CDA LEVEL II 大数据分析师模拟题

一、单选题（每题 0.5 分，共计 50 分）

1. 按下（ A ）键能终止当前运行的命令

A. Ctrl-C

B. Ctrl-F

C. Ctrl-B

D. Ctrl-D

2. ls 命令有很多的参数，显示所有文件，包括隐藏文件的参数是（ A ）

A. -a

B. -l

C. –r

D. --help

3. 若要将鼠标从 VM 中释放出来，可按（ A ）键来实现

A. Ctrl + Alt

B. Ctrl +Alt +Del

C. Ctrl +Alt +Enter

D. Ctrl +Enter

4. 除非特别指定，cp 假定要拷贝的文件在下面哪个目录下（ D ）

A. 用户目录

B. home 目录

C. root 目录

D. 当前目录

5. 用 "rm -i",系统会提示什么来让你确认（ B ）

A. 命令行的每个选项

B. 是否真的删除

C. 是否有写的权限

D. 文件的位置

6. 几位学生的某门课成绩分别是 67 分、78 分、88 分、89 分、96 分，则“成绩”

是（ B ）

A. 品质标志

B. 数量标志

C. 标志值

D. 数量指标

7. 抽样调查与重点调查的主要区别是（ D ）

A. 作用不同

B. 组织方式不同

C. 灵活程度不同

D. 选取调查单位的方法不同

8. 先对总体中的个体按主要标志加以分类，再以随机原则从各类中抽取一定的单位进行调查，这种抽样调查形式属于（ D ）

A. 简单随机抽样

B. 等距抽样

C. 整群抽样

D. 类型抽样

9. 统计指标按所反映的数量特征不同可以分为数量指标和质量指标两种。其中数量指标的表现形式是（ A ）

A. 绝对数

B. 相对数

C. 平均数

D. 小数

10. HDfS 中的 block 默认保存几份？（ C ）

A. 1 份

B. 2 份

C. 3 份

D.不确定

11.对某市全部商业企业职工的生活状况进行调查，调查对象是（ B ）

A. 该市全部商业企业

B. 该市全部商业企业的职工

C. 该市每一个商业企业

D. 该市商业企业的每一名职工

12. 在抽样推断中，可计算和控制的误差是（ D ）

A. 登记误差

B. 系统性误差（偏差）

C. 抽样实际误差

D. 抽样平均误差

13. 下面偏度系数的值表明数据分布形态是右偏的是（ A ）

A. 1.429

B. 0

C. -3.412

D. -1

14.（ B ）可以刻画离中趋势。

A. 均值

B. 全距

C. 众数

D. 中位数

15. 正态分布有两个参数μ与σ,（ C ）,分布越集中,正态曲线的形状越“扁平”。

A. μ越大

B. μ越小

C. σ越大

D. σ越小

16. HBase 依赖（ D ）提供强大的计算能力

A. Zookeeper

B. Chubby

C. RPC

D. MapReduce

17. HFile 数据格式中的 Data 字段用于（ A ）

A. 存储实际的 KeyValue 数据

B. 存储数据的起点

C. 指定字段的长度

D. 存储数据块的起点

18. HFile 数据格式中的 Magic 字段用于（ A ）

A. 存储随机数，防止数据损坏

B. 存储数据的起点

C. 存储数据块的起点

D. 指定字段的长度

19. 默认情况下，YARN 支持下面哪个调度器。（ C ）

A. Fair scheduler

B. Capacity scheduler

C. FIFO 调度器

D. DRT 调度器

20. HBase 虚拟分布式模式需要（ A ）个节点？

A. 1

B. 2

C. 3

D.最少 3 个

21. HBase 分布式模式最好需要（ C ）个节点？

A. 1

B. 2

C. 3

D.最少

22. SQL 语句中的条件用以下哪一项来表达（ C ）

A. THEN

B. WHILE

C. WHERE

D. IF

23. 下面哪项工作场景不是 MapReduce 计算框架擅长处理的？（ B ）

A. 分析 web 日志记录，分析用户的行为

B. 实时分析微博热词

C. 分析气象数据，找出有历史记录以来每年的最高气温

D. 购物篮分析，分析用户购买商品的关联度

24. 在 Hadoop 客户端提交数据到 HDFS 上时，数据文件会被分片(split)，最佳的分片策略是（ B ）

A. 等于两个 HDFS 的 block 块大小

B. 等于一个 HDFS 的 block 块大小

C. 等于操作系统的文件块大小

D. 大小无所谓

25. SELECT 语句的完整语法较复杂，但至少包括的部分是（ B ）

A. 仅 SELECT

B. SELECT，FROM

C. SELECT，GROUP

D. SELECT，INTO

26. 子表中一个列族下的所有数据统一存放在一个物理文件中，该物理文件对应于 HRegion 中的一个（ C ）

A. StoreFile

B. HFile

C. Store

D. MemStore

27. 向数据表中更新一条记录用以下哪一项（ D ）

A. CREATE

B. INSERT

C. SAVE

D. UPDATE

28. 关于 Tableau，以下说法错误的是（ D ）：

A. Tableau 支持多种格式数据，包括平面文件(比如 Execl，txt 文本)，或者是数据

库(jdbc，odbc 均可)

