中科院计算培训中心

Part 9 云环境下大数据挖掘应用

各行各业的应用

- 杨文川

1）与Docker等工具配合
2）大数据挖掘行业应用
3）大数据挖掘展望

资源虚拟化工具Docker

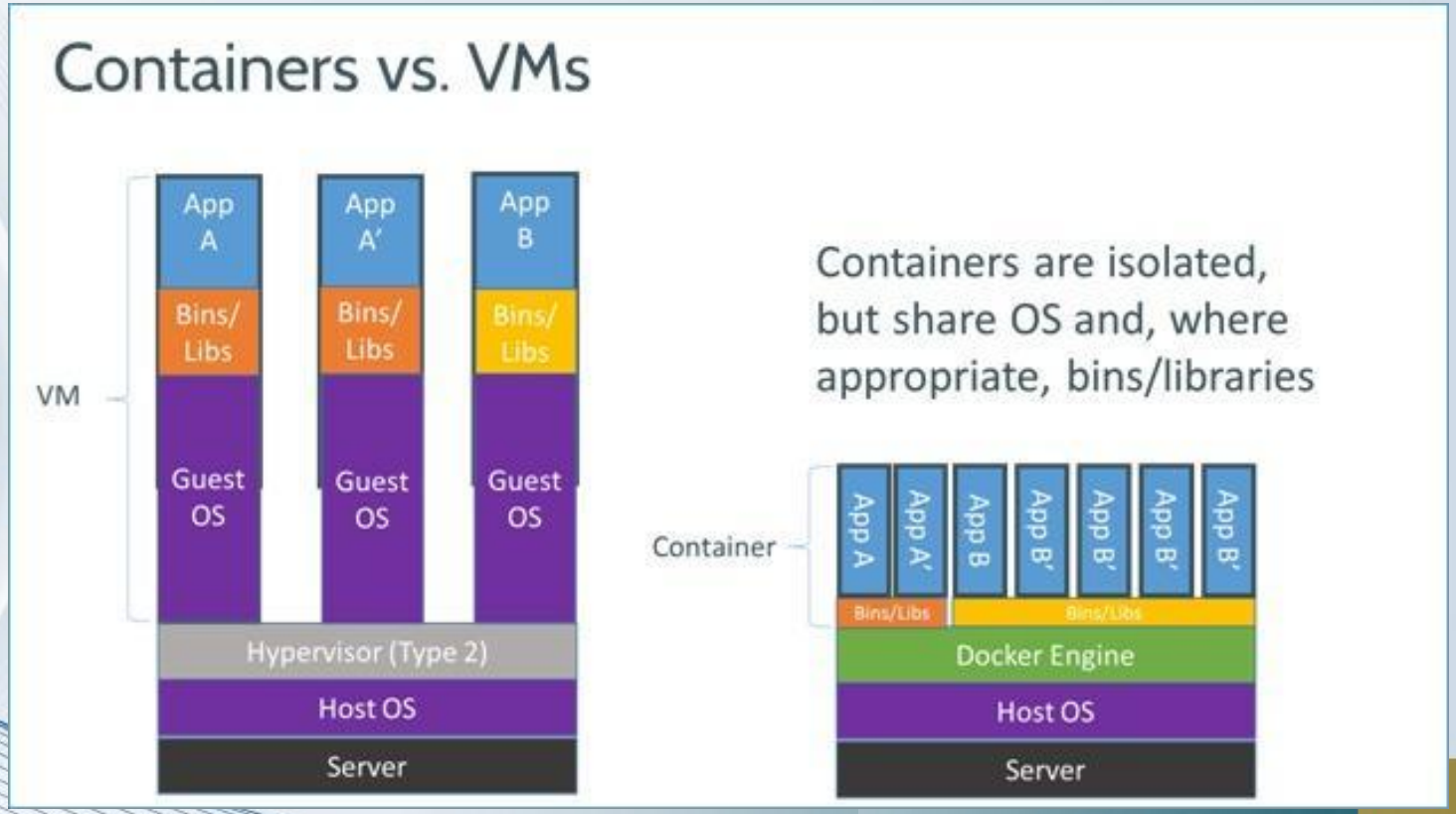
- Docker 是一个开源项目，诞生于 2013 年初，最初是 dotCloud 公司内部的一个业余项目。
  - 它基于 Google 公司推出的 Go 语言实现。 项目后来加入了 Linux 基金会，遵从了 Apache 2.0 协议，项目代码在 GitHub上进行维护。
- Docker 自开源后受到广泛的关注，甚至dotCloud 公司后来都改名为 Docker Inc。Redhat 已经在其 RHEL6.5 中集中支持 Docker；Google 也在其 PaaS 产品中广泛应用。
- Docker 项目的目标，是实现轻量级的操作系统虚拟化解决方案。

