FREETALK

数据清洗在网络安全中的应用

数据都不会清洗，你搭个毛线的社工库

演讲者：耿浩然

云南水熊科技有限公司

FREEBUF

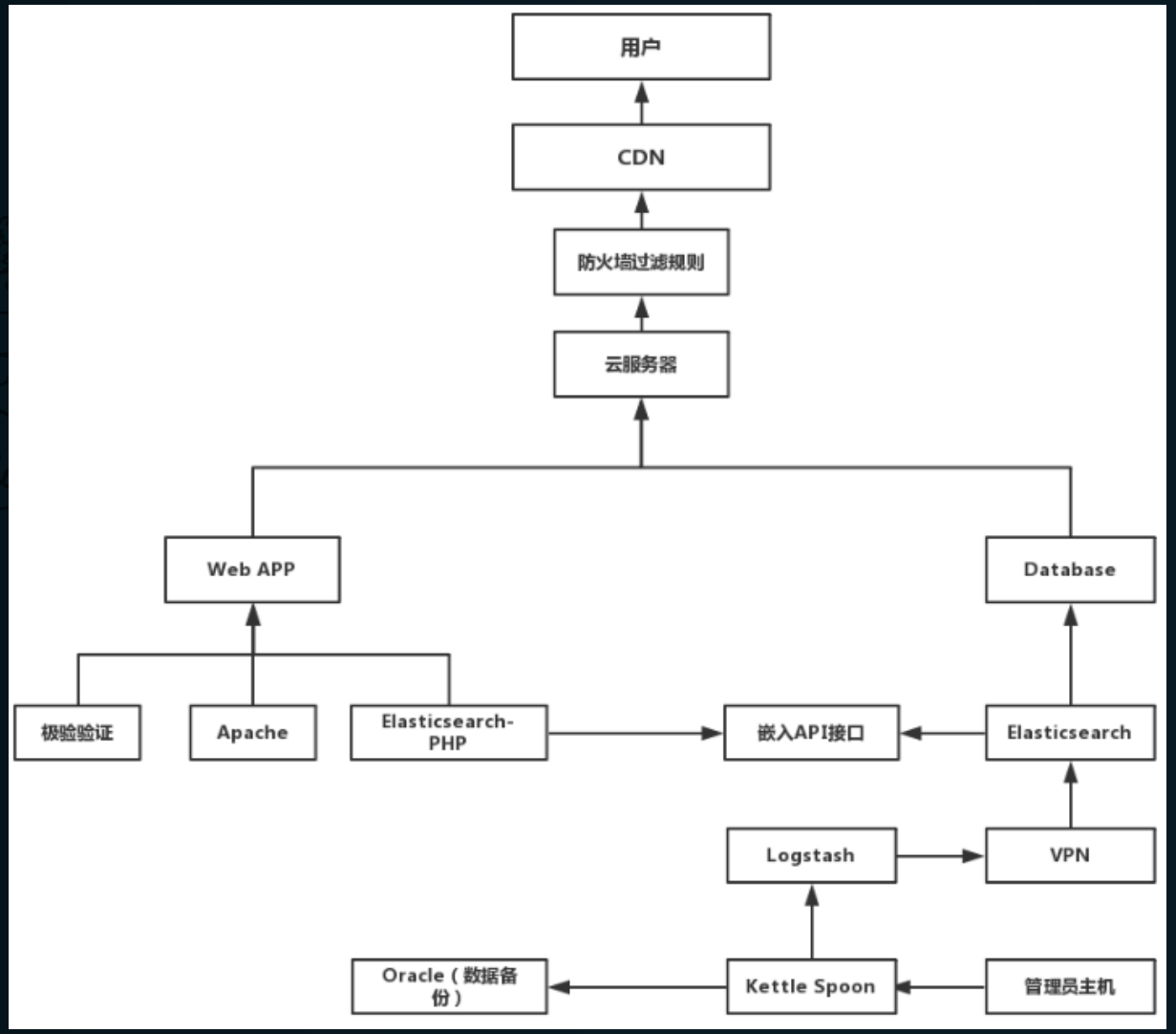
清洗数据的动机

1. 数据不正确或不一致会导致公共或者私人规模的错误结论和错误的行为
2. 减少人力成本
3. 提高数据应用效率

应用实例

某个社工库的技术实现图

ETL工具：Kettle Spoon
索引：Elasticsearch
入库：Logitech



数据分析

增：数据分析

删：删除无用字段

改：改变现有字段

查：要查询的字段

id	email	password	time
1	test@test.com	123	2017
2		3	2017
3	test@test.com		2017
3	test@test.com		2017
4	test@test.com		2017

