中国电子教育学会高教分会推荐·大数据系列教材
高等学校新工科应用型人才培养“十三五”规划教材
国信蓝桥教育科技（北京）股份有限公司推荐教材

Experiment Tutorial for
Principles and Applications of
Hadoop Big Data Technology

Hadoop大数据原理与应用实验教程

主　编　徐鲁辉
副主编　周湘贞　李月军
主　审　唐友刚

西安电子科技大学出版社
http://www.xduph.com

XDUP 584500
封面设计：佳易传播

Hadoop大数据原理与应用实验教程

内容简介

本书作为《Hadoop大数据原理与应用》教材的配套实验教程，全面介绍了Hadoop生态系统中各个开源组件的相关知识和实践技能。全书分为“基础实验篇”和“拓展实验篇”两篇，共10章，涉及数据采集、数据存储与管理、数据处理与分析等大数据应用生命周期中各阶段典型组件的部署、使用和基础编程方法。“基础实验篇”内容包括部署全分布模式Hadoop集群、实战HDFS、MapReduce编程、部署ZooKeeper集群和实战ZooKeeper、部署全分布模式HBase集群和实战HBase、部署本地模式Hive和实战Hive，“拓展实验篇”内容包括部署Spark集群和Spark编程、实战Sqoop、实战Flume、实战Kafka。

相关资源可在西安电子科技大学出版社网站下载

ISBN 978-7-5606-5543-7
9 787560 655437 >
定价：49.00元

中国电子教育学会高教分会推荐·大数据系列教材
高等学校新工科应用型人才培养“十三五”规划教材
国信蓝桥教育科技(北京)股份有限公司推荐教材

Hadoop 大数据原理与应用实验教程

Experiment Tutorial for Principles and Applications of Hadoop Big Data Technology

主　编　徐鲁辉
副主编　周湘贞　李月军
主　审　唐友刚

西安电子科技大学出版社

内 容 简 介

本书作为《Hadoop 大数据原理与应用》(本书作者编写，西安电子科技大学出版社出版)的配套实验教程，系统介绍了 Hadoop 生态系统中各个开源组件的相关知识和实践技能。全书分为“基础实验篇”和“拓展实验篇”两篇，共 10 章，涉及数据采集、数据存储与管理、数据处理与分析等大数据应用生命周期中各阶段典型组件的部署、使用和基础编程方法。“基础实验篇”内容包括部署全分布模式 Hadoop 集群、实战 HDFS、MapReduce 编程、部署 ZooKeeper 集群和实战 ZooKeeper、部署全分布模式 HBase 集群和实战 HBase、部署本地模式 Hive 和实战 Hive；“拓展实验篇”内容包括部署 Spark 集群和 Spark 编程、实战 Sqoop、实战 Flume、实战 Kafka。

本书内容翔实，案例丰富，操作过程详尽，并配有完整的立体化资源，既可作为高等院校研究生、本科生的大数据技术原理与应用课程的实验指导书，也可作为教师参考书，同时也可供相关技术人员参考。(相关资源可在西安电子科技大学出版社网站下载。)

图书在版编目(CIP)数据

Hadoop 大数据原理与应用实验教程 / 徐鲁辉主编. —西安：西安电子科技大学出版社，2020.1
ISBN 978-7-5606-5543-7

Ⅰ. ① H… Ⅱ. ① 徐… Ⅲ. ① 数据处理软件—教材 Ⅳ. ① TP274

中国版本图书馆 CIP 数据核字(2019)第 289448 号

策划编辑 李惠萍
责任编辑 唐小玉
出版发行 西安电子科技大学出版社(西安市太白南路 2 号)
电 话 (029)88242885 88201467 邮 编 710071
网 址 www.xduph.com 电子邮箱 xdupfxb001@163.com
经 销 新华书店
印刷单位 陕西天意印务有限责任公司
版 次 2020 年 1 月第 1 版 2020 年 1 月第 1 次印刷
开 本 787 毫米×1092 毫米 1/16 印张 21.5
字 数 508 千字
印 数 1～3000 册
定 价 49.00 元
ISBN 978-7-5606-5543-7 / TP
XDUP 5845001-1

前　言

大数据时代的到来，带来了信息技术发展的巨大变革，并深刻影响着社会生产和人民生活的方方面面。全球范围内，世界各国政府均高度重视大数据技术的研究与产业发展，纷纷把大数据上升为国家战略加以重点推进。大数据已经成为企业和社会关注的重要战略资源，越来越多的行业面临着海量数据存储和分析的挑战。