B.支持多表关联

C.使用 Tableau 分析数据，比传统的数据库查询快 10 到 100 倍

D. Tableau 不能查询 Hadoop

29. HDFS 的 NameNode 负责管理文件系统的命名空间，将所有的文件和文件夹的元数据保存在一个文件系统树中，这些信息也会在硬盘上保存成以下文件：（C）

A. 日志

B. 命名空间镜像

C. 两者都是

D. 两者都不是

30. 客户端在 HDFS 上进行文件写入时，namenode 根据文件大小和配置情况，返回部分 datanode 信息，然后(　D　)负责将文件划分为多个 Block，根据 DataNode 的地址信息，按顺序写入到每一个 DataNode 块。

A. Namenode

B. Datanode

C. Secondary namenode

D. 客户端

31. HDFS 的是基于流数据模式访问和处理超大文件的需求而开发的，默认的最基本的存储单位是 128M，具有高容错、高可靠性、高可扩展性、高吞吐率等特征，适合的读写任务是（C）

A. 一次写入，少次读写

B. 多次写入，少次读写

C. 一次写入，多次读写

D. 多次写入，多次读写

32. 组合多条 SQL 查询语句形成组合查询的操作符是（ D ）

A. SELECT

B. ALL

C. LINK

D. UNION

33. HBase 是面向（ B ）的数据库。

A. 行

B. 列

C. 网络

D. 内存

34. linux 中，哪个目录存放用户密码信息（ B ）

A. /boot

B. /etc

C. /var

D. /dev

35. 考虑如下场景：在 M/R 系统中，

- HDFS 块的大小是 128MB
- 输入数据格式是 FileInputFormat
- 我们有 2 个文件，大小分别为 64Mb 和 200Mb

问：Hadoop 框架将启动几个 mapper 进程进行处理?（ B ）

A. 2 个

B. 3 个

C. 4 个

D. 5 个

36. 在大数据的单位中，PB 称为”拍字节”，其中 1PB 指的是：（ D ）

A. 1024KB

B. 1024MB

C. 1024GB

D. 1024TB

37. hdfs-site.xml 中哪个主要属性决定数据存储的路径？（ B ）

A. dfs.name.dir

B. dfs.data.dir

C. fs.checkpoint.dir

D. dfs.url

38. 以下哪项用于左连接（ C ）

A. JOIN

B. RIGHT JOIN

C. LEFT JOIN

D. INNER JOIN

39. SQL 是一种（ C ）语言。

A. 函数型

B. 高级算法

C. 关系数据库

D. 人工智能

40. 一张表的主键个数为（ C ）

A. 至多 3 个

B. 没有限制

C. 至多 1 个

D. 至多 2 个

41. 在 Hadoop v2 YARN 中，负责管理一个单独节点内资源的服务是（ A ）

A. NodeManager

B. ResourceManager

C. NameNode

D. DataNode

42. 向数据表中插入一条记录用以下哪一项（ B ）

A. CREATE

B. INSERT

C. SAVE

D. UPDATE

43. Hadoop fs 中的-get 和-put 命令操作对象是：（ C ）

A. 文件

B. 目录

C. 两者都是

D. 两者都不是

44. 创建一个数据库表用以下哪一项（ B ）

A. UPDATE

B. CREATE

C. UPDATED

D. ALTER

45. HDFS 是一个分布式文件系统，它允许用户使用 shell 命令操作文件系统。其中读取/user/hduser/file1.txt 文件并打印到屏幕上的命令是：（ C ）

A. hdfs dfs -ls /user/hduser/file1.txt

B. hdfs dfs -mkdir /user/hduser/file1.txt

C. hdfs dfs -cat /user/hduser/ file1.txt

D. hdfs dfs -put /user/hduser/ file1.txt

46. LSM 更能保证哪种操作的性能？（ B ）

A. 读

B. 写

C. 随机读

D. 合并

47. HDFS 文件系统有一个/作为根目录。运行如下哪个命令来列出在 HDFS 中新创建的目录的内容：（ A ）

A. hdfs dfs -ls test

B. hdfs dfs -mkdir test

C. hdfs dfs -cat test

D. hdfs dfs -put test

48. 运行如下的命令，将本地 readme.txt 文件拷贝到 test 目录中：（ D ）

A. hdfs dfs -cat test/readme.txt

B. hdfs dfs -put test/readme.txt

C. hdfs dfs -rm -R test/readme.txt

D. hdfs dfs -copyFromLocal readme.txt test

49. 关于 MapReduce 计算框架，以下说法正确的是：（ A ）

