

《练习题》

单选题（共计 120 分）

1. Hadoop1.0 中, Hadoop 内核的主要组成是（ ）。
- A、HDFS 和 MapReduce
- B、HDFS 和 Yarn
- C、Yarn
- D、MapReduce 和 Yarn
2. 在 HDFS 中, 用于保存数据的节点是（ ）。
- A、namenode
- B、datanode
- C、secondaryNode
- D、yarn
3. 下列选项中, 关于 SSH 服务说法正确的是（ ）。
- A、SSH 服务是一种传输协议
- B、SSH 服务是一种通信协议

- C、SSH 服务是一种数据包协议
- D、SSH 服务是一种网络安全协议
4. 下列选项中, 存放 Hadoop 配置文件的目录是（ ）。
- A、include
- B、bin
- C、libexec
- D、etc
5. Hadoop 集群启动成功后, 用于监控 HDFS 集群的端口是（ ）。
- A、50010
- B、50075
- C、8485
- D、50070
6. 下列选项中, 关于 HDFS 说法错误的是（ ）。
- A、HDFS 是 Hadoop 的核心之一
- B、HDFS 源于 Google 的 GFS 论文

C、HDFS 用于存储海量大数据

D、HDFS 是用于计算海量大数据

7. 下列选项中, 若是哪个节点关闭了, 就无法访问 Hadoop 集群 ()。

A、namenode

B、datanode

C、secondary namenode

D、yarn

8. 下列选项中, 关于 HDFS 的架构说法正确的是 ()。

A、HDFS 采用的是主备架构

B、HDFS 采用的是主从架构

C、HDFS 采用的是从备架构

D、以上说法均错误

9. 下列说法中, 关于客户端从 HDFS 中读取数据的说法错误的是 ()。

A、客户端会选取排序靠前的 DataNode 来依次读取 Block 块

B、客户端会把最终读取出来所有的 Block 块合并成一个完整的最终文件

C、客户端会选取排序靠后的 DataNode 来依读取 Block 块

D、如果客户端本身就是 DataNode, 那么将从本地直接获取数据

10. 在 MapReduce 程序中, map() 函数接收的数据格式是 ()。

A、字符串

B、整型

C、Long

D、键值对

11. 每个 Map 任务都有一个内存缓冲区, 默认大小是 ()。

A、128M

B、64M

C、100M

D、32M

12. 在 MapTask 的 Combine 阶段, 当处理完所有数据时, MapTask 会对所有的临时文件进行一次 ()。

A、分片操作

B、合并操作

C、格式化操作

D、溢写操作

13. 下列选项中, 主要用于决定整个 MapReduce 程序性能高低的阶段是 ()。

A、MapTask

B、ReduceTask

C、分片、格式化数据源

D、Shuffle

14. 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、A、有向无环

B、B、无环图

C、C、有向图

D、D、无向有环图

15. 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、A、first()

B、B、count()

C、C、take(n)

D、D、reduce()

16. Spark 中 RDD 的计算函数的基本单位是 ()。

A、A、分片

B、B、数据块

C、C、Task

D、D、Job

17. 在 Spark RDD 中, 划分 Stage 的依据是 ()。

A、A、窄依赖

B、B、宽依赖

C、C、窄依赖和宽依赖

D、D、以上选项均错误

18. 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、A、filter()

B、B、groupByKey()

C、C、reduceByKey()

D、D、map()

19. 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 ()。

A、A、filter()

B、B、groupByKey()

C、C、reduceByKey()

D、D、map()

20. Spark 与 Hadoop 在基于内存的运算中, 说法正确的是 ()。

A、A、Spark 的运算效率是 Hadoop 的 10 倍

B、B、Spark 的运算效率是 Hadoop 的 100 倍

C、C、Hadoop 的运算效率是 Spark 的 100 倍

D、D、Hadoop 的运算效率是 Spark 的 10 倍

21. 下列选项中, 用于提交和监控 Task 任务的是 ()。

A、A、DAG Scheduler

B、B、Task Scheduler

C、C、Cluster Manager

D、D、SparkContext

22. 下列选项中, 可以用于退出 Spark-Shell 客户端的命令是 ()。

A、A、:quit

B、B、:wq

C、C、:q

D、D、:exit

23. Spark 集群的任务是由 () 进行调度的。

A、A、驱动器

B、B、执行器

C、C、集群管理器

D、D、应用管理器

24. 关于 SecondaryNameNode 哪项是正确的?

