初识大数据

➢ 课前准备，什么是大数据

- 大数据（BIG DATA），指无法在一定时间范围内用常规软件工具进行捕捉、管理和处理的数据集合，是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产
- 1Byte = 8 bit、1K=1024KByte、1MB = 1024K、1G = 1024MB
- 1T = 1024G、1PB = 1024TB

➢ 大数据的特征

- 容量（Volume）：数据的大小决定所考虑的数据的价值和潜在的信息；
- 种类（Variety）：数据类型的多样性；
- 速度（Velocity）：指获得数据的速度；
- 可变性（Variability）：妨碍了处理和有效地管理数据的过程。
- 真实性（Veracity）：数据的质量
- 复杂性（Complexity）：数据量巨大，来源多渠道
- 价值（value）：合理运用大数据，以低成本创造高价值

➢ 学习的路线和课程概述

- JAVA =====> 面向对象编程语言
- Linux =====> 类 Unix 操作系统
- Hadoop 生态
  - HDFS =====> 解决存储问题
  - MapReduce =====> 解决计算问题
  - Yarn =====> 资源协调者
  - Zookeeper =====> 分布式应用程序协调服务
  - Flume =====> 日志收集系统
  - Hive =====> 基于 Hadoop 的数仓工具
  - HBase =====> 分布式、面向列的开源数据库
  - Sqoop =====> 数据传递工具
- Scala =====> 多范式编程语言、面向对象和函数式编程的特性
- Spark =====> 目前企业常用的批处理离线/实时计算引擎
- Flink =====> 目前最火的流处理框架、既支持流处理、也支持批处理
- Elasticsearch =====> 大数据分布式弹性搜索引擎
- 离线/实时项目

- **学习后能增加的技能树**

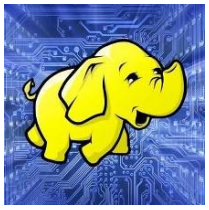
专业技能

1. 熟练使用 Hadoop，熟悉相应的常用工作流程与工作机制，根据业务需求完成 M/R 的开发
2. 熟悉 HBase 的存储原理，掌握 HBase 的常用操作
3. 熟练使用 Spark Core、Spark Sql 以及 Spark Streaming 的进行开发
4. 掌握 Spark 工作原理与 Spark 调优
5. 熟悉 Hive 工作原理，能够使用 Hive 进行海量数据的查询 清洗 分析 计算
6. 熟悉 Kafka 消息队列的工作机制，熟练掌握 Kafka 生产者、消费者的使用
7. 熟练使用 Flume 实现监听、上传、清洗，理解 Flume 框架的原理
8. 理解 Zookeeper 的存储原理，会配置 Zookeeper 集群，以及常用 API 操作
9. 熟练使用 Linux 操作命令，系统性能分析，Crontab 定时任务脚本、集群群起脚本
10. 熟悉 Redis 数据库，掌握 Redis 五大数据结构操作、持久化、事务控制、主从复制
11. 熟练掌握 MySQL 日常操作，掌握 Sql 的性能调优
12. 掌握 Scala 的基本使用，可以使用 Scala 进行 Spark 开发
13. 熟练使用 ElasticSearch 存储数据及关键字搜索
14. 了解 Flink 原理架构

- **就业大数据岗位**

  - 大数据开发工程师
  - 大数据清洗开发工程师
  - 大数据仓库开发工程师
  - 大数据运维开发工程师
  - 大数据平台开发工程师

- **起源**

  - 名字起源
    - 该项目的创建者，Doug Cutting 解释 Hadoop 的得名 ：“这个名字是我孩子给一个棕黄色的大象玩具命名的
    - 
  - 项目起源

- Hadoop 由 Apache Software Foundation 公司于 2005 年秋天作为 Lucene 的子项目 Nutch 的一部分正式引入。它受到最先由 Google Lab 开发的 Map/Reduce 和 Google File System(GFS) 的启发
- Google 是 Hadoop 的思想之源（Google 在大数据方面的三篇论文）
  - GFS ====> HDFS
  - Map-Reduce ====> MR
  - BigTable ====> HBase


