**一、单选题（共10题，每题2分）**

1. 大数据的“4V”特征中，以下哪一项不属于大数据的特点？  
   A. 数据量大  
   B. 数据类型多样  
   C. 数据处理速度快  
   D. 数据可靠性高

**答案**: D  
**解析**: 大数据的四个“V”特点是数据量大（Volume）、速度快（Velocity）、种类多（Variety）和价值密度低（Veracity）。数据可靠性并不是大数据的显著特点。

1. Hadoop的核心组件不包括以下哪一个？  
   A. HDFS  
   B. MapReduce  
   C. YARN  
   D. Spark

**答案**: D  
**解析**: Hadoop的核心组件是HDFS、MapReduce和YARN，Spark是Hadoop生态系统中的一个计算引擎，但并非Hadoop的核心组件。

1. Hadoop的作者是：  
   A. Linus Torvalds  
   B. Doug Cutting  
   C. Elon Musk  
   D. Bill Gates

**答案**: B  
**解析**: Doug Cutting是Hadoop的创始人，他从Google的MapReduce和GFS论文中得到了灵感，创建了Hadoop。

1. 以下哪一个文件是负责Hadoop集群数据存储的？  
   A. NameNode  
   B. DataNode  
   C. ResourceManager  
   D. NodeManager

**答案**: B  
**解析**: DataNode负责存储实际的数据，而NameNode则负责存储元数据。

1. Hadoop的核心组件HDFS的块大小默认值是：  
   A. 64MB  
   B. 128MB  
   C. 256MB  
   D. 512MB

**答案**: B  
**解析**: Hadoop 2.x版本的默认块大小是128MB，之前的1.x版本默认是64MB。

1. 在Hadoop安装过程中，JAVA\_HOME环境变量通常设置在哪个配置文件中？  
   A. hadoop-default.sh  
   B. hadoop-config.sh  
   C. hadoop-env.sh  
   D. hadoop-site.sh

**答案**: C  
**解析**: hadoop-env.sh是用于设置Hadoop环境变量的文件，其中包括设置JAVA\_HOME。

1. 下列哪个选项是默认情况下HDFS的数据副本数量？  
   A. 2  
   B. 3  
   C. 4  
   D. 5

**答案**: B  
**解析**: HDFS的默认副本数为3，以保证数据的高可用性和可靠性。

1. HDFS中的块大小设置主要与以下哪个资源关系最大？  
   A. CPU  
   B. 内存  
   C. 硬盘  
   D. 网络

**答案**: C  
**解析**: 块大小与硬盘资源密切相关，因为HDFS中的数据块是存储在硬盘中的。

1. HDFS的默认Web界面端口是：  
   A. 50070  
   B. 8088  
   C. 9000  
   D. 50075

**答案**: A  
**解析**: HDFS的默认Web管理界面端口是50070。

1. Hadoop的三大发行版本中，不包括以下哪一个？  
   A. Apache Hadoop  
   B. Cloudera  
   C. Hortonworks  
   D. Amazon EMR

**答案**: D  
**解析**: Amazon EMR是一个大数据处理平台，但不是Hadoop的发行版本，Hadoop的主要发行版本是Apache Hadoop、Cloudera和Hortonworks。

**二、填空题（共10题，每题1分）**

1. Hadoop NameNode的默认服务端口是\_\_\_\_\_\_。  
   **答案**: 9000
2. Hadoop HDFS Web界面的默认端口是\_\_\_\_\_\_。  
   **答案**: 50070
3. YARN的Web界面的默认端口是\_\_\_\_\_\_。  
   **答案**: 8088
4. Hadoop的两个主要核心模块是HDFS和\_\_\_\_\_\_。  
   **答案**: MapReduce
5. HDFS的默认块大小为\_\_\_\_\_\_。  
   **答案**: 128MB
6. Hadoop中负责管理资源调度的组件是\_\_\_\_\_\_。  
   **答案**: YARN
7. Hadoop系统运行时，NameNode负责管理\_\_\_\_\_\_。  
   **答案**: 元数据
8. Hadoop的三大发行版本是Apache Hadoop、\_\_\_\_\_\_ 和Hortonworks。  
   **答案**: Cloudera
9. Hadoop中使用的编程模型是\_\_\_\_\_\_。  
   **答案**: MapReduce
10. 在HDFS中，数据块的默认副本数量是\_\_\_\_\_\_。  
    **答案**: 3

