### 第一章 大数据概述

#### 单选题

1. 第一次信息化浪潮主要解决什么问题？A P3表格
2. 信息处理 一
3. 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 P16表格
4. 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和管理
5. 在大数据的计算模式中，流计算解决的是什么问题？A P17
6. 针对流数据的实时计算
7. 下列哪一个不属于第三次信息化浪潮中新兴的技术？A P3表格
8. 互联网
9. 云计算平台层（PaaS）指的是什么？A **大数据产业层次、云计算产业层次、物联网层次**
10. 操作系统和围绕特定应用的必需的服务 **PaaS**
11. 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 **SaaS**
12. 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 **大数据产业的IT基础设施层**

#### 多选题

1. 下面哪个属于大数据的应用领域？ABCD
2. 智能医疗研发
3. 监控身体情况
4. 实时掌握交通状况
5. 金融交易
6. 大数据的两个核心技术是什么？AC
7. 分布式存储
8. 分布式应用
9. 分布式处理
10. 集中式存储
11. 云计算关键技术包括什么？ABCD
12. 分布式存储
13. 虚拟化
14. 分布式计算
15. 多租户

### 第二章 大数据处理架构Hadoop

#### 单选题

1. Hadoop框架中最核心的设计是什么？A
2. **为海量数据提供存储的HDFS和对数据进行计算的MapReduce**
3. **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**
4. **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**
5. **Hadoop被视为事实上的大数据处理标准 P30**
6. 在一个基本的Hadoop集群中，DataNode主要负责什么？D
7. 负责执行由JobTracker指派的任务
8. 协调数据计算任务
9. 负责协调集群中的数据存储
10. **存储被拆分的数据块**
11. 下列哪一个不属于Hadoop的大数据层的功能？C
12. 数据挖掘
13. 离线分析
14. **实时计算**
15. 在一个基本的Hadoop集群中，SecondaryNameNode主要负责什么？A
16. 帮助NameNode收集文件系统运行的状态信息
17. 负责协调集群中的数据存储
18. 在Hadoop项目结构中，HDFS指的是什么？A
19. 分布式文件系统
20. **资源管理和调度器 YARN**
21. **Hadoop上的数据仓库 HIVE**
22. 在Hadoop项目结构中，MapReduce指的是什么？A
23. **分布式并行编程模型**
24. **流计算框架 STORM**
25. **Hadoop上的工作流管理系统 OOZIE**
26. **提供分布式协调一致性服务 zookeeper**

#### 多选题

1. 下面哪个是Hadoop2.0的组件？AD
2. ResourceManager P162
3. JobTracker
4. TaskTracker
5. NodeManager
6. 一个基本的Hadoop集群中的节点主要包括什么？ABCD
7. DataNode：存储被拆分的数据块
8. **JobTracker：协调数据计算任务 第七章**
9. **TaskTracker：负责执行由JobTracker指派的任务**
10. SecondaryNameNode：帮助NameNode收集文件系统运行的状态信息
11. 下列关于Hadoop的描述，哪些是正确的？ABCD
12. 为用户提供了系统底层细节透明的分布式基础架构
13. 具有很好的跨平台特性
14. 可以部署在廉价的计算机集群中
15. **曾经被公认为行业大数据标准开源软件**
16. Hadoop集群的整体性能主要受到什么因素影响？ABCD
17. **CPU性能**
18. **内存**
19. **网络**
20. **存储容量**

### 第三章 分布式文件系统HDFS

#### 单选题

1. 分布式文件系统指的是什么？A
2. 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**P46**
3. 用于在Hadoop与传统数据库之间进行数据传递  **sqoop**
4. 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 **flume**
5. 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  **kafka**
6. 下列哪一项不属于HDFS采用抽象的块概念带来的好处？A P50
7. 强大的跨平台兼容性
8. 在HDFS中，NameNode的主要功能是什么？B
9. **维护了block id 到datanode本地文件的映射关系**
10. **存储元数据**
11. 下面对FsImage的描述，哪个是错误的？B
12. FsImage文件包含文件系统中所有目录和文件inode的序列化形式
13. **FsImage文件记录了每个块具体被存储在哪个数据节点**
14. 下面对SecondaryNameNode第二名称节点的描述，哪个是错误的？A
15. **SecondaryNameNode一般是并行运行在多台机器上 P51**
16. **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**
17. 在Hadoop项目结构中，HDFS指的是什么？A
18. 分布式文件系统
19. **流数据读写 在P57有**
20. **资源管理和调度器 YARN**
21. **Hadoop上的数据仓库 HIVE**
22. 下列关于HDFS的描述，哪个不正确？D
23. **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 P55**
24. HDFS采用了主从（Master/Slave）结构模型
25. HDFS采用了冗余数据存储，增强了数据可靠性
26. **HDFS采用块的概念，使得系统的设计变得更加复杂 应该是更加简单**

#### 多选题

1. HDFS要实现以下哪几个目标？ABC
2. 兼容廉价的硬件设备
3. 流数据读写
4. 大数据集
5. **复杂的文件模型**
6. 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 **P49**
7. 较差的跨平台兼容性
8. 无法高效存储大量小文件
9. 不支持多用户写入及任意修改文件
10. 不适合低延迟数据访问
11. 数据节点（DataNode）的主要功能包括哪些？ P51
12. HDFS的命名空间包含什么？BCD P53
13. 磁盘
14. **文件**
15. **块**
16. **目录**
17. 下列对于客服端的描述，哪些是正确的？ABCD P54
18. 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
19. **HDFS客户端是一个库，暴露了HDFS文件系统接口**
20. **严格来说，客户端并不算是HDFS的一部分**
21. **客户端可以支持打开、读取、写入等常见的操作**
22. 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 P54
23. HDFS数据块多副本存储具备以下哪些优点？ **P55**
24. HDFS具有较高的容错性，设计哪些相应的机制检测数据错误和进行自动恢复？BCD **P56**
25. 数据源太大
26. **数据节点出错**
27. **数据出错**
28. **名称节点出错**

### 第四章 分布式数据库HBase

#### 单选题

1. 下列关于BigTable的描述，哪个是错误的？A
2. 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
3. BigTable是一个分布式存储系统
4. BigTable起初用于解决典型的互联网搜索问题
5. 网络搜索应用查询建立好的索引，从BigTable得到网页
6. 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
7. GFS与HDFS相对应
8. GFS与Zookeeper相对应
9. MapReduce与Hadoop MapReduce相对应
10. Chubby与Zookeeper相对应
11. 在HBase中，关于数据操作的描述，下列哪一项是错误的？C
12. 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
13. HBase操作不存在复杂的表与表之间的关系
14. HBase不支持修改操作
15. HBase在设计上就避免了复杂的表和表之间的关系
16. 在HBase访问接口中，Pig主要用在哪个场合？D
17. 适合Hadoop MapReduce作业并行批处理HBase表数据
18. 适合HBase管理使用
19. 适合其他异构系统在线访问HBase表数据
20. 适合做数据统计
21. 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
22. 行键
23. 关键字
24. 列族
25. 时间戳
26. 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
27. Zookeeper文件记录了用户数据表的Region位置信息
28. -ROOT-表记录了.META.表的Region位置信息
29. .META.表保存了HBase中所有用户数据表的Region位置信息
30. Zookeeper文件记录了-ROOT-表的位置信息
31. 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
32. 在Region分裂或合并后，负责重新调整Region的分布
33. 对发生故障失效的Region服务器上的Region进行迁移
34. 管理用户对表的增加、删除、修改、查询等操作
35. 不支持不同Region服务器之间的负载均衡
36. 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
37. 通过单个行健访问
38. 通过时间戳访问
39. 通过一个行健的区间来访问
40. 全表扫描
41. 下面关于Region的说法，哪个是错误的？C
42. 同一个Region不会被分拆到多个Region服务器
43. 为了加快访问速度，.META.表的全部Region都会被保存在内存中
44. 一个-ROOT-表可以有多个Region
45. 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

