一、选择题

1. 下列选项中，哪个是Scala编译后文件的扩展名？( A ) (易)  
 A. .class  
 B. .bash  
 C. .pyc  
 D. .sc  
 2. 下列方法中，哪个方法可以正确计算数组arr的长度？( D )(易)  
 A. count()  
 B. take()  
 C. tail()  
 D. length()  
 3. 下列关于List的定义，哪个是错误的？( D )(易)  
 A. val list = List(l,22,3)  
 B. val list = List("Hello”,"Scala”)  
 C. val list : String = List("A”,"B")  
 D. val list = List[Int](1,2,3)

4.下列选项中，哪个不是Spark生态系统中的组件？( D )(易)

A. Spark Streaming  
 B.  Mlib  
 C. Graphx  
 D. Spark R

5. 下面哪个端口不是Spark自带服务的端口？( C )(中)  
 A. 8080  
 B. 4040  
 C. 8090  
 D. 18080

6. 下列选项中，针对Spark运行的基本流程哪个说法是错误的？( B )(中)  
 A. Driver端提交任务，向Master申请资源  
 B. Master 与Worker进行TCP通信,使得Worker启动Executor  
 C. Executor启动会主动连接Driver,通过 Driver->Master-> WorkExecutor，从而  
得到Driver在哪里  
 D. Driver会产生Task，提交给Executor中启动Task去做真正的计算

7. 下列方法中，用于创建RDD的方法是。( C )(中)  
 A. makeRDD()  
 B. parallelize()  
 C. textFile()  
 D. testFile()

8. 下列选项中，哪个不属于转换算子操作？( C )(中)  
 A. filter(func)  
 B. map(func)  
 C. reduce(func)  
 D. reduceByKey(func)

9. 下列选项中，能使RDD产生宽依赖的是。( D )(中)   
 A. map(func)  
 B. filter(func)  
 C. union  
 D. groupByKey()

10. Spark SQL可以处理的数据源包括哪些？( D )(中)

A. Hive表  
 B.数据文件、Hive表  
 C.数据文件、Hive表、RDD  
 D.数据文件、Hive表、RDD、外部数据库

11. 下列说法正确的是哪一项？( C )(中)  
 A. Spark SQL的前身是Hive  
 B. DataFrame其实就是RDD  
 C. HiveContext继承了SqlContext  
 D. HiveContext只支持SQL语法解析器

12. Spark SQL中,mode函数可以接收的参数有哪些？( A )(中)  
 A. Overwrite、Append、Ignore、ErrorIfExists  
 B. Overwrite、Ignore  
 C. Overwrite、Append、Ignore  
 D. Append、Ignore、ErrorIfExists  
 13. 下列选项中，哪个不属于HBase的特点？( B )(易)  
 A. 面向列  
 B. 容量小  
 C. 多版本  
 D. 扩展性

14. 下列选项中，HBase是将哪个作为其文件存储系统的？( C )(易)  
 A. MySQL  
 B.  GFS  
 C. HDFS

D. MongoDB

15.  HBase官方版本不可以安装在什么操作系统上？( D )(中)  
 A. CentOS  
 B. Ubuntu  
 C. RedHat  
 D. Windows

16. 下列选项中，哪个不是Kafka的优点？( C )(易)

A. 解耦  
 B. 高吞吐量  
 C. 高延迟  
 D. 容错性

17. 下列选项中，哪个选项是每个分区消息的唯一序列标识？( D )(难)  
 A. Topic  
 B. Partition  
 C. Broker  
 D. Offset

18. 下列选项中，哪个不属于消息系统？( D )(较难)

A. Kafka  
 B. RabbitMQ  
 C. ActiveMQ  
 D. Zookeeper  
 19. 下列洗项中,说法正确的是哪个？( B )(较难)

A.窗口滑动时间间隔必须是批处理时间间隔的倍数  
 B. Kafka是Spark Streaming的基础数据源  
 C.DStream不可以通过外部数据源获取  
 D. reduce(func)是DStream的输出操作