LXC技术

- Docker 的基础是 Linux 容器（LXC）等技术。
- 在 LXC 的基础上 Docker 进行了进一步的封装，让用户不需要去关心容器的管理，使得操作更为简便。
- 用户操作 Docker 的容器，就像操作一个快速轻量级的虚拟机一样简单。

Docker容器和虚拟机区别示意图

- 容器是在操作系统层面上实现虚拟化，直接复用本地主机的操作系统
- 而传统方式则是在硬件层面实现。



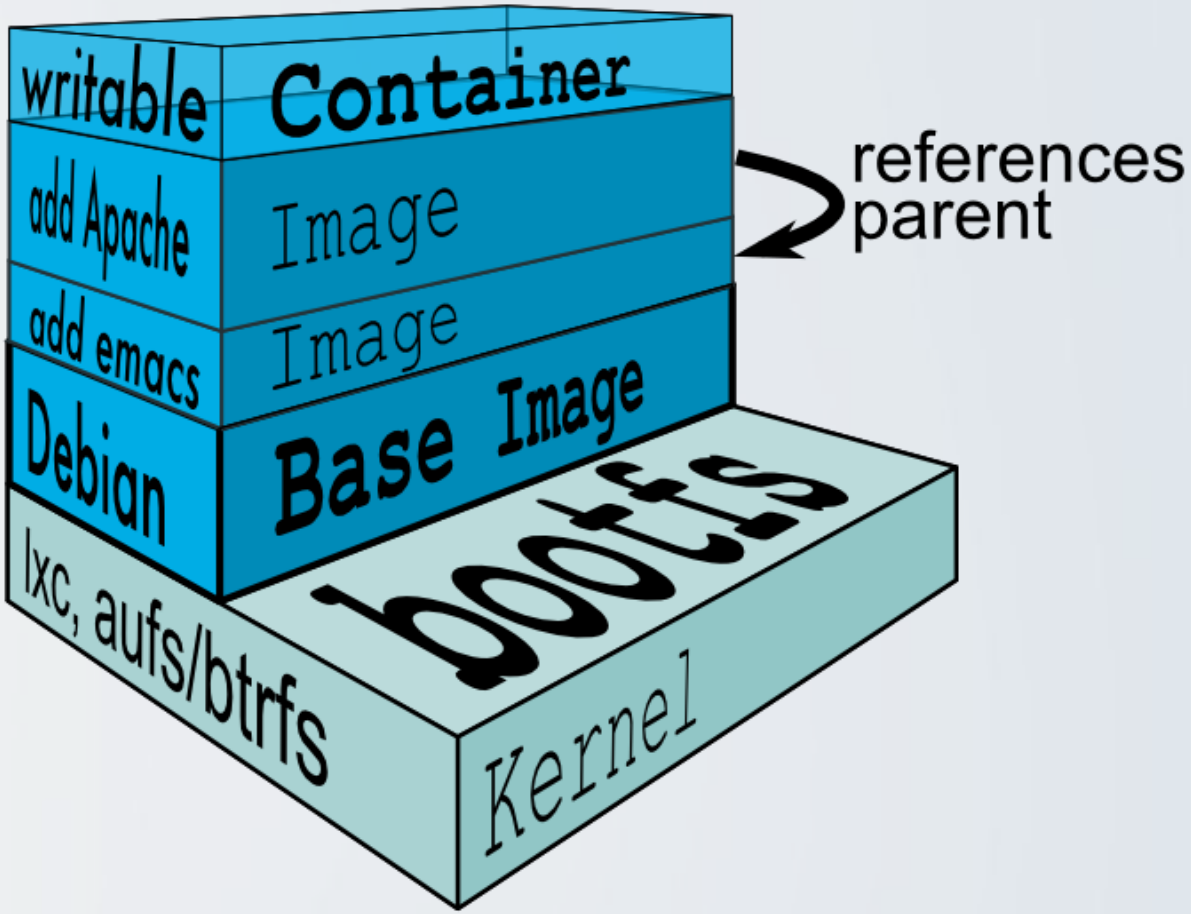
基本概念

- Docker 包括三个基本概念
- 镜像（Image）
- 容器（Container）
- 仓库（Repository）
- 理解了这三个概念，就理解了 Docker 的整个生命周期。

Docker原理

- Docker会在隔离的容器中运行进程。
  - 当运行docker run命令时，Docker会启动一个进程，并为这个进程分配其独占的文件系统、网络资源和以此进程为根进程的进程组。
- 在容器启动时，镜像可能已经定义了要运行的二进制文件、暴露的网络端口等，但是用户可以通过docker run命令重新定义
  - docker run可以控制一个容器运行时的行为，可以覆盖docker build在构建镜像时的一些默认配置。

Docker架构


writable
Container
add Apache
Image
add emacs
Image
Debian
Base Image
bootfs
lxc, aufs/btrfs
Kernel
references
parent

为什么要使用 Docker？

- 作为一种新兴的虚拟化方式，Docker 跟传统的虚拟化方式相比具有众多的优势。
  - 首先，Docker 容器的启动可以在秒级实现，这相比传统的虚拟机方式要快得多。
  - 其次，Docker 对系统资源的利用率很高，一台主机上可以同时运行数千个 Docker 容器。

