**大数据项目实战期末考试试卷**

考试时间：90分钟 总分：100分

**班级： 姓名： 学号：**

**一、单选题 （**1题1分，共50分**）**

1. Centos 中修改文件或目录的访问权限（ C ）命令

A. chown B. passwd C. chmod D. clear

2. 在文件属性中，文件的权限用第( D )列字符表示。

A. 1-5 B. 2-10 C. 3-9 D. 1-8

3. Hadoop 安装在一台计算机上，需修改相应的配置文件，用

一台计算机模拟多台主机的集群是（ B ）模式。

A. 全分布模式 B. 伪分布模式 C. 单机模式 D. 全分布 HA 模式

4. 配置 Hadoop 环境变量修改（ A ）文件

A. vi /etc/profile B. vi /etc/profiles

C. vi /etc/hosts D. vi ~/input/data

5. 配置 SSH 服务需配置（ B ）文件

A. vi /etc/ssh/sshd B. vi /etc/ssh/sshd\_config

C. C.vi /etc/sysconfig/network-scripts/ifcfg-eth0

D. D.vi ~/.bash\_profile

6. hadoop2.0 与 hadoop1.0 区别（ B ）

A. 增加 MapReduce2 B. 增加 YARN

C. 增加 HDFS2 D. 增加容错机制

7. 在 hadoop 配置中 core-site.xml 的配置是（ D ）参数。

A. 集群全局参数 B. HDFS 参数

C. Mapreduce 参数 D. 集群资源管理系统参数

8. 在 hadoop 配置中 yarn-site.xml 作用是（ D ）

A. 用于定义系统级别的参数 B. 用于名称节点和数据节点的存放位置

C. 用于配置 JobHistory Server 和应用程序参数

D. 配置 ResourceManager，NodeManager 的通信端口

9. 如没有配置 hadoop.tmp.dir 参数，此时系统默认的临时目录为( A )

A. /tmp/hadoop-hadoop B. /usr/hadoop-hadoop

C. /usr/local/hadoop-hadoop D. /usr/local/src/hadoop-hadoop

10. 以下哪个文件是能够修改 HDFS 的副本保存数量参数（ D ）

A. 修改 mapred-site.xml 配置文件 B. 修改 core-site.xml 配置文件

C. 修改 yarn-site.xml 配置文件 D. 修改 hdfs-site.xml 配置文件

11. 在 MapReduce 中 Shuffle 的主要作用是（B ）

A. 将数据进行拆分 B. 对映射后的数据进行排序, 然后输入到 Reducer

C. 经过映射后的输出数据会被排序，然后每个映射器会进行分区

D. 通过实现自定义的 Partitioner 来指定哪些数据进入哪个Reducer

12. HDFS 启动成功的标识是（ D ）

A. jps 显示 NameNode B. jps 显示 DataNode

C. jps 显示 SecondaryNameNode

D. jps显示NameNode\DataNode\SecondaryNameNode

13. Hive 适合（C ）环境

A. Hive 适合用于联机(online)事务处理 B. 提供实时查询功能

C. 适合应用在大量不可变数据的批处理作业 D. Hive 适合关系型数据环境

14. Hive 默认的元存储是保存在内嵌的（ B ）数据库中

A. MySQL B. Derby C. Oracle D. SqlServer

15. 典型的 NoSQL 数据库是（ C ）

A. Hive B. MySQL C. Hbase D. Oracle

16. HMaster 的主要作用是（ A ）

A. 用于启动任务管理多个 HRegionServer

B. 负责响应用户 I/O 请求，向 HDFS 文件系统读写数据

C. 负责协调集群中的分布式组件D. 最终保存 HBase 数据行的文件

17. ZooKeeper 的特点不包括（D ）

A. 顺序一致性 B. 原子性 C. 可靠性 D. 复合系统映像

18. Sqoop 的作用是（ A ）

**A.** 用于传统关系型数据库和 Hadoop 之间传输数据

**B.** 提供对日志数据进行简单处理的能力

**C.** 是 Hadoop 体系中数据存储管理的基础

**D.** 是一个建立在 HDFS 之上，面向列的针对结构化和半结

构化数据的动态数据库

19. 测试 Sqoop 是否能够正常连接 MySQL 数据库命令是( D )

