三、判断题

1、当join算子操作后，分区数量发生变化则为窄依赖。 （易）

参考答案: 错

2、由于Shark不依赖于Hive，因此在版本迭代时很容易添加新的优化策略。 （易）

参考答案: 错

3、创建HBase表需要指定列，因为列是不变的。 （易）

参考答案: 错

4、Scala中声明变量时，可以不给出变量的类型，因为在初始化的时候，Scala的类型推断机制能够根据变量初始化的值自动推算出来。 （易）

参考答案: 对

5、在数据科学应用中，数据工程师可以利用Spark进行数据分析与建模。 （易）

参考答案: 对

6、密集矩阵是将所有元素的值存储在⼀个列优先的双精度数组中，而稀疏矩阵则是将以列优先的非零元素压缩到稀疏列（CSC）格式中。 （易）

参考答案: 对

7、Scala方法的返回值类型必须写，编译器不能自动推断出来，但是对于递归函数来说，必须要指定返回类型。 （易）

参考答案: 错

8、在Scala中，集合有三⼤类分别是List、Set以及Map。 （易）

参考答案: 对

9、DataFrame除了提供比RDD更丰富的算子以外，更重要的特点是提升Spark框架执行效率、减少数据读取时间以及优化执行计划。 （中）

参考答案: 对

10、Scala提供了样例类，它可以对模式匹配进行优化，提高匹配的速率。 （中）

参考答案: 对

11、⼀个进程中只能有⼀个线程在工作，可以处理多个数据分区（例如运行任务、读取或者存储数据）。 （中）

参考答案: 错

12、离线计算和实时计算指的是数据处理的延迟，批量计算和流式计算指的是数据处理的方式。 （中）

参考答案: 对

13、Dataset提供了特定域对象中的强类型集合，也就是在RDD的每行数据中添加了类型约束条件，只有约束条件的数据类型才能正常运行。 （难）

参考答案: 对

14、HBase分布式数据库的数据存储在行列式的表格中，它是⼀个多维度的映射模型。 （难）

参考答案: 对

15、主题的数据分割为⼀个或多个分区，每个分区的数据使用多个segment⽂件存储，分区中的数据是有序的。 （难）

参考答案: 对

16、truncate命令用于删除整个表，数据和结构都删除。 （难）

参考答案: 错

17、⼀个消费者组里它的内部数据是有序的，消费者组与消费者组之间是无序的。（难）

参考答案: 对

18、Spark Streaming中，不可以通过RDD的转换算子生成新的DStream。 （难）

参考答案: 错

19、DAG是⼀种非常重要的图论数据结构。 （中）

参考答案: 对

20、由于HBase是Google Bigtable的开源实现，因此它们是相同的。 （中）

参考答案: 错

21、 在Linux系统下执行“nc –lk 9999”命令启动服务端且监听socket服务。（中）

参考答案: 对

22、 部署Spark高可用集群不需要用到Zookeeper服务。 （中）

参考答案: 错

23、 Task Scheduler将Worker发送给Task进程中的Executor运行并提供应用程序代码。 （中）

参考答案: 错

24、 在数据准备阶段，可直接将采集到的数据进行机器学习算法训练。 （较难）

参考答案: 错

25、 当RDD对象创建后，SparkContext会根据RDD对象构建DAG有向无环图，然后将Task提交给DAGScheduler。 （较难）

参考答案: 对

26、 DataFrame提供了两种语法风格，即DSL风格语法和SQL风格语法。 （中）

参考答案: 对

27、 Apache Spark Streaming是Apache公司非开源的实时计算框架。 （中）

参考答案: 错

28、 Scala有两种类型的变量，⼀种是使用关键字var声明的常量，值是不可变的；另⼀种是使用关键字val声明的变量，值是可变的。 （易）

参考答案: 错

29、 在Yarn模式下，Spark集群中的资源管理器是Master。 （易）

参考答案: 错

30、 HBase利用Chubby作为协调服务作用。 （易）

参考答案: 错

31、 Null是所有引用类型的子类型，主要用途是与其他JVM语言互操作，⼏乎不在Scala代码中使用。 （易）

参考答案: 对

32、 Spark集群可以包含⼀台Master节点和两台Slave节点。 （易）

参考答案: 对

33、 Executor会向SparkContext进行反向注册并申请Task。 （易）

参考答案: 对

34、 Worker Node为执行器的工作单元。 （中）

参考答案: 错

35、 RDD的持久化操作有两种方法，分别是cache()方法和persist()方法。 （中）

参考答案: 对

36、 点对点消息传递模式结构中，生产者发送⼀条数据，消息将持久化到⼀个队列中，数据可以被有序重复处理。 （中）

参考答案: 错

37、 Scala用scalac编译器把源⽂件编译成Java的class⽂件（即在JVM上运行的字节码）。 （中）

