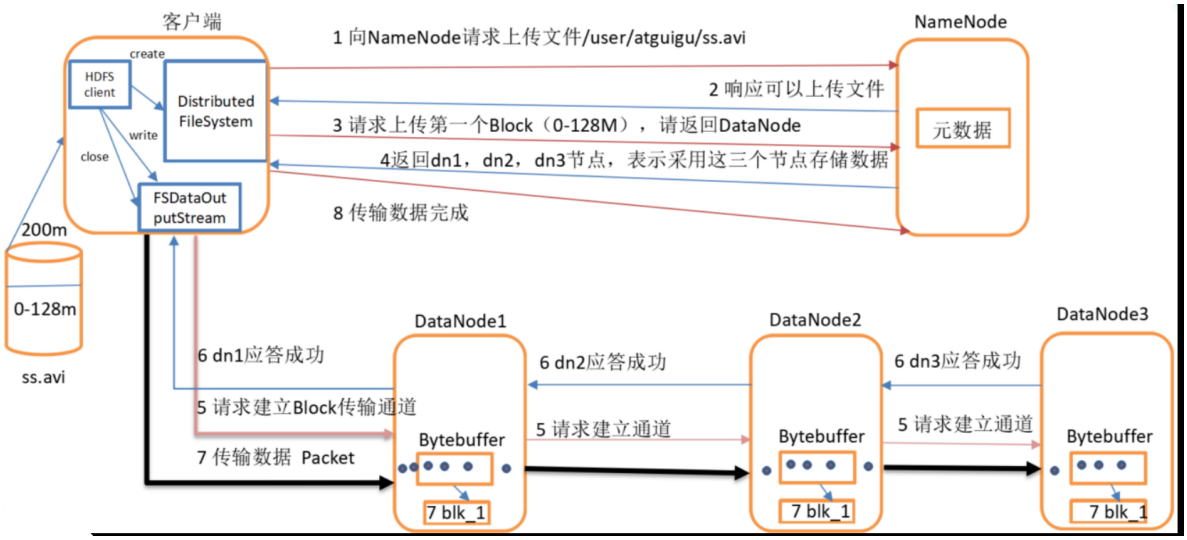
> 不断关联，不断加入，不断迭代，不断应用

HDFS数据读写流程

HDFS写数据流程



1. 客户端通过Distributed FileSystem模块向NameNode请求上传文件，NameNode检查目标文件是否已存在，父目录是否存在。
2. NameNode返回是否可以上传。
3. 客户端请求第一个 Block上传到哪几个DataNode服务器上。
4. NameNode返回3个DataNode节点，分别为dn1、dn2、dn3。
5. 客户端通过FSDataOutputStream模块请求dn1上传数据，dn1收到请求会继续调用dn2，然后dn2调用dn3，将这个通信管道建立完成。
6. dn1、dn2、dn3逐级应答客户端。
7. 客户端开始往dn1上传第一个Block（先从磁盘读取数据放到一个本地内存缓存），以Packet为单位，dn1收到一个Packet就会传给dn2，dn2传给dn3；dn1每传一个packet会放入一个应答队列等待应答。
8. 当一个Block传输完成之后，客户端再次请求NameNode上传第二个Block的服务器。（重复执行3-7步）。

NameNode：可以理解为DataNode管理器

DataNode：存储块数据，默认128M为一块

网络拓扑-节点距离计算

在HDFS写数据的过程中，NameNode会选择距离待上传数据最近距离的DataNode接收数据。那么这个最近距离怎么计算呢?