#### 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
2. 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
3. HDFS面向批量访问模式，不是随机访问模式
4. 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
5. 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
6. 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
7. 数据类型
8. 数据操作
9. 存储模式
10. 数据维护
11. HBase访问接口类型包括哪些？ABCD
12. Native Java API
13. HBase Shell
14. Thrift Gateway
15. REST Gateway
16. 下列关于数据模型的描述，哪些是正确的？ABCD
17. 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
18. 每个HBase表都由若干行组成，每个行由行键（row key）来标识
19. 列族里的数据通过列限定符（或列）来定位
20. 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
21. HBase的实现包括哪三个主要的功能组件？ABC
22. 库函数：链接到每个客户端
23. 一个Master主服务器
24. 许多个Region服务器
25. 廉价的计算机集群
26. HBase的三层结构中，三层指的是哪三层？ABC
27. Zookeeper文件
28. -ROOT-表
29. .META.表
30. 数据类型
31. 以下哪些软件可以对HBase进行性能监视？ABCD
32. Master-status(自带)
33. Ganglia
34. OpenTSDB
35. Ambari
36. 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
37. 配置维护
38. 域名服务
39. 分布式同步
40. 负载均衡服务
41. 下列关于Region服务器工作原理的描述，哪些是正确的？ABCD
42. 每个Region服务器都有一个自己的HLog 文件
43. 每次刷写都生成一个新的StoreFile，数量太多，影响查找速度
44. 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
45. Store是Region服务器的核心
46. 下列关于HLog工作原理的描述，哪些是正确的？ABCD
47. 分布式环境必须要考虑系统出错。HBase采用HLog保证
48. HBase系统为每个Region服务器配置了一个HLog文件
49. Zookeeper会实时监测每个Region服务器的状态
50. Master首先会处理该故障Region服务器上面遗留的HLog文件

### 第五章 NoSQL数据库

#### 单选题

1. 下列哪个不属于NoSQL数据库的特点？D
2. 灵活的可扩展性
3. 灵活的数据模型
4. 与云计算紧密融合
5. 数据存储规模有限
6. 下面关于NoSQL和关系数据库的简单比较，哪个是错误的？B
7. RDBMS有关系代数理论作为基础，NoSQL没有统一的理论基础
8. NoSQL很难实现横向扩展，RDBMS可以很容易通过添加更多设备来支持更大规模的数据
9. RDBMS需要定义数据库模式，严格遵守数据定义，NoSQL一般不存在数据库模式，可以自由灵活定义并存储各种不同类型的数据
10. RDBMS借助于索引机制可以实现快速查询，很多NoSQL数据库没有面向复杂查询的索引
11. 下列哪一项不属于NoSQL的四大类型？D
12. 文档数据库
13. 图数据库
14. 列族数据库
15. 时间戳数据库
16. 下列关于键值数据库的描述，哪一项是错误的？D
17. 扩展性好，灵活性好
18. 大量写操作时性能高
19. 无法存储结构化信息
20. 条件查询效率高
21. 下列关于列族数据库的描述，哪一项是错误的？A
22. 查找速度慢，可扩展性差
23. 功能较少，大都不支持强事务一致性
24. 容易进行分布式扩展
25. 复杂性低
26. 下列哪一项不属于数据库事务具有ACID四性？A
27. 间断性
28. 原子性
29. 一致性
30. 持久性
31. 下面关于MongoDB说法，哪一项是正确的？D
32. 具有较差的水平可扩展性
33. 设置个别属性的索引来实现更快的排序
34. 提供了一个面向文档存储，操作复杂
35. 可以实现替换完成的文档（数据）或者一些指定的数据字段
36. 下列关于NoSQL与关系数据库的比较，哪个说法是错误的？D
37. 在一致性方面，RDBMS强于NoSQL
38. 在数据完整性方面，RDBMS容易实现
39. 在扩展性方面，NoSQL 比较好
40. 在可用性方面，NoSQL优于RDBMS
41. 关于文档数据库的说法，下列哪一项是错误的？A
42. 数据是规则的
43. 性能好（高并发）
44. 缺乏统一的查询语法
45. 复杂性低

#### 多选题

1. 关系数据库已经无法满足Web2.0的需求，主要表现在以下几个方面？ABC
2. 无法满足海量数据的管理需求
3. 无法满足数据高并发的需求
4. 无法满足高可扩展性和高可用性的需求
5. 使用难度高
6. 下列关于MySQL集群的描述，哪些是正确的？ABCD
7. 复杂性：部署、管理、配置很复杂
8. 数据库复制：MySQL主备之间一般采用复制方式，很多时候是异步复制
9. 扩容问题：如果系统压力过大需要增加新的机器，这个过程涉及数据重新划分
10. 动态数据迁移问题：如果某个数据库组压力过大，需要将其中部分数据迁移出去
11. 关系数据库引以为傲的两个关键特性（完善的事务机制和高效的查询机制），到了Web2.0时代却成了鸡肋，主要表现在以下哪几个方面？ACD
12. Web2.0 网站系统通常不要求严格的数据库事务
13. Web2.0 网站系统基本上不用关系数据库来存储数据
14. Web2.0 并不要求严格的读写实时性
15. Web2.0 通常不包含大量复杂的SQL 查询
16. 下面关于NoSQL与关系数据库的比较，哪些是正确的？ABCD
17. 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
18. 关系数据库可扩展性较差，无法较好支持海量数据存储
19. NoSQL可以支持超大规模数据存储
20. NoSQL 数据库缺乏数学理论基础，复杂查询性能不高
21. 下列关于文档数据库的描述，哪些是正确的？AD
22. 性能好（高并发），灵活性高
23. 具备统一的查询语法
24. 文档数据库支持文档间的事务
25. 复杂性低，数据结构灵活
26. 下列关于图数据库的描述，哪些是正确的？ABCD
27. 专门用于处理具有高度相互关联关系的数据
28. 比较适合于社交网络、模式识别、依赖分析、推荐系统以及路径寻找等问题
29. 灵活性高，支持复杂的图算法
30. 复杂性高，只能支持一定的数据规模
31. NoSQL的三大基石？ABC
32. CAP
33. 最终一致性
34. BASE
35. DN8
36. 关于NoSQL的三大基石之一的CAP，下列哪些说法是正确的？ACD
37. 一致性，是指任何一个读操作总是能够读到之前完成的写操作的结果量
38. 一个分布式系统可以同时满足一致性、可用性和分区容忍性这三个需求
39. 可用性，是指快速获取数据
40. 分区容忍性，是指当出现网络分区的情况时（即系统中的一部分节点无法和其他节点进行通信），分离的系统也能够正常运行
41. 当处理CAP的问题时，可以有哪几个明显的选择？ABC
42. CA：也就是强调一致性（C）和可用性（A），放弃分区容忍性（P）
43. CP：也就是强调一致性（C）和分区容忍性（P），放弃可用性（A）
44. AP：也就是强调可用性（A）和分区容忍性（P），放弃一致性（C）
45. CAP：也就是同时兼顾可用性（A）、分区容忍性（P）和一致性（C），当时系统性能会下降很多
46. 数据库事务具有ACID四性，下面哪几项属于四性？ABD
47. 原子性
48. 持久性
49. 间断性
50. 一致性

