一、填空题（将正确的答案写入答题纸中。每空1分，共20分）

1、HDFS有 \_ NameNode\_\_ \_，\_ DataNode\_\_ \_和\_ SecondaryNameNode\_\_ \_组成；

2、Hadoop集群可以运行在\_ 单机模式\_\_ \_，\_ 伪分布式模式\_\_ \_和\_ 完全分布式模式\_\_ \_3个模式？

3、安装完全分布式Apache Hadoop的核心配置文件有\_ core-site.xml\_\_ \_,\_ hdfs-site.xml\_\_ \_,\_ mapred-site.xml\_\_ \_,\_ yarn-site.xml\_\_ \_。

4、HBase中通过\_ HDFS\_\_ \_存储底层数据,通过\_ \_Zookeeper\_ \_提供消息通信机制

5、HIVE默认采用\_ \_ Derby \_ \_\_ \_\_ \_数据库进行元数据的存储，而真实的数据是存储在\_ \_\_HDFS \_\_ \_\_ \_中。

6、如何在浏览器中查看HDFS运行状况的默认端口号是\_ 50070\_\_ \_查看yarn运行状态的默认端口号是\_ 8088\_\_ \_

7、HDFS中当前block大小为128M,如果当前要上传到HDFS中的文件大小为300M，那么在存储时会分配\_ 3\_\_ \_个block进行存储。

8、HDFS 获取根目录下的文件列表的命令shell命令是\_hadoop fs -ls -R / \_\_ \_\_ \_\_ \_\_ \_\_ \_；根目录下创建hdfs文件夹的shell命令是\_ \_ hadoop fs -mkdir /hdfs\_\_ \_\_ \_\_ \_\_ \_ 。

9、YARN架构中整个集群同一时间提供服务的ResourceManager有\_ 1\_\_ \_个，负责集群资源的统一管理和调度。

二、选择题（将正确的答案写入答题纸中。每题2分，共20分）

1、配置Hadoop时，JAVA\_HOME包含在哪一个配置文件中(B)

A． hadoop-default.xml

B． hadoop-env.sh

C． hadoop-site.xml

D． configuration.xs

2、**下面哪个程序负责 HDFS 数据存储。** （C）  
A)NameNode   
B)Jobtracker   
C)Datanode   
D)secondaryNameNode

3、下列关于Hadoop API的说法错误的是（A）

A． Hadoop的文件API不是通用的，只用于HDFS文件系统

B． Configuration类的默认实例化方法是以HDFS系统的资源配置为基础的

C． FileStatus对象存储文件和目录的元数据

D． FSDataInputStream是java.io.DataInputStream的子类

4、HDfS 中的 block 默认保存几份? （A）

A)3 份

B)2 份

C)1 份

D)不确定

5、为销售报表展示开发一个MapReduce作业，Mapper输入数据的Key是年份（IntWritable），Value表示商品标识（Text）。下列哪一项决定该Mapper的数据类型？（D）

A. JobConf.setMapInputKeyClass与JobConf.setMapInputValuesClass

B. HADOOP\_MAP\_DATATYPES环境变量

C. 随作业一起提交的mapper-specification.xml文件

D. InputFormat格式类

6、HDFS无法高效存储大量小文件，想让它能处理好小文件，比较可行的改进策略不包括（D）

A. 利用SequenceFile、MapFile、Har等方式归档小文件

B. 多Master设计

C. Block大小适当调小

D. 调大namenode内存或将文件系统元数据存到硬盘里

7、下列哪个程序通常与NameNode 在一个节点启动？（D）

a)SecondaryNameNode

b)DataNode

c)TaskTracker

d)Jobtracker

8、下面与 HDFS 类似的框架是？（C）

(A) NTFS

(B) FAT32

(C) GFS

(D) EXT3

9 、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，默认的最基本的存储单位是64M，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（C）