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
25. HDFS 中的 Block 默认保存 () 份。
- A、3 份
- B、2 份
- C、1 份
- D、不确定
26. 下列哪项通常是集群的最主要的性能瓶颈?
- A、CPU
- B、网络
- C、磁盘
- D、内存
27. Zookeeper 中的数据存储结构和标准文件系统非常类似, 两者采用的层次结构是 () 。
- A、树形

- B、星形
- C、网形
- D、分布式
28. 下列说法中, 关于 Zookeeper 说法错误的是 () 。
- A、Apache Zookeeper 旨在减轻构建健壮分布式系统的服务
- B、Zookeeper 最早起源于雅虎研究院的一个研究小组
- C、Zookeeper 是一个分布式协调服务的收费框架
- D、Zookeeper 本质上是一个分布式的小文件存储系统
29. Hive 查询语言和 SQL 的一个不同之处在于 () 操作。
- A、Group by
- B、Join
- C、Partition
- D、Union
30. 下列选项中, 不属于 Spark 的四大组件的是 () 。
- A、A、Spark Streaming

B、B、Mlib

C、C、Graphx

D、D、Spark R

31. 关于 Spark Streaming, 下列说法错误的是哪一项?

A、A、Spark Streaming 是 Spark 的核心子框架之一。

B、B、Spark Streaming 具有可伸缩、高吞吐量、容错能力强等特点。

C、C、Spark Streaming 处理的数据源可以来自 Kafka。

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

32. Task 是运行 () 中 Executor 的工作单元。

A、A、Driver program

B、B、spark master

C、C、worker node

D、D、Cluster manager

33. 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、A、transform(func)

B、B、updateStateByKey(func)

C、C、countByKey()

D、D、cogroup(otherStream, [numTasks])

34. 下列选项中, 说法正确的是哪个?

A、A、窗口滑动时间间隔必须是批处理时间间隔的倍数。

B、B、Kafka 是 Spark Streaming 的基础数据源。

C、C、DStream 不可以通过外部数据源获取。

D、D、reduce(func) 是 DStream 的输出操作。

35. 下列说法不正确的是?

A、数据源是数据仓库的基础, 通常包含企业的各种内部信息和外部信息。

B、数据存储及管理是整个数据仓库的核心。

C、OLAP 服务器对需要分析的数据按照多维数据模型进行重组、分析, 发现数据规律和趋势。

D、前端工具主要功能是将数据可视化展示在前端页面中。

36. 下列选项中, 哪项不是 Hive 系统架构的组成部分?

A、用户接口

B、跨语言服务

C、HDFS

D、底层驱动引擎

37. Hive 加载数据文件到数据表中的关键语法是?

A、LOAD DATA [LOCAL] INPATH filepath [OVERWRITE] INTO TABLE tablename

B、INSERTDATA [LOCAL] INPATH filepath [OVERWRITE] INTO TABLE tablename

C、LOAD DATA INFILE d:\car.csv APPEND INTO TABLE t_car_temp FIELDS TERMINATED BY “,”