Hadoop 由道格·卡丁(Doug Cutting)创建，起源于开源项目网络搜索引擎 Apache Nutch，于 2008 年 1 月成为 Apache 顶级项目。Hadoop 是一个开源的、可运行于大规模集群上的分布式存储和计算的软件框架，它具有高可靠、弹性可扩展等特点，非常适合处理海量数据。Hadoop 实现了分布式文件系统 HDFS 和分布式计算框架 MapReduce 等功能，允许用户可以在不了解分布式系统底层细节的情况下，使用简单的编程模型轻松编写出分布式程序，将其运行于计算机集群上，完成对大规模数据集的存储和分析。目前，Hadoop 在业内得到了广泛应用，已经是公认的大数据通用存储和分析平台，许多厂商都围绕 Hadoop 提供开发工具、开源软件、商业化工具和技术服务，例如谷歌、雅虎、微软、淘宝等都支持 Hadoop。另外，还有一些专注于 Hadoop 的公司，例如 Cloudera、Hortonworks 和 MapR 都可以提供商业化的 Hadoop 支持。

未来 5～10 年，我国大数据产业将会处于高速发展时期，社会亟需高校培养一大批大数据相关专业人才。自 2016 年以来，我国新增的大数据类专业包括“数据科学与大数据技术”本科专业(080910T)、“数据管理与应用”本科专业(120108T)、“大数据技术与应用”专科专业(610215)，以适应地方产业发展对战略性新兴产业的人才需求。因此，学会使用大数据通用存储和分析平台 Hadoop 及其生态系统对于未来适应新一代信息技术产业的发展具有重要的意义。

实践教学是高等院校知识创新和人才培养的重要环节，因此，实验实训类教材建设在学生能力培养中发挥着不可或缺的重要作用。本书面向 Hadoop 生态系统，以企业需求为导向，紧紧围绕大数据应用的闭环流程展开讲述，引导学生进行大数据技术的初级实践，旨在使读者掌握 Hadoop 的架构设计和 Hadoop 的运用能力。

本书分为上篇“基础实验篇”和下篇“拓展实验篇”，共 10 章，涉及数据采集、数据存储与管理、数据处理与分析等大数据应用生命周期中各阶段典型组件的部署、使用和基础编程方法。在“基础实验篇”中，实验 1 介绍了 Linux 基本命令、vim 编辑器、Java 基本命令、SSH 安全通信协议、Hadoop 基础知识等先修技能，然后详细讲述了部署全分布模式 Hadoop 集群的全过程，并附加了伪分布模式 Hadoop 集群的部署过程；实验 2 介绍了分布式文件系统 HDFS 的体系架构、文件存储原理、接口等基础知识，详细演示了如何通过 HDFS Web UI 和 HDFS Shell 命令使用 HDFS 以及 HDFS Java API 编程，并附加了如

何搭建 HDFS NameNode HA 环境；实验 3 在介绍了分布式计算框架 MapReduce 编程思想、作业执行流程的基础上，讲述了如何编写 MapReduce 程序，并附加了在 Windows 平台上开发 MapReduce 程序和使用 MapReduce 统计对象中某些属性的案例；实验 4 介绍了分布式协调框架 ZooKeeper 的系统模型、工作原理等基本知识，详细演示了如何部署 ZooKeeper 集群以及通过 ZooKeeper Shell 命令使用 ZooKeeper，并附加了 ZooKeeper 编程实践；实验 5 介绍了分布式数据库 HBase 的数据模型、体系架构、接口等基础知识，详细演示了如何部署全分布模式 HBase 集群以及通过 HBase Web UI 和 HBase Shell 使用 HBase，并附加了 HBase 编程实践；实验 6 介绍了数据仓库 Hive 的体系架构、数据类型、文件格式、数据模型、函数、接口等基础知识，详细演示了如何部署本地模式 Hive 以及通过 Hive Shell 使用 Hive，并附加了 Hive 编程实践。在“拓展实验篇”中，实验 7 介绍了内存型计算框架 Spark 的生态系统、体系架构、计算模型、RDD 原理等基础知识，详细演示了如何部署 Spark-Standalone 集群以及 Spark 简单编程；实验 8 介绍了数据迁移工具 Sqoop 的功能、体系架构、接口等基础知识，详细演示了如何安装 Sqoop 和使用 Sqoop Shell 完成数据的导入导出操作；实验 9 介绍了日志采集工具 Flume 的功能、体系架构、接口等基础知识，详细演示了如何安装、配置 Flume 以及使用 Flume 高效进行海量日志的收集、聚合和移动；实验 10 介绍了分布式流平台 Kafka 的功能、体系架构、Kafka Shell、Kafka API，详细演示了如何部署 Kafka 集群以及通过 Kafka Shell 使用 Kafka 进行生产和消费消息。

