习题

1. **概述大数据的价值和影响。**
大数据的价值体现在以下几个方面：

1)对许多顾客供给产品或服务的企业可以运用大数据进行精准营销

2) 做小而美形式的中小微企业可以运用大数据做服务转型

3) 面对互联网压力之下有必要转型的传统企业需求与时俱进充沛运用大数据的价值

1. **详细描述大数据的四大核心特征。**
答：
1、数据量大（Volume)

大数据的显而易见的特征就是其庞大的数据规模。随着信息技术的发展，互联网规模的不断扩大，每个人的生活都被记录在了大数据之中，由此数据本身也呈爆发性增长。其中大数据的计量单位也逐渐发展，现如今对大数据的计量已达到EB了。

2、类型多样（Variety)

在数量庞大的互联网用户等因素的影响下，大数据的来源十分广泛，因此大数据的类型也具有多样性。大数据由因果关系的强弱可以分为三种，即结构化数据、非结构化数据、半结构化数据，它们统称为大数据。资料表明，结构化数据在整个大数据中占比较大，高达百分之七十五，但能够产生高价值的大数据却是非结构化数据。

3、价值密度（Value)

大数据所有的价值在大数据的特征中占核心地位，大数据的数据总量与其价值密度的高低关系是成反比的。同时对于任何有价值的信息，都是在处理海量的基础数据后提取的。在大数据蓬勃发展的今天，人们一直探索着如何提高计算机算法处理海量大数据，提取有价值信息的的速度这一难题。

4、高速（Velocity)

大数据的高速特征主要体现在数据数量的迅速增长和处理上。与传统媒体相比，在如今大数据时代，信息的生产和传播方式都发生了巨大改变，在互联网和云计算等方式的作用下，大数据得以迅速生产和传播，此外由于信息的时效性，还要求在处理大数据的过程中要快速响应，无延迟输入、提取数据。

1. **大数据处理中所面临的数据结构类型大多数是结构化数据。判断这句话是否正确，并阐述理由。**

答：定义有误，大数据的多样性使得数据被分为三种数据结构，
分别是：1、结构化数据，是由二维表结构来逻辑表达和实现的数据；2、非结构化数据是数据结构不规则或不完整，没有预定义的数据模型，不方便用数据库二维逻辑表来表现的数据；3、半结构化数据。

1. **Had oop这个项目主要由哪两大部分组成， 这两大部分的子项目分别是什么?**

Hadoop是Apache组织正在推进的项目。这个项目主要由两大部分的子项目构成，一个是基础部分，另一个是配套部分。
一、**基础部分**
（1）Hadoop Common。Hadoop Common是支撑Hadoop的公共部分，包括文件系统、远程过程调用RPC和序列化函数库等。
（2）HDFS。HDFS是可以提供高吞吐量的可靠分布式文件系统，是Google GFS的开源实现。
（3）MapReduce。MapReduce是大型分布式数据处理模型，是Google MapReduce的开源实现。
**二、配套部分**
（1）HBase。HBase是支持结构化数据存储的分布式数据库，是Google BigTable的开源实现。
（2）Hive。Hive是提供数据摘要和查询功能的数据仓库。
（3）Pig。Pig是在MapReduce上构建的一种脚本式开发方式，大大简化了MapReduce的开发工作。
（4）Cassandre。Cassandre是由Facebook支持的开源、高可扩展分布式数据库，是Amazon库层架构Dynamo的全分布和Google BigTable的列式数据存储模型的有机结合。
（5）Chukwa。Chukwa是用来管理大型分布式系统的[数据](https://link.zhihu.com/?target=http://cda.pinggu.org/view/22342.html" \t "https://zhuanlan.zhihu.com/p/_blank)采集系统。
（6）Zookeeper。Zookeeper用于解决分布式系统中一致性问题，是Google Chubby的开源实现。

