大数据技术期末复习

第 1 章 大数据概论

一、单选题

1. 第一次信息化浪潮主要解决什么问题？（**B**）

   A、信息传输

   B、信息处理

   C、信息爆炸

   D、信息转换

2. 下面哪个选项属于大数据技术的“数据存储和管理”技术层面的功能？（**A**）

   A、利用分布式文件系统、数据仓库、关系数据库等实现对结构化、半结构化和非结构化海量数据的存储管理

   B、利用分布式并行编程模型和计算框架，结合机器学习和数据挖掘算法，实现对海量数据的处理和分析

   C、构建隐私数据保护体系和数据安全体系，有效保护个人隐私和数据安全

   D、把实时采集的数据作为流计算系统的输入，进行实时处理分析

3. 在大数据的计算模式中，流计算解决的是什么问题？（**D**）

   A、针对大规模数据的批量处理

   B、针对大规模图结构数据的处理

   C、大规模数据的存储管理和查询分析

   D、针对流数据的实时计算

4. 大数据产业指什么？（**A**）

   A、一切与支撑大数据组织管理和价值发现相关的企业经济活动的集合

   B、提供智能交通、智慧医疗、智能物流、智能电网等行业应用的企业

   C、提供数据分享平台、数据分析平台、数据租售平台等服务的企业

   D、提供分布式计算、数据挖掘、统计分析等服务的各类企业

5. 下列哪一个不属于大数据产业的产业链环节？（**A**）

   A、数据循环层

   B、数据源层

   C、数据分析层

   D、数据应用层

6. 下列哪一个不属于第三次信息化浪潮中新兴的技术？（**A**）

   A、互联网

   B、云计算

   C、大数据

   D、物联网

7. 云计算平台层（PaaS）指的是什么？（**A**）

   A、操作系统和围绕特定应用的必需的服务

   B、将基础设施(计算资源和存储)作为服务出租

   C、从一个集中的系统部署软件，使之在一台本地计算机上(或从云中远程地)运行的一个模型

   D、提供硬件、软件、网络等基础设施以及提供咨询、规划和系统集成服务

8. 下面关于云计算数据中心的描述正确的是：（**A**）

   A、数据中心是云计算的重要载体，为各种平台和应用提供运行支撑环境

   B、数据中心就是放在企业内部的一台中心服务器

   C、每个企业都需要建设一个云计算数据中心

   D、数据中心不需要网络带宽的支撑

9. 下列哪个不属于物联网的应用？（**D**）

   A、智能物流

   B、智能安防

   C、环保监测

   D、数据清洗

10. 下列哪项不属于大数据的发展历程？（**D**）

    A、成熟期

    B、萌芽期

    C、大规模应用期

    D、迷茫期

二、多选题

1. 第三次信息化浪潮的标志是哪些技术的兴起？（**BCD**）

   A、个人计算机

   B、物联网

   C、云计算

   D、大数据

2. 信息科技为大数据时代提供哪些技术支撑？（**ABC**）

   A、存储设备容量不断增加

   B、网络带宽不断增加

   C、CPU 处理能力大幅提升

   D、数据量不断增大

3. 大数据具有哪些特点？（**ABCD**）

   A、数据的“大量化”

   B、数据的“快速化”

   C、数据的“多样化”

   D、数据的“价值密度比较低”

4. 下面哪个属于大数据的应用领域？（**ABCD**）

A、智能医疗研发

B、监控身体情况

C、实时掌握交通状况

D、金融交易

5. 大数据的两个核心技术是什么？（**AC**）

A、分布式存储

B、分布式应用

C、分布式处理

D、集中式存储

6. 云计算关键技术包括什么？（**ABCD**）

A、分布式存储

B、虚拟化

C、分布式计算

D、多租户

7. 云计算的服务模式和类型主要包括哪三类？（**ABC**）

A、软件即服务（SaaS）

B、平台即服务（PaaS）

C.、基础设施即服务（IaaS）

D、数据采集即服务（DaaS）

8. 物联网主要由下列哪些部分组成的？（**ABCD**）

A、应用层

B、处理层

C、感知层

D、网络层

9. 物联网的关键技术包括哪些？（**ABC**）

A、识别和感知技术

B、网络与通信技术

C、数据挖掘与融合技术

D、信息处理一体化技术

10. 大数据对社会发展的影响有哪些？（**ABC**）

A、大数据成为一种新的决策方式

B、大数据应用促进信息技术与各行业的深度融合

C、大数据开发推动新技术和新应用的不断涌现

D、大数据对社会发展没有产生积极影响

三、填空题

1. 信息科技需要解决（**信息存储**）、（**信息传输**）和（**信息处理**）3个核心问题。
2. 大数据的基本处理流程主要包括（**采集**）、（**存储**）、（**分析**）和结果呈现等环节。
3. 云计算的关键技术包括（**虚拟化**）、（**分布式存储**）、（**分布式计算**）和多租户等。
4. 大数据的计算模式主要包括（**批处理计算**）、（**流计算**）、（**图计算**）和查询分析计算。
5. 大数据产业包括IT基础设施层、（**数据源层**）、（**数据管理层**）、（**数据分析层**）、数据平台层和数据应用层。

四、简答题

1. 试述信息技术发展史上的3次信息化浪潮及其具体内容。

信息化浪潮	发生时间	标志	解决的问题
第一次浪潮	1980年前后	个人计算机	信息处理
第二次浪潮	1995年前后	互联网	信息传输
第三次浪潮	2010年前后	物联网、云计算和大数据	信息爆炸

2. 试述数据产生方式经历的几个阶段。
    1. **运营式系统阶段**：以门户网站为代表，数据的产生方式是被动的，只有当实际的企业业务发生时，才会产生新的记录并存入数据库。
    2. **用户原创内容阶段**：以wiki、博客、微博、微信等自服务模式为主，上网用户本身就是内容的生成者。
    3. **感知式系统阶段**：物联网中的设备每时每刻都在自动产生大量数据，并且其自动数据产生方式将在短时间内产生更密集、更大量的数据。
3. 科学研究经历了哪4个阶段?
    1. **实验科学**：采用科学解决一些科学问题；
    2. **理论科学**：开始采用各种数学、几何、物理等理论。构建问题模型和解决方案；
    3. **计算科学**：计算科学主要用于对各个科学问题进行计算机模拟和其他形式的计算，借助于计算机的高速运算能力可以解决各种问题；
    4. **数据密集型科学**：先有大量已知数据，从数据中挖掘未知模式和有价值的信息，通过计算和分析得到未知的结论。
4. 试述大数据对思维方式的重要影响。