为了方便读者整体把握各个实验，本书在每个实验的一开始先给出该实验的知识地图。根据我们近几年的教学实践，建议根据本书为大数据技术原理与应用课程增加 16 学时的上机实践课，可根据具体情况灵活安排本书实验项目。

本书面向高等院校计算机、大数据、人工智能等相关专业的研究生、本科生，可以作为专业核心课程大数据技术原理与应用的辅助实验教材。本书是《Hadoop 大数据原理与应用》教材的配套实验教程，两本书配套使用，可以达到更好的学习效果。此外，本书也可以作为现有其他大数据教材的实验教材或辅助教材。

本书由校企联合完成，实验 1 由安徽信息工程学院李月军编写，实验 2 由郑州升达经贸管理学院周湘贞编写，实验 3 由国信蓝桥教育科技(北京)股份有限公司颜群工程师编写，实验 4～10 由西京学院徐鲁辉编写。此外，李月军和周湘贞还参与了本书全部架构设计和部分审阅工作。全书由国信蓝桥教育科技(北京)股份有限公司大数据专家唐友刚主审，由西京学院徐鲁辉负责策划、审校和定稿。

本书与配套教材《Hadoop 大数据原理与应用》拥有完整的立体化资源，包括教学大纲、授课计划、教案、PPT、源代码、在线题库、实验大纲、实验指导书、实验视频、项目案例库等教学资源，提供全方位的免费服务。读者可通过以下三种方式免费在线浏览或下载全部配套资源：教材官方网站 https://dzxxgcx.xijing.edu.cn/xkzy/zxkc.htm/pabigdata/index.asp，教材官方云班课“Hadoop 大数据原理与应用教材云班课”(邀请码 5962412)，教材官方 GitHub 网站 https://github.com/xuluhuixijing/pabigdata。

本书中关于图形界面元素代替符号的约定如表 1 所示。

表 1　本书图形界面元素代替符号约定

文字描述	代替符号	举　例
按钮	边框+阴影+底纹	“确定”按钮可简化为确定
菜单项	『　』	菜单项“文件”可简化为『File』
连续选择菜单项及子菜单项	→	选择『File』→『New』→『Java Project』
下拉框、单选框、复选框选项	[　]	复选框选项“启用用户”可简化为[启用用户]
窗口名	【　】	例如，进入窗口【Properties for HDFSExample】
提示信息	“　”	例如，否则会出现错误信息“bash: ****: command not found...”

本书中各实验所使用软件的名称、版本、发布日期及下载地址如表 2 所示。

表 2　本书使用软件的名称、版本、发布日期及下载地址

序号	软件名称	软件版本	发布日期	下载地址	安装文件名
1	VMware Workstation Pro	VMware Workstation 12.5.7 Pro for Windows	2017.6.22	https://www.vmware.com/products/workstation-pro.html	VMware-workstation-full-12.5.7-5813279.exe
2	CentOS	CentOS 7.6.1810	2018.11.26	https://www.centos.org/download/	CentOS-7-x86_64-DVD-1810.iso
3	Java	Oracle JDK 8u191	2018.10.16	http://www.oracle.com/technetwork/java/javase/downloads/index.html	jdk-8u191-linux-x64.tar.gz
4	Hadoop	Hadoop 2.9.2	2018.11.19	http://hadoop.apache.org/releases.html	hadoop-2.9.2.tar.gz
5	Eclipse	Eclipse IDE 2018-09 for Java evelopers	2018.9	https://www.eclipse.org/downloads/packages	eclipse-java-2018-09-linux-gtk-x86_64.tar.gz
6	ZooKeeper	ZooKeeper 3.4.13	2018.7.15	http://zookeeper.apache.org/releases.html	zookeeper-3.4.13.tar.gz
7	HBase	HBase 1.4.10	2019.6.10	https://hbase.apache.org/downloads.html	hbase-1.4.10-bin.tar.gz
8	MySQL Connector /J	MySQL Connector/J 5.1.48	2019.7.29	https://dev.mysql.com/downloads/connector/j/	mysql-connector-java-5.1.48.tar.gz

