**单选题（共10题，每题2分）**

1. Hadoop的三大发行版本包括：Apache、Cloudera 和哪一个？  
   A. Amazon  
   B. Hortonworks  
   C. Microsoft  
   D. IBM

**答案**: B  
**解析**: Apache、Cloudera和Hortonworks是Hadoop的三大发行版本。

1. 下列哪一项是Hadoop的核心组件之一？  
   A. Hive  
   B. Spark  
   C. HDFS  
   D. Pig

**答案**: C  
**解析**: Hadoop的核心组件包括HDFS、Yarn和MapReduce。

1. 在Hadoop中，NameNode的主要职责是？  
   A. 存储数据块  
   B. 处理MapReduce任务  
   C. 管理文件系统的元数据  
   D. 负责作业调度

**答案**: C  
**解析**: NameNode主要负责管理文件系统的元数据。

1. 以下哪个命令可以格式化Hadoop的NameNode？  
   A. hadoop format namenode  
   B. hdfs namenode -format  
   C. yarn format namenode  
   D. hadoop namenode -start

**答案**: B  
**解析**: hdfs namenode -format 用于格式化Hadoop的NameNode。

1. 配置Hadoop伪分布式运行模式时，通常需要修改的配置文件包括：  
   A. hdfs-site.xml 和 core-site.xml  
   B. yarn-site.xml 和 mapred-site.xml  
   C. hadoop-env.sh  
   D. 以上都是

**答案**: D  
**解析**: 配置伪分布式模式时，所有提到的文件都需要修改。

1. Hadoop完全分布式运行模式的主要特点是？  
   A. 在单个节点上模拟分布式运行  
   B. 各个节点在不同的机器上运行  
   C. 仅在本地机器上运行  
   D. 不支持大数据集处理

**答案**: B  
**解析**: 完全分布式模式是指各个节点在不同机器上运行，以处理大规模数据集。

1. Hadoop伪分布式模式的默认端口是多少？  
   A. 50070  
   B. 8020  
   C. 9000  
   D. 8088

**答案**: C  
**解析**: Hadoop伪分布式模式中，HDFS的默认端口是9000。

1. 配置Hadoop集群时，SSH免密登录的主要目的是？  
   A. 增强集群的安全性  
   B. 提高数据传输速度  
   C. 简化节点间的操作管理  
   D. 加快作业调度的速度

**答案**: C  
**解析**: SSH免密登录的目的是简化集群中各节点之间的操作管理。

1. Hadoop中哪个组件负责作业调度？  
   A. HDFS  
   B. YARN  
   C. Hive  
   D. Sqoop

**答案**: B  
**解析**: YARN负责资源管理和作业调度。

1. Hadoop伪分布式模式与完全分布式模式的主要区别是什么？  
   A. 是否需要修改配置文件  
   B. 是否支持大规模数据处理  
   C. 是否在同一台机器上运行所有组件  
   D. 是否需要安装JDK

**答案**: C  
**解析**: 伪分布式模式在同一台机器上运行所有组件，而完全分布式模式在多个节点上运行。

**二、填空题（共10题，每题1分）**

1. Hadoop运行环境需要依赖\_\_\_（软件）来保证其正常运行。  
   **答案**: JDK
2. Hadoop中，存储数据的核心组件是\_\_\_。  
   **答案**: HDFS
3. 在完全分布式模式下，每个节点上运行的资源管理组件是\_\_\_。  
   **答案**: NodeManager
4. Hadoop中NameNode的高可用配置使用了\_\_\_来避免单点故障。  
   **答案**: Zookeeper
5. 在Hadoop中，\_\_\_文件保存了HDFS的元数据。  
   **答案**: fsimage
6. 伪分布式运行模式是一种在\_\_\_（台数）机器上模拟分布式环境的模式。  
   **答案**: 一台
7. Hadoop集群的启动和停止通常通过\_\_\_（命令）来完成。  
   **答案**: start-dfs.sh / stop-dfs.sh
8. Hadoop完全分布式运行模式通常需要至少\_\_\_（数量）台计算机。  
   **答案**: 三
9. Hadoop中的任务调度依赖于\_\_\_框架。  
   **答案**: YARN
10. 配置免密登录需要生成\_\_\_密钥对。  
    **答案**: SSH

**三、判断题（共10题，每题1分）**

1. Hadoop伪分布式运行模式主要用于模拟和测试集群环境。  
   **答案**: 对
2. 在Hadoop集群中，NameNode是分布式存储的数据节点。  
   **答案**: 错  
   **解析**: NameNode管理元数据，而DataNode才是存储数据的节点。
3. Hadoop可以运行在本地模式下，但不支持大规模数据处理。  
   **答案**: 对
4. SSH免密登录配置仅限于完全分布式模式下使用。  
   **答案**: 错  
   **解析**: 免密登录也可以用于伪分布式模式下的测试和操作。
5. Hadoop可以在虚拟机环境中运行，但需要配置合适的内存和CPU。  
   **答案**: 对
6. MapReduce框架中的Reduce任务可以省略。  
   **答案**: 对  
   **解析**: 有些情况下，Map任务就足够处理数据，Reduce任务是可选的。
7. HDFS默认的块复制系数是2。  
   **答案**: 错  
   **解析**: 默认的复制系数是3。
8. YARN仅负责资源管理，不负责任务调度。  
   **答案**: 错  
   **解析**: YARN负责资源管理和任务调度。
9. 配置Hadoop集群时，所有节点都必须安装相同版本的Hadoop。  
   **答案**: 对
10. 完全分布式模式中，每个节点可以独立运行HDFS、YARN和MapReduce任务。  
    **答案**: 对

**四、分析计算题（共3题，每题5分）**

1. 假设有一个完全分布式Hadoop集群，由5个DataNode节点组成，每个DataNode上存储的数据块副本数为3。若需要存储600GB的数据，计算该数据在集群中的存储需求量。  
   **答案**:

* 副本数为3倍，存储600GB的数据需要的总存储量 = 600GB \* 3 = 1800GB。

1. Hadoop集群中，每个块的大小设置为128MB，一个2GB的文件需要分成多少块来存储？  
   **答案**:

* 2GB = 2048MB
* 2048MB ÷ 128MB = 16块。  
  因此，该文件需要分成16个块来存储。

1. 如果在完全分布式模式下，有8台机器，每台机器可以并行处理2个Map任务，共有80个Map任务，计算完成所有Map任务所需的最短时间，假设每个任务的运行时间为15分钟。  
   **答案**:

* 每台机器可以并行运行2个Map任务，8台机器总共可以并行运行16个任务。
* 80个任务需要分5批执行，每批15分钟，总时间为15分钟 \* 5批 = 75分钟。

**五、简答题（共4题，每题5分）**

1. 请简述Hadoop伪分布式模式和完全分布式模式的区别。  
   **答案**:

* 伪分布式模式是在单台机器上运行Hadoop的所有组件，用于开发和测试。
* 完全分布式模式是在多台机器上运行Hadoop的各个组件，适用于处理大规模数据。

1. 如何配置Hadoop集群中的SSH免密登录？  
   **答案**:

* 使用ssh-keygen命令生成公钥和私钥。
* 将公钥复制到远程节点的`~/.ssh/authorized\_keys