**选择题（10题，每题1分）**

1. 大数据的四个“V”特点不包括以下哪一项？ A. Volume（数据量大）  
   B. Velocity（速度快）  
   C. Variety（种类多）  
   D. Value（低价值）

**答案**: D  
**解析**: 大数据的四个特点是数据量大（Volume）、速度快（Velocity）、种类多（Variety）、价值密度低（Veracity）。

1. 以下哪一项不是Hadoop的核心组件？ A. HDFS  
   B. YARN  
   C. Hive  
   D. MapReduce

**答案**: C  
**解析**: Hive是Hadoop生态系统中的数据仓库工具，不是Hadoop的核心组件，HDFS、YARN和MapReduce才是核心部分。

1. Hadoop的默认块大小为： A. 32MB  
   B. 64MB  
   C. 128MB  
   D. 256MB

**答案**: C  
**解析**: Hadoop 2.x版本及以后的默认块大小为128MB，Hadoop 1.x版本为64MB。

1. 在HDFS中，一个文件默认会有多少份副本？ A. 1  
   B. 2  
   C. 3  
   D. 4

**答案**: C  
**解析**: HDFS默认将每个文件分为3个副本存储，以保证数据的高可用性。

1. MapReduce的两个阶段是： A. Map和Reduce  
   B. Split和Combine  
   C. Input和Output  
   D. Parse和Execute

**答案**: A  
**解析**: MapReduce由两个阶段组成，分别是Map阶段和Reduce阶段，负责数据处理和汇总。

1. 在Hadoop中，负责管理集群资源的组件是： A. HDFS  
   B. YARN  
   C. MapReduce  
   D. Hive

**答案**: B  
**解析**: YARN是Hadoop集群中的资源管理框架，它协调和管理计算资源。

1. HDFS中的文件以块的形式存储，这些块通常存储在： A. 主节点  
   B. 从节点  
   C. 客户端  
   D. 数据库

**答案**: B  
**解析**: HDFS中的块被存储在从节点（DataNode）上。

1. Hadoop的三大发行版本不包括以下哪一项？ A. Apache Hadoop  
   B. Cloudera  
   C. Hortonworks  
   D. Google BigQuery

**答案**: D  
**解析**: Apache Hadoop、Cloudera和Hortonworks都是Hadoop的发行版本，Google BigQuery是另一种大数据处理技术，不属于Hadoop发行版。

1. HDFS是如何保证数据的可靠性的？ A. 通过数据压缩  
   B. 通过数据副本  
   C. 通过数据加密  
   D. 通过数据分片

**答案**: B  
**解析**: HDFS通过在不同节点上创建多个副本来保证数据的可靠性，默认情况下是三个副本。

1. 以下哪个场景最适合使用大数据技术？ A. 个人银行交易记录管理  
   B. 社交媒体数据分析  
   C. 小型企业库存管理  
   D. 单个计算任务的实时执行

**答案**: B  
**解析**: 大数据技术适用于处理大量、多样化、快速变化的数据，社交媒体数据分析是典型的应用场景。

**填空题（10题，每题1分）**

1. Hadoop的核心组件包括\_\_\_\_\_\_、YARN和MapReduce。  
   **答案**: HDFS
2. 大数据通常被归结为四个“V”特点，分别是Volume、Velocity、Variety和\_\_\_\_\_\_。  
   **答案**: Veracity
3. Hadoop 2.x版本引入了\_\_\_\_\_\_来管理集群资源。  
   **答案**: YARN
4. HDFS的默认块大小是\_\_\_\_\_\_。  
   **答案**: 128MB
5. 在MapReduce中，负责将输入数据切分为小块并进行处理的阶段是\_\_\_\_\_\_。  
   **答案**: Map
6. YARN的全称是\_\_\_\_\_\_。  
   **答案**: Yet Another Resource Negotiator
7. 大数据技术在\_\_\_\_\_\_、金融、医疗等领域有广泛的应用。  
   **答案**: 电商
8. HDFS中的文件会被切分成多个\_\_\_\_\_\_，然后存储在不同的节点上。  
   **答案**: 块（Block）
9. MapReduce的输出通常存储在\_\_\_\_\_\_中。  
   **答案**: HDFS
10. 大数据技术常见的一个应用场景是\_\_\_\_\_\_，通过分析用户行为模式做出预测。  
    **答案**: 推荐系统

**判断题（10题，每题1分）**

1. Hadoop的核心组件是HDFS、MapReduce和Hive。  
   **答案**: 错  
   **解析**: Hive是Hadoop生态系统中的一个数据仓库工具，而非核心组件。
2. 大数据的“4V”特点包括数据量大、速度快、种类多、真实性高。  
   **答案**: 错  
   **解析**: 大数据的四个特点是数据量大（Volume）、速度快（Velocity）、种类多（Variety）、价值密度低（Veracity），不是“真实性高”。
3. YARN可以同时管理多个应用程序的计算资源。  
   **答案**: 对  
   **解析**: YARN是Hadoop的资源管理器，负责管理集群中多个应用程序的资源分配。
4. Hadoop最早由Google开发，用于内部的搜索引擎。  
   **答案**: 错  
   **解析**: Hadoop的开发灵感来源于Google的论文，但它由Doug Cutting等人开发。
5. HDFS中的每个文件会被切成多个块并分布存储在不同节点上。  
   **答案**: 对  
   **解析**: HDFS通过将文件切分为块，分布存储在不同的DataNode上，以保证数据的可靠性。
6. MapReduce算法只能处理结构化数据。  
   **答案**: 错  
   **解析**: MapReduce算法可以处理结构化、非结构化和半结构化数据。
7. YARN引入了资源管理和任务调度的分离。  
   **答案**: 对  
   **解析**: YARN将资源管理和任务调度分开，提升了集群的可扩展性。
8. Hadoop默认的文件复制策略是将文件复制到集群中最近的节点。  
   **答案**: 对  
   **解析**: Hadoop的副本放置策略通常会优先考虑物理拓扑距离，以提高效率。
9. 完全分布式运行模式意味着Hadoop在单台机器上运行所有服务。  
   **答案**: 错  
   **解析**: 完全分布式模式是指Hadoop在多台机器上分布式运行各个服务。
10. 在HDFS中，NameNode负责存储文件的实际数据。  
    **答案**: 错  
    **解析**: NameNode负责存储文件的元数据，而实际数据由DataNode存储。

**分析计算题（3题，每题5分）**

1. 假设你有一个5TB的大文件需要存储到HDFS中，块大小为128MB，HDFS默认使用3个副本。计算该文件需要多少个块，HDFS需要存储多少数据量？ **答案**: 5TB = 5000GB = 5000 \* 1024MB，块数 = 5000 \* 1024 / 128 = 40000块。由于每个块存储3个副本，因此HDFS需要存储的总数据量为40000 \* 128MB \* 3 = 15TB。
2. MapReduce处理一个1TB的数据集，输入分为1024个块，每个块的大小是1GB。如果每个Map任务的处理时间是10分钟，计算完成Map阶段需要的总时间。 **答案**: 每个Map任务处理1个块，总共有1024个块。每个任务处理时间为10分钟，因此Map阶段总时间为10分钟。
3. 假设有一个集群运行YARN，总共有100个节点，每个节点可以运行8个Map任务和8个Reduce任务。计算该集群可以同时运行的最大Map和Reduce任务数量。 **答案**: 每个节点可以运行8个Map任务和8个Reduce任务，因此整个集群可以同时运行的Map任务数是100 \* 8 = 800个，Reduce任务数也是800个。