    大数据时代对思维方式的重要影响是三种思维的转变：全样而非抽样，效率而非精确，相关而非因果。
    1. **全样而非抽样**：大数据技术的核心就是海量数据的存储和处理，因此科学分析可以完全直接针对全集数据而不是抽样数据，并且可以在短时间内迅速得到分析结果。
    2. **效率而非精确**：由于大数据采用的是全样分析，故不存在误差放大的问题，故追求高精确性不是首要目标。另一方面，由于大数据时代要求在几秒内迅速给出海量数据的分析结果，否则会丧失数据的价值，因此数据分析效率成为关键。
    3. **相关而非因果**：由于因果关系不再那么重要，人们更追求相关性。
5. 试述大数据的4个基本特性。

    数据量大(Volume)、数据类型繁多(Variety)、处理速度快(Velocity)和价值密度低(Value)。
    1. 数据量大：信息社会中，数据以自然方式增长，数据每两年增加一倍；
    2. 数据类型繁多：数据类型丰富，包括结构化数据和非结构化数据，如邮件、音频、视频等，给数据处理和分析技术提出了新的挑战；

3. 处理速度快：由于很多应用都需要基于快速生成的数据给出实时分析结果，因此新兴的大数据分析技术通常采用集群处理和独特的内部设计；
4. 价值密度低：有价值的数据分散在海量数据中。

6. 详细阐述大数据、云计算和物联网三者之间的区别与联系。
    1. 区别：大数据侧重于**对海量数据的存储、处理与分析**；云计算本质旨在**整合和优化各种IT资源**，并通过网络**以服务的方式廉价提供给用户**；物联网的发展目标是**实现物物相连。**
    2. 联系：物联网是大数据的**重要来源**，大数据**植根于**云计算，大数据分析的很多技术来源于云计算；大数据为云计算**提供用武之地**，供云计算**发挥价值**；物联网**借助于**云计算和大数据技术，实现物联网大数据的存储、分析和处理。

第 2 章 大数据处理架构

一、单选题

1. 下列哪个不属于Hadoop的特性？（**A**）

    A、成本高

    B、高可靠性

    C、高容错性

    D、运行在Linux平台上

2. Hadoop框架中最核心的设计是什么？（**A**）

    A、为海量数据提供存储的HDFS和对数据进行计算的MapReduce

    B、提供整个HDFS文件系统的NameSpace(命名空间)管理、块管理等所有服务

    C、Hadoop不仅可以运行在企业内部的集群中，也可以运行在云计算环境中

    D、Hadoop被视为事实上的大数据处理标准

3. 在一个基本的Hadoop集群中，DataNode主要负责什么？（**D**）

    A、负责执行由JobTracker指派的任务

    B、协调数据计算任务

    C、负责协调集群中的数据存储

    D、存储被拆分的数据块

4. 在一个基本的Hadoop集群中，SecondaryNameNode主要负责什么？（**A**）

    A、帮助NameNode收集文件系统运行的状态信息

    B、负责执行由JobTracker指派的任务

    C、协调数据计算任务

    D、负责协调集群中的数据存储

5. 在Hadoop项目结构中，HDFS指的是什么？（**A**）

    A、分布式文件系统

    B、分布式并行编程模型

    C、资源管理和调度器

    D、Hadoop上的数据仓库

6. 在Hadoop项目结构中，MapReduce指的是什么？（**A**）

A、分布式并行编程模型

B、流计算框架

C、Hadoop上的工作流管理系统

D、提供分布式协调一致性服务

二、多选题

1. 一个基本的Hadoop集群中的节点主要包括什么？（**ABCD**）

   A、DataNode：存储被拆分的数据块

   B、JobTracker：协调数据计算任务

   C、TaskTracker：负责执行由JobTracker指派的任务

   D、SecondaryNameNode：帮助NameNode收集文件系统运行的状态信息

2. 下列关于Hadoop的描述，哪些是正确的？（**ABCD**）

   A、为用户提供了系统底层细节透明的分布式基础架构

   B、具有很好的跨平台特性

   C、可以部署在廉价的计算机集群中

   D、曾经被公认为行业大数据标准开源软件

3. Hadoop集群的整体性能主要受到什么因素影响？（**ABCD**）

   A、CPU性能

   B、内存

   C、网络

   D、存储容量

4. 下列关于Hadoop的描述，哪些是错误的？（**AB**）

   A、只能支持一种编程语言

   B、具有较差的跨平台特性

   C、可以部署在廉价的计算机集群中

   D、曾经被公认为行业大数据标准开源软件

三、填空题

1. Hadoop的三种运行模式分别是独立（本地）模式、（**伪分布式模式**）和（**完全分布式模式**）。
2. 配置Hadoop时，Java的路径JAVA_HOME在配置文件（**hadoop-env.sh**）中进行设置；所有节点的HDFS路径通过fs.defualt.name来设置，这个选项在配置文件（**core-site.xml**）中设置。
3. Hadoop伪分布模式，通过start-dfs.sh运行启动后所具有的进程包括（**NameNode**）、（**DataNode**）和（**SecondaryNameNode**）。
4. Hadoop的核心是（**HDFS**）和（**MapReduce**）。

四、简答题

1. 试述Hadoop和谷歌的MapReduce、GFS等技术之间的关系。

   1. Hadoop的**核心**是分布式文件系统**HDFS**和**MapReduce**；
   2. HDFS是谷歌文件系统**GFS的开源实现**；
   3. MapReduces是针对**谷歌MapReduce的开源实现。**

2. 试述Hadoop在各个领域的应用情况。

Hadoop已经在各个领域得到了广泛的应用，互联网领域是其应用的主阵地，具体如下

1. **雅虎**将Hadoop主要用于**支持广告系统与网页搜索**；
2. **Facebook**主要将Hadoop平台用于**日志处理、推荐系统和数据仓库**等方面；
3. **百度**选择Hadoop主要用于**日志的存储和统计、网页数据的分析和挖掘、商业分析、在线数据反馈和网页聚类**等。

