网络流量解析器

总体思路

基于Wireshark的二次开发

一、需求

1、输入要求：

（1）离线的pcap文件（wireshark支持）

（2）实时接入网络数据流（wireshark支持）

（3）非标准pcap文件，包头添加其他信息

**需求分析**：需要对wireshark的pcap解析逻辑进行修改，支持新类型的文件解析，从而支持自定义的包头信息，在open\_info\_base结构体数组中增加新的协议解析类型，并注册协议入口函数，如libpcap\_open，在入口函数中注册协议操作函数，例如read函数和close函数，需要自行实现

**需求问题**：

非标准的pcap文件具体会是什么格式（pcap头增加固定长度）

**需求难点**：

* 修改数据结构，可能会造成兼容性的问题
* 对于新协议的匹配不止open\_info\_base结构体，还有其他地方需要修改
* 需要实现自定义的函数，则需要仿造出open，read，close函数，原函数结构较为复杂，仿写一个难度高

2、输出要求：

（1）支持协议全字段解析结果保存为文本文件（应用层全字段，五元组，固定格式用|分割，最高优先级，需要存储解析的文件路径）

**需求分析**：通过wireshark解析出每个字段内容

**需求问题**：文本文件格式类型是什么，具体内容包括什么，和json格式输出有什么区别

**需求难点**：

* + 拿到协议的全字段较为困难，wireshark对于协议中不同层的协议字段是放在不同的解析函数中，要追踪到全部的解析函数
  + 需要输出到文件中，则还需要增加文件的操作的相关模块，从而需要修改wireshark的编译体系
  + 是对于一个包一次性写入还是分开写入，全局变量的使用存在问题

（2）支持全字段解析结果输出到ES数据库或其他数据库（参考猫头鹰工具，arkime）

**需求分析**：解析出的每个字段，将其存入数据库中

**需求问题**：具体需要使用什么数据库

**需求难点**：ES数据库没有找到C语言接口

（3）解析的每个协议需有详细的字段说明文件（根据rfc规范，设计协议说明文档，对每个字段做出详细说明）

（4）支持json格式的输出文件（将文本文件内容作为json格式字典输出）

3、开发语言和平台

（1）C语言及C++

（2）Linux平台（基于libpcap进行抓包解析）（优先）

（3）Windows平台（基于winpcap进行解析，wireshark的源码同时支持两个平台）

**需求分析**：在双平台实现所有需求

**需求难点**：双平台下某些库的接口是不同的，导致代码需要作出调整

4、其他需求：

（1）单机部署解析速度能达到1000Mbit/s（需要对wireshark的解析速率进行测试，删除不必要功能，提高其解析速率）（最低100M）

（2）支持分布式部署

**需求分析**：基于DNS流量分发完成分布式部署

**需求难点**：由于wireshark原本是没有分布式框架的，改成分布式框架需要修改整体的框架，没有具体的思路实现

（3）支持用户增量开发，可扩充更多的协议

**需求分析**：wireshark支持新协议扩充，可以用C语言或lua进行协议扩充，对其接口进行利用封装为我们的接口，再封装成对于json的解析接口，通过对json文件的解析，再实现协议的扩展（刚需，最高优先级）

**需求问题**：扩展协议的json文件格式的定义

**需求难点**：扩充协议的C语言实现方案代码量大，再扩充json接口代码量也很大

（4）支持离线安装（wireshark满足）

（5）具备会话还原，会话重组功能，支持邮件协议还原原始邮件完整内容，支持基于HTTP协议还原HTML页面

**需求分析**：wireshark可以还原重组tcp会话；支持邮件还原功能，需要对其功能函数进行调用；HTML还原需要对HTTP协议进行解析，拿到负载内容，进行重组得到html文件

**需求问题**：会话还原和重组是否基于TCP协议的连接

**需求难点**：需要找到wireshark对于连接还原重组功能的支持模块，邮件还原的功能模块，并且对于HTTP协议的解析模块进行研究，找到对于html的解析模块，解析出数据

（6）协议识别的规则支持配置文件或者其他方式的动态加载（通过脚本调用实现）

**需求分析**：wireshark支持协议解析规则的修改，需要拿到其接口进行修改

**需求问题**：配置文件的格式类型，其他方式指什么

**需求难点**：需要找到wireshark对协议识别的模块，对其识别规则进行修改

（7）解析时允许用户添加额外字段（暂时不需要）

**需求分析**：需要修改对于数据包解析的模块，支持扩充额外字段，可以让用户进行手动添加

**需求问题**：需要在哪个位置增加额外字段

**需求难点**：

* 熟悉wireshark的数据解析模块
* 扩充额外字段可能会让原有的解析受到影响
* 用户能够手动添加则需要增加接口，代码量大

（8）对错误的pcap文件解析具有一定的容错能力（出现异常情况能够继续运行）

**需求问题**：需要对pcap文件提供哪种容错能力，也就是说支持哪种错误类型

（9）代码需有相对详细的注释说明（对wireshark代码模块进行注释说明，新增代码也需要注释）

（10）每个协议的解析程序需要注明依据的RFC文档编号及相关网址（对每个协议解析都是基于RFC标准文档完成，增加注明）

（11）支持各大主流语种，避免乱码和无法解码的情况

**需求分析**：wireshark支持多种解码格式，可以根据需求进行扩充

**需求问题**：需要用哪些编码格式

（12）框架有完善的文档说明

**需求分析**：完善文档撰写

**需求问题**：是全部的代码文档，还是新增功能的相关文档

**需求难点**：wireshark的相关代码文档不够完善，如果需要补充它的文档，那很难能够吧所有功能说清楚

（13）包含至少15种主流协议的解析程序

**需求分析**：wireshark支持基本全部主流协议的解析工作，可以直接沿用，并加以修改

**需求难点**：15种以上的协议解析工作，工作量很大，对代码的阅读也很复杂

（14）每一个协议的解析程序中，字段的命名尽量与RFC参考文档中的字段说明保持一致，便于阅读（代码中的变量和字段名符合RFC规范）

（15）集成wireshark的协议类型识别规则库

**需求分析**：wireshark协议类型识别解析，协议过滤功能

**需求问题**：wireshark有自己的规则识别库，是否需要新增功能或修改

（16）离线pcap文件的解析支持增量解析（基于文件夹做pcap文件的解析，记录解析位置，增加日志系统，一个pcap文件解析出问题，可以记录当前指针继续解析）

**需求分析**：对解析模块进行修改，使其可以同时解析多个pcap文件，从而支持增量解析

**需求问题**：增量解析是否是我们理解的意思

**需求难点**：暂时没有具体思路

二、开发周期

6.1-6.30 预计开发出一个Demo，覆盖基本需求

7.1-7.31 对Demo内容进行增补，覆盖全部需求

8.1-8.31 对完成内容进行补充和bug的修改，进行文档的编写，完成全部内容

三、技术路线

分析wireshark源代码，自定义界面并集成wireshark的抓包和协议解析部分；

通过修改源代码的各个协议解析模块完成对解析功能的扩展；

通过wireshark的挂钩机制实现对自定义协议解析的扩展；

分布式部署，wireshark不提供分布式部署框架，需要新增分布式调度器，部署多个解析节点协同完成解析。

四、项目难点

1、内容还原， wireshark界面上已经集成了内容还原功能，但项目组自行集成需要抛弃原有界面，需要分析wireshark代码本身的还原部分代码，大概率需要重新开发；

2、需求要求的协议解析模块工作量大，需要集成wireshark原来的解析模块，有大量的源码分析和调试工作，寻找协议解析功能函数并进行部分集成修改。