一、单选题。(本题共 20 小题，每小题 1 分，满分 20 分)

（1） 下列哪一个不属于第三次信息化浪潮中新兴的技术？（C）

A、大数据B、云计算C、互联网D、物联网

（2） 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

（3） 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

（4） 下面哪一项不是 Hadoop 的特性？(B)

A、可扩展性高B、只支持少数几种编程语言

C、成本低D、能在 linux 上运行

（5） 在 HDFS 中，默认一个块多大？(A)

A、64MB B、32KB C、128KB D、16KB

（6） 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

（7） 下列选项中关于 HBase 和 BigTable 的底层技术对应关系，哪个是错的？(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

（8） 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

（9） 下列哪个不属于 NoSQL 数据库的特点？(D)

A、灵活的可扩展性B、灵活的数据模型

C、与云计算紧密融合D、数据存储规模有限

（10） 下列哪一项不属于 NoSQL 的四大类型？(D)

A、文档数据库B、图数据库C、列族数据库 D、时间戳数据库

（11）下列关于 UMP 系统功能的说法，哪个是错误的？(D)

A、充分利用主从库实现用户读写操作的分离，实现负载均衡

B、UMP 系统实现了对于用户透明的读写分离功能

C、UMP 采用的两种（资源隔离方式（用 Cgroup 限制 MySQL 进程资源和在 Proxy

服务器端限制 QPS）

D、UMP 系统只设计了一种机制来保证数据安全

（12）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

（13）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

（14）以下哪个不是数据仓库的特性：（C）

A、面向主题的B、集成的

C、动态变化的D、反映历史变化的

（15）下面关于 Hive 的描述错误的是：（C）

A、HBase 与 Hive 的功能是互补的，它实现了 Hive 不能提供的功能

B、当采用 MapReduce 作为执行引擎时，用 HiveQL 语句编写的处理逻辑，最终都

要转化为 MapReduce 任务来运行

C、Hive 一般用于处理静态数据，主要是 BI 报表数据

D、Hive 主要是用于满足实时数据流的处理需求

（16）下列有关 Hive 和 Impala 的对比错误的是：(D)

A、Hive 与 Impala 使用相同的元数据

B、Hive 与 Impala 中对 SQL 的解释处理相似，都是通过词法分析生成执行计划

C、Hive 适合于长时间的批处理查询分析，而 Impala 适合于实时交互式 SQL 查询

D、Hive 在内存不足以存储所有数据时，会使用外存，而 Impala 也是如此

（17）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

（18）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<k1,v1 >会输出一批<k2,v2 >，<k2,v2 >是计算的中间结果。

C、Reduce 输入的中间结果<k2,List(v2)>中的 List(v2 )表示是一批属于不同 k2 的

value。

D、Reduce 输入的中间结果<k2,List(v2)>中的 List(v2)表示是一批属于同一个 k2 的

value。

（19）下列哪项不属于流计算的处理流程的三个阶段？(B）

A、数据实时采集B、数据批量采集

C、数据实时计算D、实时查询服务

（20）以下哪个不属于事件驱动型应用？(C)

A.、反欺诈B、异常检测

C、基于规则的报警D、消费者技术中的实时数据即席分析

二、多选题。(本题共 8 小题，每小题 3 分，满分 24 分)

（1）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

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

A、FsImageB、DN8

C、BlockD、EditLog

（3）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

（4）下面关于 NoSQL 与关系数据库的比较，哪些是正确的？（ABCD）

A、关系数据库以完善的关系代数理论作为基础，有严格的标准

B、关系数据库可扩展性较差，无法较好支持海量数据存储

C、NoSQL 可以支持超大规模数据存储

D、NoSQL 数据库缺乏数学理论基础，复杂查询性能不高

（5）下列关于 MapReduce 的体系结构的描述，说法正确的有？（ABD）

A、用户编写的 MapReduce 程序通过 Client 提交到 JobTracker 端

B、JobTracker 负责资源监控和作业调度

C、TaskTracker 监控所有 TaskTracker 与 Job 的健康状况

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

（6）Hive 主要由哪三个模块组成：（ACD)

A、用户接口模块

B、用户查询模块

C、驱动模块

D、元数据存储模块

（7）Flink 常见的应用场景包括：（ABC)

A、事件驱动型应用

B、数据分析应用

C、数据流水线应用

D、正反馈应用

（8）下列关于云数据库的描述，哪些是正确的？ （ABCD）

A、Amazon 是云数据库市场的先行者

B、Google Cloud SQL 是谷歌公司推出的基于 MySQL 的云数据库

C、从数据模型的角度来说，云数据库并非一种全新的数据库技术

D、云数据库并没有专属于自己的数据模型

三、简答题。（(本题共 3 小题，每小题 10 分，满分 30 分）

1、下图是 Hadoop 生态系统图，请分别阐述 Hadoop 生态系统的各个组成部分

（Zookeeper、HDFS、HBase、MapReduce、Pig、Hive、Sqoop）的主要功能。



Zookeeper：是一个分布式协调服务，主要负责管理分布式系统中的配置信息、命名服务、组服务等。

HDFS：是Hadoop的分布式文件系统，主要负责存储大数据，并提供高可靠性、高可用性和高容错性。

HBase：是一个分布式、面向列的开源数据库，主要负责存储海量结构化数据，并支持高效的读写操作。

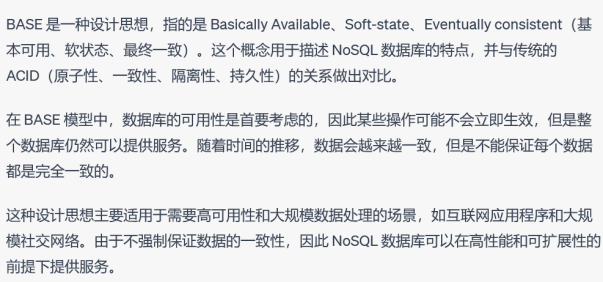
MapReduce：是Hadoop的核心计算框架，主要负责实现分布式数据的计算。