### 第六章 云数据库

#### 单选题

1. 下列哪个不属于云计算的优势？D
2. 按需服务
3. 随时服务
4. 通用性
5. 价格不菲
6. 下列关于云数据库的描述，哪个是错误的？C
7. 云数据库是部署和虚拟化在云计算环境中的数据库
8. 云数据库是在云计算的大背景下发展起来的一种新兴的共享基础架构的方法
9. 云数据库价格不菲，维护费用极其昂贵
10. 云数据库具有高可扩展性、高可用性、采用多租形式和支持资源有效分发等特点
11. 下列哪一个不属于云数据库产品？A
12. 本地安装MySQL
13. 阿里云RDS
14. Oracle Cloud
15. 百度云数据库
16. UMP系统是构建在一个大的集群之上的，下列哪一项不属于系统向用户提供的功能？D
17. 读写分离
18. 分库分表
19. 数据安全
20. 资源合并
21. 下列关于UMP系统功能的说法，哪个是错误的？D
22. 充分利用主从库实现用户读写操作的分离，实现负载均衡
23. UMP系统实现了对于用户透明的读写分离功能
24. 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）
25. UMP系统只设计了一种机制来保证数据安全
26. 下列关于阿里云RDS的说法，哪个是错误的？D
27. RDS是阿里云提供的关系型数据库服务
28. RDS由专业数据库管理团队维护
29. RDS具有安全稳定、数据可靠、自动备份
30. RDS实例，是用户购买RDS服务的基本单位，在实例中，用户只能创建一个数据库
31. 下面哪一项不是云数据库的特性？B
32. 动态可扩展
33. 高成本
34. 易用性
35. 大规模并行处理
36. 下列哪个不是UMP系统中的角色？D
37. Controller服务器
38. Proxy服务器
39. 愚公系统
40. 阿斯隆服务器
41. 关于UMP系统架构依赖的开源组件Mnesia，说法错误的是哪一项？B
42. Mnesia是一个分布式数据库管理系统
43. Mnesia的数据库模式(schema)只能在未运行前静态重配置
44. Mnesia的这些特性，使其在开发云数据库时被用来提供分布式数据库服务
45. Mnesia支持事务，支持透明的数据分片
46. 关于UMP系统架构的Controller 服务器，说法错误的是哪一项？C
47. Controller服务器向UMP集群提供各种管理服务
48. Controller服务器上运行了一组Mnesia分布式数据库服务
49. 当其它服务器组件需要获取用户数据时，不可以向Controller服务器发送请求获取数据
50. 为了避免单点故障，保证系统的高可用性，UMP系统中部署了多台Controller服务器

#### 多选题

1. 云数据库具有以下哪些特性？ABCD
2. 动态可扩展
3. 高可用性
4. 免维护
5. 安全
6. 下列关于云数据库的描述，哪些是正确的？ABCD
7. Amazon是云数据库市场的先行者
8. Google Cloud SQL是谷歌公司推出的基于MySQL的云数据库
9. 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
10. 云数据库并没有专属于自己的数据模型
11. UMP系统架构设计遵循了以下哪些原则？ABCD
12. 保持单一的系统对外入口，并且为系统内部维护单一的资源池
13. 消除单点故障，保证服务的高可用性
14. 保证系统具有良好的可伸缩，能够动态地增加、删减计算与存储节点
15. 保证分配给用户的资源也是弹性可伸缩的
16. UMP系统架构依赖的哪些开源组件？ABCD
17. Mnesia
18. LVS
19. RabbitMQ
20. ZooKeeper
21. 下列关于UMP系统架构的描述，哪些是正确的？ACD
22. 信息统计服务器定期将采集到的用户的连接数
23. Web控制台无法向用户提供系统管理界面
24. LVS(Linux Virtual Server)即Linux虚拟服务器
25. UMP系统借助于LVS来实现集群内部的负载均衡
26. 为什么说云数据库是个性化数据存储需求的理想选择？ABCD
27. 云数据库可以满足大企业的海量数据存储需求
28. 云数据库可以满足中小企业的低成本数据存储需求
29. 云数据库可以满足企业动态变化的数据存储需求
30. 前期零投入、后期免维护的数据库服务，可以很好满足它们的需求
31. 下列关于云数据库与其他数据库的关系，哪些是正确的？ABD
32. 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
33. 云数据库并没有专属于自己的数据模型，云数据库所采用的数据模型可以是关系数据库所使用的关系模型
34. 同一个公司只能提供采用不同数据模型的单个云数据库服务
35. 许多公司在开发云数据库时，后端数据库都是直接使用现有的各种关系数据库或NoSQL数据库产品
36. 以下哪些是Amazon的云数据库产品？ABC
37. Amazon RDS：云中的关系数据库
38. Amazon SimpleDB：云中的键值数据库
39. Amazon DynamoDB：云中的数据仓库
40. Amazon ElastiCache：云中的分布式内存缓存
41. Microsoft的云数据库产品SQL Azure具有以下哪些特性？ABCD
42. 属于关系型数据库：支持使用TSQL来管理、创建和操作云数据库
43. 支持存储过程：它的数据类型、存储过程和传统的SQL Server具有很大的相似性
44. 支持大量数据类型
45. 支持云中的事务：支持局部事务，但是不支持分布式事务

### 第七章 MapReduce

#### 单选题

1. 下列传统并行计算框架，说法错误的是哪一项？B
2. 刀片服务器、高速网、SAN，价格贵，扩展性差上
3. 共享式(共享内存/共享存储)，容错性好
4. 编程难度高
5. 实时、细粒度计算、计算密集型
6. 下列关于MapReduce模型的描述，错误的是哪一项？D
7. MapReduce采用“ 分而治之”策略
8. MapReduce设计的一个理念就是“ 计算向数据靠拢”
9. MapReduce框架采用了Master/Slave架构
10. MapReduce应用程序只能用Java来写
11. MapReduce1.0的体系结构中，JobTracker是主要任务是什么？A
12. 负责资源监控和作业调度，监控所有TaskTracker与Job的健康状况
13. 使用“slot”等量划分本节点上的资源量（CPU、内存等）
14. 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
15. 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）
16. 下列关于MapReduce工作流程，哪个描述是正确的？A
17. 所有的数据交换都是通过MapReduce框架自身去实现的
18. 不同的Map任务之间会进行通信
19. 不同的Reduce任务之间可以发生信息交换
20. 用户可以显式地从一台机器向另一台机器发送消息
21. 下列关于MapReduce的说法，哪个描述是错误的？D
22. MapReduce具有广泛的应用，比如关系代数运算、分组与聚合运算等
23. 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
24. 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
25. 不同的Map任务之间可以进行通信
26. 下列关于Map和Reduce函数的描述，哪个是错误的？C
27. 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
28. 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
29. 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
30. 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
31. 下面哪一项不是MapReduce体系结构主要部分？A
32. Client
33. JobTracker
34. TaskTracker以及Task
35. Job
36. 关于MapReduce1.0的体系结构的描述，下列说法错误的是？A
37. Task 分为Map Task 和Reduce Task 两种，分别由JobTracker 和TaskTracker 启动
38. slot 分为Map slot 和Reduce slot 两种，分别供MapTask 和Reduce Task 使用
39. TaskTracker 使用“slot”等量划分本节点上的资源量（CPU、内存等）
40. 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）
41. 下列说法错误的是？C
42. Hadoop MapReduce是MapReduce的开源实现，后者比前者使用门槛低很多
43. MapReduce采用非共享式架构，容错性好
44. MapReduce主要用于批处理、实时、计算密集型应用
45. MapReduce采用“ 分而治之”策略

#### 多选题

1. MapReduce相较于传统的并行计算框架有什么优势？ABCD
2. 非共享式，容错性好
3. 普通PC机，便宜，扩展性好
4. 编程简单，只要告诉MapReduce做什么即可
5. 批处理、非实时、数据密集型
6. MapReduce体系结构主要由以下那几个部分构成？ABCD
7. Client
8. JobTracker
9. TaskTracker
10. Task
11. 下列关于MapReduce的体系结构的描述，说法正确的有？ABD
12. 用户编写的MapReduce程序通过Client提交到JobTracker端
13. JobTracker负责资源监控和作业调度
14. TaskTracker监控所有TaskTracker与Job的健康状况
15. TaskTracker 使用“slot”等量划分本节点上的资源量（CPU、内存等）
16. MapReduce的作业主要包括什么？AB
17. 从磁盘或从网络读取数据，即IO密集工作
18. 计算数据，即CPU密集工作
19. 针对不同的工作节点选择合适硬件类型
20. 负责协调集群中的数据存储
21. 对于MapReduce 而言，其处理单位是split。split 是一个逻辑概念，它包含哪些元数据信息？AB
22. 数据起始位置
23. 数据长度
24. 数据所在节点
25. 数据大小
26. 下列关于Map 端的Shuffle的描述，哪些是正确的？BC
27. MapReduce默认为每个Map任务分配1000MB缓存
28. 多个溢写文件归并成一个或多个大文件，文件中的键值对是排序的
29. 当数据很少时，不需要溢写到磁盘，直接在缓存中归并，然后输出给Reduce
30. 每个Map任务分配多个缓存，使得任务运行更有效率
31. MapReduce的具体应用包括哪些？ABCD
32. 关系代数运算（选择、投影、并、交、差、连接）
33. 分组与聚合运算
34. 矩阵-向量乘法
35. 矩阵乘法
36. MapReduce执行的全过程包括以下哪几个主要阶段？ABCD
37. 从分布式文件系统读入数据
38. 执行Map任务输出中间结果
39. 通过 Shuffle阶段把中间结果分区排序整理后发送给Reduce任务
40. 执行Reduce任务得到最终结果并写入分布式文件系统
41. 下列说法正确的是？ABCD
42. MapReduce体系结构主要由四个部分组成，分别是：Client、JobTracker、TaskTracker以及Task
43. Task 分为Map Task 和Reduce Task 两种，均由TaskTracker 启动
44. 在MapReduce工作流程中，所有的数据交换都是通过MapReduce框架自身去实现的
45. 在MapReduce工作流程中，用户不能显式地从一台机器向另一台机器发送消息