模型构建

输入：起始数据全部转换为txt文本，采用统一分隔符，相同字段数

第一次过滤：过滤缺失数据

错误输出：将过滤的缺失数据输出到指定路径

第二次过滤：对数据MD5加密，然后进行排序去重

字段选择：删除和增加字段

输出：输出到logitech配置文件指定目录

Logitech config配置文件

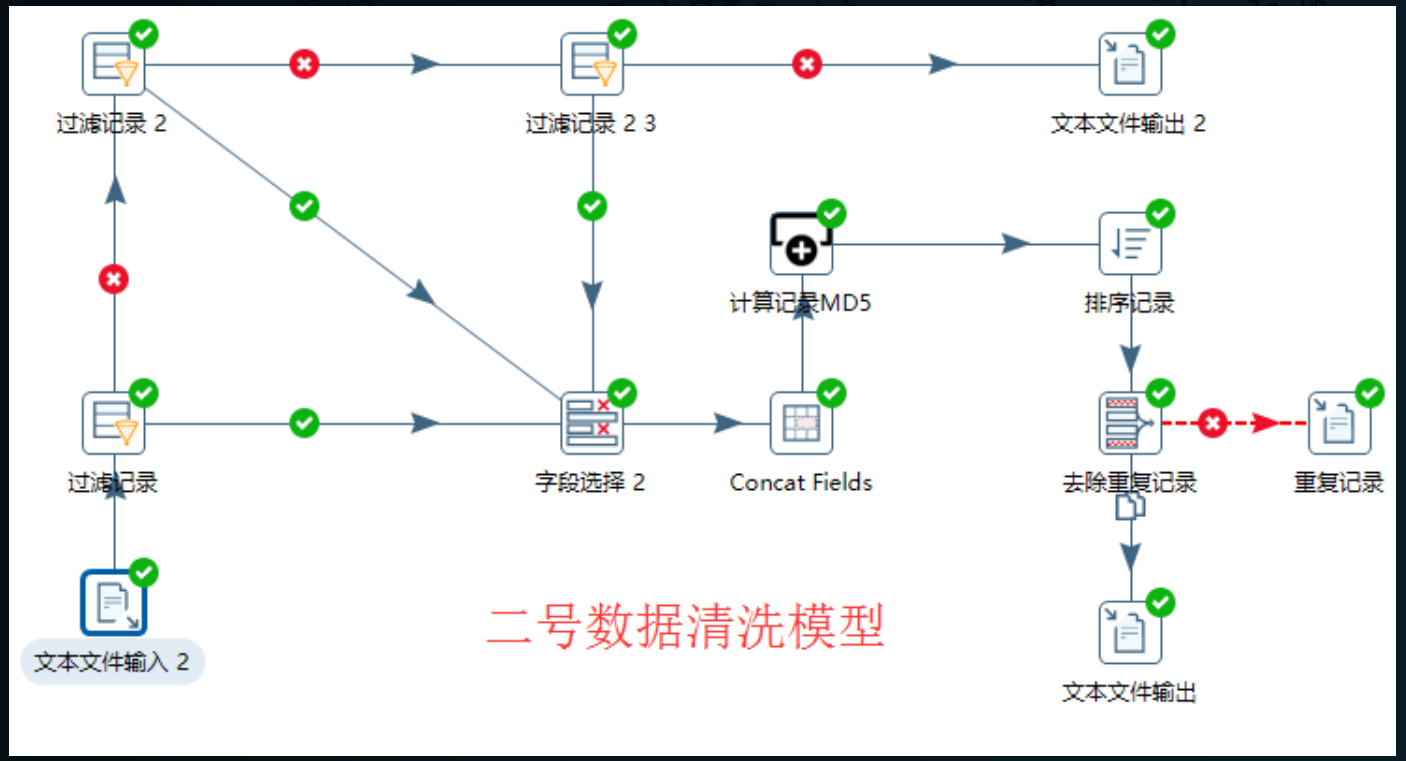
```
input {
    file {
        path => ["E:/test/updata/*.txt"]
        start_position => "beginning"
        ignore_older => 999999999
        codec => plain {charset => ["UTF-8"]}
    }
}

filter {
  csv {
      separator => "|~|^|"
      source => "message"
      columns => ["username","password","email","source"]
  }
mutate { remove_field => ["message","host","@timestamp","path","@version"] }
}
output {
    elasticsearch {
        hosts => "127.0.0.1"
        index => "test"
        document_type => "test"
        flush_size => "60000"
    }
    stdout {
        codec => json
    }
}
```

