_____ 学年 第____学期

_____ 课程 期末考试 试卷（ A） 共 页 （闭卷考试）

适用学生： _____

_____ 学院 _____ 专业 _____ 级_____ 班 姓名 _____ 学号 _____

题目	一	二	三	四	总分
得分					

一. 单选题：本大题共 15 小题，每小题 2 分，共 30 分。

1. 下列哪个不是大数据的特征？（ D ）

A. Volume B. Variety C. Velocity D. Variance

2. 下列 List 的定义中不正确的是（ C ）。

A. val number:List[Int] = 1::2::3::Nil B. val name:List[String] = "Tom"::("Jack"::("Lucy"::Nil))

C.val list:String = List(“ a ”, “ b ”, “ c ”) D. val number:List[Int] = List(1, 2, 3, 4, 5)

3. 在图的创建过程中，哪个不是图的创建方法（ C ）。

A. apply B. fromEdges C. fromVertexs D. fromEdgeTuples

4. 下列函数的定义中不正确的是（ B ）。

A. def add(x:Int, y:Int):Int = {x + y} B.def minus(x:Int, y:Int) => {return x −y}

C. def addFun(x:Int)(y:Int) = x * y D. val add = (a:Double, b:Double) => a + b

5. Spark 堆内内存中表示执行内存， 主要用于存放计算过程中的临时数据， 如执行 Shuffle 时占用的内存是 （ B ）。

A. Storage Memory B. Execution Memory C. User Memory D. Reserved Memory

6. 下列不是 Spark 的部署模式的是（ C ）。

A. 单机式 B. 单机伪分布式 C. 列分布式 D. 完全分布式

7. 下列哪个不可以作为 Spark Streaming 的输入数据流？（ D ）

A. Kafka B. Twitter C. TCP 套接字 D. Openstack

8. 下列不是划窗操作重要参数是（ D ）。

A. 批处理间隔 B. 窗口间隔 C. 滑动间隔 D. 输入流间隔

9. 下面的 Scala 语句能正确执行的是（ B ）。

A. val a = 2; a = 3 B. var a = 2; a = 3

C. var a = 2; a = “ 3 ” D. val a = 2; a = “ 3 ”

10. 下面哪个端口不是 spark 自带服务的端口。（C ）

A.8080 B.4040 C.8090 D.18080

11. 关于广播变量，下面哪个是错误的。 （ D ）

A 任何函数调用 B 是只读的 C 存储在各个节点 D 存储在磁盘或 HDFS

12. Master 的 ElectedLeader 事件后做了哪些操作。 （ D ）

A. 通知 driver　　B. 通知 worker　　C. 注册 application　　D. 直接 ALIVE

13. 下面哪个不是图内部存在的操作（ D ）。

A. 图结构转换操作　　B. 图关联操作　　C. 图近邻聚合操作　　D. 图划分操作

14. 下面哪个不是 MLlib 的特征选择方法（ B ）。

A. VectoerSlicer　　B. KafSelector　　C. Rformula　　D. ChiSqSelector

15. 按照任务划分，下列不是机器学习模型的是（ D ）。

A. 回归模型　　B. 分类模型　　C. 结构化学习模型　　D. 结构化统计模型

二. 填空题：本大题 8 小题，共 10 空，每空 2 分，共 20 分。

1. Spark 的基本计算单元是 _____。

2. Scala 运行在 _____ 之上，源代码通过 scalac 编译器编译成 Java 字节码，Scala 兼容现有的 Java 程序。

3. 如果想把一列特征进行数值化，使得相应的特征索引化要使用 _____ 方法。

4. 默认的存储级别是 _____ 。

5. 在 Spark 运行架构中，以下 _____ 组件负责任务控制。

6. Task 是运行在 _____ 中 Executor 上的工作单元。

7. Scala 使用两个关键字来声明变量： _____ 和_____ 。