### 第八章 Hadoop再探讨

#### 单选题

1. Hadoop1.0的核心组件（仅指MapReduce和HDFS，不包括Hadoop生态系统内的Pig、Hive、HBase等其他组件），下列哪项是它的不足？B
2. 抽象层次高
3. 表达能力有限，抽象层次低，需人工编码
4. 价格昂贵
5. 可维护性低
6. 下面哪个选项不属于Hadoop1.0 的问题？D
7. 单一名称节点，存在单点失效问题
8. 单一命名空间，无法实现资源隔离
9. 资源管理效率低
10. 很难上手
11. 下列哪项是Hadoop生态系统中Spark的功能？D
12. 处理大规模数据的脚本语言
13. 工作流和协作服务引擎，协调Hadoop上运行的不同任务
14. 不支持DAG作业的计算框架
15. 基于内存的分布式并行编程框架，具有较高的实时性，并且较好支持迭代计算
16. 在Hadoop生态系统中，Kafka主要解决Hadoop 中存在哪些的问题？A
17. Hadoop生态系统中各个组件和其他产品之间缺乏统一的、高效的数据交换中介
18. 不同的MapReduce任务之间存在重复操作，降低了效率
19. 延迟高，而且不适合执行迭代计算
20. 抽象层次低，需要手工编写大量代码
21. 下列哪一个不属于HDFS1.0 中存在的问题？A
22. 无法水平扩展
23. 单点故障问题
24. 单一命名空间
25. 系统整体性能受限于单个名称节点的吞吐量
26. 关于HDFS Federation 的设计的描述，哪个是错误的？A
27. 属于不同命名空间的块可以构成同一个“块池”
28. HDFS Federation中，所有名称节点会共享底层的数据节点存储资源，数据节点向所有名称节点汇报
29. 设计了多个相互独立的名称节点
30. HDFS的命名服务能够水平扩展
31. 下列关于MapReduce1.0的描述，错误的是？B
32. JobTracker“大包大揽”导致任务过重
33. 不存在单点故障
34. 容易出现内存溢出（分配资源只考虑MapReduce任务数，不考虑CPU、内存）
35. 资源划分不合理（强制划分为slot ，包括Map slot和Reduce slot）
36. 下列哪个不属于YARN体系结构中ResourceManager的功能？D
37. 处理客户端请求
38. 监控NodeManager
39. 资源分配与调度
40. 处理来自ApplicationMaster的命令
41. 下列哪个不属于YARN体系结构中ApplicationMaster的功能？D
42. 任务调度、监控与容错
43. 为应用程序申请资源
44. 将申请的资源分配给内部任务
45. 处理来自ResourceManger的命令

#### 多选题

1. 下列选项中，哪些属于Hadoop1.0的核心组件的不足之处？ABCD
2. 实时性差（适合批处理，不支持实时交互式）
3. 资源浪费（Map和Reduce分两阶段执行）
4. 执行迭代操作效率低
5. 难以看到程序整体逻辑
6. Hadoop的优化与发展主要体现在哪几个方面？ABC
7. Hadoop自身核心组件MapReduce的架构设计改进
8. Hadoop自身核心组件HDFS的架构设计改进
9. Hadoop生态系统其它组件的不断丰富
10. Hadoop生态系统减少不必要的组件，整合系统
11. 下列哪些属于Hadoop2.0相对于Hadoop1.0的改进？ABCD
12. 设计了HDFS HA
13. 提供名称节点热备机制
14. 设计了HDFS Federation，管理多个命名空间
15. 设计了新的资源管理框架YARN
16. 下面哪个属于不断完善的Hadoop生态系统中的组件？ABC
17. Pig
18. Tez
19. Kafka
20. DN8
21. HDFS1.0 主要存在哪些问题？ACD
22. 单点故障问题
23. 不可以水平扩展
24. 单个名称节点难以提供不同程序之间的隔离性
25. 系统整体性能受限于单个名称节点的吞吐量
26. HDFS Federation 相对于HDFS1.0 的优势主要体现在哪里？BCD
27. 能够解决单点故障问题
28. HDFS 集群扩展性
29. 性能更高效
30. 良好的隔离性
31. JobTracker主要包括哪三大功能？ABC
32. 资源管理
33. 任务调度
34. 任务监控
35. 数据即服务
36. YARN 体系结构主要包括哪三部分？ABD
37. ResourceManager
38. NodeManager
39. DataManager
40. ApplicationMaster
41. 在YARN体系结构中，ApplicationMaster主要功能包括哪些？ABC
42. 当用户作业提交时，ApplicationMaster与ResourceManager协商获取资源，ResourceManager会以容器的形式为ApplicationMaster分配资源
43. 把获得的资源进一步分配给内部的各个任务（Map任务或Reduce任务），实现资源的“二次分配”
44. 定时向ResourceManager发送“心跳”消息，报告资源的使用情况和应用的进度信息
45. 向ResourceManager汇报作业的资源使用情况和每个容器的运行状态

### 第九章 数据仓库Hive

#### 单选题

1. 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的解释处理比较相似，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

2. 下列关于Hive基本操作命令的解释错误的是:(C)

A create database userdb;//创建数据库userdb

B create table if not exists usr(id bigint,name string,age int);//如果usr表不存在，创建表usr，含三个属性id,name,age

C load data local inpath ‘/usr/local/data’ overwrite into table usr; //把目录’/usr/local/data’下的数据文件中的数据以追加的方式装载进usr表

D insert overwrite table student select \* from user where age>10; //向表usr1中插入来自usr表的age大于10的数据并覆盖student表中原有数据

3.下面哪个不可能是Hive的执行引擎：（C）

A．MapReduce

B.Tez

C.Storm

D.Spark

4.下面描述错误的是：（C）

A．Hive的功能十分强大，可以支持采用SQL方式查询Hadoop平台上的数据

B．在实际应用中，Hive也暴露出不稳定的问题，在极少数情况下，甚至会出现端口不响应或者进程丢失的问题

C．在Hive HA中，在Hadoop集群上构建的数据仓库是由单个Hive实例进行管理的

D．在Hive HA中，客户端的查询请求首先访问HAProxy，由HAProxy对访问请求进行转发

5.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

6.下面关于Hive的描述错误的是：（D）

A．Hive是一个构建在Hadoop之上的数据仓库工具

B．Hive是由Facebook公司开发的

C．Hive在某种程度上可以看作是用户编程接口，其本身并不存储和处理数据

D．Hive定义了简单的类似SQL的查询语言——HiveQL，它与大部分SQL语法无法兼容

7.下面关于Hive的描述错误的是：（D）

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

8.关于Hive和传统关系数据库的对比分析，下面描述错误的是：（B）

A．Hive一般依赖于分布式文件系统HDFS，而传统数据库则依赖于本地文件系统

B．传统的关系数据库可以针对多个列构建复杂的索引，Hive不支持索引

C．Hive和传统关系数据库都支持分区

D. 传统关系数据库很难实现横向扩展，Hive具有很好的水平扩展性

9. 以下哪个不是Hive的用户接口模块：（A）

A．PMI

B．HWI（Hive Web Interface）

C．JDBC/ODBC

D.Thrift Server

#### 多选题

1. 下列说法正确的是：（BCD）

A 数据仓库Hive不需要借助于HDFS就可以完成数据的存储

B Impala和Hive、HDFS、HBase等工具可以统一部署在一个Hadoop平台上

C Hive本身不存储和处理数据，依赖HDFS存储数据，依赖MapReduce处理数据

D HiveQL语法与传统的SQL语法很相似

2. Impala主要由哪几个部分组成:(ABC)

A Impalad

B State Store

C CLI

D Hive

3. 以下属于Hive的基本数据类型是:(ABCD)