A. MapReduce 是一个离线的批处理计算框架

B. MapReduce 是一个实时的流处理计算框架

C. MapReduce 是一个内存计算框架

D. 以上说法都正确

50. 在 MapReduce 的 Shuffle 阶段，每个 Reducer 使用 HTTP 协议来从 Mapper 节点获取自己的 partition。默认每个 Reducer 使用几个线程来获取 Maper 节点数据?（ C ）

A. 3 个

B. 4 个

C. 5 个

D. 6 个

51. Hadoop MapReduce 应用程序可以运行在 YARN 上，使用一个（ D ）来协调每个 job 以及一系列资源容器(resource container)来运行 Map 和 Reduce 任务。

A. NodeManager

B. ResourceManager

C. JobTracker

D. ApplicationMaster

52. 以下哪个命令可以终止一个用户的所有进程（ D ）

A. skillall

B. skill

C. kill

D. killall

53. 在基本 K 均值算法里，当邻近度函数采用（ A ）的时候，合适的质心是簇中各点的中位数

A. 曼哈顿距离

B. 平方欧几里德距离

C. 余弦距离

D. Bregman 散度

54. 关于 SecondaryNameNode 哪项是正确的？（ C ）

A. 它是 NameNode 的热备

B. 它对内存没有要求

C. 它的目的是帮助 NameNode 合并编辑日志，减少 NameNode 启动时间

D. SecondaryNameNode 应与 NameNode 部署到一个节点

55. 以 HDFS 上 master:9000/ graphdata.txt 中的数据创建图 graphhdfs，其中 graphdata.txt 文本格式：121256　132156。现将 graphhdfs 中每个节点的属性值变为原值的 3 倍，并查看其中的 10 个顶点，则以下选项正确的是（ A ）

A. val temp=graphhdfs.mapVertices((x1,x2)=>x2.toInt*3).vertices.take(10)

B. val temp=graphhdfs.mapEdges((x1,x2)=>x2.toInt*3).vertices.take(10)

C. val temp=graphhdfs.mapVertices(x=>x._2.toInt*3).vertices.take(10)

D. val temp=graphhdfs.map(x=>x._1.toInt*3).vertices.take(10)

56. Mahout 中实现的 ALS-WR 算法计算（ D ）后，就可以进行推荐了

A. 评分矩阵(user X item) A

B. 用户特征矩阵 U

C. 物品特征矩阵 M

D. U 与 M’的乘积 A_k 矩阵

57. HBase 依赖（ A ）提供消息通信机制

A. Zookeeper

B. Chubby

C. RPC

D. Socket

58. Client 端上传文件的时候下列哪项正确（ B ）

A.数据经过 NameNode 传递给 DataNode

B. Client 端将文件切分为 Block，依次上传

C. Client 只上传数据到一台 DataNode，然后由 NameNode 负责 Block 复制工作

D.以上都不对

59. 下面与 Zookeeper 类似的框架是（ D ）

A. Protobuf

B. Java

C. Kafka

D. Chubby

60. 下面与 HDFS 类似的框架是（ C ）

A. NTFS

B. FAT32

C. GFS

D. EXT3

61. 在 Hbase 中删除表 t1 的命令是（ C ）

A. drop table t1

B. truncate t1

C. drop 't1'

D. truncate table t1

62. 使用 Pig 的优势在于（ A ）

A. Pig 可以使用一个类 SQL 的语言，降低了学习成本

B. Pig 的语言编辑器可以把类 SQL 语言转换为一系列经过优化处理的 MapReduce 运算

C. 目前 Pig 主要由开源社区维护

D. Pig 是一种数据流语言

63. Spark 中的 task 分别是以（ B ）方式维护的

A. 进程

B. 线程

C. 流水线

D. 以上都不是

64. MapReduce 中的 task 是以（ A ）方式维护的

A. 进程

B. 线程

C. 流水线

D. 以上都不是

65. 配置 Standalone 模式下的 Spark 集群，Worker 节点需要在 conf 文件夹下的哪个文件中指明（ D ）

A. regionserver

B. spark-env.sh

C. spark-defaults.conf

D. slaves

66. val rdd = sc.parallelize(List(("Tom",2),("Lee",5),("Lee",6),(“Tom”,7)))