8. Dstream 的转换操作中，将 RDD 进行聚合的操作是 _____ 。

三. 简答题：本大题有 5 小题，每小题 6 分，共 30 分。

1. driver 的功能是什么？

2. spark 工作机制？

3. RDD 机制？

4.Spark 技术栈有哪些组件，每个组件都有什么功能，适合什么应用场景？

四. 程序题：共 20 分。

使用 Spark 编程对文件 E:\\hello.txt 中的单词进行统计，完成程序中的代码。

```
import org.apache.spark.rdd.RDD

import org.apache.spark.{SparkConf, SparkContext}

object SparkWordCountWithScala {

def main(args: Array[String]): Unit = {

  val conf = new SparkConf()

 //设置运行模式为 local

 1._____

 //设置任务名

 conf.setAppName("WordCount")
```

```
1. conf.setMaster("local")
2. val word: RDD[String] = file.flatMap(_.split(","))
3. val wordOne: RDD[(String, Int)] = word.map((_,1))
4. val wordCount: RDD[(String, Int)] = wordOne.reduceByKey(_+_)
5. val sortRdd: RDD[(String, Int)] = wordCount.sortBy(tuple => tuple._2, false)
```

```
//创建 SparkCore  的程序入口
val sc = new SparkContext(conf)
//读取文件生成  RDD
val file: RDD[String] = sc.textFile("E:\\hello.txt")
//把每一行数据按照“  ,”分割
2. _____
  //让每一个单词都出现一次
3._____
  //单词计数
4._____
  //按照单词出现的次数降序排序
5._____
  //将最终的结果进行保存
  sortRdd.saveAsTextFile("E:\\result")
  sc.stop()
}
```

1. conf.setMaster("local")
2. val word: RDD[String] = file.flatMap(_.split(","))
3. val wordOne: RDD[(String, Int)] = word.map((_,1))
4. val wordCount: RDD[(String, Int)] = wordOne.reduceByKey(_+_)
5. val sortRdd: RDD[(String, Int)] = wordCount.sortBy(tuple => tuple._2, false)

_____ 学年 第____学期

_____ 课程 期末考试 试卷（ B） 共 页 （闭卷考试）

适用学生： _____

_____ 学院 _____ 专业 ____级____ 班 姓名 _____ 学号 _____

题目	一	二	三	四	总分
得分					

一. 单选题：本大题共 20 小题，每小题 2 分，共 40 分。

1. 下列不属于 Spark 生态系统的是（ B ）。

A. Spark Streaming B. Storm C. Shark SQL D. Spark R

2. 以下说法正确的是（ A ）。

A. 数组中的元素都属于同一个数据类型 B. 默认情况下， Scala 使用的是可变映射

C. List 中不可以存放重复对象 D.Set 中可以有重复对象且 Set 中元素是有序的

3. spark.deploy.recoveryMode 不支持那种（ D ）。

A.ZooKeeper B. FileSystem C. NONE D. Hadoop

4. 表达式“ for(i <-1 to 3; j <- 1 to 3; if i != j)print((i + j) +" ") 的”输出结果是（ A ）。

A. 3 4 3 5 4 5 B. 1 2 3 4 5 6 C. 2 4 6 D.1 2 4 6

5. 有如下函数定义。

```
def fac(n:Int):Int={
var res=0
for(i<-1 to n){
res += i
}
res
}
```

