# 腾讯大数据面试及参考答案[2017]

原创 2017-03-31 desehawk about云

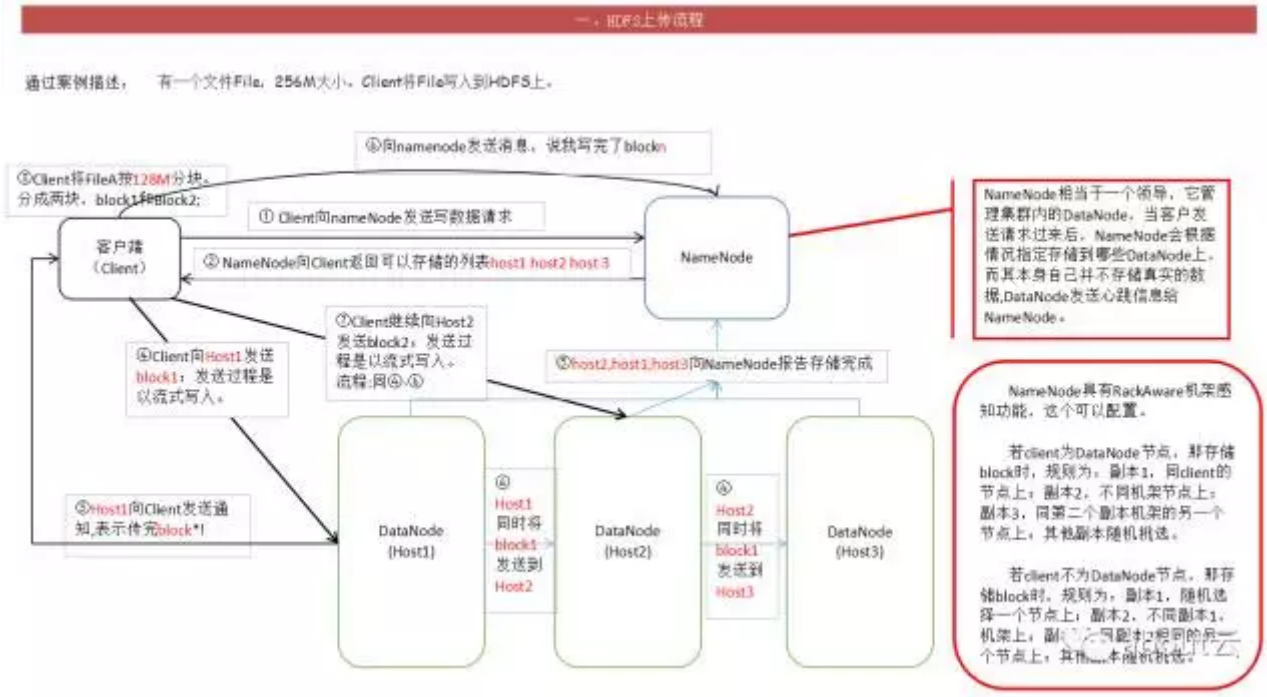
- 自我介绍
- 讲述HDFS上传文件和读文件的流程
- HDFS在上传文件的时候，如果其中一个块突然损坏了怎么办
- NameNode的作用
- NameNode在启动的时候会做哪些操作
- NameNode的HA
- NameNode和DataNode之间有哪些操作
- Innodb事务怎么实现的
- 项目介绍
- Hadoop的作业提交流程
- Hadoop怎么分片
- 如何减少Hadoop Map端到Reduce端的数据传输量
- Hadoop的Shuffle
- HMaster的作用
- HBase的操作数据的步骤
- Innodb的二进制文件和Redo日志的区别
- Redo日志的格式(不知道这个)
- 二进制日志的复制(不知道这个)

题目来自：csdn leishenop

##########
下面关于大数据的答案，个人见解，欢迎交流

**讲述HDFS上传文件和读文件的流程**

**HDFS 上传流程**




过程解析：详解
这里描述的 是一个256M的文件上传过程
① 由客户端 向 NameNode节点节点 发出请求
②NameNode 向Client返回可以可以存数据的 DataNode 这里遵循 机架感应 原则

③客户端 首先 根据返回的信息 先将 文件分块（Hadoop2.X版本 每一个block为 128M 而之前的版本为 64M）
④然后通过那么Node返回的DataNode信息 直接发送给DataNode 并且是 流式写入 同时 会复制到其他两台机器
⑤dataNode 向 Client通信 表示已经传完 数据块 同时向NameNode报告
⑥依照上面（④到⑤）的原理将 所有的数据块都上传结束 向 NameNode 报告 表明 已经传完所有的数据块

这样 整个HDFS上传流程就 走完了 （来自csdn Only、爱你）

相关文章：
HDFS文件读写及准确性介绍
http://www.aboutyun.com/forum.php?mod=viewthread&tid=6966

Hadoop学习总结：HDFS读写过程解析
http://www.aboutyun.com/forum.php?mod=viewthread&tid=14846

HDFS追本溯源：租约，读写过程的容错处理及NN的主要数据结构
http://www.aboutyun.com/forum.php?mod=viewthread&tid=17620