Pig：是一个高级分析工具，主要负责简化MapReduce的开发和运行，使得用户可以使用高级语言来描述分析任务。

Hive：是一个数据仓库工具，主要负责简化对大数据的分析和查询，使得用户可以使用类SQL语言来描述分析任务。

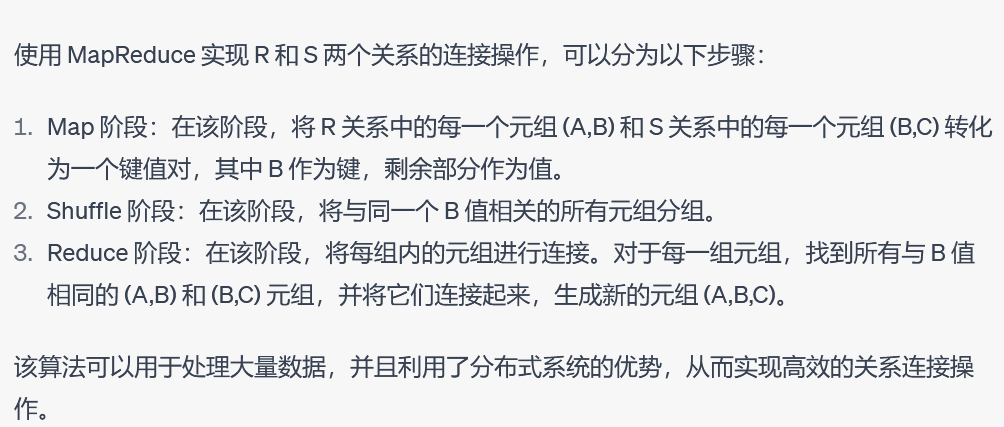
Sqoop：是一个数据迁移工具，主要负责实现从关系型数据库到Hadoop集群的数据迁移。

2、请描述作为 NoSQL 数据库的基石之一的 BASE 的含义。



3、假设关系 R(A，B)和 S(B,C)都存储在一个文件中。请阐述如何用 MapReduce 实现

R 和 S 这两个关系的连接(join)操作。



四、操作题。(本题共 2 小题，第一题 10 分，第二题 16 分，满分 26

四、程序题

1、HDFS 操作题。假设 Hadoop 的安装目录为“/usr/local/hadoop”：

（1**）为 hadoop 用户在 HDFS 中创建用户目录“/user/hadoop”。**

su hadoop

/usr/local/hadoop/bin/hdfs dfs -mkdir /user/hadoop

其中，su hadoop 用于切换到 hadoop 用户，

/usr/local/hadoop/bin/hdfs dfs -mkdir /user/hadoop 用于在 HDFS 中创建 /user/hadoop 目录。

**（2） HDFS 的目录在“/user/hadoop”下， 创建 input 文件夹，并将 Linux 系统本地的“～**

**/input/test.txt”文件上传到上述 HDFS 用户目录的 input 文件夹中。**

首先，创建 HDFS 的目录：

hadoop fs -mkdir /user/hadoop/input

然后，将本地的 test.txt 文件上传到 HDFS 目录：

hadoop fs -put ~/input/test.txt /user/hadoop/input/

1. **将 HDFS 中 input 文件夹中 test.txt 文件的内容输出到终端中。**

hadoop fs -cat /user/hadoop/input/test.txt

1. **删除 HDFS 中 input 文件夹中 test.txt 文件。**

hadoop fs -rm /user/hadoop/input/test.txt

2、Hbase 的 shell 操作。

（1）列出 HBase 所有的表的相关信息。

hbase> list

1. 创建了一个 student 表，属性有：name,sex,age,dept,course。

hbase>creat 'student','name','sex','age','dept','course'

1. 为 student 表添加了学号为 22001，名字为 SuHai 的一行数据，其行键为 22001。

hbase> put 'student', '22001', 'name', 'SuHai'

hbase> put 'student', '22001', 'sex', '男'

hbase> put 'student', '22001', 'age', '22'

hbase> put 'student', '22001', 'dept', '计算机科学与技术'

hbase> put 'student', '22001', 'course', '数据库'

1. 即为 22001 行下的 course 列族的 math 列添加了一个数据 80。

hbase> put 'student', '22001', 'course:math', '80'

1. 在终端打印出表 student 的所有记录数据。

hbase> scan 'student'

1. 删除 student 表中 22001 行下的 sex 列的所有数据。

hbase> delete 'student', '22001', 'sex'

1. 删除了 student 表中的 95001 行的全部数据。

hbase> deleteall 'student', '95001'

1. 删除表 student。

hbase> disable 'student'

hbase> drop 'student'