则 fac(5) 的输出结果是（ B ）。

A. 14 B. 15 C. 0 D. 5

6. 关于累加器，下面哪个是错误的（ D ）。

A.支持加法 B. 支持数值类型 C. 可并行 D. 不支持自定义类型

7. 在图的属性转换过程中，哪个不是图的属性转换方法（ C ）。

A. mapEdges B. mapTriplets C. mapDegree D. mapVertices

8. 下面哪一种全部是转化操作？（ C ）

A. map 、 take 、 reduceByKey B. map 、 filter 、 collect

C. map 、 zip、 reduceByKey D. map 、 join 、 take

9. Dstream 的转换操作中，将 RDD 进行聚合的操作是（ B ）。

A. flat map()　　B. reduce()　　C. count()　　D. union()

10. 下列适合 Spark 大数据处理场景的是（ D ）。

A. 复杂的批处理　　B. 基于历史数据的交互式查询

C. 基于实时数据流的数据处理　　D. PB 级的数据存储

11. 下列不属于 Spark Streaming 的输出操作的是（ B ）。

A. saveAsTextFiles　　B. saveAsStreamingFiles

C. saveAsHadoopFiles　　D. saveAsObjectFiles

12. 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

13. 下面哪个属性是图无法获取到的（ C ）。

A. VertexRDD　　B. numVertices　　C. EdgeVertex　　D. EdgeRDD

14. 下列不是 MLlib 数据类型的是（ D ）。

A. 本地向量　　B. 标记向量　　C. 本地矩阵　　D. 向量矩阵

15. 下列不属于大数据技术的是（ C ）。

A. 大数据采集技术　　B. 大数据存储及管理技术

C. 财务报表分析技术　　D. 大数据分析及挖掘技术

二. 填空题：本大题有 9 小题，共 10 空，每空 2 分，共 20 分。

1. Scala 是_____ ，每个值都是一个对象，包括基本数据类型和函数，每个操作都是方法的调用。Scala 是一门 _____ 语言，每个函数都是一个值。

2. Spark 中使用 _____ 对 RDD 的关系进行建模。

3. Spark 的一个重要特点是基于 _____ 计算的，因而更快。

4. Scala 中没有基本类型的概念，Scala 中没有原生的数据类型，所有的数据类型都是 _____ 。

5. 在特征提取过程中，通过计数方法将一组文本文档转换为向量使用 _____ 方法。

6. Stage 的 Task 的数量由 _____ 的决定。

7. Spark Job 默认的调度模式是 _____ 。

8. 如果一个 RDD 在计算过程中出错，可以直接通过它的父 RDD 重新计算得到，这就是 Spark 基于 _____ 的容错机制。

9. Spark 采用 _____ 和堆外内存（ Off-heap memory ）的规划机制。

三. 简答题：本大题有 5 小题，每小题 6 分，共 30 分。

1. spark 的有几种部署模式，每种模式特点？

2. Spark 中 Worker 的主要工作是什么？

3. 什么是 RDD 宽依赖和窄依赖？

4. spark 有哪些组件？

5. Spark 为什么比 mapreduce 快？

四．程序题：本大题有　2 小题，每小题　10 分，共 20 分。

使用 Spark 编程对 E:\\hello.txt 中的单词进行统计，完成程序中的代码。

```
import org.apache.spark.rdd.RDD
import org.apache.spark.{SparkConf, SparkContext}
object SparkWordCountWithScala {
    def main(args: Array[String]): Unit = {
        val conf = new SparkConf()
        //设置运行模式为　local
        conf.setMaster("local")
        //设置任务名
        conf.setAppName("WordCount")
        //创建 SparkCore 的程序入口
1._____
        //读取文件生成　RDD
        val file: RDD[String] = sc.textFile("E:\\hello.txt")
        //把每一行数据按照“　,”分割
        2.
        //让每一个单词都出现一次
        3._____
        //单词计数
        4._____
        //按照单词出现的次数　　降序排序
        5._____
        //将最终的结果进行保存
        sortRdd.saveAsTextFile("E:\\result")
        sc.stop()
    }
```

_____ 学年 第____学期

_____ 课程 期末考试 试卷（ 1） 共 页 （闭卷考试）

适用学生： _____

_____ 学院 _____ 专业 _____ 级_____ 班 姓名 _____ 学号 _____

题目	一	二	三	四	总分
得分					

一. 单选题：本大题共 15 小题，每小题 2 分，共 30 分。

1. D 2. C 3. C 4. B 5. B 6. C 7. D 8. D 9. B 10. C 11. D 12. D 13. D 14. B 15. D

二. 填空题：本大题 8 小题，共 10 空，每空 2 分，共 20 分。

1. 弹性分布式数据集（ Resilient Distributed Dataset ， RDD ）

2. Java 虚拟机（ JVM)、 scalac

3. StringIndexer

4. MEMORY_ONLY

5. Driver Program

6. worker node

7. val 、 var

8. reduce()

