第 1 页 （共 6 页）

**Hadoop 大数据技术 科目复习题**

## 一、单项选择题

1、Hadoop 版本演进过程中，Hadoop2.0 比 Hadoop1.0 有了很多的优化，下面哪项不属于 Hadoop2.0？（）

A、加入 HDFS 的 NameNode Federation 和 YARN

B、YARN 基于 cgroup 的内存和磁盘 IO 隔离

C、支持 NameNode HA

D、Wire-compatibility 特性

2、Hadoop 版本演进过程中，Hadoop3.0 比 Hadoop2.0 有了很多的优化，下面哪项

不属于 Hadoop3.0？（

）

A、JDK 版本的最低依赖从 1.7 变成了 1.8

B、支持多个 Standby 状态的 NameNode

C、支持 NameNode HA

D、datanode 内部添加了负载均衡

3、Hadoop 更适合哪些场景？（

）

A、离线分析

B、复杂数据

C、少量数据

D、在线分析

4、 Hadoop 的作者是？（

）

A、Doug cutting

B、Martin Fowler

C、Kent Beck

5、下面的配置项配置在 hadoop 哪个配置文件？（

）

<property>

<name>hadoop.tmp.dir</name>

<value>/home/hadoop/hadoop/tmp</value>

</property>

A、core-site.xml

B、hdfs-site.xml

C、mapred-site.xml

D、yarn-site.xml

6、端口 50070 默认是 Hadoop 哪个服务的端口？（

）

A、NameNode

B、DataNode

C、SecondaryNameNode

D、Yarn

7、Hadoop 完全分布模式配置免密登录是要（

）

A、实现主节点到其他节点免密登录。

B、实现从节点到主节点的免密登录。

C、主节点和从节点任意两个节点之间免密登录

8、安装 Hadoop 时，发现 50070 对应的页面无法打开，可以通过下面哪个命令查

看某个端口（

TCP 或 UDP）是否在监听?（

）

A、ps

B、netstat

C、ping

D、ifconfig

9、下面哪个目录保存了 Hadoop 集群的命令（比如启动 Hadoop）?（

）

A、bin

B、sbin

C、etc

D、share

10、把公钥追加到授权文件的命令是（

）

A、ssh-copy-id

B、ssh-add

C、ssh

D、ssh-keygen

11、采用用户 user1 安装 hadoop 伪分布式时，解压 hadoop 安装包采用下面命令：

sudo tar -zxvf hadoop-2.7.3.tar.gz -C ~

运行 ls -al 命令显示

drwxr-xr-x 11 root root

4096 Aug 3 01:06 hadoop-2.7.3

如果要修改hadoop-2.7.3 目录的权限，采用下面哪个命令才有效而且最佳？（

）

A、chown user1:user1

hadoop-2.7.3

B、chown -R user1:user1

hadoop-2.7.3

C、sudo chmod -R 777

hadoop-2.7.3

D、sudo chown -R user1:user1

hadoop-2.7.3

12、下列哪个属性是 hdfs-site.xml 中的配置？（

）

A、fs.defaultFS

B、dfs.replication

C、mapreduce.framework.name

D、yarn.resourcemanager.address

13、安装 Hadoop 时，配置项“

hadoop.tmp.dir ”应该配置在哪个文件？（

）

A、core-site.xml

B、hdfs-site.xml

C、mapred-site.xml

D、yarn-site.xml

14、下面哪个命令可以实现将 HDFS 中的文件下载到 Linux 本地？（

）

A、hdfs dfs -copyToLocal

B、hdfs dfs -put

C、hdfs dfs copyFromLocal

D、hdfs dfs -cp

15、通过哪个命令可以查看 hdfs 的状态？（

）

A、hdfs dfsadmin –report

B、hdfs dfsadmin -safemode

C、hdfs dfsadmin -printTopology

16、关于 SecondaryNameNode 哪项是正确的？ （

）

A、它是 NameNode 的热备

B、它对内存没有要求

