**一、单选题（共10题，每题2分）**

1. Hadoop的组成部分中，负责资源管理和作业调度的是：  
   A. HDFS  
   B. YARN  
   C. MapReduce  
   D. Zookeeper

**答案**: B  
**解析**: YARN是Hadoop的资源管理和作业调度框架。

1. 在Hadoop完全分布式模式下，每个节点上运行的DataNode的作用是：  
   A. 管理文件系统的元数据  
   B. 存储和管理实际数据块  
   C. 处理MapReduce任务  
   D. 负责作业调度

**答案**: B  
**解析**: DataNode负责存储和管理HDFS中的实际数据块。

1. Hadoop安装JDK的最低版本要求是：  
   A. JDK 6  
   B. JDK 7  
   C. JDK 8  
   D. JDK 9

**答案**: C  
**解析**: Hadoop的运行需要至少JDK 8版本。

1. HDFS的副本默认值（dfs.replication）是多少？  
   A. 1  
   B. 2  
   C. 3  
   D. 4

**答案**: C  
**解析**: HDFS默认情况下会将每个数据块存储3个副本。

1. 配置完全分布式模式时，需要在核心配置文件core-site.xml中设置的关键参数是：  
   A. fs.defaultFS  
   B. dfs.replication  
   C. yarn.resourcemanager.hostname  
   D. mapreduce.framework.name

**答案**: A  
**解析**: fs.defaultFS 用于指定Hadoop的文件系统路径。

1. 在Hadoop集群中，通过哪一个组件可以访问HDFS文件系统的元数据？  
   A. DataNode  
   B. NameNode  
   C. ResourceManager  
   D. NodeManager

**答案**: B  
**解析**: NameNode管理和存储HDFS的元数据，负责文件系统的命名空间。

1. SSH免密登录的主要作用是：  
   A. 提高数据传输效率  
   B. 实现各节点之间的安全连接  
   C. 简化节点间的远程登录和管理  
   D. 防止节点间通信被窃听

**答案**: C  
**解析**: SSH免密登录用于简化集群中节点间的操作和管理，不需要每次输入密码。

1. Hadoop伪分布式模式主要用于：  
   A. 大规模数据处理  
   B. 集群性能优化  
   C. 开发和测试环境  
   D. 提高集群的安全性

**答案**: C  
**解析**: 伪分布式模式在单台机器上运行所有Hadoop组件，主要用于开发和测试。

1. 在集群配置中，如果要让节点之间的时间保持一致，通常使用的工具是：  
   A. NTP  
   B. SSH  
   C. Zookeeper  
   D. DFS

**答案**: A  
**解析**: NTP（Network Time Protocol）用于同步集群中所有节点的时间。

1. 下面哪个是启动Hadoop集群的命令？  
   A. start-all.sh  
   B. hadoop-start.sh  
   C. dfs-format.sh  
   D. stop-all.sh

**答案**: A  
**解析**: start-all.sh 用于启动HDFS和YARN集群服务。

**二、填空题（共10题，每题1分）**

1. Hadoop的文件系统（HDFS）中，数据存储的最小单位是\_\_\_。  
   **答案**: block
2. Hadoop完全分布式模式通常需要至少\_\_\_台机器。  
   **答案**: 三
3. 在HDFS配置文件中，负责定义副本数的参数是\_\_\_。  
   **答案**: dfs.replication
4. Hadoop运行环境搭建中，虚拟机的操作系统推荐使用\_\_\_（系统）。  
   **答案**: Linux
5. NameNode是HDFS架构中的\_\_\_节点，负责管理元数据。  
   **答案**: 主
6. 完全分布式模式下，集群的每个节点都需要运行\_\_\_服务，用于存储数据。  
   **答案**: DataNode
7. Hadoop运行在\_\_\_（模式）时，所有组件都运行在本地计算机上。  
   **答案**: 本地模式
8. 集群配置中的核心文件包括core-site.xml、hdfs-site.xml和\_\_\_。  
   **答案**: yarn-site.xml
9. Hadoop中的start-dfs.sh命令用于启动\_\_\_服务。  
   **答案**: HDFS
10. 在完全分布式模式下，SSH免密登录可以通过\_\_\_命令生成密钥对。  
    **答案**: ssh-keygen