A．一次写入，少次读写

B．多次写入，少次读写

C．一次写入，多次读写

D．多次写入，多次读写

10、HBase中的批量加载底层使用（A）实现。

A MapReduce

B Hive

C Coprocessor

D Bloom Filter

三、简答题（将正确的答案写入答题纸中。每题5分，共20分）

1、简述下HDFS **数据读流程；**

**答：步骤如下，能大致描述清楚流程，没有关键错误即可**

1、跟namenode通信查询元数据（block所在的datanode节点），找到文件块所在的datanode服务器

2、挑选一台datanode（就近原则，然后随机）服务器，请求建立socket流

3、datanode开始发送数据（从磁盘里面读取数据放入流，以packet为单位来做校验）

4、客户端以packet为单位接收，先在本地缓存，然后写入目标文件，后面的block块就相当于是append到前面的block块最后合成最终需要的文件。

2、如何查看hadoop 进程并列出hadoop的几个进程名

答：使 用jps命令可查看hadoop进程。 Namenode, SecondaryNameNode，Datanode ，ResourceManager， NodeManager

3、请简述MapReduce中combiner、partition的作用

答：

combiner：

有时一个map可能会产生大量的输出，combiner的作用是在map端对输出先做一次合并，以减少网络传输到reducer的数量。

注意：mapper的输出为combiner的输入，reducer的输入为combiner的输出。

partition：

把map任务输出的中间结果按照key的范围划分成R份(R是预先定义的reduce任务的个数)，划分时通常使用hash函数，如：hash(key) mod R

这样可以保证一段范围内的key，一定会由一个reduce任务来处理。

4、HBase的检索支持3种方式是哪些？

答：

（1） 通过单个Rowkey访问，即按照某个Rowkey键值进行get操作，这样获取唯一一条记录；

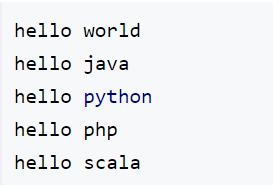
（2） 通过Rowkey的range进行scan，即通过设置startRowKey和endRowKey，在这个范围内进行扫描。这样可以按指定的条件获取一批记录；

（3） 全表扫描，即直接扫描整张表中所有行记录。

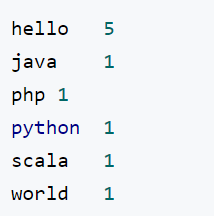
四、程序题（将正确的答案写入答题纸中。每题10分，共20分）

1、以一段文本作为输入对象，写一个WordCount程序。

例如文本数据的情况如下



统计出文本中每个单词出现的次数，输出结果如下面例子所示

这里有营业员们向顾客们示范着制作各种风格炯异的饰品，许多顾客也是学得不亦乐乎。据介绍，经常光顾“碧芝”的都是些希望得到世界上“独一无二”饰品的年轻人，他们在琳琅满目的货架上挑选，然后亲手串连，他们就是偏爱这种ＤＩＹ的方式，完全自助在现场，有上班族在里面精挑细选成品，有细心的小女孩在仔细盘算着用料和价钱，准备自己制作的原料。可以想见，用本来稀奇的原料，加上别具匠心的制作，每一款成品都必是独一无二的。而这也许正是自己制造所能带来最大的快乐吧。

参考文献与网址：

答：

**public class** WordCount {  
  
 **public static class** TokenizerMapper  
 **extends** Mapper<Object, Text, Text, IntWritable> {  
  
 **private final static** IntWritable ***one*** = **new** IntWritable(1);  
 **private** Text **word** = **new** Text();  
  
 **public void** map(Object key, Text value, Context context  
 ) **throws** IOException, InterruptedException {  
 StringTokenizer itr = **new** StringTokenizer(value.toString());  
 **while** (itr.hasMoreTokens()) {  
 **word**.set(itr.nextToken());  
 context.write(**word**, ***one***);  
 }  
 }  
 }  
  
 **public static class** IntSumReducer  
 **extends** Reducer<Text,IntWritable,Text,IntWritable> {  
 **private** IntWritable **result** = **new** IntWritable();  
  
 **public void** reduce(Text key, Iterable<IntWritable> values,  
 Context context  
 ) **throws** IOException, InterruptedException {  
 **int** sum = 0;  
 **for** (IntWritable val : values) {  
 sum += val.get();  
 }  
 **result**.set(sum);  
 context.write(key, **result**);  
 }  
 }  
  