A TINYINT

B FLOAT

C STRING

D BINARY

4.数据仓库Hive的执行引擎可以是：（ABD）

A.Tez

B.MapReduce

C.Pig

D.Spark

5.以下哪些是数据仓库的特性：（ABCD）

A．面向主题的（Subject Oriented）

B．集成的（Integrated）

C．相对稳定的（Non-Volatile）

D．反映历史变化

6.传统数据仓库面临哪些挑战：（ABD）

A．无法满足快速增长的海量数据存储需求

B．无法有效处理不同类型的数据

C.具有很强的扩展性

D. 计算和处理能力不足

7. Hadoop生态系统中Hive与其他部分的关系的描述正确的是：（ABCD）

A．HDFS作为高可靠的底层存储，用来存储Hive的海量数据

B. MapReduce对这些海量数据进行批处理，实现Hive的高性能计算

C．当采用MapRedue作为Hive的执行引擎时，用HiveQL语句编写的处理逻辑，最终都要转化为MapReduce任务来运行

D．HBase与Hive的功能是互补的，它实现了Hive不能提供的功能

8.关于Hive与传统关系数据库的对比分析，下面描述正确的是：（ABD）

A．在数据存储方面，Hive一般依赖于分布式文件系统HDFS，而传统数据库则依赖于本地文件系统

B．在索引方面，传统的关系数据库可以针对多个列构建复杂的索引，大幅度提升数据查询性能，而Hive不像传统的关系型数据库那样有键的概念，它只能提供有限的索引功能

C．在分区方面，传统的数据库提供分区功能来改善大型表以及具有各种访问模式的表的可伸缩性、可管理性，以及提高数据库效率；Hive不支持分区功能

D．在执行引擎方面，传统的关系数据库依赖自身的执行引擎，而Hive则依赖于MapReduce、Tez和Spark等执行引擎

9.Hive主要由哪三个模块组成：（ABD）

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

10.当采用MapReduce作为Hive的执行引擎时，下面描述正确的是：（ABC）

A．当用户向Hive输入一段命令或查询（即HiveQL语句）时，Hive需要与Hadoop交互工作来完成该操作

B．命令或查询首先进入到驱动模块，由驱动模块中的编译器进行解析编译，并由优化器对该操作进行优化计算，然后交给执行器去执行

C．执行器通常的任务是启动一个或多个MapReduce任务，有时也不需要启动MapReduce任务

D. 执行器通常的任务一定会包含Map和Reduce操作

### 第十章 Spark

#### 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)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

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

1. 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

9、Spark生态系统组件MLlib的应用场景是？D

(A)图结构数据的处理

(B)基于历史数据的交互式查询

(C)复杂的批量数据处理

(D)基于历史数据的数据挖掘

#### 多选题

1. Spark具有以下哪几个主要特点？ABC

(A)运行速度快

(B)容易使用

(C)通用性

(D)运行模式单一

2、Scala具有以下哪几个主要特点？ABCD

(A)Scala的优势是提供了REPL（Read-Eval-Print Loop，交互式解释器），提高程序开发效率

(B)Scala兼容Java，运行速度快，且能融合到Hadoop生态圈中

(C)Scala具备强大的并发性，支持函数式编程

(D)Scala可以更好地支持分布式系统

3、下列哪些选项属于Hadoop的缺点？ABCD

(A)表达能力有限

(B)磁盘IO开销大

(C)延迟高

(D)在前一个任务执行完成之前，其他任务就无法开始，难以胜任复杂、多阶段的计算任务

1. 下列说法中，哪些选项描述正确？AB

(A)Spark在借鉴Hadoop MapReduce优点的同时，很好地解决了MapReduce所面临的问题

(B)Spark的计算模式也属于MapReduce，但不局限于Map和Reduce操作

(C)Hadoop MapReduce编程模型比Spark更灵活

(D)Hadoop MapReduce提供了内存计算，可将中间结果放到内存中，对于迭代运算效率更高

5、在实际应用中，大数据处理主要包括以下哪三个类型？ABD

(A)复杂的批量数据处理：通常时间跨度在数十分钟到数小时之间

(B)基于历史数据的交互式查询：通常时间跨度在数十秒到数分钟之间

(C)基于实时数据流的数据处理：通常时间跨度在数十秒到数分钟之间

(D)基于实时数据流的数据处理：通常时间跨度在数百毫秒到数秒之间

6、在实际应用中，当采用多种计算架构来满足不同应用场景需求时，大数据处理难免会带来哪些问题？ABCD

(A)不同场景之间输入输出数据无法做到无缝共享，通常需要进行数据格式的转换

(B)不同的软件需要不同的开发和维护团队

(C)需要较高的使用成本

(D)比较难以对同一个集群中的各个系统进行统一的资源协调和分配

7、与Hadoop MapReduce计算框架相比，Spark所采用的Executor具有哪些优点？AB

(A)利用多线程来执行具体的任务，减少任务的启动开销

(B)Executor中有一个BlockManager存储模块，有效减少IO开销

(C)提供了一种高度受限的共享内存模型

(D)不同场景之间输入输出数据能做到无缝共享

8、Spark运行架构具有以下哪些特点？ABCD

(A)每个Application都有自己专属的Executor进程，并且该进程在Application运行期间一直驻留

(B)Executor进程以多线程的方式运行Task

(C)Spark运行过程与资源管理器无关，只要能够获取Executor进程并保持通信即可

(D)Task采用了数据本地性和推测执行等优化机制

9、Spark采用RDD以后能够实现高效计算的原因主要在于？ABC

(A)高效的容错性

(B)中间结果持久化到内存，数据在内存中的多个

(C)存放的数据可以是Java对象，避免了不必要的对象序列化和反序列化

(D)采用数据复制实现容错

10、Spark支持哪三种不同类型的部署方式？ABC

(A)Standalone（类似于MapReduce1.0，slot为资源分配单位）

(B)Spark on Mesos（和Spark有血缘关系，更好支持Mesos）

(C)Spark on YARN

(D)Spark on HDFS

### 第十一章 流计算

#### 单选题

1. 下列关于流计算的说法，哪项是错误的？D
2. 实时获取来自不同数据源的海量数据，经过实时分析处理，获得有价值的信息
3. 流计算秉承一个基本理念，即 数据的价值随着时间的流逝而降低
4. 对于一个流计算系统来说，它应该支持TB级甚至是PB级的数据规模
5. 流计算只需要保证较低的延迟时间，即只达到秒级别即可处理一切问题
6. 下列关于数据处理流程，说法有误的是？D
7. 在传统的数据处理流程中，存储的数据是旧的
8. 在传统的数据处理流程中，需要用户主动发出查询来获取结果
9. 传统的数据处理流程，需要先采集数据并存储在关系数据库等数据管理系统中
10. 流计算的处理流程一般包含三个阶段：数据实时采集、数据批量计算、实时查询服务
11. 下列哪项不属于流计算的处理流程的三个阶段？B
12. 数据实时采集
13. 数据批量采集
14. 数据实时计算
15. 实时查询服务
16. 下列哪一项不属于Storm的特点？B
17. 支持各种编程语言：Storm支持使用各种编程语言来定义任务
18. 容错性：Storm需要人工进行故障节点的重启、任务的重新分配
19. 可扩展性：Storm的并行特性使其可以运行在分布式集群中
20. 免费、开源：Storm是一款开源框架，可以免费使用
21. 下列关于Storm设计思想，描述有误的是？A
22. Storm将Streams的状态转换过程抽象为Spout
23. Storm认为每个Stream都有一个源头，并把这个源头抽象为Spout
24. Storm将Spouts和Bolts组成的网络抽象成Topology