**三、判断题（共10题，每题1分）**

1. Hadoop的块大小（Block Size）是可以配置和修改的。  
   A. 对  
   B. 错

**答案**: A  
**解析**: HDFS的块大小可以在配置文件中修改，通常可以根据数据处理需求调整。

1. 1GB等于1024MB。  
   A. 对  
   B. 错

**答案**: A  
**解析**: 1GB等于1024MB，这是计算机存储单位的常见换算。

1. Hadoop的NameNode是单点故障的潜在风险。  
   A. 对  
   B. 错

**答案**: A  
**解析**: Hadoop 1.x版本中，NameNode是单点故障，一旦它崩溃，整个集群将不可用。Hadoop 2.x版本中通过高可用性机制解决了这一问题。

1. 使用hdfs命令可以向HDFS中上传文件。  
   A. 对  
   B. 错

**答案**: A  
**解析**: hdfs dfs -put命令可以将文件从本地上传到HDFS中。

1. MapReduce的任务调度完全由YARN控制。  
   A. 对  
   B. 错

**答案**: A  
**解析**: 在Hadoop 2.x版本中，MapReduce的任务调度由YARN负责。

1. Hadoop只能用于处理结构化数据。  
   A. 对  
   B. 错

**答案**: B  
**解析**: Hadoop可以处理结构化、半结构化和非结构化数据。

1. DataNode存储的是文件的元数据。  
   A. 对  
   B. 错

**答案**: B  
**解析**: DataNode存储的是实际的数据块，元数据由NameNode存储。

1. Hadoop中的HDFS是一种分布式文件系统。  
   A. 对  
   B. 错

**答案**: A  
**解析**: HDFS（Hadoop Distributed File System）是一种分布式文件系统，负责存储和管理大规模的数据。

1. Hadoop的副本机制用于提高数据的可靠性和可用性。  
   A. 对  
   B. 错

**答案**: A  
**解析**: HDFS的副本机制通过将每个数据块复制到多个节点上来保证数据的可靠性和可用性。

1. YARN中的ResourceManager负责分配集群中的所有计算资源。  
   A. 对  
   B. 错

**答案**: A  
**解析**: ResourceManager是YARN的核心组件之一，负责全局资源的调度和分配。

**四、分析计算题（共3题，每题5分）**

1. 假设你有一个6TB的大文件需要存储到HDFS中，块大小为128MB，副本数量为3，计算需要多少个块，HDFS需要存储多少数据量？  
   **答案**:

* 文件大小为6TB = 6000GB = 6000 \* 1024MB = 6144000MB，块数 = 6144000MB / 128MB = 48000个块。
* HDFS需要存储的数据量 = 48000块 \* 3个副本 \* 128MB = 18TB。

1. 在MapReduce中，有100个输入文件，每个文件的大小为1GB，假设Map任务处理每个文件的时间为15分钟，计算完成所有Map任务的总时间。  
   **答案**:

* 总共有100个输入文件，Map任务处理时间为15分钟。假设同时运行所有Map任务，完成所有任务的总时间就是15分钟。
* 如果每次只能运行10个任务，处理100个任务需要分10批，总时间为15分钟 \* 10批 = 150分钟。

1. 假设你的集群有50个节点，每个节点可以运行2个Reduce任务，你的Reduce任务数为200，计算完成所有Reduce任务的最小时间。  
   **答案**:

* 每个节点可以同时运行2个Reduce任务，50个节点总共可以并行运行100个任务。
* Reduce任务数为200，分为两批运行，每批完成时间假设为30分钟，总时间为30分钟 \* 2批 = 60分钟。

**五、简答题（共4题，每题5分）**

1. 简述HDFS中NameNode和DataNode的作用。  
   **答案**:

* NameNode负责存储元数据，包括文件名、目录结构、文件块的位置等。
* DataNode负责存储实际的数据块，并定期向NameNode汇报其存储的块信息。

1. Hadoop中的MapReduce是如何工作的？  
   **答案**:

* MapReduce将任务分为两个阶段：Map阶段和Reduce阶段。
* Map阶段负责将输入数据分解成键值对并进行初步处理。
* Reduce阶段负责对Map阶段的输出进行汇总和处理，最终生成结果。

1. 什么是YARN？简述其作用。  
   **答案**:

* YARN（Yet Another Resource Negotiator）是Hadoop的资源管理框架。
* 它负责资源的分配和任务的调度，确保集群中的资源得到高效利用。

1. 大数据的主要应用场景有哪些？  
   **答案**:

* 大数据主要应用于电商推荐、医疗数据分析、交通预测、金融风控等领域。通过大数据技术，能够分析海量数据，挖掘出潜在的商业价值。