参考答案: 对

38、 Spark SQL使用的数据抽象并非是DataFrame，而是RDD。 （中）

参考答案: 错

39、 Apache Flink仅支持实时处理，并且支持高吞吐、高效率的处理能力。 （中）

参考答案: 错

40、 Spark计算框架在处理数据时，所有的中间数据都保存在磁盘中。 （中）

参考答案: 错

41、 对于宽依赖来说，RDD分区的转换处理是在⼀个线程里完成，所以宽依赖会被Spark划分到同⼀个Stage中。 （难）

参考答案: 错

42、 机器学习中的训练和预测过程可以看作⼈类的归纳和推测的过程。 （中）

参考答案: 错

43、 在Mac下，执行“scala -version”命令，可以测试scala安装情况。（中）

参考答案: 对

44、 take()主要用于通过函数func（输⼊两个参数并返回⼀个值）聚合数据集中的元素。 （中）

参考答案: 错

45、 Kafka采用消费者组统⼀了点对点消息传递模式和发布订阅消息传递模式，当Kafka使用点对点模式时，它可以将处理工作任务平均分配给消费组中的消费者成 （中）

参考答案: 对

46、 Partitioner函数不能决定RDD本身的分区数量，也不能决定parent RDD Shuffle输出时的分区数量。 （较难）

参考答案: 错

47、 Spark SQL不仅能够查询MySQL数据库中的数据，还可以向表中插⼊新的数据。 （较难）

参考答案: 对

48、 在启动HBase集群之前，必须要保证集群中各个节点的时间是同步的，若不同步会抛出ClockOutOfSyncException异常，导致从节点无法启动。 （较难）

参考答案: 对

49、 Scala中，定义定长数组需要使用new关键字，而定义变长数组时，则需要导包import scala.collection.mutable.ArrayBuffer。 （较难）

参考答案: 对

50、 无论是在Scala中还是Java中，对象都是类的抽象，而类都是对象的具体实例；对象不占用内存，而类占用存储空间。 （较难）

参考答案: 错

51、 Spark SQL为Spark框架在结构化数据分析方面提供重要技术支持。（中）

参考答案: 对

52、 join算子操作既可以属于窄依赖，也可以属于宽依赖。 （中）

参考答案: 对

53、 DataFrame可以从很多数据源中创建，例如结构化⽂件、外部数据库、Hive表等数据源。 （中）

参考答案: 对

54、在DataFrame的操作中，filter()用于实现对特定字段进行排序操作。 （中）

参考答案: 错

55、 Scala提供了重量级的语法用于定义匿名函数，支持高阶函数，允许嵌套多层函数，并支持柯里化。 （中）

参考答案: 错

56、 如果子类要重写⼀个⽗类中的非抽象方法，则必须使用override关键字，否则会出现语法错误。 （中）

参考答案: 对

57、 Kafka Streams是Apache Kafka开源项⽬的⼀个流处理框架，它是基于Kafka的生产者和消费者，为开发者提供了流式处理的能力，具有低延迟性、高扩展性、 弹性、容错的特点，易于集成到现有的应用程序中。 （中）

参考答案: 对

58、 Spark SQL支持访问Hive数据仓库，并在Spark引擎中进行统计分析。（中）

参考答案: 对

59、 Kafka具有故障容错的特性，每个分区都会以副本的方式复制到其它Broker节点上，副本分区数可以自定义设置。 （中）

参考答案: 对

60、 flatMap(func)与map(func)相似，应用于（Key，Value）键值对的数据集时，返回⼀个新的（Key，Iterable <Value>）形式的数据集。 （中）

参考答案: 错

61、 执行“bin/hbase shell”命令进⼊HBase Shell交互界面。 （易）

参考答案: 对

62、 HBase分布式数据库中的表可以存储成千上万的行和列组成的数据。 （易）

参考答案: 对

63、 Kafka是专门为分布式高吞吐量系统而设计开发的。 （易）

参考答案: 对

64、 HBase中写数据的流程其实就是寻址的流程。 （易）

参考答案: 错

65、 Spark Streaming能够实时接收输⼊的数据流并立即处理，在底层数据会转换成RDD交由Spark引擎处理。 （易）

参考答案: 错

66、 Scala中，使用关键字var或val声明变量时，后面紧跟的变量名称可以和Scala中的保留字重名，而且变量名不可以以字母或下划线开头，且变量名是不区分⼤⼩写的。 （易）

参考答案: 错

67、 Spark SQL与Hive相互不兼容。 （易）

参考答案: 错

68、 Cluster Manager负责整个集群的统⼀资源管理，Worker节点中的Executor是应用执行的主要进程，内部含有多个Task线程以及内存空间。 （易）