rdd.reduceByKey((x,y) => x + y).collect

上述代码的执行结果为（ C ）

A. Array((1,2), (3,10))

B. Array((9, “Tom”), (11, “Lee”))

C. Array((“Tom”,9), (“Lee”,11))

D. Array((“Tom”,2,7), (“Lee”,5,6))

67. val rdd=sc.parallelize(1 to 10).filter(_%2== 0)

rdd.collect

上述代码的执行结果为（ C ）

A. Array(1,2,3,4,5,6,7,8,9,10)

B. Array(1, 3, 5,7,9)

C. Array(2, 4, 6, 8, 10)

D. Array(1,10)

68. 基于密集向量(1.0, 0.0, 3.0)创建一个 LabledPoint，设其标识值为 1.0，以下正确的选项为（ A ）

A.val pos = LabeledPoint(1.0, Vectors.dense(1.0, 0.0, 3.0))

B.val pos = LabeledPoint(1.0, (1.0, 0.0, 3.0))

C.val pos = LabeledPoint(Vectors.dense(1.0, 0.0, 3.0), 1.0)

D.val pos = LabeledPoint((1.0, 0.0, 3.0),1.0)

69. MLlib 中创建稀疏矩阵((0.0, 2.0), (3.0, 0.0), (0.0, 6.0))的语句是（ C ）

A. val dm: Matrix = Matrices.dense(3, 2, Array(0.0, 3.0, 0.0, 2.0, 0.0, 6.0))

B. val dm: Matrix = Matrices.sparse(3, 2, Array(0.0, 2.0, 3.0, 0.0, 0.0, 6.0))

C. val sm: Matrix = Matrices.sparse(3, 2, Array(0, 1, 2), Array(1,0, 1), Array(2, 3, 6))

D. val sm: Matrix = Matrices.dense(3, 2, Array(0, 1, 2), Array(1,0, 1), Array(2, 3, 6))

70. MLlib 提供的分布式矩阵中，不包含行、列索引信息的矩阵类型是（ A ）

A. RowMatrix

B. IndexedRowMatrix

C. Matrix

D. CoordinateMatrix

71. Spark 支持的分布式部署方式中哪个是错误的（ D ）

A standalone

B spark on mesos

C spark on YARN

D Spark on local

72. 下列哪个操作能够实现“基于窗口将 DStream[(K,V)]中的值 V 按键 K 使用聚合函数 func 聚合得到新的 DStream”（ B ）

A. count

B. reduceByKeyAndWidow

C. countByValue

D. reduceByKey

73. 在 Spark Streaming 中批处理时间间隔是指（ A ）

A. 系统将获取到的数据流封装成一个 RDD 的时间间隔

B. 系统对数据流进行统计分析的时间间隔

C. 系统对数据流进行统计分析的频率

D. 系统作业处理的周期

74. DataFrame 和 RDD 最大的区别（ B ）

A. 科学统计支持

B. 多了 schema

C. 存储方式不一样

D. 外部数据源支持

75. 在使用 mkdir 命令创建新的目录时，在其父目录不存在时先创建父目录的选项是（ D ）

A. –m

B. –d

C. –f

D. –p

76. 在 Spark 中，DAG Scheduler 的输出形式为（ C ）

A. DAG 图

B. Stage

C. TaskSet

D. Task

77. Stage 的 Task 的数量由什么决定（ A ）

A. Partition

B. Job

C. Stage

D. TaskScheduler

78. 下面哪个操作是窄依赖（ B ）

A. join

B. filter

C. group

D. sort

79.下面哪个操作肯定是宽依赖（ C ）

A. map

B. flatMap

C. reduceByKey

D. sample

80. hive 的元数据存储在 derby 和 mysql 中有什么区别（ B ）

A. 没区别

B. 多会话

C. 支持网络环境

D. 数据库的区别

81. Spark SQL 组件的主要功能是（ A ）

A. 海量数据的交互式查询

B. 机器学习与数据挖掘

C. 图计算

D. 实时数据流处理

82. Spark Streaming 组件的主要功能是（ D ）

A. 海量数据的交互式查询

B. 机器学习与数据挖掘

C. 图计算

D. 实时数据流处理