**HDFS在上传文件的时候，如果其中一个块突然损坏了怎么办**
其中一个块坏了，只要有其它块存在，会自动检测还原。

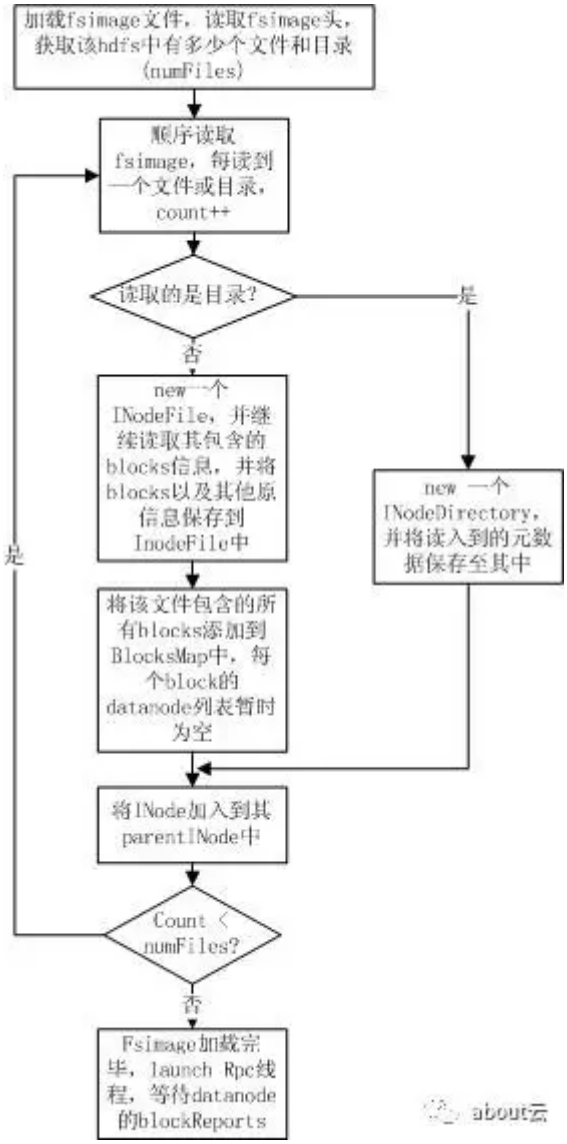
**NameNode的作用**
namenode总体来说是管理和记录恢复功能。
比如管理datanode，保持心跳，如果超时则排除。
对于上传文件都有镜像images和edits,这些可以用来恢复。更多：
深度了解namenode---其 内部关键数据结构原理简介
http://www.aboutyun.com/forum.php?mod=viewthread&tid=7388

**NameNode在启动的时候会做哪些操作**
NameNode启动的时候，会加载fsimage

更多参考下面内容
**NameNode启动过程fsimage加载过程**

Fsimage加载过程完成的操作主要是为了：
1. 从fsimage中读取该HDFS中保存的每一个目录和每一个文件
2. 初始化每个目录和文件的元数据信息
3. 根据目录和文件的路径，构造出整个namespace在内存中的镜像
4. 如果是文件，则读取出该文件包含的所有blockid，并插入到BlocksMap中。

整个加载流程如下图所示：




如上图所示，namenode在加载fsimage过程其实非常简单，就是从fsimage中不停的顺序读取文件和目录的元数据信息，并在内存中构建整个namespace，同时将每个文件对应的blockid保存入BlocksMap中，此时BlocksMap中每个block对应的datanodes列表暂时为空。当fsimage加载完毕后，整个HDFS的目录结构在内存中就已经初始化完毕，所缺的就是每个文件对应的block对应的datanode列表信息。这些信息需要从datanode的blockReport中获取，所以加载fsimage完毕后，namenode进程进入rpc等待状态，等待所有的datanodes发送blockReports。

**NameNode的HA**

NameNode的HA一个备用，一个工作，且一个失败后，另一个被激活。他们通过journal node来实现共享数据。

更多

Hadoop之NameNode+ResourceManager高可用原理分析

http://www.aboutyun.com/forum.php?mod=viewthread&tid=16024

Hadoop常见 HA方案 及如何解决HA

http://www.aboutyun.com/forum.php?mod=viewthread&tid=6724

**NameNode和DataNode之间有哪些操作**

这个问题有些歧义。操作具体可以查看hadoop命令，应该超不出命令汇总

Hadoop Shell命令字典（可收藏）

http://www.aboutyun.com/forum.php?mod=viewthread&tid=6983

hadoop高级命令详解

http://www.aboutyun.com/forum.php?mod=viewthread&tid=14829

**Hadoop的作业提交流程**

Hadoop2.x Yarn作业提交（客户端）

http://www.aboutyun.com/forum.php?mod=viewthread&tid=9498

Hadoop2.x Yarn作业提交（服务端）

http://www.aboutyun.com/forum.php?mod=viewthread&tid=9496

更多：

hadoop作业提交脚本分析（1）

http://www.aboutyun.com/forum.php?mod=viewthread&tid=6954

hadoop作业提交脚本分析（2）

http://www.aboutyun.com/forum.php?mod=viewthread&tid=6956

**Hadoop怎么分片**

如何让hadoop按文件分片

http://www.aboutyun.com/forum.php?mod=viewthread&tid=14549

Hadoop分块与分片

http://www.aboutyun.com/blog-5994-697.html

**如何减少Hadoop Map端到Reduce端的数据传输量**

减少传输量，可以让map处理完，让同台的reduce直接处理，理想情况下，没有数据传输。

**Hadoop的Shuffle**

彻底了解mapreduce核心Shuffle--解惑各种mapreduce问题

http://www.aboutyun.com/forum.php?mod=viewthread&tid=7078

hadoop代码笔记 Mapreduce shuffle过程之Map输出过程((1)
http://www.aboutyun.com/forum.php?mod=viewthread&tid=10335

**HMaster的作用**

hmaster的作用
为region server分配region.
负责region server的负载均衡。
发现失效的region server并重新分配其上的region.
Gfs上的垃圾文件回收。
处理schema更新请求。
更多
region server and hmaster server

**HBase的操作数据的步骤**

Hbase写数据，存数据，读数据的详细过程
http://www.aboutyun.com/forum.php?mod=viewthread&tid=10886

转载注明来自：about云（www.aboutyun.com）