三. 简答题：本大题有 5 小题，每小题 6 分，共 30 分。

1. driver 的功能是什么？

答： 1）一个 Spark 作业运行时包括一个 Driver 进程，也是作业的主进程， 具有 main 函数，并且有 SparkContext 的实例，是程序的人口点；

2）功能：负责向集群申请资源，向 master 注册信息，负责了作业的调度， ，负责作业的解析、生成 Stage 并调度 Task 到 Executor 上。包括 DAGScheduler ， TaskScheduler 。

2. spark 工作机制？

答：用户在 client 端提交作业后，会由 Driver 运行 main 方法并创建 spark context 上下文。

执行 add 算子，形成 dag 图输入 dagscheduler ，按照 add 之间的依赖关系划分 stage 输入 task scheduler 。 task scheduler 会将 stage 划分为 task set 分发到各个节点的 executor 中执行。

3. RDD 机制？

答： rdd 分布式弹性数据集，简单的理解成一种数据结构，是 spark 框架上的通用货币。

所有算子都是基于 rdd 来执行的，不同的场景会有不同的 rdd 实现类，但是都可以进行互相转换。

rdd 执行过程中会形成 dag 图，然后形成 lineage 保证容错性等。 从物理的角度来看 rdd 存储的是 block 和 node 之间的映射。

4.Spark 技术栈有哪些组件，每个组件都有什么功能，适合什么应用场景？

答：1）Spark core ：是其它组件的基础， spark 的内核，主要包含： 有向循环图、 RDD 、Lingage 、Cache 、broadcast 等，并封装了底层通讯框架，是 Spark 的基础。

2） SparkStreaming 是一个对实时数据流进行高通量、容错处理的流式处理系统，可以对多种数据源（如 Kdfka 、Flume 、 Twitter 、 Zero 和 TCP 套接字）进行类似 Map 、Reduce 和 Join 等复杂操作，将流式计算分解成一系列短小的批处理作业。

3） Spark sql ：Shark 是 SparkSQL 的前身， Spark SQL 的一个重要特点是其能够统一处理关系表和 RDD ，使得开发人员可以轻松地使用 SQL 命令进行外部查询，同时进行更复杂的数据分析

4）BlinkDB ：是一个用于在海量数据上运行交互式 SQL 查询的大规模并行查询引擎， 它允许用户通过权衡数据精度来提升查询响应时间，其数据的精度被控制在允许的误差范围内。

5）MLBase 是 Spark 生态圈的一部分专注于机器学习， 让机器学习的门槛更低， 让一些可能并不了解机器学习的用户也能方便地使用 MLbase 。 MLBase 分为四部分： MLlib 、 MLI 、 ML Optimizer 和 MLRuntime 。

6） GraphX 是 Spark 中用于图和图并行计算

5. spark 的优势和劣势

优势：

1)速度快

2)其次， Spark 是一个灵活的运算框架，适合做批次处理、工作流、交互式分析、流量处理等不同类型的应用，因此 spark 也可以成为一个用途广泛的运算引擎，并在未来取代 MapReduce 的地位

3)最后， Spark 可以与 Hadoop 生态系统的很多组件互相操作。 Spark 可以运行在新一代资源管理框架 YARN 上，它还可以读取已有并存放在 Hadoop 上的数据，这是个非常大的优势

劣势：

1)稳定性方面

2)不能处理大数据

3)不能支持复杂的 SQL 统计

四. 程序题：共 20 分。

```
1. conf.setMaster("local")

2. val word: RDD[String] = file.flatMap(_.split(","))

3. val wordOne: RDD[(String, Int)] = word.map((_,1))

4. val wordCount: RDD[(String, Int)] = wordOne.reduceByKey(_+_)

5. val sortRdd: RDD[(String, Int)] = wordCount.sortBy(tuple => tuple._2,false)
```