Google-File-System中文版_1.0.pd

Google-MapReduce中文版_1.0.pc

Google-Bigtable中文版_1.0.pdf

➢ **三大发行版本**

- Apache、Cloudera、Hortonworks
- Apache 版本最原始、最基础：适合零基础　大公司在用
- Cloudera
  - Cloudera's DistributionIncluding Apache Hadoop 简称 CDH
  - 中小型公司用、简单方便、自带可视化
- Hortonworks
  - 文档较好
- 注：Cloudera 和 Hortonworks 在 2018 年 10 月，国庆期间宣布合并

大数据软件环境部署

以下所有软件下载地址：

链接：https://pan.baidu.com/s/1kFmIZGlbSJ-iKR5RUEFouw
提取码：im6t

➢ **实验环境详解**

- 硬性要求：
  - 内存：最低 8G+ (建议 12+)
    - 个人电脑最大内存检测：
      - win + R 输入 cmd
      - 复制代码：wmic memphysical get maxcapacity
      - 所显示的值：MaxCapacity 除以 1024 的平方
      - MaxCapacity：33554432
      - 33554432 除以 1024 除以 1024 等于 32G
      - 即个人 PC 的最大支持内存为 32G

- 磁盘：500GB+

- **我的个人电脑：**

- **Google 浏览器**

- 图标：
- 下载：360 管家下载/百度下载
- 程序员必备：不用‘谷歌浏览器’的程序员不是好程序员（此句五毛，括号内删除）

- **Everything（文件搜索工具）**

- 图标：
- 提供对个人 PC 的快速下载
- 下载：360 管家下载/百度下载

- **Notepad++（文本工具）**


&


- 下载：360 管家下载/百度下载

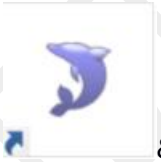
- **Pandownload（网盘高速下载工具）**

- **IDEA（集成开发工具）**
  - 图标： 
  - 后续学习用于编写 Java 和 Scala 代码
  - 全称 IntelliJ IDEA
  - 在业界被公认为最好的 java 开发工具之一
  - 支持多插件
  - 下载地址：百度下载社区版即可 https://www.jetbrains.com/

- **Vmware work station14 的安装配置 （虚拟网络环境）**
  - 图标： 

VMware workstation14安

  - 安装文档：

- **SQL yog & Navicat（数据库的图形管理软件）**
  -  & 

- **Secure CRT & Xshell（远程连接虚拟机的工具）**
  -  & 

SecureCRT破解安装配置.doc

xshell5安装配置.docx

  -

- **Winscp（Winodws 和 Linux 的传输）**
  - 

- **Linux 虚拟机的搭建**
  - 
Centos7安装配置详解.docx
- **鹏保宝解密视频**

第二天

大数据入门篇 Linux 基础

- **初认识** Linux
  - Linux 内核最初只是由芬兰人林纳斯·托瓦兹（Linus Torvalds）在赫尔辛基大学上学时出于个人爱好而编写的。
  - 

