**一、单选题（共10题，每题2分）**

1. 在大数据架构中，哪个组件负责作业调度和资源管理？  
   A. HDFS  
   B. MapReduce  
   C. YARN  
   D. Hive

**答案**: C  
**解析**: YARN（Yet Another Resource Negotiator）负责Hadoop集群的资源管理和作业调度。

1. 下列哪一项是HDFS中的核心组件？  
   A. DataNode和MapReduce  
   B. NameNode和DataNode  
   C. NameNode和YARN  
   D. HDFS和MapReduce

**答案**: B  
**解析**: HDFS由两个核心组件组成，NameNode负责元数据存储，DataNode负责数据存储。

1. 大数据技术应用的基础是什么？  
   A. 数据库管理系统  
   B. 分布式存储与计算  
   C. 云存储  
   D. 关系型数据库

**答案**: B  
**解析**: 大数据技术的基础是分布式存储和计算，能够处理海量数据。

1. Hadoop中负责数据存储的文件系统是？  
   A. NFS  
   B. HDFS  
   C. FAT32  
   D. NTFS

**答案**: B  
**解析**: HDFS（Hadoop Distributed File System）是Hadoop用于存储大数据的分布式文件系统。

1. YARN的两个主要组件是？  
   A. ResourceManager和NodeManager  
   B. NameNode和ResourceManager  
   C. DataNode和NodeManager  
   D. TaskManager和NodeManager

**答案**: A  
**解析**: YARN的两个主要组件是ResourceManager（负责全局资源管理）和NodeManager（负责每个节点上的资源管理）。

1. 下列哪种类型的数据不适合使用Hadoop处理？  
   A. 结构化数据  
   B. 半结构化数据  
   C. 非结构化数据  
   D. 低延迟的实时数据

**答案**: D  
**解析**: Hadoop适合处理大批量的数据，而低延迟的实时数据不太适合使用Hadoop。

1. HDFS的默认块大小为多少？  
   A. 32MB  
   B. 64MB  
   C. 128MB  
   D. 256MB

**答案**: C  
**解析**: HDFS的默认块大小为128MB。

1. 哪个组件在MapReduce框架中负责分配任务和监控进度？  
   A. NameNode  
   B. ResourceManager  
   C. JobTracker  
   D. TaskTracker

**答案**: C  
**解析**: JobTracker在旧的MapReduce框架中负责分配任务并监控作业的进度。

1. HDFS中将大文件分为多个小块存储的好处是？  
   A. 提高系统吞吐量  
   B. 降低网络带宽消耗  
   C. 简化数据访问  
   D. 增强数据的安全性

**答案**: A  
**解析**: 将大文件分成小块存储可以并行化处理，进而提高系统的吞吐量。

1. 以下哪一个选项是Hive的作用？  
   A. 数据存储  
   B. 数据传输  
   C. 数据查询  
   D. 数据备份

**答案**: C  
**解析**: Hive是一个基于Hadoop的数仓工具，用于数据的存储和查询。

**二、填空题（共10题，每题1分）**

1. HDFS默认的块复制系数是\_\_\_。  
   **答案**: 3
2. 在Hadoop中，MapReduce任务的执行通过\_\_\_（组件）进行资源分配。  
   **答案**: YARN
3. HDFS中的文件块（block）是由\_\_\_（组件）管理的。  
   **答案**: NameNode
4. 在MapReduce中，\_\_\_函数将输入数据分解为键值对。  
   **答案**: Map
5. HDFS是为了处理\_\_\_数据集而设计的。  
   **答案**: 大规模（或海量）
6. YARN中的\_\_\_负责分配集群资源。  
   **答案**: ResourceManager
7. NameNode保存文件系统的\_\_\_数据。  
   **答案**: 元
8. Hadoop集群的时间同步通过\_\_\_工具实现。  
   **答案**: NTP
9. 在Hadoop中，数据首先存储在\_\_\_，然后进行计算。  
   **答案**: HDFS
10. Reduce任务的输入是Map任务的\_\_\_。  
    **答案**: 输出

**三、判断题（共10题，每题1分）**

1. HDFS可以存储结构化数据、半结构化数据和非结构化数据。  
   **答案**: 对
2. 在HDFS中，数据存储不使用副本机制。  
   **答案**: 错
3. MapReduce只能处理结构化数据。  
   **答案**: 错
4. YARN中的NodeManager负责管理每个节点的资源。  
   **答案**: 对
5. Hadoop主要用于实时数据处理场景。  
   **答案**: 错
6. MapReduce的输入数据可以是HDFS中的数据，也可以是本地文件系统中的数据。  
   **答案**: 对
7. Reduce阶段在MapReduce任务中是可选的。  
   **答案**: 对
8. HDFS支持跨数据中心的文件存储。  
   **答案**: 错
9. HDFS允许用户自定义块的大小。  
   **答案**: 对
10. Hadoop生态系统中不包括Pig和Hive。  
    **答案**: 错

**四、分析计算题（共3题，每题5分）**

1. 假设Hadoop集群有10个节点，每个节点可以同时运行4个Map任务。总共有80个Map任务要执行。计算完成所有Map任务的最短时间，假设每个任务的运行时间为20分钟。  
   **答案**:

* 每个节点可以同时运行4个任务，总共可以并行运行40个任务。
* 80个任务分为两批执行，每批20分钟，总时间为20分钟 \* 2批 = 40分钟。

1. 某个Hadoop集群中的HDFS块大小设置为128MB，文件大小为1GB，计算该文件需要分成多少块存储。  
   **答案**:

* 1GB = 1024MB
* 每个块大小为128MB，因此需要的块数为1024MB ÷ 128MB = 8块。

1. 一个集群中有50个节点，每个节点可以运行2个Reduce任务，共有200个Reduce任务需要运行。计算完成这些Reduce任务的最短时间，假设每个任务的执行时间为25分钟。  
   **答案**:

* 每个节点可以同时运行2个任务，50个节点可以并行运行100个任务。
* 200个任务分为两批执行，每批25分钟，总时间为25分钟 \* 2批 = 50分钟。

**五、简答题（共4题，每题5分）**

1. 简述Hadoop中Map任务和Reduce任务的区别。  
   **答案**:

* Map任务负责将输入数据分割成键值对，并进行初步处理。
* Reduce任务负责对Map任务的输出进行汇总和处理，最终生成结果。

1. 什么是Hadoop生态系统？列举其中的三个重要组件。  
   **答案**:

* Hadoop生态系统是指围绕Hadoop分布式框架的工具和软件。
* 重要组件包括：HDFS（分布式存储）、YARN（资源管理）和MapReduce（分布式计算）。

1. 如何配置Hadoop伪分布式模式？  
   **答案**:

* 配置虚拟机并安装JDK。
* 下载并解压Hadoop，修改配置文件如hdfs-site.xml、core-site.xml和yarn-site.xml。
* 通过SSH设置无密码登录并格式化NameNode。
* 启动HDFS和YARN，验证伪分布式环境。

1. 大数据处理与传统数据处理的区别是什么？  
   **答案**:

* 大数据处理适合处理海量、多样性、高速生成的数据。
* 传统数据处理则更适合处理结构化、较小的数据集，通常使用关系型数据库。