A. sqoop list-database --connect

jdbc:mysql://127.0.0.1:3306/ --username root –P

B. sqoop list-databases --connection

jdbc:mysql://127.0.0.1:3306/ --username root –P

C. sqoop list-databases --connect

jdbc:mysql://127.0.0.1:3306/ --username root –L

D. sqoop list-databases --connect

jdbc:mysql://127.0.0.1:3306/ --username root –P

1. 大数据解决方案确认需求分析准备工作不包括（ C ）

A. 充分了解客户的要求，客户的业务，客户的关注重点

B. 和客户负责人沟通，确定客户需求小组成员

C. 编制确定客户实施计划

D. 根据事先了解的初步用户需求，列出这些难点

21. 大数据平台项目实施计划不包括（ B ）

A. 基础平台部署 B. 用户环境准备

C. 完成大数据平台的安装调试 D. 完成大数据平台操作开发

22. 大数据平台操作演示不包括（ C ）

A. 在大数据平台操作演示前需要做好准备工作

B. 分析本次演示的目的和听众群体

C. 实现呈现技巧提升 D. 准备好演示环境

23. Hbase 的主要作用是（ A ）

A. 是一个分布式的、面向列的开源数据库

B. 是一种编程模型，用于大规模数据集（大于 1TB）的并行运算

C. 是 Hadoop 集群当中的资源管理系统模块

D. 将要储存的文件分散在不同的硬盘上，并记录他们的位置

24. 在配置好 Hadoop 集群之后，通过浏览器登录查看

Hadoop 状态，以下是查看的 URL（B ）

A. http://master:8088 B. http://192.168.1.6:50070

C. http://master:16010 D. http://master:9999

25. 通过浏览器界面监控 Hive 的状态, 以下是查看的端口 （A ）

A. http://master:8088 B. http://192.168.1.6:50070

C. <http://master:16010> D. http://master:9999

26. 日志消息的级别中哪个是比较严重的情况( B )

A. ALEKKT B. CRIT C. ERR D. DEBUG

27. 大数据处理中数据处理层不包括（ C ）

A. 数据清洗 B. 数据规约 C. 数据提取 D. 数据分析

28. 大数据处理中数据应用层不包括（ B ）

A. 数据挖掘 B. 数据建模 C. 数据可视化 D. 数据分析

29. 日志消息的级别中那个是不会影响系统但值得注意( C )

A. ALEKKT B. CRIT C. NOTICE D. DEBUG

30. 在进行 Hive 组件配置时，JAR 包冲突导致异常，出现以

下错误信息( A )

A. Failed to set setXIncludeAware(true) for parser

B. mkdir: `input’: No such file or directory

C. E45: ‘readonly’ option is set (add ! to override)

D. ERROR tool.ImportTool:Imported Failed:Character

8216 is an out-of-range delimiter

31. Hadoop 官方真正支持的操作系统是（D ）

A. UNIX B. Windows C. DOS D. Linux

32. 显示当前所在目录的命令是（ A ）

A.pwd B.ls C.cd D.mkdir

33. 以下选项哪个可以修改文件或目录的访问权限（ C ）

A.passwd B.su C.chmod D.useradd

34. 以下选项哪个是 HDFS 的名称节点（ D ）

A. DataNode B. SecondaryNode C. Node D. NameNode

35. 以下选项哪个是 MapReduce 正确的运行模型（ C ）

A.Reduce-Map-Shuffle B.Shuffle-Map-Reduce

C..Map-Shuffle-Reduce D.Map-Reduce-Shuffle

36. 以下选项哪个是 YARN 中动态创建的资源容器（ D ）

A.ResourceManager B.NodeManager

C.ApplicationMaster D.Container

37. 以下选项哪个命令可以生成公钥（ A ）

A.ssh-keygen B.ifup team0 C.nmcli dev D.nmcli con sh

38. 以下选项哪个命令是用 hadoop 用户登录，通过 scp 命

令实现密钥拷贝到 slave1 节点上（C ）

A.scp ~/.ssh/id\_rsa.pub hadoop@slave2:~/

B.scp ~/.ssh/id\_rsa.pub root@slave1:~/

C.scp ~/.ssh/id\_rsa.pub hadoop@slave1:~/

D.scp ~/.ssh/id\_rsa.pub slave1@root:~/

39. 以下选项哪个命令可以查看主机信息( A )