C、它的目的是帮助 NameNode 合并编辑日志，减少 NameNode 启动时间第

D、SecondaryNameNode 应与 NameNode 部署到一个节点

17、下列哪些提法正确?（ ）

A、 Hadoop 适合数据的随机读写。

B、Hadoop 的数据块大小(Block Size) 是不可以修改的。

C、Hadoop 集群默认没有严格的权限管理和安全措施保障。

D、因为 HDFS 有多个副本，所以 NameNode 是不存在单点问题的。

18、通过 HDFS 哪个命令可以直接清空回收站？（ ）

A、hdfs dfs -expunge

B、hdfs dfs -df

C、hdfs dfs -mv

D、hdfs dfs -deleteSnapshot

19、关于 HDFS 回收站描述正确的是？（ ）

A、HDFS 回收站默认开启

B、HDFS 回收站中的文件文件像 Windows 回收站一样，如果不清空回收站，文件

会一直保留在回收站。

C、HDFS 为每一个用户都创建了回收站，这个类似操作系统的回收站。位置是/user/

用户名/.Trash/

D、用户不能手动清空回收站中的内容

20、下面哪个程序负责 HDFS 数据存储?（ ）

A、NameNode

B、ResourceManager

C、SecondaryNameNode

D、NodeManager

E、Datanode

21、HDFS 中 block 默认保存几份?（ ）

A、1

B、2

C、3

D、4

22、HDFS 检查点（

CheckPoint）的作用是可以减少下面哪个组件的启动时间？ （ ）

A、SecondaryNameNode

B、NameNode

C、DataNode D、JournalNode

23、下面哪一项不属于 DataNode 的职责？（ ）

A、存储数据块（Block）

B、负责客户端对数据块的 IO 请求

C、管理 DataNode 上文件数据块（Block）的均衡

D、定期向 Namenode 汇报自身所持有的 Block 信息

24、当 NameNode 出错时，下面哪个方案描述正确，且是最佳故障恢复和容错方案？（）

A、采用 SecondaryNameNode 定时备份 NameNode 的 fsimage 和 edits

B、采用 NameNode HA,当一个 NameNode 出错时，另一个 NameNode 接管它的工作。

C、采用 NameNode Fedaration，多个 Namenode 一起工作。

D、多增加 DataNode第 7 页 （共 6 页）

25、以下哪个不是 HDFS 的进程？（

）

A、SecondaryNamenode

B、Datanode

C、Namenode

D、MRAppMaster/YarnChild

26、假如现在 cd 到 hadoop 安装目录下，请问以下哪个命令不正确？（ ）

A、sbin/stop-dfs.sh

B、sbin/start-dfs.sh

C、bin/hdfs dfs -cat /mydemo/my.txt

D、sbin/hdfs namenode -format

27、HDFS 集群中的 NameNode 职责不包括？（ ）

A、维护 HDFS 集群的目录树结构

B、维护 HDFS 集群的所有数据块的分布、副本数和负载均衡

C、响应客户端的所有读写数据请求

D、负责保存客户端上传的数据

28、关于 HDFS 集群中的 DataNode 的描述不正确的是？（ ）

A、一个 DataNode 上存储一个数据块的多个副本

B、存储客户端上传的数据的数据块

C、响应客户端的所有读写数据请求，为客户端的存储和读取数据提供支撑

D、当 Datanode 读取数据块的时候，会计算它的校验和（

checksum），如果计算后的校验和，与数据块创建时值不一样，说明该数据块已经损坏

29、HDFS 集群中的 DataNode 的主要职责是？（ ）

A、维护 HDFS 集群的目录树结构

B、维护 HDFS 集群的所有数据块的分布、副本数和负载均衡第 8 页 （共 6 页）

C、存储数据块

D、接收客户端的请求

30、下列关于配置机架感知的相关描述哪项不正确？（ ）

A、如果一个机架出问题，不会影响数据读写和正确性

B、写入数据的时候多个副本会写到不同机架的 DataNode 中

C、MapReduce 会根据机架的拓扑获取离自己比较近的数据块

D、数据块的第一个副本会优先考虑存储在客户端所在节点

31、HDFS 的是基于流数据模式访问和处理超大文件的需求而开发的，具有高容错、高可靠性、高可扩展性、高吞吐率等特征，适合的读写任务是：（）

A、一次写入，少次读取

B、多次写入，少次读取