20. 关于Spark Streaming,下列说法错误的是哪一项？( D )(较难) A. Spark Streaming是Spark的核心子框架之  
 B. Spark Streaming具有可伸缩、高吞吐量、容错能力强等特点  
 C. Spark Streaming处理的数据源可以来自Kafka  
 D. Spark Streaming不能和Spark SQL.Mllib.GraphX无缝集成

21. DStream的转换操作方法中,哪个方法可以直接调用RDD上的操作方法？( A )(较难)  
 A. transform(func)  
 B. updateStateByKey(func)  
 C. countByKey()  
 D. cogroup(otherStream,[numTasks])

22. 下列选项中,对于机器学习的理解错误的是哪一项？( C )(较难)  
 A. 机器学习是一种让计算机利用数据来进行各种工作的方法  
 B. 机器学习是研究如何使用机器人来模拟人类学习活动的一门学科  
 C. 机器学习是一种使用计算机指令来进行各种工作的方法  
 D. 机器学习就是让机器能像人一样有学习、理解、认识的能力  
 23. 下列选项中,哪一项是不属于监督学习的方法？( A )(中)

A. KMeans  
 B. 线性回归  
 C. SVM

D. 朴素贝叶斯  
 24. 下列选项中,哪一项是最常见的评价分类器好坏的指标。( A )(中)  
 A. 准确率(auc)  
 B. 精确度(precision)  
 C. 召回率(recall)  
 D. F值  
 25. 持久化RDD的存储级别中，将RDD分区全部存储到磁盘上的存储级别是( C )。

(较难)

A、MEMORY\_ONLY

B、MEMORY\_AND\_DISK

C、DISK\_ONLY

D、MEMORY\_ONLY\_SER

26、当MemStore存储的数据达到⼀个阀值（B）时，数据就会被进行flush操作，将数据写⼊到StoreFile⽂件。 (中)

A、64M

B、128M

C、256M

D、512M

27、下列选项中，用于提交和监控Task任务的是（B）。 (较难)

A、DAG Scheduler

B、Task Scheduler

C、Cluster Manager

D、SparkContext

28、下列选项中，用于获取指定行或列的数据的HBase Shell命令是（D）。 （中）

A、create

B、describe

C、count

D、get

29、spark 1.4版本的最大变化是（B）。 （中）

A、spark sql Release 版本

B、引⼊ Spark R

C、DataFrame

D、支持动态资源分配

30、下列选项，不属于集群管理器的有（C）。 （中）

A、Hadoop Yarn

B、Apache Mesos

C、Local

D、Spark自带的独立调度器

31、SaveMode属于（C）类型。 （中）

A、整数类型

B、浮点类型

C、枚举类型

D、字符串类型

32、Spark集群的任务是由（A）进行调度的。 （难）

A、驱动器

B、执行器

C、集群管理器

D、应用管理器

33、在RDD的转换算子中，主要用于（Key，Value）键值对的数据集，将具有相同Key的Value进行分组，会返回⼀个新的（Key，Iterable）形式的数据集的转换算子是（B）。 （中）

A、filter()

B、groupByKey()

C、reduceByKey()

D、map()

34、Spark计算框架在分布式环境下对数据处理后的结果进行随机的、实时的存储归功于（D）。 （中）

A、Hive

B、Oracle

C、Mongodb

D、HBase

35、在RDD的行动算子中，用于用于返回数组的第⼀个元素的行动算子是（）。在RDD的行动算子中，用于用于返回数组的第⼀个元素的行动算子是（A）。 （中）

A、first()

B、count()

C、take(n)

D、reduce()

36、在Spark2.0版本之前，Spark SQL中创建DataFrame和执行SQL的⼊口是（C）。 （中）

A、HiveContext

B、SparkSession

C、SQLContext

D、SparkContext

37、在DataFrame的操作中，用于实现对列名进行重命名的操作是（A）。（中）

A、select()

B、show()

C、filter()

D、map()

