hadoop基础题

//Hadoop基础

Doug Cutting所创立的项目的名称都受到其家人的启发，以下项目不是由他创立的项目是
A． Hadoop
B． Nutch
C． Lucene
D． Solr
答案：D

配置Hadoop时，JAVA_HOME包含在哪一个配置文件中
A． hadoop-default.xml
B． hadoop-env.sh
C． hadoop-site.xml
D． configuration.xsl
答案：B
知识点：hadoop配置

Hadoop配置文件中，hadoop-site.xml显示覆盖hadoop-default.xml里的内容。在版本0.20中，hadoop-site.xml被分离成三个XML文件，不包括
A． conf-site.xml
B． mapred-site.xml
C． core-site.xml
D． hdfs-site.xml
答案：A
知识点：hadoop配置

HDFS默认的当前工作目录是/user/$USER，fs.default.name的值需要在哪个配置文件内说明
A. mapred-site.xml
B. core-site.xml
C. hdfs-site.xml
D. 以上均不是
答案：B
知识点：hadoop配置

关于Hadoop单机模式和伪分布式模式的说法，正确的是
A． 两者都起守护进程，且守护进程运行在一台机器上
B． 单机模式不使用HDFS，但加载守护进程
C． 两者都不与守护进程交互，避免复杂性
D． 后者比前者增加了HDFS输入输出以及可检查内存使用情况
答案：D
知识点：hadoop配置

下列关于Hadoop API的说法错误的是

A． Hadoop的文件API不是通用的，只用于HDFS文件系统
B． Configuration类的默认实例化方法是以HDFS系统的资源配置为基础的
C． FileStatus对象存储文件和目录的元数据
D． FSDataInputStream是java.io.DataInputStream的子类
答案：A

//HDFS
HDFS的NameNode负责管理文件系统的命名空间，将所有的文件和文件夹的元数据保存在一个文件系统树中，这些信息也会在硬盘上保存成以下文件：
A．日志
B．命名空间镜像
C．两者都是
答案：C
知识点：

HDFS的namenode保存了一个文件包括哪些数据块，分布在哪些数据节点上，这些信息也存储在硬盘上。
A．正确
B．错误
答案：B
知识点：在系统启动的时候从数据节点收集而成的

Secondary namenode就是namenode出现问题时的备用节点
A．正确
B．错误
答案：B
知识点：它和元数据节点负责不同的事情。其主要功能就是周期性将元数据节点的命名空间镜像文件和修改日志合并，以防日志文件过大。合并过后的命名空间镜像文件也在Secondary namenode保存了一份，以防namenode失败的时候，可以恢复。

出现在datanode的VERSION文件格式中但不出现在namenode的VERSION文件格式中的是
A. namespaceID
B. storageID
C. storageType
D. layoutVersion
答案：B
知识点：其他三项是公有的。layoutVersion是一个负整数，保存了HDFS的持续化在硬盘上的数据结构的格式版本号；namespaceID是文件系统的唯一标识符，是在文件系统初次格式化时生成的；storageType表示此文件夹中保存的是数据节点的类型

Client在HDFS上进行文件写入时，namenode根据文件大小和配置情况，返回部分datanode信息，谁负责将文件划分为多个Block，根据DataNode的地址信息，按顺序写入到每一个DataNode块
A． Client
B． Namenode
C． Datanode
D． Secondary namenode
答案：A

知识点：HDFS文件写入

HDFS的是基于流数据模式访问和处理超大文件的需求而开发的，默认的最基本的存储单位是64M，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
A．一次写入，少次读写
B．多次写入，少次读写
C．一次写入，多次读写
D．多次写入，多次读写
答案：C
知识点：HDFS特性

HDFS无法高效存储大量小文件，想让它能处理好小文件，比较可行的改进策略不包括
A． 利用SequenceFile、MapFile、Har等方式归档小文件
B． 多Master设计
C． Block大小适当调小
D． 调大namenode内存或将文件系统元数据存到硬盘里
答案：D
知识点：HDFS特性

关于HDFS的文件写入，正确的是
A． 支持多用户对同一文件的写操作
B． 用户可以在文件任意位置进行修改
C． 默认将文件块复制成三份存放
D． 复制的文件块默认都存在同一机架上
答案：C
知识点：在HDFS的一个文件中只有一个写入者，而且写操作只能在文件末尾完成，即只能执行追加操作。默认三份文件块两块在同一机架上，另一份存放在其他机架上。

