第一章 初识Hadoop

1、【单选题】下面哪个选项不属于Google的三驾马车？ A

A、HDFS B、MapReduce C、BigTable D、GFS

2、【单选题】大数据的数据量现在已经达到了哪个级别? C

A、GB B、TB C、PB D、ZB

3、【多选题】Hadoop能够使用户轻松开发和运行处理大数据的应用程序，那它主要有下面哪些特点 ABCD

A、高可靠性 B、高扩展性 C、高效性 D、高容错性

4、【多选题】下面哪些是大数据的基本特征？ ABCD

A、数据体量大 B、数据类型多 C、处理速度快 D、价值密度低

5、【判断题】Hbase是非关系型数据库，是面向列的。 V

6、【判断题】MapReduce是的最早提出是Google为了解决PageRank的问题 V

7、【判断题】Google的在大数据解决方案是开源的 X

答案：

1、A 2、C 3、ABCD 4、ABCD

5、√ 6、√ 7、×

第二章 搭建Hadoop集群

1、【单选题】以下哪一项不属于Hadoop可以运行的模式 ABD

A、单机（本地）模式 B、伪分布式模式 C、互联模式 D、分布式模式

2、【单选题】下列哪个程序通常与NameNode在同一个节点启动 D

A、TaskTracker

B、DataNode

C、SecondaryNameNode

D、Jobtracker

3、【单选题】下面哪个程序负责 HDFS 数据存储 C

A、NameNode

B、Jobtracker

C、DataNode

D、SecondaryNameNode

5、【填空题】列出Hadoop运行的模式 单机 、 分布式 、 伪分布式 。

6、【填空题】HDFS客户端执行所有的写操作都会被记录到\_\_\_\_\_\_edits\_\_\_\_\_\_\_\_\_文件中。

7、【填空题】Hadoop搭建中常用的4个配置文件为 core-site.xml 、hdfs-site.xml 、 mapred-site.xml 、 Yarn-site.xml 。

8、【填空题】HDFS中的block默认保存\_\_3\_\_份。

9、【判断题】SecondaryNameNode应与 NameNode部署到一个节点 X

10、【判断题】查看Linux IP的命令是ifconfig V

11、【判断题】每次启动Hadoop都要格式化文件系统 X

12、【判断题】启动Hadoop所有进程的命令是start-all.sh V

13、【判断题】DataNode负责存储数据 V

14、【判断题】修改~/.bashrc文件保存后，修改的内容能立即生效 X

15、【判断题】在/etc/hosts文件映射IP和主机名称 V

16、【判断题】安装Hadoop时，配置项"dfs.replication"是配置在文件core-site.xml X

17 【判断题】设置免密登录需要先用ssh-keygen生成一对密钥 V

18、【简答题】Hadoop的核心配置是什么？

Hadoop的核心配置通过两个xml文件来完成：1，hadoop-default.xml；2，hadoop-site.xml。这些文件都使用xml格式，因此每个xml中都有一些属性，包括名称和值，但是当下这些文件都已不复存在。

答案：

1、C 2、D 3、C 5、单机；伪分布式；分布式 6、edit log

7、core-site.xml ；hdfs-site xml ；mapred-site.xml ；yarn-site.xml 8、3

9、× 10、√ 11、× 12、√ 13、√ 14、× 15、√ 16、× 17、√

第三章 HDFS分布式文件系统

1、【单选题】HDFS是基于流数据模式访问和处理超大文件的需求而开发的，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 D