83. 与 MapReduce 相比，Spark 更适合处理以下哪种类型的任务（ B ）

A. 较多迭代次数的长任务

B. 较多迭代次数的短任务

C. 较少迭代次数的长任务

D. 较少迭代次数的短任务

84. 对于 SparkStreaming 与 Storm，系列叙述错误的是（ C ）

A. 二者同为大数据流式数据处理框架

B. SparkStreaming 在吞吐量与集成性方面要优于 Storm

C. SparkStreaming 在数据处理的实时性要优于 Storm

D. SparkStreming 又称为准实时处理框架，对数据的处理延迟能够达到秒级别

85. 当 HRegion 中的 StoreFile 数目达到一定阈值时，就会触发 HRegion 的（ A ）

A. compact 操作

B. split 操作

C. flush 操作

D. write 操作

86. spark 的 master 和 worker 通过什么方式进行通信的？（ D ）

A. http

B. nio

C. netty

D. Akka

87. MLlib 提供的分布式矩阵中，既有行索引，又有列索引的是（ D ）

A. RowMatrix

B. IndexedRowMatrix

C. Matrix

D. CoordinateMatrix

88. Standalone 模式下配置 Spark 集群时，master 节点的工作端口号需要在 conf 文件夹下的哪个文件指明（ B ）

A. regionserver

B. spark-env.sh

C. spark-defaults.conf

D. slaves

89. 执行如下哪个命令，用来初始化 name 目录和 data 目录（ B ）

A. hadoop namenode -jar

B. hadoop namenode -format

C. hadoop datanode -jar

D. hadoop datanode -format

90. 以下哪个命令用来启动 HDFS 系统：（ A ）

A. start-dfs.sh

B. stop-dfs.sh

C. sbin/mr-jobhistory-daemon.sh start historyserver

D. jps

91. HDFS 有一个 gzip 文件大小 75MB，客户端设置 Block 大小为 64M。当运行 MapReduce 任务读取该文件时 input split 大小为多少（ B ）

A. 64M

B. 75M

C. 一个 map 读取 64M，另外一个 map 读取 11M

D. 一个 map 读取 11M，另外一个 map 读取 64M

92. Spark Job 默认的调度模式（ A ）

A. FIFO

B. FAIR

C.无

D.运行时指定

93. 以下关于 SPARK 中的 spark context，描述错误的是：（ A ）

A. 控制整个 application 的生命周期

B. 可以控制 dagsheduler 组件

C. 可以控制 task scheduler 组件

D. SparkContext 为 Spark 的主要入口点

94. 以下对 Spark 中 Work 的主要工作描述错误的是（ C ）

A. 管理当前节点内存

B. 不会运行业务逻辑代码

C. 会运行业务逻辑代码

D. 接收 master 分配过来的资源指令

95. SPARK 默认的存储级别是（ A ）

A. MEMORY_ONLY

B. MEMORY_ONLY_SER

C. MEMORY_AND_DISK

D. MEMORY_AND_DISK_SER

96. Mahout 中进行大数据分析时，需要对数据进行聚类，其所使用的命令是（ B ）

A. seq2sparse

B. kmeans

C. trainnb

D. testnb

97. 使用 spark MLib 进行 K-means 算法分析时，我们会调用 KMeans.train 方法对数据集进行聚类训练，该函数的返回值是（ B ）

A. K_MEANS_PARALLEL

B. KMeansModel 类实例

C. kmeans.epsilon

D. kmeans.test 实例

98. 大数据的起源是以下哪个领域（ C ）

A. 金融

B. 电信

C. 互联网

D. 公共管理

99. 如果你面对的大数据都是结构化的数据，使用传统的数据库进行数据库查询和分析时，数据库的反应速度很慢，在这种大数据应用场景下，正确的大数据技术解决方案是（ B ）

A. Hadoop + oracle+ spark

B. Hadoop + sqoop + hive + spark

C. oracle + mahout

D. sql server + oracle + mahout

100. 假设需要对某个数据集使用 mahout 进行聚类，数据集共有 6 类数据，需要迭代 7 次，拟使用 mahout 进行聚类，假设输入文件为 input/part-m-0000，输出目录为 output，初始聚类中心点文件路径为 clusters，下列聚类语句正确的是（ A ）