A.nmcli dev B.nmcli con sh C.ifup team0 D.teamdctl team0 st

40. 在 Hadoop 配置文件中，以下属于 core-site.xml 文件主

要内容的是（ A ）

A、用于定义系统级别的参数，如 HDFS URL、Hadoop的临时目录等。

B、名称节点和数据节点的存放位置、文件副本的个数、文

件读取权限等。

C、配置 ResourceManager，NodeManager 的通信端

口，web 监控端口等。

1. 包括 JobHistory Server 和应用程序参数两部分，如 reduce
2. 任务的默认个数、任务所能够使用内存的默认上下限等。

41. 以下（D ）文件中主要用来配置 ResourceManager，

NodeManager 的通信端口，web 监控端口等。

A、core-site.xml B、mapred-site.xml

C、hdfs-site.xml D、yarn-site.xml

42. 在 HDFS 分布式文件系统中，一般采用冗余存储，冗余因

子通常设置为（ C ）

A、 1 B、 2 C、 3 D、 4

43. 用于检查节点自身的健康状态的进程是（ B ）

A、 ResourceManager B、 NodeManager

C、 NameNode D、 DataNode

44. ResourceManager 对 web 服务提供地址，用户可通过

ip 地址：（ A ）在浏览器中查看集群各类信息

A、 8088 B、 8033 C、 8080 D、 8032

45. 启动 HDFS 至少需要启动（A ）进程

A、 NameNode B、 ResourceManager

C、 SecondaryNameNode D、 DataNode

46. 在 HDFS 文件操作中，上传文件的命令是（A ）

A、 put B、 input C、 get D、 up

47. MySQL 数据库驱动文件放置于 hive 哪个目录下（B ）

A、 jar B、 lib C、 bin D、 sbin

48. 下列哪些不是 HBase 的特点（ D ）

A. 高可靠性 B. 高性能 C. 面向列 D. 紧密性

49. HBase 与下列哪个选项属于同一种类型的数据库（ A ）

A. MongoDB B. MariaDB C. MySQL D. Oracle

50. 在 HBase 的组件中，哪一个负责日志记录（ D ）

A. HRegion B. HFile C. MemStore D. WAL

**二、多选题** （1题2分，共20分）

1. Linux 版本分为以下哪几类（ AC ）

A. 发行版 B. 测试版 C. 内核版 D. 开发版

2. Linux 的特点是以下（ ACD ）方面

A. 开放源代码的程序软件，可自由修改。

B. 与 Unix 系统不兼容，具备几乎所有 Unix 的优秀特性。

C. 可自由传播，免费使用，无任何商业化版权制约。

D. 适合 Intel 等 x86 CPU 系列架构的计算机。

3. Hadoop 的运行模式分为以下（ABC ）种

A. 单机模式 B. 伪分布模式 C. 全分布模式 D. 主备模式

4. Hadoop 优点是以下（ABCD ）种。

A. 扩容能力强 B. 成本低 C. 高效率 D. 高可靠性

5. 验证 SSH 无密码登录需要（AC ）步骤

A. 修改 Master 节点 authorized\_keys 文件

B. 修改 Slave 节点 authorized\_keys 文件

C. 验证 Master 到每个 Slave 节点无密码登录

D. 验证两个 Slave 节点到 Master 节点无密码登录

6. 查询获取 Hadoop 全部的配置信息官方网址是（ ABCD ）

A. http://hadoop.apache.org/docs/current/hadoop

project-dist/hadoop-common/core-default.xml

B. http://hadoop.apache.org/docs/current/hadoop

project-dist/hadoop-hdfs/hdfs-default.xml

C. http://hadoop.apache.org/docs/current/hadoop

mapreduce-client/hadoop-mapreduce-client

core/mapred-default.xml

D. http://hadoop.apache.org/docs/current/hadoop

yarn/hadoop-yarn-common/yarn-default.xml

7. 以下哪些是 yarn-site.xml 配置文件主要参数（ ABC ）

A. yarn.resourcemanager.address

B. yarn.resourcemanager.scheduler.address

C. yarn.resourcemanager.admin.address

D. yarn.nodemanager.oux-services

8. Hadoop 平台的主要配置文件是（ ABCD ）

A. core-site.xml B. hdfs-site.xml

C. mapred-site.xml D. slaves.xml

9. core-site.xml 文件核心参数是（ AC ）

A. mapreduce.framework.name

B. mapreduce.jobhistory.address

C. yarn.resourcemanager.address

D. mapreduce.jobhistory.webapp.address

10. MapReduce 是 Hadoop 中集群数据处理的核心，分为（ AB ）

A. 分为映射, 映射为 Map 阶段 B. 分为减速，减速为 Reduce 阶段

C. 分为映射, 映射为 Reduce 阶段 D. 分为减速，减速为 Map 阶段