- 目前市面上较知名的发行版有：Ubuntu、RedHat、CentOS、Debain、Fedora、SuSE、OpenSUSE

➢ Windows **和** Linux **得区别**

比较	Window	Linux
界面	界面统一，外壳程序固定所有 Windows 程序菜单几乎一致，快捷键也几乎相同	圆形界面风格依发布版本不同而不同，可能互不兼容。GNU/Linux 的终端机是从 UNIX 传承下来，基本命令和操作方法也几乎一致。
驱动程序	驱动程序丰富，版本更新频繁。默认安装程序里面一般包含有该版本发布时流行的硬件驱动程序，之后所出的新硬件驱动依赖于硬件厂商提供。对于一些老硬件，如果没有了原配的驱动有时候很难支持。另外，有时硬件厂商未提供所需版本的 Windows 下的驱动，也会比较头痛。	由志愿者开发，由 Linux 核心开发小组发布，很多硬件厂商基于版本考虑并未提供驱动程序，尽管多数无需手动安装，但是涉及安装则相对复杂，使得新用户面对驱动程序问题会一筹莫展。但是在开源开发模式下，许多老硬件尽管在 Windows 下很难支持的也容易找到驱动。HP、Intel、AMD 等硬件厂商逐步不同程序支持开源驱动，问题正在得到缓解。
使用	使用比较简单，容易入门。圆形化界面对没有计算机背景知识的用户使用十分有利。	圆形界面使用简单，容易入门。文字界面，需要学习才能掌握。
学习	系统构造复杂、变化频繁、且知识、技能淘汰快，深入学习困难	系统构造简单、稳定，且知识、技能传承性好，深入学习相对容易
软件	每一种特定功能可能都需要商业软件的支持，需要购买相应的授权	大部分软件都可以自由获取，同样功能的软件选择较少。

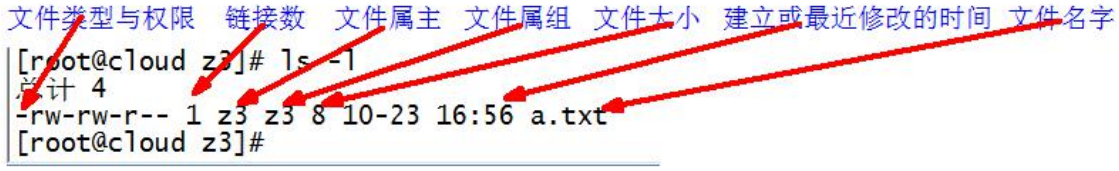
➢ Linux **的安装**

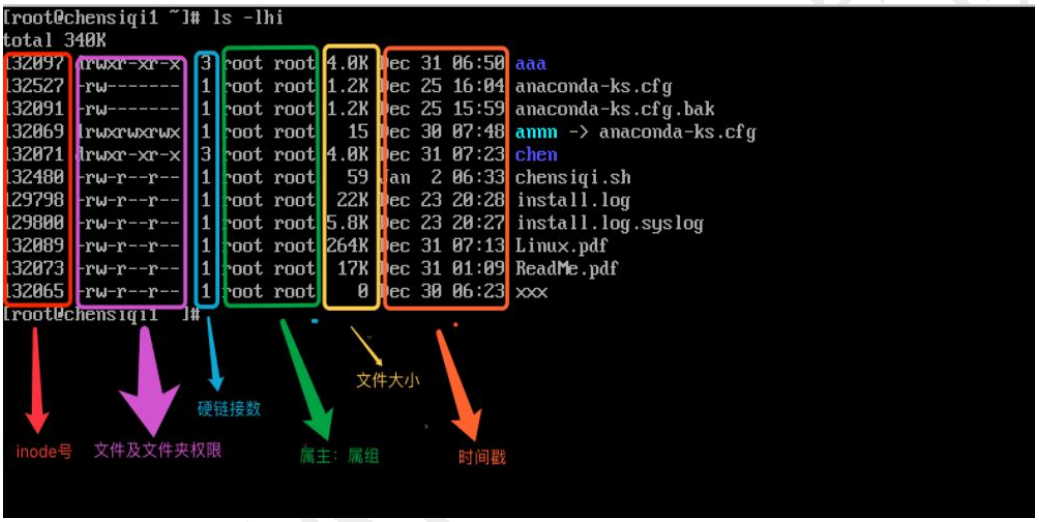
Centos7安装配置详解.docx

-

- Linux **常用的命令**

文件类型	属主权限	属组权限	其他用户权限
-	1 2 3	4 5 6	7 8 9
d	R w x	R - x	R - x
目录文件	读 写 执行	读 写 执行	读 写 执行





- **一般模式**
  - yy 复制
  - yNy 复制 N 行
  - p 粘贴
  - u 撤销
  - dd 删除一行
  - dNd 删除 N 行
  - shift + ^ 移动到行头
  - shift + & 移动到行尾
  - N + shift + g 跳到第 N 行