3. 试述Hadoop具有哪些特性。

1. **高可靠性**：采用冗余数据存储方式，即使一个副本发生故障，其他副本也可以保证正常对外提供服务；
2. **高效性**：是一个并行分布式计算平台，能够高效处理PB级数据；
3. **高可扩展性**：Hadoop的设计目标是可以高效稳定地运行在廉价的计算机集群上，可以扩展到数以千计的计算机节点上；
4. **高容错性**：采用冗余数据存储方式，自动保存数据的多个副本，并且能够自动将失败的任务进行重新分配；
5. **成本低**：Hadoop采用廉价计算机集群，成本较低，普通用户也很容易用自己的PC机搭建Hadoop运行环境；
6. **运行在Linux操作系统上**：是基于Java开发的，可以较好地运行在Linux操作系统上；
7. **支持多种编程语言**：支持Java、C++、Python等编程语言。

4. 试列举单机模式和伪分布式模式的异同点。

1. 相同点：都**只在一台单机上运行。**
2. 不同点：①**运行模式不同**：单机模式是Hadoop的默认模式。这种模式在一台单机上运行，没有分布式文件系统，而是直接读写本地操作系统的文件系统。伪分布模式这种模式也是在一台单机上运行，但用不同的Java进程模仿分布式运行中的各类结点；②**配置不同**：单机模式首次解压Hadoop的源码包时，Hadoop无法了解硬件安装环境，便保守地选择了最小配置。在这种默认模式下所有3个XML文件均为空。当配置文件为空时，Hadoop会完全运行在本地。伪分布模式在“单节点集群”上运行Hadoop，其中所有的守护进程都运行在同一台机器上；③**节点交互不同**：单机模式因为不需要与其他节点交互，单机模式就不使用HDFS，也不加载任何Hadoop的守护进程。该模式主要用于开发调试MapReduce程序的应用逻辑。伪分布模式在单机模式之上增加了代码调试功能，允许你检查内存使用情况，HDFS输入输出，以及其他的守护进程交互。

5. 试述Hadoop生态系统以及每个部分的具体功能。

1. HDFS：用于**存储和管理大规模数据集。**它将数据分散存储在多个节点上，提供高可靠性和可扩展性。
2. MapReduce：用于**处理大规模数据集。**它将数据分成小块，然后在多个节点上并行处理，最后将结果合并。
3. YARN：用于**管理Hadoop集群中的计算资源。**它允许多个应用程序在同一集群上运行，提高资源利用率。
4. Hive：用于**查询和分析大规模数据集。**它提供了类似于SQL的查询语言，使用户可以使用熟悉的语法进行数据分析。
5. Pig：用于**处理大规模数据集。**它提供了一组高级操作，如过滤、聚合和排序，使用户可以轻松地处理和分析数据。
6. HBase：用于**存储大规模结构化数据。**它提供了高可靠性、高可扩展性和高性能的数据存储和访问。
7. zookeeper：用于**管理Hadoop集群中的各种配置和状态信息。**它提供了高可用性和可靠性，使Hadoop集群的管理更加简单和可靠。

第 3 章 分布式文件系统HDFS

一、单选题

1. 分布式文件系统指的是什么？（**A**）

   A、把文件分布存储到多个计算机节点上，成千上万的计算机节点构成计算机集群

   B、用于在Hadoop与传统数据库之间进行数据传递

   C、一个高可用的，高可靠的，分布式的海量日志采集、聚合和传输的系统

   D、一种高吞吐量的分布式发布订阅消息系统，可以处理消费者规模的网站中的所有动作流数据

2. 下列哪一项不属于HDFS采用抽象的块概念带来的好处？（**C**）

   A、简化系统设计

   B、支持大规模文件存储

   C、强大的跨平台兼容性

   D、适合数据备份

3. 下面对SecondaryNameNode第二名称节点的描述，哪个是错误的？（**A**）

   A、SecondaryNameNode一般是并行运行在多台机器上

   B、它是用来保存名称节点中对HDFS元数据信息的备份，并减少名称节点重启的时间

   C、SecondaryNameNode通过HTTPGET方式从NameNode上获取到FsImage和EditLog文件，并下载到本地的相应目录下

   D、SecondaryNameNode是HDFS架构中的一个组成部分

4. 下面哪一项不属于计算机集群中的节点？（**B**）

   A、主节点(Master Node)

   B、源节点（SourceNode）

   C、名称结点(NameNode)