A. mahout kmenas -i input/part-m-0000 -o output -c clusters - k 6 -x 7

B. mahout kmenas -i input/part-m-0000 -o output -c clusters - k 7 -x 6

C. mahout kmenas -i input/part-m-0000 -o output -c clusters -x 6

D. mahout kmenas -i input/part-m-0000 -o output -c clusters - k 7

二、多选题（每题 1 分，共计 50 分）

1. 在下列分区中，哪些不是 Linux 默认的分区（ ACD ）

A. FAT32

B. EXT3

C. FAT

D. NTFS

2. 总体与样本的关系是（ AB ）

A. 样本来自于总体

B. 以样本推断总体

C. 两者可以互换角色

D. 以总体指标估计样本指标

3. 下面哪些是影响必要样本容量的因素？（ ABCD ）

A. 总体各单位标志变异程度

B. 允许的极限误差大小

C. 推断的可靠程度

D. 抽样方法和抽样组织方式

4. 下面对 HBase 的描述哪些是正确的？（ BCD ）

A. 不是开源的

B. 是面向列的

C. 是分布式的

D. 是一种 NoSQL 数据库

5. MapReduce 与 HBase 的关系，哪些描述是错误的？（ AD ）

A. 两者不可或缺，MapReduce 是 HBase 可以正常运行的保证

B. 两者不是强关联关系，没有 MapReduce，HBase 可以正常运行

C. MapReduce 可以直接访问 HBase

D. 它们之间没有任何关系

6. spark 的四大组件包括下面哪几个?（ ABC ）

A. Spark Streaming

B. Mlib

C. Graphx

D. Spark R

7. 下面哪些概念是 HBase 框架中使用的？（ AC ）

A. HDFS

B. GridFS

C. Zookeeper

D. EXT3

8. 下面对 LSM 结构描述正确的是？（ AC ）

A. 顺序存储

B. 直接写硬盘

C. 需要将数据 Flush 到磁盘

D. 是一种搜索平衡树

9. HFile 数据格式中的 KeyValue 数据格式，下列选项描述正确的是？（ AD ）

A. 是 byte[]数组

B. 没有固定的结构

C. 数据的大小是定长的

D. 有固定的结构

10.下面哪些是连续型数量标志（ ABD ）

A. 住房面积

B. 商店的商品销售额

C. 高校的院系

D. 人口的出生率

11. 下列关于舍恩伯格对大数据特点的说法中，正确的是？（ ABC ）

A. 数据规模大

B. 数据类型多样

C. 数据处理速度快

D. 数据价值密度高

12. spark 的有几种部署模式（ ABCD ）

A. 本地模式

B. standalone 模式

C. spark on yarn 模式

D. mesos 模式

13. Namenode 在启动时自动进入安全模式，在安全模式阶段，说法正确的是：（ ABC ）

A. 安全模式目的是在系统启动时检查各个 DataNode 上数据块的有效性

B. 根据策略对数据块进行必要的复制或删除

C. 当数据块最小百分比数满足的最小副本数条件时，会自动退出安全模式

D. 文件系统允许有修改

14. 下列关于聚类挖掘技术的说法中，正确的是？（ ACD ）

A. 不预先设定数据归类类目，完全根据数据本身性质将数据聚合成不同类别

B. 要求同类数据的内容相似度尽可能小

C. 要求不同类数据的内容相似度尽可能小

D. 与分类挖掘技术相似的是，都是要对数据进行分类处理

15. 我们可以通过下面哪几个配置文件来控制 Hadoop 配置。在集群重启以后，Hadoop 会从这些配置文件中重新加载配置。（ ABCD ）

A. core-site.xml

B. hdfs-site.xml

C. mapred-site.xml

D. yarn-site.xml

16. 在 MRv2 中，Container 是一个动态资源分配单位，将相关的资源封装在一起，包括（ ABC ），从而限定每个任务的资源量：

A. 内存

B. 磁盘

C. CPU

D. IP 地址

17. 下面哪个是 RDD 的特点（ ABD ）

A. 可分区

B. 可序列化

C. 可修改

D. 可持久化

18. 以下哪些项是 HDFS(Hadoop 分布式文件系统)设计的前提和目标？（ABCD）

A. 大数据

B. 硬件错误是常态

C. 流式数据访问

D. 简单一致性

19. 在 HDFS 中，NameNode 是用来管理文件系统的命名空间的。它将所有的文件和文件夹的元数据保存在一个文件系统树中。这些信息也会在硬盘上保存成以下文件：（ AB ）