25. Topology里面的每个处理组件（Spout或Bolt）都包含处理逻辑， 而组件之间的连接则表示数据流动的方向
26. 下列关于Storm和Hadoop架构组件功能对应关系，哪个是错误的？C
27. Hadoop Job 对应 Storm Topology
28. Hadoop TaskTracker对应 Storm Supervisor
29. Hadoop JobTracker对应 Storm Spout
30. Hadoop Reduce对应 Storm Bolt
31. 下列哪个说法是错误的？B
32. 从编程的灵活性来讲，Storm是比较理想的选择，它使用Apache Thrift，可以用任何编程语言来编写拓扑结构（Topology）
33. Spark Streaming可以实现毫秒级的流计算
34. Spark Streaming无法实现毫秒级的流计算
35. 在Spark上可以统一部署Spark SQL，Spark Streaming、MLlib，GraphX等组件，提供便捷的一体化编程模型
36. 下列关于Storm的描述，有误的是？A
37. Storm将流数据Stream描述成一个有限的Tuple序列
38. Storm保证每个消息都能完整处理
39. Storm认为每个Stream都有一个源头，并把这个源头抽象为Spout
40. Bolt可以执行过滤、函数操作、Join、操作数据库等任何操作
41. 下列哪个关于Stream Groupings的描述，有误的是？A
42. FieldsGrouping：广播发送，每一个Task都会收到所有的Tuple
43. GlobalGrouping：全局分组，所有的Tuple都发送到同一个Task中
44. NonGrouping：不分组，和ShuffleGrouping类似，当前Task的执行会和它的被订阅者在同一个线程中执行
45. DirectGrouping：直接分组，直接指定由某个Task来执行Tuple的处理

#### 多选题

1. 流数据具有以下哪些特征？ABC
2. 数据快速持续到达，潜在大小也许是无穷无尽的
3. 数据来源众多，格式复杂
4. 注重数据的整体价值，不过分关注个别数据
5. 系统可以控制将要处理的新到达的数据元素的顺序
6. 下列关于批量计算和实时计算的说法，正确的有？ACD
7. 批量计算：充裕时间处理静态数据，如Hadoop
8. 静态数据不适合采用批量计算，因为它不适合用传统的关系模型建模
9. 流数据必须采用实时计算
10. 流数据的响应时间一般为秒级，甚至需要毫秒级
11. 对于一个流计算系统来说，它应达到以下哪些需求？ABCD
12. 高性能：处理大数据的基本要求，如每秒处理几十万条数据
13. 实时性：保证较低的延迟时间，达到秒级别，甚至是毫秒级别
14. 分布式：支持大数据的基本架构，必须能够平滑扩展
15. 可靠性：能可靠地处理流数据
16. 下列关于流计算与Hadoop的说法，正确的有？ABCD
17. Hadoop擅长批处理，不适合流计算
18. MapReduce是专门面向静态数据的批量处理的
19. Hadoop设计的初衷是面向大规模数据的批量处理
20. MapReduce不适合用于处理持续到达的动态数据
21. 下列关于流处理系统与传统的数据处理系统的描述，正确的有？ABC
22. 流处理系统处理的是实时的数据，而传统的数据处理系统处理的是预先存储好的静态数据
23. 用户通过流处理系统获取的是实时结果，而通过传统的数据处理系统，获取的是过去某一时刻的结果
24. 流处理系统无需用户主动发出查询，实时查询服务可以主动将结果推送给用户
25. 传统的数据处理系统系统无需用户主动发出查询，批量查询服务可以主动将结果推送给用户
26. Storm具有以下哪些特点？ABCD
27. 整合性：Storm可方便地与队列系统和数据库系统进行整合
28. 简易的API：Storm的API在使用上即简单又方便
29. 容错性：Storm可自动进行故障节点的重启、任务的重新分配
30. 可扩展性：Storm的并行特性使其可以运行在分布式集群中
31. 目前，Storm中的Stream Groupings有如下哪几种方式？ABC
32. AllGrouping：广播发送，每一个Task都会收到所有的Tuple
33. GlobalGrouping：全局分组，所有的Tuple都发送到同一个Task中
34. FieldsGrouping：按照字段分组，保证相同字段的Tuple分配到同一个Task中
35. DirectGrouping：随机分组，随机分发Stream中的Tuple，保证每个Bolt的Task接收Tuple数量大致一致
36. 下列关于Storm框架，描述正确的有？ABCD
37. worker:每个worker进程都属于一个特定的Topology
38. executor：executor是产生于worker进程内部的线程
39. task:实际的数据处理由task完成
40. 在Topology的生命周期中，每个组件的task数目是不会发生变化的，而executor的数目却不一定
41. 下列关于Spark Streaming与Storm的描述，哪些是正确的的？ABD
42. Spark Streaming无法实现毫秒级的流计算
43. Storm可以实现毫秒级响应件
44. Storm无法用于实时计算
45. Spark Streaming采用的小批量处理的方式使得它可以同时兼容批量和实时数据处理的逻辑和算法

### 第十二章 Flink

#### 单选题

1.下面描述错误的是：（D）

A．Storm虽然可以做到低延迟，但是无法实现高吞吐，也不能在故障发生时准确地处理计算状态

B．Spark Streaming通过采用微批处理方法实现了高吞吐和容错性，但是牺牲了低延迟和实时处理能力

C．Flink实现了Google Dataflow流计算模型，是一种兼具高吞吐、低延迟和高性能的实时流计算框架，并且同时支持批处理和流处理

D．Spark Streaming可以实现毫秒级响应，而Flink只能实现秒级响应

2. 以下哪个不属于事件驱动型应用？D

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

3.以下哪个不属于数据分析应用？A

A．基于规则的报警

B．移动应用中的产品更新及实验评估分析

C.消费者技术中的实时数据即席分析

D. 大规模图分析

4.以下哪个属于数据流水线应用？B

A．基于规则的报警

B．实时查询索引构建

C．移动应用中的产品更新及实验评估分析

D．费者技术中的实时数据即席分析

#### 多选题

1. 企业数据架构的典型形式包括：（ABC）

A．传统数据处理架构

B．大数据Lambda架构

C．流处理架构

D．循环处理架构

2.Flink的优势包括：（ABCD）

A．同时支持高吞吐、低延迟、高性能

B．同时支持流处理和批处理

C．支持有状态计算

D．具有独立的内存管理

3. 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

4. Flink核心组件栈包括：（ACD）

A．物理部署层

B．算法库层

C．Runtime核心层

D．API&Libraries层

5. Flink系统主要由两个组件组成：（AD）

A. JobManager

B.JobTrackder

C.TaskTracker

D. TaskManager

6.Flink编程模型包括哪几层：（ABCD）

A.SQL

B.Table API

C.DataStream API / DataSet API

D.有状态数据流处理

### 第十三章 图计算

#### 单选题

1. 下列关于图结构数据的描述，错误的是？A

(A)许多非图结构的大数据，通常会被转换为关系模型后进行分析

(B)许多大数据都是以大规模图或网络的形式呈现

(C)图数据结构很好地表达了数据之间的关联性

(D)关联性计算是大数据计算的核心

1. 在Pregel计算模型中，图中的每个顶点会对应一个计算单元，下列哪一项不属于计算单元的成员变量？D
2. 顶点值（Vertex value）：顶点对应的PR值
3. 出射边（Out edge）：只需要表示一条边，可以不取值
4. 消息（Message）：传递的消息
5. 入射边（Iut edge）：只需要表示一条边，可以不取值
6. 采用MapReduce实现PageRank的计算过程包括三个阶段，下列哪一项是错的？D