   D、从节点（Slave Node）

5. 在HDFS中，NameNode的主要功能是什么？（**D**）

   A、维护了block id 到datanode本地文件的映射关系

   B、存储文件内容

   C、文件内存保存在磁盘中

   D、存储元数据

6. 下面对FsImage的描述，哪个是错误的？（**D**）

   A、FsImage文件没有记录每个块存储在哪个数据节点

   B、FsImage文件包含文件系统中所有目录和文件inode的序列化形式

   C、FsImage用于维护文件系统树以及文件树中所有的文件和文件夹的元数据

   D、FsImage文件记录了每个块具体被存储在哪个数据节点

7. 在HDFS中，默认一个块多大？（**A**）

   A、64MB

   B、32KB

   C、128KB

   D、16KB

8. HDFS采用了什么模型？（**B**）

A、分层模型

B、主从结构模型

C、管道-过滤器模型

D、点对点模型

9. 下列关于HDFS的描述，哪个不正确？（**D**）

A、HDFS还采用了相应的数据存放、数据读取和数据复制策略，来提升系统整体读写响应性能

B、HDFS采用了主从（Master/Slave）结构模型

C、HDFS采用了冗余数据存储，增强了数据可靠性

D、HDFS采用块的概念，使得系统的设计变得更加复杂

10. 在Hadoop项目结构中，HDFS指的是什么？（**A**）

A、分布式文件系统

B、流数据读写

C、资源管理和调度器

D、Hadoop上的数据仓库

二、多选题

1. HDFS要实现以下哪几个目标？（**ABC**）

A、兼容廉价的硬件设备

B、流数据读写

C、大数据集

D、复杂的文件模型

2. HDFS特殊的设计，在实现优良特性的同时，也使得自身具有一些应用局限性，主要包括以下哪几个方面？（**BCD**）

A、较差的跨平台兼容性

B、无法高效存储大量小文件

C、不支持多用户写入及任意修改文件

D、不适合低延迟数据访问

3. HDFS采用抽象的块概念可以带来以下哪几个明显的好处？（**ACD**）

A、支持大规模文件存储

B、持小规模文件存储

C、适合数据备份

D、简化系统设计

4. 在HDFS中，名称节点（NameNode）主要保存了哪些核心的数据结构？（**AD**）

A、FsImage

B、DN8

C、Block

D、EditLog

5. 数据节点（DataNode）的主要功能包括哪些？（**ABC**）

A、负责数据的存储和读取

B、根据客户端或者是名称节点的调度来进行数据的存储和检索

C、向名称节点定期发送自己所存储的块的列表

D、用来保存名称节点中对HDFS元数据信息的备份，并减少名称节点重启的时间

6. HDFS的命名空间包含什么？（**BCD**）

   A、磁盘

   B、文件

   C、块

   D、目录

7. 下列对于客服端的描述，哪些是正确的？（**ABCD**）

   A、客户端是用户操作HDFS最常用的方式，HDFS在部署时都提供了客户端

   B、HDFS客户端是一个库，暴露了HDFS文件系统接口

   C、严格来说，客户端并不算是HDFS的一部分

   D、客户端可以支持打开、读取、写入等常见的操作

8. HDFS只设置唯一一个名称节点，这样做虽然大大简化了系统设计，但也带来了哪些明显的局限性？（**ABCD**）

   A、命名空间的限制

   B、性能的瓶颈

   C、隔离问题

   D、集群的可用性

9. HDFS数据块多副本存储具备以下哪些优点？（**ABC**）

   A、加快数据传输速度

   B、容易检查数据错误

   C、保证数据可靠性

   D、适合多平台上运行

10. HDFS具有较高的容错性，设计了哪些相应的机制检测数据错误和进行自动恢复？（**BCD**）

    A、数据源太大

    B、数据节点出错

    C、数据出错

    D、名称节点出错

三、填空题

1. 与普通文件系统类似，分布式文件系统数据读写的基本单元是（**块**），只是分布式文件系统中这一基本读写单元比操作系统中的大很多。
2. HDFS只允许一个文件有一个写入者，不允许多个用户对同一文件执行写操作，而且只允许对文件执行（**追加**）操作，不能执行随机写操作。
3. HDFS是一个部署在集群上的分布式文件系统，因此很多数据需要通过网络进行传输。HDFS通信协议是构建上（**TCP/IP**）协议基础之上的。
4. HDFS文件系　统在物理结构上是由计算机集群中的多个节点构成的。这些节点分为两类，一类叫（**NameNode**），另一类叫（**DataNode**）。
5. HDFS采用“（**一次写入，多次读取**）”的简单文件模型。

6. HDFS不支持多用户写入及任意修改文件，只允许对文件执行（**追加**）操作，不能执行（**随机写**）操作。
7. HDFS采用大文件块设计是为了最小化（**寻址开销**）。
8. 在HDFS的设计中，第二名称节点只是起到了名称节点的（**检查点**）作用，并不能起到（**热备份**）的作用。
9. HDFS的数据复制策略采用（**流水线复制**）。
10. HDFS名称节点保存的数据信息中最核心的两大数据结构是（**FsImage**）和（**EditLog**）。

四、简答题

1. 试述HDFS中的名称节点和数据节点的具体功能。
    1. 名称节点负责管理**分布式文件系统的命名空间**，记录每个文件中各个块所在的数据节点的**位置信息**；
    2. 数据节点是分布式文件系统HDFS的**工作节点**，负责数据的**存储和读取**，会根据客户端或者名称节点的**调度**来进行数据的存储和检索，并向名称节点**定期**发送自己所存储的块的列表。
2. 在分布式文件系统中，中心节点的设计至关重要，请阐述HDFS是如何减轻中心节点的负担的。
    1. HDFS的文件块为大小为**64MB**，比普通文件系统中512B大小的数据块大得多，该设计使得名称节点的元数据较少，**减少了元数据占用NameNode的内存容量**；
    2. HDFS集群只有**一个名称节点**，该节点负责所有元数据的管理，这种设计大大简化了分布式文件系统的结构，从而**保证数据不会脱离名称节点的控制**；
    3. HDFS的数据块数据**不会经过名称节点**，大大减轻名称节点的负担，也方便了数据管理。
3. HDFS只设置一个名称节点，在简化系统设计的同时也带来了一些明显的局限性，请阐述局限性具体表现在哪些方面。
    1. 命名空间的限制。名称节点是保存在**内存**中的，因此名称节点能够容纳对象（文件、块）的个数受到**内存空间大小的限制**。
    2. 性能的瓶颈。整个**分布式文件系统的吞吐量**受限于**单个名称节点的吞吐量**。
    3. 隔离问题。由于集群只有一个名称节点，**只有一个命名空间**，因此无法对不同应用程序进行隔离。
    4. 集群的可用性。一旦此唯一的名称节点发生故障，会导致整个集群变得**不可用**。
4. 数据复制主要是在数据写入和恢复的时候发生，HDFS数据复制是使用流水线复制的策略，请阐述该策略的细节。
    1. 当客户端要往HDFS中写入一个文件时，此文件首先被写入**本地**，并被切分为**若干个块**，每个块的大小由HDFS的**设定值**来决定。
    2. 每个块都向HDFS集群中的名称节点发起**写请求**，名称节点会根据系统中各个数据节点的使用情况，选择一个数据节点列表返回给客户端，然后客户端就将数据首先写入列表中的第一数据节点，同时将列表传给第一个数据节点，当第一个数据节点接收到4KB数据时，写入本地，并且向列表中的第二个数据节点发起连接请求，将自己已经接收到的4KB数据和列表传给第二个数据节点，当第二个数据节点接收到4KB数据时，写入本地，并且向列表中的第三个数据节点发起连接请求，依次类推。列表中的多个数据节点形成一条数据复制贩流水线。
    3. 当文件写完时，数据复制也同时完成。
5. 试述HDFS是如何探测错误发生以及如何进行恢复的。
    1. 名称节点出错：将名称节点上的元数据信息同步存储到**其他文件系统**中；并运行一个第二名称节点，当名称节点宕机后，**利用第二名称节点进行系统恢复**。
    2. 数据节点出错：将无法接收到的**“心跳”信号**的数据节点标记为不可读，如果数据块的副本数量**小于**冗余因子，就会启动**数据冗余恢复**，为它生成新副本。
    3. 数据出错：客户端收到数据后会使用md5和sha1对数据块进行校验；如果校验出错，客户端就会请求到**另外一个数据节点读取该文件块**，并向名称节点报告这个文件块有错误，名称节点会**定期检查**并且**重新复制**这个块。
6. 请阐述HDFS在不发生故障的情况下读文件的过程。