38. 在 Scala 工程中, .idea 文件夹是用于 ()。

A、A、存放该工程的代码

B、B、存放相关依赖

C、C、存放该工程的配置信息

D、D、存放测试代码

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 ()。

A、A、DAG

B、B、Block

C、C、Application

D、D、Stage

40. 下列选项中, 可以支持 Scala 和 Python 编程的交互式解释器是 ()。

A、A、HBase-Shell

B、B、Spark-Shell

C、C、Hadoop-Shell

D、D、Hive-Shell

41. 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 ()。

A、<dependency>

B、<groupId>

C、<properties>

D、<artifactId>

42. 下列选项中, 不属于 Spark 自带的服务端口有 ()。

A、A、8080

B、B、4040

C、C、8090

D、D、18080

43. Spark 于 2009 年诞生于（）。

A、A、美国加州大学伯克利分校的 AMP 实验室

B、B、加利福尼亚大学伯克利分校

C、C、宾夕法尼亚大学

D、D、普林斯顿大学

44. Spark SQL 快速的计算效率得益于（）。

A、A、Catalyst

B、B、Execution

C、C、Parser

D、D、Analyzer

45. 在 Catalyst 优化器中, 用于将 LogicalPlan 转换成 PhysicalPlan 的组件是（）。

A、A、SqlParse

B、B、Analyze

C、C、Optimizer

D、D、Planner

46. DataFrame 的结构类似于传统数据库的（）。

A、A、一维表格

B、B、二维表格

C、C、三维表格

D、D、四维表格

47. SaveMode 属于（）类型。

A、A、整数类型

B、B、浮点类型

C、C、枚举类型

D、D、字符串类型

48. Spark SQL 的前身是（）。

A、A、SQL

B、B、Shark

C、C、Spark RDD

D、D、MapReduce

49. Client 与 HBase 进行通信是通过 ()。

A、A、RPC 协议

B、B、TCP 协议

C、C、HTTP 协议

D、D、UDP 协议

50. 在 HBase 表中, 列的限定符为 ()。

A、A、冒号

B、B、逗号

C、C、斜杠

D、D、下划线

51. 每个 Region 存储的数据是有限的, 如果当 Region 增大到一个阈值 () 时, 会被等分切成两个新的 Region。

A、A、64M

B、B、128M

C、C、256M

D、D、512M

52. 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 ()。

A、A、Hive

B、B、Oracle

C、C、Mongodb

D、D、HBase

53. 启动 HBase 集群的命令是 ()。

A、A、start-dfs.sh

B、B、zkServer.sh start

C、C、start-hbase.sh

D、D、start-yarn.sh

54. HBase 的底层依赖的是 ()。

A、A、Hive

B、B、HDFS

C、C、Mongodb

D、D、MySQL

55. 关于 MongoDB 数据库的层次结构, 下列说法正确的是 ()。

A、文档、集合、数据库。

B、文档、表、数据库。

C、字段、集合、数据库。

D、行、表、数据库。

56. 关于 MongoDB 文档的描述, 下列说法不正确的是 ()。

A、文档是以键值对的形式存储在集合中。

B、MongoDB 文档的键可以为任意文件类型。

C、MongoDB 文档中值的存储形式为 BSON。

D、每个文档都有一个默认的_id 键。

57. 关于 MongoDB 数据库中基本单元, 下列说法正确的是 ()。

A、行

B、键

C、表格

D、文档

58. 下列选项中, 关于 MongoDB 数据库使用规范说法正确的是 ()。

A、MongoDB 数据库名称不区分大小写

B、MongoDB 数据库没有长度限制

C、MongoDB 数据库名称可以包含特殊字符“-”

D、数据库名称不可与系统保留的数据库名称相同

59. 下列程序中, 可以用于启动 Redis 服务器的是 ()。

A、redis-benchmark.exe

B、redis-check-aof.exe

C、redis-check-dump.exe

D、redis-server.exe

60. 下列选项中, 属于 Redis 应用场景的有 ()。

A、社交

B、缓存

C、推荐系统

D、图存储

判断题（对的打“√”，错的打“×”；共 30 分）

61. 大数据在医疗行业中可以有效控制疾病的发生。()
62. JobTracker 只负责执行 TaskTracker 分配的计算任务。()
63. 启动 Hadoop 集群时,可能出现 NodeManager 进程无法启动或者启动后自动关闭情况,这是由于系统内存和资源分配不足导致的。()
64. 启动 Hadoop 集群,只能有一种方式启动,即单节点逐个启动。()
65. 执行“start-all.sh”指令,可以一键启动整个 Hadoop 集群的服务。()
66. Hadoop 集群执行完 MapReduce 程序后,会输出_SUCCESS 和 part-r-00000 结果文件。()
67. Secondary NameNode 可以有效缩短 Hadoop 集群的启动时间。()
68. HDFS 适用于低延迟数据访问的场景,例如毫秒级实时查询。()
69. DataNode 是 HDFS 集群的主节点,NameNode 是 HDFS 集群的从节点。()
70. MapReduce 的数据流模型可能只有 Map 过程,由 Map 产生的数据直接被写入 HDFS 中。()
71. Hadoop 提供的 Mapper 类是实现 Map 任务的一个抽象基类。()
72. 在 MapReduce 程序中,只有 Map 阶段涉及到 Shuffle 机制。()