**将以上选择题填至以下答题卡**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **单选题** | | | | |
| **1：** | **2：** | **3：** | **4：** | **5：** |
| **6：** | **7：** | **8：** | **9：** | **10：** |
| **11：** | **12：** | **13：** | **14：** | **15：** |
| **16：** | **17：** | **18：** | **19：** | **20：** |
| **21:** | **22:** | **23:** | **24:** | **25:** |
| **26:** | **27:** | **28:** | **29:** | **30:** |
| **31:** | **32:** | **33:** | **34:** | **35:** |
| **36:** | **37:** | **38:** | **39:** | **40:** |
| **41:** | **42:** | **43:** | **44:** | **45：** |
| **46:** | **47:** | **48:** | **49:** | **50：** |
| **多选题** | | | | |
| **1：** | **2：** | **3：** | **4：** | **5：** |
| **6：** | **7：** | **8：** | **9：** | **10：** |

# 实操题： 安装单机版Hadoop系统（30分）

## 实验目标

完成本实验，您应该能够：

* 掌握hadoop的安装步骤
* 掌握hadoop的运行原理
* 掌握运行hadoop的基本命令

## 实验要求

* 熟悉运行hadoop的基本命令
* 熟悉hadoop的运行原理
* 熟悉hadoop的安装步骤

## 实验环境

本实验所需之主要资源环境如表1-1所示。

表1-1 资源环境

|  |  |
| --- | --- |
| 服务器集群 | 单节点，机器最低配置：双核CPU、8GB内存、100G硬盘 |
| 运行环境 | CentOS 7.4 |
| 大数据平台 | H3C教学与实践管理系统 |
| 服务和组件 | 完成前面章节的实验，其他服务和组件根据实验需求安装 |

## 实验过程

### 实验任务一：配置Hadoop配置文件

进入Hadoop目录

[root@master ~]# cd /usr/local/src/hadoop/

配置hadoop-env.sh文件，目的是告诉Hadoop系统JDK的安装目录。

[root@master hadoop]# vi etc/hadoop/hadoop-env.sh

在文件中查找export JAVA\_HOME这行，将其改为如下所示内容。

export JAVA\_HOME=/usr/local/src/java

这样就设置好Hadoop的本地模式，下面使用官方案例来测试Hadoop是否运行正常。

### 实验任务二：测试Hadoop本地模式的运行

#### 步骤一: 切换到hadoop用户

使用hadoop这个用户来运行Hadoop软件。

[root@master hadoop]# su - hadoop

[hadoop@master ~]$

#### 步骤二: 创建输入数据存放目录

将输入数据存放在~/input目录（hadoop用户主目录下的input目录中）。

[hadoop@master ~]$ mkdir ~/input

#### 步骤三: 创建数据输入文件

创建数据文件data.txt，将要测试的数据内容输入到data.txt文件中。

[hadoop@master ~]$ vi ~/input/data.txt

输入如下内容，保存退出。

Hello World

Hello Hadoop

Hello Huasan

#### 步骤四: 测试MapReduce运行

运行WordCount官方案例，统计data.txt文件中单词的出现频度。这个案例可以用来统计年度十大热销产品、年度风云人物、年度最热名词等。命令如下：

[hadoop@master ~]$ hadoop jar /usr/local/src/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.7.1.jar wordcount ~/input/data.txt ~/output

WARN io.ReadaheadPool: Failed readahead on ifile

EBADF: Bad file descriptor

如果出现上面的警告，是因为快速读取文件的时候，文件被关闭引起，此处警告不影响单机mapreduce的运行。

运行结果保存在~/output目录中，命令执行后查看结果：

[hadoop@master ~]$ ll ~/output

**total 4**

**-rw-r--r--. 1 hadoop hadoop 34 May 2 12:08 part-r-00000**

**-rw-r--r--. 1 hadoop hadoop 0 May 2 12:08 \_SUCCESS**

文件\_SUCCESS表示处理成功，处理的结果存放在part-r-00000文件中，查看该文件。

[hadoop@master ~]$ cat ~/output/part-r-00000

**Hadoop 1**

**Hello 3**

**Huasan 1**

**World 1**

可以看出统计结果正确，说明Hadoop本地模式运行正常。读者可将这个运行结果与“3.2.3 MapReduce”中的WordCount案例运行过程进行对照，来加深对MapReduce框架的理解。

注意：输出目录不能事先创建，如果已经有~/output目录，就要选择另外的输出目录，或者将~/output目录先删除。删除命令如下所示。

[hadoop@master ~]$ rm -rf ~/output

**把成功运行截图贴在下方**