小象学院

环境搭建

主　讲: 杨真

小象学院

Part 1 基础

- 环境搭建
- HTML 基础
- 第一个10行代码的爬虫
- 内容抽取及解析
- HTTP 协议
- POSTMAN 工具详解

Part 2 爬虫

- 网站结构分析
- 抓取方案
- 多线程并行及排重
- 用 MySQL 信息存储

Part 3 进阶

- 网站服务结构
- Cookie 及 登录处理
- 控制抓取的节奏
- 日志
- 守护进程

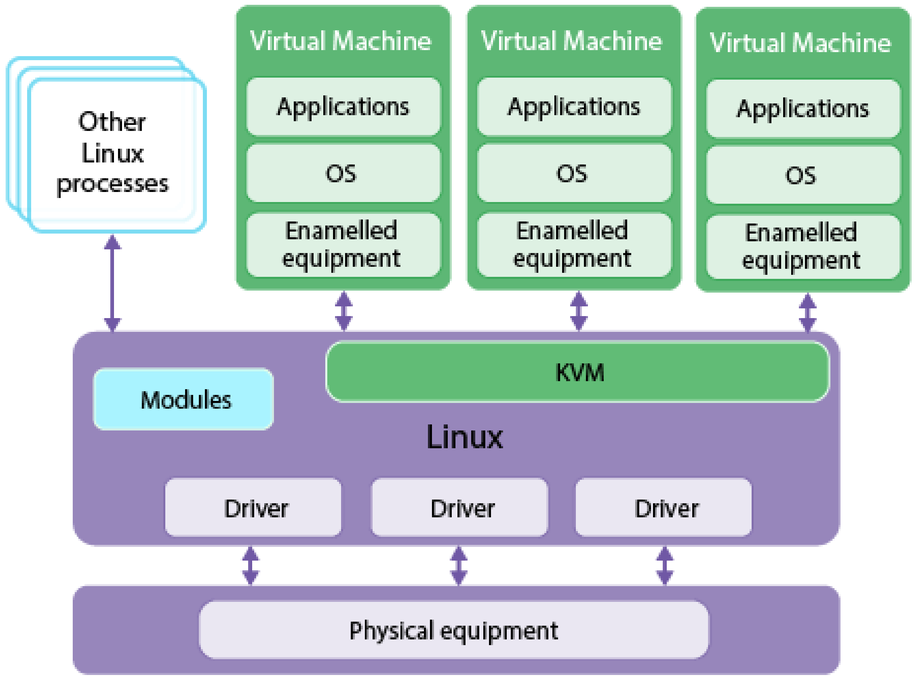
Part 4 实战

- 网站结构分析
- 网页抓取方案
- 数据提取
- 存储方案

- Linux 虚拟机安装
- 常用 Linux 命令
- Linux 环境搭建 – Python

小象学院

Ubuntu 虚拟机安装

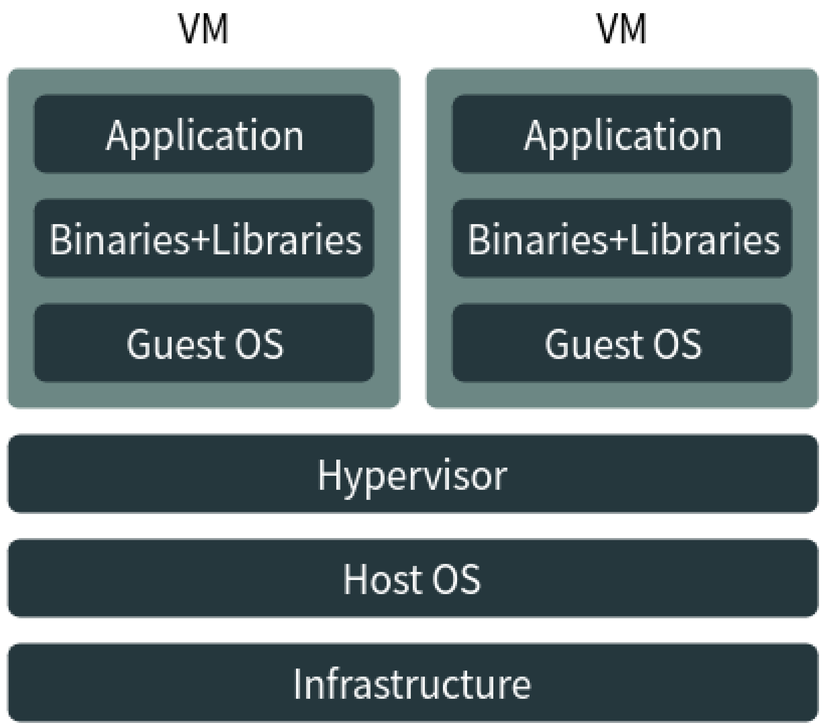


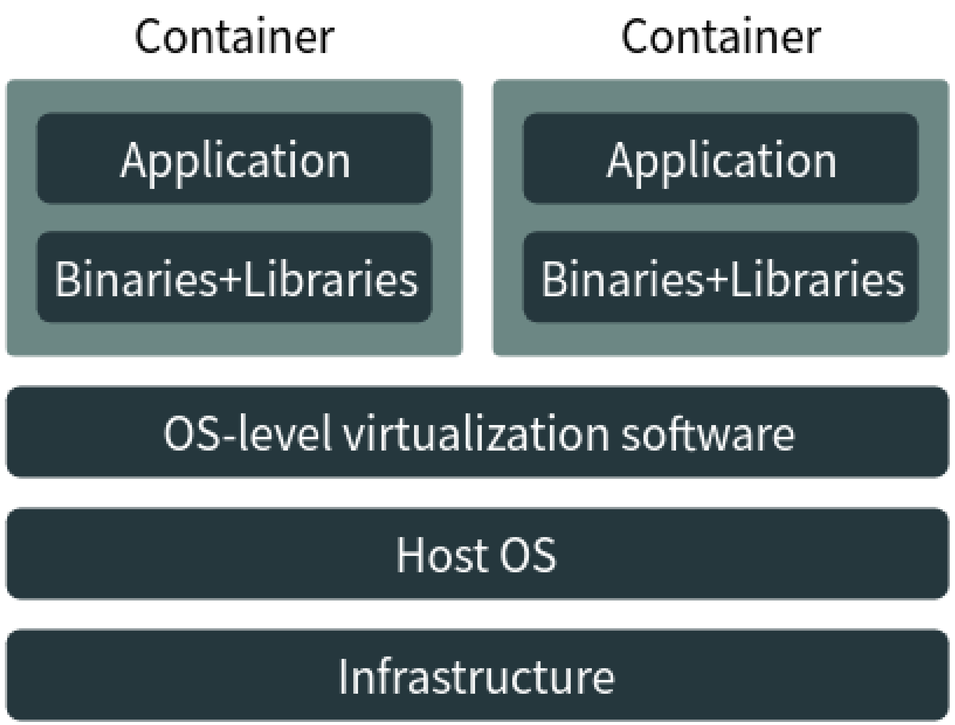
VirtualBox5.2.2:

https://www.virtualbox.org/wiki/Downloads

Ubuntu VM Image:

https://pan.baidu.com/s/1qSTZ_rGe7AcJJo7vlVTp1w


VM
VM
Application
Application
Binaries+Libraries
Binaries+Libraries
Guest OS
Guest OS
Hypervisor
Host OS
Infrastructure


Container
Container
Application
Application
Binaries+Libraries
Binaries+Libraries
OS-level virtualization software
Host OS
Infrastructure

Ubuntu 命令

apt get install/remove/search

rm

mv

find

vim (i,h,j,k,l,w,q,u,d)

ls

mkdir

ps

kill

netstat

grep

top

sudo

小象学院

Python 环境

sudo apt-get install python3.6
sudo apt install python3-venv python3-pip

课程代码：

https://github.com/suneri/junior_spider.git

替换现有上游

cd "$(brew --repo)"

git remote set-url origin https://mirrors.tuna.tsinghua.edu.cn/git/homebrew/brew.git
cd "$(brew --repo)/Library/Taps/homebrew/homebrew-core"
git remote set-url origin https://mirrors.tuna.tsinghua.edu.cn/git/homebrew/homebrew-core.git
brew update

使用homebrew-science或者homebrew-python

cd "$(brew --repo)/Library/Taps/homebrew/homebrew-science”

git remote set-url origin https://mirrors.tuna.tsinghua.edu.cn/git/homebrew/homebrew-science.git
cd "$(brew --repo)/Library/Taps/homebrew/homebrew-python"
git remote set-url origin https://mirrors.tuna.tsinghua.edu.cn/git/homebrew/homebrew-python.git
brew update

vim ~/.config/pip/pip.conf

[global]
index-url = https://pypi.tuna.tsinghua.edu.cn/simple

安装全部依赖的库：

#pip install -r requirements.txt

小象学院

只为遇见明天更优秀的你！