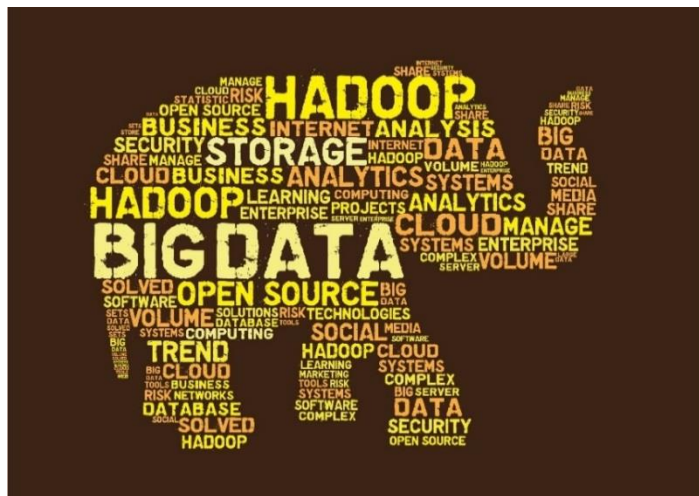




# **SOC В ЭПОХУ BIGDATA**

Сергей Рублев, руководитель SOC, CISSP  
[rublev@infosecservice.ru](mailto:rublev@infosecservice.ru)

# О нас



- SOC для компаний «Открытия»
- Реагирование 24/7
- 50 000 хостов
- Участники Spark Community
- Используем bigdata-решения в «бою»

# Постановка задачи (в разрезе данных)

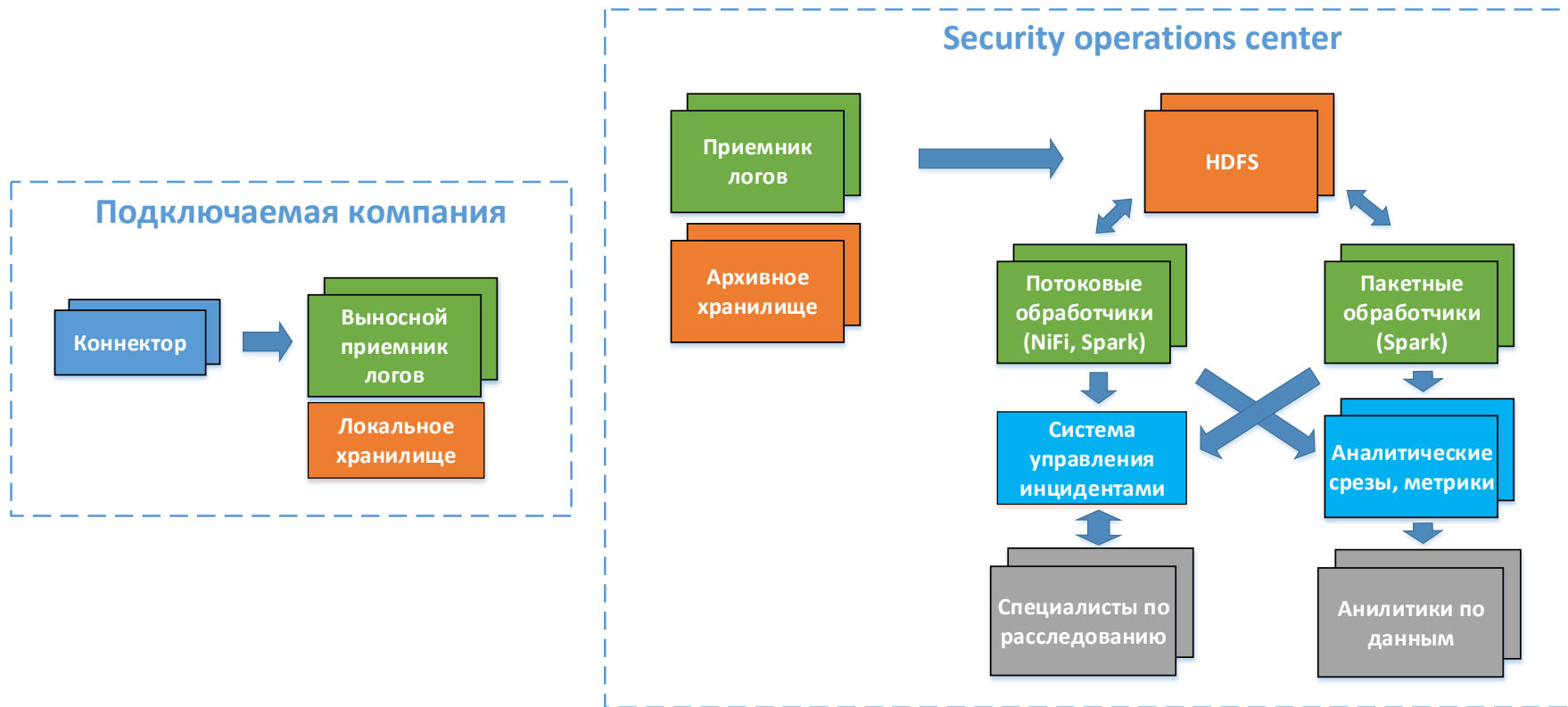
## От «Открытия»