38、Scala是⼀种纯粹的面向对象语言，每⼀个值都是（D）。 （中）

A、类

B、接口

C、特质

D、对象

39、在Spark运行过程中，每个Job可以划分为更⼩的Task集合，每组任务被称为（D）。 （中）

A、DAG

B、Block

C、Application

D、Stage

40、Spark中RDD的计算函数的基本单位是（A）。 （中）

A、分区

B、数据块

C、Task

D、Job

41、RDD在Spark中运行时，将作业的DAG划分成不同的Stage的阶段是（B）。 （中）

A、RDD Objects

B、DAGScheduler

C、TaskScheduler

D、Worker

42、DataFrame的结构类似于传统数据库的（B）。 （易）

A、⼀维表格

B、二维表格

C、三维表格

D、四维表格

43、下列选项中，哪个不是Kafka的优点。(C) (易)

A、解耦

B、高吞吐量

C、高延迟

D、容错性

44、启动HBase集群的命令是（C）。 (易)

A、start-dfs.sh

B、zkServer.sh start

C、start-hbase.sh

D、start-yarn.sh

45、Task是运行（C）中Executor的工作单元。 (中)

A、Driver program

B、spark master

C、worker node

D、Cluster manager

46、Scala中，数组的遍历方式不包含（D）。 (中)

A、for循环遍历

B、while循环遍历

C、do...while循环遍历

D、do...for循环遍历

47、下列选项中，对于机器学习的理解错误的是（C）。 (较难)

A、机器学习是⼀种让计算机利用数据来进行各种工作的方法。

B、机器学习是研究如何使用机器⼈来模拟⼈类学习活动的⼀门学科。

C、机器学习是⼀种使用计算机指令来进行各种工作的方法。

D、机器学习就是让机器能像⼈⼀样的有学习、理解、认识的能力。

48、下列说法中，关于Scala说法正确的是（）。下列说法中，关于Scala说法正确的是（A）。 (中)

A、Scala是由Martin Odersky（马丁•奥德斯基）创建

B、Scala语言和Java语言的作用是相同的

C、Scala是面向过程编程的语言

D、Scala是由杰弗逊创建

49、下列选项中，说法正确的是哪个？ (B)(难)

A、窗口滑动时间间隔必须是批处理时间间隔的倍数。

B、Kafka是Spark Streaming的基础数据源。

C、DStream不可以通过外部数据源获取。

D、reduce(func)是DStream的输出操作。

50、使用Maven Projects工具，双击（D）选项，即可自动将项目打成Jar包。 (中)

A、clean

B、test

C、deploy

D、package

51、Client与HBase进行通信是通过（A）。(中)

A、RPC协议

B、TCP协议

C、HTTP协议

D、UDP协议

52、Spark SQL的前身是（B）。 (中)

A、SQL

B、Shark

C、Spark RDD

D、MapReduce

53、下列选项中，可以用于退出Spark-Shell客户端的命令是（A）。 (中)

A、:quit

B、:wq

C、:q

D、:exit

54、下列选项中，不属于监督学习的方法的是（A）。 (中)

A、KMeans

B、线性回归

C、SVM

D、朴素贝叶斯

55、下列方法中，用于计算数组长度的是（D）。 (中)

A、count()

B、take()

C、tail()

D、length()

56、数据科学家可能会同时使用多个算法（模型）进行预测， 并且最后把这些算法的结果集成起来进行最后的预测（集成学习），以下对集成学习说法正确的是。(B) (中)

A、单个模型之间有高相关性

B、单个模型之间有低相关性

C、在集成学习中使用“平均权重”而不是“投票”会⽐较好

D、单个模型都是用的⼀个算法

57、由于Scala是运行在JVM平台上的，所以安装Scala之前必须配置好JDK环境，其中JDK版本要求（B）。 (中)

A、不高于1.5版本

B、不低于1.5版本

C、等于1.5版本

D、任何版本均可

58、在Maven工程的pom.xml⽂件中，用于设置所需依赖的版本号的标签是（C）。(难)

A、<dependency>

B、<groupId>

C、<properties>

D、<artifactId>

59、在HBase表中，列的限定符为（A）。 (难)

A、冒号

B、逗号

C、斜杠

D、下划线

60、下列选项中，不属于Spark的四大组件的是（D）。 (中)