C、一次写入，多次读取

D、多次写入，多次读取

32、Namenode 在启动时自动进入安全模式，在安全模式阶段，说法不正确的是？ （ ）

A、安全模式目的是在系统启动时检查各个 DataNode 上数据块的有效性

B、根据策略对数据块进行必要的复制或删除

C、当数据块最小百分比数满足的最小副本数条件时，会自动退出安全模式

D、文件系统允许有修改

33、关于 HDFS 的文件写入，正确的是？（ ）

A、支持多用户对同一文件的写操作

B、用户可以在文件任意位置进行修改第 9 页 （共 6 页）

C、默认将文件块复制成三份分别存放

D、复制的文件块默认都存在同一机架的多个不同节点上

34、YARN Web 界面默认占用哪个端口？（ ）

A、50070

B、8088

C、50090

D、9000

35、下面哪个 YARN 的描述不正确的是？（

）

A、YARN 指 Yet Another Resource Negotiator，另一种资源协调者

B、YARN 只支持 MapReduce 一种分布式计算模式

C、YARN 最初是为了改善 MapReduce 的实现

D、YARN 的引入为集群在利用率、资源统一管理和数据共享等方面带来了巨大好

处

36、下面哪个不属于 YARN 的架构的组成部分？（

）

A、JobTracker

B、ResourceManager

C、NodeManager

D、Application Master

E、Container

37、下面对 YARN 调度器描述正确的是？（

）

A、Capacity Scheduler 是一种单队列的调度器

B、Hadoop2.0，Fair Scheduler 是 YARN 中默认的资源调度器

C、多用户的情况下，Fair Scheduler 可以最大化集群的吞吐和利用率

D、Hadoop1.0，FIFO Scheduler 是 YARN 中默认的资源调度器第 10 页 （共 6 页）

38、YARN 中，任务进度监控是向哪个组件汇报的？（

）

A、ResourceManager

B、NodeManager

C、ApplicationMaster

D、Container

39、MapReduce 的特点不包括：（

）

A、易于编程

B、良好的扩展性

C、高容错性

D、擅长对 PB 级以上海量数据进行实时处理

40、MapReduce 更擅长：（

）

A、离线计算

B、实时计算

C、流式计算

D、DAG（有向图）计算

41、MapReduce 中，Mapper 的个数由什么决定的？（

）

A、SplitInput 的个数

B、DataNode 的个数

C、文件切分的数据块的个数

D、计算机计算能力

42、Reducer 的个数由什么决定的？（

）

A、DataNode 的个数

B、文件切分的数据块的个数

C、计算机计算能力第 11 页 （共 6 页）

D、Partition 分区的个数

43、MapTask 或 ReduceTask 向自己的哪个组件报告进度和状态？（

）

A、ResourceManager

B、NodeManager

C、MRAppMaster

D、Container

44、MapReduce 的 Shuffle 过程中哪个操作是最后做的？ （

）

A、合并

B、 溢写

C、分区

D、排序

45、下面关于 MapReduce 的描述中正确的是？（

）

A、MapReduce 程序必须包含 Mapper 和 Reducer

B、MapReduce 程序的 MapTask 可以任意指定

C、MapReduce 程序的 ReduceTask 可以任意指定

D、MapReduce 程序的默认数据读取组件是 TextInputFormat

46、MapReduce 编程模型中以下组件哪个是最后执行的？ （

）

A、Mapper

B、Partitioner

C、Reducer

D、RecordReader

47、在 MapReduce 中，哪个组件如果用户不指定，则不会默认存有的？ （

）

A、Combiner

B、OutputFormat第 12 页 （共 6 页）

C、Partitioner

D、InputFormat

48、下列哪种业务场景中，不能直接使用 Reducer 充当 Combiner 使用？（

）

A、sum 求和

B、avg 求平均

C、max 求最大值

D、count 求计数

49、以下描述不正确的是？（

）

A、SequenceFile 可以用来作为小文件的合并存储容器

B、TextInputFormat 的 key 是 LongWritable 类型的

C、TextInputFormat 的 key 是指该记录在文件中的行号

