)

```
ENT=0 STR=0.000000 TRN=0 EXP=0 INI=NOINIT TGT=NOTARGET
RAID=NORAID LUN=NOLUN LNUM=0 CDB=00000000000000000000000000000000
LLBA=0 PLBA=0 LEN=0 RT=0 STAT=0 SKEY=0 SCOD=0 DRTC=0 NRAC=0 RAP=0
WBP=0 RST=0 RET=0 TET=0 SKP=0
ENT=1 STR=1389634989.883851 TRN=0 EXP=14 INI=2100001086027d6d
TGT=celerity8fc0 RAID=raid LUN=LUN LNUM=0 CDB=
00000000000000000000000000000000 LLBA=0 PLBA=0 LEN=0 RT=0 STAT=0
SKEY=0 SCOD=0 DRTC=0 NRAC=0 RAP=0 WBP=0 RST=0 RET=0 TET=0 SKP=0
ENT=2 STR=1389634992.885059 TRN=0 EXP=12 INI=2100001086027d6d
TGT=celerity8fc0 RAID=raid LUN=LUN LNUM=0 CDB=
00000000000000000000000000000000 LLBA=0 PLBA=0 LEN=0 RT=0 STAT=0
SKEY=0 SCOD=0 DRTC=0 NRAC=0 RAP=0 WBP=0 RST=0 RET=0 TET=0 SKP=0
```

RAIDIX

2014 CEE-SEC(R)

Данные

1. Входными данными являются запросы от различных инициаторов к системе хранения данных
2. Это таблица, строка — это один запрос, столбцов — 25 (различные параметры, которые снимаются для нужд СХД)
3. Используется для идентификации только три столбца:
    - Длина запроса
    - Тип запроса (write или read)
    - Время прихода запроса

Предобработка данных

Специальный алгоритм формирования I/O сигнатур.

- I/O сигнатура формируется за временной интервал T, основываясь на длине запроса
- Интенсивность запросов выше заданного порога.
- I/O сигнатура это вектор состоящий из 8 атрибутов.

RAIDIX

2014 CEE-SEC(R)

I/O Сигнатура. Атрибуты

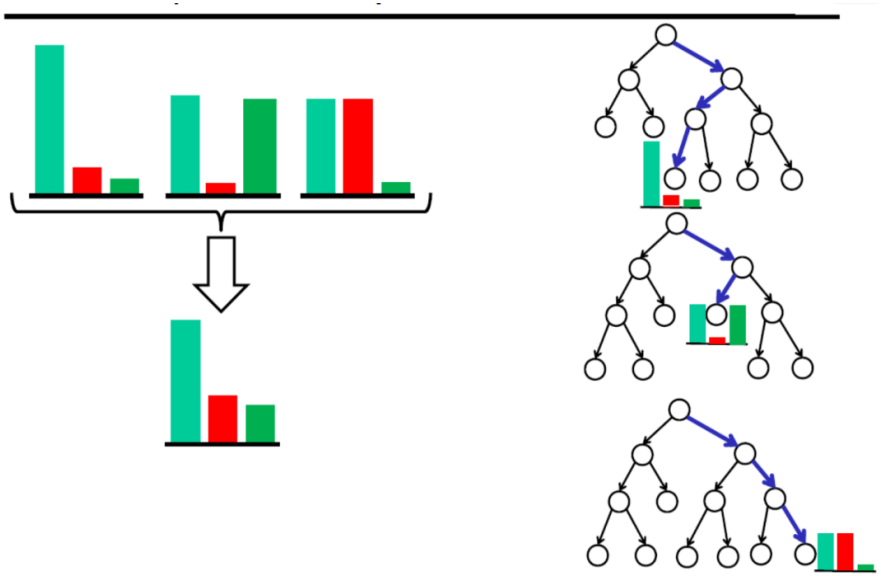
- Длина запроса
- Доля (длина)
- Интенсивность (длина)
- Среднее время между запросами (длина)
- Стековое расстояние (длина)
- Разница (длина)
- Доля записи (длина)
- Класс (имя приложения)

RAIDIX

2014 CEE-SEC(R)

Классификатор: **Random Forest**

- Один из самых эффективных алгоритмов классификации
- Вероятностное распределение на выходе
- Высокая скорость обучения и тестирования
- Относительная простота реализации
- Масштабируемость (способность обрабатывать большие массивы данных)



RAIDIX

2014 CEE-SEC(R)

Выход — имена приложений

RAIDIX

2014 CEE-SEC(R)

Результаты

1. Высокая точность идентификации **99.9%**
2. Высокая скорость идентификации менее **1 миллисекунды**

Имя приложение или паттерна	Вероятность неправильной идентификации
Autodesk Smoke	0.0012
Autodesk Smoke, обработка видео 2K	0.001
Autodesk Smoke, обработка видео 4K	0.001
Apple Finalcut Pro X	0.0009
Adobe Premiere Pro	0.0011
Compressor (Transcode)	0.001
Низкоприоритетные приложения (Backup, Antivirus Office, RemoteDesktop и т. п.)	0.0008

RAIDIX

2014 CEE-SEC(R)

Спасибо за внимание!

www.raidix.com
request@raidix.com