Hadoop fs中的-get和-put命令操作对象是
A． 文件
B． 目录
C． 两者都是
答案：C
知识点：HDFS命令

Namenode在启动时自动进入安全模式，在安全模式阶段，说法错误的是
A． 安全模式目的是在系统启动时检查各个DataNode上数据块的有效性
B． 根据策略对数据块进行必要的复制或删除
C． 当数据块最小百分比数满足的最小副本数条件时，会自动退出安全模式
D． 文件系统允许有修改
答案：D
知识点：HDFS安全模式

//MapReduce
MapReduce框架提供了一种序列化键/值对的方法，支持这种序列化的类能够在Map和Reduce过程中充当键或值，以下说

法错误的是
A. 实现Writable接口的类是值
B. 实现WritableComparable<T>接口的类可以是值或键
C. Hadoop的基本类型Text并不实现WritableComparable<T>接口
D. 键和值的数据类型可以超出Hadoop自身支持的基本类型
答案：C

以下四个Hadoop预定义的Mapper实现类的描述错误的是
A. IdentityMapper<K, V>实现Mapper<K, V, K, V>，将输入直接映射到输出
B. InverseMapper<K, V>实现Mapper<K, V, K, V>，反转键/值对
C. RegexMapper<K>实现Mapper<K, Text, Text, LongWritable>，为每个常规表达式的匹配项生成一个(match, 1)对
D. TokenCountMapper<K>实现Mapper<K, Text, Text, LongWritable>，当输入的值为分词时，生成(taken, 1)对
答案：B
知识点：InverseMapper<K, V>实现Mapper<K, V, V, K>

下列关于HDFS为存储MapReduce并行切分和处理的数据做的设计，错误的是
A. FSDataInputStream扩展了DataInputStream以支持随机读
B. 为实现细粒度并行，输入分片(Input Split)应该越小越好
C. 一台机器可能被指派从输入文件的任意位置开始处理一个分片
D. 输入分片是一种记录的逻辑划分，而HDFS数据块是对输入数据的物理分割
答案：B
知识点：每个分片不能太小，否则启动与停止各个分片处理所需的开销将占很大一部分执行时间

针对每行数据内容为”Timestamp Url”的数据文件，在用JobConf对象conf设置
conf.setInputFormat(WhichInputFormat.class)来读取这个文件时，WhichInputFormat应该为以下的
A. TextInputFormat
B. KeyValueTextInputFormat
C. SequenceFileInputFormat
D. NLineInputFormat
答案：B
知识点：四项主要的InputFormat类。KeyValueTextInputFormat以每行第一个分隔符为界，分隔符前为key，之后为value，默认制表符为\t

有关MapReduce的输入输出，说法错误的是
A. 链接多个MapReduce作业时，序列文件是首选格式
B. FileInputFormat中实现的getSplits()可以把输入数据划分为分片，分片数目和大小任意定义
C. 想完全禁止输出，可以使用NullOutputFormat
D. 每个reduce需将它的输出写入自己的文件中，输出无需分片
答案：B
知识点：分片数目在numSplits中限定，分片大小必须大于mapred.min.size个字节，但小于文件系统的块

Hadoop Streaming支持脚本语言编写简单MapReduce程序，以下是一个例子：
bin/hadoop jar contrib/streaming/hadoop-0.20-streaming.jar
—input input/filename
—output output
—mapper ‘dosth.py 5’

—file dosth.py
—D mapred.reduce.tasks=1
以下说法不正确的是
A． Hadoop Streaming使用Unix中的流与程序交互
B． Hadoop Streaming允许我们使用任何可执行脚本语言处理数据流
C． 采用脚本语言时必须遵从UNIX的标准输入STDIN，并输出到STDOUT
D． Reduce没有设定，上述命令运行会出现问题
答案：D
知识点：没有设定特殊的reducer，默认使用IdentityReducer

在高阶数据处理中，往往无法把整个流程写在单个MapReduce作业中，下列关于链接MapReduce作业的说法，不正确的是
A．Job和JobControl类可以管理非线性作业之间的依赖
B．ChainMapper和ChainReducer类可以用来简化数据预处理和后处理的构成
C．使用ChainReducer时，每个mapper和reducer对象都有一个本地JobConf对象
D．ChainReducer.addMapper()方法中，一般对键/值对发送设置成值传递，性能好且安全性高
答案：D
知识点：ChainReducer.addMapper()方法中，值传递安全性高，引用传递性能高