5**.在大数据如火如荼的今天，企业面临哪些挑战?简要阐述企业对大数据要采取什么样的态度?**

答：**企业面临以下5个挑战：**1、跨不同数据源的同步

2、处理大量数据

3、隐私权与安全性

4、大数据处理成本

5、招聘和保留大数据人才
态度：1、要做到的是重视大数据资源，一直以来数据都是一件受重视的东西，因为透过数据能够反映出很多东西，企业的决策很大程度上都要依托于对于数据的分析。
 2、要搭建属于自己的大数据系统，企业想要真正地用上大数据技术必须依托于大数据系统才行，完整地大数据系统必须包涵大数据的挖掘、存储、处理与分析等等所有的流程。
 3、挖掘真正高质量的数据，这个时代是一个数据爆炸的时代，因此才诞生了大数据的技术

4、能够将大数据融入到各项业务当中，因为大数据终究只是工具，我们需要的是使用这个工具创造价值，而不是为了使用这个工具而使用这个工具

习题

**1.选择题**

(1) Had oop采用( **B** )来整合分布式文件系统上的数据，以保证分析和处理数据的高效。

A.Map Reduce B.HDFS C.Namenode D.Datanode

(2)(  **D**  ) 程序负责HDFS数据存储。

A.Name Node B.Jobracker C. Name node D.secondary Data node

(3) HDFS默认BlockSize的是( C )。

A.32MB B.64MB C.128MB D.256MB

(4)( B )通常是集群的最主要的性能瓶颈。

A.CPU B.网络 C.磁盘 D.内存

(5)( D ) 不包含在Had oop生态系统中。

A.Hive B.Map Reduce C.HDFS D.Spark

**2.问答题**

**(1) 简要介绍Google采集系统的核心组件。**

答：Google采集系统的核心的组件有两个:第一个就是GFS(GoogleFileSystem)，一个分布式文件系统，隐藏下层负载均衡，冗余复制等细节，对上层程序提供一个统一的文件系统API接口;第二个是MapReduce计算模型，Google发现大多数分布式运算可以抽象为MapReduce操作。Map是把输入Input分解成中间的Key/Value对，Reduce把Key/Value合成最终输出Output。这两个函数由程序员提供给系统，下层设施把Map和Reduce操作分布在集群上

运行，并把结果存储在GFS上。

1. **例举Had oop的功能作用， 以及Had oop的优势是什么?**

答：功能: hadoop是通过分布式集群技术,将多个物理机或虚拟机当作一台机器运行,简单说就是做大数据的分布式计算。 优点: (1)高可靠性。Hadoop按位存储和处理数据的能力值得人们信赖; (2)高扩展性。Hadoop是在可用的计算机集簇间分配数据并完成计算任务的,这些集簇可以方便地扩展到数以千计的节点中。 (3)高效性。Hadoop能够在节点之间动态地移动数据,并保证各个节点的动态平衡,因此处理速度非常快。 (4)高容错性。Hadoop能够自动保存数据的多个副本,并且能够自动将失败的任务重新分配。

1. **Had oop的体系结构是怎样的?请简要说明。**

答：Had oop主要由两部分组成：
第一部分：HDFS：分布式⽂件系统

 1)在集群上实现了分布式⽂件系统。

 2)采⽤了主从(Master/Slave)结构模型，⼀个HDFS集群是由⼀个NameNode和若⼲个DataNode组成的。
 第二部分：NameNode(元数据服务器)：属于管理层，⽤于管理数据的存储。

 SecondaryNameNode(辅助元数据服务器)：属于管理层，备份NameNode数据。

DataNodes(块存储)：属于应⽤层，⽤户进⾏数据的存储，被NameNode进⾏管理，要定时地向NameNode进⾏⼯作汇报，执⾏NameNode分配分发的任务。
**⼆者的关系：⼆者相互作⽤，完成了hadoop分布式集群的主要任务：**

 1)HDFS在MapReduce任务处理过程中提供了⽂件操作和存储等⽀持，

 2)MapReduce在HDFS的基础上实现了任务的分发、跟踪、执⾏等⼯作，并收集结果。

**(4) Had oop与大数据、云计算之间的关系是什么，主要起什么作用?**

答：云计算因大数据问题而生,大数据驱动了云计算的发展,而 Hadoop 在大数据和云计算之间建起了一座坚实可靠的桥梁。大数据处理离不开云计算技术，云计算可为大数据提供弹性可扩展的基础设施支撑环境以及数据服务的高效模式，大数据则为云计算提供了新的商业价值，大数据技术与云计算技术必将有更完美的结合。