三分钟模型演示之数据过滤

第一道过滤条件：条数据需要有username和password，username和email，或者password和email。

第二道过滤条件：对数据进行MD5加密排序后去重。



数据处理结构图概览

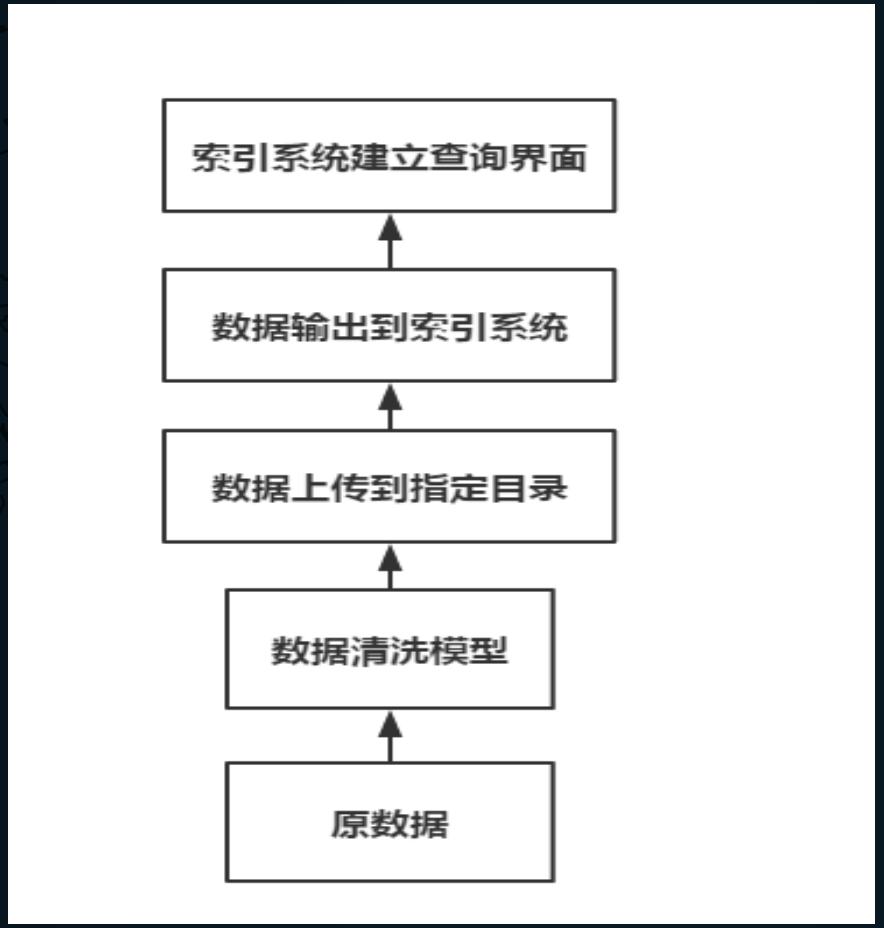
必要条件：

相同字段数量

字段value排序相同

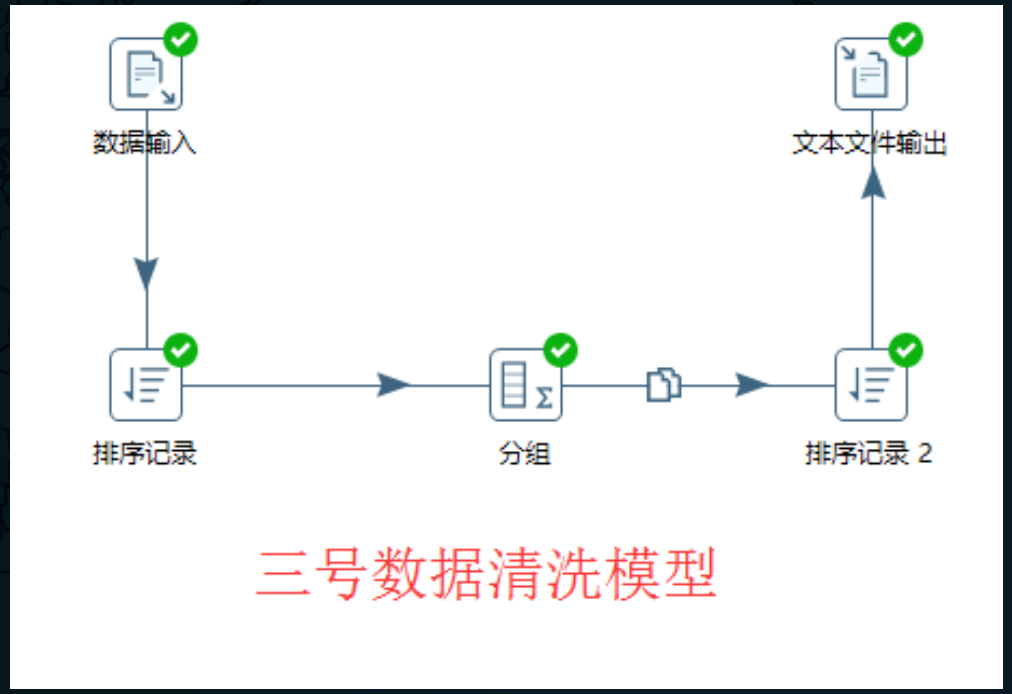
数据分隔符和封闭符相同

数据编码相同



一分钟模型演示之数据统计

弱口令字典清洗



Reference list

http://ieeexplore.ieee.org/document/7307098/ A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection

http://www.aqniu.com/news-views/9369.html BlackHat2015焦点访谈：中国黑客如何玩转深度学习

http://forums.pentaho.com/showthread.php?59651-MD5-Hash Thread: MD5 Hash

http://wiki.pentaho.com/display/EAI/PDI+Spoon+Plugin+Development PDI Spoon Plugin Development

FREETALK

谢谢观看

制作

By-耿浩然

18687117526