D、TextInputFormat 是默认 InputFormat

50、以下哪个组件可以指定对 key 进行 Reduce 分发的策略？（

）

A、RecordReader

B、Combiner

C、Partitioner

D、FileInputFormat

51、执行一个 job，如果这个 job 的输出路径已经存在，那么程序会？（

）

A、覆盖这个输出路径

B、抛出警告，但是能够继续执行

C、创建一个新的输出路径

D、抛出一个异常，然后退出

52、下列关于 MapReduce 并行切分或输入输出描述不正确的是？（

）第 13 页 （共 6 页）

A、InputFormat 中实现的 getSplits()可以把输入数据划分为输入分片(InputSplit)

B、为实现细粒度并行，输入分片(InputSplit)应该越小越好

C、一台机器可能被指派从输入文件的任意位置开始处理一个分片

D、输入分片(InputSplit)是一种记录的逻辑划分，而数据块(Block)是对输入数据的

物理分割

53、MapReduce 中，Mapper 的输出经过 Shuffle 后，Reducer 获取到的输入<k3,v3>

是有序的，且 k3 互不相同，v3 是相同 k2 的 v2 组成的集合，这相当于实现了 SQL

中哪个语句？（

）

A、group by

B、distinct

C、order by

D、以上都是

## 编程题 1（10 分）：

### 1、在一台操作系统为 Ubuntu16.04 机器部署 Hadoop 伪分布式环境。实现下面功能需要输入什么 Linux 命令？

（

1)查看是否安装了 openssh-server

rpm -qa | grep -i openssh-server

2)查看机器主机名

hostname

3)检查 Hadoop 进程是否存在

start-dfs.sh

4)查看 SSH 服务的 22 端口是否在监听（

Listen）

lsof -i:**22**

### 2、在三台操作系统为 Ubuntu16.04 机器(机器名分别是 node1、node2、node3)部署Hadoop 完全分布式环境，三台机器已经实现免密码登录。实现下面功能需要输入什么 Linux 命令？

（

1)从 node1，通过 ssh 登录到 node2

（

ssh node2

2)在 node2 上运行命令，将 node1 的/home/hadoop/hadoop-2.7.3 拷贝到

/home/hadoop 下

hdfs dfs -cp /home/hadoop/hadoop-2.7.3 /home/hadoop

3)查看当前机器的磁盘使用量

df -lh

3、通过 HDFS Shell 操作命令实现如下功能。第 14 页 （共 6 页）

1)创建一个 HDFS 目录/mydemo

hdfs dfs -mkdir /mydemo

2)创建一个空文件/mydemo/file1.txt

touch /mydemo/file1.txt

3)将 Linux 当前目录下的文件 data.txt 追加到 HDFS 文件/mydemo/file1.txt 末尾

hdfs dfs -appendToFile data.txt /mydemo/file1.txt

4)统计 HDFS 目录/mydemo 下的目录个数， 文件个数， 文件总计大小。

hadoop fs -count /mydemo

4、通过 HDFS Shell 操作命令实现如下功能。

（

1)

递归列出 HDFS 中/mydemo 文件夹下的所有子文件或子目录

hadoop fs -ls /mydemo

2)

将本地目录 data.txt 文件上传到 HDFS 的/mydemo 目录下

hdfs dfs -put data.txt /mydemo

3)

查看 HDFS 下/mydemo/data.txt 文件中的内容

hdfs dfs -cat /mydemo/data.txt

## 编程题 2（10 分）：

### 理解并默写 WordCount 全部代码。

WordCountDriver类：

public class WordCountDriver {

public static void main(String[] args) throws IOException,

ClassNotFoundException, InterruptedException {

// 1 获取配置信息以及获取job对象

Configuration conf = new Configuration();

Job job = Job.getInstance(conf);