73. 优先位置列表会存储每个 Partition 的优先位置, 对于一个 HDFS 文件来说, 就是每个 Partition 块的位置。()
74. RDD 采用了惰性调用。()
75. DAG 是一种非常重要的图论数据结构。()
76. Hadoop 的 MapReduce 进行计算时, 每次产生的中间结果都是存储在内存中; 而 Spark 在计算时产生的中间结果存储在本地磁盘中。()
77. 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。()
78. 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。()
79. Apache Spark Streaming 是 Apache 公司非开源的实时计算框架。()
80. 可以把不同结构的文件存储在同一个 MongoDB 数据库里。()
81. HDFS 目前不支持并发多用户的写操作, 写操作只能在文件末尾追加数据。()
82. Namenode 存储的是元数据信息, 元数据信息并不是真正的数据, 真正的数据是存储在 DataNode 中。()
83. MapReduce 是 Hadoop 系统核心组件之一, 它是一种可用于大数据并行处理的计算模型、框架和平台()
84. reduce() 函数会将 map() 函数输出的键值对作为输入, 把相同 key 值的 value 进行汇总, 输出新的键值对()
85. 对于 MapReduce 任务来说, 一定需要 Reduce 过程。()
86. MapReduce 通过 TextOutputFormat 组件输出到结果文件中。()
87. 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。()
88. Executor 会向 SparkContext 进行反向注册并申请 Task。()

89. 在 HBase 集群中, 不会出现单点故障的问题。()

90. 一个 HRegion Server 上只能存储一个 Region。()

填空题 (共计 60 分)

91. 【】中引入了资源管理框架 Yarn。

92. 在 HDFS 的高可用集群中, 通常有两台或两台以上的机器充当 NameNode, 在任意时间, 保证有一台机器处于 【】 状态, 一台机器处于 【】 状态。

93. NameNode 主要以 【】 的形式对数据进行管理和存储。

94. 【】 节点, 负责记录文件系统名称空间或其属性的任何更改操作, 并存储配置文件中设置备份的数量。

95. MapReduce 的核心思想是 【】 。

96. MapReduce 编程模型的实现过程是通过 【】 和 【】 函数来完成的。

97. 【】 是 MapReduce 的核心, 它用来确保每个 reducer 的输入都是按键排序的。

98. 【】 表示时间戳, 记录每次操作数据的时间, 通常记作数据的版本号。

99. 若需要停止 HBase 集群, 则执行 【】 命令。

100. 在 Spark 中, 不同的 RDD 之间具有依赖的关系。RDD 与它所依赖的 RDD 的依赖关系有两种类型, 分别是 【】 和宽依赖 (wide dependency) 。

101. 按照 【】 的理念, Spark 在进行任务调度的时候, 会尽可能地将计算任务分配到你所要处理数据块的存储位置。

102. Spark Streaming 具有很好的 【】, 在没有额外代码和配置的情况下, 可以恢复丢失的数据。

103. Spark Streaming 的特点有易用性、 【】、易整合性。

104. Spark Streaming 中对 DStream 的转换操作会变成对 【】 的转换操作。

105. 操作 HBase 常用的方式有两种, 一种是 【】, 另一种是 Java API。

106. 【】 表示行键, 每个 HBase 表中只能有一个行键, 它在 HBase 中以字典序的方式存储。

107. 在 HBase 集群中, 【】 负责为 HRegion Server 分配 HRegion。

108. Scala 语言可以运行在 Windows、 【】、Mac OS 等系统上。

109. NoSQL 是一种 【】 型数据库。

110. 列式存储数据库是以 【】 为单位存储数据

111. HBase 是面向 【】 的存储和权限控制, 并支持独立检索。

112. HBase 表的列是由 【】、限定符以及 【】 组成的。

113. HBase 中的列族由多个 【】 组成。

114. Spark Streaming 支持多种数据源, 例如 【】、Flume 以及 TCP 套接字等数据源。

115. Hadoop 的 MapReduce 在计算数据时, 计算过程必须要转化为 【】 和 Reduce 两个过程。

116. Spark 在 2013 年加入 【】, 之后获得迅猛的发展, 并于 2014 年正式成为 Apache 软件基金会的顶级项目。

117. 输入 Map 阶段的数据源, 必须经过 【】 和格式化操作。

118. ReduceTask 在 Sort 阶段, 为了将 key 相同的数据聚在一起, Hadoop 采用了基于 【】 的策略。

119. Hive 查询语句 select ceil(2.34) 输出内容是 【】 。

120. 数据仓库的结构包含了 4 部分, 分别是数据源、【】服务器、数据管理服务器和前端工具。

简答题 (共计 90 分)