**三、判断题（共10题，每题1分）**

1. Hadoop的伪分布式模式适合用于生产环境的大规模数据处理。  
   **答案**: 错  
   **解析**: 伪分布式模式主要用于开发和测试，而非生产环境。
2. 在Hadoop的本地模式下，不需要使用SSH免密登录。  
   **答案**: 对  
   **解析**: 本地模式下所有操作都在单台计算机上进行，不需要SSH配置。
3. HDFS中的数据块大小可以通过参数dfs.blocksize来设置。  
   **答案**: 对
4. 集群中的每个节点都必须安装相同版本的Hadoop。  
   **答案**: 对  
   **解析**: 集群中所有节点应保持一致的软件版本，以确保正常通信和操作。
5. Hadoop集群在配置SSH免密登录后，可以简化节点之间的操作。  
   **答案**: 对
6. 使用hdfs dfs -put命令可以将本地文件上传到HDFS。  
   **答案**: 对
7. Hadoop完全分布式模式只能在本地单机环境中运行。  
   **答案**: 错  
   **解析**: 完全分布式模式是在多台机器上运行的。
8. NameNode的崩溃会导致整个HDFS系统无法正常工作。  
   **答案**: 对  
   **解析**: NameNode管理HDFS的元数据，崩溃会影响整个文件系统的工作。
9. Hadoop运行时必须依赖Zookeeper来管理任务调度。  
   **答案**: 错  
   **解析**: Hadoop不强制依赖Zookeeper，只有在高可用配置时才需要。
10. stop-all.sh命令用于停止Hadoop集群中的所有服务。  
    **答案**: 对

**四、分析计算题（共3题，每题5分）**

1. 在一个Hadoop集群中，如果集群的副本数设置为2，每个块大小为128MB，假设需要存储4TB的数据，那么该集群至少需要多少存储空间？  
   **答案**:

* 4TB = 4000GB
* 副本数为2，因此存储需求为 4000GB × 2 = 8000GB = 8TB。

1. 在Hadoop中，有一个大小为512MB的文件，HDFS的块大小为128MB，计算该文件需要分成多少块来存储？  
   **答案**:

* 512MB ÷ 128MB = 4块。  
  因此，该文件需要分成4块存储在HDFS中。

1. 假设Hadoop集群有10个DataNode，每个DataNode可以并行运行4个Map任务，共有200个Map任务，假设每个任务运行时间为10分钟，计算完成所有任务所需的时间。  
   **答案**:

* 10个DataNode × 每个DataNode 4个任务 = 每批次可以并行运行40个任务。
* 200个任务 ÷ 40个任务/批次 = 5批次。
* 每批次运行时间为10分钟，因此总时间 = 5批次 × 10分钟 = 50分钟。

**五、简答题（共4题，每题5分）**

1. 简述Hadoop集群中配置SSH免密登录的步骤。  
   **答案**:
2. 使用ssh-keygen命令生成SSH密钥对。
3. 将生成的公钥复制到远程节点的~/.ssh/authorized\_keys文件中。
4. 测试免密登录是否成功，例如使用ssh命令登录远程节点，确保无需输入密码。
5. Hadoop伪分布式模式和完全分布式模式有何区别？  
   **答案**:

* 伪分布式模式：所有Hadoop组件在同一台机器上运行，用于开发、调试和测试。
* 完全分布式模式：Hadoop组件分布在多台机器上运行，适用于大规模数据处理和生产环境。

1. 简述Hadoop集群中如何配置NameNode的高可用性。  
   **答案**:
2. 使用Zookeeper作为分布式协调器，保证NameNode的高可用性。
3. 设置一个主NameNode和一个备用NameNode。
4. 配置Zookeeper和NameNode，使得备用NameNode可以接管主NameNode的任务，当主NameNode宕机时保证HDFS的正常运行。
5. 请简述Hadoop集群的启动和停止命令的执行流程。  
   **答案**:
6. 使用start-dfs.sh启动HDFS，启动NameNode和DataNode服务。
7. 使用start-yarn.sh启动YARN，启动ResourceManager和NodeManager服务。
8. 停止集群时，使用stop-dfs.sh停止HDFS服务，使用stop-yarn.sh停止YARN服务。