// 2 关联本Driver程序的jar

job.setJarByClass(WordCountDriver.class);

// 3 关联Mapper和Reducer的jar

job.setMapperClass(WordCountMapper.class);

job.setReducerClass(WordCountReducer.class);

// 4 设置Mapper输出的kv类型

job.setMapOutputKeyClass(Text.class);

job.setMapOutputValueClass(IntWritable.class);

// 5 设置最终输出kv类型

job.setOutputKeyClass(Text.class);

job.setOutputValueClass(IntWritable.class);

// 6 设置输入和输出路径;args[0]表示传进来的第一个参数，args[1]表示传进来的第二个参数

FileInputFormat.setInputPaths(job, new Path("要统计单词的文件"));

FileOutputFormat.setOutputPath(job, new Path("统计结束后输出的路径"));

// 7 提交job

boolean result = job.waitForCompletion(true);

System.exit(result ? 0 : 1);

}

}

WordCountMapper类：

public class WordCountMapper extends Mapper<LongWritable, Text, Text, IntWritable>{

Text k = new Text();

IntWritable v = new IntWritable(1);

@Override

protected void map(LongWritable key, Text value, Context context)

throws IOException, InterruptedException {

// 1 获取一行

String line = value.toString();

// 2 切割

String[] words = line.split(" ");

// 3 输出

for (String word : words) {

k.set(word);

context.write(k, v);

}

}

}

WordCountReducer类：

public class WordCountReducer extends Reducer<Text, IntWritable, Text, IntWritable>{

int sum;

IntWritable v = new IntWritable();

@Override

protected void reduce(Text key, Iterable<IntWritable> values,Context context)

throws IOException, InterruptedException {

// 1 累加求和

sum = 0;

for (IntWritable count : values) {

sum += count.get();

}

// 2 输出

v.set(sum);

context.write(key,v);

}

}

## 简答题（注意：只有图形作为答案是会扣分的！）

### Checkpoint 的工作流程。

SecondaryName会定期地创建命名空间的检查点（CheckPoint）操作：把edits中最新的状态信息合并到fsimage文件中，防止edits过大。也可以做冷备，对一定范围内数据做快照性备份。

SecondaryNameNode的CheckPoint流程∶

（1）NameNode生成一个名叫edits.new的文件，用于记录合并过程中产生的日志信息；

（2）SecondaryNameNode将edits文件，与fsimag文件从NameNode上下载到SecondNamenode上；

（3）SecondaryNameNode将edits文件与fsimage进行合并操作，合并成一个fsimage.ckpt文件；

（4）SecondaryNameNode将生成的合并后的文件fsimage.ckpt文件上传到NameNode上；

（5）在NameNode上，将fsimage.ckpt变成fsimage文件替换NameNode上原有的fsinage文件。将edits.new文件上变成edits文件替换NameNode上原有的edits文件。

### HDFS 的体系结构。

HDFS采用了主从（Master/Slave）结构模型，一个HDFS集群是由一个NameNode和若干个DataNode组成的。其中NameNode 作为主服务器，管理文件系统的命名空间和客户端对文件的访问操作；

集群中的DataNode管理存储的数据。HDFS允许用户以文件的形式存储数据。从内部来看，文件被分成若干个数据块，而且这若干个数据块存放在一组DataNode上。

NameNode执行文件系统的命名空间操作，比如打开、关闭、重命名文件或目录等，它也负责数据块到具体DataNode的映射。DataNode负责处理文件系统客户端的文件读写请求，并在NaneNode的统一调度下进行数据块的创建、删除和复制工作。NameNode和DataNode都被设计成可以在普通商用计算机上运行。这些计算机通常运行的是GNU/Linux操作系统。HDFS采用Java语言开发，因此任何支持Java的机器都可以部署NameNode和DataNode。一个典型的部署场景是集群中的一台机器运行一个NaneNode实例，其他机器分别运行一个 DataNode实例。当然，并不排除一台机器运行多个DataNode实例的情况。集群中单一的 NameNode的设计则大大简化了系统的架构。NameNode是所有HDFS元数据的管理者，用户数据永远不会经过NameNode。