续表

序号	软件名称	软件版本	发布日期	下载地址	安装文件名
9	MySQL Community Server	MySQL Community 5.7.27	2019.7.22	http://dev.mysql.com/get/mysql57-community-release-el7-11.noarch.rpm	mysql57-community-release-el7-11.noarch.rpm(Yum Repository)
10	Hive	Hive 2.3.4	2018.11.7	https://hive.apache.org/downloads.html	apache-hive-2.3.4-bin.tar.gz
11	Spark	Spark 2.3.3	2019.2.15	https://spark.apache.org/downloads.html	spark-2.3.3-bin-haoop2.7.tgz
12	Sqoop	Sqoop 1.4.7	2017.12	http://www.apache.org/dyn/closer.lua/sqoop/	sqoop-1.4.7.bin_hadoop-2.6.0.tar.gz
13	Flume	Flume 1.9.0	2019.1.8	http://flume.apache.org/download.html	apache-flume-1.9.0-bin.tar.gz
14	Kafka	Kafka 2.1.1	2019.2.15	http://kafka.apache.org/downloads	kafka_2.12-2.1.1.tgz

本书在编写过程中得到了很多人的帮助。国信蓝桥教育科技(北京)股份有限公司高校合作部项目经理单宝军在教材编写方面提供了帮助，西京学院校长黄文准、西京学院信息工程学院院长郭建新、副院长乌伟在学院政策方面提供了支持，西安电子科技大学出版社李惠萍编辑对本书的出版提出了很多意见和建议，在此一并表示衷心感谢。

本书在撰写的过程中参考了部分国内外教材、专著、论文和开源社区资源，在此也向这些文献作者一并致谢。由于作者水平和能力有限，书中难免有疏漏与不足之处，衷心希望广大同行和读者批评指正。

徐鲁辉

2019 年 10 月于西安

目　录

上篇　基础实验篇

下篇 拓展实验篇

上篇　基础实验篇

- ❖ 实验 1　部署全分布模式 Hadoop 集群
- ❖ 实验 2　实战 HDFS
- ❖ 实验 3　MapReduce 编程
- ❖ 实验 4　部署 ZooKeeper 集群和实战 ZooKeeper
- ❖ 实验 5　部署全分布模式 HBase 集群和实战 HBase
- ❖ 实验 6　部署本地模式 Hive 和实战 Hive