(A)第一阶段：解析网页

(B)第二阶段：PageRank分配

(C)第三阶段：收敛阶段

(D)第一阶段：收集网页

1. 下列关于PageRank 算法在MapReduce 中的实现的描述，错误的是？A
2. 解析网页的任务就是分析一个页面的链接数，但是不赋初值
3. PageRank 分配就是多次迭代计算页面的PageRank值
4. 收敛阶段的任务就是由一个非并行组件决定是否达到收敛
5. 一般判断是否收敛的条件是所有网页的PageRank值不再变化，或者运行30次以后我们就认为已经收敛了
6. 下列哪个不属于PageRank算法在Pregel和MapReduce中实现方式的区别？D
7. Pregel将PageRank处理对象看成是连通图，而MapReduce则将其看成是键值对
8. Pregel将计算细化到顶点，同时在顶点内控制循环迭代次数
9. MapReduce将计算批量化处理，按任务进行循环迭代控制
10. 图算法如果用Pregel实现，需要一系列的Pregel的调用
11. 下列哪项不属于在Vetex类中的值类型参数？A

(A)顶点的出度

(B)顶点

(C)边

(D)消息

1. 在每个超步中，Worker调用顶点上的Compute()函数，下列哪个参数是不需要传递？D

(A)该顶点的当前值

(B)一个接收到的消息的迭代器

(C)一个出射边的迭代器

(D)一个入射边的迭代器

1. 下列哪一项不属于每个超步的组件？A

(A)全局计算

(B)局部计算

(C)通讯

(D)栅栏同步

1. 在一个Worker中，它所管辖的分区的状态信息是保存在内存中的，下列哪一项不属于分区的状态信息？D
2. 顶点的当前值
3. 消息队列
4. 标志位
5. 一个接收到的消息的迭代器

#### 多选题

1. 很多传统的图计算算法都存在以下哪些典型问题？ABC
2. 常常表现出比较差的内存访问局部性
3. 针对单个顶点的处理工作过少
4. 计算过程中伴随着并行度的改变
5. 计算过程简易
6. 针对大型图（比如社交网络和网络图）的计算问题，哪些说法是正确的？ABCD
7. 为特定的图应用定制相应的分布式实现：通用性不好
8. 基于现有的分布式计算平台进行图计算：在性能和易用性方面往往无法达到最优
9. 使用单机的图算法库，但是，在可以解决的问题的规模方面具有很大的局限性
10. 使用已有的并行图计算系统，但是，对大规模分布式系统非常重要的一些方面（比如容错），无法提供较好的支持
11. 针对大型图的计算，目前通用的图计算软件主要哪些？AB
12. 基于遍历算法的、实时的图数据库，如Neo4j、OrientDB、DEX和 Infinite Graph
13. 以图顶点为中心的、基于消息传递批处理的并行引擎，如GoldenOrb、Giraph、Pregel和Hama
14. 基于遍历算法的、实时的图数据库，如GoldenOrb、Giraph、Pregel和Hama
15. 以图顶点为中心的、基于消息传递批处理的并行引擎，如Neo4j、OrientDB、DEX和 Infinite Graph
16. 一次BSP(Bulk Synchronous Parallel Computing Model，又称“大同步”模型)计算过程包括一系列全局超步（所谓的超步就是计算中的一次迭代），每个超步主要包括哪些组件？ABC
17. 局部计算：每个参与的 处理器都有自身的计算任务
18. 通讯：处理器群相互交换数据
19. 栅栏同步：当一个处理器遇到“路障”（或栅栏），会等到其他所有处理器完成它们的计算步骤
20. 基于现有的分布式计算平台进行图计算
21. 下列说法正确的是？ABCD
22. 在Pregel执行计算过程时，在每个超步中都会并行调用每个顶点上定义的Compute()函数
23. 顶点之间的通讯是借助于消息传递机制来实现的
24. 在默认情况下，Pregel计算框架并不会开启Combiner功能
25. 通常只对那些满足交换律和结合律的操作才可以去开启Combiner功能
26. 在Pregel计算框架中，图的保存格式多种多样，具体包括哪些？ABC
27. 文本文件
28. 关系数据库
29. 键值数据库
30. 语音文件
31. 下列对于Pregel的执行过程，哪些是正确的？ABCD
32. 选择集群中的多台机器执行图计算任务，每台机器上运行用户程序的一个副本
33. Master把一个图分成多个分区，并把分区分配到多个Worker
34. Master会把用户输入划分成多个部分，通常是基于文件边界进行划分
35. Master向每个Worker发送指令，Worker收到指令后，开始运行一个超步
36. 在一个Worker中，它所管辖的分区的状态信息是保存在内存中的，分区中的顶点的状态信息包括？ABCD
37. 顶点的当前值
38. 以该顶点为起点的出射边列表，每条出射边包含了目标顶点ID和边的值
39. 消息队列，包含了所有接收到的、发送给该顶点的消息
40. 标志位，用来标记顶点是否处于活跃状态
41. 在每个超步中，Worker会对自己所管辖的分区中的每个顶点进行遍历，并调用顶点上的Compute()函数，在调用时，会把以下哪些参数传递进去？ABC
42. 该顶点的当前值
43. 一个接收到的消息的迭代器
44. 一个出射边的迭代器
45. 标志位，用来标记顶点是否处于活跃状态
46. 下列说法正确的有？ABCD
47. 在Pregel中，为了获得更好的性能，“标志位”和输入消息队列是分开保存的
48. 在超步S中，当一个Worker在进行顶点处理时，用于当前超步的消息会被处理
49. 需要两个消息队列用于存放作用于当前超步S的消息和作用于下一个超步S+1的消息
50. 每个Worker上都保存了一个或多个分区的状态信息，当一个Worker发生故障时，它所负责维护的分区的当前状态信息就会丢失

### 第十四章 数据可视化

一、单选题

1、 下列关于数据可视化的描述，哪个是错误的？(D)

A.数据可视化是指将大型数据集中的数据以图形图像形式表示

B.利用数据分析和开发工具发现其中未知信息的处理过程

C.数据可视化技术的基本思想是将数据库中每一个数据项作为单个图元素表示

D.将数据的各个属性值以一维数据的形式表示

2、 下列哪个不属于可视化工具？(D)

A.Google Chart API

B.D3

C.Visual.ly

D.Spark

3、 下列说法错误的是？(B)

A.大数据魔镜是一款优秀的国产数据分析软件，可以让用户真正理解探索分析数据

B.Tableau是桌面系统中最简单的商业智能工具软件，是一个用于网页作图、生成互动图形的JavaScript函数库

C.Google Fusion Tables让一般使用者也可以轻松制作出专业的统计地图

D.Modest Maps是一个小型、可扩展、交互式的免费库，提供了一套查看卫星地图的API

4、下面关于Timetoast的描述，哪个是错误的？(D)

A.Timetoast是在线创作基于时间轴事件记载服务的网站

B.提供个性化的时间线服务

C.Timetoast基于 flash 平台，可以在类似flash时间轴上任意加入事件

D.Timetoast是一个提供复杂统计图表的工具

5、 下列关于可视化工具中高级分析工具的说法，错误的是？(B)

A.R是属于GNU系统的一个自由、免费、源代码开放的软件

B.Weka主要用于社交图谱数据可视化分析，可以生成非常酷炫的可视化图形

C.Gephi主要用于社交图谱数据可视化分析，可以生成非常酷炫的可视化图形

D.R通常用于大数据集的统计与分析

二、多选题

1、 在大数据时代，可视化技术可以支持实现哪些目标？（ABCD）

A.观测、跟踪数据

B.分析数据

C.辅助理解数据

D.增强数据吸引力

2、 信息图表是信息、数据、知识等的视觉化表达，下列哪个说法正确？(ABCD)