121. 简述 Spark 生态系统的优势。

122. 简述 RDD 的处理过程。

123. 简述 RDD 的依赖关系。

124. 简述 NoSQL 数据库的四大类型和典型产品。

125. 简述 MapReduce 编程模型的核心设计思想与关键特点。

126. 简述 HBase 和传统数据库的区别。

127. 简述 HDFS 的组成结构, 说明其主要组件的作用。

128. 什么是大数据的 4V 特征?

129. 简述流计算的基本概念, 并对比 Storm 与 Spark Streaming 的实现原理和延迟差异。

答案

单选题 (共计 120 分)

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

18. D
19. B
20. B
21. B
22. A
23. A
24. C
25. A
26. C
27. A
28. C
29. C
30. D
31. D
32. C
33. A
34. B
35. D
36. C
37. A

姓名: _____

学号: _____

年级专业: _____

凡年级专业、姓名、学号错写、漏写或字迹不清者，成绩按零分记。

.....密.....封.....线.....

38. C

39. D

40. B

41. C

42. C

43. A

44. A

45. D

46. B

47. C

48. B

49. A

50. A

51. B

52. D

53. C

54. B

55. A

56. B

57. D

- | | |
|-----|---|
| 38. | C |
| 39. | D |
| 40. | B |
| 41. | C |
| 42. | C |
| 43. | A |
| 44. | A |
| 45. | D |
| 46. | B |
| 47. | C |
| 48. | B |
| 49. | A |
| 50. | A |
| 51. | B |
| 52. | D |
| 53. | C |
| 54. | B |
| 55. | A |
| 56. | B |
| 57. | D |

58. D
59. D
60. B

判断题（共计 30 分）

- | | |
|-----|---|
| 61. | 错 |
| 62. | 错 |
| 63. | 对 |
| 64. | 错 |
| 65. | 对 |
| 66. | 对 |
| 67. | 对 |
| 68. | 错 |
| 69. | 错 |
| 70. | 对 |
| 71. | 对 |
| 72. | 错 |
| 73. | 对 |
| 74. | 对 |
| 75. | 对 |

76. 错
77. 错
78. 对
79. 错
80. 对
81. 对
82. 对
83. 对
84. 对
85. 错
86. 对
87. 错
88. 对
89. 错
90. 错

填空题 (共计 60 分)

91. 【Hadoop2. x】
92. 【活动】 【备用】
93. 【元数据】

94. 【NameNode】
95. 【分而治之】
96. 【map()】 【reduce()】
97. 【Shuffle】
98. 【Timestamp】
99. 【stop-hbase.sh】
100. 【窄依赖 (narrow dependency)】
101. 【移动数据不如移动计算】
102. 【容错性】
103. 【容错性】
104. 【RDD】
105. 【Shell 命令行】
106. 【RowKey】
107. 【HMaster】
108. 【Linux】
109. 【非关系】
110. 【列】
111. 【列】
112. 【列族名】 【列名】
113. 【列】

114. **【Kafka】**
115. **【Map】**
116. **【Apache 孵化器项目】**
117. **【分片】**
118. **【排序】**
119. **【3】**
120. **【数据存储】**