- Поток событий 4 ТБ/сутки
- Контроль потока (пропадание, задержки)
- Работа на исторических данных (старше 1 года)

## От «Инфосекьюрити»

- Высокая доступность и отказоустойчивость
- Горизонтальная масштабируемость
- Построение аналитической платформы для широкого спектра задач
- «Быстрые» ретроспективные запросы
- Поиск по «сырым» данным
- Изоляция подключенных компаний (multitenancy)

# Что построили



# Текущие показатели

**2ТБ** /сутки

поток событий

**1:50**

коэффициент  
сжатия

**2** мин

поиск  
по IP-адресу

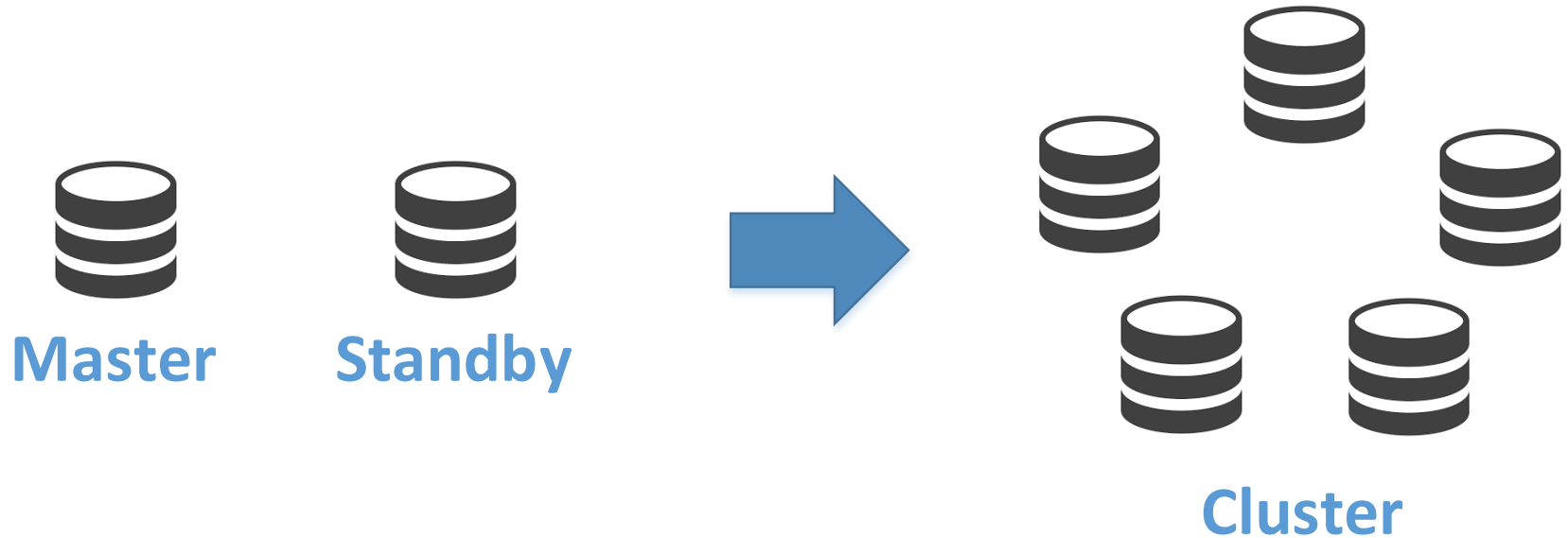
**8** мин

GroupBy + Sort

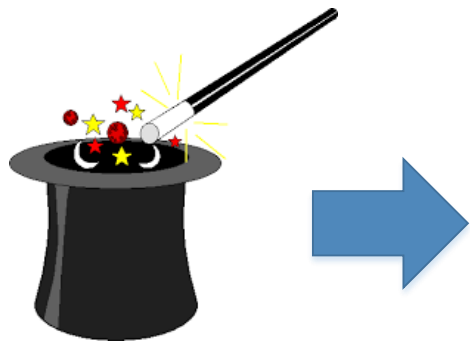
Производительность  
кластера на тестовой  
выборке 3 мес. Netflow:

17млрд событий  
24тб данных

# Отказоустойчивость всех компонент



# Возможности работы с данными



```
// Чтение HIVE-таблицы
val df = sqlContext.read.table("windows_data").
    where('sys_year === 2017
          and 'sys_month === 9
          and 'sys_day === 29)

// Чтение файлов
val json = sqlContext.read.json("jsondata.json")

// Чтение хранилища в Cassandra
val sql_data_count = sqlContext.read.format("org.apache.spark.sql.cassandra")
    .options(Map("table" -> "table", "keyspace" -> "keyspace"))
    .filter('year === 2017)
    .groupBy('dest_ip, 'dest_port)
    .count

// Аналитические запросы
val result = df.select('dvchost, 'eventid, 'sys_timestamp, 'targetusername)
    .groupBy('dvchost, 'eventid)
    .count
    .orderBy('count.desc)
```

# SIEM fail

## Неудобная таксономия событий

- Additional Data
- Agent Address
- Agent Asset ID
- Agent Asset Local ID
- Agent Asset Name
- Agent Asset Resource
- Agent Dns Domain
- Agent Host Name
- Agent ID
- Agent Mac Address
- Agent Name
- Agent Nt Domain
- Agent Receipt Time
- Agent Severity
- Agent Time Zone
- Agent Time Zone Offset
- Agent Translated Address
- Agent Translated Zone
- Agent Translated Zone External ID
- Agent Translated Zone ID
- Agent Translated Zone Name
- Agent Translated Zone Resource
- Agent Translated Zone URI
- Agent Type
- Agent Version
- Agent Zone

- Attacker Address
- Attacker Asset ID
- Attacker Asset Local ID
- Attacker Asset Name
- Attacker Asset Resource
- Attacker Dns Domain
- Attacker Fqdn
- Attacker Geo Country Code
- Attacker Geo Country Flag Url
- Attacker Geo Country Name
- Attacker Geo Latitude
- Attacker Geo Location Info
- Attacker Geo Longitude
- Attacker Geo Postal Code
- Attacker Geo Region Code
- Attacker Host Name
- Attacker Mac Address
- Attacker Nt Domain
- Attacker Port
- Attacker Process ID
- Attacker Process Name
- Attacker Service Name
- Attacker Translated Address

- Destination Address
- Destination Asset ID
- Destination Asset Local ID
- Destination Asset Name
- Destination Asset Resource
- Destination Dns Domain
- Destination Fqdn
- Destination Geo Country Code
- Destination Geo Country Flag Url
- Destination Geo Country Name
- Destination Geo Latitude
- Destination Geo Location Info
- Destination Geo Longitude
- Destination Geo Postal Code
- Destination Geo Region Code
- Destination Host Name
- Destination Mac Address
- Destination Nt Domain
- Destination Port
- Destination Process ID
- Destination Process Name
- Destination Service Name
- Destination Translated Address
- Destination Translated Port
- Destination Translated Zone
- Destination Translated Zone External ID
- Destination Translated Zone ID
- Destination Translated Zone Name





- Source Address
- Source Asset ID
- Source Asset Local ID
- Source Asset Name
- Source Asset Resource
- Source Dns Domain
- Source Fqdn
- Source Geo Country Code
- Source Geo Country Flag Url
- Source Geo Country Name
- Source Geo Latitude
- Source Geo Location Info
- Source Geo Longitude
- Source Geo Postal Code
- Source Geo Region Code
- Source Host Name
- Source Mac Address
- Source Nt Domain
- Source Port
- Source Process ID



# SIEM fail

## Неудобная таксономия событий

Данные для самых  
«стреляющих»  
корреляций здесь 😊

<input checked="" type="checkbox"/>		Device Custom String1
<input checked="" type="checkbox"/>		Device Custom String1 Label
<input checked="" type="checkbox"/>		Device Custom String2
<input checked="" type="checkbox"/>		Device Custom String2 Label
<input checked="" type="checkbox"/>		Device Custom String3
<input checked="" type="checkbox"/>		Device Custom String3 Label
<input checked="" type="checkbox"/>		Device Custom String4
<input checked="" type="checkbox"/>		Device Custom String4 Label
<input checked="" type="checkbox"/>		Device Custom String5
<input checked="" type="checkbox"/>		Device Custom String5 Label
<input checked="" type="checkbox"/>		Device Custom String6
<input checked="" type="checkbox"/>		Device Custom String6 Label

# SIEM fail

Загрузка и выгрузка исторических данных



# Неожиданные победы

- Shell — это круто 😊
- Поисквые запросы -> правила выявления инцидентов
- На Scala легко писать обычным программистам
- DevOps для правил корреляции



# Что дальше



- Больше данных!
- Применение алгоритмов машинного обучения для поиска аномалий
- UEVA
- Анализ графов (достижимость, связность)

---

gk-is.ru

+7 (499) 677-10-00

---