对比传统虚拟机总结

特性	容器	虚拟机
启动	秒级	分钟级
硬盘使用	一般为 MB	一般为 GB
性能	接近原生	弱于
系统支持量	单机支持上千个容器	一般几十个

大数据挖掘应用

- 数据已经成为了一种优质商业资本，一项重要的经济投入，可以创造新的经济利益。
  - 传统行业结合大数据可以发现无限的商机，能够预测未来。

向用户推荐电影

时间	主机	消息
Jun 20 04:50:36	Vigor	Local User: 192.168.1.38:3113 -> 211.144.92.240:80 (TCP) close connection
Jun 20 04:50:36	Vigor	Local User: 192.168.1.38:3119 -> 211.144.92.242:80 (TCP)Web
Jun 20 04:50:36	Vigor	Local User: 192.168.1.38:3112 -> 211.144.92.240:80 (TCP) close connection
Jun 20 04:50:36	Vigor	Local User: 192.168.1.38:3118 -> 211.144.92.241:80 (TCP)Web
Jun 20 04:50:36	Vigor	Local User: 192.168.1.38:3116 -> 211.144.92.243:80 (TCP)Web
Jun 20 04:50:35	Vigor	Local User: 192.168.1.38:3113 -> 211.144.92.240:80 (TCP)Web
Jun 20 04:50:35	Vigor	Local User: 192.168.1.38:3112 -> 211.144.92.240:80 (TCP)Web
Jun 20 04:50:35	Vigor	Local User: 192.168.1.38:3062 -> 211.144.92.241:80 (TCP) close connection
Jun 20 04:50:35	Vigor	Local User: 192.168.1.38:3109 -> 219.142.78.61:80 (TCP)Web
Jun 20 04:50:34	Vigor	Local User: 192.168.1.38:3106 -> 211.144.92.240:80 (TCP) close connection
Jun 20 04:50:33	Vigor	Local User: 192.168.1.38:3106 -> 211.144.92.240:80 (TCP)Web
Jun 20 04:50:31	Vigor	Local User: 192.168.1.38:3104 -> 211.144.92.240:80 (TCP) close connection
Jun 20 04:50:31	Vigor	Local User: 192.168.1.38:3104 -> 211.144.92.240:80 (TCP)Web
Jun 20 04:50:31	Vigor	Local User: 192.168.1.38:3102 -> 211.144.92.240:80 (TCP) close connection
Jun 20 04:50:31	Vigor	Local User: 192.168.1.38:3102 -> 211.144.92.240:80 (TCP)Web