简答题（共计 90 分）

121. 正确答案：1. Spark 生态系统包含的所有程序库和高级组件都可以从 Spark 核心引擎的改进中获益，全组件共享 Spark 内核优化成果。
2. Spark 一套引擎整合批处理、流处理、SQL、机器学习，不需要运行多套独立的软件系统，能够大大减少运行整个系统的资源代价。
3. Spark 组件无缝集成，构建不同处理模型的应用，降低集群资源开销与运维成本。
122. 正确答案：1. RDD 分为转换算子与行动算子，转换算子仅记录计算逻辑、惰性生成新 RDD，不实际运算。经过一系列的“转换”操作，每一次转换都会产生不同的 RDD，以供给下一次“转换”操作使用。
2. 只有执行行动算子时才触发真实计算，输出结果。也就是说，直到最后一个 RDD 经过“行动”操作才会被真正计算处理，并输出到外部数据源中。
3. 若是中间的数据结果需要复用，则可以进行缓存处理，将数据缓存到内存中。

123. 正确答案：1. 窄依赖是指单个父 RDD 分区仅对应一个子 RDD 分区，即 OneToOneDependencies。
2. 窄依赖的表现一般分为两类，第一类表现为一个父 RDD 的分区对应于一个子 RDD 的分区；第二类表现为多个父 RDD 的分区对应于一个子 RDD 的分区。也就是说，一个父 RDD 的一个分区不可能对应一个子 RDD 的多个分区。
3. 当 RDD 执行 map、filter 及 union 和 join 操作时，都会产生窄依赖。
4. 宽依赖是指子 RDD 分区依赖父 RDD 多个分区，即 OneToManyDependencies。
5. 当 RDD 做 groupByKey 和 join 操作时，会产生宽依赖。
124. 正确答案：1. 键值对存储数据库，Redis。
2. 文档存储数据库，MongoDB。
3. 列式存储数据库，HBase。
4. 图形存储数据库，Neo4j。
125. 正确答案：1. MapReduce 将分布式并行计算抽象为 Map、Reduce 两个核心函数，简化分布式编程，开发者无需关注底层集群并行细节即可开发大数据计算程序。
2. MapReduce 采用“分而治之”策略，一个存储在分布式文件系统的大规模数据集，会被切分成许多独立的分片（split），这些分片可以被多个 Map 任务并行处理。
3. MapReduce 设计的一个理念是“计算向数据靠拢”，把计算任务调度至数据所在节点运行，减少海量数据跨节点网络传输，节省传输开销。
4. MapReduce 采用 Master/Slave 架构，包括一个 Master 和若干个 Slave。Master 上运行 JobTracker，Slave 上运行 TaskTracker 。Hadoop 框架是用 Java 实现的，但是 MapReduce 应用程序不一定要用 Java 来写。

126. 正确答案: 1. 存储模式: 传统数据库中是基于行存储的, 而 HBase 是基于列进行存储的。

2. 表字段: 传统数据库中建表需要提前固定字段, 而 HBase 中的表字段不作限制, 可以动态新增。

3. 可延伸性: 传统数据库中的列是固定的, 分布式扩展难度大, 而 HBase 是原生分布式架构, 易于集群扩容。

127. 正确答案: 1. HDFS 由 NameNode、DataNode 和客户端三部分组成。

2. NameNode 用于存储元数据, 元数据保存在内存中, 用于保存文件、Block、Datanode 之间的映射关系。

3. DataNode 用以存储文件内容, 文件内容保存在磁盘, 维护了 Block ID 到 Datanode 本地文件的映射关系。

4. 客户端作为用户入口, 实现文件上传、下载等交互操作。

128. 正确答案: 1. 大数据的 4V 特征通常指 Volume 大量化, Velocity 快速化, Variety 多样化和 Value 价值密度低。

2. Volume 大量化: 数据体量巨大, 远超传统处理能力。

3. Velocity 快速化: 数据产生、传输速度快, 要求快速实时处理。

4. Variety 多样化: 数据类型多元, 包含结构化、半结构化以及非结构化数据。

5. Value 价值密度低: 海量数据里有效信息占比低, 但经过处理挖掘后整体价值很高。

129. 正确答案: 1. 流计算: 一种针对源源不断持续生成的流式数据, 实时获取、即时运算、实时输出有价值的信息的计算模式。

2. Storm: 逐条消息处理, 原生流式架构, 可达毫秒级低延迟, 但吞吐弱、状态差, 无完善状态管理。

3. Spark Streaming: 将数据流切分成秒级微批次, 依托 RDD 批处理实现流式运算, 高吞吐, 延迟为秒级, 无法做到毫秒实时。