A、一次写入，少次读写

B、多次写入，少次读写

C、多次写入，多次读写

D、一次写入，多次读写

2、【单选题】HDFS首先把大数据文件切分成若干个小的数据块，再把这些数据块分别写入不同的节点，这些负责保存文件数据的节点被称为 B

A、NameNode

B、DataNode

C、SecondaryNameNode

D、Block

3、【单选题】Hadoop更适合哪些场景 A

A、离线分析 B、复杂数据 C、少量数据 D、在线分析

4、【单选题】Hadoop的复制命令为\_\_\_\_C\_\_。

A、cp

B、hadoop fs -cp

C、hadoop fs -put

D、-put

5、【单选题】关于SecondaryNameNode哪项是正确的 C

A、它是NameNode的设备

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

6、【判断题】HDFS是整个Hadoop生态圈中的基石 V

7、【判断题】在HDFS HA集群中，两个NameNode都处于活跃状态，这样其中

一个WameWode故障时，集群仍然可用 X

8、【判断题】DataNode一旦发生故障将导致整个集群不可用 X

9、【判断题】上传到HDFS的一个数据块是1M，那么它在HDFS上占用的内存是1M V

10、【判断题】NameNode管理了两个文件，其中fsimage体现了其最新的状态 V

11、【判断题】安全模式（Safenode）是IDFS所处的一种特殊状态。处于这种状态时，HDFS只接受读数据请求，不能对文件进行写、删除等操作 V

12、【判断题】HDFS为每一个用户都创建了类似操作系统的回收站（Trash），当用户删除文件时，文件马上就会被永久性删除 X

13、【判断题】hdfs dfsadmin -disallowSnapshot是开启HDFS快照的命令 X

14、【判断题】HDFS提供了如下两种配额（Quota）命令（这两种命令是管理命令——hdfs dfsadmin） V

15、【简答题】简述HDFS的体系结构。

HDFS采用了主从（Master/Slave）结构模型，一个HDFS集群是由一个NameNode和若干个DataNode组成的。其中NameNode 作为主服务器，管理文件系统的命名空间和客户端对文件的访问操作；

集群中的DataNode管理存储的数据。HDFS允许用户以文件的形式存储数据。从内部来看，文件被分成若干个数据块，而且这若干个数据块存放在一组DataNode上。

NameNode执行文件系统的命名空间操作，比如打开、关闭、重命名文件或目录等，它也负责数据块到具体DataNode的映射。DataNode负责处理文件系统客户端的文件读写请求，并在NaneNode的统一调度下进行数据块的创建、删除和复制工作。NameNode和DataNode都被设计成可以在普通商用计算机上运行。这些计算机通常运行的是GNU/Linux操作系统。HDFS采用Java语言开发，因此任何支持Java的机器都可以部署NameNode和DataNode。一个典型的部署场景是集群中的一台机器运行一个NaneNode实例，其他机器分别运行一个 DataNode实例。当然，并不排除一台机器运行多个DataNode实例的情况。集群中单一的 NameNode的设计则大大简化了系统的架构。NameNode是所有HDFS元数据的管理者，用户数据永远不会经过NameNode。

16、【简答题】请详细描述Checkpoint操作。

SecondaryName会定期地创建命名空间的检查点（CheckPoint）操作：把edits中最新的状态信息合并到fsimage文件中，防止edits过大。也可以做冷备，对一定范围内数据做快照性备份。

SecondaryNameNode的CheckPoint流程∶

（1）NameNode生成一个名叫edits.new的文件，用于记录合并过程中产生的日志信息；

（2）SecondaryNameNode将edits文件，与fsimag文件从NameNode上下载到SecondNamenode上；

（3）SecondaryNameNode将edits文件与fsimage进行合并操作，合并成一个fsimage.ckpt文件；

（4）SecondaryNameNode将生成的合并后的文件fsimage.ckpt文件上传到NameNode上；

（5）在NameNode上，将fsimage.ckpt变成fsimage文件替换NameNode上原有的fsinage文件。将edits.new文件上变成edits文件替换NameNode上原有的edits文件。

17、【程序题】根据要求写出对应的HDFS shell命令：

（1）在HDFS上创建名为model的目录

（2）将本地的文件file上传至model目录下

（3）将本地的文件file1内容追加至model目录下的file中

（4）查看model的目录信息

（5）将model下的file文件下载至本地

正确答案∶

hdfs dfs -mkdir model

hdfs dfs -put file model

hdfs dfs -appendToFile file1 model/file

hdfs dfs -ls model

hdfs dfs -get file model

答案：

1—5：DBACC 6—10：√××√√ 11—14：√××√