- **编辑模式**
  - i 进入编辑模式
  - o 进入下一行的编辑模式

- **指令模式**
  - w 保存
  - q 退出
  - ! 感叹号强制执行

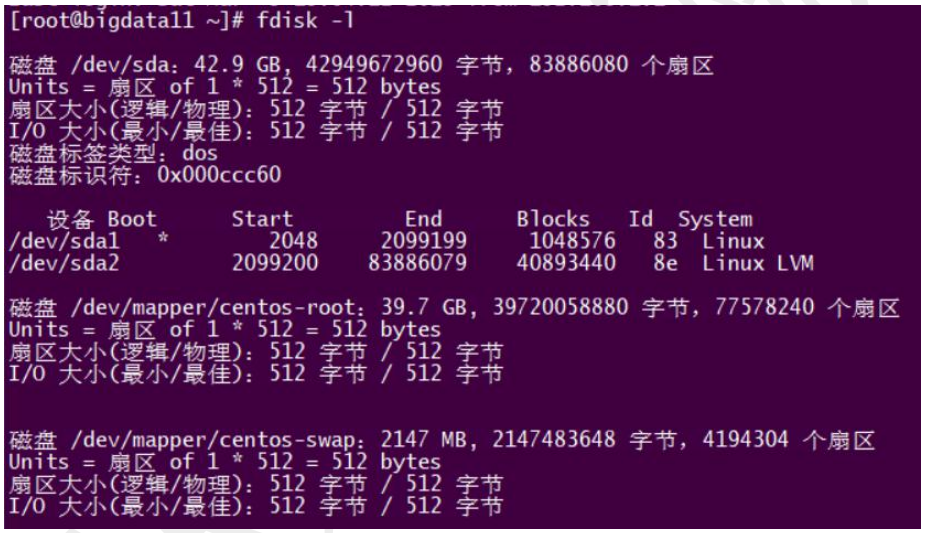
- **文件目录类**
  - pwd 显示当前工作目录
  - ls 列出目录内容
  - mkdir 创建新目录
    - mkdir -p 递归创建
  - touch 创建空文件
  - cd 切换目录
    - 绝对路径和相对路径
  - cp 复制文件或目录
    - cp -r 递归复制
  - rm 删除文件
  - mv 移动目录
  - cat 查看目录
  - more 分页查看文件
    - 空格 向下翻页
    - ctrl + B 返回上一屏
  - tail -F 监控文件
  - echo 追加文件
  - ln -s[原文件][目标文件] 软连接
  - history 历史服务器

- **时间日期类**
  - date 显示当前时间
    - date -s 设置系统时间
    - date -s '2019-03-09 23:23:23'
  - cal 查看日历
  - tab 自动补充键

- **用户管理命令**
  - useradd [用户] 添加新用户
  - userdel [用户] 删除新用户
  - passwd [用户] 设置用户密码
  - id [用户] 判断用户是否存在
  - su [用户] 切换用户
  - /etc/sudoers 设置普通用户具有 root 权限
  - usermod 修改用户
    - usermod -g dev itstar 把用户 itstar 加入到 dev 用户组

- groupadd　　itstar　新增用户组
- groupdel　　删除组
- groupmod　　修改组
- cat /etc/group　　查看创建了哪些组