A、Spark Streaming

B、MlibB、Mlib

C、Graphx

D、Spark R

61、Spark于2009年诞生于（A）。 (中)

A、美国加州大学伯克利分校的AMP实验室

B、加利福尼亚大学伯克利分校

C、宾⼣法尼亚大学

D、普林斯顿大学

62、关于Spark Streaming，下列说法错误的是哪⼀项？ (D) (中)

A、Spark Streaming是Spark的核⼼子框架之⼀。

B、Spark Streaming具有可伸缩、高吞吐量、容错能力强等特点。

C、Spark Streaming处理的数据源可以来自Kafka。

D、Spark Streaming不能和Spark SQL、Mllib、GraphX无缝集成。

63、HBase的底层依赖的是（B）。 (中)

A、Hive

B、HDFS

C、Mongodb

D、MySQL

64、DStream的转换操作方法中，哪个方法可以直接调用RDD上的操作方法？ (A) (中)

A、transform(func)

B、updateStateByKey(func)

C、countByKey()

D、cogroup(otherStream，[numTasks])

65、如果⼀个有向图无法从任意顶点出发经过若⼲条边回到该点，则这个图就是（A）。 (中)

A、有向无环

B、无环图

C、有向图

D、无向有环图

66、在Spark RDD中，划分Stage的依据是（B）。 (中)

A、窄依赖

B、宽依赖

C、窄依赖和宽依赖

D、以上选项均错误

67、下列选项中，可以支持Scala和Python编程的交互式解释器是（B）。 (中)

A、HBase-Shell

B、Spark-Shell

C、Hadoop-Shell

D、Hive-Shell

68、在Catalyst优化器中，用于将LogicalPlan转换成PhysicalPlan的组件是（D）。 (难)

A、SqlParse

B、Analyze

C、Optimizer

D、Planner

69、在Spark中，调用RDD的（C）方法，可以将RDD转换为DataFrame对象。 (中)

A、show()

B、map()

C、toDF()

D、printSchema()

70、每个Region存储的数据是有限的，如果当Region增大到⼀个阀值（）时，会被等分切成两个新的Region。 (B)(中)

A、64M

B、128M

C、256M

D、512M

71、在Scala工程中，.idea⽂件夹是用于（C）。 (中)

A、存放该工程的代码

B、存放相关依赖

C、存放该工程的配置信息

D、存放测试代码

72、Spark SQL快速的计算效率得益于（A）。 (中)

A、Catalyst

B、Execution

C、Parser

D、Analyzer

73、下列选项中，不属于RDD特点的是（C）。 (中)

A、可分区

B、可序列化B、可序列化

C、可修改

D、可持久化

74、Scala中的类不需要关键字public声明为公共的，并且⼀个Scala源⽂件中可以拥有（B）类。 (中)

A、单个

B、多个

C、两个

D、三个

75、下列选项中，哪个选项是每个分区消息的唯⼀序列标识。 (D)(中)

A、Topic

B、Partition

C、Broker

D、Offset

76、Spark Job默认的调度模式是（A）。 (中)

A、FIFO

B、FAIR

C、无

D、运行时指定

77、下列选项中，最常见的评价分类器好坏的指标是（A）。 (中)

A、准确率（auc）

B、精确度（precision）

C、召回率（recall）

D、F值

78、以下说法正确的是 : (C)(中)

1. ⼀个机器学习模型，如果有较高准确率，总是说明这个分类器是好的。

2. 如果增加模型复杂度， 那么模型的测试错误率总是会降低。

3. 如果增加模型复杂度， 那么模型的训练错误率总是会降低。

4. 我们不可以使用聚类“类别id”作为⼀个新的特征项， 然后再用监督学习分别进行学习。

A、1

B、2

C、3

D、1 and 3

79、DataFrame和RDD最大的区别是（B）。 (较难)

A、科学统计支持

B、schema

C、存储方式不⼀样

D、外部数据源支持

80、下列说法中，关于Scala数据类型说法错误的是（B）。(中)