> 节点距离：两个节点到达最近的共同祖先的距离总和。

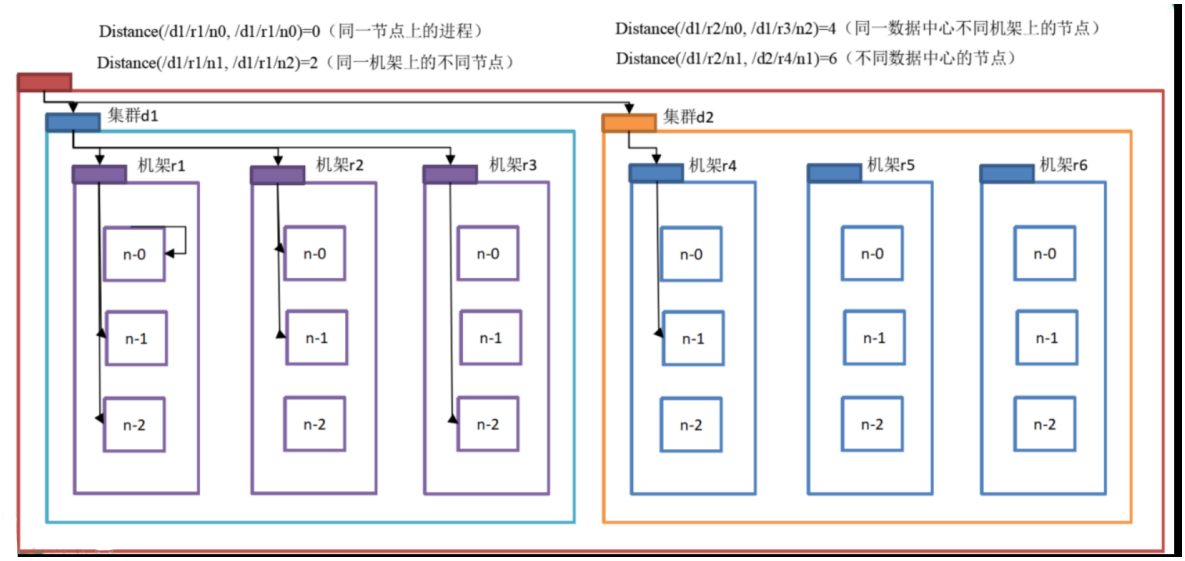
不好理解的就是共同祖先啥意思，可以理解为上级节点，节点等级如下

【数据中心（集群d）---> 机架r ---> 具体节点n】

举个例子：计算节点d1/r1/n1到节点d1/r1/n2的节点距离

确认共同祖先为r1，节点n1到它的距离为1，节点n2到它的距离也为1，两者和为2

所以它们之间的节点距离就是2



机架感知

机架感知说明

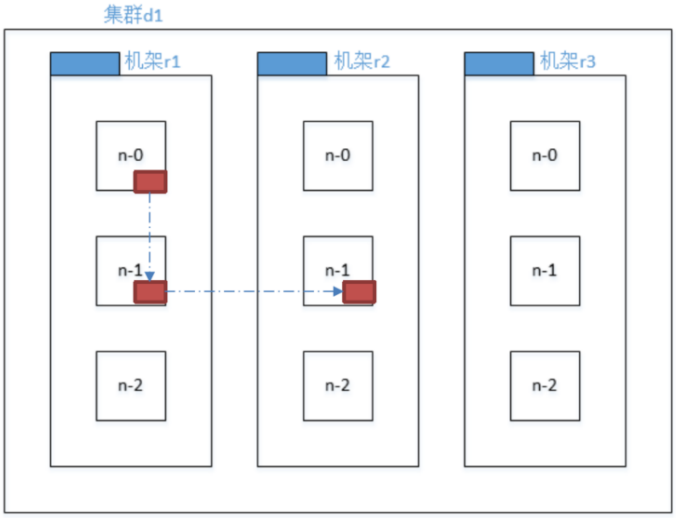
> http://hadoop.apache.org/docs/r2.7.2/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html#Data_Replication

其实机架感知听起来高大上，但可以理解为副本节点的位置的选择就行

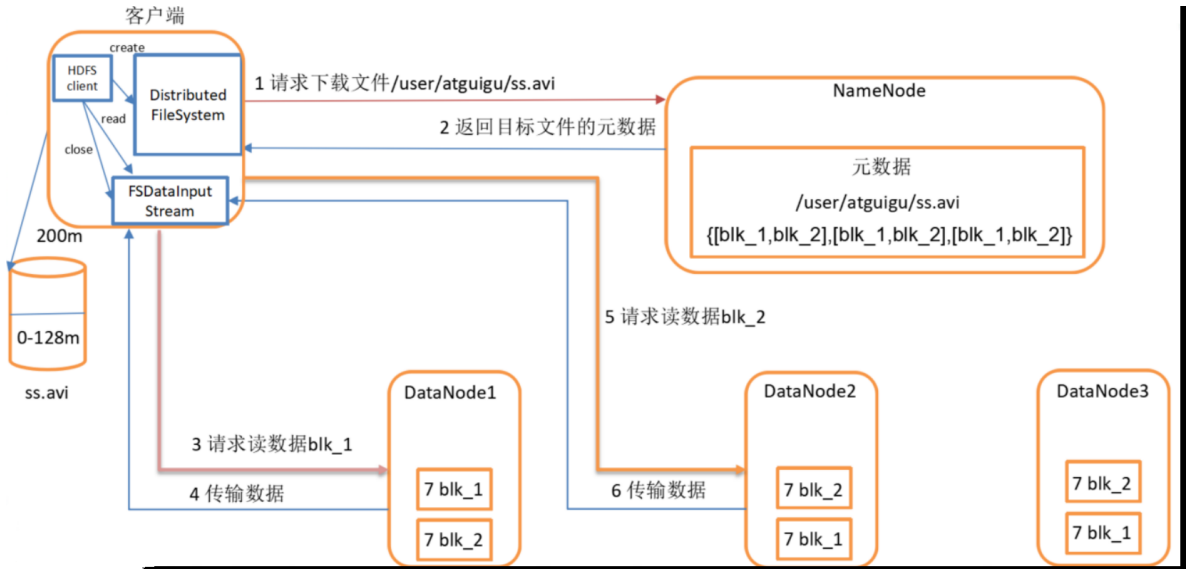
第一个副本在Client所处的节点上。
如果客户端在集群外，随机选一个。

第二个副本和第一个副本位于相同机架，随机节点。

第三个副本位于不同机架，随机节点。



HDFS读数据流程



1. 客户端通过Distributed FileSystem向NameNode请求下载文件，NameNode通过查询元数据，找到文件块所在的DataNode地址。
2. 挑选一台DataNode（就近原则，然后随机）服务器，请求读取数据。
3. DataNode开始传输数据给客户端（从磁盘里面读取数据输入流，以Packet为单位来做校验）。
4. 客户端以Packet为单位接收，先在本地缓存，然后写入目标文件。