**大数据技术基础 第三次平时作业**  
第五章 MapReduce与Spark分布式计算

**1.简述Map Reduce工作原理**  
答：工作原理：  
 1）将大规模数据集分成若干个小数据片段。

2） 将每个小数据片段传给不同的计算节点。

3）计算节点使用Map函数对每个小数据片段进行处理，将每个数据片段转换成多个键值对。

4）将输出的键值对按照键进行排序，以便于后续的Reduce操作。

5）将排序后的键值对传给不同的计算节点。

6）计算节点使用Reduce函数对每个键值对进行处理，将相同键的所有值合并在一起。

7）最终输出Reduce函数的结果。

**2.简述Spark工作原理**

答：Spark工作原理: 首先看中间是一个Spark集群,可以理解为是Spark的 standalone集群,集群中有6个节点 左边是Spark的客户端节点,这个节点主要负责向Spark集群提交任务,假设在这里我们向Spark集群提交了一个任务 那这个Spark任务肯定会有一个数据源,数据源在这我们使用HDFS,就是让Spark计算HDFS中的数据。

第六章  HBase和Hive数据管理

**1.简述HBase工作原理的理解**  
答：HBase的工作原理：

答：1. 数据存储：HBase的数据存储是基于HDFS的分布式文件系统，它将数据分成多个块并存储在不同的服务器上。每个块都有一个唯一的块ID，这个ID可以用来定位块的位置。HBase的数据存储是按照列族和列来组织的，每个列族可以包含多个列，每个列都有一个唯一的列ID。

2. 数据访问： HBase的数据访问是基于行键的，每个行键都对应着一行数据。当用户需要访问一行数据时，HBase会根据行键找到对应的块，并从块中读取数据。HBase支持随机读写和批量读写，用户可以根据自己的需求选择不同的读写方式。

3. 数据复制：HBase的数据复制是基于Hadoop的复制机制，它可以将数据复制到多个服务器上，以提高数据的可靠性和可用性。HBase的数据复制是异步的，当数据发生变化时，HBase会将变化记录到WAL（Write Ahead Log）中，并异步地将变化复制到其他服务器上。

4. 数据一致性：HBase的数据一致性是通过ZooKeeper来实现的，ZooKeeper是一个分布式的协调服务，它可以协调多个服务器之间的状态。当HBase的数据发生变化时，ZooKeeper会通知所有的服务器，以保证数据的一致性。

**2.简述Hive工作原理的理解**答：Web服务器的工作原理：

①连接过程:是Web服务器与其浏览器之间建立的连接。检查连接过程是否实现。用户可以找到并打开虚拟文件套接字。该文件的建立意味着连接过程已经成功建立。

②请求过程:Web浏览器利用socket文件向其服务器发出各种请求。

(3)响应过程:在请求过程中发出的请求通过使用HTTP协议传输到Web服务器，然后执行任务处理。然后，通过使用HTTP协议将任务处理的结果传送到网络浏览器，并且在网络浏览器上显示所请求的界面。

④关闭连接:是最后一步——响应过程完成后，Web服务器与其浏览器断开连接的过程。Web服务器的上述四个步骤联系紧密，逻辑严密，可以支持多进程、多线程以及多进程、多线程混合的技术。