A、Scala中的所有值都有⼀个类型

B、AnyVal类型是所有类型的超类型

C、AnyRef表⽰引用类型

D、Nothing所有类型的子类型，也称为底部类型

81、Scala中，用于创建单例对象的关键字是（B）。 （易）

A、new

B、object

C、super

D、def

82、变量选择是用来选择最好的判别器子集，如果要考虑模型效率，我们应该做哪些变量选择的考虑？ (C)（中）

1. 多个变量其实有相同的用处

2. 变量对于模型的解释有多大作用

3. 特征携带的信息

4. 交叉验证

A、1 和 4

B、1, 2 和 3

C、1,3 和 4

D、以上所有

83、Scala提供了强大的模式匹配机制，最常见的模式匹配就是（A）。 （易）

A、match case语句

B、switch case语句

C、if else语句

D、if语句D、if语句

84、在RDD的转换算子中，用于将每个元素传递到函数func中，并将结果返回为⼀个新的数据集的转换算子是（A）。 （中）

A、filter()

B、groupByKey()

C、reduceByKey()

D、map()

85、在Scala中，获取元组中的值是通过（C）来获取的。 （易）

A、get（）方法

B、角标

C、下划线加角标

D、下划线

86、Spark与Hadoop在基于内存的运算中，说法正确的是（B）。 （易）

A、Spark的运算效率是Hadoop的10倍

B、Spark的运算效率是Hadoop的100倍

C、Hadoop的运算效率是Spark的100倍

D、Hadoop的运算效率是Spark的10倍

87、Catalyst优化器内部包含五大组件，不包括（D）。（中）

A、SqlParse和Planner

B、Analyze和CostModel

C、Optimizer

D、Job

88、HBase表的列不是由（D）字段组成。 （中）

A、列族名

B、限定符

C、列名

D、时间戳

89、下列说法中，关于HBase集群说法错误的是（A）。 （中）

A、HBase集群中只有⼀个角色

B、HBase集群中有两个角色

C、HRegionServer是HBase集群的从节点

D、HMaster是HBase的主节点

90、下列选项中，用于Scala语言开发的工具是（D）。 （中）

A、PhpStorm

B、PyCharm

C、Notepad++

D、IDEA

91、如果想要实现自定义分区，就需要定义⼀个类，使得这个自定义的类继承org.apache.spark.Partitioner类，并实现其中的3个方法，这3个方法不包括（A）。（中）

A、RangePartitioner()

B、numPartitions()

C、getPartition()

D、equals()

92、下列说法中，关于Spark SQL说法正确的是（B）。 （中）

A、Spark SQL架构与Hive底层结构完全不同

B、Spark SQL能够通过DataFrame和Dataset操作多种数据源执行SQL查询

C、Spark SQL不可以无缝地将SQL查询与Spark程序进行结合

D、以上说法均正确

93、下列选项中，不属于Spark生态圈的组件有（A）。 （中）

A、Map Reduce

B、Spark Streaming

C、GraphX

D、MLlib

94、Spark SQL作为分布式SQL查询引擎，用户不可以通过（B）方式实现对结构化数据的处理。 （易）

A、SQL

B、Shark

C、DataFrames API

D、Datasets API

95、Spark编程支持的语言不包括（B）。 （中）

A、Java

B、JavaScript

C、R

D、ScalaD、Scala

96、下列选项中，不属于RDD行动算子的有（D）。 （中）

A、foreach(func)

B、collect()

C、take(n)

D、reduceByKey()

97、下列选项中，不属于Scala的数据结构有（D）。 （中）

A、数组

B、元组

C、集合

D、对象

98. 下面哪个操作是窄依赖 （B） （中）

A. join

B. filter

C. group

D. sort

99.下面哪个操作肯定是宽依赖 (C ) （中）

A. map

B. flatMap

C. reduceByKey

D. sample

100. 与hadoop相⽐，哪个不是Spark的主要优点（D） （中）

A. 提供多种数据集操作类型而不仅限于MapReduce

B. 数据集中式计算而更加高效

C. 提供了内存计算，带来了更高的迭代运算效率

D. 可以和Hive和Mysql整合