1. 客户端打开文件，**创建输入流。**
2. 输入流通过**远程调用名称节点**，获得文件开始部分数据块的保存位置。
3. 客户端得到位置之后开始读取数据，输入流选择**距离客户端最近的数据节点**建立连接并读取数据。
4. 数据从该数据节点读取至客户端结束时，关闭连接。
5. 输入流查找下一个数据块。
6. 找到该数据块的**最佳数据节点**，读取数据。
7. 客户端读取完毕数据时，**关闭输入流。**

7. 请阐述HDFS在不发生故障的情况下写文件的过程。

    1. 客户端**创建文件和输出流。**
    2. HDFS调用**名称节点**，在文件系统的命名空间中建一个新的文件，并执行检查；检查通过后，名称节点会构造一个新文件夹，并添加文件信息。
    3. 客户端通过输出流向**HDFS**的文件写入数据。
    4. 客户端写入的数据首先会被分成一个个的**分包**，将分包放入输入流对象的内部队列，并向名称节点**申请若干个数据节点**，然后通过**流水线复制策略**打包成数据包发送出去。
    5. 为保证所有数据节点的数据都是准确的，需要数据节点向发送者发送“**确认包**”，当客户端收到应答时，将对应的分包**从内部队列移除。**不断执行3~5直到数据写完。
    6. 客户端**关闭输出流**，通知名称节点**关闭文件。**

8. 试述HDFS的冗余数据保存策略。

    HDFS采用**多副本**方式对数据进行冗余存储。

    1. 第一个副本放置在上传文件的数据节点，如果是集群外提交，则随机挑选一台磁盘**不太满、CPU不太忙的节点。**
    2. 第二个副本放置在与第一个副本**不同的机架的节点**上。
    3. 第三个副本与第一个副本**相同机架的其他节点**上 。
    4. 更多副本的放置节点**随机选取。**

第 4 章 分布式数据库HBase

一、单选题

1. 下列关于BigTable的描述，哪个是错误的？（**A**）

    A、爬虫持续不断地抓取新页面，这些页面每隔一段时间地存储到BigTable里

    B、BigTable是一个分布式存储系统

    C、BigTable起初用于解决典型的互联网搜索问题

    D、网络搜索应用查询建立好的索引，从BigTable得到网页

2. 下列选项中，关于HBase和BigTable的底层技术对应关系，哪个是错误的？（**B**）

    A、GFS与HDFS相对应

    B、GFS与Zookeeper相对应

    C、MapReduce与Hadoop MapReduce相对应

    D、Chubby与Zookeeper相对应

3. 在HBase中，关于数据操作的描述，下列哪一项是错误的？（**C**）

    A、HBase采用了更加简单的数据模型，它把数据存储为未经解释的字符串

    B、HBase操作不存在复杂的表与表之间的关系

    C、HBase不支持修改操作

D、HBase在设计上就避免了复杂的表和表之间的关系

4. 在HBase访问接口中，Pig主要用在哪个场合？（**D**）

   A、适合Hadoop MapReduce作业并行批处理HBase表数据

   B、适合HBase管理使用

   C、适合其他异构系统在线访问HBase表数据

   D、适合做数据统计

5. HBase中需要根据某些因素来确定一个单元格，这些因素可以视为一个“四维坐标”，下面哪个不属于“四维坐标”？（**B**）

   A、行键

   B、关键字

   C、列族

   D、时间戳

6. 关于HBase的三层结构中各层次的名称和作用的说法，哪个是错误的？（**A**）

   A、Zookeeper文件记录了用户数据表的Region位置信息

   B、-ROOT-表记录了.META.表的Region位置信息

   C、.META.表保存了HBase中所有用户数据表的Region位置信息

   D、Zookeeper文件记录了-ROOT-表的位置信息

7. 下面关于主服务器Master主要负责表和Region的管理工作的描述，哪个是错误的？（**D**）

   A、在Region分裂或合并后，负责重新调整Region的分布

   B、对发生故障失效的Region服务器上的Region进行迁移

   C、管理用户对表的增加、删除、修改、查询等操作

   D、不支持不同Region服务器之间的负载均衡

8. HBase只有一个针对行健的索引，如果要访问HBase表中的行，下面哪种方式是不可行的？（**B**）

   A、通过单个行健访问

   B、通过时间戳访问

   C、通过一个行健的区间来访问

   D、全表扫描

9. 下面关于Region的说法，哪个是错误的？（**C**）

   A、同一个Region不会被分拆到多个Region服务器

   B、为了加快访问速度，.META.表的全部Region都会被保存在内存中

   C、一个-ROOT-表可以有多个Region

   D、为了加速寻址，客户端会缓存位置信息，同时，需要解决缓存失效问题

二、多选题

1. 关系数据库已经流行很多年，并且Hadoop已经有了HDFS和MapReduce，为什么需要HBase？（**ABCD**）

   A、Hadoop可以很好地解决大规模数据的离线批量处理问题，但是，受限于Hadoop MapReduce编程框架的高延迟数据处理机制，使得Hadoop无法满足大规模数据实时处理应用的需求上

   B、HDFS面向批量访问模式，不是随机访问模式

C、传统的通用关系型数据库无法应对在数据规模剧增时导致的系统扩展性和性能问题

D、传统关系数据库在数据结构变化时一般需要停机维护；空列浪费存储空间

2. HBase与传统的关系数据库的区别主要体现在以下哪几个方面？（**ABCD**）

A、数据类型

B、数据操作

C、存储模式

D、数据维护

3. HBase访问接口类型包括哪些？（**ABCD**）

A、Native Java API

B、HBase Shell

C、Thrift Gateway

D、REST Gateway

4. 下列关于数据模型的描述，哪些是正确的？（**ABCD**）

A、HBase采用表来组织数据，表由行和列组成，列划分为若干个列族

B、每个HBase表都由若干行组成，每个行由行键（row key）来标识

C、列族里的数据通过列限定符（或列）来定位

D、每个单元格都保存着同一份数据的多个版本，这些版本采用时间戳进行索引

5. HBase的实现包括哪三个主要的功能组件？（**ABC**）

A、库函数：链接到每个客户端

B、一个Master主服务器

C、许多个Region服务器

D、廉价的计算机集群

6. HBase的三层结构中，三层指的是哪三层？（**ABC**）

A、Zookeeper文件

B、-ROOT-表

C、.META.表

D、数据类型

7. 以下哪些软件可以对HBase进行性能监视？（**ABCD**）

A、Master-status(自带)