//源码分析

//Zookeeper
//Hadoop基础

Doug Cutting所创立的项目的名称都受到其家人的启发，以下项目不是由他创立的项目是
A． Hadoop
B． Nutch
C． Lucene
D． Solr
答案：D

配置Hadoop时，JAVA_HOME包含在哪一个配置文件中
A． hadoop-default.xml
B． hadoop-env.sh
C． hadoop-site.xml
D． configuration.xsl
答案：B
知识点：hadoop配置

Hadoop配置文件中，hadoop-site.xml显示覆盖hadoop-default.xml里的内容。在版本0.20中，hadoop-site.xml被分离成三个XML文件，不包括
A． conf-site.xml
B． mapred-site.xml
C． core-site.xml

D． hdfs-site.xml
答案：A
知识点：hadoop配置

HDFS默认的当前工作目录是/user/$USER，fs.default.name的值需要在哪个配置文件内说明
A. mapred-site.xml
B. core-site.xml
C. hdfs-site.xml
D. 以上均不是
答案：B
知识点：hadoop配置

关于Hadoop单机模式和伪分布式模式的说法，正确的是
A． 两者都起守护进程，且守护进程运行在一台机器上
B． 单机模式不使用HDFS，但加载守护进程
C． 两者都不与守护进程交互，避免复杂性
D． 后者比前者增加了HDFS输入输出以及可检查内存使用情况
答案：D
知识点：hadoop配置

下列关于Hadoop API的说法错误的是
A． Hadoop的文件API不是通用的，只用于HDFS文件系统
B． Configuration类的默认实例化方法是以HDFS系统的资源配置为基础的
C． FileStatus对象存储文件和目录的元数据
D． FSDataInputStream是java.io.DataInputStream的子类
答案：A

//HDFS
HDFS的NameNode负责管理文件系统的命名空间，将所有的文件和文件夹的元数据保存在一个文件系统树中，这些信息也会在硬盘上保存成以下文件：
A．日志
B．命名空间镜像
C．两者都是
答案：C
知识点：

HDFS的namenode保存了一个文件包括哪些数据块，分布在哪些数据节点上，这些信息也存储在硬盘上。
A．正确
B．错误
答案：B
知识点：在系统启动的时候从数据节点收集而成的

Secondary namenode就是namenode出现问题时的备用节点
A．正确
B．错误

答案：B
知识点：它和元数据节点负责不同的事情。其主要功能就是周期性将元数据节点的命名空间镜像文件和修改日志合并，以防日志文件过大。合并过后的命名空间镜像文件也在Secondary namenode保存了一份，以防namenode失败的时候，可以恢复。

出现在datanode的VERSION文件格式中但不出现在namenode的VERSION文件格式中的是
A. namespaceID
B. storageID
C. storageType
D. layoutVersion
答案：B
知识点：其他三项是公有的。layoutVersion是一个负整数，保存了HDFS的持续化在硬盘上的数据结构的格式版本号；namespaceID是文件系统的唯一标识符，是在文件系统初次格式化时生成的；storageType表示此文件夹中保存的是数据节点的类型

Client在HDFS上进行文件写入时，namenode根据文件大小和配置情况，返回部分datanode信息，谁负责将文件划分为多个Block，根据DataNode的地址信息，按顺序写入到每一个DataNode块
A． Client
B． Namenode
C． Datanode
D． Secondary namenode
答案：A
知识点：HDFS文件写入

HDFS的是基于流数据模式访问和处理超大文件的需求而开发的，默认的最基本的存储单位是64M，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
A．一次写入，少次读写
B．多次写入，少次读写
C．一次写入，多次读写
D．多次写入，多次读写
答案：C
知识点：HDFS特性

HDFS无法高效存储大量小文件，想让它能处理好小文件，比较可行的改进策略不包括
A． 利用SequenceFile、MapFile、Har等方式归档小文件
B． 多Master设计
C． Block大小适当调小
D． 调大namenode内存或将文件系统元数据存到硬盘里
答案：D
知识点：HDFS特性

关于HDFS的文件写入，正确的是
A． 支持多用户对同一文件的写操作
B． 用户可以在文件任意位置进行修改
C． 默认将文件块复制成三份存放
D． 复制的文件块默认都存在同一机架上

答案：C
知识点：在HDFS的一个文件中只有一个写入者，而且写操作只能在文件末尾完成，即只能执行追加操作。默认三份文件块两块在同一机架上，另一份存放在其他机架上。