A. 命名空间镜像

B. 修改日志

C. 数据块 block

D. 分片文件

20. cache 和 pesist 的描述，正确的是（ ABD ）

A. cache 和 persist 都是用于将一个 RDD 进行缓存的，这样在之后使用的过程中就不需要重新计算了，可以大大节省程序运行时间

B. cache 只有一个默认的缓存级别 MEMORY_ONLY

C. persist 可调用 cache，而 cache 可以根据情况设置缓存级别

D. executor 执行的时候，默认 60%做 cache，40%做 task 操作，persist 最根本的函数，最底层的函数

21. 如果要将一个本地用户主目录下的数据文件 a.data，上传到 HDFS 文件系统的/test-data/目录下，执行以下哪些命令可以实现？（ ABD ）

A. hadoop fs -copyFromLocal ~/a.data /test-data/

B. hadoop fs -put ~/a.data /test-data/

C. hdfs fs -copyFromLocal ~/a.data /test-data/

D. hdfs dfs -copyFromLocal ~/a.data /test-data/

22. MapReduce 框架提供了一种序列化键/值对的方法，支持这种序列化的类能够在 Map 和 Reduce 过程中充当键或值，以下说法正确的是：（ ABD ）

A. 实现 Writable 接口的类是值

B. 实现 WritableComparable<T>接口的类可以是值或键

C. Hadoop 的基本类型 Text 并不实现 WritableComparable<T>接口

D. 键和值的数据类型可以超出 Hadoop 自身支持的基本类型

23. Spark 是大数据的综合处理框架，其综合性体现为（ ABC ）

A. 能够对海量数据进行批处理

B. 能够对海量数据进行流式计算

C. 能够对海量数据进行交互式查询

D. 能够对海量数据进行存储

24. 以下对 Spark 中 RDD 叙述错误的是（ AC ）

A. RDD 是可读、写的

B. RDD 是基于内存的高度首先的数据共享模型

C. RDD 是基于磁盘的高度首先的数据共享模型

D. RDD 之间的依赖关系分为宽依赖与窄依赖

25. 对 HBase 构建二级索引的实现方式有哪些？（ AB ）

A. MapReduce

B. Coprocessor

C. Bloom Filter

D. Filter

26. 在 SparkStreaming 中，以下哪些数据可以作为 DStream 的输入源（ ABCD ）

A. socketSteam

B. kafkaSteam

C. flumeSteam

D. twitterSteam

27. 令 ds 为 SparkStreaming 中 DStream 的一个实例，下列叙述正确的是（ ACD ）

A. ds 上的操作都作用于其中的每个 RDD 上

B. ds.count 结果返回一个 RDD

C. ds.reduceByKey 结果返回一个 DStream 类型实例

D. ds 中的每个 RDD 是一个批处理时间间隔内 SparkStreaming 获取的实时数据.

28. 对 GraphX 以下描述正确的是（ ABCD ）

A. GraphX 是一种基于内存的分布式的图计算框架与图计算库

B. GraphX 中引入了弹性分布式属性图

C. GraphX 实现了表视图与图视图的统一

D. GraphX 提供了丰富的 Pregel API 用以实现经典的图计算算法

29. 以下对于 GraphX 中 triangleCount()的描述错误的是（ ACD ）

A. 用以实现三角形计数功能

B. 返回的数据是顶点集合

C. 要求边是规范的指向（srcId <dstId）

D. 返回的数据是图

30. 对 MLlib 的特点描述正确的是（ ABC ）

A. 运算速度快，适用于具有较多迭代次数的算法

B. 具有易用性，RDD 中封装了大量的操作，提供了经典机器学习算法的 API

C. 集成度高，能够与 Spark 上的其他组件进行无缝对接

D. 运行原理是将 Spark 程序转换为 MapReduce 程序运行，并行度高

31. 对于 MLlib 中向量与 LabledPoint，以下描述正确的是（ ACD ）

A. LabledPoint 是一种基于向量扩展得到的数据结构

B. 向量既可以是本地的也可以是分布式的

C. MLlib 中既可以定义稀疏向量也可以定义密集向量

D. 在 LabledPoint 中除了包含一个向量成员外，还包含一个 Double 类型的标识成员

32. 以下属于 MLlib 中能够实的接口有（ ABCD ）

A. KMeans

B. SVMWithSGD

C. ALS

D. LinearRegressionWithSGD

33. SparkStreaming 可以对多种数据源（ ABCD ）进行类似 Map、Reduce 和 Join 等复杂操作。

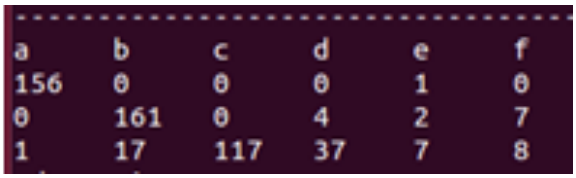
A. Kdfka

B. Flume

C. Twitter

D. Zero

34. 20Newsgroups 数据集是机器学习研究中常用的标准数据集，它使用 20 个 Usenet 新闻单位上几个月发布的 18828 个消息，共 18828 个文件，如果对该数据集使用 mahout 进行文本分类，分类后得到的混淆矩阵中，部分结果如下图所示：



a	b	c	d	e	f
156	0	0	0	1	0
0	161	0	4	2	7
1	17	117	37	7	8

图中第一行是类别名称，第二行是属于 a 类的分类情况（a 类文本原有 168 篇），第三行是属于 b 类的分类情况（b 类文本原有 180 篇），第四行是 c 类的分类情况（c 类文本原有 189 篇），根据各行的分类情况，以下分析正确的是（ ACD ）