B、Ganglia

C、OpenTSDB

D、Ambari

8. Zookeeper是一个很好的集群管理工具，被大量用于分布式计算，它主要提供什么服务？（**ABC**）

A、配置维护

B、域名服务

C、分布式同步

D、负载均衡服务

9. 下列关于Region服务器工作原理的描述，哪些是正确的？（**ABCD**）

A、每个Region服务器都有一个自己的HLog 文件

B、每次刷写都生成一个新的StoreFile，数量太多，影响查找速度

C、合并操作比较耗费资源，只有数量达到一个阈值才启动合并

D、Store是Region服务器的核心

10. 下列关于HLog工作原理的描述，哪些是正确的？（**ABCD**）

A、分布式环境必须要考虑系统出错。HBase采用HLog保证

B、HBase系统为每个Region服务器配置了一个HLog文件

C、Zookeeper会实时监测每个Region服务器的状态

D、Master首先会处理该故障Region服务器上面遗留的HLog文件

三、简答题

1. 试述在Hadoop体系架构中HBase与其他组成部分的相互关系。
    1. HBase利用Hadoop MapReduce来**处理HBase中的海量数据**，实现高性能计算；
    2. 利用Zookeeper作为**协同服务**，实现稳定服务和失败恢复；
    3. 利用HDFS作为高可靠的**底层存储**，利用廉价集群提供海量数据存储能力；
    4. Sqoop为HBase提供了高效、便捷的**RDBMS数据导入功能**；
    5. Pig和Hive为HBase提供了高层**语言支持。**
2. 请阐述HBase和传统关系数据库的区别。

项目	**传统关系数据库**	**HBase**
数据类型	关系模型	数据模型
数据操作	插入、删除、更新、查询、多表连接	插入、查询、删除、清空，**无法实现表与表之间关联**
存储模式	基于**行**模式存储，元组或行会被连续地存储在磁盘也中	基于**列**存储，每个列族都由几个文件保存，不同列族的文件是分离的
数据索引	针对不同列构建复杂的多个索引	只有**一个行键索引**
数据维护	用最新的当前值去替换记录中原来的旧值	更新操作**不会删除数据旧的版本**，而是生成一个新的版本
可伸缩性	很难实现横向扩展，纵向扩展的空间也比较有限	**轻易**地通过在集群中增加或者减少硬件数量来实现性能的伸缩

3. 请举个实例来阐述HBase的概念视图和物理视图的不同。

    在HBase的概念视图中，一个表可以视为一个**稀疏、多维的映射关系。**在物理视图中，一个表会按照属于**同一列族的数据保存在一起。**

4. HBase中的分区是如何定位的?

    通过构建的映射表的每个条目包含两项内容，一个是**Regionde标识符**，另一个是**Region服务器标识**，这个条目就标识Region和Region服务器之间的对应关系，从而就可以知道某个Region被保存在哪个Region服务器中。

5. 请阐述HBase的三层结构下，客户端是如何访问到数据的。

首先访问**Zookeeper**，获取-ROOT-表的位置信息，然后访问**-Root-**表，获得.META.表的信息，接着访问**.META.**表，找到所需的Region具体位于哪个Region服务器，最后才会到该**Region服务器**中读取数据。

6. 请阐述Region服务器向HDFS文件系统中读写数据的基本原理。

   用户写入数据时，会被分配到相应的Region服务器去执行操作。用户数据首先被写入到**MemStore和HLog**中，当写操作写入HLog之后，**commit()调用**才会将其返回给客户端。 当用户读取数据时，Region服务器会首先访问**MemStore缓存**，如果数据不在缓存中，才会到磁盘上面的**StoreFile**中去寻找。

7. 当一台Region服务器意外终止时，Master如何发现这种意外终止情况？为了恢复这台发生意外的Region服务器上的Region,Master应该做出哪些处理(包括如何使用HLog进行恢复)?

   1. Zookeeper会**实时监测每个Region服务器的状态**，当某个Region服务器发生故障时，Zookeeper会通知Master。
   2. Master首先会处理该故障Region服务器上面**遗留的HLog文件**，这个遗留的HLog文件中包含了来自多个Region对象的日志记录。
   3. 系统会根据每条日志记录所属的Region对象**对HLog数据进行拆分**，分别放到相应Region对象的**目录**下，然后，再将失效的Region**重新分配**到可用的Region服务器中，并把与该Region对象相关的**HLog日志记录**也发送给相应的Region服务器。
   4. Region服务器领取到分配给自己的Region对象以及与之相关的HLog日志记录以后，会重新做一遍**日志记录中的各种操作**，把日志记录中的数据写入到**MemStore缓存**中，然后，刷新到磁盘的**StoreFile文件**中，完成数据恢复。

8. 试述HLog的工作原理。

   HBase系统为每个Region服务器配置了一个**HLog文件**，用户更新数据必须**首先被计入日志**后才能写入MemStore缓存，并且直到MemStore缓存内容对应的日志已经被写入磁盘之后，该缓存内容才会被刷新写入磁盘。

9. 试述HStore的工作原理。

   每个Store对应了表中的一个**列族的存储。**每个Store包括**一个MenStore缓存**和**若干个StoreFile文件。**MenStore是排序的内存缓冲区，当用户写入数据时，系统首先把数据放入MenStore缓存，当MemStore缓存满时，就会刷新到磁盘中的一个StoreFile文件中。随着StoreFile文件数量的不断增加，当达到事先设定的**阈值**是触发文件**合并操作**，当单个StoreFile文件大小超过一定阈值时，就会触发文件分裂操作。

10. 在HBase中，每个Region服务器维护一个HLog，而不是为每个Region都单独维护一个HLog。请说明这种做法的优缺点。

    1. 优点： 多个Region对象的更新操作所发生的日志修改，只需要不断把日志记录追加到单个日志文件中，不需要同时打开、写入到多个日志文件中，可以**减少磁盘寻址次数**，**提高对表的写操作性能。**
    2. 缺点：如果一个Region服务器发生故障，为了恢复其上的Region对象，需要将Region服务器上的HLog按照其所属的Region对象**进行拆分**，然后**分发到其他Region服务器**上执行恢复操作。

