# Hadoop试题一

## 单选题

1、Hadoop的作者（）

A：Martin Fowler

B：Kent Beck

C：Doug Cutting

D：James Gosling

2、YARNWebUI默认端口是（）

A：50070

B：8088

C：19888

D：8020

3、HDFS主节点进程名称是（）

A：NameNode

B：DataNode

C：ResourceManager

D：NodeManager

4、Hadoop2.x中HDFS默认Block Size的值是（）

A：32M

B：64M

C：128M

D：256M

5、Hadoop目前支持很多压缩格式，（）默认就支持切分

A：GZIP

B：BZIP2

C：SNAPPY

D：LZO

6、YARN上的MapReduce实体不包括（）

A：NodeManager

B：client

C：ResourceManager

D：JobTracker

7、下列哪个属性是hdfs-site.xml中的配置？

A：dfs.replication

B：fs.defaultFS

C：mapreduce.framework.name

D：yarn.resourcemanager.address

8、如果我们现有一个安装CDH的hadoop集群，在不修改默认配置的情况下存储200个每个200M的文本文件，请问最终会在集群中产生多少个数据块（包括副本）？

A：200

B：40000

C：400

D：1200

9、配置Hadoop时，JAVA\_HOME包含在哪一个配置文件中

A： hadoop-default.xml

B： hadoop-env.sh

C： hadoop-site.xml

D： configuration.xsl

10、在map和reduce函数的输入和输出类型中，必须一致的是（）

A：map的输入和输出

B：reduce的输入和输出

C：map的输入和reduce的输出

D：map的输出和reduce的输入

11、关于SecondaryNameNode哪项（）是正确的？

A：它是NameNode的热备

B：它是内存没有要求

C：他的目的是帮助NameNode合并编辑日志，减少NameNode启动时间

D：SecondaryNameNode应与NameNode部署到一个节点

12、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

13、MapReduce框架提供了一种序列化机制，支持这种序列化的类能够在Map和Reduce过程中充当键或值，以下说法错误的是？

A：实现Writable接口的类

B：实现WritableComparable<T>接口的类可以是值或键

C：Hadoop的基本类型Text并不实现WritableComparable<T>接口

D：键和值的数据类型可以超出Hadoop自身支持的基本类型

14、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128M，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

15、与HDFS类似的框架是（）

A：GFS

B：Spark

C：Flink

D：Hive

## 多选题

1、下列（）属于NameNode的功能

A：提供名称查询服务

B：保存Block，汇报Block信息

C：保存metadata信息

D：metadata信息在启动后会加载到内存

2、以下描述正确的是（）？

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：CombineFileInputFormat是抽象类

D：TextInputFormat的key是指该记录在文件中的行号

3、下列哪种业务场景中，可以直接使用Reducer充当Combiner使用？

A：sum求和

B：max求最大值

C：count求计数

D：avg求平均

4、Namenode在启动时自动进入安全模式，在安全模式阶段，说法正确的是（）

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

5、以下描述正确的是（）

A：输入分片InputSplit其实是对数据的引用

B：MultipleInputs可以设置多个数据源以及它们对应的输入格式

C：可以通过重载isSplitable()方法来避免文件分片

D：ReduceTask需要等到所有的map输出都复制完才进行Merge

## 判断题

1. blocksize只要配置文件中配置了，就不可以修改
2. Hadoop支持数据的随机读写
3. 每个MapTask是一个线程
4. HDFS的NameNode中保存了文件包含哪些block，以及分布在哪些节点上
5. MapReduce适用于PB级别以上的海量数据的离线处理
6. MapReduce计算过程中，相同的key默认情况会被发送到同一个reduce处理
7. HDFS默认有三副本，所以集群上的NameNode是不存在单节点问题的
8. MapJoin会产生Shuffle
9. LZO默认支持分片
10. MapReduce支持读取Oracle的数据进行处理

## 问答题

1. 为什么Hadoop不适合处理小文件
2. HDFS文件读流程
3. YARN的执行流程
4. 请描述使用MapReduce完成ReduceJoin的思路
5. MapReduce中Combiner的作用是什么？不适用于什么场景