_____ 学年 第____学期

_____ 课程 期末考试 试卷（ B） 共 页 （闭卷考试）

适用学生： _____

_____ 学院 _____ 专业 ____级____ 班 姓名 _____ 学号 _____

题目	一	二	三	四	总分
得分					

一. 单选题：本大题共 20 小题，每小题 2 分，共 40 分。

1. B 2. A 3. D 4. A 5. B 6. D 7. C 8. C 9. B 10. D 11. B 12. B 13. C 14. D 15. C

二. 填空题：本大题有 9 小题，共 10 空，每空 2 分，共 20 分。

1. 纯面向对象的、函数式编程

2. DAG

3.内存

4.对象

5. Tokenizer

6. Partition

7. FIFO

8. Lineage

9.堆内内存（ On-heap memory ）

三. 简答题：本大题有 5 小题，每小题 6 分，共 30 分。

1. spark 的有几种部署模式，每种模式特点？

1）本地模式

Spark 不一定非要跑在 hadoop 集群，可以在本地，起多个线程的方式来指定。将 Spark 应用以多线程的方式直接运行在本地，一般都是为了方便调试，本地模式分三类

· local ：只启动一个 executor

· local[k]: 启动 k 个 executor

· local

：启动跟 cpu 数目相同的 executor

2)standalone 模式

分布式部署集群， 自带完整的服务，资源管理和任务监控是 Spark 自己监控，这个模式也是其他模式的基础，

3)Spark on yarn 模式

分布式部署集群，资源和任务监控交给 yarn 管理，但是目前仅支持粗粒度资源分配方式，包含 cluster 和 client 运行模式， cluster 适合生产， driver 运行在集群子节点，具有容错功能， client 适合调试， dirver 运行在客户端

4）Spark On Mesos 模式。官方推荐这种模式（当然，原因之一是血缘关系）。正是由于 Spark 开发之初就考虑到支持 Mesos，因此，目前而言， Spark 运行在 Mesos 上会比运行在 YARN 上更加灵活，更加自然。

2. Spark 中 Worker 的主要工作是什么？

答：主要功能：管理当前节点内存， CPU 的使用状况，接收 master 分配过来的资源指令，通过 ExecutorRunner 启动程序分配任务， worker 就类似于包工头，管理分配新进程，做计算的服务，相当于 process 服务。需要注意的是： 1）worker 会不会汇报当前信息给 master， worker 心跳给 master 主要只有 workid， 它不会发送资源信息以心跳的方式给 mater， master 分配的时候就知道 work，只有出现故障的时候才会发送资源。 2）worker 不会运行代码，具体运行的是 Executor 是可以运行具体 appliaction 写的业务逻辑代码，操作代码的节点，它不会运行程序的代码的。

3. 什么是 RDD 宽依赖和窄依赖？

RDD 和它依赖的 parent RDD(s) 的关系有两种不同的类型，即窄依赖（ narrow dependency ）和宽依赖（ wide dependency ）。