第 7 章 MapReduce

一、单选题

1. 下列传统并行计算框架，说法错误的是哪一项？（**B**）

   A、刀片服务器、高速网、SAN，价格贵，扩展性差上

   B、共享式(共享内存/共享存储)，容错性好

   C、编程难度高

   D、实时、细粒度计算、计算密集型

2. 下列关于MapReduce模型的描述，错误的是哪一项？（**D**）

   A、MapReduce采用“ 分而治之”策略

   B、MapReduce设计的一个理念就是“ 计算向数据靠拢”

   C、MapReduce框架采用了Master/Slave架构

   D、MapReduce应用程序只能用Java来写

3. MapReduce1.0的体系结构中，JobTracker是主要任务是什么？（**A**）

   A、负责资源监控和作业调度，监控所有TaskTracker与Job的健康状况

   B、使用“slot”等量划分本节点上的资源量（CPU、内存等）

   C、会周期性地通过“心跳”将本节点上资源的使用情况和任务的运行进度汇报给TaskTracker

   D、会跟踪任务的执行进度、资源使用量等信息，并将这些信息告诉任务（Task）

4. 下列关于MapReduce工作流程，哪个描述是正确的？（**A**）

   A、所有的数据交换都是通过MapReduce框架自身去实现的

   B、不同的Map任务之间会进行通信

   C、不同的Reduce任务之间可以发生信息交换

   D、用户可以显式地从一台机器向另一台机器发送消息

5. 下列关于MapReduce的说法，哪个描述是错误的？（**D**）

   A、MapReduce具有广泛的应用，比如关系代数运算、分组与聚合运算等

   B、MapReduce将复杂的、运行于大规模集群上的并行计算过程高度地抽象到了两个函数

   C、编程人员在不会分布式并行编程的情况下，也可以很容易将自己的程序运行在分布式系统上，完成海量数据集的计算

   D、不同的Map任务之间可以进行通信

6. 下列关于Map和Reduce函数的描述，哪个是错误的？（**C**）

   A、Map将小数据集进一步解析成一批<key,value>对，输入Map函数中进行处理

   B、Map每一个输入的<k 1 ,v 1 >会输出一批<k 2 ,v 2 >。<k 2 ,v 2 >是计算的中间结果

   C、Reduce输入的中间结果<k 2 ,List(v 2)>中的List(v 2)表示是一批属于不同k 2 的value

   D、Reduce输入的中间结果<k 2 ,List(v 2)>中的List(v 2)表示是一批属于同一个k 2 的value

7. 下面哪一项不是MapReduce体系结构主要部分？（**A**）

   A、Client

   B、JobTracker

   C、TaskTracker以及Task

   D、Job

8. 关于MapReduce1.0的体系结构的描述，下列说法错误的是？（**A**）

A、Task 分为Map Task 和Reduce Task 两种，分别由JobTracker 和TaskTracker 启动

B、slot 分为Map slot 和Reduce slot 两种，分别供MapTask 和Reduce Task 使用

C、TaskTracker 使用“slot”等量划分本节点上的资源量（CPU、内存等）

D、TaskTracker 会周期性接收JobTracker 发送过来的命令并执行相应的操作（如启动新任务、杀死任务等）

9. 下列说法错误的是？（**C**）

   A、Hadoop MapReduce是MapReduce的开源实现，后者比前者使用门槛低很多

   B、MapReduce采用非共享式架构，容错性好

   C、MapReduce主要用于批处理、实时、计算密集型应用

   D、MapReduce采用“ 分而治之”策略

二、简答题

1. MapReduce 是处理大数据的有力工具，但不是每个任务都可以使用MapReduce 来进行处理。试述适合用MapReduce来处理的任务或者数据集需满足怎样的要求。

   适合用MapReduce来处理的数据集，需要满足一个前提条件: 待处理的数据集可以分解成**许多小的数据集**，而且每一个小数据集都可以**完全并行地进行处理。**

2. MapReduce模型采用Master(JobTracker)-Slave(TaskTracker)结构，试描述JobTracker和TasKTracker的功能。

   MapReduce 框架采用了Master/Slave 架构，包括一个Master 和若干个Slave。**Master 上运行JobTracker,Slave上运行TaskTracker。**用户提交的每个计算作业，会被划分成若干个任务。JobTracker 负责**作业和任务的调度**，**监控它们的执行**，并**重新调度已经失败的任务。**TaskTracker 负责**执行由JobTracker指派的任务。**

3. MapReduce计算模型的核心是Map函数和Reduce函数，试述这两个函数各自的输入、输出以及处理过程。

函数	输入	输出	说明
Map	<k1,v1>	List(<k2,v2>)	（1）将小数据集进一步解析成一批<key,value>对，输入Map函数中进行处理；
Reduce	<k2,List(v2)>	<k3,v3>	（2）每一个输入的<k1,v1>会输出一批<k2,v2>,<k2,v2>是计算的中间结果<k2,List(v2)>中的List(v2)表示是一批属于同一个k2的value

4. 试画出使用MapReduce来对英语句子“Whatever is worth doing is worth doing well”进行单词统计的过程。

```
Map输出：<”whatever”,1>
         <”is ”,1>
         <”worth”,1>
         <”doing”,1>
         <”is”,1>
         <”worth”,1>
         <”doing”,1>
         <”well”,1>
Shuffle:
```

```
          <”doing”,<1,1>>
          <”is”,<1,1>>
          <”well”,1>
          <”whatever”,1>
          <”worth”,<1,1>>
Reduce:
          <”doing”,2>
          <”is”,2>
          <”well”,1>
          <”whatever”,1>
          <”worth”,2>
```

5. MapReduce中有这样一个原则:移动计算比移动数据更经济。试述什么是本地计算，并分析为何要采用本地计算。
    1. MapReduce设计的一个理念就是“**计算向数据靠拢**”，而不是“数据向计算靠拢”，因为移动数据需要**大量的网络传输开销**，尤其是在大规模数据环境下，这种开销尤为惊人，所以，移动计算要比移动数据更加经济。
    2. 本地计算：在一个集群中，只要有可能，MapReduce框架就会将Map程序**就近**地在HDFS数据所在的节点运行，即将**计算节点和存储节点放在一起运行**，从而减少了节点间的数据移动开销。

第 10 章 Spark

一、单选题

1. 下列关于Spark的描述，错误的是哪一项？（**D**）

   A、Spark最初由美国加州伯克利大学（UCBerkeley）的AMP实验室于2009年开发

   B、Spark在2014年打破了Hadoop保持的基准排序纪录.