Hadoop fs中的-get和-put命令操作对象是
A． 文件
B． 目录
C． 两者都是
答案：C
知识点：HDFS命令

Namenode在启动时自动进入安全模式，在安全模式阶段，说法错误的是
A． 安全模式目的是在系统启动时检查各个DataNode上数据块的有效性
B． 根据策略对数据块进行必要的复制或删除
C． 当数据块最小百分比数满足的最小副本数条件时，会自动退出安全模式
D． 文件系统允许有修改
答案：D
知识点：HDFS安全模式

//MapReduce
MapReduce框架提供了一种序列化键/值对的方法，支持这种序列化的类能够在Map和Reduce过程中充当键或值，以下说法错误的是
A． 实现Writable接口的类是值
B． 实现WritableComparable<T>接口的类可以是值或键
C． Hadoop的基本类型Text并不实现WritableComparable<T>接口
D． 键和值的数据类型可以超出Hadoop自身支持的基本类型
答案：C

以下四个Hadoop预定义的Mapper实现类的描述错误的是
A． IdentityMapper<K, V>实现Mapper<K, V, K, V>，将输入直接映射到输出
B． InverseMapper<K, V>实现Mapper<K, V, K, V>，反转键/值对
C． RegexMapper<K>实现Mapper<K, Text, Text, LongWritable>，为每个常规表达式的匹配项生成一个(match, 1)对
D． TokenCountMapper<K>实现Mapper<K, Text, Text, LongWritable>，当输入的值为分词时，生成(taken, 1)对
答案：B
知识点：InverseMapper<K, V>实现Mapper<K, V, V, K>

下列关于HDFS为存储MapReduce并行切分和处理的数据做的设计，错误的是
A． FSDataInputStream扩展了DataInputStream以支持随机读
B． 为实现细粒度并行，输入分片(Input Split)应该越小越好
C． 一台机器可能被指派从输入文件的任意位置开始处理一个分片
D． 输入分片是一种记录的逻辑划分，而HDFS数据块是对输入数据的物理分割
答案：B
知识点：每个分片不能太小，否则启动与停止各个分片处理所需的开销将占很大一部分执行时间

针对每行数据内容为”Timestamp Url”的数据文件，在用JobConf对象conf设置

conf.setInputFormat(WhichInputFormat.class)来读取这个文件时，WhichInputFormat应该为以下的
A． TextInputFormat
B． KeyValueTextInputFormat
C． SequenceFileInputFormat
D． NLineInputFormat
答案：B
知识点：四项主要的InputFormat类。KeyValueTextInputFormat以每行第一个分隔符为界，分隔符前为key，之后为value，默认制表符为\t

有关MapReduce的输入输出，说法错误的是
A． 链接多个MapReduce作业时，序列文件是首选格式
B． FileInputFormat中实现的getSplits()可以把输入数据划分为分片，分片数目和大小任意定义
C． 想完全禁止输出，可以使用NullOutputFormat
D． 每个reduce需将它的输出写入自己的文件中，输出无需分片
答案：B
知识点：分片数目在numSplits中限定，分片大小必须大于mapred.min.size个字节，但小于文件系统的块

Hadoop Streaming支持脚本语言编写简单MapReduce程序，以下是一个例子：
bin/hadoop jar contrib/streaming/hadoop-0.20-streaming.jar
—input input/filename
—output output
—mapper ‘dosth.py 5’
—file dosth.py
—D mapred.reduce.tasks=1
以下说法不正确的是
A． Hadoop Streaming使用Unix中的流与程序交互
B． Hadoop Streaming允许我们使用任何可执行脚本语言处理数据流
C． 采用脚本语言时必须遵从UNIX的标准输入STDIN，并输出到STDOUT
D． Reduce没有设定，上述命令运行会出现问题
答案：D
知识点：没有设定特殊的reducer，默认使用IdentityReducer

在高阶数据处理中，往往无法把整个流程写在单个MapReduce作业中，下列关于链接MapReduce作业的说法，不正确的是
A．Job和JobControl类可以管理非线性作业之间的依赖
B．ChainMapper和ChainReducer类可以用来简化数据预处理和后处理的构成
C．使用ChainReducer时，每个mapper和reducer对象都有一个本地JobConf对象
D．ChainReducer.addMapper()方法中，一般对键/值对发送设置成值传递，性能好且安全性高
答案：D
知识点：ChainReducer.addMapper()方法中，值传递安全性高，引用传递性能高

//源码分析

//Zookeeper