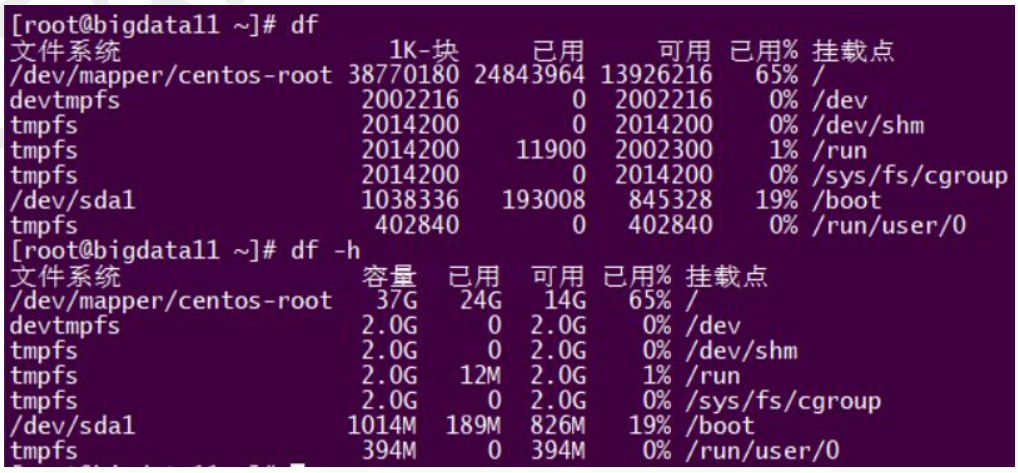
- **文件权限类**
  - chmod 改变权限 chmod -R 777 用户名
  - chown [最终用户][文件或目录]
    - chown -R itstar:itstar [文件名]
  - su [用户] 切换用户

- **磁盘分区类**
  - fdisk　　在 root 用户下查看分区
    - 
      ```
      [root@bigdata11 ~]# fdisk -l

      磁盘 /dev/sda：42.9 GB, 42949672960 字节, 83886080 个扇区
      Units = 扇区 of 1 * 512 = 512 bytes
      扇区大小(逻辑/物理)：512 字节 / 512 字节
      I/O 大小(最小/最佳)：512 字节 / 512 字节
      磁盘标签类型：dos
      磁盘标识符：0x000ccc60

         设备 Boot      Start         End      Blocks   Id  System
      /dev/sda1   *        2048     2099199     1048576   83  Linux
      /dev/sda2         2099200    83886079    40893440   8e  Linux LVM

      磁盘 /dev/mapper/centos-root：39.7 GB, 39720058880 字节, 77578240 个扇区
      Units = 扇区 of 1 * 512 = 512 bytes
      扇区大小(逻辑/物理)：512 字节 / 512 字节
      I/O 大小(最小/最佳)：512 字节 / 512 字节


      磁盘 /dev/mapper/centos-swap：2147 MB, 2147483648 字节, 4194304 个扇区
      Units = 扇区 of 1 * 512 = 512 bytes
      扇区大小(逻辑/物理)：512 字节 / 512 字节
      I/O 大小(最小/最佳)：512 字节 / 512 字节
      ```
  - df 查看硬盘
    - 
      ```
      [root@bigdata11 ~]# df
      文件系统                 1K-块     已用     可用 已用% 挂载点
      /dev/mapper/centos-root 38770180 24843964 13926216   65% /
      devtmpfs                 2002216        0  2002216    0% /dev
      tmpfs                    2014200        0  2014200    0% /dev/shm
      tmpfs                    2014200    11900  2002300    1% /run
      tmpfs                    2014200        0  2014200    0% /sys/fs/cgroup
      /dev/sda1                1038336   193008   845328   19% /boot
      tmpfs                     402840        0   402840    0% /run/user/0
      [root@bigdata11 ~]# df -h
      文件系统                 容量  已用  可用 已用% 挂载点
      /dev/mapper/centos-root   37G   24G   14G   65% /
      devtmpfs                 2.0G     0  2.0G    0% /dev
      tmpfs                    2.0G     0  2.0G    0% /dev/shm
      tmpfs                    2.0G   12M  2.0G    1% /run
      tmpfs                    2.0G     0  2.0G    0% /sys/fs/cgroup
      /dev/sda1               1014M  189M  826M   19% /boot
      tmpfs                    394M     0  394M    0% /run/user/0
      ```
  - mount /unmount 挂载/卸载

- **搜索查找类**
  - find [搜索范围][匹配条件]
    - 按文件名
      - find /opt -name *.jar

- 按拥有者
  - find /opt -user itstar
- 按文件大小（在某目录下查找大于 1M 的文件）
  - find /opt -size +1024