1）窄依赖指的是每一个 parent RDD 的 Partition 最多被子 RDD 的一个 Partition 使用

2）宽依赖指的是多个子 RDD 的 Partition 会依赖同一个 parent RDD 的 Partition

4. spark 有哪些组件？

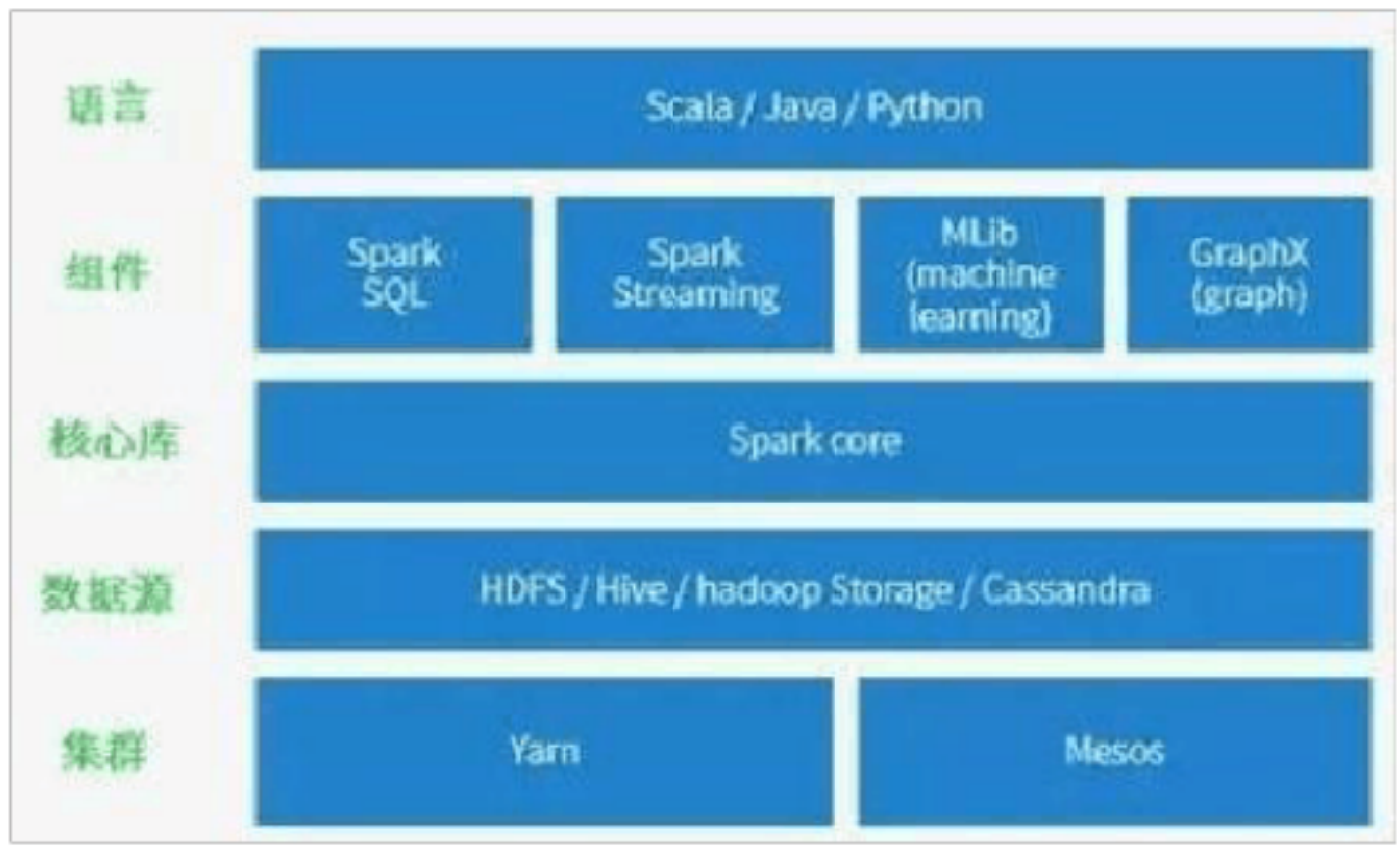
答：主要有如下组件：

1） master ：管理集群和节点，不参与计算。

2） worker ：计算节点，进程本身不参与计算，和 master 汇报。

3） Driver ：运行程序的 main 方法，创建 spark context 对象。

4） spark context ：控制整个 application 的生命周期



5. Spark 为什么比 mapreduce 快？

答： 1）基于内存计算，减少低效的磁盘交互；

2）高效的调度算法，基于 DAG ；

3) 容错机制 Linage ，精华部分就是 DAG 和 Lingae

四．程序题：本大题有 2 小题，每小题 10 分，共 20 分。

使用 Spark 编程对 E:\\hello.txt 中的单词进行统计，完成程序中的代码。

```
1.val sc = new SparkContext(conf)

2.val word: RDD[String] = file.flatMap(_.split(","))

3.val wordOne: RDD[(String, Int)] = word.map((_,1))

4.val wordCount: RDD[(String, Int)] = wordOne.reduceByKey(_+_)

5.val sortRdd: RDD[(String, Int)] = wordCount.sortBy(tuple => tuple._2,false)
```