### 理解并掌握整个 MapReduce 工作的流程。 （存疑）

1．分片、格式化数据源

输入Map阶段的数据源，必须经过分片和格式化操作。其中：

分片操作：指的是将源文件划分为大小相等的小数据块（Hadoop2.x中默认128M），也就是分片(split)，Hadoop会为每一个分片构建一个Map任务，并由该任务运行自定义的map()函数，从而处理分片里的每一条记录；格式化操作：将划分好的分片（split）格式化为键值对<key，value>形式的数据，其中，key代表偏移量，value代表每一行内容。

2．执行MapTask

每个Map任务都有一个内存缓冲区（缓冲区大小100M），输入的分片（split）数据经过Map任务处理后的中间结果，会写入内存缓冲区中。如果写入的数据达到内存缓冲的阀值（80M），会启动一个线程将内存中的溢出数据写入磁盘，同时不影响map中间结果继续写入缓冲区。在溢写过程中，MapReduce框架会对Key进行排序，如果中间结果比较大，会形成多个溢写文件，最后的缓冲区数据也会全部溢写入磁盘形成一个溢写文件，如果是多个溢写文件，则最后合并所有的溢写文件为一个文件。

3．执行Shuffle过程

MapReduce工作过程中，map阶段处理的数据如何传递给Reduce阶段，这是MapReduce框架中关键的一个过程，这个过程叫做Shuffle。Shuffle会将MapTask输出的处理结果数据，分发给ReduceTask，并在分发的过程中，对数据按key进行分区和排序。

4．执行ReduceTask

输入ReduceTask的数据流是<key，{value list}>形式，用户可以自定义reduce()方法进行逻辑处理，最终以<key，value>的形式输出。

5．写入文件

MapReduce框架会自动把ReduceTask生成的<key,value>传入OutputFormat的write方法，实现文件的写入操作。

### MapReduce 的编程模型，理解各个 k 和 v 的值及其数据类型。

MapReduce 的编程模型主要包括Mapper 和 Reducer 两部分，Mapper 负责切分，即把复杂的任务分解为若干个简单，Mapper 负责切分，即把复杂的任务分解为若干个简单。

### MapReduce 1 与 MapReduce 2 的区别

MapReduce 2：

JobTracker：主节点，单点，负责调度所有的作用和监控整个集群的资源负载。

TaskTracker：从节点，自身节点资源管理和JobTracker进行心跳联系，汇报资源和获取task。

Client：以作业为单位，规划作业计算分布，提交作业资源到[HDFS](https://so.csdn.net/so/search?q=HDFS&spm=1001.2101.3001.7020)，最终提交作业到JobTracker。

MapReduce 2:

ResourceManager：主节点，负责整个集群的资源管理。

NodeManager：ResourceManager汇报资源，管理Container生命周期，计算框架中的角色都以Container表示。

Container：默认NodeManager启动线程监控Container大小，超出申请资源额度会kill掉。支持Linux内核的Cgroup。

Client：

ResourceManager-client：请求资源创建ApplicationMaster-client。ApplicationMaster-client：与ApplicationMaster交互。

### NameNode 的元信息具体包括哪些内容？

名称空间目录树结构、文件和块的位置信息、访问权限等信息。

### YARN 三种调度器，并简要说明其工作方法

FIFO Scheduler ，Capacity Scheduler，FairS cheduler

FIFO Scheduler把应用按提交的顺序排成一个队列，这是一个先进先出队列，在进行资源分配的时候，先给队列中最头上的应用进行分配资源，待最头上的应用需求满足后再给下一个分配，以此类推。

Capacity Schedule调度器以队列为单位划分资源，队列以分层方式组织资源,设计了多层级别的资源限制条件以更好的让多用户共享一个Hadoop集群，比如队列资源限制、用户资源限制、用户应用程序数目限制。

在Fair调度器中，我们不需要预先占用一定的系统资源，Fair调度器会为所有运行的job动态的调整系统资源。