实验 1　部署全分布模式 Hadoop 集群

本实验的知识结构图如图 1-1 所示(★表示重点，▶表示难点)。

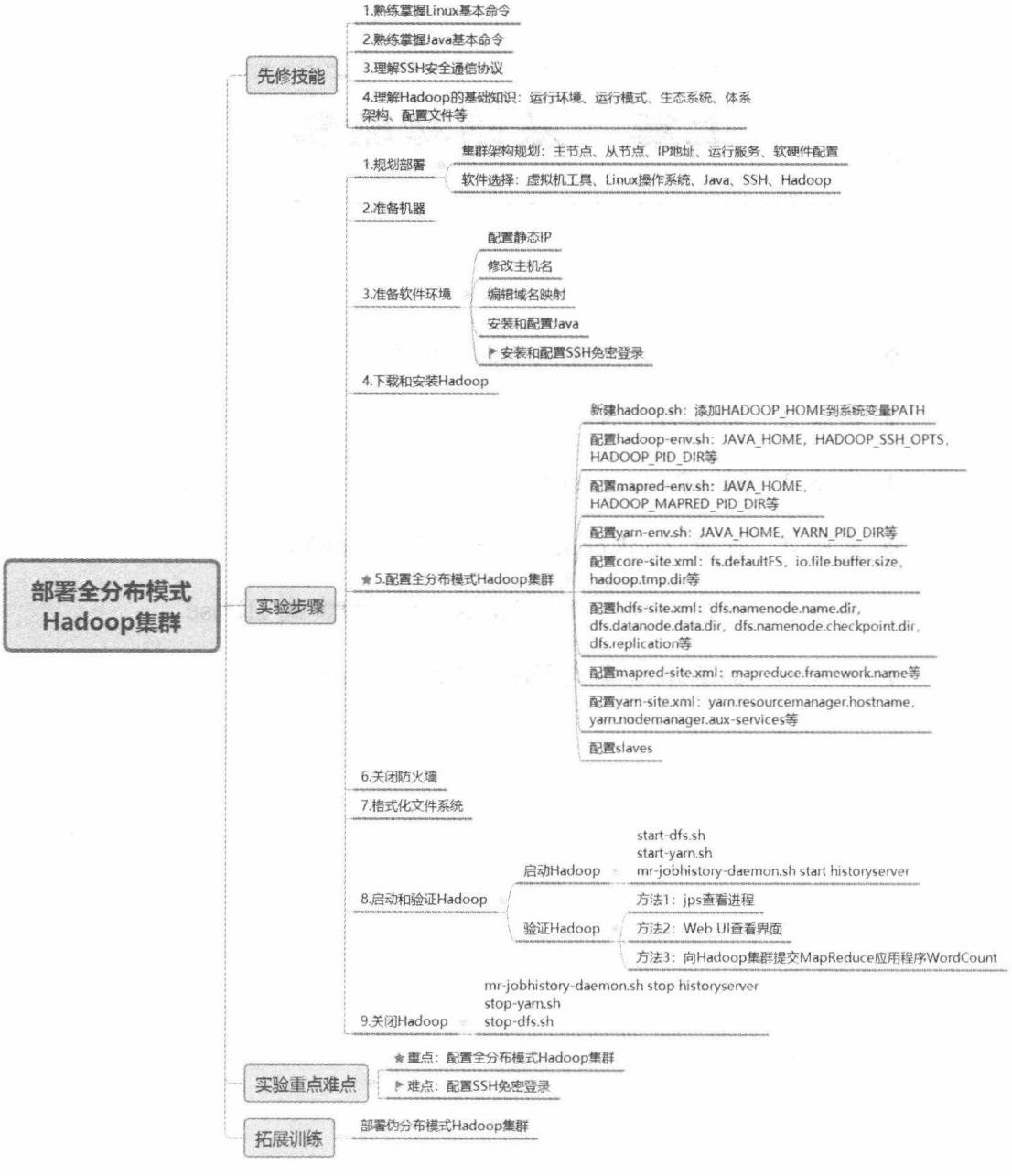

图 1-1　部署全分布模式 Hadoop 集群知识结构图

1.1　实验目的、实验环境和实验内容

一、实验目的

(1) 熟练掌握 Linux 基本命令。

(2) 掌握静态 IP 地址的配置及主机名和域名映射的修改。

(3) 掌握 Linux 环境下 Java 的安装、环境变量的配置、Java 基本命令的使用。

(4) 理解为何需要配置 SSH 免密登录，掌握 Linux 环境下 SSH 的安装、免密登录的配置。

(5) 熟练掌握在 Linux 环境下部署全分布模式 Hadoop 集群的过程。

二、实验环境

本实验所需的软硬件环境包括 PC、VMware Workstation Pro、CentOS 安装包、Oracle JDK 安装包、Hadoop 安装包。

三、实验内容

(1) 规划部署。

(2) 准备机器。

(3) 准备软件环境：配置静态 IP，修改主机名，编辑域名映射，安装和配置 Java，安装和配置 SSH 免密登录。

(4) 下载和安装 Hadoop。

(5) 配置全分布模式 Hadoop 集群。

(6) 关闭防火墙。

(7) 格式化文件系统。

(8) 启动和验证 Hadoop。

(9) 关闭 Hadoop。

1.2　实 验 原 理

1.2.1　Linux 基本命令

Linux 是一套免费使用和自由传播的类 UNIX 操作系统，是一个基于 POSIX 和 UNIX 的、多用户、多任务、支持多线程和多 CPU 的操作系统。它能运行主要的 UNIX 工具软件、应用程序和网络协议，支持 32 位和 64 位硬件。Linux 继承了 UNIX 以网络为核心的设计思想，是一个性能稳定的多用户网络操作系统。

Linux 操作系统诞生于 1991 年 10 月 5 日。Linux 存在着许多不同的版本，但它们都使用了 Linux 内核。Linux 可安装在各种计算机硬件设备中，比如手机、平板电脑、路由器、视频游戏控制台、台式计算机、大型机和超级计算机。