参考答案: 对

69、 DStream的内部结构是由⼀系列连续的RDD组成，每个RDD都是⼀⼩段时间分隔开来的数据集。 （中）

参考答案: 对

70、 在HBase集群中，Zookeeper用于监控HRegionServer的状态。 （中）

参考答案: 对

71、二元分类的相关算法包含线性支持向量机、逻辑回归、随机森林、朴素贝叶斯等。 （中）

参考答案: 错

72、 Spark生态系统已经发展成为⼀个可应用于⼤规模数据处理的统⼀分析引擎，它是基于内存计算的⼤数据并行计算框架，适用于各种各样的分布式平台的系统。（中）

参考答案: 对

73、Apache Hive是Spark上的SQL引擎，也是⼤数据系统中重要的数据仓库工具。 （中）

参考答案: 错

答案解析:

74、 由于Column是HBase表的唯⼀标识，因此Column的设计非常重要。 （中）

参考答案: 错

75、 HFile负责记录HBase修改。 （中）

参考答案: 错

76、 对于迭代式计算和交互式数据挖掘，RDD可以将中间计算的数据结果保存在本地磁盘中，若是后面需要中间结果参与计算时，则可以直接从磁盘中读取，从而可以极⼤地提高计算速度。 （中）

参考答案: 错

77、 在HBase集群中，不会出现单点故障的问题。 （中）

参考答案: 错

78、DataFrame可以看作是分布式的Row对象的集合。 （较难）

参考答案: 对

79、Consumer是数据的生产者，Producer是数据的消费者。 （较难）

参考答案: 错

80、 行动算子主要是将在数据集上运行计算后的数值返回到驱动程序，从而触发真正的计算。 （较难）

参考答案: 对

81、 flatMap(func)方法与map(func)方法相似，因此可以任意使用。 （较难）

参考答案: 错

82、 “--master”参数用于指定Master节点的地址。 （较难）

参考答案: 对

83、 Apache Storm可以简单、高效、可靠地实时处理海量数据，处理数据的速度达到毫秒级别，并将处理后的结果数据保存到持久化介质中（如数据库、HDFS） （较难）

参考答案: 对

84、 Spark不支持在各种集群管理器（Cluster Manager）上运行。 （较难）

参考答案: 错

85、 Kafka采用拉取模型，由消费者记录消费状态，根据主题、Zookeeper集群地址和要消费消息的偏移量，每个消费者互相独立地按顺序读取每个分区的消息 （较难）

参考答案: 对

86、 在处理结构化数据时，开发⼈员无需编写MapReduce程序，直接使用SQL命令就能完成更加复杂的数据查询操作。 （较难）

参考答案: 对

87、 通过反射机制推断Schema实现RDD转换DataFrame时，不需要定义⼀个case class样例类。 （中）

参考答案: 错

88、 local[\*]表⽰的是使用⼀个Worker线程本地化运行Spark。 （中）

参考答案: 错

89、 Spark2.0之后，Spark使用全新的SparkContext接⼝替代SQLContext及HiveContext接⼝完成数据的加载、转换、处理等功能。 （中）

参考答案: 错

90、 Application中包含了多个驱动器（Driver）进程和集群上的单个执行器（

Executor）进程。 （中）

参考答案: 错

91、 HBase就可以随着存储数据的不断增加而实时动态的增加列。 （中）

参考答案: 对

92、 在Scala中，在⼀个源⽂件中有⼀个类和⼀个单例对象，若单例对象名与类名相同，则把这个单例对象称作伴生对象（companion object）。 （中）

参考答案: 对

93、 资源⽂件夹中存放测试的源码⽂件，测试⽂件夹中存放开发中项⽬的源码⽂件。（中）

参考答案: 错

94、 Kafka集群部署依赖于Java环境和Zookeeper服务，因此在安装Kafka集群时，需要⾸先配置JDK环境和Zookeeper服务。 （中）

参考答案: 对

95、 本地向量分为密集向量和稀疏向量，密集向量是两个并列的数组（索引、值）支持，而稀疏向量是由Double类型的数组支持。 （中）

参考答案: 错

96、 HBase支持使用SQL语法。 （易）

参考答案: 错

97、 发布订阅模式可以有多种不同的订阅者，发布者发布的消息会被持久化到⼀个主题中，订阅者可以订阅⼀个或多个主题。 （中）

参考答案: 对

98、 标注点是⼀种带有标签的本地向量，通常用于非监督学习算法中。（易）

参考答案: 错

99、 在Kafka中，若想建立生产者和消费者互相通信，就必须提前创建⼀个“公共频道”，它就是我们所说的主题（Topic）。 （易）

参考答案: 对

100、 Spark是专为⼤规模数据处理而设计的快速通用的计算引擎，它是由Java语言开发实现的。 （易）

参考答案: 错