Jun 20 04:50:31	Vigor	Local User: 192.168.1.38:3100 -> 211.144.92.240:80 (TCP) close connection
Jun 20 04:50:30	Vigor	Local User: 192.168.1.38:3061 -> 211.144.92.242:80 (TCP) close connection
Jun 20 04:50:30	Vigor	Local User: 192.168.1.38:3092 -> 211.144.92.240:80 (TCP) close connection
Jun 20 04:50:30	Vigor	Local User: 192.168.1.38:3095 -> 211.144.92.240:80 (TCP) close connection
Jun 20 04:50:28	Vigor	Local User: 192.168.1.38:3090 -> 218.30.66.7:80 (TCP) close connection
Jun 20 04:50:28	Vigor	Local User: 192.168.1.38:3100 -> 211.144.92.240:80 (TCP)Web

NETFLIX

Netflix每天会对2700万和3600万注册用户的

3000万次“动作”（包括播放、暂停、倒退和快进等动作）、

400万次评级、

300万次搜索，

用户观看视频的时间和设备进行分析

- **挖掘用户的喜好，向用户推荐节目**

将文字进行分类

- Yahoo! Mail决定接收的信息是不是垃圾邮件，
  - 基于先前邮件和用户的垃圾邮件报告，以及邮件的特性。
  - 一些信息被分类为垃圾邮件

聚类分析刷卡数据

- **Cardlytics：解析你的刷卡足迹**
  - “她去了McDonald，然后Target，再是Babies R Us”可能是一个已婚妈妈
  - “…他的消费地点都是酒吧和Taco Bell”这可能是个单身汉
- 刷卡数据经过聚类处理为多类人群，然后可以提供给Groupon，银行，商家等。

PriceStats

- MIT的一个名为PriceStats的大数据方案，通过一个软件在互联网上收集信息，每天可以收集到50万种商品的价格。
  - 收集到的数据很混乱，也不是所有数据都能轻易比较。
  - 但是把大数据和数据挖掘方法相结合，该项目可以实时发现通货紧缩趋势

收集到的价格信息多，而且是即时的，目前被70多个国家的银行和经济决策人用到。

预测警务

基于现有的视频、安保等海量大数据，采用数据挖掘方法，实现“预测警务”决定哪些街道、群体需要更严密的监控。

- 美国的Blue CRUSH项目为警员提供情报，关于哪些地方更容易发生犯罪事件，什么时候更容易逮到罪犯。

大数据挖掘可以帮助执法部门更好地分配其有限的资源。

预测传染病

采用大数据挖掘方法，处理海量网络搜索记录。

- 例如：谷歌针对来自全球超过30亿条的搜索指令，采用上亿个挖掘模型进行处理，发现了45条检索词条的组合，准确的提前预测了H7N9。

个性化售书推荐

Amazon通过网上售书，从每一个客户身上捕获了大量的数据。并基于这些数据实现个性化售书推荐

- 例如：他们购买了什么书籍？哪些书他们只浏览却没有购买？他们浏览了多久？哪些书是他们一起购买的？

大数据挖掘总结

- **大数据挖掘为探索未知世界提供了一个工具！**
- 数据已经成为了一种优质商业资本，
  - 传统行业结合大数据和大数据挖掘技术，可以发现无限的商机，能够预测未来，激发新产品和新型服务，创造新的经济利益。

谢　谢