严格来讲，Linux 这个词本身只表示 Linux 内核，但实际上人们已经习惯用 Linux 来形容整个基于 Linux 内核且使用 GNU 工程各种工具和数据库的操作系统。

本节将介绍本实验中涉及的一些 Linux 操作系统的基本命令。

1. 查看当前目录

pwd 命令用于显示当前目录，效果如下所示：

```
[xuluhui@localhost ~]$ pwd
/home/xuluhui
```

2. 切换目录

cd 命令用来切换目录，效果如下所示：

```
[xuluhui@localhost ~]$ cd /usr/local
[xuluhui@localhost local]$ pwd
/usr/local
```

3. 罗列文件

ls 命令用于查看文件与目录，效果如下所示：

```
[xuluhui@localhost ~]$ ls
Desktop  Documents  Downloads  Music  Pictures  Public  Templates  Videos
```

4. 创建目录

mkdir 命令用于创建目录，效果如下所示：

```
[xuluhui@localhost ~]$ mkdir TestData
[xuluhui@localhost ~]$ ls
Desktop     Downloads  Pictures  Templates  Videos
Documents   Music      Public    TestData
```

5. 拷贝文件或目录

cp 命令用于拷贝文件。若拷贝的对象为目录，则需要使用-r 参数，效果如下所示：

```
[xuluhui@localhost ~]$ cp -r TestData TestData2
[xuluhui@localhost ~]$ ls
Desktop     Downloads  Pictures  Templates  TestData2
Documents   Music      Public    TestData   Videos
```

6. 移动或重命名文件或目录

mv 命令用于移动文件。在实际使用中，也常用于重命名文件或目录，效果如下所示：

```
[xuluhui@localhost ~]$ mv TestData2 TestDataxlh
[xuluhui@localhost ~]$ ls
Desktop     Downloads  Pictures  Templates  TestDataxlh
Documents   Music      Public    TestData   Videos
```

7. 删除文件或目录

rm 命令用于删除文件。若删除的对象为目录，则需要使用-r 参数，效果如下所示：

```
[xuluhui@localhost ~]$ rm -rf TestDataxlh
[xuluhui@localhost ~]$ ls
Desktop    Downloads  Pictures   Templates  Videos
Documents  Music      Public     TestData
```

8. 查看进程

ps 命令用于显示当前运行中进程的相关信息，效果如下所示：

```
[xuluhui@localhost ~]$ ps
    PID TTY      TIME CMD
  69780 pts/0    00:00:00 bash
  71680 pts/0    00:00:00 ps
```

9. 压缩与解压文件

tar 命令用于文件压缩与解压，参数中的 c 表示压缩，x 表示解压缩，效果如下所示：

```
[root@localhost local]# tar -zxvf /home/xuluhui/Downloads/hadoop-2.9.2.tar.gz
```

10. 查看文件内容

cat 命令用于查看文件内容，效果如下所示：

```
[xuluhui@localhost ~]# cat /usr/local/hadoop-2.9.2/etc/hadoop/core-site.xml
```

11. 查看机器 IP 配置

ip address 命令用于查看机器 IP 配置，效果如下所示：

```
[xuluhui@localhost ~]$ ip address
1: lo: <LOOPBACK,UP,LOWER_UP> mtu 65536 qdisc noqueue state UNKNOWN group
                                default qlen 1000
    link/loopback 00:00:00:00:00:00 brd 00:00:00:00:00:00
    inet 127.0.0.1/8 scope host lo
       valid_lft forever preferred_lft forever
    inet6 ::1/128 scope host
       valid_lft forever preferred_lft forever
2: ens33: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc pfifo_fast state UP
                                group default qlen 1000
    link/ether 00:0c:29:6d:5d:c9 brd ff:ff:ff:ff:ff:ff
    inet 192.168.18.128/24 brd 192.168.18.255 scope global noprefixroute dynamic ens33
       valid_lft 1795sec preferred_lft 1795sec
    inet6 fe80::6bb8:6e80:d029:10f2/64 scope link noprefixroute
       valid_lft forever preferred_lft forever
3: virbr0: <NO-CARRIER,BROADCAST,MULTICAST,UP> mtu 1500 qdisc noqueue state DOWN
                                group default qlen 1000
    link/ether 52:54:00:0b:74:1b brd ff:ff:ff:ff:ff:ff
    inet 192.168.122.1/24 brd 192.168.122.255 scope global virbr0
```