 **public static void** main(String[] args) **throws** Exception {  
 Configuration conf = **new** Configuration();  
 Job job = Job.*getInstance*(conf, **"word count"**);  
 job.setJarByClass(WordCount.**class**);  
 job.setMapperClass(TokenizerMapper.**class**);  
 job.setCombinerClass(IntSumReducer.**class**);  
 job.setReducerClass(IntSumReducer.**class**);  
 job.setOutputKeyClass(Text.**class**);  
 job.setOutputValueClass(IntWritable.**class**);  
 FileInputFormat.*addInputPath*(job, **new** Path(args[0]));  
 FileOutputFormat.*setOutputPath*(job, **new** Path(args[1]));  
 System.*exit*(job.waitForCompletion(**true**) ? 0 : 1);  
 }  
}

10元以下□ 10～50元□ 50～100元□ 100元以上□

根据调查资料分析：大学生的消费购买能力还是有限的，为此DIY手工艺品的消费不能高，这才有广阔的市场。

2、写一段代码，利用Java api操作HDFS文件系统，实现文件的上传和下载，

但这些困难并非能够否定我们创业项目的可行性。盖茨是由一个普通退学学生变成了世界首富，李嘉诚是由一个穷人变成了华人富豪第一人，他们的成功表述一个简单的道理：如果你有能力，你可以从身无分文变成超级富豪；如果你无能，你也可以从超级富豪变成穷光蛋。Hadoop文件系统地址为 hdfs://hadoop:8020 ，将本地hadoop.txt文件上传至根目录下的hadoop文件夹中，将HDFS中hadoop文件中的hadoop1.txt 下载到本地

尽管售价不菲，但仍没挡住喜欢它的人来来往往。这里有营业员们向顾客们示范着制作各种风格迥异的饰品，许多顾客也是学得不亦乐乎。在现场，有上班族在里面精挑细选成品，有细心的小女孩在仔细盘算着用料和价钱，准备自己制作的原料。可以想见，用本来稀奇的原料，加上别具匠心的制作，每一款成品都必是独一无二的。而这也许正是自己制造所能带来最大的快乐吧。

**public class** App {  
 **public static final** String ***HDFS\_PATH*** = **"hdfs://hadoop:8020"**;  
  
 **public static void** main(String[] args) **throws** Exception {  
 *//创建FileSystem* Configuration configuration = **new** Configuration();  
 FileSystem fileSystem = FileSystem.*get*(**new** URI(***HDFS\_PATH***), configuration, **"hadoop"**);  
 *//文件上传* Path localPath = **new** Path(**"hadoop.txt"**);  
 Path hdfsPath = **new** Path(**"/hadoop"**);  
 fileSystem.copyFromLocalFile(localPath, hdfsPath);  
 *//文件下载* hdfsPath = **new** Path(**"/hadoop/hadoop1.txt"**);  
 localPath = **new** Path(**"hadoop1.txt"**);  
 fileSystem.copyToLocalFile(hdfsPath, localPath);  
 }  
}

2、价格“适中化”五、分析题（将正确的答案写入答题纸中。每题20分，共20分）

1、有如下一个场景，有一个1G大小的一个文件，里面每一行是一个词，词的大小不超过16字节，内存限制大小是1M，要求返回频数最高的100个词，写出解决问题主要思路。

服饰□ 学习用品□ 食品□ 休闲娱乐□ 小饰品□答：

主要考察map 和reduce的运用分析，能体现出这几点即可得分

Step1：顺序读文件中，对于每个词x，取hash(x)%5000，然后按照该值存到5000个小文件（记为f0 ,f1 ,... ,f4999）中，这样每个文件大概是200k左右，如果其中的有的文件超过了1M大小，还可以按照类似的方法继续往下分，直到分解得到的小文件的大小都不超过1M；

（4） 信息技术优势

Step2：对每个小文件，统计每个文件中出现的词以及相应的频率（可以采用trie树/hash\_map等），并取出出现频率最大的100个词（可以用含100个结点的最小堆），并把100词及相应的频率存入文件，这样又得到了5000个文件；

在上海， 随着轨道交通的发展，地铁商铺应运而生，并且在重要商圈已经形成一定的气候，投资经营地铁商铺逐渐为一大热门。在人民广场地下的迪美购物中心，有一家DIY自制饰品店--“碧芝自制饰品店”

Step3：把这5000个文件进行归并（类似与归并排序）；