1 大数据的由来

参考答案

随着计算机技术的发展，互联网的普及，信息的积累已经到了一个非常庞大的地步，信息的增长也在不断的加快，随着互联网、物联网建设的加快，信息更是爆炸是增长，收集、检索、统计这些信息越发困难，必须使用新的技术来解决这些问题

2 什么是大数据

参考答案

数据指无法在一定时间范围内用常规软件工具进行捕捉、管理和处理的数据集合，需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产

是指从各种各样类型的数据中，快速获得有价值的信息

3 简述大数据特性有哪些

参考答案

Volume (大体量)：可从数百TB到数十数百PB、甚至EB的规模

Variety (多样性)：大数据包括各种格式和形态的数据

Velocity (时效性)：很多大数据需要在一定的时间限度下得到及时处理

Veracity (准确性)：处理的结果要保证一定的准确性

Value (大价值)：大数据包含很多深度的价值，大数据分析挖掘和利用将带来巨大的商业价值

4 Hadoop常用组件以及核心组件有哪些

参考答案

HDFS：Hadoop分布式文件系统（核心组件）

MapReduce：分布式计算框架（核心组件）

Yarn：集群资源管理系统（核心组件）

Zookeeper：分布式协作服务

Hbase：分布式列存数据库

Hive：基于Hadoop的数据仓库

Sqoop：数据同步工具

Pig：基于Hadoop的数据流系统

Mahout：数据挖掘算法库

Flume：日志收集工具

5 Hadoop如何实现统计词频

参考答案

```
[root@nn01 ~]# cd /usr/local/hadoop/
[root@nn01 hadoop]# mkdir /usr/local/hadoop/aa
[root@nn01 hadoop]# ls
bin etc include lib libexec LICENSE.txt NOTICE.txt aa README.txt sbin share
[root@nn01 hadoop]# cp *.txt /usr/local/hadoop/aa
[root@nn01 hadoop]# ./bin/hadoop jar \
 share/hadoop/mapreduce/hadoop-mapreduce-examples-2.7.6.jar wordcount aa bb
[root@nn01 hadoop]# cat bb/part-r-00000 //查看
```