A.谷歌公司的制图服务接口Google Chart API，可以用来为统计数据并自动生成图片

B.D3是最流行的可视化库之一，是一个用于网页作图、生成互动图形的JavaScript函数库

C. ECharts是由百度公司前端数据可视化团队研发的图表库，可以流畅地运行在PC和移动设备上

D.大数据魔镜是一款优秀的国产数据分析软件，它丰富的数据公式和算法可以让用户真正理解探索分析数据

3、 下列关于数据可视化的描述，正确的有？ABC

（A） 数据可视化是指将大型数据集中的数据以图形图像形式表示

（B） 数据可视化技术的基本思想是将数据库中每一个数据项作为单个图元素表示

（C） 利用数据分析和开发工具发现其中未知信息的处理过程

（D） 将数据的各个属性值以一维数据的形式表示

4、下列说法中，哪些是正确的？ABCD

A.Modest Maps是一个小型、可扩展、交互式的免费库

B.Leaflet是一个小型化的地图框架，通过小型化和轻量化来满足移动网页的需要

C.Google Fusion Tables让一般使用者也可以轻松制作出专业的统计地图

D.大数据魔镜是一款优秀的国产数据分析软件，它丰富的数据公式和算法可以让用户真正理解探索分析数据

### 第十五、十六、十七章

#### 单选题

1. 下列说法错误的是？C
2. UserCF算法推荐的是那些和目标用户有共同兴趣爱好的其他用户所喜欢的物品
3. ItemCF算法推荐的是那些和目标用户之前喜欢的物品类似的其他物品
4. UserCF算法的推荐更偏向个性化
5. UserCF随着用户数目的增大，用户相似度计算复杂度越来越高
6. 下面关于UserCF算法和ItemCF算法的对比，哪个是错误的？C
7. UserCF算法的推荐更偏向社会化：适合应用于新闻推荐、微博话题推荐等应用场景，其推荐结果在新颖性方面有一定的优势
8. ItemCF算法的推荐更偏向于个性化
9. ItemCF随着用户数目的增大，用户相似度计算复杂度越来越高
10. UserCF推荐结果相关性较弱，难以对推荐结果作出解释，容易受大众影响而推荐热门物品
11. 下列关于推荐系统的描述，哪一项是错误的？D
12. 推荐系统是大数据在互联网领域的典型应用
13. 推荐系统是自动联系用户和物品的一种工具
14. 推荐系统可以创造全新的商业和经济模式，帮助实现长尾商品的销售
15. 推荐系统分为基于物品的协同过滤和基于商家的协同过滤
16. 下列哪一项不属于推荐算法？A
17. 基于物品和商家的联合协同推荐
18. 基于统计的推荐
19. 专家推荐
20. 基于内容的推荐
21. 下列描述有误的是？B
22. 专家推荐：人工推荐，由资深的专业人士来进行物品的筛选和推荐，需要较多的人力成本性
23. 基于统计的推荐：通过机器学习的方法去描述内容的特征，并基于内容的特征来发现与之相似的内容
24. 协同过滤推荐：应用最早和最为成功的推荐方法之一
25. 混合推荐：结合多种推荐算法来提升推荐效果
26. 下列哪一项不属于完整推荐系统的三个模块之一？D
27. 用户建模模块
28. 推荐对象建模模块
29. 推荐算法模块
30. 数据采集模块
31. 下列关于协同过滤的说法，哪一项是错误的？C
32. 协同过滤可分为基于用户的协同过滤和基于物品的协同过滤
33. UserCF算法符合人们对于“趣味相投”的认知，即兴趣相似的用户往往有相同的物品喜好
34. 实现UserCF算法的关键步骤是计算物品与物品之间的相似度
35. 基于物品的协同过滤算法（简称ItemCF算法）是目前业界应用最多的算法
36. 下列哪个说法是错误的？D
37. 无论是亚马逊还是Netflix，其推荐系统的基础都是ItemCF算法
38. ItemCF算法是给目标用户推荐那些和他们之前喜欢的物品相似的物品
39. ItemCF算法主要通过分析用户的行为记录来计算物品之间的相似度
40. 实现UserCF算法的关键步骤是计算物品与物品之间的相似度
41. 下列哪一项说法是错误的？C
42. UserCF算法推荐的是那些和目标用户有共同兴趣爱好的其他用户所喜欢的物品
43. ItemCF算法推荐的是那些和目标用户之前喜欢的物品类似的其他物品
44. ItemCF算法的推荐更偏向社会化，而UserCF算法的推荐更偏向于个性化
45. ItemCF算法倾向于推荐与用户已购买商品相似的商品，往往会出现多样性不足、推荐新颖度较低的问题

#### 多选题

1. 下列哪些属于大数据应用？ABCD
2. 推荐系统：为用户推荐相关商品
3. 物流：基于大数据和物联网技术的智能物流
4. 智能交通：利用交通大数据，实现交通实时监控
5. 汽车：无人驾驶汽车，实时采集车辆各种行驶数据和周围环境
6. 下列关于推荐系统集群的描述，哪些是正确的？ABC
7. 为了让用户从海量信息中高效地获得自己所需的信息，推荐系统应运而生
8. 推荐系统是大数据在互联网领域的典型应用
9. 推荐系统是自动联系用户和物品的一种工具
10. 推荐系统是利用大数据调整线下门店布局、控制店内人流量
11. 下列关于推荐长尾理论的描述，哪些是正确的？ABCD
12. 推荐系统可以创造全新的商业和经济模式，帮助实现长尾商品的销售
13. “长尾”概念于2004年提出，用来描述以亚马逊为代表的电子商务网站的商业和经济模式
14. 可以通过发掘长尾商品并推荐给感兴趣的用户来提高销售额
15. 热门推荐的主要缺陷在于推荐的范围有限，所推荐的内容在一定时期内也相对固定。无法实现长尾商品的推荐
16. 推荐系统的本质是建立用户与物品的联系，根据推荐算法的不同，推荐方法包括以下哪几类？ABCD
17. 专家推荐：人工推荐，由资深的专业人士来进行物品的筛选和推荐，需要较多的人力成本
18. 基于统计的推荐：基于统计信息的推荐（如热门推荐），易于实现，但对用户个性化偏好的描述能力较弱
19. 基于内容的推荐：通过机器学习的方法去描述内容的特征，并基于内容的特征来发现与之相似的内容
20. 混合推荐：结合多种推荐算法来提升推荐效果
21. 一个完整的推荐系统通常包括以下哪三个模块？ACD
22. 用户建模模块
23. 数据采集模块
24. 推荐对象建模模块
25. 推荐算法模块
26. 下列关于推荐系统的描述，哪些是正确的？ABCD
27. 用户建模模块：对用户进行建模，根据用户行为数据和用户属性数据来分析用户的兴趣和需求
28. 推荐对象建模模块：根据对象数据对推荐对象进行建模
29. 推荐算法模块：基于用户特征和物品特征，采用推荐算法计算得到用户可能感兴趣的对象
30. 推荐算法模块：根据推荐场景对推荐结果进行一定调整，将推荐结果最终展示给用户
31. 协同过滤可分为哪几种过滤方式？AB
32. 基于用户的协同过滤
33. 基于物品的协同过滤
34. 基于用户和物品的联合协同过滤
35. 基于商家的协同过滤
36. 关于基于用户的协同过滤，下列哪些说法是正确的？ABCD
37. 基于用户的协同过滤算法（简称UserCF算法）在1992年被提出，是推荐系统中最古老的算法
38. UserCF算法符合人们对于“趣味相投”的认知
39. 实现UserCF算法的关键步骤是计算用户与用户之间的兴趣相似度
40. UserCF算法符合兴趣相似的用户往往有相同的物品喜好
41. 实现UserCF算法的关键步骤是计算用户与用户之间的兴趣相似度，下列哪些是属于计算相似度的算法？ABC
42. 泊松相关系数
43. 余弦相似度
44. 调整余弦相似度
45. 调整正弦相似度
46. 下列关于协同过滤算法的描述，哪些是正确的？ABCD
47. 基于物品的协同过滤算法（简称ItemCF算法）是目前业界应用最多的算法
48. ItemCF算法是给目标用户推荐那些和他们之前喜欢的物品相似的物品
49. ItemCF算法通过建立用户到物品倒排表（每个用户喜欢的物品的列表）来计算物品相似度
50. UserCF算法推荐的是那些和目标用户有共同兴趣爱好的其他用户所喜欢的物品