A. 分类算法对 a 类文本分类情况较好

B. 分类算法对 c 类文本分类情况较好

C. 分类算法对 b 类文本分类情况较好

D. 分类算法对 c 类文本分类情况较差

35. 数据清洗的方法包括（ ABC ）

A. 缺失值处理

B. 噪声数据清除

C. 一致性检查

D. 重复数据记录处理

36. 下列说法正确的是（ AB ）

A. 在 MySQL 中，不允许有空表存在，即一张数据表中不允许没有字段。

B. 在 MySQL 中，对于存放在服务器上的数据库，用户可以通过任何客户端进行访问。

C. 数据表的结构中包含字段名、类型、长度、记录。

D. 字符型数据其常量标志是单引号和双引号，且两种符号可以混用。

37. 下列关于大数据的分析理念的说法中，正确的是（ ABC ）

A. 在数据基础上倾向于全体数据而不是抽样数据

B. 在分析方法上更注重相关分析而不是因果分析

C. 在分析效果上更追究效率而不是绝对精确

D. 在数据规模上强调相对数据而不是绝对数据

38. 下列哪些命令是 Mahout 中用于实现贝叶斯文本分类算法（ ABC ）

A. seqdirectory

B. seq2sparse

C. trainnb

D. trainlogistic

39. 按照远近程度来聚类需要明确两个距离（ AB ）

A. 点和点之间的距离

B. 类和类之间的距离

C. 欧式距离

D. 兰氏距离

40. Mahout 实现的聚类算法（ ABCD ）

A. K-means

B. Canopy

C.模糊 K-Means 聚类

D.狄利克雷聚类

41. 使用 Hbase 的优势在于（ ABCD ）

A. 相对 Hive，Hbase 支持随机查询

B. 使用 HDFS 文件系统，让 Hbase 存储的扩展几乎随着节点数的增加线性扩展

C. Hbase 能够使用分布式计算，短时间内完成 TB、PB 级的数据搜索

D. Hbase 数据库数据块大小和 HDFS 数据库块大小一致更好

42. 关于 Hadoop 单机模式和伪分布式模式的说法，错误的是：（ ABC ）

A. 两者都起守护进程，且守护进程运行在一台机器上

B. 单机模式不使用 HDFS，但加载守护进程

C. 两者都不与守护进程交互，避免复杂性

D. 后者比前者增加了 HDFS 输入输出以及可检查内存使用情况

43. 按照涉及自变量的多少，可以将回归分析分为（ CD ）

A. 线性回归分析

B. 非线性回归分析

C. 一元回归分析

D. 多元回归分析

44. 以下适用 HDFS 的场景有：（ AC ）

A. 超大文件处理

B. 低延时的数据访问

C. 使用廉价商用硬件

D. 多用户写入，随机修改文件

45. 下列对 Sqoop 描述正确的是（ ABCD ）

A. Sqoop 可以将数据从 MySQL 转储到 HDFS 上

B. Sqoop 可以数据从 HDFS 转储到 MySQL 上

C. Sqoop 可以将数据从 Hbase 转储到 HDFS 上

D. Sqoop 可以数据从 HDFS 转储到 Hbase 上

46. 基于内容的推荐算法生成推荐的过程主要依靠（ ACD ）

A. 内容分析器

B. 推荐系统

C. 文件学习器

D. 过滤部件

47. Mahout 中实现的 kmeans 聚类命令的必选参数是（ ABC ）

A. –input 偏好数据路径

B. –output 推荐结果路径

C. –clusters 初始聚类中心点文件路径

D. --overwrite 对输出路径进行重写

48. Mahout 中实现的 canopy 聚类命令的可选参数是（ CD ）

A. –input 偏好数据路径

B. –output 推荐结果路径

C. --t1 (-t1) t1：T1 阈值

D. --t2 (-t2) t2：T2 阈值

49. 给定一个巨大的文本（如 1TB），可以编写 mapreduce 程序计算单词出现的

数目，需要经历如下哪几个步骤（ ABCD ）

A. 自动对文本进行分割

B. 对分割后的每一个 KV 对应用用户定义的 Map 进行处理，生成新的 KV 对

C. 对输出的结果集归拢、排序（系统自动完成）

D. 通过 Reduce 操作生成最后结果

50. 20Newsgroups 数据集是机器学习研究中常用的标准数据集，它使用 20 个 Usenet 新闻单位上几个月发布的 18828 个消息，共 18828 个文件，如果要对该数据集使用 mahout 进行文本分类，错误的做法是（ ABC ）

A. 直接使用 mahout 算法，在 namenode 机器的本地文件系统中调用这 18828 个文件

B. 将这 18828 个文件上传到 hdfs 上，然后使用 mahout 算法分析

C. 使用 mahout 提供的 seqdirectory 命令将 18828 个文件序列化成一个大文件上传到 hdfs 上，然后使用 mahout 算法分析

D. 将这 18828 个文件 rar 压缩软件将其压缩成一个大文件上传到 hdfs 上，然后使用 mahout 算法分析