◆ grep 管道符
- grep + 参数 + 查找内容 + 源文件
- rpm -qa|grep mysql 查找系统中是否有 mysql 的 rpm 包
- grep "C\|A" A 注：区分大小写
- grep -i "C\|A" A 是不区分大小写

■ **进程线程类**

◆ ps -aux 查看系统中的进程
◆ top 查看系统的健康状态
◆ kill 进程 kill -9 进程号、直接杀死进程

■ **压缩和解压缩**

◆ gzip + 文件 压缩文件 注：不能压缩目录
◆ gunzip + 文件.gz 解压缩文件
◆ zip + 文件名 + 要压缩的内容
◆ unzip + *.zip 解压文件
- zip a.zip a 把 a 压缩成 zip 格式的文件

➢ Linux **定时任务** Crontab

◆ 基本语法
◆ crontab -e 编辑定时任务
◆ crontab -l 查询定时任务
◆ crontable -r 删除定时任务

crontab -e 进入编辑状态，* * * * * 执行的任务

项目	含义	范围
第一个“*”	一小时当中的第几分钟（分）	0-59
第二个“*”	一天当中的第几小时（时）	0-23
第三个“*”	一个月当中的第几天（天）	1-31
第四个“*”	一年当中的第几月（月）	1-12
第五个“*”	一周当中的星期几（周）	0-7（0 和 7 都代表星期日）

特殊符号

特殊符号	含义
*	代表任何时间。比如第一个“*”就代表一小时中每分钟都执行一次的意思。
,	代表不连续的时间。比如“0 8,12,16 * * * 命令”，就代表在每天的 8 点 0 分，12 点 0 分，16 点 0 分都执行一次命令
-	代表连续的时间范围。比如“0 5 * * 1-6 命令”，代表在周一到周六的凌晨 5 点 0 分执行命令

*/n	代表每隔多久执行一次。比如“*/10 * * * * 命令”，代表每隔 10 分钟就执行一遍命令

特定时间执行命令

时间	含义
45 22 * * * 命令	在 22 点 45 分执行命令
0 17 * * 1 命令	每周 1 的 17 点 0 分执行命令
0 5 1,15 * * 命令	每月 1 号和 15 号的凌晨 5 点 0 分执行命令
40 4 * * 1-5 命令	每周一到周五的凌晨 4 点 40 分执行命令
*/10 4 * * * 命令	每天的凌晨 4 点，每隔 10 分钟执行一次命令
0 0 1,15 * 1 命令	每月 1 号和 15 号，每周 1 的 0 点 0 分都会执行命令。注意：星期几和几号最好不要同时出现，因为他们定义的都是天。非常容易让管理员混乱。

案例：

*/1 * * * * echo “1” >> /opt/Andy

翻译：每分钟把 1 追加到该目录中

- **安装** linux **版本** JDK

  - **命令：**
    tar -zxvf JDKVERSON -C 目标目录
  - **环境变量：**
    vi /etc/profile
  - **环境配置：**
    export JAVA_HOME=/opt/module/jdk1.8.0_144
    export PATH=$JAVA_HOME/bin:$PATH

- **虚拟机快照**

右键虚拟机 -> 快照

功能描述：相当于“存档”的功能

- **主机名的设置**

hostnamectl set-hostnamectl 主机名

- **虚拟机联网**

- RPM **包**

  - **概述**

    RPM（RedHat Package Manager），Rethat 软件包管理工具，类似 windows 里面的

setup.exe

是 Linux 这系列操作系统里面的打包安装工具，它虽然是 RedHat 的标志，但理念是通用的。

RPM 包的名称格式

- Apache-1.3.23-11.i386.rpm
- “apache” 软件名称
- “1.3.23-11”软件的版本号，主版本和此版本
- “i386”是软件所运行的硬件平台
- “rpm”文件扩展名，代表 RPM 包

- **常用命令**
  - 查询
    - rpm -qa | grep mysql 查询是否具有 mysql 的 RPM 包
  - 卸载
    - rpm -e --nodeps [包名] 强制卸载此包
  - 安装
    - rpm -ivh --nodeps [包名] 不检测依赖进度