   C、Spark用十分之一的计算资源，获得了比Hadoop快3倍的速度

   D、Spark运行模式单一

2. 下列关于Spark的描述，错误的是哪一项？（**C**）

   A、使用DAG执行引擎以支持循环数据流与内存计算析

   B、可运行于独立的集群模式中，可运行于Hadoop中，也可运行于Amazon EC2等云环境中

   C、支持使用Scala、Java、Python和R语言进行编程，但是不可以通过Spark Shell进行交互式编程

   D、Spark运行模式不是单一的

3. 下列关于Scala特性的描述，错误的是哪一项？（**A**）

   A、Scala语法复杂，但是能提供优雅的API计算

   B、Scala具备强大的并发性，支持函数式编程，可以更好地支持分布式系统

   C、Scala兼容Java，运行速度快，且能融合到Hadoop生态圈中

   D、Scala是Spark的主要编程语言

4. 下列说法哪项有误？（**C**）

   A、相对于Spark来说，使用Hadoop进行迭代计算非常耗资源

   B、Spark将数据载入内存后，之后的迭代计算都可以直接使用内存中的中间结果作运算，避免了从磁盘中频繁读取数据

C、Hadoop的设计遵循“一个软件栈满足不同应用场景”的理念

D、Spark可以部署在资源管理器YARN之上，提供一站式的大数据解决方案

5. 在Spark生态系统组件的应用场景中，下列哪项说法是错误的？（**C**）

   A、Spark应用在复杂的批量数据处理

   B、Spark SQL是基于历史数据的交互式查询

   C、Spark Streaming是基于历史数据的数据挖掘

   D、GraphX是图结构数据的处理

6. 下列说法错误的是？（**A**）

   A、RDD（Resillient Distributed Dataset）是运行在工作节点（WorkerNode）的一个进程，负责运行Task

   B、Application是用户编写的Spark应用程序

   C、一个Job包含多个RDD及作用于相应RDD上的各种操作

   D、Directed Acyclic Graph反映RDD之间的依赖关系

7. 下列关于RDD说法，描述有误的是？（**C**）

   A、一个RDD就是一个分布式对象集合，本质上是一个只读的分区记录集合

   B、每个RDD可分成多个分区，每个分区就是一个数据集片段

   C、RDD是可以直接修改的

   D、RDD提供了一种高度受限的共享内存模型

8. Spark生态系统组件Spark Streaming的应用场景是？（**D**）

   A、基于历史数据的数据挖掘

   B、图结构数据的处理

   C、基于历史数据的交互式查询

   D、基于实时数据流的数据处理

二、简答题

1. Spark是基于内存计算的大数据计算平台，试述Spark的主要特点。

   ①运行速度快；②容易使用；③通用性；④运行模式多样。

2. Spark的出现是为了解决Hadoop MapReduce的不足，试列举Hadoop MapReduce的几个缺陷，并说明Spark具备哪些优点。

   1. Hadoop的缺点：①表达能力有限；②磁盘I/O开销大；③延迟高。
   2. Spark的优点：①除了Map和Reduce操作，提供了**多种数据集操作类型**，编程模型比Reduce更加灵活。②提供了**内存计算**，中间结果直接存放在内存中，带来了跟高的迭代运算效率。③**基于DAG的任务调度执行机制**，优于MapReduce的迭代执行机制。

3. Spark已打造出结构一体化，功能多样化的大数据生态系统，试述Spark的生态系统。

   Spark所提供的生态系统同时支持批处理、交互式查询和流数据处理。Spark生态系统主要包括SparkCore、Spark SQL、Spark Streaming、MLlib、GraphX等组件。

   1. Spark Core。包含**Spark的基本功能**，如内存计算、任务调度、部署模式、故障恢复、存储管理等，主要面向批数据处理。
   2. Spark SQL。允许开发人员**直接处理RDD**，同时也可查询Hive、HBase等外部数据源。能够统一处理关系表和RDD。

3. Spark Streaming。支持高吞吐量、可容错处理的实时流数据处理，其核心是将流数据**分解成一系列短小的批处理作业**，每个短小的批处理作业都可以使用Spark Core进行快速处理。
4. MLlib。提供常用的**机器学习**算法的实现，包括聚类、分类、回归、协同过滤等。
5. GraphX。Spark用于**图计算**的API，可以认为是Pregel在Spark上的重写及优化。

4. 美国加州大学伯克利分校提出的数据分析的软件栈BDAS认为目前的大数据处理可以分为哪三个类型?
    1. **复杂的批量数据处理**：时间跨度通常在数十分钟到数小时之间；
    2. **基于历史数据的交互式查询**：时间跨度通常在数十秒到数分钟之间；
    3. **基于实时数据流的数据处理**：时间跨度通常在数百毫秒到数秒之间。
5. 从Hadoop+Storm架构转向Spark架构可带来哪些好处?
    1. 实现**一键式**安装和配置、**线程级别**的任务监控和告警；
    2. 降低硬件集群、软件维护、任务监控和应用开发的难度；
    3. 便于做成统一的硬件、计算平台资源池。
6. Spark对RDD的操作主要分为行动（Action）和转换（Transformation）两种类型，两种类型操作的区别是什么?
    1. 行动用于执行计算并指定输出的形式，接受RDD但是返回**非RDD**；
    2. 转换用于指定RDD之间的相互依赖关系，接受RDD并返回**RDD**。
7. 试述如下Spark的几个主要概念：RDD、DAG、阶段、分区、窄依赖、宽依赖。
    1. RDD：弹性分布式数据集（Resilient Distributed Dataset），是分布式内存的一个抽象概念，提供了一种**高度受限**的共享内存模型。
    2. DAG：**有向无环图**（Directed Acyclic Graph），反映RDD之间的依赖关系。
    3. 阶段：是作业的**基本调度单位**，一个作业会分为多个阶段，一个阶段会分为多个任务。
    4. 分区：一个RDD就是一个分布式对象集合，本质上是一个**只读**的分区记录集合，每个RDD可以分成多个分区，每个分区就是一个数据集片段。
    5. 窄依赖：一个父RDD对应一个子RDD的分区，或者是**多对一**.
    6. 宽依赖：一个父RDD对应多个子RDD的分区。